ME414 - Estatística para Experimentalistas

Parte 22

Notas de aula produzidas pelos professores **Samara Kiihl**, **Tatiana Benaglia** e **Benilton Carvalho** e modificadas pela Profa. **Larissa Avila Matos**

Testes de Independência e Homogeneidade

Tabela de contingência

Quando dois ou mais atributos são observados para cada elemento amostrado, os dados podem ser simultaneamente classificados com respeito aos níveis de ocorrência para cada um dos atributos.

Por exemplo, empregados podem ser classificados de acordo com escolaridade e tipo de ocupação, flores podem ser classificadas com respeito ao tipo de folhagem e tamanho.

Tabela de contingência: enumerar a frequência de obervações da classificação simultânea de duas ou mais características.

Podemos usar a tabela de contingência para estudar se certa característica parece se manifestar independentemente da outra ou se níveis de uma característica tendem a estar associados com níveis da outra.

Exemplo: Racionamento de energia

Uma amostra aleatória de 500 pessoas responde um questionário sobre filiação partidária (partido A ou B) e atitude mediante um programa de racionamento de energia. Os resultados estão apresentados na tabela de contingência a seguir:

	Favorável	Indiferente	Contrário
A	138	83	64
B	64	67	84

Os dados indicam que a opinião sobre racionamento de energia é independente da filiação partidária?

Podemos medir quantitativamente a associaçãoo entre as duas características?

Exemplo: Racionamento de energia

Primeiramente, consideremos a tabela de um ponto de vista descritivo, transformando as contagens em proporcões.

- Proporções por caselas

	Favorável	Indiferente	Contrário	Total
A	0.28	0.17	0.13	0.57
B	0.13	0.13	0.17	0.43
Total	0.4	0.3	0.3	1.00

Exemplo: Racionamento de energia

Primeiramente, consideremos a tabela de um ponto de vista descritivo, transformando as contagens em proporcões.

- Proporções por linhas

	Favorável	Indiferente	Contrário	Total
A	0.48	0.29	0.22	1.00
B	0.3	0.31	0.39	1.00

Exemplo: Racionamento de energia

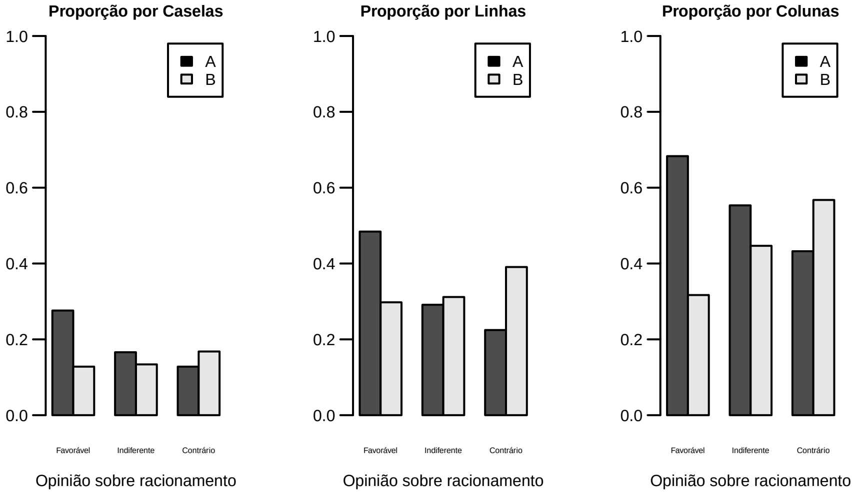
Primeiramente, consideremos a tabela de um ponto de vista descritivo, transformando as contagens em proporcões.

- Proporções por colunas

	Favorável	Indiferente	Contrário
A	0.68	0.55	0.43
B	0.32	0.45	0.57
Total	1.00	1.00	1.00

Exemplo: Racionamento de energia

- Gráficos de barras: Frequências relativas (caselas, linhas e colunas)



Exemplo: Racionamento de energia

Através das tabelas de proporções e gráficos de barras, observam-se diferenças aparentes nas distribuições ao longo das linhas, colunas ou das proporções totais das respostas.

Por exemplo, com relação às proporções por linha, observa-se que as proporções diminuem ao longo da primeira linha e aumentam ao longo da segunda.

Podemos usar um teste estatístico para avaliar possível associação entre filiação partidária e opinião com relação ao programa de racionamento de energia.

