ADC – Aula 2:Independência em Tabelas de Contingência Bidimensionais

1. Motivação

- Temos 2 variáveis: classe social e rádio preferida.
- Classe social: variável categórica ordinal, resultado da aplicação de inquérito de opinião, com adoção do critério ABEP (2003).
- Rádio preferida: variável categórica nominal, resultado da aplicação de inquérito de opinião.
- Vamos à tabela com os dados de nosso exemplo.

Classe Social	Rádio Preferida						$n_{i.}$
	Solar	Globo	Itatiaia	Cidade	Pio XII	Outras	
A	48	10	9	6	12	14	**99**
B	119	24	20	8	10	16	**197**
C	68	5	3	4	1	2	**83**
n_{+j}	**235**	**39**	**32**	**18**	**23**	**32**	$n_{++} = 379$

- Tabela *Bivariada* 3 × 6. Variável de linha *categórica ordinal* com *l* = 3 categorias (resultado da aplicação de inquérito de opinião, com adoção do critério ABEP (2003)); variável de coluna *categórica nominal* com *c* = 6 categorias (resultado da aplicação de inquérito de opinião).
- Categorias **mutuamente exclusivas** e **exaustivas**.
- Amostragem **probabilística** com tamanho de amostra fixado (*n* = 379).
- Indagação motivadora da pesquisa: *Há <u>independência</u> entre a classe social do respondente e sua rádio preferida?*

2. A estatística de Pearson

- Quando a distribuição populacional satisfaz a hipótese de independência estatística H_0: $\pi_{ij} = \pi_{i+} \times \pi_{+j}$ os valores observados podem se desviar da mesma devido à viés de amostragem. Foi com este objetivo (avaliar a magnitude do viés) que Pearson, no início do século XX, propôs a seguinte medida de desvio:

$$\chi^2 = N \sum_{i \in I} \sum_{j \in J} \frac{\left(p_{ij} - \pi_{i+}\pi_{+j}\right)^2}{\pi_{i+}\pi_{+j}}$$

A distribuição desta estatística converge para a distribuição teórica qui-quadrado com ($\ell \times c$) -1 graus de liberdade, quando a amostragem é feita de acordo com o modelo de distribuição multinomial.

- Quando as probabilidades marginais π_{i+} e π_{+j} são desconhecidas, o que é comum, podemos utilizar suas estimativas de MV: p_{i+} e p_{+j}, respectivamente, obtidas da amostra $N = n_{++}$.

Desta forma, a estatística χ^2 apresentada anteriormente se transforma em $N\mathrm{X}^2$, onde

$$\mathrm{X}^2 = \sum_{i \in I} \sum_{j \in J} \frac{\left(p_{ij} - p_{i+}p_{+j}\right)^2}{p_{i+}p_{+j}}$$

- A distribuição de $N\mathrm{X}^2$, sob as condições padrão de amostragem multinomial, converge assintóticamente para a distribuição qui-quadrado com (ℓ - 1)(c – 1) graus de liberdade, devido aos estimadores p_{i+} e p_{+j}, respectivamente. É esta convergência que permite, **em amostras suficientemente grandes**[1], realizar o teste de hipótese de independência entre as variáveis.

O número de graus de liberdade é dado pelo número de termos independentes da relação acima, dado que os totais marginais de linha e de coluna são conhecidos. O número total de células é $\ell \times c$, o conhecimento dos totais de linhas e de colunas limita o no. de células independentes.

O conhecimento dos totais marginais das ℓ linhas ($n_{i+}, i = 1,2, \ldots, \ell$)fixa ℓ das frequências n_{ij}, uma em cada linha, desta forma determinando ℓ dos termos totais $\ell \times c$. Consequentemente, reduzimos ℓ dos $\ell \times c$ termos independentes iniciais. Ainda, se a frequência fixada em cada linha for, por exemplo, aquela da última coluna, então dos totais marginais das c colunas ($n_{+j}, j = 1,2, \ldots, c$) apenas os c -1 primeiros permanecem independentes. O número de termos independentes nestas c -1 colunas é reduzido sempre em 1, devido aos marginais de coluna conhecidos. Logo, GL = ($\ell \times c$) - ℓ - (c -1) = (ℓ - 1)(c – 1).

