Análise de Dados Categorizados - Aula 5

Márcia D Elia Branco

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

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

- Intervalos de credibilidade $1 - \alpha$ são obtidos usando os quantis da distribuição *a posteriori*.
- Existem dois tipos de IC: caudas iguais e HPD .
- Testes de hipóteses podem ser feitos considerando-se as probabilidades das hipóteses serem verdadeiras ou utilizando os Intervalos de Credibilidade.

Intervalo de credibilidade de caudas iguais de probabilidade $1-\alpha$ para $d = p_{(1)1} - p_{(2)1}$

Considere as distribuições *a priori* : $p_{(1)1} \sim Beta(a_1, b_1)$ e $p_{(2)1} \sim Beta(a_2, b_2)$ independentes.

Usando a fórmula de Bayes, obtemos que as distribuições *a posteriori* também são Betas independentes com parâmetros:

$$A_1 = a_1 + n_{11} \ , \ B_1 = b_1 + n_{12} \ , \ A_2 = a_2 + n_{21} \ , \ B_2 = b_2 + n_{22}.$$

Para construção do IC para d precisamos da sua distribuição *a posteriori* . Não conseguimos obter uma distribuição de probabilidades conhecida para diferença de Betas.

- A distribuição *a posteriori* $f(d \mid n_{11}, n_{21})$ pode ser aproximada via simulação. Método de Monte Carlo.
- O método consiste em simular de cada uma das Betas de forma independente. Para cada par de valores simulados, obter o valor de d.
- Se simularmos uma grande quantidade de valores, as estatísticas amostrais devem se aproximar dos parâmetros dessa distribuição.
- Usamos os quantis da amostra de Monte Carlo para aproximar os quantis populacionais de ordem $\alpha/2$ e $1 - \alpha/2$ e obter o IC aproximado.

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

$$\hat{d} = \frac{48}{200} - \frac{68}{200} = -0.10 \quad \text{e} \quad E[d \mid n_{11}, n_{22}] = \frac{49}{202} - \frac{69}{202} \approx -0.10$$

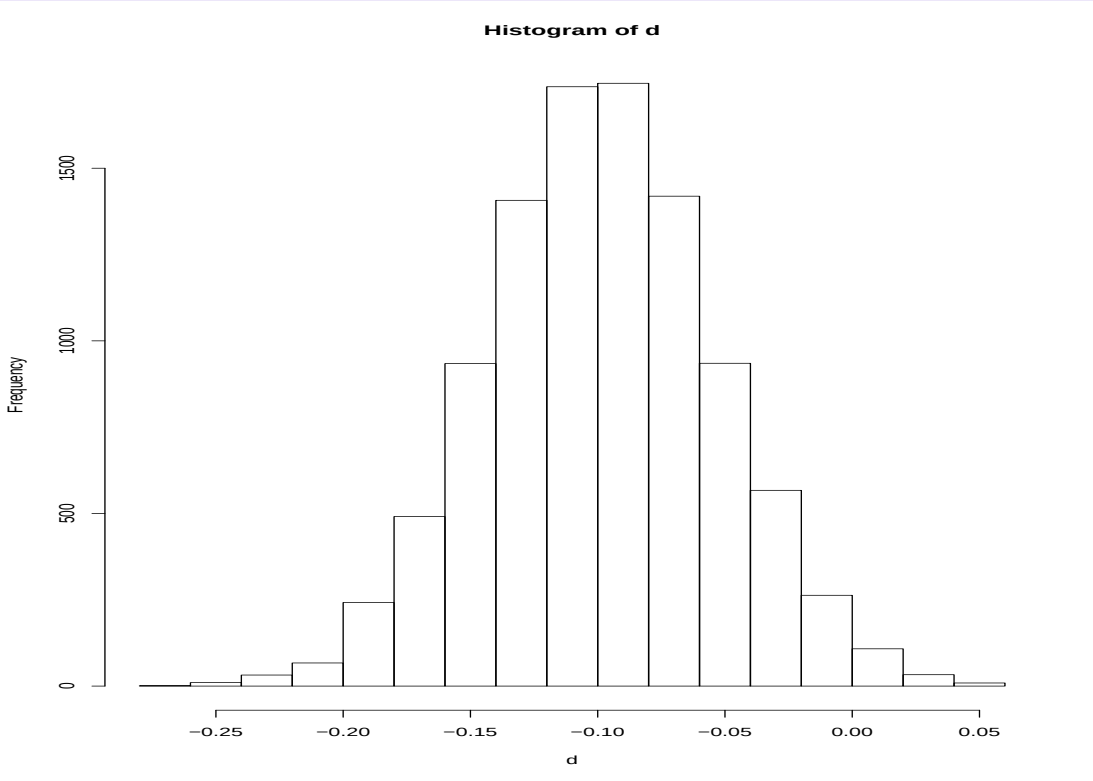
As distribuições *a posteriori* são $p_{(1)1} \mid n_{11} \sim Beta(49, 153)$ e $p_{(2)1} \mid n_{21} \sim Beta(69, 133)$, usando distribuições *a priori* uniformes.

