BIOESTATÍSTICA USANDO R

APOSTILA DE EXEMPLOS PARA O BIÓLOGO

Colin Robert Beasley

beasley@ufpa.br

Universidade Federal do Pará,

Campus de Bragança

Laboratório de Moluscos

Bragança

2004

O QUE É R?[1]

Introdução a R

R é uma linguagem e ambiente para computação estatística e gráficos. É um projeto GNU que é similar à linguagem e ambiente S que foi desenvolvida no Bell Laboratories (anteriormente AT&T, agora Lucent Technologies) por John Chambers e colegas. R pode ser considerada como uma implementação diferente da S. Há algumas diferenças importantes, mas muito código para S funciona inalterado em R.

R fornece uma ampla variedade de técnicas estatísticas (modelagem linear e não linear, testes estatísticos clássicos, análise de séries temporais, classificação, agrupamento, ...) e gráficos, e é altamente extensível. A linguagem S é muitas vezes o veículo de escolha para pesquisa em metodologia estatística, e R fornece uma rota Open Source para participação naquela atividade.

Um dos pontos fortes de R é a facilidade com que gráficos bem-desenhados com qualidade para publicação podem ser produzidos, incluindo símbolos matemáticos e fórmulas quando necessário. Muitos cuidados têm sido feitos sobre as definições padrão para as menores escolhas em desenho, entretanto o usuário retém controle total.

R é disponível como Software Livre sob os termos da Licença Pública Geral GNU da Free Software Foundation na forma de código fonte. Ela compila e funciona em uma grande variedade de plataformas UNIX e sistemas similares (incluindo FreeBSD e Linux). Ele compila e funciona em Windows 9x/NT/2000 e MacOS.

O ambiente R

R é um conjunto integrado de facilidades de software para manipulação de dados, cálculo e visualização gráfica. Ele inclui

* uma facilidade efetiva para manipulação e armazenagem de dados,
* um conjunto de operadores para cálculos sobre quadros de dados, em particular as matrizes,
* uma grande e coerente coleção integrada de ferramentas intermediárias para análise de dados,
* facilidades gráficas para análise de dados e visualização na tela ou impressa,
* uma linguagem de programação bem desenvolvida, simples e efetiva que inclui condicionais, alças, funções recursivas definidas pelo usuário, e facilidades para entrada e saída.

O termo “ambiente” pretende caracterizar R como um sistema totalmente planejado e coerente, em vez de uma aglomeração de ferramentas muito específicas e inflexíveis, como é o caso com outros softwares de análise de dados.

R, bem como S, é desenhada ao redor de uma verdadeira linguagem de computador, e permite aos usuários acrescentar funcionalidade adicional por definição de novas funções. Muito do sistema é escrita

1 Tradução da página *What is R?* (O que é R) no site do Projeto R: http://www.r-project.org/about.html

no dialeto R da S, que faz com que seja fácil para usuários seguir as escolhas algorítimicas feitas. Para tarefas computacionalmente-intensivas, C, C++ e código Fortran podem ser ligados e chamados na hora de calcular. Usuários avançados podem escrever código C para manipular objetos R diretamente.
Muitos usuários pensam em R como um sistema estatístico. Nós preferimos pensar nele como um ambiente dentro do qual técnicas estatísticas são implementadas. R pode ser estendido (facilmente) através de pacotes. Há cerca de oito pacotes fornecidos com a distribuição R e muitos outros são disponíveis através da família CRAN de sítios na Internet cobrindo uma ampla variedade de estatísticas modernas.

R tem seu próprio formato de documentação, parecido com LaTex, que é usado para fornecer documentação compreensiva, tanto on-line e em uma variedade de formatos, como impressa.

--Fim da tradução--

CONSIDERAÇÕES GERAIS

O R (Ihaka & Gentleman, 1996) é uma linguagem e ambiente estatístico que traz muitas vantagens para o usuário, embora que estas nem sempre sejam óbvias no início. Primeiro, é um Software Livre (livre no sentido de liberdade) distribuído sob a Licença Pública Geral (http://www.fsf.org/pt_home.html) e pode ser livremente copiado e distribuído entre usuários, bem como pode ser instalado em diversos computadores livremente. Isso contraste com pacotes comerciais que têm licenças altamente restritivas e não permitem que sejam feitas cópias ou que seja instalado em mais de um computador sem a devida licença (e pagamento, claro!). Segundo, a grande maioria de Softwares Livres são grátis e R não é uma exceção. Isso contrasta com os pacotes comerciais. Terceiro, sendo um Software Livre, os códigos fontes do R estão disponíveis e atualmente são gerenciados por um grupo chamado o *Core Development Team* (http://r-project.org/contributors). A vantagem de ter o código aberto é que falhas podem ser detectadas e corrigidas rapidamente e atualizações para Softwares Livres podem ser disponibilizadas em uma questão de dias. Essa sistema de revisão depende pesadamente da participação dos usuários. Em contraste, em muitos pacotes comerciais, as falhas não são corrigidas até o próximo lançamento que pode levar vários anos. Quarta, R fornece um interface de entrada por linha de comando (ELC). Todos os comandos são digitados e o *mouse* é pouco usado. Pode parecer “antigo” ou até “pobre em recursos visuais”, mas aí há o melhor recurso do R, a sua flexibilidade. Para usuários, a linguagem da R se torna clara e simples e a flexibilidade da ELC permite que uns poucos comandos simples sejam juntados para criar funções poderosas. Além disso a transparência das funções e a entrada de dados é altamente didática. O usuário é sempre consciente do que foi pedido através da ELC. Isso contraste com outros pacotes em que uma interface bonita e sofisticada pode esconder a dinâmica dos cálculos, e potencialmente pode esconder erros. Finalmente, R é disponível para muitas plataformas incluindo Unix, Linux, Macintosh e Windows. Embora seja possível baixar e compilar os códigos fontes para instalar R no seu sistema, a maioria de usuários optam para a via mais fácil de instalar R através de arquivos binários ou executáveis.
Esta apostila não é um manual completo para R e não é uma substituição para os arquivos de ajuda. O R conta com um enorme acervo de arquivos de ajuda que são disponíveis no programa e na Internet na forma de arquivos html ou pdf, além de outros formatos. Na seção *Documentation* na página do Projeto R

(http://www.r-project.org) há manuais e questões freqüentemente perguntadas (FAQ's) e estes estão também disponíveis localmente quando R é instalado. Há também um boletim, páginas de ajuda, e listas de publicações em revistas científicas. As listas de e-mail são fontes extremamente úteis de informações sobre como executar tarefas em R, além de dicas úteis de como resolver problemas em que não há ajuda documentada. Veja a Lista de Recursos de Ajuda na pagina 3 para iniciar aprendizagem do R. Referência é feita aos exemplos tirados do livro de Fowler & Cohen (1990).

OS OBJETIVOS DESTA APOSTILA

Esta apostila foi originalmente escrita para alunos da disciplina de Biometria do curso de Ciências Biológicos do Campus de Bragança, Universidade Federal do Pará, Brasil. Depois achei que também poderia ser útil para alunos de biologia (e de outros cursos) em outras instituições de ensino superior no Brasil. Assim, gostaria contar com o apoio dos leitores quanto às sugestões, críticas e melhorias no texto.

Fornecer exemplos do uso do R no contexto de biologia porque, em geral, há carência de exemplos biológicos na documentação atual do R.

Estimular a aprendizagem da estatística dando exemplos claros e simples da funcionalidade e flexibilidade do R.

Estimular os alunos aproveitarem do Software Livre e portanto evitando as restrições de softwares comerciais e o uso não autorizado destes.

LISTA DE RECURSOS DE AJUDA

Das páginas do Projeto R na Internet

http://r-project.org.	O *home page* do projeto R
http://cran.br.r-project.org	O servidor *mirror* (espelho) brasileiro (UFPr)
http://r-project.org/mail	*r-help* é a lista mais apropriado para usuários
http://cran.r-project.org/other-docs.html	documentos de ajuda e tutoriais em vários formatos

Nesta última página é particularmente recomendado o documento *R for Beginners* (Inglês) da autoria de Emanuel Paradis, ou *R para Principiantes* (a tradução da *R for Beginners* para o Espanhol feita por Jorge A. Ahumadal).

Outras páginas na Internet sobre R

http://www.agr.kuleuven.ac.be/vakken/StatisticsByR/index.htm	Introdução à análise de dados usando R
http://www.math.csi.cuny.edu/Statistics/R/simpleR/index.html	Introdução usando o pacote Simple em R
http://www.est.ufpr.br/Rtutorial/contents.html	Tutorial sobre R em Português

Algumas páginas chaves sobre Software Livre (SL)

http://www.fsf.org/home_pt.html	Free Software Foundation (GNU)
http://www.softwarelivre.rs.gov.br	Site sobre SL do Governo do RS
http://www.softwarelivre.unicamp.br/sl	Site sobre SL da UNICAMP, SP

http://gufsc.lcmi.ufsc.br Grupo de usuários de SL da UFSC

http://focalinux.cipsga.org.br/guia/inic_interm/ch-intro.htm Introdução ao GNU/Linux

Livros

Dalgaard P (2002) Introductory Statistics with R. Springer, New York, ISBN 0-387-95475-9.

Fox J (2002) An R and S-PLUS Companion to Applied Regression. Sage Publications, ISBN 0-761-92280-6 (softcover) ou 0-761-92279-2 (hardcover)

Comandos de ajuda do R

`> help.start()` inicia documentação na forma de arquivos html visualizados no seu browser

`> help (`*tópico*`)` inicia uma janela de ajuda sobre tópico

`> ?(`*tópico*`)` a mesma coisa

DICAS GERAIS ANTES DE COMEÇAR

Sempre iniciar no seu diretório de trabalho (p. ex. meuprojeto)

Em Windows: File=> Change dir e seleciona C:\meunome\meuprojeto ou alternativamente, usa

```
> setwd("C:\\Meus Documentos\\meuprojeto")
```

Em Linux:

```
> setwd("/home/meunome/meuprojeto")
```

Texto pode ser digitado após o *prompt* de comando >

Funções em R sempre são acompanhados com parênteses ()

Há uma distinção entre minúsculas e MAIÚSCULAS.

Você pode ver o histórico de comandos colocados por você durante a sua sessão pressionando a tecla da seta para cima (↑). Isso é muito útil para verificar novamente os comandos anteriores ou reeditá-los.

Pode copiar e colar na linha de comando: primeiro seleciona o texto a ser copiado e, em Windows clique com o botão direito do *mouse,* selecione Copy e clique novamente no botão direta e selecione Paste. Alternativamente use Crtl+C e Ctrl+V.

Para copiar e colar em Linux, clique sobre o texto selecionado com o botão esquerdo do *mouse* com a tecla Ctrl pressionada (ou Ctrl+C). Para colar, clique com o botão do meio (ou os dois no mesmo tempo, se não tiver um mouse com três botões).

Você pode designar nomes a objetos R usando a combinação "<−"

```
> x<-c(12,24,14)
```

Lembre que R usa um ponto "." em vez de vírgula "," quando há números com casas decimais. Se precisar importar dados que usam vírgulas em vez de pontos, troque na planilha as vírgulas por pontos usando Localizar e Substituir, se não, os dados não serão reconhecidos como números.

Isso não é a mesma coisa que dados separados por vírgulas. Por exemplo, o jogo de dados a seguir tem casas decimais definidos usando pontos, mas os valores em cada categoria são separados por vírgulas.

A, B, C
2.6, 3.8, 7.6

Neste documento, comandos a serem digitados na linha de comando serão assinalados com o prefixo > (o *prompt* do R) e estão na fonte `Courier New 12` ex. `> mean(massa)`. Texto em `Courier New 12` sem o *prompt* é o resultado. Texto escrito em Arial 10 é a explicação.

