Analisis de datos trabajo simulacion

Índice

Autores:

Maria Fernanda Calle Agudelo - mcalleag@unal.edu.co

Jaider Castañeda Villa - jcastanedavi@unal.edu.co

Mónica Sofía Restrepo León - morestrepol@unal.edu.co

Luis Alejandro Varela Ojeda - luvarelao@unal.edu.co

1. Introduccion

En este documento se llevaran a cabo los respectivos analisis sobre los datos tomados con el fin de determinar a que distribucion pertenecen estos

2. Cargar paquetes

```
library("MASS")
library("car")
library("fitdistrplus")
library("moments")
```

3. Cargar datos

```
datos<-read.csv("Datos.csv")
```

4. Tiempo entre llegadas

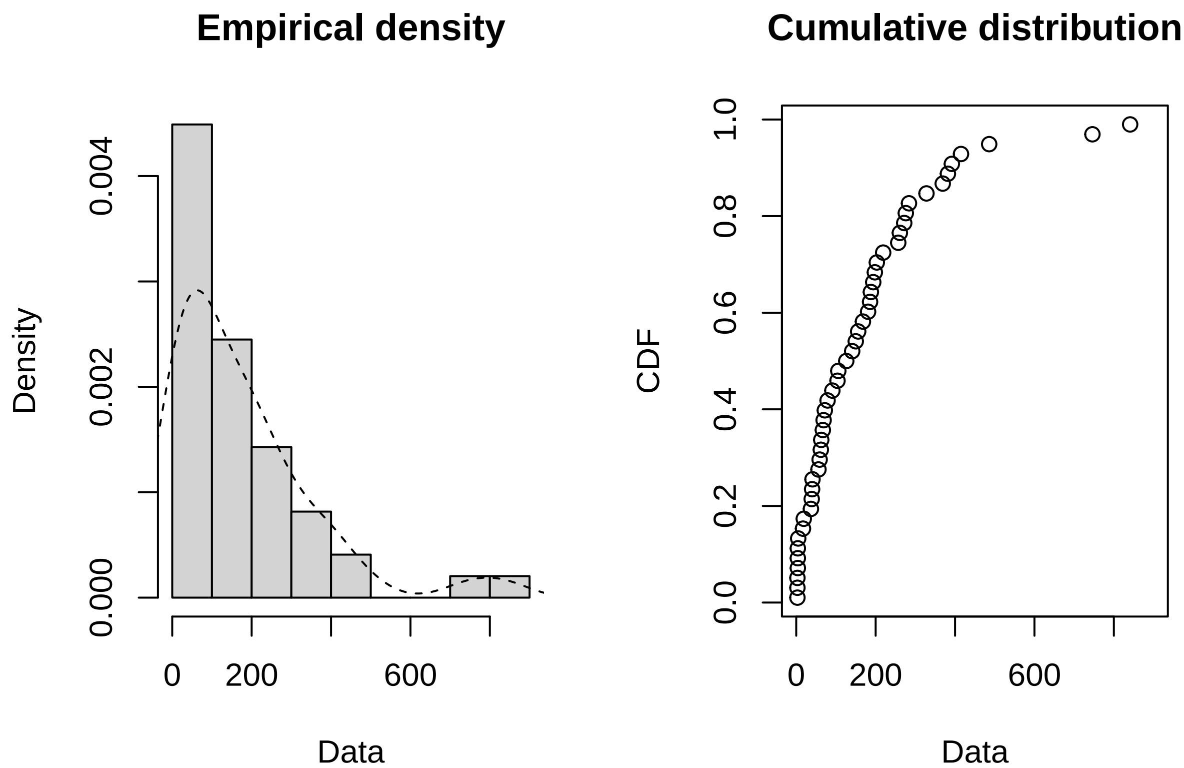
```
datos$Tiempo.de.llegada <- as.POSIXlt(datos$Tiempo.de.llegada, format = "%H:%M:%S")
tiempoDellegadas <- as.numeric(datos$Tiempo.de.llegada, units = "secs")

tiemposEntreLlegadasSec <- as.numeric(diff(datos$Tiempo.de.llegada), units = "secs")
tiemposEntreLlegadasSec <- tiemposEntreLlegadasSec[-21]
sd(tiemposEntreLlegadasSec)
```

```
## [1] 180.8375
```

Es necesaria la eliminacion del dato numero 21 ya que es el primer dato del segundo dia, el tiempo 0 es partir de que se hace el primer pedido

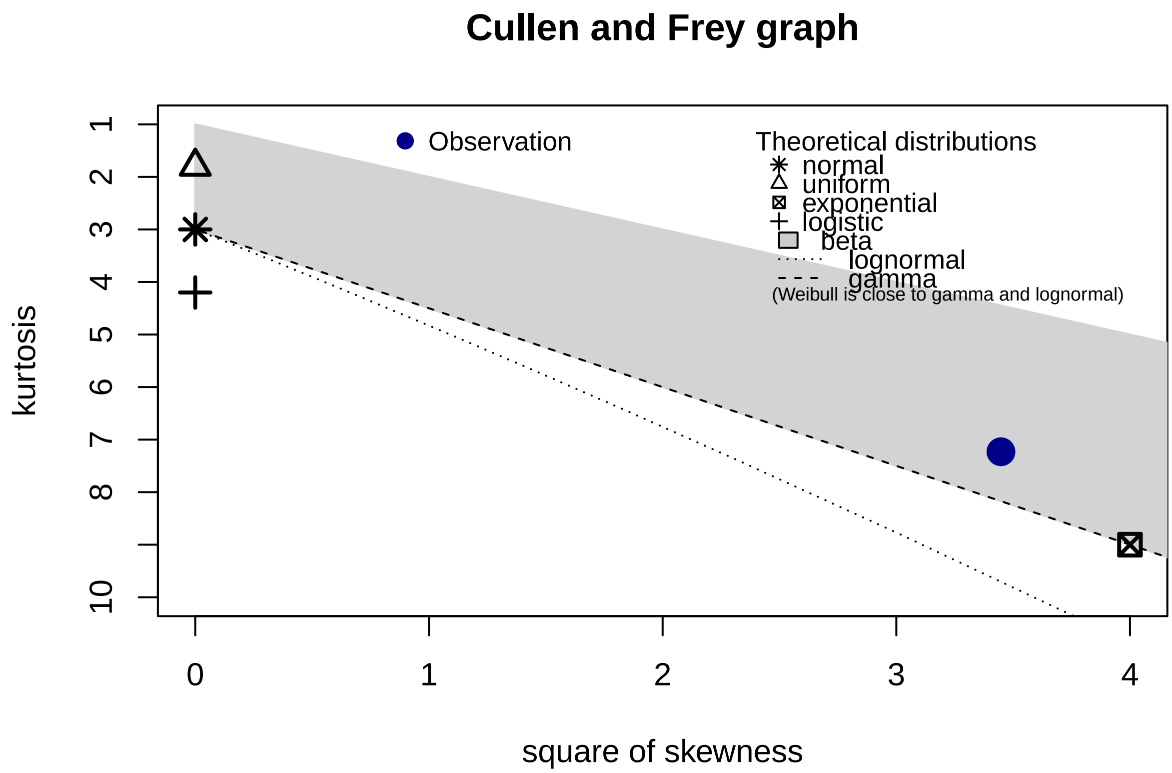
4.1. Analisis de distribucion mas apropiada

```
plotdist(tiemposEntreLlegadasSec,histo=TRUE,demp=TRUE)
```