Teste de Independência

Considere duas características designadas por A e B e suponha que existem r categorias $A_1, A_2, \ldots, A_r$ para A e c categorias $B_1, B_2, \ldots, B_c$ para B.

Suponha que uma amostra de tamanho n é classificada e distribuída nas classes de A e B, produzindo uma tabela de contingência em que:

- n_{ij} = frequência de observações com as características A_i e B_j conjuntamente.
- n_{i0} = total da i-ésima linha, ou frequência de A_i.
- n_{0j} = total da j-ésima coluna, ou frequência de B_j.

Teste de Independência

	B_1	B_2	$\ldots$	B_c	Total da linha
A_1	n_{11}	n_{12}	$\ldots$	n_{1c}	n_{10}
A_2	n_{21}	n_{22}	$\ldots$	n_{2c}	n_{20}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_r	n_{r1}	n_{r2}	$\ldots$	n_{rc}	n_{r0}
Total da coluna	n_{01}	n_{02}	$\ldots$	n_{0c}	n

Teste de Independência

Podemos usar a população classificada em termos de proporções populacionais e a tabela anterior fica:

	B_1	B_2	$\ldots$	B_c	Total da linha
A_1	p_{11}	p_{12}	$\ldots$	p_{1c}	p_{10}
A_2	p_{21}	p_{22}	$\ldots$	p_{2c}	p_{20}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_r	p_{r1}	p_{r2}	$\ldots$	p_{rc}	p_{r0}
Total da coluna	p_{01}	p_{02}	$\ldots$	p_{0c}	1

- $p_{ij} = P(A_i \cap B_j)$ é a probabilidade da ocorrência conjunta de A_i e B_j.
- $p_{i0} = P(A_i)$ é a probabilidade total da i-ésima linha.
- $p_{0j} = P(B_j)$ é a probabilidade total da j-ésima coluna.

Teste de Independência

Teste de independência: interesse é testar se as classificações nas categorias de A e B são independentes, ou seja, pretende-se avaliar se

$$P(A_i \cap B_j) = P(A_i)P(B_j),$$

para todo $i = 1, 2, \ldots, r$ e $j = 1, 2, \ldots, c$.

Teste de Independência

Hipóteses:

$H_0 : p_{ij} = p_{i0}p_{0j}$ para todas as componentes (i, j) (**independência**)

$H_a : p_{ij} \neq p_{i0}p_{0j}$ para pelo menos um par (i, j)

O modelo de independência especifica as probabilidades das componentes em termo das probabilidades marginais. **Problema:** as probabilidades marginais são parâmetros desconhecidos.

Como $p_{i0} = P(A_i)$, um estimador natural é a frequência relativa amostra de A_i,

$$\hat{p}_{i0} = \frac{n_{i0}}{n}.$$

Da mesma forma, $p_{0j} = P(B_j)$ é estimado por

$$\hat{p}_{0j} = \frac{n_{0j}}{n}.$$

Teste de Independência

Usando essas estimativas a probabilidade da componente (i, j) é estimada por

$$\hat{p}_{ij} = \hat{p}_{i0}\hat{p}_{0j} = \frac{n_{i0}n_{0j}}{n^2}.$$

Logo, a frequência relativa esperada sob o modelo de independência é

$$E_{ij} = n\hat{p}_{ij} = \frac{n_{i0}n_{0j}}{n}.$$

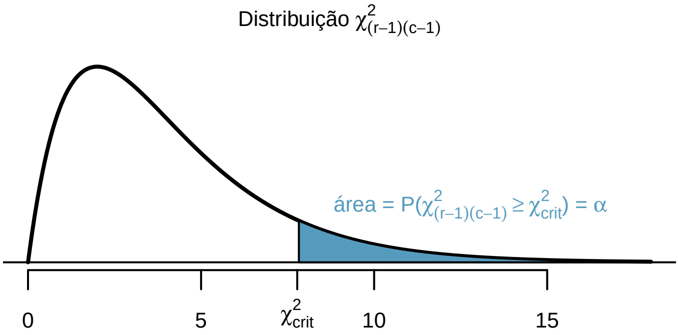
Portanto, a **estatística do teste** é dada por:

$$\chi^2 = \sum_{r\times c \text{ componentes}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{r\times c \text{ componentes}} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

que sob H_0 tem distribuição aproximadamente χ^2 com $(r-1)\times(c-1)$ graus de liberdade, para n grande.

