ME414 - Estatística para Experimentalistas

Parte 21

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

Testes Chi-Quadrado: Aderência e Independência

Introdução

Muitas vezes, a informação da amostra coletada tem a estrutura de dados categorizados, ou seja, cada membro da população pode assumir um entre k valores de uma ou mais características estudadas.

Introdução

Muitas vezes, a informação da amostra coletada tem a estrutura de dados categorizados, ou seja, cada membro da população pode assumir um entre k valores de uma ou mais características estudadas.

Dessa forma, o conjunto de dados consiste em frequências de contagens para essas categorias.

Introdução

Muitas vezes, a informação da amostra coletada tem a estrutura de dados categorizados, ou seja, cada membro da população pode assumir um entre k valores de uma ou mais características estudadas.

Dessa forma, o conjunto de dados consiste em frequências de contagens para essas categorias.

Esse tipo de dados ocorre com frequência nas áreas sociais e biomédicas.

Introdução

Muitas vezes, a informação da amostra coletada tem a estrutura de dados categorizados, ou seja, cada membro da população pode assumir um entre k valores de uma ou mais características estudadas.

Dessa forma, o conjunto de dados consiste em frequências de contagens para essas categorias.

Esse tipo de dados ocorre com frequência nas áreas sociais e biomédicas.

O objetivo aqui é estudar dados agrupados em categorias múltiplas e veremos isso através de dois tipos de testes:

- Teste de Aderência (ou Bondade de Ajuste); e
- Teste de Independência.

Introdução

Teste de Aderência: considere uma população na qual cada membro assume qualquer um de k possíveis valores. Iremos verificar quão adequado uma amostra obtida dessa população se ajusta a um modelo de probabilidade proposto.

Introdução

Teste de Aderência: considere uma população na qual cada membro assume qualquer um de k possíveis valores. Iremos verificar quão adequado uma amostra obtida dessa população se ajusta a um modelo de probabilidade proposto.

Teste de Independência: considere uma população na qual cada membro é classificado de acordo com duas características distintas. Com os dados de uma amostra dessa população, iremos verificar se essas duas características podem ser consideradas independentes.

Introdução

Teste de Aderência: considere uma população na qual cada membro assume qualquer um de k possíveis valores. Iremos verificar quão adequado uma amostra obtida dessa população se ajusta a um modelo de probabilidade proposto.

Teste de Independência: considere uma população na qual cada membro é classificado de acordo com duas características distintas. Com os dados de uma amostra dessa população, iremos verificar se essas duas características podem ser consideradas independentes.

Duas características serão independentes se a classificação de um membro da população de acordo com uma característica não interfere na probabilidade de classificação em relação à segunda característica desse mesmo membro.

Introdução

Teste de Aderência: considere uma população na qual cada membro assume qualquer um de k possíveis valores. Iremos verificar quão adequado uma amostra obtida dessa população se ajusta a um modelo de probabilidade proposto.

Teste de Independência: considere uma população na qual cada membro é classificado de acordo com duas características distintas. Com os dados de uma amostra dessa população, iremos verificar se essas duas características podem ser consideradas independentes.

Duas características serão independentes se a classificação de um membro da população de acordo com uma característica não interfere na probabilidade de classificação em relação à segunda característica desse mesmo membro.

Na aula de hoje iremos focar em Testes de Aderência.

Exemplo: Cores de Geladeira

Uma determinada marca de geladeira é vendida em cinco cores diferentes e uma pesquisa de mercado quer avaliar a popularidade das várias cores.

Exemplo: Cores de Geladeira

Uma determinada marca de geladeira é vendida em cinco cores diferentes e uma pesquisa de mercado quer avaliar a popularidade das várias cores.

As frequências abaixo são observadas para uma amostra de 300 vendas feitas num semestre.

Exemplo: Cores de Geladeira

Uma determinada marca de geladeira é vendida em cinco cores diferentes e uma pesquisa de mercado quer avaliar a popularidade das várias cores.

As frequências abaixo são observadas para uma amostra de 300 vendas feitas num semestre.

Suponha que seja de interesse testar a hipótese das cinco cores serem igualmente populares.

Exemplo: Cores de Geladeira

Uma determinada marca de geladeira é vendida em cinco cores diferentes e uma pesquisa de mercado quer avaliar a popularidade das várias cores.

As frequências abaixo são observadas para uma amostra de 300 vendas feitas num semestre.

Suponha que seja de interesse testar a hipótese das cinco cores serem igualmente populares.

Vendas das cinco cores das geladeiras da marca W.

