Trabalho 1 - Introdução ao Aprendizado Estatístico

Rodrigo Malta Esteves

24/02/2021

1. Etapa 1
    1. Custos de cancelamento na escala original.
    2. Justificando
2. Etapa 2
    1. Obtenha estimativas de máxima verossimilhança
    2. Probabilidade de cancelamento
3. Etapa 3
    1. Grade de parâmetros
    2. Gráfico da função de verossimilhança normal
    3. Comente o gráfico obtido
4. Encontrando os estimadores de forma analítica.

Etapa 1: Especificação de um modelo observacional

Visualização das primeiras linhas do arquivo e a dimensão do arquivo.

```
custos_cancel <- read.table(
  "C:\\Users\\malta\\Desktop\\Pós Graduação\\Introdução ao Aprendizado Estatístico\\Trabalho 1\\Rodrigo
names(custos_cancel) <- c("custos","log_custos")
head(custos_cancel) #Visualização das primeiras linhas do arquivo lido
```

```
##     custos log_custos
## 1 6.846861   1.923790
## 2 6.947157   1.938333
## 3 5.788298   1.755838
## 4 6.927557   1.935507
## 5 9.035326   2.201142
## 6 8.192294   2.103194
```

```
dim(custos_cancel) #Conferindo a dimensão do arquivo
```

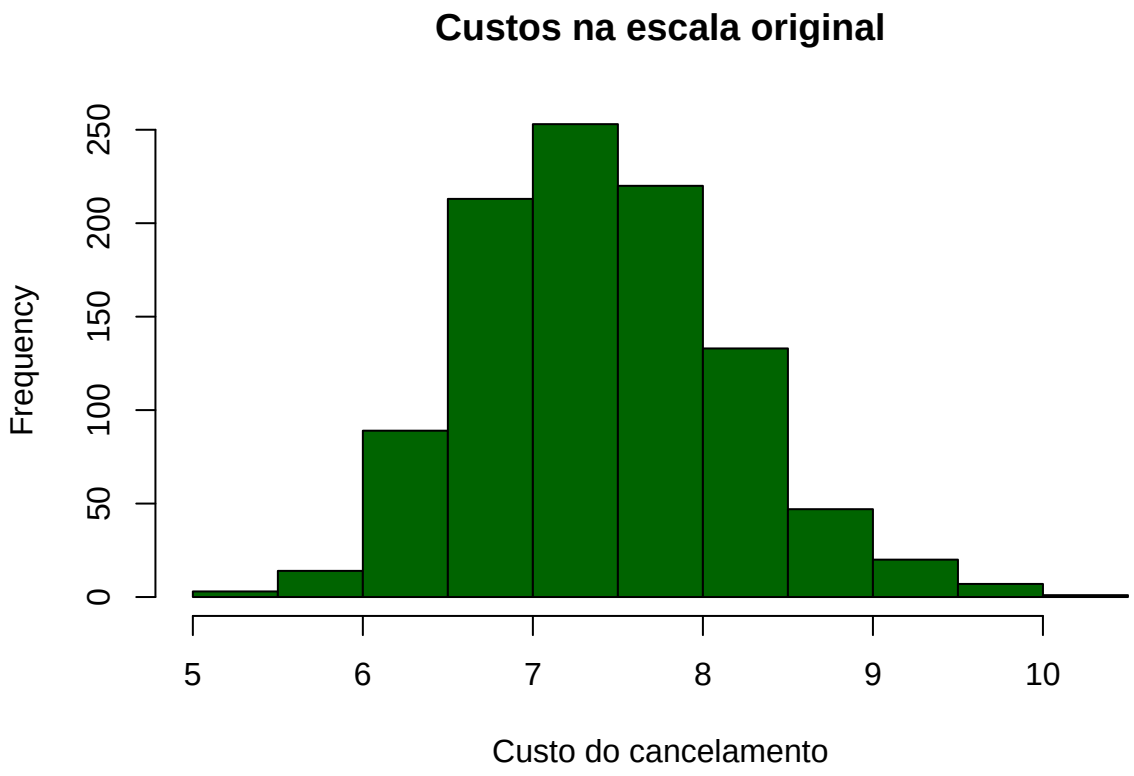
```
## [1] 1000    2
```

```
class(custos_cancel)
```

```
## [1] "data.frame"
```

a) Histograma dos custos de cancelamento na escala original.

```r
hist(custos_cancel$custos,col="darkgreen",
     main="Custos na escala original",
     xlab="Custo do cancelamento")
```



b) Justificando

O espaço amostral da distribuição Normal equivale a todo conjunto dos Reais, ou seja, se Y é uma variável aleatória tal que $Y \sim N(\mu, \sigma^2)$, temos que $Y \in (-\infty, \infty)$. Por outro lado, a transformação logarítmica dessa variável aleatória $\ln(Y) \sim N(\mu, \sigma^2)$ tem espaço amostral apenas nos reais positivos $\ln(Y) \in (0, \infty)$.

