Actividad 6 distribuciones muestrales y TLC

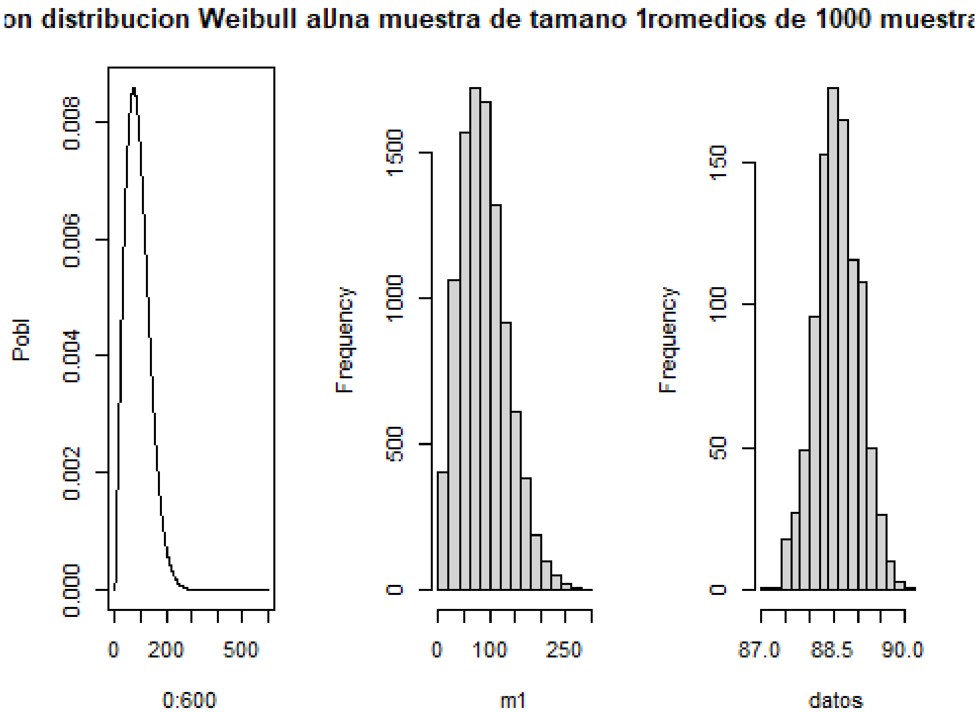
Rodolfo Jesús Cruz Rebollar

2024-08-18

Problema 1: Ensayando distribuciones

Inciso a

```r
# Generar una matriz de gráficos de 1 x 3 (1 fila y 3 columnas)

par(mfrow=c(1,3))

# Graficando una distribucion Weibull de alfa =2, beta = 100

Pobl = dweibull(0:600,2, 100)

plot(0:600, Pobl, type="l", main = "Poblacion con distribucion Weibull
alfa=2, beta = 100")

# Tomando una muestra de 10000 elementos tomados al azar

m1 = rweibull(10000, 2, 100)

hist(m1, main = "Una muestra de tamano 10000")

# Tomando 1000 promedios de las 1000 muestras como la anterior

m = rweibull(10000,2,100)

prom = mean(m)

datos = prom

for(i in 1:999) {

  m = rweibull(10000,2,100)
  prom = mean(m)
  datos = rbind(datos,prom)

}

hist(datos, main="Grafica de los promedios de 1000 muestras de tamano
10,000")
```



En cuanto a la primera gráfica resultante, la interpretación de la misma radica principalmente en el hecho de que los valores presentes en el eje horizontal de la misma hacen referencia a la cantidad de tiempo de antiguedad de un cierto componente, por ejemplo de un automóvil que en éste caso es de 0 a 600, mientras que al mismo tiempo, los valores en el eje vertical representan la proporción de fiabilidad de dicho componente del automóvil, misma que hace referencia al hecho de que el componente mencionado presente un fallo después de un determinado tiempo de uso, por lo que teniendo lo anterior en cuenta, para este caso en particular, la gráfica muestra que en las primeras 100 unidades de tiempo transcurridas de uso del componente del automóvil, éste mismo componente alcanzó su máxima proporción de fiabilidad, misma que fue superior al 0.8%, mientras que después de alcanzr su máxima fiabilidad, ésta misma fue decayendo gradualmente, esto debido a que entre más tiempo de uso tenga el componente del auto, éste tendrá una mayor probabilidad de fallar y como consecuencia, también se reducirá el nivel de eficiencia y confiabilidad del mismo hasta tal punto que después de un momento dado, ya no será suficientemente efectivo o confiable para seguir utilizándose.

Además de lo anterior, en relación a la segunda gráfica (histograma), se puede observar que los datos representados en las diferentes clases del histograma tienen en general un sesgo positivo, dado que la algomeración de datos se produce en la región izquierda del histograma, mientras que en la región derecha del gráfico, la cantidad presente de datos es mínima, casi nula, por lo cual la distribución de los datos es en su mayoría asimétrica, además se evidencia que existe simlitud entre el histograma y la primer gráfica, motivo por el cual, es posible establecer una cierta relación entre ambas gráficos, por lo que tiene sentido que los valores que indican un nivel adecuado de fiabilidad, sucedan de forma más frecuente (valores en el eje vertical del histograma) sobretodo en cuando el tiempo de uso de un componente (podríamos considerarlo como el eje horizontal del histograma) no es tan grande ya que entre menos tiempo de uso tenga, mayor será su efectividad (barras del histograma), por lo que en resumen, el histograma también puede ser útil para representar gráficamente la relación existente entre el tiempo de uso transcurrido del componente del automóvil y su porcentaje de fiabilidad para tener un rendimiento efectivo al ser utilizado.

