CE-210: Inferência Estatística II
Aulas Práticas

Paulo J. Ribeiro Jr. & Ricardo Sanders Ehlers

Primeiro Semestre de 2003
Última atualização: 19 de agosto de 2003

Sumário

1 Introdução

Nas aulas práticas deste curso vamos utilizar o programa estatístico R. Vamos começar "experimentando o R", para ter uma idéia de seus recursos e a forma de trabalhar. Para isto vamos rodar e estudar os comandos mostrados no texto e seus resultados para nos familiarizar com o programa. Nas sessões seguintes iremos ver com mais detalhes o uso do programa R. Siga os seguintes passos:

1. inicie o R em seu computador;
2. voce verá uma janela de comandos com o símbolo >,
   este é o *prompt* do R indicando que o programa está pronto para receber comandos;
3. a seguir digite (ou "recorte e cole") os comandos mostrados abaixo.

No restante deste texto vamos seguir as seguintes convenções.

- comandos do R são sempre mostrados em fontes do tipo `typewriter como esta`,
- linhas iniciadas pelo símbolo # são comentários e são ignoradas pelo R.

2 Distribuições de Probabilidade

O programa R inclui funcionalidade para operações com distribuições de probabilidades. Para cada distribuição há 4 operações básicas indicadas pelas letras:

`d` calcula a densidade de probabilidade $f(x)$ no ponto

`p` calcula a função de probabilidade acumulada $F(x)$ no ponto

`q` calcula o quantil correspondente a uma dada probabilidade

`r` amostra da distribuição

2.1 Distribuição Normal

A funcionalidade para distribuição normal é implementada por argumentos que combinam as letras acima com o termo `norm`. Vamos ver alguns exemplos com a distribuição normal padrão.

```
> dnorm(-1)
[1] 0.2419707

> (1/sqrt(2*pi)) * exp((-1/2)*(-1)^2)
[1] 0.2419707

> pnorm(-1)
[1] 0.1586553

> qnorm(0.975)
[1] 1.959964

> rnorm(10)
 [1] -0.0442493 -0.3604689  0.2608995 -0.8503701 -0.1255832  0.4337861
 [7] -1.0240673 -1.3205288  2.0273882 -1.7574165
```

O primeiro valor acima corresponde ao valor da densidade da normal

$$f(x) = \frac{1}{2\pi\sigma^2} \exp\{\frac{1}{2\sigma^2}(x-\mu)^2\}$$

com parâmetros ($\mu = 0, \sigma^2 = 1$) no ponto -1, como pode ser confirmado com o cálculo apresentado no segundo comando.
A função `pnorm(-1)` calcula a probabilidade $P(X < -1)$.
O comando `qnorm(0.975)` calcula o valor de a tal que $P(X < a) = 0.975$.
Finalmente o comando `rnorm(10)` gera uma amostra de 10 elementos da normal padrão.

Nota: cada vez que o comando `rnorm` é chamado diferentes elementos da amostra são produzidos, porque a *semente* do gerador é modificada. Para gerar duas amostras idênticas deve-se usar o comando `set.seed` como ilustrado abaixo.

```
> set.seed(214)     # define o valor da semente
> rnorm(5)          # amostra de 5 elementos
[1] -0.46774980  0.04088223  1.00335193  2.02522505  0.30640096
> rnorm(5)          # outra amostra de 5 elementos
```

```
[1]  0.4257775  0.7488927  0.4464515 -2.2051418  1.9818137
> set.seed(214)     # retorna o valor da semente ao valor inicial
> rnorm(5)          # gera novamente a primeira amostra de 5 elementos
[1] -0.46774980  0.04088223  1.00335193  2.02522505  0.30640096
```

As funções acima possuem argumentos adicionais, para os quais valores padrão (*default*) foram assumidos, e que podem ser modificados. Usamos `args` para ver os argumentos de uma função e `help` para visualisar a documentação detalhada:

```
> args(rnorm)
function (n, mean = 0, sd = 1)
```

As funções relacionadas à distribuição normal tem (entre outros) os argumentos `mean` e `sd` para definir média e desvio padrão da distribuição que podem ser modificados como nos exemplos a seguir.

```
> qnorm(0.975, mean = 100, sd = 8)
[1] 115.6797

> qnorm(0.975, m = 100, s = 8)
[1] 115.6797

> qnorm(0.975, 100, 8)
[1] 115.6797
```

Para informações mais detalhadas pode-se usar a função `help`. O comando

```
> help(rnorm)
```

irá exibir em uma janela a documentação da função que pode também ser chamada com `?rnorm`. Note que ao final da documentação são apresentados exemplos que podem ser rodados pelo usuário e que auxiliam na compreensão da funcionalidade.
Note que as 4 funções relacionadas à distribuição normal são documentadas conjuntamente, portanto `help(rnorm)`, `help(qnorm)`, `help(dnorm)` e `help(pnorm)` vão exibir a mesma documentação.

Estas funções aceitam também vetores em seus argumentos como ilustrado nos exemplo abaixo.

```
> qnorm(c(0.05, 0.95))
[1] -1.644854  1.644854
> rnorm(4, mean=c(0, 10, 100, 1000))
[1]   0.1599628   9.0957340 100.5595095 999.9129392
> rnorm(4, mean=c(10, 20, 30, 40), sd=c(2, 5))
[1] 10.58318 21.92976 29.62843 42.71741
```

Note que no último exemplo a *lei da reciclagem* foi utilizada no vetor de desvios padrão, i.e. os desvios padrão utilizados foram $(2, 5, 2, 5)$.

Cálculos de probabilidades usuais, para os quais utilizávamos tabelas estatísticas podem ser facilmente obtidos como no exemplo a seguir.

Seja X uma v.a. com distribuição $N(100, 100)$. Calcular as probabilidades:

1. $P[X < 95]$
2. $P[90 < X < 110]$

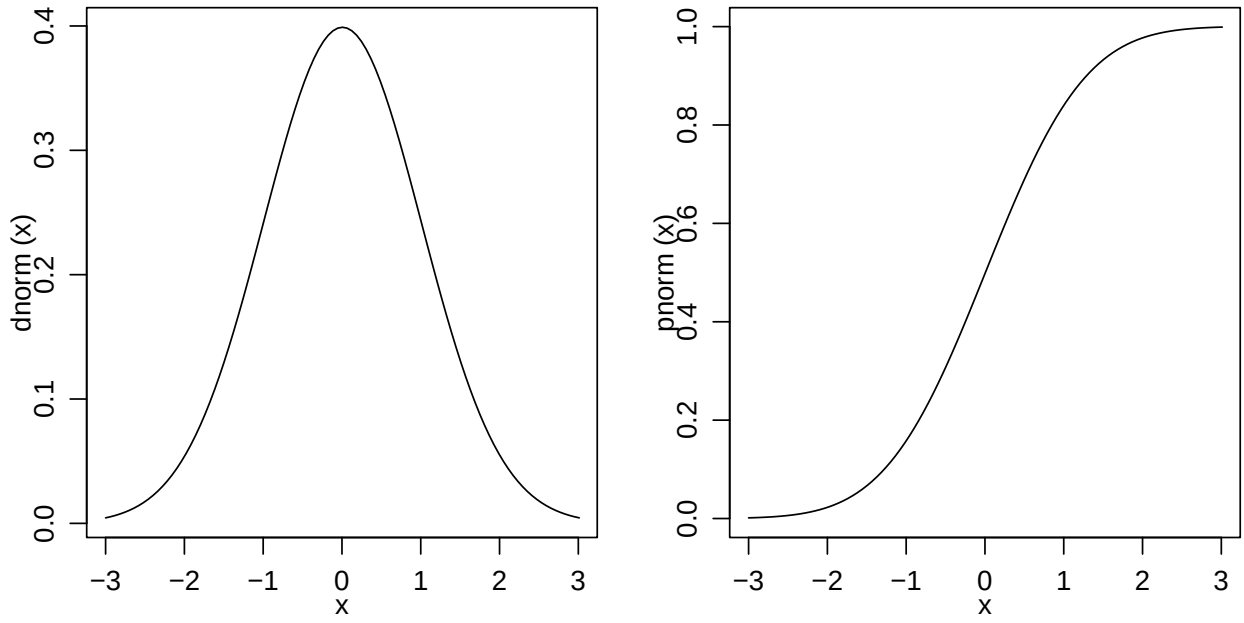

Figura 1: Funções de densidade e probabilidade da distribuição normal padrão.

3. $P[X > 95]$

Os comandos do R são:

```
> pnorm(95, 100, 10)
[1] 0.3085375

> pnorm(110, 100, 10) - pnorm(90, 100, 10)
[1] 0.6826895

> 1 - pnorm(95, 100, 10)
[1] 0.6914625
> pnorm(95, 100, 10, lower=F)
[1] 0.6914625
```

Note a última probabilidade foi calculada de duas formas diferentes, sendo a segunda usando o argumento `lower` mais estável numericamente.

A seguir vamos ver comandos para fazer gráficos de distribuições de probabilidade. Vamos fazer gráficos de funções de densidade e de probabilidade acumulada. Estude cuidadosamente os comandos abaixo e verifique os gráficos por eles produzidos. A Figura 1 mostra gráficos da densidade (esquerda) e probabilidade acumulada da normal padrão produzidos com os comandos:

```
> plot(dnorm, -3, 3)
> plot(pnorm, -3, 3)
```

A Figura 2 mostra gráficos da densidade (esquerda) e probabilidade acumulada da $N(100, 64)$ produzidos com os comandos:

```
> plot(function(x) dnorm(x, 100, 8), 70, 130)
> plot(function(x) pnorm(x, 100, 8), 70, 130)
```

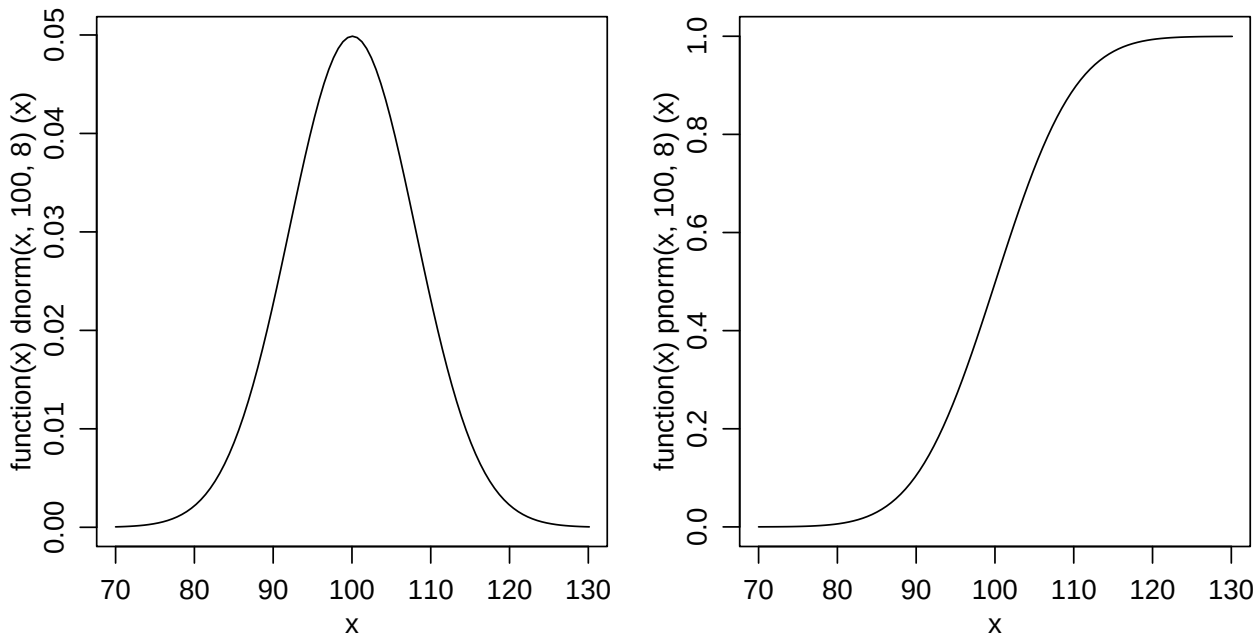

Figura 2: Funções de densidade e probabilidade da $N(100, 64)$.

