UNIVERSIDADE FEDERAL DO PARANÁ
Setor de Ciências Agrárias
Curso de Engenharia Florestal
Departamento de Ciências Florestais

JULIA E EXPERIMENTOS FLORESTAIS

PRIMEIRO CURSO



Henrique Soares Koehler | Sylvio Péllico Netto
Liz Lorena Cabrera Hong | Renilson Lisboa Júnior
Alexandre Behling

CURITIBA
2026

UNIVERSIDADE FEDERAL DO PARANÁ
Setor de Ciências Agrárias
Curso de Engenharia Florestal
Departamento de Ciências Florestais

JULIA E EXPERIMENTOS FLORESTAIS

PRIMEIRO CURSO

Henrique Soares Koehler | Sylvio Péllico Netto
Liz Lorena Cabrera Hong | Renilson Lisboa Júnior
Alexandre Behling

CURITIBA
EDIÇÃO DOS AUTORES
2026

Os materiais podem ser utilizados livremente para reprodução das análises apresentadas, estudo dos procedimentos estatísticos, adaptação dos códigos a diferentes conjuntos de dados e desenvolvimento de novas aplicações. O repositório poderá ser atualizado com correções, aprimoramentos e novos materiais relacionados ao livro.

Todas as figuras e tabelas apresentadas neste livro foram elaboradas pelos autores.

Foto Capa: Johannes Plenio - Forest Covered in White Fog, disponível em Pexels.

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
UNIVERSIDADE FEDERAL DO PARANÁ / SISTEMA DE BIBLIOTECAS
BIBLIOTECA DE CIÊNCIAS FLORESTAIS E DA MADEIRA

Julia e experimentos florestais: primeiro curso / Henrique Soares Koehler… [et al.]. Curitiba : Ed. dos Autores, 2026.
1 recurso online : PDF

ISBN 978-65-02-38666-8
Autores: Henrique Soares Koehler, Sylvio Péllico Netto, Liz Lorena Cabrera Hong, Renilson Lisboa Júnior, Alexandre Behling

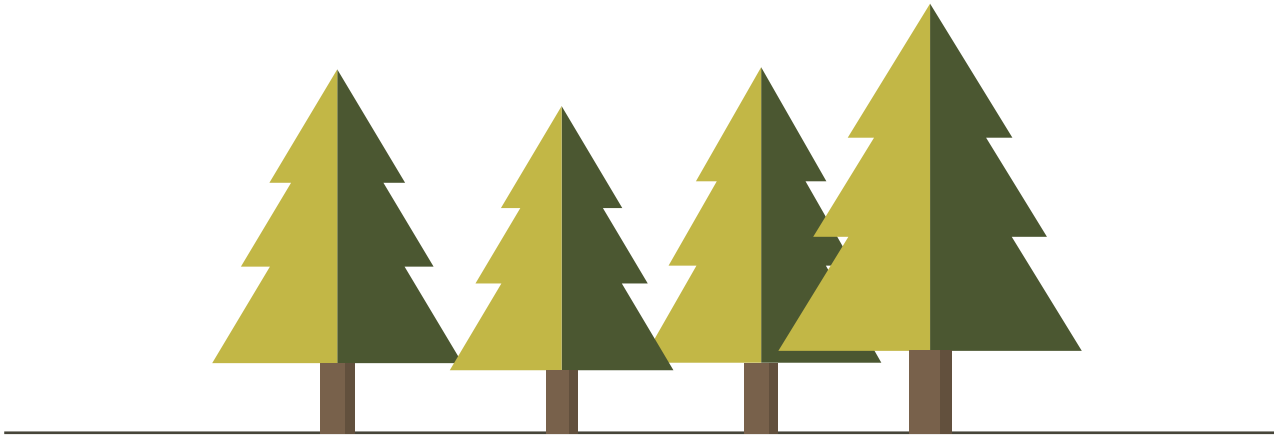
1. Julia (Linguagem de programação de computador). 2. Análise de dados. 3. Pesquisa florestal - Métodos estatísticos. 4. Florestas - Métodos estatísticos. 5. Engenharia florestal - Métodos estatísticos. I. Koehler, Henrique Soares. II. Péllico Netto, Sylvio. III. Hong, Liz. IV. Lisboa Júnior, Renilson. V. Behling, Alexandre. VI. Universidade Federal do Paraná. Setor de Ciências Agrárias. Departamento de Ciências Florestais. Curso de Engenharia Florestal. VII. Título.

CDD - 634.90285513
CDU - 630:519.22:004.43
CDU/GFDC - 634.0.51

Bibliotecária: Marilene do Rocio Veiga CRB9/424

Qualquer parte desta publicação pode ser reproduzida, desde que citada a fonte.

Os autores expressam seus sinceros agradecimentos ao Prof. Fabiano de Oliveira Fortes (in memoriam), pela generosa colaboração na revisão deste texto. Sua contribuição foi valiosa e permanece registrada neste trabalho, ao qual dedicamos, com carinho, respeito e gratidão este trabalho à sua memória.



SUMÁRIO

1. EXPERIMENTOS FLORESTAIS. PRIMEIRO CURSO.....15

1.1. POR QUE UM PRIMEIRO CURSO?..... 15
1.2. POR QUE CONDUZIR EXPERIMENTOS?..... 16
1.3. POR QUE UTILIZAR A LINGUAGEM DE PROGRAMAÇÃO JULIA NA TAREFA DE ANÁLISE DE DADOS? 17
1.4. ASPECTOS BÁSICOS DA LINGUAGEM 18
1.5. DISPONIBILIDADE DOS ARQUIVOS E UTILIZAÇÃO DO JULIA..... 23

2. MODELO GENÉRICO PARA EXPERIMENTOS FLORESTAIS..... 28

2.1. BASES PARA TESTAR HIPÓTESES ESTATÍSTICAS EM EXPERIMENTOS FLORESTAIS – CASO PARA RAZÃO DE VARIÂNCIAS..... 31

3. TERMOS BÁSICOS EM ESTATÍSTICA 38

3.1. EXERCÍCIOS PROPOSTOS PARA SEREM RESOLVIDOS UTILIZANDO A LINGUAGEM JULIA 40

4. DISTRIBUIÇÕES DE PROBABILIDADES BÁSICAS 59

4.1. DISTRIBUIÇÃO NORMAL..... 59
4.2. DISTRIBUIÇÃO NORMAL PADRONIZADA..... 62
4.3. DISTRIBUIÇÃO *t* DE STUDENT 68

5. ESTATÍSTICAS DESCRITIVAS 70

5.1. ESTATÍSTICA DESCRITIVA 70
5.2. SIMETRIA DE UM CONJUNTO DE DADOS 84
5.3. SIMETRIA DAS MÉDIAS AMOSTRAIS..... 87
5.4. RESUMO GRÁFICO DE DADOS..... 91

6. TESTES DE HIPÓTESES..... 99

6.1. TESTE DE HIPÓTESES ESTATÍSTICAS E NÍVEL DE SIGNIFICÂNCIA 103
6.2. TESTES PARAMÉTRICOS PARA A MÉDIA DE UMA AMOSTRA..... 110
6.3. TESTES T PARA DUAS AMOSTRAS INDEPENDENTES..... 113
6.4. TESTE T PARA DUAS AMOSTRAS PAREADAS..... 119

7. TESTE SOBRE UMA PROPORÇAO 124

8. TESTE DE QUI-QUADRADO 126

8.1. TESTE DE QUI-QUADRADO PARA AJUSTE 127
8.2. TESTE DE QUI-QUADRADO PARA ASSOCIAÇÃO 130
8.3. TESTE PARA A GRANDEZA DO DESVIO-PADRÃO 132

9. COMPARAÇÃO DE DUAS VARIÂNCIAS – A DISTRIBUIÇÃO F 135

10. MEDIDAS DE ASSOCIAÇÃO 138

10.1. COVARIÂNCIA 138
10.2. CORRELAÇÃO 139
10.3. INTERVALO DE CONFIANÇA E TESTE DE HIPÓTESE PARA A CORRELAÇÃO 145
10.4. TESTE DOS COEFICIENTES DE CORRELAÇÃO 147

11. TERMOS E CONCEITOS EM DELINEAMENTOS EXPERIMENTAIS 152

11.1. TERMOS E CONCEITOS 152

12. PRINCÍPIOS BÁSICOS DA EXPERIMENTAÇÃO 155

13. BASES PARA TESTE DE HIPÓTESES E ANÁLISE DE VARIÂNCIA 156

13.1 CONDICIONANTES PARA A ANÁLISE DE VARIÂNCIA 159
13.2. AVALIAÇÃO DA HOMOGENEIDADE DAS VARIÂNCIAS DOS TRATAMENTOS 159

14. DELINEAMENTO INTEIRAMENTE CASUALIZADO 162

15. DELINEAMENTO EM BLOCOS AO ACASO 172

16. DELINEAMENTO EM QUADRADO LATINO 183

17. DELINEAMENTO EM ENSAIOS FATORIAIS 195

18. DELINEAMENTO EM PARCELAS SUBDIVIDIDAS 216

19. TESTES DE COMPARAÇÕES MÚLTIPLAS DE MÉDIAS 239

19.1. TESTES DE COMPARAÇÃO DE MÉDIAS 239
19.2. CONTRASTES PLANEJADOS: A PRIORI 240
19.3. CONTRASTES A POSTERIORI (POST HOC) 240
19.4. TESTE DMS 241
19.5. TESTE DE BONFERONI 245
19.6. TESTE DE TUKEY 247

19.7. TESTE DE DUNCAN 248
19.8. TESTE DE SCHEFFÉ 250
19.9. TESTE DE SCOTT-KNOTT 251
19.10. TESTE DE DUNNETT 255

20. ANÁLISE DE REGRESSÃO 257

20.1. FUNÇÕES 257
20.2. REGRESSÃO LINEAR SIMPLES 260
20.3. ESTATÍSTICAS DE AJUSTES DA EQUAÇÃO 265
20.4. REGRESSÃO PELA ORIGEM 279
20.5. AJUSTE DE UMA REGRESSÃO LINEAR SIMPLES POR MATRIZES 283
20.6. REGRESSÃO LINEAR MÚLTIPLA 289
20.7. AJUSTE DE MODELOS LINEARIZÁVEIS 293
20.8. REGRESSÃO PONDERADA 301
20.9. DETERMINAÇÃO DE EQUAÇÕES DE TENDÊNCIA 304

21. ASPECTOS NA CONDUÇÃO DE EXPERIMENTOS 305

REFERÊNCIAS 313

LISTA DE FIGURAS

FIGURA 1.1. Passo 1 para o uso de REPL no ambiente Visual Studio Code. Depois de instalado o Visual Studio (https://code.visualstudio.com), clicar em Extensions, digite julia na caixa superior esquerda e selecione a primeira opção (Julia). 18

FIGURA 1.2. Passo 2 para o uso de REPL no ambiente Visual Studio Code. O Julia está instalado. 19

FIGURA 1.3. Passo 3 para o uso de REPL no ambiente Visual Studio Code. Selecione View e Command Palette. 19

FIGURA 1.4. Passo 4 para o uso de REPL no ambiente Visual Studio Code. Digite Julia e selecione Julia: Start REPL 20

FIGURA 1.5. Passo 5 para o uso de REPL no ambiente Visual Studio Code. Para iniciar um novo arquivo em Julia, selecione File, New File. Para salvar o arquvo selevione File, Save As com o nome do arquivo que você deseja - Resumo_rlm. O arquivo é um bloco de notas. 20

FIGURA 1.6. Passo 6 para o uso de REPL no ambiente Visual Studio Code. Digite os comandos do Julia. 21

FIGURA 1.7. Passo 7 para o uso de REPL no ambiente Visual Studio Code. Resultado dos dados importados. 22

FIGURA 1.8. Passo 8 para o uso de REPL no ambiente Visual Studio Code. Para executar os comandos escritos no arquivo em Julia, selecione os comandos digitados (ou posione o cursor na linha desejada), selecione View, Command Pallete e digite o seguinte comando na caixa superior - Julia: Send Current Line or Selection to REPL. Após isso, os demais comandos podem ser executados por meio das teclas de atalho Ctrl + Enter. 22

FIGURA 1.9. Acesso dos arquivos de dados e funções ao Google Colab. 23

FIGURA 1.10. Acesso ao GitHub. 24

FIGURA 1.11. Pesquisa no GitHub: Julia e Experimentos Florestais. 24

FIGURA 1.12. Acesso ao repositório dos arquivos de dados e funções ao GitHub. 25

FIGURA 1.13. Acesso dos arquivos de dados e funções ao GitHub. 25

FIGURA 1.14. Baixar função e arquivo de dados. 26

FIGURA 1.15. Acessar a função e arquivo de dados via VSCODE. 27

FIGURA 2.1. Processo de teste de hipótese 30

FIGURA 2.2. Modelo amostral 30

FIGURA 2.3. Plano experimental 31

FIGURA 2.4. Curva da distribuição em função de 3 doses de adubo NPK 33

FIGURA 2.5. Curva da distribuição em função de 3 doses de adubo NPK 33

FIGURA 2.6. Curva da distribuição em função de 3 doses de adubo NPK com erro experimental 36

FIGURA 2.7. Tabela de tratamento 36

FIGURA 2.8. Curva da distribuição F 36

FIGURA 3.1. Classificação do desenvolvimento do sistema radicular de mudas. 40

FIGURA 3.2. Classificação da severidade da doença. 41

FIGURA 3.3. Folha com aspecto de deficiência nutricional de nitrogênio. 42

FIGURA 3.4. Folha com aspecto de deficiência nutricional de fósforo. 42

FIGURA 3.5. Folha com aspecto de deficiência nutricional de cálcio. 42

FIGURA 3.6. a) Folha com aspecto de deficiência nutricional de magnésio. b) Folha com aspecto de deficiência nutricional de potássio. 43

FIGURA 3.7. Padrões exigidos da qualidade de mudas. 44

FIGURA 3.8. Variáveis coletadas no inventário florestal de acácia-negra. 55

FIGURA 3.9. Condição da árvore: sem defeitos. 56

FIGURA 4.1 Distribuição da variável altura de mudas da espécie louro-pardo, 12 meses após a emergência e cultivadas em um viveiro (n=500). 60

FIGURA 4.2. À esquerda mudas de guapuruvu em estufa e à direita mudas de guapuruvu em ambiente natural. 60

FIGURA 4.3. Distribuição de altura de mudas da espécie guapuruvu cultivadas por 5 meses em estufa e, também, em ambiente natural (sem proteção para as mudas). 61

FIGURA 4.4. Distribuição normal padronizada. 62

FIGURA 4.5. Mudas de acácia-negra para avaliação de área foliar. 63

FIGURA 4.6. Obtenção de amostra para avaliação de densidade básica da madeira de Araucária. 65

FIGURA 4.7. Obtenção de amostra para avaliação de densidade básica da madeira de Pau-ferro. 66

FIGURA 4.8. Propriedades da distribuição normal padronizada 67

FIGURA 4.9. Distribuição Normal e distribuição t. 69

FIGURA 5.1. Resultado das simulações. 75

FIGURA 5.2. Diâmetro à altura do peito de acácia-negra. 77

FIGURA 5.3. Intervalo de confiança para a média da amostra 80

FIGURA 5.4. Intervalo de confiança para cada uma das 100 amostras tomadas 82

FIGURA 5.5. Simetria e assimetria na distribuição dos dados. 85

FIGURA 5.6. Histograma de frequências para a variável diâmetro à altura do peito obtidas utilizando-se o pacote Gadfly. 92

FIGURA 5.7. Histograma de frequências acumuladas para a variável diâmetro à altura do peito utilizando o pacote PyPlot. 92

FIGURA 5.8. Diagrama de caixa para a variável diâmetro à altura do peito de florestas de acácia-negra com mesma idade, utilizando o pacote StatsPlots. 93

FIGURA 5.9. Comportamento da temperatura média do ar dentro de uma estufa de produção de mudas ao longo de 10 dias, utilizando o pacote Plots. 94

FIGURA 5.10. Mortalidade do Pinheiro-do-Paraná e de várias espécies de canelas observada em uma Floresta Ombrófila Mista Montana ao longo da década de monitoramento, utilizando o pacote Plots. 94

FIGURA 5.11. Número de espécie por família botânica observada em uma Floresta Ombrófila Mista Montana ao longo da década de monitoramento, utilizando o pacote Plots. 95

FIGURA 5.12. Relação entre diâmetro à altura do peito e altura das árvores de uma floresta de acácia-negra com 10 anos, utilizando o pacote Plots. 96

FIGURA 5.13. Proporção de árvores ingressas entre os grupos ecológicos Araucária, Canelas e Brancas em uma área de 3 hectares de Floresta Ombrófila Mista durante uma década, utilizando o pacote Plots. 96

FIGURA 5.14. Correlação entre as variáveis diâmetro à altura do peito, altura, comprimento da copa, volume do fuste e fator de forma para árvores de acácia-negra com 10 anos, utilizando o pacote StatsPlots. 97

FIGURA 6.1. Amostra para coleta de dados de densidade média de tábuas de Araucária. 100

FIGURA 6.2. Distribuição de médias amostrais de densidade de tábuas (g.cm³) de Araucária. 102

FIGURA 6.3. Teste de hipótese t bilateral. 106

FIGURA 6.4. Teste de hipótese t unilateral direita. 107

FIGURA 6.5. Teste de hipótese t unilateral esquerda. 108

FIGURA 6.6. Medição de altura de mudas de acácia-negra. 111

FIGURA 6.7. Parcela instalada no povoamento de eucalipto. 114

FIGURA 6.8. Parcela instalada no povoamento de acácia-negra. 115

FIGURA 6.9. Medição de altura total com hipsômetro e LiDAR. 121

FIGURA 7.1. Processo de germinação da semente de louro-pardo. 125

FIGURA 8.1. Espécies florestais para recuperação de área degradada. 128

FIGURA 8.2. Colheita florestal com harvester. 133

FIGURA 10.1. Relação entre as variáveis biomassa e circunferência à altura do peito para árvores de acácia-negra. com 10 anos. 141

FIGURA 10.2. Coleta de variáveis altura, cap e biomassa do fuste da espécie florestal acácia-negra. 142

FIGURA 10.3. Obtenção da biomassa de copa e volume do fuste de árvores de acácia-negra. 148

FIGURA 11.1. Experimento para avaliar a produção em volume de madeira do fuste oito anos após a instalação de diferentes espécies florestais com delineamento em blocos ao acaso. 152

FIGURA 13.1. Área ocupada pela estatística F para testar as médias dos tratamentos. 159

FIGURA 14.1. Recipientes utilizados como tratamento para a avaliação do crescimento em altura de mudas de *Pinus taeda*. 164

FIGURA 14.2. Unidade experimental com os tratamentos testados. 165

FIGURA 14.3. Experimento para avaliar o efeito de diferentes tipos de recipientes no crescimento em altura de mudas de *Pinus taeda* em delineamento inteiramente casualizado. 165

FIGURA 14.4 Experimento para avaliar o efeito de tipos de recipientes no crescimento em altura de mudas de *Pinus taeda* em delineamento inteiramente casualizado, aos 120 dias após o plantio. 166

FIGURA 14.5. Resumo do delineamento inteiramente casualizado. 171

FIGURA 15.1. Espécies (mudas e árvores com 8 anos) utilizadas como tratamento para avaliar a produção em volume de madeira do fuste. 175

FIGURA 15.2. Unidade experimental das diferentes espécies utilizadas para avaliar a produção em volume de madeira do fuste 175

FIGURA 15.3. Experimento para avaliar a produção em volume de madeira do fuste, de diferentes espécies florestais, em delineamento com blocos ao acaso. 176

FIGURA 15.4. Unidade experimental das diferentes espécies utilizadas para avaliar a produção em volume de madeira do fuste oito anos após a instalação do experimento. 176

FIGURA 15.5. Experimento em campo para avaliar a produção em volume de madeira do fuste de diferentes espécies florestais em delineamento com blocos ao acaso oito anos após a instalação. 178

FIGURA 15.6. Resumo do Delineamento em blocos ao acaso. 182

FIGURA 16.1. Substratos utilizados como tratamentos para avaliar o crescimento de mudas de *Araucaria angustifolia*. 185

FIGURA 16.2. Árvores de *Araucaria angustifolia* para seleção de sementes a serem submetidas à avaliação de efeito de tipos de substratos no crescimento das mudas, para 5 idades e tamanhos de sementes. 186

FIGURA 16.3. Experimento para avaliar o efeito de tipos de substratos no crescimento de mudas de *Araucaria angustifolia* com delineamento em quadrado latino. 187

FIGURA 16.4. Experimento para avaliar o efeito de tipos de substratos no crescimento de mudas de *Araucaria angustifolia* com delineamento em quadrado latino um ano após a instalação. 188

FIGURA 16.5. Resumo do Delineamento em quadrado latino. 194

FIGURA 17.1. Níveis testados para o preparo de solo (fator a) para avaliar o crescimento inicial de *Eucalyptus urophylla*. 199

FIGURA 17.2. Níveis testados para os métodos de adubação (fator b) para avaliar o crescimento inicial de E*ucalyptus urophylla*. 199

FIGURA 17.3. Tratamentos resultantes da combinação de diferentes níveis de preparo do solo e métodos de adubação testados para avaliar o crescimento inicial de *Eucalyptus urophylla*. 200

FIGURA 17.4. Unidade experimental com mudas de *Eucalyptus urophylla* utilizadas para avaliar o crescimento inicial sob efeito de diferentes tipos de preparo de solo e métodos de adubação. 201

FIGURA 17.5. Experimento para avaliar o efeito de diferentes tipos de preparo de solo e métodos de adubação no crescimento inicial de mudas de *Eucalyptus urophylla*. 201

FIGURA 17.6. Experimento para avaliar o efeito de diferentes tipos de preparo de solo e métodos de adubação no crescimento inicial de mudas de *Eucalyptus urophylla* - bloco 1 202

FIGURA 17.7. Experimento para avaliar o efeito de diferentes tipos de preparo de solo e métodos de adubação no crescimento inicial de mudas de *Eucalyptus urophylla* - bloco 2 202

FIGURA 17.8. Experimento para avaliar o efeito de diferentes tipos de preparo de solo e métodos de adubação no crescimento inicial de mudas de *Eucalyptus urophylla* - bloco 3 203

FIGURA 17.9. Experimento para avaliar o efeito de diferentes tipos de preparo de solo e métodos de adubação no crescimento inicial de mudas de *Eucalyptus urophylla* - bloco 4 203

FIGURA 17.10. Experimento para avaliar o efeito de diferentes tipos de preparo de solo e métodos de adubação no crescimento inicial de mudas de *Eucalyptus urophylla* - bloco 5 204

FIGURA 17.11. Unidade experimental com mudas de *Eucalyptus urophylla* utilizadas para avaliar o crescimento inicial sob efeito de diferentes tipos de preparo de solo e métodos de adubação após a instalação do experimento. 204

FIGURA 17.12. Experimento para avaliar o efeito de diferentes tipos de preparo de solo e métodos de adubação no crescimento inicial de mudas de *Eucalyptus urophylla* um ano após a instalação com delineamento de blocos ao acaso com arranjo fatorial..... 205

FIGURA 17.13. Resumo do Delineamento em ensaios fatoriais. 215

FIGURA 18.1. Unidade amostral com mudas de *Tectona grandis* utilizadas para avaliar o crescimento inicial sob efeito de diferentes tipos de preparo de solo e métodos de adubação..... 219

FIGURA 18.2. Tratamentos testados para avaliar o crescimento inicial da espécie florestal *Tectona grandis* sob o efeito de diferentes tipos de preparo de solo e métodos de adubação..... 220

FIGURA 18.3. Tratamentos testados para avaliar o crescimento inicial da espécie florestal *Tectona grandis* sob o efeito de diferentes tipos de preparo de solo e métodos de adubação..... 222

FIGURA 18.4. Experimento instalado em campo para avaliar o crescimento inicial da espécie florestal *Tectona grandis* sob o efeito de diferentes tipos de preparo de solo e métodos de adubação com delineamento de blocos ao acaso em esquema de parcelas subdivididas. 222

FIGURA 18.5. Experimento instalado em campo para avaliar o crescimento inicial da espécie florestal *Tectona grandis* sob o efeito de diferentes tipos de preparo de solo e métodos de adubação com delineamento de blocos ao acaso em esquema de parcelas subdivididas - bloco 1. 223

FIGURA 18.6. Experimento instalado em campo para avaliar o crescimento inicial da espécie florestal *Tectona grandis* sob o efeito de diferentes tipos de preparo de solo e métodos de adubação com delineamento de blocos ao acaso em esquema de parcelas subdivididas - bloco 2. 223

FIGURA 18.7. Experimento instalado em campo para avaliar o crescimento inicial da espécie florestal *Tectona grandis* sob o efeito de diferentes tipos de preparo de solo e métodos de adubação com delineamento de blocos ao acaso em esquema de parcelas subdivididas - bloco 3. 224

FIGURA 18.8. Experimento instalado em campo para avaliar o crescimento inicial da espécie florestal *Tectona grandis* sob o efeito de diferentes tipos de preparo de solo e métodos de adubação com delineamento de blocos ao acaso em esquema de parcelas subdivididas - bloco 4. 224

FIGURA 18.9. Experimento instalado em campo para avaliar o crescimento inicial da espécie florestal *Tectona grandis* sob o efeito de diferentes tipos de preparo de solo e métodos de adubação com delineamento de blocos ao acaso em esquema de parcelas subdivididas - bloco 5. 225

FIGURA 18.10. Experimento instalado em campo para avaliar o crescimento inicial da espécie florestal *Tectona grandis* sob o efeito de diferentes tipos de preparo de solo e métodos de adubação com delineamento de blocos ao acaso em esquema de parcelas subdivididas - bloco 6. 225

FIGURA 18.11. Experimento instalado em campo para avaliar o crescimento inicial da espécie florestal *Tectona grandis* sob o efeito de diferentes tipos de preparo de solo e métodos de adubação com delineamento de blocos ao acaso em esquema de parcelas subdivididas 2 anos após a instalação. 226

FIGURA 18.12. Parcelas selecionadas para avaliar o efeito do tipo e intensidade de desbaste sobre a produção da espécie *Pinus caribaea*..... 233

FIGURA 18.13. Parcela instalada em campo para avaliar o efeito do tipo e intensidade de desbaste sobre a produção em área basal da espécie *Pinus caribaea* – fator a: desbaste sistemático 20%..... 234

FIGURA 18.14. Parcela instalada em campo para avaliar o efeito do tipo e intensidade de desbaste sobre a produção em área basal da espécie *Pinus caribaea* – fator a: desbaste sistemático 25%..... 234

FIGURA 18.15. Parcela instalada em campo para avaliar o efeito do tipo e intensidade de desbaste sobre a produção em área basal da espécie *Pinus caribaea* – fator a: desbaste sistemático 33%..... 235