Cor	marrom	creme	vermelho	azul	branco	total
Frequência	88	65	52	40	55	300

Modelo Multinomial

Distribuição Multinomial

Para acomodar dados como no Exemplo 1, precisamos estender o modelo Bernoulli de forma que os resultados possam ser classificados em mais de duas categorias. Esse modelo é chamado de **distribuição multinomial**.

Distribuição Multinomial

Para acomodar dados como no Exemplo 1, precisamos estender o modelo Bernoulli de forma que os resultados possam ser classificados em mais de duas categorias. Esse modelo é chamado de **distribuição multinomial**.

Modelo Multinomial

Distribuição Multinomial

Para acomodar dados como no Exemplo 1, precisamos estender o modelo Bernoulli de forma que os resultados possam ser classificados em mais de duas categorias. Esse modelo é chamado de **distribuição multinomial**.

Modelo Multinomial

1. O resultado de cada amostra pode ser classificado em uma de k respostas denotadas por $1, 2, \ldots, k$.
2. A probabilidade da amostra assumir o valor i é p_i, $i = 1, 2, \ldots, k$, com
$$\sum_{i=1}^{k} p_i = 1.$$
3. As observações são independentes.

Distribuição Multinomial

Considere uma amostra de uma população que consiste de elementos em diversas categorias, por exemplo, k valores possíveis.

Distribuição Multinomial

Considere uma amostra de uma população que consiste de elementos em diversas categorias, por exemplo, k valores possíveis.

Denotaremos por $n_1, n_2, \ldots, n_k$, com $\sum_{i=1}^{k} n_i = n$ as frequências e $p_1, p_2, \ldots, p_k$ as probabilidades.

Distribuição Multinomial

Considere uma amostra de uma população que consiste de elementos em diversas categorias, por exemplo, k valores possíveis.

Denotaremos por $n_1, n_2, \ldots, n_k$, com $\sum_{i=1}^{k} n_i = n$ as frequências e $p_1, p_2, \ldots, p_k$ as probabilidades.

A distribuição conjunta de $n_1, n_2, \ldots, n_k$ é chamada de distribuição multinomial e tem função de probabilidade dada por:

$$f(n_1, n_2, \ldots, n_k) = \frac{n!}{n_1! \ldots n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k},$$

em que $\displaystyle\sum_{i=1}^{k} n_i = n$ e com $\displaystyle\sum_{i=1}^{k} p_i = 1$.

Distribuição Multinomial

Se designarmos a componente n_1 como "sucesso" e juntarmos as demais numa mesma que designamos "fracasso", a variável aleatória n_1 é o número de sucessos em n ensaios de Bernoulli, ou seja, $n_1 \sim Bin(n, p_1)$.

Distribuição Multinomial

Se designarmos a componente n_1 como "sucesso" e juntarmos as demais numa mesma que designamos "fracasso", a variável aleatória n_1 é o número de sucessos em n ensaios de Bernoulli, ou seja, $n_1 \sim Bin(n, p_1)$.

Portanto, $\mathbb{E}(n_1) = np_1$, $Var(n_1) = np_1(1 - p_1)$.

Distribuição Multinomial

Se designarmos a componente n_1 como "sucesso" e juntarmos as demais numa mesma que designamos "fracasso", a variável aleatória n_1 é o número de sucessos em n ensaios de Bernoulli, ou seja, $n_1 \sim Bin(n, p_1)$.

Portanto, $\mathbb{E}(n_1) = np_1$, $Var(n_1) = np_1(1 - p_1)$.

Analogamente aplicando o mesmo argumento a cada n_i temos:

$$\mathbb{E}(n_i) = np_i \quad \text{e} \quad Var(n_i) = np_i(1 - p_i).$$

Distribuição Multinomial

Se designarmos a componente n_1 como "sucesso" e juntarmos as demais numa mesma que designamos "fracasso", a variável aleatória n_1 é o número de sucessos em n ensaios de Bernoulli, ou seja, $n_1 \sim Bin(n, p_1)$.

Portanto, $\mathbb{E}(n_1) = np_1$, $Var(n_1) = np_1(1 - p_1)$.

Analogamente aplicando o mesmo argumento a cada n_i temos:

$$\mathbb{E}(n_i) = np_i \quad \text{e} \quad Var(n_i) = np_i(1 - p_i).$$

Iremos usar o valor esperado de n_i nos testes que veremos a seguir.

Teste de Aderência

Teste de Aderência

Objetivo: Testar quão adequado é assumir um modelo probabilístico para descrever um determinado conjunto de dados.