Podriamos decir que los datos tienen forma cercana a la distribucion exponencial

```
descdist(tiemposEntreLlegadasSec)
```

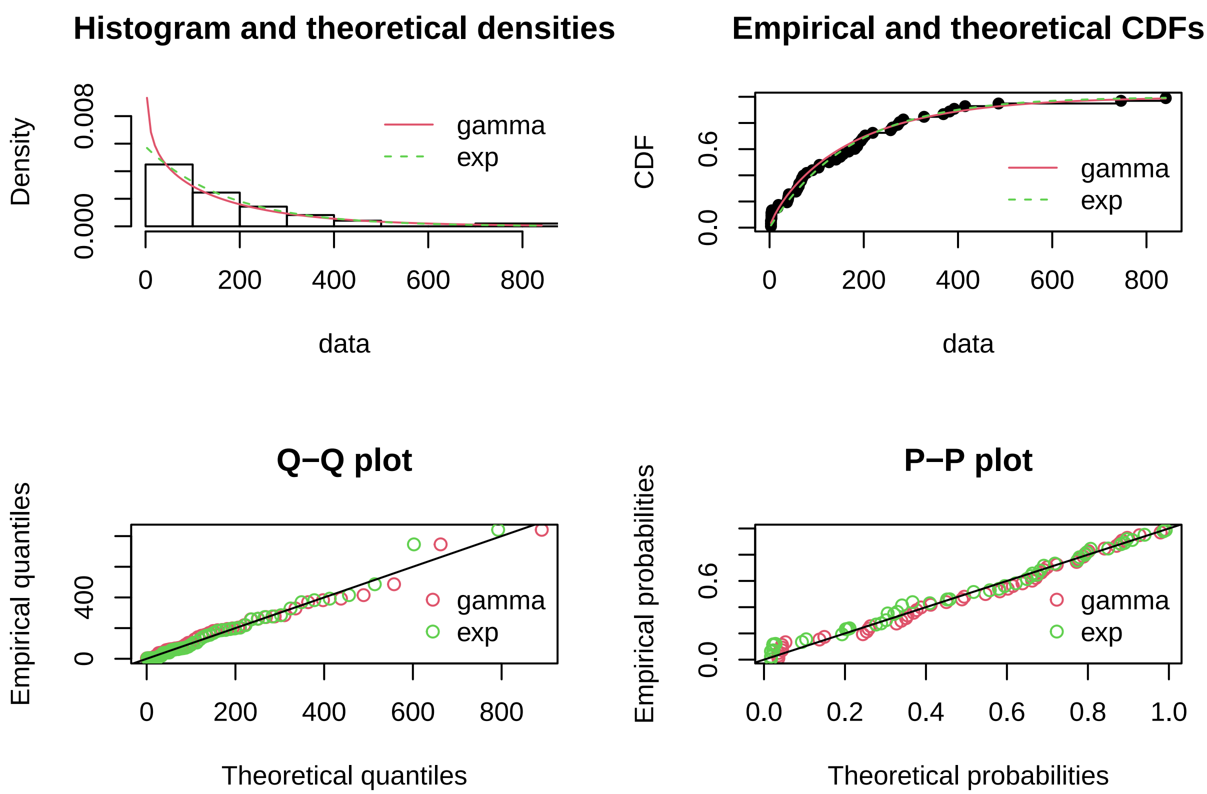



```
## summary statistics
## ------
## min:  3   max:  841
## median:  126
## mean:  172.7755
## estimated sd:  180.8375
## estimated skewness:  1.856844
## estimated kurtosis:  7.231298
```

Se puede ver que los datos estan cerca tanto de la distribucion gamma como de la distribucion exponencial asi que sobre esas dos haremos las pruebas respectivas para ver cual es la mas adecuada.

```
fit_gamma <- fitdist(tiemposEntreLlegadasSec,"gamma")
fit_exp <- fitdist(tiemposEntreLlegadasSec,"exp")
```

```
par(mfrow=c(2,2))
plot.legend <- c("gamma","exp")
denscomp(list(fit_gamma,fit_exp), legendtext = plot.legend)
cdfcomp (list(fit_gamma,fit_exp), legendtext = plot.legend)
qqcomp (list(fit_gamma,fit_exp), legendtext = plot.legend)
ppcomp (list(fit_gamma,fit_exp), legendtext = plot.legend)
```



4.2. Prueba de bondad de ajuste

Observando las gráficas, se puede notar que la distribución exponencial se asemeja más a la distribución de los datos. Se llevará a cabo una prueba de Kolmogorov-Smirnov con un nivel de confianza del 95 %, utilizando un valor crítico de D=0.1943, esto con el fin de determinar si los datos si se distribuyen como tales distribuciones

```r
gofstat(list(fit_gamma,fit_exp))
```

```
## Goodness-of-fit statistics
##                              1-mle-gamma 2-mle-exp
## Kolmogorov-Smirnov statistic  0.08982890 0.1143326
## Cramer-von Mises statistic    0.05708485 0.0653053
## Anderson-Darling statistic    0.48468038 0.8941509
##
## Goodness-of-fit criteria
##                                1-mle-gamma 2-mle-exp
## Akaike's Information Criterion    605.0543  604.8953
## Bayesian Information Criterion    608.8380  606.7871
```

Como se puede ver los datos se distribuyen como ambas distribuciones, por lo tanto se usaran los criterios de Akaike y Bayesiano para determinar cual distribucion se usara en el modelo, siendo esta la distribucion exponencial.

```r
summary(fit_exp)
```

```
## Fitting of the distribution ' exp ' by maximum likelihood
## Parameters :
##         estimate   Std. Error
## rate 0.005787857 0.0008012352
## Loglikelihood:  -301.4477   AIC:  604.8953   BIC:  606.7871
```

