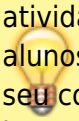
Análise de Dados

alunos2.dat



Na próxima aula você deverá entregar um relatório sobre as atividades desenvolvidas neste roteiro, em grupo de dois alunos. Faça a atividade com muita atenção, discutindo com seu companheiro/a de grupo. Anote todos os pontos importantes e aquilo que foi solicitado para a construção do relatório.

Um bom trabalho científico começa com (1) uma boa pergunta, (2) um bom planejamento experimental, (3) coleta de dados e segue com (4) análise desses dados para responder a pergunta definida e (5) um bom relato do que foi feito e as implicações dos resultados obtidos. Em nossa atividade, pulamos as fases (1) e (2) para focarmos na experiência de campo, com a coleta dos dados e agora nessa fase onde os dados já foram coletados e precisamos entender o que eles estão nos dizendo e relatá-los. Essa foi uma decisão didática, não deve ser seguida! Nunca inicie a coleta de dados sem ter passado um bom tempo se dedicando às fases (1) e (2). Elas são essenciais para o bom procedimento científico. Não há análise complexa ou escrita sedutora que conserte um trabalho mal planejado ou um pergunta irrelevante. Como a análise está intrinsecamente acoplada à pergunta, vamos revisitar uma pergunta simples que podemos responder e sua hipótese relacionada.

Perguntas

- Palmitos tem crescimento diferente quando em solos com diferentes regimes de alagamento?

Alternativamente, podemos pensar em uma pergunta mais específicas, por exemplo, sabendo que os palmitos tem preferência por áreas alagadas, podemos perguntar se:

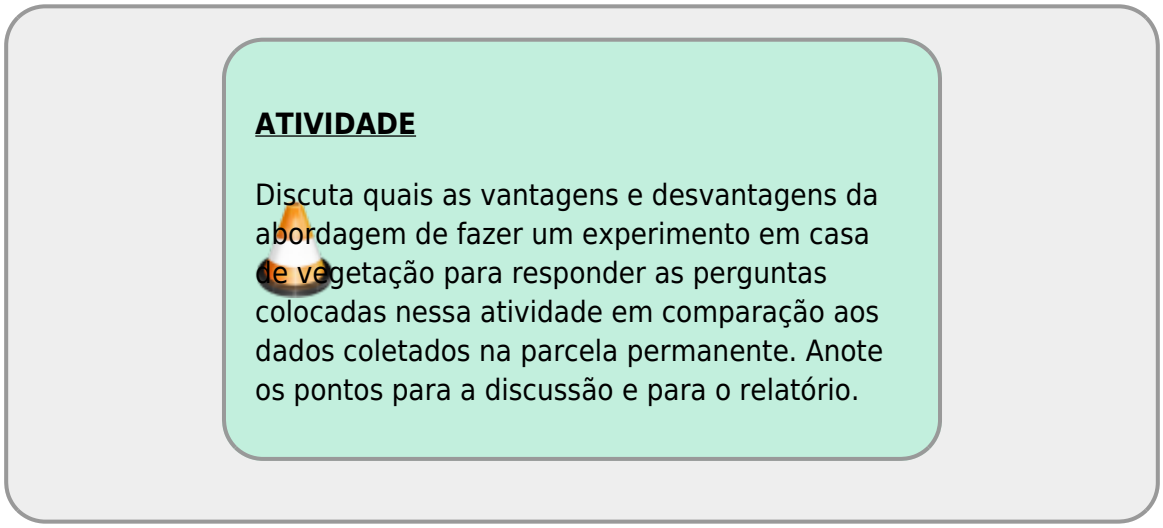
- Palmitos crescem melhor em solos alagados?

ATIVIDADE

1. Discuta com seu colega de grupo e relate sucintamente a diferença que existe entre as perguntas acima, do ponto de vista biológico e na análise de dados;
2. Quais premissas estão relacionadas a cada um delas ou a ambas?

Desenho experimental

Nessa fase, definimos quais informações iremos coletar, como essa informação esta distribuída nos diferentes tratamentos e quais variáveis de confusão iremos controlar. No caso de experimentos em que manipulamos os objetos, algumas dessas etapas ficam mais fáceis de planejar. Por exemplo, podemos cultivar palmitos em vasos na casa de vegetação, deixado parte dos vasos sem aberturas no fundo e o restante com aberturas. Como a água fica retida no vaso sem drenagem, simulamos um efeito de alagamento. Além disso, podemos regar os vasos todos os dias para garantir o alagamento e medir o crescimento dos palmitos nos dois tratamentos. No nosso trabalho de campo, apenas sorteamos parcelas em áreas alagadas e sem alagamento para coletar os dados sobre os palmitos.