Por último, en cuanto a la tercera gráfica (segundo histograma), se aprecia que los promedios de las muestras aleatorias se distribuyen de forma aproximadamente normal, principalmente debido a que las muestras al tener un tamaño que supera por mucho los 30 elementos, pasan a tener un comportamiento normal en cuanto a su distribución que se atribuye al terorema central del límite (TCL), mismo que establece que sin importar el tipo de distribución de los datos, la distribución de los mismos se aproximará a la normal si el tamaño de la muestra es grande, lo cual sucede en el caso particular de los datos involucrados en el segundo histograma, por lo que se evidencia un comportamiento creciente en la frecuencia de ocurrencia de aquellos datos que tienen valores entre 87 y 89, mientras que después de éste último valor, dicha frecuencia disminuye gradualmente, por lo que nuevamente, el comportamiento general de los datos observados en este último histograma comparte cierta similitud con el comportamiento de los datos del primer gráfico analizado, ya que en el tercer gráfico, entre mayores sean los datos, la frecuencia de ocurrencia de los mismos disminuye de forma gradual, lo cual también ocurre en el primer gráfico.

Inciso b

```
# Librería para calcular sesgo y curtosis

library(moments)

# Calcular sesgo de la muestra de tamaño 10000

sesgo_muestra = skewness(m1)

# Calcular curtosis de la muestra de tamaño 10000

curtosis_muestra = kurtosis(m1)

cat("Sesgo: ", sesgo_muestra, "\n")

## Sesgo:  0.6345273

cat("Curtosis: ", curtosis_muestra)

## Curtosis:  3.258676
```

Prueba de hipótesis de normalidad

H_0: los datos de la muestra siguen una distribución normal.

H_1: los datos de la muestra no siguen una distribución normal.

Dado que el tamaño de la muestra de datos es bastante mayor a 50, se utilizará el test de normalidad de Anderson Darling, mismo que se recomienda para muestras de tamaño grande (mayor a 30).

```
# Importar librería para tests de normalidad

library(nortest)

# Realizar test de normalidad de Anderson Darling a los datos de la
muestra grande

ad.test(m1)

##
##  Anderson-Darling normality test
##
## data:  m1
## A = 55.299, p-value < 2.2e-16
```

En base al test de normalidad de Anderson Darling, se aprecia que el p valor es inferior a 2.2e-16, por lo cual dado que el valor p es mucho menor que 0.05, se rechaza H_0, por lo cual, se conclye que los datos de la muestra no siguen una distribución normal, por lo que es probable que el tamaño de muestra aún no sea lo suficientemente grande como para cumplir con el Teorema Central del Límite (TCL).

Inciso c

```
# Calcular el sesgo de las medias de las 1000 muestras

sesgo_1000 = skewness(datos)

# Calcular curtosis de medias de las 1000 muestras

curtosis_1000 = kurtosis(datos)

cat("Sesgo: ", sesgo_1000, "\n")

## Sesgo:  0.04079246

cat("Curtosis: ", curtosis_1000)

## Curtosis:  2.902571
```

H_0: las medias de las muestras siguen una distribución normal.

H_1: las medias de las muestras no siguen una distribución normal.

```
# Aplicar test de normalidad de Anderson Darling a las medias de 1000
muestras
```

```
ad.test(datos)

##
##  Anderson-Darling normality test
##
## data:  datos
## A = 0.31896, p-value = 0.5344
```

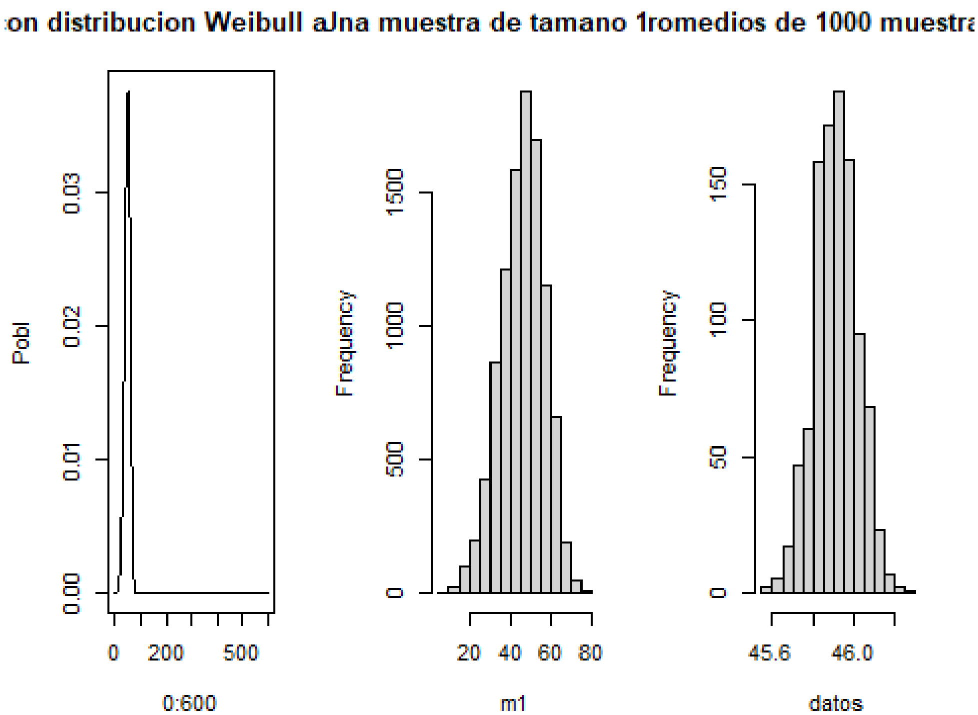
En base al test de Anderson Darling previo, se puede apreciar que el valor p de la prueba es de 0.3553, lo cual resulta ser mayor que 0.05, motivo por el cual no se rechaza H_0 y se concluye que las medias de las 1000 muestras siguen una distribución normal.

