Modelos Estatísticos com Parâmetros Constantes

ou: como ajustar distribuições de probabilidades aos seus dados

Paulo Inácio Prado e João L.F. Batista

BIE5781 - Pós-Graduação em Ecologia USP

novembro de 2021

Objetivo da Aula

Os objetivos dessa aula são:

- Mostrar o ciclo de trabalho para ajuste de modelos probabilísticos com parâmetros constantes.
- Utilizar o método da Máxima Verossimilhança para obter estimativas dos parâmetros.
- Realizar inferências com base na curva/superfície de verossimilhança.
- Apresentar as ferramentas básicas para avaliação do ajuste.

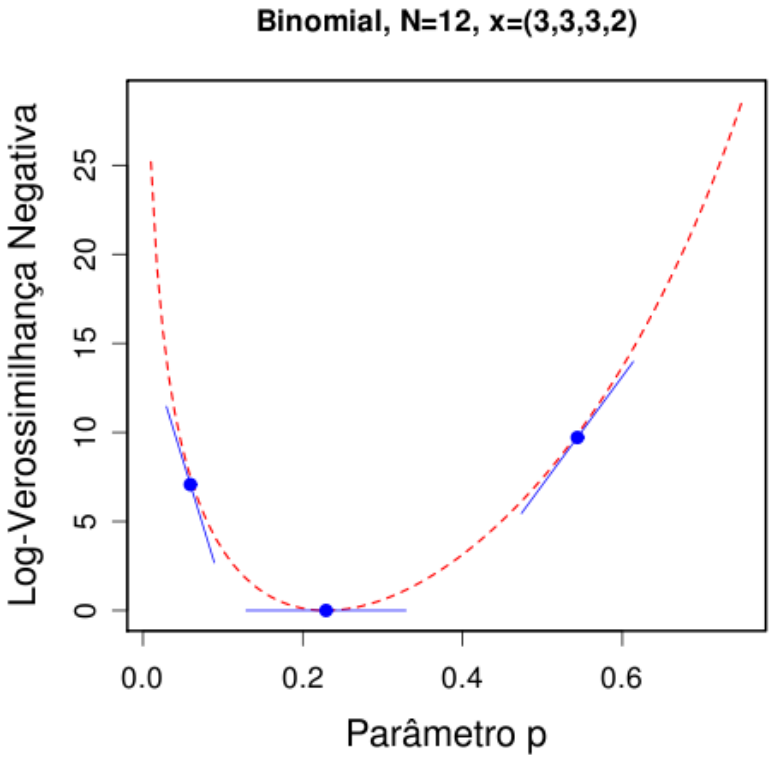
os três passos para o ajuste

1. Escolha o modelo
2. Encontre a função de log-verossimilhança negativa
3. Encontre o mínimo da função de log-verossimilhança negativa: MLEs

Caso raro: solução analítica

1. Escolha o modelo
$$X \sim f(x \mid \Theta)$$
2. Encontre a função de log-verossimilhança negativa
$$\mathbf{L}\{\Theta \mid x_i\} = -\sum_{i}^{N} \ln f(x_i \mid \Theta)$$
3. Encontre o mínimo da log-verossimilhança negativa
$$\frac{\partial \mathbf{L}\{\Theta \mid x_i\}}{\partial \Theta} = 0$$

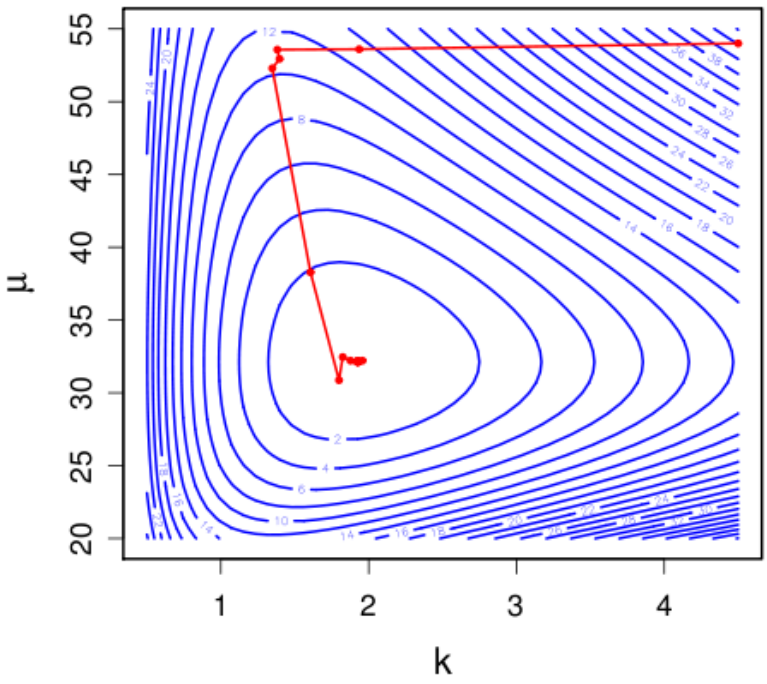
Soluções analíticas: derivadas



A regra: soluções numéricas

1. Escolha o modelo
   - `dgeom`, `dexp`, `dnbinom`, `dweibull`, ...
2. Encontre a função de log-verossimilhança negativa
   - `LG <- function(theta) -sum (dgeom(dados, prob=theta, log=T))`
   - `LDB <- function(t1,t2) -sum (dnbinom(dados, mu=t1, size=t2, log=T))`
   - ...
3. Encontre o mínimo da log-verossimilhança negativa
   - `m1 <- mle2(LG, start=list(theta1=0.5))`
   - ...