Teste de Independência

Valor Crítico: Para um nível de significância α, encontrar o valor crítico χ^2_{crit} na tabela Chi-quadrado tal que $P(\chi^2_{(r-1)(c-1)} \geq \chi^2_{crit}) = \alpha$.



Conclusão: Rejeitamos H_0 se $\chi^2_{obs} \geq \chi^2_{crit}$.

Exemplo: Racionamento de energia

Frequências observadas (n_{ij}):

	Favorável	Indiferente	Contrário
A	138	83	64
B	64	67	84

Frequências esperadas (E_{ij}), segundo hipótese de independência:

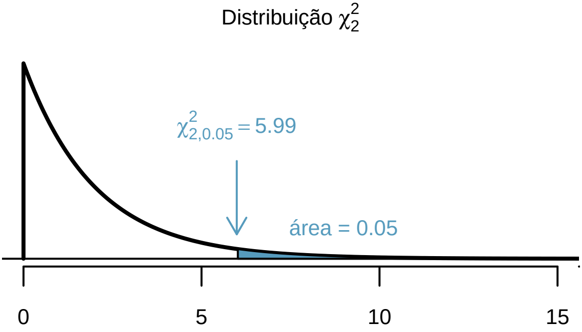
	Favorável	Indiferente	Contrário
A	115.14	85.5	84.36
B	86.86	64.5	63.64

Exemplo: Racionamento de energia

A estatística χ^2 tem o valor observado de

$$\chi^2_{obs} = 4.539 + 0.073 + 4.914 + 6.016 + 0.097 + 6.514 = 22.15 \overset{H_0}{\sim} \chi^2_2$$

Usando o nível de significância $\alpha = 0.05$, o valor crítico é $\chi^2_{crit} = \chi^2_{2,0.05} = 5.99$.



Exemplo: Racionamento de energia

Como $\chi^2_{obs} = 22.15 > 5.99 = \chi^2_{crit}$, rejeitamos a hipótese nula de indepêndencia.

Concluímos que os dados trazem evidências de **associação** entre as duas características (filiação e opinião).

CUIDADO!!! Associação não implica CAUSA.

Não podemos afirmar que existe uma relação de causa e efeito, pois os dados são observacionais, isto é, não aleatorizamos as pessoas para serem do partido A ou B, por exemplo.

Exemplo: Gênero e escolha da carreira

Existe associação entre sexo e a carreira escolhida por 200 alunos de Economia e Administração?

Frequências observadas (n_{ij}):

	Masculino	Feminino
Economia	85	35
Administração	55	25

Frequências esperadas (E_{ij}), segundo hipótese de independência:

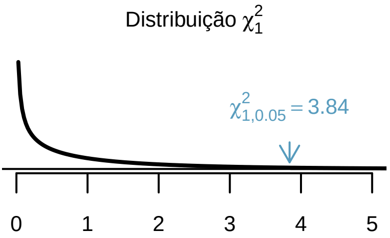
	Masculino	Feminino
Economia	84	36
Administração	56	24

Exemplo: Gênero e escolha da carreira

A estatística χ^2 tem o valor observado de

$$\chi^2_{obs} = \frac{(85-84)^2}{84} + \frac{(35-36)^2}{36} + \frac{(55-56)^2}{56} + \frac{(25-24)^2}{24} = 0.099 \overset{H_0}{\sim} \chi^2_1$$

Usando o nível de significância $\alpha = 0.05$, o valor crítico é $\chi^2_{crit} = \chi^2_{1,0.05} = 3.84$.



Como $\chi^2_{obs} = 0.099 < 3.84 = \chi^2_{crit}$, não rejeitamos a hipótese nula de indepêndencia.

Exemplo: Exercícios do Moodle e nota da Prova 1

Existe associação entre obter no mínimo 5 nos exercícios do Moodle e obter no mínimo 5 na prova 1 de ME414?

