

- Testes Qui-quadrado - Aderência e Independência

1. Testes de Aderência

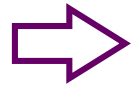
Objetivo: Testar a adequação de um modelo probabilístico a um conjunto de dados observados.

Exemplo 1: Segundo Mendel (geneticista famoso), os resultados dos cruzamentos de ervilhas amarelas redondas com ervilhas verdes enrugadas ocorrem na proporção de 9:3:3:1, ou seja, seguem uma distribuição de probabilidades dada por:

Resultado	Amarela redonda	Amarela enrugada	Verde redonda	Verde enrugada
Probabilidade	9/16	3/16	3/16	1/16

Uma amostra de **556 ervilhas** resultantes de cruzamentos de ervilhas amarelas redondas com ervilhas verdes enrugadas foi classificada da seguinte forma:

Resultado	Amarela redonda	Amarela enrugada	Verde redonda	Verde enrugada
Freq. observada	315	101	108	32



Há evidências de que os resultados desse experimento estão de acordo com a distribuição de probabilidades proposta por Mendel?

4 categorias para os resultados dos cruzamentos:

Amarelas redondas (AR), Amarelas enrugadas (AE), Verdes redondas (VR), Verdes enrugadas (VE).

Segundo Mendel, a probabilidade de cada categoria é dada por:

Probabilidades:	<i>AR</i>	<i>AE</i>	<i>VR</i>	<i>VE</i>
(de Mendel)	9/16	3/16	3/16	1/16

No experimento, 556 ervilhas foram classificadas segundo o tipo de resultado, fornecendo a tabela a seguir:

Tipo de resultado	Frequência observada
<i>AR</i>	315
<i>AE</i>	101
<i>VR</i>	108
<i>VE</i>	33
Total	556

Objetivo: Verificar se o modelo probabilístico proposto é adequado aos resultados do experimento.

Se o modelo probabilístico for adequado, a **frequência esperada** de ervilhas do tipo **AR**, dentre as 556 observadas, pode ser calculada por:

$$556 \times P(AR) = 556 \times \mathbf{9/16} = 312,75$$

Da mesma forma, temos para o tipo **AE**,

$$556 \times P(AE) = 556 \times \mathbf{3/16} = 104,25$$

Para o tipo **VR** temos

$$556 \times P(VR) = 556 \times \mathbf{3/16} = 104,25$$

E para o tipo **VE**,

$$556 \times P(VE) = 556 \times \mathbf{1/16} = 34,75$$

Podemos expandir a tabela de frequências dada anteriormente:

Tipo de resultado	Frequência observada	Frequência esperada (por Mendel)
<i>AR</i>	315	312,75
<i>AE</i>	101	104,25
<i>VR</i>	108	104,25
<i>VE</i>	32	34,75
Total	556	556

→ **Pergunta:** Podemos afirmar que os valores observados estão suficientemente próximos dos valores esperados, de tal forma que o modelo probabilístico proposto por Mendel é adequado aos resultados desse experimento?

Testes de Aderência – Metodologia

Considere uma tabela de frequências, com $k \geq 2$ categorias de resultados:

Categorias	Frequência observada
1	O_1
2	O_2
3	O_3
\vdots	\vdots
k	O_k
Total	n

em que O_i é o total de indivíduos **observados** na categoria i , $i = 1, \dots, k$.

Seja p_i a probabilidade associada à categoria i , $i = 1, \dots, k$.

O objetivo do teste de aderência é testar as hipóteses

$$H_0: p_1 = p_{o1}, \dots, p_k = p_{ok}$$

H_1 : existe pelo menos uma diferença

sendo p_{oi} a probabilidade especificada para a categoria i , $i = 1, \dots, k$, **fixada através do modelo probabilístico de interesse.**

Se E_i é o total de indivíduos **esperados** na categoria i , quando a hipótese H_0 é verdadeira, então:

$$E_i = n \times p_{oi}, i = 1, \dots, k$$

Expandindo a tabela de frequências original, temos:

Categorias	Frequência observada	Frequência esperada, sob H_0
1	O_1	E_1
2	O_2	E_2
3	O_3	E_3
\vdots	\vdots	\vdots
k	O_k	E_k
Total	n	n

Quantificação da distância entre as colunas de frequências:

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

← Estatística do teste de aderência

Supondo H_0 verdadeira,

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \sim \chi_q^2, \text{ aproximadamente,}$$

sendo que $q = k - 1$ representa o número de graus de liberdade.