Soluções numéricas: otimização computacional



Um Exemplo Simples: Distribuição Exponencial

A inferência por Verossimilhança na distribuição exponencial é simples porque:

- possui apenas um parâmetro,
- a função de verossimilhança é simples,
- a MLE pode ser obtida analiticamente.

Distribuição Exponencial: O Modelo

- Função de Densidade Probabilística:

$$f(x) = \lambda\, e^{-\lambda x}$$

- Função de Verossimilhança

$$\mathcal{L}\{\lambda\} = \prod_{i=1}^{n} f(\lambda | x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \lambda\, e^{-\lambda x_i}$$

- Função de Log-Verossimilhança

$$\mathbf{L}\{\lambda\} = \ln \left[\prod_{i=1}^{n} \lambda\, e^{-\lambda x_i} \right]$$

Distribuição Exponencial: O Modelo

- Função de Log-Verossimilhança

$$\begin{aligned} \mathbf{L}\{\lambda\} &= \ln \left[\prod_{i=1}^{n} \lambda\, e^{-\lambda x_i} \right] = \sum_{i=1}^{n} \ln \left[\lambda\, e^{-\lambda x_i} \right] \\ &= \sum_{i=1}^{n} \ln \lambda - \sum_{i=1}^{n} \lambda x_i \\ &= n\, \ln(\lambda) - \lambda \sum_{i=1}^{n} x_i \end{aligned}$$

Distribuição Exponencial: Estimativa

- Primeira Derivada da Função de Log-Verossimilhança:

$$\frac{\delta \mathbf{L}\{\lambda\}}{\delta \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$$

- Estimativa de Máxima Verossimilhança (MLE):

$$\frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0 \qquad \Longrightarrow \qquad \widehat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}$$

MLEs: soluções analíticas mais conhecidas

Exponencial

$$\hat{\lambda} = \bar{x}^{-1} = \frac{n}{\sum x_i}$$

Poisson

$$\hat{\lambda} = \bar{x} = \frac{\sum x_i}{n}$$

Normal

$$\hat{\mu} = \bar{x} = \frac{\sum x_i}{n}$$

$$\widehat{\sigma^2} = s^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

Um exemplo simulado no R

Dados

```
> x <- rexp(100, rate=0.1)
```

MLE analítico conhecido: basta calcular

```
> (L1 <- 1/mean(x))

[1] 0.09446453
```

Conferindo valor analítico no R

Log-verossimilhança analítica

```
> nllexp <- function(lambda){
+     N <- length(x)
+     sx <- sum(x)
+    -(N*log(lambda) - lambda*sx)
+ }
```

Calculo analítico da log-verossimilhança negativa mínima

```
> nllexp(lambda = L1)

[1] 335.9531
```

Ajuste numérico no R

Log-verossimilhança como função da densidade

```
> LL.e2 <- function(lambda){
+     -sum(dexp(x, rate=lambda, log=TRUE))
+ }
```

Minimização numérica com a função mle2

```
> library(bbmle) #basta uma vez

Loading required package:  stats4

> m1 <- mle2( LL.e2, start=list(lambda=L1) )
```

Comparação das duas soluções

Analítica

```
> L1

[1] 0.09446453

> nllexp(lambda=L1)

[1] 335.9531
```

Numérica

```
> coef(m1)

    lambda
0.09446453

> logLik(m1)

'log Lik.' -335.9531 (df=1)
```

Um Exemplo Complexo Contínuo: Distribuição Weibull

- A dist. Weibull tem dois parâmetros:

$$\text{Escala:} \quad \beta \qquad \text{Forma:} \quad \gamma$$

- A função de log-verossimilhança é complexa:

$$\mathbf{L}\{\beta, \gamma\} \quad = \quad \sum_{i=1}^{n} \ln\left[(\gamma/\beta)\,(x/\beta - 1)^{\gamma} \exp\left(-(x/\beta)^{\gamma}\right)\right]$$

- De fato, é complexa:

$$\begin{aligned} \mathbf{L}\{\beta, \gamma\} \quad = \quad & n\,\ln(\gamma) - n\,\ln(\beta) + (\gamma - 1)\sum_{i=1}^{n} \ln(x_i) \\ & -n(\gamma - 1)\ln(\beta) - \sum_{i=1}^{n} (x_i/\beta)^{\gamma} \end{aligned}$$

Um Exemplo Complexo Contínuo: Distribuição Weibull

- A MLE da escala depende da forma:

$$\widehat{\beta} \quad = \quad \left(\sum_{i=1}^{n} \frac{x_i^{\gamma}}{n}\right)^{\gamma}$$

- A MLE da forma depende da escala:

$$\widehat{\gamma} \quad = \quad \frac{n\,\beta}{\sum_{i=1}^{n} x_i^{\gamma}\,\ln(x_i) - \sum_{i=1}^{n} \ln(x_i)}$$

- Somente a *solução numérica* é possível.

