Curso de Extensão - R

Paulo Justiniano Ribeiro Junior

Última atualização: 12 de março de 2003

1 Uma primeira sessão com o R

Antes de mais nada vamos "experimentar o R", para ter uma amostra de seus recursos e a forma de trabalhar no programa.
Vamos rodar e estudar os resultados dos comandos abaixo para nos familiarizar com o programa. Depois iremos ver com mais detalhe o uso do R.

Inicie o R em seu computador.
Voce verá uma *janela de comandos* com o símbolo >.
Este é o *prompt* do R indicando que o programa está pronto para receber comandos.

A seguir digite os comandos abaixo. Qualquer texto iniciando com o símbolo "#" é entendido pelo programa como um comentário e portanto ignorado.

```
# gerando dois vetores de coordenadas x e y de números psedo-aleatórios
x <- rnorm(50)
y <- rnorm(x)

# colocando os pontos em um gráfico.
# Note que a janela gráfica se abrirá autometicamente
plot(x, y)

# verificando os objetos existentes na área de trabalho
ls()

# removendo objetos que não são mais necessários
rm(x, y)

# criando um vetor com uma sequência de números de 1 a 20
x <- 1:20

# um vetor de pesos com os desvios padrões de cada observação
w <- 1 + sqrt(x)/2

# montando um 'data-frame' de 2 colunas, x e y, e inspecionando o objeto
dummy <- data.frame(x=x, y= x + rnorm(x)*w)
dummy

# Ajustando uma regressão linear simples de y em x e examinando os resultados
fm <- lm(y ~ x, data=dummy)
summary(fm)
```

```
# como nós sabemos os pesos podemos fazer uma regressão ponderada
fm1 <- lm(y ~ x, data=dummy, weight=1/w^2)
summary(fm1)

# tornando visíveis as colunas do data-frame
attach(dummy)

# fazendo uma regressão local não-paramétrica
lrf <- lowess(x, y)

# plotando os pontos
plot(x, y)

# adicionando a linha de regressão local ...
lines(x, lrf$y)

# ... e a linha de regressão verdadeira (intercepto 0 e inclinação 1)
abline(0, 1, lty=3)

# a linha da regressão sem ponderação
abline(coef(fm))

# e a linha de regressão ponderada.
abline(coef(fm1), col = "red")

# removendo o objeto do caminho de procura
detach()

# O gráfico diagnóstico padrão para checar homocedasticidade.
plot(fitted(fm), resid(fm),
     xlab="Fitted values",
     ylab="Residuals",
     main="Residuals vs Fitted")

# gráficos de escores normais para checar assimetria, curtose e outliers (não muito útil
qqnorm(resid(fm), main="Residuals Rankit Plot")

# ‘‘limpando’’ novamente (apagando objetos)
rm(fm, fm1, lrf, x, dummy)
```

Agora vamos inspecionar dados do experimento clássico de Michaelson e Morley para medir a velocidade da luz.
Clique aqui para ver o arquivo de dados.
Gravar este arquivo no diretório `temp`.

```
# para ver o arquivo digite:
file.show(‘‘c:\\temp\\morley.tab.txt’’)

# Lendo so dados como um  data-frame e olhando os dados.
```

```
# Há 5 experimentos (coluna Expt) e cada um com 20 ``rodadas''(coluna
# Run) e sl é o valor medido da velocidade da luz numa escala apropriada
mm <- read.table("c:\\temp\\morley.tab.txt")
mm

# definindo Expt e Run como fatores
mm$Expt <- factor(mm$Expt)
mm$Run <- factor(mm$Run)

# tornando o data-frame visível na posição 2 do caminho de procura (default)
attach(mm)

# comparando os 5 experimentos
plot(Expt, Speed, main="Speed of Light Data", xlab="Experiment No.")

# análisando como blocos ao acaso com `runs' and `experiments' como
fatores e inspecionando resultados
fm <- aov(Speed ~ Run + Expt, data=mm)
summary(fm)
names(fm)
fm$coef

# ajustando um sub-modelo sem ``runs''e comparando via análise de variância
fm0 <- update(fm, . ~ . - Run)
anova(fm0, fm)

# desanexando o objeto e limpando novamente
detach()
rm(fm, fm0)
```

Finalmente, vamos ver alguns gráficos: contour e image plots.

```
# x é um vetor de 50 valores igualmente espaçados  no intervalo [-pi\, pi]. y idem.
x <- seq(-pi, pi, len=50)
y <- x

# f é uma matrix quadrada com linhas e colunas indexadas por x e y respectivamente
# com os valores da função cos(y)/(1 + x^2).
f <- outer(x, y, function(x, y) cos(y)/(1 + x^2))

# gravando parâmetros gráficos e definindo a região gráfica como quadrada
oldpar <- par(no.readonly = TRUE)
par(pty="s")

# fazendo um mapa de contorno de f e depois adicionando mais linhas para maiores detalhes
contour(x, y, f)
contour(x, y, f, nlevels=15, add=TRUE)

#  fa 's a ``parte assimétrica''. (t() é transposição).
fa <- (f-t(f))/2
```

```
# fazendo um mapa de contorno
contour(x, y, fa, nlevels=15)

#    ...  e restaurando parâmetros gráficos iniciais
par(oldpar)

# Fazendo um gráfico de imagem
image(x, y, f)
image(x, y, fa)

# e apagando objetos novamente.
objects(); rm(x, y, f, fa)
... and clean up before moving on.


# O R pode fazer operação com complexos
th <- seq(-pi, pi, len=100)
#  1i denota o número complexo  i.
z <- exp(1i*th)

# plotando complexos significa parte imaginária versus real
# Isto deve ser um círculo:
par(pty="s")
plot(z, type="l")

# Suponha que desejamos amostrar pontos dentro do círculo de raio unitário.
# uma forma simples de fazer isto é tomar números complexos com parte
# real e imaginária padrão
w <- rnorm(100) + rnorm(100)*1i

#  ...  e para mapear qualquer externo ao círculo no seu recíproco:
w <- ifelse(Mod(w) > 1, 1/w, w)

# todos os pontos estão dentro do círculo unitário, mas a distribuição
# não é uniforme.
plot(w, xlim=c(-1,1), ylim=c(-1,1), pch="+",xlab="x", ylab="y")
lines(z)

# este segundo método usa a distribuição uniforme.
# os pontos devem estar melhor distribuídos sobre o círculo
w <- sqrt(runif(100))*exp(2*pi*runif(100)*1i)
plot(w, xlim=c(-1,1), ylim=c(-1,1), pch="+", xlab="x", ylab="y")
lines(z)

# apagando os objetos
rm(th, w, z)

# saindo do R
q()
```

2 Recursos do R

O projeto R

O programa R é gratuito e de codigo aberto e a página oficial do projeto está em:
http://www.r-project.org.

