

# Estatística aplicada

Lino Costa

Departamento de Produção e Sistemas  
Escola de Engenharia  
lac@dps.uminho.pt

Ano letivo 2018/2019

# Sumário

1. dados categoriais
2. probabilidades completamente especificadas
  - hipóteses
  - frequências esperadas
  - estatística de qui-quadrado
3. probabilidades não completamente especificadas
  - estimação de parâmetros
  - hipóteses
  - frequências esperadas
  - estatística de qui-quadrado

# Testes de bom ajuste

## Dados categoriais

Uma variável categórica é usada para representar um conjunto de categorias. Dois tipos de variáveis categóricas podem ser definidos:

- **variável nominal** - a medição apenas define a que classe a unidade pertence, em relação àquela propriedade (e.g., sexo)
- **variável ordinal** - a medição também esclarece quando uma unidade tem mais da propriedade do que outra unidade (e.g., opinião)

|            |       |       |     |       |
|------------|-------|-------|-----|-------|
| Classe     | 1     | 2     | ... | $k$   |
| Frequência | $f_1$ | $f_2$ | ... | $f_k$ |

## Teste de bom ajuste

A distribuição da variável categórica é desconhecida. O objetivo do teste de bom ajuste é testar se uma determinada distribuição se ajusta à população com base nos dados de uma grande amostra ( $n \geq 30$ ).

# Testes de bom ajuste

## Probabilidades completamente especificadas

A distribuição está completamente especificada na hipótese nula para as  $k$  classes, sem haver necessidade de estimar parâmetros.

$$\begin{aligned} H_0 : p_1 &= p_{10}, p_2 = p_{20}, \dots, p_k = p_{k0} & (p_{10} + p_{20} + \dots + p_{k0} &= 1) \\ H_1 : p_i &\neq p_{i0} & \exists i \in \{1, \dots, k\} \end{aligned}$$

## Estatística de teste

$$E.T. : Q = \sum_{i=1}^k \frac{(f_i - e_i)^2}{e_i} \text{ onde } f_i \text{ são as frequências observadas e } e_i = np_i$$

são as frequências esperadas. Se a frequência esperada de uma classe  $i$  for pequena ( $e_i < 5$ ), a classe deve ser agrupada com a classe adjacente e o número de classes  $k$  reduzido de uma unidade.

## Região de rejeição

$$R.R. : Q > \chi_{\alpha, k-1}^2 \text{ onde } \alpha \text{ é o nível de significância.}$$

# Testes de bom ajuste

## Exemplo 1

Em 2003, o número de AVCs masculinos no concelho de Braga foram os reportados na tabela, de acordo com a estação do ano. Teste se as probabilidades de ocorrência de AVCs é idêntica nas quatro estações do ano ( $\alpha = 0.05$ ).

| Estação do ano | Primavera | Verão | Outono | Inverno |
|----------------|-----------|-------|--------|---------|
| Número de AVCs | 64        | 81    | 39     | 28      |

- Estação do ano é uma variável categórica nominal com  $k = 4$  classes
- Teste do bom ajuste para testar se as probabilidades de AVCs são idênticas (probabilidades completamente especificadas), i.e.,

$$H_0 : p_1 = 1/4, p_2 = 1/4, p_3 = 1/4, p_4 = 1/4 \quad H_1 : \neg H_0$$

| Estação do ano              | Primavera | Verão | Outono | Inverno | $\Sigma$    |
|-----------------------------|-----------|-------|--------|---------|-------------|
| $f_i$                       | 64        | 81    | 39     | 28      | $212 = n$   |
| $p_i$                       | $1/4$     | $1/4$ | $1/4$  | $1/4$   | 1           |
| $e_i = np_i$                | 53        | 53    | 53     | 53      | 212         |
| $\frac{(f_i - e_i)^2}{e_i}$ | 2.28      | 14.79 | 3.70   | 11.79   | $32.56 = Q$ |

- $E.T. : Q = \sum_{i=1}^4 \frac{(f_i - e_i)^2}{e_i} = 32.56$
- $R.R. : Q > \chi_{0.05,3}^2 \Leftrightarrow Q > 7.81$  (Tabela 7)
- Rejeita-se  $H_0$  para  $\alpha = 0.05$ , pelo que a probabilidade de AVCs não é idêntica nas quatro estações do ano.

