Análise de Dados Categorizados - Aula2

Márcia D Elia Branco

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

Exemplo 4: 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

Modelo Produto de Poisson

$$N_{ij} \sim Poisson(\mu_{ij}) \quad i = 1, \ldots, r \quad j = 1, \ldots, c$$

Sob a suposição de independência, a função de verossimilhança é proporcional a

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

A Logverossimilhança é dada por

$$C + \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \log(\mu_{ij}) - \mu_{ij}$$

Derivando em μ_{ij} e igualando a zero obtemos

$$\frac{n_{ij}}{\mu_{ij}} = 1 \Leftrightarrow \mu_{ij} = n_{ij}$$

Portanto, os emv são $N_{ij} \quad i = 1, \ldots, r \;\; j = 1, \ldots, c.$

Hipótese de multiplicatividade

$$H_0 : \frac{\mu_{1j}}{\mu_{1+}} = \frac{\mu_{2j}}{\mu_{2+}} = \cdots = \frac{\mu_{rj}}{\mu_{r+}} \;,\; j = 1, \ldots, c$$

H_A : Existe pelo menos uma diferente.
Notação:

$$\mu_{i+} = \sum_{j=1}^{c} \mu_{ij} \;,\; \mu_{+j} = \sum_{i=1}^{r} \mu_{ij} \;,\; \mu = \sum_{i=1}^{r} \sum_{j=1}^{c} \mu_{ij}.$$

Modelo Produto de Poisson

A hipótese de multiplicidade pode ser reescrita como

$$H_0 : \mu_{ij} = \frac{\mu_{i+}\mu_{+j}}{\mu} \qquad \forall i, j$$

Sob H_0, $E[N_{ij}] = \frac{\mu_{i+}\mu_{+j}}{\mu}$. Sua estimativa é dada por

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

Analogamente ao caso anterior podemos definir a estatística de teste de Pearson

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

Modelo Produto de Poisson

- Valores grandes de Q_p são indicativos contra H_0 (independência). Portanto indicam associação 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 4 temos $Q_p = 6.5$. Para $\alpha = 0.05$ e $\nu = 1$ temos $\chi^2_{0.05,1} = 3.84$.
- Rejeita-se a independência entre Tipo de Estrada e Limite de Velocidade (valor-P < 0.011)

Modelo Produto de Poisson

No exemplo 4, avaliando as proporções em cada uma das caselas

$$\hat{p}_{11} = 0.038 < \hat{p}_{12} = 0.197 < \hat{p}_{21} = 0.267 < \hat{p}_{22} = 0.498$$

- Maior incidência de acidentes em estradas menores (não auto-estradas) sem limite de velocidade.
- Fixado o fato de estarmos em auto-estrada (primeira coluna) temos que a chance de acidente é $57/8 = 7.12$ para estradas sem limite de velocidade relativamente as estradas com limite.
- Fixando a segunda coluna (não auto-estrada) temos que a chance de acidente é $106/42 = 2.53$ para estradas sem limite de velocidade relativamente as estradas com limite.

Inferência para Proporção

Considere $X \sim Binomial(n, \pi)$ com n conhecido.

- O emv para π é $\hat{\pi} = \frac{X}{n}$ (proporção amostral).
- $E[\hat{\pi}] = \pi$ e $ep(\hat{\pi}) = \sqrt{\frac{\pi(1-\pi)}{n}}$ (erro padrão).
- $Z = \frac{\hat{\pi}-\pi}{ep(\hat{\pi})}$ converge para $N(0,1)$ quando $n \to \infty$.
- O Intervalo de $(1-\alpha)$ x 100 % de confiança (aproximado) para π é dada por

$$[\hat{\pi} - z_\alpha ep(\hat{\pi}); \hat{\pi} + z_\alpha ep(\hat{\pi})]$$

Teste de Hipóteses

$$H_0 : \pi = p_0 \quad versus \quad H_a : \pi \neq p_0$$

Estatística do teste (aproximada normal)

$$Z = \frac{\hat{\pi} - p_0}{ep(\hat{\pi})} \approx N(0,1)$$

Equivalentemente

$$Z^2 \approx \chi^2_1.$$

Região critica:

$$\{|Z| > z_\alpha\} \quad \text{ou} \quad \{Z^2 > q_\alpha\}$$

- **Como estimar o erro padrão?**

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

- 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 fixado em H_0.

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

Exemplo 1: Para cada um dos Vermífugos:
(a) Construir o IC(0.90) para a probabilidade do animal ter verminose (π);
(b) Testar a hipótese $H_0 : \pi = 0.25$.

