

BIOESTATÍSTICA

Parte 6 – Métodos não Paramétricos

Aulas Teóricas de 19/05/2011 a 26/05/2011

6.1. O teste do qui-quadrado como teste de ajustamento

[Testes de ajustamento:

⇒ avaliam se a população se ajusta a um determinado modelo;

⇒ podem ser aplicados a dados *qualitativos* ou *quantitativos*];

Dados Qualitativos

Suponhamos que uma população é caracterizada por um atributo qualitativo, estando então a população dividida em $k \geq 2$ categorias disjuntas e exaustivas.

Para $i = 1, \dots, k$ seja p_i a proporção (desconhecida) de elementos da população pertencentes à categoria A_i , tendo-se $\sum_{i=1}^k p_i = 1$. Considere-se então o modelo probabilístico

Categoria	A_1	\dots	A_i	\dots	A_k
Probabilidade	p_1	\dots	p_i	\dots	p_k

Pretendemos testar

H_0 : o modelo ajusta-se aos dados r 's H_1 : o modelo não se ajusta

Recolhe-se da população em estudo uma amostra de dimensão n .

Seja O_i o número de observações pertencentes à categoria A_i , ou seja, a frequência observada da classe A_i , para $i = 1, \dots, k$.

A frequência esperada da classe A_i , se H_0 for verdadeira, é dada por $e_i = np_i$.

Estatística de teste: é obtida à custa das diferenças entre a frequência observada de cada classe e a respectiva frequência esperada, sob H_0 .

$$X^2 = \sum_{i=1}^k \frac{(O_i - e_i)^2}{e_i}$$

Se H_0 for verdadeira, X^2 tem distribuição assintótica de Qui-quadrado com $k - 1$ graus de liberdade.

- Quanto maior a diferença entre os valores observados e esperados, maior X^2 , e maior a evidência de que o modelo considerado não se ajusta aos dados.
- X^2 tem distribuição assintótica de Qui-quadrado se H_0 for verdadeira:
 - Distribuição assintótica significa aproximada;

◦ O teste só é válido para grandes amostras
Se a amostra for pequena utiliza-se a aproximação de Welch-Satterthwaite

Quanto maior for o valor observado de X^2 , menos plausível é a hipótese nula e portanto rejeita-se H_0 para valores grandes da estatística.

Para o nível de significância α , rejeita-se H_0 se $X^2 > \chi_{k-1, 1-\alpha}^2$.

[k - número de classes]

Para uma correcta aplicação do teste, exige-se que $e_i \geq 5$, para qualquer i . Caso tal não aconteça, as classes correspondentes devem agrupar-se com as adjacentes, sendo o número de graus de liberdade reduzido em conformidade.

Não é consensual

Dados Quantitativos

[discretos ou contínuos]

Dada a amostra observada (x_1, \dots, x_n) de dados quantitativos, proveniente de uma população X , pretende-se testar a hipótese de que X segue uma determinada distribuição.

H_0 : X segue um certo modelo paramétrico r 's

H_1 : X não segue esse modelo

[Exemplos de modelos paramétricos: são modelos que dependem de parâmetros

- Binomial (n, p)
- Poisson (λ)
- Normal (μ, σ)

- Os parâmetros podem ser ou não desconhecidos.]

Considerem-se então classes A_1, \dots, A_k que constituam uma partição do domínio de variação da v.a. X . Seja O_i a frequência observada da classe A_i , para $i = 1, \dots, k$. Supondo H_0 verdadeira, calculemos $p_i = P(X \in A_i)$.

A frequência esperada da classe A_i , sob a validade de H_0 , é dada por $c_i = np_i$.

[partição do domínio de variação: a soma das probabilidades das partições deve ser 1]

Exemplo: Número de acidentes num cruzamento por mês num ano.

Dados:

Acidentes por mês	0	1	2	3
Número de meses				

Aplica-se a hipótese de Poisson?

Domínio de variação do parâmetro λ de Poisson: $[0, +\infty[$

Classes utilizadas: $n=0$; $n=1$; $n=2$; $n \geq 3$.

p_i : $P(X=0)$; $P(X=1)$; $P(X=2)$; $P(X \geq 3)$

$P(X \geq 3) = 1 - P(X < 3) = 1 - P(X=0) - P(X=1) - P(X=2)$]

Estatística de teste:

$$X^2 = \sum_{i=1}^k \frac{(O_i - c_i)^2}{c_i}$$

Sob a validade de H_0

☐ se o modelo está completamente especificado (não há parâmetros desconhecidos) então X^2 tem distribuição assintótica χ^2_{k-1}

Ao nível α , rejeitar H_0 se $X^2 > \chi^2_{k-1, 1-\alpha}$.