Explorando os dados

Conhecer os dados antes de iniciar as análises é essencial. Para isso vamos explorá-los antes de qualquer análise estatística.

A natureza das variáveis

Até agora nossa pergunta ou hipótese está no campo da teoria. Falamos de “crescimento” e “regime de alagamento” e precisamos conectar essas variáveis teóricas com aquilo que medimos em campo e que melhor representa o que estamos querendo descrever. É preciso estabelecer a conexão entre as variáveis medidas e a hipótese que se quer testar. Normalmente, uma hipótese é construída de forma que há uma **variável resposta** e uma ou mais **variáveis preditoras**. Essa nomenclatura está associada ou fato de que pressupomos um relação causal entre elas. Além disso, devemos entender qual a natureza dessas variáveis, se são provenientes de uma contagem, de uma medida contínua, de um probabilidade, quais são seus limites máximos e mínimos teóricos, entre outras características, Uma outra questão é qual a nossa unidade amostral e como calcular as variáveis para esse nível amostral. Por exemplo: estamos representando o “regime de alagamento” pelo contraste da restinga melhor drenada (“restinga seca”) daquela mais alagada (“restinga alagada”). Ou seja, uma variável categórica com dois níveis. Essa característica foi definida para a unidade amostral da subparcela de 20×20 m.

- Baixe o arquivo de dados palmito2016;
- reconheça cada uma das variáveis e o que os seus valores significam. Veja a aba **metadados**;
- separe dados que não fazem sentido ou que são erros para testar a hipótese em questão;
- discuta com o seus colega de grupo qual medida melhor representa o crescimento;
- ajuste os valores coletados para que sejam comparáveis aos valores do censo de 2009;
- calcule o crescimento para cada árvore no período de 2009 a 2016;
- faça um boxplot que represente a hipótese que queremos responder, mostrando a variação entre tipos de solo;

Representação do Boxplot

O boxplot é uma boa representação da distribuição de variáveis contínuas cujas preditoras são variáveis categóricas. Os gráficos de diferentes níveis da variável categórica pode ser colocado lado a lado as distribuições comparadas.

- construa uma planilha limpa apenas com os dados para analisar contendo cada linha uma observação e nas colunas as variáveis;
- salve esse arquivo em formato texto (csv ou txt) e reconheça qual o símbolo usado para separar os campos no arquivo texto [1)]

- Para limpar os dados e fazer tabelas sínteses use as ferramentas de filtro e tabelas dinâmicas em um programa de planilha eletrônica (excel, libreoffice ou planilha google).
- Construir boxplot é um pouco mais complexo nesses programas. Uma boa forma de começar é pensar o boxplot como um conjunto de

informações, os quartis, ordenados em gráfico de barras justapostas.

ATIVIDADE

Para o relatório, descreva os dados de crescimento de palmito para a parcela de modo geral e as diferenças que existam entre os tipos de solo. Uso o boxplot como gráfico base para essa apresentação.

R

A partir desse ponto o roteiro vai usar algumas ferramentas de interface para o R, um ambiente de programação para análise de dados e produção gráfica bastante poderosa. Caso tenha uma versão antiga do R antiga ou não tenha, instale a versão mais recente, baixando o linque abaixo:

Instalando o R

- baixe os arquivo de instalação
  - windows
  - Mac OS X
- execute os arquivos
- siga as intruções

Rcmdr

Rcommander (Rcmdr) é uma ferramenta de interface gráfica que facilita o uso do R. O objetivo do idealizador, John Fox, foi criar uma ferramenta para o ensino de estatística no ensino superior sem a necessidade de ensinar a linguagem de programação do R. Ele mesmo, no artigo que apresenta o Rcmdr, desaconselha o seu uso para análise de dados, pois acredita que o aprendizado da linguagem de programação auxilia no entendimento da própria estatística e modelagem dos dados. Vamos no valer dela para fazer a análise exploratória dos dados de palmito.