Como no caso dos custos de cancelamento de contrato teremos apenas valores positivos, faz mais sentido usar a distribuição Lognormal dado que seu espaço amostral compreende todos os valores possíveis.

Um outro modelo probabilístico possível é a Gama, uma vez que os dados são sempre positivos e apresenta uma pequena assimetria a direita. Estimamos os parâmetros utilizando as funções de log-verossimilhança e comparamos com os dados.

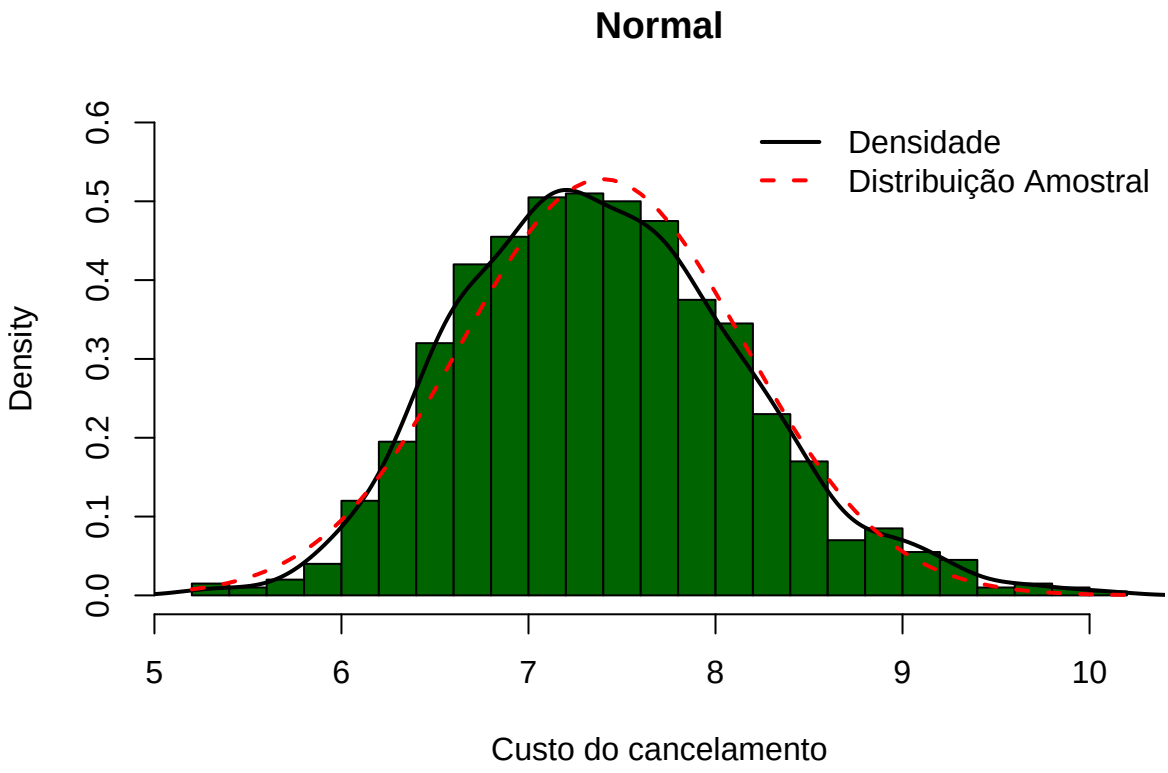
Nos gráficos a seguir podemos ver que tanto Gama quanto a Normal dão suporte aos dados da primeira coluna de forma semelhante e a Lognormal aos dados da segunda coluna.

```r
n <- length(custos_cancel$custos)
a <- NULL
b <- NULL
# -logverossimilhança
theta<-c(a,b)
```

