IPS Instituto Politécnico de Setúbal
Escola Superior de Tecnologia de Setúbal

MÉTODOS ESTATÍSTICOS

Testes de Hipóteses Não Paramétricos - Parte 2

Teste de Independência

Licenciatura em Engenharia Informática

Departamento de Matemática
Escola Superior de Tecnologia de Setúbal
Instituto Politécnico de Setúbal
2023-2024

Testes de Hipóteses Não Paramétricos:

Teste de Independência do Qui-Quadrado

- Pretende-se verificar se existe ou não independência entre duas variáveis, ou seja, este teste é usado para descobrir se existe associação entre duas variáveis qualitativas que se apresentem agrupadas numa tabela de contingência.
- Apenas vamos considerar tabelas de contingência bidimensionais (mas é possível analisar a independência de variáveis em tabelas de dimensão superior a 2 - não será abordado).

Teste de independência do Qui-Quadrado

Dados Bivariados

- Por vezes a população que se pretende estudar, aparece sob a forma de pares de valores, isto é, cada indivíduo ou resultado experimental, contribui com um conjunto de dois valores.
- É o que acontece quando se pretende estudar dois atributos da mesma população visando investigar em que medida eles se relacionam, isto é, de que modo a variação de um deles exerce influencia na variação do outro.
- Quando os atributos são ambos **quantitativos**, como iremos ver, podemos recorrer à **Regressão Linear Simples**.
- Quando os atributos são ambos **qualitativos** vamos recorrer ao **Teste de Independência do Qui-Quadrado**.

Observação:

Uma variável originalmente quantitativa pode ser recolhida ou transformada em qualitativa.

Por exemplo, a variável idade, medida em anos é quantitativa (contínua), mas, se for obtida ou transformada em níveis etários (0 a 5 anos, 6 a 10 anos,...), é qualitativa (ordinal).

Teste de independência do Qui-Quadrado

Objetivo

Estudar a relação entre duas **variáveis qualitativas**.

Para atingir este objetivo vamos investigar a presença ou ausência de **associação** entre as duas variáveis. Essa investigação será feita em duas etapas:

- etapa 1 $\rightarrow$ resumir os dados
  - tabelas de dupla entrada: **tabelas de contingência** também chamadas de tabelas de informação cruzada;
- etapa 2 $\rightarrow$ testar, estatisticamente, se existe associação entre as variáveis: **teste de independência do Qui-Quadrado**.

Teste de independência do Qui-Quadrado

Tabelas de Contingência

É uma tabela de dupla entrada:

- as r categorias de uma das variáveis definem as linhas,
- as c categorias da outra variável definem as colunas,
- a tabela tem $r \times c$ células.

	Variável B				
Variável A	$\mathbf{B_1}$	$\mathbf{B_2}$	$\cdots$	$\mathbf{B_c}$	**TOTAL**
$\mathbf{A_1}$	O_{11}	O_{12}	$\cdots$	O_{1c}	$n_{1.}$
$\mathbf{A_2}$	O_{21}	O_{22}	$\cdots$	O_{2c}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$\mathbf{A_r}$	O_{r1}	O_{r2}	$\cdots$	O_{rc}	$n_{r.}$
TOTAL	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.c}$	n

O_{ij}, $i = 1, \ldots, r$ e $j = 1, \ldots, c$ $\rightarrow$ representa o número de elementos observados na amostra que foram classificados simultaneamente nas categorias A_i da variável A e B_j da variável B.

$n_{i.} = \sum_{j=1}^{c} O_{ij}$ $\rightarrow$ representa o número de elementos da amostra classificados na categoria A_i da variável A, ou seja, representa o total marginal de linha.

$n_{.j} = \sum_{i=1}^{r} O_{ij}$ $\rightarrow$ representa o número de elementos da amostra classificados na categoria B_j da variável B, ou seja, representa o total marginal de coluna.