☐ se o modelo contém r parâmetros desconhecidos (que são estimados pelo método da máxima verosimilhança) então X^2 tem distribuição assintótica χ^2_{k-r-1}

Ao nível α , rejeitar H_0 se $X^2 > \chi^2_{k-r-1, 1-\alpha}$.

[Parâmetros Desconhecidos - Método da Máxima Verosimilhança:

$$X \sim N(\mu, \sigma)$$

$$\lambda = E(X) \Rightarrow \hat{\lambda} = \bar{x}$$

$$X \sim N(\mu, \sigma)$$

$$\hat{\mu} = \bar{x} \quad \text{e} \quad \hat{\sigma} = s$$

Exercício: Segundo a teoria de Mendel a distribuição de flores, segundo as suas cores, na segunda geração é:

- Cor amarela: $\frac{3}{16} \rightarrow p_1$
- Cor branca: $\frac{4}{16} \rightarrow p_2$
- Cor vermelha: $\frac{9}{16} \rightarrow p_3$

Recolheram-se 145 flores e observaram-se 35 flores amarelas, 38 flores brancas e 72 flores vermelhas. Os resultados são compatíveis com a Teoria de Mendel, ao nível de 1% de significância?

$$H_0: p_1 = \frac{3}{16} \quad \wedge \quad p_2 = \frac{4}{16} \quad \wedge \quad p_3 = \frac{9}{16}$$

vs

H_1 : pelo menos uma é diferente

Testes de Homogeneidade

Suponhamos que temos $p \geq 2$ amostras independentes, provenientes das p populações X_1, \dots, X_p . Seja n_i a dimensão da amostra proveniente da população X_i .

Pretende-se testar as hipóteses

H_0 : as populações são homogêneas vs

H_1 : pelo menos uma fornece observações maiores/menores

Consideremos uma partição da recta real em k classes

A_1, \dots, A_k .

Seja n_{ij} o número de elementos da amostra proveniente da população X_i que pertencem à classe A_j . Vamos construir a seguinte tabela de contingência:

	A_1	...	A_k
X_1	n_{11}	...	n_{1k}
\vdots	\vdots		\vdots
X_p	n_{p1}	...	n_{pk}
	$n_{.1}$...	$n_{.k}$
	$n_{.1}$...	$n_{.k}$

A tabela pode ser "rodada" e as margens trocadas de posição

A margem vertical da tabela é fixa, constituída pelas dimensões das várias amostras.

n_{ij} é a frequência observada da célula (i, j) da tabela. Supondo H_0 verdadeira, a respectiva frequência esperada é dada por

$$e_{ij} = n_i \times \frac{n_{.j}}{n}$$

onde $n_{.j}$ é o número total de elementos na categoria A_j .

[Em testes de homogeneidade uma das margens da tabela é fixa à partida, ou seja, há grupos em que o tamanho da amostra é conhecido à partida. Os valores correspondentes à outra variável resultam de uma contagem]

Estatística de teste:

$$\chi^2 = \sum_{i=1}^p \sum_{j=1}^k \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

	Cor amarela	Cor branca	Cor vermelha	Total
O_i	35	38	72	145
e_i	27,19	36,5	81,56	145
$O_i - e_i$	7,5	1,5	-9,56	-
$(O_i - e_i)^2 / e_i$	2,0688	0,0616	1,1201	3,25

$$e_i = n \cdot p_i$$

Rejeito H_0 se $\chi^2 > \chi^2_{k-1; 1-\alpha}$

$$\alpha = 0,01\%$$

$$\chi^2_{2; 0,99} = 9,21$$

$$\chi^2 = 3,25$$

Conclusão: não se rejeita H_0 , logo há evidência de que os resultados são compatíveis com a teoria de Mendel

No caso discreto, as classes coincidem com os pontos a_i tais que $P(X = a_i) > 0$, alguns dos quais podem ser agrupados.

No caso contínuo, procede-se ao agrupamento dos dados originais em classes $A_i = [a_{i-1}, a_i]$, o que dá origem a perda de informação.

Pode ser necessário agrupar classes adjacentes de modo que $c_i \geq 5$, para $i = 1, \dots, k$.