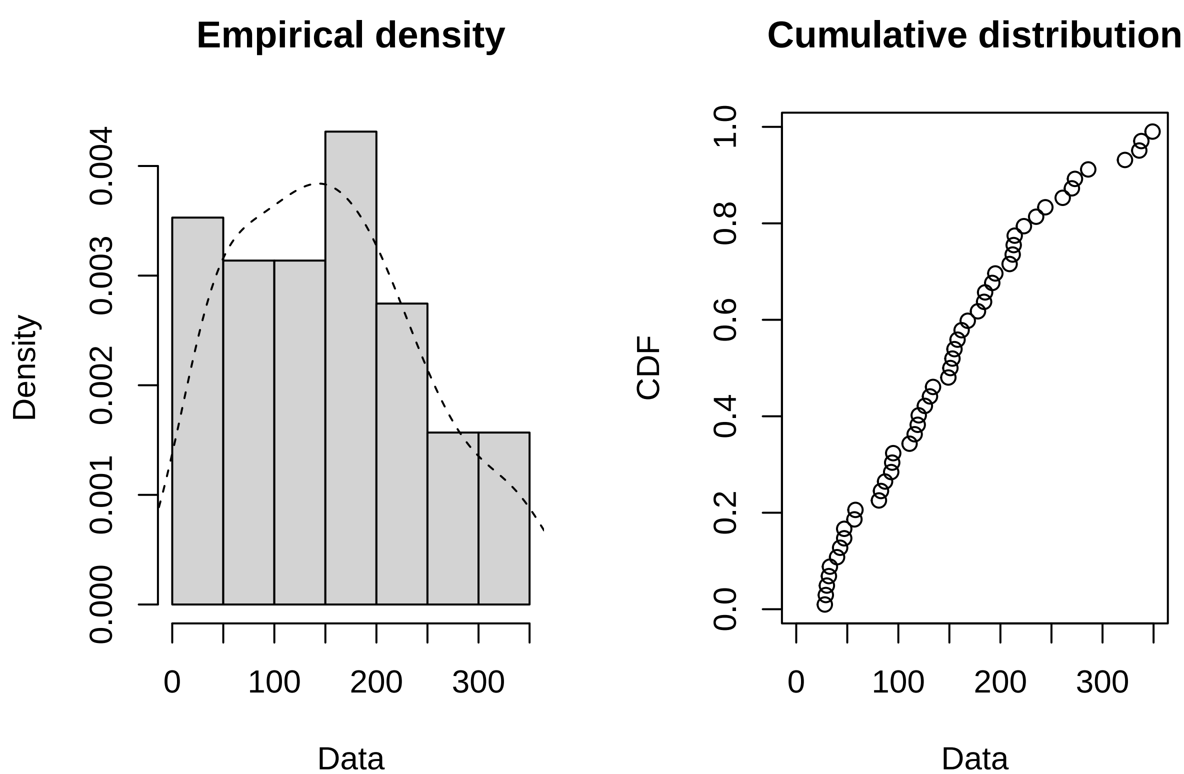
5. Tiempo de pedido

Esta variable trata sobre cuanto tiempo tarda una persona en hacer su pedido sin tener en cuenta que tipo de pedido esta realizando.

```r
datos$Termina.de.hacer.el.pedido <- as.POSIXlt(datos$Termina.de.hacer.el.pedido, format = "%H:%M:%S")
tiempoPedidoSec <- as.numeric(difftime(datos$Termina.de.hacer.el.pedido, datos$Tiempo.de.llegada, units
```

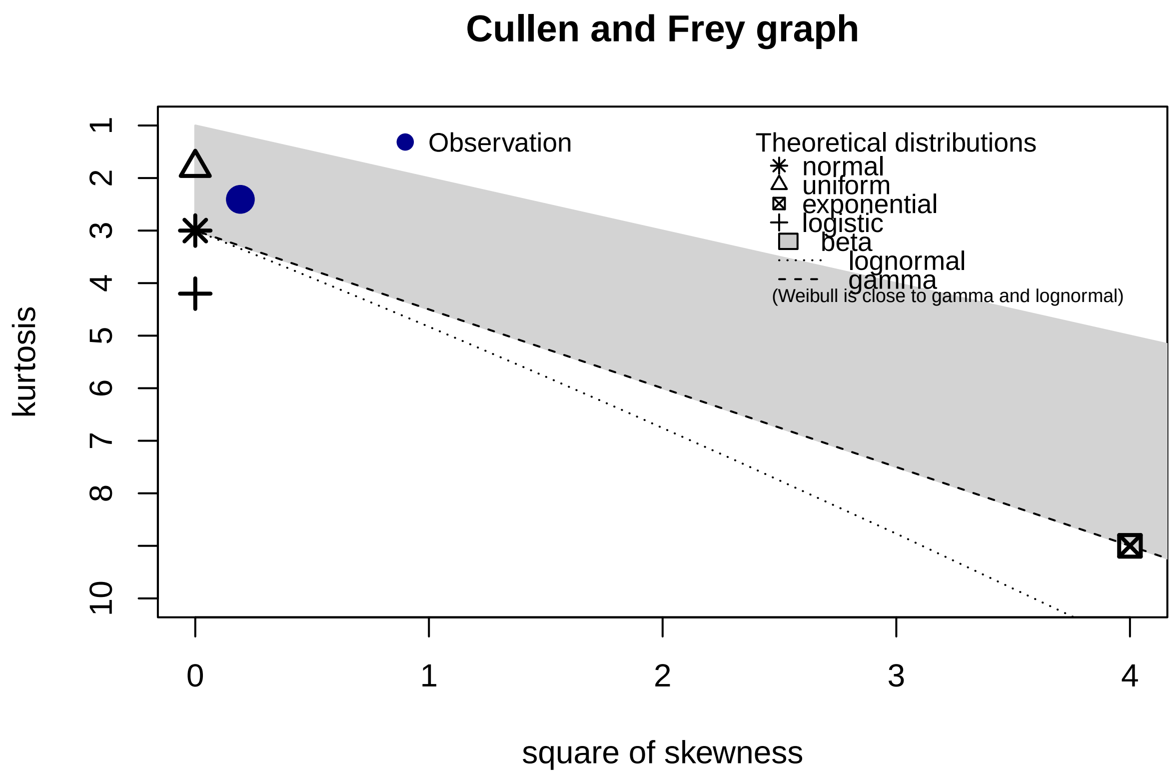
5.1. Analisis de distribucion mas apropiada

```r
plotdist(tiempoPedidoSec,histo=TRUE,demp=TRUE)
```



Podriamos decir que los datos se asemejan a algo como una normal pero no se puede decir con certeza.

```r
descdist(tiempoPedidoSec)
```

