

Universidade Federal do Rio Grande do Norte
Instituto Metrópole Digital
IMD0601 – Bioestatística

Teste Qui-Quadrado (χ^2)

Renata Lilian Dantas Cavalcante
Mestranda PPG-Bioinformática - UFRN



Estatística Inferencial

Fundamentos da Estatística Inferencial

- Teoria da Amostragem;
- Distribuição Amostral;
- Estimativa de Parâmetros;
- Bootstrapping;

Teste de Hipótese

Técnica para realizar uma **inferência estatística** sobre uma população a partir de uma amostra.



Teoria Popperiana

Não se pode provar nada, apenas “desaprovar”;

Só aprendemos quando erramos;

É mais fácil refutar do que provar alguma assertiva;

Os estatísticos não perguntam qual é a probabilidade de estarem certos, mas a probabilidade de estarem errados;

Para fazerem isso estabelecem uma hipótese nula.

A decorative graphic at the bottom of the slide consists of a series of overlapping, semi-transparent colored shapes in shades of purple, pink, red, and blue, resembling a liquid or paint splash.

(Karl Popper)

O que é o Teste Qui-Quadrado ?



Testes de adequação de amostras e associação entre variáveis;

Alternativa estatística para testar hipóteses;

Comparação entre os Valores observados (V_o) com os Valores esperados (V_e);

Erros aleatórios \neq Erros sistemáticos (viés);

H_0 Hipótese nula $\rightarrow o_i = e_i$ (hipótese estatística a ser testada);

H_1 Hipótese alternativa $\rightarrow o_i \neq e_i$;

$$X^2 = \sum_{i=1}^n \frac{(o_i - e_i)^2}{e_i}$$

- o_i : valores observados.
- e_i : valores esperados.

Teste não-paramétrico (não depende de parâmetros populacionais - variância e média).

Divide-se em 2 categorias:

- Teste de adequação/aderência (hipótese ou concordância):
Verifica uma contagem observada a uma contagem esperada.

- Teste de independência/associação entre duas variáveis:
Observa se o comportamento de uma variável depende de outra.



Quando utilizar ?



Condições

Os grupos analisados neste teste devem ser independentes;

Os itens de cada grupo devem ser selecionados aleatoriamente;

As observações devem ser frequências ou contagens;

Cada observação pertence a uma e somente uma categoria;

Restrições ao uso:

Se o número de classes é $k = 2$, o valor esperado mínimo deve ser ≥ 5 .

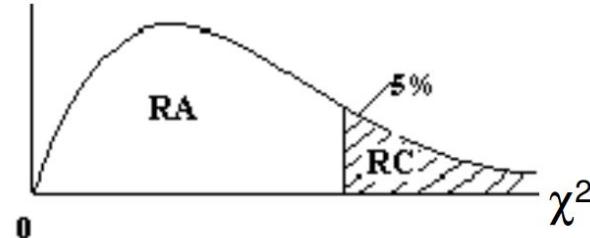
Se $k > 2$, o teste não deve ser usado se mais de 20% dos valores esperados forem abaixo de 5 ou se qualquer um deles for inferior a 1.

Procedimentos do Teste Qui-Quadrado

1º Enunciar as hipóteses (H_0 e H_1);

2º Fixar α , escolher a variável $\chi^2 \varphi - df$ (graus de liberdade) $\rightarrow (k-1)$. Onde k é o número de eventos;

3º Com auxílio da tabela de χ^2 , determinar RA (região de aceitação de H_0) e RC (região de rejeição de H_0).