FIGURA 18.16. Bloco instalado em campo para avaliar o efeito do tipo e intensidade de desbaste sobre a produção em área basal da espécie *Pinus caribaea* – bloco de área basal média de 14,20 $m^2.ha^{-1}$ com os 9 tratamentos testados..... 235
FIGURA 18.17. Resumo do Delineamento em parcelas subdivididas. 238
FIGURA 19.1. Aplicação dos tratamentos nos corpos de prova. 242
FIGURA 20.1. Modelos volumétricos ajustados. 290
FIGURA 21.1. Desenhando o experimento. 308
FIGURA 21.2. Fluxograma do planejamento experimental. 308
FIGURA 21.3. Passos da coleta e organização dos dados..... 309
FIGURA 21.4. Análise dos dados. 309
FIGURA 21.5. Composição do título do experimento. 310
FIGURA 21.6. Ciclo de pesquisa..... 312

LISTA DE TABELAS

TABELA 3.1. Classificação do desenvolvimento do sistema radicular de mudas. 41

TABELA 3.2. Variáveis e padrões estabelecidos para avaliação da qualidade de mudas de eucalipto. 44

TABELA 3.3. Prática: variáveis e padrões estabelecidos para avaliação da qualidade de mudas de eucalipto. 45

TABELA 3.4. Variáveis medidas no inventário florestal de acácia-negra..... 55

TABELA 6.1. Dados amostrais de altura de mudas (em mm) de acácia-negra. 111

TABELA 6.2. Dados experimentais de volume de madeira ($m^3.ha^{-1}$) de eucalipto (X1) e acácia-negra (X2). 115

TABELA 6.3. Dados experimentais de altura (m) mensurada com hipsômetro (X_1) e por LiDAR (X_2) em uma floresta de Pinus. 120

TABELA 8.1. Número de indivíduos por espécie observados em uma área de 5.000 m^2 no estado do Paraná. 128

TABELA 8.2. Número de acidentes observado em 100 viveiros florestais durante um período de 10 anos em função do turno de trabalho e idade dos colaboradores. 131

TABELA 10.1. Dados de produção de biomassa de árvores e radiação fotossinteticamente ativa interceptada por elas. 139

TABELA 13.1. Tabela resumo da análise de variância. 156

TABELA 14.1. Esquema para análise de variância para o delineamento inteiramente casualizado..... 164

TABELA 14.2. Dados de altura média de mudas de *Pinus taeda* (em cm) obtidas em um experimento para avaliar o efeito de recipientes no seu crescimento em delineamento inteiramente casualizado (continua). 166

TABELA 14.3. Análise de variância aplicada para avaliar os efeitos de tipos de recipiente sobre o crescimento da altura (em cm) de mudas de *Pinus taeda*. 167

TABELA 14.4. Sumarização dos resultados obtidos pelo teste de Tukey, na comparação de médias dos diferentes tipos de recipientes sobre o crescimento em altura (em cm) de mudas de *Pinus taeda*..... 170

TABELA 15.1. Esquema para análise de variância para o delineamento em blocos ao acaso. 174

TABELA 15.2. Dados de volume de madeira (em $m^3.ha^{-1}$) obtidos em um experimento para avaliar a produção de espécies florestais em delineamento com blocos ao acaso. 177

TABELA 15.3. Resultados da análise de variância aplicada para avaliar a produção em volume de madeira (em $m^3.ha^{-1}$) do fuste de diferentes espécies florestais em delineamento em blocos ao acaso. 180

TABELA 15.4. Resultado do teste de Tukey para comparação das médias das diferentes espécies, na produção em volume de madeira (em m^3 ha^{-1}) do fuste de diferentes espécies florestais..... 182

TABELA 16.1. Esquema para análise de variância para o delineamento em quadrado latino. 185

TABELA 16.2. Resultados experimentais obtidos para a altura (em cm) e diâmetro de colo (em mm) da espécie *Araucaria angustifolia*, sob efeito de cinco diferentes substratos. 189

TABELA 16.3. Resultados da análise de variância realizada para testar o efeito de cinco substratos no crescimento em altura (em cm) de mudas de *Araucaria angustifolia*. 190

TABELA 16.4. Resultados do teste de Tukey para comparação das médias dos efeitos do substratos no crescimento em altura (em cm) de mudas de *Araucaria angustifolia*. 193

TABELA 17.1. Esquema para análise de variância um experimento fatorial a·b conduzido em delineamento em delineamento inteiramento casualizado. 198

TABELA 17.2. Resultados obtidos para a altura (em m), diâmetro de copa (em m) e diâmetro à altura do peito (em cm) da espécie *Eucalyptus urophylla* sob efeito de diferentes tipos de preparo de solo e métodos de adubação (continua). 205

TABELA 17.3. Resultado da análise de variância para avaliar o crescimento em altura (em m) da espécie florestal *Eucalyptus urophylla* sob o efeito de diferentes tipos de preparo de solo e métodos de adubação. 209

TABELA 17.4. Resultados do teste de Tukey na comparação de médias para avaliar o crescimento inicial *Eucalyptus urophylla* sob o efeito de diferentes tipos de preparo de solo e métodos de adubação. 214

TABELA 18.1. Esquema para análise de variância um experimento conduzido em delineamento em blocos ao acaso com parcelas subdivididas.....218

TABELA 18.2. Resultados experimentais obtidos para a altura (em m), diâmetro de copa (em m) e diâmetro à altura do peito (em cm) da espécie *Tectona grandis* submetidas a diferentes tipos de preparo de solo e métodos de adubação. 221

TABELA 18.3. Resultados da análise de variância para avaliar o crescimento em diâmetro à altura do peito (em cm) da espécie florestal *Tectona grandis* sob o efeito de diferentes tipos de preparo de solo e métodos de adubação. 230

TABELA 18.4. Resultados do teste de Tukey para comparação das médias dos diferentes tipos de preparo do solo (fator A) para o crescimento em diâmetro à altura do peito (em cm) da espécie florestal *Tectona grandis*..... 231

TABELA 18.5. Resultados obtidos para a área basal ($m^2\ ha^{-1}$) da espécie *Pinus caribaea* aos sete anos de idade, obtidos nas parcelas, nas quais foram efetuados os tratamentos para testar o efeito e a intensidade do desbaste sobre a produção em área basal. 233

TABELA 18.6. Resultados obtidos para a área basal (G, em $m^2.ha^{-1}$) da espécie *Pinus caribaea* aos 14 anos de idade para testar o efeito do tipo e intensidade do desbaste sete anos após a instalação do experimento (continua). 236

TABELA 20.1. Estatísticas para avaliação da qualidade do ajuste. 291

1 EXPERIMENTOS FLORESTAIS. PRIMEIRO CURSO.

Uma tarefa cotidiana do Engenheiro Florestal é coletar dados, organizá-los em tabelas, descrevê-los detalhadamente para analisá-los, interpretá-los e utilizá-los na tomada de decisão. A tarefa de coletar, organizar e descrever os dados é um campo da estatística descritiva e a de analisar, interpretar e realizar inferências e conclusões da estatística indutiva. Nesse contexto, métodos computacionais para realização de cálculos e utilização de gráficos passam a auxiliar a tarefa da análise dos dados.

O Engenheiro Florestal deve saber analisar e interpretar dados utilizando a estatística para resolver problemas comuns, possibilitando efetuar, por exemplo, diagnósticos para uma empresa e formular soluções apropriadas. Deve, ainda, planejar experimentos e analisar os dados obtidos quer em trabalhos para a academia quer para fins de pesquisas científicas.

Assim, o propósito é propiciar métodos estatísticos para analisar dados e obter conclusões, sendo direcionado para os Engenheiros Florestais que estão em primeiro contato com a área de estatística. Pretende também apresentar aos estudantes como analisar dados, por meio de programas estatísticos.

Os cálculos e os gráficos serão efetuados utilizando a linguagem Julia. A decisão em utilizá-la foi baseada no seu desempenho, flexibilidade e simplicidade para implementação dos cálculos e construção dos gráficos. Os códigos apresentados são simples e têm como objetivo realizar as tarefas dos exemplos propostos.

1.1 POR QUE UM PRIMEIRO CURSO?

Engenheiros Florestais conduzem experimentos em florestas nativas, em plantios florestais, em viveiros florestais e em laboratórios de tecnologia da madeira, entre outros. São aplicados testes de espaçamento (qual é o melhor espaçamento de cultivo?), fertilizantes (qual é o melhor fertilizante?), regeneração (a regeneração é influenciada por doses de adubação?), germinação (as espécies apresentam a mesma taxa de germinação?), crescimento de mudas (qual é a reposta do crescimento de mudas em diferentes recipientes?), propriedades físicas da madeira (as propriedades para clones de uma espécie são as mesmas?), entre outros.

Este texto destina-se a auxiliar os estudantes a projetar experimentos, analisá-los apropriadamente e responder as perguntas que foram propostas para a execução do experimento. Apenas serão apresentados experimentos comparativos, ou seja, quando se deseja comparar alguns tratamentos.

Tratamentos serão aplicados às unidades experimentais e nelas serão mensuradas uma ou mais variáveis-respostas. Por exemplo, pode-se comparar o crescimento da espécie florestal acácia-negra sob efeito de 4 doses de fertilizante.

A unidade experimental pode ser 20 mudas plantadas em campo com espaçamento de 3 m x 1,75 m. Os tratamentos são avaliados pelas respostas obtidas nas unidades experimentais que receberam as doses de adubação. Isso poderá revelar se há alguma diferença nas respostas entre os tratamentos, qual é o tamanho desta diferença e qual é a dose de adubação mais apropriada para propiciar a melhor taxa de crescimento.

Basicamente, um experimento é uma investigação planejada visando testar tratamentos, determinar a forma e constituição das unidades experimentais, como os tratamentos serão atribuídos às unidades e os tipos de respostas que serão obtidas nas unidades experimentais.

Os experimentos permitem estabelecer uma comparação direta entre os tratamentos. Para tanto, projeta-se experimentos para minimizar tendências e obter o erro resultante desta comparação.

1.2 POR QUE CONDUZIR EXPERIMENTOS?

Os estudos florestais podem ser categorizados em três projetos experimentais conceituais:

descritivo | **correlacional** | **manipulativo**

Nas pesquisas delineadas usando uma abordagem descritiva, são observados e descritos processos a partir de amostras. Estes são importantes para descrever o processo de crescimento de uma espécie ou, por exemplo, a estrutura de uma floresta nativa. Nesses estudos não são testadas hipóteses, porque, como o próprio termo indica, a informação útil deriva de uma estruturação descritiva das variáveis medidas. Os inventários florestais se enquadram nessa categoria de estudo.

Quando se formula pelo menos uma hipótese, a pesquisa é conduzida sobre um estudo correlacional, também denominados de estudos observacionais. Usando esta abordagem, o pesquisador coleta dados sobre o assunto em uma ampla gama de fatores ambientais, por exemplo, um estudo correlacional da ocorrência de araucárias jovens. As densidades de araucárias são quantificadas em áreas que haviam sido manejadas (com corte de árvores) e em áreas sem manejo (sem corte de árvores). Como as áreas tinham sido previamente tratadas (com corte de árvores e sem corte de árvores) sem aleatorização dos tratamentos entre áreas, a análise pode indicar apenas se houve diferenças observadas nas densidades de araucária entre os dois tipos de áreas.

A partir desse tipo de estrutura experimental, não se pode inferir que os tratamentos (com e sem manejo florestal) como a causa de diferença na densidade de araucárias, porque fatores além do tratamento poderiam ser responsáveis pelas respostas observadas devido as escolhas das áreas estudadas. Neste tipo de estudo, mede-se características específicas, mas não se controla ou se modifica os elementos a serem estudados.

A inferência sobre as relações entre tratamentos é formalizada a partir de estudos com uma abordagem sistemática. O estudo deve ser projetado de maneira a determinar se existem relações hipotéticas, por meio de uma abordagem experimental com tratamentos apropriadamente

definidos, aleatorização, repetição e controle das unidades experimentais.

Os tratamentos podem envolver mais de um tipo de sistematização e comparados entre si ou com um tratamento de controle (testemunha). No experimento, os resultados de áreas tratadas da mesma forma (repetição) são comparados com os de áreas tratadas de maneira diferente (tratamentos). Se os procedimentos experimentais apropriados forem aplicados e os dados diferirem entre os tratamentos, pode-se fazer inferências que os tratamentos são diferentes. No entanto, se todos os tratamentos apresentarem resultados semelhantes, pode-se concluir que eles não manifestaram efeitos diferenciais expressos nas variáveis medidas. Assim, em um experimento aplica-se tratamentos e se observa seus efeitos sobre as variáveis a serem pesquisadas.

Embora os experimentos sistemáticos tenham algumas vantagens, os estudos observacionais também são úteis e podem produzir resultados importantes. Note que em um estudo observacional também têm tratamentos, unidades e respostas, entretanto, não se tem o controle das unidades experimentais, porque estas foram obtidas ao acaso.

As unidades experimentais são definidas cuidadosamente e, muitas vezes, para seu controle, blocos são estabelecidos de modo que as unidades dentro de cada bloco sejam o mais semelhantes possível, o que permite a redução do erro do erro experimental.

Quando é utilizada uma abordagem científica para se avaliar efeitos, quaisquer impactos ou benefícios que possam ser medidos durante o projeto experimental podem ser atribuídos às práticas de manejo florestal, quando estas, por exemplo, representam os tratamentos, dado que a maioria do público deseja identificar a técnica manejo florestal que deve ser aplicada para maximizar o uso dos recursos.

Muitos modelos estatísticos estão disponíveis para analisar dados experimentais. Neste contexto, apenas serão apresentados os modelos básicos, para ilustrar possíveis abordagens estatísticas.

1.3 POR QUE UTILIZAR A LINGUAGEM DE PROGRAMAÇÃO JULIA NA TAREFA DE ANÁLISE DE DADOS?

Diversas linguagens de programação foram desenvolvidas para computação numérica e têm sido aplicadas em pesquisas em instituições públicas ou privadas, bem como no ensino de disciplinas relacionadas com cálculo numérico e análise de dados. Em 2012 o MIT (*Massachusetts Institute of Tecnology*) lançou uma nova linguagem de programação: JULIA, de livre acesso e código aberto, que tem como principal proposta ser uma alternativa às outras linguagens de computação científica já existentes. Várias universidades já estão utilizando-a em suas disciplinas e existem conferências internacionais anuais sobre a linguagem (*JuliaCon*).

A linguagem SAS foi predominante por muitos anos e, recentemente, uma versão para acadêmicos e universidades foi disponibilizada (*SAS OnDemand for Academics*). Alguns engenheiros optaram por desenvolver suas tarefas em outras linguagens de programação, como, por exemplo, Python e Matlab. Recentemente, a linguagem R passou a ser um consenso na comunidade acadêmica que trabalha com análise de dados.

Dada a consolidação de linguagens de programação para a tarefa de análise de dados, por que utilizar a recente linguagem de programação Julia?

Entre outros objetivos, Julia foi criada para computação numérica científica. A tarefa de fazer contas em Julia é simples!

Uma importante obrigação do Engenheiro Florestal atual é efetuar cálculos em computador, porque esta será a situação encontrada na prática. A intenção de apresentação dos cálculos com a linguagem Julia é despertar o interesse dos alunos para criar novas habilidades e resolver seus próprios problemas de cálculos em estatística, além de representações gráficas.

Nós identificamos na linguagem Julia uma oportunidade para isso. A ideia é desenvolver nos alunos a habilidade de uso de uma linguagem estatística para a execução de tarefas de análise de diferentes tipos de dados existentes nas empresas ou na academia.

1.4 ASPECTOS BÁSICOS DA LINGUAGEM

A linguagem de programção Julia está disponível em *https://julialang.org/*. Os cálculos serão apresentados na versão REPL (*Read Evaluate Print Loop*) *command line interface*.

Alternativamente, outros ambientes podem ser considerados, por exemplo, a integração de Julia em *Visual Studio Code* disponível em *https://code.visualstudio.com/*. A sequência de Figuras de 1.1 a 1.8 demonstra os passos para o uso de REPL no ambiente *Visual Studio Code*.

FIGURA 1.1. Passo 1 para o uso de REPL no ambiente Visual Studio Code. Depois de instalado o Visual Studio (https://code.visualstudio.com), clicar em Extensions, digite julia na caixa superior esquerda e selecione a primeira opção (Julia).

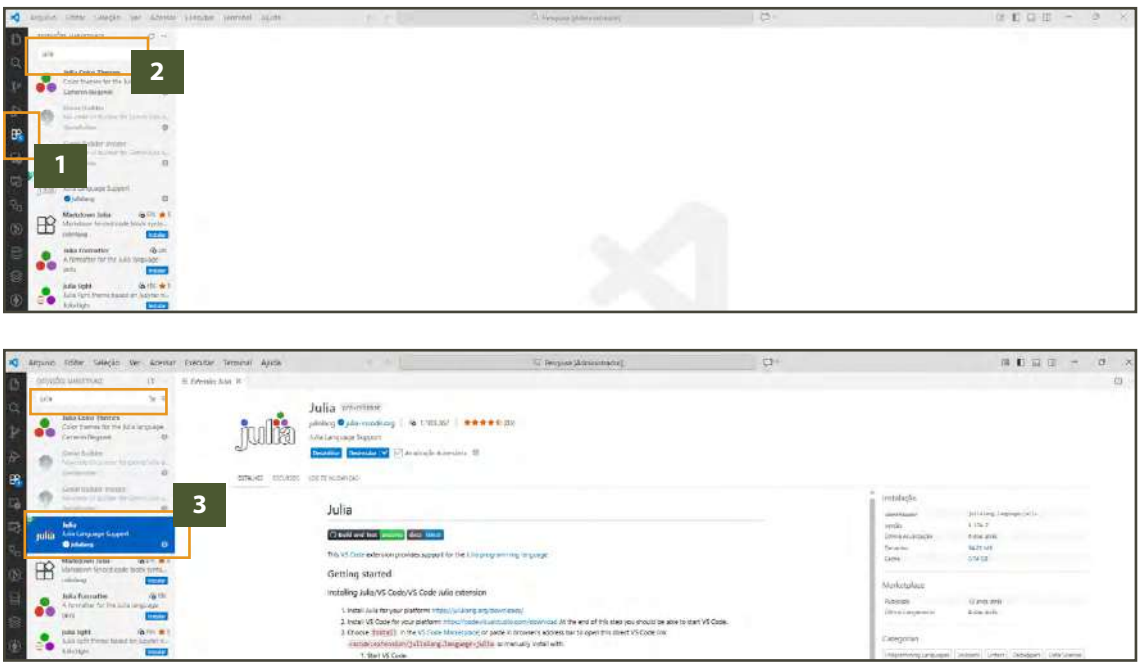


FIGURA 1.2. Passo 2 para o uso de REPL no ambiente Visual Studio Code. O Julia está instalado.

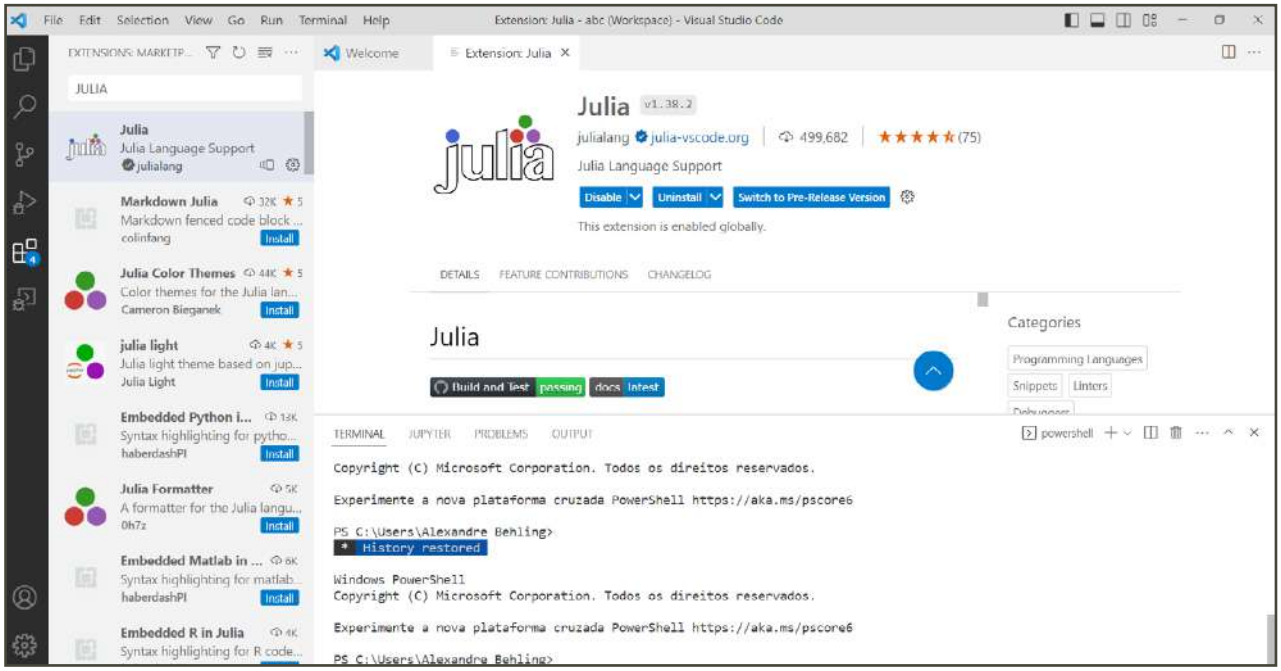


FIGURA 1.3. Passo 3 para o uso de REPL no ambiente Visual Studio Code. Selecione View e Command Palette.

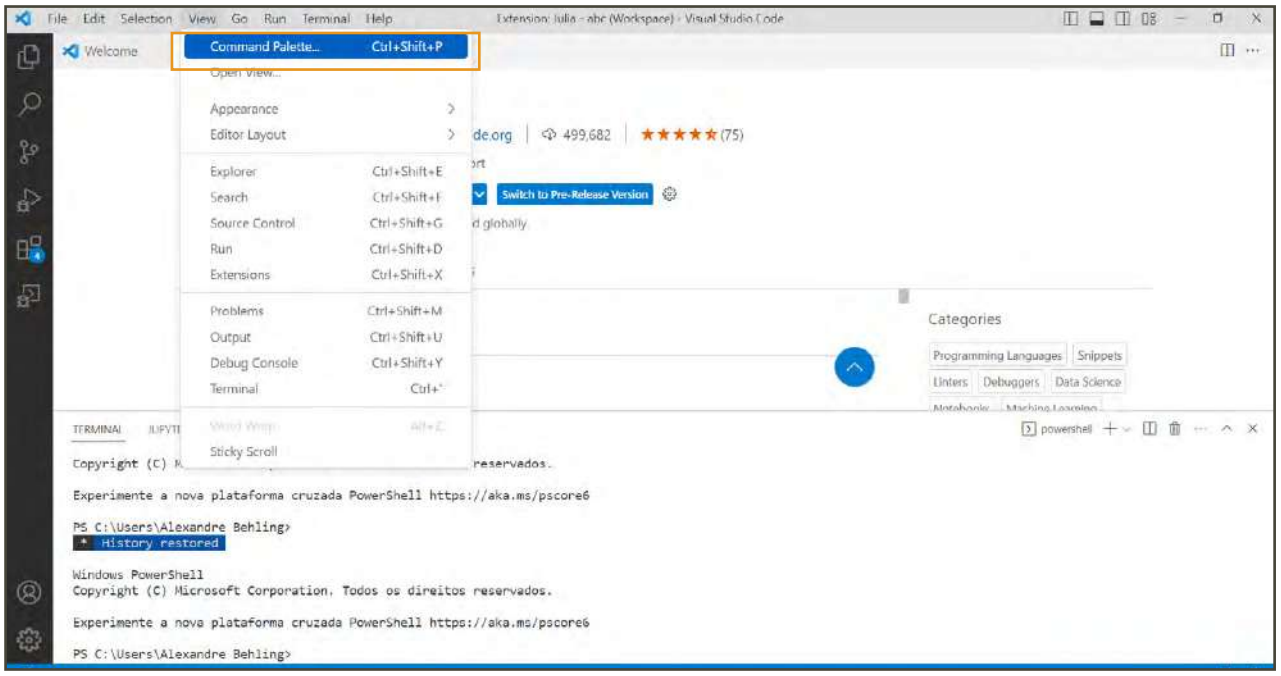


FIGURA 1.4. Passo 4 para o uso de REPL no ambiente Visual Studio Code. Digite Julia e selecione Julia: Start REPL.

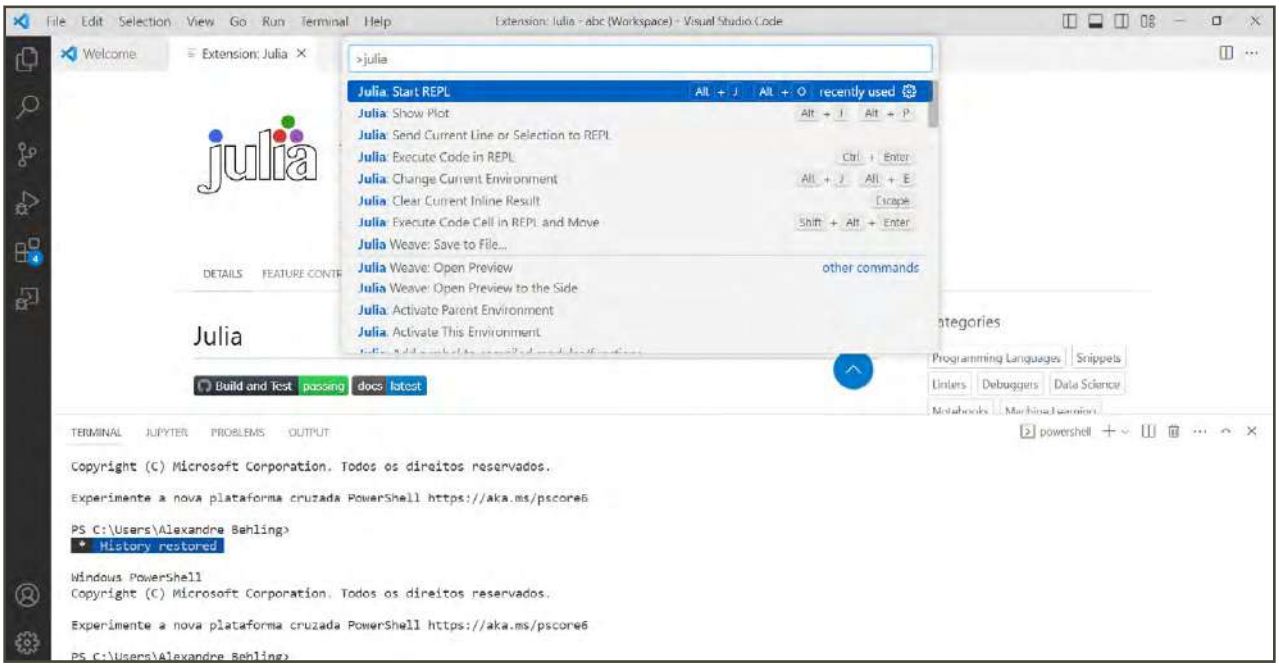


FIGURA 1.5. Passo 5 para o uso de REPL no ambiente Visual Studio Code. Para iniciar um novo arquivo em Julia, selecione File, New File. Para salvar o arquvo selevione File, Save As com o nome do arquivo que você deseja - Resumo_rlm. O arquivo é um bloco de notas.