Teste de Aderência

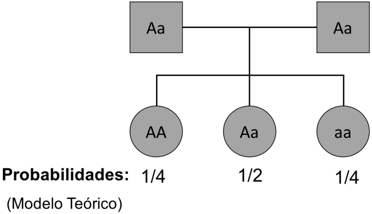
Objetivo: Testar quão adequado é assumir um modelo probabilístico para descrever um determinado conjunto de dados.

Exemplo: Vocês já devem ter visto em alguma aula de Biologia o seguinte:

Teste de Aderência

Objetivo: Testar quão adequado é assumir um modelo probabilístico para descrever um determinado conjunto de dados.

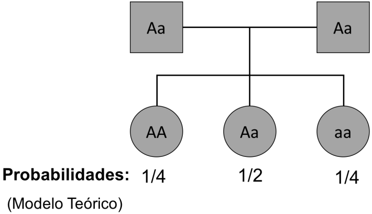
Exemplo: Vocês já devem ter visto em alguma aula de Biologia o seguinte:



Teste de Aderência

Objetivo: Testar quão adequado é assumir um modelo probabilístico para descrever um determinado conjunto de dados.

Exemplo: Vocês já devem ter visto em alguma aula de Biologia o seguinte:



3 genótipos (categorias): AA, Aa e aa.

Teste de Aderência

Em uma certa população, 100 descendentes foram estudados, fornecendo a tabela a seguir:

Teste de Aderência

Em uma certa população, 100 descendentes foram estudados, fornecendo a tabela a seguir:

Genótipo	AA	Aa	aa	Total
Frequência Observada	26	45	29	100

Teste de Aderência

Em uma certa população, 100 descendentes foram estudados, fornecendo a tabela a seguir:

Genótipo	AA	Aa	aa	Total
Frequência Observada	26	45	29	100

Objetivo: Verificar se o modelo genético proposto (Equilíbrio de Hardy-Weinberg) é adequado para essa população.

Teste de Aderência

Se o modelo teórico for adequado, a freqüência esperada de descendentes para o genótipo AA, dentre os 100 indivíduos, pode ser calculada por:

$$100 \times P(AA) = 100 \times \frac{1}{4} = 25.$$

Teste de Aderência

Se o modelo teórico for adequado, a freqüência esperada de descendentes para o genótipo AA, dentre os 100 indivíduos, pode ser calculada por:

$$100 \times P(AA) = 100 \times \frac{1}{4} = 25.$$

Da mesma forma para o genótipo Aa:

$$100 \times P(Aa) = 100 \times \frac{1}{2} = 50.$$

Teste de Aderência

Se o modelo teórico for adequado, a freqüência esperada de descendentes para o genótipo AA, dentre os 100 indivíduos, pode ser calculada por:

$$100 \times P(AA) = 100 \times \frac{1}{4} = 25.$$

Da mesma forma para o genótipo Aa:

$$100 \times P(Aa) = 100 \times \frac{1}{2} = 50.$$

E para o genótipo aa:

$$100 \times P(aa) = 100 \times \frac{1}{4} = 25.$$

Teste de Aderência

Podemos expandir a tabela de frequências dada anteriormente com as frequências esperadas sob o modelo teórico:

Teste de Aderência

Podemos expandir a tabela de frequências dada anteriormente com as frequências esperadas sob o modelo teórico:

Genótipo	AA	Aa	aa	Total
Frequência Observada	26	45	29	100
Frequência Esperada	25	50	25	100

Teste de Aderência

Podemos expandir a tabela de frequências dada anteriormente com as frequências esperadas sob o modelo teórico:

Genótipo	AA	Aa	aa	Total
Frequência Observada	26	45	29	100
Frequência Esperada	25	50	25	100

Pergunta: Podemos afirmar que os valores observados estão suficientemente próximos dos valores esperados, de tal forma que o modelo genético teórico é adequado a esta população?

Teste de Aderência

Podemos expandir a tabela de frequências dada anteriormente com as frequências esperadas sob o modelo teórico:

Genótipo	AA	Aa	aa	Total
Frequência Observada	26	45	29	100
Frequência Esperada	25	50	25	100

Pergunta: Podemos afirmar que os valores observados estão suficientemente próximos dos valores esperados, de tal forma que o modelo genético teórico é adequado a esta população?

O procedimento que responde esse tipo de pergunta é chamado de **teste de bondade de ajuste** ou **teste de aderência**.

Teste de Aderência - Procedimento

Considere uma tabela de freqüências, com $k \geq 2$ categorias de resultados:

Categorias	1	2	...	k	Total
Frequência Observada	O_1	O_2	...	O_k	n

Teste de Aderência - Procedimento

Considere uma tabela de freqüências, com $k \geq 2$ categorias de resultados:

Categorias	1	2	...	k	Total
Frequência Observada	O_1	O_2	...	O_k	n

Sendo O_i o total de indivíduos observados na categoria i, $i = 1, 2, \ldots, k$.