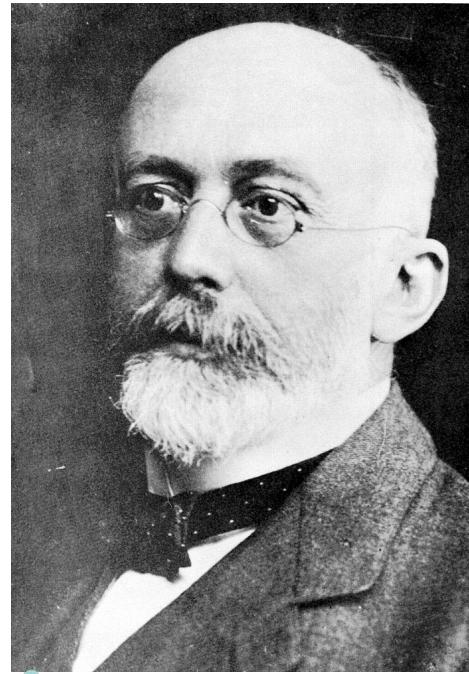
Equilíbrio de Hardy - Weinberg





Godfrey Harold Hardy

Inglaterra



Wilhelm Weinberg

Alemanha

Equilíbrio de Hardy – Weinberg

Teorema de Equilíbrio de Hardy - Weinberg

Em uma população mendeliana, sob determinadas condições, as frequências gênicas permanecem constantes com o passar das gerações.



Equilíbrio de Hardy – Weinberg

População mendeliana (população ideal)

- População grande;
- Reprodução sexuada;
- Acasalamentos aleatórios;
- Diplóide;
- É infinita;
- Todos os casais sendo férteis e possuindo o mesmo número de proles;

Sob a condição de NÃO sofrer:

- Seleção natural;
- Mutação;
- Migração;
- Deriva genética;



Equilíbrio de Hardy – Weinberg

$$p + q = 1$$

p = dominante

q = recessivo

$$p^2 + 2pq + q^2 = 1$$

Onde:

p^2 = genótipo AA

$2pq$ = genótipo Aa ou aA

q^2 = genótipo aa



Ex.: 1 gene, 2 alelos, 3 genótipos:

A1A1, A1A2, A2A2

p = frequência de A1

q = frequência de A2

$$p^2 + 2pq + q^2 = 1$$

A1 (p)	A1A1 (p^2)	A1A2 (pq)
A2 (q)	A1A2 (pq)	A2A2 (q^2)

Caso a população não esteja em equilíbrio, as frequências genotípicas irão ser diferentes das esperadas em p^2 , $2pq$ e q^2 .

Para conferir → **Teste do Qui-quadrado**

Onde:

Grau de liberdade 3 (genótipos) - 2 (alelos) = 1

Nível de significância de 5%

Teste de hipótese

Você está estudando uma espécie de besouro cujo padrão de coloração intraespecífica é variável e é codificado pelo **gene A** que possui dois alelos, sendo os genótipos **A1A1** (cor branca), **A1A2** (variegado) e **A2A2** (cor verde). Sua hipótese é que o fenótipo verde possui um maior valor adaptativo. Você observa 700 indivíduos verdes, 200 variegados e 100 brancos. Avalie se esta população está em equilíbrio de Hardy –Weinberg. Se não, qual o provável fator evolutivo operante.

→ 1º passo:

$$\Phi df: \text{graus de liberdade} = 3 - 2 = 1$$

P<0,05: resultado observado significativamente diferente do resultado esperado.

→ 2º passo:

$$\text{Total da população} = 1000$$

$$\text{Freq. de A1} = (2 \times 100 + 200) / 2000 = 0,2$$

$$\text{Freq. de A2} = (2 \times 700 + 200) / 2000 = 0,8$$

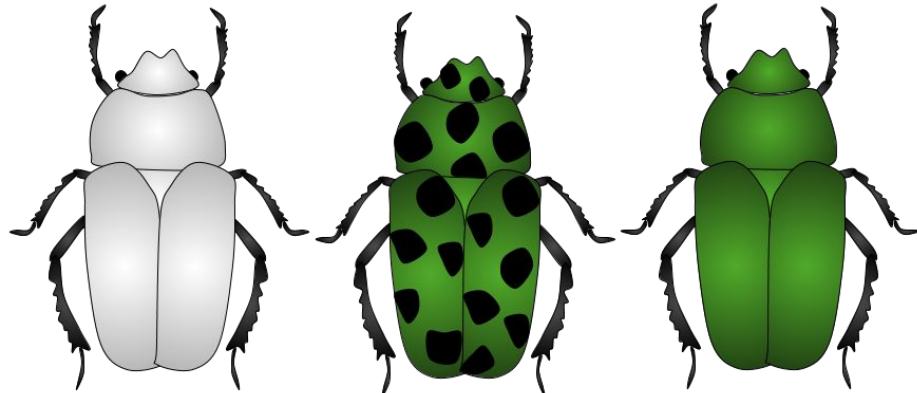
→ 3º passo:

Valores esperados de:

$$A1A1 = (0,2)^2 \times 1000 = 40$$

$$A1A2 = 2 \times (0,2 \times 0,8) \times 1000 = 320$$

$$A2A2 = (0,8)^2 \times 1000 = 640$$



Teste de hipótese

Fórmula do Qui-Quadrado:

$$X^2 = \sum_{i=1}^n \frac{(o_i - e_i)^2}{e_i}$$

→ 4º passo:

$$X^2 = \frac{(100-40)^2}{40} + \frac{(200-320)^2}{320} + \frac{(700-640)^2}{640}$$

$$X^2 = 90 + 45 + 5.625 = 140,625$$

X^2 tabela p/ α 0,05 e 1 GL = 3.841

Conclusões:

O valor de X^2 ultrapassou o tabelado, portanto há diferença significativa indicando que a população não está em equilíbrio de Hardy – Weinberg!
O possível fator operante é a Seleção Natural favorecendo o alelo A2.

Table 3.5 Critical values of the χ^2 distribution

df	P								
	0.995	0.975	0.9	0.5	0.1	0.05*	0.025	0.01	0.005
1	0.000	0.000	0.016	0.455	2.706	3.841	5.024	6.635	7.879
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838
4	0.207	0.484	1.064	3.357	7.779	9.488	11.143	13.277	14.860
5	0.412	0.831	1.610	4.351	9.236	11.070	12.832	15.086	16.750
6	0.676	1.237	2.204	5.348	10.645	12.592	14.449	16.812	18.548
7	0.989	1.690	2.833	6.346	12.017	14.067	16.013	18.475	20.278
8	1.344	2.180	3.490	7.344	13.362	15.507	17.535	20.090	21.955
9	1.735	2.700	4.168	8.343	14.684	16.919	19.023	21.666	23.589
10	2.156	3.247	4.865	9.342	15.987	18.307	20.483	23.209	25.188
11	2.603	3.816	5.578	10.341	17.275	19.675	21.920	24.725	26.757
12	3.074	4.404	6.304	11.340	18.549	21.026	23.337	26.217	28.300
13	3.565	5.009	7.042	12.340	19.812	22.362	24.736	27.688	29.819
14	4.075	5.629	7.790	13.339	21.064	23.685	26.119	29.141	31.319
15	4.601	6.262	8.547	14.339	22.307	24.996	27.488	30.578	32.801

P, probability; df, degrees of freedom.

*Most scientists assume that, when $P < 0.05$, a significant difference exists between the observed and the expected values in a chi-square test.

Teste de Aderência



Objetivo

- Aferir a adequabilidade de um modelo probabilístico a um conjunto de dados observados.