Podemos incluir títulos e mudar texto dos eixos conforme mostrado na gráfico da esquerda da Figura 3 e nos dois primeiros comandos abaixo. Os demais comandos mostram como colocar diferentes densidades em um um mesmo gráfico como ilustrado à direita da mesma Figura.

```
> plot(dnorm, -3, 3, xlab='valores de X', ylab='densidade de probabilidade')
> title('Distribuicão Normal\nX ~ N(100, 64)')

> plot(function(x) dnorm(x, 100, 8), 60, 140, ylab='f(x)')
> plot(function(x) dnorm(x, 90, 8), 60, 140, add=T, col=2)
> plot(function(x) dnorm(x, 100, 15), 60, 140, add=T, col=3)
> legend(120, 0.05, c("N(100,64)","N(90,64)","N(100,225)"), fill=1:3)
```

2.2 Distribuição Binomial

Cálculos para a distribuição binomial são implementados combinando as *letras básicas* vistas acima com o termo `binom`. Vamos primeiro investigar argumentos e documentação com os comandos `args` e `binom`.

```
> args(dbinom)
function (x, size, prob, log = FALSE)

> help(dbinom)
```

Seja X uma v.a. com distribuição Binomial com $n = 10$ e $p = 0.35$. Vamos ver os comandos do R para:

1. fazer o gráfico das função de densidade
2. idem para a função de probabilidade
3. calcular $P[X = 7]$

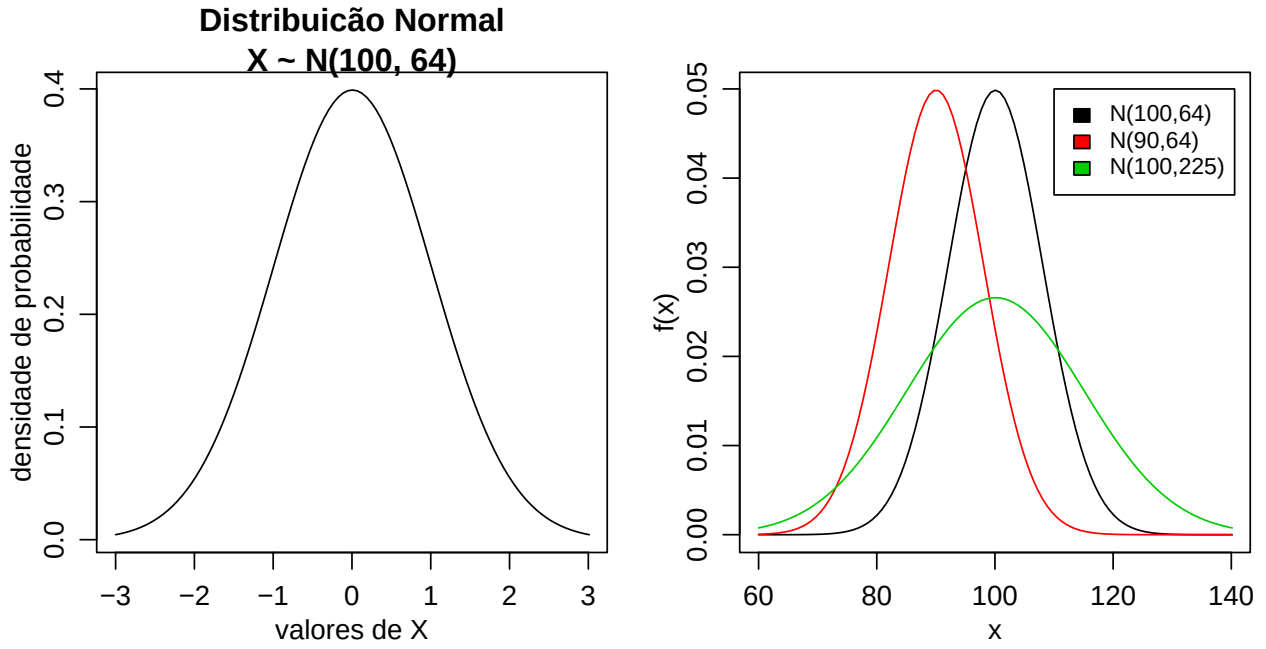

Figura 3: Gráfico com texto nos eixos e título (esquerda) e várias distribuições em um mesmo gráfico (direita).

4. calcular $P[X < 8] = P[X \leq 7]$
5. calcular $P[X \geq 8] = P[X > 7]$
6. calcular $P[3 < X \leq 6] = P[4 \geq X < 7]$

Note que sendo uma distribuição discreta de probabilidades os gráficos são diferentes dos obtidos para distribuição normal e os cálculos de probabilidades devem considerar as probabilidades nos pontos. Os gráficos das funções de densidade e probabilidade são mostrados na Figura 4.

```
> x <- 0:10

> fx <- dbinom(x, 10, 0.35)
> plot(x, fx, type='h')

> Fx <- pbinom(x, 10, 0.35)
> plot(x, Fx, type='S')

> dbinom(7, 10, 0.35)
[1] 0.02120302

> pbinom(7, 10, 0.35)
[1] 0.9951787
> sum(dbinom(0:7, 10, 0.35))
[1] 0.9951787

> 1-pbinom(7, 10, 0.35)
[1] 0.004821265
> pbinom(7, 10, 0.35, lower=F)
```

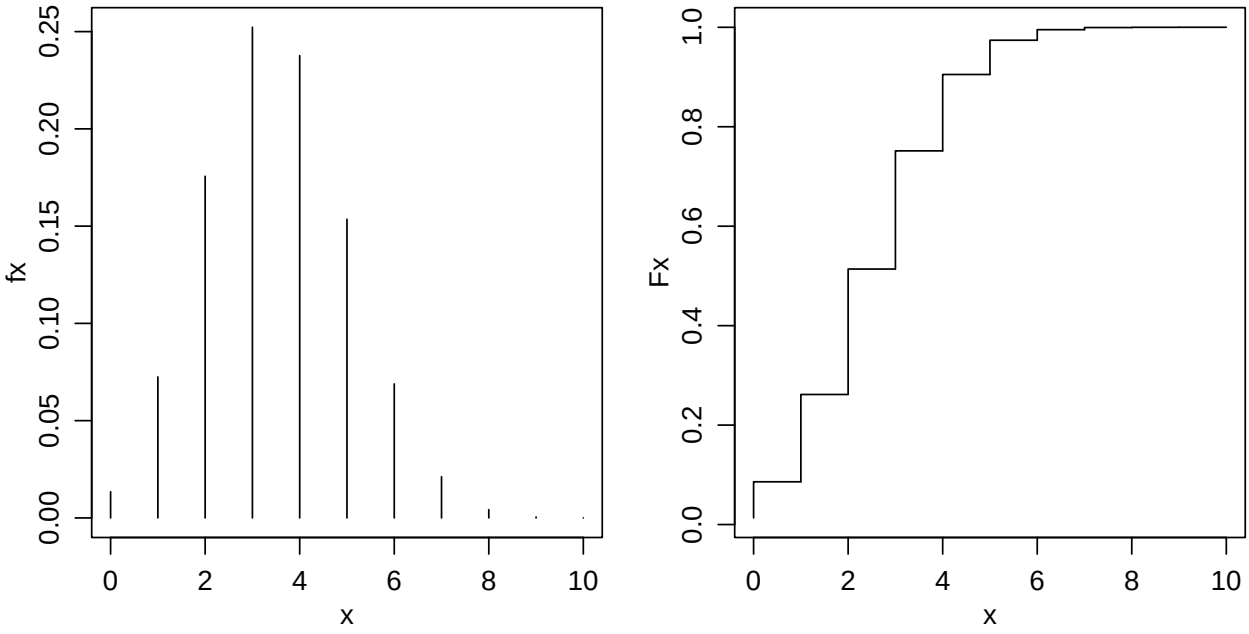

Figura 4: Funções de densidade e probabilidade da $B(10, 0.35)$.

```
[1] 0.004821265

> pbinom(6, 10, 0.35) - pbinom(3, 10, 0.35)
[1] 0.4601487
> sum(dbinom(4:6, 10, 0.35))
[1] 0.4601487
```

2.3 Exercícios

Nos exercícios abixo iremos também usar o R como uma calculadora estatística para resolver alguns exemplos/exercícios de probabilidade tipicamente apresentados em um curso de estatística básica.

Os exercícios abaixo com indicação de página foram retirados de:

Magalhães, M.N. & Lima, A.C.P. (2001) **Noções de Probabilidade e Estatística**. 3 ed. São Paulo, IME-USP. 392p.

1. (Ex 1, pag 67) Uma moeda viciada tem probabilidade de cara igual a 0.4. Para quatro lançamentos independentes dessa moeda, estude o comportamento da variável *número de caras* e faça um gráfico de sua função de distribuição.

2. (Ex 3.6, pag 65) Num estudo sobre a incidência de câncer foi registrado, para cada paciente com este diagnóstico o número de casos de câncer em parentes próximos (pais, irmãos, tios, filhos e sobrinhos). Os dados de 26 pacientes são os seguintes:

Paciente	1	2	3	4	5	6	7	8	9	10	11	12	13
Incidência	2	5	0	2	1	5	3	3	3	2	0	1	1
Paciente	14	15	16	17	18	19	20	21	22	23	24	25	26
Incidência	4	5	2	2	3	2	1	5	4	0	0	3	3

Estudos anteriores assumem que a incidência de câncer em parentes próximos pode ser modelada pela seguinte função discreta de probabilidades:

Incidência	0	1	2	3	4	5
p_i	0.1	0.1	0.3	0.3	0.1	0.1

- os dados observados concordam com o modelo teórico?
- faça um gráfico mostrando as frequências teóricas (esperadas) e observadas.

3. (Ex 5, pag 77) Sendo X uma variável seguindo o modelo Binomial com parâmetro $n = 15$ e $p = 0.4$, pergunta-se:
   - $P(X \geq 14)$
   - $P(8 < X \leq 10)$
   - $P(X < 2 \text{ ou } X \geq 11)$
   - $P(X \geq 11 \text{ ou } X > 13))$
   - $P(X > 3 \text{ e } X < 6)$
   - $P(X \leq 13 \mid X \geq 11)$

4. (Ex 8, pag 193) Para $X \sim N(90, 100)$, obtenha:
   - $P(X \leq 115)$
   - $P(X \geq 80)$
   - $P(X \leq 75)$
   - $P(85 \leq X \leq 110)$
   - $P(|X - 90| \leq 10)$
   - P valor de a tal que $P(90 - a \leq X \leq 90 + a) = \gamma,\ \gamma = 0.95$

5. Faça os seguintes gráficos:
   - da função de densidade de uma variável com distribuição de Poisson com parametro $lambda = 5$
   - da densidade de uma variável $X \sim N(90, 100)$
   - sobreponha ao gráfico anterior a densidade de uma variável $Y \sim N(90, 80)$ e outra $Z \sim N(85, 100)$
   - densidades de distribuições *Chi-quadrado* com 1, 2 e 5 graus de liberdade.

3 Alguns recursos do R

3.1 O projeto R

O programa R é gratuito e de código aberto que propicia excelente ambiente para análises estatísticas e com recursos gráficos de alta qualidade. Detalhes sobre o projeto, colaboradores, documentação e diversas outras informações podem ser encontradas na página oficial do projeto em:
http://www.r-project.org.

O programa pode ser copiado livremente pela internet. Há um espelho (*mirror*) brasileiro da área de *downloads* do programa no *Departamento de Estatística da UFPR*:
http://www.est.ufpr.br/R
ou então via FTP:
ftp://est.ufpr.br/R

Será feita uma apresentação rápida da página do R durante o curso onde os principais recursos serão comentados assim como as idéias principais que governam o projeto e suas direções futuras.

3.2 Demos

O R vem com algumas demonstrações (*demos*) de seus recursos "embutidas" no programa. Para listar as demos disponíveis digite na linha de comando:

```
demo()
```

E para rodar uma delas basta colocar o nome da escolhida entre os parênteses. Por exemplo, vamos rodar a de recursos gráficos. Note que os comandos vão aparecer na janela de comandos e os gráficos serão automaticamente produzidos na janela gráfica. Voce vai ter que teclar ENTER para ver o próximo gráfico.

- inicie o programa R
- no "prompt" do programa digite:

  ```
  demo(graphics)
  ```

- Voce vai ver a seguinte mensagem na tela:

  ```
  demo(graphics)
  ---- ~~~~~~~~

  Type  <Return>   to start :
  ```

- pressione a tecla ENTER
- a "demo" vai ser iniciada e uma tela gráfica irá se abrir. Na tela de comandos serão mostrados comandos que serão utilizados para gerar um gráfico seguidos da mensagem:

  ```
  Hit <Return> to see next plot:
  ```

- inspecione os comandos e depois pressione novamente a tecla ENTER.
  Agora voce pode visualizar na janela gráfica o gráfico produzido pelos comandos mostrados anteriormente. Inspecione o gráfico cuidadosamente verificando os recursos utilizados (título, legendas dos eixos, tipos de pontos, cores dos pontos, linhas, cores de fundo, etc).