$n = \sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij} = \sum_{i=1}^{r} n_{i.} = \sum_{j=1}^{c} n_{.j}$ $\rightarrow$ representa o total da tabela, o número total de elementos da amostra.

Exemplo 1

Foi efetuado um estudo onde se procurou analisar a relação existente entre a prática desportiva dos filhos quando os pais praticam ou não desporto. A amostra do presente estudo é constituída por 82 alunos do sexo masculino que frequentavam o 10^o ano de escolaridade de uma dada escola e pelos respetivos pais (ficheiro dados1.txt). Neste caso as variáveis em análise são:

- **Pai** - com as categorias:
  - **Não** - não pratica desporto regularmente,
  - **Sim** - pratica desporto regularmente.
- **Filho** - com as categorias:
  - **Não** - não pratica desporto regularmente,
  - **Sim** - pratica desporto regularmente.

Dados:

Pai	Filho
Sim	Não
Sim	Não
Não	Não
Não	Sim
Sim	Sim
⋮	⋮

2 variáveis qualitativas nominais.

Tabela de contingência:

- $r = 2$ linhas, correspondem às 2 categorias da variável "Pai".
- $c = 2$ colunas, correspondem às 2 categorias da variável "Filho".
- $r \times c = 2 \times 2 = 4$ células.

	Filho		
Pai	**Não**	**Sim**	**TOTAL**
Não	24	41	65
Sim	6	11	17
TOTAL	30	52	82

Teste de independência do Qui-Quadrado

Objetivo

Avaliar a existência de associação entre atributos de uma população, estudando a independência entre as variáveis qualitativas que representam esses atributos.

Princípios Básicos na Realização do Teste de Independência do Qui-Quadrado

1. São definidas duas **hipóteses**:
   - **Hipótese Nula** $= H_0$ - é a hipótese que indica que as duas variáveis são independentes.
   - **Hipótese Alternativa** $= H_1$ - é a hipótese que se contrapõe à hipótese nula, ou seja, que indica que o que foi colocado na hipótese nula não se verifica.
2. É definida uma **Estatística Teste**, que é a base da realização do teste e consiste em comparar o observado com o previsto caso as variáveis sejam independentes.
3. São construídas duas regiões:
   - **Região de Aceitação** $= RA$ - conjunto de valores para os quais H_0 é admissível.
   - **Região de Rejeição ou Região Crítica** $= RC$ - conjunto de valores para os quais H_0 não é admissível.

Princípios Básicos na Realização do Teste de Independência do Qui-Quadrado

4. A **regra de decisão** define as condições de rejeição ou não rejeição da hipótese nula:
   - Se o Valor Observado da Estatística de Teste sob a hipótese H_0 pertencer à Região de Aceitação, então Não se Rejeita H_0
   - Se o Valor Observado da Estatística de Teste sob a hipótese H_0 pertencer à Região Crítica, então Rejeita-se H_0

5. **Erros de decisão** - um teste de hipóteses nem sempre conduz a decisões corretas, a análise de uma amostra pode falsear as conclusões quanto à população. Como já vimos, um dos erros é o chamado **Erro de 1^a espécie** ou **Nível de significância do teste**:

$$\alpha = P\left[\text{rejeitar } H_0 \mid H_0 \text{ verdadeira}\right]$$

   para minimizar este erro fixa-se o seu valor.

6. As regiões de aceitação e de rejeição (RA e RC) são definidas à custa do valor fixado para o nível de significância (α).

Na prática, em vez de calcular a região crítica (RC) e a região de aceitação (RA), é usual calcular-se o **Valor-p** (ou **p-value**).

Valor-p (ou p-value)

É a probabilidade associada ao valor da estatística de teste, considerando H_0 verdadeira.

- Se o valor-p for pequeno significa que, no caso de H_0 ser verdadeira, estamos perante um evento muito raro, pouco provável de ocorrer, então deve optar-se por rejeitar H_0.