Para sair com segurança da R usa >`q()`. Um diálogo aparecerá perguntando se quer salvar o espaço de trabalho. `Save workspace image? [y/n/c]:`

Para estas sessões, não é necessário e pode responder `n` para não salvar a imagem do espaço de trabalho.

ARQUIVOS ACOMPANHANDO ESTE APOSTILA

Dados

larvas.csv	o comprimento de larvas de bivalves (Beasley, dados não publicados)
massa.csv	a massa de pássaros em quatro locais (Fowler & Cohen, 1990)
starling.csv	a massa de pássaros por local e mês (Fowler & Cohen, 1990)
plasma.csv	o efeito de tratamento e sexo sobre níveis de plasma (Zar, 1999)
dadosmv.csv	associações de macrofauna no estuário do Caeté, Pará (Beasley, dados não publicados)

AGRADECIMENTOS

Gostaria agradecer Profa. Dra. Claudia Helena Tagliaro da Universidade Federal do Pará (UFPA) para leitura crítica do documento. Obrigado também ao Dr. Ulf Mehlig do Centre for Marine Ecology (ZMT), Universidade de Bremen, para leitura crítica, melhorias no código para plotar a média e o desvio padrão, e permissão para usar material não-publicado na seção cálculos simples.

CÁLCULOS SIMPLES

Usando a linha de comando de R podemos somar..

```
> 1+1
[1] 2
```

..subtrair,

```
> 14-6
[1] 8
```

..multiplicar,

```
> 3*4
[1] 12
```

..dividir,

```
> 47/11
[1] 4.272727
```

e realizar cálculos mais complexos como se fosse uma calculadora científica.

```
> sin(5)
[1] -0.9589243
```

Considere o seguinte cálculo:

```
> 2.3*2
[1] 4.6
```

Erros podem ocorrer quando R não entende o que foi digitado

```
> 2,3*2
Error: syntax error
> sib(5)
Error: couldn't find function "sib"
```

Lembrando resultados: podemos designar um nome ao resultado do cálculo

```
> 47/11-> resultado
> resultado
[1] 4.272727
> resultado*11
[1] 47
> resultado+42->resultado
> resultado
[1] 46.27273
```

Cálculos repetitivos podem ser automatizados

```
> x<-17
> x
[1] 17
> for (i in 1:5) {x+2->x}
> x
[1] 27
```

Para ir de x=17 e chegar a x=27, o número 2 foi somado cinco vezes ao valor de x que foi armazenado cada vez como x. O cálculo a mão seria: 17+2=19, 19+2=21, 21+2=23, 23+2=25, 25+2=27

A seqüência de números 1 a 5 pode ser obtida usando

```
> 1:5
[1] 1 2 3 4 5
```

Podemos designar x como a seqüência 1 a 5

```
> x<-1:5
> x
[1] 1 2 3 4 5
```

Podemos calcular com a seqüência x

```
> x+2
[1] 3 4 5 6 7
```

É possível somar duas seqüências..

```
> y<-6:10
> y
[1] 6 7 8 9 10
> x+y
[1] 7 9 11 13 15
```

..ou multiplica-las.

```
> x
[1] 1 2 3 4 5
> y
[1]  6  7  8  9 10
> x*y
[1] 6 14 24 36 50
```

Duas seqüências podem ser juntadas para criar uma terceira usando c(). O "c" significa concatenar, ou seja, juntar.

```
> z=c(x,y)
> z
[1]  1  2  3  4  5  6  7  8  9 10
```

Seqüências arbitrárias podem ser criadas.

```
> x<-c(1, 2, 17, 42, 4.5)
> x
[1] 1.0  2.0  17.0  42.0  4.5


> x*0.3
[1] 0.3 0.6 5.10 12.60 1.35
```

Podemos extrair informações sobre elementos específicos de um jogo de dados. Por exemplo, a extração do terceiro elemento de uma seqüência

```
> x <-c(11, 22, 33, 44, 55)
> x[3]
[1] 33
```

A extração do terceiro a quinto elemento da seqüência

```
> x[3:5]
[1] 33 44 55
```

Criar um jogo de dados chamado y

```
> y<-c(0, 4, 2, 1, 0, 4, 0, 3, 0, 3, 3, 3, 4, 4, 2, 2, 0)
> y
[1] 0 4 2 1 0 4 0 3 0 3 3 3 4 4 2 2 0
```

Podemos fazer buscas condicionais dentro de y como, por exemplo, procurar valores menor que 3

```
> y[y<3]
[1] 0 2 1 0 0 0 2 2 0
```

Procurar valores menor ou igual a 3

```
> y[y<=3]
[1] 0 2 1 0 0 3 0 3 3 3 2 2 0
```

Procurar valores igual a zero (sim, a operadora “igual a” ou "==" é assim mesmo!)

```
> y[y==0]
[1] 0 0 0 0 0
```

Procurar valores não igual a zero (a combinação "!=" é o operadora neste caso)

```
> y[y!=0]
[1] 4 2 1 4 3 3 3 3 4 4 2 2
```

Determinar o número de observações em um jogo de dados

```
> length (y)
[17]
```

Determinar o número de valores igual a zero

```
> length(y[y==0])
[1] 5
```

EXPLORAÇÃO PRELIMINAR DOS DADOS

Estatística descritiva, *boxplots*, gráficos de barra e histogramas.

A teoria e os cálculos detalhados de procedimentos nesta seção podem ser encontrados nos Capítulos 3, 4 e 5 de Vieira (1980), 4 e 5 de Levin (1985), 3 a 6 de Fowler & Cohen (1990) e 3 e 4 de Zar (1999).

Importar os dados no arquivo de valores separados por virgulas, larvas.csv

```
> x<-read.csv("larvas.csv")
```

Renomear as colunas com títulos com acentos e depois mostrar o novo jogo de dados x

```
> names(x)<-c("Tocantins", "Melgaço", "Ourém", "Irituia", "Guamá")
> x
```

Obter um resumo das estatísticas descritivas: a observação mínima, 1ª quartil (25%), mediana (50%), média aritmética, 3ª quartil (75%) e a máxima são exibidos para cada amostra

```
> summary(x)
```

```
 Tocantins       Melgaço         Ourém           Irituia         Guamá
Min. :  232.0   Min. :  218.0   Min. :  232.0   Min. :  232.0  Min. :  218.0
1st Qu.:255.0   1st Qu.:255.0   1st Qu.:255.0   1st Qu.:255.0  1st Qu.:218.0
Median :255.0   Median :278.0   Median :255.0   Median :255.0  Median :232.0
Mean :  251.2   Mean :  265.8   Mean :  251.2   Mean :  251.2  Mean :  229.3
3rd Qu.:255.0   3rd Qu.:278.0   3rd Qu.:255.0   3rd Qu.:255.0  3rd Qu.:232.0
Max. :  255.0   Max. :  289.0   Max. :  255.0   Max. :  255.0  Max. :  255.0
```

Podemos representar as estatísticas em uma forma gráfica, o boxplot, mas primeiro vamos importar um novo jogo de dados: massa.csv e mostrar as estatísticas (não mostrados aqui)

```
> y<-read.csv("massa.csv")
> summary(y)
```

Boxplot mostrando a mediana (linha horizontal, quartis 25 % e 75 % (caixa verde) e observações máxima e mínima (linhas verticais) – Figura1. ylab e xlab são as etiquetas dos eixos y e x, respectivamente.

```
> boxplot(y, col=3, ylab="Massa(g)", xlab="Local do ninho")
```

Usa `range=0` para não mostrar os *outliers* (observações extremas representadas como pontos, p. ex. em amostra D) - Figura 2

```
> boxplot(y, range=0, col=3, ylab="Massa(g)", xlab="Local do
     ninho")
```

Brincadeira: tente usar diferentes cores variando o parâmetro col

Ajuda geral sobre a função plot utiliza `help(plot)`

Mostrar todas as observações cruas da amostra A na forma de gráfico barplot. Anote `y$A` significa a coluna A de observações no arranjo de dados y - Figura 3

```
> barplot(y$A,ylab="Valores crus", xlab="Valores crus",
     names=as.character(y$A),cex.names=0.7, ylim=c(0,100))
```

Para mostrar um histograma de freqüências das observações em amostra A (mais informativo sobre a distribuição das observações do que gráfico anterior!) - Figura 4

```
> hist(y$A, col=2, main="Histograma de amostra A", xlab="Classe de
     massa (g)", ylab="Freqüência")
```

Podemos variar o número de intervalos com breaks e os limites do eixo x com `xlim`. Reparou as diferenças entre os dois histogramas? Lembre, são os mesmos dados - Figura 5

```
> hist(y$A, breaks=2, xlim=(range(70,100)), col=2,
     main="Histograma de amostra A", xlab="Classe de massa (g)",
     ylab="Freqüência")
```

Algumas estatísticas úteis:

Desvio padrão (*s*) da amostra A

```
> sd(y$A)
[1] 4.033196
```

variância (s^2) da amostra A

```
> var(y$A)
[1] 16.26667
```

n, o número de observações da amostra A

```
> length(y$A)
[1] 10
```

Desvio interquartílico (a diferença entre a 3a e a 1a quartil) da amostra A

```
> IQR(y$A)
[1] 7.5
```

Reconhece a formula para calcular o erro padrão da média?

Erro padrão da média da amostra A

```
> sd(y$A)/sqrt(length(y$A))
[1] 1.275408
```

Simule um jogo de dados de 100 observações e a média e o desvio padrão especificados

```
> sim.dados<-round(rnorm(n=100,mean=74,sd=2.34));sim.dados
```

Os dados não são mostrados aqui. Lembra que os valores (e as suas estatísticas!) serão ligeiramente diferentes devido a maneira aleatória de gerar as observações.

Calcule o erro padrão e nomeie esta como o objeto R errpad

```
> errpad<-sd(sim.dados)/sqrt(length(sim.dados))
```

Calcule o Intervalo de Confiança 95 %. A amostra é grande (n>30) então basta multiplicar o erro padrão pelo valor crítico de z para a probabilidade p=0,05

```
> errpad*1.96
[1] 0.4419679
```
(ou um valor muito parecido)

Calcule o Intervalo de Confiança 95 % para uma pequena amostra (n<30). Neste caso, amostra A do jogo de dados y, é preciso usar o valor crítico de t para a probabilidade 0,05 e com o número de graus de liberdade (gl) apropriado. Lembre: pequenas amostras de dados de contagem talvez precisem ser transformadas antes!

O erro padrão da amostra A

```
> errpad<-sd(y$A)/sqrt(length(y$A))
```

Obter o valor crítico de t com 9 graus de liberdade. O valor de 2,262 representa o valor crítico da cauda superior com P=0,025. Olhe para um gráfico da curva normal para ver a distribuição da probabilidade em duas caudas.

```
> qt(0.025,9,lower.tail=FALSE)
[1] 2.262157
```

Intervalo de confiança 95% da amostra A

```
> errpad*2.262
[1] 2.884974
```

Se quiser, pode também calcular estatísticas descritivas em planilhas como Gnumeric ou OpenOffice.

Mais detalhes sobre o erro padrão e intervalos de confiança podem ser obtidos em Capítulo 14 de Vieira (1980), 7 de Levin e 11 de Fowler & Cohen (1990).

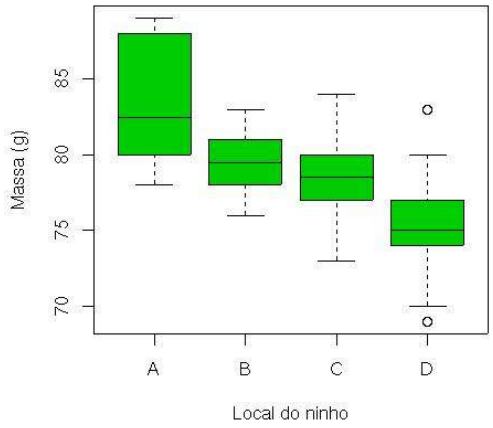

Figura 1. *Boxplot* dos dados sobre massa (g) de pássaros em quatro locais (A-D).