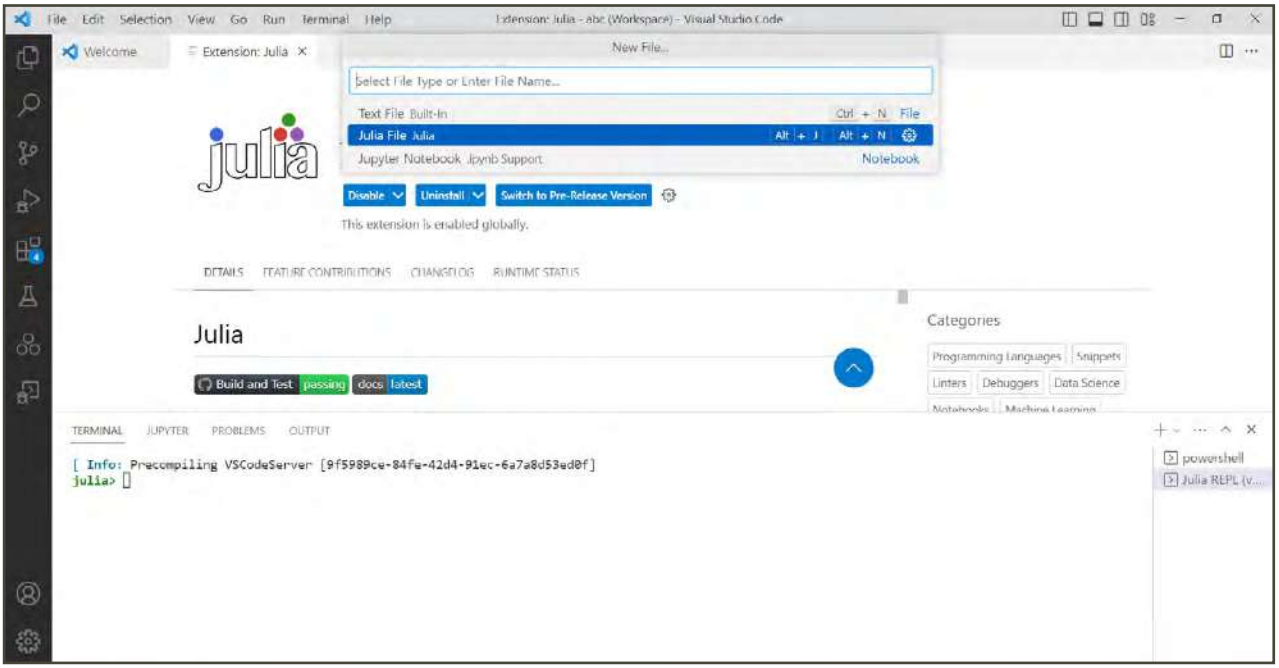
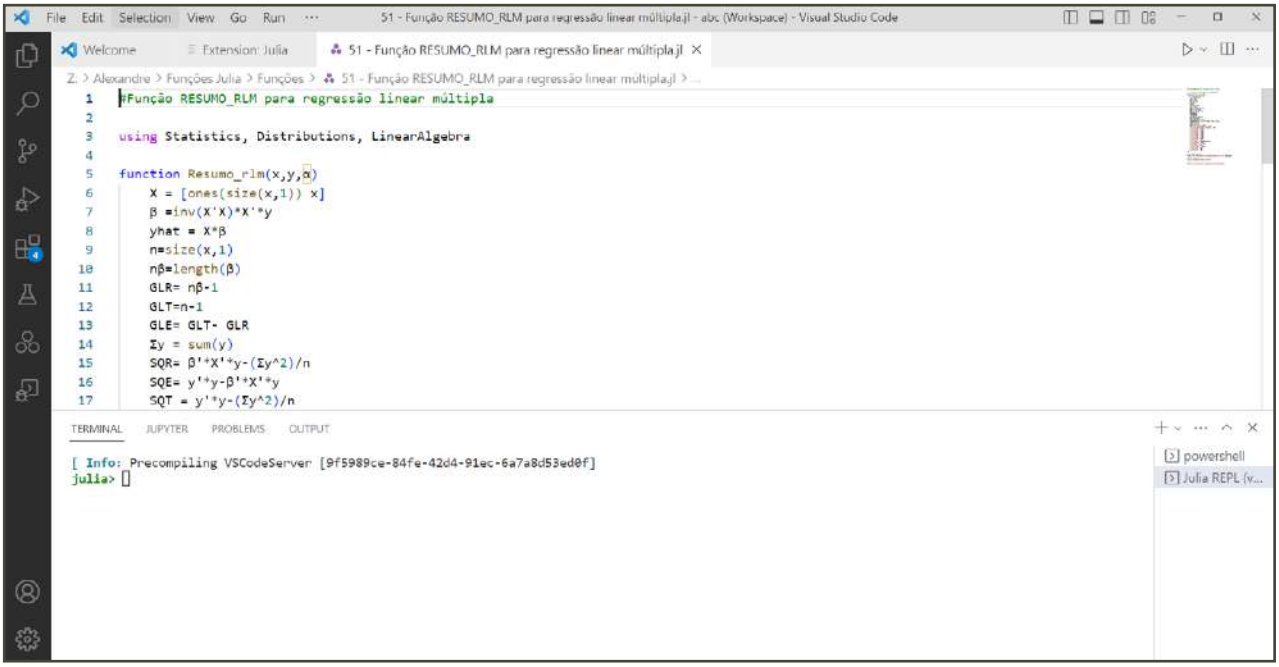


FIGURA 1.6. Passo 6 para o uso de REPL no ambiente Visual Studio Code. Digite os comandos do Julia.



Todas as funções que serão apresentadas nos exemplos estarão disponibilizadas em arquivos de bloco de notas, uma vez que é possível utilizá-las neste ambiente de programação. Os dados utilizados também estarão disponibilizados desta forma ou em formato .CSV. Ao iniciar o processamento via REPL, o primeiro passo é instalar os pacotes de análise que serão utilizados para o desenvolvimento dos exercícios com os respectivos comandos, como por exemplo:

```
julia> using Pkg
julia>Pkg.add("NomeDoPacote")
```

Para utilizar um package e, por exemplo, importar dados armazenados em um arquivo .CSV (Figura 1.7), os seguintes comandos podem ser utilizados (# indica uma linha de comentário):

ARQUIVO: 1 - LEITURA DE ARQUIVO CSV

```
#Leitura de um arquivo de dados em formato CSV

using DataFrames, CSV
X=CSV.read("Z:/Alexandre/Julia/Exemplos/mudas.csv", DataFrame)

#Ativação dos pacotes: using DataFrames, CSV
# X variável onde serão armazenados os dados
#Para leitura de uma tabela de dados em formato .CSV: CSV.read
#Local em que o arquivo está armazenado: Z:/Alexandre/Julia/Exemplos
#Nome do arquivo: mudas
#Formato do arquivo: .csv
#DataFrame: armazenamento dos dados em uma tabela.
```

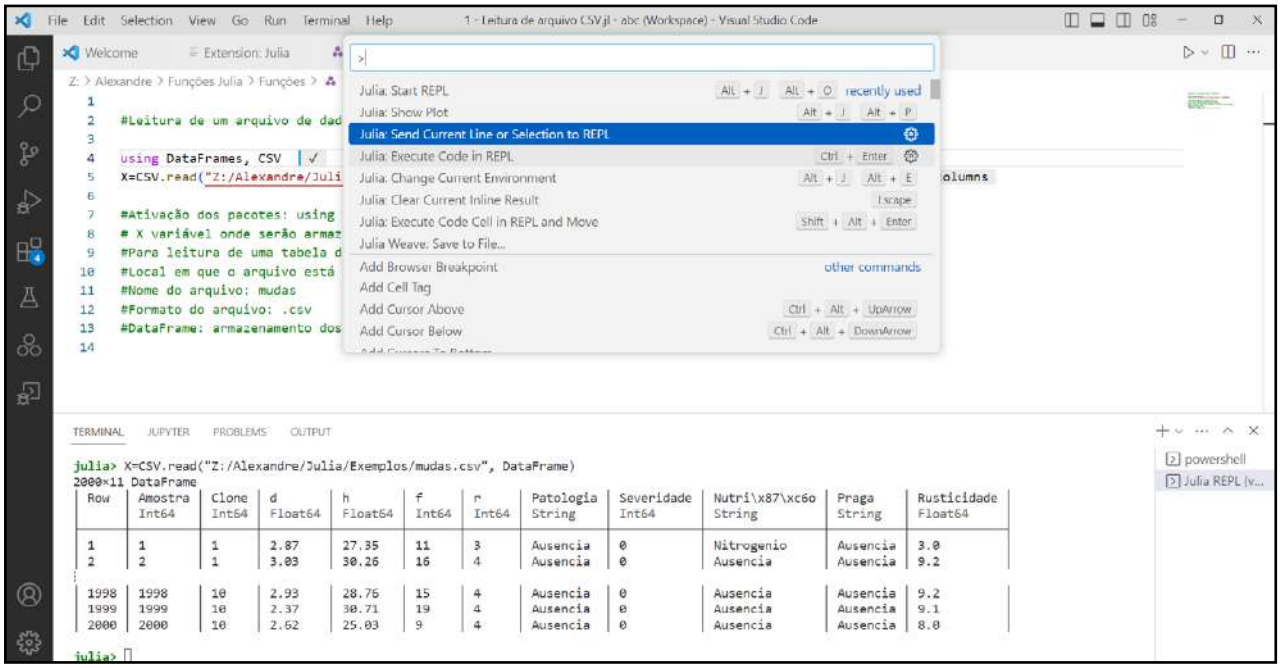
FIGURA 1.7. Passo 7 para o uso de REPL no ambiente Visual Studio Code. Resultado dos dados importados.



Para execução dos exercícios, vários pacotes deverão ser instalados. Não se preocupe no início, pois pode ser que seja difícil nesta fase lidar com a linguagem para operacionalizar os cálculos, construir gráficos e tabelas. No decorrer da solução dos exercícios, isso tornar-se-á mais compreensível e agradável e você terá condições de dominar uma linguagem de programação moderna e eficiente para resolver problemas de estatística aplicada à Engenharia Florestal. Pode ser que isto será provavelmente até muito estimulante.

Os comandos anteriores também podem ser executados no próprio editor do Visual Studio Code e podem ser salvos, por exemplo, em um bloco de notas. Todos os comandos, armazenados neste arquivo, poderão ser processados utilizando a linguagem Julia. Para tanto, selecione as linhas de comando e execute o código (Figura 1.8).

FIGURA 1.8. Passo 8 para o uso de REPL no ambiente Visual Studio Code. Para executar os comandos escritos no arquivo em Julia, selecione os comandos digitados (ou posione o cursor na linha desejada), selecione View, Command Pallete e digite o seguinte comando na caixa superior - Julia: Send Current Line or Selection to REPL. Após isso, os demais comandos podem ser executados por meio das teclas de atalho Ctrl + Enter.



Row	Amostra	Clone	d	h	f	r	Patologia	Severidade	Nutri\x87\xc6o	Praga	Rusticidade
	Int64	Int64	Float64	Float64	Int64	Int64	String	Int64	String	String	Float64
1	1	1	2.87	27.35	11	3	Ausencia	0	Nitrogenio	Ausencia	3.0
2	2	1	3.03	30.26	16	4	Ausencia	0	Ausencia	Ausencia	9.2
1998	1998	10	2.93	28.76	15	4	Ausencia	0	Ausencia	Ausencia	9.2
1999	1999	10	2.37	30.71	19	4	Ausencia	0	Ausencia	Ausencia	9.1
2000	2000	10	2.62	25.03	9	4	Ausencia	0	Ausencia	Ausencia	8.0

1.5 DISPONIBILIDADE DOS ARQUIVOS E UTILIZAÇÃO DO JULIA

Acesso aos dados e códigos em JULIA em GOOGLE COLAB e GITHUB.

GOOGLE COLAB

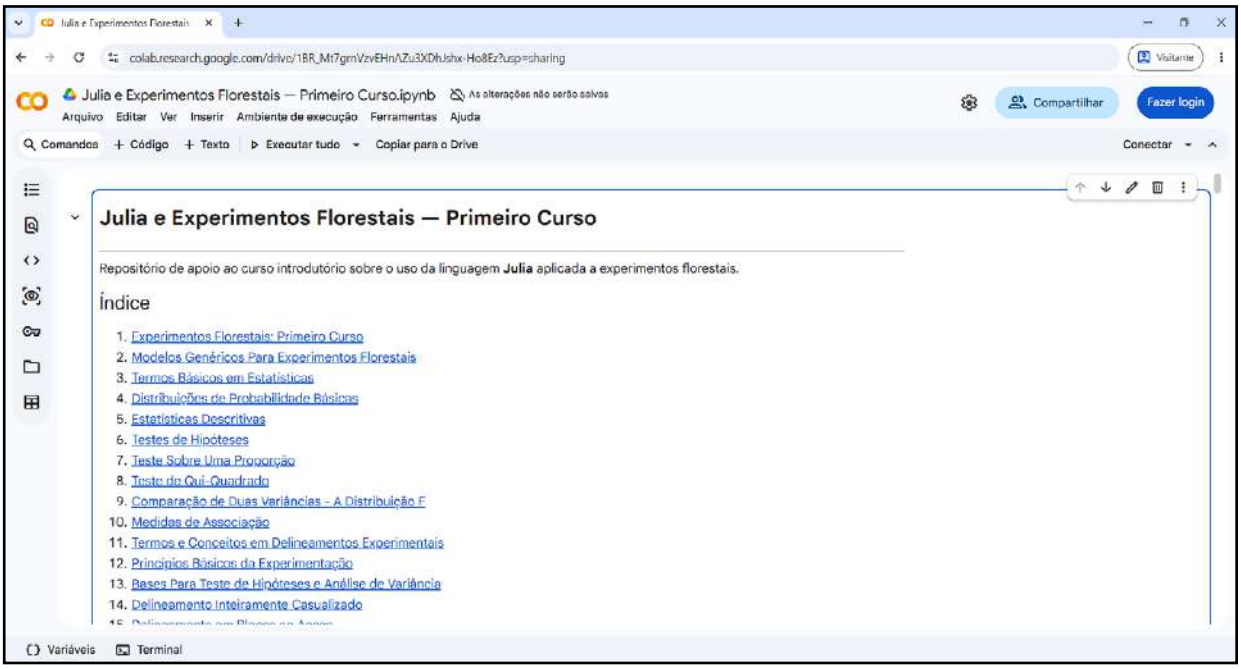
Embora o Google Colab seja amplamente utilizado para aplicações em Python, também é possível utilizar a linguagem Julia diretamente nesse ambiente.

Os arquivos, códigos e exemplos utilizados no presente texto estão também disponibilizados na plataforma Google Colab. Na plataforma pode ser realizada a execução dos códigos e modificações nos exemplos diretamente no ambiente virtual, sem a necessidade de instalar previamente todos os programas e pacotes necessários em seu computador.

O acesso aos arquivos pode ser realizado por meio do endereço:
https://colab.research.google.com/ ou scaneie o QR code acima.

Ressalta-se, portanto, que os arquivos disponibilizados constituem uma extensão prática deste livro. O leitor é incentivado a utilizá-los como ponto de partida para a exploração dos métodos apresentados, modificando os exemplos e aplicando-os a diferentes conjuntos de dados e situações experimentais, particularmente aquelas relacionadas às Ciências Florestais.

FIGURA 1.9. Acesso dos arquivos de dados e funções ao Google Colab.



A utilização das funções pode ser feita diretamente da plataforma google colab, ou então copiando as funções desejadas e colando diretamente no terminal da linguagem Julia no computador do usuário.

O usuário poderá interagir como índice para navegar por todos os 21 tópicos apresentados no presente texto.

GITHUB

Os scripts e conjuntos de dados apresentados no presente texto também estão disponibilizados na plataforma GitHub, sendo possível acessar pesquisando pelo nome do livro no site oficial do GitHub:

https://github.com/

diretamente através do link do repositório:

Julia-e-Experimentos-Florestais---Primeiro-Curso ou scaneie o QR code acima.

FIGURA 1.10. Acesso ao GitHub.

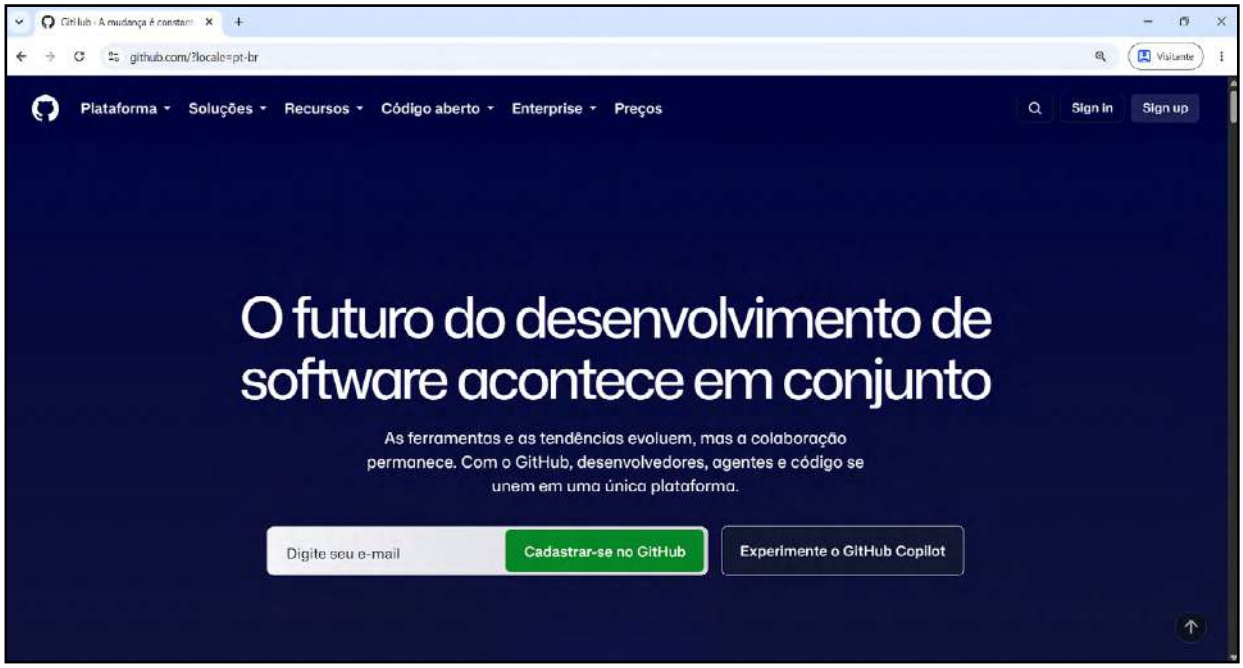


FIGURA 1.11. Pesquisa no GitHub: Julia e Experimentos Florestais.

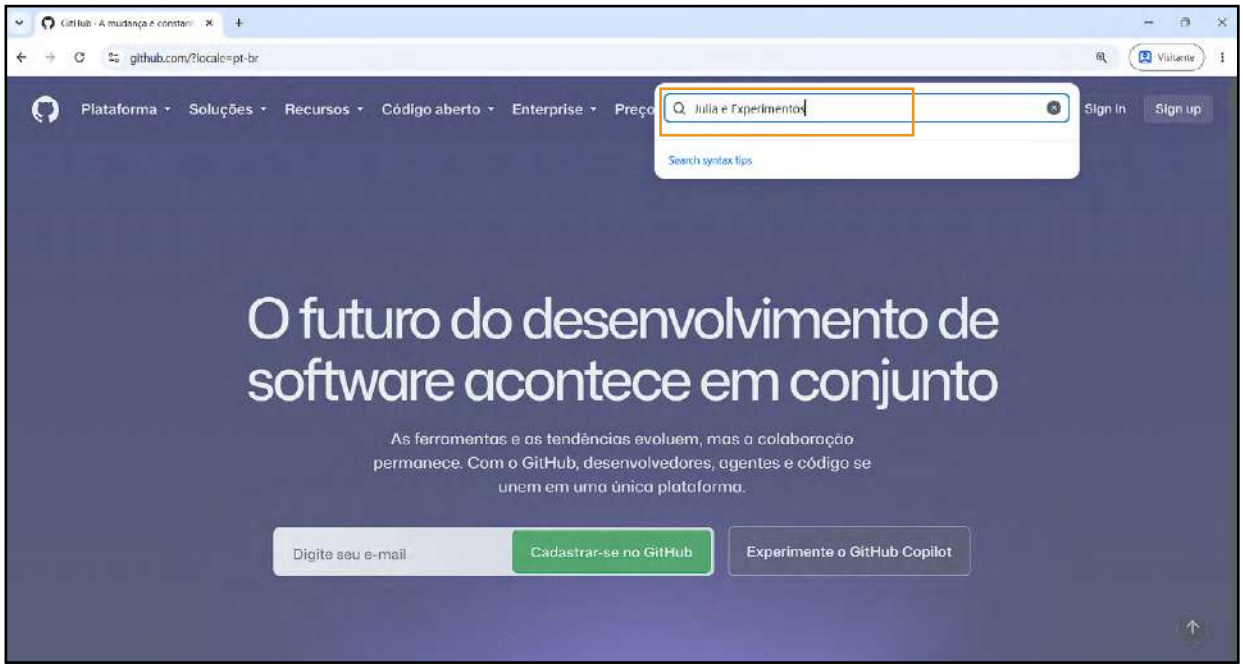


FIGURA 1.12. Acesso ao repositório dos arquivos de dados e funções ao GitHub.

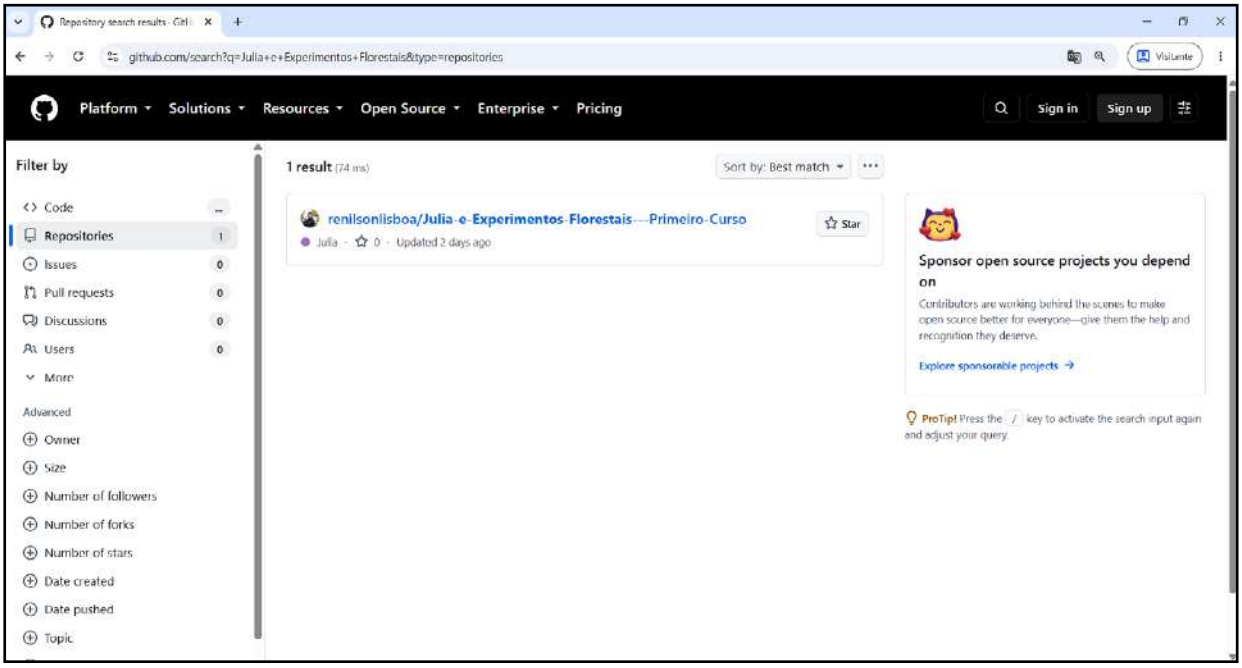
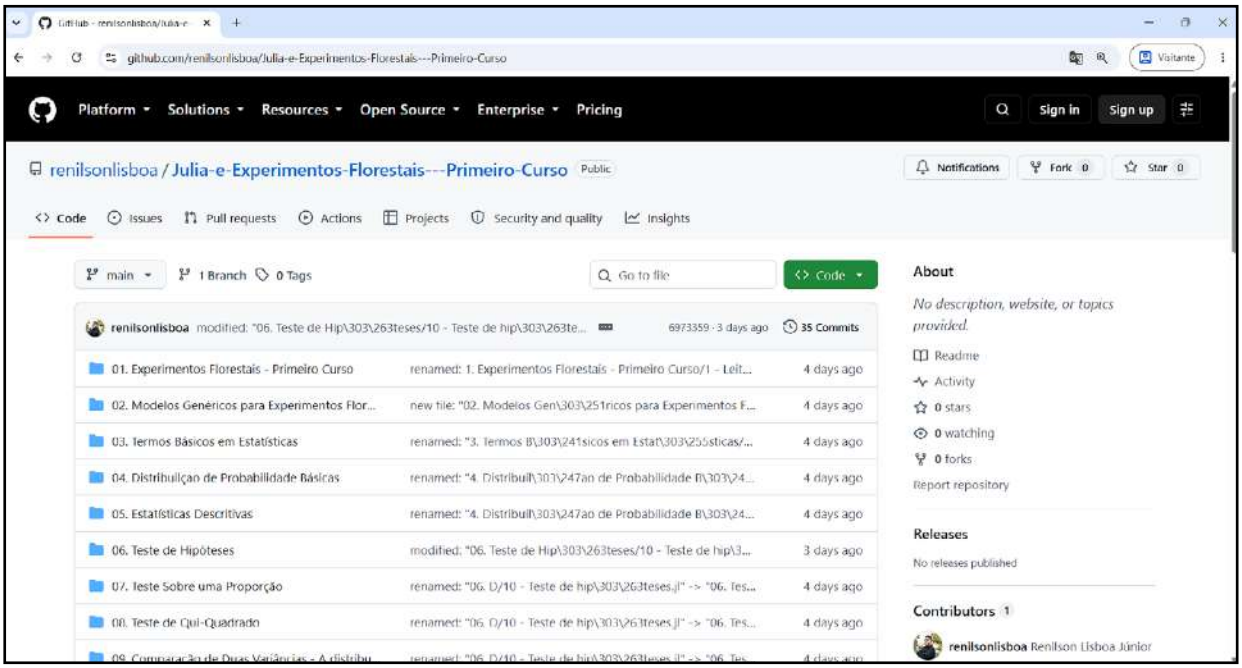


FIGURA 1.13. Acesso dos arquivos de dados e funções ao GitHub.