As notas de 453 alunos matriculados nas turmas de ME414 no 2S2015 foram consideradas. Os seguintes resultados foram obtidos:

	< 5 na P1	>= 5 na P1	Total
< 5 no Moodle	21	44	65
>= 5 no Moodle	37	351	388
Total	58	395	453

Exemplo: Exercícios do Moodle e nota da Prova 1

Tabela de frequências esperadas, segundo a hipótese nula de independência:

$$E_{ij} = n\hat{p}_{ij} = \frac{n_{i0}n_{0j}}{n}$$

	$<$ 5 na P1	$>=$ 5 na P1
$<$ 5 no Moodle	8.32	56.68
$>=$5 no Moodle	49.68	338.32

A estatística χ^2 tem o valor observado de

$$\chi^2_{obs} = \frac{(21-8.32)^2}{8.32} + \frac{(44-56.68)^2}{56.68} + \frac{(37-49.68)^2}{49.68} + \frac{(351-338.32)^2}{338.32} = 25.86 \overset{H_0}{\sim} \chi^2_1$$

Usando o nível de significância $\alpha = 0.05$, o valor crítico é $\chi^2_{crit} = \chi^2_{1,0.05} = 3.84$.

Como $\chi^2_{obs} = 25.86 > 3.84 = \chi^2_{crit}$, rejeitamos a hipótese nula de indepêndencia.

Teste de Homogeneidade

Teste de Homogeneidade

Nas situações em que utilizamos os testes de independência, o esquema de amostragem utizado foi baseado numa amostra aleatória de tamanho n que é classificada com respeito a duas características simultaneamente.

Nesse caso, as frequências marginais totais (totais por linhas e totais por colunas) são variáveis aleatórias, pois a cada nova amostragem, não temos como saber de antemão quais serão os valores dos totais por linhas/colunas.

Se o esquema de amostragem for de dividir a população em duas subpopulações de acordo com as categorias de uma característica e selecionar uma amostra de um tamanho pré-determinado para cada subpopulação, então esta será uma situação de tabela de contingência com margens fixas.

Teste de Homogeneidade

Por exemplo, no caso do problema de filiação partidária, poderíamos selecionar amostras aleatórias de tamanho 200 entre afiliados do partido A e 300 dentre os afiliados do partido B e se classificaria essas amostras de acordo com a atitude (favorável, indiferente ou contrário).

O interesse então é estudar as proporções nessas categorias para determinar se elas são aproximadamente iguais para as diferentes subpopulações. Ou seja, queremos testar se as subpopulações são homogêneas.

Teste de Homogeneidade

Suponha que amostras aleatórias independentes de tamanho $n_{10}, \ldots, n_{r0}$ são selecionadas de r subpopulações $A_1, \ldots, A_r$ respectivamente. Classificando cada amostra em uma das categorias $B_1, \ldots, B_c$, obtemos uma tabela de contigência $r \times c$ onde os totais das linhas são tamanhos de amostras fixos.

Tabelas de contingência $r \times c$ com totais das linhas fixos:

	B_1	B_2	$\ldots$	B_c	Total da linha
A_1	n_{11}	n_{12}	$\ldots$	n_{1c}	n_{10}
A_2	n_{21}	n_{22}	$\ldots$	n_{2c}	n_{20}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_r	n_{r1}	n_{r2}	$\ldots$	n_{rc}	n_{r0}
Total da coluna	n_{01}	n_{02}	$\ldots$	n_{0c}	n

Teste de Homogeneidade

As probabilidades das várias categorias de B dentro de cada subpopulação de A também são apresentadas a seguir, onde cada w representa uma probabilidade condicional,

$$w_{ij} = P(B_j|A_i) = \text{ probabilidade de } B_j \text{ dentro da população } A_i.$$

Probabilidades das categorias de B dentro de cada subpopulação:

	B_1	B_2	$\ldots$	B_c	Total da linha
A_1	w_{11}	w_{12}	$\ldots$	w_{1c}	1
A_2	w_{21}	w_{22}	$\ldots$	w_{2c}	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_r	w_{r1}	w_{r2}	$\ldots$	w_{rc}	1

Teste de Homogeneidade

A hipótese nula de iqualdade das categorias B para as r subpopulações é:

$$H_0 : w_{1j} = w_{2j} = \cdots = w_{rj}, \text{ para todo } j = 1, 2, \ldots c.$$

