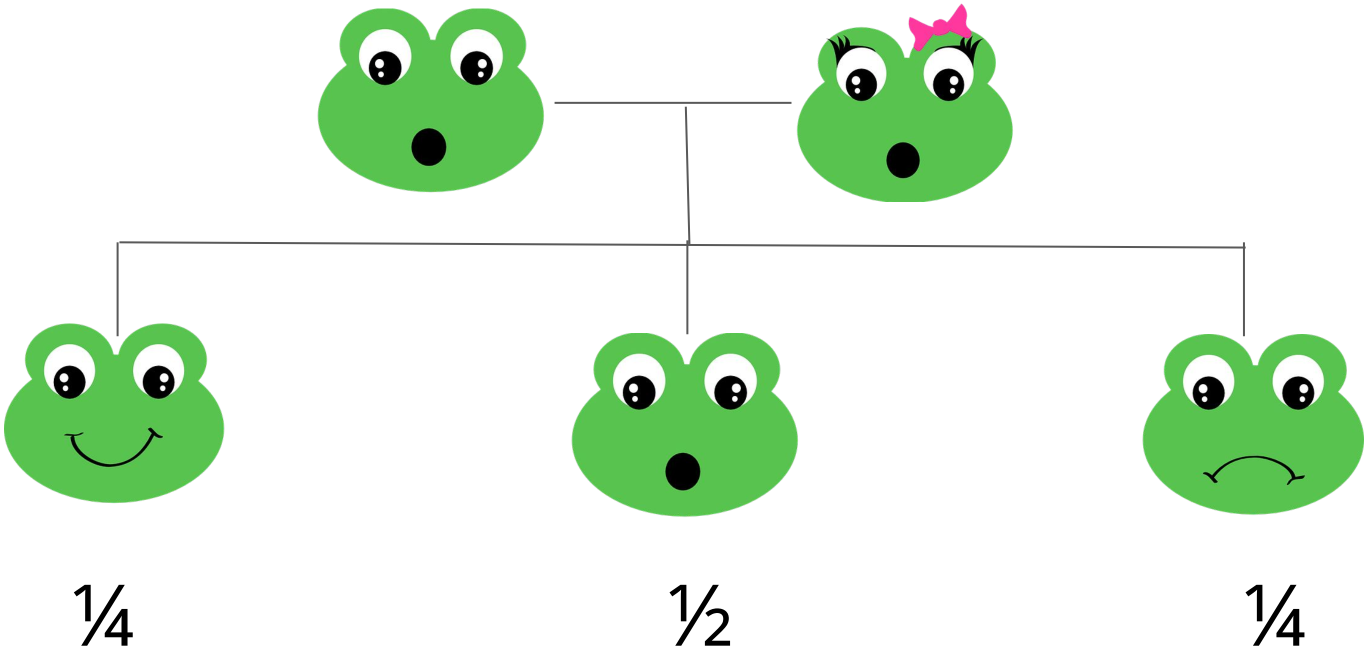
TESTES DE ADERÊNCIA

6.2

MOTIVAÇÃO

Modelo genético de Hardy-Weinberg supõe que dois seres heterozigotos quando se cruzam vão ter descendentes heterozigotos, homozigotos dominantes e homozigotos recessivos em certas proporções.

Modelo de Hardy-Weinberg



Pergunta: Esse modelo está correto?

Experimento

Genótipo	Frequência observada
AA	22
Aa	52
aa	26
Total	100

Pergunta: O modelo de Hardy-Weinberg é válido para esta espécie?

IDEIA

Se o modelo for adequado, as frequências observadas e esperadas devem ser próximas. Assim, se as diferenças forem muito grandes, admitimos que o modelo não é razoável.

Genótipo	Frequência observada	Frequência esperada
AA	22	25
Aa	52	50
aa	26	25
Total	100	100

TESTE DE ADERÊNCIA

Serve para testar a adequabilidade do modelo.

Vamos ver dois testes: teste qui-quadrado para modelos com parâmetros conhecidos e desconhecidos e o teste de Kolmogorov-Smirnov

TESTE QUI-QUADRADO (parâmetros conhecidos)

H_0: Os dados seguem o modelo proposto (parâmetros conhecidos)
H_1: Os dados não seguem tal modelo.

Estatística de teste

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

- o_i é a frequência observada da categoria i
- e_i é a frequência esperada da categoria i

Resultado: Se as frequências esperadas forem ao menos iguais a 5, então

$$Q^2 \sim \chi^2_{k-1}$$

Região crítica

$$RC = \{Q^2 : Q^2 \geq q_c\},$$

em que

$$P\left(\chi^2_{k-1} \geq q_c\right) = \alpha$$

α é o nível de significância do teste.