Inciso d

Procedimiento para distribución con alfa = 5, beta = 50

```
# Generar una matriz de gráficos de 1 x 3 (1 fila y 3 columnas)

par(mfrow=c(1,3))

# Graficando una distribucion Weibull de alfa =5, beta = 50

Pobl = dweibull(0:600,5, 50)

plot(0:600, Pobl, type="l", main = "Poblacion con distribucion Weibull
alfa=5, beta = 50")

# Tomando una muestra de 10000 elementos tomados al azar

m1 = rweibull(10000, 5, 50)

hist(m1, main = "Una muestra de tamano 10000")

# Tomando 1000 promedios de las 1000 muestras como la anterior

m = rweibull(10000, 5, 50)

prom = mean(m)

datos = prom

for(i in 1:999) {

  m = rweibull(10000, 5, 50)
  prom = mean(m)
  datos = rbind(datos,prom)

}
```

```
hist(datos, main="Grafica de los promedios de 1000 muestras de tamano
10,000")
```



En cuanto a la primera gráfica resultante, la interpretación de la misma radica principalmente en el hecho de que los valores presentes en el eje horizontal de la misma hacen referencia a la cantidad de tiempo de antiguedad de un cierto componente, por ejemplo de un automóvil que en éste caso es de 0 a 600, mientras que al mismo tiempo, los valores en el eje vertical representan la proporción de fiabilidad de dicho componente del automóvil, misma que hace referencia al hecho de que el componente mencionado presente un fallo después de un determinado tiempo de uso, por lo que teniendo lo anterior en cuenta, para este caso en particular, la gráfica muestra que en las primeras 50 unidades de tiempo transcurridas de uso del componente del automóvil, éste mismo componente alcanzó su máxima proporción de fiabilidad, misma que fue superior al 3%, mientras que después de alcanzr su máxima fiabilidad, ésta misma fue decayendo gradualmente, esto debido a que entre más tiempo de uso tenga el componente del auto, éste tendrá una mayor probabilidad de fallar y como consecuencia, también se reducirá el nivel de eficiencia y confiabilidad del mismo hasta tal punto que después de un momento dado, ya no será suficientemente efectivo o confiable para seguir utilizándose.

Además de lo anterior, en relación a la segunda gráfica (histograma), se puede observar que los datos representados en las diferentes clases del histograma tienen en general un sesgo ligeramente negativo, dado que la aglomeración de datos se produce ligeramente más en la región derecha del histograma, mientras que en la región izquierda del gráfico, la cantidad presente de datos es mínima, casi nula, por lo cual la distribución de los datos es en su mayoría asimétrica, además se evidencia que el comportamiento del histograma esopuesto al del primer gráfico, motivo por el cual, es posible establecer una cierta relación inversa entre ambas gráficos, por lo que ya no tiene sentido que los valores que indican un nivel adecuado de fiabilidad, sucedan de forma más frecuente (valores en el eje vertical del histograma) sobretodo en cuando el tiempo de uso de un componente (podríamos considerarlo como el eje horizontal del histograma) no es tan grande ya que entre menos tiempo de uso tenga, mayor será su efectividad (barras del histograma), por lo que en resumen, el histograma ya no puede ser útil para representar gráficamente la relación existente entre el tiempo de uso transcurrido del componente del automóvil y su porcentaje de fiabilidad para tener un rendimiento efectivo al ser utilizado.

Por último, en cuanto a la tercera gráfica (segundo histograma), se aprecia que los promedios de las muestras aleatorias se distribuyen de forma aproximadamente normal, principalmente debido a que las muestras al tener un tamaño que supera por mucho los 30 elementos, pasan a tener un comportamiento normal en cuanto a su distribución que se atribuye al terorema central del límite (TCL), mismo que establece que sin importar el tipo de distribución de los datos, la distribución de los mismos se aproximará a la normal si el tamaño de la muestra es grande, lo cual sucede en el caso particular de los datos involucrados en el segundo histograma, por lo que se evidencia un comportamiento creciente en la frecuencia de ocurrencia de aquellos datos que tienen valores entre 45.6 y 45.9, mientras que después de éste último valor, dicha frecuencia disminuye gradualmente, por lo que nuevamente, el comportamiento general de los datos observados en este último histograma no comparte similitud con el comportamiento de los datos del primer gráfico analizado, ya que en el tercer gráfico, entre mayores sean los datos, la frecuencia de ocurrencia de los mismos disminuye de forma gradual, lo cual también ocurre en el primer gráfico.

```r
# Calcular sesgo de la muestra de tamaño 10000

sesgo_muestra2 = skewness(m1)

# Calcular curtosis de la muestra de tamaño 10000

curtosis_muestra2 = kurtosis(m1)

cat("Sesgo: ", sesgo_muestra2, "\n")

## Sesgo:  -0.2639676

cat("Curtosis: ", curtosis_muestra2)

## Curtosis:  2.848107

# Realizar test de normalidad de Anderson Darling a los datos de la
muestra grande

ad.test(m1)
```

```
##
##  Anderson-Darling normality test
##
## data:  m1
## A = 13.31, p-value < 2.2e-16
```