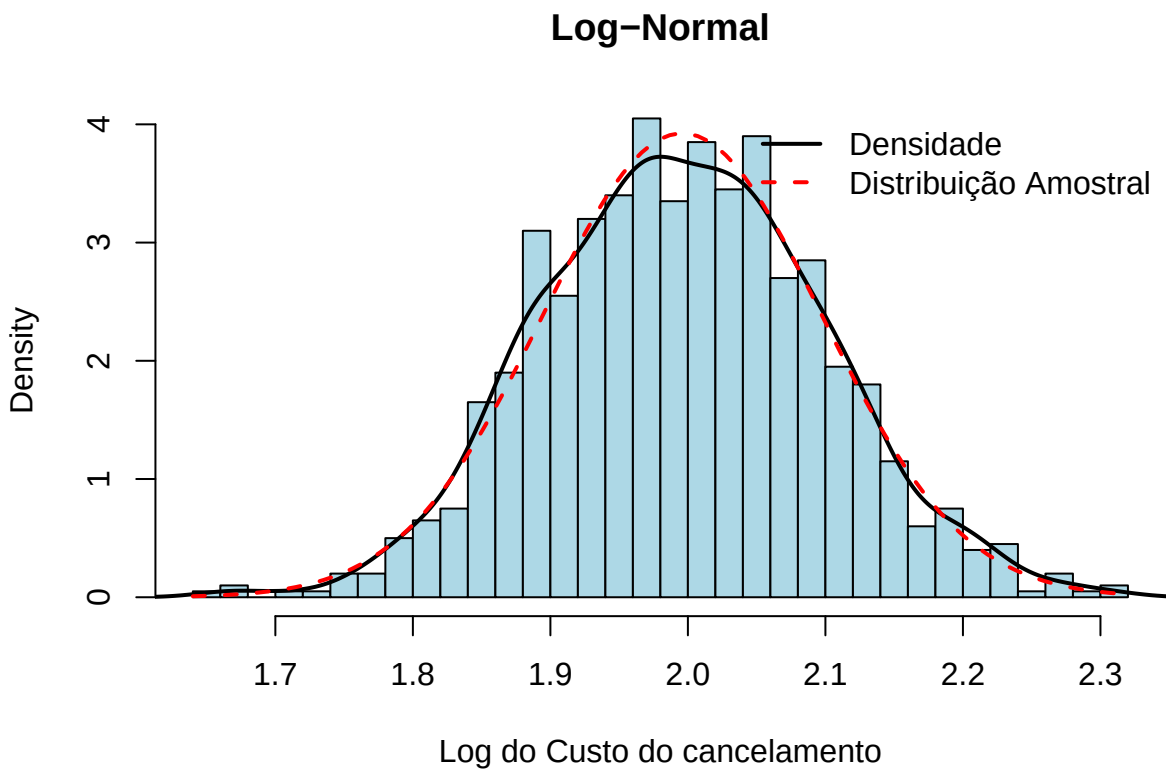
```r
neglogvero<-function(theta){-sum(log(dgamma(custos_cancel$custos,theta[1],theta[2])))}
#resultado:
saida<-nlm(neglogvero,p=c(0.5,0.5))
saida$estimate #estimativas de maxima verossimilhanca
```

```
## [1] 96.87731 13.09618
```

