Análise de Dados Categóricos

Dados Pareados

Enrico A. Colosimo

Departamento de Estatística
Universidade Federal de Minas Gerais

Dados Pareados

Pareamento é frequentemente utilizado quando deseja-se evitar possíveis vícios nos resultados.

Observações pareadas são correlacionadas e, portanto, devem receber tratamento diferenciado.

Duas situações:

1. Mudança (antes e depois) ou pareamento, propriamente dito.
2. Concordância entre observadores.

Caso 1: Exemplos

- Efeito de propaganda eleitoral (antes e depois) para um candidato A.
- Efeito de um filme educativo sobre aborto na opinião (antes e depois) das pessoas.
- Resposta dicotômica (diabete: sim e não) com pareamento por idade e sexo entre fumantes e não fumantes (análise mais complexa).

Caso 1: Eleitores do candidato A

Efeito de propaganda eleitoral para o candidato A.
Duas pesquisas eleitorais foram realizadas em um período de um mês com os mesmos eleitores para avaliar o efeito da propaganda eleitoral para o candidato A.

	Segunda Pesquisa		
Primeira Pesquisa	A	$\bar{A}$	Totais
A	380	70	n_{1+}
$\bar{A}$	150	400	n_{2+}
Totais	n_{+1}	n_{+2}	1000

A proporção de eleitores do candidato A se alterou?
Primeira pesquisa: $\widehat{\pi}_{1+} = \frac{450}{1000}$
Segunda pesquisa: $\widehat{\pi}_{+1} = \frac{530}{1000}$

Teste de McNemar

$$H_0 : \pi_{1+} = \pi_{+1}$$
$$H_1 : \pi_{1+} \neq \pi_{+1}$$

Sob H_0: As caselas discordantes devem ser iguais

$$X_M^2 = \frac{\left(n_{12} - \hat{E}_{12}\right)^2}{\hat{E}_{12}} + \frac{(n_{21} - \hat{E}_{21})^2}{\hat{E}_{21}}, \text{ em que } \hat{E}_{12} = \hat{E}_{21} = \frac{n_{12} + n_{21}}{2}$$

$$X_M^2 = \frac{\left(n_{12} - \frac{n_{12}+n_{21}}{2}\right)^2}{\frac{n_{12}+n_{21}}{2}} + \frac{\left(n_{21} - \frac{n_{12}+n_{21}}{2}\right)^2}{\frac{n_{12}+n_{21}}{2}}$$

$$X_M^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} \sim \chi_1^2 \quad \text{(sob } H_0\text{)}$$

Exemplo

$$X_M^2 = \frac{(150 - 70)^2}{150 + 70} = \frac{6400}{220} = 29,1$$

$$\text{valor-p} = 6,9 \cdot 10^{-8}$$

Teste McNemar

$$H_0 : \pi_{1+} = \pi_{+1} \qquad \text{ou} \qquad H_0 : R = \frac{\pi_{+1}}{\pi_{1+}} = 1$$

	Depois		
Antes	A	Ā	
A	n_{11}	n_{12}	n_{1+}
Ā	n_{21}	n_{22}	n_{2+}
	n_{+1}	n_{+2}	1000

$$\hat{Var}(\log(\widehat{R})) = \frac{(n_{12} + n_{21})}{(n_{12} + n_{11})(n_{21} + n_{11})}$$

Intervalo 95% de confiança para $\frac{P_D}{P_A}$

$$exp\{\log \hat{R} \pm 1,96 \cdot \sqrt{\widehat{Var}(log(\widehat{R}))}\}$$

Para os dados do exemplo:

$$\hat{R} = \frac{\frac{530}{1000}}{\frac{450}{1000}} = 1,18$$

$$\begin{aligned} \hat{Var}(log(\hat{R})) &= \frac{(70+150)}{(70+380)(150+380)} \\ &= \frac{220}{(450)(530)} = 0,0009 = 0,0304^2 \end{aligned}$$

$$\begin{aligned} IC(95\%; \hat{R}) &= exp\{\log 1,18 \pm 1,96 \cdot 0,0304\} \\ &= [1,11 \quad ; \quad 1,25] \end{aligned}$$

A propaganda para o Candidato A aumentou o percentual de eleitores em 18% com um intervalo de 95% de confiança entre 11 e 25%.

Caso 2: Kappa - Concordância entre observadores

Na presença de erro de medida, observadores analisando o mesmo exame podem obter classificações diferentes.

Exemplo: Dois patologistas analisam biópsias de 149 mulheres com o objetivo de classificá-las com relação a existência ou não de um determinado tipo de câncer.

	Patologista 1		
Patologista 2	-	+	
-	71	6	77
+	13	59	72
	84	65	149

Qual é a força de concordância entre os patologistas?

Soluções Extremas

- Caso 01 - Perfeita

Patologista 2	Patologista 1 -	+	
-		0	
+	0		

- Caso 02 - Nula

Patologista 2	Patologista 1 -	+	
-	$\frac{n}{4}$	$\frac{n}{4}$	
+	$\frac{n}{4}$	$\frac{n}{4}$	
			n

Kappa (Colen(1960)): Coeficiente de Concordância k

$$k = \frac{\pi_0 - \pi_e}{1 - \pi_e} \text{ , em que } \quad \pi_0 = \sum_{i=1}^{2} \pi_{ii} \text{ e } \pi_e = \sum_{i=1}^{2} \pi_{i+}\pi_{+i}.$$

em que, $\widehat{\pi}_{ii} = n_{ii}/n$ e $\widehat{\pi}_{i+} = n_{i+}/n$.

O Kappa varia de 0 (tabela nula) a 1 (tabela perfeita).

Para o exemplo dos patologistas:

$$\begin{aligned} \hat{k} &= \frac{\frac{71+59}{149} - \frac{(77\cdot 84)+(72\cdot 65)}{149^2}}{1 - \frac{(77\cdot 84)+(72\cdot 65)}{149^2}} \\ &= 0,74 \end{aligned}$$

Variância Assintótica e Intervalo de Confiança para k

$$\widehat{Var}(\hat{k}) = \frac{A + B - C}{n(1 - \hat{\pi}_e)^2}$$

$$A = \sum_{i=1}^{2} \hat{\pi}_{ii}[1 - (\hat{\pi}_{i+} + \hat{\pi}_{+i})(1 - \hat{k})]^2$$

$$B = (1 - \hat{k})^2 \sum_{i \neq j} \hat{\pi}_{ij}(\hat{\pi}_{i+} + \hat{\pi}_{+j})^2$$

$$C = [\hat{k} - \hat{\pi}_e(1 - \hat{k})]^2$$

$$IC(95\%, k) = \{\hat{k} \pm 1,96 \cdot \sqrt{\hat{Var}(\hat{k})}\}$$

Variância Assintótica e Intervalo de Confiança para k

No exemplo dos patologistas:

$$A = 0,4756$$
$$B = 0,0089$$
$$C = 0,3714$$
$$\hat{Var}(\hat{k}) = 0,00306$$

E o respectivo Intervalo de 95% de Confiança :

$$IC(95\%, \hat{k}) = \{\hat{k} \pm 1,96 \cdot \sqrt{0,0030}\}$$
$$= [0,64 \quad ; \quad 0,85]$$

A concordância entre os patologistas é 74% com um intervalo de 95% de confiança entre (63; 85%).

O pacote vcd do R calcula o Kappa e seu respectivo IC.

- print(K, CI = TRUE)
- value ASE z Pr($>|z|$) lower upper
- 0.7439 0.05463 13.62 3.232e-42 0.6368 0.8509