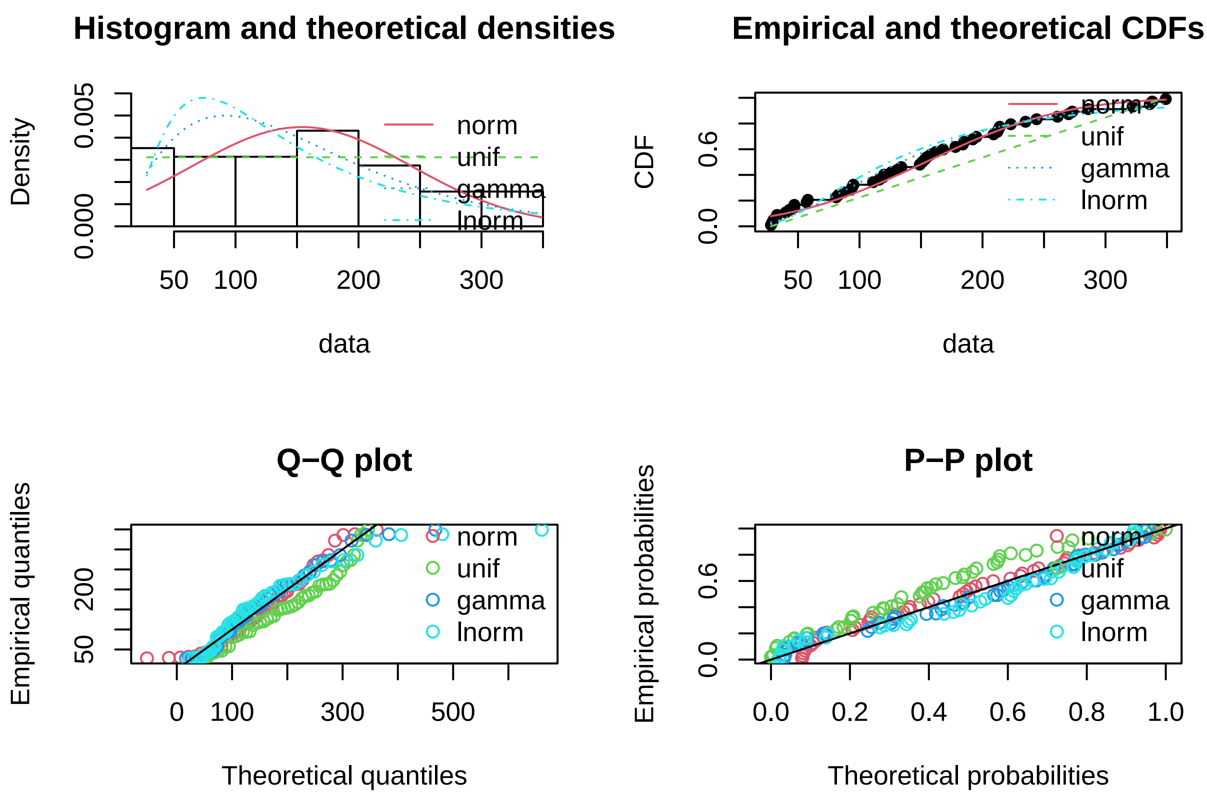



```
## summary statistics
## ------
## min:  28   max:  349
## median:  151
## mean:  153.902
## estimated sd:  89.97494
## estimated skewness:  0.4393988
## estimated kurtosis:  2.405434
```

Se puede ver que los datos estan cerca tanto de la distribucion uniforme, normal, gamma y lognormal asi que sobre estas distribuciones se haran las pruebas de ajuste.

```r
fit_normal <- fitdist(tiempoPedidoSec,"norm")
fit_unif <- fitdist(tiempoPedidoSec,"unif")
fit_gamma <- fitdist(tiempoPedidoSec,"gamma")
fit_lnorm <- fitdist(tiempoPedidoSec,"lnorm")
```

```r
par(mfrow=c(2,2))
plot.legend <- c("norm","unif","gamma","lnorm")
denscomp(list(fit_normal,fit_unif,fit_gamma,fit_lnorm), legendtext = plot.legend)
cdfcomp (list(fit_normal,fit_unif,fit_gamma,fit_lnorm), legendtext = plot.legend)
qqcomp (list(fit_normal,fit_unif,fit_gamma,fit_lnorm), legendtext = plot.legend)
ppcomp (list(fit_normal,fit_unif,fit_gamma,fit_lnorm), legendtext = plot.legend)
```



5.2. Prueba de bondad de ajuste

A ojo se puede decir que la que mas se adapta a los datos es la distribucion normal pero aun asi haremos la prueba Kolmogorov-Smirnov con un valor de confianza del 95 % para un valor critico de D = 0.1904.

```
gofstat(list(fit_normal,fit_unif,fit_gamma,fit_lnorm))
```

```
## Goodness-of-fit statistics
##                              1-mle-norm 2-mle-unif 3-mle-gamma 4-mle-lnorm
## Kolmogorov-Smirnov statistic 0.07907930  0.2048745  0.09422004   0.1302051
## Cramer-von Mises statistic   0.05847433  0.7305584  0.08093470   0.1873759
## Anderson-Darling statistic   0.52738318        Inf  0.58635855   1.2010255
##
## Goodness-of-fit criteria
##                                1-mle-norm 2-mle-unif 3-mle-gamma 4-mle-lnorm
## Akaike's Information Criterion   606.6740   592.6870    601.0315    606.4516
## Bayesian Information Criterion   610.5376   596.5506    604.8951    610.3152
```

Se descarta la distribucion uniforme por el criterio de Kolmogorov-Smirnov y para determinar cual es la distribucion mas adecuada nos basaremos en el criterio visual ya que sus resultados son bastante similares dando por escogida la distribucion normal.

```
summary(fit_normal)
```

```
## Fitting of the distribution ' norm ' by maximum likelihood
## Parameters :
##        estimate Std. Error
## mean 153.90196  12.474873
## sd    89.08847   8.821072
## Loglikelihood:  -301.337   AIC:  606.674   BIC:  610.5376
## Correlation matrix:
##      mean sd
## mean    1  0
## sd      0  1
```

6. Tiempo de servicio

Aqui el tiempo de servicio se divide en dos categorias aquellas en las que el pedido debe ser enviado a la cocina y aquellos donde compran algo y se les entrega sin tener que esperar demasiado a esta categoria se le llama estante.