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).

Exemplo

i	Genótipo	Frequência observada
1	AA	22
2	Aa	52
3	aa	26
k = 3	Total	100

H_0: Os dados seguem o modelo proposto: $p_1 = ¼$, $p_2 = ½$, $p_3 = ¼$
H_1: Os dados não seguem tal modelo.

Adotamos $\boldsymbol{\alpha} = 5\%$

Exemplo

i	Genótipo	Frequência observada	Frequência esperada	$\frac{(o_i - e_i)^2}{e_i}$
1	AA	22	25	0,36
2	Aa	52	50	0,08
3	aa	26	25	0,04
k = 3	Total	100	100	0,48

$$P\left(\chi^2_{3-1} \geq q_c\right) = 0,05 \Rightarrow q_c = 5,991$$

$$\{Q^2 : Q^2 \geq 5,991\} \Rightarrow Q^2_{\text{obs}} = 0,48 \notin RC.$$

Conclusão: Não rejeitamos H_0, ou seja, não há evidências para rejeitarmos o modelo proposto.

Exemplo (continuação)

No R

```
> o = c(22,52, 26)
> e = c(25,50,25)
>
> chisq.test(o,p=e/100)

        Chi-squared test for given probabilities

data:  o
X-squared = 0.48, df = 2, p-value = 0.7866
```

Conclusão: Não rejeitamos H_0, ou seja, não há evidências para rejeitarmos o modelo proposto.

Exemplo (continuação)

No R

```
> o = c(22,52, 26)
> e = c(25,50,25)
>
> chisq.test(o,p=e, rescale.p = T)

        Chi-squared test for given probabilities

data:  o
X-squared = 0.48, df = 2, p-value = 0.7866
```

Conclusão: Não rejeitamos H_0, ou seja, não há evidências para rejeitarmos o modelo proposto.

TESTE QUI-QUADRADO (parâmetros desconhecidos)

Procedimento para parâmetros não conhecidos

O procedimento é análogo ao anterior. Há mudança no número de g.l. da estatística de teste.

Ideia

Utilizaremos a amostra para obter as estimativas dos parâmetros desconhecidos.

Consequência

O número de g.l. da estatística de teste reduz para k - q - 1, em que q é quantidade de parâmetros desconhecidos.

TESTE QUI-QUADRADO (parâmetros desconhecidos)

H_0: Os dados seguem o modelo proposto (q parâmetros desconhecidos)
H_1: Os dados não seguem tal modelo.

Estatística de teste

- o_i é a frequência observada da categoria i
- e_i é a frequência esperada da categoria i

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

Resultado: Se todas as frequências esperadas forem ao menos iguais a 5, então

$$Q^2 \sim \chi^2_{k-q-1}$$

Exemplo

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

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

Solução

Cinzas (em %)	$9,5 \vdash 10,5$	$10,5 \vdash 11,5$	$11,5 \vdash 12,5$	$12,5 \vdash 13,5$	$13,5 \vdash 14,5$	$14,5 \vdash 15,5$	$15,5 \vdash 16,5$	$16,5 \vdash 17,5$	$17,5 \vdash 18,5$	$18,5 \vdash 19,5$
Freq. observada	2	5	16	42	69	51	32	23	9	1
Freq. esperada	1,80	7,07	19,74	39,22	55,50	55,92	40,11	20,49	9,38	1,93

$$e_1 = nP(10,5 < X \leq 9,5 | X \sim \mathcal{N}(\bar{x}, s^2))$$

Solução

Cinzas (em %)	$9,5 \vdash 11,5$	$11,5 \vdash 12,5$	$12,5 \vdash 13,5$	$13,5 \vdash 14,5$	$14,5 \vdash 15,5$	$15,5 \vdash 16,5$	$16,5 \vdash 17,5$	$17,5 \vdash 19,5$
Freq. observada	7	16	42	69	51	32	23	10
Freq. esperada	8,88	19,74	39,22	55,50	55,92	40,11	20,49	9,38

