UNIVERSIDADE FEDERAL DA PARAÍBA - UFPB

CENTRO DE CIÊNCIAS SOCIAIS APLICADAS - CCSA

DEPARTAMENTO DE FINANÇAS E CONTABILIDADE – DFC

PROF. FILIPE COELHO DE LIMA DUARTE

JÉSSICA EVELIN SILVA DAMACENA

MATRICULA: 20170000459

RELATÓRIO

ANALISAR A DISTRIBUIÇÃO DO BANCO DE DADOS

2019

Analisar a distribuição do banco de dados

Jessica Evelin

7/10/2019

fitditsrplus

O pacote `fitdistrplus` estima os parâmetros dos modelos pela máxima verossimilhança.

```
library(fitdistrplus)
```

```
## Warning: package 'fitdistrplus' was built under R version 3.6.1
```

```
## Loading required package: MASS
```

```
## Loading required package: survival
```

```
## Loading required package: npsurv
```

```
## Loading required package: lsei
```

```
library(actuar)
```

```
## Warning: package 'actuar' was built under R version 3.6.1
```

```
##
## Attaching package: 'actuar'
```

```
## The following object is masked from 'package:grDevices':
##
##     cm
```

O banco de dados escolhido é sobre o custo de seguros de carros.

```
#Carregando dados:
bancodedados <- read.csv("sinistros.csv", header = TRUE, sep=",")

#Visualizando os dados
head(bancodedados)
```

```
##   age sex    bmi steps children smoker region   charges insuranceclaim
## 1  19   0 27.900  3009        0      1      3 16884.924              1
## 2  18   1 33.770  3008        1      0      2  1725.552              1
## 3  28   1 33.000  3009        3      0      2  4449.462              0
## 4  33   1 22.705 10009        0      0      1 21984.471              0
## 5  32   1 28.880  8010        0      0      1  3866.855              1
## 6  31   0 25.740  8005        0      0      2  3756.622              0
```

```
#Criando um vetor apenas com os valores dos sinistros
dados <- bancodedados$charges
```

Estatísticas descritivas

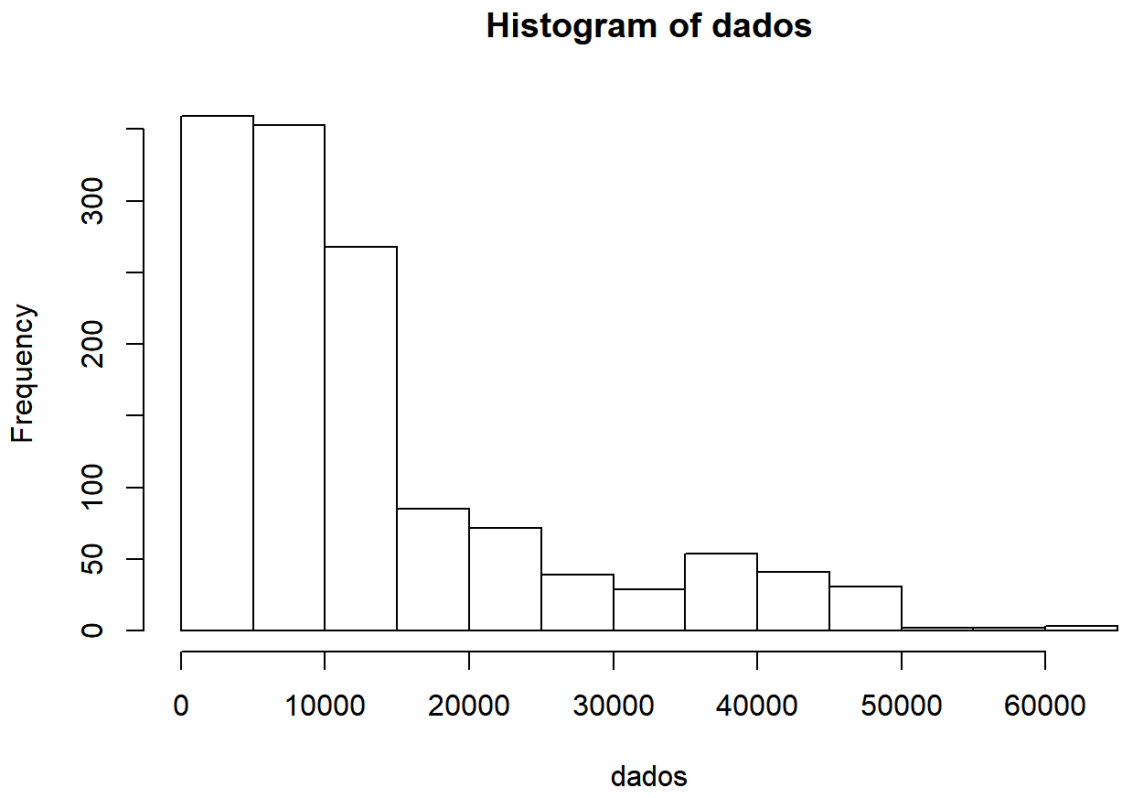
```
#summary
summary(dados)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1122    4740    9382   13270   16640   63770
```

```
#desvio padrão
sd(dados)
```

```
## [1] 12110.01
```

```
# histograma
hist(dados)
```

