LCE0216
Introdução à Bioestatística Florestal

10. Teste de hipóteses

Profa. Dra. Clarice Garcia Borges Demétrio

Monitor: Silvio Gomes

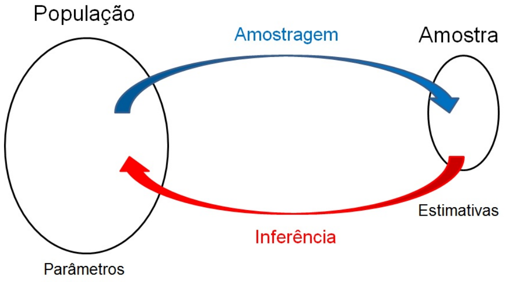
Escola Superior de Agricultura "Luiz de Queiroz"
Universidade de São Paulo

Piracicaba, 17 de maio de 2020

Inferência estatística

Dada a caracterização da população por meio de uma distribuição de probabilidades, há dois principais objetivos:

- **Estimar** os parâmetros dessa população;
- **Testar** hipóteses (afirmações) sobre esses parâmetros.



Testes de hipóteses

Hipóteses: Estabelecem as crenças (afirmações) a serem testadas. São definidas a partir do conhecimento do problema.

$$X \sim \text{Distribuição}(\theta)$$

$$\begin{cases} \mathrm{H}_0\colon \theta = \theta_0 & \text{(hipótese nula)} \\ \mathrm{H}_a\colon \theta \underset{>}{\overset{<}{\neq}} \theta_0 & \text{(hipótese alternativa)} \end{cases}$$

Tipos de erros:

- Rejeitar H_0, quando H_0 é verdadeira (erro tipo I);
- Não rejeitar H_0, quando H_0 é falsa (erro tipo II).

	Realidade	
Decisão	H_0 **verdadeira**	H_0 **falsa**
Não rejeitar H_0	Decisão correta	Erro tipo II (β)
Rejeitar H_0	Erro tipo I (α)	Decisão correta

Exemplo

Hoje à noite, você vai a uma festa. A previsão do tempo diz que a probabilidade de chover é de 0,8. Você leva guarda-chuva?

$$\begin{cases} \mathrm{H}_0: & \text{vai chover hoje à noite} \\ \mathrm{H}_a: & \text{Não vai chover hoje a noite} \end{cases}$$

Erro tipo I: Você rejeita H_0, acredita que não vai chover, vai sem guarda-chuva e se molha.

Erro tipo II: Você não rejeita H_0, acredita que vai chover, leva guarda-chuva e passa a noite toda carregando um guarda-chuva sem usá-lo.

	Realidade	
Decisão	H_0: **Chove**	H_a: **Não chove**
Leva guarda-chuva	Decisão correta	Erro tipo II (β)
Não leva guarda-chuva	Erro tipo I (α)	Decisão correta

Ver slide 55 da Aula 3 - Probilidades

Diagnósticos de doenças

Para que os exames laboratoriais auxiliem no diagnóstico de doenças é importante conhecer a capacidade dos exames em acertar o diagnóstico (ou seja, a sua acurácia). Duas medidas que auxiliam a mensurar a acurácia de um exame são:

- **sensibilidade** - capacidade que o teste diagnóstico/triagem apresenta de detectar os indivíduos verdadeiramente positivos, ou seja, de diagnosticar corretamente os doentes.
- **especificidade** - capacidade que o teste diagnóstico/triagem tem de detectar os verdadeiros negativos, isto é, de diagnosticar corretamente os indivíduos sadios.

Se se considerar a condição do indivíduo como sendo doente ou não doente e o resultado do exame como sendo positivo ou negativo, têm-se quatro combinações possíveis, como se segue.

	Status do paciente	
Resultado do exame	Doente (D_+)	Sadio (D_-)
Positivo (T_+)	Decisão correta	Erro tipo II (β)
Negativo (T_-)	Erro tipo I (α)	Decisão correta

Exemplo

Uma serraria vem utilizando um método tradicional de processar toras. O Engenheiro Florestal deseja implantar um novo método que seria mais eficiente, que reduziria as sobras de madeira (perdas).

$$\begin{cases} \mathrm{H}_0: & \text{O novo método é tão eficiente como o tradicional.} \\ \mathrm{H}_a: & \text{O novo método é mais eficiente que o tradicional.} \end{cases}$$

	Realidade	
Decisão	H_0: Método tradicional	H_a: Novo método
Método tradicional	Decisão correta	Erro tipo II (β)
Novo método	Erro tipo I (α)	Decisão correta

- Rejeitar H_0 significa, geralmente, mudar um procedimento tradicional em favor de um novo procedimento.
- Essa mudança sempre acarreta custos e evita-se mudar quando não se tem certeza que o novo procedimento será de fato melhor que o tradicional.
- Na maioria das pesquisas preocupa-se mais em controlar a probabilidade α de se cometer o Erro Tipo I.

Floresta Plantada ou Floresta Nativa?

Situação:

- No final do século passado, Manuel Gomes Acher foi incumbido por D. Pedro II de reflorestar as encostas dos morros na região na Tijuca (Rio de Janeiro).
- As repetidas secas que a cidade do Rio de Janeiro vinha sofrendo, com a consequente falta de água era atribuída ao desmatamento dos morros.
- Acher cumpriu sua missão com bastante eficiência e um visitante passeando hoje pelo Parque Nacional da Tijuca terá dificuldade em saber se a floresta que observa é nativa ou foi plantada pelo "major"Acher.

Floresta Plantada ou Floresta Nativa?

Um pesquisador deseja iniciar um projeto sobre o impacto urbano da cidade do Rio de Janeiro sobre a Floresta da Tijuca. Entretanto, o projeto deve ser instalado em área de floresta nativa. O pesquisador possui, como informação inicial, gráficos que apresentam a frequência de árvores observadas em diferentes classes de tamanhos, em áreas plantadas e áreas nativas. Na escolha da área apropriada para o projeto, o pesquisador trabalha com as hipóteses:

$$\begin{cases} \mathrm{H}_0: & \text{A área escolhida é de floresta nativa.} \\ \mathrm{H}_a: & \text{A área escolhida é de floresta plantada.} \end{cases}$$

A partir de mapas, o pesquisador selecionou uma certa área.

	Realidade	
Decisão	H_0: Floresta nativa.	H_a: Floresta plantada
Floresta nativa	Decisão correta	Erro tipo II (β)
Floresta plantada	Erro tipo I (α)	Decisão correta

Consequências dos erros que podem ocorrer:

Erro tipo I: Procura uma nova área, embora a área selecionada seja nativa.

Erro tipo II: Estabelece o projeto na área errada.

Regra da decisão: Regra que estabelece, com base em dados obtidos, quando a hipótese nula (H_0) é rejeitada.

- Selecionada a área, o pesquisador sorteia uma árvore e a mede (amostra de tamanho $n = 1$).
- Suponha que a árvore sorteada tenha diâmetro de 80cm. Você concluiria que a área é nativa ou plantada? Por quê? E se a árvore sorteada tivesse 20cm?
- Várias regras de decisão diferentes são possíveis e os níveis de α e β dependem de qual regra de decisão é utilizada.
- Verificando a chance de se sortearem árvores de diferentes tamanhos em ambas as áreas de floresta pode-se verificar isso.

Floresta Plantada ou Floresta Nativa?

Tamanho da Árvore Sorteada (diâmetro em cm)	Chance de a árvore ser da área	
	Nativa	Plantada
20	22 / 40	3 / 40
40	10 / 40	7 / 40
60	4 / 40	12 / 40
80	2 / 40	11 / 40
100	2 / 40	5 / 40
120	0 / 40	2 / 40
Total	40 / 40	40 / 40

Suponha que a regra de decisão fosse:
1) RD1: Rejeitar H_0 se o diâmetro $\geq$ 60cm
$\alpha = 8/40 = 0,20 \quad \beta = 10/40 = 0,25$
2) RD2: Rejeitar H_0 se o diâmetro $\geq$ 100cm
$\alpha = 2/40 = 0,05 \quad \beta = 33/40 = 0,83$
Pergunta: É possível manter, simultaneamente, α e β pequenos ?
Resposta: Não. Ao se diminuir a chance de um erro tipo I (α) aumenta-se a chance de um erro tipo II (β). A única maneira de melhorar ambos é aumentar o tamanho da amostra.

Nível de Significância: Nível de Significância α associado a uma regra de decisão, é a probabilidade de se cometer um erro tipo I.

- Para que as decisões tomadas por diferentes pessoas tenha uma certa coerência, cada ramo de atividade costuma estabelecer níveis máximos para a ocorrência do erro tipo I (α).
- Esse nível é chamado de nível de significância.
- Na área florestal, as decisões devem ser tomadas com um nível de significância (α) de no máximo 0,05 (5%).
- Um teste de hipótese é chamado significativo quando H_0 é rejeitada por uma regra de decisão que possui nível de significância α (5% na área florestal).

Nível descritivo (*p-value*): Probabilidade de se obterem estatísticas mais extremas para rejeição de H_0 do que aquela fornecida pela amostra.

Conclusão:

- *p-value* $<$ Nível de significância $\alpha \Rightarrow H_0$ é rejeitada
- *p-value* $>$ Nível de significância $\alpha \Rightarrow H_0$ não é rejeitada

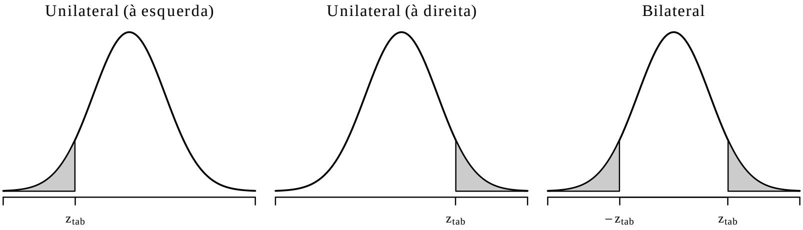
Conceitos básicos - Resumo

- **Hipóteses:** Estabelecem as crenças (afirmações) a serem testadas. São definidas a partir do conhecimento do problema e podem ser do tipo simples ou composta.
- **Nível de significância (α):** associado à regra de decisão. É a probabilidade de se cometer erro tipo I, $\alpha = P(\text{rejeitar } \mathrm{H}_0 \mid \mathrm{H}_0 \text{ é verdadeira})$.
- **Estatística do teste:** Uma estatística que depende do parâmetro de interesse, mas que tem distribuição conhecida que independe desse parâmetro.
- **Regra da decisão:** Regra que estabelece, com base nos dados obtidos e no nível de significância α, quando H_0 será rejeitada.
- **Nível descritivo (*p-value*):** Probabilidade de se obterem estatísticas mais extremas para rejeição de H_0 do que aquela fornecida pela amostra.

Tipos de hipóteses

- **Hipóteses simples**:
  H_0: $\theta = \theta_0$ versus H_a: $\theta = \theta_a$.

- **Hipóteses compostas**:
  - Unilateral (à esquerda): H_0: $\theta = \theta_0$ versus H_a: $\theta < \theta_0$;
  - Unilateral (à direita): H_0: $\theta = \theta_0$ versus H_a: $\theta > \theta_0$;
  - Bilateral: H_0: $\theta = \theta_0$ versus H_a: $\theta \neq \theta_0$.



Passos para construção de um teste de hipóteses

1. Formule as hipóteses, nula (H_0) e alternativa (H_a).
2. Identifique a estatística (estimador) adequada. Conhecer, ao menos assintoticamente, a distribuição amostral desse estimador.
3. Fixe o nível de significância (α), probabilidade do erro tipo I e construa a regra de decisão.
4. Use as observações da amostra para calcular o valor da estatística do teste.
5. Compare o valor calculado da estatística, com base em uma amostra, com as regiões da regra de decisão, para rejeitar ou não a hipótese nula.

Tipos de testes de hipóteses

Nesse curso, serão apresentados os procedimentos para testar hipóteses relacionadas a:

- Testes para proporção;
- Testes para média:
  - com variância conhecida;
  - com variância desconhecida;
- Testes para comparação de variâncias;
- Testes para comparação de duas médias:
  - dados pareados (dependentes);
  - dados não-pareados (independentes);
- Testes em tabelas de contigência.

1

Testes para proporção

Construção do teste

- População: $X = \{0(\text{fracasso}), 1(\text{sucesso})\}$; $X \sim \text{Bernoulli}(\pi)$;
- Hipóteses:

$$\begin{cases} \mathrm{H}_0\colon \pi = \pi_0 & \text{(hipótese nula)} \\ \mathrm{H}_a\colon \pi \underset{>}{\overset{<}{\neq}} \pi_0 & \text{(hipótese alternativa)} \end{cases}$$

- Com base em uma amostra de tamanho n, $n \to \infty$, sob H_0, sabe-se que

$$\hat{P} \sim N(\pi_0, \pi_0(1-\pi_0)/n);$$

$$Z = \frac{\hat{P} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} \sim N(0,1)$$

- Compara-se a estatística do teste

$$z_{calc} = \frac{\hat{p} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$$

com um quantil z_{tab} da distribuição normal padrão, que represente o indicado pela hipótese alternativa.

