Aula 22 - Teste de Aderência

Professor: Jorge L. Bazán Monitora: Patrícia Stülp

03/07/2023

1 Teste Qui-Quadrado (parâmetros conhecidos)

Estatística do teste:

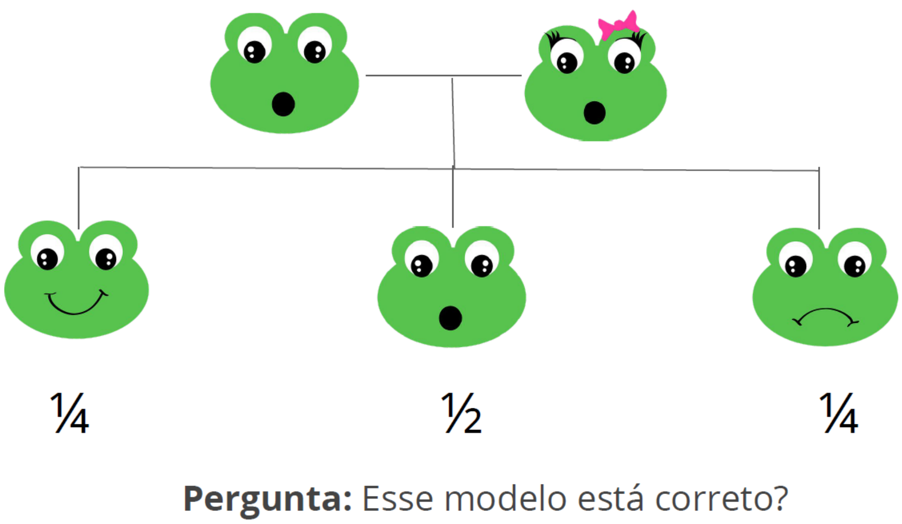
$$Q^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \tag{1}$$

em que, o_i é a frequência observada da categoria i e e_i é a frequência esperada da categoria i. Se as frequências esperadas são maiores ou iguais a 5, então $Q^2 \sim X^2_{k-1}$.

NOTAS:

- Se a suposição não ocorre para alguma categoria, devemos agrupá-la de forma conveniente.
- Os dados podem ser quantitativos contínuos ou discretos ou qualitativos.
- É necessário um número grande de observações.
- O modelo proposto não precisa estar completamente especificado (parâmetros conhecidos ou desconhecidos).

1.1 Exemplo: Modelo de Hardy-Weinberg



1.1.1 PASSO 1: Especificar as hipóteses H0 e Ha

Para testar a hipótese de pesquisa propomos as seguintes hipóteses estatísticas:

- H0: Os dados seguem o modelo proposto (parâmetros conhecidos): $p_1 = 0.25$, $p_2 = 0.5$, e $p_3 = 0.25$.
- Ha: Os dados não seguem tal modelo.

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

Dados amostrais

Considerando as informações amostrais, temos

```
# Amostra: frequencia observada
o = c(22, 52, 26)
```

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

Temos que,

```
# Parametros
p = c(0.25, 0.5, 0.25)
n = sum(o)

# Frequencia esperada
e = n*p
e

## [1] 25 50 25

Q2 = sum( (o - e)^2 / e )
Q2

## [1] 0.48
```

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

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, $\mathbb{P}(X^2_{k-1} \geq q_c) = \alpha$, 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 parámetros envolvidos
k = 3

# hipoteses de uma cauda
# valor de qui-quadrado para o erro tipo 1
qc = round(qchisq(1 - alfa, k - 1), 3)
qc

## [1] 5.991
```

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

$$RC = \{Q^2 : Q^2 \geq 5.991\} \Rightarrow Q^2 = 0.48 \notin RC$$

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

```
valorp = round(1 - pchisq(Q2, 2), 4)
valorp

## [1] 0.7866
```

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 $Q^2 = 0.48 < q_c = 5.991$ não rejeitamos H0 e concluímos que não há evidências para rejeitarmos o modelo proposto.
- Usando valor p: encontramos que o valor-p $= 0.7866 > \alpha = 0.05$ e, portanto, não rejeitamos H0 e concluímos que não há evidências para rejeitarmos o modelo proposto com um nível de confiança de 95 % ou probabilidade do erro tipo 1 de 5 %.
- Portanto, usando região crítica ou valor p, concluímos que o modelo de Hardy-Weinberg é válido para esta espécie.

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

```
chisq.test(o, p = e/100)

##
## Chi-squared test for given probabilities
##
## data: o
## X-squared = 0.48, df = 2, p-value = 0.7866
```

OU

```
chisq.test(o, p = e, rescale.p = TRUE)

