Análise de Dados Categorizados - Aula1

Márcia D Elia Branco

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

Conceitos Introdutórios e Notações

- **Variáveis categóricas:** são variáveis qualitativas, com um número finito de categorias disjuntas de respostas. Podem estar na escala *nominal* ou *ordinal*.
- Variáveis com apenas duas categorias são denominadas *dicotômicas* ou *binárias*.
- Variáveis com três ou mais categorias são denominadas *politômicas*.
- Variáveis numéricas discretas ou contínuas também podem ser categorizadas.
- Variáveis respostas x Variáveis explicativas (fatores ou covariáveis).
- As técnicas de dados categorizados estudadas aqui referem-se basicamente à característica categórica da variável resposta.

Conceitos Introdutórios e Notações

- Tabela de contingência de dupla entrada ($r \times c$): envolve apenas duas variáveis categóricas.
- n_{ij} representa a frequência observada na linha i coluna j da tabela.
- n_{i+} representa a soma da linha i.
- n_{+j} representa a soma da coluna j.
- $\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} = n_{++} = n$
- N_{ij} denota a variável aleatória associada a observação n_{ij}.
- $p_{ij} = P(X = i, Y = j)$, $p_{i+} = P(X = i)$ e $p_{+j} = P(Y = j)$.
- Tabelas com multiplas entradas ($q \times r \times c$): três ou mais variáveis categóricas .

Exemplos

Exemplo 1: O interesse é comparar dois vermífugos. Para isso o pesquisador selecionou 400 carneiros da mesma raça, todos sem verminose, mantendo-os sob o mesmo manejo em pastos com condições similares. A seguir, separam-se os animais aleatoriamente em dois grupos de tamanhos iguais e, para cada um, administrou-se um de dois vermífugos. Decorridos 4 meses os animais foram examinados.

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

Exemplos

Exemplo 2: 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 verificar 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

1. **Estudos Longitudinais (de Coorte ou Prospectivos):**
   - Inicia-se com um grupo de indivíduos todos livres da doença sob investigação.
   - Esse grupo é classificado como *Exposto* ou *Não Exposto* a um determinado agente (fator de risco), obtendo-se dois grupos (coortes).
   - Observa-se os dois grupos por um período de tempo, registra-se o números de indivíduos que desenvolveram a doença nesse período.
   - Nota-se que os totais n_{1+} (expostos) e n_{2+} (não expostos) são fixados previamente.

2. Estudos caso-controle:

- Seleciona-se um grupo de indivíduos com uma determinada doença, denominado *casos* .
- Escolhe-se um outro grupo de indivíduos sem a doença, com características similares ao primeiro, denomindado *controle*.
- Estabelecidos os grupos, registram-se os indivíduos *expostos* e *não expostos* ao fator de risco sob investigação.
- Esses estudos são denomindados *retrospectivos*, pois usam informações do passado.
- Nota-se que os totais n_{+1} (casos) e n_{+2} (controle) são fixados previamente.

3. Estudos transversais:

- Informações sobre várias características são obtidas simultaneamente.
- Difícil inferir causualidade.
- Neste caso, somente n é fixado, as demais quantidades são observações de variáveis aleatórias.

4. Ensaios clínicos aleatorizados:

- Usualmente usado para comparar dois ou mais tratamentos.
- Inicia-se com um grupo de indivíduos elegíveis, todos livres da doença sob investigação.
- Os tratamentos de interesse são alocados aleatoriamento aos indivíduos elegíveis.
- Observa-se os grupos (definidos pelo tratamento alocado) por um período de tempo, registra-se o números de indivíduos que desenvolveram a doença nesse período. Estudo *prospectivo*
- Nota-se que os totais n_{1+} (expostos) e n_{2+} (não expostos) são fixados previamente.

Exemplo 3: Um estudo foi realizado para pesquisar a associação entre tabagismo e câncer de pulmão. Foram considerados dois grupos de indivíduos: o primeiro com indivíduos com câncer e o segundo com indivíduos sem a doença. Em seguida, investigou-se sobre o histórico de exposição ao tabaco dos dois grupos. Os dados são apresentados na tabela a seguir.

Exposição ao tabaco	Câncer		Totais
	Sim	Não	
Sim	75	45	120
Não	21	56	77
Totais	96	101	197

Exemplos

Exemplo 4: Na Faculdade de Medicina de Ribeirão Preto selecionou-se 8135 fichas hospitalares referentes aos nascimentos ocorridos num certo período de tempo. Os nascimentos foram classificados em relação as variáveis: Classe Social (em 5 níveis, decrescentemente ordenada da mais favorecida pra menos favorecida); Hábito de Fumar (em 2 níveis) da parturiente e Peso do recém-nascido (em 3 níveis).

Peso RN	< 2.5 kg		2.5 - 3.0 kg		> 3.0 kg	
	HFumar		HFumar		HFumar	
Classe Social	Sim	Não	Sim	Não	Sim	Não
A	2	5	11	24	31	95
B	3	11	32	57	91	238
C	15	25	58	105	134	445
D	94	105	225	339	340	1053

Modelos probabilísticos em tabelas de dupla entrada

- Modelo produto de binomiais: os totais marginais das linhas (ou colunas) são fixados pelo plano amostral. Associado a tabelas $r \times 2$ (ou $2 \times c$).
- Modelo produto de multinomias: os totais marginais das linhas (ou colunas) são fixados pelo plano amostral. Associado a tabelas $r \times c$.
- Modelo multinomial: somente o tamanho total da amostra (n) é fixado pelo plano.
- Modelo produto de Poisson: observações são realizadas num período de tempo. O tamanho da amostra não é fixado previamente.