Exemplo 1: Um produtor precisa decidir pela compra ou não de sementes fornecidas por um distribuidor, que afirma que a proporção de germinação de sementes é $\pi = 0,94$. Para tanto ele observou a proporção de germinação de uma amostra aleatória simples de 100 sementes e encontrou o valor $\hat{p} = 0,93$. Com base nesse resultado, para as questões que se seguem, assuma um nível de significância de 5%, estabeleça regras de decisão e calcule os valores de p ($p - value$) com base nas proporções amostrais.

(a) Teste H_0 versus se o produtor poderia discordar do distribuidor (isto é, a desigualdade $\neq$).

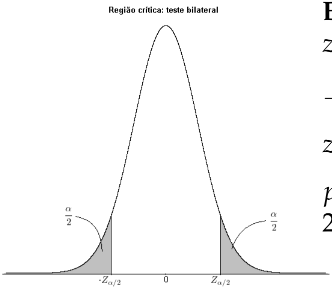
(b) Teste H_0 versus se o distribuidor está sendo muito otimista ($<$).

Exemplo 1: Solução

(a) $\pi_0 = 0,94 \quad \hat{p} = 0,93 \quad n = 100 \quad \alpha = 0,05$

Hipóteses: $\begin{cases} \text{H}_0\text{: } \pi = 0,94 & \text{(hipótese nula)} \\ \text{H}_a\text{: } \pi \neq 0,94 & \text{(hipótese alternativa)} \end{cases}$

Regra de decisão



Estatística

$z_{calc} = \frac{\hat{p}-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0{,}93-0{,}94}{\sqrt{\frac{0{,}94\times(1-0{,}94)}{100}}} = -0,42$

$z_{0,025} = -1,96$

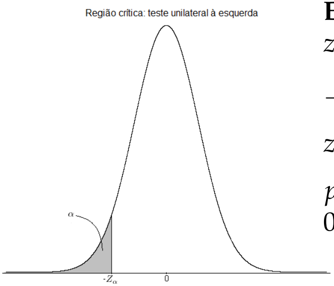
$p - value = 2P(Z < -0,42) = 2 \times 0,3372 = 0,6744 > 0,05$

Conclusão: Como $-1,96 < -0,42 < 1,96$ (ou $p - value = 0,6744 > 0,05$), não se rejeita H_0 ao nível de 5% de significância. Portanto, existem evidências de que o poder germinativo das sementes não difere de 94%.

(b) $\pi_0 = 0,94 \quad \hat{p} = 0,93 \quad n = 100 \quad \alpha = 0,05$

Hipóteses: $\begin{cases} H_0\colon \pi = 0,94 & \text{(hipótese nula)} \\ H_a\colon \pi < 0,94 & \text{(hipótese alternativa)} \end{cases}$

Regra de decisão



Estatística

$z_{calc} = \frac{\hat{p} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0,93 - 0,94}{\sqrt{\frac{0,94 \times (1-0,94)}{100}}} = -0,42$

$z_{0,05} = -1,64$

$p - value = P(Z < -0,42) = 0,3372 > 0,05$

Conclusão: Como $-0,42 > -1,96$ (ou $p - value = 0,3372 > 0,05$), não se rejeita H_0 ao nível de 5% de significância. Portanto, existem evidências de que o poder germinativo das sementes não é menor do que 94%.

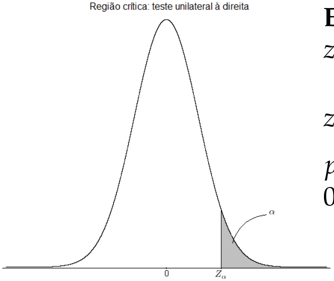
Exemplo 2: Após vários anos de acompanhamento de parcelas permanentes, uma engenheira florestal concluiu que, das árvores que morrem num fragmento florestal, 75% são devido ao abafamento da copa por cipós. No ano seguinte, ocorreu um período prolongado de intensa seca e 30 árvores morreram durante o ano, das quais 24 mortes podem ser atribuídas ao efeito dos cipós. A engenheira florestal afirma que a proporção de árvores que morreram devido ao abafamento por cipós foi maior nesse ano de seca intensa do que nos anos anteriores. Verifique a afirmação, usando o nível de 5% de significância.

Exemplo 2: Solução

$\pi_0 = 0,75 \quad n = 30 \quad \hat{p} = \frac{24}{30} = 0,80 \quad \alpha = 0,05$

Hipóteses: $\begin{cases} H_0\colon \pi = 0,75 & \text{(hipótese nula)} \\ H_a\colon \pi > 0,75 & \text{(hipótese alternativa)} \end{cases}$

Regra de decisão



Estatística

$$z_{calc} = \frac{\hat{p} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0{,}80 - 0{,}75}{\sqrt{\frac{0{,}75 \times (1-0{,}75)}{30}}} = 0,63$$

$$z_{0{,}05} = 1,64$$

$p - value = P(Z > 0,63) = 0,2643 > 0,05$

Conclusão: Como $0,63 < 1,64$ (ou $p - value = 0,2643 > 0,05$), não se rejeita H_0 ao nível de 5% de significância. Portanto, existem evidências de que a proporção de árvores que morreram devido ao abafamento por cipós não é maior do que 75%.

2

Teste para média

Construção do teste (variância conhecida)

- População: $X \sim N(\mu, \sigma^2)$, σ^2 conhecida;
- Hipóteses:

$$\begin{cases} \mathrm{H}_0\colon \mu = \mu_0 & \text{(hipótese nula)} \\ \mathrm{H}_a\colon \mu \underset{>}{\overset{<}{\neq}} \mu_0 & \text{(hipótese alternativa)} \end{cases}$$

- Com base em uma amostra de tamanho n, sob H_0, sabe-se que

$$\bar{X} \sim N(\mu_0, \sigma^2/n);$$

$$Z = \frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}} \sim N(0,1)$$

- Compara-se a estatística do teste

$$z_{calc} = \frac{\bar{x} - \mu_0}{\sqrt{\sigma^2/n}} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

com um quantil z_{tab} da distribuição normal padrão, que represente o indicado pela hipótese alternativa.

Exemplo 3: Uma balança para encher pacotes de sementes automaticamente está programada para produzir pacotes com peso médio de 20 kg e desvio padrão de 0,20 kg.

Periodicamente, é feita uma inspeção para verificar se o peso médio está sob controle. Para esse fim, foi selecionada uma amostra de oito pacotes de sementes, cujos resultados foram:

20,3	19,8	20,3	19,7	19,8	19,7	19,8	19,8

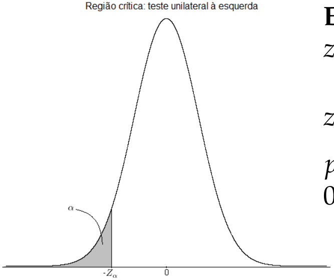
Teste a hipótese nula versus a hipótese que a balança se desregulou e está produzindo um peso médio inferior a 20 kg, assumindo que a variância permanece inalterada. Use o nível de significância de 5%.

Exemplo 3: Solução

$\mu_0 = 20\text{kg} \quad \sigma = 0,20\text{kg} \quad n = 8 \quad \hat{\mu} = \frac{159,2}{8} = 19,9 \quad \alpha = 0,05$

Hipóteses: $\begin{cases} H_0\text{: } \mu = 20 & \text{(hipótese nula)} \\ H_a\text{: } \mu < 20 & \text{(hipótese alternativa)} \end{cases}$

Regra de decisão



Estatística

$$z_{calc} = \frac{\hat{\mu} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} = \frac{19,9 - 20}{\frac{0,20}{\sqrt{8}}} = -1,41$$

$$z_{0,05} = -1,64$$

$p - value = P(Z < -1,41) = 0,0793 > 0,05$

Conclusão: Como $-1,41 > -1,64$ (ou $p - value = 0,0793 > 0,05$), não se rejeita H_0 ao nível de 5% de significância. Portanto, existem evidências de que o peso médio de pacotes de sementes está sob controle e não é menor do que 20kg.

Construção do teste (variância desconhecida)

- População: $X \sim N(\mu, \sigma^2)$;
- Hipóteses:

$$\begin{cases} \mathrm{H}_0\colon \mu = \mu_0 & \text{(hipótese nula)} \\ \mathrm{H}_a\colon \mu \underset{>}{\overset{<}{\neq}} \mu_0 & \text{(hipótese alternativa)} \end{cases}$$

- Com base em uma amostra de tamanho n, sob H_0, sabe-se que

$$\bar{X} \sim N(\mu_0, \sigma^2/n); \text{e como } \sigma^2 \text{ é desconhecida,}$$

$$T = \frac{\bar{X} - \mu_0}{\sqrt{S^2/n}} \sim t_{(n-1)}, \quad S^2 = \frac{\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2/n}{n-1}$$

- Compara-se a estatística do teste

$$t_{calc} = \frac{\bar{x} - \mu_0}{\sqrt{s^2/n}}$$

  com um quantil t_{tab} da distribuição t de Student com $n-1$ graus de liberdade, que represente o indicado pela hipótese alternativa.

Exemplo 3 (revisitado): Uma balança para encher pacotes de sementes automaticamente está programada para produzir pacotes com peso médio de 20 kg e desvio padrão de 0,20 kg. Periodicamente, é feita uma inspeção para verificar se o peso médio está sob controle. Para este fim, foi selecionada uma amostra de oito pacotes de sementes, cujos resultados foram:

20,3	19,8	20,3	19,7	19,8	19,7	19,8	19,8

Teste a hipótese nula versus a hipótese que a balança se desregulou e está produzindo um peso médio inferior a 20 kg. Use o nível de significância de 5%.

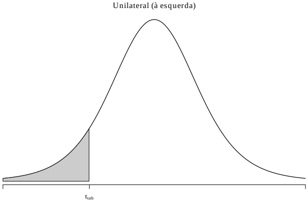
Exemplo 3 (revisitado): Solução

$\mu_0 = 20\text{kg} \quad n = 8 \quad \alpha = 0,05 \quad \hat{\mu} = \frac{159,2}{8} = 19,9kg$

$\hat{\sigma}^2 = s^2 = \frac{1}{7}\left[3168,52 - \frac{159,2^2}{8}\right] = 0,0628\text{kg}^2$

Hipóteses: $\begin{cases} \text{H}_0\text{: } \mu = 20 & \text{(hipótese nula)} \\ \text{H}_a\text{: } \mu < 20 & \text{(hipótese alternativa)} \end{cases}$

Regra de decisão



Estatística

$t_{calc} = \frac{\hat{\mu}-\mu_0}{\sqrt{\frac{s^2}{n}}} = \frac{19,9-20}{\sqrt{\frac{0,0628}{8}}} = -1,13$

$t_{7;0,05} = -1,89$

$p - value = P(t < -1,13) = 0,1482 > 0,05$

Conclusão: Como $-1,13 > -1,89$ (ou $p - value = 0,1482 > 0,05$), não se rejeita H_0 ao nível de 5% de significância. Portanto, existem evidências de que o peso médio de pacotes de sementes está sob controle e não é menor do que 20kg.

Exemplo 4: Um cientista desejava saber se o pH de um solo era ácido. Ele tomou uma amostra de 10 unidades e obteve os valores de pH:

5,8	6,0	7,0	6,2	6,2
7,1	6,4	5,5	5,8	5,9

Tire uma conclusão supondo nível de significância de 5%.

Exemplo 4: Um cientista desejava saber se o pH de um solo era ácido. Ele tomou uma amostra de 10 unidades e obteve os valores de pH:

5,8	6,0	7,0	6,2	6,2
7,1	6,4	5,5	5,8	5,9

Tire uma conclusão supondo nível de significância de 5%.

Observação: A escala de pH vai de 1 a 14, sendo 7 a neutralidade, abaixo de 7 a acidez e acima de 7 a alcalinidade.

