Análise de Dados Categorizados - Aula4

Márcia D Elia Branco

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

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 e homogeneidade. Testes de Pearson e da Razão de Verossimilhança.
- Teste exato de Fisher.
- Inferência Bayesiana.

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

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

- $RR = 1$ 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

- 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.

- 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

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

- 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!

Vamos considerar o modelo produto de binomiais com os totais das linhas previamente fixados pelo plano amostral.

Neste caso, o interesse é testar a hipótese de homogeneidade

$$H_0 : p_{(1)1} = p_{(2)1} \quad versus \quad H_a : p_{(1)1} \neq p_{(2)1}$$

A estatística de Pearson é dada por

$$Q_P = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

em que

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

Teste da Razão de Verossimilhança para homogeneidade.
A função de verossimilhança

$$L(p_{(1)1}, p_{(2)1}) \propto p_{(1)1}^{n_{11}} p_{(1)2}^{n_{12}} p_{(2)1}^{n_{21}} p_{(2)2}^{n_{22}}$$

Sob a hipótese alternativa, o ponto de máximo é emv sem restrição. Assim, a verossimilhança no ponto de máximo é

$$L_1 \propto \left(\frac{n_{11}}{n_{1+}}\right)^{n_{11}} \left(\frac{n_{12}}{n_{1+}}\right)^{n_{12}} \left(\frac{n_{21}}{n_{2+}}\right)^{n_{21}} \left(\frac{n_{22}}{n_{2+}}\right)^{n_{22}}$$

Sob a hipótese nula, o ponto de máximo é obtido maximizando-se a função restrita a H_0. Considere $p_{+1} = p_{(1)1} = p_{(2)1}$, temos a função de verossimilhança

$$L_0(p_{+1}) \propto p_{+1}^{n_{11}+n_{21}}(1 - p_{+1})^{n_{12}+n_{22}}$$

Derivando a log-verossimilhança e igualando a zero obtemos o ponto de máximo $\hat{p}_{+1} = \frac{n_{+1}}{n}$. Então

$$L_0 \propto \left(\frac{n_{+1}}{n}\right)^{n_{+1}} \left(\frac{n_{+2}}{n}\right)^{n_{+2}}$$

Usando a notação $e_{ij} = \frac{n_{i+}n_{+j}}{n}$, resulta em

$$\frac{L_1}{L_0} = \left(\frac{n_{11}}{e_{11}}\right)^{n_{11}} \left(\frac{n_{12}}{e_{12}}\right)^{n_{12}} \left(\frac{n_{21}}{e_{21}}\right)^{n_{21}} \left(\frac{n_{22}}{e_{22}}\right)^{n_{22}} .$$

Portanto, o valor da estatística do teste de RV é dada por:

$$2log(\frac{L_1}{L_0}) = 2\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}\log(\frac{n_{ij}}{e_{ij}})$$

Os graus de liberdades associados a distribuição assintótica dessa estatística são $\nu = 2 - 1 = 1$.

Para os dados do exemplo 2 vamos testar a homogeneidade entre as populações de indivíduos com câncer e controle.

$$H_0 : p_{1(1)} = p_{1(2)} \quad versus \quad H_a : p_{1(1)} \neq p_{1(2)}$$

Usando $\alpha = 0.05$ e $\nu = 1$ a região critica é $RC = \{Q_{RV} > 3.84\}$.

$$Q_{RV} = 2\left[688\log(\frac{688}{669}) + 650\log(\frac{650}{669}) + 21\log(\frac{21}{40}) + 59\log(\frac{59}{40})\right]$$

$$Q_{RV} = 19.88 \in RC \qquad Valor - P < 10^{-5}$$

Rejeita-se a hipótese de homogeneidade.

Exemplo 3: Durante 18 semanas de determinado ano foi contado o número de acidentes de carros registrados na Suécia, avaliando-se o tipo de estrada e o fato de haver ou não um limite de velocidade. O objetivo é avaliar se o limite de velocidade influencia de maneira diferente o número de acidentes dependendo do tipo de estrada.

Limite de velocidade	Tipo de estrada		Totais
	Auto-estrada	Outra	
Sim	8	42	50
Não	57	106	163
Totais	65	148	213

Vamos realizar um teste de multiplicatividade considerando o modelo produto de Poisson.

