



Análise de dados categóricos

Universidade Estadual de Santa Cruz

Ivan Bezerra Allaman

Cronograma

1. Introdução
2. Estimadores
3. Estatística de teste
4. Exemplo
5. Aplicações

Introdução

- Lembrando lá da estatística básica, quando nos referimos a dados categóricos, estamos nos referindo a variáveis que são mensuradas em escala nominal e ordinal.
- Neste tópico veremos métodos inferenciais aplicados a tabelas de contingências, técnica esta abordado em introdução a estatística e que será brevemente revisado.

Tabela de contingência de ordem $n \times p$

- A tabela de contingência é uma técnica utilizada para avaliar se há associação ou não entre duas variáveis categóricas.
- Como exemplo poderíamos estudar se há relação entre:
 - Sexo e o hábito de fumar
 - Grau de escolaridade dos pais e grau de escolaridade dos filhos
 - Grau de stress e profissão
 - Escolha partidária política e classe social
 - Etc.

Hipóteses

- As hipóteses que iremos formular são as seguintes:

H_0 : As variáveis são independentes, ou seja, não há associação entre elas

H_a : As variáveis não são independentes, ou seja, há associação entre elas

Estatística de teste

- Utilizaremos a estatística qui-quadrado para testarmos as hipóteses acima.
- A estatística de teste é a seguinte:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

com $(coluna - 1) * (linha - 1)$ graus de liberdade.

- Em que O representa a frequência observada na amostra e E a frequência esperada.
- A frequência esperada é calculada da seguinte forma:

$$E = \frac{\text{Total linha} \cdot \text{Total coluna}}{\text{Tamanho da amostra}}$$

Aplicação

1. Um estudo foi realizado com o intuito de estudar a associação entre sexo (masculino e feminino) e tabagismo (fumante e não fumante) em uma população. Coletou-se uma amostra de 300 pessoas adultas dessa população, fazendo-se a classificação segundo a tabela abaixo:

Tabagismo	Sexo/Masculino	Sexo/Feminino	Total
Fumante	92 (46)	38 (38)	130
Não fumante	108 (54)	62 (62)	170
Total	200 (100)	100 (100)	300

Se observarmos a proporção entre parênteses, tudo nos leva a crer que não há associação entre *tabagismo* e *sexo*, pois a proporção de fumantes não se alterou com o sexo, ou seja, seja masculino ou feminino, a proporção de fumantes é menor.

Percebamos que esta é uma conclusão subjetiva, ou seja, estamos afirmando algo sem fornecer uma probabilidade de certeza.

Logo, utilizaremos a estatística de teste qui-quadrado para avaliarmos a associação entre variáveis qualitativas provenientes de uma tabela de contingência.

As hipóteses são:

H_0 : Sexo e tabagismo são variáveis independentes, na população em estudo

H_a : Existe associação entre as variáveis sexo e tabagismo, na população em estudo

Vamos mostrar o cálculo da frequência esperada apenas para a célula que corresponde a masculino e fumante, pois o raciocínio é o mesmo para as demais células.

$$E_{fu-masc} = \frac{(92 + 38) \cdot (92 + 108)}{300} \\ = 86,67$$

Logo, tem-se a seguinte organização dos cálculos:

Tabagismo	Sexo	Freq.observada	Freq.esperada	OmE	OmE2	OmE2dE
fumante	Mascu	92	86.67	5.33	28.4089	0.3277824
nfumante	Mascu	108	113.33	-5.33	28.4089	0.2506741
fumante	Femi	38	43.33	-5.33	28.4089	0.6556404
nfumante	Femi	62	56.67	5.33	28.4089	0.5013040

As abreviações **OME**, **OME2**, **OME2dE** significam respectivamente **Freq. observada menos a freq. esperada**, **Freq. observada menos freq. esperada ao quadrado** e **Freq. observada menos freq. esperada ao quadrado dividido pela freq. esperada**.