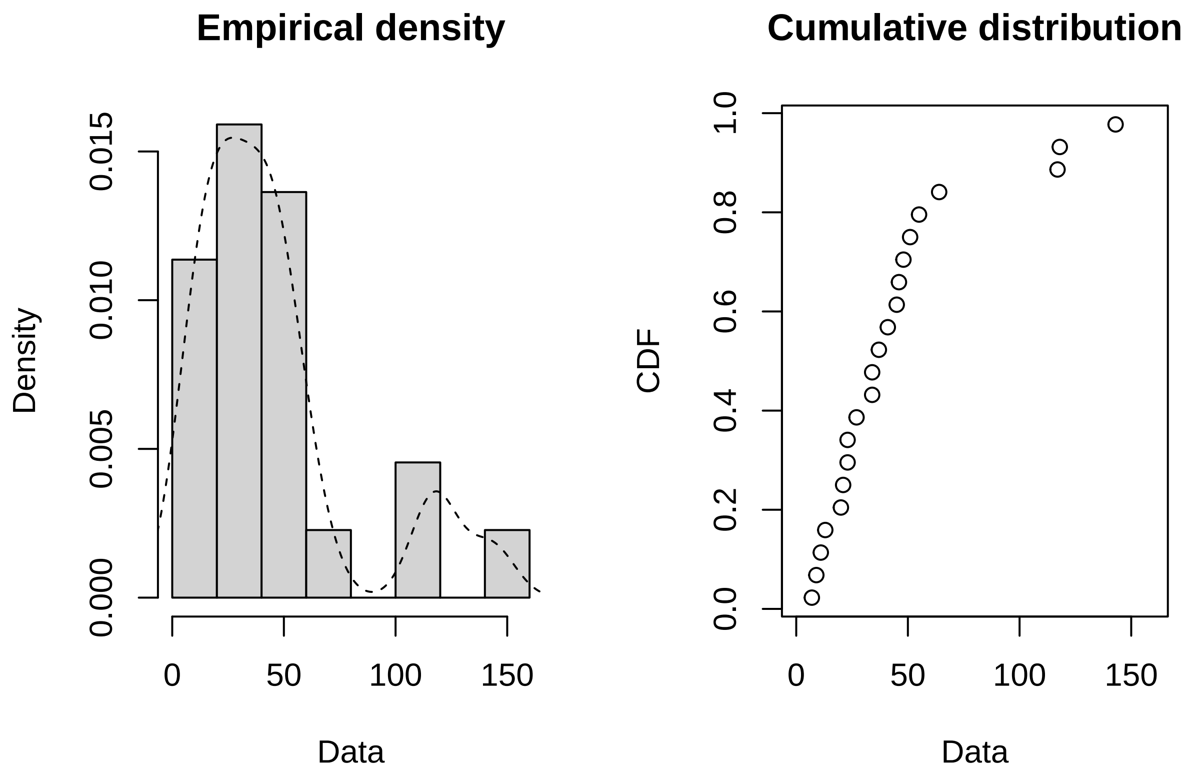
```
datos$Le.entregan.el.pedido <- as.POSIXlt(datos$Le.entregan.el.pedido, format = "%H:%M:%S")
tiempoDeServicio <- as.numeric(difftime(datos$Le.entregan.el.pedido, datos$Termina.de.hacer.el.pedido, u
datosASeparar <- data.frame(
  tiemposSec = tiempoDeServicio,
  tipo       = datos$Tipo
)
```

6.1. Tiempo de servicio para el area del estante

Esto se refiere a cuanto tiempo se tarda en entregarle el producto a alguien que va a comprar algo del estante tipo mecato y similares que ya estan hechos.

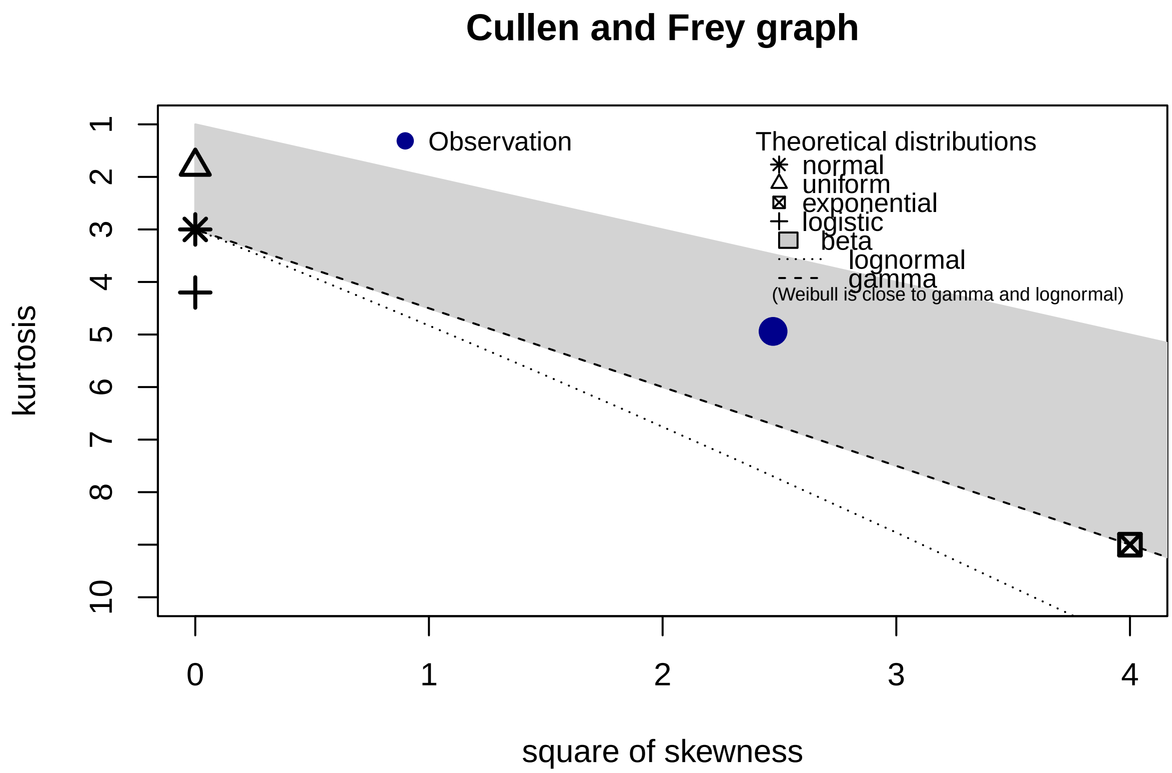
6.1.1. Analisis de distribucion mas apropiada

```
datosEstante <- datosASeparar[datosASeparar$tipo=="estante",]
datosEstante$tiemposSec <- as.numeric(datosEstante$tiemposSec)
plotdist(datosEstante$tiemposSec,histo=TRUE,demp=TRUE)
```



