ME111 - Laboratório de Estatística

Aula 12 - Testes Chi-Quadrado

Profa. Larissa Avila Matos

Testes Chi-Quadrado: Aderência e Independência

- 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);
  - 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.

- **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.

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

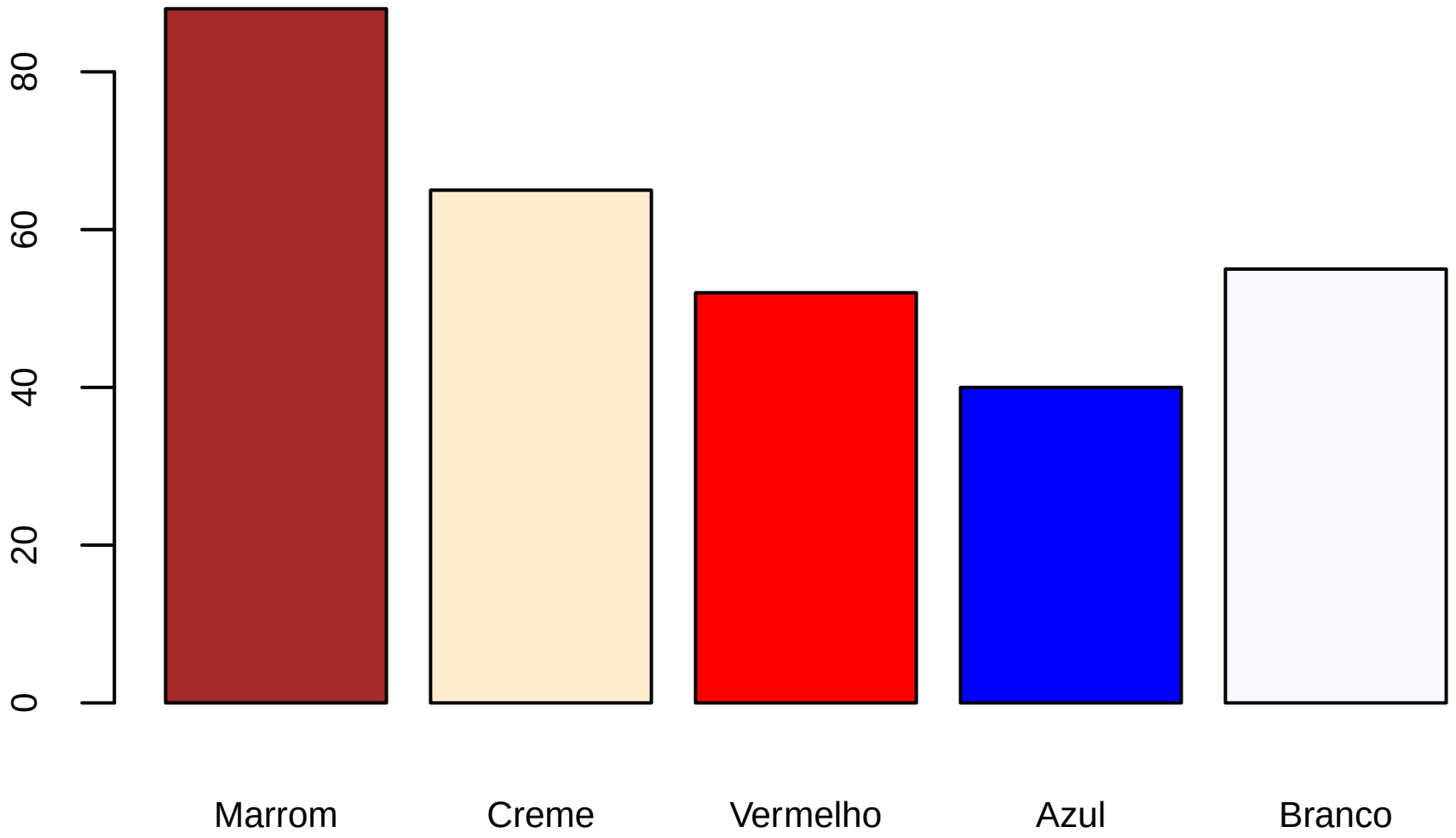
```r
obs <- c(G1=88, G2=65, G3=52, G4=40, G5=55)
obs
```

```
G1 G2 G3 G4 G5
88 65 52 40 55
```

```r
names(obs) <- c("Marrom","Creme","Vermelho","Azul","Branco")
obs
```

```
  Marrom    Creme Vermelho     Azul   Branco
      88       65       52       40       55
```

```
barplot(obs,cex.axis = 0.7,cex.names = 0.7,
        col=c("brown","blanchedalmond","red","blue","ghostwhite"))
```



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

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

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

- **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:

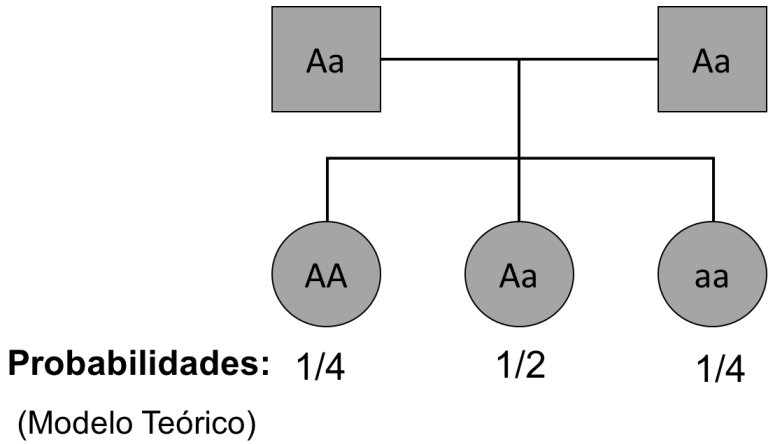


Figure 1: 3 genótipos (categorias): AA, Aa e aa

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.

- 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:

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

- 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

- 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

- 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:

$$H_0 : p_1 = p_{01}, \ldots, p_k = p_{0k},$$

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

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

- 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

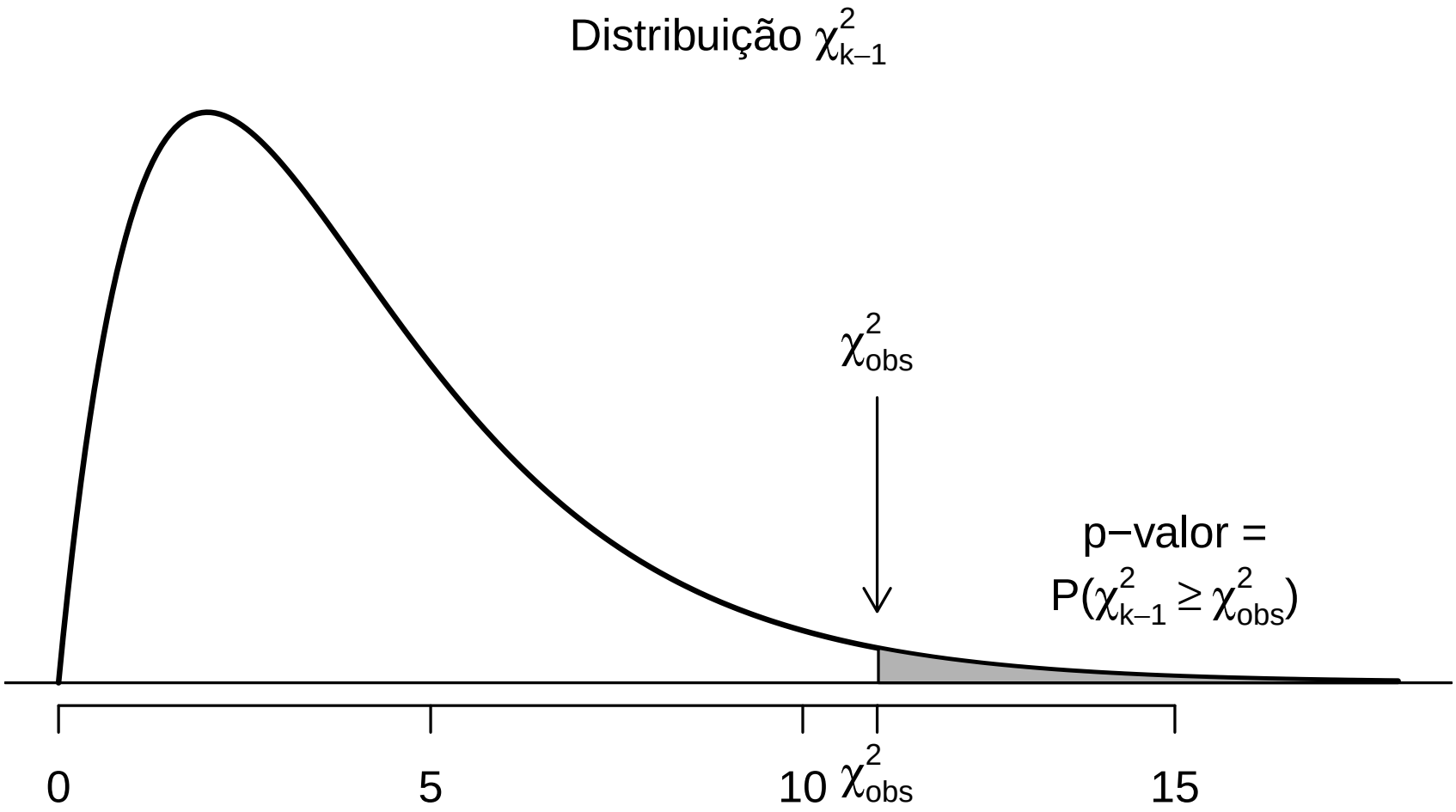
- Para quantificar quão distante os frequências observadas estão das frequências esperadas, usamos a seguinte estatística:

Estatística do Teste:

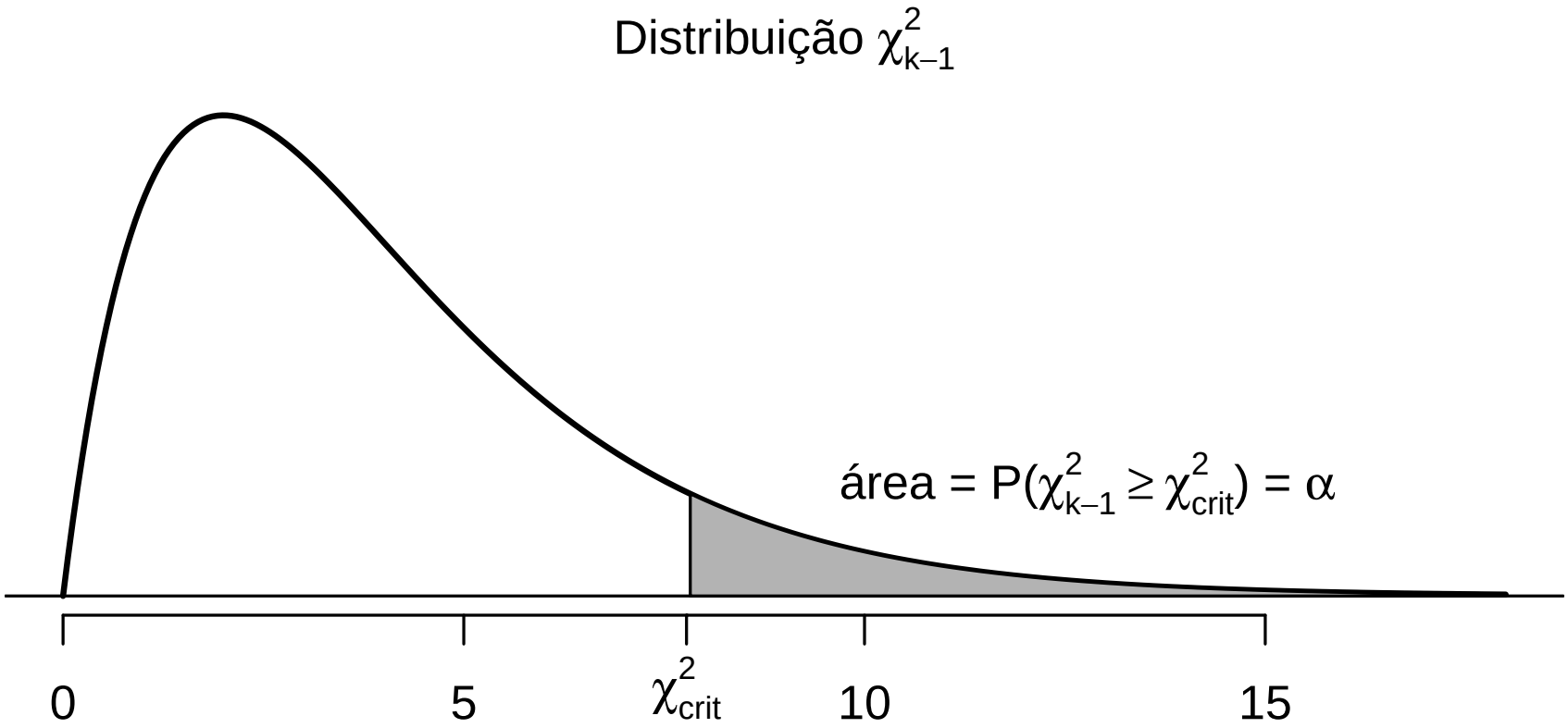
$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{k} \frac{(n_i - np_{0i})^2}{np_{0i}}.$$

- 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.

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



- **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

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

Tabela da Distribuição Chi-Quadrado

Quantis da Distribuição χ^2. Graus de liberdade na margem esquerda da tabela e probabilidades p dadas no topo da tabela tal que $p = P[\chi^2 \geq \chi^2_t]$.