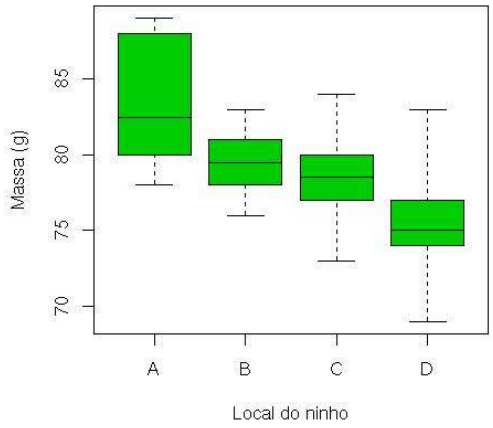

Figura 2. O mesmo *boxplot* sem os outliers da amostra D.

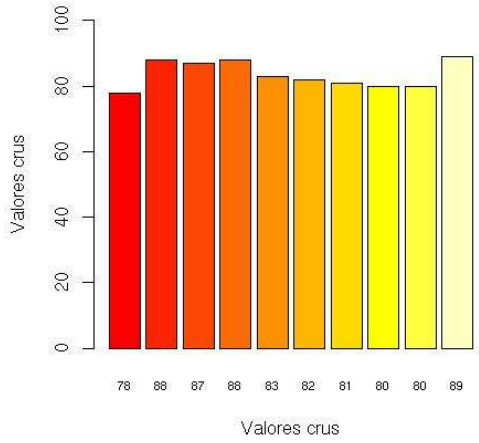

Figura 3. *Barplot* mostrando as observações cruas individuais de massa (g) da amostra A.

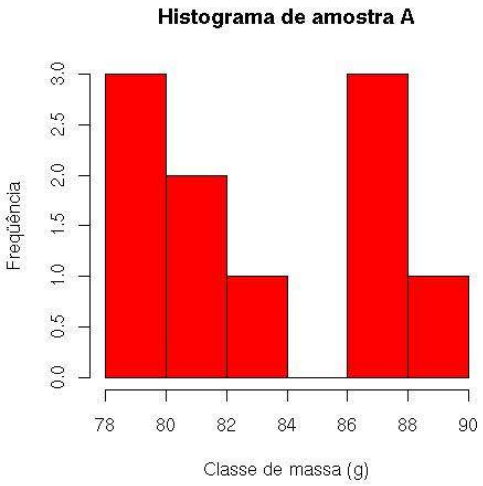

Figura 4. Os dados de amostra A organizados em um histograma.

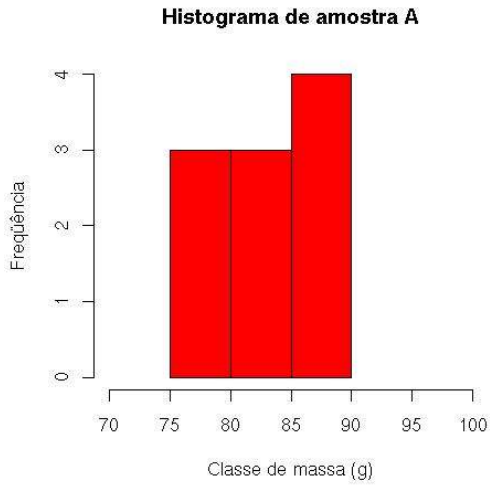

Figura 5. Histograma dos dados de amostra A, mas com classes maiores e um aumento na largura do eixo-x.

Plote a média ±desvio padrão (s) de uma série de amostras (comprimento, em µm, de larvas de bivalves de água doce, chamadas gloquídios, dos rios Tocantins, Melgaço, Ourém, Guamá, e Irituia) no arquivo larvas.csv no diretório de trabalho.

Importe os dados.

```
> z<-read.csv("larvas.csv")
```

Renomeie as colunas (por algum motivo, a importação não traz os acentos e as vírgulas)

```
> names(z)<-c("Tocantins", "Melgaço", "Ourém", "Irituia", "Guamá")
> z
   Tocantins Melgaço Ourém Irituia Guamá
1        232     218   232     232   218
2        232     218   232     232   218
3        232     218   232     232   218
.          .       .     .       .     .
.          .       .     .       .     .
.          .       .     .       .     .
30       255     289   255     255   255
```

Deve apresentar os dados mais ou menos assim. O jogo inteiro de dados não é mostrado aqui por razões de clareza e espaço.

Infelizmente não há uma função direta para plotar a média ± desvio padrão e talvez a maior razão para isso seja porque não são consideradas estimativas robustas, em contraste com a mediana ± os quartis

Inicie uma ligação com o jogo de dados z para permitir acesso fácil aos dados.

```
> attach(z)
```

NOTA IMPORTANTE: Sempre use `detach ()` antes de ligar um novo jogo de dados, especialmente se as colunas dos dois jogos tenham nomes idênticas, se não haverá problemas!

Plote as médias primeiro com os eixos devidamente etiquetados. Criamos um objeto do gráfico chamado centros que lembra a posição do centro das barras de cada rio. O tamanho das etiquetas dos eixos é reduzida pelo cex.names enquanto o ylim controla o valor mínimo e máximo do eixo-y.

```
> centros<-barplot(mean(z), cex.names=0.7, xlab="Rios",
ylab="Comprimento médio (±sd)", ylim=c(0,max(mean(z)+sd(z)*2)))
```

Acrescente as barras de erro como valores de desvio padrão da média

```
> arrows(centros,mean(z)-sd(z), centros,mean(z)+sd(z), length=0.1,
     angle=90, code=3)
```

Brincadeira: experimente valores diferentes de `angle`, `length` e `code`. Faça o gráfico com uma única cor para cada barra (p.ex. `col=3`)

Você deve ver no final um gráfico mostrando as médias de cada amostra e com barras de erro mostrando o desvio padrão – Figura 6

Termine a ligação com o jogo de dados z

```
> detach(z)
> q()
```

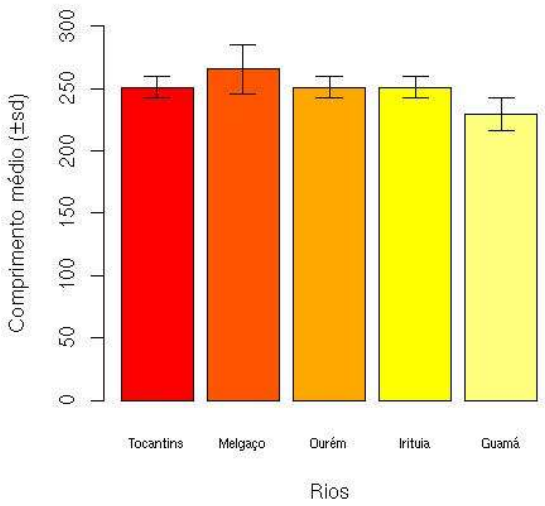

Figura 6. Média (± desvio padrão) dos dados (larvas.csv) sobre comprimento de larvas de bivalves de água doce em cinco diferentes rios amazônicos.

TRANSFORMAÇÃO DE OBSERVAÇÃOES

Excelentes discussões sobre transformações de dados podem ser encontrados em Capítulo 13 de Zar (1999) e 10 de Fowler & Cohen (1990).

No R, log() é o logaritimo natural (ln) enquanto log10() é o logaritmo à base de 10. Na verdade, log*b*() é o logaritmo à base do número *b*.

Crie uma amostra chamada meusdados e verifique os dados em seguida

```
> meusdados<-c(12,19,0,15,145,0,32,5,17); meusdados

[1]  12  19   0  15 145   0  32   5  17
```

Calcule a média e a variância da amostra. A variância da amostra é muito maior que a média, então a transformação log (x+1) é apropriada. Lembre a razão que é log (x+1) e não apenas log (x)?

```
>mean(meusdados); var(meusdados)

[1] 27.22222
[1] 2052.944
```

Transforme usando log10 (x+1) e crie um novo jogo de dados transformados chamado logmeusdados. Em seguida digite logmeusdados para visualizar as observções transformadas

```
> logmeusdados<-log10(meusdados+1); logmeusdados

[1] 1.1139434 1.3010300 0.0000000 1.2041200 2.1643529 0.0000000
     1.5185139
[8] 0.7781513 1.2552725
```

Calcule a média e a variância da amostra transformada. O que aconteceu com a razão entre a média e a variância? Experimente outras funções como a raiz quadrada: sqrt()

```
> mean(logmeusdados); var(logmeusdados)

[1] 1.037265
[1] 0.4839655
```

Para vocês que estão com pressa para entregar aquele relatório ou TCC (!!), tente a seguinte:

```
> mean(log10(meusdados+1))
> var(log10(meusdados+1))
```

Uma transformação simples é a de BoxCox que transforma a observação x para x^{λ}. Cada observação é elevada ao poder de lambda (λ).

Cria um jogo de dados de arranjos em três locais (fator Localidade) com três réplicas em cada local.

```
> bcx<-data.frame(Localidade=gl(3,3), Dados=c(meusdados))
```

Plote os dados e anote a grande diferença em variabilidade entre as amostras.

```
> boxplot (Dados~Localidade, data=bcx)
```

Carregue a biblioteca MASS

```
> library(MASS)
```

Rodar a transformação boxcox. Acrescentamos um pequeno valor 0.01 para evitar valores negativos

```
> boxcox(Dados+0.01~Localidade,data=bcx)
```

Aparecerá um gráfico mostrando a curva e o valor de lambda é o ponto onde a linha pontilhada do pico da curva intercepta o eixo-x. Podemos ler o valor de lambda usando o mouse.

Iniciar propriedade de localização de pontos do mouse

```
> locator()
```

Coloque o cursor "+" no ponto da intercepção da linha pontilhada do pico da curva com o eixo-x. Clique uma vez com o botão esquerda. Depois clique uma vez com o botão do meio (se não tiver um mouse com três botões, clique em ambas botões no mesmo tempo).

Aparecerá os coordinados do ponto na linha de comando. O valor x é o valor de lambda.

```
$x
[1] 0.2175

$y
[1] -116.7279
```

Pronto, agora para transformar os dados, basta elevar cada valor x ao poder de 0.2175. Em R, esta operação é feita usando o "chapeu", a seguir: `x^0.2175`.

Plotar os dados transformados e anote a maior similaridade em variabilidade entre amostras em relação aos dados não-transformados.

```
> boxplot (Dados^0.2175~Localidade, data=bcx)
```

```
> q()
```

ANÁLISE DE FREQÜÊNCIAS (QUI-QUADRADO)

Mais exemplos podem ser obtidos nos Capítulos 11 de Vieira (1980), 10 de Levin (1985), 13 de Fowler & Cohen (1990) e 22 e 23 de Zar (1999).

Freqüências observadas dos 10 integers 0 a 9 obtidos em uma amostra gerada aleatoriamente (n=100) pelo computador:

```
> obs<-c(10, 7, 10, 6, 14, 8, 11, 11, 12, 11)
```

A freqüência esperada ($f_{esperada}$) para cada integer é 10, mas vocês já devem ter pensado nisso ;-)

Rode qui-quadrado

```
> chisq.test(obs)

        Chi-squared test for given probabilities

data:  obs
X-squared = 5.2, df = 9, p-value = 0.8165
```

O valor de probabilidade indica que as freqüências observadas não diferem das freqüências esperadas, ou seja as observações são verdadeiramente aleatórias.