[1] Discutiremos depois sobre o que deve ser considerado como valores esperados suficientemente grandes para que a aproximação assintótica seja satisfatória.

Alternativamente, poderíamos calcular os graus de liberdade como o número de células menos o número de parâmetros π_{ij} estimados, conhecidos os valores de N e n_{i+} e n_{+j}. Nesse caso, teremos GL = ($\ell \times c$) -1 – [(ℓ -1) +(c -1)] = ($\ell \times c$) -1- ℓ +1-c+1 = ($\ell \times c$) - ℓ - c +1=(ℓ - 1)(c – 1).

A estatística qui-quadrado $\chi^2 = N\mathrm{X}^2$ é mais comumente representada por:

$$\chi^2 = \sum_{i \in I} \sum_{j \in J} \frac{\left(n_{ij} - n_{i+}n_{+j}/N\right)^2}{n_{i+}n_{+j}/N}$$

ou com referência a $E_{ij} = (n_{i+} \times n_{+j})/N$:

$$\chi^2 = \sum_{i \in I} \sum_{j \in J} \frac{\left(n_{ij} - E_{ij}\right)^2}{E_{ij}}$$

- No caso de amostragem estratificada, quando podemos associar o modelo teórico de produto de multinomiais, p_{i+} ou p_{+j} não são uma característica da amostra O, pois foram pré-determinados. Isto poderia ter sido o caso, em nosso exemplo, se o número de respondentes de cada classe social (n_{i+}), ou seja, os marginais de linha, tivesse sido fixado a priori. Neste caso a tabela seria apenas um conjunto de linhas (estratos) e estaríamos interessados o quanto homogêneas são as distribuições destas linhas.

O modelo para teste de homogeneidade é que existe uma distribuição para os marginais de colunas, π_{+j} e que **os dados são homogêneos se $\mathbf{p_{(i \mid j)} = \pi_{+j}}$** para qualquer $i \in I$ e $j \in J$, ou seja, **quando**:

$$\frac{p_{ij}}{p_{i+}} = \pi_{+j}$$

Em cada linha $i \in I$, a estatística qui-quadrado de Pearson para bondade do ajuste é:

$$\sum_{j \in J} \frac{\left(\frac{p_{ij}}{p_{i+}} - \pi_{+j}\right)^2}{\pi_{+j}}$$

ou utilizando as estimativas de MV p_{+j} no lugar de π_{+j}:

$$\sum_{j \in J} \frac{\left(\frac{p_{ij}}{p_{i+}} - p_{+j}\right)^2}{p_{+j}}$$

Se tomarmos a soma ponderada de cada somatório acima como uma medida agregada de bondade de ajuste:

$$\sum_{i \in I} p_{i+} \sum_{j \in J} \frac{\left(\frac{p_{ij}}{p_{i+}} - p_{+j}\right)^2}{p_{+j}}$$

Este valor será igual à estatística X^2 apresentada anteriormente, o que indica ser a mesma uma medida de homogeneidade. Esta estatística X^2, equivalente à χ^2/N, é conhecida por **coeficiente de associação Φ^2** ou então por **inércia** da tabela de contingência.

4. Aplicação ao exemplo:

- Temos uma tabela bidimensional 3 × 6, ou seja, com 18 células e 379 observações.
- Para a hipótese de independência H_0: $\pi_{ij} = \pi_{i+} \times \pi_{+j}$ temos GL = 2 × 5 = 10
- Para a forma mais usual de apresentarmos o cálculo da estatística qui-quadrado de Pearson temos:

nij	Eij	nij-Eij	(nij-Eij)^2	(nij-Eij)^2/Eij
48	61,39	-13,39	179,16	2,92
119	122,15	-3,15	9,93	0,08
68	51,46	16,54	273,43	5,31
10	10,19	-0,19	0,04	0,00
24	20,27	3,73	13,90	0,69
5	8,54	-3,54	12,54	1,47
9	8,36	0,64	0,41	0,05
20	16,63	3,37	11,34	0,68
3	7,01	-4,01	16,06	2,29
6	4,70	1,30	1,69	0,36
8	9,36	-1,36	1,84	0,20
4	3,94	0,06	0,00	0,00
12	6,01	5,99	35,91	5,98
10	11,96	-1,96	3,82	0,32
1	5,04	-4,04	16,30	3,24
14	8,36	5,64	31,82	3,81
16	16,63	-0,63	0,40	0,02
2	7,01	-5,01	25,08	3,58
SOMA	-	**0**	-	**30,99**

