Análise de Dados Categorizados - Aula 3

Márcia D Elia Branco

Universidade de São Paulo
Instituto de Matemática e Estatística
www.ime.usp.br/ mbranco - sala 295-A -

Inferência para Proporção: $X \sim Binomial(n, \pi)$

- Na última aula falamos de TH para π considerando amostras grandes ($n \to \infty$).
- Os Testes de Wald e Escore são baseados na distribuição assintótica qui-quadrado com 1 grau de liberdade da estatística

$$Z^2 = \frac{(\hat{\pi} - p_0)^2}{(ep(\hat{\pi}))^2}$$

- A diferença entre eles está em como estimar o erro padrão.
- Estatística de Wald: substitui π pelo seu emv .

$$ep(\hat{\pi}) = \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}$$

- Estatística de Escore: substitui π pelo valor em $H_0 : \pi = p_0$.

$$ep(\hat{\pi}) = \sqrt{\frac{p_0(1 - p_0)}{n}}$$

Intervalos de Confiança para Proporção

Analogamente, podemos definir dois tipos de Intervalos:

- O Intervalo de Wald de $(1-\alpha)$ x 100 % de confiança para π é obtido por

$$\left[\hat{\pi} - z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}; \hat{\pi} + z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right]$$

- O Intervalo Escore de $(1-\alpha)$ x 100 % de confiança para π é dada por todos os valores p_0 que satisfazem

$$\frac{\mid \hat{\pi} - p_0 \mid}{\sqrt{\frac{p_0(1-p_0)}{n}}} < z_{1-\alpha/2}$$

Em que z_p é o quantil de ordem p da $N(0,1)$

Teste de Hipóteses Exato

$$H_0 : \pi \leq p_0 \quad versus \quad H_a : \pi > p_0$$

Considera-se a distribuição exata sob H_0, isto é,

$$X \sim Binomial(n, p_0)$$

O valor-P deveria ser obtido usando-se

$$P(X \geq x_{obs}) = \sum_{j=x_{obs}}^{n} P(X = j).$$

No entanto, esse teste é muito conservador, isto é, rejeita pouco.

Uma proposta alternativa para testes exatos é usar o *mid* Valor-P

$$\frac{P(X = x_{obs})}{2} + P(X > x_{obs})$$

Exemplo 1: $H_0 : \pi = 0.5$ versus $H_a : \pi > 0.5$

Resultado amostral: $n = 10$ e $x = 9$.

Então, mid valor-P é

$$\frac{P(X = 9)}{2} + P(X = 10) = \frac{0.01}{2} + 0.001 = 0.006.$$

Rejeita-se H_0

Comparação de duas proporções

Para o caso de duas amostras independentes, temos

$$X_1 \sim Bin(n_1, \pi_1) \quad \text{ind.} \quad X_2 \sim Bin(n_2, \pi_2)$$

Deseja-se obter um Intervalo de Confiança γ para $\pi_1 - \pi_2$.

Usamos os estimadores de máxima verossimilhança (as proporções amostrais):

$$\hat{\pi}_1 = \frac{X_1}{n_1} \quad \text{e} \quad \hat{\pi}_2 = \frac{X_2}{n_2}$$

Os quais tem distribuições assintóticas normais e são independentes.

Assim

$$\hat{\pi}_1 - \hat{\pi}_2 \approx N\left(\pi_1 - \pi_2, \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right)$$

O Intervalo de Wald é obtido pela substituição de π_i por $\hat{\pi}_i$ $i = 1, 2$. Assim, o $IC(1-\alpha)$ é

$$\left[(\hat{\pi}_1 - \hat{\pi}_2) - z_{1-\alpha/2} ep(\hat{\pi}_1 - \hat{\pi}_2); (\hat{\pi}_1 - \hat{\pi}_2) + z_{1-\alpha/2} ep(\hat{\pi}_1 - \hat{\pi}_2)\right]$$

com

$$ep(\hat{\pi}_1 - \hat{\pi}_2) = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}.$$

A distribuição multinomial

$$(X_1, X_2, \ldots, X_m) \sim Mult(n, \pi)$$

Em que $\pi = (\pi_2, \pi_2, \ldots, \pi_m)$ com $\sum_{i=1}^{m} \pi_i = 1$.

$$f(x_1, x_2, \ldots, x_m) = \frac{n!}{x_1! \ldots x_m!} \prod_{i=1}^{m} \pi_i^{x_i}$$

Propriedades:

1. $E[X_i] = n\pi_i$ e $Var[X_i] = n\pi_i(1 - \pi_i)$
2. $Cov(X_i, X_j) = -n\pi_i\pi_j$
3. $X_i \sim Bin(n, \pi_i)$
4. As componentes do vetor são correlacionadas negativamente.