Há também 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 em
ftp://est.ufpr.br/R

A página do R possui uma diversidade de recursos que serão comentados no curso.

Demos

O R vem com algumas demostrações (*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 netre 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 menssagem:

  ```
  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 menssagem:

  ```
  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!

Um tutorial sobre o R

Há um Tutorial de Introdução ao Rdisponível em http://www.est.ufpr.br/Rtutorial.

RWeb

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

Para acessar o **RWeb** vá até a página do R e 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. Os participantes do curso são estimulados a explorar os outros recursos da página.

3 Uma análise descritiva

Vamos agora efetuar algumas análises em um conjunto de dados.
Para isto vamos utilizar um conjunto de dados que já vem disponível com o R - o conjunto `airquality`.

Estes dados são medidas de: concentração de ozônio, radiação solar, velocidade de ventos e temperatura coletados diariamente por cinco meses.

Primeiramente vamos carregar e visualisar os dados com os comandos:

```
data(airquality)       # carrega os dados
airquality             # mostra os dados
```

Vamos agora usar alguns comandos para "conhecer melhor"os dados:

```
is.data.frame(airquality)    # verifica se é um data-frame
names(airquality)            # nome das colunas (variáveis)
dim(airquality)              # dimensões do data-frame
help(airquality)             # mostra o ``help''que explica os dados
```

Bem agora que conhecemos melhor o conjunto `airquality`, sabemos o número de dados, seu formato, o número de nome das variáveis podemos começar a analisá-los.

Veja por exemplo alguns comandos:

```
summary(airquality)             # rápido sumário das variáveis
summary(airquality[,1:4])       # rápido sumário apenas das 4 primeiras variáveis
mean(airquality$Temp)           # média das temperaturas no período
mean(airquality$Ozone)          # média do Ozone no período - note a resposta NA
airquality$Ozone                # a razão é que existem ``dados perdidos'' na variável Oz
mean(airquality$Ozone, na.rm=T) # média do Ozone no período - retirando valores perdidos
```

Note que os útimos tres comandos são trabalhosos de serem digitados pois temos que digitar `airquality$` a cada vez!
Mas há um mecanismo no R para facilitar isto: o *caminho de procura* ("search path"). Começe digitando e vendo s saída de:
`search()`

O programa vai mostrar o caminho de procura dos objetos. Ou seja, quando voce usa um nome do objeto o R vai procurar este objeto nos caminhos indicado, na ordem apresentada.

Pois bem, podemos "adicionar" um novo local neste caminho de procura e este novo local pode ser o nosso objeto `airquality`. Digite o seguinte e compara com o anterior:

```
attach(airquality)   # anexando o objeto airquality no caminho de procura.
search()             # mostra o caminho agora com o airquality incluído
mean(Temp)           # e ... a digitação fica mais fácil e rápida !!!!
mean(Ozone, na.rm=T) # pois com o airquality anexado o R acha as variáveis
```

NOTA: Para retirar o objeto do caminho de procura basta digitar `detach(airquality)`.

Bem, agora é com voce!
Reflita sobre os dados e use seus conhecimentos de estatística para fazer uma análise descritiva interessante destes dados.

Pense em questões relevantes e veja como usar medidas e gráficos para respondê-las. Por exemplo:

- as médias mensais variam entre si?

- como mostrar a evolução das variáveis no tempo?
- as variáveis estão relacionadas?
- etc, etc, etc

Para a análise exploratória/descritiva dos dados voce vai precisar conhecer alguns comandos do R.
Lembre-se que há vários materiais para consulta:

- O "Tutorial de Introdução ao R" contém alguns exemplos de comandos para análise descritiva.
- A página do Rweb contém a sessão Rweb modules onde voce pode fazer algumas análises através dos "menus" disponíveis e verificar os resultados para aprender os comandos.
- Lembre-se do Cartão de Referência que contém os comandos mais frequentemente utilizados.
- a `demo(graphics)` ilustra comandos para fazer diversos tipos de gráficos.

4 Calculando e fazendo alguns gráficos

Tente fazer os exercícios abaixo usando o R

1. Calcular as seguinte somas:

   (a) $10^2 + 11^2 + \ldots + 20^2$

   (b) $\sqrt{\log(1)} + \sqrt{\log(10)} + \sqrt{\log(100)} + \ldots + \sqrt{\log(1000000)}$, onde log é o logarítmo neperiano.

   Solução

2. Faça um gráfico para cada uma das funções a seguir.

   (a) $f(x) = 1 - \frac{1}{x} sin(x)$ para $0 \leq x \leq 50$

   (b) $f(x) = \frac{1}{\sqrt{50\pi}} \exp[-\frac{1}{50}(x-100)^2)]$ para $85 \leq x \leq 115$

   Solução

3. Seja uma v.a. X com distribuição exponencial com densidade dada por $f(x) = \lambda \exp(-\lambda x)$ com parâmetro $\lambda = 5$.

   (a) Faça um gráfico da função de densidade de probabilidade $f(x)$.

   (b) Faça um gráfico da função de distribuição acumulada $F(x)$.

   Solução

4. As funções `rep` e `seq` do R são úteis para criar vetores de dados que seguem um certo padrão.
   Clique aqui para ver um arquivo de dados.
   Mostre os comandos que podem ser usados para criar vetores para cada uma das três colunas iniciais deste arquivo.

   Solução

   Note que há mais detalhes do uso destas funções no Tutorial de Introdução ao R.