  Portanto, o valor-p também permite tomar decisões:

  - se valor-p $\leq \alpha$, então rejeita-se H_0
  - se valor-p $> \alpha$, então não se rejeita H_0

Teste de independência do Qui-Quadrado

Objetivo

Avaliar a existência de associação entre atributos de uma população, estudando a independência entre as variáveis qualitativas que representam esses atributos.

Formulação das Hipóteses a Testar:

H_0- Não há relação entre as variáveis

vs

H_1- Há relação entre as variáveis

ou de forma equivalente

H_0- As variáveis são independentes

vs

H_1- As variáveis não são independentes

Teste de independência do Qui-Quadrado

Estatística de Teste

A estatística de teste tem por base os desvios entre as frequências observadas (O_{ij}) e esperadas (E_{ij}):

$$Q = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)\times(c-1)}$$

onde r é o número de linhas da tabela de contingência e c é o número de colunas da tabela de contingência.

Teste de independência do Qui-Quadrado

Cálculo do Valor Observado da Estatística de Teste sob a Hipótese H_0

$$Q_{obs} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- r corresponde ao número de linhas da tabela de contingência
- c corresponde ao número de colunas da tabela de contingência
- **frequências observadas** $= O_{ij} \rightarrow$ corresponde às frequências absolutas observadas (amostra) da tabela de contingência;
- **frequências esperadas** $= E_{ij} = \frac{n_{i.} \times n_{.j}}{n} \rightarrow$ frequências absolutas esperadas se as variáveis são independentes
  - n é a dimensão da amostra
  - $n_{i.}$ totais das linhas
  - $n_{.j}$ totais das colunas

Observação 1: os acontecimentos A e B dizem-se independentes sse $P(A \cap B) = P(A) \times P(B)$

Observação 2: Tem-se $\sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij} = \sum_{i=1}^{r}\sum_{j=1}^{c} E_{ij} = n$

Teste de independência do Qui-Quadrado

Cálculo do Valor Observado da Estatística de Teste sob a Hipótese H_0

O teste de independência do Qui-Quadrado compara as frequências observadas, O_{ij}:

	Variável B				
Variável A	$\mathbf{B_1}$	$\mathbf{B_2}$	$\cdots$	$\mathbf{B_c}$	TOTAL
$\mathbf{A_1}$	O_{11}	O_{12}	$\cdots$	O_{1c}	$n_{1.}$
$\mathbf{A_2}$	O_{21}	O_{22}	$\cdots$	O_{2c}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$\mathbf{A_r}$	O_{r1}	O_{r2}	$\cdots$	O_{rc}	$n_{r.}$
TOTAL	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.c}$	n

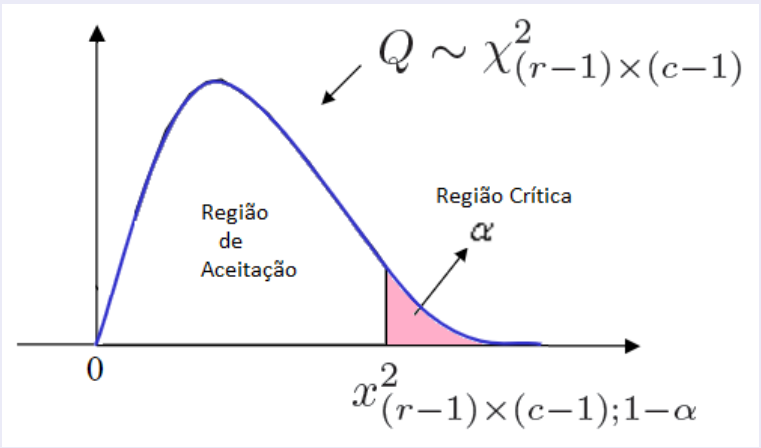
com as frequências esperadas, caso as variáveis fossem independentes, $E_{ij} = \frac{n_{i.} \times n_{.j}}{n}$:

	Variável B				
Variável A	$\mathbf{B_1}$	$\mathbf{B_2}$	$\cdots$	$\mathbf{B_c}$	TOTAL
$\mathbf{A_1}$	$E_{11} = \frac{n_{1.} \times n_{.1}}{n}$	$E_{12} = \frac{n_{1.} \times n_{.2}}{n}$	$\cdots$	$E_{1c} = \frac{n_{1.} \times n_{.c}}{n}$	$n_{1.}$
$\mathbf{A_2}$	$E_{21} = \frac{n_{2.} \times n_{.1}}{n}$	$E_{22} = \frac{n_{2.} \times n_{.2}}{n}$	$\cdots$	$E_{2c} = \frac{n_{2.} \times n_{.c}}{n}$	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$\mathbf{A_r}$	$E_{r1} = \frac{n_{r.} \times n_{.1}}{n}$	$E_{r2} = \frac{n_{r.} \times n_{.2}}{n}$	$\cdots$	$E_{rc} = \frac{n_{r.} \times n_{.c}}{n}$	$n_{r.}$
TOTAL	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.c}$	n

Teste de independência do Qui-Quadrado

Definição da Região de Aceitação e de Região Crítica

Um valor da estatística de teste elevado indica discrepância entre os valores observados e os respetivos valores esperados indicando associação entre as variáveis, ou seja, as variáveis não podem ser consideradas independentes:



- a Região de Aceitação é $RA = \left[0, x^2_{1-\alpha;(r-1)\times(c-1)}\right[$
- a Região Crítica é $RC = \left[x^2_{1-\alpha;(r-1)\times(c-1)}, +\infty\right[$

Teste de independência do Qui-Quadrado

Regra de Decisão com base na Região Crítica

- Se o valor observado da estatística de teste não pertencer à Região Crítica,

$$Q_{obs} \notin RC$$

então, ao nível de significância α, **a hipótese H_0 não é rejeitada**, isto é, com base na amostra há evidências estatísticas que as variáveis são independentes.

- Se o valor observado da estatística de teste pertencer à Região Crítica,

$$Q_{obs} \in RC$$

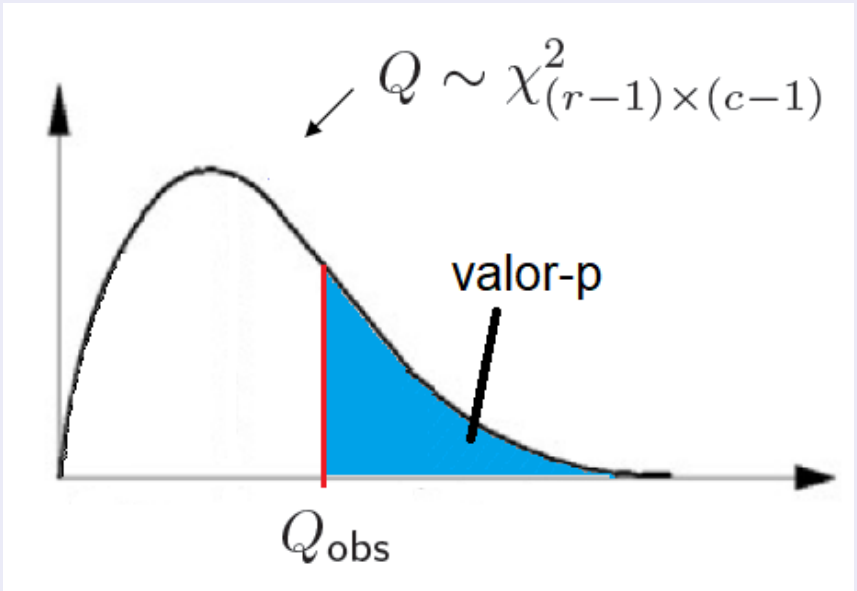
então, ao nível de significância α, **a hipótese H_0 é rejeitada**, isto é, com base na amostra não há evidências estatísticas que as variáveis são independentes.