Exemplo de ajuste da Weibull no R: dados

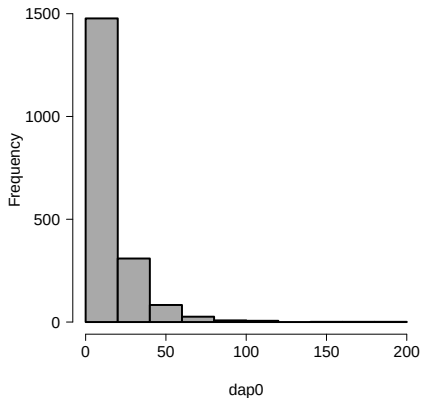
- Conjunto de árvores com mais de DAP $\geq$ 25 cm, Paragominas, PA.
- DAP = diâmetro à altura do peito.
- Exemplo:

  dap
  33.7
  44.6
  83.1
  34.7
  27.6
  ...
- Variável Resposta: X = DAP - 25 (cm)
- Tamanho da amostra: $n = 1912$.

Exemplo de ajuste da Weibull no R: dados

Comandos

```
> parag <- read.table(
+ "parago-sobrev.csv", header=T)
> dap <- parag$dap[parag$dap>25]
> dap0 <- dap - 25
> hist(dap0, main="")
```



Exemplo de ajuste da Weibull no R: otimização

Log-verossimilhança como função da densidade

```
> nllweibull = function(escala, forma){
+   -sum( dweibull(dap0, shape=forma, scale=escala,
+                  log=TRUE) )
+ }
```

Minimização numérica com a função mle2

```
> parag.wei = mle2(nllweibull,
+   start=list(escala=20, forma=1))
```

Exemplo de ajuste da Weibull no R: resultados

MLEs

```
> coef(parag.wei)

    escala      forma
13.4946178  0.9161387
```

Log-verossimilhança

```
> logLik(parag.wei)

'log Lik.' -6958.119 (df=2)
```

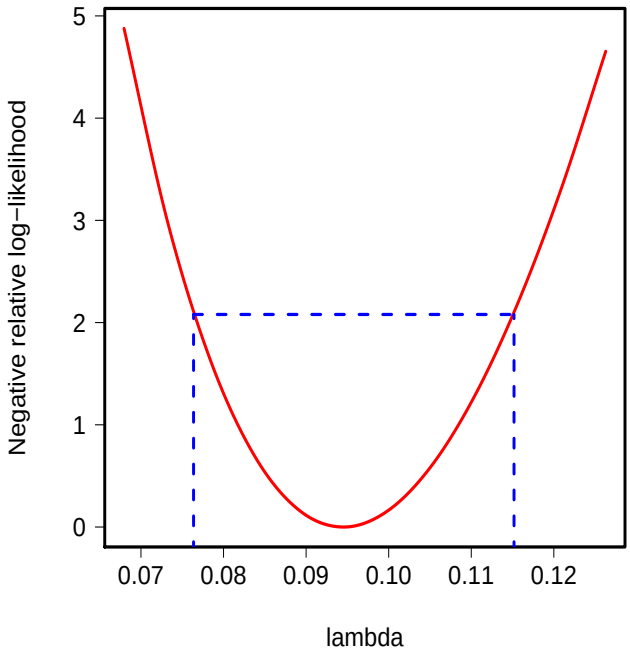
Diagnósticos: qual a qualidade de seu modelo?

- A otimização é confiável?
  - Perfis de verossimilhança
- As estimativas dos parâmetros (MLEs) são boas?
  - Intervalos de plausibilidade
  - Intervalos de confiança (aproximação)
- O modelo descreve bem os dados?
  - Gráficos de valores observados e previstos
  - Gráficos quantil-quantil
- Há algum modelo melhor?
  - Seleção de modelos