5 Estatística Básica - Probabilidades

Nesta sessão iremos também usar o R como uma calculadora estatística para resolver alguns exemplos/exercícios de probabilidades.

Os exercícios abaixo com indicação de paginas foram retirados de:
Magalhês, 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.

   Solução

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.

   Solução

3. (Ex 5, pag 77) Sendo X uma variável seguindo o modelo Binomial com parâmetro $n = 15$ e $p = 0.4$, pergunta-se:

   - $P(X \geq 14)$
   - $P(8 < X \leq 10)$
   - $P(X < 2 \text{ ou } X \geq 11)$
   - $P(X \geq 11 \text{ ou } X > 13))$
   - $P(X > 3 \text{ e } X < 6)$
   - $P(X \leq 13 \mid X \geq 11)$

   Solução

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 $P(90 - a \leq X \leq 90 + a) = \gamma,\ \gamma = 0.95$

Solução

5. Seja uma v.a. X com distribuição exponencial com densidade dada por $f(x) = \lambda \exp(-\lambda x)$ com parâmetro $\lambda = 5$.

   (a) Encontre $P[X < 0.6]$

   (b) Encontre $P[0.3 < X < 0.6]$

   (c) Encontre $P[X \geq 0.5]$

   (d) Encontre a tal que $P[X \geq a] = 0.05$

   Solução

6. Considere uma v.a. $X \sim Bin(12, 0.3)$. Calcule:

   (a) $P[X < 6]$

   (b) $P[5 \leq X < 10]$

   (c) $P[X \geq 8]$

   (d) $P[X = 7]$

   (e) o valor de a tal que $P[X \leq a] \approx 0.25$

   Solução

7. 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.

   Solução

6 Estatística Básica - Medidas descritivas

1. Dados de cancer de esôfago

   Carregue o conjunto de dados `esoph` com o comando `data(esoph)`.

   Este conjunto mostra o número da casos (`ncases`) e controles (`ncontrols`) de cancer de esôfago para diferentes grupos de idade (`agegp`), consumo de alcool (`alcgp`) e de tabaco (`tobgp`). Digite `help(esoph)` para saber mais sobre estes dados.

   Inspecione o conjunto de dados e mostre os comandos para se obter:

   (a) O número de casos e controles para cada faixa de idade
   (b) O número de casos e controles para cada grupo de consumo de alcool
   (c) O número de casos e controles para cada combinação de faixa etária e grupo de consumo de tabaco
   (d) O perfil (idade, consumo de alcool e tabaco) do grupo com maior número de casos
   (e) O perfil (idade, consumo de alcool e tabaco) do grupo com maior proporção de casos/controles

   Solução

2. Dados de comprimentos de rios

   Carregue o conjunto de dados `rivers` com o comando `data(rivers)`. Este arquivo contém o comprimento (em minhas) dos principais rios dos Estados Unidos. Digite `help(rivers)` para saber mais sobre estes dados.

   (a) Calcule a média e variância dos comprimentos dos rios
   (b) Quantos são os rios com comprimento superior a 1000 milhas?
   (c) Quais os comprimentos do maior e menor rio?
   (d) Obtenha os quartis para os comprimentos dos rios.
   (e) Qual o desvio padrão para rios com comprimento inferior a 700 milhas
   (f) Mostre algum gráfico que ilustre bem este cojunto de dados.

   Solução

3. Dados de crime nos EUA

   Carregue o conjunto de dados `USAarrests` com o comando `data(USArrests)`. Este conjunto contém o número dos seguinte crimes por estado: assasinato (Murder), assalto (Assault), estupro (Rape) e além disto a percentagem de população morando em área urbana (UrbanPop). Digite `help(rivers)` para saber mais sobre estes dados.

   (a) Quais os estados com o maior e menor número de estupros?
   (b) Somando-se os três crimes, qual o melhor e qual o pior estado?
   (c) Há relação entre a percentagem de população urbana e cada um dos três crimes? Justifique sua resposta.
   (d) Qual o número médio de assasinatos para estados com percentagem de população urbana acima da 60%? e abaixo de 60%?
   (e) Faça um gráfico para mostrar a relação entre número de homicídios e estupros.

(f) Obtenha o número mediano de cada tipo de crime para estados com mais de 75% de população urbana.

Solução

7 Explorando *arrays*

O conceito de `array` generaliza a idéia de matrix. Enquanto em uma matrix os elementos são organizados em duas dimensões (linhas e colunas), em um array os elementos podem ser organizados em um número arbitrário de dimensões.

No R um array é definido utilizando a função `array()`.

1. Defina um array com o comando a seguir e inspecione o objeto certificando-se que voce entendeu como arrays são criados.

   ```
   ar1 <- array(1:24, dim=c(3,4,2))
   ar1
   ```

2. Inspecione o "help" da função array (digite `help(array)`), rode e inspecione os exemplos contidos na documentação.

3. Veja agora um exemplo de dados já incluido no R no formato de array. Para "carregar" e visualizar os dados digite:

   ```
   data(Titanic)
   Titanic
   ```

   Para maiores informações sobre estes dados digite:

   ```
   help(Titanic)
   ```

   Agora responda às seguintes perguntas, mostrando os comandos do R utilizados:

   (a) quantas pessoas havia no total?
   (b) quantas pessoas havia na tripulação (crew)?
   (c) quantas crianças sobreviveram?
   (d) qual a proporção (em %) entre pessoas do sexo masculino e feminino entre os passageiros da primeira classe?
   (e) quais sao as proporções de sobreviventes entre homens e mulheres?

   Solução

4. O conjunto de dados `HairEyeColor` contém informações sobre cor do cabelo, olhos e sexo de 592 indivíduos.
   As cores de cabelo (Hair) são: preto (Black), castanho (Brown), ruivo (Red) e louro (Blond).
   As cores dos olhos são: castanho (Brown), azul (Blue), mel (Hazel) e verde (Green).
   Os indivíduos são classificados em sexo masculino (Male) o feminino (Female).

   Carregue o conjunto de dados com o comando
   `data(HairEyeColor)`

   e responda as seguintes perguntas fornecendo também o comando do R para obter a resposta:

   (a) Qual a proporção de homens e mulheres na amostra?

