Teste-t no R

A pergunta que nos propusemos a responder em aula foi de quanto pesa uma andorinha. Para isso, fomos à campo, coletamos e pesamos 100 indivíduos. Os dados obtidos foram tabulados e se encontram nesta planilha.

Entrando os dados

Vamos começar carregando os dados da planilha no R. Para isso, use a função *read.csv*. Execute o comando `help(read.csv)` caso seja necessário.

Para que sua planilha fique salva e você possa utilizá-la inúmeras vezes, você deve atribuir a leitura do .csv para um objeto. Exemplo:

```
andorinhas = read.csv("andorinhas.csv")
```

Note que para esse comando funcionar você deve ter ajustado seu diretório de trabalho para o diretório onde o arquivo se encontra. Utilize a função *setwd* para esse fim.

Calculando medidas de tendência central e variabilidade

Primeiramente, vamos dar uma olhada geral nos dados utilizando o comando *head*. Agora que você já sabe a estrutura do seu *data.frame* podemos prosseguir. Podemos calcular a média usando a função *mean*, a variância com *var* e o desvio padrão com *sd* ou tirando a raíz quadrada da variância com *sqrt*. Note que nosso *data.frame* possui duas colunas, mas o que nos interessa no momento é apenas a coluna **peso**. Uma maneira de extrair apenas essa coluna é utilizando *$*:

```
andorinhas$peso #imprime o vetor dos pesamos
media = mean(andorinhas$peso) # calcula a média dos pesos e atribui à variável média
dp = sd(andorinhas$peso) # calcula o desvio padrão e atribui à variável dp
cv = dp/media # calcula o coeficiente de variação e atribui à variável cv
```

Tente calcular o desvio padrão tirando a raíz quadrada da variância e salve num objeto chamado `dp2`. Agora execute o comando abaixo e interprete-o:

```
dp == dp2
```

Qual foi o resultado? As duas maneiras de calcular o desvio padrão são equivalentes? O que o operador `==` faz?

Teste-t

Depois de explorarmos um pouco nossos dados, podemos realizar um teste estatístico. Suponha que previamente se sabia que andorinhas do México pesavam 12g, mas sabemos que as andorinhas brasileiras são maiores. Queremos testar

então se as andorinhas mexicanas e brasileiras possuem pesos diferentes. A estatística de interesse que podemos usar é a média. Então, nossa hipótese nula estatística deve ser de que andorinhas brasileiras pesam 12g. Se refutarmos H0, ficamos com a hipótese alternativa de que andorinhas brasileiras têm, em média, peso (μ_b) diferente das andorinhas mexicanas.

$$H_0 : \mu_b = 12$$

$$H_1 : \mu_b \neq 12$$

Dê uma olhada na função *t.test* do R antes de prosseguir (`?t.test`). Precisamos de um vetor x com os valores de peso, hipótese alternativa *two.sided* e $mu = 12$. Tente rodar sozinho antes de prosseguir.

```
t.test(x = andorinhas$peso, alternative = "two.sided", mu = 12)
```

Segue o resultado que deverá aparecer no seu console:

```
One Sample t-test

data:  andorinhas$peso
t = 14.171, df = 99, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 12
95 percent confidence interval:
14.67358 15.54420
sample estimates:
mean of x
15.10889
```

Tente interpretar o que o R nos deu como resultado. Utilize o R apenas como calculadora e calcule a estatística t e a média dos pesos das andorinhas. O novo resultado bate com o que obtivemos anteriormente? Você refutaria a sua hipótese nula com esse p-valor? Por quê?

Um outro resultado interessante que o R mostra é a estimativa intervalar, ou intervalo de confiança. A idéia do tamanho amostral e do desvio padrão são incorporados nessa estimativa. O que você acha que aconteceria se o tamanho amostral fosse menor? E se o desvio padrão fosse maior?

Teste-t com duas populações

No exercício anterior comparamos a média do peso das andorinhas brasileiros com um valor conhecido a priori de 12g. No entanto, com dados reais quase nunca teremos um valor conhecido a priori. Quando quisermos comparar duas populações devemos utilizar o teste-t de duas amostras.

Uma outra questão que pode surgir ao olharmos mais atentamente aos nossos dados é a de dimorfismo sexual. Anteriormente, fizemos uma média pra todas as andorinhas, sem levar em conta o sexo. Será que há diferença no peso médio entre os sexos? Veremos!

Primeiramente, vamos separar os dados de machos e fêmeas em dois objetos distintos. Tente usar a função *subset* para fazer isso.

```
machos = subset(andorinhas, andorinhas$sexo == "M")
femeas = subset(andorinhas, andorinhas$sexo == "F")
```

Não prossiga até ter certeza que entendeu o que os dois comandos acima fazem. Vamos agorar explorar nossos dados, fazendo as medidas de tendência central e variabilidade. Olhando só para essas medidas, você diria que há diferença entre o peso médio de machos e fêmeas?

Pense sozinho na hipótese nula estatística e realize o teste-t para ver se há diferença entre os sexos. Lembre-se de olhar o *help* da função *t.test* para realizar o teste. Você diria que essa espécie apresenta dimorfismo sexual pra característica peso?

Explorando graficamente

Antes mesmo de fazer um teste estatístico temos que conhecer os dados com que estamos lidando. Uma maneira bem prática e eficaz de fazer isso é com gráficos exploratórios. Vamos aprender como fazer alguns gráficos com o R, que é uma ferramenta poderosa para esse fim.

Boxplot

Um dos gráficos mais úteis e que melhor resume variáveis contínuas é o *boxplot*. Esse gráfico sumariza algumas estatísticas de uma maneira visual bem intuitiva. O boxplot é composto por 5 medidas: limite superior, 3^o quartil, mediana, 1^o quartil e limite inferior. Como interpretamos esse gráfico? Na caixa central, encontra-se 50% das nossas observações. Observando a simetria entre os dois lados a partir da mediana podemos saber se nossos dados são simétricos ou assimétricos. Só aqui já temos uma boa noção da distribuição dos nossos dados.

Fazer um *boxplot* no R é bem simples. Tente utilizar a função *boxplot* para fazer um boxplot do peso das andorinhas.

```
boxplot(andorinhas$peso)
```

Podemos ainda dividir as observações em classes determinadas por uma varíavel categórica. Nosso *data.frame* possui duas colunas: peso e sexo. Podemos dividir em dois boxplots, um para fêmeas e outro para machos. Para fazermos isso temos que usar uma notação de fórmula do R: `y ~ x`. A nossa varíavel y seria o peso das andorinhas e a variável x o sexo. Tente fazer sozinho o boxplot por sexo.

Histograma

Outro gráfico bastante utilizado e que nos mostra mais detalhadamente a distribuição dos dados é o histograma. Para fazermos um histograma no R basta utilizar a função *hist*. Tente fazer o histograma para a distribuição dos pesos das andorinhas. Você acha que a distribuição se aproxima de uma normal? Faça agora dois histogramas, um para machos e outro para fêmeas. As distribuições se aproximam da normal?