

Mais alguns exemplos usados na formação BILLIG

Diana Santos

Maio de 2020

1 O perigo das tabelas de contingência

O exemplo de Agresti sobre os julgamentos e a cor do defensor nos EUA:

```
simpson<-read.table("http://folk.uio.no/dssantos/cursosR/simpson.txt",  
header=TRUE)  
attach(simpson)  
xtabs(tamanho~vitima+veredito)
```

Aproveito para mostrar uns comandos muito úteis sobre tabelas de contingência, que tornam os resultados mais legíveis, transformando os números em proporções, e mesmo em percentagens:

```
prop.table(xtabs(tamanho~vitima+veredito))  
prop.table(xtabs(tamanho~vitima+veredito),2)  
prop.table(xtabs(tamanho~vitima+veredito),1)*100  
prop.table(xtabs(tamanho~veredito+defensor),2)*100
```

Da última tabela podemos pois concluir (?) que os advogados negros são melhores, visto que em apenas 7,8% dos casos levam à morte do seu cliente, enquanto que os advogados brancos têm quase 11% de condenados à morte. (De facto, não poderíamos concluir isso, se fizéssemos um teste de chi-quadrado, mas adiante).

Mas se olharmos para cada caso separadamente (de vítimas brancas e vítimas negras), vemos outra coisa muito diferente:

```
vitimasbrancas<-simpson[simpson$vitima=="branca",]  
prop.table(xtabs(vitimasbrancas$tamanho~vitimasbrancas$veredito+  
vitimasbrancas$defensor),2)*100  
vitimasnegras<-simpson[simpson$vitima=="negra",]
```

```
prop.table(xtabs(vitimasnegras$tamanho~vitimasnegras$veredito+
vitimasnegras$defensor),2)*100
```

Olhando para cada caso em separado, vemos que os advogados negros para cada caso têm mais números de condenados à morte.

O mesmo caso se dá nos exemplos de proporção de homens/mulheres que entraram numa universidade,

```
entraram<-matrix(c(233,88,324,194),ncol=2)
colnames(entraram)<-c("sim", "não")
row.names(entraram)<-c("homens","mulheres")
entraramLiteratura<-matrix(c(82,289,72,192),nrow=2,byrow=TRUE)
colnames(entraramLiteratura)<-c("sim", "não")
row.names(entraramLiteratura)<-c("homens","mulheres")
prop.table(entraramLiteratura,1)*100
entraramEngenharia<-rbind(c(151,35),c(16,2))
colnames(entraramEngenharia)<-c("sim", "não")
row.names(entraramEngenharia)<-c("homens","mulheres")
prop.table(entraramEngenharia,1)*100
```

Aqui a variável com que não se entrou inicialmente foi: quantas pessoas se candidataram a cada área.

2 Um modelo com vários fatores e níveis de fatores

Para falar de ANOVA, vamos observar se podemos modelar o número de palavras de cor com base nos fatores década, variante e tema.

```
condiv<-read.table("http://folk.uio.no/dssantos/cursorR/CondivCompleto.txt",
header=TRUE)
condiv$decada<-factor(condiv$decada)
condiv$correl<-condiv$PalavrasCor/condiv$Tamanho
attach(condiv)
summary(lm(correl~variante*tema*decada))
summary(lm(correl~variante+tema+decada))
```

Uma forma (parcial) de visualizar este tipo de dados

```
assocplot(xtabs(correl~tema+decada))
assocplot(xtabs(correl~variante+tema))
```

3 Exemplos exploratórios não relacionados com a Literateca

Algumas palavras em holandês classificadas pela frequência de acordo com três fatores: homem/mulheres, nível de educação formal, Bélgica/Holanda.

```
library(languageR)
head(variationLijk)
chisq(variationLijk)
variationLijk.ca = corres.fnc(variationLijk)
plot(variationLijk.ca)
```