→ Em outras palavras, se H_0 é verdadeira, a v.a. χ^2 tem distribuição aproximada qui-quadrado com q graus de liberdade.

IMPORTANTE.: Esta resultado é válido para **n grande** e para

$$E_i \geq 5, i = 1, \dots, k.$$

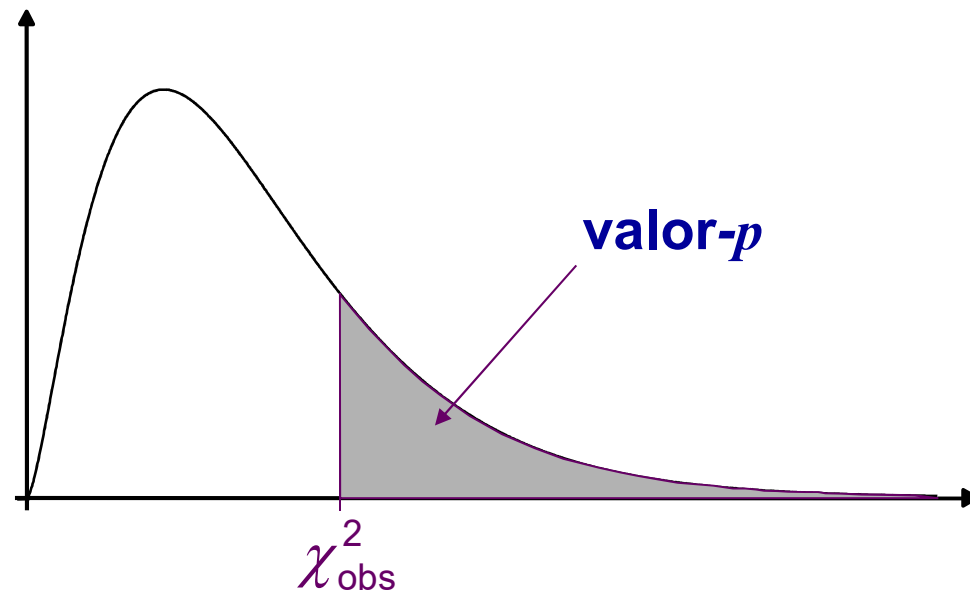
Regra de decisão:

Pode ser baseada no **nível descritivo** ou **valor- p** , neste caso

$$\text{valor-}p = P(\chi_q^2 \geq \chi_{obs}^2),$$

em que χ_{obs}^2 é o valor calculado, a partir dos dados, usando a expressão apresentada para χ^2 .

Graficamente:



Se, para α fixado, obtemos **valor- $p \leq \alpha$** , **rejeitamos a hipótese H_0** .

Exemplo (continuação): Cruzamentos de ervilhas

Hipóteses:

H_0 : O modelo probabilístico proposto por Mendel é adequado.

H_1 : O modelo proposto por Mendel não é adequado.

De forma equivalente, podemos escrever:

H_0 : $P(AR) = 9/16$; $P(AE) = 3/16$; $P(VR) = 3/16$; $P(VE) = 1/16$.

H_1 : ao menos uma das igualdades não se verifica.

A tabela seguinte apresenta os valores observados e esperados (calculados anteriormente).

Resultado	O_i	E_i
<i>AR</i>	315	312,75
<i>AE</i>	101	104,25
<i>VR</i>	108	104,25
<i>VE</i>	32	34,75
Total	556	556

Cálculo do valor da estatística do teste ($k = 4$):

$$\chi^2_{obs} = \sum_1^4 \frac{(O_i - E_i)^2}{E_i} = \frac{(315 - 312,75)^2}{312,75} + \frac{(101 - 104,25)^2}{104,25} + \frac{(108 - 104,25)^2}{104,25} + \frac{(32 - 34,75)^2}{34,75} = 0,016 + 0,101 + 0,135 + 0,218 = 0,470.$$

Usando a distribuição de qui-quadrado com $q = k-1 = 3$ graus de liberdade, o nível descritivo é dado por: **valor- p** = $P(\chi^2_3 \geq 0,470) \cong 0,925$.

Conclusão: Para $\alpha=0,05$, como **valor- p** = **0,925 > 0,05**, não há evidências para rejeitarmos a hipótese H_0 , isto é, ao nível de significância de 5%, concluimos não há evidências para contestar o modelo de probabilidades de Mendel.