- agora na tela de comandos apareceram novos comandos para produzir um novo gráfico e a mensagem:

  ```
  Hit <Return> to see next plot:
  ```

- inspecione os novos comandos e depois pressione novamente a tecla ENTER.
  Um novo gráfico surgirá ilustrando outros recursos do programa.
  Prossiga inspecionando os gráficos e comandos e pressionando ENTER até terminar a "demo".
  Experimente outras demos como `demo(pers)` e `demo(image)`, por exemplo.

3.3 Um tutorial sobre o R

Além dos materiais disponíveis na página do programa há também um *Tutorial de Introdução ao* R disponível em http://www.est.ufpr.br/Rtutorial.

3.4 RWeb

Este é um mecanismo que permite rodar o R pela web, sem que voce precise ter o R instalado no seu computador. Para usá-lo basta estar conectado na internet.

Para acessar o **RWeb** vá até a página do Re no menu à esquerda da página siga os links:
R GUIs . . . R Web
Nesta página selecione primeiro o link **R Web** e examine seu conteúdo.

Há ainda uma diversidade de recursos disponíveis na página do projeto. Os participantes do curso são estimulados a explorar detalhadamente ao final do curso os outros recursos da página.

3.5 Cartão de referência

Como voce já pode perceber, para utilizar o R é necessário conhecer e digitar comandos. Isto pode trazer alguma dificuldade no inicio até que o usuário de familiarize com os comandos mais comuns. Uma boa forma de aprender e memorizar os comandos básicos é utilizar o Cartão de Referência que contém os comandos mais frequentemente utilizados. Imprima o conteúdo deste arquivo (1 folha) e carregue sempre com voce.

4 Intervalos de confiança – exemplos iniciais

Nesta sessão vamos verificar como utilizar o R para obter intervalos de confiança para parâmetros de distribuições para as quais os resultados são bem conhecidos.

Para fins didáticos de demonstração dos recursos do R vamos mostrar três possíveis soluções:

1. fazendo as contas passo a passo, utilizando o R como uma calculadora
2. escrevendo uma função
3. usando uma função já existente no R

4.1 Média de uma distribuição normal com variância desconhecida

Considere resolver o seguinte problema:

Exemplo

O tempo de reação de um novo medicamento pode ser considerado como tendo distribuição Normal e deseja-se fazer inferência sobre a média que é desconhecida obtendo um intervalo de confiança. Vinte pacientes foram sorteados e tiveram seu tempo de reação anotado. Os dados foram os seguintes (em minutos):

2.9	3.4	3.5	4.1	4.6	4.7	4.5	3.8	5.3	4.9
4.8	5.7	5.8	5.0	3.4	5.9	6.3	4.6	5.5	6.2

Entramos com os dados com o comando

```
> tempo <- c(2.9, 3.4, 3.5, 4.1, 4.6, 4.7, 4.5, 3.8, 5.3, 4.9,
           4.8, 5.7, 5.8, 5.0, 3.4, 5.9, 6.3, 4.6, 5.5, 6.2)
```

Sabemos que o intervalo de confiança para média de uma distribuição normal com média desconhecida é dado por:

$$\left(\bar{x} + t_{\alpha/2}\sqrt{\frac{S^2}{n}} \quad , \quad \bar{x} + t_{1-\alpha/2}\sqrt{\frac{S^2}{n}}\right)$$

Vamos agora obter a resposta das três formas diferentes mencionadas acima.

4.1.1 Fazendo as contas passo a passo

Nos comandos a seguir calculamos o tamanho da amostra, a média e a variância amostral.

```
> n <- length(tempo)
> n
[1] 20
> t.m <- mean(tempo)
> t.m
[1] 4.745
> t.v <- var(tempo)
> t.v
[1] 0.992079
```

A seguir montamos o intervalo utilizando os quantis da dritribuição t.

```
> t.ic <- t.m + qt(c(0.025, 0.975), df = n-1) * sqrt(t.v/length(tempo))
> t.ic
[1] 4.278843 5.211157
```

4.1.2 Escrevendo uma função

Podemos generalizar a solução acima agrupando os comandos em uma função. Nos comandos abaixo primeiro definimos a função e a seguir utilizamos a função criada definindo intervalos a 95% e 99%.

```
> ic.m <- function(x, nivel = 0.95){
+   n <- length(x)
+   media <- mean(x)
+   variancia <- var(x)
+   quantis <- qt(c((1-nivel)/2, 1 - (1-nivel)/2), df = n-1)
+   ic <- media + quantis * sqrt(variancia/n)
+   return(ic)
+ }
> ic.m(tempo)
[1] 4.278843 5.211157

> ic.m(tempo, nivel=0.99)
[1] 4.107814 5.382186
```

Escrever uma função é particularmente útil quando um procedimento vai ser utilizados várias vezes.

4.1.3 Usando a função `t.test`

Mostramos as soluções acima para ilustrar a flexibilidade e o uso do programa. Entretanto não precisamos fazer isto na maioria das vezes porque o R já vem com várias funções para procedimentos estatísticos já escritas. Neste caso a função `t.test` pode ser utilizada como vemos no resultado do comando a sequir que coincide com os obtidos anteriormente.

```
> t.test(tempo)

        One Sample t-test

data:  tempo
t = 21.3048, df = 19, p-value = 1.006e-14
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 4.278843 5.211157
sample estimates:
mean of x
    4.745
```

4.2 Exercícios

Em cada um dos exercícios abaixo tente obter os intervalos das três formas mostradas acima.

1. (Ex 7.21, pag 233) Pretende-se estimar a proporção p de cura, através de uso de um certo medicamento em doentes contaminados com cercária, que é uma das formas do verme da esquitosomose. Um experimento consistiu em aplicar o medicamento em 200 pacientes, escolhidos ao acaso, e observar que 160 deles foram curados. Montar o intervalo de confiança para a proporção de curados.
   Note que há duas expressões póssíveis para este IC: o "otimista"e o "conservativo". Encontre ambos intervalos.

2. Os dados abaixo são uma amostra aleatória da distribuição $Bernoulli(p)$. Obter IC's a 90% e 99%.

   0 0 0 1 1 0 1 1 1 1 0 1 1 0 1 1 1 1 0 1 1 1 1 1 1

3. Encontre intervalos de confiança de 95% para a média de uma distribuição Normal com variância 1 dada a amostra abaixo

   9.5 10.8 9.3 10.7 10.9 10.5 10.7 9.0 11.0 8.4
   10.9 9.8 11.4 10.6 9.2 9.7 8.3 10.8 9.8 9.0

4. Queremos verificar se duas máquinas produzem peças com a mesma homogeneidade quanto a resistência à tensão. Para isso, sorteamos dias amostras de 6 peças de cada máquina, e obtivemos as seguintes resistências:

Máquina A	145	127	136	142	141	137
Máquina B	143	128	132	138	142	132

   Obtenha intervalos de confiança para a razão das variâncias e para a diferença das médias dos dois grupos.

5 Funções de verossimilhança

A função de verossimilhança é central na inferência estatística. Nesta sessão vamos ver como traçar funções de verossimilhança utilizando o programa R.

5.1 Exemplo 1: Distribuição normal com variância conhecida

Seja o vetor $(12, 15, 9, 10, 17, 12, 11, 18, 15, 13)$ uma amostra aleatória de uma distribuição normal de média μ e variância conhecida e igual a 4. O objetivo é fazer um gráfico da função de log-verossimilhança.

Solução:

Vejamos primeiro os passos da solução analítica:

1. Temos que $x_1, \ldots, x_n$ é uma a.a. de $X \sim N(\mu, 4)$,
2. a densidade para cada observação é dada por $f(x_i) = \frac{1}{2\sqrt{2\pi}} \exp\{\frac{1}{8}(x_i - \mu)^2\}$,
3. a verossimilhança é dada por $L(\mu) = \prod_1^{10} f(x_i)$,
4. e a log-verossimilhança é dada por

$$\begin{aligned} l(\mu) &= \sum_1^{10} \log(f(x_i)) \\ &= -5\log(8\pi) - \frac{1}{8}(\sum_1^{10} x_i^2 - 2\mu \sum_1^{10} x_i + 10\mu^2), \end{aligned} \tag{1}$$

5. que é uma função de μ e portanto devemos fazer um gráfico de $l(\mu)$ *versus* μ tomando vários valores de μ e calculando os valores de $l(\mu)$.

Vamos ver agora uma primeira possível forma de fazer a função de verossimilhança no R.

1. Primeiro entramos com os dados que armazenamos no vetor `x`

```
> x <- c(12, 15, 9, 10, 17, 12, 11, 18, 15, 13)
```

2. e calculamos as quantidades $\sum_1^{10} x_i^2$ e $\sum_1^{10} x_i$

```
> sx2 <- sum(x^2)
> sx <- sum(x)
```

3. agora tomamos uma sequência de valores para μ. Sabemos que o estimador de máxima verossimilhança neste caso é $\hat{\mu} = 13.2$ (este valor pode ser obtido com o comando `mean(x)`) e portanto vamos definir tomar valores ao redor deste ponto.

```
> mu.vals <- seq(11, 15, l=100)
```

4. e a seguir calculamos os valores de $l(\mu)$ de acordo com a equação acima

```
> lmu <- -5 * log(8*pi) - (sx2 - 2*mu.vals*sx + 10*(mu.vals^2))/8
```

5. e finalmente fazemos o gráfico visto na Figura 5

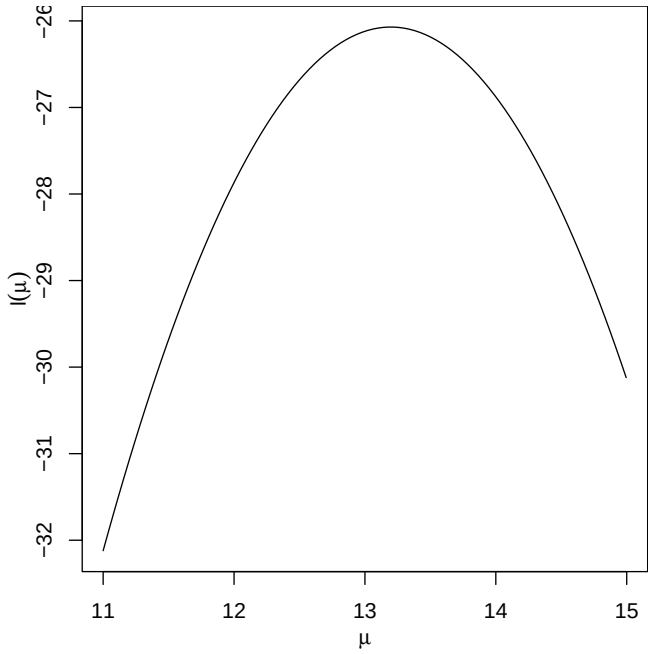

Figura 5: Função de verossimilhança para o parâmetro μ da distribuição normal com variância $\sigma^2 = 4$ com os dados do Exemplo 1.

```
> plot(mu.vals, lmu, type='l', xlab=expression(mu), ylab=expression(l(mu)))
```

Entretanto podemos obter a função de verossimilhança no R de outras forma mais geral e menos trabalhosas. Sabemos que a função `dnorm` calcula a densidade $f(x)$ da distribuição normal e podemos usar este fato para evitar a digitação da expressão acima.

- Primeiro vamos criar uma função que calcula o valor da log-verossimilhança para um certo valor do parâmetro e para um certo conjunto de dados,

```
> logvero <- function(mu, dados){
     sum(dnorm(dados, mean = mu, sd = 2, log = TRUE))
  }
```

- a seguir criamos uma sequência adequada de valores de μ e calculamos $l(\mu)$ para cada um dos valores

```
> mu.vals <- seq(11, 15, l=100)
> mu.vals
> lmu <- sapply(mu.vals, logvero, dados = x)
> lmu
```

  Note na sintaxe acima que a função `sapply` aplica a função `logvero` anteriormente definida em cada elemento do vetor `mu.vals`.

- Finalmente fazemos o gráfico.

```
> plot(mu.vals, lmu, type='l', xlab=expression(mu), ylab=expression(l(mu)))
```

Para encerrar este exemplo vamos apresentar uma solução ainda mais genérica que consiste em criar uma função que vamos chamar de `vero.norm.v4` para cálculo da verossimilhança de distribuições normais com σ^2=4. Esta função engloba os comandos acima e pode ser utilizada para obter o gráfico da log-verossimilhança para o parâmetro μ para qualquer amostra obtida desta distribuição.

```
> vero.normal.v4 <- function(mu, dados){
    logvero <- function(mu, dados)
      sum(dnorm(dados, mean = mu, sd = 2, log = TRUE))
    sapply(mu, logvero, dados = dados)
  }
> curve(vero.normal.v4(x, dados = x), 11, 15,
      xlab=expression(mu), ylab=expression(l(mu)))
```

5.2 Exercícios

1. Seja a amostra abaixo obtida de uma distribuição Poisson de parâmetro λ.
   5 4 6 2 2 4 5 3 3 0 1 7 6 5 3 6 5 3 7 2
   Obtenha o gráfico da função de log-verossimilhança.