(b) Quantos são os homens de cabelos pretos?

(c) Quantos mulheres tem cabelos loiros?

(d) Qual a proporção de homens e mulheres entre as pessoas ruivas?

(e) Quantas pessoas tem olhos verdes?

Solução

8 Estatística Básica - Intervalos de confiança

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.

   Solução

2. (Ex 1, pag 235) Por analogia a produtos similares, o tempo de reação de um novo medicamento pode ser considerado como tendo distribuição Normal com desvio padrão a 2 minutos (a média é desconhecida). 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

   Obtenha um intervalo de confiança a 95% para o tempo médio de reação.

   Solução

3. Considere os dados a seguir de uma a.a. de uma distribuição Normal.

   $$23, 36, 30, 21, 23, 45, 36, 17, 34, 22, 21, 30.$$

   Faça os ítem abaixo mostrando o comandos do R necessários para obter as respostas

   (a) Encontre um IC a 95% para a média

   (b) Encontre um IC a 99% para a variância

   Solução

9 Estatística Básica - Teste de Hipóteses

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?

   Solução

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.

   Solução

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?

   Solução

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?

   Solução

5. 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?

   Solução

6. 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?
Solução

10 Análise de experimentos

EXPERIMENTOS EM ESQUEMA FATORIAL

Este experimento descrito na apostila do curso de Planejamento de Experimentos II comparou a resposta de mudas a diferentes recipientes e espécies de eucalipto.

No restante destas notas as linhas que começam com o símbolo `>` são comandos a serem digitados no R. Outros textos com a font `typewriter como esta` são saídas produzidas pelo programa.

1. **Lendo os dados**

   Clique aqui para ver e copiar o arquivo com conjunto de dados.

   A seguir vamos ler (importar) os dados para R com o comando `read.table`:

   ```
   > ex04 <- read.table("exemplo04.txt", header=T)
   > ex04
   ```

   Antes de começar as análise vamos inspecionar o objeto que contém os dados para saber quantas observações e variáveis há no arquivo, bem como o nome das variáveis. Vamos tembém pedir o R que exiba um rápido resumo dos dados.

   ```
   > dim(ex04)
   [1] 24  3

   > names(ex04)
   [1] "rec"  "esp"  "resp"

   > attach(ex04)

   > is.factor(rec)
   [1] TRUE
   > is.factor(esp)
   [1] TRUE
   > is.factor(resp)
   [1] FALSE
   > is.numeric(resp)
   [1] TRUE
   ```

   Nos resultados acima vemos que o objeto `ex04` que contém os dados tem 24 linhas (observações) e 3 colunas (variáveis). As variáveis tem nomes `rec`, `esp` e resp, sendo que as duas primeiras são *fatores* enquanto `resp` é uma variável numérica, que neste caso é a variável resposta. O objeto `ex04` foi incluído no caminho de procura usando o comando `attach` para facilitar a digitação.

2. **Análise exploratória**

   Inicialmente vamos obter um resumo de nosso conjunto de dados usando a função `summary`.

   ```
   > summary(ex04)
    rec    esp          resp
   ```

```
 r1:8   e1:12   Min.   :18.60
 r2:8   e2:12   1st Qu.:19.75
 r3:8           Median :23.70
                Mean   :22.97
                3rd Qu.:25.48
                Max.   :26.70
```

Note que para os fatores são exibidos o número de dados em cada nível do fator. Já para a variável numérica são mostrados algumas medidas estatísticas. Vamos explorar um pouco mais os dados

```
> ex04.m <- tapply(resp, list(rec,esp), mean)
> ex04.m
       e1     e2
r1 25.650 25.325
r2 25.875 19.575
r3 20.050 21.325

> ex04.mr <- tapply(resp, rec, mean)
> ex04.mr
     r1      r2      r3
25.4875 22.7250 20.6875

> ex04.me <- tapply(resp, esp, mean)
> ex04.me
      e1       e2
23.85833 22.07500
```

Nos comandos acima calculamos as médias para cada fator, assim como para os cruzamentos entre os fatores. Note que podemos calcular outros resumos além da média. Experimente nos comandos acima substituir `mean` por `var` para calcular a variância de cada grupo, e por `summary` para obter um outro resumo dos dados.

Em experimentos fatoriais é importante verificar se existe interação entre os fatores. Inicialmente vamos fazer isto graficamente e mais a frente faremos um teste formal para presença de interação. Os comandos a seguir são usados para produzir os gráficos exibidos na Figura 1.

```
> par(mfrow=c(1,2))
> interaction.plot(rec, esp, resp)
> interaction.plot(esp, rec, resp)
```

Pode-se usar o R para obter outros tipos de gráficos de acordo com o interesse de quem está analisando os dados. Por exemplo, os comandos abaixo ilustram outros tipos de gráficos. Experimente estes comandos, verifique os gráficos produzidos e certifique-se que voce entendeu cada comando.

```
> plot.default(rec, resp, ty="n")
> points(rec[esp=="e1"], resp[esp=="e1"], col=1)
> points(ex04.m[,1], pch="x", col=1, cex=1.5)
```

