Análise de Dados Categorizados - Aula 15

Márcia D Elia Branco

Universidade de São Paulo
Instituto de Matemática e Estatística
mdbranco@usp.br - sala 295-A -

Tabelas com amostras dependentes

Exemplo 1: Um estudo foi realizado para investigar a taxa de aprovação de um político antes e após o anúncio de certas medidas.

Antes	Após		Totais
	Aprova	Reprova	
Aprova	20	05	25
Reprova	10	10	20
Totais	30	15	45

Interesse: Comparar as proporções de aprovação antes e depois das medidas.

Tabelas com amostras dependentes

- Em muitas tabelas a suposição de independência entre as amostras não é razoável.
- Um exemplo são estudos de caso-controle em que as amostras são pareadas.
- Outro exemplo é quando temos dados longitudinais, isto é, onde observações de um mesmo indivíduo são realizadas ao longo do tempo.
- Observações dependentes também ocorrem em questionários onde o indivíduo pode responder mais de uma alternativa.
- Iniciamos nosso estudo desse tipo de dados em tabelas 2×2 .

Tabelas 2×2 com amostras pareadas

No exemplo a hipótese de interesse é a de homogeneidade marginal, isto é,

$$H_0 : p_{1+} = p_{+1} \quad versus \quad H_a : p_{1+} \neq p_{+1}$$

Notamos que $p_{1+} = p_{11} + p_{12}$ e $p_{+1} = p_{11} + p_{21}$.
Portanto, a hipótese de homogeneidade pode ser reescrita como

$$H_0 : p_{12} = p_{21} \quad versus \quad H_a : p_{12} \neq p_{21}$$

Definindo um teste de simetria.

Tabelas 2×2 pareadas: Teste de McNemar

- McNemar propos um teste de hipóteses baseado na ideia de que sob H_0, N_{12} tem uma distribuição binomial com probabilidade $\frac{1}{2}$ e número de ensaios $n^* = n_{12} + n_{21}$.
- Neste caso o valor esperado é $\frac{n^*}{2}$ e a variância $n^*(\frac{1}{2})(\frac{1}{2})$.
- Considerando a aproximação normal para binomial, temos $Z \approx N(0,1)$ em que

$$Z = \frac{N_{12} - n^*/2}{\sqrt{n^*/4}} = \frac{2N_{12} - n^*}{\sqrt{n^*}}$$

- Equivalentemente, Z^2 tem distribuição aproximadamente qui-quadrado com 1 grau de liberdade.

Tabelas 2×2 pareadas: Teste de McNemar

A estatística do teste de McNemar é dada por

$$Q_{Mc} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

No exemplo 1, temos que

$$Q_{Mc} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} = \frac{(5 - 10)^2}{5 + 10} = 1.67$$

Considerando a qui-quadrado com 1 g.l., temos valor-P $= 0.1967$.

Conclusão: Não é possível afirmar que a taxa de aprovação do político tenha se alterado após o anúncio das medidas.

Qual é a região critica do teste?

Exemplo 2: 1144 indivíduos foram questionados se para ajudar o meio ambiente estariam dispostos a: (1) pagar maiores impostos; (2) aceitar cortes no padrão de vida. A partir dos dados da tabela abaixo, deseja-se comparar as probabilidades de *Sim* nas duas questões.

Pagar mais impostos	Corte no padrão de vida		Totais
	Sim	Não	
Sim	227	132	359
Não	107	678	785
Totais	334	810	1144

Tabelas 2×2 pareadas: Teste de McNemar

Vamos primeiro testar as hipóteses:

$$H_0 : p_{1+} = p_{+1} \quad \Leftrightarrow \quad p_{12} = p_{21}$$

$$H_a : p_{1+} \neq p_{+1} \quad \Leftrightarrow \quad p_{12} \neq p_{21}$$

A estatística do teste de McNemar resultou em $Q_{Mc} = 2.62$ com valor-P $= 0.1059$.

Logo, não rejeita-se a igualdade entre as probabilidades.
No entanto, a evidência em favor de H_0 é bem fraca (valor-P pequeno).

Para complementar a análise, vamos construir um intervalo de confiança para a diferença $d = p_{1+} - p_{+1}$.

