Análise de Dados Categorizados - Aula 5

Márcia D Elia Branco

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

Inferência em Tabelas $r \times c$

- Os testes de homogeneidade, independência e multiplicatividade usam extensões simples das estatísticas de Pearson e Razão de Verossimilhanças:

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

$$Q_{RV} = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} \log\left(\frac{n_{ij}}{e_{ij}}\right)$$

- 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.

Inferência em Tabelas $r \times c$: variáveis ordinais

Se as categorias de respostas de Y e X são ordenáveis, podemos substituir por escores.

Considere $u_1 \leq u_2 \leq \cdots \leq u_r$ e $v_1 \leq v_2 \leq \cdots \leq v_c$ os escores associados a X e Y respectivamente.

Seja R o coeficiente de correlação linear de Pearson entre u e v, obtido por

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

Em que

$$\bar{u} = \sum_{i=1}^{r} u_i \frac{n_{i+}}{n} \quad \text{e} \quad \bar{v} = \sum_{j=1}^{c} v_j \frac{n_{+j}}{n}$$

Considere $M^2 = (n-1)R^2$, esta estatística tem uma distribuição assintótica qui-quadrado com $\nu = 1$ graus de liberdade e pode ser usada para testar coeficiente de correlação linear populacional, denotado por ρ.

Assim, o teste de correlação é estabelecido

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

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

Tabelas $r \times c$

- Também é possível realizar um teste unilateral com $H_a : \rho > 0$ ou $H_a : \rho < 0$.
- Nestes casos usamos a estatística $M = \sqrt{M^2}$ que tem distribuição assintótica normal padrão.
- A região crítica é dada por $\{M > z_{1-\alpha}\}$.
- A conclusão irá depender do sinal do coeficiente de correlação R.

Exemplo 1: 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 $v = (0, 1)$.

- 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.

Razão de chances de caselas adjacentes

$$OR_{ij} = \frac{p_{ij}p_{(i+1)(j+1)}}{p_{i(j+1)}p_{(i+1)j}}$$

A estimativa é dada por

$$\hat{OR}_{ij} = \frac{n_{ij}n_{(i+1)(j+1)}}{n_{i(j+1)}n_{(i+1)j}}$$

$i = 1, \ldots, r-1 \quad \text{e} \quad j = 1, \ldots, c-1$

Exemplo 2: Em um estudo sobre tratamento de artrite reumatóide observou-se os seguintes resultados.

Tratamento	Melhora			Totais
	Nenhuma	Alguma	Acentuada	
Placebo	29	07	07	43
Ativo	13	07	21	41
Totais	42	14	28	84

$$\hat{OR}_{11} = \frac{29 \times 7}{7 \times 13} = 2.23$$

A chance de nenhuma melhora relativa a haver alguma melhora no grupo placebo é aproximadamente 2 vezes a mesma chance no grupo que recebeu o tratamento.

$$\hat{OR}_{12} = \frac{7 \times 21}{7 \times 7} = 3$$

A chance de haver alguma melhora relativa a melhora acentuada no grupo placebo é 3 vezes a mesma chance no grupo que recebeu o tratamento.

Outras razões de chances podem ser obtidas, como por exemplo, comparar *(Nenhum)* com *(Alguma + Acentuada)*

$$\hat{OR}_{1(2+3)} = \frac{29 \times 28}{14 \times 13} = 4.46$$

A chance de nenhuma melhora no grupo placebo é aproximadamente 4.5 vezes a chance de nenhuma melhora no grupo que recebeu o tratamento.

1. O coeficiente de contingência

É baseado na estatística do teste qui-quadrado de Pearson Q_P e dado por

$$C = \sqrt{\frac{Q_P}{Q_P + n}}$$

- $C = 0$ indica não associação.
- Esse coeficiente é limitado superiormente pelo valor

$$C_{max} = \sqrt{\frac{min(r-1, c-1)}{1 + min(r-1, c-1)}}$$

2. Coeficiente de Incerteza

$$U = \frac{-\sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} \log\left(\frac{p_{ij}}{p_{i+}p_{+j}}\right)}{\sum_{j=1}^{c} p_{+j} \log(p_{+j})}$$

- $0 \leq U \leq 1$. $U = 0$ representa não associação.
- A motivação para construção dessa medida é o coeficiente de determinação R^2 que mede a porcentagem da variação de Y que é explicada por X.
- Para medir variabilidade de uma variável categórica usa-se a medida de entropia.

Motivação para construção de U

Lembre que R^2 pode ser escrita como

$$R^2 = \frac{Var(Y) - E_X[Var(Y \mid X)]}{Var(Y)}$$

e interpretada como a porcentagem da variação de Y que é explicada por X .