Solução

Cinzas (em %)	$9,5 \vdash 11,5$	$11,5 \vdash 12,5$	$12,5 \vdash 13,5$	$13,5 \vdash 14,5$	$14,5 \vdash 15,5$	$15,5 \vdash 16,5$	$16,5 \vdash 17,5$	$17,5 \vdash 19,5$
Freq. observada	7	16	42	69	51	32	23	10
Freq. esperada	8,88	19,74	39,22	55,50	55,92	40,11	20,49	9,38
$(o_i - e_i)^2/e_i$	0,39	0,71	0,20	3,28	0,43	1,64	0,31	0,04

$$e_1 = nP(11,5 < X \leq 9,5 | X \sim \mathcal{N}(\bar{x}, s^2))$$

$$Q^2_{\text{obs}} = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} = 7,01$$

Solução (continuação)

$$Q^2 \sim \chi^2_{k-q-1=8-2-1=5}$$

$$RC = \{Q^2 > q_c\} = \{Q^2 > 11,64\}$$

$$Q^2_{\text{obs}} = 7,01 \notin RC \Rightarrow \text{não rejeitamos } H_0$$

Conclusão: É aceitável assumir que a distribuição dos dados é normal com média igual a sua média amostral e desvio padrão igual ao amostral.

TESTE DE KOLMOGOROV-SMIRNOV

H_0: Os dados seguem o modelo proposto
H_1: Os dados não seguem tal modelo.

Ideia: Comparar a distribuição acumulada empírica com a distribuição acumulada teórica (proposta).

TESTE DE KOLMOGOROV-SMIRNOV

H_0: Os dados seguem o modelo proposto
H_1: Os dados não seguem tal modelo.

Estatística de teste

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

- $F_n(x) = \frac{1}{n}\text{número}\{x_i : x_i \leq x\}$
- $F(x)$ é a fda do modelo proposto
- n é o tamanho amostral

NOTAS

- A estatística de teste precisa ser comparada a um valor estipulado que depende do tamanho da amostra e do nível de significância, $D_{n,\alpha}$, e a região crítica é dada por

$$RC = \{D_n : D_n > D_{n,\alpha}\}$$

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

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%.

Solução

Vamos aplicar o teste Kolmogorov-Smirnov (teste K-S), pois temos apenas 10 observações.

As hipóteses a serem testadas são:

$$H_0 : X \sim \mathcal{N}(30, 16^2) \text{ versus } H_1 : X \text{ segue outra distribuição}$$

Solução (continuação)

$x_{(i)}$	$F_{10}(x_i)$	$F_{10}(x_{i-1})$	z_i	$F(x_i) = F(z_i)$	$\|F(x_i)-F_{10}(x_i)\|$	$\|F(x_i)-F_{10}(x_{i-1})\|$
2	0,1	0	-1,7500	0,040	0,060	0,040
5	0,2	0,1	-1,5625	0,059	0,141	0,041
17	0,3	0,2	-0,8125	0,208	0,092	0,008
21	0,4	0,3	-0,5625	0,287	0,113	0,013
25	0,5	0,4	-0,3125	0,377	0,123	0,023
33	0,6	0,5	0,1875	0,574	0,026	0,074
34	0,7	0,6	0,2500	0,599	0,101	0,001
37	0,8	0,7	0,4375	0,669	0,131	0,031
40	0,9	0,8	0,6250	0,734	0,166	0,066
56	1,0	0,9	1,6250	0,948	0,052	0,048

Solução (continuação)

$x_{(i)}$	$F_{10}(x_i)$	$F_{10}(x_{i-1})$	z_i	$F(x_i) = F(z_i)$	$\|F(x_i)-F_{10}(x_i)\|$	$\|F(x_i)-F_{10}(x_{i-1})\|$
2	0,1	0	-1,7500	0,040	0,060	0,040
5	0,2	0,1	-1,5625	0,059	0,141	0,041
17	0,3	0,2	-0,8125	0,208	0,092	0,008
21	0,4	0,3	-0,5625	0,287	0,113	0,013
25	0,5	0,4	-0,3125	0,377	0,123	0,023
33	0,6	0,5	0,1875	0,574	0,026	0,074
34	0,7	0,6	0,2500	0,599	0,101	0,001
37	0,8	0,7	0,4375	0,669	0,131	0,031
40	0,9	0,8	0,6250	0,734	0,166	0,066
56	1,0	0,9	1,6250	0,948	0,052	0,048

Solução

$D_{10} = 0,166 < 0,410$ (valor crítico do teste K-S para n = 10 e $\alpha = 0,05$). Assim $D_{10} \notin RC$ e, portanto, não rejeitamos H_0.

```
>
> 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	$\frac{1,07}{\sqrt{N}}$	$\frac{1,22}{\sqrt{N}}$	$\frac{1,36}{\sqrt{N}}$	$\frac{1,63}{\sqrt{N}}$