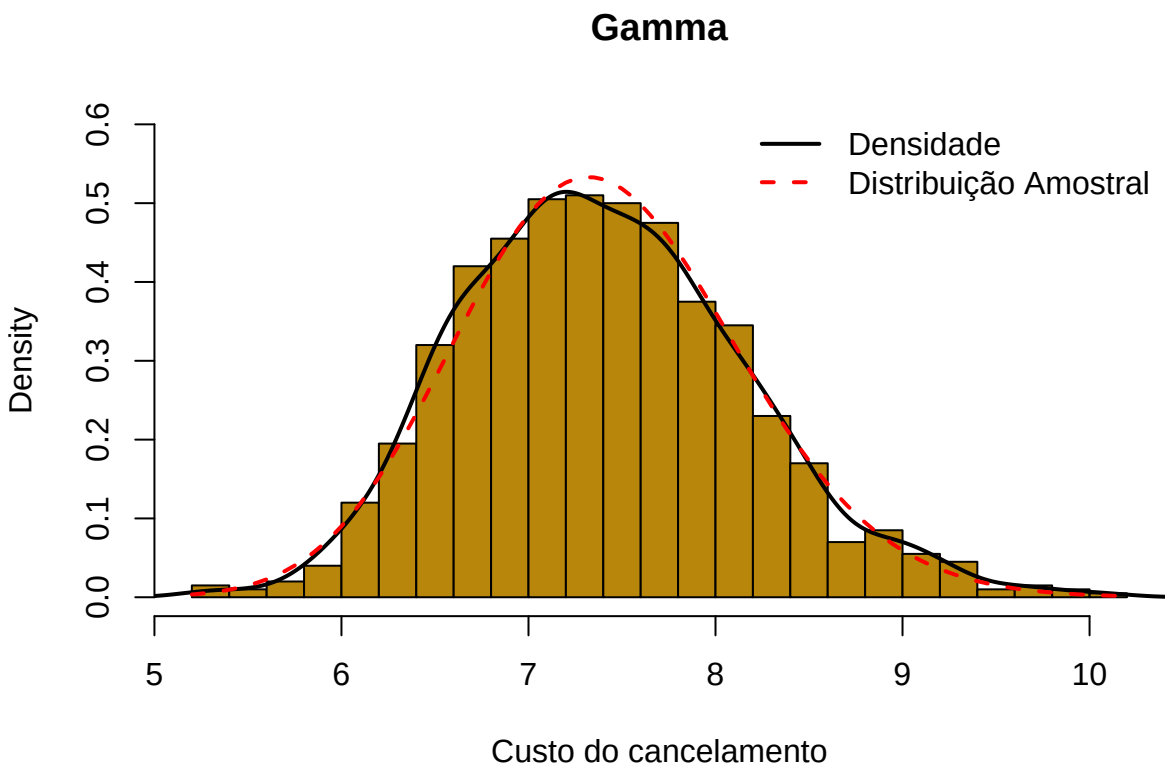
```r
hist(custos_cancel$custos,col="darkgreen",prob=TRUE,ylim=c(0,.6),
     main="Normal",xlab="Custo do cancelamento",breaks=25)
lines(density(custos_cancel$custos),lwd=2)
curve(dnorm(x,mean(custos_cancel$custos),sd(custos_cancel$custos)),
      col=2,lty=2,lwd=2,add=TRUE)
legend("topright",legend=c("Densidade","Distribuição Amostral"),lty=c(1,2),lwd=2,
       col=c(1,2),bty="n")
```




```r
hist(custos_cancel$log_custos,col="lightblue",prob=TRUE,ylim=c(0,4),
     main="Log-Normal",xlab="Log do Custo do cancelamento",breaks=25)
lines(density(custos_cancel$log_custos),lwd=2)
curve(dnorm(x,mean(custos_cancel$log_custos),sd(custos_cancel$log_custos)),
      col=2,lty=2,lwd=2,add=TRUE)
legend("topright",legend=c("Densidade","Distribuição Amostral"),lty=c(1,2),lwd=2,
       col=c(1,2),bty="n")
```




```
hist(custos_cancel$custos,prob=TRUE,ylim=c(0,.6),
     main="Gamma",xlab="Custo do cancelamento",col="darkgoldenrod",breaks=25)
lines(density(custos_cancel$custos),lwd=2)
curve(dgamma(x,96.87731,13.09618),col=2,lty=2,lwd=2,add=TRUE)
legend("topright",legend=c("Densidade","Distribuição Amostral"),lty=c(1,2),lwd=2,
       col=c(1,2),bty="n")   # legenda
```

