Testes Qui-quadrado

Teste de Aderência

Teste de Independência

1. Testes de Aderência

Objetivo: Testar a adequabilidade 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:

	AR	*AE*	*VR*	*VE*
Probabilidades: (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
AR	315
AE	101
VR	108
VE	33
Total	556

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

- ***Teste da "Aderência" do Modelo Probabilístico com as observações experimentais***

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)
AR	315	312,75
AE	101	104,25
VR	108	104,25
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?

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

Estas duas tabelas "são parecidas"?

Como medir "distância entre tabelas"?

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, $i = 1,\ldots,k$.

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

Queremos confrontar duas hipóteses

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

H_1: 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_0 é verdadeira, então:

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

proporção esperada na categoria i
SE o Modelo Probabilístico for verdadeiro

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
⋮	⋮	⋮
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

"Estatística Qui-quadrado"

Supondo H_0 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.

Em outras palavras, se H_0 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

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

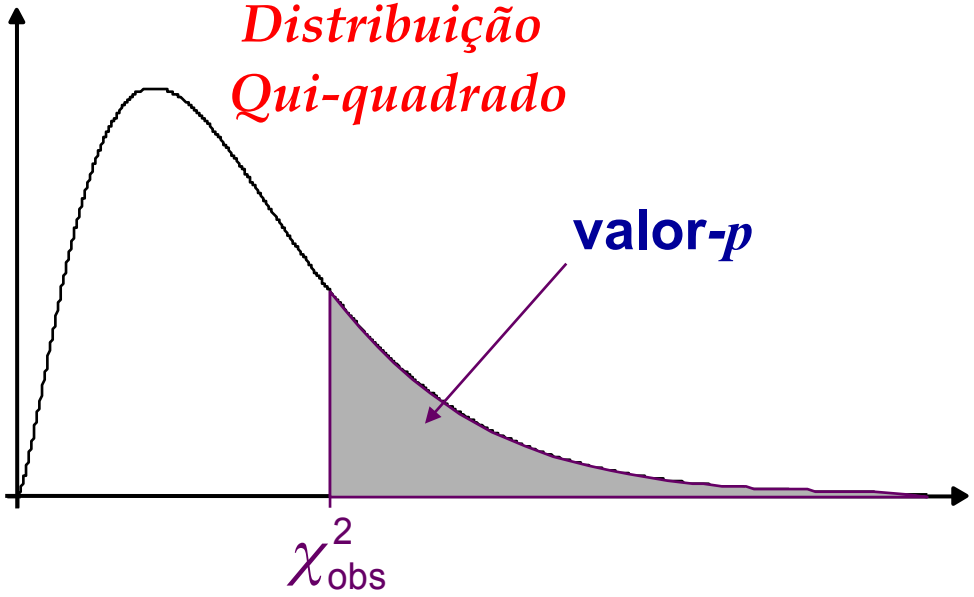
Regra de decisão:

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

$$\text{valor-}p = P(\chi_q^2 \geq \chi^2_{obs} \mid \text{"hipótese } H \text{ é verdadeira"})$$

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 **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
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 dado por: **valor-*p*** $= P(\chi^2_3 \geq 0{,}470) \cong 0{,}925$.

Conclusão: Para α=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%, concluímos 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

Pergunta:

Há associação entre

Número de acidentes
e
Dia da semana?

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 $i = 1,\ldots, 7$

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,

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

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:

$$\chi^2_{obs} = \sum_{1}^{4} \frac{(O_i - E_i)^2}{E_i} = \textit{27,5 (Verifique!)}$$

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

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

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).

(Exercício: verifique o valor-p encontrado usando Rcmdr, Rstudio ou outro software estatístico)

Conclusão: Para α = 0,05, temos que **valor-p = 0,0001 < α.**

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 interar 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.

Perfil global 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 300x0,0385=**11,54**(=300x50/1300),
- **Médio e preferem Internet** é 300x0,3462=**103,85** (=300x450/1300),
- **Superior e preferem Internet** é 300x0,6154=**184,62** (=300x800/1300).