# Testes de bom ajuste

## Probabilidades não completamente especificadas

A distribuição não está completamente especificada na hipótese nula para as  $k$  classes, havendo a necessidade de estimar parâmetros com base nos dados.

$H_0$  : as probabilidades das classes provêm de uma distribuição da família...

$H_1$  : as probabilidades das classes não provêm de uma distribuição da família...

## Estatística de teste

$E.T. : Q = \sum_{i=1}^k \frac{(f_i - e_i)^2}{e_i}$  onde  $f_i$  são as frequências observadas e  $e_i = np_i$

são as frequências esperadas. Se a frequência esperada de uma classe  $i$  for pequena ( $e_i < 5$ ), a classe deve ser agrupada com a classe adjacente e o número de classes  $k$  reduzido de uma unidade.

## Região de rejeição

$R.R. : Q > \chi_{\alpha, gl}^2$  onde  $gl = k - 1 - \text{número de parâmetros estimados}$  e  $\alpha$  é o nível de significância.

# Testes de bom ajuste

## Exemplo 2

Julga-se que o número de emails por hora ( $X$ ) que chegam a uma instituição segue uma distribuição de Poisson. Os seguintes dados foram obtidos durante 100 horas. Teste se o número de emails por hora que chegam à instituição segue uma distribuição de Poisson ( $\alpha = 0.05$ ).

|                               |    |    |   |   |
|-------------------------------|----|----|---|---|
| Número de emails/hora ( $X$ ) | 0  | 1  | 2 | 3 |
| Frequência                    | 60 | 28 | 7 | 5 |

- Número de emails/hora ( $X$ ) é uma variável categórica ordinal com  $k = 4$  classes
- Teste do bom ajuste para testar  $X$  segue uma distribuição de Poisson sem  $\lambda$  especificado (probabilidades não completamente especificadas), i.e.,

$H_0$  : as probabilidades das classes provém de uma distribuição de Poisson     $H_1$  :  $\neg H_0$

- estimar parâmetro  $\lambda$ :  $\hat{\lambda} = \bar{x} = \frac{0 \times 60 + 1 \times 28 + 2 \times 7 + 3 \times 5}{100} = 0.57$
- logo  $X \sim P(0.57)$  e  $p_1 = P(X = 0) = \frac{e^{-0.57} 0.57^0}{0!} = 0.57$ ,  
 $p_2 = P(X = 1) = 0.32$ ,  $p_3 = P(X = 2) = 0.09$  e  
 $p_4 = P(X \geq 3) = 1 - (0.57 + 0.32 + 0.09) = 0.02$

|                               |      |      |      |           |           |
|-------------------------------|------|------|------|-----------|-----------|
| Número de emails/hora ( $X$ ) | 0    | 1    | 2    | 3 ou mais | $\Sigma$  |
| $f_i$                         | 60   | 28   | 7    | 5         | $100 = n$ |
| $p_i$                         | 0.57 | 0.32 | 0.09 | 0.02      | 1         |
| $e_i = np_i$                  | 57   | 32   | 9    | 2         | 100       |

- como  $e_4 = 2 < 5$  tem de se agrupar classes

# Testes de bom ajuste

## Exemplo 2

| Número de emails/hora ( $X$ ) | 0    | 1    | 2 ou mais | $\sum$     |
|-------------------------------|------|------|-----------|------------|
| $f_i$                         | 60   | 28   | 12        | $100 = n$  |
| $p_i$                         | 0.57 | 0.32 | 0.11      | 1          |
| $e_i = np_i$                  | 57   | 32   | 11        | 100        |
| $\frac{(f_i - e_i)^2}{e_i}$   | 0.16 | 0.50 | 0.09      | $0.75 = Q$ |

- com o agrupamento, passa-se a ter  $k = 3$
- $E.T. : Q = \sum_{i=1}^3 \frac{(f_i - e_i)^2}{e_i} = 0.75$
- $R.R. : Q > \chi^2_{\alpha, gl}$  com  
 $gl = k - 1 - \text{número de parâmetros estimados} = 3 - 1 - 1 = 1$ , logo  
 $Q > \chi^2_{0.05, 1} \Leftrightarrow Q > 3.84$  (Tabela 7)
- Não se rejeita  $H_0$  para  $\alpha = 0.05$ , pelo que o número de emails por hora que chegam à instituição poderá seguir uma distribuição de Poisson.