Exemplo de curva de verossimilhança: Distribuição Exponencial

Comandos

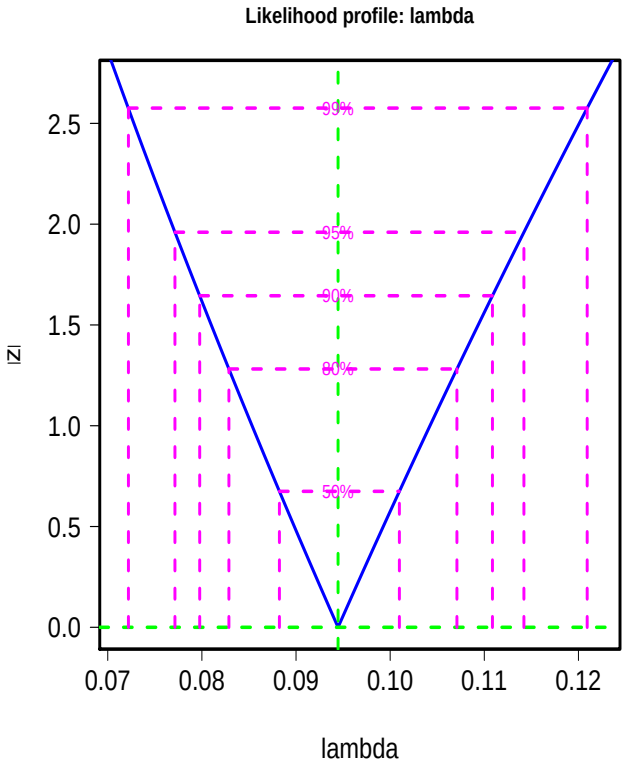
```
> library(sads)
> m1.p <- profile(m1)
> plotprofmle(m1.p)
```



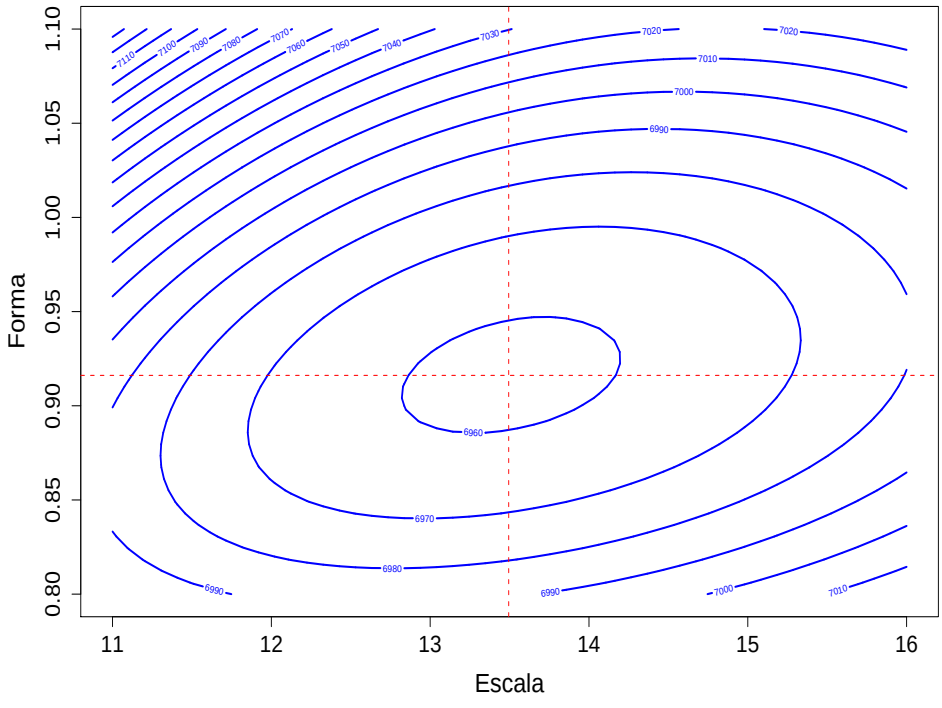
Exemplo de perfil normalizado: Distribuição Exponencial

