USP – ICMC – SME0811 - Análise de Dados Categorizados

1ª prova – 1º/2016 – S O L U Ç Ã O

1. Em uma tabela bidimensional $I \times J$, $I \geq 2$ e $J \geq 2$, as probabilidades são $\pi_{ij} = P(X = i, Y = j) = \dfrac{1}{IJ}$, $i = 1, \dots, I$ e $j = 1, \dots, J$.
   Prove que as variáveis X e Y são independentes.
   Solução. Como as probabilidades π_{ij} são constantes e iguais a $\frac{1}{IJ}$, a distribuição marginal de X tem probabilidades
   $$\pi_{i+} = \sum_{j=1}^{J} \pi_{ij} = \underbrace{\frac{1}{IJ} + \cdots + \frac{1}{IJ}}_{J \text{ colunas}} = \frac{1}{I}, \text{ para } i = 1, \dots, I.$$
   Analogamente, a distribuição marginal de Y tem probabilidades
   $$\pi_{+j} = \sum_{i=1}^{I} \pi_{ij} = \underbrace{\frac{1}{IJ} + \cdots + \frac{1}{IJ}}_{I \text{ linhas}} = \frac{1}{J}, \text{ para } j = 1, \dots, J.$$
   Portanto, $\pi_{ij} = \pi_{i+}\pi_{+j}$, para $i = 1, \dots, I$, $j = 1, \dots, J$, e concluímos que X e Y são independentes.

2. Considere uma tabela de contingências 2×2. Representando $\pi_{1|2}$ no eixo horizontal e $\pi_{1|1}$ no eixo vertical, apresente os gráficos correspondentes a cada uma das seguintes situações: (a) risco relativo $= 1/2$, (b) razão de chances $= 1/2$ e (c) diferença de proporções $= -1/2$.
   Solução. (a) Da relação $RR = \pi_{1|1}/\pi_{1|2} = 1/2$ obtemos $\pi_{1|1} = \pi_{1|2}/2$.
   (b) Da relação
   $$RC = \frac{\dfrac{\pi_{1|1}}{1 - \pi_{1|1}}}{\dfrac{\pi_{1|2}}{1 - \pi_{1|2}}} = \frac{1}{2}$$
   obtemos
   $$2\pi_{1|1}(1 - \pi_{1|2}) = \pi_{1|2}(1 - \pi_{1|1}),\ 2\pi_{1|1} - 2\pi_{1|1}\pi_{1|2} = \pi_{1|2} - \pi_{1|2}\pi_{1|1} \quad \text{e} \quad \pi_{1|1} = \frac{\pi_{1|2}}{2 - \pi_{1|2}}.$$
   (c) Da relação $D_{12} = \pi_{1|1} - \pi_{1|2} = -1/2$ obtemos $\pi_{1|1} = \pi_{1|2} - 1/2$. As expressões são representadas na Figura 1. No item (b) a função é não linear e a curva pode ser aproximada selecionando alguns valores de $\pi_{1|2}$, por exemplo, $\{0, 1/4, 1/2, 3/4, 1\}$.

3. Em uma tabela de contingências 2×2 com variáveis X e Y, prove que se o risco relativo é igual a 1, então X e Y são independentes.
   Solução. Se $RR = 1$, então $\pi_{1|1} = \pi_{1|2}$, ou seja, $P(Y = 1|X = 1) = P(Y = 1|X = 2) = a$, de modo que $1 - P(Y = 1|X = 1) = 1 - P(Y = 1|X = 2) = 1 - a$, ou seja, $P(Y = 2|X = 1) = P(Y = 2|X = 2) = 1 - a$. Portanto, a distribuição condicional de $Y|X = 1$ coincide com a distribuição condicional de $Y|X = 2$. Já pode ser afirmado que X e Y são independentes.
   Para obter a distribuição marginal de Y calculamos
   $$\begin{aligned} P(Y = 1) &= P(Y = 1|X = 1)P(X = 1) + P(Y = 1|X = 2)P(X = 2) \\ &= aP(X = 1) + aP(X = 2) = a. \end{aligned}$$

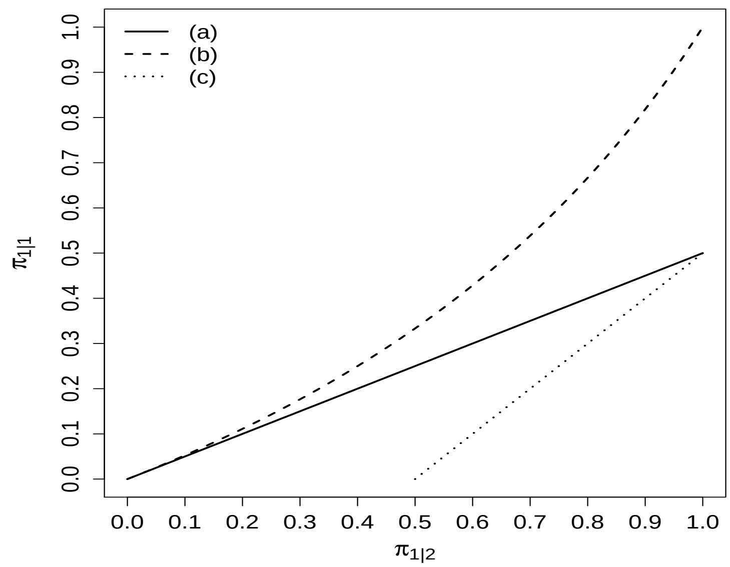

Figura 1: Gráficos da questão 2.