Modelo Produto de Binomiais

Ilustramos com o Exemplo 1.

N_{11} é o número de indivíduos com verminose no grupo 1 e
N_{21} é o número de indivíduos com verminose no grupo 2.

$$N_{i1} \sim Binomial(200, p_{(i)1}) \quad i = 1, 2$$

Em que $p_{(i)1}$ é uma probabilidade condicional dada por

$$p_{(i)1} = \frac{p_{i1}}{p_{i+}}$$

Sob a hipótese de amostras independentes temos a seguinte função de verossimilhança

$$L(p_{(1)1}, p_{(2)1}) = \prod_{i=1}^{2} C_{n_{i1}}^{n_{i+}} p_{(i)1}^{n_{i1}} (1 - p_{(i)1})^{n_{i+} - n_{i1}}$$

Exemplo 5: A tabela a seguir apresenta o resultado de um ensaio clínico aleatorizado.

	Melhora do paciente			
Tratamento	Nenhuma	Alguma	Acentuada	Totais
Ativo	13	07	21	41
Placebo	29	07	07	43
Totais	42	14	28	84

Modelo Produto de Multinomiais

Para o grupo 1 (ativo) temos

$$(N_{11}, N_{12}) \sim Multinomial(41, (p_{(1)1}, p_{(1)2}, p_{(1)3}))$$

Isto é,

$$P(N_{11} = n_{11}, N_{12} = n_{12}) = \frac{n_{1+}!}{n_{11}!n_{12}!n_{13}!} \prod_{j=1}^{3} p_{(1)j}^{n_{1j}}$$

Em que
$p_{(1)3} = 1 - p_{(1)1} - p_{(1)2}$ e $n_{13} = n_{1+} - n_{11} - n_{12}$

Equivalmentemente para o grupo 2 (placebo).

Modelo Produto de Multinomiais

A função de verossimilhança (supondo independência)

$$L(p_{(1)1}, p_{(1)2}) \propto \prod_{i=1}^{2} \prod_{j=1}^{3} p_{(i)j}^{n_{ij}}$$

Para tabelas $r \times c$ a verossimilhança é proporcional a

$$\prod_{i=1}^{r} \prod_{j=1}^{c} p_{(i)j}^{n_{ij}}$$

em que

$$p_{(i)c} = 1 - \sum_{j=1}^{c-1} p_{(i)j} \quad \text{e} \quad n_{ic} = n_{i+} - \sum_{j=1}^{c-1} n_{ij}.$$

Exemplo 6: Foi realizada uma entrevista com 1100 indivíduos para avaliar a opinião em relação a legalização do aborto. Os resultados, considerando o sexo do indivíduo, são apresentados na tabela a seguir.

Sexo	Favoráveis	Contrários	Totais
Fem	309	191	500
Mas	319	281	600
Totais	628	472	1100

Modelo Multinomial

$$(N_{11}, N_{12}, N_{21}) \sim Multinomial(1100, (p_{11}, p_{12}, p_{21}, p_{22}))$$

Para tabelas $r \times c$ a verossimilhança é proporcional a

$$\prod_{i=1}^{r} \prod_{j=1}^{c} p_{ij}^{n_{ij}}$$

Os estimadores de máxima verossimilhança são:

$$\hat{p}_{ij} = \frac{N_{ij}}{n}$$

Hipótese de independência:

$$H_0 : p_{ij} = p_{i+} p_{+j} \quad i = 1, \ldots, r \ , \ j = 1, \ldots, c$$

Sob H_0 (independência) temos que

$$E[N_{ij}] = np_{ij} = np_{i+}p_{+j}$$

Substituindo p_{i+} e p_{+j} por suas estimativas, temos que o valor esperado na casela(i, j) é

$$e_{ij} = n\left(\frac{n_{i+}}{n}\right)\left(\frac{n_{+j}}{n}\right) = \frac{n_{i+}n_{+j}}{n}$$

A estatística Qui-quadrado de Pearson é dada por

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

Modelo Multinomial

- Valores grandes de Q_p são indicativos contra H_0 (independência). Portanto, indicam dependência entre as variáveis.
- Para amostras "grandes"a estatística Q_p tem distribuição Qui-quadrado com $\nu = (r-1)(c-1)$ graus de liberdade
- A região crítica do teste é dada por $RC = \{Q_p > \chi^2_{\alpha,\nu}\}$
- No exemplo 3 temos $Q_p = 8.3$. Para $\alpha = 0.05$ e $\nu = 1$ temos $\chi^2_{0.05,1} = 3.84$.
- Rejeita-se a independência entre Sexo e Opinião sobre o aborto (valor-P < 0.004)

Tarefa

1. Para cada um dos exemplos apresentados aqui, estabelecer o tipo de estudo (se possível) e o modelo probabilístico associado

2. Estudar e fazer os exercícios dos capítulos 1 e 2 do livro da Suely Giolo.