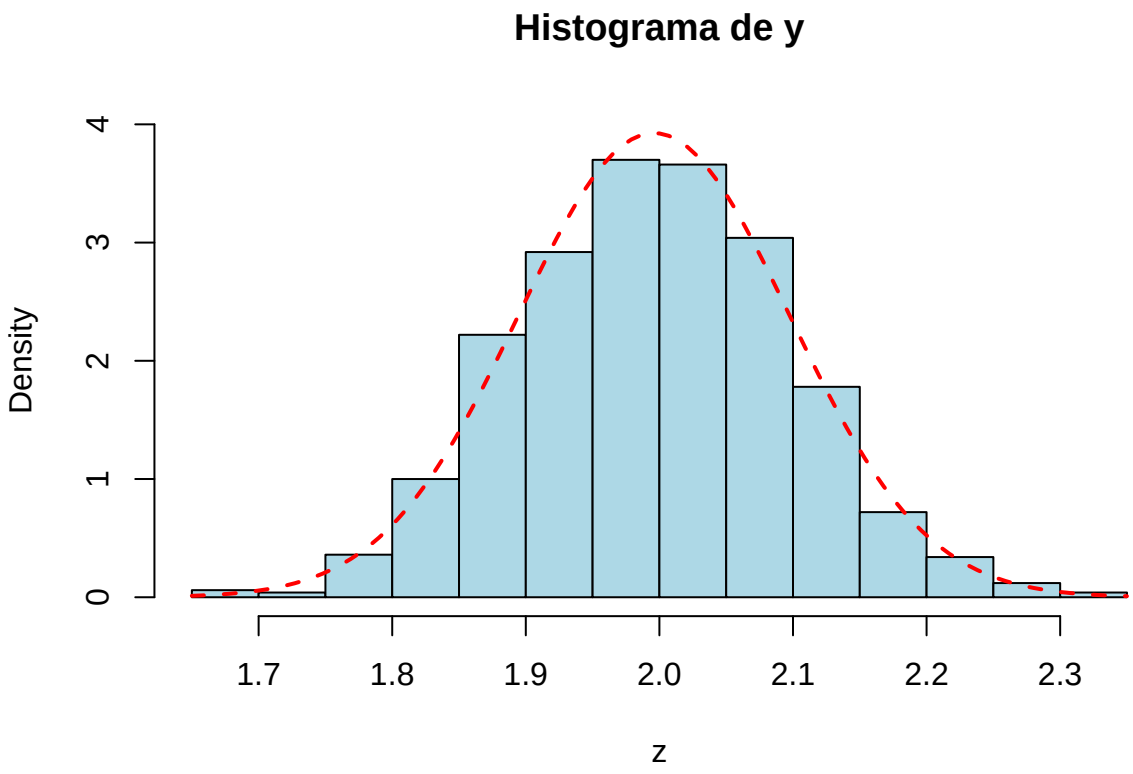


Etapa 2: Obtenção numérica de estimas de máxima verossimilhança e seu uso para alimentar o modelo probabilístico

c) Obtenha estimativas de máxima verossimilhança para μ e σ^2

As estimativas de máxima verossumilhança para os parâmetros são alpha = 1.9959550 e beta = 0.1016071.

```
# Estimativa de max. veros para o log dos dados
#Entrada de dados
z <- custos_cancel$log_custos
#Visualização da distribuição da amostra observada
hist(z, main="Histograma de y",right=FALSE, prob=TRUE,ylim=c(0,4),col="lightblue")
curve(dnorm(x,1.9959550,0.1016071),col=2,lty=2,lwd=2,add=TRUE)
```




```
#tamanho amostral
n <- length(z)
a <- NULL
b <- NULL
# -logverossimilhança
theta<-c(a,b)
neglogvero<-function(theta){-sum(log(dnorm(z,theta[1],theta[2])))}
#resultado:
saida<-nlm(neglogvero,p=c(1,1))
saida$estimate #estimativas de maxima verossimilhanca
```

```
## [1] 1.9959550 0.1016071
```

d) Probabilidade de cancelamento futuro superior a 9 mil reais.

A probabilidade do custo de cancelamento de contrato futuro nessa empresa ser superior a 9 mil Reais é de 0.02380341.

```
1 - pnorm(log(9),1.9959550,0.1016071) # 1 - P(X<=9)
```

```
## [1] 0.02380341
```

