Mais exercícios de aquecimento em R

Diana Santos

Formação BILLIG, maio de 2020

1 Roupa na Literateca

Leia a folha de registo de `https://www.linguateca.pt/formacaoBILLIG/roupaPorAutorVariante.tsv`, que tem as menções de roupa por obra, nas obras da Literateca v. 4.9.

```
roupa <- read.table("roupaPorAutorVariante.tsv")
colnames(roupa) <- c("obra","palroupa","palobra","variante")
summary(roupa)
```

Veja a variação entre as diferentes obras. Para as poder comparar, tem de usar uma medida/quantidade que não dependa do tamanho da obra: o número relativo de palavras de roupa.

```
roupa$rouparel<-roupa$palroupa/roupa$palobra
plot(roupa$rouparel)
boxplot(roupa$rouparel)
```

Veja quais os casos com mais roupa:

```
roupa[which(roupa$rouparel>0.010),]$obra
[1] Filosofia_de_um_par_de_botas História_comum
[3] Três_consequências           A_carteira
```

(Todos contos de Machado de Assis).

Comparando entre obras portuguesas e brasileiras

```
boxplot(roupa$rouparel~roupa$variante)
boxplot(roupa[which(roupa$rouparel<0.010),]$rouparel
 ~roupa[which(roupa$rouparel<0.010),]$variante)
```

A visualização dos diagramas de caixa parece indicar que as obras brasileiras falam mais de roupa. Mais tarde veremos como verificar isto com um teste estatístico.

Antes disso, vamos ver se o autor tem importância. Leiam a folha de registo

```
autores<-read.table("https://www.linguateca.pt/formacaoBILLIG/AutoresObras.tsv")
colnames(autores)<-c("autor","tam_autor","obra","tam_obra","lixo")
```

que indica o autor para cada obra.

Juntando ambas as folhas de registo, podemos ter uma nova folha de registo com informação do tamanho da roupa por autor.

```
novaroupa<-merge(roupa,autores,by.x=c("palobra","obra"),
by.y=c("tam_obra","obra"))
attach(novaroupa)
boxplot(rouparel~autor)
```

Parece haver bastante variação entre autores. Escolhendo, por exemplo, apenas quatro

```
selecao<-novaroupa[novaroupa$autor=="JulDin"|novaroupa$autor=="CamCBra"|
novaroupa$autor=="CoeNet"|novaroupa$autor=="MacAss",]
boxplot(selecao$rouparel~selecao$autor)
```

vemos que todos os valores de autor ainda se encontram acessíveis... (nos níveis (levels) desse fator). Se quisermos reduzir a folha de registo apenas aos quatro autores, é preciso executar o seguinte comando:

```
selecao$autor<-selecao$autor[drop=TRUE]
boxplot(selecao$rouparel~selecao$autor)
```

Voltando agora à folha de registo que tem todos os dados, ordenemos os autores pelo número de palavras de roupa (relativas) que usam

```
totalRoupaPorAutor<-xtabs(novaroupa$palroupa~novaroupa$autor)
novoautores<-merge(autores,totalRoupaPorAutor,by.x=c("autor"),
by.y=c("novaroupa.autor"))
autoresRoupa<-unique(subset(novoautores, TRUE, c("autor","Freq","tam_autor"))
autoresRoupa$rouparelautor<-autoresRoupa$Freq/autoresRoupa$tam_autor
barplot(autoresRoupa[order(autoresRoupa$rouparelautor),]$rouparelautor,
names=autoresRoupa[order(autoresRoupa$rouparelautor),]$autor,las=2)
```

Vemos que os autores que mencionam menos roupa são: ManLar, MJTM e OslLim e os que mencionam mais são ChiBua, DomPel e MoaScl.

Voltando à distribuição por variante,

```
ultimoautores<-merge(autoresRoupa,novaroupa)
autoresfinal<-unique(subset(ultimoautores,TRUE,c("autor","rouparelautor",
"variante"))
boxplot(autoresfinal$rouparelautor~autoresfinal$variante)
```

embora pareça visualmente que os autores brsieleiros em média usam mais roupa, temos de usar um teste estatístico para verificar isso.

Quando queremos comparar apenas duas amostras, usa-se o t-test:

```
t.test(autoresfinal$rouparelautor~autoresfinal$variante)

Welch Two Sample t-test

data:  autoresfinal$rouparelautor by autoresfinal$variante
t = 3.6088, df = 62.86, p-value = 0.0006105
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.0003475811 0.0012102334
sample estimates:
mean in group BR mean in group PT
     0.002049547      0.001270639
```

Que nos diz que a probabilidade de ter os resultados que obtivemos vindos de uma mesma população é .0006105, ou seja, tão baixa que podemos rejeitar a hipótese. Ou por outras palavras, que a diferença entre estes dois grupos de autores é estatisticamente significativa.

Claro está que, para fazer um estudo como deve ser, teríamos de comparar autores comparáveis, e em particular não misturar, como fizemos aqui, textos completos com excertos de um ou dois capítulos!

Se quiséssemos simplesmente comparar as proporções (total) das palavras de cor nos textos brasileiros e portugueses (algo muito menos informativo), o teste seria um teste de proporções

```
tamRoupaBR<-sum(roupa[variante=="BR",]$palroupa)
tamBR<-sum(roupa[variante=="BR",]$palobra)
tamRoupaPT<-sum(roupa[variante=="PT",]$palroupa)
tamPT<-sum(roupa[variante=="PT",]$palobra)
prop.test(c(tamRoupaBR,tamRoupaPT),c(tamBR,tamPT))
```

```
2-sample test for equality of proportions with continuity correction

data:  c(tamRoupaBR, tamRoupaPT) out of c(tamBR, tamPT)
X-squared = 111.02, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 0.0001192440 0.0001747403
sample estimates:
     prop 1      prop 2
0.001618312 0.001471320
```

O resultado é que podemos rejeitar a hipótese de que a proporção é igual.