Teste de Aderência - Procedimento

Considere uma tabela de freqüências, com $k \geq 2$ categorias de resultados:

Categorias	1	2	...	k	Total
Frequência Observada	O_1	O_2	...	O_k	n

Sendo O_i o total de indivíduos observados na categoria i, $i = 1, 2, \ldots, k$.

Seja p_i a probabilidade associada à categoria i, $i = 1, 2, \ldots, k$.

Teste de Aderência - Procedimento

Considere uma tabela de freqüências, com $k \geq 2$ categorias de resultados:

Categorias	1	2	...	k	Total
Frequência Observada	O_1	O_2	...	O_k	n

Sendo O_i o total de indivíduos observados na categoria i, $i = 1, 2, \ldots, k$.

Seja p_i a probabilidade associada à categoria i, $i = 1, 2, \ldots, k$.

O objetivo do teste de aderência é testar as hipóteses

$$\begin{aligned} H_0 : & \quad p_1 = p_{01}, \ldots, p_k = p_{0k} \\ H_a : & \quad \text{existe pelo menos uma diferença,} \end{aligned}$$

sendo p_{0i} a probabilidade da categoria i sob o modelo teórico e $\sum_{i=1}^{k} p_{0i} = 1$.

Teste de Aderência - Procedimento

Se E_i é o total de indivíduos esperados na categoria i, quando a hipótese nula H_0 é verdadeira, então:

Teste de Aderência - Procedimento

Se E_i é o total de indivíduos esperados na categoria i, quando a hipótese nula H_0 é verdadeira, então:

$$E_i = n \times p_{0i}, \quad i = 1, 2, \ldots, k.$$

Teste de Aderência - Procedimento

Se E_i é o total de indivíduos esperados na categoria i, quando a hipótese nula H_0 é verdadeira, então:

$$E_i = n \times p_{0i}, \quad i = 1, 2, \ldots, k.$$

Então, expandindo a tabela de freqüências original, temos

Categorias	1	2	...	k	Total
Frequência Observada	O_1	O_2	...	O_k	n
Frequência Esperada	E_1	E_2	...	E_k	n

Teste de Aderência - Procedimento

Para quantificar quão distante os frequências observadas estão das frequências esperadas, usamos a seguinte estatística:

Teste de Aderência - Procedimento

Para quantificar quão distante os frequências observadas estão das frequências esperadas, usamos a seguinte estatística:

Estatística do Teste:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{k} \frac{(n_i - np_{0i})^2}{np_{0i}}.$$

Teste de Aderência - Procedimento

Para quantificar quão distante os frequências observadas estão das frequências esperadas, usamos a seguinte estatística:

Estatística do Teste:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{k} \frac{(n_i - np_{0i})^2}{np_{0i}}.$$

Se H_0 é verdadeira: $\chi^2 \sim \chi^2_{k-1}$.

Teste de Aderência - Procedimento

Para quantificar quão distante os frequências observadas estão das frequências esperadas, usamos a seguinte estatística:

Estatística do Teste:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{k} \frac{(n_i - np_{0i})^2}{np_{0i}}.$$

Se H_0 é verdadeira: $\chi^2 \sim \chi^2_{k-1}$.

Em outras palavras, se H_0 é verdadeira, a v.a. χ^2 segue uma distribuição aproximadamente Qui-quadrado com $k-1$ graus de liberdade.

Teste de Aderência - Procedimento

Para quantificar quão distante os frequências observadas estão das frequências esperadas, usamos a seguinte estatística:

Estatística do Teste:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{k} \frac{(n_i - np_{0i})^2}{np_{0i}}.$$

Se H_0 é verdadeira: $\chi^2 \sim \chi^2_{k-1}$.

Em outras palavras, se H_0 é verdadeira, a v.a. χ^2 segue uma distribuição aproximadamente Qui-quadrado com $k-1$ graus de liberdade.

Condição: Este resultado é válido para n grande e para frequências esperadas maiores ou iguais a 5.

Teste de Aderência - Procedimento

Calcular o **valor-de-p** ou encontrar o **valor crítico**.

