Introdução à análise de dados discretos

Prof. Caio Azevedo

Exemplo 1: comparação de métodos de detecção de cárie

- Suponha que um pesquisador lhe apresente a seguinte tabela de contingência, resumindo os dados coletados por ele, oriundos de um determinado experimento:

		Risco de cárie segundo o método convencional			
		Baixo	Médio	Alto	Total
Risco de cárie segundo o método simplificado	Baixo	11	5	0	16
	Médio	14	34	7	55
	Alto	2	13	11	26
Total	-	27	52	18	97

Aspectos relevantes

- Pergunta: sendo o método simplificado mais barato, vale à pena usá-lo no lugar do método covencional?
- O que significa "valer à pena"?
- Como os dados foram coletados?
- População ou amostra?
- Inferência paramétrica: qual modelo estatístico (de regressão) deve ser utilizado? Verificar a qualidade do ajuste do modelo.
- Como responder à pergunta de interesse?
- Como apresentar a resposta obtida à pergunta de interesse?

Um pouco de discussão

- Suponha que: selecionou-se, ao acaso, através de um processo de amostragem aleatória simples sem reposição, 97 indivídios (de um possível grupo de interesse?). Em cada um deles, as duas técnicas foram aplicadas.
- Podemos considerar que a tabela obtida é uma dentre várias possíveis obtidas, ao se replicar o experimento. Ou seja, ela é uma amostra de uma população de interesse.
- Possível modelo probabilístico apropriado: multinominal com 8 classes (não exaustivas).

Distribuição multinomial

- Seja $\mathbf{N} = (N_{11}, N_{12}, N_{13}, N_{21}, N_{22}, N_{23}, N_{31}, N_{32}, N_{33}) \sim$ Multinomial$(n, \boldsymbol{\theta})$, $\sum_{i=1}^{3}\sum_{j=1}^{3} n_{ij} = n$, em nosso caso n $= 97$ e $\boldsymbol{\theta} = (\theta_{11}, \theta_{12}, \theta_{13}, \theta_{21}, \theta_{22}, \theta_{23}, \theta_{31}, \theta_{32}, \theta_{33})$, $\sum_{i=1}^{3}\sum_{j=1}^{3} \theta_{ij} = 1$.
- Para n$=$ 97, observamos $\mathbf{n} = (n_{11}, n_{12}, n_{13}, n_{21}, n_{22}, n_{23}, n_{31}, n_{32}, n_{33}$.

Assim, temos

$$
\begin{aligned}
P(\mathbf{N} = \mathbf{n}|\boldsymbol{\theta}, n) &= \frac{n!}{\prod_{i=1}^{3}\prod_{j=1}^{3} n_{ij}!} \prod_{i=1}^{3}\prod_{j=1}^{3} \theta_{ij}^{n_{ij}} \\
&\times \left(\prod_{i=1}^{3}\prod_{j=1}^{3} \mathbb{1}_{\{0,1,..,n\}}(n_{ij})\right)\left(\mathbb{1}_{\{n\}}\left(\sum_{i=1}^{3}\sum_{j=1}^{3} n_{ij}\right)\right) \rightarrow \\
L(\boldsymbol{\theta}) &\propto \prod_{i=1}^{3}\prod_{j=1}^{3} \theta_{ij}^{n_{ij}}; \text{ em que } \sum_{i=1}^{3}\sum_{j=1}^{3} \theta_{ij} = 1
\end{aligned}
$$

Um pouco de discussão: distribuição multinomial

		Risco de cárie segundo o método convencional			
		Baixo	Médio	Alto	Total
Risco de cárie segundo o método simplificado	Baixo	n_{11}	n_{12}	n_{13}	$n_{1.}$
	Médio	n_{21}	n_{22}	n_{23}	$n_{2.}$
	Alto	n_{31}	n_{32}	n_{33}	$n_{3.}$
Total	-	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{..} = n$

- $n_{i.} = \sum_{j=1}^{3} n_{ij}, i = 1, 2, 3$, $n_{.j} = \sum_{i=1}^{3} n_{ij}, j = 1, 2, 3$ e $n_{..} = \sum_{i=1}^{3} \sum_{j=1}^{3} n_{ij} = n$. Em nosso caso, $n = 97$.