No nosso contexto, Y e X são qualitativas. Temos que usar outra medida de variabilidade para substituir a variância. Usamos a medida de entropia associada a distribuição de probabilidades.

Medidas de associação em Tabelas $r \times c$

Entropia marginal de Y :

$$En_Y = -\sum_{j=1}^{c} p_{+j} \log(p_{+j}) \tag{1}$$

Entropia condicional de Y dado $X = i$

$$En_{Y|X=i} = -\sum_{j=1}^{c} \frac{p_{ij}}{p_{i+}} \log\left(\frac{p_{ij}}{p_{i+}}\right)$$

A entropia média considerando todos valores de X é dada por

$$-\sum_{i=1}^{r} p_{i+} \sum_{j=1}^{c} \frac{p_{ij}}{p_{i+}} \log\left(\frac{p_{ij}}{p_{i+}}\right) \tag{2}$$

Subtraindo (1) de (2) temos

$$\sum_{i=1}^{r} p_{i+} \sum_{j=1}^{c} \frac{p_{ij}}{p_{i+}} \log\left(\frac{p_{ij}}{p_{i+}}\right) - \sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} \log(p_{+j})$$

Simplificando, obtemos

$$\sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} \log\left(\frac{p_{ij}}{p_{i+}p_{+j}}\right) \tag{3}$$

Dividindo (3) por (1) o resultado segue.

A estimativa do coeficiente de incerteza é dada por:

$$\hat{U} = \frac{-\sum_{i=1}^{r}\sum_{j=1}^{c} \frac{n_{ij}}{n} \log\left(\frac{n_{ij}n}{n_{i+}n_{+j}}\right)}{\sum_{j=1}^{c} \frac{n_{+j}}{n} \log\left(\frac{n_{+j}}{n}\right)}$$

3. Coeficiente de Spearman

- Considere $u_1 < u_2 < \cdots < u_n$ e $v_1 < v_2 < \cdots < v_n$ os postos associados a X e Y respectivamente.
- Supondo que não tenha empates, o coeficiente de Spearman é dado por

$$\rho = 1 - \frac{6\sum_{i=1}^{n}(u_i - v_i)^2}{n(n^2-1)}$$

- $-1 \leq \rho \leq 1$ e tem a mesma interpretação do coeficiente de Pearson.
- Questões:
  (a) Como adaptar para tabelas de dupla entrada?
  (b) Relacionar com a correlação de Pearson.

4. Coeficientes de Concordância

4.1. Coeficiente de concordância Gama de Goodman:

$$\frac{C - D}{C + D}$$

em que C é o número de pares concordantes e D o número de pares discordantes.

4.2. Coeficiente de concordância Tau-b de Kendall:

$$\tau_b = \frac{C - D}{\sqrt{(C + D + E_X)(C + D + E_Y)}}$$

em que E_X é o número de empares na variável X e E_Y é o número de empates na variável Y.

Exemplo 3: Considere os dados hipotéticos na tabela abaixo. Y representa o grau de satisfação com determinado projeto e X a classe econômica do indivíduo.

Classe Economica	Satisfação			
	NãoS	PoucoS	ModeradoS	Satisfeito
Baixa	01	03	10	06
MBaixa	02	03	10	07
Média	01	06	14	12
Alta	00	01	09	11

$$C = 1331 \quad D = 849$$

$$\hat{\gamma} = \frac{1331 - 849}{1331 + 849} = 0.221$$

Tabelas $q \times r \times c$

Em tabelas $q \times 2 \times 2$, podemos indicar independência usando a razão de chances.

Se X e Y são independentes (marginalmente) então

$$P(Y = i \mid X = 1) = P(Y = i \mid X = 2) \quad i = 1, 2.$$

Logo,

$$\left[\frac{P(Y = 1 \mid X = 1)}{P(Y = 2 \mid X = 1)}\right] = \left[\frac{P(Y = 1 \mid X = 2)}{P(Y = 2 \mid X = 2)}\right]$$

Portanto,

$$OR_{XY} = 1$$

A razão de chances condicional à $Z = k$ é denotada por $OR_{(k)XY}$ e dada por

$$OR_{(k)XY} = \left[\frac{P(Y = 1 \mid X = 1, Z = k)}{P(Y = 2 \mid X = 1, Z = k)}\right] \left[\frac{P(Y = 2 \mid X = 2, Z = k)}{P(Y = 1 \mid X = 2, Z = k)}\right]$$

Razões de chances condicionais iguais a 1, para todo k, não implicam em razão de chances marginal igual a 1 .

