- Testes Qui-quadrado - Aderência e Independência

1. Testes de Aderência

Objetivo: Testar a validade de um modelo probabilístico discreto a um conjunto de dados observados

Exemplo 1: Segundo Mendel (geneticista famoso), os resultados dos cruzamentos de ervilhas amarelas redondas com ervilhas verdes enrugadas seguem uma distribuição de probabilidades dada por:

Resultados	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:

Resultados	Amarela redonda	Amarela enrugada	Verde redonda	Verde enrugada
Frequência observada	315	101	108	32

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

Temos 4 categorias para os resultados dos cruzamentos:

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

Segundo Mendel, a probabilidade de cada categoria é dada pela seguinte tabela:

Resultados	AR	AE	VR	VE
Probabilidade	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
$\boldsymbol{AR}$	315
$\boldsymbol{AE}$	101
$\boldsymbol{VR}$	108
$\boldsymbol{VE}$	33
Total	556

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

Se o modelo probabilístico (de Mendel) for adequado, a ***frequência esperada*** de ervilhas do tipo $\boldsymbol{AR}$, dentre as 556 observadas, pode ser calculada por:

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

Da mesma forma, temos para o tipo $\boldsymbol{AE}$,

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

Para o tipo $\boldsymbol{VR}$ temos

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

E, para o tipo $\boldsymbol{VE}$,

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

Podemos expandir a tabela de frequências dada anteriormente:

Tipo de resultado	Frequência observada	Frequência esperada
$\boldsymbol{AR}$	315	312,75
$\boldsymbol{AE}$	101	104,25
$\boldsymbol{VR}$	108	104,25
$\boldsymbol{VE}$	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
⋮	⋮
k	O_k
Total	n

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

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

O objetivo do teste de aderência é testar as hipóteses

$$H : p_1 = p_{o1} , \quad .... , p_k = p_{ok}$$

$$A : \text{existe pelo menos uma diferença}$$

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

Se $\boldsymbol{E_i}$ é o total de indivíduos *esperados* na categoria i, quando a hipótese H é verdadeira, então:

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

Expandindo a tabela de frequências original, temos

Categorias	**Frequência observada**	**Frequência esperada sob H**
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 discrepância 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 verdadeira,

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

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

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

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

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

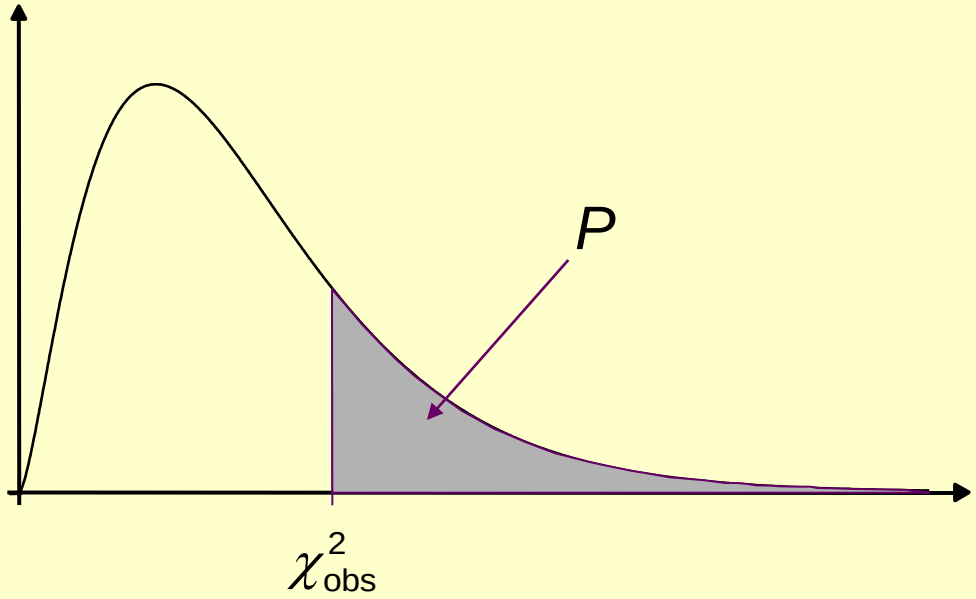
Regra de decisão:

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

$$P = \mathrm{P}\left(\chi_q^2 \geq \chi_{obs}^2\right),$$

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

Graficamente:



Se, para α fixado, obtemos $P \leq \alpha$, **rejeitamos a hipótese *H*.**

Exemplo (continuação): Cruzamentos de ervilhas

Hipóteses:

H : O modelo probabilístico proposto por Mendel é adequado.

A : O modelo proposto por Mendel não é adequado.

De forma equivalente, podemos escrever:

H: $\mathrm{P}(AR)$ =9/16, $\mathrm{P}(AE)$ = 3/16,

$\mathrm{P}(VR)$ = 3/16 e $\mathrm{P}(VE)$ = 1/16.

A: ao menos uma das igualdades não se verifica.

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

Resultado	O_i	E_i
AR	315	312,75
AE	101	104,25
VR	108	104,25
VE	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 calculado por

$$P = \mathrm{P}(\chi^2_3 \geq 0{,}470) = 0{,}925.$$

Conclusão: Para α = 0,05, como P **= 0,925 > 0,05**, não há evidências para rejeitarmos a hipótese H, isto é, ao nível de significância de 5%, concluímos o modelo de probabilidades de Mendel se aplica aos resultados do experimento.

O cálculo do *nível descritivo* ***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

$\Rightarrow$ O que pode ser dito?

Hipóteses a serem testadas:

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

A: 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,

H: $p_i = 1/7$ para todo $i = 1,\ldots, 7$
A: $p_i \neq 1/7$ para pelo menos um valor de i.

Total de acidentes na semana: n =140.

Logo, se H for verdadeira,

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

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:

$$\chi^2_{obs} = \sum_{1}^{7} \frac{(O_i - E_i)^2}{E_i} = \frac{(20-20)^2}{20} + \frac{(10-20)^2}{20} + \frac{(10-20)^2}{20} + \frac{(15-20)^2}{20} + \frac{(30-20)^2}{20} + \frac{(20-20)^2}{20} + \frac{(35-20)^2}{20} = 27{,}50$$

Neste caso, temos $\chi^2 \sim \chi^2_6$, aproximadamente.

O nível descritivo é dado por $P = \mathrm{P}(\chi^2_6 \geq 27{,}50) \cong 0{,}00012$,

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 α = 0,05, temos que P **= 0,0001 <** α**.** Assim, há evidências para **rejeitarmos** H, ou seja, concluímos ao nível de significância de 5% que o número de acidentes não é o mesmo em todos os dias da semana.

2. Testes de Independência

Objetivo: Verificar se existe independência entre duas variáveis medidas nas mesmas unidades experimentais.

Exemplo 3: A Associação de Imprensa do Estado de São Paulo fez um levantamento com 1300 leitores, para verificar se a preferência por leitura de um determinado jornal é independente do nível de instrução do indivíduo. Os resultados obtidos foram:

	Tipo de Jornal				
Grau de instrução	Jornal A	Jornal B	Jornal C	Outros	Total
1º Grau	10	8	5	27	50
2º Grau	90	162	125	73	450
Universitário	200	250	220	130	800
Total	300	420	350	230	1300

Proporções segundo os totais das colunas são apresentadas na seguinte tabela:

	Tipo de Jornal				
Grau de instrução	Jornal A	Jornal B	Jornal C	Outros	Total
1º Grau	3,33%	1,90%	1,43%	11,74%	3,85%
2º Grau	30,00%	38,57%	35,71%	31,74%	34,62%
Universitário	66,67%	59,52%	62,86%	56,52%	61,54%
Total	100,00%	100,00%	100,00%	100,00%	100,00%

As diferenças entre proporções em cada coluna e proporções correspondentes na última coluna não são muito pequenas, indicando que as variáveis (Tipo de Jornal e Grau de Instrução) parecem não ser independentes.