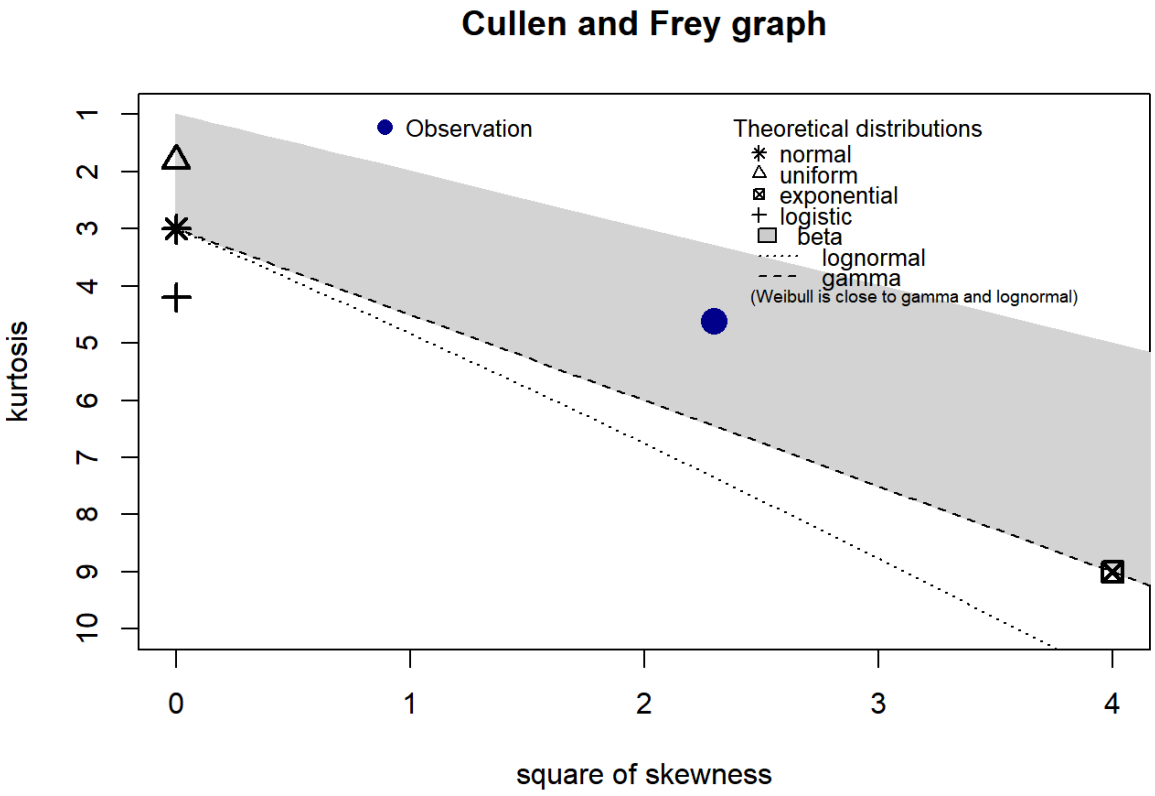


A estatística descritiva dos dados diz que o valor médio é de 13270, a mediana que é o valor central é de 9382 e o valor máximo e minimo são respectivamente 63770 e 1122. Esses valores mostram que os custos de seguros estão concentrados em valores menores. o gráfico aparenta ser de calda pesada.