Teste de Aderência - Procedimento

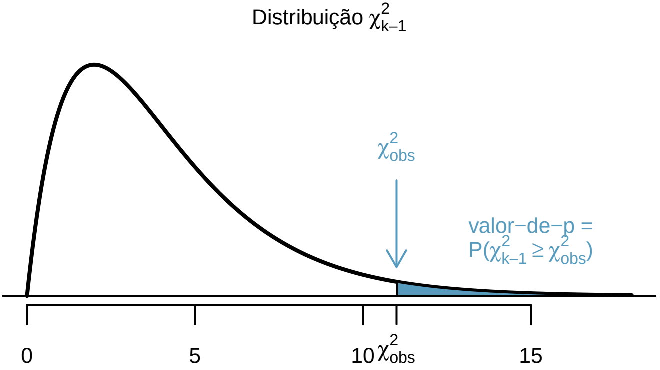
Calcular o **valor-de-p** ou encontrar o **valor crítico**.

Valor-de-p: $P(\chi^2_{k-1} \geq \chi^2_{obs})$, em que χ^2_{obs} é o valor da estatística do teste calculada a partir dos dados.

Teste de Aderência - Procedimento

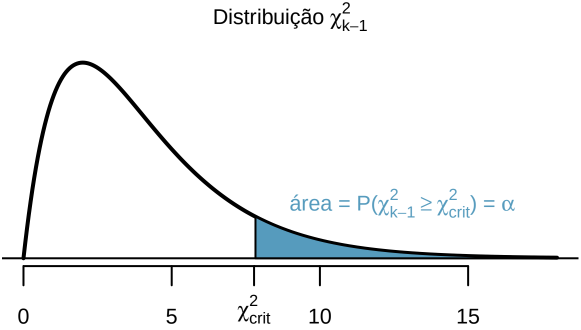
Calcular o **valor-de-p** ou encontrar o **valor crítico**.

Valor-de-p: $P(\chi^2_{k-1} \geq \chi^2_{obs})$, em que χ^2_{obs} é o valor da estatística do teste calculada a partir dos dados.



Teste de Aderência - Procedimento

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



Conclusão: Rejeitamos H_0 se

$$\text{valor-de-p} \leq \alpha \quad \text{ou} \quad \chi^2_{obs} \geq \chi^2_{crit}.$$

Tabela da Distribuição Chi-Quadrado

Tabela IV: Distribuição Qui-Quadrado

Fornece o quantil χ^2_p em função do nº de g.l. ν (linha) e de $p = P(\chi^2 \leq \chi^2_p)$ (coluna). χ^2 tem distribuição qui-quadrado com ν g.l.

ν \ p	**0,005**	**0,010**	**0,025**	**0,050**	**0,100**	**0,250**	**0,500**	**0,750**	**0,900**	**0,950**	**0,975**	**0,990**	**0,995**
1	0,000	0,000	0,001	0,004	0,016	0,102	0,455	1,323	2,706	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	0,211	0,575	1,386	2,773	4,605	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	0,584	1,213	2,366	4,108	6,251	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	1,064	1,923	3,357	5,385	7,779	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	1,610	2,675	4,351	6,626	9,236	11,070	12,833	15,086	16,750
6	0,676	0,872	1,237	1,635	2,204	3,455	5,348	7,841	10,645	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	2,833	4,255	6,346	9,037	12,017	14,067	16,013	18,475	20,278
8	1,344	1,646	2,180	2,733	3,490	5,071	7,344	10,219	13,362	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	4,168	5,899	8,343	11,389	14,684	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	4,865	6,737	9,342	12,549	15,987	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	5,578	7,584	10,341	13,701	17,275	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	6,304	8,438	11,340	14,845	18,549	21,026	23,337	26,217	28,300
13	3,565	4,107	5,009	5,892	7,042	9,299	12,340	15,984	19,812	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	7,790	10,165	13,339	17,117	21,064	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	8,547	11,037	14,339	18,245	22,307	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	9,312	11,912	15,338	19,369	23,542	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	10,085	12,792	16,338	20,489	24,769	27,587	30,191	33,409	35,718

Exemplo: Genética

Voltando no exemplo da Genética.

Exemplo: Genética

Voltando no exemplo da Genética.

Hipóteses:

$$\begin{aligned} H_0 : &\quad \text{o modelo proposto é adequado} \\ H_a : &\quad \text{o modelo proposto não é adequado.} \end{aligned}$$

Exemplo: Genética

Voltando no exemplo da Genética.

Hipóteses:

$$\begin{aligned} H_0 &: \quad \text{o modelo proposto é adequado} \\ H_a &: \quad \text{o modelo proposto não é adequado.} \end{aligned}$$

Que de forma equivalente, podem ser escritas como:

$$\begin{aligned} H_0 &: \quad p_1 = 1/4, p_2 = 1/2, p_3 = 1/4 \\ H_a &: \quad \text{ao menos umas das desigualdades não verifica,} \end{aligned}$$

sendo $p_1 = P(AA), p_2 = P(Aa)$ e $p_3 = P(aa)$.