O cálculo do *nível descritivo* (**valor- p**) pode ser feito no *Rcmdr*, via menu, através do seguinte caminho:

Distribuições → Distribuições contínuas → Distribuição Qui-Quadrado → Probabilidades da Qui-Quadrado → Cauda Superior

Inserindo o **valor 0,470** e o número de **graus de liberdade igual a 3**, o **valor- p** será igual a **0,925431**.

Exemplo 2: Deseja-se verificar se o número de acidentes em uma estrada muda conforme o dia da semana. O número de acidentes observado para cada dia de uma semana escolhida aleatoriamente foram:

Dia da semana	No. de acidentes
Seg	20
Ter	10
Qua	10
Qui	15
Sex	30
Sab	20
Dom	35

⇒ O que pode ser dito?

Hipóteses a serem testadas:

H_0 : O número de acidentes não muda conforme o dia da semana;

H_1 : Pelo menos um dos dias tem número diferente dos demais.

Se p_i representa a probabilidade de ocorrência de acidentes no i -ésimo dia da semana, temos as hipóteses estatísticas,

$H_0: p_i = 1/7$ para todo $i = 1, \dots, 7$

$H_1: p_i \neq 1/7$ para pelo menos um valor de i .

Total de acidentes na semana: $n = 140$.

Logo, se H_0 for verdadeira,

$$E_i = 140 \times 1/7 = 20, i = 1, \dots, 7,$$

ou seja, esperamos 20 acidentes por dia.

Dia da semana	Nº. de acidentes observados (O_i)	Nº. esperado de acidentes (E_i)
Seg	20	20
Ter	10	20
Qua	10	20
Qui	15	20
Sex	30	20
Sab	20	20
Dom	35	20

Cálculo da estatística de qui-quadrado:

$$\begin{aligned}
 \chi_{obs}^2 &= \sum_1^7 \frac{(O_i - E_i)^2}{E_i} = \frac{(20-20)^2}{20} + \frac{(10-20)^2}{20} + \frac{(10-20)^2}{20} + \frac{(15-20)^2}{20} + \\
 &\quad \frac{(30-20)^2}{20} + \frac{(20-20)^2}{20} + \frac{(35-20)^2}{20} \\
 &= 0 + 5 + 5 + 1,25 + 5 + 0 + 11,25 = 27,50
 \end{aligned}$$

Neste caso, temos $\chi^2 \sim \chi_6^2$, aproximadamente.

O nível descritivo é dado por: $\text{valor-}p = P(\chi_6^2 \geq 27,50) \cong 0,00012$,
que pode ser obtido no *Rcmdr* pelo caminho (via menu):

Distribuições → Distribuições contínuas → Distribuição Qui-Quadrado → Probabilidades da Qui-Quadrado → Cauda Superior

(inserindo o valor 27,50 e o número de graus de liberdade igual a 6).

Conclusão: Para $\alpha = 0,05$, temos que **$\text{valor-}p = 0,0001 < \alpha$** .

Assim, há evidências para **rejeitarmos H_0** , ou seja, concluímos, ao nível de significância de 5%, de que o número de acidentes se altera ao longo das semanas.

2. Testes de Independência

Objetivo: Verificar se há dependência entre duas variáveis medidas nas mesmas unidades experimentais.

Exemplo 3: Uma grande empresa de comunicação no Brasil fez um levantamento com 1300 usuários de seus recursos midiáticos, para verificar se a preferência por um determinado canal de informação para se inteirar de notícias é independente do nível de instrução do indivíduo. Os resultados obtidos foram:

	Tipo de mídia				
Grau de instrução	Internet	TV	Rede Social	Outras	Total
Fundamental	10	27	5	8	50
Médio	90	73	125	162	450
Superior	200	130	220	250	800
Total	300	230	350	420	1300

Vamos calcular *proporções segundo os totais das colunas* (poderiam também ser calculadas pelos totais das linhas). Temos a seguinte tabela:

	Tipo de mídia				
Grau de instrução	Internet	TV	Rede Social	Outras	Total
Fundamental	3,33%	11,74%	1,90%	1,43%	3,85%
Médio	30,00%	31,74%	38,57%	35,71%	34,62%
Superior	66,67%	56,52%	59,52%	62,86%	61,54%
Total	100,00%	100,00%	100,00%	100,00%	100,00%

⇒ O que representam as porcentagens na colunas?