	Tipo de mídia				
Grau de instrução	**Internet**	**TV**	**Rede Social**	**Outras**	**Total**
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

A \ 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?

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

$A \setminus B$	B_1 ... B_j	...	B_s	Total
A_1	O_{11} ... O_{1j}	...	O_{1s}	$O_{1.}$
...				
A_i	O_{i1} ... O_{ij}	...	O_{is}	$O_{i.}$
...	... ...	...	...	...
A_r	O_{r1} O_{rj}	...	O_{rs}	$O_{r.}$
Total	$O_{.1}$ $O_{.j}$	...	$O_{.s}$	n

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

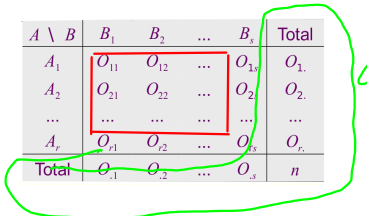
$$\chi^2 = \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Supondo H_0 verdadeira,

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

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

r linhas
s colunas



Os valores marginais estão fixos: temos (r-1)(s-1) "campos livres"

Regra de decisão:

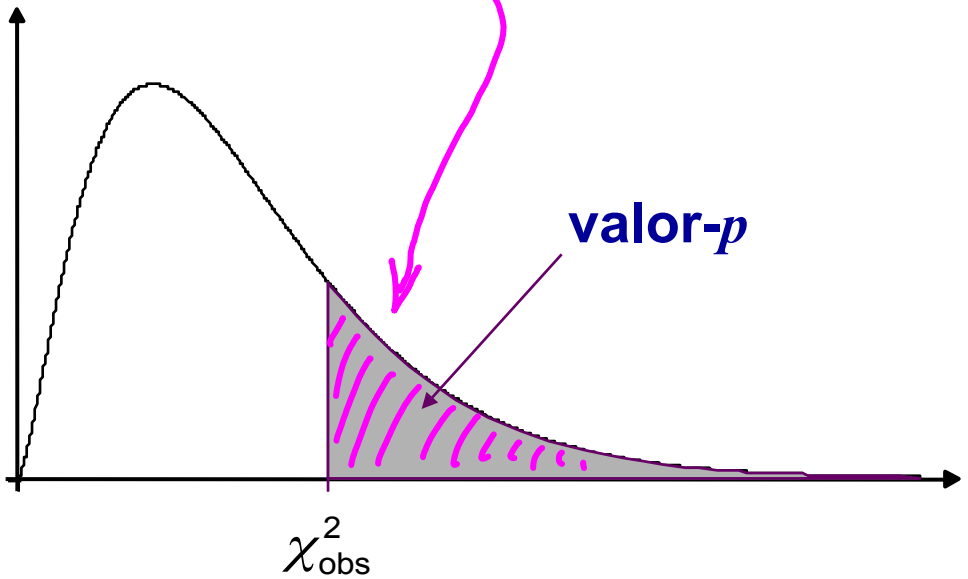
baseada no **valor-*p*** (nível descritivo)

valor-p $= P(\chi_q^2 \geq \chi^2_{obs} \mid$ *"vale independência"*$)$

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

Graficamente:

Distribuição Qui-quadrado com q "graus de liberdade"



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**

Médio e prefere TV:

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

Superior e prefere outras mídias:

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

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

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**

$$\chi^2_{obs} = \frac{(10-11{,}54)^2}{11{,}54} + \frac{(27-8{,}85)^2}{8{,}85} + \frac{(5-13{,}46)^2}{13{,}46} + \frac{(8-16{,}15)^2}{16{,}15}$$

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

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

$$= 0{,}21 + 37{,}25 + 5{,}32 + 4{,}12 + 1{,}85 + 0{,}55 + 0{,}12 + 1{,}90 + 1{,}28 + 0{,}94 + 0{,}10 + 0{,}28$$

$$= 53{,}91.$$

Determinação do **número de graus de liberdade:**

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

O nível descritivo (valor-p): valor-$p = P(\chi^2_6 \geq 53{,}910) < 0{,}0001$

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

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