Verifique as freqüências esperadas

```
> chisq.test(obs) $expected
[1] 10 10 10 10 10 10 10 10 10 10
```

Um exemplo sobre as freqüências de moscas em uma pequena lagoa

```
> moscas<-c("D. autumnalis"=24, "D. aestivalis"=32, "D.
     amphibia"=10, "D. attica"=9)
```

Verifique as freqüências de quatro espécies de mosca

```
> moscas

D. autumnalis D. aestivalis   D. amphibia     D. attica
           24            32            10             9
```

Plote as freqüências – Figura 7

```
> barplot(moscas,xlab="Espécie",ylab="Freqüência",cex.names=0.7)
```

As freqüências observadas são significativamente diferentes da homogeneidade? (lembra como se calcula a $f_{esperada}$=18.75?)

```
>chisq.test(moscas)

Chi-squared test for given probabilities

data:  moscas
X-squared = 19.9867, df = 3, p-value = 0.0001708
```

As freqüências das quatro espécies são diferentes de uma distribuição homogênea.

UM GRAU DE LIBERDADE

Quando há apenas duas categorias, há um grau de liberdade e a correção de Yates é usada. Podemos testar a hipótese nula que a razão entre macho : fêmea não é diferente de 1 : 1. Uma amostra de 16 larvas coletadas e criadas até adulto contém 12 machos e 4 fêmeas (Fowler & Cohen, 1990). Esta razão é significativamente diferente de 1 : 1?

Coloque as freqüências observadas numa matriz com duas colunas

```
> x<-matrix (c (12,4), nc=2)
> x
      [,1]   [,2]
[1,]   12      4
```

Rode o teste para proporções iguais e com correção de Yates (`correct=TRUE`)

```
> prop.test (x, correct=TRUE)

        1-sample proportions test with continuity correction

data:  x, null probability 0.5
X-squared = 3.0625, df = 1, p-value = 0.08012
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4740838 0.9166722
sample estimates:
   p
0.75
```

O valor de qui-quadrado 3,0625 é menor que o valor crítico com v=1, portanto aceitamos a hipótese nula. Importante: os indivíduos de cada sexo devem ser dispersos em uma maneira independente.

TESTE PARA CONCORDÂNCIA ENTRE DADOS OBSERVADOS E UM MODELO MATEMÁTICO

Dados de contagem sobre o número de nematódeos vistos em uma amostra de 60 observações em uma câmera de contagem sob o microscópio. A probabilidade (P) para cada valor de x foi calculada usando a distribuição de probabilidade de Poisson, estimando o parâmetro lambda (λ) a partir da amostra.

No. nematódeos (x):	0	1	2	3	4	5	6 ou mais
Freqs. Observadas (n=60):	3	12	17	13	9	3	3
P (Poisson):	0.0743	0.193	0.251	0.218	0.141	0.074	0.0478
Freqs esperadas (P x n):	4.458	11.58	15.06	13.08	8.46	4.44	2.87

```
> Obs<-c(3,12,17,13,9,3,3)
> Ppoisson<-c(0.0743, 0.193, 0.251, 0.218, 0.141, 0.074, 0.0478)
```

Usamos a lista de probabilidades (Ppoisson) gerada pelo modelo Poisson para esta amostra de dados de contagem para calcular as freqüências esperadas. Indicamos isso no teste pelo componente p=Ppoisson.

```
> chisq.test(Obs,p=Ppoisson)

        Chi-squared test for given probabilities

data:  Obs
X-squared = 1.25, df = 6, p-value = 0.9743

Warning message:
Chi-squared approximation may be incorrect in: chisq.test(Obs, p =
Ppoisson)
```

Sabemos que devemos retirar mais um grau de liberdade quando usamos o modelo Poisson devido a estimativa do parâmetro lambda (λ). Portanto, o número de graus de liberdade é *v*=(7-2)=5. Consultando a tabela de qui-quadrado com *v*=5, vemos que o resultado permanece não significativo, ou seja, não há uma diferença entre as freqüências observadas e esperadas. O modelo Poisson descreve adequadamente a dispersão (aleatória, claro!) dos nematódeos na câmera de contagem.

Verifique as freqüências esperadas só para confirmar!

```
> chisq.test(Obs,p=Ppoisson)$expected
>[1]  4.458 11.580 15.060 13.080  8.460  4.440  2.868
```

ANÁLISE DE FREQÜÊNCIAS USANDO TABELAS DE CONTINGÊNCIA

Em duas amostras tiradas de solos diferentes foram encontradas duas espécies de tatuzinho:

	Oniscus	***Armadilidium***
Solo argiloso	14	6
Solo calcáreo	22	46

Coloque os dados numa matriz e verifique a matriz x. Pode separar comandos na mesma linha com ";"

```
> x<-matrix(c(14,22,6,46),nc=2);x

     [,1] [,2]
[1,]   14    6
[2,]   22   46
```

Rode qui-quadrado com a correção de Yates, mesmo se você esquecer, ele vai fazer a correção automaticamente ;-)

```
> chisq.test(x, correct=TRUE)

    Pearson's Chi-squared test with Yates'
        continuity correction

data:  x
X-squared = 7.5706, df = 1, p-value = 0.005933
```

Há uma associação significativa entre as variáveis, podemos inferir que *Oniscus* é associada com solo calcáreo e *Armadilidium* com solo argiloso.

Pode até plotar os dados! - Figura 8

```
> barplot(x, names=c("Oniscus","Armadilidium"),
     font=3,beside=TRUE)
> legend(1,30,c("Argiloso","Calcáreo"),fill=c(2,7))
```

MAIS UM EXEMPLO DE UMA TABELA DE CONTINGÊNCIA

Dados sobre as freqüências de espécies de mosca do gênero *Dixa* em três riachos com diferentes graus de eutrofização

	D. nebulosa	***D. submaculata***	***D. dilatata***	***D. nubilipennis***
Oligotrófico	12	7	5	17
Mesotrófico	14	6	22	9
Eutrófico	35	12	7	11

Coloque os dados em uma matriz e verifique a matriz eutrof

```
> eutrof<-matrix(c(12,14,35,7,6,12,5,22,7,17,9,11),nc=4);eutrof

     [,1]  [,2]  [,3]  [,4]
[1,]  12    7     5     17
[2,]  14    6    22      9
[3,]  35   12     7     11
```

Rode o teste qui-quadrado

```
> chisq.test(eutrof)

        Pearson's Chi-squared test

data:  eutrof
X-squared = 30.9545, df = 6, p-value = 2.586e-05
```

O resultado mostra que há uma associação significativa entre as espécies e o nível de eutrofização.

Verifique os valores esperados

```
> chisq.test(eutrof) $expected
         [,1]        [,2]        [,3]       [,4]
[1,] 15.92994    6.528662    8.878981    9.66242
[2,] 19.81529    8.121019   11.044586   12.01911
[3,] 25.25478   10.350318   14.076433   15.31847
```

As freqüências podem ser plotadas - Figura 9

```
> barplot (eutrof, names=c("D. nebulosa", "D. submaculata", "D.
          dilatata", "D. nubilipennis"), col=c(4,7,3),
          cex.names=0.6, beside=TRUE)
> legend (7,35, c("Oligotrófico", "Mesotrófico", Eutrófico"),
     fill=c(4,7,3))
```

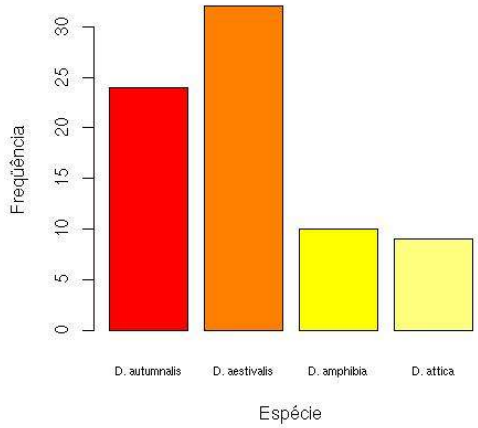

Figura 7. Barplot mostrando as freqüências de espécies de mosca amostrados na lagoa.

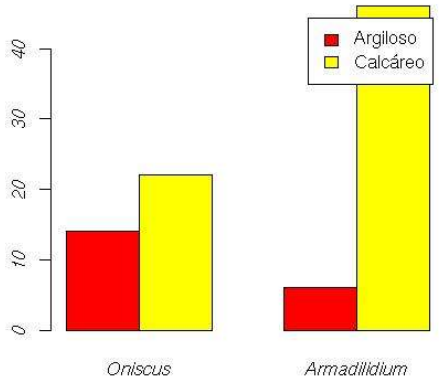

Figura 8. Barplot mostrando as freqüências de duas espécies de tatuzinho em dois tipos de solo.

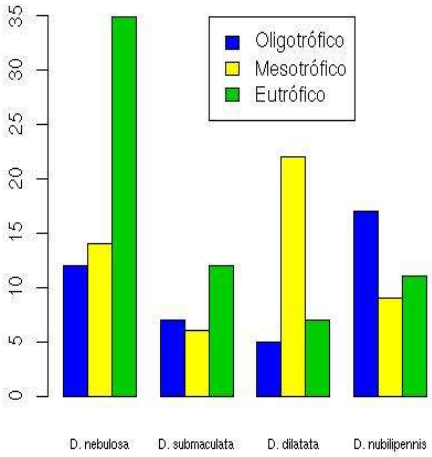

Figura 9. Barplot mostrando as freqüências de espécies de *Dixa* em diferentes graus de poluição.

TESTE-t

Mais detalhes sobre o o teste de F e o teste-t podem ser encontradas em Capítulo 12 da Vieira (1980), 16 de Fowler & Cohen (1990) e 8 de Levin (1985). Capítulo 9 de Zar (1999) apresenta uma boa discussão do uso do teste t.

PARA AMOSTRAS INDEPENDENTES

Exemplo do livro da Viera (1980) p.122-124 sobre a perda de peso (kg) em dois grupos de pacientes; cada paciente seguindo a dieta designada para seu grupo.

```
> Dieta1<-c(12,8,15,13,10,12,14,11,12,13)
> Dieta2<-c(15,19,15,12,13,16,15)
```

Verifique normalidade dos dados

```
> shapiro.test(Dieta1)

        Shapiro-Wilk normality test

data:  Dieta1
W = 0.9615, p-value = 0.8029

> shapiro.test(Dieta2)

        Shapiro-Wilk normality test

data:  Dieta2
W = 0.926, p-value = 0.5178
```

Verifique homogeneidade das variâncias usando o teste de F

```
> var.test(Dieta1,Dieta2)

        F test to compare two variances

data:  Dieta1 and Dieta2
F = 0.8, num df = 9, denom df = 6, p-value = 0.7325
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1448382 3.4557775
sample estimates:
ratio of variances
               0.8
```

Os dados são normais e as variâncias não são significativamente diferentes. Podemos prosseguir com o

teste-t para as duas amostras independentes, mas com variâncias iguais. A hipótese nula é que não há uma diferença na perda de massa média e a alternativa é que há uma diferença. O teste é de duas caudas.

```
> t.test(Dieta1,Dieta2, var.equal=TRUE,alternative="two.sided")

        Two Sample t-test

data:  Dieta1 and Dieta2
t = -2.9021, df = 15, p-value = 0.01095
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.2033162 -0.7966838
sample estimates:
mean of x mean of y
       12        15
```

Como pode ser visto no resultado, há uma diferença significativa em perda de massa média e a perda de peso é maior para os pacientes seguindo a dieta 2.