Sob H_0, a probabilidade comum da categoria B_j pode ser estimada do conjunto de amostras notando que de um total de n elementos amostrados, n_{0j} possuem a característica B_j, daí a probabilidade estimada fica

$$\hat{w}_{1j} = \hat{w}_{2j} = \cdots = \hat{w}_{rj} = \frac{n_{0j}}{n}.$$

A frequência esperada estimada na componente (i, j) sob H_0 é:

$$\begin{aligned} E_{ij} &= (\text{Número de } A_i \text{ amostrados}) \times (\text{Probabilidade de } B_j \text{ dentro de } A_i) \\ &= n_{i0}\hat{w}_{ij} = \frac{n_{i0}n_{0j}}{n}. \end{aligned}$$

Teste de Homogeneidade

A **estatística do teste** é dada por:

$$\chi^2 = \sum_{r \times c \text{ componentes}} \frac{(n_{ij} - E_{ij})^2}{E_{ij}},$$

que sob H_0 segue uma distribuição χ^2 com $(r-1) \times (c-1)$ graus de liberdade.

Pode-se observar que as fórmulas e os graus de liberdade dessa seção são iguais ao da seção anterior, somente o método de amostragem e a formalização da hipótese nula são diferentes.

Valor Crítico: Para um nível de significância α, encontrar o valor crítico χ^2_{crit} na tabela Chi-quadrado tal que $P(\chi^2_{(r-1)(c-1)} \geq \chi^2_{crit}) = \alpha$.

Conclusão: Rejeitamos H_0 se $\chi^2_{obs} \geq \chi^2_{crit}$.

Exemplo: Alcoolismo e profissões

Foi feita uma pesquisa para determinar a incidência de alcoolismo em diferentes grupos profissionais.

Separadamente, um amostra aleatória entre religiosos, educadores, executivos e comerciantes foi coletada.
Os dados são apresentados na tabela:

	Alcoólatras	Não Alcoólatras
Religiosos	32	268
Educadores	51	199
Executivos	67	233
Comerciantes	83	267

Exemplo: Alcoolismo e profissões

$$w_{ij} = P(B_j|A_i) = \text{ probabilidade de } B_j \text{ dentro da subpopulação } A_i.$$

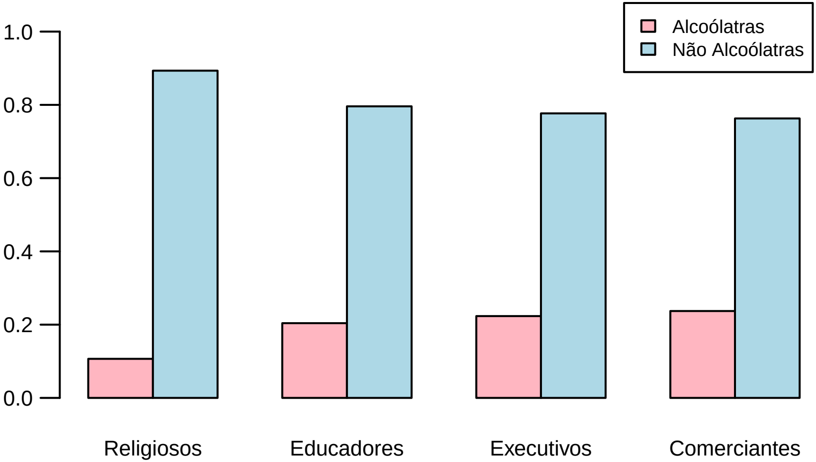
$H_0 : w_{1j} = w_{2j} = \cdots = w_{rj}$, para todo $j = 1, 2, \ldots c$.

Tabela de contingência de alcoolismo vs profissão: frequência relativa por linha.

	Alcoólatras	Não Alcoólatras
Religiosos	0.11	0.89
Educadores	0.20	0.80
Executivos	0.22	0.78
Comerciantes	0.24	0.76

Exemplo: Alcoolismo e profissões

Gráfico de barras de alcoolismo vs profissão: frequência relativa por linha.