Logo, para calcularmos o chi-quadrado calculado basta somarmos a última coluna. Portanto, tem-se:

$$\chi^2 = 1.735401$$

Agora nos resta encontrar o p-valor com 1 graus de liberdade. Este grau só serve para tabelas do tipo 2x2. Lembrando que já foi ensinado duas opções para encontrar quantis e probabilidades de funções de probabilidade e de densidades. Portanto, o p-valor é igual a

$$\text{p-valor} = 0.1877226$$

Como o p-valor foi maior do que α , não rejeita-se H_0 , ou seja, não há razões para crermos que há uma relação entre sexo e tabagismo.

Correção de continuidade de Yates

- Quando existe alguma frequência esperada entre 5 e 10, há uma necessidade de se fazer uma correção na estatística qui-quadrado.
- Esta correção só é necessário quando a tabela tem dimensões 2 x 2.
- Nestes casos a correção é feita da seguinte forma:

$$\chi^2 = \sum \frac{(|O - E| - 0,5)^2}{E}$$

Aplicação

2. Considere a seguinte amostra quanto a classificação de 38 indivíduos à ansiedade e tabagismo.

	Fumante	Não fumante
Ansioso	15	6
Nansioso	7	10

Tem-se as seguintes hipóteses:

H_0 : Ansiedade e tabagismo são variáveis independentes, na população em estudo.

H_a : Existe associação entre as variáveis ansiedade e tabagismo, na população em estudo.

Montando a tabela com os cálculos tem-se:

Tabagismo	Sexo	Freq.observada	Freq.esperada
fumante	Ansioso	15	12.157895
nfumante	Ansioso	7	9.842105
fumante	Nansioso	6	8.842105
nfumante	Nansioso	10	7.157895

Calculando a estatística qui-quadrado com a correção de Yates, pois houve 3 células com frequência esperada entre 5 e 10 tem-se: $\chi^2 = 2.3952604$.

Logo, o p-valor com 1 graus de liberdade é igual a 0.1217035.

Portanto, como o p-valor é maior do que α não rejeita-se H_0 , ou seja, não há relação entre ansiedade e tabagismo.

Tabelas de grandes dimensões

- São tabelas com dimensões acima de 2×2 , ou seja, com n linhas e p colunas.
- As técnicas apresentadas acima servem para qualquer dimensão de tabela.

Aplicação

3. Considerem os dados disponíveis no link <http://nbcgib.uesc.br/lec/download/R/dados/barbetta.txt> de uma pesquisa realizada em Santa Catarina a respeito do perfil de famílias de três localidades distintas.

Vamos avaliar se existe associação entre "programa de alimentação popular (p.a.p)" e "local".

Portanto, tem-se as seguintes hipóteses:

H_0 : As variáveis p.a.p e localidade são independentes.