PARA AMOSTRAS PAREADAS

A massa de 10 pássaros migratórios foi medida em duas ocasiões, primeiro em agosto e os mesmos pássaros (marcados individualmente e recapturados) foram remedidos em setembro

```
> ago<-c(10.3,11.4,10.9,12.0,10.0,11.9,12.2,12.3,11.7,12.0)
> set<-c(12.2,12.1,13.1,11.9,12.0,12.9,11.4,12.1,13.5,12.3)
```

Plotamos os duas amostras – Figura 10

```
> boxplot(ago,set,names=c("Agosto","Setembro"))

> shapiro.test(ago)

        Shapiro-Wilk normality test

data:  ago
W = 0.8701, p-value = 0.1002

> shapiro.test(set)

        Shapiro-Wilk normality test

data:  set
```

```
W = 0.9302, p-value = 0.45

> var.test(ago,set)

        F test to compare two variances

data:  ago and set
F = 1.6496, num df = 9, denom df = 9, p-value = 0.4674
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4097496 6.6414787
sample estimates:
ratio of variances
          1.649649
```

Os dados são normais e as variâncias são homogêneas. Rode o teste-t com as duas amostras pareadas. O teste é de duas caudas e com as variâncias iguais.

```
> t.test(ago,set,paired=TRUE,alternative="two.sided",
var.equal=TRUE)

        Paired t-test

data:  ago and set
t = -2.6119, df = 9, p-value = 0.02818
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.6421526 -0.1178474
sample estimates:
mean of the differences
                  -0.88
```

O resultado indica que há uma diferença significativa entre as médias das duas amostras e concluímos que o aumento em massa média entre agosto e setembro é significativo.

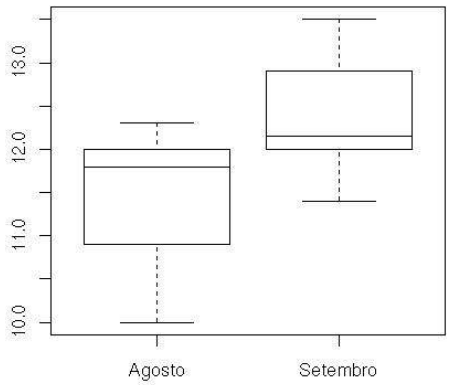

Figura 10. *Boxplot* das duas amostras pareadas de massa (g) dos pássaros em Agosto e Setembro.

ANÀLISE DE VARIÂNCIA (ANOVA)

O cálculo de ANOVA pode ser encontrado em Capítulo 10 e 11 de Zar (1999), 17 de Fowler & Cohen (1990), 13 de Vieira (1980) e 9 de Levin (1985). Uma das melhores discussões sobre as suposições do ANOVA é encontrada em Capítulos 7 e 8 de Underwood (1998).

Teste a hipótese de que a massa média (g) de uma espécie de pássaro é igual entre as quatro localidades de coleta (A-D, com n=10 indivíduos medidos em cada local).

Os dados estão contidos no arquivo massa.csv na seguinte forma:

A	B	C	D
78	78	79	77
88	78	73	69
87	83	79	75
88	81	75	70
83	78	77	74
82	81	78	83
81	81	80	80
80	82	78	75
80	76	83	76
89	76	84	75

Importe os dados

```
> massa<-read.csv("massa.csv")
```

Facilite acesso às variáveis

```
> attach(massa)
```

Calcule a média e desvio padrão dos dados massa. Existe uma diferença significativa entre as médias?

```
> mean(massa)
   A    B    C    D
83.6 79.4 78.6 75.4

> sd(massa)
       A        B        C        D
4.033196 2.503331 3.306559 4.141927
```

Verifique a maior e a menor variância. Precisamos testar se a variância maior (D) é significativamente diferente da variância menor (B). Se não for o caso então nenhuma das variâncias é significativamente diferente das outras.

```
> var(D)
[1] 17.15556

> var(B)
[1] 6.266667
```

Realize o teste de F sobre as amostras D e B

```
> var.test(D,B)

F test to compare two variances

data:  D and B
F = 2.7376, num df = 9, denom df = 9, p-value = 0.1496
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  0.6799783 11.0215159
sample estimates:
ratio of variances
          2.737589
```

O resultado mostra que não há uma diferença significativa entre as variâncias.

Alternativamente, pode ser usado o teste de Bartlett.

```
> bartlett.test(massa)
```

```
Bartlett test for homogeneity of variances

data:  massa
Bartlett's K-squared = 2.5279, df = 3, p-value = 0.4703
```

O resultado é igual ao teste de F: não há uma diferença entre as variâncias do grupo

Para verificar a normalidade das quatro amostras, usaremos o teste de Shapiro-Wilk

```
> shapiro.test(A)
```

Dados de A são normais

```
    Shapiro-Wilk normality test

data:  A
W = 0.893, p-value = 0.1835
```

```
> shapiro.test(B)
```

Dados de B são normais

```
    Shapiro-Wilk normality test

data:  B
W = 0.8992, p-value = 0.2148
```

```
> shapiro.test(C)
```

Dados de C são normais

```
        Shapiro-Wilk normality test

data:  C
W = 0.9658, p-value = 0.8494
```

```
>shapiro.test(D)
```

Dados de D são normais

```
        Shapiro-Wilk normality test

data:  D
W = 0.9463, p-value = 0.625
```

Não esqueça desprender o jogo de dados massa

```
> detach(massa)
```

Prosseguimos com ANOVA para verificar diferenças entre médias, mas precisamos modificar o jogo de dados para ser lido pelo programa e os dados serão colocados em forma vertical com uma coluna fatorial (Localidade) e uma da variável (Massa)

```
> massa.vert<-data.frame(Localidade=gl(4,10),
     Massa=c(massa$A,massa$B,massa$C,massa$D))
```

```
> massa.vert
   Localidade Massa
1           1    78
2           1    88
3           1    87
.           .     .
.           .     .
.           .     .
40          4    75
```

Deveria apresentar os mesmos dados, mas a diferença é que as observações de massa (g) são arranjados em uma coluna Massa e a primeira coluna Localidade tem os códigos dos locais de coleta (A=1, B=2, C=3, D=4). Portanto, cada observação é ainda associada com sua respectiva localidade. Compare a tabela de dados horizontais massa com os dados verticais massa.vert para confirmar.

Facilite o acesso às variáveis

```
> attach(massa.vert)
```

Define a coluna Localidade como um fator

```
> Localidade<-factor(Localidade)
```

Verifique se Localidade é mesmo um fator

```
> is.factor(Localidade)
[1] TRUE
```

Rodamos a análise de variância pelo teste aov para verificar diferenças em massa média (variável Massa) entre as amostras tiradas em diferentes ninhos (fator Localidade)

```
> massa.aov<-aov(Massa~Localidade)
```

Mostramos a tabela de ANOVA

```
> summary(massa.aov)
              Df   Sum Sq    Mean Sq   F value   Pr(>F)
Localidade     3    341.90    113.97    9.0053    0.0001390 ***
Residuals     36   455.60    12.66
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

A tabela de ANOVA mostra que a(s) diferença(s) entre as médias é (são) altamente significativa(s). Podemos concluir que há uma diferença significativa entre a massa média dos pássaros nos quatro localidades.

Entretanto, a análise não terminou ainda e para verificar entre exatamente quais pares de amostras (A-D) ocorrem diferenças signficativas usamos o teste de Tukey HSD.

```
> TukeyHSD(massa.aov, ordered = TRUE)
```

Os valores das diferenças *diff* entre as médias de pares de amostras. Localidade A=1, B=2, C=3, D=4

```
 Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = Massa ~ Localidade, data = massa.vert)

$Localidade
    diff         lwr       upr
3-4  3.2 -1.08478062  7.484781
2-4  4.0 -0.28478062  8.284781
1-4  8.2  3.91521938 12.484781
2-3  0.8 -3.48478062  5.084781
1-3  5.0  0.71521938  9.284781
1-2  4.2 -0.08478062  8.484781
```

Podemos plotar as diferenças – Figura 11

```
> plot(TukeyHSD(massa.aov,ordered=TRUE))
```

O gráfico mostra as diferenças (±intervalo de confiança 95%) entre as médias das pares de amostras. Os pares com diferenças significativas são aqueles com limites inferiores (lwr) positivos. Os detalhes do cálculo do teste de Tukey são descritos por Zar (1999) e Levin (1985). Apenas as diferenças entre A e D e A e C são significativas no nível de 5% ($p<0,05$). Portanto, foram estas amostras que contribuiram para as diferenças detectadas pela ANOVA.

Exercício: tenta plotar as médias com o desvio padrão (±s) das amostras A-D como na Figura 6. Como é possível plotar o erro padrão em vez do desvio padrão?

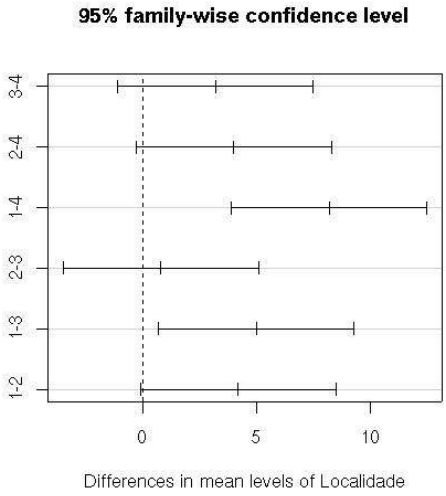

Figura 11. Diferenças significativas do teste de TukeyHSD. Os pares de amostras com diferenças significativas são A-C e A-D. A=1, B=2, C=3, D=4.

CALCULANDO ESTATÍSTICAS USANDO `tapply()`

Quando as observações estão em uma coluna vertical com uma ou mais colunas de fatores, podemos usar a função `tapply` para calcular a média, desvio padrão, variância, etc.

Calcular a massa média por localidade a partir do jogo de dados `massa.vert`:

```
> tapply(Massa, Localidade, mean)
```

Substituir `sd`, `var` e `length` no lugar de mean para calcular o desvio padrão, variância e número de observações destes dados.

No final da sessão não esqueça

```
> detach(massa.vert)
> q()
```

ANÁLISE DE VARIÂNCIA COM DOIS FATORES MODELO I (ambos fatores fixos)

Os detalhes de ANOVA com dois fatores podem ser obtidos em Capítulo 17 de Fowler & Cohen (1990) e 12 de Zar (1999).

Importe os dados (mostrados aqui parcialmente para facilitar visualização). Dados sobre a Massa (g) de pássaros medidos em dois meses (Mes=Fator 1) em 4 diferentes locais (Ninho=Fator 2). Há 10 réplicas em cada local ou seja o número total de réplicas é 80. Apenas as primeiras três réplicas de cada amostra estão mostradas aqui.