Um pouco de discussão: distribuição multinomial

		Risco de cárie segundo o método convencional			
		Baixo	Médio	Alto	Total
Risco de cárie segundo o método simplificado	Baixo	θ_{11}	θ_{12}	θ_{13}	$\theta_{1.}$
	Médio	θ_{21}	θ_{22}	θ_{23}	$\theta_{2.}$
	Alto	θ_{31}	θ_{32}	θ_{33}	$\theta_{3.}$
Total	-	$\theta_{.1}$	$\theta_{.2}$	$\theta_{.3}$	$\theta_{..} = 1$

- $\theta_{i.} = \sum_{j=1}^{3} \theta_{ij}, i = 1, 2, 3$, $\theta_{.j} = \sum_{i=1}^{3} \theta_{ij}, j = 1, 2, 3$ e $\theta_{..} = \sum_{i=1}^{3} \sum_{j=1}^{3} \theta_{ij} = 1.$

"Situação ideal" do ponto de vista do pesquisador

		Risco de cárie segundo o método convencional			
		Baixo	Médio	Alto	Total
Risco de cárie segundo o método simplificado	Baixo	θ_{11}	0	0	$\theta_{1.}$
	Médio	0	θ_{22}	0	$\theta_{2.}$
	Alto	0	0	θ_{33}	$\theta_{3.}$
Total	-	$\theta_{.1}$	$\theta_{.2}$	$\theta_{.3}$	$\theta_{..} = 1$

- Ou seja $\theta_{i.} = \theta_{.j} = \theta_{ij}, \forall i = j$ (concordância absoluta).

Possível configuração de interesse do ponto de vista do pesquisador

		Risco de cárie segundo o método convencional			
		Baixo	Médio	Alto	Total
Risco de cárie segundo o método simplificado	Baixo	θ_{11}	θ_{12}	θ_{13}	$\theta_{1.}$
	Médio	θ_{21}	θ_{22}	θ_{23}	$\theta_{2.}$
	Alto	θ_{31}	θ_{32}	θ_{33}	$\theta_{3.}$
Total	-	$\theta_{.1}$	$\theta_{.2}$	$\theta_{.3}$	$\theta_{..} = 1$

- $\theta_{i.} = \theta_{.j}, \forall i = j$ (concordância marginal). Como testar tal conjunto de hipóteses?

Exemplo 2: comparação do número de acidentes

- Descrição: número de acidentes (com algum tipo de trauma para as pessoas envolvidas) em 92 dias (correspondentes) em dois anos distintos (1961 e 1962), medidos em algumas regiões da Suécia.
- Considerou-se apenas 43 dias, correspondendo a dias de 1961 em que não havia limite de velocidade e de 1962 em que havia limites de velocidade (90 ou 100 km/h).
- Questão de interesse: a imposição dos limites de velocidade levou à redução do número de acidentes?

Aspectos relevantes

- O que significa "levou à redução"?
- Como os dados foram coletados?
- População ou amostra?
- Inferência paramétrica: qual modelo estatístico (de regressão) deve ser utilizado? Verificar a qualidade do ajuste do modelo.
- Como responder à pergunta de interesse?
- Como apresentar a resposta obtida à pergunta de interesse?

Discussão

- Seja Y_{ij} o número de acidentes ocorridos no dia j = 1,2,...,43 do ano i=1,2 e y_{ij} os respectivos valores observados.

Discussão

- Seja Y_{ij} o número de acidentes ocorridos no dia j = 1,2,...,43 do ano i=1,2 e y_{ij} os respectivos valores observados.
- Seja o modelo:

$$\begin{aligned} Y_{ij} &= \mu + \alpha_i + \xi_{ij} \\ \xi_{ij} &\overset{i.i.d.}{\sim} N(0, \sigma^2), \alpha_1 = 0 \end{aligned}$$

Discussão

- Seja Y_{ij} o número de acidentes ocorridos no dia j = 1,2,...,43 do ano i=1,2 e y_{ij} os respectivos valores observados.
- Seja o modelo:

$$\begin{aligned} Y_{ij} &= \mu + \alpha_i + \xi_{ij} \\ \xi_{ij} &\overset{i.i.d.}{\sim} N(0, \sigma^2), \alpha_1 = 0 \end{aligned}$$

- O modelo acima é apropriado?

Discussão

- Seja Y_{ij} o número de acidentes ocorridos no dia j = 1,2,...,43 do ano i=1,2 e y_{ij} os respectivos valores observados.
- Seja o modelo:

$$\begin{aligned} Y_{ij} &= \mu + \alpha_i + \xi_{ij} \\ \xi_{ij} &\overset{i.i.d.}{\sim} N(0, \sigma^2), \alpha_1 = 0 \end{aligned}$$

- O modelo acima é apropriado?
- Se não, quais problemas ele apresenta?