Inferência para várias proporções: modelo multinomial

Intervalo aproximado de confiança $1-\alpha$ para $\pi_i - \pi_j$

$$\left[(\hat{\pi}_i - \hat{\pi}_j) - z_{1-\alpha/2}ep(\hat{\pi}_i - \hat{\pi}_j), (\hat{\pi}_i - \hat{\pi}_j) + z_{1-\alpha/2}ep(\hat{\pi}_i - \hat{\pi}_j)\right]$$

Com

$$ep(\hat{\pi}_i - \hat{\pi}_j) = \sqrt{\frac{\hat{\pi}_i(1-\hat{\pi}_i) + \hat{\pi}_j(1-\hat{\pi}_j) + 2\hat{\pi}_i\hat{\pi}_j}{n}}$$

Exemplo 2: Considere uma amostra de 309 ingressantes no mestrado. Seu desempenho em uma disciplina básica é apresentado na tabela a seguir

Conceito	A	B	C	R	Total
Frequência	84	80	112	33	309

Deseja-se comparar as proporções populacionais (ou probabilidades) associada a cada um dos conceitos nesta disciplina.

Modelo probabilístico:

$$(X_A, X_B, X_C, X_R) \sim Mult(309, (\pi_A, \pi_B, \pi_C, \pi_R))$$

Comparando A com B : $IC(\pi_A - \pi_B, 0.90)$

$\hat{\pi}_A = 0.272$, $\hat{\pi}_A = 0.259$

$$ep = \sqrt{\frac{0.272(1-0.272) + 0.259(1-0.259) + 2 \times 0.272 \times 0.259}{309}} = 0.0414$$

Resulta no seguinte intervalo

$$(0.013 - 1.645 \times 0.0414, 0.013 + 1.645 \times 0.0414) = (-0.0551, 0.0811)$$

Como o intervalo contêm o zero, não há diferença significativa entre as probabilidades de conceitos A e B.

Teste de aderência ou bondade de ajuste

$$H_0 : Y \sim F_0 \quad \text{versus} \quad H_a : Y \nsim F_0$$

Alternativamente

$$H_0 : \pi_i = p_{0i} \; (\forall i) \text{ versus } \; H_a : \pi_j \neq p_{0j} \;\; \text{(pelo menos um j)}$$

Sob H_0 as frequências esperadas são $E_i = np_{0i}$.

Estatística do teste qui-quadrado de aderência:

$$Q = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i}$$

Esta estatística tem distribuição assintótica qui-quadrado com $\nu = m - 1$ graus de liberdades.

Exemplo 3: Verifique se o número de gols em uma partida de futebol segue uma distribuição de Poisson com taxa $\lambda = 0.7$.

No. Gols	Frequência
0	33
1	17
2	07
3 ou mais	03

Primeiro obtemos as probabilidades no modelo Poisson

$$p_0 = 0.4965 \ , \ p_1 = 0.3476 \ , \ p_2 = 0.1217$$

Teste de aderência ou bondade de ajuste

$$H_0 : \pi_0 = 0.4965, \pi_1 = 0.3476, \pi_2 = 0.1217, \pi_3 = 0.0342$$

Valores esperados

$$E_0 = 60 \times 0.4965 = 29.79 \ , \ E_1 = 60 \times 0.3476 = 20.856$$

$$E_2 = 60 \times 0.1217 = 7.302 \ , \ E_3 = 60 \times 0.0342 = 2.052$$

Resulta em $Q = 1.51$.

Região crítica do teste: $RC = \{Q \geq \chi^2_{3,0.05} = 7.814\}$.

Decisão: Não rejeita-se H_0.