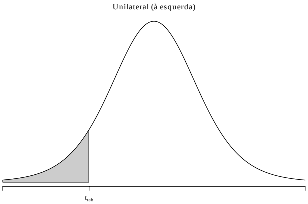
Exemplo 4: Solução

$\mu_0 = 7 \quad n = 10 \quad \alpha = 0,05 \quad \hat{\mu} = \frac{61,9}{10} = 6,19$

$\hat{\sigma}^2 = s^2 = \frac{1}{9}\left[385,59 - \frac{61,9^2}{8}\right] = 0,2699$

Hipóteses: $\begin{cases} \mathrm{H}_0\colon \mu = 7 & \text{(hipótese nula)} \\ \mathrm{H}_a\colon \mu < 7 & \text{(hipótese alternativa)} \end{cases}$

Regra de decisão



Estatística

$t_{calc} = \frac{\hat{\mu}-\mu_0}{\sqrt{\frac{s^2}{n}}} = \frac{6,19-7}{\sqrt{\frac{0,2699}{10}}} = -4,93$

$t_{9;0,05} = -1,83$

$p - value = P(t < -1,13) = 0,0004 < 0,05$

Conclusão: Como $-4,93 < -1,83$ (ou $p - value = 0,0004 > 0,05$), rejeita-se H_0 ao nível de 5% de significância. Portanto, existem evidências de que o solo é ácido.

Inferência para duas populações

Conceitos iniciais

Interesse

Dadas duas populações, caracterizadas em uma mesma família de distribuições, o objetivo nesse caso é **testar afirmações comparativas sobre os parâmetros das duas populações**.

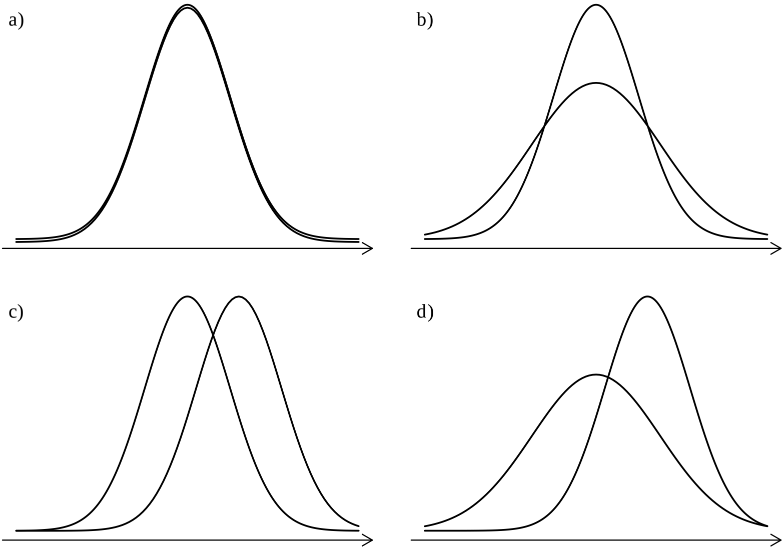
- Descrição das populações,

$$X_1 \sim \text{Distribuição}(\theta_1)$$
$$X_2 \sim \text{Distribuição}(\theta_2)$$

- Hipóteses de interesse,

$$\begin{cases} \mathrm{H}_0\colon \theta_1 = \theta_2 & \text{(hipótese nula)} \\ \mathrm{H}_a\colon \theta_1 \underset{>}{\overset{<}{\neq}} \theta_2 & \text{(hipótese alternativa)} \end{cases}$$

Conceitos iniciais



- Com relação a alguma variável contínua, podemos considerar que as populações são iguais?

3

Teste para comparação de variâncias

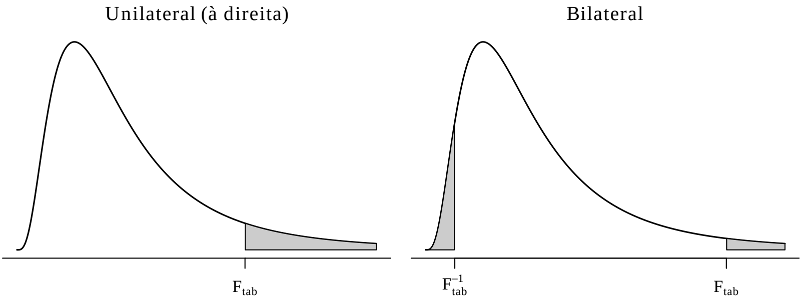
Construção do teste

- População:

$$X_1 \sim N(\mu_1, \sigma_1^2)$$
$$X_2 \sim N(\mu_2, \sigma_2^2);$$

- Hipóteses:

$$\begin{cases} \mathrm{H}_0\colon \sigma_1^2 = \sigma_2^2 \Rightarrow \dfrac{\sigma_1^2}{\sigma_2^2} = 1 & \text{(hipótese nula)} \\ \mathrm{H}_a\colon \sigma_1^2 \overset{<}{\underset{>}{\neq}} \sigma_2^2 \Rightarrow \dfrac{\sigma_1^2}{\sigma_2^2} \overset{<}{\underset{>}{\neq}} 1 & \text{(hipótese alternativa)} \end{cases}$$



Construção do teste

- Considerando amostras aleatórias de X_1 e X_2 independentes, de tamanho n_1 e n_2, respectivamente. Sob H_0, sabe-se que

$$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1); \quad \text{em que}$$

$$S_j^2 = \frac{\sum_{i=1}^{n} X_{ji}^2 - (\sum_{i=1}^{n} X_{ji})^2/n}{n-1}, \quad j = 1, 2.$$

  Na prática, considera-se o cálculo da estatística F de tal forma que $F = S_1^2/S_2^2 > 1$, ou seja, utiliza-se como numerador a maior variância amostral observada.

- Compara-se a estatística do teste

$$F_{calc} = s_1^2/s_2^2,$$

  com a região critíca (RC) construída de tal forma que $P(F \in \text{RC}) = \alpha$. Para hipóteses bilaterais, a RC é dada por $(0, F_{tab}^{-1}) \cup (F_{tab}, \infty)$, em que $P(F > F_{tab}) = \alpha/2$.

Exemplo 5: Um pesquisador tem interesse em comparar se a proliferação de fungos em plantas tem relação com determinados tipos de solo. Para isso, o pesquisador mensurou as áreas atacadas por fungos em 13 árvores plantadas em solo tipo A e nove árvores plantadas em solo tipo B. Os resultados obtidos foram

Tipo de solo A	**Tipo de solo B**
$\sum_{i=1}^{13} x_{1i} = 104\text{cm}^2$	$\sum_{i=1}^{9} x_{2i} = 63\text{cm}^2$
$\sum_{i=1}^{13} x_{1i}^2 = 997\text{cm}^4$	$\sum_{i=1}^{9} x_{2i}^2 = 497\text{cm}^4$

Pode-se considerar que as áreas atacadas por fungos em árvores plantadas em diferentes tipos de solo possuem a mesma variabilidade?

Exemplo 4: Solução

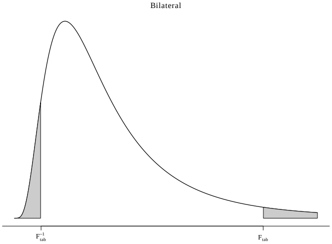
Teste de variâncias $\alpha = 0,05$

$n_A = 13 \quad \hat{\mu}_A = \frac{104}{13} = 8 \quad \hat{\sigma}^2_A = s^2_A = \frac{1}{12}\left[997 - \frac{104^2}{13}\right] = 13,75$

$n_B = 9 \quad \hat{\mu}_B = \frac{63}{9} = 7 \quad \hat{\sigma}^2_B = s^2_B = \frac{1}{8}\left[497 - \frac{63^2}{9}\right] = 7,00$

Hipóteses: $\begin{cases} \text{H}_0\text{: } \sigma^2_A = \sigma^2_B \Rightarrow \frac{\sigma^2_A}{\sigma^2_B} = 1 & \text{(hipótese nula)} \\ \text{H}_a\text{: } \sigma^2_A \neq \sigma^2_B \Rightarrow \frac{\sigma^2_A}{\sigma^2_B} \neq 1 & \text{(hipótese alternativa)} \end{cases}$

Regra de decisão



Estatística

$F_{calc} = \frac{s^2_1}{s^2_2} = \frac{13{,}75}{7{,}00} = 1,96$

$F_{12;8;0{,}975} = 4,20$ e
$F_{8;12;0{,}025} = \frac{1}{F_{12;8;0{,}975}} = \frac{1}{4{,}20} = 0,24$

$p - value = 2P(F > 1,96) =$
$2 \times 0,3827 = 0,7654 > 0,05$

Conclusão: Como $0,24 < 1,96 < 4,20$ ($p - value = 0,7654 > 0,05$), não se rejeita H_0 ao nível de 5% de significância. Logo, existem evidências que as variâncias não diferem estatisticamente.

4

Teste para comparação de médias

Dados pareados *vs* dados não pareados

- **Dados pareados (dependentes)**: São casos em que é razoável supor que há correlação entre as observações das diferentes populações. Nesse curso, apresentam-se testes apenas para quando essa dependência se dá em pares. Exemplos:
  - Experimentos do tipo antes e depois;
  - Diferentes mensurações em uma mesma unidade amostral.

- **Dados não pareados (independentes)**: São casos em que é razoável supor a independência entre as observações das diferentes populações.

Construção do teste (dados pareados)

- População: Nesse caso, como a amostra se dá por pares (X_1, X_2), consideram-se as diferenças $D = X_2 - X_1$,

$$D = X_2 - X_1 \sim N(\mu_D, \sigma_D^2),$$

retornando-se a um problema de uma única população, conforme visto anteriormente.

- Hipóteses:

$$\begin{cases} \mathrm{H}_0\colon \mu_D = \mu_2 - \mu_1 = 0 & \text{(hipótese nula)} \\ \mathrm{H}_a\colon \mu_D = \mu_2 - \mu_1 \overset{<}{\underset{>}{\neq}} 0 & \text{(hipótese alternativa)} \end{cases}$$

- Com base em uma amostra pareada de tamanho n, sob H_0, sabe-se que

$$\bar{D} = \bar{X}_2 - \bar{X}_1 \sim N(0, \sigma_D^2/n) \text{ e como } \sigma_D^2 \text{ é desconhecido,}$$

$$T = \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{S_D^2}{n}}} \sim t_{(n-1)}, \quad S_D^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} D_i^2 - \frac{(\sum_{i=1}^{n} D_i)^2}{n}\right]$$

Exemplo

Exemplo 6: Um pesquisador deseja verificar se a altura (m) de uma árvore medida em pé, usando-se o método trigonométrico (aproximado), não difere da altura da árvore medida no chão. Com esse objetivo, mediu as alturas das árvores pelo método trigonométrico, derrubou-as e mediu novamente suas alturas.

n^o	Árvore em pé (A)	Árvore no chão (B)	Diferença (D=B-A)
1	20,4	21,7	1,3
2	25,4	26,3	0,9
3	25,6	26,8	1,2
4	26,6	26,2	-0,4
5	28,6	27,3	-1,3
6	28,7	29,5	0,8
7	29,0	32,0	3,0
8	29,8	30,9	1,1
9	30,5	32,3	1,8
10	30,9	32,3	1,4
11	31,1	31,7	0,6
12	25,6	28,1	2,5

Pode-se considerar que a altura obtida pelo método aproximado equivale à altura real (obtida derrubando-se a árvore)?

Exemplo

Exemplo 6: Um pesquisador deseja verificar se a altura (m) de uma árvore medida em pé, usando-se o método trigonométrico (aproximado), não difere da altura da árvore medida no chão. Com esse objetivo, mediu as alturas das árvores pelo método trigonométrico, derrubou-as e mediu novamente suas alturas.

n^o	Árvore em pé (A)	Árvore no chão (B)	Diferença (D=B-A)
1	20,4	21,7	1,3
2	25,4	26,3	0,9
3	25,6	26,8	1,2
4	26,6	26,2	-0,4
5	28,6	27,3	-1,3
6	28,7	29,5	0,8
7	29,0	32,0	3,0
8	29,8	30,9	1,1
9	30,5	32,3	1,8
10	30,9	32,3	1,4
11	31,1	31,7	0,6
12	25,6	28,1	2,5

Pode-se considerar que a altura obtida pelo método aproximado equivale à altura real (obtida derrubando-se a árvore)?
$\sum_{i=1}^{n} D_i = 12,9$ e $\sum_{i=1}^{n} D_i^2 = 28,45$.

