UNIVERSIDADE FEDERAL DA FRONTEIRA SUL
Campus CERRO LARGO

PROJETO DE EXTENSÃO
Software R:
Capacitação em análise estatística de dados utilizando um software livre.

Fonte: `https://www.r-project.org/`

Módulo IV
Teste Qui-Quadrado

Ministrante: Tatiane Chassot

Blog do projeto: `https://softwarelivrer.wordpress.com/equipe/`

Equipe:
Coordenadora:
Profe. Iara Endruweit Battisti (iara.battisti@uffs.edu.br)
Colaboradores:
Profa. Denize Reis
Prof. Erikson Kaszubowski
Prof. Reneo Prediger
Profa. Tatiane Chassot
Mestrando Felipe Smolski
Bolsista:
Djaina Rieger - aluna de Engenharia Ambiental (djaina.rieger@outlook.com)

Sumário

1 Teste Qui-quadrado para verificar associação entre duas variáveis qualitativas

O teste do Qui-quadrado é usado quando se quer comparar frequências observadas com frequências esperadas.

Requisitos:

- os dados amostrais devem ser selecionados **aleatoriamente** e são representados como **contagens de frequências** em tabela de dupla entrada.
- em toda célula da tabela, a frequência esperada (E) deve ser, no mínimo, 5.
- não há exigência quanto à frequência observada (O).
- não há exigência de que a população deva ter distribuição normal. Por isso, o teste Qui-quadrado é classificado como um teste não-paramétrico.

A tabela de dupla entrada ou tabela de contingência (Tabela 1) é utilizada para representar as frequências observadas. Cada célula ou casela da tabela de dupla entrada é usada para representar uma frequência observada (Oij), onde i corresponde a linha e j corresponde a coluna.

Exemplo 1 - Num estudo da associação entre a ocorrência de tromboembolismo e grupo sanguíneo, 200 mulheres usuárias de contraceptivo oral foram classificadas quanto à presença de tromboembolismo (doente ou sadia) e quanto ao grupo sanguíneo (A, B, AB ou O). Os resultados dessa classificação foram reproduzidos na Tabela 1.

Tabela 1 – Classificação de 200 mulheres usuárias de contraceptivo oral quanto à presença de tromboembolismo e quanto ao grupo sanguíneo.

Grupo sanguíneo	Tromboembolismo		Total
	Doente	Sadia	
A	32	47	79
B	8	19	27
AB	7	14	21
O	9	64	73
Total	56	144	200

Existem evidências estatísticas suficientes nesses dados para verificar a hipótese de que a presença do tromboembolismo e o grupo sanguíneo estejam associados?

- Formular as hipóteses H0 e H1:

H0 = as variáveis são independentes (não existe associação entre grupo sanguíneo e presença de tromboembolismo)

H1 = as variáveis não são independentes (existe associação entre grupo sanguíneo e presença de tromboembolismo)

- Sintaxe no software RStudio:

Digitar os dados da tabela cruzada (tabela de contingência) no formato de uma matriz, valor ij, considerando i=linha e j=coluna, em sequência por coluna (por exemplo, digita-se todos os valores da primeira coluna, depois digita-se todos os valores da segunda coluna e assim sucessivamente).

```
tromboembolismo=matrix(c(32,8,7,9,47,19,14,64),nc=2)

# Em Que:
# -matrix = indica que os dados serão organizados em uma matriz
# -nc = número de colunas
```

Para visualizar a matrix no RStudio:

```
tromboembolismo

##      [,1] [,2]
## [1,]   32   47
## [2,]    8   19
## [3,]    7   14
## [4,]    9   64
```

Primeiramente, deve-se verificar a existência de alguma casela com frequência esperada menor que 5.

```
chisq.test(tromboembolismo)$expected

##       [,1]  [,2]
## [1,] 22.12 56.88
## [2,]  7.56 19.44
## [3,]  5.88 15.12
## [4,] 20.44 52.56
```

Caso não exista, utiliza-se o teste qui-quadrado com o comando chisq.test:

```
chisq.test(tromboembolismo)

##
##  Pearson's Chi-squared test
##
## data:  tromboembolismo
## X-squared = 15.354, df = 3, p-value = 0.001538
```

Conclusão:

Como p<0,01, rejeita-se H0 com nível de significância de 1% e conclui-se que as variáveis não são independentes (existe associação entre grupo sanguíneo e presença de tromboembolismo).