```
>x<-read.csv("starling.csv");x

Massa Mes Ninho Replicates
1     78   1      1           1
2     88   1      1           2
3     87   1      1           3
     .    .       .           .
11     78   1      2           1
12     78   1      2           2
13     85   1      2           3
     .    .       .           .
21     79   1      3           1
22     73   1      3           2
23     79   1      3           3
     .    .       .           .
31     77   1      4           1
32     68   1      4           2
33     75   1      4           3
     .    .       .           .
41     85   2      1           1
42     88   2      1           2
43     86   2      1           3
     .    .       .           .
51     84   2      2           1
52     88   2      2           2
53     91   2      2           3
     .    .       .           .
61     91   2      3           1
62     90   2      3           2
63     87   2      3           3
      .   .       .           .
71     90   2      4           1
72     87   2      4           2
```

```
73    85   2     4          3
    .    .     .          .
80    77   2     4         10
```
A última réplica. O jogo completo não é mostrado aqui.

Agora podemos ligar os dados em x para facilitar acesso

```
> attach(x)
```

Definimos e verificamos os fatores

```
> Mes<-factor(Mes); Ninho<-factor(Ninho)
> is.factor(Mes); is.factor(Ninho)
[1] TRUE
[1] TRUE
```

Plotamos as massas médias das 8 localidades para cada mês e o resultado mostra que o padrão de variação da média ao longo dos dois meses é parecida entre as localidades – Figura 12

```
> interaction.plot (Ninho,Mes,Massa)
```

Roda ANOVA sobre a variável Massa com dois fatores Mes e Ninho e com um teste para interação entre Mes e Ninho

```
> x.aov<-aov(Massa~Mes+Ninho+Mes:Ninho)
```

Exibe a tabela de ANOVA após rodar o teste

```
> summary(x.aov)
              Df   Sum Sq    Mean Sq   F value   Pr(>F)
Mes           1    1656.20   1656.20   93.6000   1.172e-14 ***
Ninho         6    608.60    101.43    5.7325    6.455e-05 ***
Mes:Ninho     3    34.20     11.40     0.6443    0.5891
Residuals     72   1274.00   17.69
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

A localidade do ninho afeta significativamente a massa média. O efeito da data de coleta (Mes) sobre massa é também significativo. O resultado mostra que não há interação significativa, confirmando a interpretação do *interaction plot* acima

Podemos verificar entre quais amostras há diferenças significativas.

```
 TukeyHSD(x.aov,ordered=T)
  Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered
```

```
Fit: aov(formula = x$Massa ~ Mes + Ninho + Mes:Ninho)

$Mes
    diff      lwr      upr
2-1  9.1 7.224955 10.97505

$Ninho
    diff        lwr      upr
3-4  3.6  0.1014801  7.09852
2-4  5.0  1.5014801  8.49852
1-4  7.4  3.9014801 10.89852
2-3  1.4 -2.0985199  4.89852
1-3  3.8  0.3014801  7.29852
1-2  2.4 -1.0985199  5.89852

$"Mes:Ninho"
      diff        lwr       upr
 [1,]  3.2 -2.6727261  9.072726
 [2,]  4.0 -1.8727261  9.872726
 [3,]  8.2  2.3272739 14.072726
 [4,]  8.8  2.9272739 14.672726
 [5,] 12.8  6.9272739 18.672726
 [6,] 14.8  8.9272739 20.672726
 [7,] 15.4  9.5272739 21.272726
 [8,]  0.8 -5.0727261  6.672726
 [9,]  5.0 -0.8727261 10.872726
[10,]  5.6 -0.2727261 11.472726
[11,]  9.6  3.7272739 15.472726
[12,] 11.6  5.7272739 17.472726
[13,] 12.2  6.3272739 18.072726
[14,]  4.2 -1.6727261 10.072726
[15,]  4.8 -1.0727261 10.672726
[16,]  8.8  2.9272739 14.672726
[17,] 10.8  4.9272739 16.672726
[18,] 11.4  5.5272739 17.272726
[19,]  0.6 -5.2727261  6.472726
[20,]  4.6 -1.2727261 10.472726
[21,]  6.6  0.7272739 12.472726
[22,]  7.2  1.3272739 13.072726
```

```
[23,]  4.0 -1.8727261  9.872726
[24,]  6.0  0.1272739 11.872726
[25,]  6.6  0.7272739 12.472726
[26,]  2.0 -3.8727261  7.872726
[27,]  2.6 -3.2727261  8.472726
[28,]  0.6 -5.2727261  6.472726
```

Há apenas dois níveis no fator Mes, então qualquer diferença significativa tem que ser entre os dois meses. Em termos do fator Ninho, o ninho 4 é diferente dos demais ninhos e ninho 1 é diferente de ninho 3.

Plote um gráfico mostrando as diferenças entre os ninhos do Tukey HSD – Figura 13

```
> plot(TukeyHSD(x.aov,"Ninho", ordered=T)
```

Tanto na tabela, como no plot, os pares de amostras com diferenças significativas são aquelas cujos limites inferiores (lwr) são positivos. Neste caso, apenas três pares de ninhos têm diferenças significativas: 1-3, 1-4, 2-4, e 3-4.

Calcular a massa média por mês e ninho a partir do jogo de dados x:

```
> tapply(Massa, Mes:Ninho, mean)
```

Desligar o jogo de dados x

```
> detach (x)
```

Termina a sessão

```
> q()
```

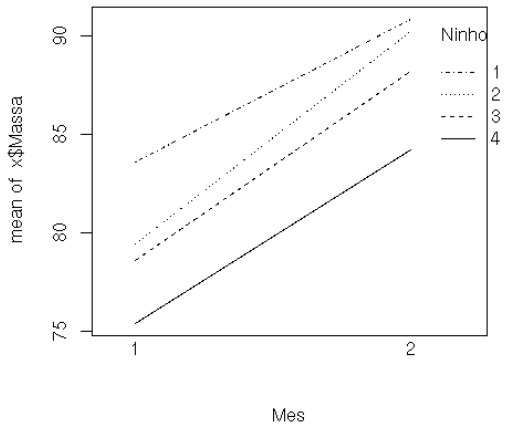

Figura 12. Gráfico de interação: a variação na massa média entre os locais é parecida entre os dois meses.

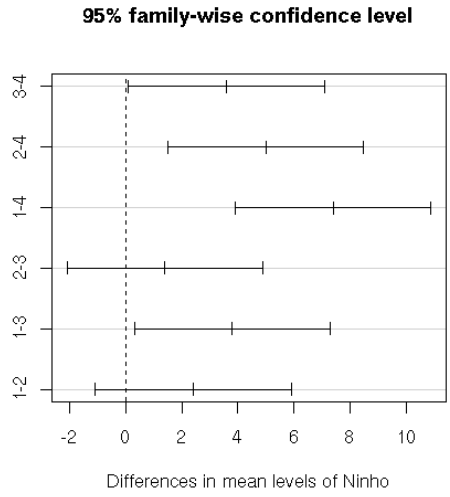

Figura 13. Diferenças significativas do teste TukeyHSD. Os pares de ninhos com diferenças significativas são 1-3, 1-4, 2-4 e 3-4.

MAIS UM EXEMPLO DE ANOVA MODELO I DE DOIS FATORES (ambos fatores fixos)

Exemplo de Zar (1999), p.233. Dados sobre o efeito de tratamento e sexo sobre níveis de plasma. Fator 1 = tratamento, 1 = sem hormônio, 2 = com hormônio. Fator 2 = sexo, 1 = fêmea, 2 = macho.

Importe os dados.

```
> x<-read.csv("plasma.csv"); x
   plasma tratmento sexo replicas
1    16.5         1    1         1
2    18.4         1    1         2
3    12.7         1    1         3
4    14.0         1    1         4
5    12.8         1    1         5
6    14.5         1    2         1
7    11.0         1    2         2
8    10.8         1    2         3
9    14.3         1    2         4
10   10.0         1    2         5
11   39.1         2    1         1
12   26.2         2    1         2
13   21.3         2    1         3
14   35.8         2    1         4
15   40.2         2    1         5
16   32.0         2    2         1
17   23.8         2    2         2
18   28.8         2    2         3
19   25.0         2    2         4
20   29.3         2    2         5
```

Agora podemos ligar os dados em x para facilitar acesso

```
> attach(x)
```

Definimos os fatores

```
> tratmento<-factor(tratmento); sexo<-factor(sexo)
```

Verificamos a interação visualmente – Figura 14

```
> interaction.plot(tratmento,sexo,plasma)
```

Não há interação entre tratamento e sexo e ainda há um grande efeito de tratamento enquanto o efeito de sexo é pouco (pouca distância entre as linhas paralelas). Rode ANOVA sobre a variável plasma com dois fatores tratamento e sexo e com um teste para interação entre os dois fatores

```
> x.aov<-aov(plasma~tratmento+sexo+tratmento:sexo)

> summary(x.aov)
               Df   Sum Sq  Mean Sq     F value    Pr(>F)
tratmento      1   1386.11   1386.11    60.5336    7.943e-07 ***
sexo           1   70.31      70.31     3.0706     0.09886
tratmento:sexo 1   4.90       4.90      0.2140     0.64987
Residuals      16  366.37    22.90
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

O resultado mostra que não há interação significativa, não há efeito de sexo e que o efeito de tratamento é altamente significativo, confirmando a interpretação do gráfico de interação (*interaction.plot*) – Figura 14

Calcular a média e desvio padrão por tratamento e sexo a partir do jogo de dados x:

```
> tapply(plasma, tratamento:sexo, mean)
> tapply(plasma, tratamento:sexo, sd)

> detach(x)
> q()
```

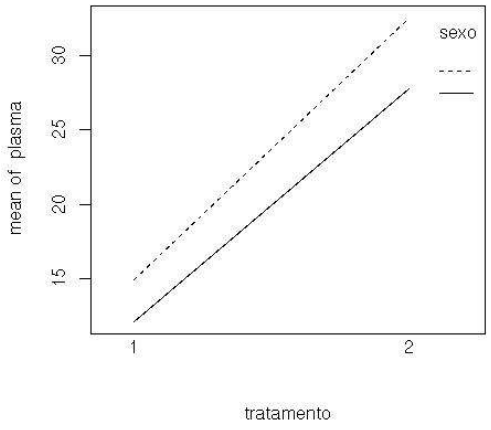

Figura 14. Gráfico de interação mostrando a ausência de interação (linhas paralelas), pouco efeito de sexo (distância curta entre linhas) e um grande efeito de tratamento (diferença grande na média entre tratamentos 1 e 2). Sexo: ----fêmea, ___macho.

TESTE DE WILCOXON RANK SUM TEST (equivalente a TESTE U DE MANN-WHITNEY)

PARA OBSERVAÇÕES INDEPENDENTES

Detalhes do teste podem ser encontrados em Zar (1999) p. 145, Capítulo 10 de Levin (1985) e 16 de Fowler & Cohen (1990).

Para testar para uma diferença entre as medianas de duas amostras: números de besouros capturados durante a noite em 8 armadilhas no hábitat A e em 7 armadilhas no hábitat B

```
> A<-c(8,12,15,21,25,44,44,60)
> B<-c(2,4,5,9,12,17,19)
```

Verifique as estatísticas das amostras usando o boxplot (Figura 15)

```
> boxplot(A,B, names=c("A","B"))
```

Rode o teste de Wilcoxon

```
> wilcox.test(A,B)

        Wilcoxon rank sum test with continuity
        correction

data:  A and B
W = 47.5, p-value = 0.02761
alternative hypothesis: true mu is not equal to 0

Warning message:
Cannot compute exact p-value with ties in: wilcox.test.default(A,
B)
```

O resultado do teste mostra que há uma diferença significativa entre as medianas das amostras A e B. O Wilcoxon Rank Sum Test (equivalente ao Teste U de Mann-Whitney) testa a hipótese nula na qual as medianas das distribuições de A e B diferem pelo valor de μ. Desde que o valor padrão de μ=0, signifique que a hipótese nula é que as medianas são iguais

TESTE DE WILCOXON SIGNED RANK TEST

PARA OBSERVAÇÕES PAREADAS

Mais informações podem ser obtidas em Zar (1999) p. 145, Capítulo 16 de Fowler & Cohen (1990).
O Wilcoxon Signed Rank Test é igual ao Wilcoxon Rank Sum Test com a diferença que as duas amostras têm observações pareadas.

A massa de 10 pássaros migratórios foi medida em duas ocasiões, primeiro em agosto e os mesmos pássaros (marcados individualmente e recapturados) foram medidos novamente em setembro

```
> ago<-c(10.3,11.4,10.9,12.0,10.0,11.9,12.2,12.3,11.7,12.0)
> set<-c(12.2,12.1,13.1,11.9,12.0,12.9,11.4,12.1,13.5,12.3)
```

Plote as estatísticas das amostras (Figura 9)