A utilização dos scripts pode ser realizada selecionando o arquivo desejado, baixando-o e posteriormente utilizando o mesmo no VSCode, ou então, diretamente no terminal da linguagem Julia.

FIGURA 1.14. Baixar função e arquivo de dados.

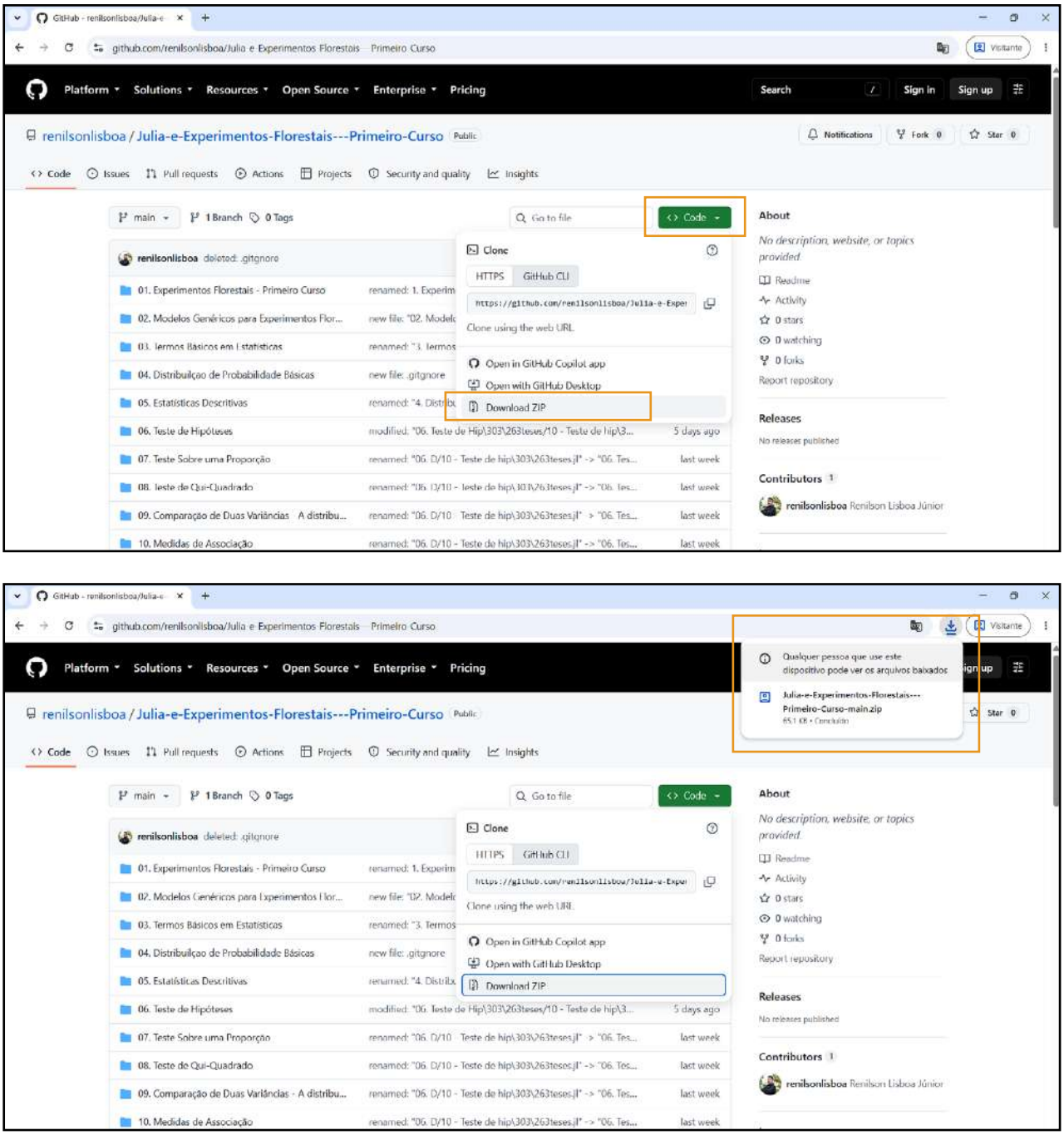
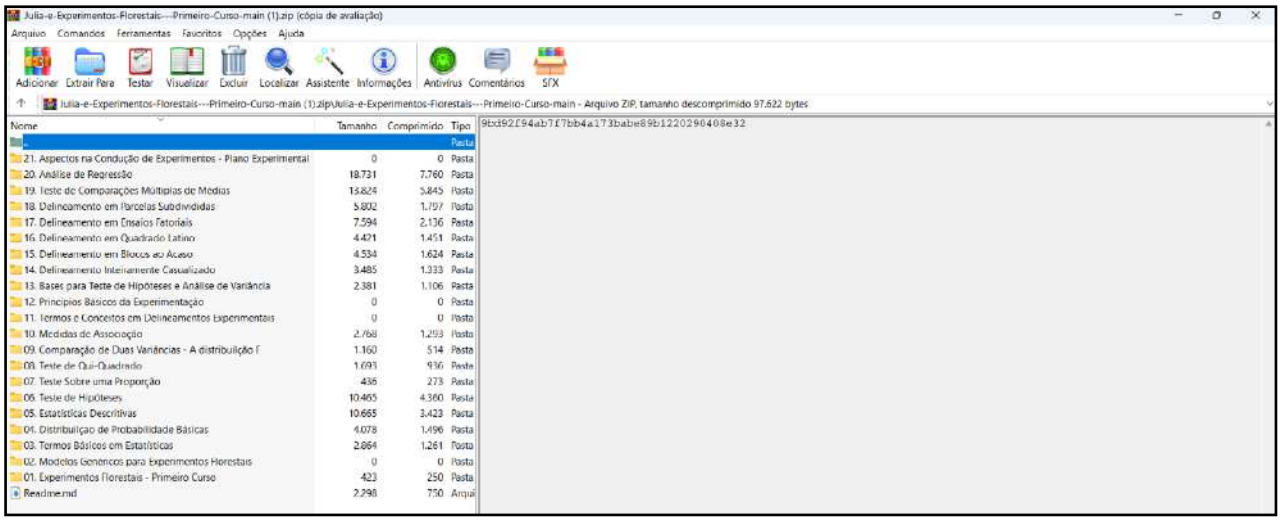
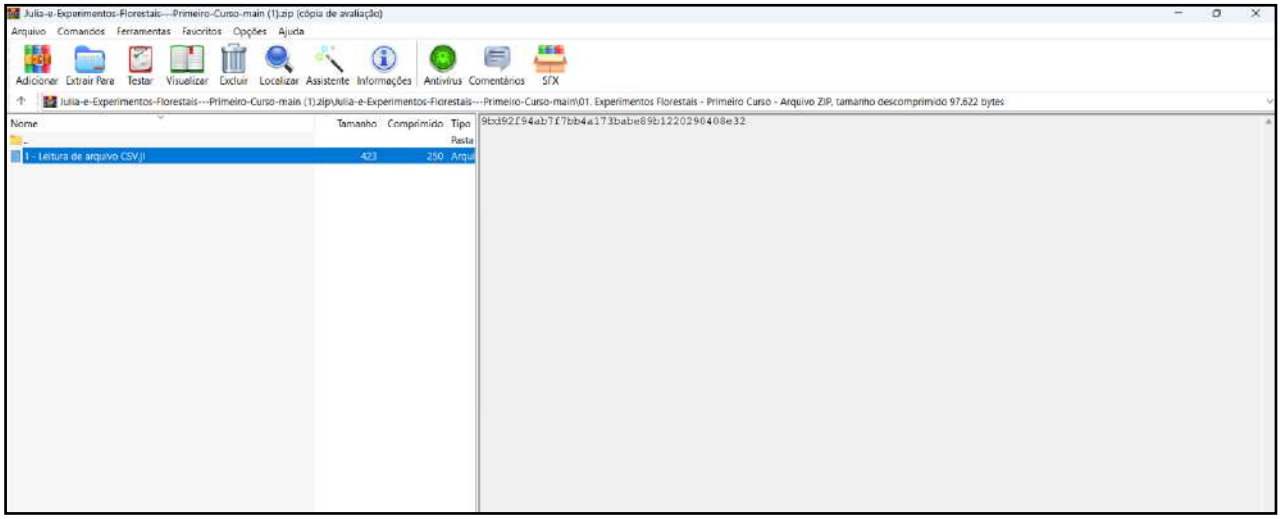
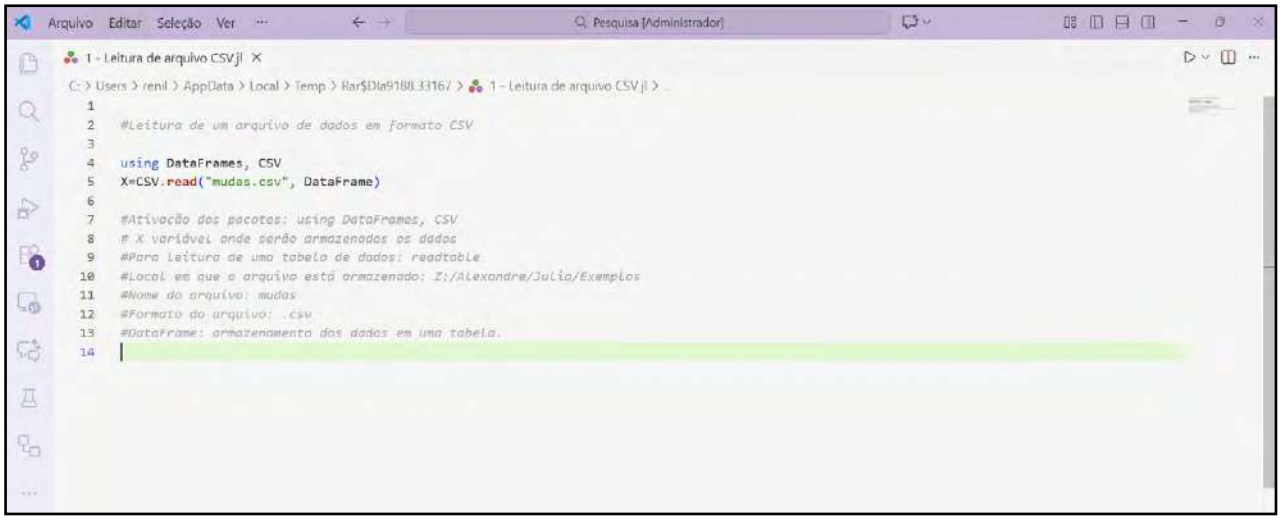


FIGURA 1.15. Acessar a função e arquivo de dados via VSCODE.







2 MODELO GENÉRICO PARA EXPERIMENTOS FLORESTAIS

Engenheiros Florestais investem tempo e recursos consideráveis conduzindo e monitorando experimentos. Estes profissionais se preocupam em analisar o comportamento dos elementos (árvores, mudas, solo, tábuas, etc.) integrantes dos tratamentos. Para que as conclusões obtidas sejam válidas, os métodos experimentais são utilizados.

O método experimental consiste em comparar dois ou mais tratamentos, ou seja, condições nas quais os elementos são submetidos. Os elementos de cada tratamento são considerados de maneira idêntica, enquanto elementos que recebem os diferentes tratamentos são considerados de acordo com o protocolo estabelecido para cada um deles.

Considere a seguinte denominação:

Tratamentos → Variável independente X

Para avaliar o efeito dos tratamentos sobre os elementos, características (variáveis) são mensuradas. Nos experimentos tem-se como objetivo prático o estabelecimento de tratamentos projetados visando maximizar ou minimizar os valores de uma característica dos elementos. As variáveis mensuradas podem ser denominadas de:

Características dos elementos → Variável dependente Y

Qualquer diferença entre os valores de uma variável Y relacionada a uma variável X é denominado efeito de tratamento, causada pelos explícitos efeitos de cada um deles. A execução de um experimento é feita para avaliar a relação de efeito, ou seja, concluir que a diferença observada (maximização ou minimização dos valores de Y) foi causada pelos efeitos dos tratamentos e não como um fato meramente casual. Para expressar tal relação, um modelo genérico pode ser estabelecido:

$$Y_{ij} = f(X_j)$$

Em que:
Y_{ij} é a variável dependente para o i- ésimo elemento do j- ésimo tratamento
X_j é a variável independente, referente aos J níveis de tratamento
f é uma função de correspondência aplicada para testar que os efeitos de X_j sobre Y_i são nulos

A maioria dos experimentos consiste em amostrar $i = 1,2,...,I$ elementos da população. Existe uma fonte de variação não controlada pelo pesquisador, dado que os elementos I no inicío do experimento não são perfeitamente idênticos e, também, podem ser submetidos à pequenas diferenças no meio experimental (mesmo que todo cuidado seja observado). Portanto, existe uma incerteza associada (e) e deve ser acrescentada ao modelo geral:

$$y_{ij} = f(x_j) + e_{ij}$$

Métodos estatísticos são utilizados para evidenciar se um mesmo experimento for replicado *K*- vezes, as respostas de Y_i em função de X_j seriam as mesmas. Raramente os experimentos são repetidos *K*- vezes, devido aos seus custos de execução e, também, ao tempo exigido. Diversamente, um experimento simples é conduzido e, por meio de métodos estatísticos, inferências são formalizadas para decidir se as mesmas diferenças de y_i em função de x_j seriam observadas, caso o experimento seja repetido. Isto quer dizer que, além de dados experimentais, as conclusões finais dependem de probabilidade.

A literatura estatística viabiliza para o Engenheiro Florestal várias técnicas e procedimentos para a análise de dados experimentais, formalização de inferências e apresentação de conclusões sobre as hipóteses que estão sendo testadas no experimento. Nos experimentos as hipóteses testadas referem-se a:

1. Hipótese do experimento (do pesquisador):

O aumento dos valores de X_j maximiza os valores de sobre Y_i

2. Hipótese estatística (do método científico):

Os efeitos de X_j sobre Y_i são nulos (não existe diferenças entre os tratamentos; não é possível maximizar Y_i em função de X_j).

O que motiva a condução de um experimento é maximizar (ou minimizar) os valores de Y_i em função da sistematização de X_j, o que dá origem a hipótese do experimento. Essa hipótese é definida antes da execução do experimento, uma vez que toda a seleção de níveis *j* de *X* e as variáveis *Y* são delineadas em função da hipótese do experimento.

O estabelecimento da hipótese estatística revela protocolos sobre o plano experimental, tais como:

- número de repetições *I*
- definição dos tratamentos X_j sobre os elementos Y_i
- testes estatísticos visando testar a hipótese da nulidade que os efeitos dos tratamentos X_j são nulos sobre Y_i, ou seja, que na prática não é possível maximizar ou minimizar os valores de *Y* em função dos níveis testados de *X*.

As conclusões do experimento são inferenciais, o pesquisador está sujeito a um erro α relativa à conclusão formalizada sobre a hipótese da nulidade. Assim, as hipóteses estatísticas estabelecidas são:

$$\begin{cases} H_0\text{: os efeitos dos tratamentos são nulos} \\ H_1\text{: os efeitos dos tratamentos são diferentes de zero} \end{cases}$$
$$\alpha = 0{,}05 = 1 - P = 1 - 0{,}95$$

Métodos estatísticos são aplicadas para inferir se a hipótese da nulidade (H_0) é verdadeira ou é falsa, por meio de uma técnica estatística que é denominada de análise dos dados. O processo de análise dos dados experimentais pode ser iniciado por meio de sua sumarização, ou seja, por meio da apresentação das medidas de tendência central e de variabilidade.

De modo geral, se houver interesse em se aplicar uma função *f* para descrever o desempenho dos *I* elementos (se os valores de *Y* são maximizados ou minimizados) submetidos aos *J* tratamentos, com objetivo de estabelecer inferências sobre o efeito de tratamentos nos valores paratmétricos de *Y* em função dos valores médios amostrais de $\overline{y}$ e concluir a respeito de H_0.

Portanto:

- **ESTIMATIVA (INFERÊNCIA):** nos experimentos florestais, a inferência refere-se a estimar a influência dos tratamentos sobre os elementos da população (*Y*), não apenas aos *I* elementos amostrais (*y*). A estimativa pode ser pontual (média) ou intervalar (intervalo de confiança da média).
- **TESTE DA HIPÓTESE:** o processo de teste de hipótese pode ser efetuado considerando o experimento como um todo (Figura 2.1):

FIGURA 2.1. Processo de teste de hipótese



MOTIVAÇÕES - FIGURA 2.1

Inicialmente **(1)** um problema prático é estabelecido por um Engenheiro Florestal. Por exemplo, este profissional deseja aumentar a área basal (G, em m². ha^{-1}) em um povoamento florestal na idade de rotação (7 anos), com espaçamento (3 m x 1,75 m).

Maximizar G, em m². ha^{-1}

O segundo passo **(2)** é verificar a existência de resultados experimentais publicados em artigos científicos a respeito do assunto pesquisado. O Engenheiro Florestal identifica métodos silviculturais potenciais para aumentar a produtividade de G. Entre os métodos indicados na literatura, o engenheiro observa que a adubação em cobertura aplicada a cada ano e durante os 3 primeiros anos após o plantio é apropriada para maximizar os valores de G, propondo as hipóteses do experimento **(3)**[1] . Em seguida **(4)** o modelo paramétrico $Y_{ij} = f(X_j)$ é definido e a estrutura experimental passa a ser formalizada valendo-se do modelo amostral (Figura 2.2):

FIGURA 2.2. Modelo amostral

$$y_{ij} = f(X_j) + e_{ij}$$

variável independente

erro associado

variável dependente

[1] Os passos para formalizar as hipóteses estão apresentados na seção 21.

No modelo amostral em **(4)** são definidas as variáveis dependentes (valor que se quer maximizar ou minimizar) e independentes (tratamentos). O projeto para este item é importante, porque é quando se efetua o planejamento do número de repetições I para os níveis J, a composição das unidades experimentais, detalhes sobre a instalação, condução e monitoramento (mensuração) do experimento.

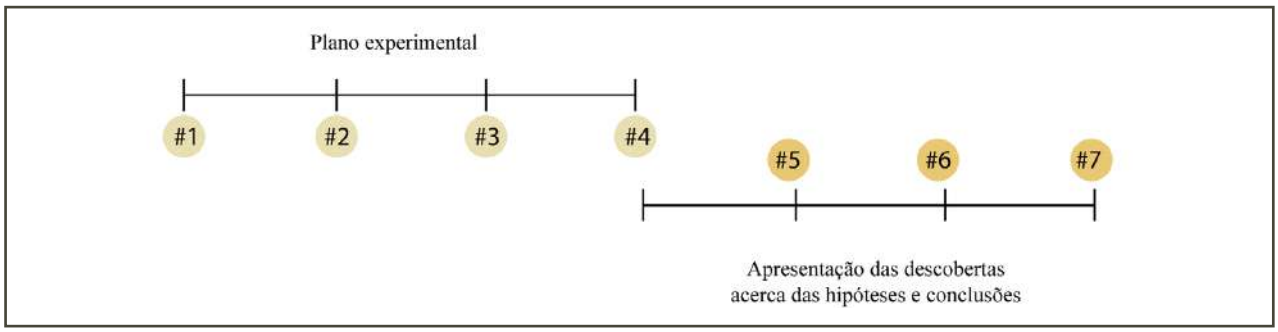
Uma vez coletado os dados **(5)** a função de correspondência f (modelo estatístico) é estabelecida, bem como as hipóteses estatísticas que serão testadas. Aspectos de estatística descritiva também são definidos, pois auxiliam a compreender o domínio dos valores de y_i. A motivação para a apresentação dos capítulos subsequentes é a de apresentar alguns modelos da função f utilizados com frequência na Experimentação Florestal para testar hipóteses. A decisão de qual modelo f deve ser utilizado depende de algumas características da variável y e, também, de x, daí a necessidade de se conhecer diversos modelos. Destaca-se o uso de programas para efetuação de cálculos, os quais foram apresentados em linguagem Julia de maneira a aproximar dos procedimentos manuais de cálculos.

A seguir em **(6)** são apresentados os resultados das estatísticas descritivas e do modelo estatístico f, bem como sobre a hipótese da nulidade.

Finalmente em **(7)** as recomendações gerais são apresentadas, concernentes ao problema que deu origem ao experimento para maximizar ou minimizar valores de uma variável Y.

Observe que a apresentação dos resultados de um experimento requer todos os itens, ou seja:

FIGURA 2.3. Plano experimental



Um resumo do plano experimental está disponibilizado na seção 21. Nos exercícios apresentados o foco refere-se aos itens 4 a 7. Os modelos f apresentados são básicos para a Experimentação Florestal. Para os mesmos exemplos existem outros f que podem ser utilizados para melhorar, sobretudo, os resultados em 6 e, consequentemente, em 7.

2.1 BASES PARA TESTAR HIPÓTESES ESTATÍSTICAS EM EXPERIMENTOS FLORESTAIS – CASO PARA RAZÃO DE VARIÂNCIAS

Os experimentos florestais são planejados tal que os elementos I nas diferentes condições X_j a que são submetidos, sejam tratados igualitariamente. Infelizmente, mesmo com todos os cuidados possíveis, dificilmente o Engenheiro Florestal terá condição para controlar todas as fontes de variação no ambiente experimental, os quais afetam os valores da variável resposta y. Em um experimento sob condições perfeitas, somente as fontes y e x variariam, de tal forma que x é controlado pelo pesquisador e y em função de x.

Entretanto, nos experimentos reais uma fonte adicional de variação é admitida (e_{ij}), chamada de erro experimental, devido a características que não são controladas. Assim, no planejamento experimental também são discutidas formas para minimizar as fontes de erro, cuja minimização implica na escolha do modelo *f*.

Uma vez que a hipótese do experimento tenha sido definida, assim como o modelo genérico $y_{ij}=f(x_j)+e_{ij}$, para efeito fixo, hipóteses estatísticas são formalizadas:

H_0: $\mu_1 = \mu_2 = ... = \mu_I$ (as médias não diferem entre si, os efeitos dos tratamentos são nulos)
H_1: $\mu_i \neq \mu_{i'}$ (as médias diferem entre si, os efeitos dos tratamentos não são nulos)

A hipótese da nulidade H_0, no presente caso para a média, revela que *X* não afeta os valores de *Y*, portanto, o efeito sobre as médias dos tratamentos é nulo. Se as estimativas obtidas por meio de *f* revelar que *X* afeta os valores de *Y*, então, a hipótese da nulidade é falsa e, portanto, rejeitada e a hipótese alternativa H_1 é adotada.

A decisão em rejeitar H_0, o que implica em aceitar H_1, provém suporte à hipótese do experimento. Por meio dos dados experimentais e de um teste estatístico (modelo *f*) a hipótese da nulidade é testada. Em boa parte dos experimentos florestais são testados níveis de $J > 2$ e as condições permitem aplicar um modelo *f* que corresponde a uma razão de duas estimativas:

$$\frac{\textit{diferença entre as médias dos tratamentos}}{\textit{diferença entre os elementos tratados da mesma forma}}$$

Ou seja:

$$\frac{\textit{diferença entre tratamentos}}{\textit{diferença dentro dos tratamentos}}$$

Por meio da hipótese da nulidade, a expectativa para estes experimentos, se eles forem repetidos, é obter um valor da razão que se aproxima de 1. Então, a seguinte razão pode ser considerada:

$$\frac{\textit{erro experimental}}{\textit{erro experimental}}$$

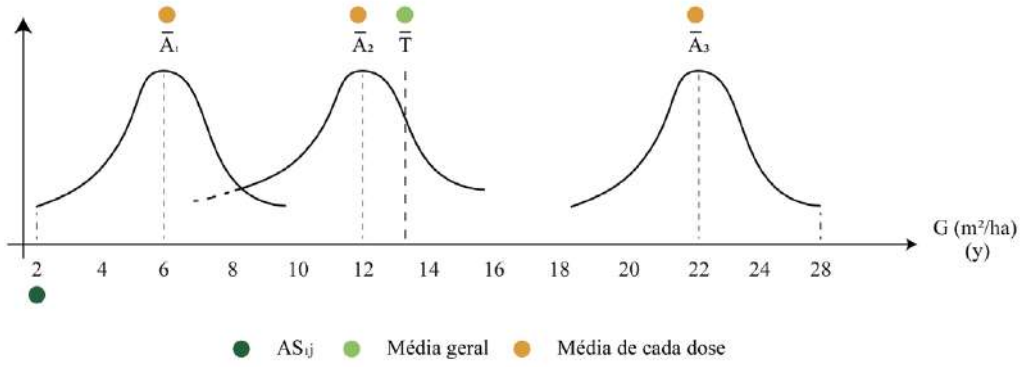
Se a hipótese da nulidade for falsa, a expectativa é obter razão com valores grandes, geralmente bem maiores que 1. Assim, para atender tal condição, a razão pode ser reescrita:

$$\frac{\textit{erro experimental + efeito de tratamento}}{\textit{erro experimental}}$$

Se a hipótese H_0 é verdadeira (as médias não diferem) o valor da razão se aproxima de 1. Se H_0 for falsa, ou seja H_1 é verdadeira (as médias são diferentes), o valor da razão é tanto maior que 1 quanto mais claras são as evidências de efeito dos tratamentos. Para avaliar se o valor da razão obtida revela evidência para aceitar ou rejeitar H_0, o valor é comparado com um valor oriundo de uma distribuição de probabilidade (lembre-se: o que aconteceria se os experimentos fossem repetidos?).