Em caso contrário, utiliza-se o teste **Exato de Fisher** (ou seja, se existir casela com frequência esperada menor que 5).

Exemplo 2 - Um psicólogo submeteu um grupo de pacientes a um teste, ministrando sonífero a um grupo e pílulas de farinha (placebo) a outro grupo. Perguntado aos pacientes se o medicamento ajudou ou não a dormir melhor, as respostas foram as seguintes:

Tabela 2 – Levantamento em pacientes que tomaram sonífero e placebo e relação à qualidade do sono.

Pílulas	Dormiram melhor	Dormiram pior	Indiferente
Sonífero	10	2	10
Placebo	5	7	5

Testar a hipótese de não haver diferença entre o fato do doente tomar sonífero e dormir melhor.

H0 = as variáveis são independentes (não existe associação entre pílulas e qualidade do sono)
H1 = as variáveis não são independentes (existe associação entre pílulas e qualidade do sono)

```
medicamento=matrix(c(10,5,2,7,10,5),nc=3)
medicamento
```

```
##      [,1] [,2] [,3]
## [1,]   10    2   10
## [2,]    5    7    5
```

```
chisq.test(medicamento)$expected
```

```
## Warning in chisq.test(medicamento): Chi-squared approximation may be incorrect
```

```
##          [,1]     [,2]     [,3]
## [1,] 8.461538 5.076923 8.461538
## [2,] 6.538462 3.923077 6.538462
```

```
fisher.test(medicamento)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  medicamento
## p-value = 0.09433
## alternative hypothesis: two.sided
```

Conclusão:

Como p>0,05, não rejeita-se HO e conclui-se que as variáveis são independentes (não existe associação entre pílulas e qualidade do sono).

Caso a **tabela** seja **2x2**, então usa-se o teste de qui-quadrado com o comando 'chisq.test' acrescido de 'correct=TRUE' indicando a utilização da correção de continuidade **(correção de Yates). No entanto, no RStudio, esta correção já é feita automaticamente.**

Exemplo 3 – Com o objetivo de verificar se existe associação entre prática de esportes e tabagismo, foram entrevistados jovens e as respostas estão apresentadas abaixo:

Tabela 3 - Associação entre prática de esportes e tabagismo.

Prática de esportes	Tabagismo	
	Presente	Ausente
Presente	50	15
Ausente	10	25

H0 = as variáveis são independentes (não existe associação entre prática de esportes e tabagismo)
H1 = as variáveis não são independentes (existe associação entre prática de esportes e tabagismo)

```
tabagismo=matrix(c(50,10,15,25),nc=2)
tabagismo

##      [,1] [,2]
## [1,]   50   15
## [2,]   10   25

chisq.test(tabagismo)$expected

##      [,1] [,2]
## [1,]   39   26
## [2,]   21   14

chisq.test(tabagismo)

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tabagismo
## X-squared = 20.192, df = 1, p-value = 7.003e-06

chisq.test(tabagismo,correct-TRUE)

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tabagismo
## X-squared = 20.192, df = 1, p-value = 7.003e-06
```

Conclusão:

Como p<0,01, rejeita-se H0 com nível de significância de 1% e conclui-se que as variáveis não são independentes (existe associação entre prática de esportes e tabagismo).

No caso de amostras **pareadas** (dependentes), utiliza-se o teste de McNemar para testar a associação.

Importante para o teste de McNemar: no software R os dados na matriz (tabela de contingência) devem ser distribuídos da mesma maneira tanto nas linhas quanto nas colunas. Isto é,

a e d devem expressar o mesmo comportamento. Por exemplo: aprovado, desaprovado, aprovado, desaprovado.