Testes de Independência – Metodologia

Em geral, os dados referem-se a observações de duas variáveis (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: A e B são variáveis independentes

A: As variáveis A e B não são independentes

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

Sendo O_{ij} o total de observações na casela (i, j), se A e B forem independentes, esperamos que, para todos os possíveis pares (A_i e B_j):

$$O_{i1}/O_{.1} = O_{i2}/O_{.2} = \ldots = O_{is}/O_{.s} = O_{i.}/n,\ i = 1, \ldots, r$$

ou ainda

$$O_{ij}/O_{.j} = O_{i.}/n = 1, \ldots, r,\ j = 1, \ldots, s$$

de onde se deduz, finalmente, que

$$O_{ij} = (O_{i.} \times O_{.j})/n,\ i = 1, 2, \ldots, r \text{ e } j = 1, 2, \ldots, s.$$

Logo, o *número esperado de observações com as características* (A_i e B_j), entre as n observações, sob a hipótese de independência, é dado por

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

A medida que representa as diferenças entre os valores observados e os valores esperados em todas as caselas (i,j) sob a suposição de independência é:

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

Estatística do teste de independência

Supondo H verdadeira,

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

aproximadamente,

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

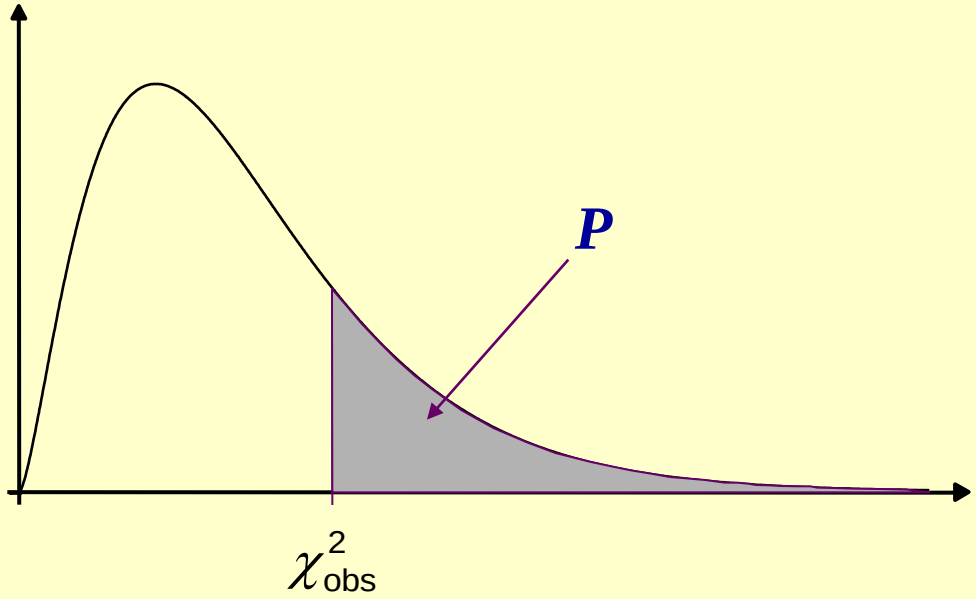
Regra de decisão:

Pode ser baseada no valor P (nível descritivo), neste caso

$$P = \mathrm{P}\left(\chi_q^2 \geq \chi_{obs}^2 \right)$$

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

Graficamente:



Se, para α fixado, obtemos $P \leq \alpha$, **rejeitamos a hipótese H de independência.**

Exemplo (continuação): Estudo da independência entre preferência por um tipo de jornal e grau de instrução. 1300 eleitores foram entrevistados ao acaso.

	Tipo de Jornal				
Grau de instrução	Jornal A	Jornal B	Jornal C	Outros	Total
1º Grau	10	8	5	27	50
2º Grau	90	162	125	73	450
Universitário	200	250	220	130	800
Total	300	420	350	230	1300

Hipóteses H: As variáveis preferência por um tipo de jornal e grau de instrução são independentes.