ν/p	99.5%	99%	97.5%	95%	90%	50%	10%	5%	2.5%	1%	0.5%
1	0.00	0.00	0.00	0.00	0.02	0.45	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	1.39	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	2.37	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	3.36	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	4.35	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	5.35	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	6.35	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	7.34	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	4.17	8.34	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	9.34	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	10.34	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	11.34	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	12.34	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	13.34	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	14.34	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	15.34	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	16.34	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	17.34	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	18.34	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	19.34	28.41	31.41	34.17	37.57	40.00
30	13.79	14.95	16.79	18.49	20.60	29.34	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	39.34	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	49.33	63.17	67.50	71.42	76.15	79.49

Distribuição Chi-Quadrado no R

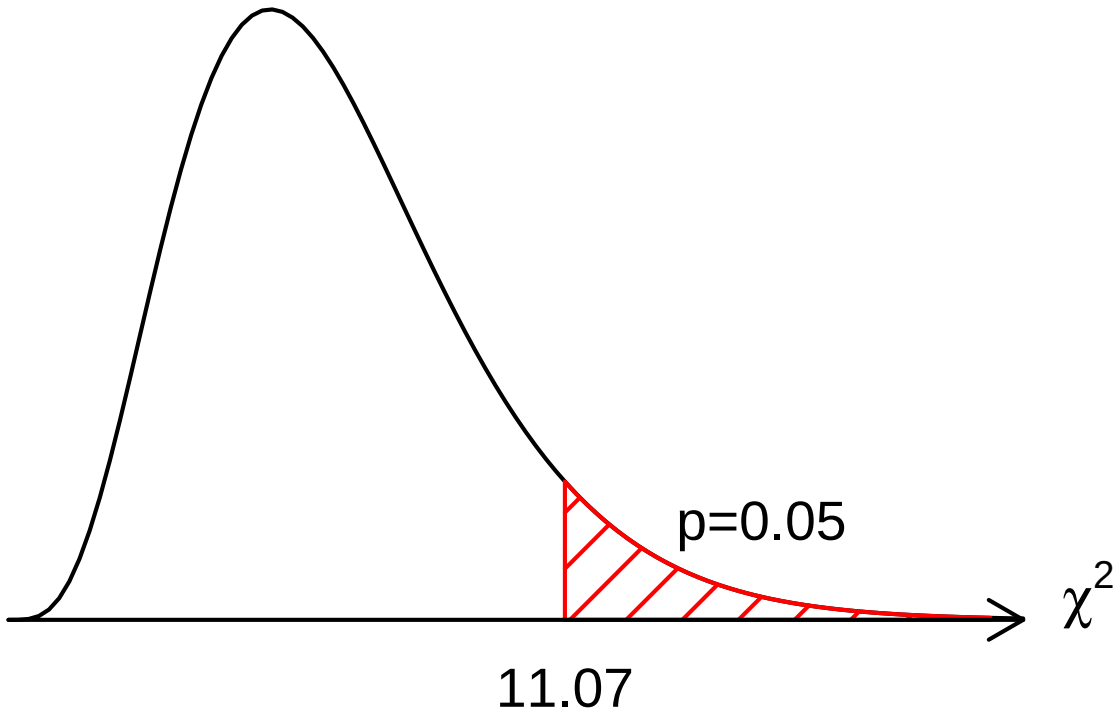
```r
dchisq(x, df, ncp = 0, log = FALSE)
pchisq(q, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)
qchisq(p, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)
rchisq(n, df, ncp = 0)
```

```r
q=qchisq(0.95, 5, ncp = 0, lower.tail = TRUE, log.p = FALSE)
q
```

```
[1] 11.0705
```

```r
1-pchisq(q, 5, ncp = 0, lower.tail = TRUE, log.p = FALSE)
```

```
[1] 0.05
```



Exemplo: Genética

- Voltando no exemplo da Genética:
- **Hipóteses:**

$$H_0 : \text{o modelo proposto é adequado}$$
$$H_a : \text{o modelo proposto não é adequado}$$

- Que de forma equivalente, podem ser escritas como:

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

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

- 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**:

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

$$= 0.04 + 0.5 + 0.64 = 1.18$$

- 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.

No R

```
obs <- c(AA=26, Aa=45, aa=29)
obs
```

```
AA Aa aa
26 45 29
```

```
p0 <- c(0.25, 0.5, 0.25)
xsq <- chisq.test(obs, p=p0)
xsq
```

```
    Chi-squared test for given probabilities

data:  obs
X-squared = 1.18, df = 2, p-value = 0.5543
```

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,

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

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

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,

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

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:**

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:**

$$\chi^2 = \sum_{i=1}^{5} \frac{(O_i - E_i)^2}{E_i} = 13.07 + 0.42 + 1.07 + 6.67 + 0.42 = 21.63$$