Antes	Depois	
	Aprovado	Desaprovado
Aprovado	a	b
Desaprovado	c	d

Exemplo 4 - Uma pesquisa foi realizada para verificar o efeito de uma propaganda sobre a satisfação de um produto. Para isso foram selecionados aleatoriamente 76 indivíduos com características semelhantes para avaliarem o produto antes e após a propaganda. Abaixo é apresentada a satisfação dos indivíduos pesquisados. Teste a hipótese de que existe diferença na satisfação antes e após a propaganda.

Antes	Depois	
	MS ou S	+-, I ou MI
MS ou S	34	2
+-, I ou MI	25	15

Em que:

MS = muito satisfeito
S = satisfeito
+- = mais ou menos satisfeito
MI = muito insatisfeito
I = insatisfeito

Hipóteses:

H0 = as frequências das diferentes categorias ocorrem na mesma proporção, as mudanças não são significativas.
H1 = as frequências das diferentes categorias ocorrem em proporções diferentes, as mudanças são significativas.

```
propaganda=matrix(c(34,25,2,15),nc=2)
propaganda

##      [,1] [,2]
## [1,]   34    2
## [2,]   25   15

chisq.test(propaganda)$expected

##          [,1]     [,2]
## [1,] 27.94737 8.052632
## [2,] 31.05263 8.947368

mcnemar.test(propaganda)
```

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  propaganda
## McNemar's chi-squared = 17.926, df = 1, p-value = 2.297e-05
```

Conclusão:

Como p<0,01, rejeita-se H0 com nível de significância de 1% e conclui-se que as frequências das diferentes categorias ocorrem em proporções diferentes, as mudanças são significativas.

2 Teste Qui-quadrado para verificar aderência a uma distribuição

Para verificar se o conjunto de dados segue uma distribuição teórica específica.

Exemplo 5 - Uma bibliotecária observa que o número de livros emprestados durante certa semana foi:

Dias da semana	Segunda	Terça	Quarta	Quinta	Sexta	Sábado	Domingo
Nº livros emprestados:	15	12	16	14	19	30	34

Use esta amostra para testar se o número de livros retirados é mesmo durante os dias da semana.

H0: os livros circulam com a mesma frequência em cada dia da semana.
H1: os livros não circulam com a mesma frequência em cada dia da semana.

```
livros=c(15,12,16,14,19,30,34)
livros

## [1] 15 12 16 14 19 30 34

chisq.test(livros)$expected

## [1] 20 20 20 20 20 20 20

chisq.test(livros)

##
##  Chi-squared test for given probabilities
##
## data:  livros
## X-squared = 21.9, df = 6, p-value = 0.001262
```

Conclusão:

Como p<0,01, rejeita-se H0 com nível de significância de 1% e conclui-se que os livros não circulam com a mesma frequência em cada dia da semana.

Ou para verificar se as frequências de um conjunto de dados segue uma **distribuição específica**, informado em 'dist'. Lembrando que os valores no vetor 'dist' devem estar no formato de proporção.

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

Exemplo 6 - Segundo Mendel (geneticista famoso), os resultados dos cruzamentos de ervilhas amarelas redondas com ervilhas verdes enrugadas 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
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?

H0: os dados se justam ao modelo de probabilidade proposto por Mendel
H1: os dados não se justam ao modelo de probabilidade proposto por Mendel

```
ervilhas=c(315,101,108,32)
dist=c(0.5625,0.1875,0.1875,0.0625)
chisq.test(ervilhas)$expected

## [1] 139 139 139 139

chisq.test(ervilhas,p=dist)

##
##  Chi-squared test for given probabilities
##
## data:  ervilhas
## X-squared = 0.47002, df = 3, p-value = 0.9254
```

Conclusão:

Como p>0,05, não rejeita-se H0 e conclui-se que os dados se justam ao modelo de probabilidade proposto por Mendel.