En base al test de normalidad de Anderson Darling, se aprecia que el p valor es inferior a 2.2e-16, por lo cual dado que el valor p es mucho menor que 0.05, se rechaza H_0, por lo cual, se conclye que los datos de la muestra no siguen una distribución normal, por lo que es probable que el tamaño de muestra aún no sea lo suficientemente grande como para cumplir con el Teorema Central del Límite (TCL).

```
# Calcular el sesgo de las medias de las 1000 muestras

sesgo_1000 = skewness(datos)

# Calcular curtosis de medias de las 1000 muestras

curtosis_1000 = kurtosis(datos)

cat("Sesgo: ", sesgo_1000, "\n")

## Sesgo:  -0.02432775

cat("Curtosis: ", curtosis_1000)

## Curtosis:  2.968749
```

H_0: las medias de las muestras siguen una distribución normal.

H_1: las medias de las muestras no siguen una distribución normal.

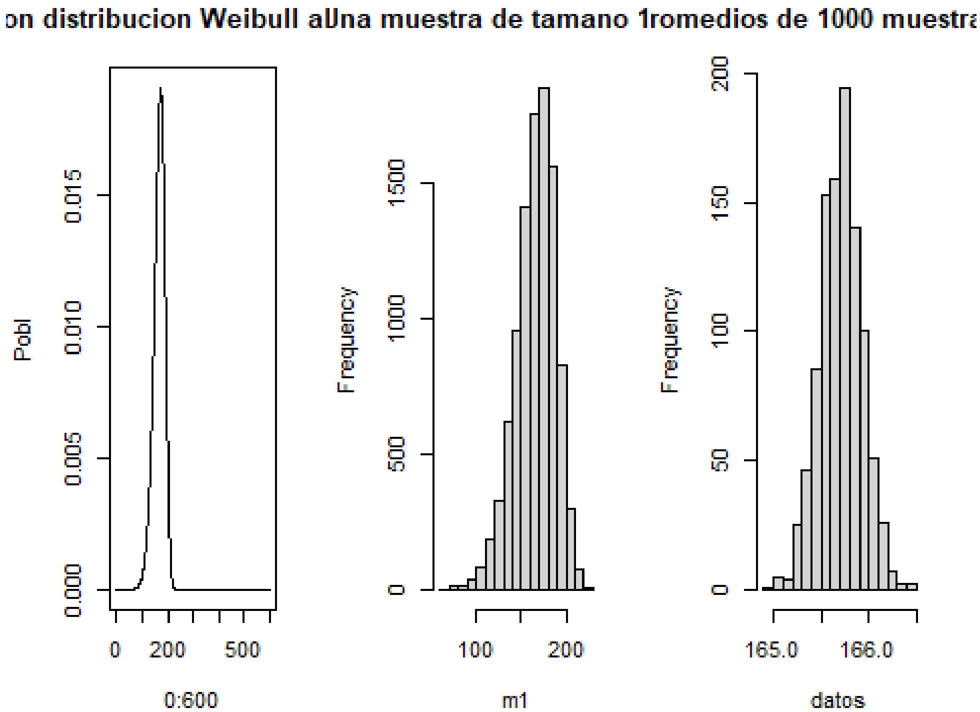
```
# Aplicar test de normalidad de Anderson Darling a las medias de 1000
muestras

ad.test(datos)

##
##  Anderson-Darling normality test
##
## data:  datos
## A = 0.34179, p-value = 0.4929
```

En base al test de Anderson Darling previo, se puede apreciar que el valor p de la prueba es de 0.4582, lo cual resulta ser mayor que 0.05, motivo por el cual no se rechaza H_0 y se concluye que las medias de las 1000 muestras siguen una distribución normal.

Procedimiento para distribución con alfa = 9, beta = 175

```
# Generar una matriz de gráficos de 1 x 3 (1 fila y 3 columnas)
```

```r
par(mfrow=c(1,3))

# Graficando una distribucion Weibull de alfa =9, beta = 175

Pobl = dweibull(0:600,9, 175)

plot(0:600, Pobl, type="l", main = "Poblacion con distribucion Weibull
alfa=9, beta = 175")

# Tomando una muestra de 10000 elementos tomados al azar

m1 = rweibull(10000, 9, 175)

hist(m1, main = "Una muestra de tamano 10000")

# Tomando 1000 promedios de las 1000 muestras como la anterior

m = rweibull(10000, 9, 175)

prom = mean(m)

datos = prom

for(i in 1:999) {

  m = rweibull(10000, 9, 175)
  prom = mean(m)
  datos = rbind(datos,prom)

}

hist(datos, main="Grafica de los promedios de 1000 muestras de tamano
10,000")
```



