Introdução à estatística

Aula 8 - Análise de Dados Bidimensional

Felipe Nunes, Ph.D.

November 7, 2016

Curso de aperfeiçoamento BR040

Teste do Qui-quadrado

Teste do Qui-quadrado

- O teste do qui-quadrado de independência está associado a duas variáveis qualitativas, ou seja, a uma análise bidimensional.
- Muitas vezes, queremos verificar a relação de dependência entre as duas variáveis qualitativas a serem analisadas.
- Nesse caso, procuramos calcular a frequência de ocorrência das características dos eventos a serem estudados. Por exemplo, podemos estudar a relação entre o sexo de pessoas (masculino e feminino) e o grau de aceitação do governo estadual (ruim, médio e bom).

Teste do Qui-quadrado

- Veja na tabela abaixo:

Sexo	Ruim	Médio	Bom	Total
Masculino	157	27	74	258
Feminino	206	0	10	216
Total	363	27	84	474

- Podemos determinar o grau de associação entre essas duas variáveis, ou seja, determinar se o grau de aceitação do governo depende do sexo ou se existe uma relação de dependência.

Teste do Qui-quadrado

- As hipóteses a serem testadas são:

  H_0: variável linha independe da variável coluna
  H_1: variável linha está associada à variável coluna

- A estatística de qui-quadrado será dada por meio da seguinte expressão:

$$\chi^2 = \frac{\sum (f_o - f_e)^2}{f_e}$$

  onde f_o é a frequencia observada e f_e é a frequência esperada (valor se as variáveis fossem independentes).

Teste do Qui-quadrado

- f_e vai ser calculado da seguinte forma:

$$f_e = \frac{(total\ da\ linha)(total\ da\ coluna)}{total\ geral}$$

- E os graus de liberdade para que possamos olhar na tabela de qui-quadrado são dados por:

$$GL = (h-1).(k-1)$$

  sendo h linhas e k colunas

- No exemplo usado aqui, teremos $GL = (2-1) \times (3-1) = 2$

Teste do Qui-quadrado

- Então, para cada célula da tabela de contingências, você irá calcular a diferença entre f_e e f_o.
- Essa diferença é elevada ao quadrado para evitar que as diferenças positivas e negativas se anulem.
- A divisão pela frequência esperada é feita para obtermos diferenças em termos relativos.

Teste do Qui-quadrado

- **Exemplo**: A UFMG fez uma pesquisa e observou que a distribuição de frequencia entre tipo de curso e tipo de escola foi a seguinte:

Escola	Elite	Não-elite	Total
Publica	30	120	150
Privada	200	150	350
Total	230	270	500

Vamos testar a hipótese de que não há relação entre tipo de curso e tipo de escola que estudou nessa amostra de alunos da UFMG.

Teste do Qui-quadrado

- Se as variáveis fossem independentes, teríamos:

$$\begin{aligned} P(\text{Publica} \cap \text{Elite}) &= P(\text{Publica}) \times P(\text{Elite}) \\ &= \frac{150}{500} \times \frac{230}{500} = 0.138 \end{aligned}$$

Se os eventos fossem independentes deveria haver 0,138 x 500 na célula que cruza pública e elite, ou seja, 69 alunos.

Teste do Qui-quadrado

- Como essa tabela só tem 1 grau de liberdade, posso completar o restante das células descobrindo apenas uma delas.

Escola	Elite	Não-elite	Total
Publica	69	81	150
Privada	161	189	350
Total	230	270	500

Como completamos o restante da tabela?

Teste do Qui-quadrado

- Agora que já tenho os valores esperados e os observados é só calcular o qui-quadrado!