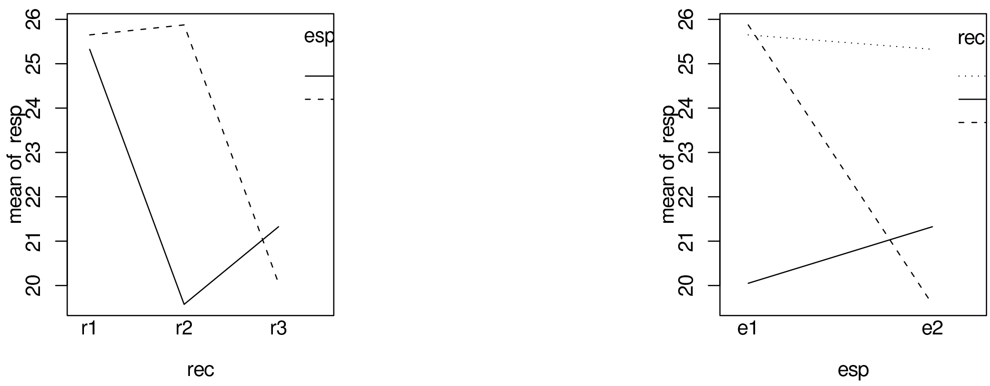

Figura 1: Gráficos de interação entre os fatores.

```
> points(rec[esp=="e2"], resp[esp=="e2"], col=2)
> points(ex04.m[,2], pch="x", col=2, cex=1.5)

> plot.default(esp, resp, ty="n")
> points(esp[rec=="r1"], resp[rec=="r1"], col=1)
> points(ex04.m[1,], pch="x", col=1, cex=1.5)
> points(esp[rec=="r2"], resp[rec=="r2"], col=2)
> points(ex04.m[2,], pch="x", col=2, cex=1.5)
> points(esp[rec=="r3"], resp[rec=="r3"], col=3)
> points(ex04.m[3,], pch="x", col=3, cex=1.5)

> coplot(resp ~ rec|esp)
> coplot(resp ~ esp|rec)
```

3. **Análise de variância**

Seguindo o modelo adequado, o análise de variância para este experimento inteiramente casualizado em esquema fatorial pode ser obtida com o comando:

```
> ex04.av <- aov(resp ~ rec + esp + rec * esp)
```

Entretanto o comando acima pode ser simplificado produzindo os mesmos resultados com o comando

```
> ex04.av <- aov(resp ~ rec * esp)
> summary(ex04.av)
```

```
            Df Sum Sq Mean Sq F value    Pr(>F)
rec          2 92.861  46.430  36.195 4.924e-07 ***
esp          1 19.082  19.082  14.875  0.001155 **
rec:esp      2 63.761  31.880  24.853 6.635e-06 ***
Residuals   18 23.090   1.283
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Isto significa que no R, ao colocar uma interação no modelo, os efeitos principais são incluídos automaticamente. Note no quadro de análise de variância que a interação é denotada por `rec:esp`. A análise acima ostra que este efeito é significativo, confirmando o que verificamos nos gráficos de interação vistos anteriormente.

O objeto `ex04.av` guarda todos os resultados da análise e pode ser explorado por diversos comandos. Por exemplo a função `model.tables` aplicada a este objeto produz tabelas das médias definidas pelo modelo. O resultado mostra a média geral, médias de cada nível fatores e das combinações dos níveis dos fatores. Note que no resultado está incluído também o número de dados que gerou cada média.

```
> ex04.mt <- model.tables(ex04.av, ty="means")
> ex04.mt
Tables of means
Grand mean

22.96667

 rec
       r1    r2    r3
    25.49 22.73 20.69
rep  8.00  8.00  8.00

 esp
       e1    e2
    23.86 22.07
rep 12.00 12.00

 rec:esp
     esp
rec   e1     e2
  r1  25.650 25.325
  rep  4.000  4.000
  r2  25.875 19.575
  rep  4.000  4.000
  r3  20.050 21.325
  rep  4.000  4.000
```

Mas isto não é tudo! O objeto `ex04.av` possui vários elementos qaue guardam informações sobre o ajuste.

```
> names(ex04.av)
 [1] "coefficients"  "residuals"     "effects"       "rank"
```

```
 [5] "fitted.values" "assign"       "qr"           "df.residual"
 [9] "contrasts"     "xlevels"      "call"         "terms"
[13] "model"

> class(ex04.av)
[1] "aov" "lm"
```

O comando `class` mostra que o objeto `ex04.av` pertence às classes `aov` e `lm`. Isto significa que devem haver *métodos* associados a este objeto que tornam a exploração do resultado mais fácil. Na verdade já usamos este fato acima quando digitamos o comando `summary(ex04.av)`. Existe uma função chamada `summary.aov` que foi utilizada já que o objeto é da classe `aov`. Iremos usar mais este mecanismo no próximo passo da análise.

4. **Análise de resíduos**

   Após ajustar o modelo devemos proceder a análise dos resíduos para verificar os pressupostos. O R produz automaticamente 4 gráficos básicos de resíduos conforme a Figura 2 com o comando `plot`.

```
> par(mfrow=c(2,2))
> plot(ex04.av)
```

Os gráficos permitem uma análise dos resíduos que auxiliam no julgamento da adequacidade do modelo. Evidentemente voce não precisa se limitar os gráficos produzidos automaticamente pelo R – voce pode criar os seus próprios gráficos muito facilmente. Neste gráficos voce pode usar outras variáveis, mudar texto de eixos e títulos, etc, etc, etc. Examine os comandos abaixo e os gráficos por eles produzidos.

```
> par(mfrow=c(2,1))
> residuos <- resid(ex04.av)

> plot(ex04$rec, residuos)
> title("Resíduos vs Recipientes")

> plot(ex04$esp, residuos)
> title("Resíduos vs Espécies")