En cuanto a la primera gráfica resultante, la interpretación de la misma radica principalmente en el hecho de que los valores presentes en el eje horizontal de la misma hacen referencia a la cantidad de tiempo de antiguedad de un cierto componente, por ejemplo de un automóvil que en éste caso es de 0 a 600, mientras que al mismo tiempo, los valores en el eje vertical representan la proporción de fiabilidad de dicho componente del automóvil, misma que hace referencia al hecho de que el componente mencionado presente un fallo después de un determinado tiempo de uso, por lo que teniendo lo anterior en cuenta, para este caso en particular, la gráfica muestra que en las primeras 200 unidades de tiempo transcurridas de uso del componente del automóvil, éste mismo componente alcanzó su máxima proporción de fiabilidad, misma que fue superior al 1.5%, mientras que después de alcanzr su máxima fiabilidad, ésta misma fue decayendo gradualmente, esto debido a que entre más tiempo de uso tenga el componente del auto, éste tendrá una mayor probabilidad de fallar y como consecuencia, también se reducirá el nivel de eficiencia y confiabilidad del mismo hasta tal punto que después de un momento dado, ya no será suficientemente efectivo o confiable para seguir utilizándose.

Además de lo anterior, en relación a la segunda gráfica (histograma), se puede observar que los datos representados en las diferentes clases del histograma tienen en general un sesgo ligeramente negativo, dado que la aglomeración de datos se produce ligeramente más en la región derecha del histograma, mientras que en la región izquierda del gráfico, la cantidad presente de datos es mínima, casi nula, por lo cual la distribución de los datos es en su mayoría asimétrica, además se evidencia que el comportamiento del histograma es opuesto al del primer gráfico, motivo por el cual, es posible establecer una cierta relación inversa entre ambas gráficos, por lo que ya no tiene sentido que los valores que indican un nivel adecuado de fiabilidad, sucedan de forma más frecuente (valores en el eje vertical del histograma) sobretodo en cuando el tiempo de uso de un componente (podríamos considerarlo como el eje horizontal del histograma) no es tan grande ya que entre menos tiempo de uso tenga, mayor será su efectividad (barras del histograma), por lo que en resumen, el histograma ya no puede ser útil para representar gráficamente la relación existente entre el tiempo de uso transcurrido del componente del automóvil y su porcentaje de fiabilidad para tener un rendimiento efectivo al ser utilizado, esto debido a que la naturaleza de los datos provoca que tenga un comportamiento contrario al del primero.

Por último, en cuanto a la tercera gráfica (segundo histograma), se aprecia que los promedios de las muestras aleatorias se distribuyen de forma aproximadamente normal, principalmente debido a que las muestras al tener un tamaño que supera por mucho los 30 elementos, pasan a tener un comportamiento normal en cuanto a su distribución que se atribuye al terorema central del límite (TCL), mismo que establece que sin importar el tipo de distribución de los datos, la distribución de los mismos se aproximará a la normal si el tamaño de la muestra es grande, lo cual sucede en el caso particular de los datos involucrados en el segundo histograma, por lo que se evidencia un comportamiento creciente en la frecuencia de ocurrencia de aquellos datos que tienen valores entre 165 y 165.75, mientras que después de éste último valor, dicha frecuencia disminuye gradualmente, por lo que nuevamente, el comportamiento general de los datos observados en este último histograma no comparte similitud con el comportamiento de los datos del primer gráfico analizado, ya que en el tercer gráfico, entre mayores sean los datos, la frecuencia de ocurrencia de los mismos disminuye de forma gradual, lo cual también ocurre en el primer gráfico.

```
# Calcular sesgo de la muestra de tamaño 10000

sesgo_muestra3 = skewness(m1)

# Calcular curtosis de la muestra de tamaño 10000

curtosis_muestra3 = kurtosis(m1)

cat("Sesgo: ", sesgo_muestra3, "\n")

## Sesgo:  -0.599679

cat("Curtosis: ", curtosis_muestra3)

## Curtosis:  3.577138

# Realizar test de normalidad de Anderson Darling a los datos de la
muestra grande

ad.test(m1)

##
##  Anderson-Darling normality test
```

```
##
## data:  m1
## A = 40.485, p-value < 2.2e-16
```

En base al test de normalidad de Anderson Darling, se aprecia que el p valor es inferior a 2.2e-16, por lo cual dado que el valor p es mucho menor que 0.05, se rechaza H_0, por lo cual, se conclye que los datos de la muestra no siguen una distribución normal, por lo que es probable que el tamaño de muestra aún no sea lo suficientemente grande como para cumplir con el Teorema Central del Límite (TCL).

```r
# Calcular el sesgo de las medias de las 1000 muestras

sesgo_1000 = skewness(datos)

# Calcular curtosis de medias de las 1000 muestras

curtosis_1000 = kurtosis(datos)

cat("Sesgo: ", sesgo_1000, "\n")

## Sesgo:  -0.02229046

cat("Curtosis: ", curtosis_1000)

## Curtosis:  3.124432
```

H_0: las medias de las muestras siguen una distribución normal.

H_1: las medias de las muestras no siguen una distribución normal.

```r
# Aplicar test de normalidad de Anderson Darling a las medias de 1000
muestras

ad.test(datos)

##
##  Anderson-Darling normality test
##
## data:  datos
## A = 0.24427, p-value = 0.7626
```

En base al test de Anderson Darling previo, se puede apreciar que el valor p de la prueba es de 0.1781, lo cual resulta ser mayor que 0.05, motivo por el cual no se rechaza H_0 y se concluye que las medias de las 1000 muestras siguen una distribución normal.