##
## Chi-squared test for given probabilities
##
## data: o
## X-squared = 0.48, df = 2, p-value = 0.7866
```

2 Teste Qui-Quadrado (parâmetros desconhecidos)

O procedimento é análogo ao anterior. Há apenas uma mudança no número de graus de liberdade da estatística de teste, na qual se reduz para $k-q-1$, onde q é a quantidade de parâmetros desconhecidos. Assim, se todas as frequências esperadas forem ao menores ou iguais a 5, então

Estatística do teste:

$$Q^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \sim X^2_{k-q-1}$$

em que, o_i e e_i é a frequência observada e esperada da categoria i, respectivamente. Se as frequências esperadas são maiores ou iguais a 5, então Q^2.

2.1 Exemplo: Cinzas no carvão

Supõe-se que a porcentagem de cinzas contidas em um carvão segue a distribuição normal. Considere um nível de significância de $\alpha = 4\,\%$ para verificar esta informação.

Cinzas (em %)	9,5 ⊢ 10,5	10,5 ⊢ 11,5	11,5 ⊢ 12,5	12,5 ⊢ 13,5	13,5 ⊢ 14,5	14,5 ⊢ 15,5	15,5 ⊢ 16,5	16,5 ⊢ 17,5	17,5 ⊢ 18,5	18,5 ⊢ 19,5
Freq. observada	2	5	16	42	69	51	32	23	9	1

2.1.1 PASSO 1: Especificar as hipóteses H0 e Ha

Para testar a hipótese de pesquisa propomos as seguintes hipóteses estatísticas:

- H0: Os dados seguem a distribuição normal
- Ha: Os dados não seguem tal distribuição.

*Adotamos $\alpha = 4\,\%$.

Dados amostrais

Considerando as informações amostrais, temos

```
# Frequencia observada
o = c(2, 5, 16, 42, 69, 51, 32, 23, 9, 1)

# tamanho amostra
n = sum(o)

# Classes
cl = seq(9.5, 19.5, 1)
```

A frequência esperada é dada por

$$e_i = n\mathbb{P}(LI_i < X \leq LS_i \mid X \sim N(\overline{x}, s^2)$$

em que LI_i e LS_i são os limites inferiores e superiores da classe i, respectivamente, $\overline{x}$ é a média amostral e s^2 é a variância amostral.

```
# Ponto medio
x = c()
for(i in 1:length(o))
{
x[i] = (cl[i] + cl[i+1])/2
}

# Media amostral
xbarra = sum(x*o)/n

# variancia amostral
aux = (x-xbarra)^2
s2 = sum(o*aux)/(n-1)

# Desvio padrao
s = sqrt(s2)

# Frequencia esperada
e = c()
for(i in 1:length(o))
{
e[i] = round(n*(pnorm(cl[i+1], xbarra, s) - pnorm(cl[i], xbarra, s)), 2)
}
e

## [1] 1.54 6.52 19.25 39.65 57.02 57.27 40.16 19.67 6.72 1.60
```

Note que, temos frequências esperadas menores que 5, assim precisamos unir essas frequências com a classe mais próxima. Assim, obtemos

```
obs = c(o[1]+o[2], o[3:8], o[9]+o[10])
esp = c(e[1]+e[2], e[3:8], e[9]+e[10])

cbind(obs, esp)

##     obs esp
## [1,] 7 8.06
## [2,] 16 19.25
## [3,] 42 39.65
## [4,] 69 57.02
## [5,] 51 57.27
## [6,] 32 40.16
## [7,] 23 19.67
## [8,] 10 8.32
```

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

Temos que,

```
# Estatistica do teste
Q2 = round(sum( (obs - esp)^2 / esp ), 4)
Q2

## [1] 6.5918
```

em que, aproximadamente $Q^2 \sim X^2_{8-2-1}$.

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

O erro tipo 1, o nível de significância do teste α, neste problema é 0.04, como a hipótese é de uma cauda, $\mathbb{P}(X^2_{k-1} \geq q_c) = \alpha$, então o correspondente valor da estatística para este nível de significância é obtido da seguinte forma.

```
# erro tipo 1 = alfa
alfa = 0.04

# numero de parámetros envolvidos
k = 8
q = 2

# hipoteses de uma cauda
# valor de qui-quadrado para o erro tipo 1
qc = round(qchisq(1 - alfa, k - q - 1), 2)
qc

## [1] 11.64
```

2.1.4 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

$$RC = \{Q^2 : Q^2 \geq 11.64\} \Rightarrow Q^2 = 6.5918 \notin RC$$

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