> par(mfrow=c(2,2))
> preditos <- (ex04.av$fitted.values)
> plot(residuos, preditos)
> title("Resíduos vs Preditos")
> s2 <- sum(resid(ex04.av)^2)/ex04.av$df.res
> respad <- residuos/sqrt(s2)
> boxplot(respad)
> title("Resíduos Padronizados")
> qqnorm(residuos,ylab="Residuos", main=NULL)
> qqline(residuos)
> title("Grafico Normal de \n Probabilidade dos Resíduos")
```

Além disto há alguns testes já programados. Como exemplo vejamos e teste de Shapiro-Wilk para testar a normalidade dos resíduos.

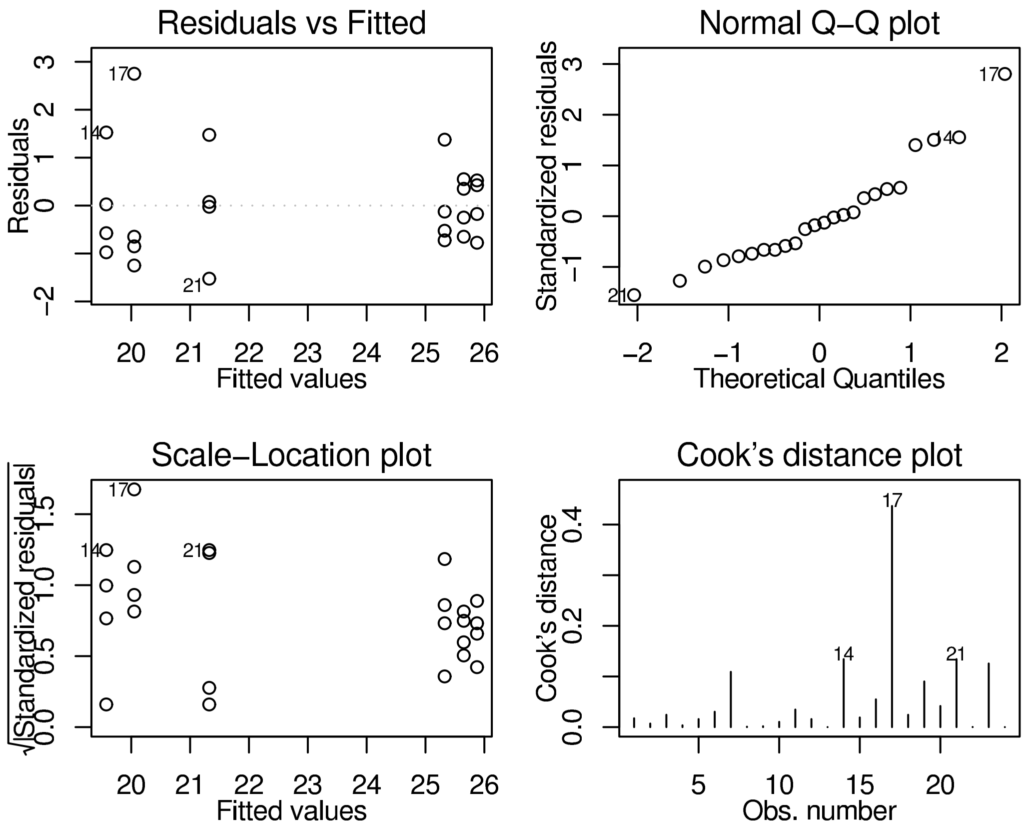

Figura 2: Gráficos de resíduos produzidos automaticamente pelo R.

```
> shapiro.test(residuos)

        Shapiro-Wilk normality test

data:  residuos
W = 0.9293, p-value = 0.09402
```

5. **Desdobrando interações**

   Conforma visto na apostila do curso, quando a interação entre os fatores é significativa podemos desdobrar os graus de liberdade de um fator dentro de cada nível do outro. A forma de fazer isto no R é reajustar o modelo utilizando a notação `/` que indica efeitos aninhados. Desta forma podemos desdobrar os efeitos de espécie dentro de cada recipiente e vice versa conforme mostrado a seguir.

```
> ex04.avr <- aov(resp ~ rec/esp)
> summary(ex04.avr, split=list("rec:esp"=list(r1=1, r2=2, r3=3)))
              Df Sum Sq Mean Sq F value    Pr(>F)
rec            2 92.861  46.430 36.1952 4.924e-07 ***
rec:esp        3 82.842  27.614 21.5269 3.509e-06 ***
  rec:esp: r1  1  0.211   0.211  0.1647    0.6897
  rec:esp: r2  1 79.380  79.380 61.8813 3.112e-07 ***
  rec:esp: r3  1  3.251   3.251  2.5345    0.1288
Residuals     18 23.090   1.283
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> ex04.ave <- aov(resp ~ esp/rec)
> summary(ex04.ave, split=list("esp:rec"=list(e1=c(1,3), e2=c(2,4))))
              Df  Sum Sq Mean Sq F value    Pr(>F)
esp            1  19.082  19.082  14.875  0.001155 **
esp:rec        4 156.622  39.155  30.524 8.438e-08 ***
  esp:rec: e1  2  87.122  43.561  33.958 7.776e-07 ***
  esp:rec: e2  2  69.500  34.750  27.090 3.730e-06 ***
Residuals     18  23.090   1.283
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

6. **Teste de Tukey para comparações múltiplas**

   Há vários testes de comparações múltiplas disponíveis na literatura, e muitos deles implementados no R. Os que não estão implementados podem ser facilmente calculados utilizando os recursos do R.

   Vejamos por exemplo duas formas de usar o *Teste de Tukey*, a primeira usando uma implementação com a função `TukeyHSD` e uma segunda fazendo ops cálculos necessários com o R.

   Poderíamos simplesmente digitar:

```
> ex04.tk <- TukeyHSD(ex04.av)
> plot(ex04.tk)
> ex04.tk
```

   e obter diversos resultados. Entretanto nem todos nos interessam. Como a interação foi significativa na análise deste experimento a comparação dos níveis fatores principais não nos interessa.

   Podemos então pedir a função que somente mostre a comparação de médias entre as combinações dos nívies dos fatores.

```
> ex04.tk <- TukeyHSD(ex04.ave, "esp:rec")
> plot(ex04.tk)
> ex04.tk
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = resp ~ esp/rec)

$"esp:rec"
        diff        lwr       upr
 [1,] -0.325 -2.8701851  2.220185
 [2,]  0.225 -2.3201851  2.770185
 [3,] -6.075 -8.6201851 -3.529815
 [4,] -5.600 -8.1451851 -3.054815
 [5,] -4.325 -6.8701851 -1.779815
 [6,]  0.550 -1.9951851  3.095185
 [7,] -5.750 -8.2951851 -3.204815
 [8,] -5.275 -7.8201851 -2.729815
 [9,] -4.000 -6.5451851 -1.454815
[10,] -6.300 -8.8451851 -3.754815
[11,] -5.825 -8.3701851 -3.279815
[12,] -4.550 -7.0951851 -2.004815
[13,]  0.475 -2.0701851  3.020185
[14,]  1.750 -0.7951851  4.295185
[15,]  1.275 -1.2701851  3.820185
```

Mas ainda assim temos resultados que não interessam. Mais especificamente estamos intessados nas comparações dos níveis de um fator dentro dos nívies de outro. Por exemplo, vamos fazer as comparações dos recipientes para cada uma das espécies.