A: As variáveis não são independentes.

Tabela de valores observados e esperados (entre parênteses)

Grau de instrução	Tipo de Jornal				
	Jornal A	Jornal B	Jornal C	Outros	Total
1º Grau	10 (11,54)	8 (16,15)	5 (13,46)	27 (8,85)	50
2º Grau	90 (103,85)	162 (145,38)	125 (121,15)	73 (79,62)	450
Universitário	200 (184,62)	250 (258,46)	220 (215,38)	130 (141,54)	800
Total	300	420	350	230	1300

Lembre-se, sob a hipótese de independência, para todas caselas (i,j):

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

Leitores que têm 1º Grau e preferem o jornal A: $E_{11} = \frac{300 \times 50}{1300} = 11,54$

2º Grau e prefere jornal B: $E_{22} = \frac{420 \times 450}{1300} = 145,38$

Universitário e prefere outros jornais: $E_{34} = \frac{230 \times 800}{1300} = 141,54$

Cálculo da estatística de qui-quadrado:

	Tipo de Jornal				
Grau de instrução	Jornal A	Jornal B	Jornal C	Outros	Total
1º Grau	10 (11,54)	8 (16,15)	5 (13,46)	27 (8,85)	**50**
2º Grau	90 (103,85)	162 (145,38)	125 (121,15)	73 (79,62)	**450**
Universitário	200 (184,62)	250 (258,46)	220 (215,38)	130 (141,54)	**800**
Total	**300**	**420**	**350**	**230**	**1300**

$$\chi^2_{obs} = \sum_{i=1}^{3}\sum_{j=1}^{4}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}} = \frac{(10-11,54)^2}{11,54} + \frac{(8-16,15)^2}{16,15} + \frac{(5-13,46)^2}{13,46} + \frac{(27-8,85)^2}{8,85}$$

$$+\frac{(90-103,85)^2}{103,85} + \frac{(162-145,38)^2}{145,38} + \frac{(125-121,15)^2}{121,15} + \frac{(73-79,62)^2}{79,62}$$

$$+\frac{(200-184,62)^2}{184,62} + \frac{(250-258,46)^2}{258,46} + \frac{(220-215,38)^2}{215,38} + \frac{(130-141,54)^2}{141,54}$$

$$= 53,910.$$

Determinação do número de graus de liberdade:

- Categorias de Grau de instrução: $s = 3$
- Categorias de Tipo de jornal: $r = 4$

$\Rightarrow q = (r - 1)\times(s - 1) = 3 \times 2 = 6$

O nível descritivo (valor P):

$$P = \mathrm{P}\left(\chi^2_6 \geq 53{,}910\right) < 0{,}0001$$

$\therefore$ Supondo $\alpha = 0{,}05$, temos $\boldsymbol{P} < \alpha$. Assim, temos evidências para **rejeitar a independência** entre as variáveis grau de instrução e preferência por tipo de jornal ao nível de 5% de significância.

Os cálculos podem ser feitos diretamente no Rcmdr:

Estatísticas $\rightarrow$ Tabelas de Contingência $\rightarrow$ 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
          1         2         3          4
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
     1    2    3     4
1 0.21 4.12 5.32 37.25
2 1.85 1.90 0.12  0.55
3 1.28 0.28 0.10  0.94
```

Exemplo 4: 1237 indivíduos adultos classificados segundo a pressão sanguínea ($mm\ Hg$) e o nível de colesterol ($mg/100cm^3$).

Verificar se existe independência 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

H: Pressão sanguínea e nível de colesterol são independentes;

A: Nível de colesterol e pressão sanguínea não são independentes.

Os cálculos podem ser feitos diretamente no Rcmdr: Estatísticas $\rightarrow$ Tabelas de Contingência $\rightarrow$ 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 α = 0,05, temos ***P*** < α. Assim, temos evidências para rejeitar a hipótese de independência entre as variáveis pressão sanguínea e nível de colesterol ao nível de 5% de significância.