Testes Qui-quadrado

Hipóteses :

$$H_0 : \mu_{11} = \frac{\mu_{1+}\mu_{+1}}{\mu}; \mu_{12} = \frac{\mu_{1+}\mu_{+2}}{\mu}; \mu_{21} = \frac{\mu_{2+}\mu_{+1}}{\mu}; \mu_{22} = \frac{\mu_{2+}\mu_{+2}}{\mu}$$

$$H_a : \text{Existe pelo menos uma diferente.}$$

- É possível usar a estatística de Pearson Q_P .
- Alternativamente, podemos usar a estatística baseada na Razão de Verossimilhança Q_{RV}.
- É possivel mostrar que $Q_{RV} = 2\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}\log(\frac{n_{ij}}{e_{ij}})$ (exercício).

  Com graus de liberdades $\nu = 4 - 3 = 1$.

Regiões Criticas dos testes :

$RC = \{Q_{RV} > 3.84\}$ e $RC = \{Q_P > 3.84\}$.

Valores esperados das caselas são $e_{11} = 15.26$, $e_{12} = 34.74$, $e_{21} = 49.74$ e $e_{22} = 113.26$.

Resultando em $Q_P = 6.49$ e $Q_{RV} = 7.09$.

Em ambos os casos rejeita-se a hipótese de multiplicidade e conclui-se pela dependência entre as variáveis. O limite de velocidade influencia de forma diferente o número de acidentes dependendo do tipo de estrada.

Teste Exato de Fisher

- O Teste exato de Fisher é um teste não paramétrico usado para pequenas amostras.
- A hipótese nula consiste em não associação entre as variáveis. A hipótese alternativa pode ser unilateral ou bilateral.
- Para construção do teste ambas marginais da tabelas são supostas conhecidas.
- É preciso calcular as probabilidades associada a ocorrência da tabela observada e de outras tabelas "mais extremas"do que essa.

Teste Exato de Fisher

Suponha que $N_{11} \mid n_{1+} \sim Binomial(n_{1+}, p_{(1)1})$ e $N_{21} \mid n_{2+} \sim Binomial(n_{2+}, p_{(2)1})$, independentes.

Se condicionarmos aos totais das colunas, temos que :

(i) apenas uma variável é livre, a segunda fica totalmente determinada pelo conhecimento da primeira.
(ii) a distribuição condicional $N_{11} \mid n_{1+}, n_{+1}, n$ é uma hipergeométrica.
(ii) a probabilidade de ocorrer o resultado de uma particular tabela é equivalente a $P(N_{11} = n_{11})$ e dada por

$$\frac{C_{n_{11}}^{n1+} \times C_{n_{+1}-n_{11}}^{n-n_{1+}}}{C_{n_{+1}}^{n}}$$

Exemplo 4: O proplema proposto por Fisher em 1935 consiste em comprovar ou refutar a afirmação feita por uma senhora que diz ser capaz de diferenciar, pelo paladar, a ordem que foi adicionado o leite à sua xicará de chá.
O experimento consistiu em considerar 8 xícaras de chá. Em 4 delas o leite foi colocado primeiro e nas outras 4, o chá foi colocado antes do leite. Antes da senhora fazer as provas, foi informado a ela que haviam 4 xicaras de cada tipo.

Ordem correta	Resposta		Totais
	Leite	Chá	
Leite	3	1	4
Chá	1	3	4
Totais	4	4	8

Teste Exato de Fisher

H_0 : Não há associação entre Resposta e Ordem Correta.
H_a : Existe associação e o sentido desta é de que a senhora tem a sensibilidade anunciada.

Quais seriam as tabelas mais extremas no sentido de H_a ?

Ordem correta	Resposta		Totais
	Leite	Chá	
Leite	4	0	4
Chá	0	4	4
Totais	4	4	8

Teste Exato de Fisher

Calculando as probabilidades das duas tabelas

$$P(N_{11} = 3) = \frac{C_3^4 \times C_1^4}{C_4^8} = 0.229$$

$$P(N_{11} = 4) = \frac{C_4^4 \times C_0^4}{C_4^8} = 0.014$$

O valor-P associado ao teste é dado por 0.243 . Não rejeita-se H_0.
Conclusão: Não há evidências de que a Senhora tenha a sensibilidade anunciada.

- Para conduzir um teste bilateral, deve-se calcular as probabilidades de tabelas mais extremas nos dois sentidos.
- Se iniciarmos com o modelo Multinomial ou Produto de Poisson, também é possível mostrar que condicionando aos totais marginais obtemos a mesma distribuição hipergeométrica.