Tabelas 2 $\times$ 2 pareadas: Intervalo

- Um $IC(\gamma)$ para $d = p_{1+} - p_{+1}$ pode ser obtido usando a aproximação assintótica normal para $\hat{d} = \hat{p}_{1+} - \hat{p}_{+1}$.
- O erro padrão do estimador $\hat{d}$ é dado por

$$ep(\hat{d}) = \frac{1}{n}\sqrt{(n_{12} + n_{21}) - (n_{12} - n_{21})^2/n}$$

- Para o exemplo 2, temos que $IC(0.95)$ para diferença das probabilidades marginais é

$$(0.022 - 1.96(0.0135)) = (-0.005, 0.048)$$

- Se as probabilidades diferem, essa diferença é muito pequena.

Ideia da prova para obter o $ep(\hat{d})$ (Feito em Aula!)

$$Var[\hat{d}] = Var[\hat{p}_{1+}] + Var[\hat{p}_{+1}] + 2Cov[\hat{p}_{1+}, \hat{p}_{+1}] \tag{1}$$

Sabemos que

$$Var[\hat{p}_{1+}] = \frac{1}{n}[p_{1+}(1 - p_{1+})] \ \text{ e } \ Var[\hat{p}_{+1}] = \frac{1}{n}[p_{+1}(1 - p_{+1})]$$

Vamos indicar algumas etapas do cálculo da covariância.

$$Cov[\hat{p}_{1+}, \hat{p}_{+1}] = Cov\left[\frac{N_{11} + N_{12}}{n}, \frac{N_{11} + N_{21}}{n}\right]$$

Para n fixado,

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

Usando as propriedades da distribuição multinomial obtemos

$$Cov[\hat{p}_{1+}, \hat{p}_{+1}] = \frac{1}{n}[p_{11}p_{22} - p_{12}p_{21}]$$

Para obtenção do resultado basta voltar a (1) e fazer as devidas simplificações. Resultando em

$$Var[\hat{d}] = \frac{1}{n}[(p_{12} + p_{21}) - (p_{12} - p_{21})^2]$$

Finalmente, deve-se substituir os valores dos parâmetros por suas estimativas para obter uma aproximação dessa variância.

Tabelas 2×2 pareadas: regressão marginal

- Vamos modelar as probabilidades marginais p_{1+} e p_{+1}.
- Considere (Y_1, Y_2) um par de variáveis que será observada para cada unidade amostral. Cada uma delas com apenas dois valores (1 ou 2), com $P(Y_1 = 1) = p_{1+}$ e $P(Y_2 = 1) = p_{+1}$
- O primeiro modelo a ser considerado é com ligação linear. Seja $x_t = 1$ se $t = 1$ e $x_t = 0$ se $t = 2$ e

$$P(Y_t = 1) = \alpha + \delta x_t$$

- Para este modelo $\delta = P(Y_1 = 1) - P(Y_2 = 1) = p_{1+} - p_{+1}$.
- Testar a hipótese de homogeneidade marginal é equivalente a testar $H_0 : \delta = 0$.

Tabelas 2×2 pareadas: regressão marginal

- Alternativamente podemos modelar os logitos. Assim

$$logito[P(Y_t = 1)] = \alpha + \beta x_t$$

- Neste caso, β representa o logaritmo da razão de chances.
- Note que $\beta = 0 \Leftrightarrow OR = 1$ e então $p_{1+} = p_{+1}$.
- O modelo de regressão marginal não pertence aos MLG e não pode ser ajustado usando a função *glm* do **R**.
- Para a inferência nesse tipo de modelos é necessário o uso de equações de estimação generalizadas.
- Para implementação no **R** usamos *library(gee)* e a função *gee* .

Tabelas 2 × 2 pareadas: regressão marginal

```
> Opinions <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Opinions.dat",
+                          header=TRUE)
> Opinions # data file at text website has 2 lines for each person
     person question y # y variable is y1 when question=1, y2 when question=0
1         1        1 1 # y1 for person 1
2         1        0 1 # y2 for person 1
3         2        1 1
4         2        0 1
...
2287   1144        1 0 # y1 for person 1144
2288   1144        0 0 # y2 for person 1144
> library(gee)
> fit <- gee(y ~ question, id=person, family=binomial(link="identity"),
+            data=Opinions) # id identifies variable on which observe y1, y2
> summary(fit) # question para. for identity link is difference of proportions
             Estimate  Naive S.E.   Naive z  Robust S.E.  Robust z
(Intercept)   0.29196     0.01345  21.70970      0.01344  21.71920
question      0.02185     0.01922   1.13725      0.01350   1.61897

> fit2 <- gee(y ~ question, id=person, family=binomial(link=logit),
+             data=Opinions)
> summary(fit2) # question parameter for logit link is log odds ratio
             Estimate  Naive S.E.    Naive z  Robust S.E.   Robust z
(Intercept)  -0.88589     0.06506  -13.61740      0.06503  -13.62336
question      0.10353     0.09108    1.13674      0.06398    1.61824
```