Intervalo de probabilidade 0.90 para d é $[-0.172; -0.024]$.

$P(p_{(1)1} > p_{(2)1} \mid n_{11}, n_{21}) \approx 0.015.$

O codigos em **R** são consistem em :

$> p1 = rbeta(10000, 49, 153)$
$> p2 = rbeta(10000, 69, 133)$
$> d = p1 - p2$
$> quantile(d, c(0.05, 0.95))$
$> mean(p1 > p2)$


Histogram of d
Frequency
0
500
1000
1500
−0.25
−0.20
−0.15
−0.10
−0.05
0.00
0.05
d

Distribuição *a posteriori* **aproximada para** *OR* .

- A IB também tem uma teoria para grandes amostras.
- Diferente da IC, a aproximação normal não é obtida para a distribuição do estimador.
- A aproximação é obtida para a distribuição *a posteriori* do parâmetro θ .
- Sob condições de regularidade $f(\theta \mid x)$ é aproximada por uma $N(Mo, V)$ em que Mo é a moda da posteriori e V é o negativo da segunda derivada da log posteriori no ponto Mo.

Exemplo: Obter a aproximação normal para o logaritmo da chance $\log\left(\frac{\pi}{1-\pi}\right)$.

A função de verossimilhança associada à uma amostra da binomial é proporcional a

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

Vamos considerar *a priori* $Beta(a, b)$ então *a posteriori* é proporcional a

$$\pi^{x+a-1}(1-\pi)^{n-x+b-1}.$$

No entanto, o nossa parâmetro de interesse é $\theta = \log\left(\frac{\pi}{1-\pi}\right)$.

Para obter a distribuição de θ podemos usar o método Jacobiano de transformação de variáveis.

Fazendo a transformação inversa temos que:
$\pi = \frac{e^\theta}{1+e^\theta}$ e $1 - \pi = \frac{1}{1+e^\theta}$.

O Jacobiano da transformação é

$$\frac{d\pi}{d\theta} = \frac{e^\theta}{(1+e^\theta)^2}.$$

Assim

$$f(\theta \mid x) \propto \frac{e^\theta}{(1+e^\theta)^2} \left[\frac{e^\theta}{(1+e^\theta)}\right]^{A-1} \left[\frac{1}{(1+e^\theta)}\right]^{B-1}$$

E

$$\log f(\theta \mid x) = C + A\theta + (A+B)\log(1+e^\theta).$$

Derivando $\log f(\theta \mid x)$ e igualando a zero, obtemos

$$Mo = \log\left(\frac{A}{B}\right) = \log\left(\frac{a+x}{b+n-x}\right).$$

Fazendo a segunda derivada da log-posteriori e substituindo θ por Mo e alterando o sinal, temos

$$V = \frac{1}{A} + \frac{1}{B}.$$

Resulta que

$$\theta \mid x \approx N(Mo, V)$$

Resultado: Sob o modelo produto de binomiais e com prioris Betas independentes, a Distribuição *a posteriori* aproximada para OR é $N(m_{OR}, v_{OR})$ em que

$$m_{OR} = \log\left(\frac{a_1 + n_{11}}{b_1 + n_{12}}\right) - \log\left(\frac{b_2 + n_{22}}{a_2 + n_{21}}\right)$$

$$v_{OR} = \frac{1}{a_1 + n_{11}} + \frac{1}{b_1 + n_{12}} + \frac{1}{a_2 + n_{21}} + \frac{1}{b_2 + n_{22}}$$

Prova: Para mostrar o resultado temos que usar o resultado do exemplo anterior e o fato que diferença de duas v.a. normais independentes é também normal com a média dada pela diferença das médias e variância pela soma das variâncias.

