Aula 23 - Teste de Independência

Professor: Jorge L. Bazán Monitora: Patrícia Stülp

05/07/2023

1 Teste Qui-Quadrado para independência

Estatística do teste:

$$Q^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \quad \text{com} \quad e_{ij} = \frac{\text{total da linha } i \times \text{total da coluna } j}{\text{total geral}} \tag{1}$$

em que, o_{ij} e e_{ij} são as frequências observada e esperada, respectivamente, da linha i e coluna j, onde r representa o número de linhas e s o número de colunas. Se as frequências esperadas são maiores ou iguais a 5, então $Q^2 \sim X^2_{(r-1)(s-1)}$.

1.1 Exemplo: Notas de matemática e física de 528 alunos

		Matemática			
		Alta	Média	Baixa	Total
Física	Alta	56	71	12	139
	Média	47	163	38	248
	Baixa	14	42	85	141
	Total	117	276	135	528

Pergunta: Essas variáveis são independentes?

1.1.1 PASSO 1: Especificar as hipóteses H0 e Ha

Questremos testar as seguintes hipóteses estatísticas:

- H0: As variáveis são independentes.
- Ha: As variáveis não são independentes.

*Adotamos $\alpha = 5\,\%$.

Dados amostrais

Considerando as informações amostrais, temos

```
# Dados
dados = matrix(c(56,71,12,47,163,38,14,42,85),
               nrow = 3, byrow = T)
```

```
colnames(dados) = c("M Alta", "M Media", "M Baixa")
rownames(dados) = c("F Alta", "F Media", "F Baixa")

dados
```

```
##         M Alta  M Media  M Baixa
## F Alta    56       71       12
## F Media   47      163       38
## F Baixa   14       42       85
```

```
# Tabela frequencia observada
obs = matrix(0, ncol = 4, nrow = 4)
obs[1:3,1:3] = dados
obs[, 4]     = rowSums(obs)
obs[4,]      = colSums(obs)
colnames(obs) = c("M Alta", "M Media", "M Baixa", "Total")
rownames(obs) = c("F Alta", "F Media", "F Baixa", "Total")

obs
```

```
##         M Alta  M Media  M Baixa  Total
## F Alta    56       71       12     139
## F Media   47      163       38     248
## F Baixa   14       42       85     141
## Total    117      276      135     528
```

1.1.2 PASSO 2: Especificar a estatística do teste e sua distribuição, sob H0

Temos que,

```
# Frequencia esperada
e = matrix(nrow = 3, ncol = 3)
colnames(e) = c("M Alta", "M Media", "M Baixa")
rownames(e) = c("F Alta", "F Media", "F Baixa")
for(i in 1:3)
{
    for(j in 1:3)
    {
        e[i, j] = round((obs[i,4] * obs[4,j])/obs[4,4], 1)
    }
}
e
```

```
## M Alta M Media M Baixa
## F Alta 30.8 72.7 35.5
## F Media 55.0 129.6 63.4
## F Baixa 31.2 73.7 36.1
```

```
o  = dados
Q2 = round(sum( colSums((o - e)^2 / e) ), 1)
```

```
Q2
```

```
## [1] 145.5
```

em que, aproximadamente, $Q^2 \sim X^2_{(3-1)(3-1)}$.

1.1.3 PASSO 3: Fixar o nível de significância do teste (α)

O erro tipo 1, o nível de significância do teste α, neste problema é 0.05. Como a hipótese é de uma cauda, $\mathbb{P}(X^2_{(r-1)(s-1)} \geq q_c) = \alpha$, então o correspondente valor da estatística para este nível de significância é obtido da seguinte forma.

```
# erro tipo 1 = alfa
alfa = 0.05

# numero de linhas e colunas
r = 3
s = 3

# hipoteses de uma cauda
# valor de qui-quadrado para o erro tipo 1
qc = round(qchisq(1 - alfa, (r-1)*(s-1)), 2)
qc
```

```
## [1] 9.49
```

1.1.4 PASSO 4: Calcular o p-valor (ou a região crítica do teste)

Este passo pode ser feito de duas formas.

a) Encontrando a região crítica do teste
Neste caso, a região crítica é: rejeitar H0 se $Q^2 \geq q_c$. Temos que

$$RC = \{Q^2 : Q^2 \geq 9.49\} \Rightarrow Q^2 = 145.5 \in RC$$

b) Encontrando o valor p
Necessitamos encontrar a probabilidade de rejeitar H0, isto é, $\mathbb{P}(X^2_{(r-1)(s-1)} \geq q_c)$, quando de fato a hipótese nula é verdadeira. Então, temos que

```
valorp = 1 - pchisq(Q2, (r-1)*(s-1))
valorp
```

```
## [1] 0
```

1.1.5 PASSO 5: Decidir entre H0 e Ha, comparando o valor p com α (ou verificando se a estatística do teste pertence ou não à região crítica)

Este passo pode ser feito de duas formas, dependendo do passo anterior.