En conclusión, después de realizar las pruebas enunciadas en los incisos A, B y C, se observa que a diferencia de la primera distribución con la que se realizaron dichas pruebas, las otras 2 distribuciones puestas a prueba arrojaron como resultado que el patrón que seguían en torno a los datos implicados en ellas era opuesto al que arroja la primer distribución puesta a prueba (la que guarda similitud de comportamiento con el gráfico que simboliza la curva de la distribución Weibull), por lo que a excepción de la primera distribución probada, los resultados derivados de las otras 2 no siguen el mismo patrón que los resultados derivados de la primera, sin embargo los resultados de las 3 distribuciones probadas exhibieron un comportamiento en primera instancia creciente y después de alcanzar el máximo valor, dicho comportamiento se volvió decreciente.

Inciso e

Finalmente, después de llevar a cabo los gráficos en las 3 distribuciones teóricas, es posible concluir que se observaron ciertas semejanzas entre ellos, entre las que destaca el hecho de que al tener una muestra de datos o medias considerablemente grande (10000 datos o 1000 medias), se cumple parcialmente (debido a que las muestras aún no son lo suficientemente grandes) el Teorema Central del Límite (TCL), motivo por el cual, aunque la distribución de los datos en cuestión aún siga conservando cierta asimetría en algunas de sus regiones, conforme se aumente el tamaño de las muestras, dicha distribución se parecerá cada vez más a una normal. No obstante, también se observaron ciertas diferencias entre los gráficos de las 3 distribuciones teóricas, entre las que se encuentra el hecho de que en los gráficos de la primera distribución teórica, se aprecia que los datos que ocurren con mayor frecuencia son aquellos que tienen valores en su mayoría relativamente bajos, mientras que por el contrario, en los gráficos derivados de las otras 2 distribuciones teóricas, sucede el comportamiento contrario, por lo que es probable que lo descrito anteriormente se deba en parte a la propia naturaleza de los datos, ya que al ser éstos generados de forma aleatoria, se pueden obtener una gran variedad de valores numéricos cuyo comportamiento puede seguir igualmente una gran variedad de distribuciones que en la gran mayoría de las ocasiones son asimétricas por la naturaleza misma de los datos y no porque hayan sido alterados de alguna manera antes de procesarlos y analizarlos.

Problema 2: remaches

La resistencia a la ruptura de un remache tiene un valor medio de 10,000 lb/pulg 2 y una desviación estándar de 500 lb/pulg2. Si se sabe que la población se distribuye normalmente:

X: resistencia a la ruptura de un remache

$$X \sim N(\mu_x = 10000, \sigma_x = 500)$$

Inciso A

a) ¿Cuál es la probabilidad de que al tomar un remache al azar de esa población, éste tenga una resistencia a la ruptura que esté a 100 unidades alrededor de la media? ¿a cuántas desviaciones estándar está de la media?

$P(9900 < X < 10100) = ?$

```r
# Calcular la probabilidad de que la resistencia de un remache esté 100 unidades
# alrededor de la media (ya sea 100 por debajo o por encima de 10000)

p_9900_10000 = pnorm(10100, 10000, 500) - pnorm(9900, 10000, 500)

# Mostrar la probabilidad de que la resistencia esté 100 unidades alrededor de la media

cat("P(9900 < X < 10100) = ", p_9900_10000)
```

```
## P(9900 < X < 10100) =  0.1585194
```

Desviaciones lejos de la media (z)

```r
# Calcular la cantidad de desviaciones estándar lejos de la media a la que se ubica la
# resistencia de un remache que esté 100 unidades alrededor de dicha media

z_A = 100 / 500

# Mostrar a cuantas desviaciones estándar está la resistencia de un remache que está a 100
# unidades alrededor de la media

cat("z = ", z_A)
```

```
## z =  0.2
```

Inciso B

b) ¿Cuál es la probabilidad de que la resistencia media a la ruptura de la muestra aleatoria de 120 remaches esté 100 unidades alrededor de su media? ¿a cuántas desviaciones estándar está de la media?

$$P(9900 < \bar{X} < 10100) = ?$$

$$\bar{X} \sim N\left(\mu_{\bar{x}} = 10000, \sigma_{\bar{x}} = \frac{500}{\sqrt{120}}\right)$$

```r
# Calcular la probabilidad de que la resistencia promedio de una muestra de tamaño 120
# esté 100 unidades alrededor de la resistencia media de la población de remaches

p_120 = pnorm(10100, 10000, 500/sqrt(120)) - pnorm(9900, 10000,
500/sqrt(120))

# Mostrar probabilidad anterior para una muestra de tamaño 120
```

```r
cat("P(9900 < X_b < 10100) = ", p_120)

## P(9900 < X_b < 10100) =  0.9715403
```

Desviaciones lejos de la media (z)

```r
# Calcular el número de desviaciones estándar a la que la resistencia
media de la muestra
# está lejos de la resistencia media de la población de remaches

z_120 = 100 / (500/sqrt(120))

# Mostrar cantidad de desviaciones estándar a la que la resistencia
promedio muestral está
# lejos de la resistencia media poblacional

cat("z = ", z_120)

## z =  2.19089
```