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

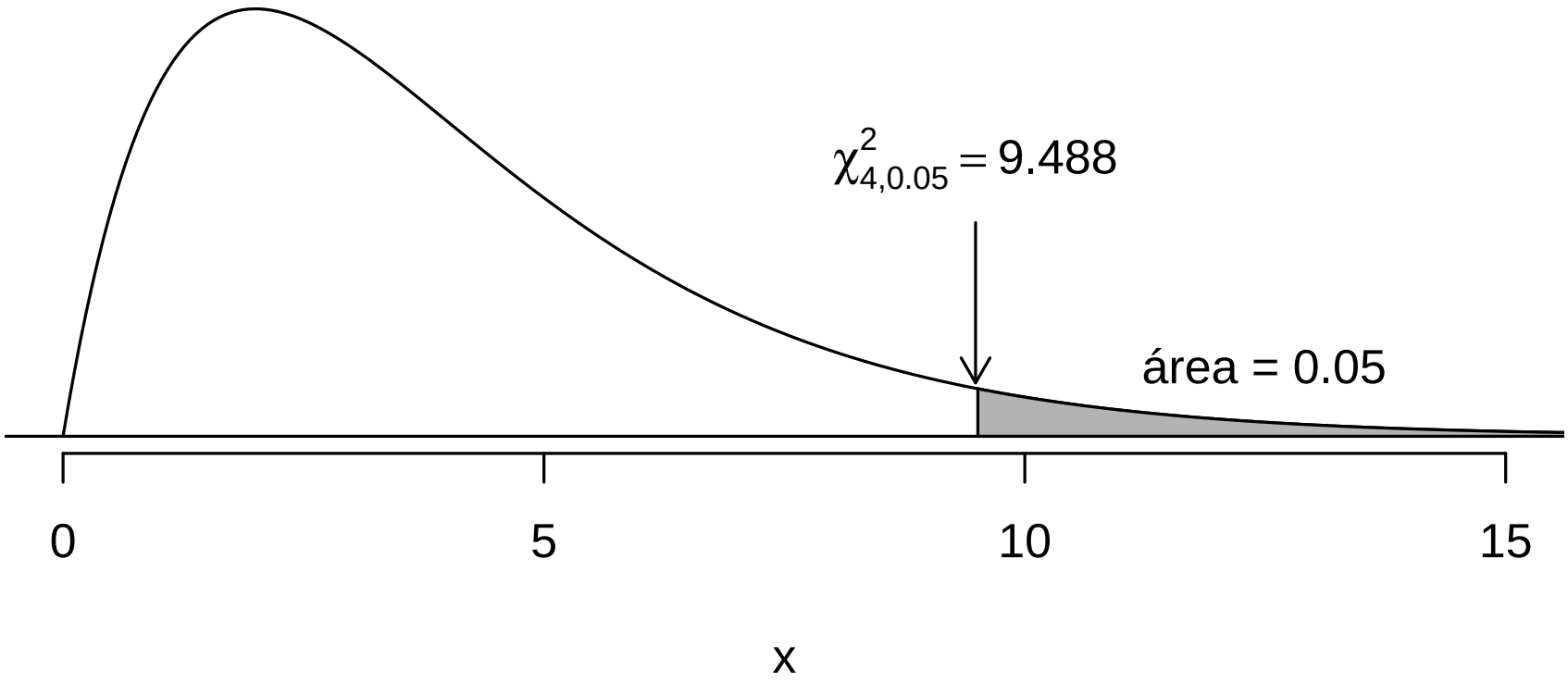
- **Estatística do Teste:**

$$\chi^2 = \sum_{i=1}^{5} \frac{(O_i - E_i)^2}{E_i} = 13.07 + 0.42 + 1.07 + 6.67 + 0.42 = 21.63$$

```
xcrit <- round(qchisq(0.95, df=4), 3)
xcrit
```

```
[1] 9.488
```

- 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.

No R

```
obs <- c(G1=88, G2=65, G3=52, G4=40, G5=55)
obs
```

```
G1 G2 G3 G4 G5
88 65 52 40 55
```

```
p0 <- rep(1/5, 5)
xsq <- chisq.test(obs, p=p0)
xsq
```

```
    Chi-squared test for given probabilities

data:  obs
X-squared = 21.633, df = 4, p-value = 0.0002371
```