Ajuste das distribuições

Gráficos

```
descdist(dados)
```




```
## summary statistics
## ------
## min:  1121.874   max:  63770.43
## median:  9382.033
## mean:  13270.42
## estimated sd:  12110.01
## estimated skewness:  1.51588
## estimated kurtosis:  4.606299
```

Analisando o gráfico de Cullen e Frey e desconsiderando beta porque ele é mais usado na teoria Bayesiana a distribuição que está mais próxima da observação é a gamma, porém não podemos tomar isso como uma única verdade é necessário fazer testes mais rigorosos como Kolmogorov.

Vamos testar as distribuições que mostram no gráfico de Cullen and Frey:

```
fitn <- fitdist(dados, "norm", method = "mle") #normal
fitu <- fitdist(dados, "unif", method = "mle") #uniforme
fite <- fitdist(dados, "exp", method = "mme") #exponencial
fitl <- fitdist(dados, "llogis", method = "mle") #logistica
fitln <- fitdist(dados, "lnorm", method = "mle") #lognormal
fitg <- fitdist(dados, "gamma", method = "mme") #gamma
fitw <- fitdist(dados, "weibull", method = "mle") #weibull
```

Fitdist encontra por meio do método dos momentos ou da máxima verossimilhança os parâmetros da distribuição.

Os parâmetros das distribuições acima de forma mais simplificada:

```
Parametros <-
  list(
    "Normal" = fitn$estimate,
    "Uniforme" = fitu$estimate,
    "Exponencial" = fite$estimate,
    "Logística" = fitl$estimate,
    "Lognormal" = fitln$estimate,
    "Gama" = fitg$estimate,
    "Weibull" = fitw$estimate
  )
Parametros
```

```
## $Normal
##     mean       sd
## 13270.42 12105.48
##
## $Uniforme
##       min       max
##  1121.874 63770.428
##
## $Exponencial
##         rate
## 7.535555e-05
##
## $Logística
##        shape        scale
##    0.1887473 8951.1129374
##
## $Lognormal
##   meanlog     sdlog
## 9.0986587 0.9191834
##
## $Gama
##        shape         rate
## 1.201725e+00 9.055665e-05
##
## $Weibull
##        shape        scale
##     1.175642 14099.780657
```

Os parâmetros de forma mais completa:

```
summary(fitn)
```

```
## Fitting of the distribution ' norm ' by maximum likelihood
## Parameters :
##      estimate Std. Error
## mean 13270.42   331.5888
## sd   12105.48   228.8180
## Loglikelihood:  -14477.63   AIC:  28959.26   BIC:  28969.66
## Correlation matrix:
```

```
##      mean sd
## mean    1  0
## sd      0  1
```

```
summary(fitu)
```

```
## Fitting of the distribution ' unif ' by maximum likelihood
## Parameters :
##      estimate Std. Error
## min  1121.874         NA
## max 63770.428         NA
## Loglikelihood:  NA   AIC:  NA   BIC:  NA
## Correlation matrix:
## [1] NA
```

```
summary(fite)
```

```
## Fitting of the distribution ' exp ' by matching moments
## Parameters :
##          estimate
## rate 7.535555e-05
## Loglikelihood:  -14040.03   AIC:  28082.05   BIC:  28087.25
```

```
summary(fitl)
```

```
## Fitting of the distribution ' llogis ' by maximum likelihood
## Parameters :
##           estimate Std. Error
## shape    0.1887473         NA
## scale 8951.1129374         NA
## Loglikelihood:  -16269.81   AIC:  32543.63   BIC:  32554.03
## Correlation matrix:
## [1] NA
```

```
summary(fitln)
```

```
## Fitting of the distribution ' lnorm ' by maximum likelihood
## Parameters :
##          estimate Std. Error
## meanlog 9.0986587 0.02512894
## sdlog   0.9191834 0.01776875
## Loglikelihood:  -13959.79   AIC:  27923.58   BIC:  27933.98
## Correlation matrix:
##         meanlog sdlog
## meanlog       1     0
## sdlog         0     1
```

```
summary(fitg)
```

```
## Fitting of the distribution ' gamma ' by matching moments
## Parameters :
##           estimate
## shape 1.201725e+00
## rate  9.055665e-05
## Loglikelihood:  -14006.09   AIC:  28016.18   BIC:  28026.57
```