Valores observados

Escola	Elite	Não-elite	Total
Publica	30	120	150
Privada	200	150	350
Total	230	270	500

Valores esperados

Escola	Elite	Não-elite	Total
Publica	69	81	150
Privada	161	189	350
Total	230	270	500

Teste do Qui-quadrado

- O resultado pode ser obtido assim:

$$\begin{aligned}
\chi^2 &= \frac{\sum(f_o - f_e)^2}{f_e} \\
&= \frac{(30-69)^2}{69} + \frac{(120-81)^2}{81} + \frac{(200-161)^2}{161} + \frac{(150-189)^2}{189} \\
&= \frac{1521}{69} + \frac{1521}{81} + \frac{1521}{161} + \frac{1521}{189} \\
&= 22.04 + 18.78 + 9.45 + 8.05 = 58.32
\end{aligned}$$

Teste do Qui-quadrado

- Para fazer o teste preciso do valor crítico do qui-quadrado. Esse valor é dado pela tabela de distribuição de qui-quadrado.
- Para encontrar o valor nós precisamos conhecer:
  1. O nível de significância do teste (95%) $\rightarrow \alpha = 0{,}05$
  2. Os graus de liberdade [(R-1)(K-1)] $\rightarrow$ GL = (2-1)(2-1) = 1
- Pela tabela, o qui-quadrado crítico $(\alpha = 0,05; GL = 1) =$ 3,84.

Teste do Qui-quadrado

- Se $\chi^2_{calculado} \leq \chi^2_{critico}$ aceitamos a hipótese nula
- Se $\chi^2_{calculado} > \chi^2_{critico}$ rejeitamos a hipótese nula
- Quando comparo, observo que meu qui-quadrado calculado (58,32) é maior do que o qui-quadrado crítico (3,84), então rejeito a hipótese nula de independência das variáveis.
- Vai dizer que não parecia mais difícil no começo? ;)