- O valor da estatística χ^2 de Pearson é 30,99 com 10 GL.
- Alternativamente, diríamos que o valor de $X^2 = \chi^2/N$, seria 0,082
- Com 379 observações, $\chi^2 = 30{,}99$, para 10 GL, o teste seria realizado:

No R: qchisq(0.95,10) nos retorna o valor 18,30, ou seja, $\chi^2_{tabelado;\, 10GL;0{,}05} = 18{,}30$. A hipótese de independência estatística entre as duas variáveis categóricas é rejeitada com significância de 0,000589 , dado pela função chisq.test(N_{ij}).

```
#Definindo a tabela de contingência no programa "R"
Nij= matrix(c(48,119,68,10,24,5,9,20,3,6,8,4,12,10,1,14,16,2),ncol=6)
#Fazendo teste qui-quadrado para a matriz Nij
chisq.test(Nij)
```

5. Comentários

- Verificamos que a associação, neste caso é significativa. Mesmo que não tivesse sido, haveria a possibilidade de analisarmos os dados (AED) de uma forma mais detalhada, e não apenas com base em uma estatística de teste.

- Podemos utilizar alguns conceitos estatísticos básicos, como o de **resíduos,** para analisarmos com mais detalhes onde, dentre as células da tabela de contingência, pode haver um distanciamento maior da independência. Em outras palavras, "quais células contribuem mais para a associação?" Podemos também utilizar de recursos gráficos para analisarmos os perfis das categorias das variáveis, as frequências relativas de cada categoria da variável, e também para verificarmos as similaridades entre as categorias das variáveis. Além disso, podemos usar medidas de associação (p/ex. inércia) para medir a grandeza da associação e também análise multivariada de redução de dimensionalidade para analisarmos similaridades entre classes de uma mesma variável e associação entre categorias de variáveis distintas (p/ex. análise de correspondências).

- Em resumo, podemos utilizar:

  1) Análise de Resíduos

  2) Gráficos de perfis

  3) Medidas de associação

  3) Análise de Correspondências: simples e múltipla.

6. Análise de Resíduos em tabelas de contingência

- Verificamos no exemplo (classe social × rádio preferida) que a associação, através da análise da estatística qui-quadrado, era significativa. Mesmo que não tivesse sido, haveria a possibilidade de analisarmos os dados de uma forma mais detalhada, e não apenas com base em uma estatística de teste. Vamos realizar tal análise.

- Podemos utilizar alguns conceitos estatísticos básicos, como o de **resíduos,** para analisarmos com mais detalhes onde, dentre as células da tabela de contingência, pode haver um distanciamento maior da independência. Em outras palavras, "quais células contribuem mais para a associação?" ou "quais parcelas do cálculo da estatística qui-quadrado de Pearson contribuem mais para o seu valor total?"

1) Análise de resíduos

- De maneira geral, chamamos de residuos o que resta, ou seja, a diferenca entre os dados e o resultado de um "modelo" que tenha sido ajustado aos dados, de acordo com a equação esquematica (Hoaglin, Mosteller&Tukey, 2006[2]) :

  resíduo = dados observados – resultados do "modelo"

  (resíduo = $\boldsymbol{\theta - \widehat{\theta}}$***)***

- No nosso caso, partimos do principio que, se as variáveis são independentes, as frequências observadas em cada célula da tabela de contingencia ($p_{ij} \times N$) dependem

[2] Hoaglin, D.C., Mosteller, F. e Tukey, J.W. 2006. Exploring Data Tables, Trends, and Shapes. Wiley Series in Probability and Statistics. New Jersey: Wiley.

apenas das distribuições marginais (de linhas e colunas, p_{i+} e p_{+j}), como vimos anteriormente.

- Veremos três formas de calcular tais resíduos:

  1- Resíduos brutos
  2- Resíduos de Pearson (padronizados)
  3- Resíduos (de Pearson) ajustados.