EX.: De acordo com as Lei de segregação proposta por Mendel, os resultados dos cruzamentos de ervilhas amarelas-lisas com ervilhas verdes-rugosas seguem uma distribuição de probabilidades dada por:

Fenótipo	Amarela-lisa	Amarela-rugosa	Verde-lisa	Verde-rugosa
Probabilidade	9/16	3/16	3/16	1/16

Em um determinado experimento uma amostra de 556 ervilhas resultantes de cruzamentos de ervilhas amarelas-lisas com ervilhas verdes-rugosas foi classificada da seguinte forma:

Fenótipo	Amarela-lisa	Amarela-rugosa	Verde-lisa	Verde-rugosa
Resultado	315	101	108	32

$H_0 \rightarrow v_o = v_e$: o experimento está de acordo com a distribuição de probabilidades mendelianas;

$H_1 \rightarrow v_o \neq v_e$: o experimento não segue a distribuição de probabilidades mendelianas;

Baseando-se nessas informações, existem evidências de que os resultados desse experimento estão de acordo com a distribuição de probabilidades proposta por pela Lei de Segregação de Mendel?

→ Se o modelo probabilístico for adequado, o valor esperado de ervilhas do tipo AMARELA LISA, dentre as 556 observadas, pode ser calculada da seguinte forma:

$$556 \times P(AL) \rightarrow 556 \times 9/16 = 312,75$$

Da mesma forma, para o tipo AMARELA RUGOSA:

$$556 \times P(AR) \rightarrow 556 \times 3/16 = 104,25$$

Para o tipo VERDE LISA temos:

$$556 \times P(VL) \rightarrow 556 \times 3/16 = 104,25$$

E, para o tipo VERDE RUGOSA:

$$556 \times P(VR) \rightarrow 556 \times 1/16 = 34,75$$

Feito isso, é hora de utilizar a fórmula do Qui-Quadrado

Fenótipo	Observado	Esperado
AL	315	312,75
AR	101	104,25
VL	108	104,25
VR	32	34,75
Total	556	556

$$\chi^2 = \sum_{i=1}^n \frac{(o_i - e_i)^2}{e_i}$$

$$X^2 = \frac{(315-312,75)^2}{312,75} + \frac{(101-104,25)^2}{104,25} + \frac{(108-104,25)^2}{104,25} + \frac{(32-34,75)^2}{34,75}$$

$$\chi^2 = 0.016 + 0.101 + 0.135 + 0.218 = 0.47$$

Baseando-se na tabela de χ^2 Graus de Liberdade:

Número de linhas -1 → 4 - 1 = 3

χ^2 tabela p/ α 0,05 e 3 GL = **7,82**

Conclusão:

O valor de χ^2 calculado não ultrapassou o tabelado,
portanto não há diferença significativa ao nível de 5 %,
mantendo-se a hipótese nula (H_0)!

Table 3.5 Critical values of the χ^2 distribution

<i>df</i>	<i>P</i>								
	0.995	0.975	0.9	0.5	0.1	0.05*	0.025	0.01	0.005
1	0.000	0.000	0.016	0.455	2.706	3.841	5.024	6.635	7.879
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838
4	0.207	0.484	1.064	3.357	7.779	9.488	11.143	13.277	14.860
5	0.412	0.831	1.610	4.351	9.236	11.070	12.832	15.086	16.750
6	0.676	1.237	2.204	5.348	10.645	12.592	14.449	16.812	18.548
7	0.989	1.690	2.833	6.346	12.017	14.067	16.013	18.475	20.278
8	1.344	2.180	3.490	7.344	13.362	15.507	17.535	20.090	21.955
9	1.735	2.700	4.168	8.343	14.684	16.919	19.023	21.666	23.589
10	2.156	3.247	4.865	9.342	15.987	18.307	20.483	23.209	25.188
11	2.603	3.816	5.578	10.341	17.275	19.675	21.920	24.725	26.757
12	3.074	4.404	6.304	11.340	18.549	21.026	23.337	26.217	28.300
13	3.565	5.009	7.042	12.340	19.812	22.362	24.736	27.688	29.819
14	4.075	5.629	7.790	13.339	21.064	23.685	26.119	29.141	31.319
15	4.601	6.262	8.547	14.339	22.307	24.996	27.488	30.578	32.801

P, probability; *df*, degrees of freedom.

*Most scientists assume that, when $P < 0.05$, a significant difference exists between the observed and the expected values in a chi-square test.