Tabelas 2×2

Covariável X	Resposta Y		Totais
	j=1	j=2	
i=1	n_{11}	n_{12}	n_{1+}
i=2	n_{21}	n_{22}	n_{2+}
Totais	n_{+1}	n_{+2}	n

- Medidas de associação: Risco atribuível; Risco relativo e Razão de chances.
- Testes qui-quadrado de independência, homogeneidade e multiplicatividade.
- Teste exato de Fisher.

1. **Risco atribuível ou diferença entre proporções**

$$d = p_{(1)1} - p_{(2)1}$$

Estimador é dado por

$$\hat{d} = \frac{N_{11}}{n_{1+}} - \frac{N_{21}}{n_{2+}}$$

- Essa medida varia no intervalo $[-1, 1]$. Se $d = 0$ não há diferença entre os grupos.
- Sob a suposição de indendência entre as amostras, o erro padrão do estimador é

$$ep(\hat{d}) = \sqrt{\frac{\hat{p}_{(1)1}(1 - \hat{p}_{(1)1})}{n_{1+}} + \frac{\hat{p}_{(2)1}(1 - \hat{p}_{(2)1})}{n_{2+}}}$$

- Considerando a aproximação normal para binomial, temos o seguinte $IC(1-\alpha)$ aproximado para d

$$[\hat{d} - z_{\alpha/2} ep(\hat{d}); \hat{d} + z_{\alpha/2} ep(\hat{d})]$$

2. Risco Relativo

$$RR = \frac{P(Y=1 \mid X=1)}{P(Y=1 \mid X=0)} = \frac{p_{(1)1}}{p_{(2)1}}$$

Estimador do RR

$$\hat{RR} = \frac{N_{11} n_{2+}}{n_{1+} N_{21}}$$

Medidas de associação em tabelas 2×2

- $RR = 1$ não há diferença entre os grupos.
- A distribuição amostral de $\hat{RR}$ é bastante assimétrica, indicando que aproximação normal é obtida apenas para amostras muito grandes.
- Para melhorar essa aproximação contruímos os IC para o logaritmo de RR.
- O estimador $\log(\hat{RR})$ tem erro padrão dado por

$$\sqrt{\frac{1-p_{(1)1}}{(n_{1+})p_{(1)1}} + \frac{1-p_{(2)1}}{(n_{2+})p_{(2)1}}}$$

Exemplo 1: O interesse é comparar dois vermífugos. Modelo Produto de Binomiais.

Vermífugo	Verminose		Totais
	Sim	Não	
1	48	152	200
2	68	132	200
Totais	116	284	400

$$\hat{d} = \frac{48}{200} - \frac{68}{200} = -0.10 \quad \text{e} \quad \hat{RR} = \frac{48 \times 200}{68 \times 200} = 0.71$$

$$ep(\hat{d}) = \sqrt{[0.24(1-0.24)]/200 + [0.34(1-0.34)]/200} = 0.045$$

IC(0.90) para o risco atribuível:

$$[-0.1 - 1.645 ep(\hat{d}); -0.1 + 1.645 ep(\hat{d})] = [-0.174; -0.026]$$

$$ep(\log(\hat{RR})) = \sqrt{\frac{(1-0.24)}{200\times 0.24} + \frac{(1-0.34)}{200\times 0.34}} = 0.16$$

IC(0.90) para o logaritmo de RR:

$$[\log(0.71) - 1.645(0.16); \log(0.71) + 1.645(0.16)] = [-0.605; -0.080]$$

- O IC(0.90) para d contêm apenas valores negativos indicando que $p_{(1)1} < p_{(2)1}$ com uma confiança de 0.90.
- O IC(0.90) para o RR é dado por

$$[e^{-0.605}; e^{-0.080}] = [0.546; 0.923]$$

contendo apenas valores menores que 1. Mais uma vez, confirma-se a superioridade do Vermífugo 1.

Lembrando:

Chance de um evento ocorrer é a razão entre a probabilidade do evento ocorrer e a probabilidade dele não ocorrer.