Discussão

- Seja Y_{ij} o número de acidentes ocorridos no dia j = 1,2,...,43 do ano i=1,2 e y_{ij} os respectivos valores observados.
- Seja o modelo:

$$\begin{aligned} Y_{ij} &= \mu + \alpha_i + \xi_{ij} \\ \xi_{ij} &\overset{i.i.d.}{\sim} N(0, \sigma^2), \alpha_1 = 0 \end{aligned}$$

- O modelo acima é apropriado?
- Se não, quais problemas ele apresenta?
- Qual seria uma alternativa?

Conhecimento

- Necessário: Probabilidade I e II, Inferência (e noções de inferência), Análise de regressão.
- Auxiliar: Técnicas de Amostragem e Planejamento e Pesquisa.

Definições e Notações

- Consideramos um modelo estatístico formado por $(\Omega, \mathcal{A}, \mathcal{P})$, (espaço amostral, sigma álgebra de eventos de Ω, família de medidas de probabilidade).
- $\mathcal{P} = \{\mathcal{P}_\theta : \theta \in \Theta\}$, $\Theta \subset \mathcal{R}^p$ (em geral infinito não-enumerável). Basicamente: $\mathcal{P}_\theta = F_X(;\theta)$ (postulado).
- Realizar inferência (frequentista) : estimação pontual, intervalar e testar hipóteses estatísticas, em um dado espaço estatístico.

Continuação

- Exemplo: Seja $X_1, ..., X_n$ uma amostra aleatória de $X \sim \text{Bernoulli}(\theta), \theta \in (0, 1)$. Obter uma estimativa pontual e uma intervalar para θ e realizar testes de hipótese acerca dele.
- Suporte da fdp:
  - $\Omega = \{x : f_X(x; \theta) \geq 0\}$, $f_X(.; \theta)$ é a fdp associada à fda $F_X(; \theta)$. Eventualmente os subíndices podem ser suprimidos. Pode-se usar $p_X(; \theta)$ no lugar de $f_X(; \theta)$.
- Em geral $f_X(; \theta)$ será uma função de probabilidade (associada à uma variável aleatória discreta ou um vetor aleatório discreto). Nesee caso, $f_X(; \theta) = P(X = x; \theta)$.

Continuação

- Variável aleatória (va) discreta: é uma va cujo suporte corresponde à um conjunto finito ou infinito enumerável.
- Amostra aleatória (aa): $X_1, ..., X_n$ são estatísticamente independentes e identicamente distribuídas segundo $X \sim F_X(;\theta)$.
- Em alguns casos podemos não ter mesma distribuição e/ou independência.

Continuação

- Seja X uma vad (variável aleatória discreta). Então
  - $\mathcal{E}(\phi(X)) = \sum_x \phi(x) f_x(x;\theta)$.
  - A função geradora de momentos (fgm) é dada por:

    $\Psi_X(t) = \mathcal{E}(e^{tX}) = \sum_x e^{tx} f_x(x;\theta)$.
  - Temos que: $\mathcal{E}(X^r) = \frac{\partial^r}{\partial t^r} \Psi_X(t)|_{t=0}, r > 0$.

Revisão de Cálculo de probabilidades

- Variável aleatória: X_i. Valor observado: x_i.
- Sejam X, Y e Z va's definidas em um mesmo espaço de probabilidade.

$$p(x|y) = \begin{cases} \dfrac{p(x,y)}{p(y)}, & \text{se } p(y) > 0, \\ 0, & \text{se } p(y) \end{cases}$$

$$p(x) = \begin{cases} \sum_y p(x,y), & \text{se y for discreto,} \\ \int_{\Omega_y} p(x,y)dy, & \text{se y for contínuo.} \end{cases}$$

Revisão de Cálculo de probabilidades

- $E_X(X) = E_Y(E_{X|Y}(X|Y))$ e
  $\mathcal{V}_X(X) = \mathcal{E}_Y[\mathcal{V}_{X|Y}(X|Y)] + \mathcal{V}_Y[\mathcal{E}_{X|Y}[X|Y]]$.
- $p(x, y) = p(x|y)p(y) = p(y|x)p(x)$
- X e Y são independentes se e somente se $p(x, y) = p(x)p(y)$
  ($\rightarrow p(x|y) = p(x)$ e $p(y|x) = p(y)$).
- X e Y são condicionalmente independentes dado Z se e somente se
  $p(x, y|z) = p(x|z)p(y|z)$ ($\rightarrow p(x|y, z) = p(x|z)$ e
  $p(y|x, z) = p(y|z)$).