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

2021-2022

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 já vimos, 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 → resumir os dados
  - tabelas de dupla entrada: **tabelas de contingência** também chamadas de tabelas de informação cruzada;
- etapa 2 → 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. 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.

Formulação das Hipóteses a Testar:

$$H_0 - \text{ Não há relação entre as variáveis}$$
$$vs$$
$$H_1 - \text{ Há relação entre as variáveis}$$

ou de forma equivalente

$$H_0 - \text{ As variáveis são independentes}$$
$$vs$$
$$H_1 - \text{ As variáveis não são independentes}$$

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}). Supondo verdadeira a hipótese H_0, então

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

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

Observação:
Recordar das probabilidades: os acontecimentos A e B dizem-se independentes sse

$$P(A \cap B) = P(A) \times P(B)$$

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}	$n_{1\cdot}$
$\mathbf{A_2}$	O_{21}	O_{22}	$\cdots$	O_{2c}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$\mathbf{A_r}$	O_{r1}	O_{r2}	$\cdots$	O_{rc}	$n_{r\cdot}$
TOTAL	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot c}$	n

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

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

Teste de independência do Qui-Quadrado

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

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

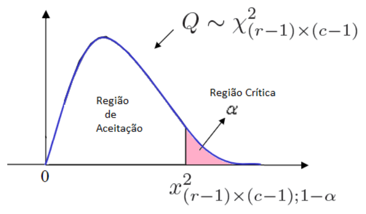
- r corresponde ao número de linhas da tabela de contingência
- c corresponde ao número de colunas da tabela de contingência
- **frequências observadas** $= O_{ij} \rightarrow$ corresponde às frequências observadas (amostra) da tabelas de contingência;
- **frequências esperadas** $= E_{ij} = \frac{n_{i.}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

Observação: 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

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

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,

$$Q_{obs} \in RC$$

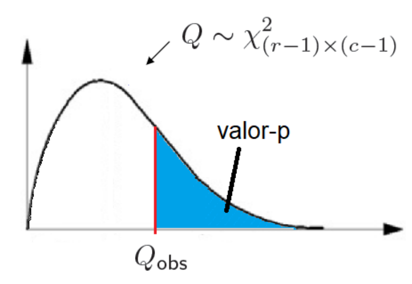
então, ao nível de significância α, **a hipótese H_0 é rejeitada**, isto é, com base na amostra não há evidências estatísticas que as variáveis são independentes.

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 $j = 1, \ldots, c$.

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. 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

$Q_{obs} = 0.0154$ e $RC = \left[x^2_{(r-1)\times(c-1);1-\alpha}, +\infty\right[= \left[x^2_{(1);0.95}, +\infty\right[= [3.84, +\infty[$

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

Regra de Decisão através do valor-p

valor-$p = P(Q \geq Q_{obs}) = P(Q \geq 0.0154) = 1 - P(Q < 0.0154) = 1 - F(0.0154)$

<u>R</u>: valor-$p = 1 - F(0.0154) = 1 - 0.0988 = 0.9012$

<u>Tabela em papel</u>: valor-$p = 1 - F(0.0154) \approx 1 - F(0.0158) = 1 - 0.10 = 0.90$

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 $chisq.test()$

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. 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	3
Uma	18	4	4
Duas ou mais	10	5	5

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

$$vs$$

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	3	192
Uma	18	4	4	26
Duas ou mais	10	5	5	20
TOTAL	160	66	12	238

- Tabela de contingência das frequências Esperadas:

	Número de faltas			
Número de reprovações	**Nenhuma**	**Algumas**	**Muitas**	**TOTAL**
Nenhuma	$129.0756 = \frac{160\times192}{238}$	$53.2437 = \frac{66\times192}{238}$	$9.6807 = \frac{12\times192}{238}$	192
Uma	$17.4790 = \frac{160\times26}{238}$	$7.2101 = \frac{66\times26}{238}$	$1.3109 = \frac{12\times26}{238}$	26
Duas ou mais	$13.4454 = \frac{160\times20}{238}$	$5.5462 = \frac{66\times20}{238}$	$1.0084 = \frac{12\times20}{238}$	20
TOTAL	160	66	12	238

- Estatística de teste:

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

$$= \frac{(132 - 129.0756)^2}{129.0756} + \frac{(57 - 53.2437)^2}{53.2437} + \cdots + \frac{(5 - 5.5462)^2}{5.5462} + \frac{(5 - 1.0084)^2}{1.0084} = 28.639$$

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

$$Q_{obs} = 28.639 \qquad \text{e} \qquad RC = \left[x^2_{(r-1)\times(c-1);1-\alpha}, +\infty\right[= \left[x^2_{(4);0.99}, +\infty\right[= [13.3, +\infty[$$

Como $Q_{obs} = 28.639 \in RC$ então rejeita-se a hipótese H_0

Regra de Decisão através do valor-p

$$\begin{aligned}\text{valor-}p &= P(Q \geq Q_{obs}) = P(Q \geq 28.639) = 1 - P(Q < 28.639) = \\ &= 1 - F(28.639) = 1 - 1 = 0\end{aligned}$$

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

- $Q_{obs} = 28.639$
- graus de liberdade $= 4$
- valor-$p = 9.254e - 06 = 0.000009254$

Como valor-$p = 0.000009254 \leq 0.01 = \alpha$ 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).