Distribuição de grau de instrução por tipo de mídia.

⇒ Independentemente da preferência por um tipo de mídia:

3,85% dos usuários têm ensino fundamental,

34,62% têm ensino médio e

61,54% têm ensino superior.

Sob independência entre grau de instrução e preferência por um tipo de mídia, o número esperado de usuários que têm:

- **Fundam. e preferem Internet** é igual a $300 \times 0,0385 = \mathbf{11,54}$ ($=300 \times 50/1300$),
- **Médio e preferem Internet** é $300 \times 0,3462 = \mathbf{103,85}$ ($=300 \times 450/1300$),
- **Superior e preferem Internet** é $300 \times 0,6154 = \mathbf{184,62}$ ($=300 \times 800/1300$).

Grau de instrução	Tipo de mídia				Total
	Internet	TV	Rede Social	Outras	
Fundamental	10 11,54 (3,85%)	27 8,85 (3,85%)	5 13,46 (3,85%)	8 16,15 (3,85%)	50 (3,85%)
Médio	90 103,85 (34,62)%	73 79,62 (34,62%)	125 121,15 (34,62%)	162 145,38 (34,62%)	450 (34,62%)
Superior	200 184,62 (61,54%)	130 141,54 (61,54%)	220 215,38 (61,54%)	250 258,46 (61,54%)	800 (61,54%)
Total	300	230	350	420	1300

As diferenças entre os valores observados e os esperados não são muito pequenas. Preferência por um tipo de mídia e grau de instrução parecem não ser *independentes*.

Testes de Independência – Metodologia

Em geral, os dados referem-se a mensurações de duas características (A e B) feitas em n unidades experimentais, que são apresentadas conforme a seguinte tabela:

$A \setminus B$	B_1	B_2	...	B_s	Total
A_1	O_{11}	O_{12}	...	O_{1s}	$O_{1.}$
A_2	O_{21}	O_{22}	...	O_{2s}	$O_{2.}$
...
A_r	O_{r1}	O_{r2}	...	O_{rs}	$O_{r.}$
Total	$O_{.1}$	$O_{.2}$...	$O_{.s}$	n

Hipóteses a serem testadas – **Teste de independência:**

H_0 : A e B são variáveis independentes

H_1 : As variáveis A e B não são independentes

→ Quantas observações devemos esperar em cada casela, se A e B forem independentes?

$$E_{ij} = \frac{O_{i.} \times O_{.j}}{n}$$

Distância entre os valores observados e os valores esperados sob a suposição de independência:

$$\chi^2 = \sum_{i=1}^s \sum_{j=1}^r \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Supondo H_0 verdadeira,

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi_q^2$$

aproximadamente, sendo $q = (r - 1) \times (s - 1)$ o número de **graus de liberdade**.

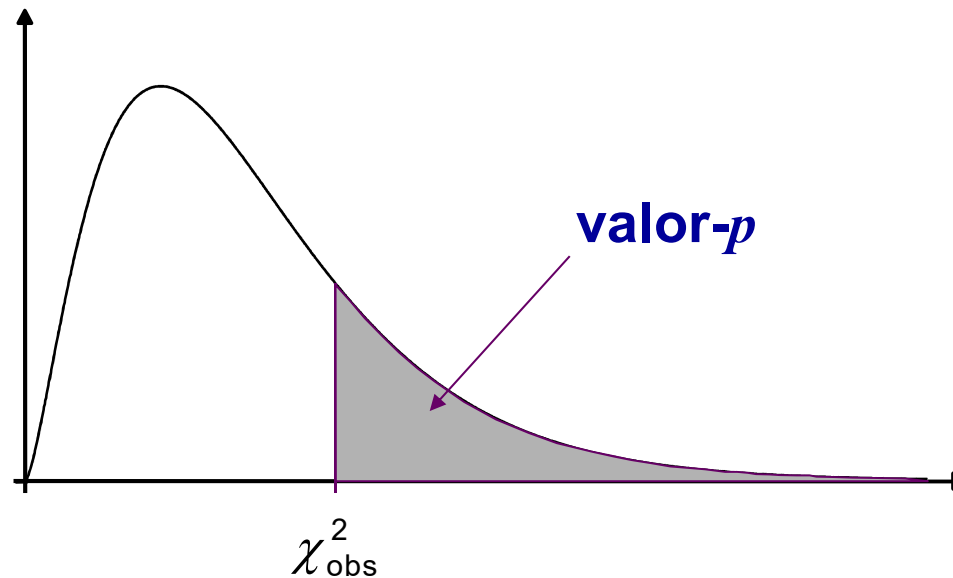
Regra de decisão:

Pode ser baseada no **valor- p** (nível descritivo), neste caso

$$\text{valor-}p = P(\chi_q^2 \geq \chi_{obs}^2),$$

em que χ_{obs}^2 é o valor calculado, a partir dos dados, usando a expressão apresentada para χ^2 .

Graficamente:



Se, para α fixado, obtemos **valor- $p \leq \alpha$** , rejeitamos a hipótese **H_0 de independência.**

Exemplo 3 (continuação): Estudo da dependência entre preferência por um tipo de mídia e grau de instrução. Foram selecionados ao acaso, e entrevistados, 1300 usuários.

Hipóteses: H_0 : As variáveis preferência por um tipo de mídia e grau de instrução são independentes.

H_1 : Existe dependência entre as variáveis.

Tabela de valores observados

	Tipo de mídia				
Grau de instrução	Internet	TV	Rede social	Outras	Total
Fundamental	10	27	5	8	50
Médio	90	73	125	162	450
Superior	200	130	220	250	800
Total	300	230	350	420	1300

→ Exemplo do cálculo dos valores esperados sob H_0 (independência):

- Número esperado de usuários que têm fundamental e preferem internet:

$$E_{11} = \frac{300 \times 50}{1300} = 11,54$$

Tabela de valores observados e esperados (entre parênteses)

	Tipo de Mídia				
Grau de instrução	Internet	TV	Rede social	Outras	Total
Fundamental	10 (11,54)	27 (8,85)	5 (13,46)	8 (16,15)	50
Médio	90 (103,85)	73 (79,62)	125 (121,15)	162 (145,38)	450
Superior	200 (184,62)	130 (141,54)	220 (215,38)	250 (258,46)	800
Total	300	230	350	420	1300

Médio e prefere TV:

$$E_{22} = \frac{230 \times 450}{1300} = 79,62$$

Superior e prefere outras mídias:

$$E_{34} = \frac{420 \times 800}{1300} = 258,46$$

Lembre-se:

$$E_{ij} = \frac{O_{i.} \times O_{.j}}{n_{..}}$$

Cálculo da estatística de qui-quadrado:

	Tipo de Mídia				
Grau de instrução	Internet	TV	Rede social	Outras	Total
Fundamental	10 (11,54)	27 (8,85)	5 (13,46)	8 (16,15)	50
Médio	90 (103,85)	73 (79,62)	125 (121,15)	162 (145,38)	450
Superior	200 (184,62)	130 (141,54)	220 (215,38)	250 (258,46)	800
Total	300	230	350	420	1300

$$\begin{aligned}
 \chi^2_{obs} &= \frac{(10-11,54)^2}{11,54} + \frac{(27-8,85)^2}{8,85} + \frac{(5-13,46)^2}{13,46} + \frac{(8-16,15)^2}{16,15} \\
 &+ \frac{(90-103,85)^2}{103,85} + \frac{(73-79,62)^2}{79,62} + \frac{(125-121,15)^2}{121,15} + \frac{(162-145,38)^2}{145,38} \\
 &+ \frac{(200-184,62)^2}{184,62} + \frac{(130-141,54)^2}{141,54} + \frac{(220-215,38)^2}{215,38} + \frac{(250-258,46)^2}{258,46} \\
 &= 0,21 + 37,25 + 5,32 + 4,12 + 1,85 + 0,55 + 0,12 + 1,90 + 1,28 + 0,94 + 0,10 + 0,28 \\
 &= 53,91.
 \end{aligned}$$

Determinação do número de graus de liberdade:

- Categorias de Grau de instrução: $s = 3$
 - Categorias de Tipo de mídia: $r = 4$
- $\Rightarrow q = (r - 1) \times (s - 1) = 3 \times 2 = 6$

O nível descritivo (valor- p): $\text{valor-}p = P(\chi_6^2 \geq 53,910) < 0,0001$

Supondo $\alpha = 0,05$, temos **valor- $p < \alpha$** .