Exemplo: Genética

A tabela seguinte apresenta os valores observados e esperados (calculados anteriormente).

Genótipo	AA	Aa	aa	Total
Frequência Observada	26	45	29	100
Frequência Esperada	25	50	25	100

Exemplo: Genética

A tabela seguinte apresenta os valores observados e esperados (calculados anteriormente).

Genótipo	AA	Aa	aa	Total
Frequência Observada	26	45	29	100
Frequência Esperada	25	50	25	100

Estatística do Teste:

$$\chi^2_{obs} = \sum_{i=1}^{3} \frac{(O_i - E_i)^2}{E_i} = \frac{(26-25)^2}{25} + \frac{(45-50)^2}{50} + \frac{(29-25)^2}{25}$$
$$= 0.04 + 0.5 + 0.64 = 1.18.$$

Exemplo: Genética

Sob H_0, a estatística $\chi^2 \sim \chi^2_2$. Veja que os graus de liberdade é o número de categorias menos 1. Então o valor-de-p é dado por:

$$\text{valor-de-p} = P(\chi^2_2 \geq \chi^2_{obs}) = P(\chi^2_2 \geq 1.18) = 0.554.$$

Exemplo: Genética

Sob H_0, a estatística $\chi^2 \sim \chi^2_2$. Veja que os graus de liberdade é o número de categorias menos 1. Então o valor-de-p é dado por:

$$\text{valor-de-p} = P(\chi^2_2 \geq \chi^2_{obs}) = P(\chi^2_2 \geq 1.18) = 0.554.$$

Para um nível de significância $\alpha = 0.1$, olhando na Tabela Qui-Quadrado, o valor crítico é: $\chi^2_{crit} = 7.779$.

Exemplo: Genética

Sob H_0, a estatística $\chi^2 \sim \chi^2_2$. Veja que os graus de liberdade é o número de categorias menos 1. Então o valor-de-p é dado por:

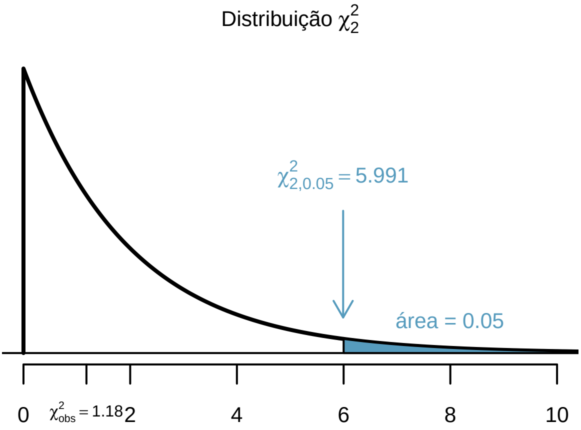
$$\text{valor-de-p} = P(\chi^2_2 \geq \chi^2_{obs}) = P(\chi^2_2 \geq 1.18) = 0.554.$$

Para um nível de significância $\alpha = 0.1$, olhando na Tabela Qui-Quadrado, o valor crítico é: $\chi^2_{crit} = 7.779$.

Conclusão: Para $\alpha = 0.1$, como valor-de-p$= 0.554 > 0.1$, não rejeitamos a hipótese H_0, isto é, essa população segue o modelo genético proposto.

Ou como $\chi^2_{obs} = 1.18 < 7.779 = \chi^2_{crit}$, não rejeitamos a hipótese H_0.

Exemplo: Genética



Exemplo: Cores de Geladeira

Voltando aos dados do exemplo das cores da geladeira, cujas componentes têm frequências multinomiais, a hipótese nula especifica que as cinco cores são igualmente populares. Ou seja,

$$
\begin{aligned}
H_0 : &\quad p_1 = p_2 = \ldots = p_k = 1/5 \\
H_a : &\quad \text{existe pelo menos uma diferença.}
\end{aligned}
$$

Exemplo: Cores de Geladeira

Voltando aos dados do exemplo das cores da geladeira, cujas componentes têm frequências multinomiais, a hipótese nula especifica que as cinco cores são igualmente populares. Ou seja,

$$H_0: \quad p_1 = p_2 = \ldots = p_k = 1/5$$

$$H_a: \quad \text{existe pelo menos uma diferença.}$$

Componente	marrom	creme	vermelho	azul	branco	total
Frequência Observada	88	65	52	40	55	300