Dificil de saber solo a ojo.

```
descdist(datosEstante$tiemposSec)
```

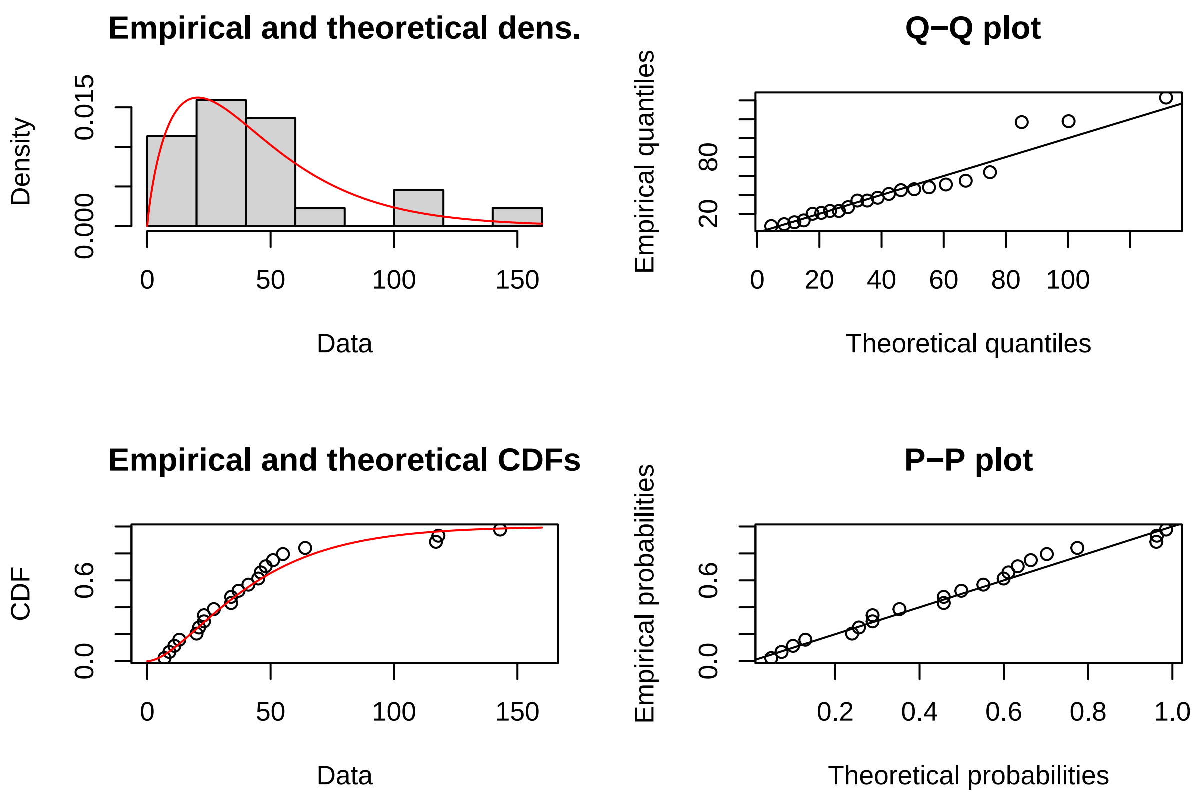



```
## summary statistics
## ------
## min:  7   max:  143
## median:  35.5
## mean:  44.86364
## estimated sd:  36.74791
## estimated skewness:  1.572489
## estimated kurtosis:  4.940581
```

Se puede ver que los datos estan cerca de la distribucion gamma asi que sobre esa haremos las pruebas respectivas para ver cual es la mas adecuada.

```r
fit_gamma <- fitdist(datosEstante$tiemposSec,"gamma")
```

```r
plot(fit_gamma)
```



Si, los datos se ven bastante cerca de la distribucion gamma, ahora haremos la Prueba de bondad de ajuste.

6.1.2. Prueba de bondad de ajuste

Se llevará a cabo una prueba de Kolmogorov-Smirnov con un nivel de confianza del 95 %, n=22 utilizando un valor crítico de D=0.270, esto con el fin de determinar si los datos si se distribuyen como tales distribuciones

```
gofstat(fit_gamma)
```

```
## Goodness-of-fit statistics
##                              1-mle-gamma
## Kolmogorov-Smirnov statistic  0.11611550
## Cramer-von Mises statistic    0.04761528
## Anderson-Darling statistic    0.37114924
##
## Goodness-of-fit criteria
##                                1-mle-gamma
## Akaike's Information Criterion    211.2545
## Bayesian Information Criterion    213.4366
```

Se puede decir que los datos se distribuyen como Gamma ya que estan dentro del area de no rechazo en la prueba Kolmogorov.

```r
summary(fit_gamma)
```

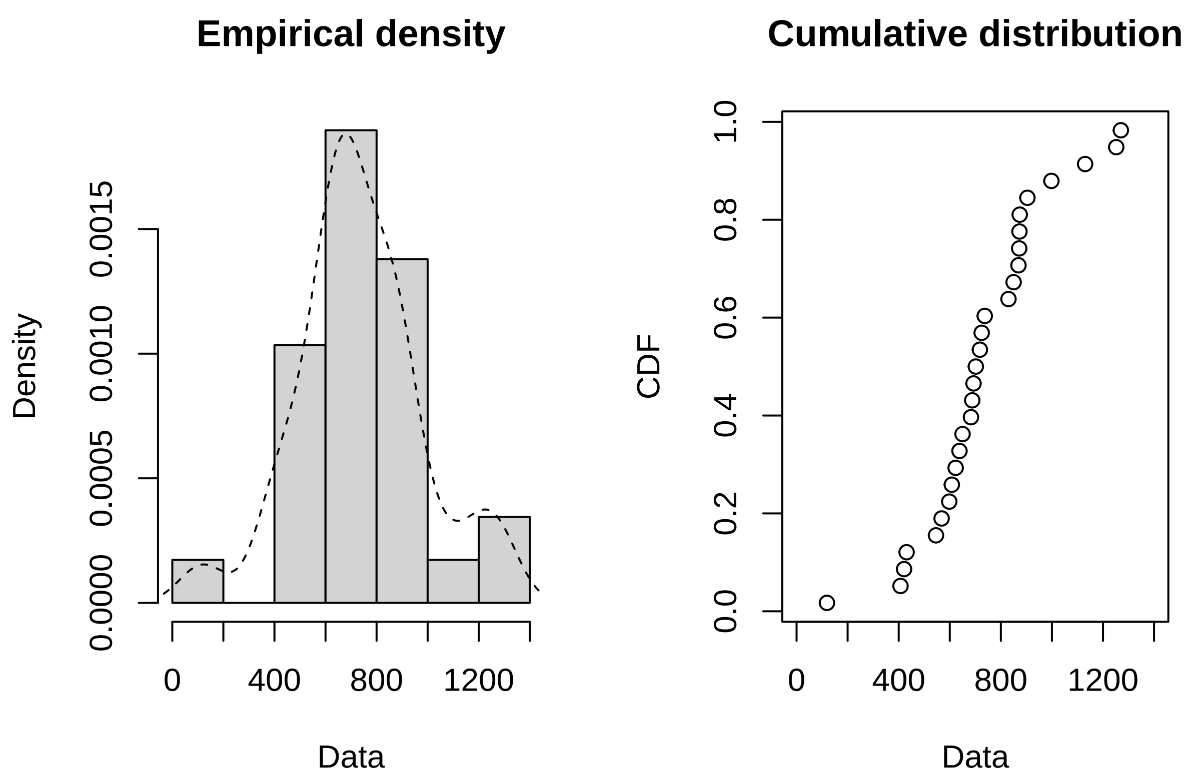
```
## Fitting of the distribution ' gamma ' by maximum likelihood
## Parameters :
##          estimate Std. Error
## shape 1.84390389 0.51295299
## rate  0.04110727 0.01311948
## Loglikelihood:  -103.6273   AIC:  211.2545   BIC:  213.4366
## Correlation matrix:
##           shape      rate
## shape 1.0000000 0.8707877
## rate  0.8707877 1.0000000
```

6.2. Tiempo de servicio para el area del restaurante

Esto se refiere a cuanto tiempo se tarda en entregarle el producto a alguien que va a comprar algo del estante tipo mecato y similares que ya estan hechos.