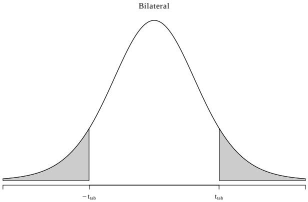
Exemplo 6: Solução

$n = 12 \quad \alpha = 0,05 \quad \hat{\mu}_D = \frac{\sum_i^n D_i}{n} = \frac{12,9}{12} = 1,075m$

$\hat{\sigma}^2_D = s^2_D = \frac{1}{n-1}\left[\sum_{i=1}^n D_i^2 - \frac{(\sum_{i=1}^n D_i)^2}{n}\right] = \frac{1}{11}\left[28,45 - \frac{12,9^2}{12}\right] = 1,325682m^2$

Hipóteses: $\begin{cases} \text{H}_0\text{: } \mu_D = \mu_A - \mu_B = 0 & \text{(hipótese nula)} \\ \text{H}_a\text{: } \mu_D = \mu_A - \mu_B \neq 0 & \text{(hipótese alternativa)} \end{cases}$

Regra de decisão



Estatística

$t_{calc} = \frac{\hat{\mu}_D - \mu_D}{\sqrt{\frac{s^2_D}{n}}} = \frac{1,075-0}{\sqrt{\frac{1,325682}{12}}} = 3,23$

$t_{11;0,025} = -2,20 \quad t_{11;0,975} = 2,20$

$p - value = 2P(t < -2,20) = 2P(t < 2,20) = 2 \times 0,0040 = 0,0080 < 0,05$

Conclusão: Como $3,23 > 2,20$ (ou $p - value = 0,0080 < 0,05$), rejeita-se H_0 ao nível de 5% de significância. Portanto, existem evidências de que a diferença entre médias de alturas medidas pelo dois métodos difere de zero.

- Populações:

$$X_1 \sim N(\mu_1, \sigma^2) \quad \text{e} \quad X_2 \sim N(\mu_2, \sigma^2);$$

- Hipóteses:

$$\begin{cases} \mathrm{H}_0\colon \mu_1 = \mu_2 \;\Rightarrow\; \mu_d = \mu_2 - \mu_1 = 0 & \text{(hipótese nula)} \\ \mathrm{H}_a\colon \mu_1 \overset{<}{\underset{>}{\neq}} \mu_2 \;\Rightarrow\; \mu_d = \mu_2 - \mu_1 \overset{<}{\underset{>}{\neq}} 0 & \text{(hipótese alternativa)} \end{cases}$$

- Considere amostras aleatórias de X_1 e X_2 independentes, de tamanho n_1 e n_2, respectivamente, e os estimadores

$$\bar{X}_j = \frac{\sum_{i=1}^{n} X_{ji}}{n_j}; \quad S_j^2 = \frac{\sum_{i=1}^{n} X_{ji}^2 - (\sum_{i=1}^{n} X_{ji})^2/n}{n-1}, \quad j = 1, 2.$$

Sob H_0 e como σ^2 é desconhecido, sabe-se que

$$\bar{X}_d = \bar{X}_2 - \bar{X}_1 \sim N\left(0, \sigma^2\left(\frac{1}{n_2} + \frac{1}{n_1}\right)\right);$$

$$T = \frac{\bar{X}_d - \mu_d}{\hat{\sigma}_{\bar{X}_d}} = \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{S^2}{\frac{1}{n_2} + \frac{1}{n_1}}}} \sim t_{(n_2+n_1-2)}$$

$$\text{em que} \quad S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_2 + n_1 - 2}$$

- Portanto, considera-se como estatística do teste

$$t_{calc} = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{s^2}{\frac{1}{n_2}+\frac{1}{n_1}}}}, \quad \text{em que} \quad s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_2+n_1-2}.$$

- Para um nível de significância α, a região crítica (RC) é construída de tal forma que $P(t_{(n_1+n_2-2)} \in \text{RC}) = \alpha$.
  - Para hipóteses uniterais:
    $P(t_{(n_1+n_2-2)} > t_{tab}) = \alpha$;
  - Para hipóteses bilaterais:
    $P(-t_{tab} < t_{(n_1+n_2-2)} < t_{tab}) = 1 - \alpha$.

Exemplo 5 (revisitado): Um pesquisador tem interesse em comparar se a proliferação de fungos em plantas tem relação com determinados tipos de solo. Para isso, o pesquisador mensurou as áreas atacadas por fungos em 13 árvores plantadas em solo tipo A e nove árvores plantadas em solo tipo B. Os resultados obtidos foram

Tipo de solo A	**Tipo de solo B**
$\sum_{i=1}^{13} x_{1i} = 104\text{cm}^2$	$\sum_{i=1}^{9} x_{2i} = 63\text{cm}^2$
$\sum_{i=1}^{13} x_{1i}^2 = 997\text{cm}^4$	$\sum_{i=1}^{9} x_{2i}^2 = 497\text{cm}^4$

Pode-se considerar que a área média com fungo obtida em árvores plantadas em solo tipo A é maior do que a obtida se plantadas do tipo B?

Exemplo 5 (revisitado): Solução $\alpha = 0,05$

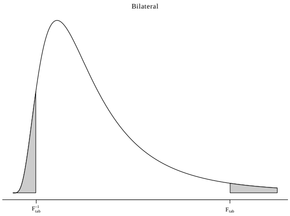
Teste de variâncias

$n_A = 13 \quad \hat{\mu}_A = \frac{104}{13} = 8cm^2 \quad \hat{\sigma}^2_A = s^2_A = \frac{1}{12}\left[997 - \frac{104^2}{13}\right] = 13,75cm^4$

$n_B = 9 \quad \hat{\mu}_B = \frac{63}{9} = 7cm^2 \quad \hat{\sigma}^2_B = s^2_B = \frac{1}{8}\left[497 - \frac{63^2}{9}\right] = 7,00cm^4$

Hipóteses: $\begin{cases} \text{H}_0\text{: } \sigma^2_A = \sigma^2_B \Rightarrow \frac{\sigma^2_A}{\sigma^2_B} = 1 & \text{(hipótese nula)} \\ \text{H}_a\text{: } \sigma^2_A \neq \sigma^2_B \Rightarrow \frac{\sigma^2_A}{\sigma^2_B} \neq 1 & \text{(hipótese alternativa)} \end{cases}$

Regra de decisão



Estatística

$F_{calc} = \frac{s^2_1}{s^2_2} = \frac{13{,}75}{7{,}00} = 1,96$

$F_{12;8;0,975} = 4,20$ e

$F_{8;12;0,025} = \frac{1}{F_{12;8;0,975}} = \frac{1}{4{,}20} = 0,24$

$p - value = 2P(F > 1,96) =$
$2 \times 0,3827 = 0,7654 > 0,05$

Conclusão: Como $0,24 < 1,96 < 4,20$ ($p - value = 0,7654 > 0,05$), não se rejeita H_0 ao nível de 5% de significância. Logo, existem evidências que as variâncias não diferem estatisticamente.

Exemplo 5 (revisitado): Solução (cont.)

Teste de médias $\alpha = 0,05$

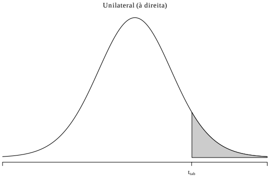
$n_A = 13 \quad \hat{\mu}_A = \frac{104}{13} = 8cm^2 \quad \hat{\sigma}^2_A = s^2_A = \frac{1}{12}\left[997 - \frac{104^2}{13}\right] = 13,75cm^4$

$n_B = 9 \quad \hat{\mu}_B = \frac{63}{9} = 7cm^2 \quad \hat{\sigma}^2_B = s^2_B = \frac{1}{8}\left[497 - \frac{63^2}{9}\right] = 7,00cm^4$

$s^2 = \frac{(n_A-1)s^2_A+(n_B-1)s^2_B}{n_A+n_B-2} = \frac{(13-1)13{,}75+(9-1)7{,}00}{13+9-2} = 11,05cm^4$

Hipóteses: $\begin{cases} \mathrm{H}_0\text{: } \mu_A = \mu_B \Rightarrow \mu_A - \mu_B = 0 & \text{(hipótese nula)} \\ \mathrm{H}_a\text{: } \mu_A > \mu_B \Rightarrow \mu_A - \mu_B > 0 & \text{(hipótese alternativa)} \end{cases}$

Regra de decisão



Estatística

$t_{calc} = \frac{\hat{\mu}_A - \hat{\mu}_B}{\sqrt{\frac{s^2}{\frac{1}{n_A}+\frac{1}{n_B}}}} = \frac{13{,}75-7{,}00}{\sqrt{\frac{11{,}05}{\frac{1}{13}+\frac{1}{9}}}} = 0,13$

$t_{20;0{,}95} = 1,72$

$p - value = P(t > 0,13) = 0,4488 > 0,05$

Conclusão: Como $0,13 < 1,72$ (ou $p - value = 0,4488 > 0,05$), rejeita-se H_0 ao nível de 5% de significância. Portanto, existem evidências de que a proliferação de fungos em árvores não tem relação com os tipos de solo em que são plantadas.

Exemplo 7: Um pesquisador quer saber se uma espécie B de árvores produz vigas de madeira com maior resistência que uma espécie A. Tomando-se 15 vigas de madeira de uma espécie A e 20 de uma espécie B, obtêm-se os resultados

	Espécie A	**Espécie B**
Média	$70,5$Mpa	$84,3$Mpa
Variância	$81,6$Mpa^2	$210,8$Mpa^2

Com base nessas observações, verifique se o pesquisador tem razão. Use um nível de significância de 10%.

Exemplo 7: Solução

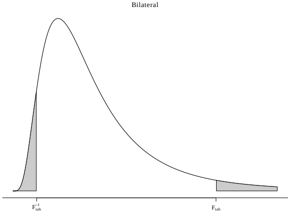
Teste de variâncias $\alpha = 0,05$

$n_A = 15 \quad \hat{\mu}_A = 70,5Mpa \quad \hat{\sigma}_A^2 = s_A^2 = 81,6Mpa^2$

$n_B = 20 \quad \hat{\mu}_B = 84,3Mpa \quad \hat{\sigma}_B^2 = s_B^2 = 210,8Mpa^2$

$$\textbf{Hipóteses}: \begin{cases} \text{H}_0\text{: } \sigma_A^2 = \sigma_B^2 \Rightarrow \frac{\sigma_A^2}{\sigma_B^2} = 1 \quad \text{(hipótese nula)} \\ \text{H}_a\text{: } \sigma_A^2 \neq \sigma_B^2 \Rightarrow \frac{\sigma_A^2}{\sigma_B^2} \neq 1 \quad \text{(hipótese alternativa)} \end{cases}$$

Regra de decisão



Estatística

$F_{calc} = \frac{s_B^2}{s_A^2} = \frac{210,6}{81,6} = 2,58$

$F_{19;14;0,975} = 2,86$ e

$F_{19;14;0,025} = \frac{1}{F_{14;19;0,975}} = \frac{1}{2,86} = 0,35$

$p - value = 2P(F > 2,58) =$
$2 \times 0,0379 = 0,0758 > 0,05$

Conclusão: Como $0,35 < 2,58 < 2,86$ ($p - value = 0,7654 > 0,05$), não se rejeita H_0 ao nível de 5% de significância. Logo, existem evidências que as variâncias não diferem estatisticamente.

Exemplo 7: Solução (cont.)

Teste de médias $\alpha = 0,10$

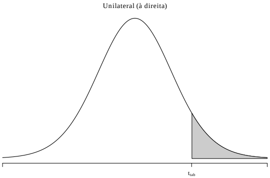
$n_A = 15 \quad \hat{\mu}_A = 70,5Mpa \quad \hat{\sigma}_A^2 = s_A^2 = 81,6Mpa^2$

$n_B = 20 \quad \hat{\mu}_B = 84,3Mpa \quad \hat{\sigma}_B^2 = s_B^2 = 210,8Mpa^2$

$s^2 = \frac{(n_A-1)s_A^2+(n_B-1)s_B^2}{n_A+n_B-2} = \frac{(15-1)81,6+(20-1)210,8}{15+9-2} = 155,9879Mpa^2$

Hipóteses: $\begin{cases} \text{H}_0\text{: } \mu_B = \mu_A \Rightarrow \mu_B - \mu_A = 0 & \text{(hipótese nula)} \\ \text{H}_a\text{: } \mu_B > \mu_A \Rightarrow \mu_B - \mu_A > 0 & \text{(hipótese alternativa)} \end{cases}$

Regra de decisão



Estatística

$t_{calc} = \frac{\hat{\mu}_B - \hat{\mu}_A}{\sqrt{\frac{s^2}{\frac{1}{n_A}+\frac{1}{n_B}}}} = \frac{81,6-70,5}{\sqrt{\frac{155,9879}{\frac{1}{20}+\frac{1}{15}}}} = 0,35$

$t_{33;0,90} = 1,31$

$p - value = P(t > 0,35) = 0,3643 > 0,10$

Conclusão: Como $0,35 < 1,31$ (ou $p - value = 0,3643 > 0,10$), rejeita-se H_0 ao nível de 5% de significância. Portanto, existem evidências que as espécies A e B não diferem estatisticamente em termos de resistência das vigas de madeira.