Teste de independência do Qui-Quadrado

Cálculo do valor-p

Considerando que H_0 é verdadeira, o valor-p indica a probabilidade do valor observado da estatística de teste ocorrer:

$$\textbf{valor-p} = P\left(Q \geq Q_{\text{obs}}\right)$$



O valor-p pode ser visto como o menor valor de α (nível de significância) para o qual os dados observados indicam que H_0 deve ser rejeitada.

Teste de independência do Qui-Quadrado

Regra de Decisão com base no valor-p

- Se

$$\text{valor-p} > \alpha$$

então, ao nível de significância α, **a hipótese H_0 não é rejeitada**, isto é, com base na amostra há evidências estatísticas que as variáveis são independentes.

- Se

$$\text{valor-p} \leq \alpha$$

então, ao nível de significância α, **a hipótese H_0 é rejeitada**, isto é, com base na amostra não há evidências estatísticas que as variáveis são independentes.

Teste de independência do Qui-Quadrado

Condições de aplicação do teste

- Não há mais de 20% das frequências esperadas inferiores a 5, isto é, $E_{ij} < 5$ no máximo em 20% das células dos E_{ij}.
- Todas as frequências esperadas devem ser maiores ou iguais a 1, isto é, $E_{ij} \geq 1$ para todo $i = 1, \ldots, r$ e $j = 1, \ldots, c$.

Observaçção:

- As condições de aplicação do teste devem ser, tanto quanto possível verificadas, sob pena do teste não ser rigoroso. Ou seja, o teste não tem qualquer pressuposto obrigatório, a infração das condições de aplicação apenas leva à perda de rigor.

Teste de independência do Qui-Quadrado

Teste de Independência do Qui-Quadrado no R

- chisq.test()

Observações

- Vamos utilizar a função chisq.test() com o campo "correct=FALSE".

- Quando as condições de aplicação do teste são violadas há a possibilidade de fazer correções nos resultados ou recorrer a outros testes. Por exemplo, a correção de Yates para tabelas de bidimensionais 2×2 (basta colocar "correct=TRUE" na função chisq.test()), o Teste Exato de Fisher (o R tem a função fisher.test()) que pode ser utlizado em tabelas bidimensionais (muito usado em tabelas bidimensionais 2×2) e não exige que as frequências esperadas sejam grandes.

Teste de independência do Qui-Quadrado

Exemplo 1

Foi efetuado um estudo onde se procurou analisar a relação existente entre a prática desportiva dos filhos quando os pais praticam ou não desporto. A amostra do presente estudo é constituída por 82 alunos do sexo masculino que frequentavam o 10^o ano de escolaridade de uma dada escola e pelos respetivos pais (ficheiro dados1.txt). As variáveis em análise e a respetiva tabela de contingência são:

Pai - com as categorias:

- **Não** - não pratica desporto regularmente,
- **Sim** - pratica desporto regularmente.

Filho - com as categorias:

- **Não** - não pratica desporto regularmente,
- **Sim** - pratica desporto regularmente.

	Filho	
Pai	**Não**	**Sim**
Não	24	41
Sim	6	11

Será que o facto dos pais praticarem ou não desporto regularmente influencia o facto dos filhos praticarem ou não desporto regularmente? Ou seja, para um nível de significância de 5%, será que as variáveis são independentes?