```
> boxplot(ago,set,names=c("Agosto","Setembro"))
```

Parece que os pássaros ganharam peso entre agosto e setembro, mas essa diferença é significativa?

Rode o teste de Wilcoxon agora com paired=TRUE, para indicar se as observações são pareadas

```
> wilcox.test(ago,set, paired=TRUE)

 Wilcoxon signed rank test

data:  ago and set
V = 8, p-value = 0.04883
alternative hypothesis: true mu is not equal to 0
```

O resultado mostra que há uma diferença significativa entre as massas medianas das duas amostras e podemos inferir que os pássaros têm uma massa significativamente maior em setembro

```
> q()
```

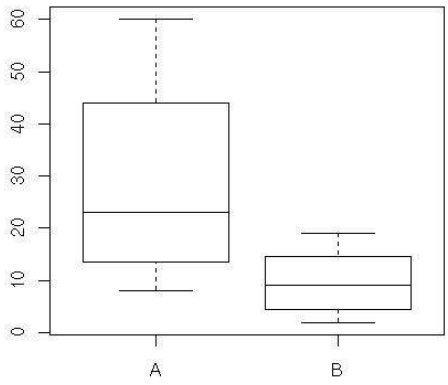

Figure 15. Número mediano de besouros capturados em duas amostras independentes em hábitat A e hábitat B.

TESTE DE KRUSKAL-WALLIS (ANOVA NÃO PARAMÉTRICA DE UM FATOR)

Mais informações podem ser obtidas na p. 195 de Zar (1999), Capítulo 10 de Levin (1985) e 16 de Fowler & Cohen (1990).

Os dados não precisam ser distribuídos normalmente, mas as distribuições das amostra devem ser semelhantes entre si. Há diferenças de opinião sobre a questão da homogeneidade das variâncias (veja Zar,1999 e Fowler & Cohen, 1990 versus Underwood, 1998).

O número de orquídeas em cinco quadrados em quatro campos (A-D).

```
> A<-c(27,14,8,18,7)
> B<-c(48,18,32,51,22)
> C<-c(11,0,3,15,8)
> D<-c(44,72,81,55,39)
```

Mostre as estatísticas das amostras

```
> summary(A)
 Min.     1st Qu.   Median    Mean 3rd Qu.   Max.
7.0  8.0  14.0 14.8 18.0 27.0
> summary(B)
Min. 1st Qu.   Median    Mean 3rd Qu.   Max.
18.0 22.0 32.0 34.2 48.0 51.0
> summary(C)
Min. 1st Qu.   Median    Mean 3rd Qu.   Max.
0.0  3.0  8.0  7.4  11.0 15.0
> summary(D)
Min. 1st Qu.   Median    Mean 3rd Qu.   Max.
39.0 44.0 55.0 58.2 72.0 81.0
```

Plote as estatísticas das amostras – Figura 16

```
> boxplot(A,B,C,D, names=c("A","B","C","D"), ylab="No. orquídeas",
     col=3)
```

Coloque os dados em uma forma vertical. Há um fator, Campo, e uma variável, o número de Orquideas

```
> orq.dados<-data.frame(Campo<-gl(4,5), Orquideas<-c(A,B,C,D))
```

Verifique os dados

```
> orq.dados
```

```
      Campo    Orquideas
1       1         27
2       1         14
3       1          8
4       1         18
5       1          7
6       2         48
7       2         18
8       2         32
9       2         51
10      2         22
11      3         11
12      3          0
13      3          3
14      3         15
15      3          8
16      4         44
17      4         72
18      4         81
19      4         55
20      4         39
```

Execute o teste Kruskal-Wallis

```
> kruskal.test(Orquideas~Campo, orq.dados)

        Kruskal-Wallis rank sum test

data:  Orquideas by Campo
Kruskal-Wallis chi-squared = 14.602, df = 3, p-value = 0.002190
```

O teste mostra que há uma diferença (ou diferenças) significativa(s) entre as medianas do grupo de quatro amostras. Há vários métodos para efetuar comparações múltiplas não-paramétricas (Zar, 1999, p. 223).

Alternativamente, o teste pode ser rodado assim, dando o mesmo resultado!

```
> kruskal.test (list(A,B,C,D))

        Kruskal-Wallis rank sum test

data:  list(A, B, C, D)
```

```
Kruskal-Wallis chi-squared = 14.602, df= 3, p-value = 0.002190
```

Se tiver replicação desigual, use este método mais flexível para arranjar os dados

```
> orq.dados<-c(A,B,C,D)

> orq.dados

[1] 27 14  8 18  7 48 18 32 51 22 11  0  3 15
[15]  8 44 72 81 55 39
```

Deve ser observado que o `[1]` e `[15]` indicam a primeira e a décima-quinta observação, respectivamente.

Organize um arranjo de 1 a 4 amostras com 5 réplicas em cada e com as etiquetas “A” a “D”.
Nota: SE TIVESSE uma réplica a mais na amostra A, então o número de réplicas seria c(6,5,5,5).

```
> g<-factor(rep(1:4, c(5, 5, 5, 5)),labels=c("A","B","C","D"))

> kruskal.test (orq.dados, g)

Kruskal-Wallis rank sum test

data:  orq.dados and g
Kruskal-Wallis chi-squared = 14.602, df = 3, p-value = 0.002190
```

Novamente o resultado é idêntico aos anteriores!

```
> q()
```

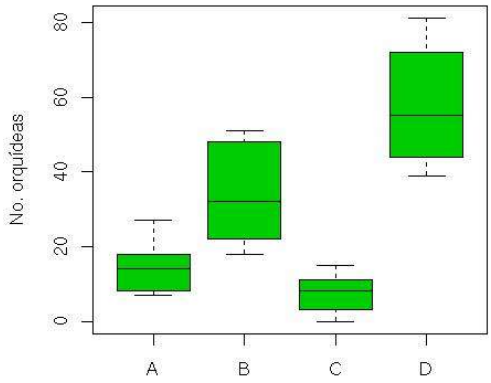

Figura 16. Boxplot dos dados A, B, C, D.

CORRELAÇÂO E REGRESSÂO

Veja Capítulo 6 e 7 de Vieira (1980), 11 de Levin (1985), 14 e 15 de Fowler & Cohen (1990) e 17 e 19 de Zar (1999).

Coeficiente de correlação de Spearman (não paramétrico)

Diversidade de gafanhotos (y) em relação ao número de anos depois da aplicação de pesticidas (x)

```
> x<-c(0,1,3,5,9,12,13,15,21,25)
> y<-c(0,0.19,0.15,1.49,1.10,1.12,1.61,1.42,1.48,1.92)

> x
 [1]  0  1  3  5  9 12 13 15 21 25
> y
 [1] 0.00 0.19 0.15 1.49 1.10 1.12 1.61 1.42 1.48 1.92
```

Plotamos as duas variáveis em um gráfico de dispersão – Figura 17

```
> plot(x,y, ylab=”Diversidade de gafanhotos”, xlab=”No. anos após
     aplicação de pesticida”)
```

Parece que há uma correlação positiva entre as duas variáveis e podemos testar se esta correlação é real ou não usando a função `cor.test()`.

Testamos a significância da relação

```
> cor.test(x,y,method="spearman",alternative="two.sided")

        Spearman's rank correlation rho
```

```
data:  x and y
S = 32, p-value = 0.008236
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.8060606
```

O valor da estimativa do coeficiente de Spearman r_s é 0,806 e a correlação é altamente significativa. Podemos inferir que há uma correlação positiva e significativa entre diversidade de gafanhotos e o tempo após aplicação de pesticidas.

Coeficiente de correlação de Pearson (paramétrico)

```
x <-c(6.6,6.9,7.3,7.5,8.2,8.3,9.1,9.2,9.4,10.2)
y <-c(86,92,71,74,185,85,201,283,255,222)
```

Plotamos a relação em um gráfico de dispersão – Figura 18

```
> plot(x,y, ylab="Massa do peixe (g)", xlab="Comprimento do
otólito")
```

Testamos a significância da relação

```
> cor.test(x,y,method="pearson",alternative="two.sided")

        Pearson's product-moment correlation

data:  x and y
t = 4.3546, df = 8, p-value = 0.00243
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4428186 0.9608852
sample estimates:
      cor
0.8386271
```

O valor da estimativa do coeficiente de Pearson r é 0,8386 é a correlação é altamente significativa. Podemos inferir que há uma correlação positiva e significativa entre comprimento do otólito e a massa do peixe.

Um exemplo de regressão simples linear: dados sobre a massa de fertilizante (gm^2) aplicada (x) e a

colheita de grama obtida (gm^{-2}) para cada aplicação de fertilizante (y).

```
x <-c(25,50,75,100,125,150,175,200,225,250)
y <-c(84,80,90,154,148,169,206,244,212,248)
```

Aplique a regressão linear simples pelo método de quadrados mínimos e mostre os resultados

```
> fertilizante.lm<-lm(y~x)
> summary(fertilizante.lm)

Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q Median     3Q    Max
-22.79 -11.07  -5.00  12.00  29.79

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 51.93333   12.97904   4.001  0.00394 **
x            0.81139    0.08367   9.697 1.07e-05 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 19 on 8 degrees of freedom
Multiple R-Squared: 0.9216,     Adjusted R-squared: 0.9118
F-statistic: 94.04 on 1 and 8 DF,  p-value: 1.067e-05
```

O intercepto y é 51,933 e o gradiente é 0,81139. A regressão tem a fórmula y = 0,81139 x + 51,933

No R, a significância da regressão é testada pela ANOVA. O valor de F é significativo com 1 e 8 graus de liberdade e concluamos que a regressão é altamente significativa.

Detalhes da ANOVA para testar a significância da regressão:

```
> anova(fertilizante.lm)

Analysis of Variance Table

Response: y
           Df  Sum Sq   Mean Sq   F value    Pr(>F)
x          1   33947     33947    94.041     1.067e-05 ***
Residuals  8   2888       361
---
```

```
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

As quantidades importantes são a variância da regressão (33947) e a variância residual (361) com 1 e 8 graus de liberdade, respectivamente.

Plote os dados

```
> plot(x,y,ylim=c(0,300),xlab="Massa de
     fertilizante",ylab="Colheita de grama")
```

Acrescente a reta da regressão ao gráfico

```
> abline (fertilizante.lm)
```

Alternativamente, se quiser uma linha mais “enxuta” use

```
> lines (x,0.81139*x+51.933)
```

Adicione a fórmula e a significância da regressão ao gráfico – Figura 19

```
> text(100,200,"y=0,81139x+51,933, p<0,01")

> q()
```

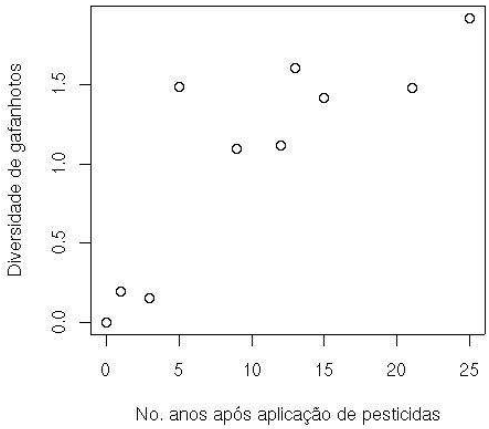

Figura 17. Diversidade de gafanhotos em relação ao número de anos após aplicação de pesticidas.

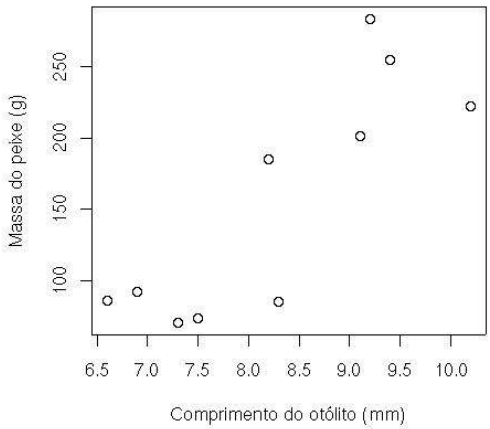

Figura 18. Massa do peixe em relação ao comprimento do otólito.