Inciso C

c) Si el tamaño muestral hubiera sido 15, en lugar de 120, ¿cuál es la probabilidad de que la resistencia media a la ruptura esté 100 unidades alrededor de la media? ¿a cuántas desviaciones estándar está de la media?

```r
# Calcular la probabilidad de que la resistencia media de la muestra esté
a 100 unidades
# alrededor de la media poblacional modficando el tamaño muestral a 15

p_15 = pnorm(10100, 10000, 500/sqrt(15)) - pnorm(9900, 10000,
500/sqrt(15))

# Mostrar probabilidad de que la resistencia promedio esté 100 unidades
alrededor de la
# media si el tamaño de la muestra es de 15 en lugar de 120

cat("P(9900 < X_b < 10100) = ", p_15)

## P(9900 < X_b < 10100) =  0.561422
```

Desviaciones lejos de la media (z)

```r
# Calcular la cantidad de desviaciones estándar a la que la resistencia
media de la muestra
# está lejos de la resistencia promedio de la población si el tamaño
muestral es 15 en vez
# de 120

z_15 = 100 / (500/sqrt(15))
```

```r
# Mostrar cantidad de desviaciones estándar

cat("z = ", z_15)

## z =  0.7745967
```

Inciso D

$$H_0: \mu = 10000$$

$$H_1: \mu < 10000$$

d) Un ingeniero recibió un lote muy grande de remaches. Antes de aceptarlo verificó si efectivamente tiene una media de 10 000 lb/pulg 2. Para ello tomó una muestra de 120 remaches elegidos al azar tenía media de 9800 lb/pulg 2 y rechazó el pedido, ¿hizo lo correcto? ¿por qué?. Si la media hubiera sido 9925, ¿recomendarías rechazarlo?

```r
# Calcular la cantidad de desviaciones estándar a la que la media de la
muestra está lejos
# de la media deseada por el ingeniero

z_D = (9800 - 10000)/(500/sqrt(120))

# Mostrar la cantidad de desviaciones estándar que representa la lejanía
de la media
# muestral respecto a la media deseada

cat("z = ", z_D)

## z =  -4.38178

# Probabilidad de que la media muestral sea de 9800 si la media deseada
es de 10000 y la
# desviación estándar de la poblción es de 500 y tamaño muestral de 120

p_9800 = pnorm(9800, 10000, 500/sqrt(120))

# Mostrar probabilidad calculada

cat("P(X_barra = 9800) = ", p_9800)

## P(X_barra = 9800) =  5.88567e-06
```

En base a los resultados previos, el ingeniero hizo lo correcto al rechazar el pedido, debido a que en base al número de desviaciones estándar calculado, la media de la muestra que el ingeniero tomó está a 4.38178 desviaciones estándar de la media deseable, lo cual implica que en realidad la media de la muestra de remaches se encuentra muy alejada de la deseable para que el pedido recibido por el ingeniero cumpliera con el criterio para considerarlo adecuado en cuanto a calidad, por lo que la decisión de rechazar dicho pedido por parte del ingeniero fue la correcta.

Inciso E

e) ¿Qué decisión recomiendas al ingeniero si la media obtenida en la media hubiera sido 9925? ¿recomendarías rechazarlo?

```
# Calcular nuevamente el número de desviaciones estándar al que la nueva
media muestral
# está lejos de la media deseada por el ingeniero

z_9925 = (9925 - 10000) / (500/sqrt(120))

# Imprimir el número de desviaciones estándar (z) calculado

cat("z = ", z_9925)

## z =  -1.643168

# Probabilidad de que la media muestral sea de 9925 si la media deseada
es de 10000 y la
# desviación estándar de la poblción es de 500 y tamaño muestral de 120

p_9925 = pnorm(9925, 10000, 500/sqrt(120))

# Mostrar la probabilidad de que la media muestral sea 9925 si la media
deseada es 10000 y
# la desviación estándar poblacional es 500 con tamaño muestral de 120

cat("P(X_barra = 9925) = ", p_9925)

## P(X_barra = 9925) =  0.05017412
```

Por otra parte, si la media de la muestra hubiera sido de 9925 en lugar de 9800, no sería recomendable rechazar el pedido, principalmente debido a que la media de la muestra estaría a tan solo 1.643 desviaciones estándar de la media deseada para considerar al pedido de buena calidad, lo cual al no representar una lejanía significativa entre ambas medias, es posible afirmar que en este caso particular, no es recomendable rechazar el pedido, ya que la media muestral al no estar muy alejada de la ideal, garantiza que la calidad del pedido recibido sea adecuada para aceptarlo.

Problema 3. Embotellando

Una máquina embotelladora puede ser regulada para que se descargue un promedio de μ onzas por botella. Se ha observado que la cantidad de líquido dosificado por una máquina embotelladora está distribuida normalmente con $\sigma = 1$ onza. La máquina embotelladora se calibra cuando la media de una muestra tomada al azar está fuera del 95% central de la distribución muestral. La media de la cantidad de líquido deseada requiere que μ sea de 15 onzas.