Teste de Independência



Objetivo:

- Verifica se existe independência entre duas variáveis ou mais variáveis medidas nas mesmas unidades experimentais

A representação dos valores observados é dada por uma tabela de dupla entrada ou tabela de contingência.

Procedimentos:

1º Definir as hipóteses:

$H_0 \rightarrow v_o = v_e$ (não há associação entre os grupos, casualidade);

$H_1 \rightarrow v_o \neq v_e$ (os grupos estariam associados);

2º Fixar α . Escolher a variável qui-quadrado com φ (graus de liberdade) = $(L-1) \times (C-1)$, onde L = número de linhas da tabela de contingência e C número de colunas.

3º Com auxílio da tabela calculam-se RA e RC.

Teste de independência

Ex.: Calcule o teste de independência entre os atributos etnia e o tipo sanguíneo (fator Rh), considerando uma amostra de 1160 indivíduos, classificados de acordo com as duas características simultaneamente.

→ 1º passo:

$H_0 \rightarrow$ não há associação entre tipagem Rh e etnia

$H_1 \rightarrow$ os grupos etnia e tipagem Rh estariam associados

	Fator Rh		
	Rh+	Rh-	Total
Subsaariano	300	40	340
Não Subsaariano	737	83	820
Total	1037	123	1160

	Fator Rh		
	Rh ⁺	Rh ⁻	Total
Subsaariano	300	40	340
Não Subsaariano	737	83	820
Total	1037	123	1160

→ 2º passo:

Para subsaariano fator Rh⁺

$$1160 \text{ pessoas} \longrightarrow 1037 \text{ Rh}^+$$

$$340 \text{ subsaarianos} \longrightarrow x$$

$$f_e = \frac{340 \times 1037}{1160} = 303,95$$

Para subsaariano fator Rh⁻

$$1160 \text{ pessoas} \longrightarrow 123 \text{ Rh}^-$$

$$340 \text{ subsaarianos} \longrightarrow x$$

$$f_e = \frac{340 \times 123}{1160} = 36,05$$

Para Não subsaariano fator Rh⁺

$$1160 \text{ pessoas} \longrightarrow 1037 \text{ Rh}^+$$

$$820 \text{ Não subsaarianos} \longrightarrow x$$

$$f_e = \frac{820 \times 1037}{1160} = 733,05$$

Para Não subsaariano fator Rh⁻

$$1160 \text{ pessoas} \longrightarrow 123 \text{ Rh}^-$$

$$820 \text{ Não subsaarianos} \longrightarrow x$$

$$f_e = \frac{820 \times 123}{1160} = 86,95$$



	Fator Rh			
	Rh ⁺ Observado	Rh ⁺ Esperado	Rh ⁻ Observado	Rh ⁻ Esperado
Subsaariano	300	303,95	40	36,05
Não Subsaariano	737	733,05	83	86,95

$$X^2 = \frac{(300-303,95)^2}{303,95} + \frac{(40-36,05)^2}{36,05} + \frac{(737-733,05)^2}{733,05} + \frac{(83-86,95)^2}{86,95}$$

$$\chi^2 = 0.051 + 0.433 + 0.021 + 0.179 = 0,684$$

Baseando-se na tabela de χ^2 Graus de Liberdade:

$$L(\text{linhas}) \times C(\text{colunas}) \rightarrow (L - 1) \times (C-1) \rightarrow (2- 1) \times (2-1) = 1$$

$$\chi^2 \text{ tabela p/ } \alpha 0,05 \text{ e } 1 \text{ GL} = 3,841$$

Conclusão:

O valor de χ^2 calculado não ultrapassou o tabelado,

portanto não há diferença significativa ao nível de 5 %, mantendo-se a hipótese nula (H_0)!