O teste do Qui-quadrado apenas pode ser utilizado para grandes amostras.

6.1. Tabelas de Contingência: o teste do Qui-quadrado como teste de independência e teste de homogeneidade

[Tabelas de contingência:

	A	B	C	Total
M				$n_{M.}$
N				$n_{N.}$
K				$n_{K.}$
Total	$n_{.A}$	$n_{.B}$	$n_{.C}$	n

Exemplo: (faltam as somas)

Grupo	Situação do Doente Após o Tratamento	
	Curado	Não Curado
Grupo A	75	25
Grupo B	65	35

Se H_0 verdadeira, X^2 tem distribuição assintótica $\chi^2_{(p-1)(k-1)}$

Ao nível de significância α , rejeitar H_0 se $X^2 > \chi^2_{(p-1)(k-1), 1-\alpha}$

Regra prática: no máximo, 20% das células podem ter frequência esperada < 5 e nenhuma célula deve ter frequência esperada < 1 .

[distribuição assintótica \Rightarrow só é uma boa aproximação para grandes amostras]

Exemplo: Pretende-se estudar se há relação entre a poluição atmosférica e as doenças pulmonares. Para tal, escolheram-se duas cidades com problemas de poluição e uma área rural pouco poluída. Estudaram-se 400 adultos em cada área, tendo-se obtido os seguintes resultados:

Área	Nº de indivíduos c/ doença pulmonar	Nº de indivíduos c/ doença pulmonar	
Cidade A	34	366	400
Cidade B	42	358	400
Campo	21	379	400
	97	1103	1200

[Objectivo: as populações serão homogêneas relativamente à incidência de doenças pulmonares?]

P(ter doença pulmonar na cidade A) = P(ter doença pulmonar na cidade B) = P(ter doença pulmonar no campo)?



Teste de Homogeneidade

Testes de Independência

Cada elemento da população é classificado de acordo com dois critérios (ou atributos) X e Y . Face a uma amostra de dimensão n , pretendemos testar a independência do par (X, Y) .

Considerem-se então p classes (ou categorias) A_1, \dots, A_p que constituam uma partição do domínio de variação da v.a. X e k classes (ou categorias) B_1, \dots, B_k que constituam uma partição do domínio de variação da v.a. Y .

Pretende-se testar as hipóteses

H_0 : existe independência entre X e Y vs

H_1 : não existe independência entre X e Y

Seja n_{ij} o número de elementos da amostra que pertencem simultaneamente às classes A_i e B_j . Vamos construir a seguinte tabela de contingência:

	B_1	...	B_k	
A_1	n_{11}	...	n_{1k}	$n_{1.}$
\vdots	\vdots		\vdots	
A_p	n_{p1}	...	n_{pk}	$n_{p.}$
	$n_{.1}$...	$n_{.k}$	n

Num teste de independência nenhuma das margens é fixa.

$n_{i.}$ é o número total de elementos na categoria A_i

$n_{.j}$ é o número total de elementos na categoria B_j .

n_{ij} é a frequência observada da célula (i, j) da tabela. Supondo H_0 verdadeira, a respectiva frequência esperada é dada por

$$e_{ij} = \frac{n_{i.} \times n_{.j}}{n}$$

[Nenhuma das margens é fixa \Rightarrow não existem definidas amostras para cada grupo, só uma amostra global]

Estatística de teste:

$$X^2 = \sum_{i=1}^p \sum_{j=1}^k \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

Se H_0 verdadeira, X^2 tem distribuição assintótica $\chi^2_{(p-1)(k-1)}$

Ao nível de significância α , rejeitar H_0 se $X^2 > \chi^2_{(p-1)(k-1), 1-\alpha}$

Aplica-se também a regra prática referida anteriormente.

[o teste só é válido para grandes amostras]

Exemplo:

As Farmácias Hospitalares têm estado em grande evidência nos meios de comunicação social, com base numa alegada incapacidade de satisfazer as expectativas dos seus utentes. O Director de uma destas unidades de saúde decidiu verificar se as alegações seriam fundadas, no caso da Farmácia Hospitalar que dirige. Para tanto, organizou um questionário, validou-o, e inquiriu os 1000 primeiros utentes que foram atendidos no balcão. A tabela abaixo foi extraída dos dados recolhidos.

Opinião dos Utes	Período de Funcionamento	
	Manhã	Tarde
Sobre o Funcionamento		
Positiva	395	405
Negativa	90	110