Inciso a

a) ¿A cuántas desviaciones estándar alrededor de la verdadera media μ puede estar la media de una muestra para que esté dentro del estándar establecido del 95% central?

```
# Calcular el valor de z correspondiente a una probabilidad de 0.025, con una media de 0
# y desviación estándar de 1

# Esto representa el valor z correspondiente al 2.5% de probabilidad de que la media miu
# se encuentre fuera de la zona central del 95% de los datos

z_2.5 = qnorm(0.025, 0, 1)

# Mostrar el valor de z, es decir, la cantidad de desviaciones estándar a la que la media
# muestral está lejos de la media poblacional miu

cat("z = ", abs(z_2.5)) # se toma el valor positivo del resultado obtenido para z

## z =  1.959964
```

Inciso b

b) ¿Cuál es la probabilidad de que en una muestra aleatoria de tamaño 10 botellas se obtenga una media mayor a 16 onzas?

```
# Calcular la probabilidad de que en la muestra de tamaño 10 se obtenga una media superior
# a 16 onzas

# En la función de abajo, 16 representa la media muestral, 15 es la media deseada o también
# llamada media de la población y 1/sqrt(10) se refiere al error estándar
# (desviación estándar poblacional entre raíz cuadrada del tamaño de la muestra)

# lower.tail = FALSE es para obtener la probabilidad de que la media se mayor a 16 onzas

prob_16 = pnorm(16, 15, 1/(sqrt(10)), lower.tail = FALSE)

# Mostrar la probabilidad de que la media muestral sobrepase 16 onzas

cat("P(X_barra > 16) = ", prob_16)

## P(X_barra > 16) =  0.0007827011
```

Inciso c

c) Si en una muestra aleatoria de tamaño 10 botellas se obtuvo una media de 16 onzas, ¿se detendría la producción para calibrar la máquina?

```
# Calcular el valor z correspondiente a la diferencia entre la media de
la muestra menos
# la media poblacional deseada, para lo cual, se resta la media deseada
de 15 a la media
# de la muestra 16 y la diferencia se divide entre la desviación estándar
de la muestra o
# error estándar

z_16 = (16 - 15) / (1 / sqrt(10))

# Imprimir el valor de z correspondiente a la media muestral de 16 onzas

cat("z = ", z_16)

## z =  3.162278
```

Sí se detendría la producción para calibrar la máquina, debido a que la media muestral de 16 onzas se ubica a 3.162278 desviaciones estándar de la media poblacional $\mu = 15$, lo cual sobrepasa el límite máximo de desviaciones estándar a las que puede estar la media de la muestra para estar todavía dentro del 95% central de la distribución muestral, motivo por el cual, sí se requiere detener la producción temporalmente para calibrar la máquina embotelladora.

Inciso d

d) ¿Cuál es la probabilidad de que en una muestra aleatoria de tamaño 10 botellas se obtenga una media menor a 14.5 onzas?

```
# Calcular la probabilidad de que se obtenga un promedio de onzas menor a
14.5 en una
# muestra

# 14.5 representa la media de onzas de la muestra, 15 representa la media
poblacional (miu)
# y 1/sqrt(10) es el error estándar o desviación estándar de la muestra

p_14.5 = pnorm(14.5, 15, 1/sqrt(10))

# Mostrar la probabilidad de que en una muestra de 10 botellas, se
obtenga una media de
# onzas inferior a 14.5 onzas

cat("P(X_barra < 14.5) = ", p_14.5)

## P(X_barra < 14.5) =  0.05692315
```

Inciso e

e) Si en una muestra aleatoria de tamaño 10 botellas se obtuvo una media de 15.5 onzas, ¿se detendría la producción para calibrar la máquina?

```
# Calcular el número de desviaciones estándar z a las que se encuentra la
media muestral de
# 15.5 onzas de la media poblacional o media deseada de 15 onzas

# Calcular diferencia de medias y dividir el resultado entre la
desviación estándar
# de la muestra (error estándar)

z_15.5 = (15.5 - 15) / (1 / sqrt(10))

# Mostrar la cantidad de desviaciones estándar a la que la media muestral
está lejos de
# la media poblacional

cat("z = ", z_15.5)

## z =  1.581139
```

De acuerdo a la cantidad de desviaciones estándar a la que la media de la muestra está lejos de la media deseada (media poblacional o miu), se concluye que no se debe detener la producción para calibrar la máquina ebotelladora, debido a que la media de la muestra se encuentra a 1.581139 desviaciones estándar de la media poblacional $\mu = 15$, lo cual es inferior a las 1.95996 desviaciones estándar a las que puede estar como máximo la media de la muestra para que aún se encuentre dentro del 95% central de la distribución muestral, por lo tanto, teniendo lo anterior en consideración, la producción no se debe detener para calibrar la máquina.

Inciso f

f) Hacer una gráfica del inciso 1.

```
# Crear un conjunto de posibles valores de onzas para realizar el gráfico
de la
# distribución normal (los valores irán desde 0 hasta 30 con incremento
de 0.5)

onzas = seq(from = 0, to = 30, by = 0.5)

# Utilizar la función dnorm para evaluar cada uno de los valores de onzas
definidos con
# anterioridad en la función de densidad de la distribución normal

# 0 representa la media de la población para graficar la distribución
normal estándar

# 1 representa la desviación estándar igualmente de la población normal
estándar
```

```r
y_onzas = dnorm(onzas, 0, 1)

# Graficar la distribución normal estándar de los datos anteriores

plot(onzas, y_onzas, main = "Distribución normal de la cantidad de 
onzas",
     xlab = "Cantidad de onzas", ylab = "f(onzas)")

# Graficar una línea vertical sobre la cantidad máxima de desviaciones 
estándar a la que
# puede estar la media de la muestra para estar dentro del 95% central de 
la distribución
# muestral

abline(v = abs(z_2.5), col = "blue")
```