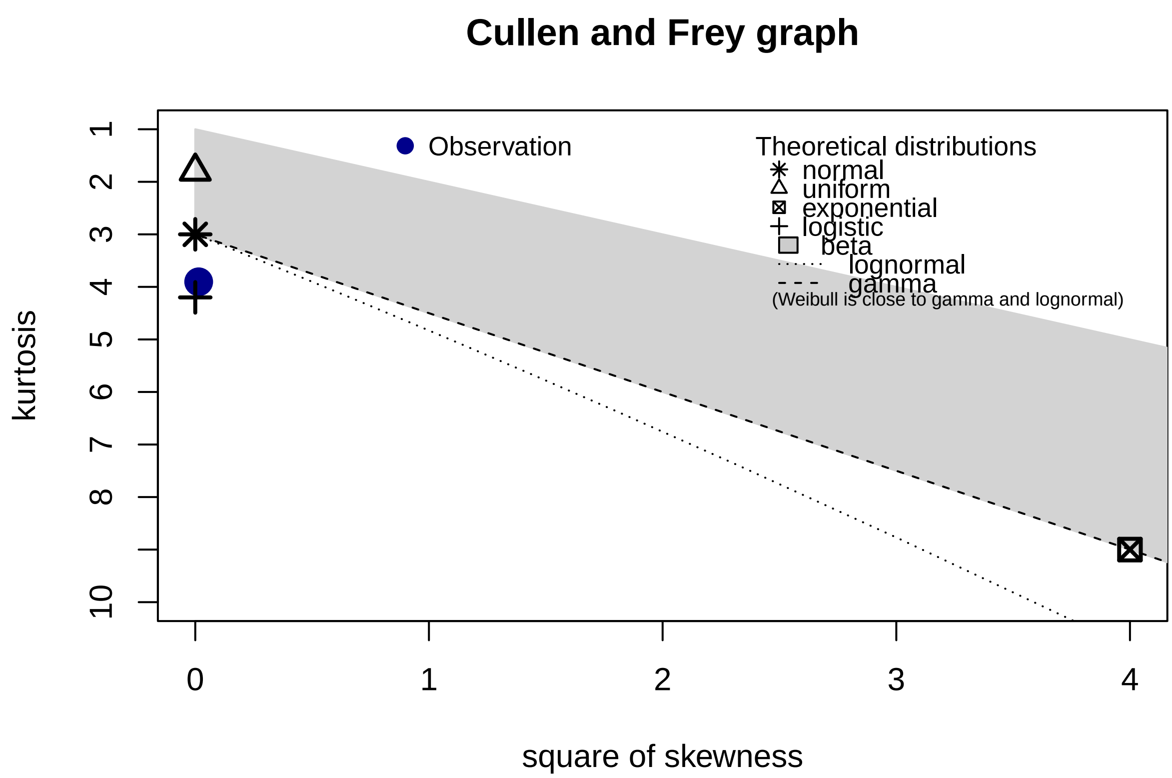
6.2.1. Analisis de distribucion mas apropiada

```r
datosRestaurante <- datosASeparar[datosASeparar$tipo=="restaurante",]
datosRestaurante$tiemposSec <- as.numeric(datosRestaurante$tiemposSec)
plotdist(datosRestaurante$tiemposSec,histo=TRUE,demp=TRUE)
```



Puede que se distribuya como una normal si nos basamos en el diagrama de densidad.

```
descdist(datosRestaurante$tiemposSec)
```



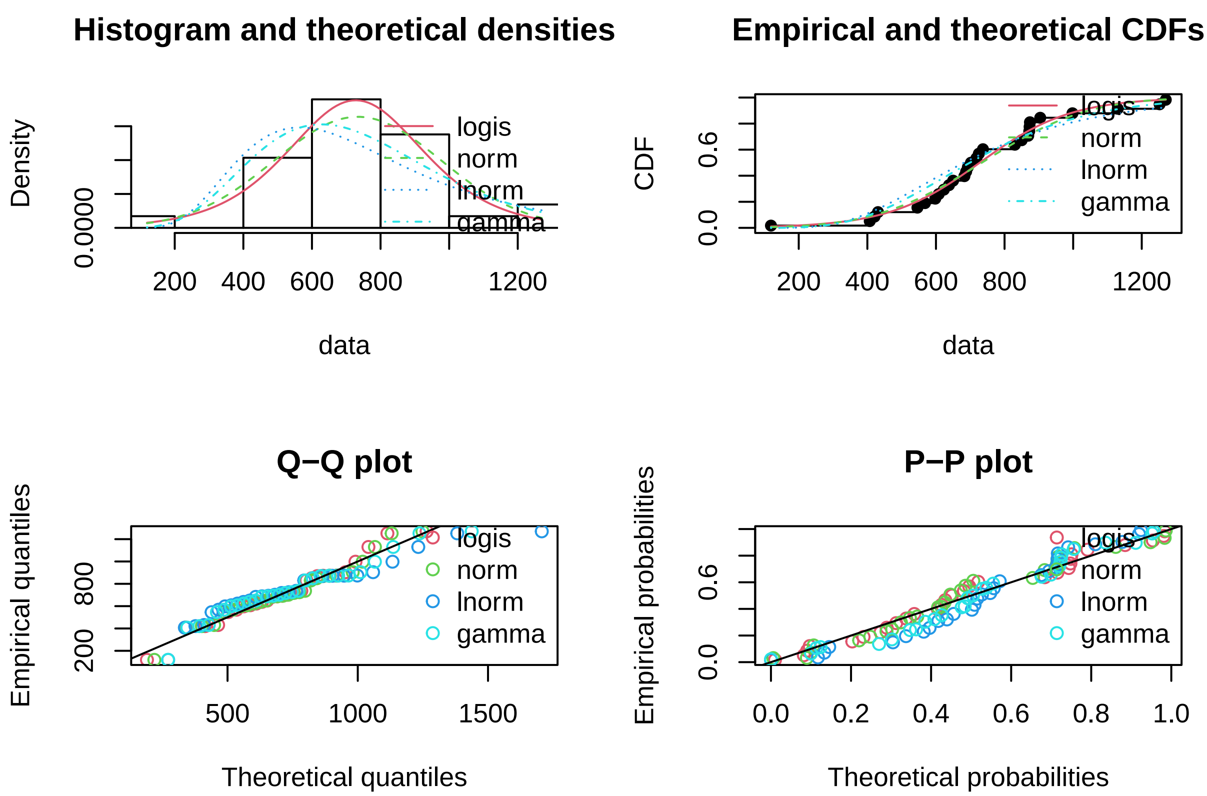

```
## summary statistics
## ------
## min:  119   max:  1270
## median:  702
## mean:  733.6897
## estimated sd:  247.7366
## estimated skewness:  0.1210631
## estimated kurtosis:  3.903518
```