Exemplo 4 : O objeto é verificar o efeito do tipo de tratamento na eficência da cura do cálculo renal. Uma terceira variável foi considerada na análise, o tamanho da pedra.

Tratamento	Pedra grande		Pedra pequena	
	Cura	Não cura	Cura	Não cura
A	192	71	81	06
B	55	25	234	36

Y : resultado do tratamento (0 ou 1) ,

X : tipo de tratamento (0 ou 1) e

Z : tamanho da pedra renal (0 ou 1)

Razão de chance para pedra grande (condicional a Z=0):

$$\hat{OR}_{XY(0)} = \frac{192 \times 25}{71 \times 55} = 1.23$$

Razão de chance para pedra pequena (condicional a Z=1):

$$\hat{OR}_{XY(1)} = \frac{81 \times 36}{6 \times 234} = 2.08$$

Ambas medidas favorecem o tratamento A. No entanto, para cálculo com pedra pequena o aumento na chance de cura dado pelo tratamento A é mais evidente.

Suponha agora que a variável Z não tenha sido considerada.

Tabelas $q \times r \times c$

Podemos agrupar os valores em uma única tabela 2×2 e obtemos

Tratamento	Resultado	
	Cura	Não cura
A	273	77
B	289	61

Portanto,

$$\hat{OR}_{XY} = \frac{273 \times 61}{77 \times 289} = 0.75$$

Com os dados agrupados a chance de cura com o tratamento A é menor que com o tratamento B.

Paradoxo!!

Tabelas $q \times r \times c$

Entendendo o problema.

A tabela a seguir relaciona X e Z

X	Z	
	Grande	Pequena
A	263	87
B	80	270

A razão de chances é

$$\hat{OR}_{XZ} = \frac{263 \times 270}{87 \times 80} = 10.2$$

Evidência de forte associação entre as variáveis.

Tabelas $q \times r \times c$: Paradoxo de Simpson

- Z é denominada variável de confundimento. Na verdade, é a variável importante que não está sendo considerada na tabela conjunta e que influe a decisão final.
- Esse fenômeno é conhecido como Paradoxo de Simpson.
- No exemplo 1 ocorre uma mudança no sentido da associação.
- Há variações desse paradoxo, uma delas refere-se a independência e será analisado a seguir.

Independência condicional versus marginal

Considere a notação $p_{kij} = P(Z = k, X = i, Y = j)$.

Independência marginal entre X e Y:

$$P(X = i, Y = j) = P(X = i)P(Y = j) = p_{+i+}p_{++j} \quad \forall i, j$$

Independência condicional entre X e Y dado $Z = k$:

$$P(X = i, Y = j \mid Z = k) = P(X = i \mid Z = k)P(Y = j \mid Z = k) =$$

$$= p_{(k)i+}p_{(k)+j}$$

Independência condicional versus marginal

Exemplo 5: Considere uma tabela $2 \times 2 \times 2$ com probabilidades :

$p_{(1)11} = 0.36$, $p_{(1)12} = p_{(1)21} = 0.24$ e $p_{(1)22} = 0.16$
$p_{(2)11} = 0.04$, $p_{(2)12} = p_{(2)21} = 0.16$ e $p_{(2)22} = 0.64$
$p_{1++} = p_{2++} = P(Z = 1) = P(Z = 2) = 0.5$,

Condicional à $Z = 1$, temos que

$$p_{(1)11} = 0.36 = 0.6 \times 0.6 = p_{(1)1+} \times p_{(1)+1}$$

Analogamente para as outras probabilidades.

Assim, X e Y são independentes dado $Z = 1$.

Independência condicional versus marginal

Condicional à $Z = 2$, temos que

$$p_{(2)11} = 0.04 = 0.2 \times 0.2 = p_{(2)1+} \times p_{(2)+1}$$

Analogamente para as outras probabilidades.

Assim, X e Y são independentes dado $Z = 2$.

Além disso, temos que:

$p_{+11} = 0.36 \times 0.5 + 0.04 \times 0.5 = 0.20.$
$p_{+12} = p_{+21} = 0.24 \times 0.5 + 0.16 \times 0.5 = 0.20.$
$p_{+22} = 0.16 \times 0.5 + 0.64 \times 0.5 = 0.40.$

Independência condicional versus marginal

Também

$p_{+1+} = (0.36 + 0.24) \times 0.5 + (0.04 + 0.16) \times 0.5 = 0.40$ e
$p_{++1} = (0.36 + 0.24) \times 0.5 + (0.04 + 0.16) \times 0.5 = 0.40$

Portanto,

$$p_{+11} = 0.2 \neq 0.4 \times 0.4$$

Resulta que X e Y não são independentes marginalmente.

Observação: Independência condicional não implica em independência marginal