2. Seja a amostra abaixo obtida de uma distribuição Binomial de parâmetro p e com $n = 10$.
   7 5 8 6 9 6 9 7 7 7 8 8 9 9 9
   Obtenha o gráfico da função de log-verossimilhança.

3. Seja a amostra abaixo obtida de uma distribuição χ^2 de parâmetro ν.
   8.9 10.1 12.1 6.4 12.4 16.9 10.5 9.9 10.8 11.4
   Obtenha o gráfico da função de log-verossimilhança.

6 Intervalos de confiança baseados na deviance

Vamos fazer aqui os exemplos vistos em sala de aula sobre intervalos de confiança baseado na deviance.

6.1 Média da distribuição normal com variância conhecida

Seja $X_1, \ldots, X_n$ a.a. de uma distribuição normal de média θ e variância 1. Vimos que:

1. A função de log-verossimilhança é dada por $l(\theta) = \text{cte} + \frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2$;
2. o estimador de máxima verossimilhança é $\hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$;
3. a função deviance é $D(\theta) = n(\bar{x} - \theta)^2$;
4. e neste caso a deviance tem distribuição exata $\chi^2_{(1)}$;
5. e os limites do intervalo são dados por $\bar{x} \stackrel{+}{-} \sqrt{c^*/n}$, onde c^* é o quantil $1 - \alpha/2$ da distribuição $\chi^2_{(1)}$.

Vamos considerar que temos uma amostra onde $n = 20$ e $\bar{x} = 32$. Neste caso a função deviance é como mostrada na Figura 6 que é obtida com os comandos abaixo onde primeiro definimos uma função para calcular a deviance que depois é mostrada em um gráfico para valores entre 30 e 34. Para obtermos um intervalo a 95% de confiança escolhemos o quantil correspondente na distribuição $\chi^2_{(1)}$ e mostrado pela linha tracejada no gráfico. Os pontos onde esta linha cortam a função são, neste exemplo, determinados analiticamente pela expressão dada acima e indicados pelos setas verticais no gráfico.

```
> dev.norm.v1 <- function(theta, n, xbar){n * (xbar - theta)^2}
> thetaN.vals <- seq(31, 33, l=101)
> dev.vals <- dev.norm.v1(thetaN.vals, n=20, xbar=32)
> plot(thetaN.vals, dev.vals, ty='l',
+      xlab=expression(theta), ylab=expression(D(theta)))
> corte <- qchisq(0.95, df = 1)
> abline(h = corte, lty=3)
> limites <- 32 + c(-1, 1) * sqrt(corte/20)
> limites
[1] 31.56174 32.43826
> segments(limites, rep(corte,2), limites, rep(0,2))
```

Vamos agora examinar o efeito do tamanho da amostra na função. A Figura 7 mostra as funções para três tamanhos de amostra, $n = 10, 20$ e 50 que são obtidas com os comandos abaixo. A linha horizontal mostra o efeito nas amplitudes dos IC's.

```
> dev10.vals <- dev.norm.v1(thetaN.vals, n=10, xbar=32)
> plot(thetaN.vals, dev10.vals, ty='l',
+      xlab=expression(theta), ylab=expression(D(theta)))
> dev20.vals <- dev.norm.v1(thetaN.vals, n=20, xbar=32)
> lines(thetaN.vals, dev20.vals, lty=2)
> dev50.vals <- dev.norm.v1(thetaN.vals, n=50, xbar=32)
> lines(thetaN.vals, dev50.vals, lwd=2)
> abline(h = qchisq(0.95, df = 1), lty=3)
> legend(31, 2, c('n=10', 'n=20', 'n=50'), lty=c(1,2,1), lwd=c(1,1,2), cex=0.7)
```

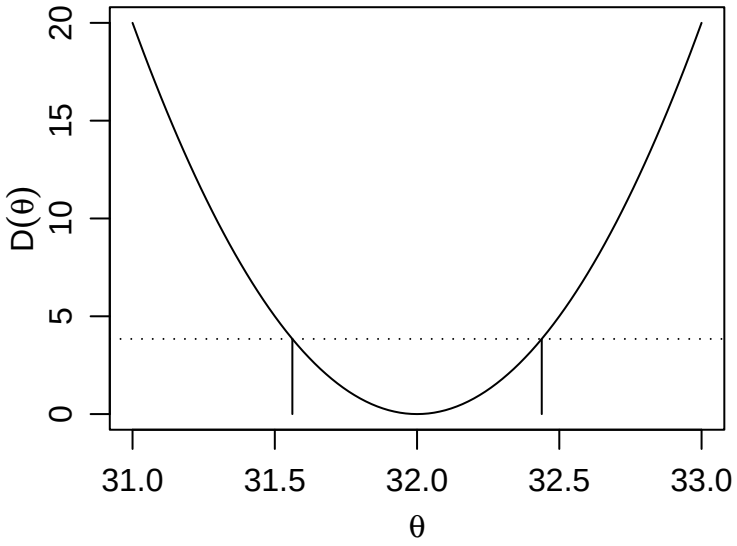

Figura 6: Função deviance para $\mathrm{N}(\theta, 1)$ para uma amostra de tamanho 20 e média 32.

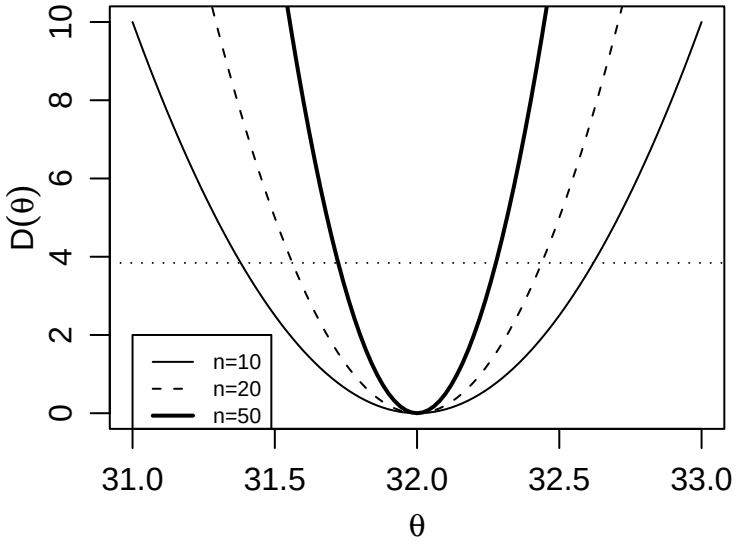

Figura 7: Funções deviance para o parâmetro θ da $\mathrm{N}(\theta, 1)$ para amostras de média 32 e tamanhos de amostra $n = 10, 20$ e 50.

6.2 IC para o parâmetro da distribuição exponencial

Seja $X_1, \ldots, X_n$ a.a. de uma distribuição exponencial de parâmetro θ com função de densidade $f(x) = \theta \exp\{-\theta x\}$. Vimos que:

1. A função de log-verossimilhança é dada por $l(\theta) = n \log(\theta) - \theta n \bar{x}$;
2. o estimador de máxima verossimilhança é $\hat{\theta} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}$;
3. a função deviance é $D(\theta) = 2n \left[\log(\hat{\theta}/\theta) + \bar{x}(\theta - \hat{\theta})\right]$;
4. e neste caso a deviance tem distribuição assintótica $\chi^2_{(1)}$;
5. e os limites do intervalo não podem ser obtidos analiticamente, devendo ser obtidos por:

- métodos numéricos ou gráficos, ou,
- pela aproximação quadrática da verossimilhança por série de Taylor que neste caso fornece uma expressão da deviance aproximada dada por $D(\theta) \approx n\left(\frac{\theta-\hat{\theta}}{\hat{\theta}}\right)^2$.

A seguir vamos ilustrar a obtenção destes intervalos no R. Vamos considerar que temos uma amostra onde $n = 20$ e $\bar{x} = 10$ para a qual a função deviance é mostrada na Figura 8 e obtida de forma análoga ao exemplo anterior.

```
> dev.exp <- function(theta, n, xbar){
+   2*n*(log((1/xbar)/theta) + xbar*(theta-(1/xbar)))
+ }
> thetaE.vals <- seq(0.04,0.20, l=101)
> dev.vals <- dev.exp(thetaE.vals, n=20, xbar=10)
> plot(thetaE.vals, dev.vals, ty='l',
+      xlab=expression(theta), ylab=expression(D(theta)))
```

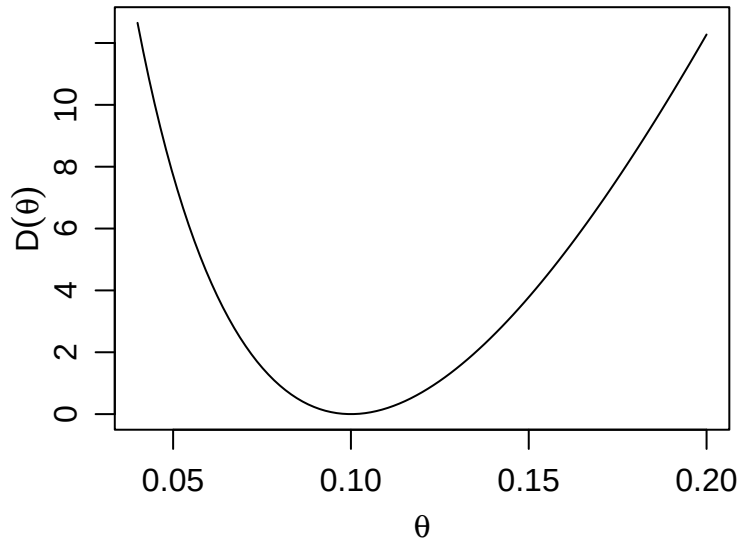

Figura 8: Função deviance da Exp(θ) para uma amostra de tamanho 20 e média 10.

Neste exemplo, diferentemente do anterior, não determinamos a distribuição exata da deviance e usamos a distribuição assintótica $\chi^2_{(1)}$ na qual se baseia a linha de corte tracejada mostrada no gráfico para definir o IC do parâmetro ao nível de 95% de confiança.

Para encontrar os limites do IC precisamos dos valores no eixo dos parâmetros nos pontos onde a linha de corte toca a função deviance o que corresponde a resolver a equação $D(\theta) = 2n\left[\log(\hat{\theta}/\theta) + \bar{x}(\theta - \hat{\theta})\right] = c^*$ onde c^* é quantil da distribuição da χ^2 com 1 grau de liberdade correspondente ao nível de confiança desejado. Por exemplo, para 95% o valor de $\chi^2_{1,0.95}$ é 3.84. Como esta equação não tem solução analítica (diferentemente do exemplo anterior) vamos examinar a seguir duas possíveis soluções para encontrar os limites do intervalo.

6.2.1 Solução numérica/gráfica simplificada

Iremos aqui considerar uma solução simples baseada no gráfico da função deviance para encontrar os limites do IC que consiste no seguinte: Para fazermos o gráfico da deviance criamos uma sequência de valores do parâmetro θ. A cada um destes valores corresponde um valor de $D(\theta)$. Vamos então localizar os valores de θ para os quais $D(\theta)$ é o mais próximo possível do ponto de corte. Isto é feito com o código abaixo e o resultado exibido na Figura 9.

```
> plot(thetaE.vals, dev.vals, ty='l',
+      xlab=expression(theta), ylab=expression(D(theta)))
> corte <- qchisq(0.95, df = 1)
> abline(h = corte, lty=3)
> dif <- abs(dev.vals - corte)
> linf <- thetaE.vals[thetaE.vals<(1/10)][which.min(dif[thetaE.vals<(1/10)])]
> lsup <- thetaE.vals[thetaE.vals>(1/10)][which.min(dif[thetaE.vals>(1/10)])]
> limites.dev <- c(linf, lsup)
> limites.dev
[1] 0.0624 0.1504
> segments(limites.dev, rep(corte,2), limites.dev, rep(0,2))
```

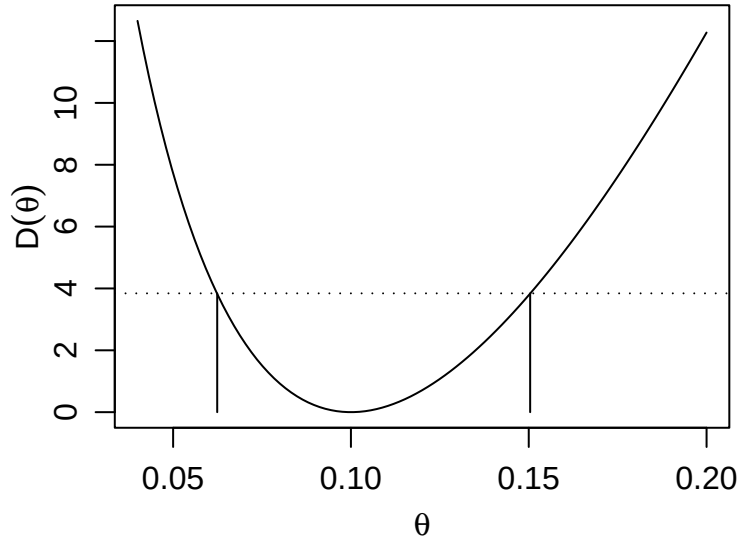

Figura 9: Obtenção gráfica do IC para o parâmetro θ da Exp(θ) para uma amostra de tamanho 20 e média 10.