```
valorp = round(1 - pchisq(Q2, k - q - 1), 4)
valorp

## [1] 0.2528
```

2.1.5 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 $Q^2 = 6.5918 < q_c = 11.64$ não rejeitamos H0 e concluímos que é aceitável assumir que a distribuição dos dados é normal com média igual a sua média amostral e desvio padrão igual ao amostral.
- Usando valor p: encontramos que o valor-p $= 0.2528 > \alpha = 0.04$ e, portanto, não rejeitamos H0 e concluimos que não há evidências para rejeitarmos que os dados tem distribuição normal com um nivel de confiança de 96 % ou probabilidade do erro tipo 1 de 4 %.
- Portanto, usando região crítica ou valor p, concluimos que a distribuição dos dados é normal com média igual a sua média amostral e desvio padrão igual ao amostral.

3 Teste de Kolmogorov-Smirnov

A ideia geral deste teste é comparar a distribuição acumulada empírica com a distribuição acumulada teórica.

Estatística do teste:

$$D_n = max_i\{\mid F(x_{(i)}) - F_n(x_{(i)}) \mid, \mid F(x_{(i)}) - F_n(x_{(i-1)} \mid\}$$

em que $F_n(x) = \frac{1}{n}$ número $\{x_i : x_i \leq x\}$, $F(x)$ é a função distribuição acumulada do modelo proposto e n é o tamanho amostral.

NOTAS:

- Conhecido como K-S teste.
- Os dados devem ser quantitativos contínuos.
- Tamanhos amostrais pequenos são permitidos.
- O modelo proposto precisa estar completamente especificado.

3.1 Exemplo:

A análise de fatores estruturais mais leves se mostrou importante para alcançar o nível de segurança desejado em aeronaves. Neste estudo foram analisados a movimentação vertical da asa da aeronave para diferentes localizações, em relação ao centro da asa, de um peso extra. Os dados da movimentação em decímetro são dados a seguir.

2	5	17	21	25	33	34	37	40	56

Os especialistas supuseram que os dados seguem uma distribuição normal com média 30 e desvio padrão 16. Essa suposição é válida? Considere o nível de significância de 5 %.

As hipóteses a serem testadas são:

- H0: $X \sim N(30, 162)$
- Ha: X segue outra distribuição.

Considerando as informações amostrais, temos

```
mu    = 30
sigma = 16
x     = c(2,5,17,21,25,33,34,37,40,56)
n     = length(x)

F10   = 1:10/n
F10_1 = (1:10 - 1) / n

# Função acumulada da normal(30,16^2)
Fx = round(pnorm(x, mu, sigma), 3)
```

```
aux1 = abs(Fx-F10)
aux2 = abs(Fx-F10_1)

# Tabela com os valores
Tab = cbind(x, F10, F10_1, aux1, aux2)
colnames(Tab) = c("x", "F10", "F10_1", "|Fx - F10|", "|Fx - F10_1|")
Tab
```

```
##        x  F10  F10_1 |Fx - F10| |Fx - F10_1|
## [1,]   2  0.1   0.0    0.060       0.040
## [2,]   5  0.2   0.1    0.141       0.041
## [3,]  17  0.3   0.2    0.092       0.008
## [4,]  21  0.4   0.3    0.113       0.013
## [5,]  25  0.5   0.4    0.123       0.023
## [6,]  33  0.6   0.5    0.026       0.074
## [7,]  34  0.7   0.6    0.101       0.001
## [8,]  37  0.8   0.7    0.131       0.031
## [9,]  40  0.9   0.8    0.166       0.066
## [10,] 56  1.0   0.9    0.052       0.048
```

Logo, a estatística do teste é dada por

```
# Estatistica do teste
Dn = max(aux1, aux2)
Dn
```

```
## [1] 0.166
```

Neste caso, a região crítica é: rejeitar H0 se $D_n > D_{n,\alpha}$, em que $D_{n,\alpha} = 0.410$ (valor crítico do K-S teste para $n = 10$ e $\alpha = 0.05$). Assim, temos que

$$RC = \{D_n : D_n > 0.410\} \Rightarrow D_n = 0.166 \notin RC$$

Portanto, não rejeitamos H0 e concluímos que é aceitável assumir que a distribuição dos dados é normal com média igual a 30 e desvio padrão igual a 16, com um nivel de confiança de 96 % ou probabilidade do erro tipo 1 de 4 %.

NOTAS: O teste de Kolmogorov-Smirnov pode ser realizado no R utilizando os seguintes comandos

```
ks.test(x, "pnorm", 30, 16)
```

```
##
## One-sample Kolmogorov-Smirnov test
##
## data: x
## D = 0.16599, p-value = 0.9052
## alternative hypothesis: two-sided
```

Valores críticos do teste K-S

n	nível de significância			
	0,200	0,100	0,050	0,010
1	0,900	0,950	0,975	0,995
2	0,684	0,776	0,842	0,929
3	0,565	0,597	0,708	0,828
4	0,494	0,525	0,624	0,733
5	0,446	0,474	0,565	0,669
6	0,410	0,436	0,521	0,618
7	0,381	0,405	0,486	0,577
8	0,358	0,381	0,457	0,543
9	0,339	0,360	0,432	0,514
10	0,322	0,342	0,410	0,490
11	0,307	0,326	0,391	0,468
12	0,295	0,313	0,375	0,450
13	0,284	0,302	0,361	0,433
14	0,274	0,292	0,349	0,418
15	0,266	0,283	0,338	0,404

n	nível de significância			
	0,200	0,100	0,050	0,010
16	0,258	0,295	0,328	0,392
17	0,250	0,286	0,318	0,381
18	0,244	0,278	0,309	0,371
19	0,237	0,272	0,301	0,363
20	0,231	0,264	0,294	0,356
25	0,21	0,24	0,27	0,32
30	0,19	0,22	0,24	0,29
35	0,18	0,21	0,23	0,27
> 35	$\frac{1,07}{\sqrt{N}}$	$\frac{1,22}{\sqrt{N}}$	$\frac{1,36}{\sqrt{N}}$	$\frac{1,63}{\sqrt{N}}$