- Populações:

$$X_1 \sim N(\mu_1, \sigma_1^2) \quad \text{e} \quad X_2 \sim N(\mu_2, \sigma_2^2);$$

- Hipóteses:

$$\begin{cases} \mathrm{H}_0\colon \mu_1 = \mu_2 \;\Rightarrow\; \mu_d = \mu_2 - \mu_1 = 0 & \text{(hipótese nula)} \\ \mathrm{H}_a\colon \mu_1 \underset{>}{\overset{<}{\neq}} \mu_2 \;\Rightarrow\; \mu_d = \mu_2 - \mu_1 \underset{>}{\overset{<}{\neq}} 0 & \text{(hipótese alternativa)} \end{cases}$$

- Considere amostras aleatórias de X_1 e X_2 independentes, de tamanho n_1 e n_2, respectivamente, e os estimadores

$$\bar{X}_j = \frac{\sum_{i=1}^{n} X_{ji}}{n_j}; \quad S_j^2 = \frac{\sum_{i=1}^{n} X_{ji}^2 - \frac{(\sum_{i=1}^{n} X_{ji})^2}{n}}{n-1}, \quad j = 1, 2.$$

Sob H_0 e como σ^2 é desconhecido, sabe-se que

$$\bar{X}_d = \bar{X}_2 - \bar{X}_1 \sim N\left(0, \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)\right);$$

$$T = \frac{\bar{X}_d - \mu_d}{\hat{\sigma}_{\bar{X}_d}} = \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \overset{\text{aprox}}{\sim} t_{(v)}$$

$$\text{em que} \quad v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$$

- Portanto, considera-se como estatística do teste

$$t_{calc} = \frac{\bar{x}_d - \mu_d}{\hat{\sigma}_{\bar{X}_d}} = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- Para um nível de significância α, a região crítica (RC) é construída de tal forma que $P(t_v \in \text{RC}) = \alpha$.
  - Para hipóteses uniterais:
    $P(t_{(v)} > t_{tab}) = \alpha$;
  - Para hipóteses bilaterais:
    $P(-t_{tab} < t_{(v)} < t_{tab}) = 1 - \alpha$.

sendo o número de graus de liberdade v, dado pelo inteiro mais próximo de

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Exemplo 8: Métodos de Resinagem Uma Engenheira Florestal testou dois métodos de resinagem em matrizes de *Pinus elliottii*. Um grupo de 18 das melhores matrizes foi selecionado. Usando-se casualização, aplicou-se em nove matrizes o tratamento com ácido sufúrico a 30% (tratamento A), enquanto que as demais nove matrizes receberam o tratamento de ácido sufúrico a 15% (Tratamento B). Os resultados obtidos foram:

Trat.	Produção de Resina(g)								
A	2326	2206	1835	1434	1629	1761	1511	2146	1548
B	6006	3455	3115	3376	2609	2582	3674	2648	2012

Exemplo 8: Solução

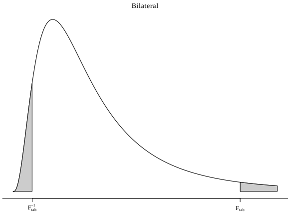
Teste de variâncias: $\alpha = 0,05$

$n_A = 9 \quad \hat{\mu}_A = 1821,778g \quad \hat{\sigma}^2_A = s^2_A = 108740,944g^2$

$n_B = 9 \quad \hat{\mu}_B = 3275,222g \quad \hat{\sigma}^2_B = s^2_B = 1324733,194g^2$

Hipóteses: $\begin{cases} \text{H}_0\text{: } \sigma^2_A = \sigma^2_B \Rightarrow \frac{\sigma^2_A}{\sigma^2_B} = 1 & \text{(hipótese nula)} \\ \text{H}_a\text{: } \sigma^2_A \neq \sigma^2_B \Rightarrow \frac{\sigma^2_A}{\sigma^2_B} \neq 1 & \text{(hipótese alternativa)} \end{cases}$

Regra de decisão



Estatística

$F_{calc} = \frac{s^2_B}{s^2_A} = \frac{1324733,194}{108740,944} = 12,18$

$F_{8;8;0,975} = 4,433$ e

$F_{8;8;0,025} = \frac{1}{F_{7;7;0,975}} = \frac{1}{4,357} = 0,2256$

$p - value = 2P(F > 12,182) =$

$2 \times 0,00096 = 0,00192 < 0,05$

Conclusão: Como $12,18 > 4,43$ (ou $p - value = 0,00192 < 0,05$), rejeita-se H_0 ao nível de 5% de significância. Logo, existem evidências que as variâncias diferem estatisticamente.

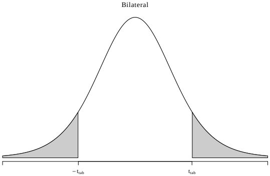
Teste de médias: $\alpha = 0,05$

$n_A = 9 \quad \hat{\mu}_A = 1821,778g \quad \hat{\sigma}^2_A = s^2_A = 108740,944g^2$

$n_B = 9 \quad \hat{\mu}_B = 3275,222g \quad \hat{\sigma}^2_B = s^2_B = 1324733,194g^2$

Hipóteses: $\begin{cases} \text{H}_0\text{: } \mu_B = \mu_A \Rightarrow \mu_B - \mu_A = 0 & \text{(hipótese nula)} \\ \text{H}_a\text{: } \mu_B \neq \mu_A \Rightarrow \mu_B - \mu_A \neq 0 & \text{(hipótese alternativa)} \end{cases}$

Regra de decisão



Estatística

$t_{calc} = \frac{\hat{\mu}_A - \hat{\mu}_B - 0}{\sqrt{\frac{s^2_A}{n_1} + \frac{s^2_B}{n_2}}} = \frac{3275{,}222 - 1821{,}778}{\sqrt{\frac{1324733{,}194}{9} + \frac{108740{,}944}{9}}} = 3,64$

$v = \frac{(s^2_A/n_A + s^2_B/n_B)^2}{(s^2_A/n_A)^2/(n_A-1) + (s^2_B/n_B)^2/(n_B-1)} = 9$

$t_{9;0{,}975} = 2,26$

$p - value = 2P(t > 3,64) = 2 \times 0,0025 = 0,0050 < 0,05$

Conclusão: Como $3,64 > 2,26$ ($p - value = 0,0050 < \alpha = 0,05$), rejeita-se H_0 ao nível de 5% de significância. Portanto, existem evidências que os métodos de resinagem são diferentes.

5
Testes em tabelas de contigência

Teste de qui-quadrado para tabelas de contingência

Tabelas de contingência: As tabelas de contingência são usadas para registrar observações independentes de duas ou mais variáveis aleatórias, normalmente qualitativas.

Testes referentes a tabelas de contingência:

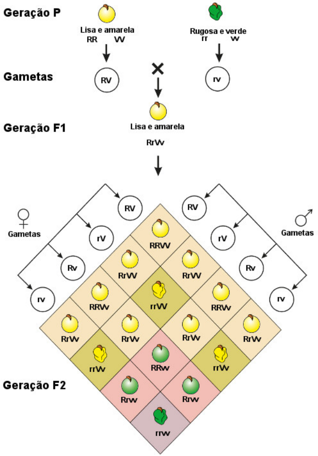
- Aderência (concordância)
- Homogeneidade
- Independência (contingência)

Leis de Mendel

Primeira lei de Mendel ou Lei da Segregação: As características dos indivíduos são determinadas por pares de fatores, os quais se separam na formação dos gametas, indo apenas um fator para cada gameta.
A Primeira Lei de Mendel aplica-se para o estudo de uma única característica.

Segunda Lei de Mendel ou Lei da Segregação Independente dos Genes: Os pares de fatores para duas ou mais características segregam-se de forma independente na formação dos gametas.

Segunda Lei de Mendel



- Mendel queria saber como ocorria a transmissão de duas ou mais características em simultâneo.
- Cruzou plantas de sementes amarelas e lisas com plantas de sementes verdes e rugosas.
- Esperava que a geração F_1 teria 100% de sementes amarelas e lisas, pois são características de caráter dominante.
- Fez autofecundação da geração F1, pois imaginava que surgiriam sementes nas proporções 9:3:3:1.

$$\left.\begin{array}{l}\text{ervilhas com sementes amarelas lisas}\\ \text{ervilhas com sementes verdes rugosas}\end{array}\right\} \Rightarrow \text{ervilhas amarelas lisas } (F_1)$$

$$\text{Autofecundação} \Rightarrow F_2 \left\{\begin{array}{l}\text{amarelas lisas } (9/16)\\ \text{verdes lisas } (3/16)\\ \text{amarelas rugosas } (3/16)\\ \text{verdes rugosas } (1/16)\end{array}\right.$$

Teste de Aderência

Objetivo: Testar a adequabilidade de um modelo probabilístico a um conjunto de dados observados.

Exemplo 9: Segundo Mendel, os resultados dos cruzamentos de ervilhas amarelas lisas com ervilhas verdes rugosas seguem uma distribuição de probabilidades dada por:

Resultado	Amarela lisa	Amarela rugosa	Verde lisa	Verde rugosa
Probabilidade	9/16	3/16	3/16	1/16

Resultados observados em 556 ervilhas na geração F2 do experimento de Mendel, envolvendo os genes para cor e textura de sementes foram:

Resultado	Amarela lisa	Amarela rugosa	Verde lisa	Verde rugosa
Fenótipos	315	101	108	32

Teste de Aderência

- Há evidências de que os resultados desse experimento estão de acordo com a distribuição de probabilidades proposta por Mendel?
- Avaliar se o padrão de segregação dos caracteres envolvidos segue a segunda lei de Mendel equivale a testar:

H_0: $\pi_1 = 9/16, \quad \pi_2 = 3/16, \; \pi_3 = 3/16, \quad \pi_4 = 1/16$
H_a: pelo menos uma das igualdades é falsa

- Se o modelo probabilístico for adequado (sob H_0), a frequência esperada é calculada por

$$E_i = n\pi_{0i}$$

em que $\pi_0 = (9/16, 3/16, 3/16, 1/16)$ e n é o número total de observações.

Teste de qui-quadrado de aderência

Frequências observadas e esperadas das quatro classes fenotópicas geradas por autofecundação de plantas dihíbridas da geração F_1

Tipos de ervilhas	Frequências observadas	Frequências esperadas sob H_0
Amarelas lisas	315	$312{,}75 = 556 \times (9/16)$
Verdes lisas	108	$104{,}25 = 556 \times (3/16)$
Amarelas rugosas	101	$104{,}25 = 556 \times (3/16)$
Verdes rugosas	32	$34{,}75 = 556 \times (1/16)$
Total	556	556

Pergunta: Pode-se afirmar que os valores observados estão suficientemente próximos dos valores esperados, isto é, o modelo probabilístico proposto por Mendel pode explicar os resultados desse experimento?

Teste de Aderência - Teste qui-quadrado

De forma genérica, seja uma tabela de frequências, com $k \times 2$ categorias de resultados:

Categorias	Frequências observadas
1	O_1
2	O_2
$\dots$	$\dots$
k	O_k
Total	n

em que O_i é o total de indivíduos observados na categoria i, $i = 1, \dots, k$ e $n = \sum_{i=1}^{k} O_i$.

Teste de Aderência - Teste qui-quadrado

- Seja π_i a probabilidade associada à categoria $i = 1, \ldots, k$. O objetivo do teste de aderência é testar as hipóteses

H_0: $\pi_1 = \pi_{01}, \;\; \pi_2 = \pi_{02}, \; \ldots, \;\; \pi_k = \pi_{0k}$

H_a: pelo menos uma das igualdades é falsa,

sendo π_{0i} a probabilidade especificada para a categoria i, $i = 1, \ldots, k$, fixada pelo modelo probabilístico de interesse.

- Se E_i é o número esperado de indivíduos na categoria i, quando a hipótese H_0 é verdadeira, então:

$$E_i = n\pi_{0i}, \;\; i = 1, \ldots, k$$

Teste de Aderência - Teste qui-quadrado

- Logo, tem-se

Categorias	Frequências observadas	Frequências esperadas
1	O_1	E_1
2	O_2	E_2
...	...	
k	O_k	E_k
Total	n	n

em que O_i é o número observado de indivíduos e $E_i = n\pi_{0i}$ é o número esperado (sob H_0) de indivíduos na categoria i, $i = 1, \ldots, k$.

- Deseja-se quantificar a diferença entre as colunas dos O_i's e dos E_i's.

Teste de qui-quadrado de aderência

Estatística do teste

$$X^2_{cal} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-1},$$

em que
k é o número de categorias da variável qualitativa

O_i é a frequência observada

E_i é a frequência esperada, sob H_0.

- **IMPORTANTE**: Esse resultado é válido para n grande e para $E_i \geq 5, i = 1, \ldots, k$.
- X^2 é sempre positivo.
- Rejeita-se H_0, a um nivel α de significância se $X^2_{cal} > \chi^2_{tab}$ ($p - value < \alpha$).
- A distribuição χ^2_ν é unimodal e assimétrica à esquerda com um único parâmetro ν que é o número de graus de liberdade.