Assim, temos evidências para rejeitar a independência entre as variáveis grau de instrução e preferência por tipo de mídia para informação, ao nível de 5% de significância, i.é, há evidências amostrais de que a preferência por uma mídia depende do grau de instrução do usuário.

Os cálculos podem ser feitos diretamente no *Rcmdr*:

Estatísticas → Tabelas de Contingência → Digite e analise tabela de dupla entrada

Saída do *Rcmdr*:

```
data:  .Table
```

```
X-squared = 53.9099, df = 6, p-value = 7.692e-10
```

```
> .Test$expected # Expected Counts
```

	net	tv	re_soc	outras
1	11.53846	16.15385	13.46154	8.846154
2	103.84615	145.38462	121.15385	79.615385
3	184.61538	258.46154	215.38462	141.538462

```
> round(.Test$residuals^2, 2) # Chi-square Components
```

	net	tv	re_soc	outras
1	0.21	37.25	5.32	4.12
2	1.85	0.55	0.12	1.90
3	1.28	0.94	0.10	0.28

Exemplo 4: 1237 indivíduos adultos classificados segundo a pressão sanguínea (*mm Hg*) e o nível de colesterol (*mg/100cm³*).

Verificar se existe dependência entre essas variáveis.

Colesterol	Pressão			Total
	< 127	127 a 166	> 166	
< 200	117	168	22	307
200 a 260	204	418	63	685
> 260	67	145	33	245
Total	388	731	118	1237

Hipóteses:

H_0 : Pressão sanguínea e nível de colesterol são independentes;

H_1 : Nível de colesterol e pressão sanguínea são variáveis dependentes

Rcmdr: Estatísticas → Tabelas de Contingência → Digite e analise tabela de dupla entrada

Saída do Rcmdr: data: .Table
X-squared = 13.5501, df = 4, **p-value = 0.008878**

> .Test\$expected # Expected Counts

	1	2	3
1	96.29426	181.4204	29.28537
2	214.85853	404.7979	65.34357
3	76.84721	144.7817	23.37106

> round(.Test\$residuals^2, 2) # Chi-square Components

	1	2	3
1	4.45	0.99	1.81
2	0.55	0.43	0.08
3	1.26	0.00	3.97

Para $\alpha = 0,05$, temos **valor- $p < \alpha$** . Assim, temos evidências para rejeitar a hipótese de independência entre as variáveis pressão sanguínea e nível de colesterol ao nível de 5% de significância.

Exemplo 5: Uma indústria, desejando melhorar o nível de seus funcionários em cargos de chefia, montou 2 cursos experimentais de inglês utilizando 2 metodologias distintas (MA , MB). Os dados referentes ao conceito obtido no curso (A , B ou C) e metodologia utilizada estão na tabela a seguir:

- (a) Identifique as variáveis em estudo. Classifique-as.
- (b) Construa uma tabela de contingência para as variáveis “metodologia” e “conceito”.
- (c) Conclua se existe associação entre essas variáveis ($\alpha = 10\%$).