Note que neste código procuramos primeiro e limite inferior entre os valores menores que a estimativa do parâmetro (1/10) e depois o limite superior entre os valores maiores que esta estimativa. Embora este procedimento bastante simples e sujeito a imprecisão podemos torná-lo quão preciso quanto quisermos bastando para isto definir um vetor com menor espaçamento para os valores para o parâmetro, por exemplo poderiamos usar `thetaE.vals <- seq(0.04,0.20,l=1001)`.

6.2.2 Aproximação quadrática da verossimilhança

Nesta abordagem aproximamos a função deviance por uma função quadrática obtida pela expansão por série de Taylor ao redor do estimador de máxima verossimilhança:

$$D(\theta) \approx n\left(\frac{\theta-\hat{\theta}}{\hat{\theta}}\right)^2.$$

A Figura 10 obtida com os comandos mostra o gráfico desta função deviance aproximada. A Figura também mostra os IC's obtido com esta função. Para a aproximação quadrática os limites dos intervalos são facilmente determinados analiticamente e neste caso dados por:

$$\left(\hat{\theta}(1-\sqrt{c^*/n}) \;\;,\;\; \hat{\theta}(1+\sqrt{c^*/n})\right).$$

```
> devap.exp <- function(theta, n, xbar){n * (xbar *(theta - (1/xbar)))^2}
> devap.vals <- devap.exp(thetaE.vals, n=20, xbar=10)
> plot(thetaE.vals, devap.vals, ty='l',
+      xlab=expression(theta), ylab=expression(D(theta)))
> corte <-  qchisq(0.95, df = 1)
> abline(h = corte, lty=3)
> limites.devap <- c((1/10)*(1 - sqrt(corte/20)), (1/10)*(1 + sqrt(corte/20)))
> limites.devap
[1] 0.05617387 0.14382613
> segments(limites.devap, rep(corte,2), limites.devap, rep(0,2))
```

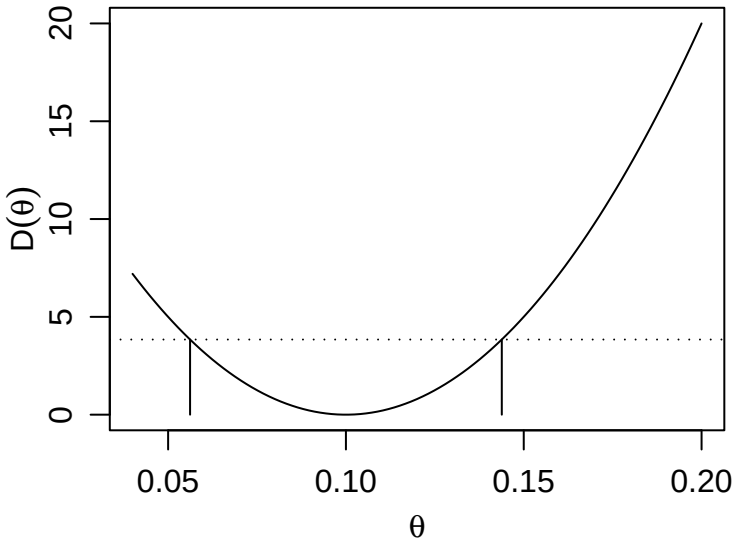

Figura 10: Função deviance obtida pela aproximação quadrática para Exp(θ) e uma amostra de tamanho 20 e média 10.

6.3 Comparando as duas estratégias

Examinando os limites dos intervalos encontrados anteriormente podemos ver que são diferentes. Vamos agora colocar os resultados pelos dois métodos em um mesmo gráfico (Figura 11) para comparar os resultados.

```
> plot(thetaE.vals, dev.vals, ty='l',
+      xlab=expression(theta), ylab=expression(D(theta)))
> lines(thetaE.vals, devap.vals, lty=2)
> abline(h = corte, lty=3)
> segments(limites.dev, rep(corte,2), limites.dev, rep(0,2))
> segments(limites.devap, rep(corte,2), limites.devap, rep(0,2), lty=2)
> legend(0.07, 12, c('deviance', 'aproximacão quadrática'), lty=c(1,2), cex=0.7)
```

Vamos agora examinar o efeito do tamanho da amostra na função deviance e sua aproximação quadrática. A Figura 7 mostra as funções para três tamanhos de amostra, $n = 10, 30$ e 100 que são obtidas com os comandos abaixo onde vemos que a aproximação fica cada vez melhor com o aumento do tamanho da amostra.

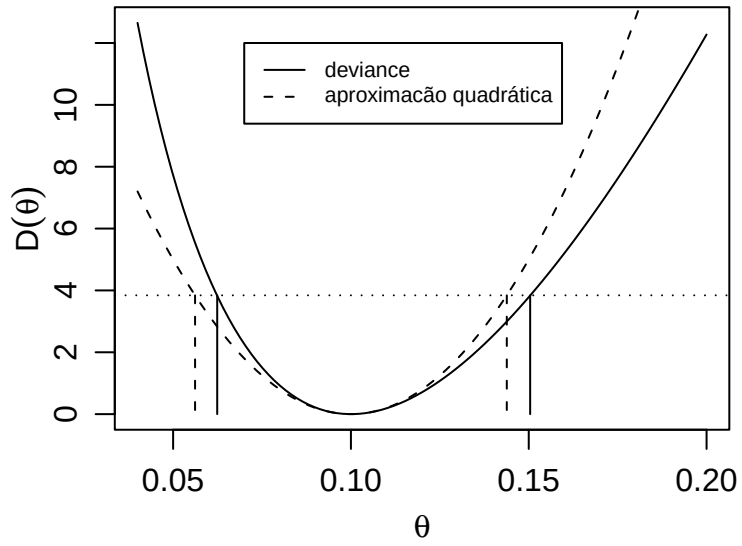

Figura 11: Comparação dos IC's de confiança obtidos pela solução gráfica/numérica (linha sólida) e pela aproximação quadrática (linha tracejada) para o parâmetro θ da Exp(θ) para uma amostra de tamanho 20 e média 10.

```
> thetaE.vals <- seq(0.04, 0.20, l=101)
> dev10.vals <- dev.exp(thetaE.vals, n=10, xbar=10)
> plot(thetaE.vals, dev10.vals, ty='l',
+      xlab=expression(theta), ylab=expression(D(theta)))
> devap10.vals <- devap.exp(thetaE.vals, n=10, xbar=10)
> lines(thetaE.vals, devap10.vals, lty=2)
> abline(h = qchisq(0.95, df = 1), lty=3)
>
> dev30.vals <- dev.exp(thetaE.vals, n=30, xbar=10)
> plot(thetaE.vals, dev30.vals, ty='l',
+      xlab=expression(theta), ylab=expression(D(theta)))
> devap30.vals <- devap.exp(thetaE.vals, n=30, xbar=10)
> lines(thetaE.vals, devap30.vals, lty=2)
> abline(h = qchisq(0.95, df = 1), lty=3)
>
> dev100.vals <- dev.exp(thetaE.vals, n=100, xbar=10)
> plot(thetaE.vals, dev100.vals, ty='l',
+      xlab=expression(theta), ylab=expression(D(theta)))
> devap100.vals <- devap.exp(thetaE.vals, n=100, xbar=10)
> lines(thetaE.vals, devap100.vals, lty=2)
> abline(h = qchisq(0.95, df = 1), lty=3)
```

6.4 Exercícios

1. Seja $14.1, 30.0, 19.6, 28.2, 12.5, 15.2, 17.1, 11.0, 25.9, 13.2, 22.8, 22.1$ a.a. de uma distribuição normal de média 20 e variância σ^2.

   (a) Obtenha a fução deviance para σ^2 e faą o seu gráfico.

   (b) Obtenha a fução deviance para σ e faą o seu gráfico.

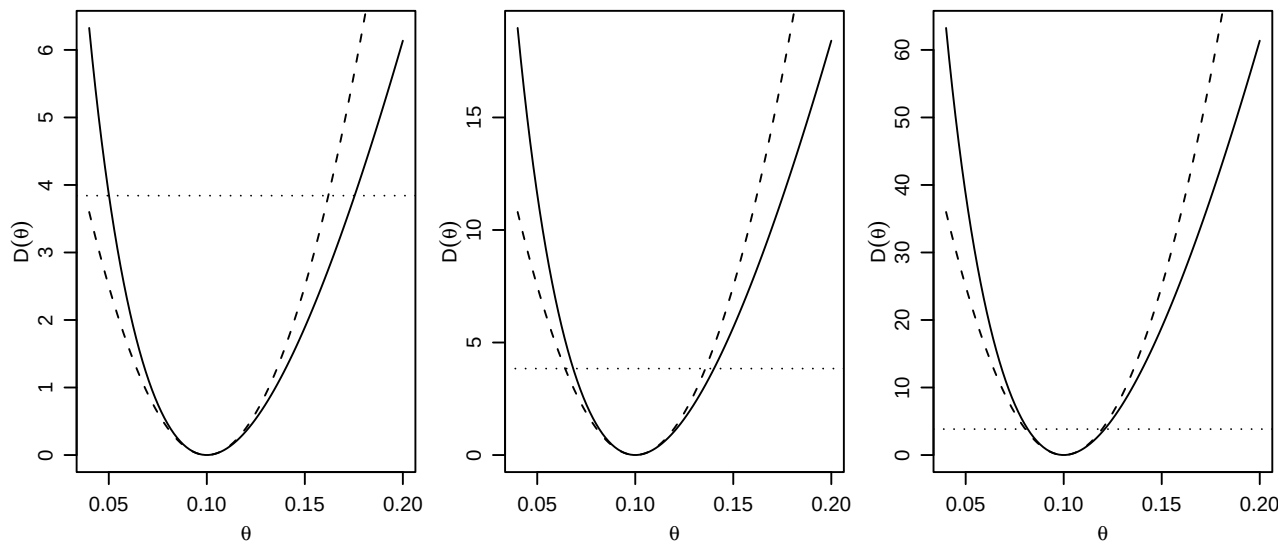

Figura 12: Funções deviance e deviance aproximada para o parâmetro θ da Exp(θ) em amostras de média 10 e tamanhos $n = 10$ (esquerda), 30 (centro) e 100 (direita).

(c) Obtenha os IC's a 90% de confiança.

2. Repita as análises mostradas no exemplo acima da distribuição exponencial mas agora utilizando a seguinte parametrização para a função de densidade:

$$f(x) = \frac{1}{\lambda} \exp \frac{x}{\lambda}.$$

Discuta as diferenças entre os resultados obtidos nas duas parametrizações.

7 Mais sobre intervalos de confiança

Nesta aula vamos nos aprofundar um pouco mais na teoria de intervalos de confiança. São ilustrados os conceitos de:

- obtenção de intervalos de confiança pelo método da quantidade pivotal,
- resultados diversos da teoria de verossimilhança,
- intervalos de cobertura.

Voce vai precisar conhecer de conceitos do método da quantidade pivotal, a propriedade de normalidade assintótica dos estimadores de máxima verossimilhança e a distribuição limite da função deviance.

7.1 Inferência para a distribuição Bernoulli

Os dados abaixo são uma amostra aleatória da distribuição $Bernoulli(p)$.