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

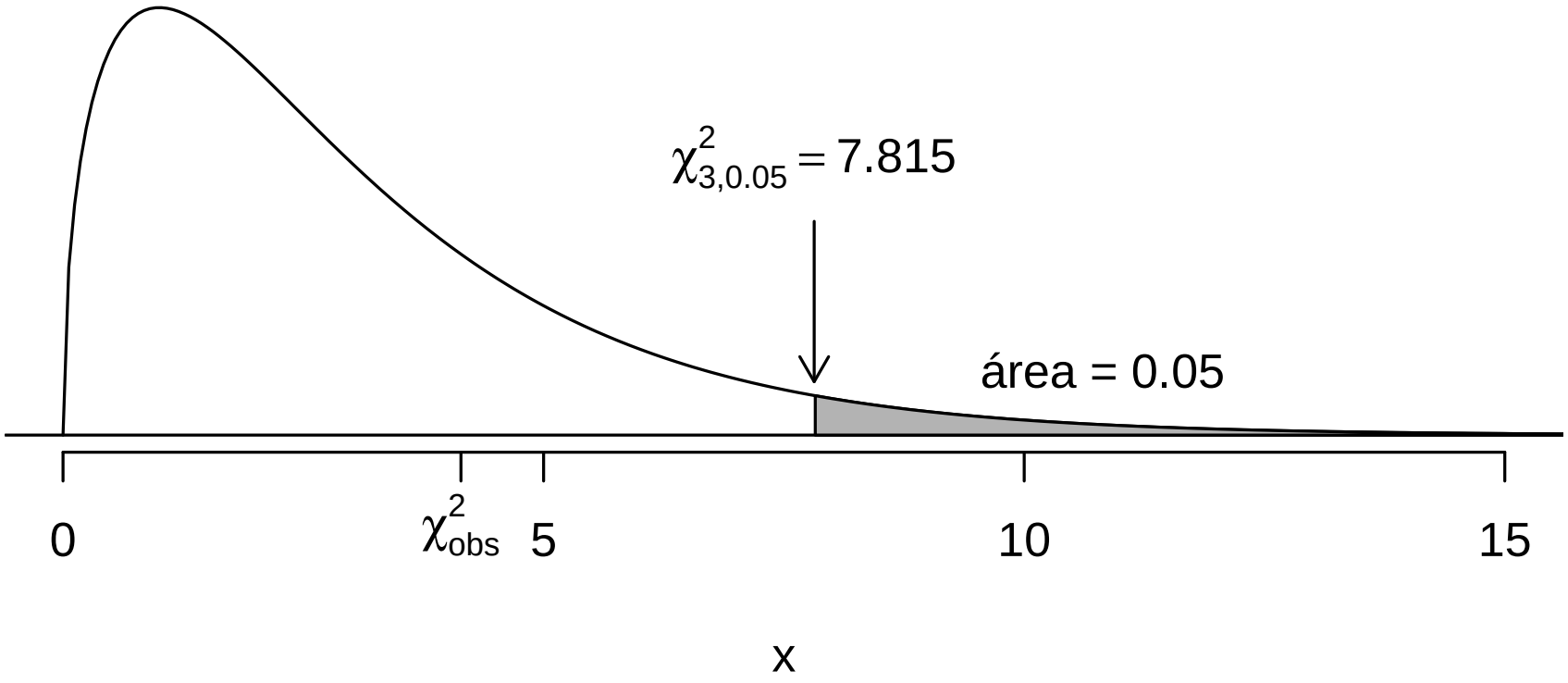
- Essas frequências observadas trazem evidência contra a hipótese de que a distribuição do tipo sanguíneo dos pacientes é igual à distribuição dos tipos sanguíneos na população geral americana? Use nível de significância $\alpha = 0.05$.