Tabelas $r \times c$

- Os testes de homogeneidade, independência e multiplicatividade usam extensões simples das estatísticas Q_P e Q_{RV}.
- Ambas estatísticas tem distribuição assintótica qui-quadrado com $\nu = (r-1)(c-1)$ graus de liberdades.
- Outras estatísticas de teste podem ser construídas para o caso das variáveis serem do tipo ordinal.
- Um desafio é a busca de medidas de associação em dimensões maiores.

Teste de independência com variáveis ordinais

$$H_0 : p_{ij} = p_{i+}p_{+j} \quad versus \quad H_a : p_{ij} \neq p_{i+}p_{+j}.$$

Se as categorias de respostas de Y e X são ordenáveis, podemos substituir por escores. Considere $u = (u_1, \ldots, u_r)$ e $v = (v_1, \ldots, v_c)$ os escores associados a X e Y respectivamente. O escore médio é dado por

$$\bar{F} = \sum_{i=1}^{r}\sum_{j=1}^{c} u_i v_j p_{ij}$$

Seu estimador é

$$\bar{f} = \sum_{i=1}^{r}\sum_{j=1}^{c} u_i v_j \frac{N_{ij}}{n}$$

Sob H_0,

$$E[\bar{f}] = \sum_{i=1}^{r}\sum_{j=1}^{c} u_i v_j \frac{E[N_{ij}]}{n} = \sum_{i=1}^{r} u_i \frac{n_{i+}}{n} \sum_{j=1}^{c} v_j \frac{n_{+j}}{n} = \mu_u \mu_v$$

$$Var[\bar{f}] = \sum_{i=1}^{r} (u_i - \mu_u)^2 \frac{n_{i+}}{n} \sum_{j=1}^{c} (v_j - \mu_v)^2 \frac{n_{+j}}{n} = \mu_u \mu_v$$

Pelo Teorema do Limite Central

$$M = \frac{\bar{f} - E[\bar{f}]}{\sqrt{Var[\bar{f}]}} \rightarrow N(0,1)$$

Então M^2 tem uma distribuição assintótica qui-quadrado com $\nu = 1$ graus de liberdade. Podemos reescrever

$$M^2 = (n-1)R^2$$

Em que R é o coeficiente de correlação linear de Pearson entre u e v, obtido por

$$R = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c}(u_i - \mu_u)(v_j - \mu_v)n_{ij}}{\left(\sqrt{\sum_{i=1}^{r}(u_i - \mu_u)^2 n_{i+}}\right)\left(\sqrt{\sum_{j=1}^{c}(v_j - \mu_v)^2 n_{+j}}\right)}$$

As hipóteses podem ser reescritas em função do coeficiente de correlação linear populacional ρ. Assim

$$H_0 : \rho = 0 \quad \textit{versus} \quad H_a : \rho \neq 0$$

A região crítica é dada por $RC = \{M^2 > \chi^2_1\}$

Também é possível realizar um teste unilateral com $H_a : \rho > 0$ ou $H_a : \rho < 0$. Neste caso usamos a estatística M e a distribuição normal.

Exemplo : Estudo sobre o uso do tabaco por adolescentes.

Consciência do risco	Uso do tabaco		Totais
	Não	Sim	
Mínima	70	33	103
Moderada	202	40	242
Substancial	218	11	229
Totais	490	84	574

Vamos considerar os escores $u = (1, 2, 3)$ e $v = (0, 1)$.

Tabelas $r \times c$

- O valor de $R = -0.274$ indicando uma associação negativa.
- A estatística do teste bilateral é $M^2 = 42.94$, associada a um Valor-P < 0.0001 . Rejeita-se H_0.
- Conclusão: Há evidência de associação entre consciência de risco e uso do tabaco pelos adolescentes.
- A estatística para o teste unilateral $H_a : \rho < 0$ é dada por $M = 6.55$ com $Valor - P < 10^{-10}$
- Conclusão: O uso do tabaco diminui à medida que a consciência do risco aumenta.

Resíduos padronizados

$$r_{ij} = \frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}(1 - \frac{n_{i+}}{n})(1 - \frac{n_{+j}}{n})}}$$

Sob H_0, r_{ij} tem uma distribuição assistótica $N(0, 1)$.

A análise desses resíduos permite verificar o ajuste dos dados à hipótese H_0 (independência) .

Tarefa: Buscar as funções no **R** que permitem obter as estatísticas de testes e os resíduos.