# Testes de bom ajuste

## Exemplo 3

Recolheu-se uma amostra aleatória de 100 operários de uma grande empresa, tendo-se obtido para o vencimento por hora uma média e um desvio padrão de 132 e 5, respetivamente. Pretende-se testar se o vencimento por hora segue uma distribuição normal ( $\alpha = 0.05$ ) tendo em conta os dados da seguinte tabela:

| Vencimento         | Nº de operários |
|--------------------|-----------------|
| $x < 125$          | 10              |
| $125 \leq x < 130$ | 20              |
| $130 \leq x < 135$ | 38              |
| $135 \leq x < 140$ | 25              |
| $x \geq 140$       | 7               |

- Vencimento por hora ( $X$ ) é uma variável agrupada em  $k = 5$  classes
- Teste do bom ajuste para testar  $X$  segue uma distribuição normal sem  $\mu$  e  $\sigma$  especificados (probabilidades não completamente especificadas), i.e.,  
 $H_0$  : as probabilidades das classes provêm de uma distribuição normal     $H_1$  :  $\neg H_0$
- estimativas dos parâmetros  $\mu$  e  $\sigma$ :  $\hat{\mu} = \bar{x} = 132$  e  $\hat{\sigma} = s = 5$ , logo  
 $X \sim N(132, 5^2)$  e  $Z = \frac{X-132}{5} \sim N(0, 1)$
- $p_1 = P(X < 125) = P(Z < -1.4) = 0.0808$ ,  
 $p_2 = P(125 \leq X < 130) = P(Z < -0.4) - P(Z < -1.4) = 0.2638$ ,  
 $p_3 = P(130 \leq X < 135) = P(Z < 0.6) - P(Z < -0.4) = 0.3811$ ,  
 $p_4 = P(135 \leq X < 140) = P(Z < 1.6) - P(Z < 0.6) = 0.2195$  e  
 $p_5 = P(X \geq 140) = P(Z \geq 1.6) = 1 - P(Z < 1.6) = 0.0548$  (Tabela 5)

# Testes de bom ajuste

## Exemplo 3

| Vencimento         | $f_i$     | $p_i$  | $e_i = np_i$ | $\frac{(f_i - e_i)^2}{e_i}$ |
|--------------------|-----------|--------|--------------|-----------------------------|
| $x < 125$          | 10        | 0.0808 | 8.080        | 0.4562                      |
| $125 \leq x < 130$ | 20        | 0.2638 | 26.38        | 1.5430                      |
| $130 \leq x < 135$ | 38        | 0.3811 | 38.11        | 0.0003                      |
| $135 \leq x < 140$ | 25        | 0.2195 | 21.95        | 0.4238                      |
| $x \geq 140$       | 7         | 0.0548 | 5.480        | 0.4216                      |
| $\sum$             | $n = 100$ | 1      | 100          | $Q = 2.845$                 |

- $E.T. : Q = \sum_{i=1}^5 \frac{(f_i - e_i)^2}{e_i} = 2.845$
- $R.R. : Q > \chi_{\alpha, gl}^2$  com  
 $gl = k - 1 - \text{número de parâmetros estimados} = 5 - 1 - 2 = 2$ , logo  
 $Q > \chi_{0.05, 2}^2 \Leftrightarrow Q > 5.99$  (Tabela 7)
- Não se rejeita  $H_0$  para  $\alpha = 0.05$ , pelo que o vencimento por hora poderá seguir uma distribuição normal.