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

(a) IC(0.90) usando Wald

Para Vermífugo 1: $ep_1 = 0.030$ e

$$\left[\frac{48}{200} - 1.645x0.03; \frac{48}{200} - 1.645x0.03\right] = [0.191; 0.289]$$

Para Vermífugo 2: $ep_2 = 0.033$ e

$$\left[\frac{68}{200} - 1.645x0.033; \frac{68}{200} - 1.645x0.033\right] = [0.286; 0.394]$$

(b) Teste $H_0 : \pi = 0.25 \times H_a : \pi \neq 0.25$.

Região crítica do Teste com $\alpha = 0.1 : \{Q_p > 2.7\}$

Para Vermífugo 1: usando Escore $ep_1 = 0.031$ e a estatística do teste

$$z^2 = \frac{(0.24 - 0.25)^2}{(0.031)^2} = 0.11$$

Decisão: Não rejeita-se H_0.
$Valor - P = 0.74$

Obs: Usando Wald $ep_1 = 0.030$ *e todos os resultados são praticamente iguais.*

Para Vermífugo 2: usando Escore

$$z^2 = \frac{(0.34 - 0.25)^2}{(0.031)^2} = 8.64$$

Decisão: Rejeita-se H_0 .
$Valor - P < 0.004$.

Estatística usando Wald com $ep_2 = 0.033$

$$z^2 = \frac{(0.34 - 0.25)^2}{(0.033)^2} = 7.44$$

$Valor - P < 0.007$.
Mantém-se a decisão de rejeitar H_0.

Inferência para Proporção

- Fazendo uma análise separada das duas amostras, notamos que o Vermífugo 1 parece mais eficiente que o Vermífugo 2. Os IC indicam uma probabilidade de verminose menor para os animais que foram tratados com Vermífugo 1.

- No entanto, para uma comparação mais adequada deveria ser feito um teste de comparação de proporções:
  $H_0 : \pi_1 \geq \pi_2$ *versus* $H_a : \pi_1 < \pi_2$.

- Alternativamente, poderiamos realizar um teste Qui-quadrado de homogeneidade e depois comparar as proporções amostrais.

O Teste da Razão de Verossimilhança: uma alternativa a Wald e Escore

Estatística do teste:

$$2 \log \left(\frac{L_1}{L_0} \right)$$

L_0 é a função de verossimilhança restrita a H_0 e
L_1 é a função de verossimilhança restrita a H_a,
ambas avaliadas no ponto de máximo.

Sob condições de regularidade esta estatística converge para uma distribuição Qui-quadrado com ν graus de liberdades, em que $\nu =$ (parâmetros livres em H_a) - (parâmetros livres em H_0) .

Teste da Razão de Verossimilhança

No caso especial de $H_0 : \pi = p_0$ versus $H_a : \pi \neq p_0$, temos que

$$L_1 \propto \hat{\pi}^x (1 - \hat{\pi})^{n-x}$$

$$L_0 \propto p_0^x (1 - p_0)^{n-x}$$

A estatística do teste é dada por

$$2 \left\{ x \left[\log \hat{\pi} - \log p_0 \right] + (n - x) \left[\log(1 - \hat{\pi}) - \log(1 - p_0) \right] \right\} =$$

$$2 \left[x \log \frac{\hat{\pi}}{p_0} + (n - x) \log \frac{(1 - \hat{\pi})}{(1 - p_0)} \right].$$

Teste da Razão de Verossimilhança

O número de graus de liberdades é $\nu = 1 - 0 = 1$.

No exemplo, para o Vermífugo 2, o valor observado da estatística é

$$2\left[68\log\frac{0.34}{0.25} + 132\log\frac{0.66}{0.75}\right] = 8.07$$

Região crítica do teste com $\alpha = 0.1 : \{Q_{RV} > 2.7\}$.
Portanto, rejeita-se $H_0 : \pi = 0.25$ com $Valor - P < 0.005$.

Teste da Razão de Verossimilhança

- Para modelos normais os três testes (Wald, Escore e RV) são equivalentes.
- Sem a suposição de normalidade sua equivalência ocorre apenas para amostras grandes (assintótico).
- Para amostras moderadas o teste de Wald é o menos confiável.
- Para amostras pequenas deve-se buscar outras alternativas de testes. Testes exatos (binomial) ou Inferência Bayesiana.