Exemplo 9: Segundo Mendel, os resultados dos cruzamentos de ervilhas amarelas lisas com ervilhas verdes rugosas seguem uma distribuição de probabilidades dada por:

Resultado	Amarela lisa	Amarela rugosa	Verde lisa	Verde rugosa
Probabilidade	9/16	3/16	3/16	1/16

Resultados observados em 556 ervilhas na geração F2 do experimento de Mendel, envolvendo os genes para cor e textura de sementes foram:

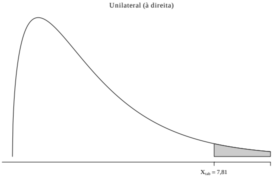
Resultado	Amarela lisa	Amarela rugosa	Verde lisa	Verde rugosa
Freq. observada	315	101	108	32
Freq. esperada	312,75	104,25	104,25	34,75

$k = 4 \quad \alpha = 0,05$

Hipóteses:

$$\begin{cases} \text{H}_0\text{: } \pi_1 = 9/16,\ \pi_2 = 3/16,\ \pi_3 = 3/16,\ \pi_4 = 1/16 & \text{(hipótese nula)} \\ \text{H}_a\text{: pelo menos uma das igualdades é falsa} & \text{(hipótese alternativa)} \end{cases}$$

Regra de decisão



Estatística

$$X^2_{calc} = \frac{(315-312{,}75)^2}{312{,}75} + \frac{(101-104{,}25)^2}{104{,}25} + \frac{(108-104{,}25)^2}{104{,}25} + \frac{(32-34{,}75)^2}{34{,}75} = 0,470.$$

$\chi_{3;0,95} = 7,81$

$p - value = P(X^2 > 0,470) = 0,925 > \alpha = 0,05$

Conclusão: Como $0,470 < 7,81$ ($p - value = 0,925 > \alpha = 0,05$), não se rejeita H_0 ao nível de 5% de significância. Portanto, existem evidências que o modelo de Mendel explica os resultados obtidos.

Exemplo 10: Um engenheiro florestal deseja saber se a distribuição normal explica o diâmetro de árvores de uma floresta plantada de *Eucalyptus camaldulensis*. A tabela que se segue apresenta a frequência *observada* para cada classe de diâmetro de uma amostra de $n = 1500$ árvores.

Classes de Diâmetros (cm)		Frequência
Lim. Inferior	Lim. Superior	(arv/ha)
10,0	12,5	4
12,5	15,0	23
15,0	17,5	50
17,5	20,0	132
20,0	22,5	224
22,5	25,0	286
25,0	27,5	295
27,5	30,0	243
30,0	32,5	161
32,5	35,0	73
35,0	37,5	9
		$n = 1500$

Exemplos

- As hipóteses sendo testadas são:

  H_0: $\pi_i = \pi_{0i} \Rightarrow \widehat{f_i} = n\pi_{0i}, i = 1, 2, \ldots, k$
  H_a: pelo menos uma das igualdades é falsa,

  em que f_i é a frequência observada e $\widehat{f_i}$ é a frequência esperada supondo que os dados provêem de uma distribuição normal.
- Isso é equivalente a testar se $X \sim \mathrm{N}(\bar{x}_{dap}, s^2_{dap})$, sendo $\bar{x}_{dap}$ e s^2_{dap} a média e a variância estimadas.
- Para se verificar se as frequências observadas e as esperadas estão próximas, devem-se calcular as *frequências esperadas*.
- Para cada classe de diâmetro fazem-se os seguintes cálculos:
  - **Valores da normal padronizada para os limites de classes:** $z_{I,i} = \frac{dap_{I,i} - \bar{x}_{dap}}{\sqrt{s^2_{dap}}}$ e $z_{S,i} = \frac{dap_{S,i} - \bar{x}_{dap}}{\sqrt{s^2_{dap}}}$
  - **Probabilidade por classe da Tabela:** $\widehat{p}_i = P(Z < z_{I,i}) - P(Z < z_{S,i})$
  - **Frequência esperada** (N=1500): $\widehat{f_i} = \widehat{p}_i\, n$

Classes	Diâmetros (cm)		Frequência	Normal Padronizada		Probab.	Frequência
(i)	Limite Inferior	Limite Superior	Observada (arv/ha)	Limite Inferior	Limite Superior	por Classe	Esperada (arv/ha)
	$dap_{I,i}$	$dap_{S,i}$	f_i	$z_{I,i}$	$z_{S,i}$	$\widehat{p}_i$	$\widehat{f}_i$
1	10,0	12,5	4	-3,23	-2,69	0,0029	4,33
2	12,5	15,0	23	-2,69	-2,16	0,0117	17,59
3	15,0	17,5	50	-2,16	-1,63	0,0361	54,18
4	17,5	20,0	132	-1,63	-1,10	0,0843	126,52
5	20,0	22,5	224	-1,10	-0,57	0,1494	224,06
6	22,5	25,0	286	-0,57	-0,04	0,2006	300,96
7	25,0	27,5	295	-0,04	0,50	0,2044	306,61
8	27,5	30,0	243	0,50	1,03	0,1580	236,93
9	30,0	32,5	161	1,03	1,56	0,0926	138,86
10	32,5	35,0	73	1,56	2,09	0,0411	61,72
11	35,0	37,5	9	2,09	2,62	0,0139	20,80
			1500			0,9950	1492,55

Note que $\sum_{i=1}^{k} \hat{f}_i \neq \sum_{i=1}^{k} f_i$, pois

$\hat{p}_0 = P(-\infty < Z < -3,23) = 0,0006 \Rightarrow \hat{f}_0 = 0,94$

$\hat{p}_{12} = P(2,62 < Z < \infty) = 0,0043 \Rightarrow \hat{f}_{12} = 6,51$

Estatística do teste

A estatística utilizada para comparar as frequências é

$$X^2 = \sum_{i=1}^{k} \frac{(f_i - \widehat{f_i})^2}{\widehat{f_i}}$$

- Grandes valores de X^2 indicam que as frequências não são semelhantes, levando a rejeitar H_0.
- Sob H_0, a estatística

  $$X^2 = \sum_{i=1}^{k} \frac{(f_i - \widehat{f_i})^2}{\widehat{f_i}} \sim \chi^2_\nu$$

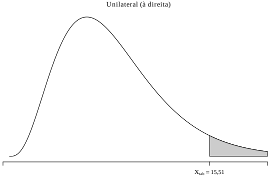
  em que $\nu = n - 1 - p$ é o número de graus de liberdade, sendo p o número de parâmetros estimados para encontrar as frequências esperadas.
- Rejeita-se H_0, a um nivel α de significância se $X^2_{cal} > \chi^2_{tab}$ $(p - value < \alpha)$.

$k = 11 \quad p = 2 \quad \nu = 11 - 1 - 2 = 8 \quad \alpha = 0,05$

$\bar{x}_{dap} = \frac{\sum_{i=1}^{k} \bar{x}_i f_i}{\sum_{i=1}^{k} f_i} = 25,16 \quad s^2_{dap} = \frac{1}{\sum_{i=1}^{k} f_i - 1}\left[\sum_{i=1}^{k} \bar{x}_i f_i - \frac{(\sum_{i=1}^{k} \bar{x}_i f_i)^2}{\sum_{i=1}^{k} f_i}\right] = 22,0822$

Hipóteses: $\begin{cases} \mathrm{H}_0\text{: } \pi_i = \pi_{0i} \Rightarrow \widehat{f_i} = n\pi_{0i}, i = 1, 2, \ldots, k & \text{(hipótese nula)} \\ \mathrm{H}_a\text{: pelo menos uma das igualdades é falsa} & \text{(hipótese alternativa)} \end{cases}$

Regra de decisão



Estatística

$X^2_{calc} = 15,8734$

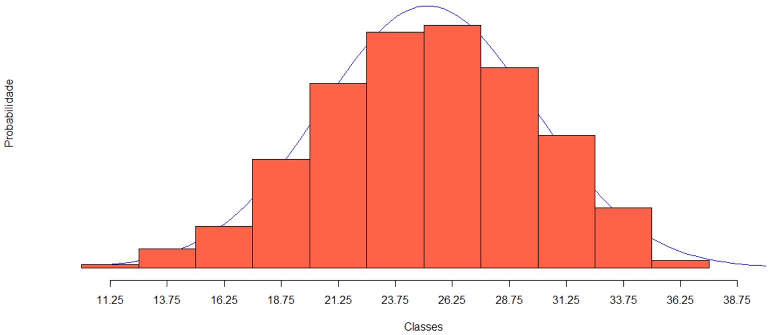
$\chi_{8;0,95} = 15,507$

$p - value = P(X^2 > 15,507) = 0,0442 < 0,05$

Conclusão: Como $15,8734 > 15,507$ ($p - value = 0,0442 < \alpha = 0,05$), rejeita-se H_0 ao nível de 5% de significância. Portanto, existem evidências que a distribuição normal não explica os resultados obtidos para o diâmetro das árvores da floresta.

Classes (i)	Diâmetros (cm) Limite Inferior $dap_{I,i}$	Limite Superior $dap_{S,i}$	Frequência Observada (arv/ha) f_i	Frequência Esperada (arv/ha) $\widehat{f_i}$	X^2 $\frac{(f_i - \widehat{f_i})^2}{\widehat{f_i}}$
1	10,0	12,5	4	4,33	0,0252
2	12.5	15.0	23	17,59	1,6631
3	15,0	17,5	50	54,18	0,3221
4	17,5	20,0	132	126,52	0,2375
5	20,0	22,5	224	224,06	0,0001
6	22,5	25,0	286	300,96	0,7432
7	25,0	27,5	295	306,61	0,4396
8	27,5	30,0	243	236,93	0,1556
9	30.0	32.5	161	138,86	3,5301
10	30,0	32,5	73	61,72	2,0619
11	35,0	37,5	9	20,80	6,6951
			1500	1492,55	15,8734

Note que $\sum_{i=1}^{k} \hat{f}_i \neq \sum_{i=1}^{k} f_i$, pois
$\hat{p}_0 = P(-\infty < Z < -3,23) = 0,0006 \Rightarrow \hat{f}_0 = 0,94$ (< 5, deve ser incorporada na primeira classe)
$\hat{p}_{01} = P(-\infty < Z < -2,69) = 0,0035 \Rightarrow \hat{f}_{01} = 5,26$ (incorporada na primeira classe)
$\hat{p}_{12} = P(2,62 < Z < \infty) = 0,0043 \Rightarrow \hat{f}_{12} = 6,51$ (deve-se criar uma classe adicional)
$X^2_{mod} = 15,8734 - 0,0252 + \frac{(4-5,26)^2}{5,26} + \frac{(0-6,51)^2}{6,51} = 22,66$ com 12-1-2 = 9 graus de liberdade.


Probabilidade
11.25
13.75
16.25
18.75
21.25
23.75
26.25
28.75
31.25
33.75
36.25
38.75
Classes

Exemplos

Exemplo 11: Os dados que se seguem referem-se a um estudo da distribuição da espécie *Primula simenses* selvagem em uma certa região florestal que foi dividida em 109 áreas. Verifique se a distribuição de Poisson ajusta-se bem a esses dados.

$X_i = x_i$	$f(x_i)$	$P(X_i = x_i)$	$\hat{f}(x_i)$
0	26		
1	21		
2	23		
3	14		
4	11		
5	4		
6	5		
7	4		
8	1		
Total	$n = 109$		

x_i: número de plantas por área
$f(x_i)$: número de áreas com x_i plantas (frequência observada)
$\hat{f}(x_i) = n \times P(X_i = x_i)$: frequência esperada sob a P($\hat{\lambda}$) em que $\hat{\lambda} = \frac{\sum_{i=1}^{k} f(x_i)x_i}{n}$

Exemplos

Exemplo 12: Um engenheiro florestal deseja saber se a distribuição Weibull explica o diâmetro de árvores de uma floresta inequiânea. A tabela que se segue apresenta a frequência *observada* para cada classe de diâmetro de uma amostra de $n =$ árvores. considerando os parâmetros alfa=10,0; beta=4,851943; gama=0,586629, a um nível de significância de 5

Classes (i)	Diâmetros (cm) Limite Inferior $dap_{I,i}$	Limite Superior $dap_{S,i}$	Frequência Observada (arv/ha) f_i	Frequência Esperada (arv/ha) $\widehat{f}_i$	X^2 $\frac{(f_i - \widehat{f}_i)^2}{\widehat{f}_i}$
1	10	20	336	406	
2	20	30	77	60	
3	30	40	42	24	
4	40	50	14	12	
5	50	60	12	6	
6	60	70	9	4	
7	70	80	3	2	
8	80	90	4	1	
9	90	100	1	1	
10	100	110	1	1	

As hipóteses testadas neste exemplo, são:
H_0:. A distribuição de Weibull explica a distribuição diamétrica observada.

H_a:. Não H_0.