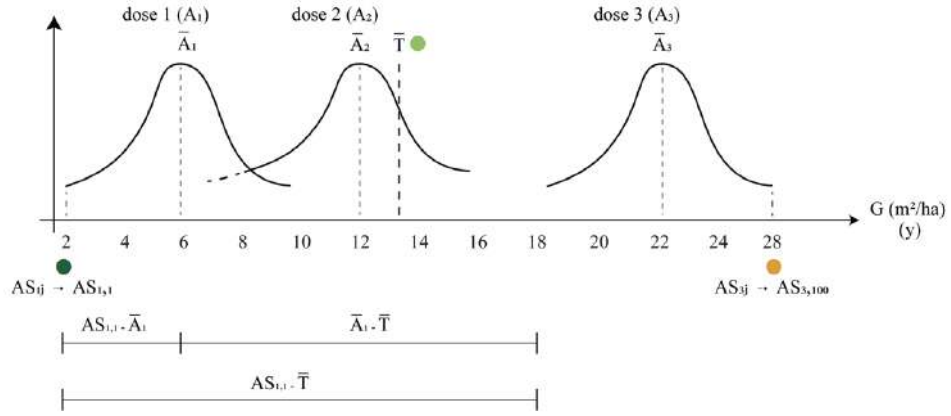
Considere os seguintes valores experimentais *y* da variável G ($m^2.ha^{-1}$) em função de 3 doses de adubo NPK (x,j = 1,2,3), para i = 100 elementos para cada nível de *x*, portanto, I = 300, ou seja, número de repetições é igual a 100 para cada dose, o que resulta em 300 unidades experimentais. Cada unidade experimental (elemento) *i* refere-se a uma parcela com 20 árvores sob espaçamento 1,75 m x 3 m, na qual foi mensurado o diâmetro à altura do peito 8 anos após a instalação do experimento.

FIGURA 2.4. Curva da distribuição em função de 3 doses de adubo NPK



Parte-se do princípio de que não existem diferenças entre os tratamentos testados (Figura 2.4). Isto implica em considerar que os valores das 3 médias ($\bar{A}_1, \bar{A}_2, \bar{A}_3$) não são diferentes, mesmo que numericamente diferentes, ou seja, que não existe efeito de *X* sobre *Y*. As diferenças entre essas médias podem ser oriundas somente do erro experimental (devido as fontes de variação não controladas). Para avaliar se existe evidências para aceitar H_0, a razão (erro experimental + efeito de tratamento)/(erro experimental) deve ser calculada, conforme apresentada na sequência.

FIGURA 2.5. Curva da distribuição em função de 3 doses de adubo NPK



Considere a média para cada dose (tratamento) $\bar{y}_1 = \bar{A}_1$, $\bar{y}_2 = \bar{A}_2$, $\bar{y}_3 = \bar{A}_3$, e adicionalmente a média geral de todos os valores $\bar{y} = \bar{T}$ (Figura 2.5). O primeiro valor $A_{(j,i)} = y_{(1,1)} = 2$ pode ser descrito como:

$$y_{(1,1)} = \bar{y} + (\bar{y}_1 - \bar{y}) + (y_{(1,1)} - \bar{y}_1)$$

Ou:

$$(y_{(1,1)} - \bar{y}) = (\bar{y}_1 - \bar{y}) + (y_{(1,1)} - \bar{y}_1)$$

Como representado na Figura 2.5, estes valores correspondem à seguinte composição:

Desvio total = $(y_{(1,1)} - \bar{y})$
Desvio entre tratamento = $(\bar{y}_1 - \bar{y})$
Desvio dentro do tratamento = $(y_{(1,1)} - \bar{y}_1)$

Esta partição dos valores experimentais em desvios pode ser aplicada para todas as observações ($I = 300$), ou seja:

Dose 1
$$\begin{cases} (y_{(1,1)} - \bar{y}) = (\bar{y}_1 - \bar{y}) + (y_{(1,1)} - \bar{y}_1) \\ \dots \\ (y_{(1,100)} - \bar{y}) = (\bar{y}_1 - \bar{y}) + (y_{(1,100)} - \bar{y}_1) \end{cases}$$

Dose 2
$$\begin{cases} (y_{(2,1)} - \bar{y}) = (\bar{y}_2 - \bar{y}) + (y_{(2,1)} - \bar{y}_2) \\ \dots \\ (y_{(2,100)} - \bar{y}) = (\bar{y}_2 - \bar{y}) + (y_{(2,100)} - \bar{y}_2) \end{cases}$$

Dose 3
$$\begin{cases} (y_{(3,1)} - \bar{y}) = (\bar{y}_3 - \bar{y}) + (y_{(3,1)} - \bar{y}_3) \\ \dots \\ (y_{(3,100)} - \bar{y}) = (\bar{y}_3 - \bar{y}) + (y_{(3,100)} - \bar{y}_3) \end{cases}$$

Esse conjunto de equações de particionamento dos valores em desvios pode ser expresso por:

Soma de desvio total= soma de desvios entre tratamentos + soma de desvios dentro dos tratamentos

O problema em efetuar a soma de desvios é que o valor resultante é sempre igual a zero. Entretanto, pode-se conservar a relação de aditividade das expressões anteriores se a soma dos desvios for tomada ao quadrado, portanto:

$$\begin{cases} (y_{(1,1)} - \bar{y})^2 = (\bar{y}_1 - \bar{y})^2 + (y_{(1,1)} - \bar{y}_1)^2 \\ \dots \\ (y_{(3,100)} - \bar{y})^2 = (\bar{y}_3 - \bar{y})^2 + (y_{(3,100)} - \bar{y}_3)^2 \end{cases}$$

Ou seja:

Soma de quadrado total=
soma de quadrado entre tratamentos + soma de quadrado dentro dos tratamentos

Portanto:

variabilidade total = efeito de tramentos + erro experimental

Note que na hipótese da nulidade H_0 supõe-se que a soma de quadrados entre tratamentos seja próxima de zero, ou seja:

$$H_0 = (\mu_1 - \mu)^2 = (\mu_2 - \mu)^2 = (\mu_3 - \mu)^2$$

Onde:
$\bar{y}_1$ estima μ_1, $\bar{y}_2$ estima μ_2, $\bar{y}_3$ estima μ_3 e $\bar{y}$ estima μ.

O problema que remanesce para ser avaliado é a soma de quadrados entre tratamentos $\sum(\bar{y}_i - \bar{y})^2$, uma vez que as médias entre tratamentos também estão sujeitas a um erro experimental. Portanto, se a soma de quadrados entre tratamentos resultar em valores grandes, também não se pode afirmar que isto é devido a efeito dos tratamentos.

Estatísticas adicionais podem ser calculadas para o modelo em **(1)** (Figura 2.1), o qual pode ser reescrito em termos de variância:

$$\frac{\textit{Soma de total de quadrados}}{\textit{Graus de liberdade do total}} = \frac{\textit{soma de quadrados entre tratamentos}}{\textit{Graus de liberdade entre tratamentos}} + \frac{\textit{soma de quadrado dentro dos tratamentos}}{\textit{Graus de liberdade dentro dos tratamentos}}$$

Ou:

$$\textit{Variância total} = \textit{variância de tratamentos} + \textit{variância do erro}$$

Uma vez que a variância $s^2 = \frac{\textit{soma de quadrados}}{\textit{Grau de liberdade}} = \textit{quadrado médio}$

Portanto, a variância total pode ser decomposta em variância entre os tratamentos e variância dentro dos tratamentos, o que também é denominado de quadrado médio dos tratamentos e quadrado médio do resíduo (ou erro). A variância entre tratamentos combina a presença do efeito dos tratamentos e o erro experimental, ou seja, a variabilidade devido aos tratamentos e ao erro experimental. A variância dentro dos tratamentos é devida somente a variabilidade existente entre as unidades experimentais tratadas da mesma forma. Dado que estas duas fontes de variação são independes sob a hipótese H_0, uma razão F entre elas pode ser calculada:

$$F = \frac{\textit{Variância entre tratamentos}}{\textit{Variância dentro dos tratamentos}} = \frac{\textit{Quadrado médio dos tratamentos}}{\textit{Quadrado médio do resíduo}}$$

O valor esperado para a estatística F é que se aproxime de 1 (hipótese da nulidade é verdadeira), enquanto um valor de $F > 1$ e acima de um valor de probabilidade especificado, indica que H_0 é falsa. O valor de F obtido no experimento é comparado com um valor de uma distribuição de razões de variâncias (distribuição de F de Fisher), com objetivo de revelar evidências probabilísticas para aceitar ou rejeitar H_0. Em termos de composição de variância para os dados, F pode ser reescrito como:

$$F = \frac{\textit{Variância entre tratamentos} + \textit{Variância dentro dos tratamentos}}{\textit{Variância dentro dos tratamentos}}$$

Note que os mesmos dados experimentais estimam as variâncias do numerador e do denominador. A variância do numerador refere-se o quanto as médias dos tratamentos variam em torno da média geral, sendo oportuno destacar que as médias são obtidas em função de valores experimentais que carregam efeito do erro experimental e efeito dos tratamentos. A variância do denominador é obtida calculando-se o quanto os valores de cada tratamento variam em torno de sua média, portanto, carrega somente efeito do erro (Figura 2.7).

FIGURA 2.6. Curva da distribuição em função de 3 doses de adubo NPK com erro experimental

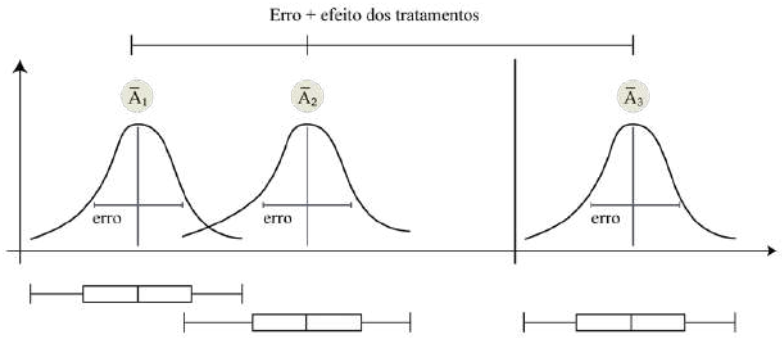


FIGURA 2.7. Tabela de tratamento

Tratamento	Média	Repetição			
		1	2	...	100
A_1	$\bar{A}_1$	(cada valor - $\bar{A}_1$)²			
A_2	$\bar{A}_2$	(cada valor - $\bar{A}_2$)²			
A_3	$\bar{A}_3$	(cada valor - $\bar{A}_3$)²			
Média Geral M					

I = 3

n = 100

Erro

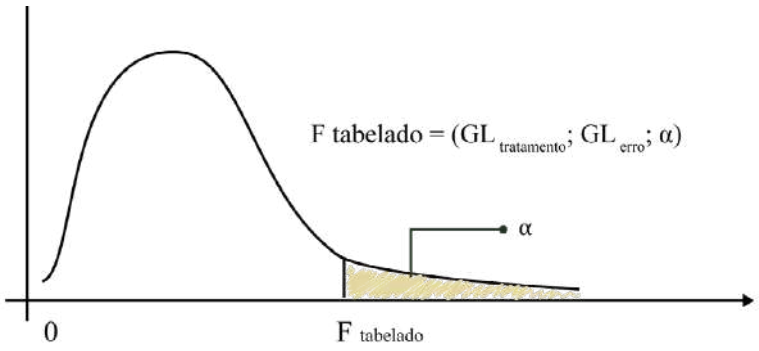
$$\frac{(y_{1i} - \bar{y}_1)^2 + (y_{2i} - \bar{y}_2)^2 + (y_{3i} - \bar{y}_3)^2}{n}$$

$$S_t^2 = \frac{n \,.\, [\,(\bar{A}_1 - M)^2 + (\bar{A}_2 - M)^2 + (\bar{A}_3 - M)^2\,]}{I - 1}$$

Erro + efeito de tratamento

O valor de F calculado (F_c) por meio dos dados experimentais é comparado com um valor de uma família da distribuição de F (F_t). O valor de F_t depende do tamanho da amostra (graus de liberdade) que deu origem à variância do numerador e à variância do denominador da estatística Fc e, também, do valor de α (proporção de área da curva à direita da distribuição de *F*) (Figura 2.8).

FIGURA 2.8. Curva da distribuição F



Considere a distribuição F apresentada na Figura 2.8. O pesquisador conduz experimentos com expectativas de calcular um valor de F_c maior que o valor de F_t, ou seja, obter um valor de F_c na região α, o que revela evidências de efeitos dos tratamentos, o que na prática significa a possibilidade de maximizar ou minimizar os valores de *Y*.

Entretanto, na hipótese estatística testada H_0, os valores das médias dos tratamentos não diferem, portanto, a expectativa inicial em um teste estatístico é obter valores de F_c menores que F_t, ou seja, localizados na região β. Normalmente, usa-se $\alpha = 0,05$ (5%), sendo uma probabilidade denominada nível de significância. Se $\alpha = 0,05$ (controlado pelo pesquisador), então, $\beta = 1 - \alpha = 0,95$. Em outros casos, pode-se também determinar $\alpha = 0,01$.

O valor de α é definido à priori nos experimentos, é uma escolha pessoal, e revela qual é a probabilidade de declarar H_0 como verdadeira (sem efeitos significantes dos tratamentos) quando deveria ser rejeitada (H_1 deveria ser aceita), porque existem evidências reais de efeito dos tratamentos.

O que foi explicado compreende a análise de experimentos para evidenciar a maximização ou minimização de valores de Y inclusos em uma função f no modelo genérico que compreende a análise de duas fontes de variação (devido aos tratamentos e ao erro experimental), mas existem outras funções f que são utilizadas com o mesmo propósito. Observe que o pesquisador conduz experimentos para testar a hipótese de que os tratamentos revelem efeito sobre as médias, enquanto por meio do método científico e dos testes estatísticos se conduz experimentos para evidenciar que os tratamentos não possuem efeito sobre as médias. O método científico pressupõe que a hipótese da nulidade H_0 é sempre verdadeira. O pesquisador ao propor um experimento tem usualmente como expectativa obter valores dos testes estatísticos na região α, mas os testes priorizam a região β. Isto decorre da relação das proporções de probabilidade definidas e auxilia na redução de erros nas inferências sobre a recomendação dos tratamentos.

Nos capítulos subsequentes serão apresentadas informações que auxiliarão a compreensão dos aspectos básicos de estatística, de experimentação e também de aplicação dos cálculos, valendo-se de uma linguagem de computador. Almeja-se, após a realização dos exercícios, que este texto básico seja mais compreensivo e útil para refletir sobre o todo de um experimento antes de executá-lo.

3 TERMOS BÁSICOS EM ESTATÍSTICA

Um aspecto essencial da estatística é utilizar métodos de inferência para obter conclusões que estão latentes nos dados amostrados, geralmente obtidos por amostragem, no cotidiano do Engenheiro Florestal. Assim, objetiva-se por meio da estatística apresentar conclusões gerais a partir dos dados, que são armazenados em tabelas.

O estudo é iniciado por meio do reconhecimento e definição de um problema, planejamento de um experimento, sua condução, análise dos dados, para então formalizar conclusões e tomar decisões acerca do problema definido. Estas etapas serão evidenciadas no decorrer do texto quando alguns exemplos serão apresentados. Para compreendê-las, algumas definições são importantes:

- **População:** conjunto de elementos que apresentam uma característica comum. Refere-se a todos os elementos pertencentes a população.
- **Censo:** dados de todos os elementos da população.
- **Parâmetro:** medida numérica que descreve uma característica da população. São representados por letras gregas e são valores fixos. Por exemplo, pode ser a média (μ), a variância (σ^2) e o desvio padrão (σ) da população.
- **Amostra:** subconjunto de elementos a serem observados ou medidos na população.
- **Estatística:** medida numérica que descreve uma característica da amostra. São representados por letras romanas e podem variar. Pode ser a média ($\overline{x}$), a variância (s^2), o desvio padrão (s) e o desvio padrão da média ($s_{\overline{x}}$) da amostra ou erro padrão.
- **Inferência estatística:** processo que utiliza as estatísticas obtidas na amostra para estimar os parâmetros da população, de modo a permitir generalização sobre sua essência. Os Engenheiros Florestais geralmente trabalham com amostras e uma das preocupações é que ela seja representativa da população.
- **Variável aleatória:** é uma característica comum aos elementos de uma população e é obtida por meio da atribuição de um número ou uma categoria. Ela é desconhecida e varia de um elemento para outro em uma população. Se assumir um valor fixo, então é denominada de constante. A variável pode ser mensurada quando se atribui a ela um número ou uma categoria utilizando algum instrumento, ou observada quando a avaliação é efetuada sem auxílio instrumental.

Quanto à sua essência, as variáveis podem ser:

i. **Variáveis qualitativas:** quando os resultados possíveis são atributos ou categorias. Esses resultados não são numéricos, mas são definidos por categorias de modo a classificar os indivíduos e geralmente são observadas.

ii. **Variáveis quantitativas:** quando os resultados possíveis são valores numéricos, de contagens, e geralmente são medidas.

As variáveis qualitativas podem ser:

- **Variáveis nominais:** quando as categorias não obedecem a uma ordem natural, ou seja, as categorias não podem ser ordenadas. Por exemplo, o sexo das árvores de araucária.
- **Variáveis ordinais:** quando as categorias obedecem a uma ordem natural. Por exemplo, o estágio de crescimento das plantas: jovem, intermediária e adulta.

As variáveis quantitativas podem ser:

- **Variáveis discretas:** quando são obtidos apenas determinados valores dentro de uma escala. Por exemplo, o número de árvores por hectare.
- **Variáveis contínuas:** quando podem assumir qualquer valor dentro de uma escala. Por exemplo, a altura das árvores.

As variáveis podem ser interpretadas conforme suas escalas de mensuração. Os métodos para resumir os dados de uma variável dependem da natureza do tipo da variável e de sua escala. As escalas compreendem 4 níveis:

1. **Escala nominal:** quando os elementos são classificados em categorias e um ordenamento delas não constitui sentido prático. Por exemplo, as categorias de sobrevivência (viva ou morta) ou as espécies mais plantadas no Brasil (Eucalipto, Pinus, Plátano e Acácia-Negra). As categorias não precisam ser ordenadas.
2. **Escala ordinal:** quando os indivíduos são classificados em categorias que podem ser ordenadas e tem importância prática. Por exemplo, a qualidade da operação de um plantio de mudas (ruim, razoável e excelente) ou o estágio do ataque de doenças em plantas (inicial, intermediário ou terminal). Nesta escala uma mesma diferença não representa a mesma quantidade.
3. **Escala de intervalo:** os dados apresentam propriedade ordinal e, adicionalmente, pode-se determinar as diferenças entre elas. Os resultados obtidos são numéricos e designados em intervalos iguais a partir de uma origem arbitrária, como, por exemplo, atribuir a importância das disciplinas de estatística, manejo florestal e inventário florestal para o curso de Engenharia Florestal, ou a aptidão de um profissional para ser operador de máquinas florestais. A importância pode ser avaliada por meio da atribuição de uma nota que varia de zero a cinco, em que zero significa muita importância e cinco não significa importância e os demais valores são intermediários. Pode-se notar que os dados apresentam propriedade ordinal (do mais importante para o menos importante) e o intervalo entre as observações é uma unidade fixa de medida. Nesta escala não existe o zero absoluto, sendo apenas relativo e este foi escolhido para representar um ponto referencial de medida. A categoria que representa muita importância pode ser referenciada no ponto zero ou também em outro ponto. A temperatura (em graus Celsius) é um exemplo de variável medida na escalda de intervalo.
4. **Escala de razão:** refere-se aos resultados numéricos obtidos e o valor zero tem significado de ausência de conteúdo da variável medida, ou seja, existe o zero absoluto. Por exemplo, a quantidade em gramas de impureza por quilograma de sementes de eucalipto. Se em um determinado lote de sementes o valor zero for observado, isto indica a inexistência de impureza. A maioria das variáveis que os Engenheiros Florestais mensuram é deste tipo: diâmetro à altura do peito (cm), altura (m), volume (m^3) e biomassa (kg) das árvores. A presença de um zero com significado real permite calcular a razão entre duas medidas, o que não faz sentido para a escala de intervalo. As demais operações aritméticas podem ser efetuadas para esses dois tipos de escala.

3.1 EXERCÍCIOS PROPOSTOS PARA SEREM RESOLVIDOS UTILIZANDO A LINGUAGEM JULIA

A seguir estão apresentados os comandos necessários na linguagem Julia para a execução dos exercícios propostos. O objetivo destes exercícios é familiarizar o usuário com os comandos iniciais da linguagem e sua posterior utilização.

EXERCÍCIO PROPOSTO 1: PRODUÇÃO DE MUDAS EM UM VIVEIRO FLORESTAL

Um Engenheiro Florestal é responsável pela produção anual de 10.000.000 de mudas clonais do híbrido *Eucalyptus grandis* x *Eucalyptus urophylla*, utilizando-se 20 clones e propagação via mini-estacas. O Engenheiro Florestal necessita apresentar mensalmente um relatório sobre a qualidade das mudas aptas à expedição e plantadas em campo, normalmente mudas com 120 dias após a estaquia.

As variáveis que devem ser analisadas são:

- **Variáveis morfométricas:**

1. Diâmetro de colo (d): a referência do local de medição é o colo da muda, sendo o diâmetro medido com paquímetro digital, em mm.
2. Altura das mudas (h): distância entre o colo da muda e a gema apical que deu origem a última folha, medida com régua graduada, em cm.
3. Número de folhas maduras (f): obtido pela contagem realizada após 120 dias da estaquia.
4. Enraizamento (R): a avaliação da emissão das raízes é efetuada de forma visual. Ele é avaliado se o sistema radicular está bem desenvolvido, sem enovelamento, com raiz principal reta, raízes secundárias bem distribuídas e agregadas ao substrato. Na avaliação, utiliza-se a seguinte codificação:

FIGURA 3.1. Classificação do desenvolvimento do sistema radicular de mudas.



TABELA 3.1. Classificação do desenvolvimento do sistema radicular de mudas.

Avaliação	Resultado
Sistema radicular bem desenvolvido, sem enovelamento, com raiz principal reta e raízes secundárias bem distribuídas.	5
Sistema radicular sem enovelamento, com raiz principal reta e raízes secundárias mau distribuídas.	4
Sistema radicular, sem enovelamento, com raiz principal torta.	3
Sistema radicular com enovelamento.	2
Sistema radicular mau desenvolvido.	1

- **Variáveis fitossanitárias**

1. Agentes patológicos

A variável patologia, definida pela análise de agentes patológicos nas plantas, é avaliada pela percepção de sintomas visuais nas folhas como: manchas, coloração e desfolha. São analisados os seguintes agentes, os quais são as respostas.

Fungo: *Cylindrocladium gracilis*
Bactéria: *Pseudomonas cichorii*

A classificação da severidade da doença causada por esses agentes é efetuada considerando quatro níveis:

0 – Folha sadia
1 – Baixa severidade
2 – Média severidade
3 – Alta severidade

FIGURA 3.2. Classificação da severidade da doença.



0 1 2 3

2. Aspectos nutricionais

A variável nutrição, dada pela análise do aspecto visual nutricional das mudas é efetuada pela percepção de sintomas visuais nas folhas como: manchas, coloração e desfolha. É realizado um indicativo sobre o suposto mineral que deve ser aplicado nas mudas.

FIGURA 3.3. Folha com aspecto de deficiência nutricional de nitrogênio.



FIGURA 3.4. Folha com aspecto de deficiência nutricional de fósforo.



FIGURA 3.5. Folha com aspecto de deficiência nutricional de cálcio.



FIGURA 3.6. a) Folha com aspecto de deficiência nutricional de magnésio. b) Folha com aspecto de deficiência nutricional de potássio.



3. Aspectos de pragas

A variável praga, dada pela análise do aspecto visual de incidência de pragas nas mudas é efetuada pela percepção visual de desfolha e/ou corte de raiz. É apresentado um indicativo sobre a suposta praga que atacou as mudas.

4. Aspectos de rusticidade

A variável rusticidade é avaliada pelo seu aspecto visual de flexibilidade. Espera-se mudas com haste sem nenhuma flexibilidade e firme em toda sua extensão, incluindo um substrato bem estruturado. Esta variável é avaliada atribuindo-se valores de 0 a 10, em que 0 é atribuído às mudas que não atingem os padrões esperados e 10 o padrão perfeito da muda.

- Padrões exigidos da qualidade de mudas

As mudas que são expedidas para campo devem ter os seguintes padrões (Figura 3.7):

» Diâmetro do colo acima de 2 mm;
» Altura, 25 a 35 cm;
» Mudas com pelo menos três pares de folhas maduras;
» Sistema radicular desenvolvido, sem enovelamento, com raiz principal reta, raízes secundárias bem distribuídas e agregadas ao substrato;
» Bom aspecto fitossanitário, sem deficiências minerais e isentas de pragas e doenças;
» Mudas rústicas, com notas superiores a 8.

FIGURA 3.7. Padrões exigidos da qualidade de mudas.



Um resumo das variáveis é apresentado na Tabela 3.2.

TABELA 3.2. Variáveis e padrões estabelecidos para avaliação da qualidade de mudas de eucalipto.

Variável	Padrão
Diâmetro	Maior que 2 mm
Altura	De 25 a 35 cm
Folhas	6 folhas
Raíz	5
Patologia	Ausência
Severidade	0
Nutrição	Ausência
Praga	Ausência
Rusticidade	A partir de 8

Uma avaliação das mudas em um viveiro foi efetuada no mês de agosto. No viveiro é efetuado um controle da quantidade de produção de mini-estacas enraizadas e com 120 dias, ou seja, aptas para serem expedidas para o campo. Entretanto, é necessário efetuar a avaliação das variáveis apresentadas na Tabela 3.2, para decidir se as mudas atingiram a qualidade desejada (padrões). O resultado do controle da produção das mudas revelou que no viveiro existem 4.500.017 mudas com 120 dias, divididas em 10 lotes, cada um deles composto por um clone.

O Engenheiro Florestal decidiu efetuar uma amostragem de 0,05% das mudas em cada lote e avaliar as variáveis listadas na Tabela 3.2. Uma amostragem aleatória simples foi efetuada, na qual cada muda de cada lote tinha a mesma chance de compor a amostra.

Os resultados amostrais estão apresentados em um arquivo denominado *mudas*, em formato CSV.

Classificação das variáveis

O conjunto de dados, conforme o exemplo descrito, é composto de informações de diversas naturezas, as quais podem ser classificados em variáveis qualitativas e quantitativas e também nas escalas nominal, ordinal, intervalo e razão. O exercício aqui proposto visa explicar o motivo da classificação das variáveis na Tabela 3.3.

TABELA 3.3. Prática: variáveis e padrões estabelecidos para avaliação da qualidade de mudas de eucalipto.

Variável	Tipo de variável	Escala de medida	Motivos
Diâmetro			
Altura			
Folhas			
Raíz			
Patologia			
Severidade			
Nutrição			
Praga			
Rusticidade			

Tabelas de dados

As tabelas são utilizadas para anotar, organizar e armazenar os dados de um experimento e, também, para representá-los em forma final para serem incluídos, por exemplo, em um artigo ou relatório.