Inferência Bayesiana para proporção

Seja $X \sim Binomial(n, \pi)$. Para conduzir a inferência bayesiana é necessário primeiro estabelecer uma distribuição *a priori* para π. Considerando o fato de $0 < \pi < 1$ uma possível proposta é

$$\pi \sim Beta(a, b).$$

A distribuição *a posteriori* é obtida via fórmula de Bayes e será proporcional à

$$f(\pi \mid x) \propto \pi^{a+x-1}(1-\pi)^{b+n-x-1}$$

Resultando em

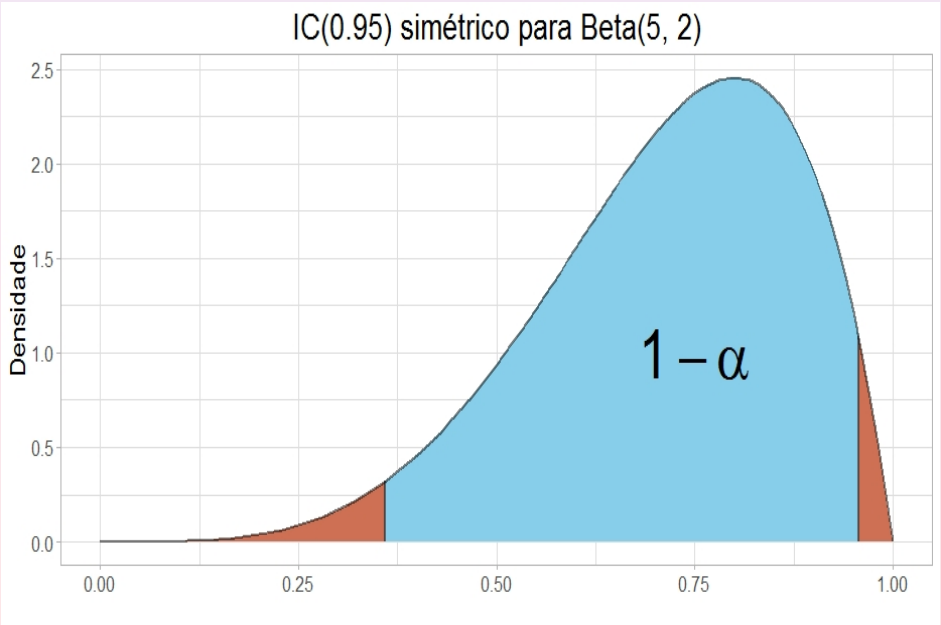
$$\pi \mid x \sim Beta(a + x, b + n - x).$$

Inferência Bayesiana para proporção

Um Intervalo de Credibilidade (ou probabilidade) $(1-\alpha)$ pode ser construído usando os quantis da distribuição *a posteriori*. Assim,

$$[qbeta(\alpha/2, a+x, b+n-x); qbeta(1-\alpha/2, a+x, b+n-x)]$$

Esse intervalo é denominado intervalo de caudas iguais.



- Quando a distribuição *a posteriori* apresenta muita assimetria, o IC de caudas iguais não é o mais adequado e pode ser substituído pelo intervalo HPD (Highest Probability Density) .

Voltando ao exemplo dos Vermífugos.
Vamos considerar $a = b = 1$, resultando em uma distribuição *a priori* Uniforme. Portanto,

$$\pi_1 \mid x_1 \sim Beta(1 + 48, 1 + 152) \quad e \quad \pi_2 \mid x_2 \sim Beta(1 + 68, 1 + 132)$$

As médias *a posteriori* são dadas por

$$E[\pi_1 \mid x_1] = \frac{49}{202} = 0.2426 \approx 0.24 \quad \text{e} \quad E[\pi_2 \mid x_2] = \frac{69}{202} \approx 0.34$$

Valores muito próximos das estimativas de máxima verossimilhança.

As medianas *a posteriori* são dadas por

$$Med[\pi_1 \mid x_1] = qbeta(0.5, 49, 153) = 0.2417$$

$$Med[\pi_2 \mid x_2] = qbeta(0.5, 69, 133) = 0.3411$$

Os limites para os intervalos de probabilidade 0.90 (quantis de ordem 0.05 e 0.95) são

$$qbeta(0.05, 49, 153) = 0.1946 \quad , \quad qbeta(0.95, 49, 153) = 0.2935$$

$$qbeta(0.05, 69, 133) = 0.2877 \quad , \quad qbeta(0.95, 69, 133) = 0.3972$$

Inferência Bayesiana para proporção

Intervalo de credibilidade 0.90 para Vermífugo 1 : $[0.19; 0.29]$.

Intervalo de credibilidade 0.90 para Vermífugo 2: $[0.29; 0.40]$.

- Os valores numéricos dos intervalos são muito próximos dos intervalos clássicos, no entanto, a interpretação é diferente.
- Na inferência clássica esses valores são a realização de uma variável aleatória. O valor do parâmetro π é fixo.
- Na inferência bayesiana, a incerteza sobre π é modelada usando uma distribuição de probabilidades. A probabilidade *a posteriori* de π_1 (π_2) pertencer ao respectivo intervalo é 0.90.
- Para comparação entre os dois Vermífugos pode-se calcular $P(\pi_1 < \pi_2 \mid x_1, x_2)$. Como?