Tabelas 2×2 pareadas: regressão marginal

- A saída do **R** apresenta resultados para os dados do Ex2.
- Y_1 = aumento de imposto e Y_2 = corte no padrão de vida.
- Para o ajuste linear temos $\hat{\delta} = 0.02185$. O mesmo valor obtido anteriormente.
- Também são apresentados dois valores de erros padrões: 0.01922 (Naive) e 0.0135 (Robusto). O erro padrão Naive não considera a dependência entre as observações da mesma pessoa. O erro padrão robusto é o mais apropriado.
- Para o ajuste usando o logito temos $\hat{\beta} = 0.10353$. Portanto, $exp(0.10353) = 1.11$ é a estimativa da razão de chances.
- A chance da pessoa aceitar pagar mais imposto é 11 % maior que a de aceitar cortes no seu padrão de vida.

Tabelas $c \times c$ pareadas: regressão marginal

- Para c> 2 , homogeneidade marginal é caracterizada por

$$P(Y_1 = i) = P(Y_2 = i) \quad i = 1, \ldots, c$$

- O modelo de regressão baseado nos logitos categorias de referência marginais, $j = 1, \ldots, c-1$, são dados por

$$\log\left[\frac{P(Y_1 = j)}{P(Y_1 = c)}\right] = \alpha_j + \beta_j \quad , \quad \log\left[\frac{P(Y_2 = j)}{P(Y_2 = c)}\right] = \alpha_j$$

- Testar a homogeneidade marginal é equivalente a testar $H_0 : \beta_1 = \cdots = \beta_{c-1} = 0$.
- A implementação desse teste é obtida com o uso de equações de estimação generalizadas (GEE).

Tarefa 1: Verifique se para tabelas quadradas com c>2 a hipótese de homogeneidade é equivalente a hipótese de simetria. Isto é,

$$H_0 : p_{i+} = p_{+i} \; \forall i \;\; \Leftrightarrow \;\; H_0 : p_{ij} = p_{ji} \quad i \neq j$$

Tarefa 2: Estudar o exemplo da seção 8.3.2, pag. 235/236 do livro do Agresti.

Exemplo 3: 149 pacientes com esclerose multipla foram avaliados por 2 neurologistas em relação ao estágio da sua doença. Para cada paciente foi atribuída uma nota de 1 à 4, em que 1 representa o estágio inicial e 4 um estágio avançado.
Verifique se há concordância de opinião entre os neurologistas.

N1	N2				Total
	1	2	3	4	
1	38	5	0	1	44
2	33	11	3	0	47
3	10	14	5	6	35
4	3	7	3	10	23
Total	84	37	11	17	149

O coeficiente de concordância *Kappa* é dado por

$$\kappa = \frac{\Pi_0 - \Pi_E}{1 - \Pi_E}$$

em que

$$\Pi_0 = \sum_{i=1}^{c} p_{ii} \qquad \text{é a probabilidade de concordância.}$$

$$\Pi_E = \sum_{i=1}^{c} (p_{i+})(p_{+i}) \qquad \text{é a probabilidade esperada sob independência}$$

A estimativa desse coeficiente, $\hat{\kappa}$, é obtida substituindo as probabilidades pelas frequências relativas observadas.

No exemplo 3, temos que

$$\hat{\Pi}_0 = \sum_{i=1}^{4} \frac{n_{ii}}{n} = \frac{38 + 11 + 5 + 10}{149} = 0.43$$

$$\hat{\Pi}_E = \sum_{i=1}^{4} \frac{(n_{i+})(n_{+i})}{n^2} = \frac{(44)(84) + (47)(37) + (35)(11) + (23)(17)}{149^2}$$

$$\hat{\Pi}_E = 0.28$$

Resulta em

$$\hat{\kappa} = \frac{0.43 - 0.28}{1 - 0.28} = 0.208$$

Tabelas $c \times c$ pareadas: coeficiente de concordância

Tarefa 3:

- O que ocorre com $\hat{\kappa}$ se sempre houver concordância?
- E se nunca houver concordância?
- Busque informações para construção de um IC para κ e completa o exemplo construindo esse intervalo.