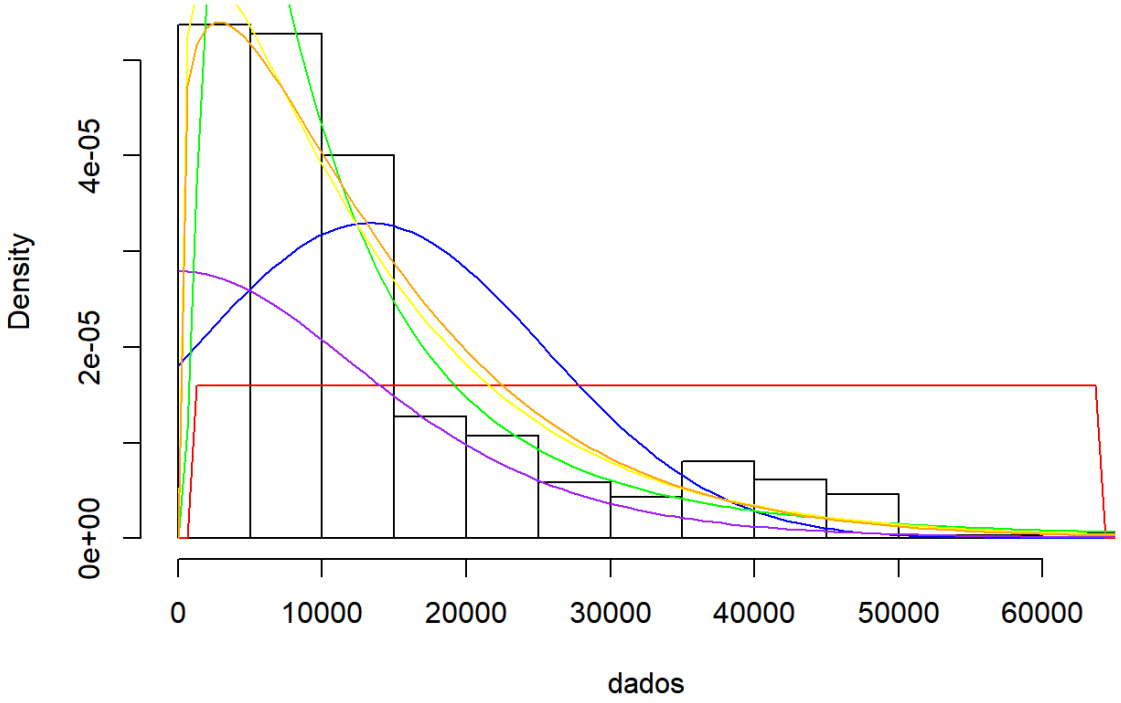
```
summary(fitw)
```

```
## Fitting of the distribution ' weibull ' by maximum likelihood
## Parameters :
##           estimate   Std. Error
## shape     1.175642   0.02417155
## scale 14099.780657 342.32575184
## Loglikelihood:  -14011.66   AIC:  28027.32   BIC:  28037.72
## Correlation matrix:
##           shape     scale
## shape 1.0000000 0.3262273
## scale 0.3262273 1.0000000
```

Plotando o histograma com a curva gerada a partir da estimação:

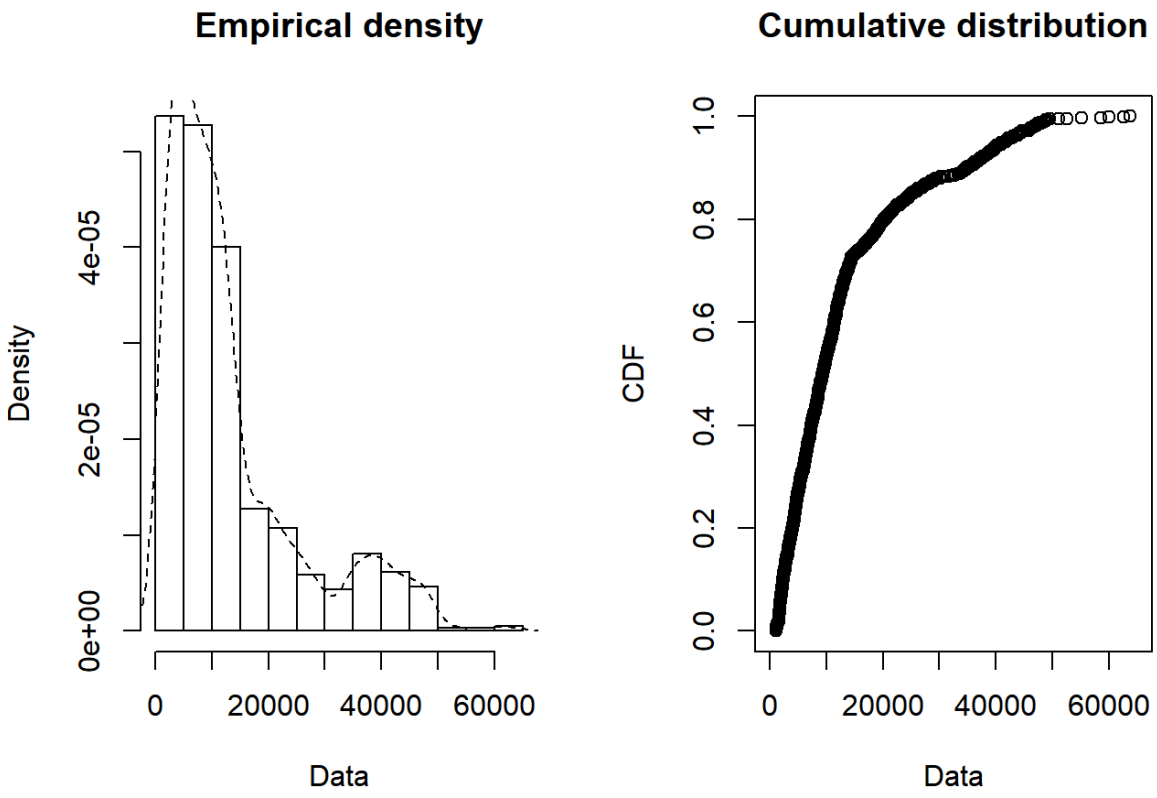
```
hist(
  dados,
  pch = 20,
  breaks = 20,
  prob = TRUE,
  main = ""
)
curve(dnorm(x, fitn$estimate[1], fitn$estimate[2]),
      add = TRUE,
      col = "blue")
curve(dunif(x, fitu$estimate[1], fitu$estimate[2]),
      add = TRUE,
      col = "red")
curve(dexp(x, fite$estimate[1], fite$estimate[2]),
      add = TRUE,
      col = "brown")
curve(dlogis(x, fitl$estimate[1], fitl$estimate[2]),
      add = TRUE,
      col = "purple")
curve(dlnorm(x, fitln$estimate[1], fitln$estimate[2]),
      add = TRUE,
      col = "green")
curve(dgamma(x, fitg$estimate[1], fitg$estimate[2]),
      add = TRUE,
      col = "yellow")
curve(dweibull(x, fitw$estimate[1], fitw$estimate[2]),
```

```
       add = TRUE,
       col = "orange")
```