A maioria dos dados coletados em um experimento é chamada de dados brutos, ou seja, não foram tratados estatisticamente. Estes são apenas registrados em uma planilha na forma de uma tabela, os quais precisam ser organizados, resumidos e reestruturados em uma tabela ou gráfico final, com o objetivo de disponibilizá-los aos usuários.

A estrutura da tabela de dados deve permitir que os dados sejam organizados para análises adicionais, ou seja, que grandes quantidades de dados brutos possam ser classificados e/ou reorganizados. Assim, prioriza-se construir tabelas de modo que se possa apresentar os resultados a serem exibidos em colunas, as quais se referem às variáveis e em linhas com os seus respectivos valores.

O uso de tabelas, visando apresentar os resultados dos dados que foram submetidos a tratamento estatístico, devem priorizar o destaque de valores de dados específicos, bem como das comparações entre variáveis, como, por exemplo, apresentações de subtotais, de totais, de médias, de percentagens, de frequências, de resultados estatísticos, entre outros. Muitas vezes as tabelas não são a melhor opção quando o intuito é destacar tendências ou fazer comparações, porque estas podem ser melhor apresentadas em um gráfico. Lembrem-se: nas boas tabelas a leitura das linhas e colunas inferiores deve ser de fácil entendimento consulta no texto do seu artigo ou relatório e devem incluir apenas resultados relevantes.

Nas tabelas de dados brutos o conjunto destes é exibido. O objetivo delas é o de facilitar a organização das informações para facilitar a digitalização e obter informações por meio dos dados.

Na construção de tabelas de dados prioriza-se os seguintes aspectos:

- **Organização:** nas tabelas de dados o seu conteúdo deve ser disponibilizado com lógica, como hierarquia ou alfabetização.
- **Interação:** as tabelas de dados devem permitir a interação do usuário, para que possa personalizar como os dados são exibidos, ou mesmo obter conjuntos específicos deles.

- **Intuitividade:** As tabelas de dados devem ser de fácil uso, com uma estrutura lógica que facilite a compreensão do seu conteúdo.

Como exemplo de como os dados são armazenados e tratados no computador, será utilizada a tabela de dados do exercício *viveiro florestal com produção anual de 10 milhões de mudas*, como uma maneira comum de organizar as informações coletadas (dados de texto ou numéricos) em estrutura retangular da tabela.

Os dados foram organizados da seguinte maneira:

- **Colunas:** nas colunas foram armazenados os nomes das variáveis: Amostra, Clone, d, h, f, r, Patologia, Severidade, Nutrição, Praga, Rusticidade.
- **Linhas:** cada linha corresponde aos dados de uma muda amostrada.
- **Tamanho da tabela:** 11 colunas e 2.000 linhas.

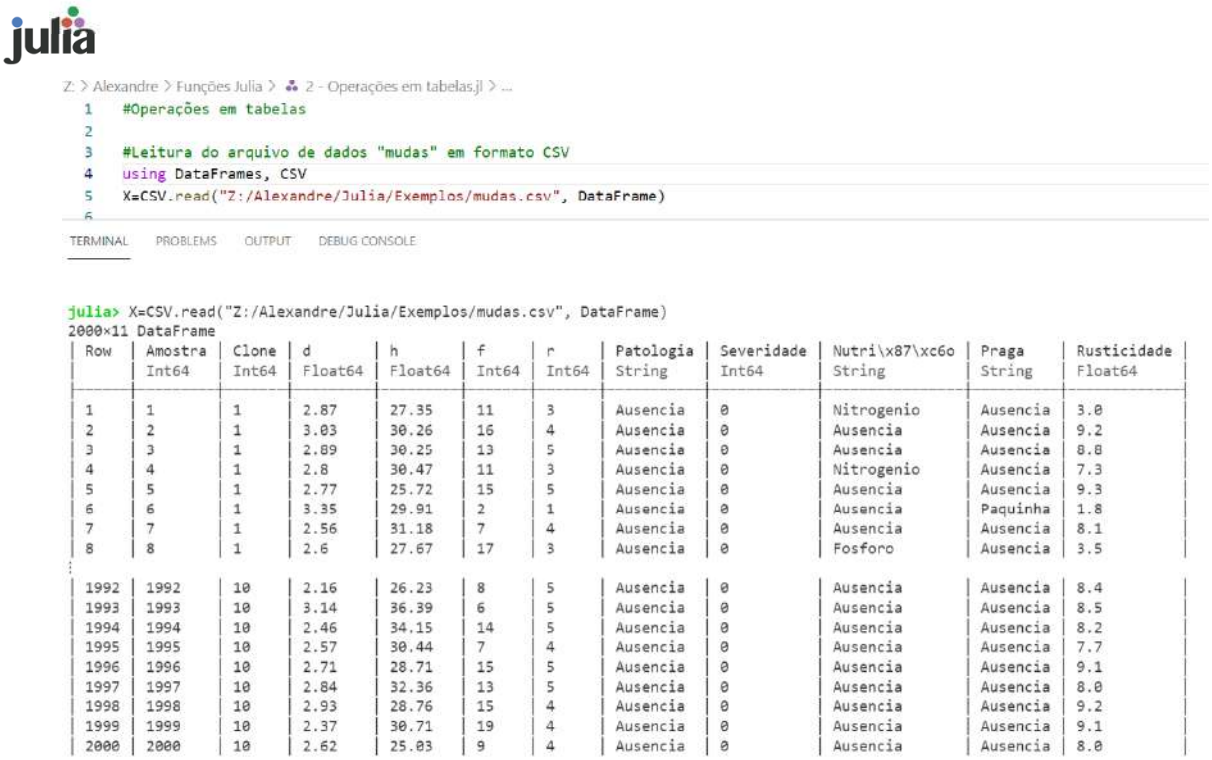
Muitas das informações armazenadas em computadores são usadas nesta estrutura de tabela. O número de colunas (que representam as variáveis) geralmente é pequeno, mas o número de linhas pode ser bastante grande – milhares ou milhões. O armazenamento de dados em uma planilha geralmente usa essa estrutura de tabela.

Existem várias operações de tratamento que podem ser efetuadas em tabelas de dados. Algumas delas estão apresentadas na sequência, as quais são úteis para revelar informações para o Engenheiro Florestal compor relatórios, como, por exemplo, sobre a qualidade das mudas aptas a serem expedidas para plantio em campo.

Utilizando a linguagem Julia, a tabela de dados será armazenada na variável *X*. Cumpre salientar que a linguagem Julia utiliza um ponto para separar as casas decimais.

ARQUIVO: 2 - OPERAÇÕES EM TABELA.TXT

```
> using DataFrames, CSV
> X=CSV.read("Z:/Alexandre/Julia/Exemplos/mudas.csv", DataFrame)
```



Row	Amostra	Clone	d	h	f	r	Patologia	Severidade	Nutri\x87\xc6o	Praga	Rusticidade
	Int64	Int64	Float64	Float64	Int64	Int64	String	Int64	String	String	Float64
1	1	1	2.87	27.35	11	3	Ausencia	0	Nitrogenio	Ausencia	3.0
2	2	1	3.03	30.26	16	4	Ausencia	0	Ausencia	Ausencia	9.2
3	3	1	2.89	30.25	13	5	Ausencia	0	Ausencia	Ausencia	8.8
4	4	1	2.8	30.47	11	3	Ausencia	0	Nitrogenio	Ausencia	7.3
5	5	1	2.77	25.72	15	5	Ausencia	0	Ausencia	Ausencia	9.3
6	6	1	3.35	29.91	2	1	Ausencia	0	Ausencia	Paquinha	1.8
7	7	1	2.56	31.18	7	4	Ausencia	0	Ausencia	Ausencia	8.1
8	8	1	2.6	27.67	17	3	Ausencia	0	Fosforo	Ausencia	3.5
⋮											
1992	1992	10	2.16	26.23	8	5	Ausencia	0	Ausencia	Ausencia	8.4
1993	1993	10	3.14	36.39	6	5	Ausencia	0	Ausencia	Ausencia	8.5
1994	1994	10	2.46	34.15	14	5	Ausencia	0	Ausencia	Ausencia	8.2
1995	1995	10	2.57	30.44	7	4	Ausencia	0	Ausencia	Ausencia	7.7
1996	1996	10	2.71	28.71	15	5	Ausencia	0	Ausencia	Ausencia	9.1
1997	1997	10	2.84	32.36	13	5	Ausencia	0	Ausencia	Ausencia	8.0
1998	1998	10	2.93	28.76	15	4	Ausencia	0	Ausencia	Ausencia	9.2
1999	1999	10	2.37	30.71	19	4	Ausencia	0	Ausencia	Ausencia	9.1
2000	2000	10	2.62	25.03	9	4	Ausencia	0	Ausencia	Ausencia	8.0

O comando para se obter os nomes das colunas na tabela de dados é:

```
> names(X)
```

```
julia> names(X)
11-element Vector{String}:
 "Amostra"
 "Clone"
 "d"
 "h"
 "f"
 "r"
 "Patologia"
 "Severidade"
 "Nutri\x87\xc6o"
 "Praga"
 "Rusticidade"
```

Para se obter o número de linhas e colunas na tabela de dados:

```
> nrow(X), ncol(X)
```

```
julia> nrow(X), ncol(X)
(2000, 11)
```

Para imprimir as primeiras 7 linhas da tabela de dados:

```
> first(X, 7)
```

```
julia> first(X, 7)
7×11 DataFrame
```

Row	Amostra	Clone	d	h	f	r	Patologia	Severidade	Nutri\x87\xc6o	Praga	Rusticidade
	Int64	Int64	Float64	Float64	Int64	Int64	String	Int64	String	String	Float64
1	1	1	2.87	27.35	11	3	Ausencia	0	Nitrogenio	Ausencia	3.0
2	2	1	3.03	30.26	16	4	Ausencia	0	Ausencia	Ausencia	9.2
3	3	1	2.89	30.25	13	5	Ausencia	0	Ausencia	Ausencia	8.8
4	4	1	2.8	30.47	11	3	Ausencia	0	Nitrogenio	Ausencia	7.3
5	5	1	2.77	25.72	15	5	Ausencia	0	Ausencia	Ausencia	9.3
6	6	1	3.35	29.91	2	1	Ausencia	0	Ausencia	Paquinha	1.8
7	7	1	2.56	31.18	7	4	Ausencia	0	Ausencia	Ausencia	8.1

Para imprimir as últimas 7 últimas linhas:

```
> last(X, 7)
```

```
julia> last(X, 7)
7×11 DataFrame
```

Row	Amostra	Clone	d	h	f	r	Patologia	Severidade	Nutri\x87\xc6o	Praga	Rusticidade
	Int64	Int64	Float64	Float64	Int64	Int64	String	Int64	String	String	Float64
1	1994	10	2.46	34.15	14	5	Ausencia	0	Ausencia	Ausencia	8.2
2	1995	10	2.57	30.44	7	4	Ausencia	0	Ausencia	Ausencia	7.7
3	1996	10	2.71	28.71	15	5	Ausencia	0	Ausencia	Ausencia	9.1
4	1997	10	2.84	32.36	13	5	Ausencia	0	Ausencia	Ausencia	8.0
5	1998	10	2.93	28.76	15	4	Ausencia	0	Ausencia	Ausencia	9.2
6	1999	10	2.37	30.71	19	4	Ausencia	0	Ausencia	Ausencia	9.1
7	2000	10	2.62	25.03	9	4	Ausencia	0	Ausencia	Ausencia	8.0

Para se obter informações resumidas sobre a tabela de dados, pode-se utilizar a função describe:

```
> describe(X)
```

```
julia> describe(X)
11×8 DataFrame
 Row │ variable    mean      min       median  max      nunique  nmissing  eltype
     │ Symbol      Union…    Any       Union…  Any      Union…   Nothing   DataType
─────┼──────────────────────────────────────────────────────────────────────────────
   1 │ Amostra     1000.5    1         1000.5  2000                        Int64
   2 │ Clone       6.159     1         7.0     10                          Int64
   3 │ d           2.79258   1.03      2.81    4.57                        Float64
   4 │ h           28.0155   15.03     28.16   42.08                       Float64
   5 │ f           11.8835   1         12.0    30                          Int64
   6 │ r           3.9545    1         4.0     5                           Int64
   7 │ Patologia             Ausencia          Fungo    3                  String
   8 │ Severidade  0.06      0         0.0     3                           Int64
   9 │ Nutri\x87\xc6o        Ausencia          fosforo  7                  String
  10 │ Praga                 Ausencia          Paquinha 5                  String
  11 │ Rusticidade 7.5849    -0.4      8.5     10.3                        Float64
```

Para se obter uma determinada coluna, por exemplo, a coluna clone:

```
> X.Clone
```

```
julia> X.Clone
2000-element Vector{Int64}:
  1
  1
  1
  1
  1
  1
  1
  1
  ⋮
 10
 10
 10
 10
 10
 10
 10
```

Para se obter um subconjunto da tabela de dados, contendo as 7 primeiras linhas e as 4 primeiras colunas:

```
> X[1:7, 1:4]
```

```
julia> X[1:7, 1:4]
7×4 DataFrame
 Row │ Amostra  Clone  d        h
     │ Int64    Int64  Float64  Float64
─────┼──────────────────────────────────
   1 │ 1        1      2.87     27.35
   2 │ 2        1      3.03     30.26
   3 │ 3        1      2.89     30.25
   4 │ 4        1      2.8      30.47
   5 │ 5        1      2.77     25.72
   6 │ 6        1      3.35     29.91
   7 │ 7        1      2.56     31.18
```

Para se criar uma tabela de dados classificada pela variável *d*.

```
> sort(X, :d)
```

```
julia> sort(X, :d)
2000×11 DataFrame
```

Row	Amostra Int64	Clone Int64	d Float64	h Float64	f Int64	r Int64	Patologia String	Severidade Int64	Nutri\x87\xc6o String	Praga String	Rusticidade Float64
1	771	6	1.03	34.96	16	5	Ausencia	0	Ausencia	Ausencia	9.3
2	1386	8	1.13	23.15	9	4	Ausencia	0	Ausencia	Ausencia	8.3
3	1300	7	1.21	28.74	20	5	Ausencia	0	Ausencia	Ausencia	9.4
4	927	7	1.22	26.5	8	4	Ausencia	0	Ausencia	Ausencia	8.1
5	1015	7	1.26	21.13	15	5	Ausencia	0	Ausencia	Ausencia	9.1
⋮											
1995	77	1	4.1	22.48	15	3	Ausencia	0	Fosforo	Ausencia	1.6
1996	1025	7	4.12	28.29	14	4	Ausencia	0	Potassio	Ausencia	9.6
1997	464	4	4.19	20.27	16	5	Ausencia	0	Ausencia	Ausencia	9.3
1998	547	5	4.31	35.86	10	5	Ausencia	0	Ausencia	Ausencia	8.4
1999	1147	7	4.35	29.24	18	4	Ausencia	0	Ausencia	Ausencia	9.1
2000	1925	10	4.57	24.46	9	5	Ausencia	0	Ausencia	Ausencia	7.7

Para retornar uma nova tabela de dados com apenas linhas que atendam a condição de valores maiores que 4 para a variável *d*:

```
> filter(row -> row.d > 4, X)
```

```
julia> filter(row -> row.d > 4, X)
12×11 DataFrame
```

Row	Amostra Int64	Clone Int64	d Float64	h Float64	f Int64	r Int64	Patologia String	Severidade Int64	Nutri\x87\xc6o String	Praga String	Rusticidade Float64
1	77	1	4.1	22.48	15	3	Ausencia	0	Fosforo	Ausencia	1.6
2	133	2	4.06	26.0	16	5	Ausencia	0	Ausencia	Ausencia	9.2
3	444	4	4.02	31.61	8	5	Ausencia	0	Ausencia	Ausencia	8.2
4	464	4	4.19	20.27	16	5	Ausencia	0	Ausencia	Ausencia	9.3
5	547	5	4.31	35.86	10	5	Ausencia	0	Ausencia	Ausencia	8.4
6	880	6	4.03	27.3	10	4	Ausencia	0	Ausencia	Ausencia	8.4
7	1025	7	4.12	28.29	14	4	Ausencia	0	Potassio	Ausencia	9.6
8	1147	7	4.35	29.24	18	4	Ausencia	0	Ausencia	Ausencia	9.1
9	1148	7	4.01	30.9	14	4	Ausencia	0	Ausencia	Ausencia	9.1
10	1553	8	4.08	24.75	17	5	Ausencia	0	Ausencia	Ausencia	8.9
11	1586	8	4.01	27.46	12	5	Ausencia	0	Ausencia	Ausencia	7.8
12	1925	10	4.57	24.46	9	5	Ausencia	0	Ausencia	Ausencia	7.7

Para retornar uma nova tabela de dados com apenas linhas que atendam a condição de valores maiores que 3.5 para a variável d e valores maiores do que 25 e menores que 26 para a variável *h*:

```
> X[(X.d .>3.5) .& (25 .< X.h .<26), :]
```

```
julia> X[(X.d .> 3.5) .& (25 .< X.h .< 26), :]
13×11 DataFrame
```

Row	Amostra Int64	Clone Int64	d Float64	h Float64	f Int64	r Int64	Patologia String	Severidade Int64	Nutri\x87\xc6o String	Praga String	Rusticidade Float64
1	330	4	3.7	25.52	13	4	Ausencia	0	Ausencia	Ausencia	8.5
2	334	4	4.0	25.86	7	5	Ausencia	0	Ausencia	Ausencia	9.1
3	562	5	3.99	25.34	17	4	Ausencia	0	Ausencia	Ausencia	9.4
4	687	6	3.79	25.03	11	3	Ausencia	0	Nitrogenio	Ausencia	5.6
5	718	6	3.56	25.4	19	4	Ausencia	0	Ausencia	Ausencia	9.2
6	897	6	3.51	25.64	19	5	Ausencia	0	Ausencia	Ausencia	8.9
7	898	6	3.72	25.07	8	5	Ausencia	0	Ausencia	Ausencia	8.3
8	1072	7	3.58	25.74	14	5	Ausencia	0	Ausencia	Ausencia	7.4
9	1290	7	3.69	25.64	16	4	Ausencia	0	Ausencia	Ausencia	9.0
10	1373	8	3.78	25.46	6	5	Ausencia	0	Ausencia	Ausencia	9.1
11	1440	8	3.61	25.6	20	5	Ausencia	0	Ausencia	Ausencia	8.7
12	1515	8	3.83	25.4	26	4	Ausencia	0	Potassio	Ausencia	2.6
13	1774	9	3.54	25.35	5	2	Ausencia	0	Ausencia	Ausencia	8.2

Para retornar um subconjunto específico de valores, como *Formiga* para a variável *Praga*:

```
> X[in.(X.Praga, Ref(["Formiga"])), :]
```

```
julia> X[in.(X.Praga, Ref(["Formiga"])), :]
46×11 DataFrame
```

Row	Amostra	Clone	d	h	f	r	Patologia	Severidade	Nutri\x87\xc6o	Praga	Rusticidade
	Int64	Int64	Float64	Float64	Int64	Int64	String	Int64	String	String	Float64
1	18	1	2.72	26.09	1	3	Ausencia	0	fosforo	Formiga	6.4
2	22	1	2.65	35.9	1	1	Fungo	3	Ausencia	Formiga	1.7
3	99	2	3.07	23.1	3	2	Ausencia	0	Ausencia	Formiga	2.2
4	102	2	2.22	21.85	1	1	Bacteria	3	Ausencia	Formiga	2.3
5	105	2	2.74	23.93	2	1	Ausencia	0	Ausencia	Formiga	1.8
⋮											
41	1877	9	3.69	22.11	1	1	Bacteria	2	Ausencia	Formiga	3.2
42	1891	10	2.88	21.69	1	2	Ausencia	0	Ausencia	Formiga	2.1
43	1893	10	2.66	25.57	1	2	Ausencia	0	Ausencia	Formiga	2.4
44	1894	10	2.39	26.83	1	2	Ausencia	0	Ausencia	Formiga	1.0
45	1914	10	3.13	24.27	1	1	Bacteria	3	Ausencia	Formiga	1.4
46	1972	10	3.39	30.81	1	1	Bacteria	2	Ausencia	Formiga	2.8

Um subconjunto específico de valores, como 2, para a variável *f*:

```
> X[in.(X.f, Ref([2])), :]
```

```
julia> X[in.(X.f, Ref([2])), :]
37×11 DataFrame
```

Row	Amostra	Clone	d	h	f	r	Patologia	Severidade	Nutri\x87\xc6o	Praga	Rusticidade
	Int64	Int64	Float64	Float64	Int64	Int64	String	Int64	String	String	Float64
1	6	1	3.35	29.91	2	1	Ausencia	0	Ausencia	Paquinha	1.8
2	30	1	3.37	26.74	2	2	Ausencia	2	Ausencia	Paquinha	2.0
3	105	2	2.74	23.93	2	1	Ausencia	0	Ausencia	Formiga	1.8
4	140	2	2.96	30.27	2	1	Ausencia	0	Ausencia	Colchonilha	1.3
5	207	2	3.52	29.82	2	1	Ausencia	0	Ausencia	Paquinha	1.7
6	257	3	2.36	27.38	2	2	Fungo	2	Ausencia	Paquinha	2.0
⋮											
31	1658	8	2.91	22.88	2	2	Ausencia	0	Ausencia	Colchonilha	1.8
32	1665	8	3.14	33.58	2	1	Ausencia	0	Ausencia	Colchonilha	2.5
33	1758	9	2.52	33.46	2	1	Ausencia	0	Ausencia	Colchonilha	2.2
34	1777	9	2.69	26.15	2	2	Fungo	2	Ausencia	Paquinha	2.0
35	1794	9	2.68	29.08	2	2	Ausencia	0	Ausencia	Paquinha	1.7
36	1867	9	2.97	24.68	2	1	Ausencia	0	Ausencia	Colchonilha	2.0
37	1953	10	2.87	31.72	2	2	Fungo	2	Ausencia	Paquinha	1.5

Para calcular a soma da variável *h* por *Clone* e armazenar o resultado na coluna x:

```
> by(X, :Clone, x = :h=>sum)
```

```
julia> by(X, :Clone, x = :h=>sum)
10×2 DataFrame
```

Row	Clone	x
	Int64	Float64
1	1	2720.31
2	2	4183.59
3	3	1029.09
4	4	7188.3
5	5	3660.0
6	6	6675.76
7	7	11884.6
8	8	10582.8
9	9	4963.95
10	10	3142.72

Para ordenar os valores de uma variável:

```
> sort!(X.Patologia)
```

```
julia> sort!(X.Patologia)
2000-element PooledArrays.PooledVector{String, UInt32, Vector{UInt32}}:
 "Ausencia"
 "Ausencia"
 "Ausencia"
 "Ausencia"
 "Ausencia"
 "Ausencia"
 "Ausencia"
 "Ausencia"
 ⋮
 "Fungo"
 "Fungo"
 "Fungo"
 "Fungo"
 "Fungo"
 "Fungo"
 "Fungo"
```

Para converter as colunas de 1 a 7 em uma matriz:

```
> Matrix(X[:, 1:7])
```

```
julia> Matrix(X[:, 1:7])
2000×7 Matrix{Any}:
    1   1  2.87  27.35  11  3  "Ausencia"
    2   1  3.03  30.26  16  4  "Ausencia"
    3   1  2.89  30.25  13  5  "Ausencia"
    4   1  2.8   30.47  11  3  "Ausencia"
    5   1  2.77  25.72  15  5  "Ausencia"
    6   1  3.35  29.91   2  1  "Ausencia"
    7   1  2.56  31.18   7  4  "Ausencia"
    8   1  2.6   27.67  17  3  "Ausencia"
    ⋮                          ⋮
 1994  10  2.46  34.15  14  5  "Fungo"
 1995  10  2.57  30.44   7  4  "Fungo"
 1996  10  2.71  28.71  15  5  "Fungo"
 1997  10  2.84  32.36  13  5  "Fungo"
 1998  10  2.93  28.76  15  4  "Fungo"
 1999  10  2.37  30.71  19  4  "Fungo"
 2000  10  2.62  25.03   9  4  "Fungo"
```

Algumas estatísticas também podem ser calculadas. A sequência de comandos seria:

```
> using Statistics
```

Para ordenar os valores com base nas variáveis *Clone* e *Rusticidade*:

```
> sort!(X, [:Clone, :Rusticidade])
```

```
julia> sort!(X, [:Clone, :Rusticidade])
2000×11 DataFrame
 Row   Amostra  Clone  d        h        f      r      Patologia  Severidade  Nutri\x87\xc6o  Praga     Rusticidade
       Int64    Int64  Float64  Float64  Int64  Int64  String     Int64       String          String    Float64
-------------------------------------------------------------------------------------------------------------------
 1     65       1      2.95     30.93    18     3      Ausencia   0           Nitrogenio      Ausencia  0.4
 2     45       1      2.17     21.83    4      3      Ausencia   0           Potassio        Ausencia  1.4
 3     61       1      2.53     30.41    23     3      Ausencia   0           Potassio        Ausencia  1.5
 4     77       1      4.1      22.48    15     3      Ausencia   0           Fosforo         Ausencia  1.6
 5     9        1      3.34     33.95    3      3      Ausencia   0           Ausencia        Ausencia  1.7
 ⋮
 1995  1968     10     2.61     31.73    24     4      Fungo      0           Ausencia        Ausencia  9.3
 1996  1980     10     3.22     24.39    19     5      Fungo      0           Ausencia        Ausencia  9.3
 1997  1981     10     2.57     35.03    25     4      Fungo      0           Ausencia        Ausencia  9.3
 1998  1988     10     3.05     25.87    16     5      Fungo      0           Ausencia        Ausencia  9.3
 1999  1915     10     3.2      37.38    15     4      Ausencia   0           Ausencia        Ausencia  9.4
 2000  1933     10     2.95     29.2     21     4      Bacteria   0           Ausencia        Ausencia  9.5
```

Obtenção da média da variável *d* por *Clone*:

```
> by(X, :Clone, :d => mean)
```

```
julia> by(X, :Clone, :d => mean)
10×2 DataFrame
 Row │ Clone  d_mean
     │ Int64  Float64
─────┼────────────────
   1 │     1  2.78051
   2 │     2  2.77733
   3 │     3  2.77778
   4 │     4  2.83685
   5 │     5  2.79511
   6 │     6  2.81143
   7 │     7  2.78047
   8 │     8  2.77204
   9 │     9  2.83317
  10 │    10  2.73378
```

Obtenção da média e da variância da variável *d* por *Clone*:

```
> by(X, :Clone) do df
(m = mean(df.d), s² = var(df.d))
end
```

```
julia> by(X, :Clone) do df
       (m = mean(df.d), s² = var(df.d))
       end
10×3 DataFrame
 Row │ Clone  m        s²
     │ Int64  Float64  Float64
─────┼─────────────────────────
   1 │     1  2.78051  0.202211
   2 │     2  2.77733  0.250787
   3 │     3  2.77778  0.216126
   4 │     4  2.83685  0.238103
   5 │     5  2.79511  0.256622
   6 │     6  2.81143  0.204109
   7 │     7  2.78047  0.24814
   8 │     8  2.77204  0.274116
   9 │     9  2.83317  0.211478
  10 │    10  2.73378  0.22012
```