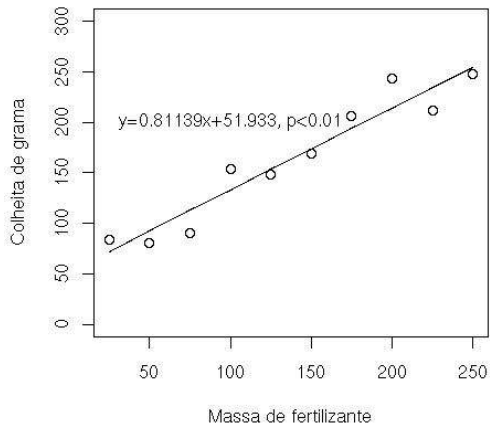

Figura 19. Regressão simples linear entre a massa de fertilizante e a colheita de grama.

FUNÇÕES MISCELÂNEAS ÚTEIS

Para planejar uma amostragem aleatória, usa a função `sample()`

Tire uma amostra aleatória de 10 observações/indivíduos/posições a partir de uma população de 100

```
> x<-sample(1:100,10) ;x
```

Arranje as observações em ordem crescente. Brincadeira: qual é a mediana?

```
> sort(x)
```

Crie um ranking das observações.

```
> rank(x)
```

Qualquer gráfico produzido em R pode ser exportado em várias formas incluindo postscript, pdf, jpeg,png,bmp, etc.

Para ver as opções dos diferentes formatos p.ex. png use

```
> help(png)
```

Exemplo

```
> x<-c("A"=2,"B"=5,"C"=12);x
 A  B  C
 2  5 12
> barplot(x)
> png (filename="meugráfico.png")
```

A extensão .png significa *Portable Network Graphic* ou Gráfico da Rede Portátil e é mais limpa que o formato .jpeg. Este último só deveria ser usado com fotos.

Repite os comandos para desenhar o gráfico

```
> barplot(x)
```

Termine o gráfico e fecha o dispositivo jpeg

```
> dev.off()
```

Agora deve ter um arquivo gráfico chamado meugráfico.jpg no seu diretório de trabalho que pode ser inserido em um documento de texto - Figura 20

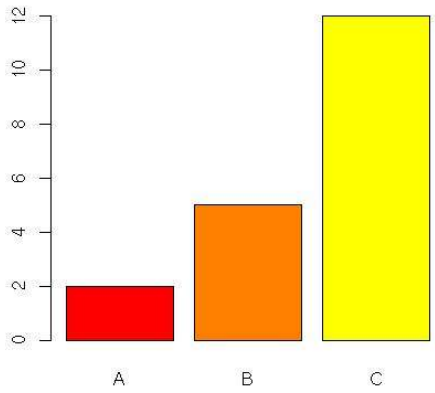

Figura 20. Barplot do jogo de dados x no exemplo anterior.

ANÁLISE MULTIVARIÁVEL BÁSICA

Análise multivariável é um tópico complexo e há muitas técnicas complementares e alternativas. Consulte a bibliografia para decidir quais são apropriados para seus dados.

Importar dados (dadosmv.csv), especifique a presença de um cabeçalho e que a primeira coluna contem os nomes das fileiras

```
> x<-read.csv("dadosmv.csv",header=T,row.names=1)
```

Para usar algumas funções é preciso carregar a biblioteca `vegan`. Esta pode ser baixada do site do R e instalada antes de prosseguir.

Carregar a biblioteca `vegan`

```
> library(vegan)
```

Calcular o índice de diversidade Shannon usando o logaritmo natural (opção padrão)

```
> diversity(x,"shannon")
```

```
Sítio A        Sítio B        Sítio C        Sítio D
0.6365142      1.0296530      0.6931472      1.1637233
```

Calcular a riqueza de espécies

```
> specnumber(x)
Sítio A Sítio B Sítio C Sítio D
      2       3       2       6
```

Para ver métodos de padronização use o comando `decostand`. R tem muitas funções integradas para a transformação de dados

Calcular a matriz de dissimilaridade Bray-Curtis sobre dados não transformados, não padronizados

```
> x.dist<-vegdist(x,"bray")
```

Criar o agrupamento hierárquico com ligação média sobre a matriz de dissimilaridade Bray-Curtis

```
x.clust<-hclust(x.dist,"average")
```

Plotar o dendrograma (veja opções de plot digitando `?hclust`)

```
> plot(x.clust, hang=-0.5)
```

Carregar a biblioteca MASS

```
> library(MASS)
```

Rodar análise nmMDS (escalonamento multidimensional não-métrico)

```
> x.mds<-isoMDS(x.dist,k=3,maxit=9999)
initial  value 0.000000
final  value 0.000000
converged
```

Anote que é recomendado o uso de `initMDS` e `postMDS` que estão disponíveis na biblioteca `vegan`.

O valor final de estresse (neste caso é muito pequeno, quase zero) é uma medida de quanto a ordenação representa os valores de dissimilaridade entre Sítios na matriz. Quanto menor o estresse melhor é essa representação.

Plotar a ordenação MDS em branco (tipo = nula), inicialmente

```
> plot(x.mds$points, type="n")
```

Acrescentar as etiquetas dos locais

```
> text(x.mds$points, labels=as.character(row.names(x)))
```

Pode embelezar os gráficos (etiquetas de eixos, cores etc.) usando as opções de `plot`. Veja `demo (graphics)` para ver a capacidade gráfica de R. Há também a possibilidade de realizar ANOSIM, PCA, e outras técnicas multivariáveis no R.

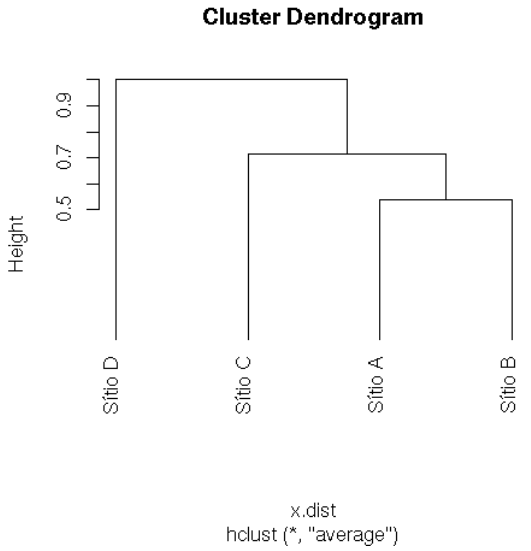

Figura 21. Agrupamente hierárquico (ligação média)

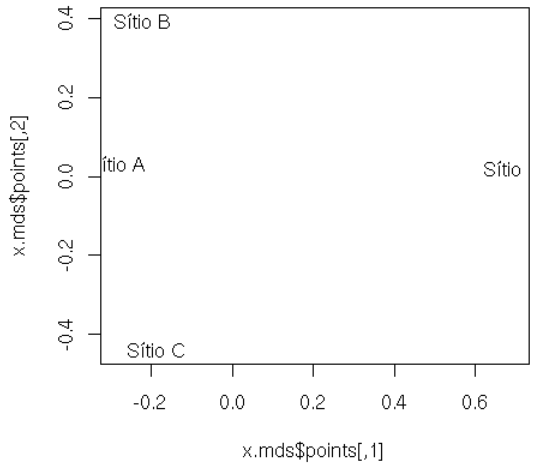

Figura 22. Ordinação nmMDS. Estresse<0.0001

DISTRIBUIÇÕES DE PROBABILIDADE

Podemos criar distribuições de probabilidade de diferentes famílias (Gaussiano ou Normal, Binomial, Poisson e Binomial Negativa) usando valores personalizados dos parâmetros das distribuições.

Distribuição normal

Para visualizar uma distribuição normal de probabilidade (parâmetros: μ e σ) com a média da amostra =74.0 e s =2,34 (estimativas de μ e σ, respectivamente) para as observações entre 68 e 80 use:

```
> barplot(dnorm(68:80, 74.0, 2.34), col=2, names=c(68:80),
     xlab="Comprimento da planta (mm)", ylab="Probabilidade")
```

Distribuição Binomial

Para visualizar uma distribuição Binomial de probabilidade com k=8 ensaios, p=q=0,5 e para o número de resultados nomeados entre 0 a 8 use:

```
> barplot(dbinom(0:8,8,0.5), names=c(0:8), xlab="Número de
     machos", ylab="Probabilidade")
```

Para visualizar uma distribuição Binomial de probabilidade com k=8 ensaios, p=0,8 e q=0,2 e para o número de resultados nomeados entre 0 e 8 use:

```
> barplot(dbinom(0:8,8,0.8), names=c(0:8),xlab="Número de
     indivíduos de D. autumnalis", ylab="Probabilidade")
```

Distribuição Poisson

Para visualizar uma distribuição Poisson de probabilidade com λ=4 (a média é uma estimativa do parâmetro λ) e para observações entre 0 e 10 use:

```
> barplot(dpois(0:10,4.0),col=3,names=c(0:10), xlab="Número de
     indivíduos por unidade de amostragem", ylab="Probabilidade")
```

Distribuição Binomial Negativa

Para visualizar uma distribuição Binomial Negativa de probabilidade com=μ=2, σ^2=s^2=5.0, kappa k=1.33 (veja Elliott, 1983 para estimação de parâmetros) e para observaçoes entre 0 e 8 use:

```
> barplot(dnbinom(0:8, mu=2, size=1.33), col=3, names=c(0:8),
     xlab="Número de indivíduos por unidade de amostragem",
     ylab="Probabilidade")
```

Criando distribuições de probabilidade permitem nós testar se dados biológicos sobre a dispersão de organismos no seu ambiente conformem aos modelos Binomial (dispersão regular), Poisson (dispersão aleatória) ou Binomial Negativa (dispersão contagiosa ou agregada). Veja Elliott (1983) ou Fowler & Cohen (1990) para mais detalhes.

REFERÊNCIAS BIBLIOGRÁFICAS

Elliott, JM (1983) Some methods for the statistical analysis of samples of benthic invertebrates. Freshwater Biological Association Scientific Publication No. 25.

Fowler J & Cohen L (1990) Practical statistics for field biology. John Wiley & Sons, Chichester.

Levin J (1985) Estatística aplicada a ciências humanas. Harper & Row do Brasil, SP.

Ihaka R & Gentleman R (1996) R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics 5 (3): 299-314.

Jongman RHG, Ter Braak CJF & Van Tongeren, OFR (1995) Data analysis in community and landscape ecology. Cambridge University Press.

Underwood AJ (1998) Experiments in ecology. Their logical design and interpretation using analysis of variance. Cambridge University Press.

Vieira S (1980) Introdução à bioestatística. Editora Campus

Zar JH (1999) Biostatistical analysis. Prentice Hall, Upper Saddle River, NJ.

Copyright (c) 2004 Colin Robert Beasley.
É dada permissão para copiar, distribuir e/ou modificar este documento sob os termos da Licença de Documentação Livre GNU, Versão 1.2 ou qualquer versão posterior publicada pela Free Software Foundation; sem nenhum Seção Invariante, sem nenhum Texto da Capa da Frente, e sem nenhum Texto da Quarta-Capa. Uma cópia da licença pode ser consultada no endereço http://www.fsf.org/licence/. Uma tradução não-oficial desta para o português do Brasil pode ser encontrado no endereço:

http://www.ead.unicamp.br/minicurso/bw/texto/fdl.pt.html.