0 0 0 1 1 0 1 1 1 1 0 1 1 0 1 1 1 1 0 1 1 1 1 1 1

Desejamos obter:

(a) o gráfico da função de verossimilhança para p com base nestes dados

(b) o estimador de máxima verossimilhança de p, a informação observada e a informação de Fisher

(c) um intervalo de confiança de 95% para p baseado na normalidade assintótica de $\hat{p}$

(d) compare o intervalo obtido em (b) com um intervalo de confiança de 95% obtido com base na distribuição limite da função deviance

(e) a probabildade de cobertura dos intervalos obtidos em (c) e (d). (O verdadeiro valor de p é 0.8)

Primeiramente vamos entrar com os dados na forma de um vetor.

```
> y <- c(0,0,0,1,1,0,1,1,1,1,0,1,1,0,1,1,1,1,0,1,1,1,1,1,1)
```

(a)
Vamos escrever uma função para obter a função de verossimilhança.

```
> vero.binom <- function(p, dados){
+   n <- length(dados)
+   x <- sum(dados)
+   return(dbinom(x, size = n, prob = p, log = TRUE))
+ }
```

Esta função exige dados do tipo 0 ou 1 da distribuição Bernoulli. Entretanto às vezes temos dados Binomiais do tipo n e x (número x de sucessos em n observações). Por exemplo, para os dados acima teríamos $n = 25$ e $x = 18$. Vamos então escrever a função acima de forma mais geral de forma que possamos utilizar dados disponíveis tanto em um quanto em ou outro formato.

```
> vero.binom <- function(p, dados, n = length(dados), x = sum(dados)){
+   return(dbinom(x, size = n, prob = p, log = TRUE))
+ }
```

Agora vamos obter o gráfico da função de verossimilhança para estes dados. Uma forma de fazer isto é criar uma sequência de valores para o parâmetro p e calcular o valor da verossimilhança para cada um deles. Depois fazemos o gráfico dos valores obtidos contra os valores do parâmetro. No R isto pode ser feito com os comandos abaixo que produzem o gráfico mostrado na Figura 14.

```
> p.vals <- seq(0.01,0.99,l=99)
> logvero <- sapply(p.vals, vero.binom, dados=y)
> plot(p.vals, logvero, type="l")
```

Note que os três comandos acima podem ser substituídos por um único que produz o mesmo resultado:

```
> curve(vero.binom(x, dados=y), from = 0, to = 1)
```

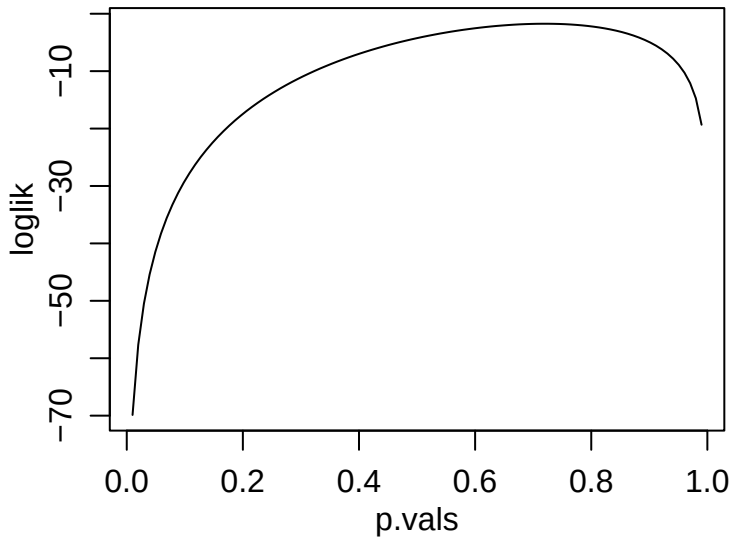

Figura 13: Função de verossimilhança para o parâmetro p da distribuição Bernoulli.

(b)
Dos resultados para distribuição Bernoulli sabemos que o estimador de máxima verossimilhança é dado por

$$\hat{p} = \frac{\sum_{i=1}^{n} y_i}{n}$$

e que a informação esperada coincide com a esperança observada e sendo iguais a:

$$I(\hat{p}) = \frac{n}{\hat{p}(1-\hat{p})}$$

. Para obter os valores numéricos para a amostra dada utilizamos os comandos:

```
> p.est <- mean(y)
> arrows(p.est, vero.binom(p.est,dados=y), p.est, min(logvero))
> io <- ie <- length(y)/(p.est * (1 - p.est))
```

```
> io
[1] 124.0079
> ie
[1] 124.0079
```

(c)

O intervalo de confiança baseado na normalidade assintótica do estimador de máxima verossimilhança é dado por:

$$\left(\hat{p} - z_{\alpha/2}\sqrt{I(\hat{p})} \ , \ \hat{p} + z_{\alpha/2}\sqrt{I(\hat{p})}\right)$$

e para obter o intervalo no R usamos os comandos a seguir.

```
> ic1.p <- p.est + qnorm(c(0.025, 0.975)) * sqrt(1/ie)
> ic1.p
[1] 0.5439957 0.8960043
```

(d)

Vamos agora obter o intervalo baseado na função deviance graficamente. Primeiro vamos escrever uma função para calcular a deviance que vamos chamar de `dev.binom`, lembrando que a deviance é definida pela expressão:

$$D(p) = 2\{l(\hat{p}) - l(p)\}.$$

```
> dev.binom <- function(p, dados, n = length(dados), x =sum(dados)){
+   p.est <- x/n
+   vero.p.est <- vero.binom(p.est, n = n, x = x)
+   dev <- 2 * (vero.p.est - vero.binom(p, n = n, x = x))
+   dev
+ }
```

E agora vamos fazer o gráfico de forma similar ao que fizemos para função de verossimilhança, definindo uma sequência de valores, calculando as valores da deviance e traçando a curva.

```
> p.vals <- seq(0.3, 0.95, l=101)
> dev.p <- dev.binom(p.vals, dados=y)
> plot(p.vals, dev.p, typ="l")
```

Agora usando esta função vamos obter o intervalo gráficamente. Para isto definimos o ponto de corte da função usando o fato que a função deviance $D(p)$ tem distribuição assintótica χ^2. Nos comandos a seguir primeiro encontramos o ponto de corte para o nível de confiança de 95%. Depois traçamos a linha de corte com o comando `abline`. Os comandos seguintes consistem em uma forma simples e aproximada para encontrar os pontos onde a linha conta a função, que definem o intervalo de confiança.

```
> corte <- qchisq(0.95, df=1)
> abline(h=corte)
> dif <- abs(dev.p - corte)
> inf <- ifelse(p.est==0, 0, p.vals[p.vals<p.est][which.min(dif[p.vals<p.est])])
> sup <- ifelse(p.est==1, 1, p.vals[p.vals>p.est][which.min(dif[p.vals>p.est])])
> ic2.p <- c(inf, sup)
> ic2.p
[1] 0.5275 0.8655
> segments(ic2.p, 0, ic2.p, corte)
```

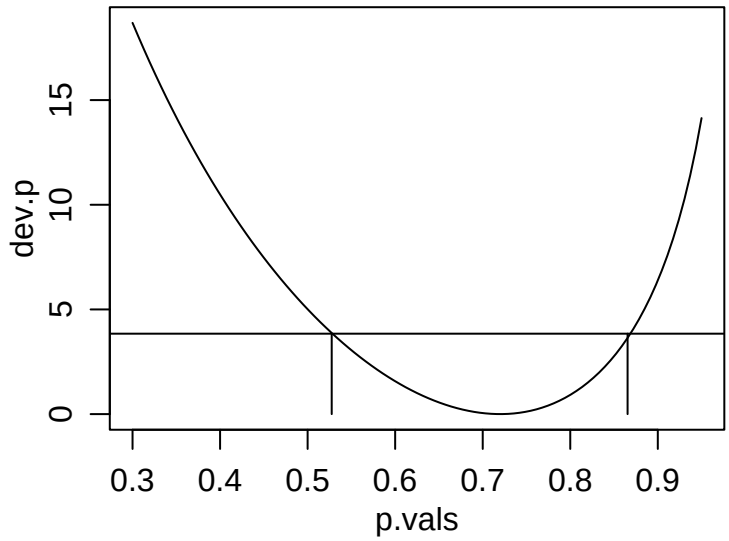

Figura 14: Função deviance para o parâmetro p da distribuição Bernoulli.

Agora que já vimos as duas formas de obter o IC passo a passo vamos usar os comandos acima para criar uma função geral para encontrar IC para qualquer conjunto de dados e com opções para os dois métodos.

```
> ic.binom <- function(dados, n=length(dados), x=sum(dados),
+                      nivel = 0.95,
+                      tipo=c("assintotico", "deviance")){
+   tipo <- match.arg(tipo)
+   alfa <- 1 - nivel
+   p.est <- x/n
+   if(tipo == "assintotico"){
+     se.p.est <- sqrt((p.est * (1 - p.est))/n)
+     ic <- p.est + qnorm(c(alfa/2, 1-(alfa/2))) * se.p.est
+   }
+   if(tipo == "deviance"){
+     p.vals <- seq(0,1,l=1001)
+     dev.p <- dev.binom(p.vals, n = n, x = x)
+     corte <- qchisq(nivel, df=1)
+     dif <- abs(dev.p - corte)
+     inf <- ifelse(p.est==0, 0, p.vals[p.vals<p.est][which.min(dif[p.vals<p.est])])
+     sup <- ifelse(p.est==1, 1, p.vals[p.vals>p.est][which.min(dif[p.vals>p.est])])
+     ic <- c(inf, sup)
+   }
+   names(ic) <- c("lim.inf", "lim.sup")
+   ic
+ }
```

E agora vamos utilizar a função, primeiro com a aproximação assintótica e depois pela deviance. Note que os intervalos são diferentes!

```
> ic.binom(dados=y)
  lim.inf   lim.sup
0.5439957 0.8960043
```

```
> ic.binom(dados=y, tipo = "dev")
lim.inf lim.sup
  0.528   0.869
```

(e)
O cálculo do intervalo de cobertura consiste em:

1. simular dados com o valor especificado do parâmetro;
2. obter o intervalo de confiança;
3. verificar se o valor está dentro do intervalo
4. repetir (1) a (3) e verificar a proporção de simulações onde o valor está no intervalo.

Espera-se que a proporção obtida seja o mais próximo possível do nível de confiança definido para o intervalo.

Para isto vamos escrever uma função implementando estes passos e que utiliza internamente `ic.binom` definida acima.

```
> cobertura.binom <- function(n, p, nsim, ...){
+   conta <- 0
+   for(i in 1:nsim){
+     ysim <- rbinom(1, size = n, prob = p)
+     ic <- ic.binom(n = n, x = ysim, ...)
+     if(p > ic[1] & p < ic[2]) conta <- conta+1
+   }
+   return(conta/nsim)
+ }
```

E agora vamos utilizar esta função para cada um dos métodos de obtenção dos intervalos.

```
> set.seed(123)
> cobertura.binom(n=length(y), p=0.8, nsim=1000)
[1] 0.885
> set.seed(123)
> cobertura.binom(n=length(y), p=0.8, nsim=1000, tipo = "dev")
[1] 0.954
```

Note que a cobertura do método baseado na deviance é muito mais próxima do nível de 95% o que pode ser explicado pelo tamanho da amostra. O IC assintótico tende a se aproximar do nível nominal de confiança na medida que a amostra cresce.

7.2 Exercícios

1. Re-faça o ítem (e) do exemplo acima com $n = 10$, $n = 50$ e $n = 200$. Discuta os resultados.
2. Seja $X_1, X_2, \cdots, X_n$ uma amostra aleatória da distribuição $U(0, \theta)$. Encontre uma quantidade pivotal e:
   (a) construa um intervalo de confiança de 90% para θ
   (b) construa um intervalo de confiança de 90% para $\log \theta$

(c) gere uma amostra de tamanho $n = 10$ da distribuição $U(0, \theta)$ com $\theta = 1$ e obtenha o intervalo de confiança de 90% para θ. Verifique se o intervalo cobre o verdadeiro valor de θ.

(d) verifique se a probabilidade de cobertura do intervalo é consistente com o valor declarado de 90%. Para isto gere 1000 amostras de tamanho $n = 10$. Calcule intervalos de confiança de 90% para cada uma das amostras geradas e finalmente, obtenha a proporção dos intervalos que cobrem o verdadeiro valor de θ. Espera-se que este valor seja próximo do nível de confiança fixado de 90%.

(e) repita o item (d) para amostras de tamanho $n = 100$. Houve alguma mudança na probabilidade de cobertura?

Note que se $-\sum_i^n \log F(x_i; \theta) \sim \Gamma(n, 1)$ então $-2\sum_i^n \log F(x_i; \theta) \sim \chi^2_{2n}$.