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

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

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



- **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

Figure 2: Ervilhas de Mendel

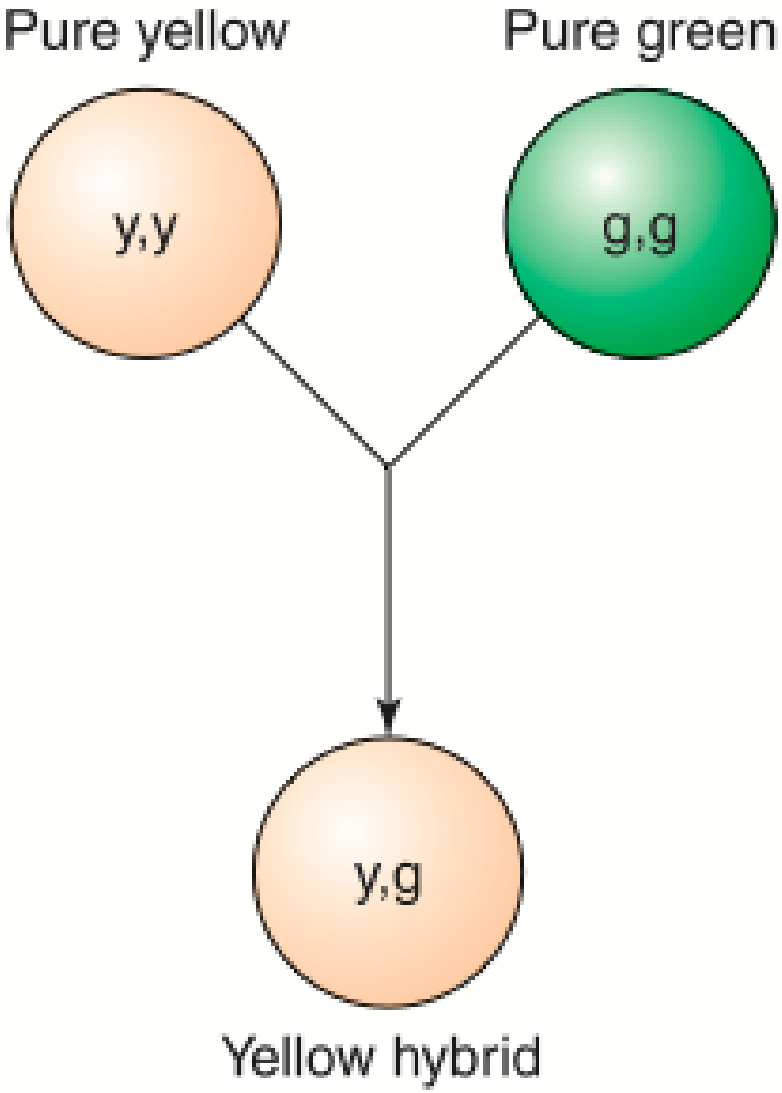


Figure 3: Cruzamento de ervilhas puramente amarelas e puramente verdes

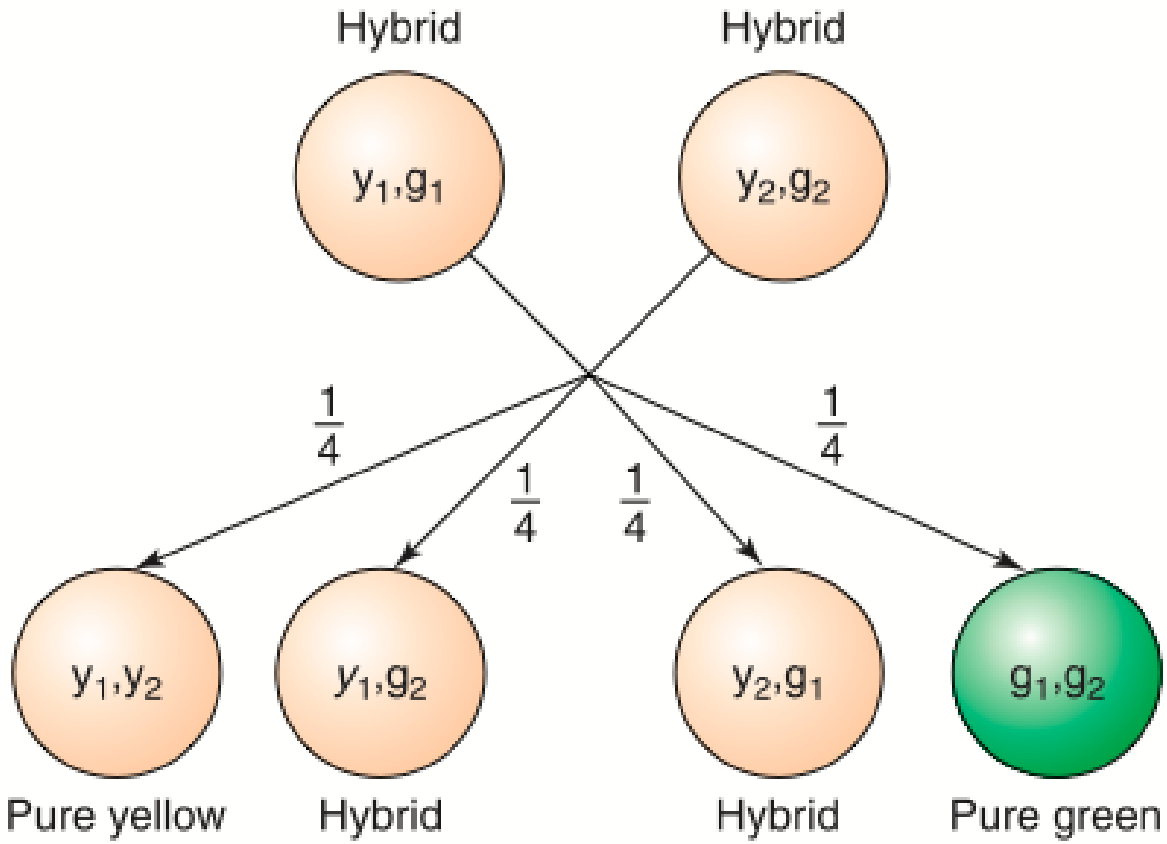


Figure 4: Cruzamento de ervilhas puramente amarelas e puramente verdes

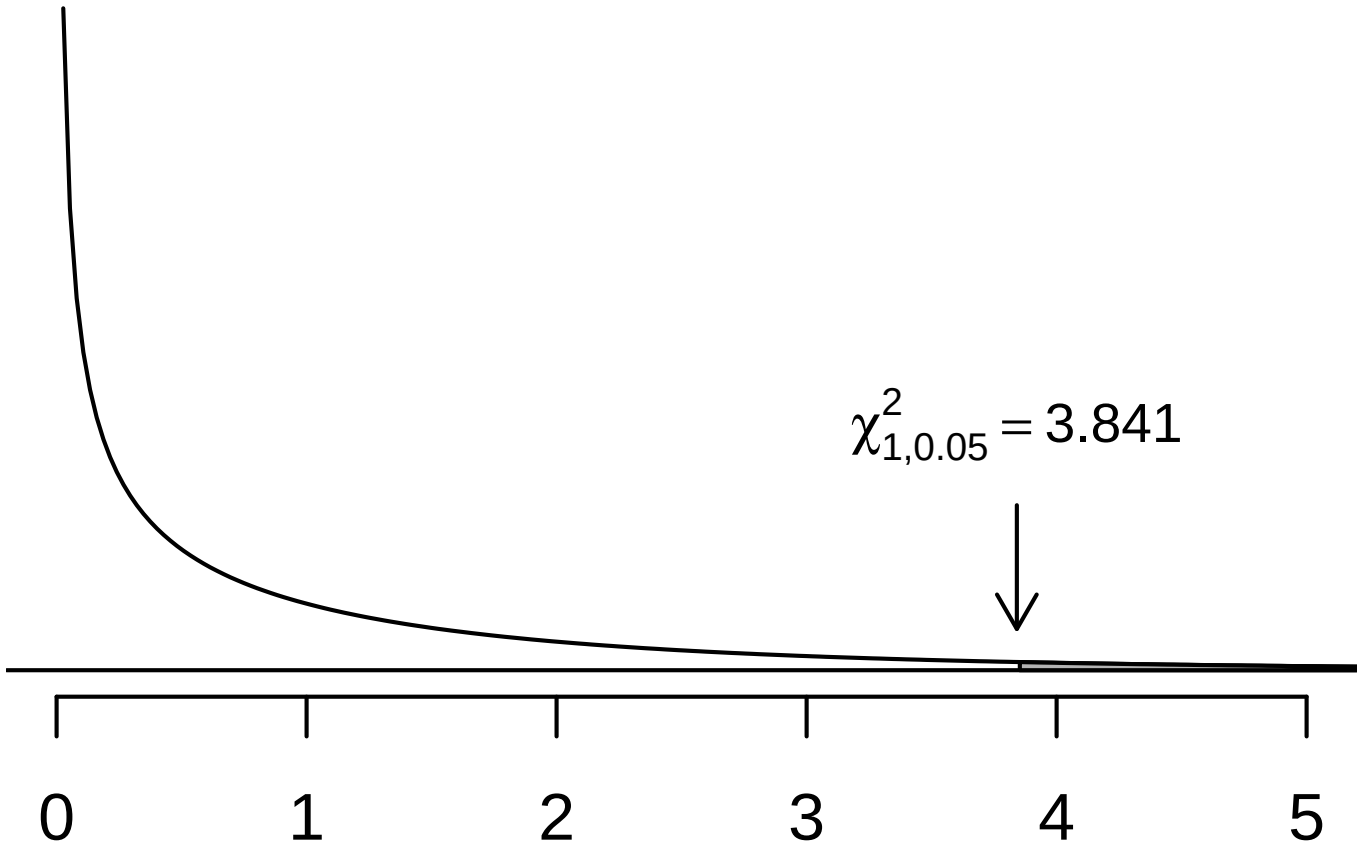
- Mendel fez o cruzamento de 8023 ervilhas híbridas e o resultado foram 6022 ervilhas amarelas e 2001 ervilhas verdes. Teoricamente, cada cruzamento deve resultar em ervilha amarela com probabilidade 3/4 e verde com probabilidade 1/4.

$$H_0 : p_1 = 3/4 \text{ e } p_2 = 1/4$$

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

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

- **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.

No R

```
obs <- c(Amarelas=6022, Verdes=2001)
obs
```

```
Amarelas   Verdes
    6022     2001
```

```
p0 <- c(3/4, 1/4)
xsq <- chisq.test(obs, p=p0)
xsq
```

```
    Chi-squared test for given probabilities

data:  obs
X-squared = 0.014999, df = 1, p-value = 0.9025
```

Referências

- Notas de aula da Profa. Samara F. Kiihl, Profa. Tatiana Benaglia e do Prof. Benilton Carvalho - ME414

- Wardrop, R. L. (1995). *Statistics: Learning in the presence of variation.*

- *Bussab, W. O. & Morettin, P. A. (1987). Estatística Básica. Atual Editora Ltda., São Paulo.*