1.1 Resíduos Brutos (*eij*$^{(B)}$)**:**
A frequência observada em cada célula menos a frequência esperada (na hipótese de independência), ou seja: *eij*$^{(B)}$ *= nij - Eij*

1.2 Resíduos de Pearson (*eij*$^{(P)}$):

A frequência observada em cada célula menos a frequência esperada, dividido pela raiz quadrada da frequência esperada, conforme equação abaixo:

$$e_{ij}^{(P)} = \frac{(n_{ij} - E_{ij})}{\sqrt{E_{ij}}}$$

Estes resíduos são **padronizados**, mas sua distribuição apresenta valores para as variâncias bem **menores** do que aqueles da distribuição normal padrão Z ~ N(0,1)[3].

1.3 Resíduos de Pearson Ajustados (*eij*$^{(Pa)}$):

Como os resíduos de Pearson não apresentam variâncias compatíveis com a distribuição normal padrão, sugere-se um ajuste a este resíduo (ver Haberman, 1973[4]):

$$e_{ij}^{(Pa)} = \frac{1}{\sqrt{\left[\left(1 - \frac{n_{i+}}{N}\right)\left(1 - \frac{n_{+j}}{N}\right)\right]}} \times e_{ij}^{(P)}$$

(Neste caso, podemos comparar os valores do resíduo de Pearson ajustados aos valores da normal padrão)

Obs: Os resíduos de Pearson ao quadrado, somados em todas as células da tabela, nos fornecem a estatística qui-quadrado de Pearson. Logo, o que os resíduos de Pearson ao quadrado indicam é a contribuição individual de cada célula da tabela ao valor da estatística.

No R, podemos solicitar os valores dos diferentes resíduos através dos seguintes "values":
Xsq$observed # frequências observadas (tabela observada)
Xsq$expected # frequências esperadas sob a hipótese nula
Xsq$residuals # resíduos de Pearson
Xsq$stdres # resíduos padronizados

[3] Everitt, B.S. 1992. The Analysis of Contingency Tables – 2nd Edition. Monographs on Statistics and Applied Probability 45. London: Chapman & Hall. Páginas 47-48.
[4] Haberman, S.J. 1973. The Analysis of Residuals in Cross-classified Tables. *Biometrics*, **29**, 205-220.

			Rádio Preferida						
			Solar	Globo	Itatiaia	Cidade	Pio XII	Outras	Total
Classe Social	A	Valor Observado	48	10	9	6	12	14	99
		Frequencia Esperada	61,4	10,2	8,4	4,7	6,0	8,4	99,0
		Resíduo Bruto	-13,4	-,2	,6	1,3	6,0	5,6	
		Resíduo de Pearson	-1,7	-,1	,2	,6	2,4	2,0	
		Residuo Ajustado	-3,2	-,1	,3	,7	2,9	2,4	
	B	Valor Observado	119	24	20	8	10	16	197
		Frequencia Esperada	122,2	20,3	16,6	9,4	12,0	16,6	197,0
		Resíduo Bruto	-3,2	3,7	3,4	-1,4	-2,0	-,6	
		Resíduo de Pearson	-,3	,8	,8	-,4	-,6	-,2	
		Residuo Ajustado	-,7	1,3	1,2	-,7	-,8	-,2	
	C	Valor Observado	68	5	3	4	1	2	83
		Frequencia Esperada	51,5	8,5	7,0	3,9	5,0	7,0	83,0
		Resíduo Bruto	16,5	-3,5	-4,0	,1	-4,0	-5,0	
		Resíduo de Pearson	2,3	-1,2	-1,5	,0	-1,8	-1,9	
		Residuo Ajustado	4,2	-1,4	-1,8	,0	-2,1	-2,2	
Total		Valor Observado	235	39	32	18	23	32	379
		Frequencia Esperada	235,0	39,0	32,0	18,0	23,0	32,0	379,0

- O que o sinal do resíduo bruto nos indica?
- O que significa um resíduo de Pearson ao quadrado com valor alto?
- Resíduos de Pearson **ajustados** são significativos, para o nível de significância $\alpha = 0,05$, se tiverem quais valores?

- **Problema**: à medida que temos mais células, fica mais difícil identificarmos e analisarmos os resíduos significativos de forma sistemática e entendermos a estrutura dos dados.
- São necessárias outras metodologias de análise: modelos log-lineares para tabelas de contingência; análise de correspondências (que veremos mais adiante no curso).