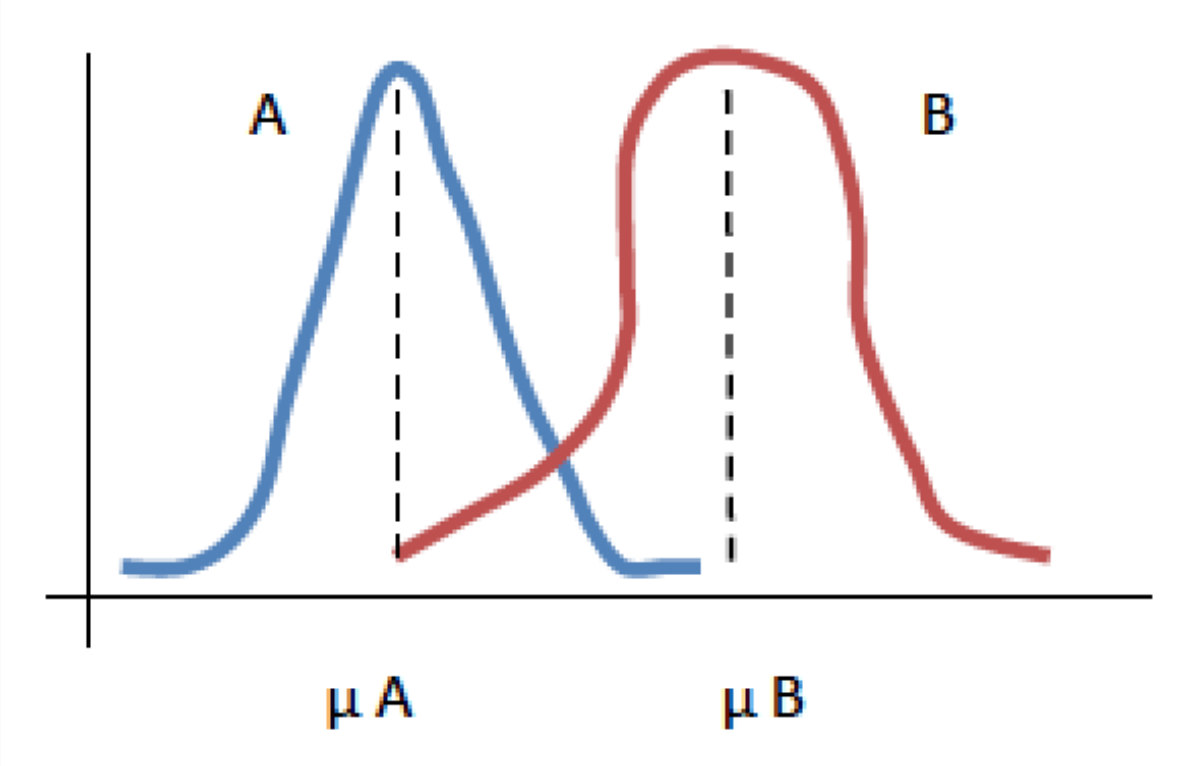
Análise de Variância

Análise de Variância

- O teste Z para duas amostras tem a limitação de só poder ser utilizado quando temos duas variáveis dicotômicas. A ANOVA nos permite testar as médias de 2 ou mais grupos.
- Pressupostos da ANOVA:
    1. **normalidade**: a variável quantitativa deve ter uma distribuição aproximadamente normal
    2. **homocedasticidade**: a variável quantitativa deve ter variâncias semelhantes entre os grupos da variável qualitativa
    3. **aleatoriedade**: amostras devem ser probabilísticas
    4. **independência**: os grupos da variável qualitativa devem ser independentes

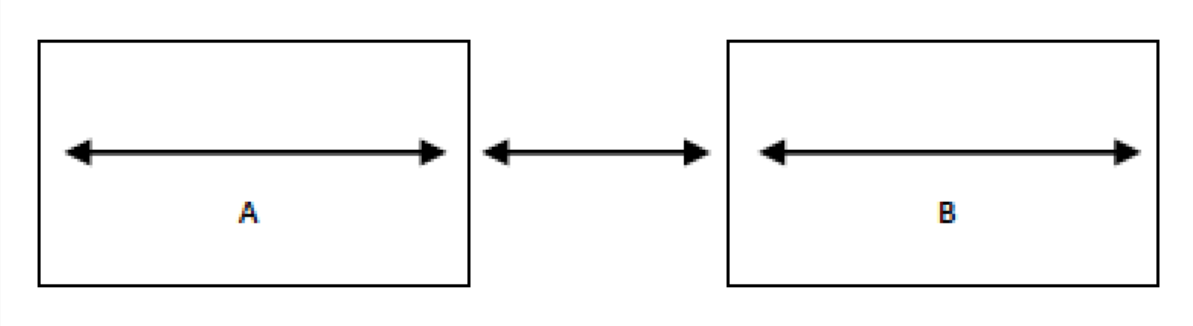
Análise de Variância

- Se tenho dois grupos de eleitores numa cidade. Aqueles que aprovam o governo (grupo A) e aqueles que desaprovam o governo (grupo B). Observamos as rendas médias dos dois grupos.



Análise de Variância

- Entre os que aprovam, todos ganham valores diferentes. Entre os que não-aprovam, também há diferenças de renda – então os valores variam dentro e entre os grupos.
- A ANOVA vai comparar a variância dentro dos grupos com a variância entre os grupos.



Análise de Variância

- Este teste vai comparar as médias das variâncias entre e dentro dos grupos.
- Se a variância dentro dos grupos é pequena e a variância entre os grupos é grande, ou, se as variâncias dentro dos grupos é grande e a variância entre os grupos é pequena; concluiremos que as médias das duas variâncias são diferentes.
- Do contrário, se ambas são pequenas ou grandes, concluiremos que as médias das variâncias são iguais.
- O teste da ANOVA é sempre bilateral.
- A variável independente é qualitativa e a dependente quantitativa.

Análise de Variância

- **Exemplo** Há tres turmas numa escola. Quero comparar as médias das notas dos alunos nas últimas provas. De todos os alunos das turmas, faço uma amostra de A = 5, B = 4 e C = 4. Pego as notas desses alunos e pergunto: há diferença significativa entre as médias das notas de alguma das turmas?