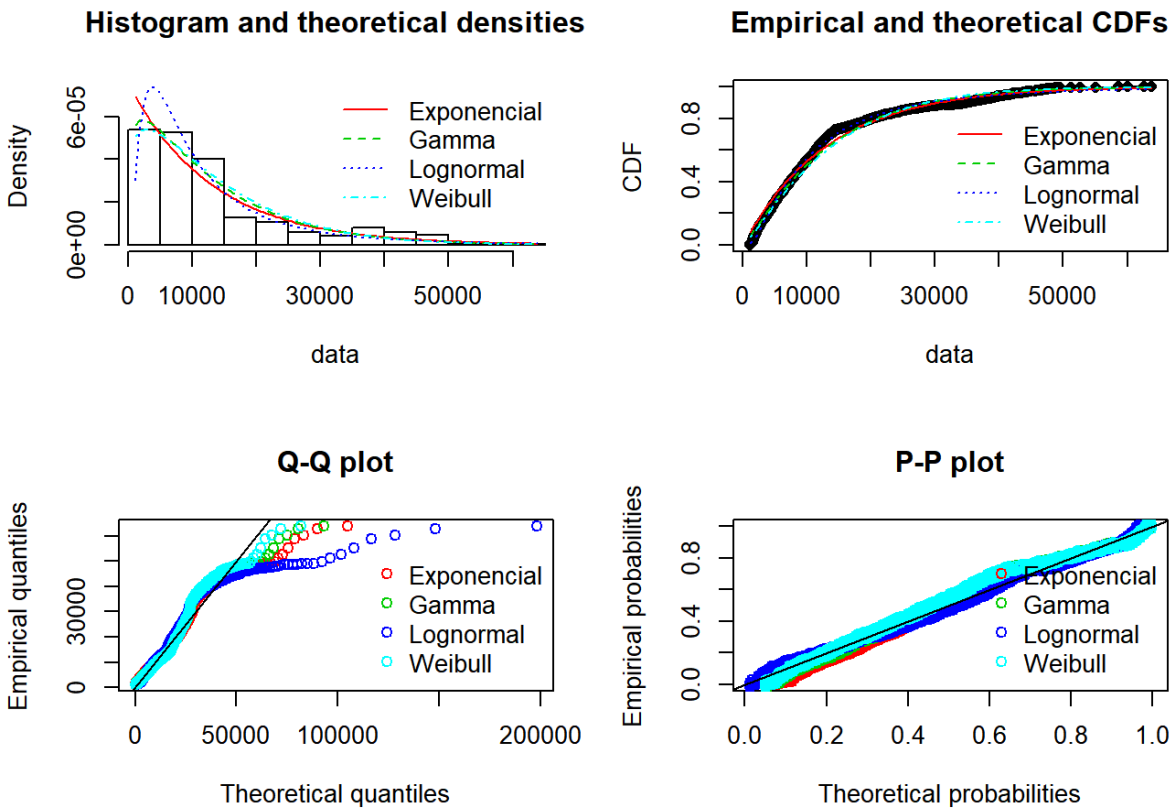
Vamos fazer o histograma e a distribuição acumulada:

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



Esses quatro gráficos comparam as distribuições teóricas com a distribuição empíricas de acordo com as características de cada uma delas. As quatro distribuições mais comuns e que são mais parecidas com a distribuição empírica:

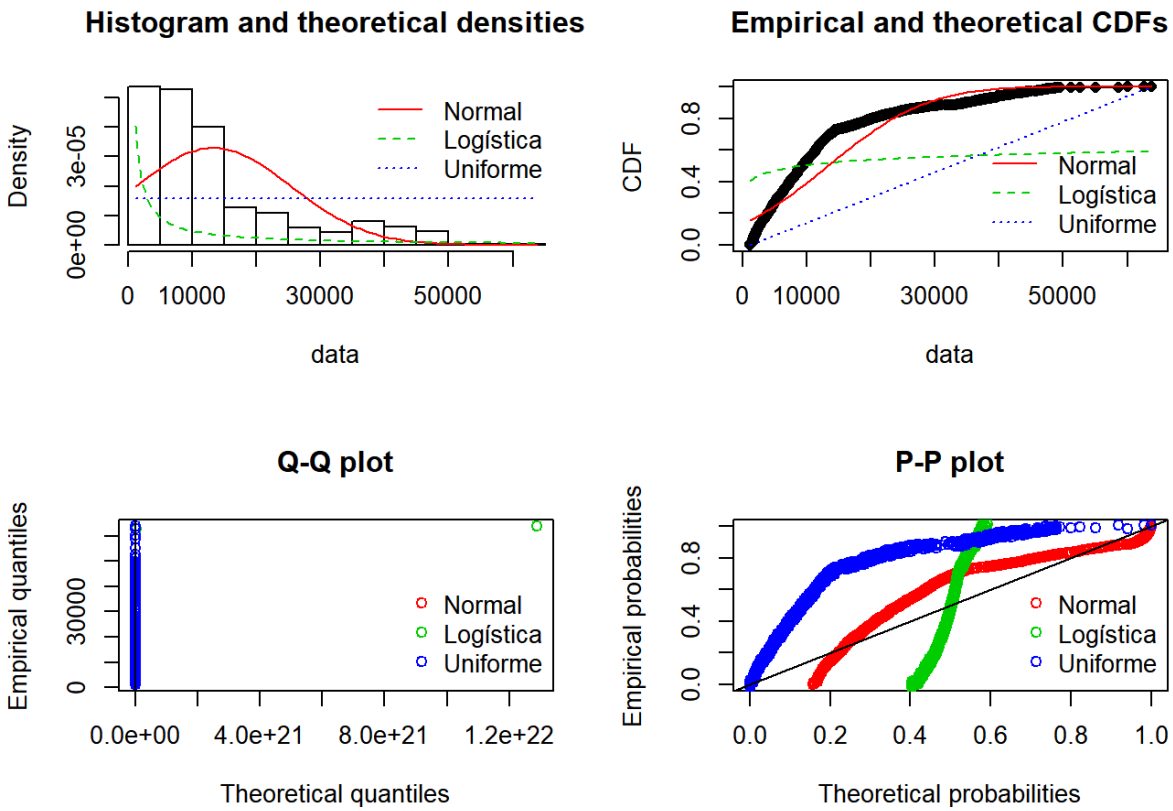
```
par(mfrow = c(2, 2))
plot.legend <- c("Exponencial", "Gamma", "Lognormal", "Weibull")
denscomp(list(fite, fitg, fitln, fitw), legendtext = plot.legend)
cdfcomp (list(fite, fitg, fitln, fitw), legendtext = plot.legend)
qqcomp  (list(fite, fitg, fitln, fitw), legendtext = plot.legend)
ppcomp  (list(fite, fitg, fitln, fitw), legendtext = plot.legend)
```