3. Razão de Chances

Em tabelas 2×2 onde o evento de interesse esta associado a $j = 1$, a chance desse evento é dada por $\frac{p_{(1)1}}{1-p_{(1)1}}$ para a linha 1 e $\frac{p_{(2)1}}{1-p_{(2)1}}$ para linha 2.
A razão das chances é obtida por

$$OR = \frac{p_{(1)1}p_{(2)2}}{p_{(1)2}p_{(2)1}}$$

O estimador pontual dessa medida é dado por

$$\hat{OR} = \frac{N_{11}N_{22}}{N_{21}N_{12}}$$

denominado razão dos produtos cruzados.

- Para estudos do tipo coorte, OR representa a razão entre as chances da doença entre os expostos ao fator de risco e a chance de ocorrência da doença entre os não expostos
- Para estudos caso-controle (retrospectivo), OR representa a razão entre a chance de exposição entre os casos e a chance de exposição entre os controles. Neste caso é calculada como

$$OR = \frac{p_{1(1)} p_{2(2)}}{p_{2(1)} p_{1(2)}}$$

- Em estudos transversais onde não é fixado previamente os totais marginais (linhas ou colunas), há controvérsia a respeito da interpretação dessa medida.

Medidas de associação em tabelas 2×2

No exemplo, para o Tratamento 1 chance do animal ter verminose é $\frac{48}{152}$; enquanto que para o Tratamento 2 essa chance é de $\frac{68}{132}$.

$$\hat{OR} = \frac{48 \times 132}{68 \times 152} = 0.613$$

- Como este valor é menor que 1, concluímos que a chance de verminose é menor para o Tratamento 1.
- Notamos que $\frac{1}{\hat{OR}} = 1.63$. Então, podemos dizer que a chance de verminose para os animais submetidos ao Tratamento 2 é 1.6 vezes a chance de verminose dos animais submetidos ao Tratamento 1.
- Note que o planejamento do exemplo é do tipo Prospectivo. Neste caso o condicionamento é feito por linha (dado $X = i$).
- Intervalos de confiança aproximados também podem ser obtidos. Tarefa!

Exemplo 2: Estudo do tipo caso-controle (retrospectivo).
Na tabela a seguir apresentamos resultado de um estudo realizado em Londres com 709 casos de câncer de pulmão e 709 indivíduos sem câncer de pulmão (controle).

Fumante	Câncer de pulmão		Totais
	Casos	Controle	
Sim	688	650	1338
Não	21	59	80
Totais	709	709	1418

Medidas de associação em tabelas 2×2

- Para os casos, a chance do indivíduo ter sido exposto ao fumo é $\frac{688}{21} = 32.76$. Para os controle, a chance do indivíduo ter sido exposto ao fumo é $\frac{650}{59} = 11.02$.
- A estimativa da razão de chances é $\hat{OR} = 2.97 \approx 3$.
- Interpretação associada ao planejamento (dado $Y = j$) : a chance de exposição ao fumo para os indivíduos com câncer é aproximadamente 3 vezes à dos indivíduos no grupo controle.
- Interpretação de interesse: a chance de câncer de pulmão em indivíduos expostos ao fumo é 3 vezes à dos indivíduos não expostos.
- Devido a simetria da medida OR a interpretação pode ser feita na direção do nosso interesse.

Medidas de associação em tabelas 2×2

- A medida de RR não possue a mesma simetria da OR.
- No exemplo, se considerarmos o planejamento deveriamos comparar as probabilidades condicionais as colunas.

$$RR = \left(\frac{688}{709}\right)\left(\frac{709}{650}\right) = 1.06$$

- Mas o interesse real do estudo é falar do risco de doença (não risco de estar exposto ao fator). Este é obtido de forma diferente

$$RR = \left(\frac{688}{1338}\right)\left(\frac{80}{21}\right) = 1.96$$

Medidas de associação em tabelas 2×2

- Podemos mostrar (exercício) que a relação entre RR e OR é dada por

$$OR = RR \times \left(\frac{1 - p_{1(2)}}{1 - p_{1(1)}} \right)$$

- Nota-se que se $p_{1(1)}$ e $p_{1(2)}$ forem muito pequenas então $OR \approx RR$.
- Para estudos do tipo retrospectivos em que a probabilidade p_{1+} é pequena, podemos usar o valor estimado de OR como uma aproximação para a estimativa do RR.
- Fique atento para diferença na interpretação dessas duas medidas!