Abra o ambiente de programação R e digite a seguinte linha de comando:

```
library(Rcmdr)
```

Caso o pacote não esteja instalado é necessário uma linha adicional de comando:

```
install.packages("Rcmdr")
library(Rcmdr)
```

Depois de digitar o código acima uma janela de interface do Rcmdr deve-se abrir na sua área de trabalho como abaixo:

Atividade

- no menu **Data** importe os dados do arquivo construído no tópico anterior denominando-o de *palmito*;
- no menu **Statistics** abra a sessão **Summary** e explore as variáveis;
- no menu **Graphs** faça um boxplot para a taxa de crescimento e defina o *group variable* em **Plot by** como sendo o solo;
- aguarde maiores informações dos professores;

Nesse tópico passamos pela análise exploratória dos dados brutos e terminamos com dados processados que podem ser usados para a análise estatística.

Teste de hipótese estatística

Um bom gráfico, em geral, representa os dados coletados indicando posição de valores como: média, mediana e quantis. Essas posições permitem visualizar a distribuição dos dados em cada nível de tratamento. Ou seja, a variabilidade associada à nossa amostra para cada tipo de solo. Biólogos, mais que outros profissionais, sabem que indivíduos de uma população diferem entre si. Então as médias de palmitos provenientes de dois tipos de solos podem diferir porque por acaso em uma delas amostramos algumas plantas com taxas maiores do que na outra. Essa é explicação mais parcimoniosa para qualquer diferença observada, pois não pede mais nada além da variação que ocorre sempre que tomamos amostras. Chamamos esta explicação de hipótese nula estatística. A questão estatística aqui é: a variação encontrada entre os tipos de solo pode ser gerada apenas pelo fato de termos tomado amostras da população de palmito? Para responder essa hipótese vamos usar um método de permutação para construir cenários onde qualquer variação encontrada entre tipos de solo seja devida à variação associada ao processo de amostrar uma população de palmito. Ou seja, o solo não é responsável por nenhuma variação além da variação ao acaso naturalmente encontrada em amostras de palmito.

Os passos para fazer isso são:

Sequencia de um teste de permutação

- escolher uma estatística de interesse que represente a hipótese a ser testada;
- definir o cenário da hipótese nula;
- calcular a estatística de interesse nesse cenário nulo;
- refazer o passo anterior para criar uma distribuição de valores no cenário nulo (pseudovalores);
- comparar o valor observado com a distribuição dos pseudovalores;
- calcular o quanto improvável é encontrar valores iguais ou extremos ao valor observado neste cenário nulo;

Rsampling

Para executar o teste acima vamos usar uma ferramenta chamada Rsampling que conecta o programa R a um navegador web. Para evitar problemas de conexão iremos trabalhar com o Rsampling instalado no computador. Para isso precisamos seguir as instruções descritas no repositório do Rsampling e descritas no próximo tópico.

Instalando

- baixar o Rsampling para windows ou mac/linux
- extrair os arquivos compactados;
- abrir o programa R no computador e digitar as seguinte linha de comando:

```
install.packages(c("Rsampling", "shiny", "PerformanceAnalytics"))
library(shiny)
language <- "pt"
runApp("endereço da pasta Rsampling-shiny")
```

No código acima mude “endereço da pasta Rsampling-shiny” para a localização da pasta que foi descompactado o Rsampling-shiny. Ao descompactar automaticamente, o nome padrão desta pasta é **Rsampling-shiny-1.6.1**, mas pode ser renomeada da maneira que desejar.

O endereço da pasta na ultima linha de comando acima precisa estar entre aspas (“”). Além disso no Windowns a localização das pastas usa como padrão a barra invertido \ para separar a hierarquia do caminho. Por exemplo:

```
runApp("C:\Users\Alunos\Downloads\Rsampling-shiny-1.6.1")
```

Rsampling

Se não houve erros na sessão anterior abriu-se uma aba no seu navegador como abaixo:

ATIVIDADES

- navegue para a aba **Entrada de dados** e escolha a opção **upload file**;
- selecione o arquivo que foi salvo na primeira parte desse roteiro;
- defina:
  - se há cabeçalho [2)];
  - qual o separador de campo do arquivo de texto;
  - qual o marcador de decimal numérico;
  - se há marcadores para definir campos de texto;
- siga para a aba **Estatística**
- defina:
  - a estatística de interesse;
  - a coluna com a variável categórica contendo os níveis de tratamento;
  - a coluna com a variável numérica resposta;
- confira se a estatística de interesse observada foi corretamente calculada;
- siga para a aba **Reamostragem**
  - defina o tipo de aleatorização;
  - deixe a caixa **Com reposição?** sem ser ativada;
  - na caixa **Alternativa:** defina que tipo de teste quer fazer: bicaudal ou unicaudal com a respectiva direção;
  - discuta e relate qual a diferença entre fazer um teste uni ou bicaudal em termos da hipótese que é testada;
  - não selecione **reamostragem estratificada**;
  - desmarque as caixas: **Mostrar extremos?** e **Mostrar regiões de aceitação?**

- clique em atualizar gráfico;

Você verá um histograma dos valores da estatística de interesse ser construído à medida que as randomizações acontecem.