De acordo com os gráficos acima a distribuição que melhor explica os dados ou que pelo menos se parece com os reais é a distribuição Lognormal seguida da Gamma.

As distribuições que tem menos características parecidas como banco de dados, mas está aqui para comparação visual:

```
par(mfrow = c(2, 2))
plot.legend <- c("Normal", "Logística", "Uniforme")
denscomp(list(fitn, fitl, fitu), legendtext = plot.legend)
cdfcomp (list(fitn, fitl, fitu), legendtext = plot.legend)
qqcomp  (list(fitn, fitl, fitu), legendtext = plot.legend)
ppcomp  (list(fitn, fitl, fitu), legendtext = plot.legend)
```



É fácil ver que nenhuma dessas distribuições é adequada aos dados.

Testes

Os testes que serão analisados é o kolmogorov e o Akaike para esses quanto menor o valor melhor, pois eles medem a diferença numérica entre a distribuição empírica e teórica.

```
gofstat(
  list(fitn, fitu, fitln, fitw, fite, fitl, fitg),
  fitnames = c(
    "Normal",
    "Uniforme",
    "Lognormal",
    "Weibull",
    "Exponencial",
    "Logística",
    "Gamma"
  )
)
```

```
## Goodness-of-fit statistics
##                                 Normal    Uniforme Lognormal     Weibul
l
## Kolmogorov-Smirnov statistic  0.188462   0.5147556 0.0365844  0.0843750
4
## Cramer-von Mises statistic   14.829729 155.5524079 0.3973136  1.9537467
1
```

```
## Anderson-Darling statistic   85.138872         Inf 3.9424972 13.5818358
9
##                              Exponencial  Logística       Gamma
## Kolmogorov-Smirnov statistic  0.09614826   0.414691  0.07157911
## Cramer-von Mises statistic    2.41235060  81.055959  1.30005636
## Anderson-Darling statistic   19.59919126 394.450454 10.98410082
##
## Goodness-of-fit criteria
##                                  Normal Uniforme Lognormal  Weibull
## Akaike's Information Criterion 28959.26       NA  27923.58 28027.32
## Bayesian Information Criterion 28969.66       NA  27933.98 28037.72
##                                Exponencial Logística    Gamma
## Akaike's Information Criterion    28082.05  32543.63 28016.18
## Bayesian Information Criterion    28087.25  32554.03 28026.57
```

De acordo com os testes Kolgomorov e Akaike o melhor é Lognormal (0.0365844, 27923.58) e em segundo Gamma (0.07157911, 28016.18), pois essa distribuições teóricas tem menor diferenças da distribuição real dos dados.