Primeiro vamos obter

```
> s2 <- sum(resid(ex04.av)^2)/ex04.av$df.res
> dt <- qtukey(0.95, 3, 18) * sqrt(s2/4)
> dt
[1] 2.043945
>
> ex04.m
       e1     e2
r1 25.650 25.325
r2 25.875 19.575
r3 20.050 21.325
>
> m1 <- ex04.m[,1]
> m1
    r1     r2     r3
25.650 25.875 20.050
> m1d <- outer(m1,m1,"-")
> m1d
       r1     r2    r3
r1  0.000 -0.225 5.600
```

```
r2  0.225  0.000 5.825
r3 -5.600 -5.825 0.000
> m1d <- m1d[lower.tri(m1d)]
> m1d
    r2     r3   <NA>
 0.225 -5.600 -5.825
>
> m1n <- outer(names(m1),names(m1),paste, sep="-")
> names(m1d) <- m1n[lower.tri(m1n)]
> m1d
 r2-r1  r3-r1  r3-r2
 0.225 -5.600 -5.825
>
> data.frame(dif = m1d, sig = ifelse(abs(m1d) > dt, "*", "ns"))
         dif sig
r2-r1  0.225  ns
r3-r1 -5.600   *
r3-r2 -5.825   *
>
> m2 <- ex04.m[,2]
> m2d <- outer(m2,m2,"-")
> m2d <- m2d[lower.tri(m2d)]
> m2n <- outer(names(m2),names(m2),paste, sep="-")
> names(m2d) <- m2n[lower.tri(m2n)]
> data.frame(dif = m2d, sig = ifelse(abs(m2d) > dt, "*", "ns"))
        dif sig
r2-r1 -5.75   *
r3-r1 -4.00   *
r3-r2  1.75  ns
```

11 Escrevendo funções

Agora sim, o `filé mignon`!

Uma das grandes vantagens de uma linguagem como o R é a facilidade para escrever as suas próprias `funções`. Desta forma voce não precisa se limitar aos recursos disponíveis do programa e pode implementar algum procedimento que voce queira programando o facilmente R usando sua estrutura de programação extremamente intuitiva.

Vamos começar com dois exemplos muito simples. Nas sessões seguintes usaremos esta idéia de escrever as nossas funções em vários exemplos.

1. A média harmônica H para um conjunto de números $x_1, x_2, \ldots, x_n$ é definida por

   $$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_n}}.$$

   Escreva uma função para calcular a média harmônica, crie um vetor de dados e exemplifique o uso como desta função.

   Solução

2. Considere os dados a seguir de uma a.a. de uma distribuição Normal.

   $$23, 36, 30, 21, 23, 45, 36, 17, 34, 22, 21, 30.$$

   Faça os ítem abaixo mostrando o comandos do R necessários para obter as respostas

   (a) Encontre um IC a 95% para a média
   (b) Encontre um IC a 99% para a variância
   (b) Escreva uma função que receba dados de uma a.a. da normal e retorne IC's para média e variância.

   Solução

12 Explorando verossimilhanças

Nesta sessão são ilustrados:

- gráficos da função de verossimilhança
- obtenção de intervalos de confiança pelo método da quantidade pivotal,
- resultados diversos da teoria de verossimilhança

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

1. Considere agora o Exercício 3 da sessão **Estatística Básica - Teste de Hipóteses**
   (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.

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 $-2 \sum_i^n \log F(x_i; \theta) \sim \chi^2_{2n}$.
   Solução

3. 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
   ```

   Obtenha:
   (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)

Solução

4. Acredita-se que o número de trens atrasados para Lancaster 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).

Solução

5. 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

Solução

6. 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 $\hat{\theta}$

Solução

13 Explorando a função poder de teste

Nesta sessão vamos utilizar o R para ilustrar o conceito de *função poder do teste.*

EXEMPLO

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 1.

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:

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

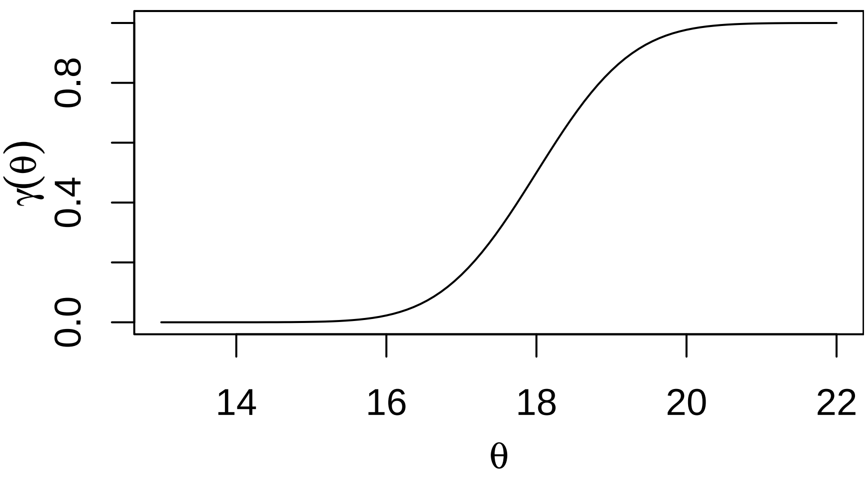

Figura 3: Função poder para $n = 25$.

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 2.

```
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))
```

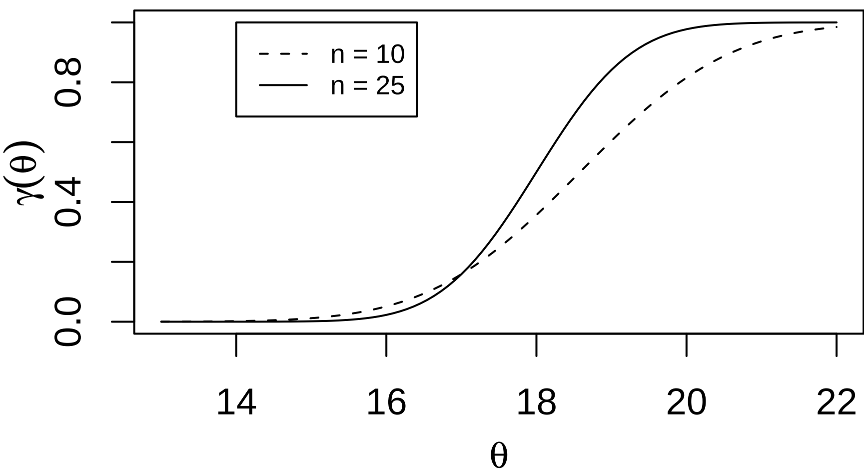

Figura 4: Função poder para $n = 10$ e $n = 25$.

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 3.