A	B	C
41	56	21
54	61	29
49	48	34
38	65	42
43		

Análise de Variância

$$H_0 : \mu_A = \mu_B = \mu_C$$

$$H_1 : \text{pelo menos uma média é diferente}$$

$$F_e = \frac{\text{Variâncias entre as amostras (VEA)}}{\text{Variância dentro das amostras (VDA)}}$$

Análise de Variância

$$VEA = \frac{\sum n_i (x_a - x_t)^2}{n_t - 1}$$

onde

n_1 é o tamanho de cada amostra

x_a é a média de cada amostra

x_t é a média das médias

n_t é o número total de casos de todas as amostras

Análise de Variância

$$VDA = \frac{\sum (n_i - 1)^2 s_i^2}{n_t - k}$$

onde

n_1 é o tamanho de cada amostra

s_i^2 é a variância de cada amostra

n_t é o número total de casos de todas as amostras

k é o número de grupos sendo comparados

Análise de Variância

$$\bar{x}_A = \frac{41 + 54 + 49 + 38 + 43}{5} = 45$$

$$\bar{x}_B = \frac{56 + 61 + 48 + 65}{4} = 57.5$$

$$\bar{x}_C = \frac{21 + 29 + 34 + 42}{4} = 31.5$$

$$x_t = \frac{(45 \times 5) + (57.5 \times 4) + (31.5 \times 4)}{13} = 44.69$$

Análise de Variância

$$
\begin{aligned}
VEA &= \frac{\sum n_i (x_a - x_t)^2}{k-1} \\
&= \frac{5.(45-44.69)^2 + 4.(57.5-44.69)^2 + 4.(31.5-44.69)^2}{13-1} \\
&= \frac{1352.72}{12} = 112.73
\end{aligned}
$$

Análise de Variância

$$
\begin{aligned}
s_A^2 &= \frac{\sum (x_a - \bar{x_a})^2}{n_a - 1} \\
&= \frac{(41-45)^2 + (54-45)^2 + (49-45)^2 + (38-45)^2 + (43-45)^2}{5-1} \\
&= \frac{166}{4} = 41.5
\end{aligned}
$$

Análise de Variância

$$
\begin{aligned}
s_B^2 &= \frac{\sum (x_a - \bar{x_a})^2}{n_a - 1} \\
&= \frac{(56 - 57.5)^2 + (61 - 57.5)^2 + (48 - 57.5)^2 + (65 - 57.5)^2}{4 - 1} \\
&= \frac{161}{3} = 53.7
\end{aligned}
$$

Análise de Variância

$$
\begin{aligned}
s_C^2 &= \frac{\sum (x_a - \bar{x_a})^2}{n_a - 1} \\
&= \frac{(21 - 31.5)^2 + (29 - 31.5)^2 + (34 - 31.5)^2 + (42 - 31.5)^2}{4 - 1} \\
&= \frac{233}{3} = 77.67
\end{aligned}
$$

Análise de Variância

$$
\begin{aligned}
VDA &= \frac{\sum (n_i - 1)^2 s_i^2}{n_t - k} \\
&= \frac{(5-1)^2.41.5 + (4-1)^2.53.7 + (4-1)^2.77.7}{13 - 3} \\
&= 56
\end{aligned}
$$

Análise de Variância

$$H_0 : \mu_A = \mu_B = \mu_B$$

$$H_1 : \text{pelo menos uma média é diferente}$$

$$\begin{aligned} F_e &= \frac{\text{Variâncias entre as amostras (VEA)}}{\text{Variância dentro das amostras (VDA)}} \\ &= \frac{112.73}{56} = 2.01 \end{aligned}$$

Análise de Variância

- Para encontrar o F crítico precisamos:
  1. $\alpha = 0,05 \rightarrow 95\%$
  2. Graus de liberdade do numerador: $(k-1) \rightarrow 3-1=2$
  3. Graus de liberdade do denominador: $(n_t - k) \rightarrow 13-3=10$
- Com esses dados olhamos na tabela da ANOVA e encontramos o $F_{critico} = 4,10$
- Como o $F_{calculado}$ (2.01) é menor do que $F_{critico}$ (4,10), então eu aceito a hipótese nula de que as médias dos três grupos são iguais.