$$X^2 = \sum_{i=1}^n \frac{(o_i - e_i)^2}{e_i}$$

Table 3.5 Critical values of the χ^2 distribution

df	P								
	0.995	0.975	0.9	0.5	0.1	0.05*	0.025	0.01	0.005
1	0.000	0.000	0.016	0.455	2.706	3.841	5.024	6.635	7.879
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838
4	0.207	0.484	1.064	3.357	7.779	9.488	11.143	13.277	14.860
5	0.412	0.831	1.610	4.351	9.236	11.070	12.832	15.086	16.750
6	0.676	1.237	2.204	5.348	10.645	12.592	14.449	16.812	18.548
7	0.999	1.690	2.833	6.346	12.017	14.067	16.013	18.475	20.278
8	1.344	2.180	3.490	7.344	13.362	15.507	17.535	20.090	21.955
9	1.735	2.700	4.168	8.343	14.684	16.919	19.023	21.666	23.589
10	2.156	3.247	4.865	9.342	15.987	18.307	20.483	23.209	25.188
11	2.603	3.816	5.578	10.341	17.275	19.675	21.920	24.725	26.757
12	3.074	4.404	6.304	11.340	18.549	21.026	23.337	26.217	28.300
13	3.565	5.009	7.042	12.340	19.812	22.362	24.736	27.688	29.819
14	4.075	5.629	7.790	13.339	21.064	23.685	26.119	29.141	31.319
15	4.601	6.262	8.547	14.339	22.307	24.996	27.488	30.578	32.801

P, probability; df, degrees of freedom.

*Most scientists assume that, when P < 0.05, a significant difference exists between the observed and the expected values in a chi-square test.

Ex.: Calcule se há dependência entre a preferência entre diferentes tipos de ninho com a sazonalidade.

→ 1º passo:

H_0 → preferência pelo tipo de ninho independe do período do ano;

H_1 → preferência pelo tipo e ninho está associada ao período do ano;

$\alpha = 5\%$;

χ^2 tabela → $\phi = L(4-1) \times C(3-1) = 6$ graus de liberdade;

Tipos de Ninho	Sazonalidade			
	Dez-Março	Abril-Julho	Agosto-Nov	Total
Ovos colocados em cavidades de árvores	70	44	86	200
Construídos com gravetos	50	30	45	125
Ovos colocados no solo	10	6	34	50
Construídos com barro	20	20	85	125
Total	150	100	250	500

→ 2º passo:

$$O_{e1.1} = 200 \times 150/500 = 60$$

$$O_{e1.2} = 200 \times 100/500 = 40$$

$$O_{e1.3} = 200 \times 250/500 = 100$$

$$O_{e2.1} = 125 \times 150/500 = 37,5$$

$$O_{e2.2} = 125 \times 100/500 = 25$$

$$O_{e2.3} = 125 \times 250/500 = 62,5$$

$$O_{e3.1} = 50 \times 150/500 = 15$$

$$O_{e3.2} = 50 \times 100/500 = 10$$

$$O_{e3.3} = 50 \times 250/500 = 25$$

$$O_{e4.1} = 125 \times 150/500 = 37,5$$

$$O_{e4.2} = 125 \times 100/500 = 25$$

$$O_{e4.3} = 125 \times 250/500 = 62,5$$

Tipos de Ninho	Sazonalidade			
	Dez-Março	Abril-Julho	Agosto-Nov	Total
(1)Ovos colocados em cavidades de árvores	60	40	100	200
(2)Construídos com gravetos	37,5	25	62,5	125
(3)Ovos colocados no solo	15	10	25	50
(4)Construídos com barro	37,5	25	62,5	125
Total	150	100	250	500

→ 3º passo:

Calcule o valor do Qui-Quadrado

$$X^2 = \sum_{i=1}^n \frac{(o_i - e_i)^2}{e_i}$$

Valores Observados

Tipos de Ninho	Sazonalidade			
	Dez-Março	Abril-Julho	Agosto-Nov	Total
Ovos colocados em cavidades de árvores	70	44	86	200
Construídos com gravetos	50	30	45	125
Ovos colocados no solo	10	6	34	50
Construídos com barro	20	20	85	125
Total	150	100	250	500