Comandos

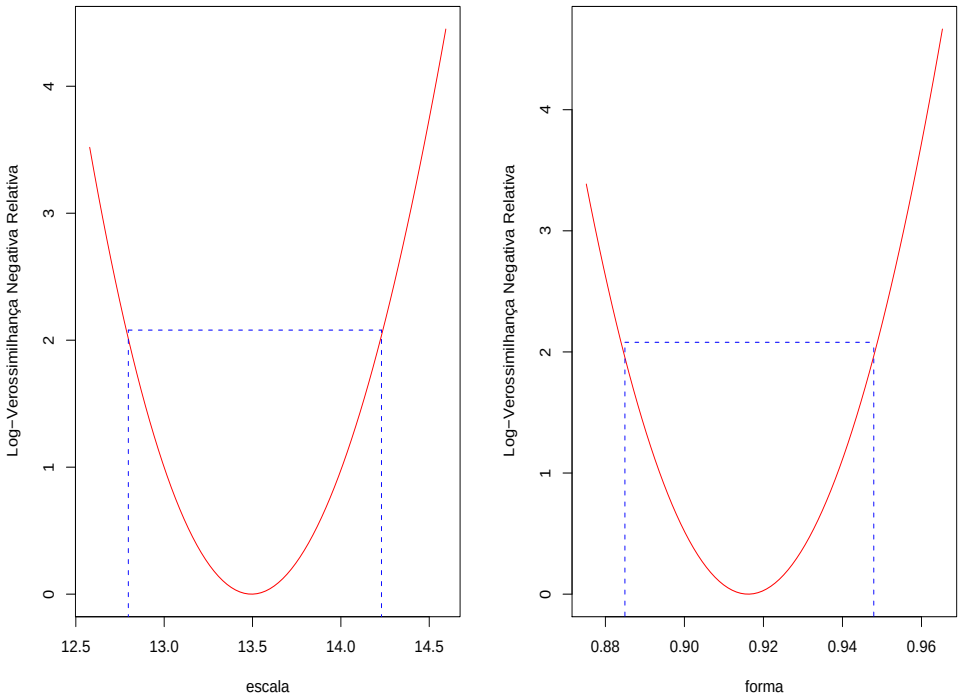
```
> plot(m1.p)
> confint(m1.p)
```



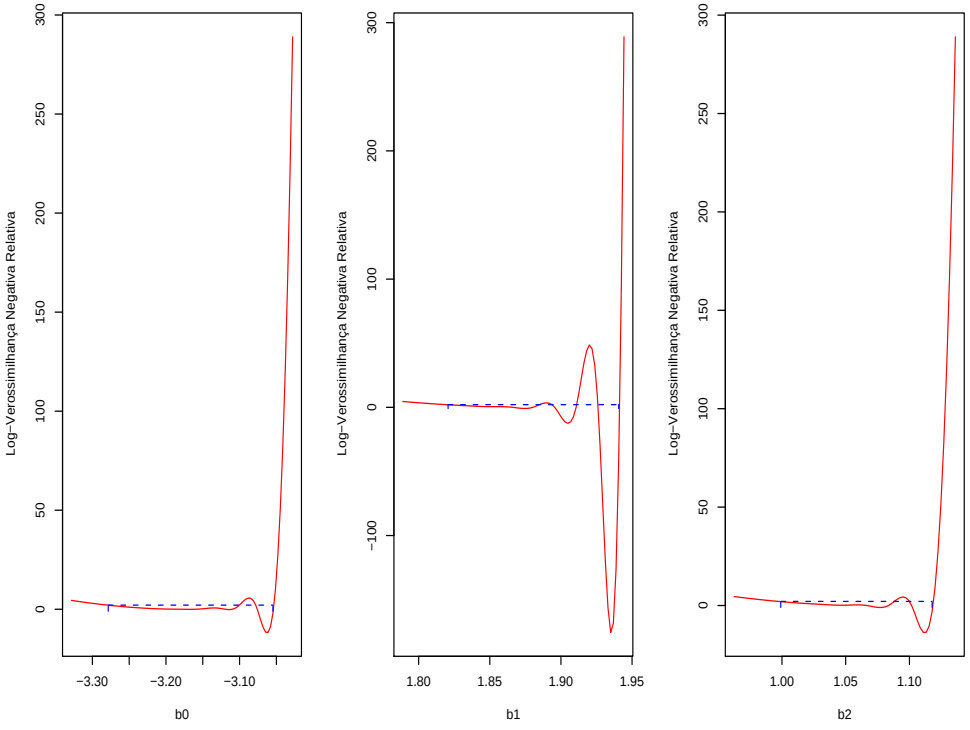
Exemplo de superfície de verossimilhança: Weibull