3. Acredita-se que o número de trens atrasados para uma certa estação de trem por dia segue uma distribuição Poisson(θ), além disso acredita-se que o número de trens atrasados em cada dia seja independente do valor de todos os outros dias. Em 10 dias sucessivos, o número de trens atrasados foi registrado em:

```
5 0 3 2 1 2 1 1 2 1
```

Obtenha:

(a) o gráfico da função de verossimilhança para θ com base nestes dados

(b) o estimador de máxima verossimilhança de θ, a informação observada e a informação de Fisher

(c) um intervalo de confiança de 95% para o número médio de trens atrasados por dia baseando-se na normalidade assintótica de $\hat{\theta}$

(d) compare o intervalo obtido em (c) com um intervalo de confiança obtido com base na distribuição limite da função deviance

(e) o estimador de máxima verossimilhança de ϕ, onde ϕ é a probabilidade de que não hajam trens atrasados num particular dia. Construa intervalos de confiança de 95% para ϕ como nos itens (c) e (d).

4. Encontre intervalos de confiança de 95% para a média de uma distribuição Normal com variância 1 dada a amostra

```
9.5 10.8 9.3 10.7 10.9 10.5 10.7 9.0 11.0 8.4
10.9 9.8 11.4 10.6 9.2 9.7 8.3 10.8 9.8 9.0
```

baseando-se:

(a) na distribuição assintótica de $\hat{\mu}$

(b) na distribuição limite da função deviance

5. Acredita-se que a produção de trigo, X_i, da área i é normalmente distribuída com média θz_i, onde z_i é quantidade (conhecida) de fertilizante utilizado na área. Assumindo que as produções em diferentes áreas são independentes, e que a variância é conhecida e igual a 1, ou seja, $X_i \sim N(\theta z_i, 1)$, para $i = 1, \cdots, n$:

(a) simule dados sob esta distribuição assumindo que $\theta = 1.5$, e $z = (1, 2, 3, 4, 5)$. Visualize os dados simulados através de um gráfico de $(z \times x)$

(b) encontre o EMV de θ, $\hat{\theta}$

(c) mostre que $\hat{\theta}$ é um estimador não viciado para θ (lembre-se que os valores de z_i são constantes)

(d) obtenha um intervalo de aproximadamente 95% de confiança para θ baseado na distribuição assintótica de $\hat{\theta}$

8 Testes de hipótese - exemplos iniciais

Os exercícios abaixo são referentes ao conteúdo de *Testes de Hipóteses* conforme visto na disciplina de Estatística Geral II. Eles devem ser resolvidos usando como referência qualquer texto de Estatística Básica. Procure resolver primeiramente na *"na mão"*, sem o uso de programa estatístico.

A idéia é relembrar como são feitos alguns testes de hipótese básicos e corriqueiros em estatística. Vamos também em alguns exercícios explorar as funções de verossimilhança e deviance.

Nesta sessão vamos verificar como utilizar o R para fazer teste de hipóteses sobre parâmetros de distribuições para as quais os resultados são bem conhecidos.

Os comandos e cálculos são bastante parecidos com os vistos em intervalos de confiança e isto nem poderia ser diferente visto que intervalos de confiança e testes de hipótese são relacionados.

Assim como fizemos com intervalos de confiança, aqui sempre que possível e para fins didáticos mostrando os recursos do R vamos mostrar três possíveis soluções:

1. fazendo as contas passo a passo, utilizando o R como uma calculadora
2. escrevendo uma função
3. usando uma função já existente no R

8.1 Comparação de variâncias de uma distribuição normal

Queremos verificar se duas máquinas produzem peças com a mesma homogeneidade quanto a resistência à tensão. Para isso, sorteamos dias amostras de 6 peças de cada máquina, e obtivemos as seguintes resistências:

Máquina A	145	127	136	142	141	137
Máquina B	143	128	132	138	142	132

O que se pode concluir fazendo um teste de hipótese adequado?

Solução:

Da teoria de testes de hipótese sabemos que, assumindo a distribuição normal, o teste para a hipótese:

$$H_0 : \sigma_A^2 = \sigma_B^2 \quad versus \quad H_a : \sigma_A^2 \neq \sigma_B^2$$

que é equivalente à

$$H_0 : \frac{\sigma_A^2}{\sigma_B^2} = 1 \quad versus \quad H_a : \frac{\sigma_A^2}{\sigma_B^2} \neq 1$$

é feito calculando-se a estatística de teste:

$$F_{calc} = \frac{S_A^2}{S_B^2}$$

e em seguida comparando-se este valor com um valor da tabela de F e/ou calculando-se o P-valor associado com $n_A - 1$ e $n_B - 1$ graus de liberdade. Devemos também fixar o nível de significância do teste, que neste caso vamos definir como sendo 5%.

Para efetuar as análises no R vamos primeiro entrar com os dados nos objetos que vamos chamar de `ma` e `mb` e calcular os tamanhos das amostras que vão ser armazenados nos objetos `na` e `nb`.

```
> ma <- c(145, 127, 136, 142, 141, 137)
> na <- length(ma)
> na
[1] 6
> mb <- c(143, 128, 132, 138, 142, 132)
> nb <- length(mb)
> nb
[1] 6
```

8.1.1 Fazendo as contas passo a passo

Vamos calcular a estatística de teste. Como temos o computador a disposição não precisamos de da tabela da distribuição F e podemos calcular o P-valor diretamente.

```
> ma.v <- var(ma)
> ma.v
[1] 40
> mb.v <- var(mb)
> mb.v
[1] 36.96667
> fcalc <- ma.v/mb.v
> fcalc
[1] 1.082056
> pval <- 2 * pf(fcalc, na-1, nb-1, lower=F)
> pval
[1] 0.9331458
```

No cálculo do P-valor acima multiplicamos o valor encontrado por 2 porque estamos realizando um teste bilateral.

8.1.2 Escrevendo uma função

Esta fica por sua conta!
Escreva a sua própria função para testar hipóteses sobre variâncias de duas distribuições normais.

8.1.3 Usando uma função do R

O R já tem implementadas funções para a maioria dos procedimentos estatísticos "usuais". Por exemplo, para testar variâncias neste exemplo utilizamos a função `var.test`. Vamos verificar os argumentos da função.

```
> args(var.test)
function (x, ...)
NULL
```

Note que esta saída não é muito informativa. Este tipo de resultado indica que `var.test` é um método com mais de uma função associada. Portanto devemos pedir os argumentos da função "default".

```
> args(var.test.default)
function (x, y, ratio = 1, alternative = c("two.sided", "less",
    "greater"), conf.level = 0.95, ...)
NULL
```

Neste argumentos vemos que a função recebe dois vertores de de dados (x e y), que por "default" a hipótese nula é que o quociente das variâncias é 1 e que a alternativa pode ser bilateral ou unilateral. Como `''two.sided''` é a primeira opção o "default" é o teste bilateral. Finalmente o nível de confiança é 95% ao menos que o último argumento seja modificado pelo usuário. Para aplicar esta função nos nossos dados basta digitar:

```
> var.test(ma, mb)

        F test to compare two variances

data:  ma and mb
F = 1.0821, num df = 5, denom df = 5, p-value = 0.9331
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1514131 7.7327847
sample estimates:
ratio of variances
          1.082056
```

e note que a saída inclui os resultados do teste de hipótese bem como o intervalo de confiança. A decisão baseia-se em verificar se o P-valor é menor que o definido inicialmente.

8.2 Exercícios

Note que nos exercícios abaixo nem sempre voce poderá usar funções de teste do R porque em alguns casos os dados brutos não estão disponíveis. Netes casos voce deverá fazer os cálculos usando o R como calculadora.

1. Uma máquina automática de encher pacotes de café enche-os segundo uma distribuição normal, com média μ e variância $400g^2$. O valor de μ pode ser fixado num mostrador situado numa posição um pouco inacessível dessa máquina. A máquina foi regulada para $\mu = 500g$. Desejamos, de meia em meia hora, colher uma amostra de 16 pacotes e verificar se a produção está sob controle, isto é, se $\mu = 500g$ ou não. Se uma dessas amostras apresentasse uma média $\bar{x} = 492g$, voce pararia ou não a produção para verificar se o mostrador está na posição correta? Construa os gráficos da funções de verossimilhança e deviance e indique no gráfico o valor referente à hipótese nula.

2. Uma companhia de cigarros anuncia que o índice médio de nicotina dos cigarros que fabrica apresenta-se abaixo de $23mg$ por cigarro. Um laboratório realiza 6 análises desse índice, obtendo: 27, 24, 21, 25, 26, 22. Sabe-se que o índice de nicotina se distribui normalmente, com variância igual a $4,86mg^2$. Pode-se aceitar, ao nível de 10%, a afirmação do fabricante. Construa os gráficos da funções de verossimilhança e deviance e indique no gráfico o valor referente à hipótese nula.

3. Uma estação de televisão afirma que 60% dos televisores estavam ligados no seu programa especial de última segunda feira. Uma rede competidora deseja contestar essa afirmação, e decide, para isso, usar uma amostra de 200 famílias obtendo 104 respostas afirmativas.

   - Qual a conclusão ao nível de 5% de significância?

   (a) Construa a curva de verossimilhança indicando o valor associado à informação do fabricante.

(b) Indique também no gráfico valores de verossimilhança associados à hipóteses de p=0.45, p=0.50 e p=0.55.

(c) Construa novamente a curva para uma situação onde são entrevistadas 100 famílias com 52 respostas afirmativas. Compare esta curva com a anterior e tire conclusões.

(d) Construa novamente a curva para uma situação onde são entrevistadas 200 famílias com 98 respostas afirmativas. Compare esta curva com as anteriores e tire conclusões.

4. O tempo médio, por operário, para executar uma tarefa, tem sido 100 minutos, com um desvio padrão de 15 minutos. Introduziu-se uma modificação para diminuir esse tempo, e, após certo período, sorteou-se uma amostra de 16 operários, medindo-se o tempo de execução de cada um. O tempo médio da amostra foi de 85 minutos, o o desvio padrão foi 12 minutos. Estes resultados trazem evidências estatísticas da melhora desejada?

5. Num estudo comparativo do tempo médio de adaptação, uma amostra aleatória, de 50 homens e 50 mulheres de um grande complexo industrial, produziu os seguintes resultados:

Estatísticas	Homens	Mulheres
Médias	3,2 anos	3,7 anos
Desvios Padrões	0,8 anos	0,9 anos

Pode-se dizer que existe diferença significativa entre o tempo de adaptação de homens e mulheres?

A sua conclusão seria diferente se as amostras tivessem sido de 5 homens e 5 mulheres?

9 Função Poder

Nesta aula vamos utilizar o R para ilustrar alguns conceitos da teoria de testes de hipóteses e em particular relacionados com a função poder do teste.

9.1 Distribuição normal com variância conhecida

Seja $X_1, X_2, \ldots, X_n$ uma amostra aleatória da distribuição normal com média θ e variância conhecida igual à 25. Considere a hipótese nula $H_0 : \theta \leq 17$ e o teste:
Rejeita-se H_0 se e somente se $\bar{x} > 17 + \frac{5}{\sqrt{n}}$.

1. Construa a função poder e calcule o tamanho do teste para $n = 25$.
2. Compare graficamente a função poder para diferentes valores de tamanho de amostra, $n = 5, 10, 20, 30, 50$.

A função poder $\gamma(\theta)$ é dada por

$$\begin{aligned}\gamma(\theta) &= P_\theta[Rej.\ H_0] = P_\theta\left[\bar{x} > 17 + \frac{5}{\sqrt{n}}\right] \\ &= 1 - P_\theta\left[\bar{x} \leq 17 + \frac{5}{\sqrt{n}}\right]\end{aligned}$$

Como $X_i \sim N(\theta, 25)$ sabemos que $\bar{X} \sim N(\theta, 5)$ e, padronizando a variável temos que $z = \frac{\bar{x}-\theta}{s/\sqrt{n}} \sim N(0,1)$. Portanto podemos escrever a função poder como

$$\begin{aligned}\gamma(\theta) &= 1 - P\left[Z \leq \frac{17 + \frac{5}{\sqrt{n}} - \theta}{5/\sqrt{n}}\right] \\ &= 1 - P[Z \leq q] = 1 - \Phi(q)\end{aligned}$$

onde $q = \frac{17+\frac{5}{\sqrt{n}}-\theta}{5/\sqrt{n}}$.