Exemplo: Cores de Geladeira

Voltando aos dados do exemplo das cores da geladeira, cujas componentes têm frequências multinomiais, a hipótese nula especifica que as cinco cores são igualmente populares. Ou seja,

$$
\begin{aligned}
H_0: &\quad p_1 = p_2 = \ldots = p_k = 1/5 \\
H_a: &\quad \text{existe pelo menos uma diferença.}
\end{aligned}
$$

Componente	marrom	creme	vermelho	azul	branco	total
Frequência Observada	88	65	52	40	55	300

Como as probabilidades das componentes na hipótese nula são todas iguais, as frequências esperadas também serão todas iguais, ou seja,

$$E_i = n \times \frac{1}{5} = 300 \times \frac{1}{5} = 60, \quad i = 1, 2, 3, 4, 5.$$

Exemplo: Cores de Geladeira

Componente	marrom	creme	vermelho	azul	branco	total
Frequência Observada	88	65	52	40	55	300
Frequência Esperada	60	60	60	60	60	300
$\frac{(O-E)^2}{E}$	13.07	0.42	1.07	6.67	0.42	21.63

Exemplo: Cores de Geladeira

Componente	marrom	creme	vermelho	azul	branco	total
Frequência Observada	88	65	52	40	55	300
Frequência Esperada	60	60	60	60	60	300
$\frac{(O-E)^2}{E}$	13.07	0.42	1.07	6.67	0.42	21.63

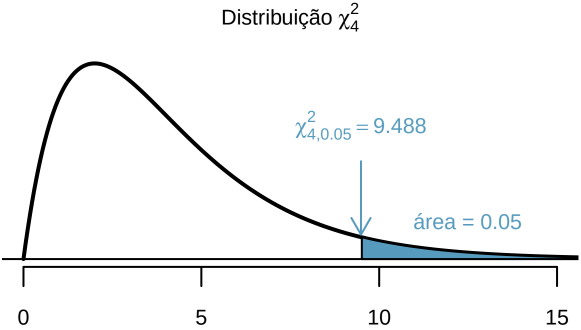
Estatística do Teste:

$$\chi^2 = \sum_{i=1}^{5} \frac{(O_i - E_i)^2}{E_i} \quad = \quad 13.07 + 0.42 + 1.07 + 6.67 + 0.42$$

$$= \quad 21.63.$$

Exemplo: Cores de Geladeira

Olhando na tabela Qui-quadrado com 4 graus de liberdade, para $\alpha = 0.05$, o valor crítico é $\chi^2_{crit} = \chi^2_{4,0.05} = 9.488$.



Conclusão: Para $\alpha = 0.05$, como $\chi^2_{obs} = 21.63 > 9.488 = \chi^2_{crit}$, rejeitamos a hipótese de que as cinco cores são igualmente populares.

Exemplo: Tipo Sanguíneo

Entre os americanos, 41% tem sangue do tipo A, 9% tem sangue tipo B, 4% tipo AB e 46% tem sangue tipo O.

Em uma amostra aleatória de 200 pacientes americanos com câncer de estômago, 92 pacientes têm sangue do tipo A, 20 do tipo B, 4 do tipo AB e 84 do tipo O.

Tipo	A	B	AB	O	total
Frequência Observada	92	20	4	84	200

Essas frequências observadas trazem evidência contra a hipótese de que a distribuição do tipo sanguíneo dos pacientes é igual à distribuição dos tipos sanguíneos na população geral americana? Use nível de significância $\alpha = 0.05$.

Exemplo: Tipo Sanguíneo

$$
\begin{aligned}
H_0: &\quad p_1 = 0.41, p_2 = 0.09, p_3 = 0.04, p_4 = 0.46 \\
H_a: &\quad \text{existe pelo menos uma diferença.}
\end{aligned}
$$

Exemplo: Tipo Sanguíneo

$$H_0: \quad p_1 = 0.41, p_2 = 0.09, p_3 = 0.04, p_4 = 0.46$$

$$H_a: \qquad \text{existe pelo menos uma diferença.}$$

Tipo	A	B	AB	O	total
Frequência Observada	92	20	4	84	200
Frequência Esperada	82	18	8	92	200
$\dfrac{(O-E)^2}{E}$	1.22	0.22	2	0.7	4.14