Exemplo de perfil de verossimilhança: Weibull

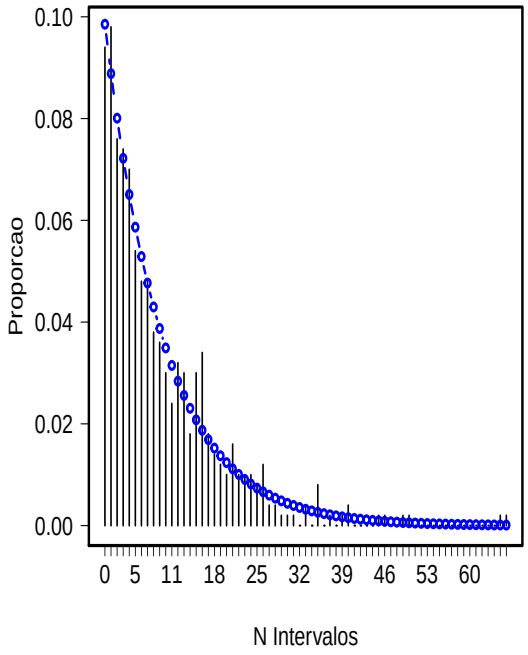


Exemplo de perfis preocupantes

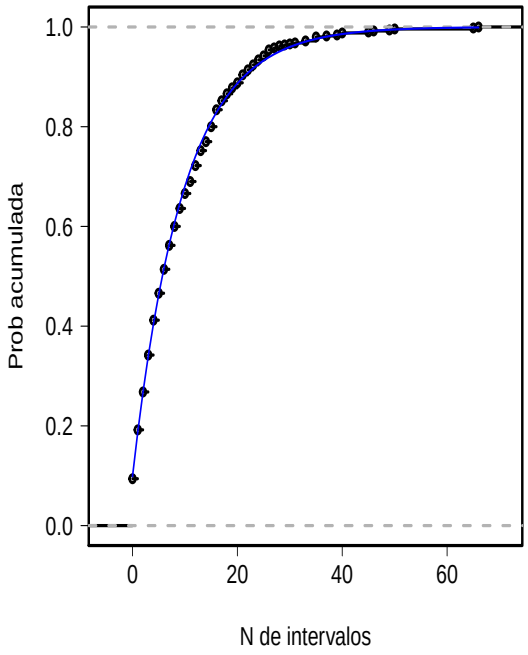


Distribuição Geométrica: previstos e observados

Probabilidades



Prob. acumuladas



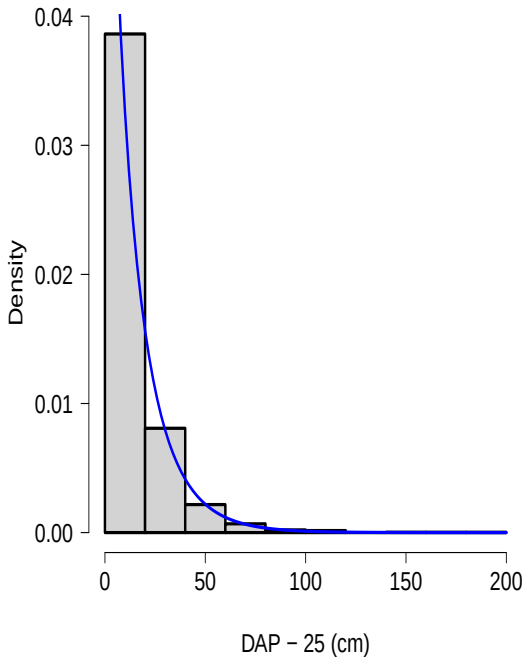
Distribuição Geométrica: previstos e observados

Comandos

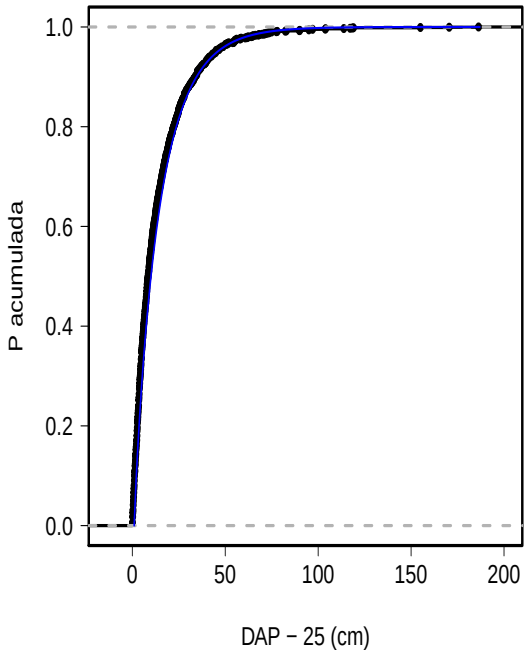
```
> ## Gera dados para um exemplo
> x2 <- rgeom(500, prob = 0.1)
> ## mle analitico
> p1 <- length(x2)/(length(x2)+sum(x2))
> ## Para construir o grafico
> obs <- table(factor(x2,levels=0:max(x2)))
> plot(obs/sum(obs), xlab="N Intervalos",
+      ylab="Proporcao")
> lines(0:max(x2), dgeom(0:max(x2), p1),
+       col="blue", lty=2, type="b")
```

Distribuição Weibull: previstos e observados

Probabilidades



Prob. acumuladas



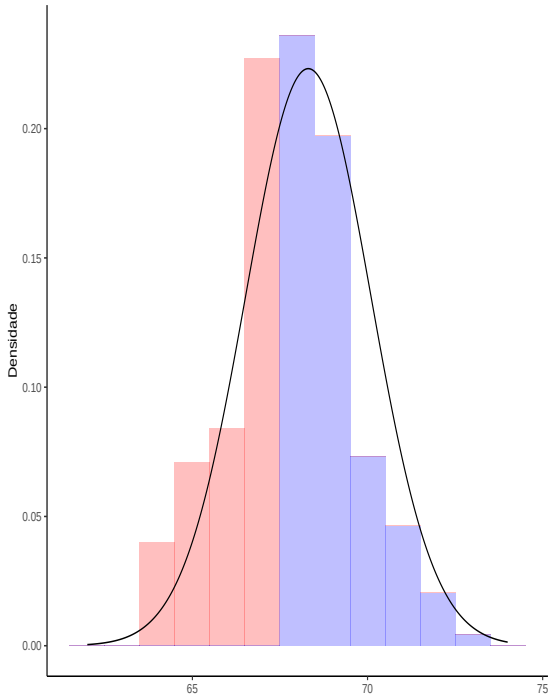
Distribuição Weibull: previstos e observados