Se puede ver que los datos estan cerca de la distribucion logistic, normal, lognormal y gamma asi que haremos el analisis de distribucion sobre estas.

```
fit_logis <- fitdist(datosRestaurante$tiemposSec,"logis")
fit_norm <- fitdist(datosRestaurante$tiemposSec,"norm")
fit_lnorm <- fitdist(datosRestaurante$tiemposSec,"lnorm")
fit_gamma <- fitdist(datosRestaurante$tiemposSec,"gamma")
```

```
par(mfrow=c(2,2))
plot.legend <- c("logis","norm","lnorm","gamma")
denscomp(list(fit_logis,fit_norm,fit_lnorm,fit_gamma), legendtext = plot.legend)
```

```r
cdfcomp (list(fit_logis,fit_norm,fit_lnorm,fit_gamma), legendtext = plot.legend)
qqcomp (list(fit_logis,fit_norm,fit_lnorm,fit_gamma), legendtext = plot.legend)
ppcomp (list(fit_logis,fit_norm,fit_lnorm,fit_gamma), legendtext = plot.legend)
```



Diria que las que mas se ajustan serian la distribucion gamma y la distribucion normal.

6.2.2. Prueba de bondad de ajuste

Se llevará a cabo una prueba de Kolmogorov-Smirnov con un nivel de confianza del 95 %, n=29 utilizando un valor crítico de D=0.240, esto con el fin de determinar si los datos si se distribuyen como tales distribuciones

```r
gofstat(list(fit_logis,fit_norm,fit_lnorm,fit_gamma))
```

```
## Goodness-of-fit statistics
##                              1-mle-logis 2-mle-norm 3-mle-lnorm 4-mle-gamma
## Kolmogorov-Smirnov statistic  0.10280259 0.11526466   0.1746822   0.1404285
## Cramer-von Mises statistic    0.04675648 0.06778987   0.1927668   0.1134065
## Anderson-Darling statistic    0.31726448 0.44043497   1.2333841   0.7390273
##
## Goodness-of-fit criteria
##                                1-mle-logis 2-mle-norm 3-mle-lnorm 4-mle-gamma
## Akaike's Information Criterion    404.0842   404.9980    416.3055    409.8933
## Bayesian Information Criterion    406.8188   407.7326    419.0401    412.6279
```

Ninguna de las distribuciones es decartada por medio de la prueba Kolmogorov Smirnov asi que nos basaremos en los criterios Akaike y Bayesiano, e incluso si usamos el criterio grafico la normal y la logistic se ven bastante similares dando esto como resultado decidir que los datos se distribuyen como normal debido a que simul8 no tiene la distribucion logistic

```
summary(fit_norm)
```

```
## Fitting of the distribution ' norm ' by maximum likelihood
## Parameters :
##      estimate Std. Error
## mean 733.6897   45.20338
## sd   243.4278   31.96362
## Loglikelihood:  -200.499   AIC:  404.998   BIC:  407.7326
## Correlation matrix:
##      mean sd
## mean    1  0
## sd      0  1
```

7. Calculo de tamaños de muestra para las respectivas variables

```
tamanhoN <- function(datos,presicion) {
  desviation <- sd(datos)
  Value_Z <- qnorm(0.975)
  n<-(Value_Z*desviation/presicion)^2
  return(n)
}
```

7.1. Tamaño de muestra el tiempo entre llegadas

Tomando una presicion de 53 segundos con una confianza del 95 % obtenemos el siguiente valor para el tamaño de muestra para el tiempo entre llegadas.

```
nTiempoEntreLlegadas <- tamanhoN(tiemposEntreLlegadasSec,53)
ceiling(nTiempoEntreLlegadas)
```

```
## [1] 45
```

7.2. Tamaño de muestra para el tiempo de pedidos

Tomando una presicion de 25 segundos con una confianza del 95 % obtenemos el siguiente valor para el tamaño de muestra para el tiempo entre pedidos.

```
nTiempoPedido <- tamanhoN(tiempoPedidoSec,25)
ceiling(nTiempoPedido)
```

```
## [1] 50
```

7.3. Tamaño de muestra tiempo de servicio estante

Tomando una presicion de 17 segundos con una confianza del 95 % obtenemos el siguiente valor para el tamaño de muestra para el tiempo de servicio del estante.

```
nTiempoServicioEstante <- tamanhoN(datosEstante$tiemposSec,17)
ceiling(nTiempoServicioEstante)
```

```
## [1] 18
```

7.4. Tamaño de muestra tiempo de servicio restaurante

Tomando una presicion de 93 segundos con una confianza del 95 % obtenemos el siguiente valor para el tamaño de muestra para el tiempo de servicio del restaurante.

```
nTiempoServicioRestaurante <- tamanhoN(datosRestaurante$tiemposSec,93)
ceiling(nTiempoServicioRestaurante)
```

```
## [1] 28
```

8. Arbol de probabilidades - elementos simplificados

Probabilidades de que compre algo del estante o del restaurante

```
tipoDePedido<-table(datosASeparar$tipo)
tipoDePedido/sum(tipoDePedido)
```

```
##
##     estante restaurante
##   0.4313725   0.5686275
```

Las probabilidades de los productos del estante

```
probEstantes <- datos[datos$Tipo=="estante",]
tableProbEst <- table(probEstantes$Tipo.de.pedido)
tableProbEst/sum(tableProbEst)
```

```
##
##   brownie      cafe     hatsu      jugo
## 0.3181818 0.2272727 0.1363636 0.3181818
```

Las probabilidades de los productos del restaurante

```
probRestaurante <- datos[datos$Tipo=="restaurante",]
tableProbRest <- table(probRestaurante$Tipo.de.pedido)
tableProbRest/sum(tableProbRest)
```

```
##
##    ensalada hamburguesa       papas       perro
##  0.41379310  0.48275862  0.03448276  0.06896552
```