Hipótese a ser testada

H_0 : os pais praticarem ou não desporto regularmente **não influencia**

o facto dos filhos praticarem ou não desporto regularmente

$$vs$$

H_1 : os pais praticarem ou não desporto regularmente **influencia**

o facto dos filhos praticarem ou não desporto regularmente

Dados

- Variáveis: 2 variáveis qualitativas nominais
- Tabela de contingência: $r = 2$ linhas e $c = 2$ colunas
- nível de significância $= \alpha = 0.05$

- Tabela de contingência das frequências Observadas:

	Filho		
Pai	**Não**	**Sim**	**TOTAL**
Não	24	41	65
Sim	6	11	17
TOTAL	30	52	82

- Tabela de contingência das frequências Esperadas:

	Filho		
Pai	**Não**	**Sim**	**TOTAL**
Não	$23.7805 = \frac{30 \times 65}{82}$	$41.2195 = \frac{52 \times 65}{82}$	65
Sim	$6.2195 = \frac{30 \times 17}{82}$	$10.7805 = \frac{52 \times 17}{82}$	17
TOTAL	30	52	82

- Estatística de teste:

$$Q_{obs} = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} =$$

$$= \frac{(24 - 23.7805)^2}{23.7805} + \frac{(41 - 41.2195)^2}{41.2195} + \frac{(6 - 6.2195)^2}{6.2195} + \frac{(11 - 10.7805)^2}{10.7805} = 0.0154$$

A estatística de teste, sob a hipótese H_0, tem distribuição Qui-Quadrado com

$$(r-1)\times(c-1) = (2-1)\times(2-1) = 1 \quad \text{graus de liberdade}$$

$$Q \sim \chi^2_{(1)}$$

Regra de Decisão através da Região Crítica

$$RC = \left[x^2_{1-\alpha;(r-1)\times(c-1)}, +\infty\right[= \left[x^2_{0.95;(1)}, +\infty\right[= [3.84, +\infty[$$

Como $Q_{obs} = 0.0154 \notin RC$ então não se rejeita a hipótese H_0

Regra de Decisão através do valor-p

$$\text{valor-}p = P(Q \geq Q_{obs}) = P(Q \geq 0.0154) \underset{\text{v.a. contínua}}{=} 1 - F(0.0154) = 0.9012$$

Como valor-$p > 0.05 = \alpha$ então não se rejeita a hipótese H_0

Conclusão: Com base na amostra e para um nível de significância de 5%, existem evidências estatísticas, que o facto dos pais praticarem ou não desporto habitualmente não influencia o facto dos filhos praticarem ou não desporto habitualmente (ou seja, são independentes).

R

usar a função $chisq.test(..., correct = FALSE)$

e obtém-se

- $Q_{obs} = 0.015412$
- graus de liberdade $= 1$
- valor-$p = 0.9012$

Como valor-$p = 0.9012 > 0.05 = \alpha$ então não se rejeita a hipótese H_0

Conclusão: Com base na amostra e para um nível de significância de 5%, existem evidências estatísticas, que o facto dos pais praticarem ou não desporto habitualmente não influencia o facto dos filhos praticarem ou não desporto habitualmente (ou seja, são independentes).

Teste de independência do Qui-Quadrado

Exemplo 2

Com o objetivo de tentar "explicar as causas" do insucesso escolar foram inquiridos vários alunos do ensino básico. Aos alunos foram colocadas diversas questões, entre as quais uma sobre o número de reprovações e outra sobre o número de faltas (ficheiro dados2.txt). As variáveis em análise e a respetiva tabela de contingência são:

Número de reprovações - com as categorias:

- **Nenhuma**
- **Uma**
- **Duas ou mais**

Número de faltas - com as categorias:

- **Nenhuma**
- **Algumas**
- **Muitas**

	Número de faltas		
Número de reprovações	**Nenhuma**	**Algumas**	**Muitas**
Nenhuma	132	57	13
Uma	28	15	15
Duas ou mais	20	17	17

Será que existe relação entre as variáveis "Número de faltas" e "Número de reprovações"? Ou seja, para um nível de significância de 1%, será que as variáveis são independentes?