Para obtenção do número de valores para cada categoria da variável *Patologia*:

```
> by(X, :Patologia, N = :Patologia => length)
```

```
julia> by(X, :Patologia, N = :Patologia => length)
3×2 DataFrame
 Row │ Patologia  N
     │ String     Int64
─────┼──────────────────
   1 │ Ausencia   1928
   2 │ Bacteria   39
   3 │ Fungo      33
```

Para obtenção da altura (h) média por categoria da variável *Patologia*:

```
> by(X, :Patologia, N = :Patologia => length, Altura = :h => mean)
```

```
julia> by(X, :Patologia, N = :Patologia => length, Altura = :h => mean)
3×3 DataFrame
 Row │ Patologia  N      Altura
     │ String     Int64  Float64
─────┼──────────────────────────
   1 │ Ausencia   1928   27.9998
   2 │ Bacteria   39     27.929
   3 │ Fungo      33     29.0394
```

Para identificar as categorias da variável *Praga*:

```
> levels(X.Praga)
```

```
julia> levels(X.Praga)
5-element Vector{String}:
 "Ausencia"
 "Colchonilha"
 "Formiga"
 "Lagarta"
 "Paquinha"
```

Observe que para nutrição existem dois tipos de escrita para o fósforo: Fosforo e fosforo, devido a um erro de digitação. A linguagem considera cada uma das escritas como únicas. A linguagem Julia, devido ter sido desenvolvida na linguagem inglesa, não permite acentuação e caracteres especiais nas palavras nos conjuntos de dados em CSV. Isso pode ser observado na coluna 9, referente a variável "Nutrição". Assim, os níveis para esta coluna podem ser obtidos conforme:

```
> levels(X[:,9])
```

```
julia> levels(X[:,9])
7-element Vector{String}:
 "Ausencia"
 "Calcio"
 "Fosforo"
 "Magnesio"
 "Nitrogenio"
 "Potassio"
 "fosforo"
```

Para substituir a categoria fosforo por Fosforo e atribuir os novos valores em uma nova variável o comando *replace* pode ser utilizado

```
> Nutrição=replace!(X[:,9], "fosforo" => "Fosforo")
```

```
julia> Nutrição=replace!(X[:,9], "fosforo" => "Fosforo")
2000-element PooledArrays.PooledVector{String, UInt32, Vector{UInt32}}:
 "Nitrogenio"
 "Ausencia"
 "Ausencia"
 "Nitrogenio"
 "Ausencia"
 "Ausencia"
 "Ausencia"
 "Fosforo"
 ⋮
 "Ausencia"
 "Ausencia"
 "Ausencia"
 "Ausencia"
 "Ausencia"
 "Ausencia"
 "Ausencia"
```

```
> levels(Nutrição)
```

```
julia> levels(Nutrição)
6-element Vector{String}:
 "Ausencia"
 "Calcio"
 "Fosforo"
 "Magnesio"
 "Nitrogenio"
 "Potassio"
```

Com base nos exemplos anteriores, pede-se:

1. Qual é o tamanho da amostra de cada clone?
2. Calcule as estatísticas valor máximo, mínimo e média para as variáveis *d, h* e *f* para todo o conjunto de dados. Calcule essas estatísticas também por clone.
3. Quais são as categorias de respostas para as variáveis: *r, Patologia, Severidade* e *Nutrição, Praga*?
4. Quais são os clones que foram identificados com a ocorrência de alguma nutrição?
5. Quais são os clones em que foi identificada a ocorrência de alguma patologia?
6. Para a variável patologia, quais foram os níveis de classificação da severidade da doença? Qual a percentagem total de mudas atacadas e por clone? Qual a percentagem por nível de classificação da severidade da doença por clone?
7. Quais são os minerais indicados para a aplicação nas mudas por clone, concernentes à variável nutrição?
8. Qual é a percentagem de mudas com valores 0 (que não atingiram os padrões esperados) e qual é a percentagem de mudas com valores 10 (padrão perfeito) por clone, concernentes à variável rusticidade?

Os padrões exigidos de qualidade das mudas para expedição ao campo são:

- Diâmetro do colo acima de 2 mm;
- Altura entre 25 e 35 cm;
- Mudas com pelo menos três pares de folhas maduras;
- Sistema radicular desenvolvido, sem enovelamento, com raiz principal reta, raízes secundárias bem distribuídas e agregadas ao substrato;
- Bom aspecto fitossanitário, sem deficiências minerais e isentas de pragas e doenças;
- Mudas rústicas, com notas superiores a 8.

Com base nesses padrões calcule a percentagem de mudas que podem ser expedidas para campo por clone. Com base nos valores da quantidade de mudas produzidas no viveiro e na percentagem de mudas que atendem ao padrão desejado, quantas mudas no viveiro podem ser expedidas para campo?

EXERCÍCIO PROPOSTO 2: INVENTÁRIO FLORESTAL DE ACÁCIA-NEGRA

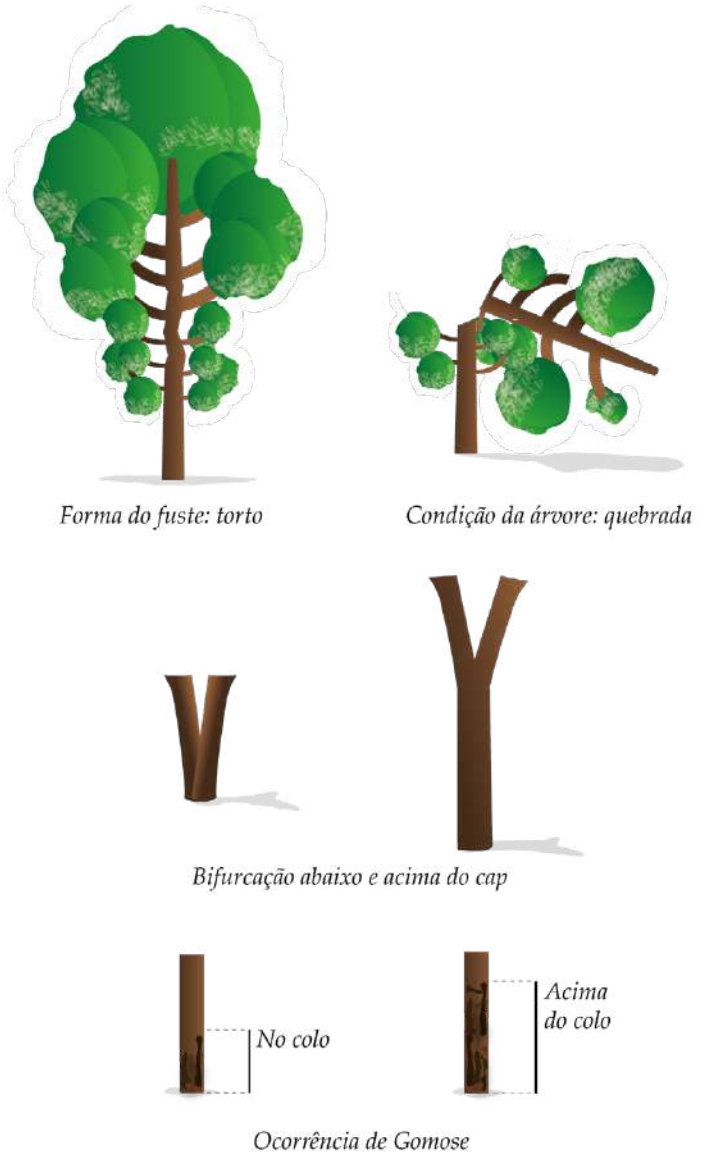
Um inventário florestal foi efetuado em junho de 2014, em um povoamento comercial de acácia -negra (*Acacia mearnsii* de Wild.), situado na região de Piratini, Rio Grande do Sul. O plantio foi realizado com mudas produzidas a partir de sementes utilizando um espaçamento de 3 m x 1,5 m.

De maneira aleatória, parcelas circulares com área de 450 m² foram instaladas. Nas parcelas foram feitas avaliações das variáveis listadas na Tabela 3.4. Em cada parcela foram avaliadas as covas existentes, com objetivo de identificar as falhas de plantio, e a presença de árvores vivas ou mortas. A variável circunferência à altura do peito foi mensurada com uma fita métrica e efetuada para todas as árvores. A altura foi obtida com um hipsômetro e realizada somente para as árvores das 4 primeiras linhas das parcelas. Todas as árvores das parcelas foram avaliadas visualmente para obtenção das variáveis qualitativas. Os dados obtidos estão armazenados no arquivo denominado inventario em formato CSV. Observe que para as árvores que não se enquadraram nos valores de classificação das variáveis, conforme os níveis definidos, não foram atribuídos nenhum valor.

TABELA 3.4. Variáveis medidas no inventário florestal de acácia-negra.

Variável	Descrição	Níveis
X1	Circunferência à altura do peito (cap)	Domínio real maior que zero
X2	Altura (m)	Domínio real maior que zero
X3	Classificação da cova	0: Falha, 1: árvore viva, 2: árvore morta
X4	Forma do fuste	1: reto, 2: torto
X5	Condição da árvore	1: sem defeitos, 2: quebrada
X6	Bifurcação	1: acima do cap, 2: abaixo do cap
X7	Árvore atacada pelo cascudo serrador	1
X8	Ocorrência da gomose	1: no colo, 2: acima do colo

FIGURA 3.8. Variáveis coletadas no inventário florestal de acácia-negra.



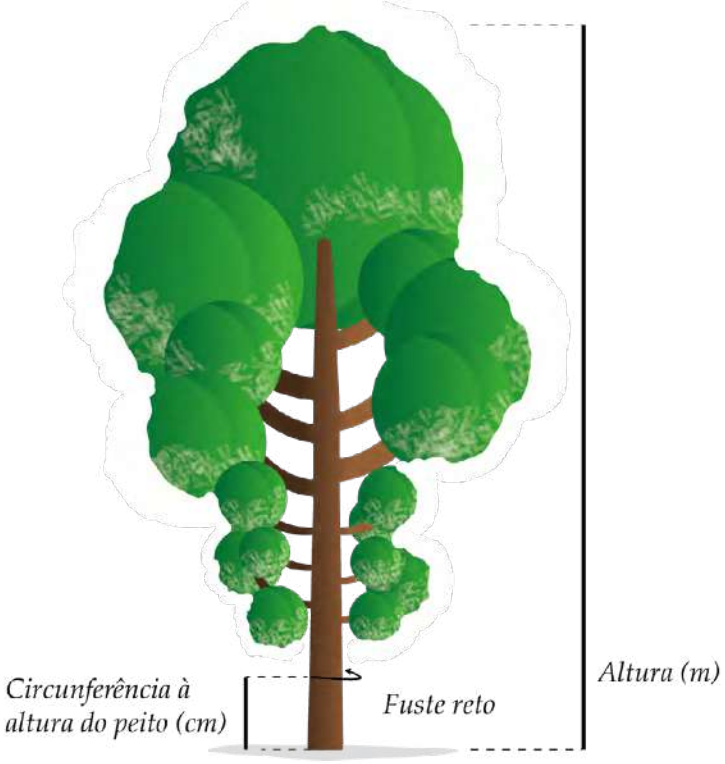
Forma do fuste: torto

Condição da árvore: quebrada

Bifurcação abaixo e acima do cap

Ocorrência de Gomose

FIGURA 3.9. Condição da árvore: sem defeitos.



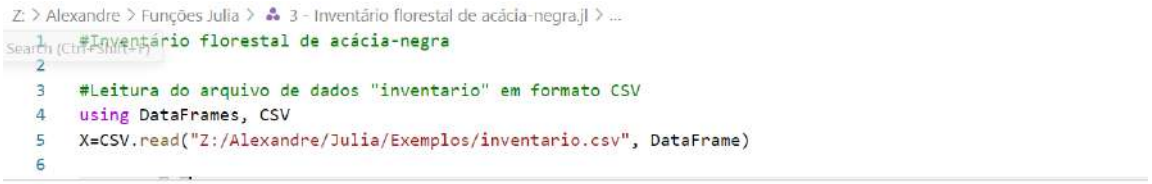
PEDE-SE:

Produza um relatório de inventário florestal, o qual deve conter as seguintes informações:

1. Identifique se existe algum valor estranho (*outlier*) para todas as variáveis. Quantas são as observações consideradas como potenciais *outliers* e para quais variáveis?
2. Quantas parcelas foram instaladas em campo? Qual é o número mínimo, máximo e médio de covas? Qual é o tamanho da tabela de dados?
3. Qual é o número de árvores total por hectare? Qual é o número de árvores vivas por hectare? Qual é o número de árvores mortas por hectare? Qual é o número de falhas por hectare?
4. Qual é o valor mínimo, máximo e médio das variáveis circunferência à altura do peito e altura total?
5. Qual é o número de árvores com fuste reto e fuste torto por hectare?
6. Qual é o número de árvores sem defeitos e de árvores quebradas por hectare?
7. Qual é o número de árvores bifurcadas acima e abaixo da cap por hectare?
8. Qual é o número de árvores atacadas e não atacadas pelo cascudo serrador por hectare?
9. Qual é o número de árvores com presença de gomose por hectare?
10. Qual é o número de árvores com presença de gomose no colo e acima do colo por hectare?

ARQUIVO: 3 - INVENTÁRIO FLORESTAL DE ACÁCIA-NEGRA

```
> using DataFrames, CSV
> X=CSV.read("Z:/Alexandre/Julia/Exemplos/inventario.csv", DataFrame)
```



```
Z: > Alexandre > Funções Julia > 3 - Inventário florestal de acácia-negra.jl > ...
1  #Inventário florestal de acácia-negra
2
3  #Leitura do arquivo de dados "inventario" em formato CSV
4  using DataFrames, CSV
5  X=CSV.read("Z:/Alexandre/Julia/Exemplos/inventario.csv", DataFrame)
6
```

TERMINAL PROBLEMS OUTPUT DEBUG CONSOLE

```
julia> X=CSV.read("Z:/Alexandre/Julia/Exemplos/inventario.csv", DataFrame)
9762×10 DataFrame
```

Row	PARCELA	COVA	X1	X2	X3	X4	X5	X6	X7	X8
	Int64	Int64	Float64?	Float64?	Int64?	Int64	Int64	Int64?	Int64?	Int64?
1	1	1	40.0	missing	2	2	1	missing	missing	missing
2	1	2	missing	missing	0	2	1	missing	missing	missing
3	1	3	missing	missing	0	2	1	missing	missing	missing
4	1	4	52.0	missing	1	2	1	missing	missing	missing
5	1	5	missing	missing	0	2	1	missing	missing	missing
6	1	6	missing	missing	0	2	1	missing	missing	missing
7	1	7	missing	missing	0	2	1	missing	missing	missing
8	1	8	54.5	missing	1	2	1	missing	missing	missing
⋮										
9754	94	103	missing	missing	0	2	1	missing	missing	missing
9755	94	104	42.5	missing	1	2	1	missing	missing	missing
9756	94	105	29.5	missing	2	2	1	missing	missing	missing
9757	94	106	49.5	missing	1	2	2	1	missing	2
9758	94	107	49.5	missing	1	2	1	missing	missing	missing
9759	94	108	67.0	missing	1	2	1	missing	missing	missing
9760	94	109	missing	missing	0	2	1	missing	missing	missing
9761	94	110	missing	missing	0	2	1	missing	missing	missing
9762	94	111	27.5	missing	2	2	1	missing	missing	missing

```
> describe(X)
```

```
julia> describe(X)
10×8 DataFrame
```

Row	variable	mean	min	median	max	nunique	nmissing	eltype
	Symbol	Float64	Real	Float64	Real	Nothing	Union…	Type
1	PARCELA	48.1326	1	49.0	94			Int64
2	COVA	53.2197	1	53.0	127			Int64
3	X1	46.1541	4.0	46.0	98.0		3158	Union{Missing, Float64}
4	X2	17.6211	1.1	18.8	32.6		7341	Union{Missing, Float64}
5	X3	0.845602	0	1.0	2		8	Union{Missing, Int64}
6	X4	1.98105	1	2.0	2			Int64
7	X5	1.03975	1	1.0	2			Int64
8	X6	1.31653	1	1.0	2		9405	Union{Missing, Int64}
9	X7	1.0	1	1.0	1		9735	Union{Missing, Int64}
10	X8	1.15726	1	1.0	2		9514	Union{Missing, Int64}

Note que algumas variáveis contém valores faltantes/perdidos (*missing*). Observa-se que na planilha de dados os valores faltantes estão representados por uma célula vazia no arquivo em .CSV. Para algumas variáveis estes valores devem ser contabilizados e para outras não. Por exemplo, para obter a informação do número de árvores não atacadas pelo cascudo serrador por hectare, deve-se levar em consideração apenas o número de valores faltantes, pois o valor de resposta 1 indica somente as árvores que foram atacadas.

Quando os valores faltantes não devem ser computados, o seguinte comando pode ser utilizado:

```
> cap_ médio=mean(skipmissing(X.X1))
```

```
julia> cap_médio=mean(skipmissing(X.X1))
46.154112658994514
```

Se for necessário, pode-se eliminar os valores faltantes de uma determinada variável:

```
> cap=collect(skipmissing(X.X1))
```

4 DISTRIBUIÇÕES DE PROBABILIDADES BÁSICAS

Existem várias distribuições de probabilidade. As duas básicas para os Engenheiros Florestais são a curva de distribuição Normal de Gauss e a curva de distribuição de *t* de Student, uma vez que procedimentos da estatística descritiva e da estatística inferencial são baseados nestas distribuições.

4.1 DISTRIBUIÇÃO NORMAL

Um modelo fundamental de probabilidades e usado amplamente para inferência estatística é a distribuição de probabilidades Normal, também conhecida como distribuição de Gauss, que se aplica a populações. As principais razões que tornam a distribuição Normal importante para os Engenheiros Florestais são:

1. Muitas das variáveis obtidas em experimentos conduzidos em florestas, viveiros ou em laboratórios, se distribuem segundo a distribuição Normal ou se aproximam da normalidade.
2. Muitos testes e modelos estatísticos têm como pressuposição a normalidade dos valores.
3. A maioria das variáveis contínuas segue uma distribuição normal.

A distribuição normal é definida por dois parâmetros, μ e σ, e é dada por:

$$f_{(x)} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Uma variável aleatória contínua segue uma distribuição Normal, quando ela se comporta com simetria em forma de sino e com o aspecto representado na Figura 4.1. Nesta Figura está apresentada a distribuição da variável altura de mudas de louro pardo. Uma população de 500 mudas foi obtida, na qual o valor mínimo foi de 10 cm, o máximo de 20 cm, a média de 15 cm e o desvio-padrão de 2 cm.

Para obtenção do gráfico de distribuição da altura das mudas, referente as estatísticas calculadas, utiliza-se os seguintes comandos:

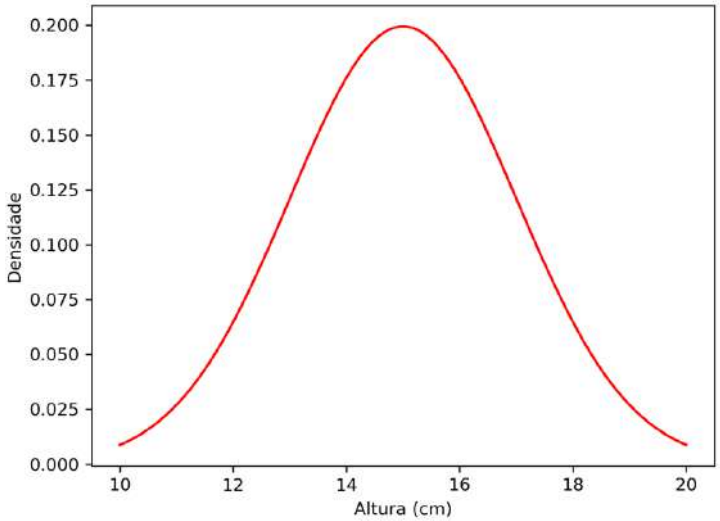
Arquivo: 4 - Distribuição Normal

```
>using PyPlot
>xGrid = 10:0.01:20
>mu, s = 15, 2
>f(x) = (1/(s*sqrt(2*pi))*exp(-0.5*(((x-mu). ^2)/(s ^2))))
```

```
>plot(xGrid,f.(xGrid),"r",label="Distribuição Normal")
>xlabel("Altura (cm)")
>ylabel("Densidade")
```

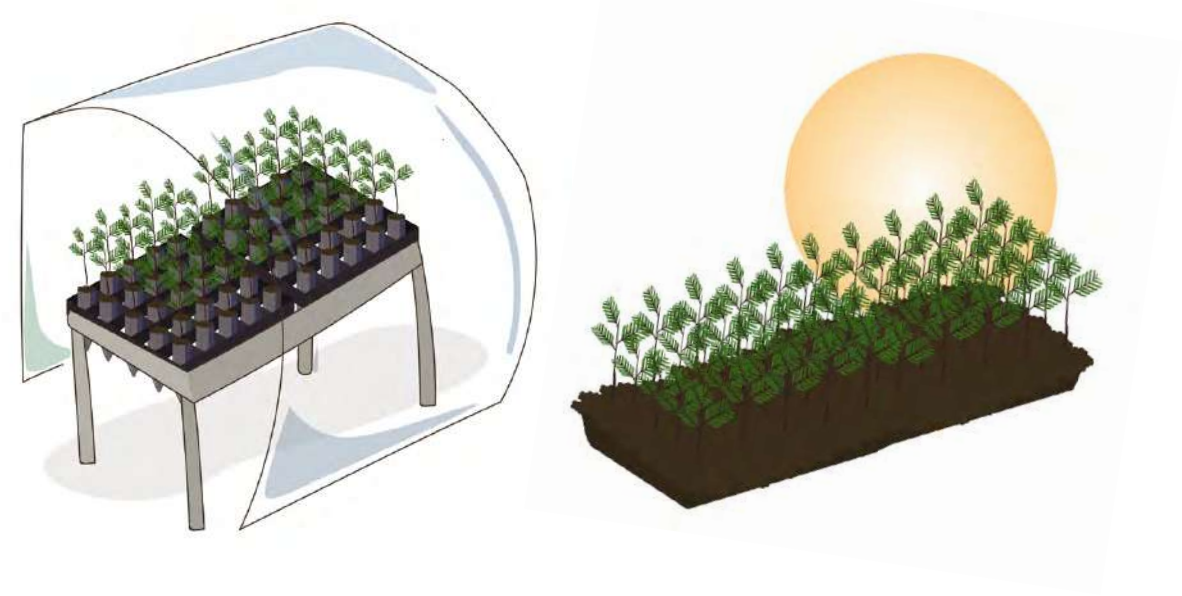
O resultado obtido está apresentado na Figura 4.1

FIGURA 4.1 Distribuição da variável altura de mudas da espécie louro-pardo, 12 meses após a emergência e cultivadas em um viveiro (n=500).



Um viveirista cultivou mudas da espécie guapuruvu por 5 meses em estufa e em ambiente natural (sem proteção para as mudas) (Figura 4.2). Em cada um dos ambientes foram obtidas 350 mudas e medida a variável altura, em cm. Os dados oriundos dos dois ambientes foram analisados para avaliar se seguiam a distribuição normal.

FIGURA 4.2. À esquerda mudas de guapuruvu em estufa e à direita mudas de guapuruvu em ambiente natural.



Para a obtenção da distribuição dos valores das alturas das mudas cultivadas em estufa e em ambiente natural, foram executados os seguintes comandos, obtendo-se o resultado apresentado na Figura 4.3.

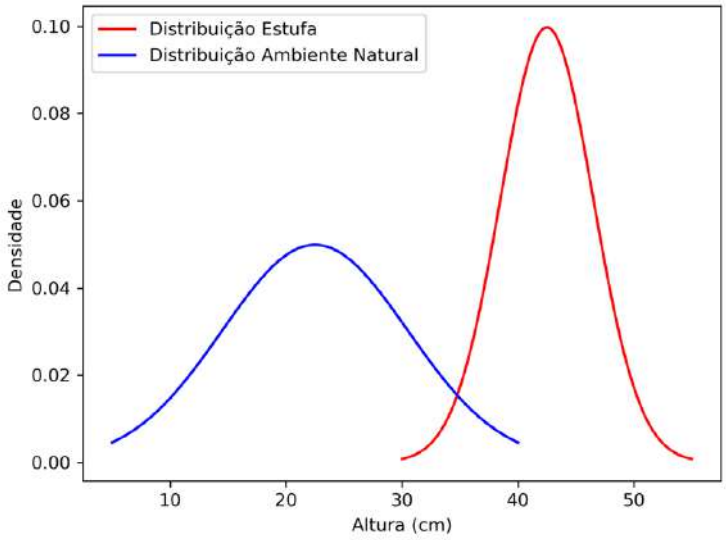
Arquivo: 4 - Distribuição Normal

```
>using PyPlot

>#Guapuruvu cultivado em estufa
>xGrid = 30:0.01:55
>mu, s = 42.5, 4
>f(x) = (1/(s*sqrt(2*pi))*exp(-0.5*(((x-mu).^2)/(s^2))))
>plot(xGrid,f.(xGrid),"r",label="Distribuição Estufa")
>xlabel("Altura (cm)")
>ylabel("Densidade")

>#Guapuruvu cultivado em ambiente natural
>xGrid = 5:0.01:40
>mu, s = 22.5, 8
>f(x) = (1/(s*sqrt(2*pi))*exp(-0.5*(((x-mu).^2)/(s^2))))
>plot(xGrid,f.(xGrid),"b",label="Distribuição Ambiente Natural")
```

FIGURA 4.3. Distribuição de altura de mudas da espécie guapuruvu cultivadas por 5 meses em estufa e, também, em ambiente natural (sem proteção para as mudas).



Note que no ambiente estufa a média foi maior do que a do ambiente natural. A variabilidade dos dados oriundos do ambiente natural foi maior do que a obtida em estufa. Isto pode ser verificado pela maior dispersão dos dados obtidos no ambiente natural (curva mais achatada) do que no ambiente de estufa (curva mais pontiaguda). Ambas seguem a distribuição normal com médias e variâncias diferentes.