Exemplos

Exercício: Os dados que se seguem referem-se a números de estacas enraizadas em oitenta vasos com 8 estacas de hortência. Teste se a distribuição binomial se ajusta a esses dados, ao nível de 5% de significância.

$X_i = x_i$	$f(x_i)$	$P(X_i = x_i)$	$\hat{f}(x_i)$
0	6		
1	20		
2	28		
3	12		
4	8		
5	0		
6	0		
7	0		
8	0		
Total	$n = 80$		

x_i: número de estacas enraizadas
$f(x_i)$: número de vasos com x_i estacas enraizadas (frequência observada)
$\hat{p}_i$: probabilidade estimada de x_i estacas enraizadas sob Bin(8, $\hat{\pi}$)
$\hat{f}_i$: número esperado de vasos com x_i estacas enraizadas, sob Bin(8, $\hat{\pi}$)
$\hat{\pi} = \frac{\sum_{i=0}^{k} f(x_i) x_i}{n}$

Exemplos

Exercício: Suponha que em um experimento de cultura de tecidos, um pesquisador colocou $n = 40$ recipientes com $m = 3$ explantes em um mesmo meio de cultura. Após uma semana, ele observou os dados que se seguem. Considere as situações:

1. Por já ter feito muitos experimentos desse tipo, ele espera que a probabilidade de um explante regenerar é $\pi = 0,3$; verifique se uma distribuição Bin(3; 0,3) se ajusta a esses dados;
2. É a primeira vez que faz esse tipo de experimento e está interessado em estimar a probabilidade de regeneração ($\hat{\pi}$) e verificar se a distribuição binomial Bin(3, $\hat{\pi}$) ajusta-se aos dados obtidos;
3. Verifique se a distribuição Bin(3, $\hat{\pi}$) não difere da Bin(3;0,3).

$X_i = x_i$	$f(x_i)$	p_i	$f_{0,i}$	$\hat{p}_i$	$\hat{f}_i$
0	11				
1	20				
2	7				
3	2				
Total	$n = 40$				

x_i: número de explantes regenerados
$f(x_i)$: número de recipientes com x_i explantes regenerados (frequência observada)
p_i: probabilidade de x_i explantes regenerados sob Bin(3; 0,3)
$f_{0,i}$: número esperado de recipientes com x_i explantes regenerados, sob Bin(3; 0,3);
$\hat{p}_i$: probabilidade de x_i explantes regenerados sob Bin(3, $\hat{\pi}$)
$\hat{f}_i$: número esperado de recipientes com x_i explantes regenerados, sob Bin(3, $\hat{\pi}$)
$\hat{\pi} = \frac{\sum_{i=0}^{k} f(x_i) x_i}{n}$

Exercício: Cruzando-se a variedade de tomate "Santa Cruz" de hipocótilo roxo e folha recortada (AACC) com a variedade "Folha de Batata", de hipocótilo verde e folha normal (aacc) obteve-se a geração F1 de hipocótilo roxo e folha recortada (AaCc) que autofecundada deu origem à geração segregante F2 na proporção. Verifique se as frequências observadas estão na proporção 9 : 3 : 3 : 1 (Lei de Mendel), ao nível de 5% de significância.

Fenótipo	f_i	$\hat{f}_i$
Hipocótilo roxo, folha recortada	105	
Hipocótilo roxo, folha normal	37	
Hipocótilo verde, folha recortada	22	
Hipocótilo verde, folha normal	8	
Total	$n = 172$	

f_i: frequência observada
$\hat{f}_i$: frequência esperada sob a lei de Mendel.

Testes de Independência: Teste de qui-quadrado

- **Objetivo**: Verificar se existe independência entre duas variáveis, ou seja, se não existe associação entre as duas variáveis.
- Em geral, os dados referem-se a mensurações de duas características (A e B) feitas em n unidades experimentais, que são apresentadas na tabela com r linhas e s colunas que se segue.

Tabela: Tabela de contingência com duas entradas

	B				
A	1	2	...	s	
1	O_{11}	O_{12}	...	O_{1s}	$O_{1.}$
2	O_{21}	O_{22}	...	O_{2s}	$O_{2.}$
...	...	...	...	...	...
r	O_{r1}	O_{r2}	...	O_{rs}	$O_{r.}$
	$O_{.1}$	$O_{.2}$	...	$O_{.s}$	$n = O_{..}$

Hipóteses:

$$\begin{cases} H_0\text{: A e B são independentes} & \text{(hipótese nula)} \\ H_a\text{: A e B não são independentes} & \text{(hipótese alternativa)} \end{cases}$$

Então, sob independência, tem-se que:

Hipóteses: $\begin{cases} \text{H}_0\text{: } \pi_{ij} = \pi_{i.}\pi_{.j} & \text{(hipótese nula)} \\ \text{H}_a\text{: } \pi_{ij} \neq \pi_{i.}\pi_{.j} & \text{(hipótese alternativa)} \end{cases}$

A	B				Total
	B_1	B_2	...	B_s	
A_1	π_{11}	π_{12}	...	π_{1s}	$\pi_{1.}$
A_2	π_{21}	π_{22}	...	π_{2s}	$\pi_{2.}$
...	...	...	...	...	...
A_r	π_{r1}	π_{r2}	...	π_{rs}	$\pi_{r.}$
Total	$\pi_{.1}$	$\pi_{.2}$	...	$\pi_{.s}$	$1 = \pi_{..}$

Como $\hat{\pi}_{i.} = \frac{O_{i.}}{n}$ e $\hat{\pi}_{.j} = \frac{O_{.j}}{n}$, então, a frequência esperada, sob independência, é calculada por

$$E_{ij} = n\hat{\pi}_{ij} = \frac{O_{i.}O_{.j}}{n}$$

isto é, o número esperado de observações no cruzamento das categorias A_i e B_j, sob independência, é igual ao produto dos totais marginais correspondentes dividido pelo total geral.

Testes de Independência: Teste de qui-quadrado

Estatística do teste

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, i = 1, 2, \ldots r; \;\; j = 1, 2, \ldots, s.$$

- Valores grandes de X^2 indicam que as frequências observadas e esperadas não são semelhantes, levando a rejeitar H_0.
- Sob H_0, isto é, sob independência, a estatística

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_\nu$$

  em que $\nu = (r-1)(s-1)$ é o número de graus de liberdade.
- Rejeita-se H_0, a um nivel α de significância se $X^2_{cal} > \chi^2_{tab}$ $(p - value < \alpha)$.

Exemplo 13: Contagem de plantas segregando para dois caracteres: ciclo e virescência (formação de cloroplastos nas pétalas, originando plantas verdes), numa progênie da espécie X

Ciclo	Virescência		Total
	Normal	Virescente	
Tardio	**3470**	910	**4380**
Precoce	1030	290	1320
Total	**4500**	1200	**5700**

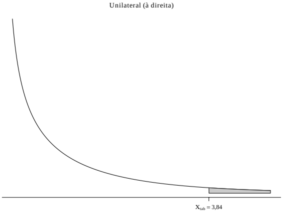
$$\begin{cases} E_{11} = \frac{O_{1.}O_{.1}}{n} = \frac{4380 \times 4500}{5700} = 3457,8947 & E_{12} = \frac{O_{1.}O_{.2}}{n} = \frac{4380 \times 1200}{5700} = 922,1053 \\ E_{21} = \frac{O_{2.}O_{.1}}{n} = \frac{1320 \times 4500}{5700} = 1042,1053 & E_{22} = \frac{O_{2.}O_{.2}}{n} = \frac{1320 \times 1200}{5700} = 277,8947 \end{cases}$$

$r = 2 \quad s = 2 \quad \nu = (2-1)(2-1) = 1 \;\; \alpha = 0,05$

Hipóteses:

$$\begin{cases} \text{H}_0\text{: } \pi_{ij} = \pi_{i.}\pi_{.j}, i = 1, 2, \ldots, r, \;\; j = i = 1, 2, \ldots, s & \text{(hipótese nula)} \\ \text{H}_a\text{: pelo menos uma das igualdades é falsa} & \text{(hipótese alternativa)} \end{cases}$$

Regra de decisão



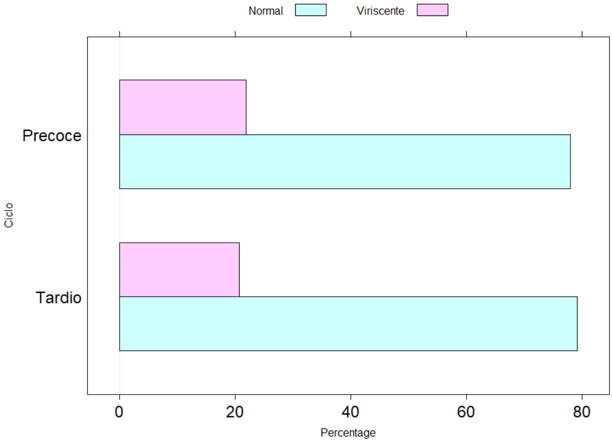
Estatística

$$X^2_{calc} = \frac{(3470-33457{,}8947)^2}{3457{,}8947} + \frac{(910-922{,}1053)^2}{922{,}1053} + \frac{(1030-1042{,}1053)^2}{1042{,}1053} + \frac{(290-277{,}8947)^2}{277{,}8947} = 0,8692.$$

$\chi_{1;0,95} = 3,84$

$p - value = P(X^2 > 0,8692) = 0,3511 > 0,05$

Conclusão: Como $0,8692 < 3.84$ ($p - value = 0,3511 > \alpha = 0,05$), não se rejeita H_0 ao nível de 5% de significância. Portanto, existem evidências que não existe associação entre viriscência e ciclo da planta.


Normal
Viriscente
Precoce
Tardio
Ciclo
0
20
40
60
80
Percentage

Exemplo 14: Os dados que se seguem referem-se a um total de 73 populações de espécies de plantas em dois ambientes de restinga (duna e floresta), classificadas quanto à herbivoria, isto é, quanto à perda de área foliar (alta ou baixa). Existe relação entre o ambiente e a perda de área foliar, ao nível de 5% de significância?

Herbivoria	Ambiente		Total
	Floresta	Duna	
Alta	**15**	32	**47**
Baixa	18	8	26
Total	**33**	40	**73**

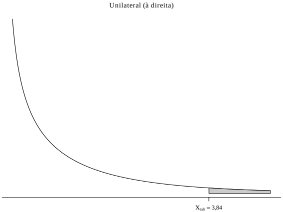
$$\begin{cases} E_{11} = \frac{O_{1.}O_{.1}}{n} = \frac{47\times 33}{73} = 21,24658 & E_{12} = \frac{O_{1.}O_{.2}}{n} = \frac{47\times 40}{73} = 25,75342 \\ E_{21} = \frac{O_{2.}O_{.1}}{n} = \frac{26\times 33}{73} = 11,75342 & E_{22} = \frac{O_{2.}O_{.2}}{n} = \frac{26\times 40}{73} = 14,24658 \end{cases}$$

$r = 2 \quad s = 2 \quad \nu = (2-1)(2-1) = 1 \;\; \alpha = 0,05$

Hipóteses:

$$\begin{cases} \text{H}_0\text{: } \pi_{ij} = \pi_{i.}\pi_{.j}, i = 1,2,\ldots,r, \;\; j = i = 1,2,\ldots,s & \text{(hipótese nula)} \\ \text{H}_a\text{: pelo menos uma das igualdades é falsa} & \text{(hipótese alternativa)} \end{cases}$$

Regra de decisão



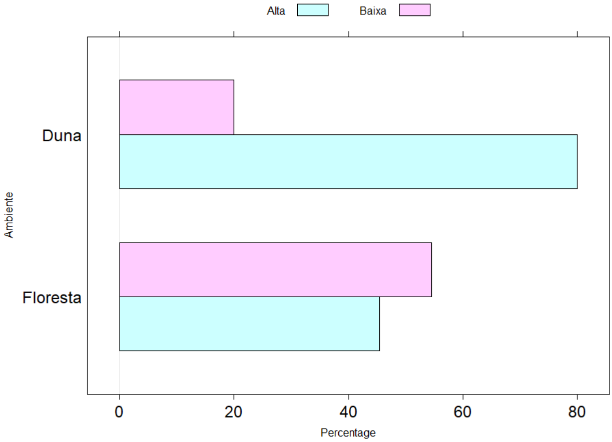
Estatística

$X^2_{calc} = \frac{(15-21,24658)^2}{21,24658} + \frac{(32-25,75342)^2}{25,75342} + \frac{(18-11,75342)^2}{11,75342} + \frac{(8-14,24658)^2}{14,24658} = 9,41.$

$\chi_{1;0,95} = 3,84$

$p - value = P(X^2 > 9,41) = 0,0022 < 0,05$

Conclusão: Como $9,41 < 3.84$ ($p - value = 0,0022 < \alpha = 0,05$), rejeita-se H_0 ao nível de 5% de significância. Portanto, existem evidências que existe associação entre o ambiente e a perda de área foliar.