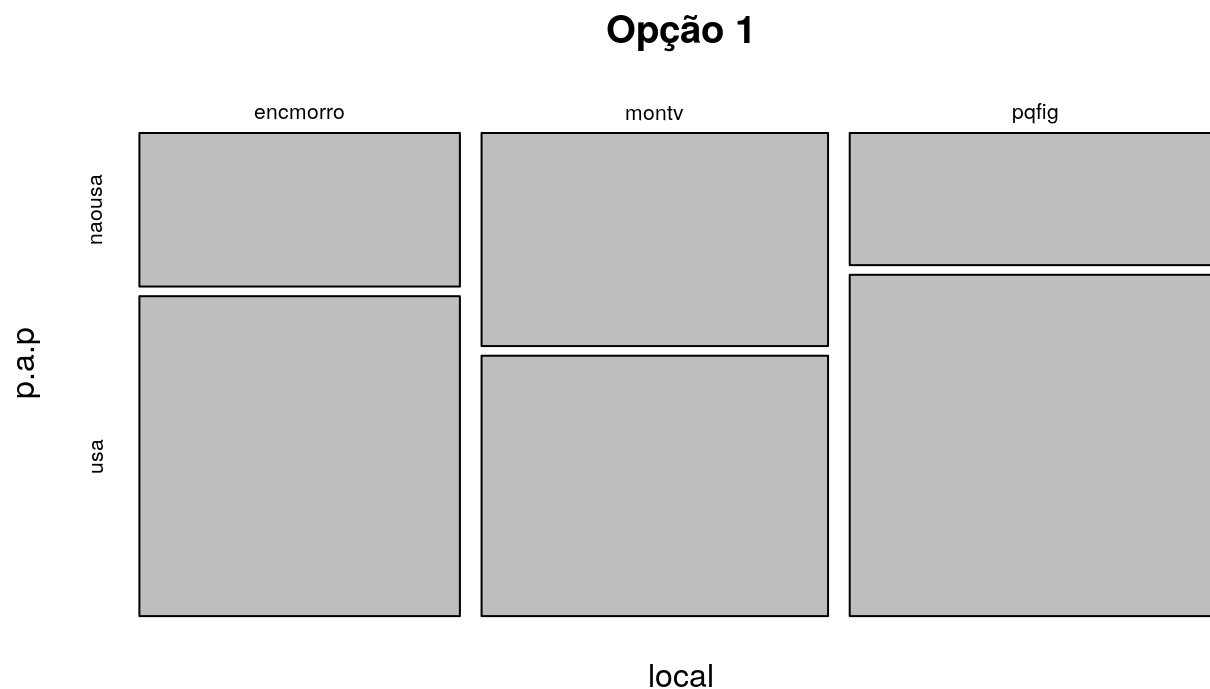
H_a : Existe associação entre as variáveis p.a.p e localidade.

Segue a tabela auxiliar com os cálculos.

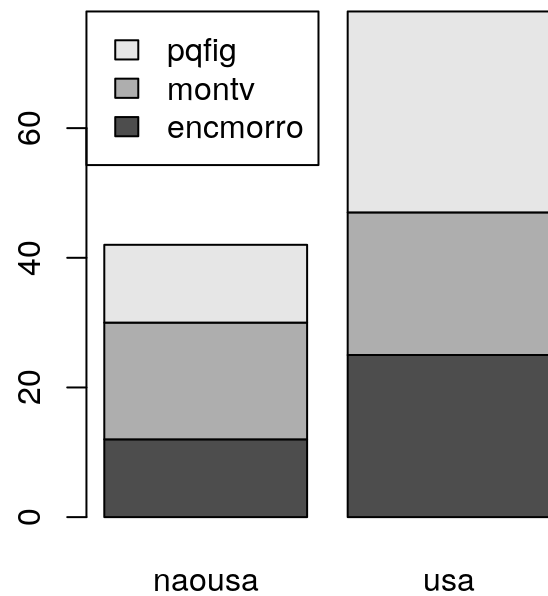
Local	Pap	O	E	(O-E)	$(O - E)^2$	$(O - E)^2/E$
Enc. morro	Não usa	12	12.95	-0.95	0.9	0.07
Monte V	Não usa	18	14.00	4.00	16.0	1.14
Pq. figu.	Não usa	12	15.05	-3.05	9.3	0.62
Enc. morro	Usa	25	24.05	0.95	0.9	0.04
Monte V	Usa	22	26.00	-4.00	16.0	0.62
Pq. figu.	Usa	31	27.95	3.05	9.3	0.33

Somando-se a última coluna da tabela tem-se a estatística qui-quadrado igual a 2.8163917. O graus de liberdade é 2 conforme fórmula já explicada anteriormente. O p-valor é igual a 0.2445842. Logo, como o p-valor é maior do que α , não rejeita-se H_0 .

Podemos ainda fazer um gráfico proveniente de uma tabela de contingência. Segue três opções de gráficos.



Opção 2



Opção 3