Valores Esperados

Tipos de Ninho	Sazonalidade			
	Dez-Março	Abril-Julho	Agosto-Nov	Total
(1)Ovos colocados em cavidades de árvores	60	40	100	200
(2)Construídos com gravetos	37,5	25	62,5	125
(3)Ovos colocados no solo	15	10	25	50
(4)Construídos com barro	37,5	25	62,5	125
Total	150	100	250	500

Table 3.5 Critical values of the χ^2 distribution

df	P								
	0.995	0.975	0.9	0.5	0.1	0.05*	0.025	0.01	0.005
1	0.000	0.000	0.016	0.455	2.706	3.841	5.024	6.635	7.879
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838
4	0.207	0.484	1.064	3.357	7.779	9.488	11.143	13.277	14.860
5	0.412	0.831	1.610	4.351	9.236	11.070	12.832	15.086	16.750
6	0.676	1.237	2.204	5.348	10.645	12.592	14.449	16.812	18.548
7	0.989	1.690	2.833	6.346	12.017	14.067	16.013	18.475	20.278
8	1.344	2.180	3.490	7.344	13.362	15.507	17.535	20.090	21.955
9	1.735	2.700	4.168	8.343	14.684	16.919	19.023	21.666	23.589
10	2.156	3.247	4.865	9.342	15.987	18.307	20.483	23.209	25.188
11	2.603	3.816	5.578	10.341	17.275	19.675	21.920	24.725	26.757
12	3.074	4.404	6.304	11.340	18.549	21.026	23.337	26.217	28.300
13	3.565	5.009	7.042	12.340	19.812	22.362	24.736	27.688	29.819
14	4.075	5.629	7.790	13.339	21.064	23.685	26.119	29.141	31.319
15	4.601	6.262	8.547	14.339	22.307	24.996	27.488	30.578	32.801

P, probability; df, degrees of freedom.

*Most scientists assume that, when P < 0.05, a significant difference exists between the observed and the expected values in a chi-square test.

$$X^2_{\text{cal}} = 37.88$$

$$X^2_{\text{tab}} = 12.592$$

Logo rejeita-se H_0 , ou seja, o tipo de construção de ninho está associado a sazonalidade.

Table 3.5 Critical values of the χ^2 distribution

df	P								
	0.995	0.975	0.9	0.5	0.1	0.05*	0.025	0.01	0.005
1	0.000	0.000	0.016	0.455	2.706	3.841	5.024	6.635	7.879
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838
4	0.207	0.484	1.064	3.357	7.779	9.488	11.143	13.277	14.860
5	0.412	0.831	1.610	4.351	9.236	11.070	12.832	15.086	16.750
6	0.676	1.237	2.204	5.348	10.645	12.592	14.449	16.812	18.548
7	0.989	1.690	2.833	6.346	12.017	14.067	16.013	18.475	20.278
8	1.344	2.180	3.490	7.344	13.362	15.507	17.535	20.090	21.955
9	1.735	2.700	4.168	8.343	14.684	16.919	19.023	21.666	23.589
10	2.156	3.247	4.865	9.342	15.987	18.307	20.483	23.209	25.188
11	2.603	3.816	5.578	10.341	17.275	19.675	21.920	24.725	26.757
12	3.074	4.404	6.304	11.340	18.549	21.026	23.337	26.217	28.300
13	3.565	5.009	7.042	12.340	19.812	22.362	24.736	27.688	29.819
14	4.075	5.629	7.790	13.339	21.064	23.685	26.119	29.141	31.319
15	4.601	6.262	8.547	14.339	22.307	24.996	27.488	30.578	32.801

P, probability; df, degrees of freedom.

*Most scientists assume that, when P < 0.05, a significant difference exists between the observed and the expected values in a chi-square test.

Por hoje é isto!



@chihuahua_chloei