Exemplo: Alcoolismo e profissões

A frequência esperada estimada na componente (i, j) sob H_0 é

$$E_{ij} = \frac{n_{i0} n_{0j}}{n}.$$

Tabela de frequências esperadas, segundo a hipótese nula de homogeneidade:

	Alcoólatras	Não Alcoólatras
Religiosos	58.25	241.75
Educadores	48.54	201.46
Executivos	58.25	241.75
Comerciantes	67.96	282.04

Exemplo: Alcoolismo e profissões

Representando por p_1, p_2, p_3 e p_4 as proporções de alcoólatras na subpopulação de religiosos, educadores, executivos e comerciantes, respectivamente, queremos testar a hipótese:

$$H_0 : p_1 = p_2 = p_3 = p_4 \quad \text{vs} \quad H_a : \text{pelo menos uma proporção é diferente.}$$

A estatística observada é:

$$\chi^2_{obs} = \frac{(32 - 58.25)^2}{58.25} + \cdots + \frac{(267 - 282.04)^2}{282.04} = 20.6 \overset{H_0}{\sim} \chi^2_3$$

Usando o nível de significância $\alpha = 0.05$, o valor crítico é $\chi^2_{crit} = \chi^2_{3,0.05} = 7.81$. Como $\chi^2_{obs} = 20.6 > 7.81 = \chi^2_{crit}$, rejeitamos a hipótese nula de homogeneidade.

Como a hipótese nula foi rejeitada verificamos que há indícios de que a proporção de alcoólatras nas classes profissionais não é homogênea.

Exemplo: Google

O Google está constantemente elaborando experimentos para testar novos algoritmos de busca. Por exemplo, o Google pode estar interessado em testar 3 algoritmos usando uma amostra aleatória para cada um: 5000 buscas feitas com o algoritmo atual foram selecionadas ao acaso, 2500 buscas feitas com o algoritmo teste 1 foram selecionadas ao acaso e 2500 buscas feitas com o algoritmo teste 2 foram selecionadas ao acaso.

Como avaliar qual o melhor algoritmo? É preciso definir alguma medida.

No caso, o Google irá avaliar se o usuário clicou em um dos links da busca e depois não realizou uma nova tentativa de busca ou se ele depois realizou nova tentativa (indicando que a primeira busca não foi bem sucedida).

Objetivo: 3 algoritmos têm a mesma performance, isto é, a proporção de buscas que não são refeitas é a mesma para os três algoritmos?

Exemplo: Google

Suponha que o Google tenha obtido os seguintes resultados:

	Atual	Teste 1	Teste 2
Sem nova busca	3511	1749	1818
nova busca	1489	751	682

Exemplo: Google

Tabela de frequências esperadas, segundo a hipótese nula de homogeneidade:

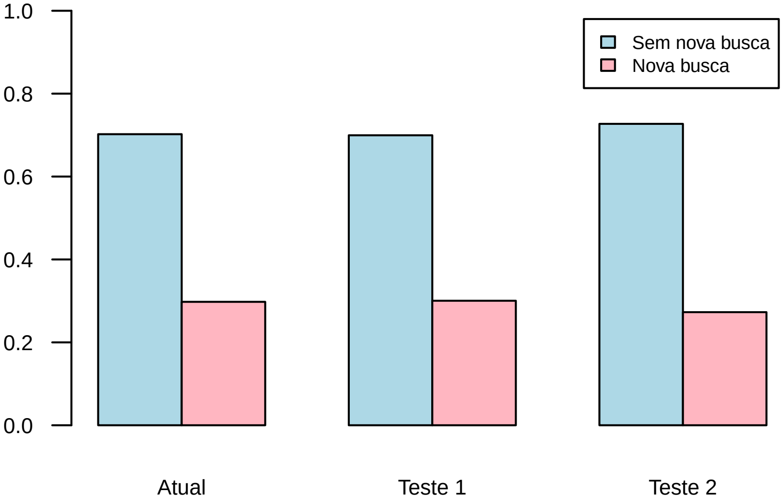
	Atual	Teste 1	Teste 2
Sem nova busca	3539	1769.5	1769.5
nova busca	1461	730.5	730.5

A estatística χ^2 tem o valor observado de

$\chi^2_{obs} = \frac{(3511-3539)^2}{3539} + \frac{(1749-1769.5)^2}{1769.5} + \frac{(1489-1461)^2}{1461} + \frac{(751-730.5)^2}{730.5} = 6.12 \overset{H_0}{\sim} \chi^2_2$

Usando o nível de significância $\alpha = 0.01$, o valor crítico é $\chi^2_{crit} = \chi^2_{2,0.01} = 9.21$. Como $\chi^2_{obs} = 6.12 < 9.21 = \chi^2_{crit}$, não rejeitamos a hipótese nula de homogeneidade.

Exemplo: Google



Leituras

- Ross: capítulo 13.
- OpenIntro: seção 6.4