Comandos

```
> pw.cf <- coef(parag.wei)
> hist(dap0, prob=T, main="", xlab = "DAP - 25 (cm)")
> curve(dweibull(x, shape=pw.cf[2], scale=pw.cf[1]),
+       add=T, col="blue")
```

Gráfico quantil-quantil

Quantil empírico



Quantil teórico

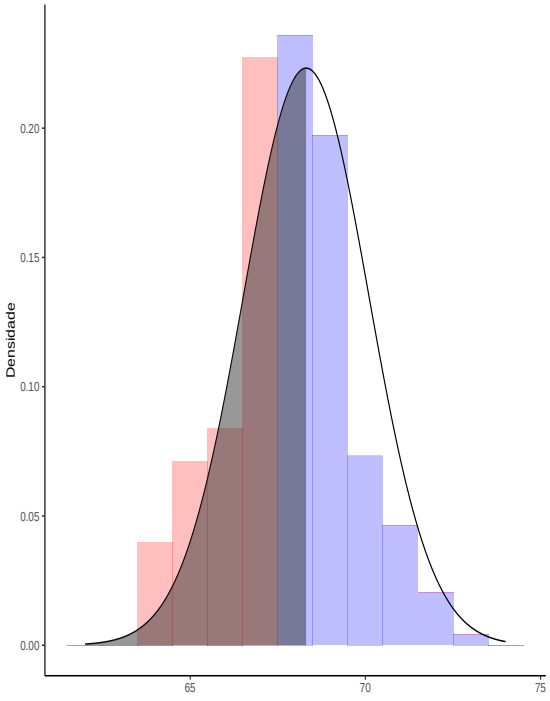
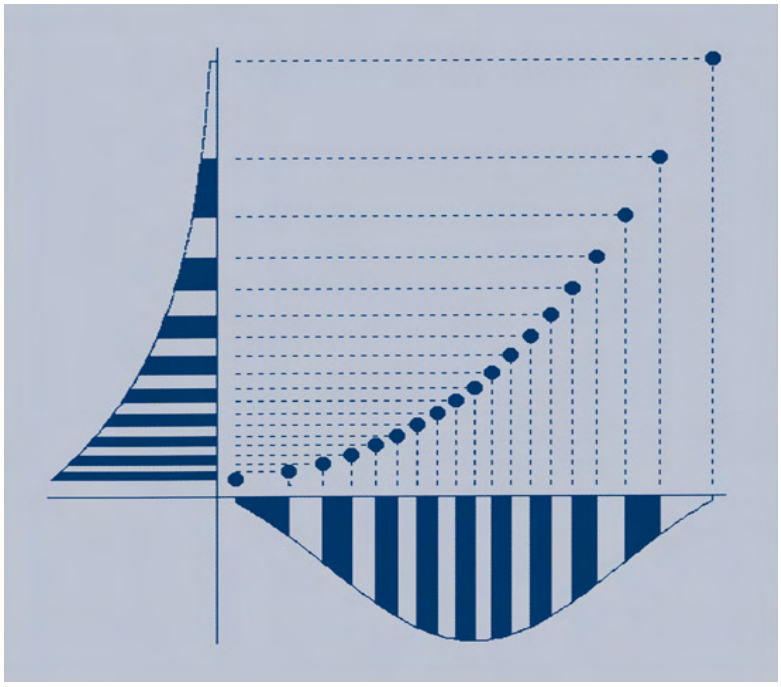