- Usando a região crítica: como $Q^2 = 145.5 > q_c = 9.49$ rejeitamos H0 e concluímos que há evidências de que as variáveis não são independentes.

- Usando valor p: encontramos que o valor-p $= 0 < \alpha = 0.05$ e, portanto, rejeitamos H0 e concluímos que há evidências de que as variáveis não são independentes com um nível de confiança de 95 %, ou probabilidade do erro tipo 1 de 5 %.
- Portanto, usando região crítica ou valor p, concluímos que as variáveis são dependentes

NOTA: Todos os passos anteriores podem ser resumidos utilizando os seguintes comandos.

```
chisq.test(dados)
```

```
##
## Pearson's Chi-squared test
##
## data: dados
## X-squared = 145.78, df = 4, p-value < 2.2e-16
```

2 Gráfico de probabilidade normal

Passo a passo:

- Ordenar o conjunto de dados.
- Para cada observação, calcular:

$$w_j = \frac{j - c}{n + 1 - 2c}$$

em que $c = \frac{3}{8}$ se $n \leq 10$ e $c = \frac{1}{2}$ caso contrário.

- Construir o gráfico dos pontos $(x_{(j)}, \phi^{-1}(w_{(j)}))$.
- Se os pontos do gráfico se aproximarem de uma reta, então é aceitável o modelo normal.

2.1 Exemplo:

Foram coletadas dez observações sobre o tempo (em minutos) efetivo de serviço de uma bateria de um computador pessoal:

176	191	214	220	205	192	201	190	183	185

Podemos dizer que é razoável supor que o tempo de serviço dessa bateria tem distribuição normal?
Considerando as informações amostrais, temos

```
x = c(176, 183, 185, 190, 191, 192, 201, 205, 214, 220)

n = length(x)
j = 1:n

# como n = 10, segue que
c = 3/8

w = round((j - c)/(n + 1 - 2*c), 2)
Prob = round(qnorm(w), 2)

Tab = cbind(j, x, w, Prob)
Tab
```

```
##        j  x   w    Prob
## [1,]   1 176 0.06 -1.55
## [2,]   2 183 0.16 -0.99
## [3,]   3 185 0.26 -0.64
## [4,]   4 190 0.35 -0.39
## [5,]   5 191 0.45 -0.13
## [6,]   6 192 0.55  0.13
## [7,]   7 201 0.65  0.39
## [8,]   8 205 0.74  0.64
```

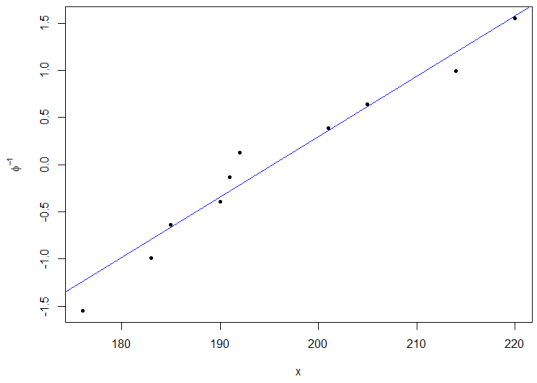
```
## [9,]   9 214 0.84  0.99
## [10,] 10 220 0.94  1.55
```

Assim, para construir o gráfico utilizamos os seguintes comandos

```
par(mar = c(5,5,2,2))
plot(x, Prob, pch = 20, ylab = expression(phi^{-1}))
(z <- line(Prob~x))

##
## Call:
## line(Prob ~ x)
##
## Coefficients:
## [1] -12.48912  0.06392

abline(coef(z), col = "blue", lwd = 1.5)
```

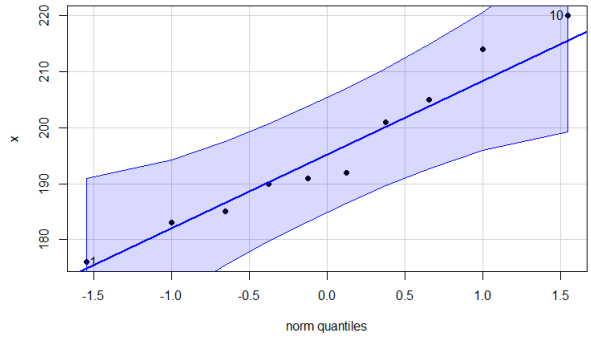


OU

```
# Caso nao tenha o pacote instalado
# install.packages("car", dep = T)

library(car)

qqPlot(x, distribution = "norm", pch = 20, cex = 1.5)
```




```
## [1] 10  1
```