Vamos agora utilizar o R para fazer o gráfico da função poder. Primeiro definimos valores de θ, depois calculamos os quantis q correspondentes a estes valores, para cada quantil usamos a função `pnorm()` para calcular o poder e por fim fazemos o gráfico. Podemos usar os seguintes comandos.

```
theta <- seq(13, 22, l=100)
q <- (17 + (5/sqrt(25)) - theta)/(5/sqrt(25))
poder <- 1 - pnorm(q)
plot(theta, poder, ty="l", xlab = expression(theta),
     ylab = expression(gamma(theta)))
```

O gráfico da função poder é mostrado na Figura 15.

Vamos agora calcular o tamanho do teste α que é dado por

$$\alpha = \sup_{\theta \in \Theta_0} \gamma(\theta)$$

Portanto para este exemplo temos:

$$\begin{aligned}\alpha &= \sup_{\theta \leq 17}\left[P_\theta(\bar{x} > 17 + \frac{5}{\sqrt{n}}\right] \\ &= \sup_{\theta \leq 17}\left[1 - P_\theta(\bar{x} \leq 17 + \frac{5}{\sqrt{n}}\right] \\ &= 1 - P[Z < 1] = 1 - \Phi(1) = 0.159\end{aligned}$$

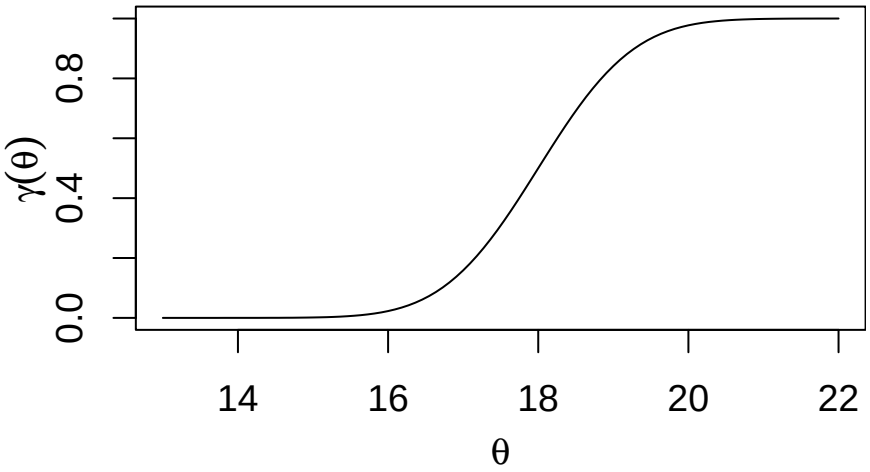

Figura 15: Função poder para $n = 25$.

Pode-se ainda usar a função `lines` para adicionar a este gráfico uma outra função com um outro valor n de tamanho de amostra. Por exemplo, para $n = 10$, executando os comandos abaixo obtemos o gráfico indicado na Figura 16.

```
q <- (17 + (5/sqrt(10)) - theta)/(5/sqrt(10))
poder <- 1 - pnorm(q)
lines(theta, poder, lty=2)
legend(14, 1, c("n = 10", "n = 25"), lty=c(2,1))
```

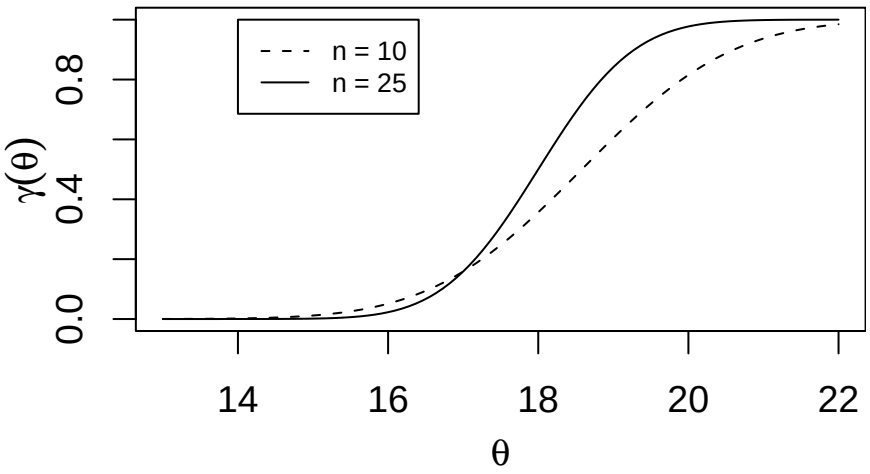

Figura 16: Função poder para $n = 10$ e $n = 25$.

Uma solução um pouco mais elegante no R é escrever uma função para plotar o função poder e depois rodar esta função:

```
poder.f <- function(n, t.min, t.max){
  theta <- seq(t.min, t.max, l=100)
  q <- (17 + (5/sqrt(n)) - theta)/(5/sqrt(n))
  poder <- 1 - pnorm(q)
  plot(theta, poder, ty = "l", xlab = expression(theta),
       ylab = expression(gamma(theta)))
}
```

```
poder.f(25, 10, 25)
poder.f(25, 14, 22)
```

A função acima tem 3 argumentos: o tamanho da amostra e os valores mínimos e máximos para θ. Ao chamar esta função o gráfico é automaticamente mostrado na janela gráfica.

Agora vamos sofisticar a função mais um pouco. Vamos adicionar o argumento `add` para permitir adicionar uma função a um gráfico já existente. Além disto vamos usar o mecanismo de ... para poder passar argumentos de tipo de linhas, cores, etc.

```
poder.f <- function(n, t.min, t.max, add = FALSE, ...){
  theta <- seq(t.min, t.max, l=100)
  q <- (17 + (5/sqrt(n)) - theta)/(5/sqrt(n))
  poder <- 1 - pnorm(q)
  if(add)
  lines(theta, poder, ...)
    else
  plot(theta, poder, ty="l", xlab=expression(theta),
       ylab=expression(gamma(theta)), ...)
}
```

E usando a função com os comandos abaixo obtemos o gráfico mostrado na Figura 17.

```
poder.f(5, 14, 24)
poder.f(10, 14, 24, add = T, lty = 2)
poder.f(20, 14, 24, add = T, col = 2)
poder.f(30, 14, 24, add = T, lty = 2, col = 2)
poder.f(50, 14, 24, add = T, col = 3)
legend(20, 0.3, c("n = 5","n = 10","n = 20","n = 30","n = 50"),
       lty=c(1,2,1,2,1), col=c(1,1,2,2,3))
```

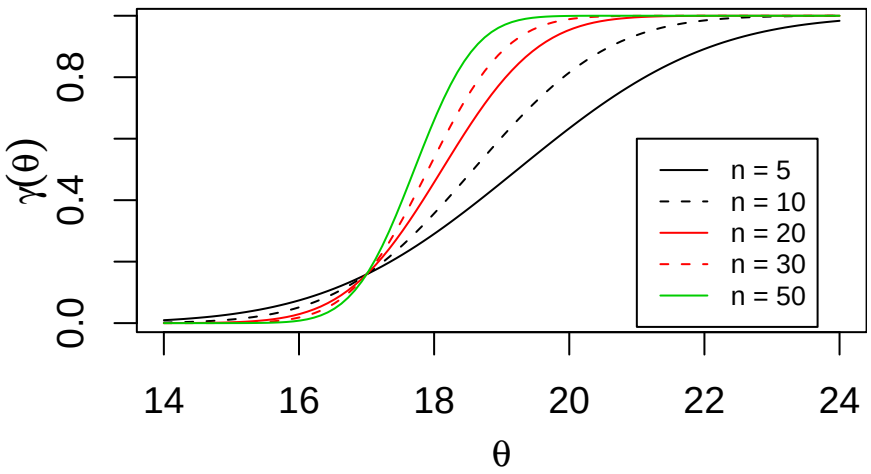

Figura 17: Função poder para diferentes tamanhos de amostra.

9.2 Exercícios

1. Um valor y amostrado de uma variável aleatória com distribuição $N(\theta, 1)$ é usado para testar a hipótese $H_0 : \theta \leq 0$ *vs* $H_1 : \theta > 0$. Define-se que a região de aceitação é dada por $y : y < 1/2$. Faça o gráfico da função poder e calcule o tamanho do teste.

2. Suponha que uma amostra $X_1, X_2, \ldots X_3$ é retirada de uma distribuição uniforme no intervalo $(0, \theta)$, onde o valor de θ $(\theta > 0)$ é desconhecido. Deseja-se testar a hipótese:

$$\begin{cases} H_0 : & 3 \leq \theta \leq 4 \\ H_1 : & \theta < 3 \text{ ou } \theta > 4 \end{cases}$$

   Sabemos que o EMV de θ é $Y_n = \max(X_1, X_2, \ldots X_n)$. Define-se a região de crítica do teste como $\{Y_n : Y_n < 2.9 \text{ ou } Y_n > 4\}$. Obtenha para $n = 68$ um gráfico da função poder e calcule o tamanho do teste.

3. Suponha que a proporção p de ítens defeituosos em uma população de ítens é desconhecida, e deseja-se testar a seguinte hipótese:

$$\begin{cases} H_0 : & p = 0.2 \\ H_1 : & p \neq 0.2. \end{cases}$$

   Suponha ainda que uma amostra aleatória de 20 ítens é retirada desta população. Denote por y o número de ítens defeituosos na amostra e considere um teste cuja a região crítica é definida por $\{y : y \geq 7 \text{ ou } y \leq 1\}$. Faça um gráfico da função poder e determine o tamanho do teste.

10 Testes mais poderosos

O objetivo desta aula é revisar os conceitos de testes mais poderosos e testes uniformemente mais poderosos utilizando o programa R quando necessário.

10.1 Exercícios

1. Seja $X_1, X_2, \cdots, X_n$ uma amostra aleatória da distribuição Exp(θ), onde $\theta = 0.50$ ou $\theta = 0.75$.

   (a) Deseja-se testar $H_0 : \theta = 0.50 \times H_1 : \theta = 0.75$ ao nível de significância de 5%. Construa um teste adequado.

   (b) Tomando-se uma amostra de tamanho 10 dessa distribuição obteve-se os seguintes valores:

   ```
   0.97 0.36 1.45 0.66 1.73 1.44 0.41 1.87 2.13 0.18
   ```

   Com base nesta amostra e utilizando o teste construído em (a) temos ou não temos evidência para rejeitar H_0?

   (c) Qual o poder deste teste?

   (d) O teste das hipóteses $H_0 : \theta \leq 0.50 \times H_1 : \theta > 0.5$ seria diferente do teste obtido em (a)?

2. Seja $X_1, X_2, \cdots, X_n$ uma amostra aleatória da $N(\theta, 25)$.

   (a) Encontre o teste mais poderoso de tamanho $\alpha = 0.1$ com $n = 100$ para testar $H_0 : \theta = 17 \times H_1 : \theta = 19$

   (b) Com base no teste construído em (a) qual seria sua decisão quanto à rejeição ou não da hipótese H_0 caso uma amostra de tamanho $n = 100$ dessa distribuição apresentasse uma média amostral $\bar{x} = 18.23$?

   (c) Qual o poder desse teste?

   (d) Encontre o teste uniformemente mais poderoso de tamanho $\alpha = 0.1$ com $n = 100$ para testar $H_0 : \theta \leq 17 \times H_1 : \theta > 17$.

   (e) Construa a curva de poder desse teste. Compare essa curva com a de um teste alternativo onde H_0 é rejeitada se $X_1 \geq 17$.

   (f) Com base no teste construído em (d) a sua decisão quanto à rejeição ou não da hipótese H_0 mudaria com base na amostra de tamanho $n = 100$ dessa distribuição de média amostral $\bar{x} = 18.23$?

3. Seja $X_1, X_2, \cdots, X_n$ uma amostra aleatória da Bernoulli(θ).

   (a) Encontre o teste mais poderoso de tamanho $\alpha = 0.05$ com $n = 10$ para testar $H_0 : \theta = 0.75 \times H_1 : \theta = 0.5$.

   (b) Com base no teste construído em (a) qual seria sua decisão quanto à rejeição ou não da hipótese H_0 caso tivessemos uma amostra de tamanho $n = 10$ dessa distribuição, com proporção amostral de sucesso $\hat{\theta} = 0.6$?

   (c) Encontre o teste UMP de tamanho $\alpha = 0.05$ com $n = 10$ para testar $H_0 : \theta \geq 0.75 \times H_1 : \theta < 0.75$.

4. Seja $X_1, X_2, \cdots, X_n$ uma amostra aleatória da distribuição Poisson(θ). Encontre o teste UMP de $H_0 : \theta \geq \theta_0 \times H_1 : \theta < \theta_0$, e trace a função poder para $\theta_0 = 1$ e $n = 25$ (com $\alpha = 0.05$).