Alta
Baixa
Duna
Floresta
Ambiente
0
20
40
60
80
Percentage

Teste de homogeneidade - Teste qui-quadrado para duas ou mais proporções

- **Objetivo**: Testar a igualdade de duas ou mais proporções.

Hipóteses: $\begin{cases} H_0\text{: } \pi_{1j} = \pi_{2j} = \ldots = \pi_{rj} = \pi_j, \;\; j = 1,2 & \text{(hipótese nula)} \\ H_a\text{: pelo menos uma das igualdades é falsa} & \text{(hipótese alternativa)} \end{cases}$

A	B		Total
	B_1	B_2	
A_1	π_{11}	π_{12}	$\pi_{1.}$
A_2	π_{21}	π_{22}	$\pi_{2.}$
...	...	...	...
A_r	π_{r1}	π_{r2}	$\pi_{r.}$
Total	$\pi_{.1}$	$\pi_{.2}$	$\pi_{..}$

Como $\hat{\pi}_{ij} = \frac{O_{ij}}{O_{i.}} = \frac{O_{.j}}{n}$, então, a frequência esperada, sob homogeneidade, é calculada por

$$E_{ij} = O_{i.}\hat{\pi}_{ij} = \frac{O_{i.}O_{.j}}{n}$$

isto é, o número esperado de observações no cruzamento das categorias A_i e B_j, sob homogeneidade de proporções, é igual ao produto dos totais marginais correspondentes dividido pelo total geral.

Teste de homogeneidade - Teste qui-quadrado para duas ou mais proporções

Estatística do teste

$$X^2_{cal} = \sum_{i=1}^{s}\sum_{j=1}^{r}\frac{(O_{ij}-E_{ij})^2}{E_{ij}},$$

em que
O_{ij} é a frequência observada de elementos na categoria i da variável A e categoria j da variável B,
E_{ij} é a frequência esperada de elementos nessa categoria, dada por:

$$E_{ij} = \frac{O_{i\cdot}.O_{\cdot j}}{O_{\cdot\cdot}},$$

$O_{i\cdot}$, $O_{\cdot j}$ e $O_{\cdot\cdot}$ representam as frequências marginais e o total da tabela de contingência a ser analisada.

- Valores grandes de X^2 indicam que as frequências observadas e esperadas não são semelhantes, levando a rejeitar H_0.
- Sob H_0, isto é, sob independência, a estatística

  $$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \sim \chi^2_{\nu}$$

  em que $\nu = r - 1$ é o número de graus de liberdade.
- Rejeita-se H_0, a um nivel α de significância se $X^2_{cal} > \chi^2_{tab}$ ($p - value < \alpha$).

Teste de qui-quadrado para duas proporções

Exemplo 15: Um experimento foi conduzido com 480 estacas de ameixeira enxertadas na primavera e 480 estacas enxertadas fora da primavera para comparar duas épocas de plantio.

Distribuição conjunta das frequências das variáveis época de plantio e sobrevivência de enxertos de ameixeiras

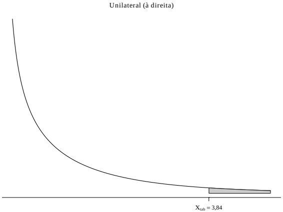
Época	Raízes		Total
	Sobreviventes	Mortas	
Fora da primavera	**263**	217	**480**
Na primavera	**115**	365	**480**
Total	378	582	**960**

$$\begin{cases} E_{11} = \frac{O_{1.}O_{.1}}{n} = \frac{480 \times 378}{960} = 189 & E_{12} = \frac{O_{1.}O_{.2}}{n} = \frac{480 \times 582}{960} = 291 \\ E_{21} = \frac{O_{2.}O_{.1}}{n} = \frac{480 \times 378}{960} = 189 & E_{22} = \frac{O_{2.}O_{.2}}{n} = \frac{480 \times 582}{960} = 291 \end{cases}$$

$r = 2 \quad \nu = 2 - 1 = 1 \quad \alpha = 0,05$

Hipóteses: $\begin{cases} \text{H}_0\text{: } \pi_{1j} = \pi_{2j} \;\; j = 1,2 & \text{(hipótese nula)} \\ \text{H}_a\text{: } \pi_{1j} \neq \pi_{2j} & \text{(hipótese alternativa)} \end{cases}$

Regra de decisão



Estatística

$X^2_{calc} = \frac{(263-189)^2}{189} + \frac{(217-291)^2}{291} + \frac{(115-189)^2}{189} + \frac{(365-291)^2}{291} = 95,58.$

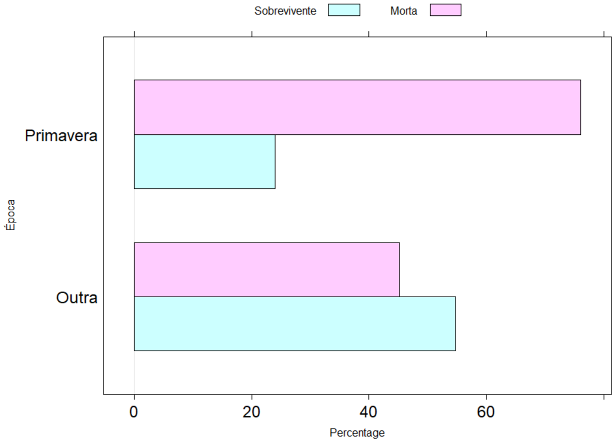
$\chi_{1;0,95} = 3,84$

$p - value = P(X^2 > 95,58) < 0,05$

Conclusão: Como $95,58 > 3,84$ ($p - value < \alpha = 0,05$), rejeita-se H_0 ao nível de 5% de significância. Portanto, existem evidências que a proporção de sobreviventes na primavera difere da proporção de sobreviventes fora dela. As proporções estimadas sâo:

263/480 = 0,5479 proporção de sobreviventes se plantadas fora da primavera

115/480 = 0,2396 proporção de sobreviventes se plantadas na primavera


Sobrevivente
Morta
Primavera
Outra
Época
0
20
40
60
Percentage

Teste de qui-quadrado para três proporções

Exemplo 16: Os dados que se seguem referem-se a um experimento, em que (240, 240 e 300) luvas de três marcas (A, B e C) foram submetidas a tensão, com o objetivo de verificar se são homogêneas quanto à permeabilidade a vírus. Existem evidências estatísticas ao nível de 5% de significância a favor da hipótese de que as três marcas possuem a mesma permeabilidade?

Distribuição conjunta das frequências das variáveis marca da luva e permeabilidade a vírus.

Marca da luva	Deixou passar vírus? Sim	Não	Total
A	**151**	89	**240**
B	**134**	106	**240**
C	**177**	123	**300**
Total	462	318	**780**

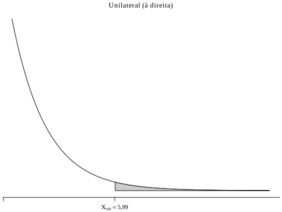
$$\begin{cases} E_{11} = \frac{O_{1.}O_{.1}}{n} = \frac{240\times 462}{780} = 142,1538 & E_{12} = \frac{O_{1.}O_{.2}}{n} = \frac{240\times 318}{780} = 97,8462 \\ E_{21} = \frac{O_{2.}O_{.1}}{n} = \frac{240\times 462}{780} = 142,1538 & E_{22} = \frac{O_{2.}O_{.2}}{n} = \frac{240\times 318}{780} = 97,8462 \\ E_{31} = \frac{O_{3.}O_{.1}}{n} = \frac{300\times 462}{780} = 177,6923 & E_{32} = \frac{O_{3.}O_{.2}}{n} = \frac{300\times 318}{780} = 122,3077 \end{cases}$$

$r = 3 \quad \nu = 3 - 1 = 2 \quad \alpha = 0,05$

Hipóteses:

$$\begin{cases} H_0: \pi_{1j} = \pi_{2j} == \pi_{3j}, \ \ j = 1,2 & \text{(hipótese nula)} \\ H_a: \text{pelo menos uma das igualdades é falsa} & \text{(hipótese alternativa)} \end{cases}$$

Regra de decisão



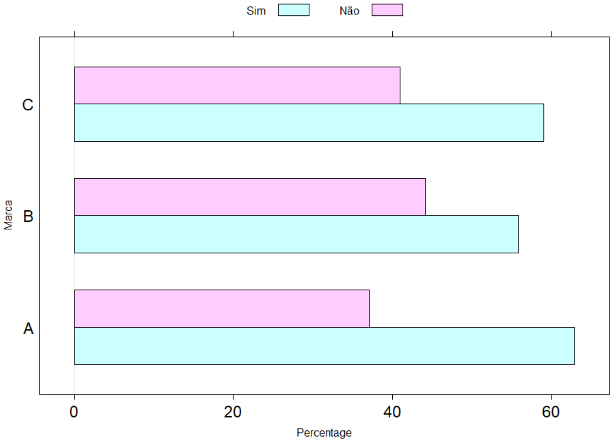
Estatística

$$X^2_{calc} = \frac{(151-142{,}1538)^2}{142{,}1538} + \frac{(89-97{,}8462)^2}{97{,}8462} + \frac{(134-142{,}1538)^2}{142{,}1538} + \frac{(106-97{,}8462)^2}{97{,}8462} + \frac{(177-177{,}6923)^2}{177{,}6923} + \frac{(123-122{,}3077)^2}{122{,}3077} = 2,50.$$

$\chi_{2;0,95} = 5,99$

$p - value = P(X^2 > 2,50) = 0,2859 > 0,05$

Conclusão: Como $2,50 < 5,99$ ($p - value = 0,2859 > \alpha = 0,05$), não se rejeita H_0 ao nível de 5% de significância. Portanto, existem evidências que as marcas de luvas não diferem quanto à permeabilidade a vírus.



Sim
Não
C
B
A
Marca
0
20
40
60
Percentage

Correção de continuidade ou correção de Yates

- Ao usar o teste χ^2, supõe-se amostragem aleatória e que o tamanho amostral será relativamente grande.
- Quando a amostra é pequena e/ou a frequência esperada em uma das caselas é pequena (≤ 5) pode ser que $X^2_{calc} > \chi^2_{Tab}$, sendo maior do que o valor real.
- Nesses casos, Fisher recomenda o uso de um fator de correção de continuidade para cada casela, para evitar eventuais conclusões erradas.

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(|O_{ij} - E_{ij}| - 0,5)^2}{E_{ij}}, i = 1, 2, \ldots r; \;\; j = 1, 2, \ldots, s.$$

- De modo geral, usa-se a correção de Yates quando:
  1. $X^2_{calc} > \chi^2_{Tab}$ e $n \leq 40$ ou
  2. $X^2_{calc} > \chi^2_{Tab}$ e há pelo menos uma casela com frequência esperada ≤ 5.

- Em uma tabela 2×2, nenhuma frequência esperada deve ser ≤ 5.
- Em tabelas maiores do que 2×2, recomenda-se que as caselas de frequências esperadas não tenham nenhum valor ≤ 5.
- Sempre que em uma tabela 2×2, qualquer das frequências esperadas for < 10, usa-se a correção de Yates.
- **Regra geral**: O teste χ^2 deve ser evitado quando a soma das frequências esperadas (= soma das frequências observadas), seja < 30. Caso seja menor e ocorra em uma tabela 2×2 deve-se usar o Teste Exato de Fisher.

Exercício: Com o objetivo de verificar o efeito da exposição do solo sobre a microfauna do solo, uma pesquisadora fez levantamentos de microfauna antes e depois da exposição do solo em áreas desmatadas de vários tipos de ecossistemas conforme tabela que se segue. Teste a 5% de significância se existe associação entre ecossistemas e exposição do solo em áreas desmatadas.

Ecossistema	No. de Micro-organismos/cm^3	
	Antes	Depois
Campo Limpo	1430	780
Cerrado	2500	1020
Restinga	732	640
Caatinga	640	680
Floresta Estacional	10530	2520
Floresta Pluvial	21883	2302

Exercício: Os dados que se seguem, referem-se à distribuição conjunta das frequências de (592) estudantes de uma universidade segundo as variáveis cor dos cabelos e cor de olhos.

Cor do cabelo	Cor de olhos: Verde	Acinzentado	Azul	Castanho	Total
Preto	5	15	20	68	108
Castanho	29	54	84	119	286
Ruivo	14	14	17	26	71
Loiro	16	10	94	7	127
Total	64	93	215	220	592

Pede-se:

1. Existe associação entre a cor de olhos e a cor dos cabelos de uma pessoa? (teste de independencia)
2. A distribuição percentual das cores de olhos é igual em cada cor de cabelo (e vice-versa)? (teste de homogeneidade)