Hipótese a ser testada

H_0 : as variáveis "Número de faltas" e "Número de reprovações" **não estão** relacionadas

contra

H_1 : as variáveis "Número de faltas" e "Número de reprovações" **estão** relacionadas

Dados

- Variáveis: 2 variáveis qualitativas ordinais
- Tabela de contingência: $r = 3$ linhas e $c = 3$ colunas
- nível de significância $= \alpha = 0.01$

- Tabela de contingência das frequências Observadas:

	Número de faltas			
Número de reprovações	**Nenhuma**	**Algumas**	**Muitas**	**TOTAL**
Nenhuma	132	57	13	202
Uma	28	15	15	58
Duas ou mais	20	17	17	54
TOTAL	180	89	45	314

- Tabela de contingência das frequências Esperadas:

	Número de faltas			
Número de reprovações	**Nenhuma**	**Algumas**	**Muitas**	**TOTAL**
Nenhuma	$115.7962 = \frac{180\times 202}{314}$	$57.2548 = \frac{89\times 202}{314}$	$28.9490 = \frac{45\times 202}{314}$	202
Uma	$33.2484 = \frac{180\times 58}{314}$	$16.4395 = \frac{89\times 58}{314}$	$8.3121 = \frac{45\times 58}{314}$	58
Duas ou mais	$30.9554 = \frac{180\times 54}{314}$	$15.3057 = \frac{89\times 54}{314}$	$7.7389 = \frac{45\times 54}{314}$	54
TOTAL	180	89	45	314

- Estatística de teste:

$$Q_{obs} = \sum_{i=1}^{3}\sum_{j=1}^{3} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} =$$

$$= \frac{(132 - 115.7962)^2}{115.7962} + \frac{(57 - 57.2548)^2}{57.2548} + \cdots + \frac{(17 - 15.3057)^2}{15.3057} + \frac{(17 - 7.7389)^2}{7.7389} = 32.539$$

A estatística de teste, sob a hipótese H_0, tem distribuição Qui-Quadrado com

$$(r-1)\times(c-1) = (3-1)\times(3-1) = 4 \quad \text{graus de liberdade}$$

$$Q \sim \chi^2_{(4)}$$

Regra de Decisão através da Região Crítica

$$RC = \left[x^2_{1-\alpha;(r-1)\times(c-1)}, +\infty\right[= \left[x^2_{0.99;(4)}, +\infty\right[= [13.3, +\infty[$$

Como $Q_{obs} = 32.539 \in RC$ então rejeita-se a hipótese H_0

Regra de Decisão através do valor-p

$$\text{valor-}p = P(Q \geq Q_{obs}) = P(Q \geq 32.539) \underset{\text{v.a. contínua}}{=} 1 - F(32.539) = 1 - 1 = 0$$

Como valor-$p = 0 \leq 0.01 = \alpha$ então rejeita-se a hipótese H_0

Conclusão: Com base na amostra e para um nível de significância de 1%, existem evidências estatísticas, que o número de reprovações e o número de faltas estão relacionados (ou seja, não são independentes).

R

usar a função $chisq.test()$

e obtém-se

- $Q_{obs} = 32.539$
- graus de liberdade $= 4$
- valor-$p = 1.484e-06 = 0.000001484$

Como valor-$p = 0.000001484 \leq 0.01 = \alpha$ então rejeita-se a hipótese H_0

Conclusão: Com base na amostra e para um nível de significância de 1%, existem evidências estatísticas, que o número de reprovações e o número de faltas estão relacionados (ou seja, não são independentes).

Medidas de Associação

Depois de tomada a decisão e nos casos em que se rejeita a hipótese nula, tem interesse em saber se a associação existente entre as variáveis é forte ou fraca.