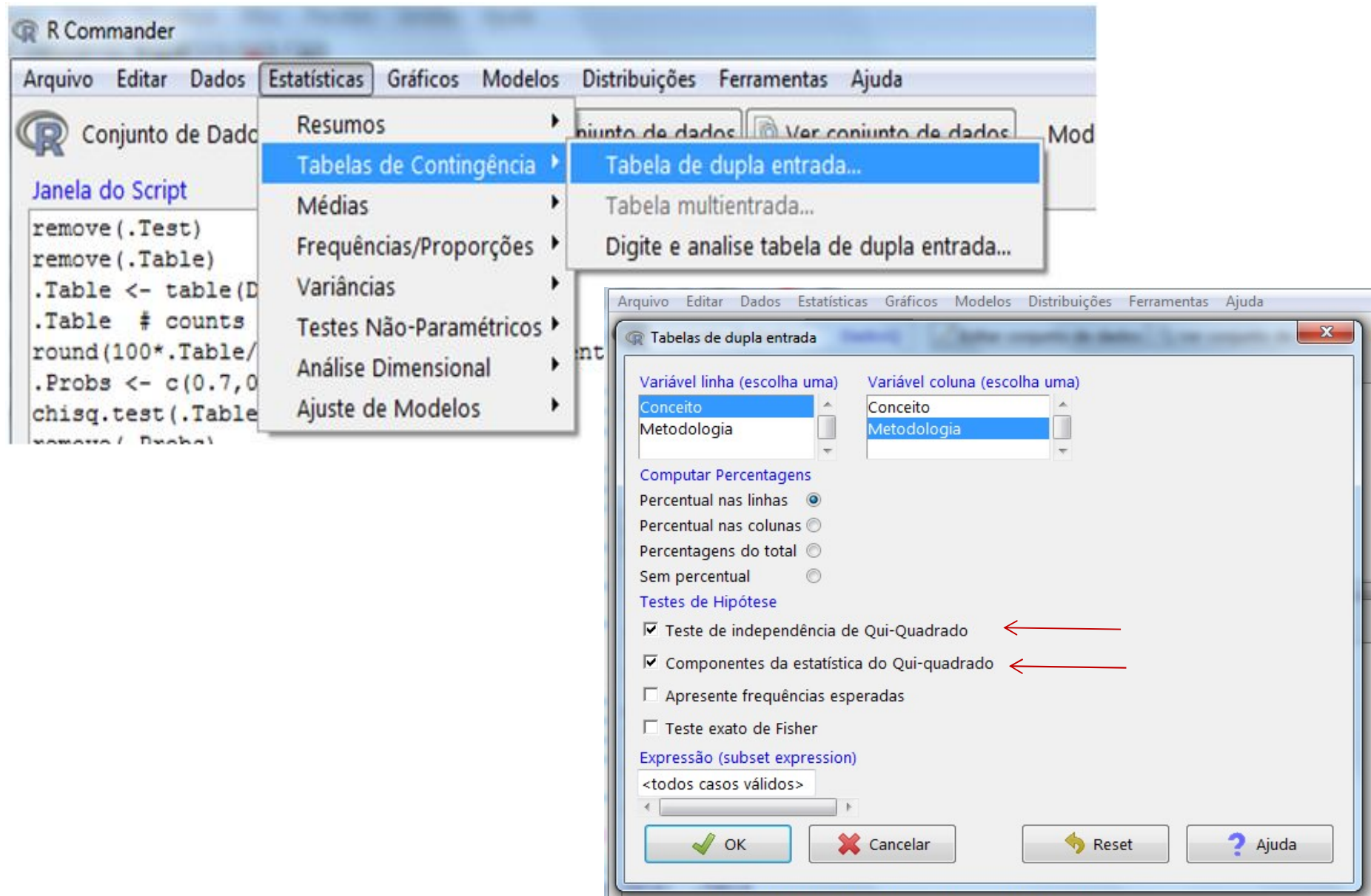
Dados:

Funcionário	Metodologia	Conceito
1	MA	A
2	MA	B
3	MB	A
4	MB	B
5	MA	A
6	MA	B
7	MA	C
8	MB	B
9	MB	B
10	MA	B
11	MB	C
12	MB	A
13	MB	B
14	MB	A
15	MB	C
16	MA	A
17	MA	B
18	MB	C
19	MA	C
20	MB	C
21	MB	A
22	MA	C
23	MB	C
24	MA	A
25	MA	B
26	MB	B
27	MA	A
28	MB	C
29	MA	A
30	MA	B
31	MA	A
32	MA	A
33	MB	B
34	MB	B
35	MA	A
36	MA	A
37	MA	A
38	MB	B
39	MB	C
40	MB	C

Variáveis:

- Metodologia: qualitativa nominal
- Conceito: qualitativa ordinal

Rcmdr: Construção da tabela de contingência (ou tabela de frequências conjuntas)



Saída do *Rcmdr*:

```
> .Table
```

	Metodologia	
Conceito	MA	MB
A	11	4
B	6	8
C	3	8


```
> rowPercents(.Table) # Row Percentages
```

	Metodologia			
Conceito	MA	MB	Total	Count
A	73.3	26.7	100	15
B	42.9	57.1	100	14
C	27.3	72.7	100	11

X-squared = 5.8251, df = 2, **p-value = 0.05434**

```
> round(.Test$residuals^2, 2) # Chi-square Components
```

	Metodologia	
Conceito	MA	MB
A	1.63	1.63
B	0.14	0.14
C	1.14	1.14

Para $\alpha = 0,10$, temos **valor- $p < \alpha=0,10$** , então, H_0 é rejeitada, ou seja, os dados mostram evidências de que o conceito no curso depende da metodologia de ensino, ao nível de 10% de significância.