Etapa 3: Estudo visual do comportamento da função de verossimilhança

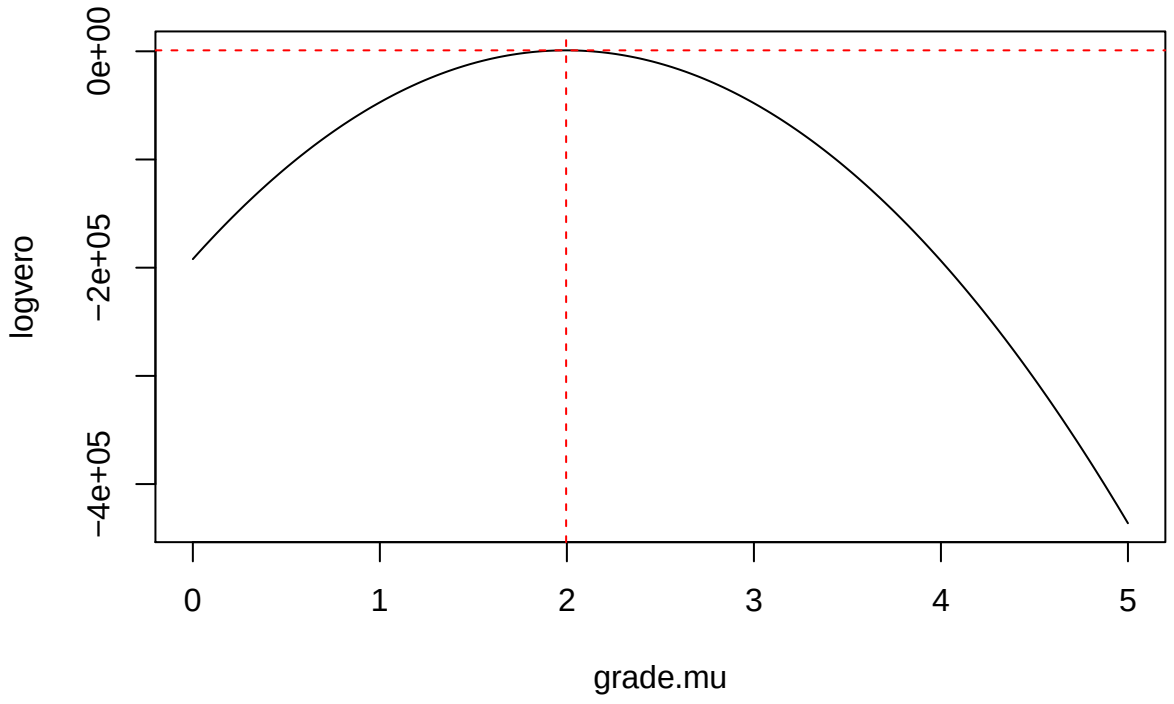
e) Grade de parâmetros

```
grade.mu <- seq(0,5,0.01)
```

f) Gráfico da função de verossimilhança normal

```
#Sigma estimado na Etapa 2
dp_est <- saida$estimate[2]
var_est <- (saida$estimate[2])^2
#Tamanho da grade de parâmetros
n = length(z)
a <- NULL
b <- NULL
#Estatísticas suficientes
t1 = sum(z)
t2 = sum(z^2)
#Verossimilhança
logvero <- (-n/2)*log(var_est*2*pi) - (t2 - 2*grade.mu*t1 + n*(grade.mu^2))/(2*var_est)

#Gráfico
plot(grade.mu,logvero,type="l")
abline(v=(saida$estimate[1]),lty=2,col="red")
abline(h=max(logvero),lty=2,col="red")
```



g) Comente o gráfico obtido

O gráfico apresenta no eixo X os valores da grade para o parâmetro μ e no eixo Y os valores da função de log-verossimilhança para essa grade. A linha vertical em vermelho indica o Estimador de Máxima Verossimilhança para μ encontrado na Etapa 2 e a linha horizontal em vermelho indica o valor máximo entre os valores da log-verossimilhança obtida para os valores da grade. Podemos perceber que, de fato, o estimador representa o máximo da verossimilhança uma vez que intecepta a curva ponto mais alto.

Encontrando os estimadores de forma analítica.

Se considerarmos X uma v.a. tal que $X \sim Normal(\mu, \sigma^2)$, temos que a sua Função de Densidade de Probabilidade tem a forma: $f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e(\frac{-(X-u)^2}{2\sigma^2})$. Considerando $X_i, i = 1, 2, 3, ...n$, podemos encontrar sua Função de Verossimilhança fazendo $l(\mu; \underline{x}) = \prod_{i=1}^{n} f(x_i|\mu) = (\frac{1}{\sqrt{2\pi\sigma^2}})^{\frac{-n}{2}} e(\frac{-\sum_{i=1}^{n}(x-u)^2}{2\sigma^2})$.

Calculando a Logverossimilhança: $L(\mu; \underline{x}) = ln(l(\mu; \underline{x})) = \frac{-n}{2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n}(x_i - \mu)^2$.

Se derivarmos essa função e igualarmos a zero, obteremos os Estimadores de Máxima Verossimilhança para μ e σ.

$\frac{d(L(\mu;\underline{x}))}{d(\mu)} = -(\frac{1}{2\sigma^2})(2)(-1) \sum_{i=1}^{n}(x_i - \mu) = \frac{1}{\sigma^2}(-n\mu + \sum_{i=1}^{n} x_i) = 0 \implies \sum_{i=1}^{n} x_i = n\mu \implies \hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}$.

$\frac{d(L(\sigma;\underline{x}))}{d(\sigma)} = \frac{-n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n}(x_i - \mu)^2 = 0 \implies \sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n} \implies \hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}}$.