Curso de Introdução à Estatística usando o R

Lista de Exercícios 12

Exemplo 1

Considere 15 sujeitos divididos em 3 grupos. Cada grupo é designado para um mês e coletamos quantas calorias são consumidas por cada indivíduo. A pergunta: será que a quantidade de calorias consumidas é maior em alguns meses e menor em outros? Ou será que são iguais? Podemos manualmente computar um teste F:

```
maio = c(2166, 1568, 2233, 1882, 2019)
setembro = c(2279, 2075, 2131, 2009, 1793)
dezembro = c(2226, 2154, 2583, 2010, 2190)
xmedia = mean(c(maio, setembro, dezembro))
SQT = 5*((mean(maio)-xmedia)^2+(mean(setembro)-xmedia)^2+(mean(dezembro)-xmedia)^2)
SQT ## soma dos quadrados totais
```

```
## [1] 174664.1
```

```
SQE = (5-1)*var(maio)+(5-1)*var(setembro)+(5-1)*var(dezembro)
SQE # soma dos quadrados explicados
```

```
## [1] 586719.6
```

```
F.obs=(SQT/(3-1)) / (SQE/(15-3)) #computamos a estatistica F
pf(F.obs,3-1,15-3, lower.tail = FALSE) # achamos p p-valor do teste
```

```
## [1] 0.2093929
```

O p-valor não é significativo a 5%, o que indica que as médias de consumo dos dados coletados não são diferentes em diferentes meses do ano. Portanto, a diferença observada é atribuída à amostragem.

Podemos avaliar isso com um ANOVA, pela função `oneway.test()`. Precisamos antes criar um dataframe com as medidas e um fator indicando de que mês cada medida é. Felizmente - e para nossa conveniência - a função `stack()` faz exatamente isso. Basta alimentar à ela um objeto de classe `list` com nomes que ela devolve um objeto de classe `data.frame` apropriado.

```
 d = stack(list(maio = maio,
                setembro = setembro,
                dezembro = dezembro))
names(d) #retornando dois valores
```

```
## [1] "values" "ind"
```

```
oneway.test(values ~ ind, data = d, var.equal = TRUE)
```

```
##
##  One-way analysis of means
##
## data:  values and ind
## F = 1.7862, num df = 2, denom df = 12, p-value = 0.2094
```

Encontramos o mesmo p-valor, como esperado. Podemos também usar a função `aov()` para realizar um ANOVA.

```
anova = aov(values ~ ind, data = d)
summary(anova)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## ind          2 174664   87332   1.786  0.209
## Residuals   12 586720   48893
```

É reconfortante ver o mesmo resultado aparecer de novo. Fizemos tudo certo até aqui. Essas são três maneiras de computar um teste de um sentido.

Exemplo 2

Será que a administração de um aeroporto prefere uma empresa à outra e induz tempos diferentes de espera nos vôos? Vamos usar os dados da base `ewr`, contida no pacote `UsingR` e nossas ferramentas para averiguar isso.

```
library(UsingR)
data("ewr")
head(ewr)
```

```
##   Year Month  AA  CO  DL   HP  NW   TW  UA  US inorout
## 1 2000   Nov 8.6 8.3 8.6 10.4 8.1  9.1 8.4 7.6      in
## 2 2000   Oct 8.5 8.0 8.4 11.2 8.2  8.5 8.5 7.8      in
## 3 2000   Sep 8.1 8.5 8.4 10.2 8.3  8.6 8.2 7.6      in
## 4 2000   Aug 8.9 9.1 9.2 14.5 9.0 10.3 9.2 8.7      in
## 5 2000   Jul 8.3 8.9 8.2 11.5 8.8  9.1 9.2 8.2      in
## 6 2000   Jun 8.8 9.0 8.8 14.9 8.4 10.8 8.9 8.3      in
```

```
ewr.saidas = subset(ewr, subset=  inorout == "out", select = 3:10) # só saídas
saidas = stack(ewr.saidas) # usando stack()
names(saidas) = c("tempo", "empresa")  #nomeando o dataframe
# agora rodamos um modelo linear com fatores
reg = lm(tempo ~ empresa, data = saidas)
summary(reg)
```

```
##
## Call:
## lm(formula = tempo ~ empresa, data = saidas)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5913 -2.8043 -0.6109  2.0239 10.0174
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27.05652    0.72041  37.557  < 2e-16 ***
## empresaCO    3.83478    1.01881   3.764 0.000228 ***
## empresaDL   -2.05217    1.01881  -2.014 0.045503 *
## empresaHP    1.52609    1.01881   1.498 0.135949
## empresaNW   -4.06087    1.01881  -3.986 9.84e-05 ***
## empresaTW   -1.65217    1.01881  -1.622 0.106665
## empresaUA   -0.03913    1.01881  -0.038 0.969406
## empresaUS   -3.83043    1.01881  -3.760 0.000231 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.455 on 176 degrees of freedom
```

```
## Multiple R-squared:  0.3548, Adjusted R-squared:  0.3291
## F-statistic: 13.82 on 7 and 176 DF,  p-value: 3.265e-14
```

Encontramos alguns parâmetros estatisticamente significantes e a regressão como um todo é significante - como aponta a estatística F. A empresa CO tem um tempo maior de espera, a NW menor, por exemplo.

Exemplo 3

Será que mães fumantes têm bebês com menor peso? Vamos usar os dados da base `babies`, do pacote `UsingR` e ANCOVAs para responder isso.

```
data(babies)
reg = lm(wt ~ wt1 + factor(smoke), data = babies)
# explicando peso do bebê com o peso da mãe e se fuma ou não
reg2 = lm(wt ~ wt1, data = babies)
# somente pelo peso da mãe
anova(reg, reg2)
```

```
## Analysis of Variance Table
##
## Model 1: wt ~ wt1 + factor(smoke)
## Model 2: wt ~ wt1
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   1230 385256
## 2   1234 409823 -4    -24568 19.609 1.166e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

De fato, o p-valor baixíssimo apoia fortemente a hipótese de que hábitos de fumo de mães afetam o peso de seus filhos.

Questão 1

Os dados da base `ToothGrowth`, do pacote `UsingR` contém medidas de tamanho de dentes (`len`) para diferentes dosagens de vitamina C (`dose`) e métodos de entrega (`supp`). Com análise de variância, é possível dizer que o método de entrega afeta o tamanho do dente.

Questão 2

A base `grip` contém dados de performance de skiis diferentes. O tipo está na variável `grip.type`. É possível dizer que as médias de performance são diferentes para tipos diferentes de skiis?

Questão 3

Para a base `mtcars`, execute uma análise de variância unidirecional da variável `mpg` modelada por `cyl`, o número de cilindros. Use a função `factor()`, pois `cyl` é armazenada como variável numérica

Questão 4

A base `Traffic` do pacote `MASS` tem dados de morte no trânsito, o ano da coleta de dados e uma variável categórica indicado se havia um limite legal de velocidade. É possível dizer que o limite alterou o número de mortes no trânsito?