Gráfico quantil-quantil



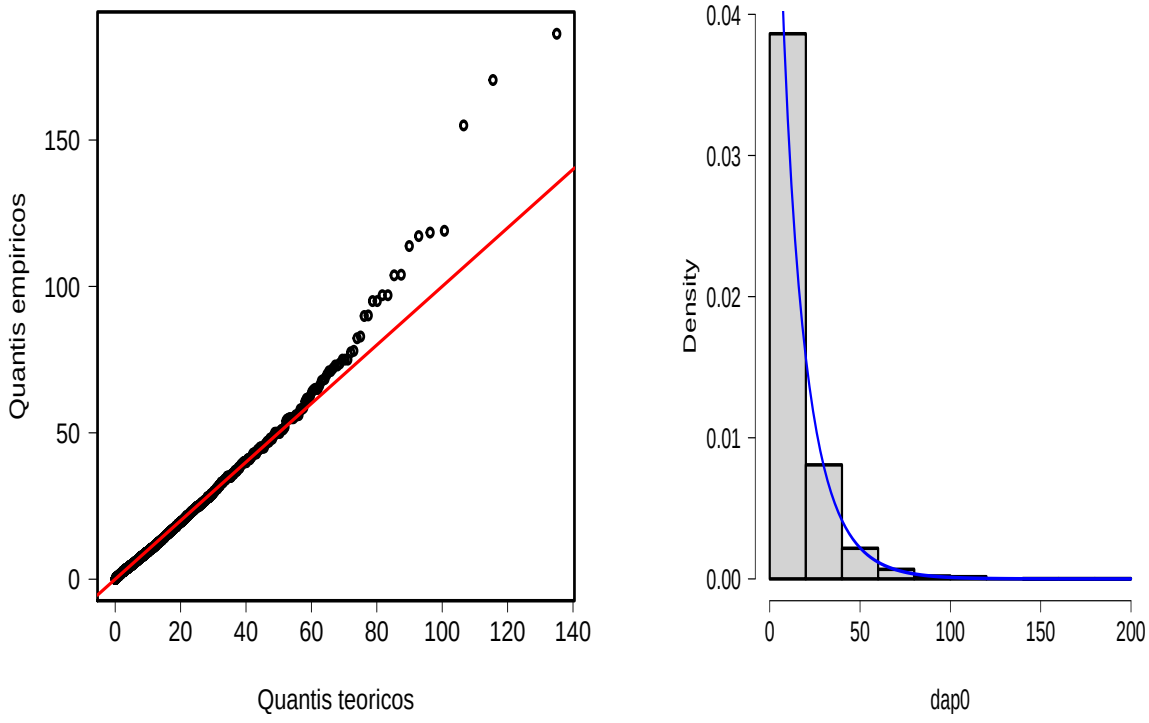
Capa de Verzani, 2005.
Using R for Introductory Statistics.

Gráfico de quantis do ajuste à Weibull

Comandos

```
> dap.P <- ppoints(dap0)
> dap.pred <- qweibull(dap.P, shape=pw.cf[2],
+                          scale=pw.cf[1])
> plot(sort(dap0)~dap.pred, xlab="Quantis teoricos",
+      ylab="Quantis empiricos")
> abline(0,1, col="red")
```

Gráfico de quantis do ajuste à Weibull



Testes de aderência: exemplo

Dados:

200 valores sorteados de uma distribuição gama com forma $= 1.3$ e escala $= 0.5$

MLEs:

exponencial: $\lambda = 0.41$

gama: forma $= 1.21$ e escala $= 0.50$

Teste K-S:

exponencial: $D = 0.049$, $p = 0.73$

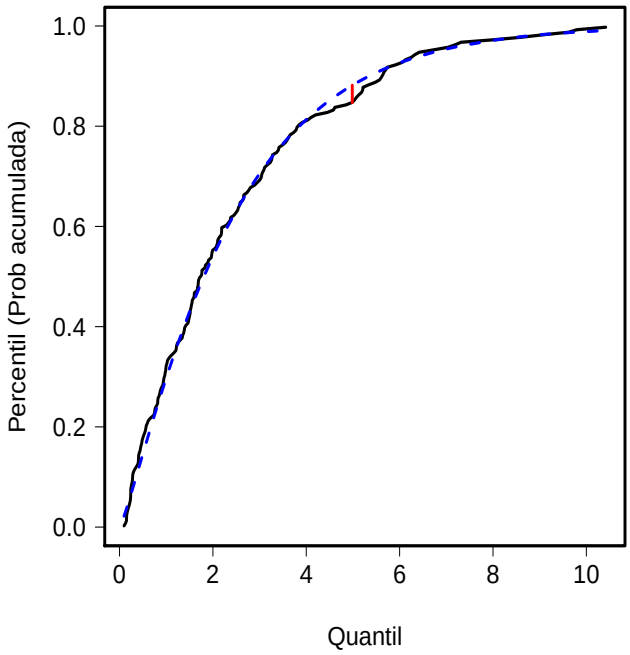
gama: $D = 0.041$, $p = 0.89$

Testes de aderência: exemplo

Teste de Kolmogorov-Smirnov

exponencial: $D = 0.049$, $p = 0.73$

gama: $D = 0.041$, $p = 0.89$



Seleção de modelos

	LogLik	gl	AIC	dAIC
gama	-375.8	2	755.7	0.0
exponencial	-378.0	1	758.1	2.4

Resumo

- Alguns modelos de distribuição são simples:
  1. tem apenas um parâmetro, ou
  2. as MLEs podem ser obtidas analiticamente.
- Nos modelos simples a *curva de verossimilhança* pode ser facilmente construída e analisada.
- Nos modelos complexos:
  1. existem mais de um parâmetro,
  2. as MLEs não podem ser obtidas analiticamente.
- Nos modelos complexos é difícil analisar a *superfície* de verossimilhança.
- A inferência é realizada *perfilhando* superfície de verossimilhança para cada parâmetro de interesse.

Resumo

- Após o ajuste faça um diagnóstico cuidadoso com:
  1. perfis de verossimilhança
  2. gráficos de valores observados x previstos
  3. gráficos de quantis e percentis
- Para a inferência estatística sobre o ajuste:
  1. Testes de aderência são problemáticos
  2. Comparação de modelos alternativos com AIC é mais adequada e flexível