Alguns testes estatísticos em R

Diana Santos

Abril de 2019

Este mini-curso é para servir de continuação ao curso da EBraLC de 2012, indicando em mais pormenor como fazer alguns dos exercícios mencionados, mas não explicados no texto anterior, usando a linguagem R.

1 Diferenças estatisticamente significativas

A estatística trabalha com amostras. Duas amostras (ou medições) diferentes é bem provável que sejam diferentes. Por isso é preciso outro conceito, o de diferença estatisticamente significativa.

A ideia subjacente é a seguinte: só se a diferença for extremamente improvável, ou seja se a sua probabilidade de acontecer for muito baixa, é que dizemos que não pode ser devida ao acaso.

Há duas maneiras de calcular esta probabilidade:

1. usando métodos paramétricos, ou seja, postulando uma dada distribuição (função estatística) e fazendo as contas,

2. ou usando métodos não paramétricos, por exemplo através de simulação.

1.1 Proporções

Vejamos a questão da comparação entre duas proporções: a sua diferença é estatisticamente significativa?

Vamos ver uns dados sobre o medo em diferentes géneros textuais: ler os dados, e visualizar num gráfico de barras.

```
medo<-read.table("http://folk.uio.no/dssantos/cursoR/medo.txt",
header=TRUE)
barplot(medo$fear)
```

É preciso indicar quais os identificadores:

```
barplot(medo$fear,names.arg=medo$genre)
```

Mas de facto deveríamos usar proporções, visto que há mais texto em alguns géneros do que noutros, por isso criamos uma nova coluna e visualisamos outra vez:

```
medo$densidade<-medo$fear/medo$total
barplot(medo$densidade*10000,names.arg=medo$genre,ylab="medo
por dez mil palavras",ylim=c(0,12))
```

Vamos ordenar pela densidade

```
ordem <-order(medo$densidade)
medoOrdenado <-medo[ordem,]
barplot(medoOrdenado$densidade*10000, names.arg=medoOrdenado$genre,
ylab="medo por dez mil palavras", ylim=c(0,10))
```

Agora que temos uma ideia das diferenças, temos de testar se são estatisticamente significativas:

```
prop.test(x=c(8150,429), n=c(29821714,2193638), alternative="greater")
```

Podemos rejeitar a hipótese nula de que são duas amostras da mesma população.

Comparando enciclopédia e livro didático, por outro lado

```
prop.test(x=c(32,52), n=c(286559,426766), alternative="greater")
```

Para estes dois géneros, não é possível rejeitar a hipótese nula de que vêm da mesma população.

Nestes exemplos, usámos uma aproximação paramétrica que pressupõe uma distribuição binomial. O que é uma aproximação, visto que postula que a escolha de cada palavra (ser de medo ou não) é independente das outras palavras.

1.2 Correlação

Duas propriedades numéricas variam de forma relacionada? Ou são independentes uma da outra? Vejamos a menção do Natal (festa cristã) em Portugal e no Brasil, por semana.

```
natal<-read.table("http://folk.uio.no/dssantos/cursoR/NatalCHAVEvar.txt")
colnames(natal)<-c("semana","BR","PT")
natal$semana<-factor(natal$semana)
attach(natal)
```

Vejamos diferentes formas de visualizar isto

```
plot(semana,PT)
points(semana,BR)
plot(BR,PT)
```

Calculemos agora a correlação:

```
cor(BR,PT)
cor(PT,BR)
```

Mas o que interessa é a significância estatística da mesma, por isso deve fazer-se sempre `cor.test`:

```
cor.test(BR,PT)
```

De facto, a função `lm` (linear model) calcula a reta que representa a covariação

```
summary(lm(PT~BR))
lines(x=c(0,450),y=c(-9,-9+450*1.34400))
```

Podem entreter-se com os dados em relação à menção ao verão no Brasil e em Portugal:

```
veraoBrasil<-read.table("http://folk.uio.no/dssantos/cursoR/veraoBrasil.txt")
veraoPortugal<-read.table("http://folk.uio.no/dssantos/cursoR/veraoPortugal.txt"
colnames(veraoBrasil)<-c("semana","veraoBR")
colnames(veraoPortugal)<-c("semana","veraoPT")
verao<-merge(veraoBrasil,veraoPortugal, by.x=c("semana"),by.y=c("semana"))
```

1.3 Tabelas de contingência

Quando estamos a cruzar propriedades/qualidades, temos contagens. Não é uma folha de registo! (dataframe).

```
tabela <- matrix(c(38,14,11,51), nrow=2)
tabela
colnames(tabela) <- c("Olhos azuis", "Olhos castanhos")
rownames(tabela) <- c("Loiro", "Cabelo castanho")
tabela
chisq.test(tabela)
```

O teste do chi quadrado testa se há alguma relação entre estas propriedades, ou se elas são independentes. Quanto mais pequeno o p, mais relação há.