Exemplo: Tipo Sanguíneo

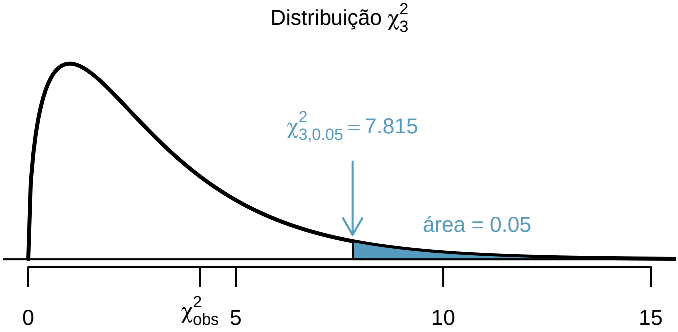
$$H_0: \quad p_1 = 0.41, p_2 = 0.09, p_3 = 0.04, p_4 = 0.46$$

$$H_a: \qquad \text{existe pelo menos uma diferença.}$$

Tipo	A	B	AB	O	total
Frequência Observada	92	20	4	84	200
Frequência Esperada	82	18	8	92	200
$\frac{(O-E)^2}{E}$	1.22	0.22	2	0.7	4.14

Estatística do Teste: $\chi^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = 4.14.$

Exemplo: Tipo Sanguíneo

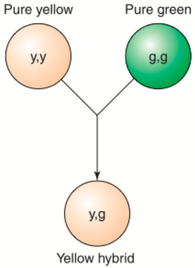


Conclusão: Como $\chi^2_{obs} = 4.14 \leq 7.815 = \chi^2_{3,0.05}$, não temos evidência para rejeitar a hipótese nula.

Portanto, concluímos que não há discrepância significativa entre o que foi observado e a distribuição sanguínea da população americana.

Exemplo: Ervilhas de Mendel

Exemplo: Ervilhas de Mendel



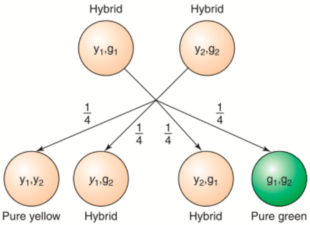


Figura: (Esquerda) Cruzamento de ervilhas puramente amarelas e puramente verdes e (Direta) cruzamento de ervilhas híbridas (Direita)

Exemplo: Ervilhas de Mendel

Mendel fez o cruzamento de 8023 ervilhas híbridas e o resultado foram 6022 ervilhas amarelas e 2001 ervilhas verdes. Teoricamente, cada cruzamento deve resultar em ervilha amarela com probabilidade 3/4 e verde com probabilidade 1/4.

$$\begin{aligned} H_0 : &\quad p_1 = 3/4 \ \text{ e } \ p_2 = 1/4 \\ H_a : &\quad \text{existe pelo menos uma diferença.} \end{aligned}$$

Exemplo: Ervilhas de Mendel

Mendel fez o cruzamento de 8023 ervilhas híbridas e o resultado foram 6022 ervilhas amarelas e 2001 ervilhas verdes. Teoricamente, cada cruzamento deve resultar em ervilha amarela com probabilidade 3/4 e verde com probabilidade 1/4.

$$
\begin{aligned}
H_0 : &\quad p_1 = 3/4 \text{ e } p_2 = 1/4 \\
H_a : &\quad \text{existe pelo menos uma diferença.}
\end{aligned}
$$

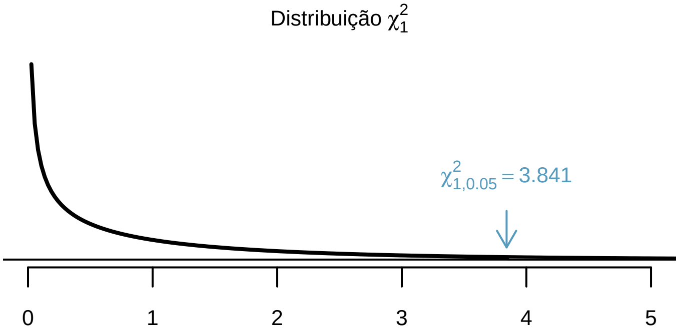
Tipo	Amarela	Verde	total
Frequência Observada	6022	2001	8023
Frequência Esperada	6017.25	2005.75	8023
$\frac{(O-E)^2}{E}$	0.004	0.011	0.015

Exemplo: Ervilhas de Mendel

Estatística do Teste: $\chi^2 = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i} = 0.015.$

Exemplo: Ervilhas de Mendel

Estatística do Teste: $\chi^2 = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i} = 0.015.$



Conclusão: Como $\chi^2_{obs} = 0.015 \leq 3.841 = \chi^2_{1,0.05}$, não temos evidência para rejeitar a hipótese nula. Concluímos que não há discrepância significativa entre o que foi observado e a hipótese nula.

Leituras

- Ross: capítulo 13.
- OpenIntro: seção 6.3