Nível de significância

No histograma construído pelo Rsampling você verá a proporção das randomizações que tiveram um valor igual ou mais extremo que a diferença observada. Se essa probabilidade é baixa, há pouca chance que o cenário nulo tenha gerado os dados. Neste caso, podemos tomar a decisão de rejeitar a hipótese nula. Mas quão baixa deve ser esta probabilidade para rejeitarmos a hipótese nula? Isso é uma convenção, e varia entre áreas de conhecimento. Nas ciências naturais usamos p < 0.05, ou uma probabilidade menor que 5%. Clique na caixa **Mostrar regiões de aceitação?** e você verá em cinza a região que corresponde a 95% da distribuição dos valores no cenário nulo. Qual sua conclusão?

Unidade amostral

Nosso teste até aqui foi feito considerando cada árvores como sendo uma unidade da amostra e permutamos árvores entre os solos. Entretanto, as árvores foram amostradas e o tipo de solo definido dentro de subparcelas de 20×20 m, o que define uma dependência da informação retirada de árvores na mesma parcela. Essa falta de independência fere um principio fundamental nos testes estatísticos tradicionais.

- discuta e relate a importância da aleatorização das amostras e quais problemas pode trazer para as análises feitas;
- planeje e execute um teste de randomização em que o problema da dependência das árvores na mesma parcela é superado;
- discuta diferenças encontradas;
- faça gráficos comparando os daps em 2009 com os de 2016 para cada solo, onde cada ponto é uma árvore;
- faça o mesmo representando cada ponto como sendo uma subparcela e a média do crescimento nela;
- interprete e discuta os resultados destes gráficos.

RELATÓRIO

Orientações para a elaboração do relatório

Para o relatório da análise de dados coletados na viagem de campo, responda os itens abaixo:

1) Em relação às duas perguntas apresentadas:

- *Palmitos têm crescimento diferente quando ocorrem em solos com diferentes regimes de alagamento?*
- *Palmitos crescem melhor em solos alagados?*

1.a) Relate sucintamente a diferença que existe entre as perguntas acima, do ponto de vista biológico e na análise de dados.(máximo 10 linhas)

1.b) Quais premissas estão relacionadas a cada um delas e/ou a ambas?(máximo 10 linhas)

2) Discuta quais as vantagens e desvantagens da abordagem de fazer um experimento em casa de vegetação para responder as perguntas colocadas nessa atividade em comparação com os dados coletados na parcela permanente.(máximo 10 linhas)

3) Liste todas as decisões que foram tomadas no momento de criar a "planilha limpa". Por exemplo: indique se foram retiradas as plantas cortadas, mortas, inclinadas e/ou com epífitas; ou ainda, indique se retiraram valores que consideraram errados, discrepantes, etc.(máximo 10 linhas) Apresente o número de observações que permaneceram na "planilha limpa" para o solo alagado e para o solo seco.(2 linhas)

4) Apresente os boxplots (com os dados da "planilha limpa") que respondam a pergunta que queremos responder, mostrando a variação entre tipos de solo. A partir desse gráfico, descreva os dados de crescimento de palmito para a parcela de modo geral e as diferenças que existam entre os tipos de solo.(máximo 10 linhas)

5) Após a realização das análises estatísticas no pacote Rsampling, qual a sua conclusão? (máximo 10 linhas)

6) Discuta e relate qual a diferença entre fazer um teste uni ou bicaudal em termos da hipótese que é testada (máximo 8 linhas).

Os textos devem ser formatados com a fonte Times New Roman, tamanho 12 e margens da folha com, no máximo, 1,5cm.

Atenção: O relatório deve ser enviado para o e-mail: amzmartini@usp.br até o dia 26/10, às 13:59 para a turma do integral e às 18:59 para a turma do noturno.

[1)]
use um editor básico de texto como o notepad após salvar o arquivo
[2)]
primeira linha de dados com nome das variáveis