```
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))
```

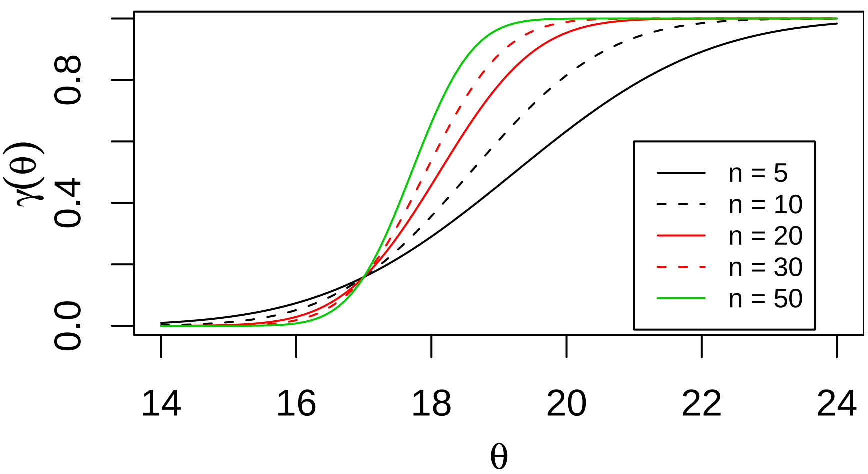

Figura 5: Função poder para diferentes tamanhos de amostra.

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.

14 Um exemplo de simulação

Nesta sessão iremos misturar diversos elementos do uso do R como geração de números aleatórios, manipulação de objetos, seleção de elementos e gráficos na solução de um problema matemático simples: a estimação do valor do valor de uma expressão matemática por simulação.

Suponha que nós esquecemos a equação que calcula a área do círculo. Com um computador na mão nem tudo está perdido! Vamos usar simulação para calcular a área do círculo.

Vamos ver como fazer isto:
Por conveniência vamos calcular a área de um círculo de raio unitário ($r = 1$) e vamos chamar esta área de π. Considere um quadrado definido pelos pontos $(-1, 1)$, $(1, 1)$, $(1, -1)$ e $(-1, -1)$. Sabemos que a área deste quadrado é igual a 4. Este quadrado contém um círculo de raio unitário.

Considere um ponto qualquer Z selecionado aleatoriamente dentro do quadrado. O ponto Z é definido por coordenadas (x, y) que possuem distribuições uniformes independentes no intervalo $(-1, 1)$. Podemos calcular a probabilidade deste ponto estar também dentro do círculo:

$$P(Z\ dentro\ do\ circulo) = \frac{area\ do\ circulo}{area\ do\ quadrado} = \pi/4$$

Se estimarmos a probabilidade por simulação, o valor da área do círculo π é dada por:

$$\pi = 4 * P(Z\ dentro\ do\ circulo).$$

Portanto, tudo que temos que fazer é estimar esta probabilidade. Para isto basta gerar um grande número de pontos no quadrado e verificar a proporção deles que está contida no círculo.

Exercício proposto:

Escreva um programa no R para o problema acima e estime o valor de π usando 100, 1000 e 10000 pontos escolhidos ao acaso.
Verifique a aproximação ao valor real de π.

DICA: Voce vai precisar checar se um ponto (x, y) está dentro do círculo. Lembre-se que um círculo é definido por sua origem e seu raio. Calcule a distância do ponto até o centro do círculo usado o teorema de Pitágoras.

15 Um pacote ilustrando conceitos estatísticos

Nesta sessão iremos examinar um pacote que foi escrito para ilustrar conceitos estatísticos utilizando o R. Este pacote serviu de material de apoio para um curso de estatística geral para estudantes de Pós-Graduação de diversos cursos na Universidade de Lancaster, Inglaterra.

Os objetivos são vários. Poderemos ver a flexibilidade do R para criar funções e materiais específicos. Alguns conceitos estatísticos serão revisados enquanto usamos as funções do pacote. Além disto podemos examinar as funções para ver como foram programadas no R.

O pacote se chama `gsse401` e tem uma página em http://www.est.ufpr.br/ paulojus/gsse401. Nas aulas práticas no Laboratório ele pode ser chamado da seguinte forma:

- **Usuários de LINUX**
  O pacote já está instalado na maquina *gauss* e portanto basta iniciar o R e digitar `require(gsse401)`

- **Usuários de WINDOWS**
  Neste caso o pacote deve ser instalado no seu computator. Inicie o R e digite os comandos

  ```
  install.packages("gsse401",
         cont="http://www.est.ufpr.br/~paulojus/gsse401")
  require(gsse401)
  ```

O pacote possui várias funções e conjuntos de dados. Para exibir os nomes dos conjuntos de dados e funções do pacote digite:

```
gsse401.data()
gsse401.functions()
```

para ver os nomes das funções e arquivos de dados e/ou explore a página do pacote.

Sugere-se as seguites atividades:

1. Carregue o conjunto de dados `ansc` e rode os exemplos de sua documentação. Discuta os resultados. Lembre-se que para carregar este conjunto de dados e ver sua documentação deve-se usar os comandos:

   ```
   data(ansc)
   help(asnc)
   ```

2. Explore a função `clt`. Veja a sua documentação, rode os exemplos e veja como foi programada digitando `clt` (sem os parênteses). Tente tambem usar a função digitando `clt()`.

3. Carregue o conjunto de dados `gravity`, veja sua documentação, rode e discuta os exemplos.

4. explore a função `mctest`, veja sua documentação e exemplos.

5. explore a função `queue`

6. explore a função `reg`. Tente também digitar `reg()` para o funcionamento interativo da função.