Tabela 1: Dados para a questão 4.

País	Cônjuge fumante (X)	Casos ($Y = 1$)	Controles ($Y = 2$)
Japão	Não	21	82
	Sim	73	188
Inglaterra	Não	5	16
	Sim	19	38
EUA	Não	71	249
	Sim	137	363

Logo, $P(Y = 2) = 1 - a$. Assim, as distribuições marginal de Y e condicional de $Y|X$ são as mesmas, provando que X e Y são independentes.

4. A Tabela 1 mostra as contagens de estudos caso-controle em três diferentes países relacionando fumantes passivos (ou seja, não fumantes cônjuges de fumantes) e câncer de pulmão (caso: fumante passivo com câncer, controle: fumante passivo sem câncer).
Comente a associação entre X e Y.

Solução. As estimativas da razão de chances nas tabelas parciais são dadas por

$$\widehat{RC}_{XY(\text{Japão})} = \frac{21 \times 188}{73 \times 82} = 0,660, \quad \widehat{RC}_{XY(\text{Inglaterra})} = \frac{5 \times 38}{19 \times 16} = 0,625 \tag{1}$$

e

$$\widehat{RC}_{XY(\text{EUA})} = \frac{71 \times 363}{137 \times 363} = 0,756. \tag{2}$$

Nas três tabelas parciais a estimativa da razão de chances é menor do que 1. Significa que comparando pessoas tendo cônjuge não fumante e pessoas tendo cônjuge fumante, há uma redução na chance de câncer de pulmão em não fumantes, sendo menor no estudo realizado nos EUA.

Pelas estimativas em (1) e (2), o padrão de associação não varia tanto entre as tabelas parciais. Passamos ao cálculo de uma estimativa comum para a razão de chances. Da Tabela 1 obtemos $n_{++1} = 364$ (Japão), $n_{++2} = 78$ (Inglaterra) e $n_{++3} = 820$ (EUA). Utilizando estes totais e a Tabela 1, a estimativa de Mantel-Haenszel para a razão de chances comum é

$$\widehat{RC}_{\text{MH}} = \frac{\sum_{k=1}^{3} n_{11k}n_{22k}/n_{++k}}{\sum_{k=1}^{3} n_{12k}n_{21k}/n_{++k}} = \frac{\frac{21 \times 188}{364} + \frac{5 \times 38}{78} + \frac{71 \times 363}{820}}{\frac{73 \times 82}{364} + \frac{19 \times 16}{78} + \frac{137 \times 249}{820}} = 0,721.$$

Poderiam ser adotadas estimativas corrigidas adicionando 1/2 a cada frequência da Tabela 1.

5. A Tabela 2 mostra as contagens obtidas em um estudo transversal relacionando o exame de mamografia e a detecção de câncer de mama em mulheres. Foi perguntado quão provável a realização de uma mamografia poderia identificar um novo caso de câncer de mama. Classifique as variáveis e apresente a estimativa de uma medida de associação que você considerar adequada.

Tabela 2: Dados para a questão 5.

Realização de mamomografia	Detecção de câncer de mama		
	Improvável	Algo provável	Muito provável
Nunca	13	77	144
Há mais de um ano	4	16	54
No ano passado	1	12	91

Solução. Como o estudo é transversal e nenhuma outra informação foi apresentada, podemos afirmar que as duas variáveis são resposta. As duas variáveis são ordinais com níveis crescentes na Tabela 2.

Como as variáveis são ordinais, a medida de associação gama (γ) de Goodman e Kruskal pode ser utilizada. Os números de pares concordantes (C) e discordantes (D) são

$$C = 13 \times (16 + 54 + 12 + 91) + 77 \times (54 + 91) + 4 \times (12 + 91) + 16 \times 91 = 15282$$

e

$$D = 77 \times (4 + 1) + 144 \times (4 + 16 + 1 + 12) + 16 \times 1 + 54 \times (1 + 12) = 5855.$$

A estimativa de γ é dada por $\widehat{\gamma} = (C - D)/(C + D) = 0,446$, indicando uma associação positiva, ou seja, quanto mais frequente a realização de uma mamografia, mais alta a percepção de que um novo caso de câncer de mama pode ser detectado.

A medida de associação tau (τ) de Goodman e Kruskal também poderia ser utilizada.
A associação entre as duas variáveis poderia ser verificada testando a independência com as estatísticas X^2 ($= 24,1$) e G^2 ($= 26,8$), tendo 4 graus de liberdade (conclusão?).