Medidas de Associação

- **coeficiente de contingência**: assume valores entre 0 e 1, mas nunca atinge o valor 1. O valor 0 corresponde a ausência de associação entre as variáveis, valores próximos de zero correspondem a fraca associação e valores elevados correspondem a associação mais forte.
- **coeficiente V de Crámer**: assume valores entre 0 e 1. O valor 0 corresponde à ausência de associação entre as variáveis, valores próximos de zero correspondem a fraca associação e valores próximos de 1 correspondem a associação forte.

Estas medidas são muito usadas quando **pelo menos 1 das variáveis é qualitativa nominal**.

Uma possível interpretação dos coeficientes pode ser a apresentada na tabela seguinte, mas estes limites não são rígidos, são apenas linhas de orientação:

	Associação		
	fraca	moderada	elevada
coeficiente de contingência	$[0.10, 0.30[$	$[0.30, 0.50[$	≥ 0.50
V de Crámer ($k = 2^{(*)}$)	$[0.10, 0.30[$	$[0.30, 0.50[$	≥ 0.50
V de Crámer ($k = 3^{(*)}$)	$[0.07, 0.20[$	$[0.20, 0.35[$	≥ 0.35
V de Crámer ($k = 4^{(*)}$)	$[0.06, 0.17[$	$[0.17, 0.29[$	≥ 0.29

$^{(*)}k$ representa o número mínimo de categorais nas linhas ou nas colunas

Medidas de Associação

Medidas de Associação $\mapsto$ as 2 das variáveis são qualitativas ordinais

- **coeficiente τ_b de Kendall**: assume valores entre -1 e 1, mas os valores -1 e 1 só são atingidos em tabelas em que o número de linhas é igual ao número de colunas. Valores próximos de -1 ou de 1 indicam forte associação. Valores próximos de zero indicam fraca associação.

  Sinal do coeficiente:

  - sinal positivo indica que o "aumento" de uma das variáveis é acompanhado pelo "aumento" da outra variável;
  - sinal negativo indica que o "aumento" de uma das variáveis é acompanhado pela "diminuição" da outra variável.

Teste de independência do Qui-Quadrado

Medidas de Associação no R

library(DescTools)

- coeficiente de contingência: ContCoef()
- coeficiente V de Crámer: CramerV()
- coeficiente τ_b de Kendall: KendallTauB()

Teste de independência do Qui-Quadrado

Exemplo 2

Com o objetivo de tentar "explicar as causas" do insucesso escolar foram inquiridos vários alunos do ensino básico. Aos alunos foram colocadas diversas questões, entre as quais uma sobre o número de reprovações e outra sobre o número de faltas (ficheiro dados2.txt). As variáveis em análise e a respetiva tabela de contingência são:

Número de reprovações - com as categorias:

- **Nenhuma**
- **Uma**
- **Duas ou mais**

Número de faltas - com as categorias:

- **Nenhuma**
- **Algumas**
- **Muitas**

	Número de faltas		
Número de reprovações	**Nenhuma**	**Algumas**	**Muitas**
Nenhuma	132	57	13
Uma	28	15	15
Duas ou mais	20	17	17

Como rejeitámos a hipótese das variáveis "Número de faltas" e "Número de reprovações" serem independentes, então interessa saber como é a associação.

R

da library(DescTools) usar as funções:

- ContCoef()
- CramerV()
- KendallTauB()

  Observ: se não funcionar, converter a tabela numa matriz: as.matrix()

e obtém-se

- coeficiente de contingência = 0.3064
- coeficiente V de Crámer = 0.2276
- coeficiente τ_b de Kendall= 0.2544

- A associação existente pode ser considerada moderada mas perto de fraca: o coeficiente de contingência $= 0.3064 \in [0.30, 0.50[$ e V de Crámer $= 0.2276 \in [0.20, 0.35[$.
- A associação existente é positiva, pois o valor do tau-b de Kendall é positivo, $\tau_b = 0.2544 > 0$, mas não parece ser forte pois está afastado de 1.
- A associação ser positiva significa que quando aumenta o número de faltas parece aumentar o número de reprovações.