4.2 DISTRIBUIÇÃO NORMAL PADRONIZADA

A distribuição normal padrão tem média zero ($\mu = 0$) e variância um ($\sigma^2 = 1$). Para transformar qualquer distribuição normal na distribuição normal padrão (média igual a zero e variância igual a um) utiliza-se a seguinte transformação:

$$f_{(x)} = \frac{1}{\sigma\sqrt{2\,\pi}}\, e^{-\frac{z^2}{2}}$$

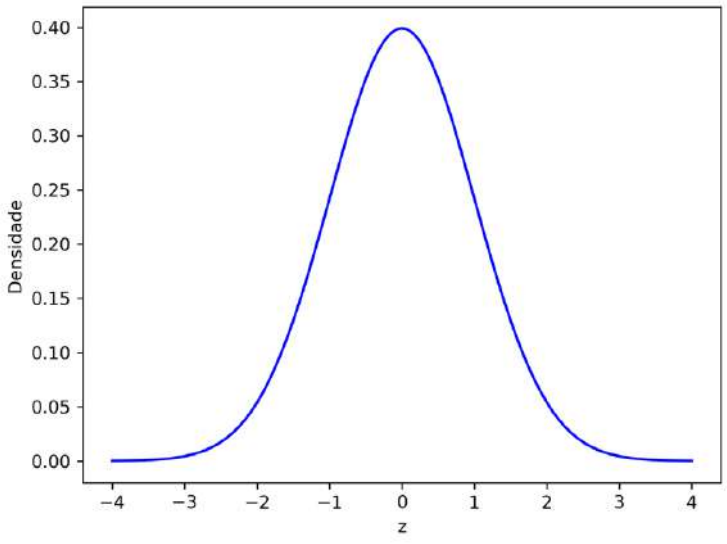
Em que: $z = \frac{x_i - \mu}{\sigma}$

Z é o valor padronizado da variável X e tem $\mu = 0$ e $\sigma = 1$. Isto implica que as variáveis transformadas estão na mesma escala do que a distribuição normal padrão, variando, geralmente, de -3 a 3. A partir destes valores a distribuição torna-se assintótica no eixo x. Seu valor indica quantos desvios-padrões a observação original x_i está afastada da média e, também, sua direção (positiva ou negativa). A forma da distribuição normal padronizada está representada na Figura 4.4, e foi obtida pelos seguintes comandos.

ARQUIVO: 4 - DISTRIBUIÇÃO NORMAL

```
>using PyPlot
>xGrid = -4:0.01:4
>f(x) =( 1 / (sqrt(2*pi)))*exp(-0.5*x.^2)
>plot(xGrid,f.(xGrid),"b",label="Distribuição Normal Padronizada")
>xlabel("z")
>ylabel("Densidade")
```

FIGURA 4.4. Distribuição normal padronizada.



4.2.1 ESCORES Z: MÉDIA ZERO E VARIÂNCIA UM

O escore *z*, também denominado de escore padronizado, que é utilizado para obtenção da curva normal padrão, corresponde ao número de desvios-padrões que um valor *x* se afasta da média na mesma escala do que a distribuição normal padrão. Pode ser calculado como:

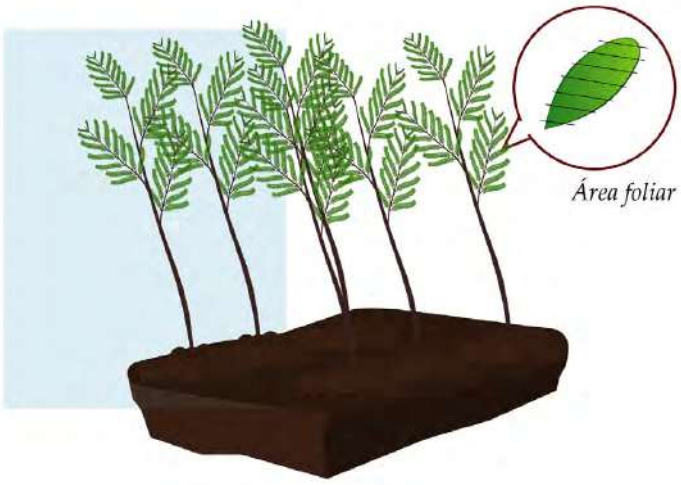
Para a população: $z = \frac{(X - \mu)}{\sigma}$

Para amostras: $z = \frac{(x - \overline{x})}{s}$

Os valores de escores *z* permitem identificar valores comuns e valores raros. Normalmente, considera-se valores comuns escores entre -2,00 e 2,00 e valores raros inferiores a -2,00 e superiores a 2,00.

Considere a população apresentada no vetor X, correspondente a 20 valores de área foliar, em cm², de mudas de acácia-negra com 4 meses (Figura 4.5). A média da área foliar foi de 5,10 cm² e o desvio padrão de 2,096 cm². O resultado para a variável padronizada ($z = (X - \mu) / \sigma$) resultou em média zero e desvio-padrão um. Note que ao se aproximar da média da área foliar os valores de z se aproximam de zero e que todos são menores que dois desvios-padrões.

FIGURA 4.5. Mudas de acácia-negra para avaliação de área foliar.



Os comandos utilizados para a obtenção dos valores originais para a distribuição normal padrão são:

ARQUIVO: 4 – DISTRIBUIÇÃO NORMAL

```
 #Padronização de valores
>X=[ 7.33, 2.01, 2.23, 3.54, 6.48, 5.11, 5.00, 1.70, 5.78, 2.79, 4.89, 6.60, 4.05, 4.82,
9.08, 3.90, 8.45, 7.73, 5.30, 5.31];

>using DataFrames, Statistics
> function z(X)
    X = sort(X)
    Média=mean(X)
    DesvioPadrão=std(X)
```

```
    z=(X.-Média)./DesvioPadrão
    df = DataFrame()
    df.X =X
    df.z=z
    return df
end
>Tabela=z(X)
```

julia

```
julia> Tabela=z(X)
20×2 DataFrame
| Row | X       | z         |
|     | Float64 | Float64   |
|-----|---------|-----------|
| 1   | 1.7     | -1.6248   |
| 2   | 2.01    | -1.47687  |
| 3   | 2.23    | -1.37189  |
| 4   | 2.79    | -1.10467  |
| 5   | 3.54    | -0.746786 |
⋮
| 15  | 6.48    | 0.656122  |
| 16  | 6.6     | 0.713384  |
| 17  | 7.33    | 1.06173   |
| 18  | 7.73    | 1.2526    |
| 19  | 8.45    | 1.59617   |
| 20  | 9.08    | 1.89679   |
```

Na Tabela acima a segunda coluna corresponde aos valores originais e a terceira aos valores transformados. As estatísticas referentes aos valores originais e transformados são obtidas pelos comandos:

Arquivo: 4 – Distribuição normal

```
>Média_X=mean(Tabela.X)
5.10
>Variância_X=var(Tabela.X)
4.39
>Desvio_Padrão_X=std(Tabela.X)
2.09
>Média_z=mean(Tabela.z)
0.00
>Variância_z=var(Tabela.z)
1.00
>Desvio_Padrão_z=std(Tabela.z)
1.00
```

Note que os valores transformados têm média igual a zero e variância igual a um.

EXERCÍCIO PROPOSTO 3: DISTRIBUIÇÃO NORMAL

A densidade básica da madeira de duas espécies foi avaliada: araucária e pau-ferro (Figuras 4.6 - 4.7). Cinquenta amostras de madeira foram tomadas para cada espécie, obtidas na posição diâmetro à altura do peito. As 100 árvores foram selecionadas aleatoriamente em florestas nativas no Rio Grande do Sul, com diâmetro variando de 50 cm a 100 cm, e consideradas como uma população. Os dados estão apresentados a seguir. Construa os gráficos para os valores originais e para os valores padronizados para a distribuição normal padrão, para a variável densidade básica da madeira das duas espécies.

FIGURA 4.6. Obtenção de amostra para avaliação de densidade básica da madeira de Araucária.

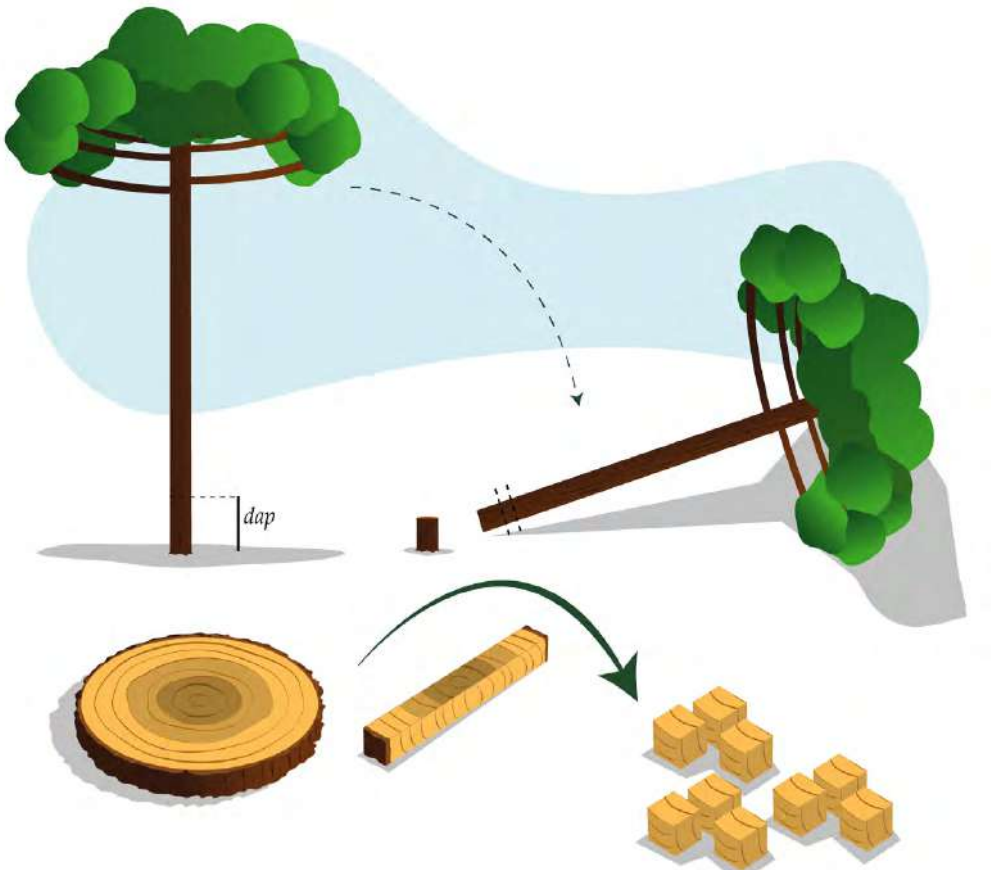
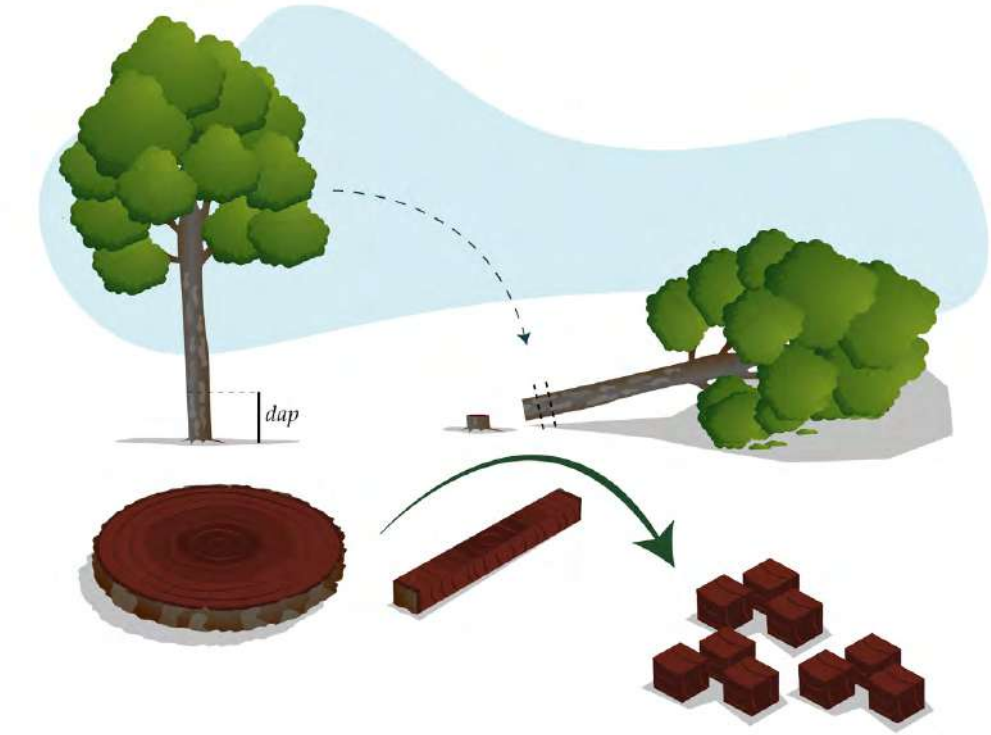


FIGURA 4.7. Obtenção de amostra para avaliação de densidade básica da madeira de Pau-ferro.



Os dados foram digitados utilizando os seguintes comandos:

Arquivo: 4 – Distribuição normal

```
>XAraucaria=[ 0.34, 0.54, 0.39, 0.49, 0.39, 0.32, 0.67, 0.74, 0.46, 0.19, 0.51, 0.16,
0.55, 0.40, 0.26, 0.32, 0.14, 0.37, 0.66, 0.34, 0.33, 0.28, 0.52, 0.37, 0.52,0.16, 0.34, 0.46,
0.48, 0.28, 0.43, 0.46, 0.45, 0.67, 0.64, 0.44, 0.29, 0.46, 0.40,0.63]

>PauFerro=[ 1.01, 1.02, 0.59, 1.17, 1.04, 0.71, 0.83, 1.14, 0.85, 1.18, 1.14, 0.84,0.55,
0.80, 0.89, 0.95, 0.94, 0.57, 0.64, 0.46, 0.62, 0.92, 0.29, 0.75, 1.11, 0.79, 0.66, 1.21, 0.94,
1.05, 0.96, 0.44, 1.33, 0.72, 1.10, 0.49, 0.77, 0.91, 0.82,0.83]
```

A distribuição normal padrão caracteriza-se por possuir:

- Aproximadamente 68% de todos os valores entre ± 1 desvio-padrão da média.
- Aproximadamente 95% de todos os valores entre ± 2 desvios-padrões da média.
- Aproximadamente 99,7% de todos os valores entre ± 3 desvios-padrões da média.

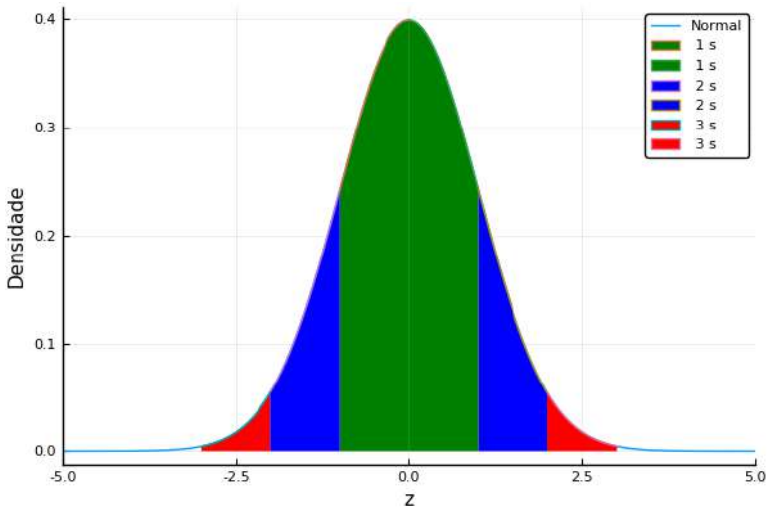
Na Figura 4.8 estas três características estão representadas e na sequência foi calculada a área de cada uma, respectivamente, pelo uso dos seguintes comandos:

Arquivo: 5 – Distribuição normal padrão e distribuição t

```
>using Distributions, Plots; pyplot()
>function NZ(zInferior,zSuperior)
    z = zInferior:0.01:zSuperior
```

```
    plot!(z, pdf.(Normal(), z), label="Normal", xlims=(zInferior, zSuperior),
    xlabel="z", ylabel="Densidade", dpi=300)
    AI=-1:0.0001:0
    AS=0:0.0001:1
    BI=-2:0.0001:-1
    BS=1:0.0001:2
    CI=-3:0.0001:-2
    CS=2:0.0001:3
    plot!(AI, pdf.(Normal(),AI), fill=(0, :green), label=" 1 s")
    plot!(AS, pdf.(Normal(),AS), fill=(0, :green), label=" 1 s")
    plot!(BI, pdf.(Normal(),BI), fill=(0, :blue), label=" 2 s")
    plot!(BS, pdf.(Normal(),BS), fill=(0, :blue), label=" 2 s")
    plot!(CI, pdf.(Normal(),CI), fill=(0, :red), label=" 3 s")
    plot!(CS, pdf.(Normal(),CS), fill=(0, :red), label=" 3 s")
end
>NZ(-5,5)
```

FIGURA 4.8. Propriedades da distribuição normal padronizada




```
>ÁreaVerde=1-ccdf(Normal(),1)*2
0.6827
>ÁreaAzul=1-ccdf(Normal(),2)*2
0.9545
>ÁreaVermelha=1-ccdf(Normal(),3)*2
0.9973
```

Note que a área verde engloba 68% dos valores amostrados (± 1 desvio-padrão), a soma das áreas verde e azul é de 95% dos valores amostrados (± 2 desvios-padrões) e a área total 99% (± 3 desvios-padrões).

4.3 DISTRIBUIÇÃO *t* DE STUDENT

A distribuição *t de Student* é importante ao se fazer a inferência sobre as médias populacionais. É um modelo de distribuição para variáveis contínuas que se aproxima da distribuição normal padrão (N~ 0,1), à medida que o tamanho da amostra aumenta (teorema do limite central). É utilizada para se fazer inferências estatísticas (testes de hipóteses e intervalos de confiança), ou seja, quando se trabalha com amostras.

$$f_{(x)} = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

Em que: Γ é a função gamma e n o tamanho da amostra.

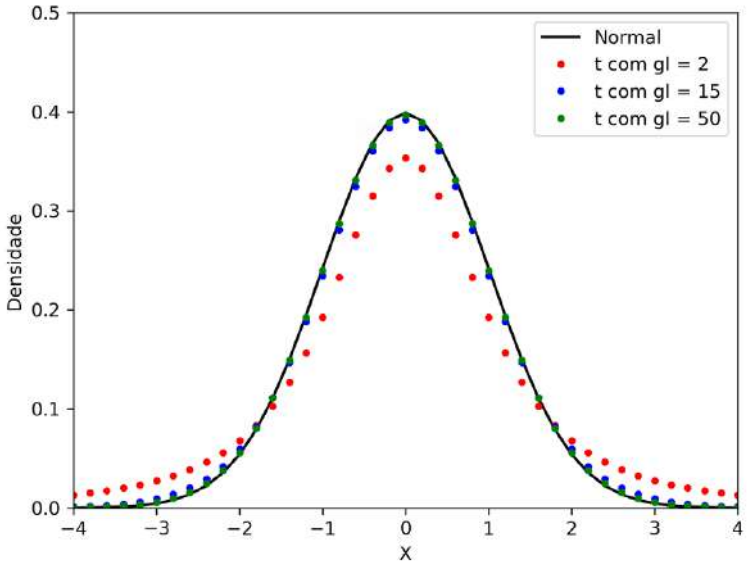
Cada tamanho de amostra dá origem a uma distribuição de *t* diferente. Em pequenas amostras a densidade resultante apresenta caudas mais alongadas. À medida que o número de observações aumenta, a distribuição *t* se aproxima da distribuição normal (Figura 4.9).

A Figura 4.9 foi obtida para amostras de tamanho 3, 16 e 51, utilizando-se os comandos:

ARQUIVO: 5 – DISTRIBUIÇÃO NORMAL PADRÃO E DISTRIBUIÇÃO t

```
>using Distributions, PyPlot
>x = -4:0.2:4
>plot(x, pdf(Normal(), x), "k", label="Normal")
>plot(x, pdf(TDist(2) ,x), "r.", label="t com gl = 2")
>plot(x, pdf(TDist(15), x), "b.", label="t com gl = 15")
>plot(x, pdf(TDist(50),x),"g.", label="t com gl = 50")
>xlim(-4,4)
>ylim(0,0.5)
>xlabel("X")
>ylabel("Densidade")
```

FIGURA 4.9. Distribuição Normal e distribuição t.



Na Figura 4.9 algumas características devem ser destacadas:

- Para cada tamanho de amostra se obtém uma distribuição de *t*.
- A média das duas distribuições (*t* e normal padronizada) é zero.
- A variância da distribuição normal padronizada é 1, já para a distribuição de *t* é maior do que 1 e depende do tamanho da amostra.
- A distribuição de *t* se aproxima da distribuição normal à medida que aumenta o tamanho da amostra. A diferença entre os valores da distribuição de *t* em relação à normal para amostras maiores que 30 é pequena.

5 ESTATÍSTICAS DESCRITIVAS

A tarefa inicial em qualquer análise de dados é a exploração dos dados por meio das estatísticas descritivas. É um processo de descoberta de informações, na qual a investigação dos dados tem o intuito de compreender as características específicas do assunto em questão. Além de viabilizar a obtenção de informações que os Engenheiros Florestais usam para fazer inferências, tal como a média da variável, auxiliam na escolha das técnicas estatísticas a serem utilizadas na análise dos dados.

Os dados coletados em um experimento são organizados em uma tabela e a partir dela é necessário resumi-los usando-se números e gráficos para facilitar sua interpretação e análise subsequente. Pode-se atribuir a esta organização o nome de estatística descritiva.

Na maioria dos experimentos florestais os pesquisadores mensuram variáveis aleatórias contínuas, assim a descrição dos conjuntos de dados é necessária para resumi-los da melhor forma possível. As estatísticas devem revelar as características de tendência central, de variabilidade e, também, os limites de confiança para a média, informações estas importantes para os Engenheiros Florestais tomarem decisões. Ainda, existem métodos gráficos que são utilizados para explorar os dados, pois permitem obter muitas informações de maneira rápida, especialmente para variáveis qualitativas.

Os métodos numéricos e gráficos quando utilizados em conjunto para resumir dados constituem-se em excelentes técnicas para auxiliar os Engenheiros Florestais na organização e descrição dos dados experimentais.

5.1 ESTATÍSTICA DESCRITIVA

O resumo numérico de dados consiste em obter as medidas de tendência central, de dispersão dos valores em torno da média e limites de confiança para a média.

Um conjunto de dados, referente ao diâmetro à altura do peito (em centímetros) obtidos em uma amostra de um povoamento de Pinus com 30 anos, foi utilizado para exemplificar os códigos dos cálculos estatísticos em Julia. Ressalta-se que todos os códigos apresentados na sequência foram desenvolvidos para amostras, método com o qual o Engenheiros Florestais trabalham rotineiramente.

```
>x=[36.4, 44.9, 61.3, 20.0, 49.5, 70.2, 29.4, 33.1, 44.9, 25.4, 44.9]
```

5.1.1 MEDIDAS DE TENDÊNCIA CENTRAL

As medidas de tendência central comumente calculadas são a média, a mediana, a moda e o ponto médio. Estas estatísticas (tudo que se calcula para uma amostra é chamado de estatísticas da amostra) são os valores típicos do conjunto de dados e, geralmente, estão localizados no centro da série. São indicadores do comportamento de uma variável aleatória e seu resumo permite a comparação com dados de outros estudos.

Média Aritmética

A média da amostra corresponde ao valor central de todas as observações. Esta é a mais importante estatística, pois é o valor mais representativo de uma amostra, de fácil compreensão e é também utilizada para cálculo das estatísticas de dispersão (que medem quanto as observações variam em torno dela). O principal problema desta estimativa é ela ser influenciada por dados extremos da amostra.

$$\bar{x} = \frac{\sum_{i=1}^{n} X_i}{n} \quad ou \quad \mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

Em que:
$\bar{x}$ é a média de um conjunto de valores amostrais - estatística
μ é a média dos valores da população - parâmetro
n é o número de elementos da amostra e N é o número de elementos da população
X_i é o valor observado da variável para o i – ésimo elemento

Para que se possa efetuar diferentes tipos de análises estatísticas, inicialmente utilizam-se os comandos:

Arquivo: 6 – Estatísticas descritivas

Para se obter o tamanho da amostra utiliza-se o seguinte comando:

```
>using Statistics, Distributions
>n= length(x)
11
```

Para obtenção da média utiliza-se o comando:

```
>Média=mean(x)
41.82

>pm=(maximum(x)+minimum(x))/2
>moda=mode(x)
>AT=maximum(x)-minimum(x)
>s²=var(x)
```

Ponto médio

O ponto médio (*pm*) é o valor que está no centro entre os valores máximo (X_{max}) e mínimo ($X_{mín}$). É uma estatística que também indica o centro de um conjunto de dados.

$$pm = \frac{(X_{max} + X_{mín})}{2}$$

Para obtenção do ponto médio utiliza-se o comando:

```
>pm=(maximum(x)+minimum(x))/2
45.1
```

Mediana

A mediana ($\tilde{x}$) é o valor central de uma amostra, ou a média dos dois valores centrais (se *n* for um número par), quando os valores estão ordenados. Ela é útil para dividir os dados em duas partes iguais, metade dos dados acima da mediana e a outra metade abaixo. A mediana é, conceitualmente, uma estatística fácil de ser calculada.

Para obtenção da mediana utiliza-se o comando:

```
>Mediana=median(x)
44.90
```

Moda

A moda (*M*) é o valor mais frequente em uma amostra. Assim como a mediana, a moda é uma estatística fácil de ser calculada.

Para obtenção da moda utiliza-se o comando:

```
>Moda=mode(x)
44.90
```

As medidas de tendência central apresentadas têm vantagens e desvantagens. A média é a estatística mais utilizada por Engenheiros Florestais, porque em seu cálculo todos os valores do conjunto de dados são considerados. Entretanto, deve-se ter cuidado com valores extremos, os quais afetam o valor da média.

5.1.2 MEDIDAS DE DISPERSÃO

As medidas de tendência central são úteis, mas revelam apenas o valor central de um conjunto de dados. Além do valor central é desejável obter um valor que releve a dispersão dos valores, ou seja, o quão espalhados ou agrupados estão os dados em torno da média.

1. a variação é uma medida de quanto os valores diferem entre si;
2. valores próximos uns dos outros resultam em baixa variação, enquanto mais dispersos uns dos outros resultam em alta variação.

Amplitude total

A medida de variabilidade mais simples é a amplitude total, que correspondente à diferença entre o maior valor X_{max} e o menor valor $X_{mín}$ do conjunto de dados. Ela depende somente dos valores extremos, portanto, é pouco informativa.

$$AT = X_{max} - X_{mín}$$

Para obtenção da amplitude total utiliza-se o comando:

```
>AT=maximum(x)-minimum(x)
50.20
```

Variância

A medida de variação que melhor explica a dispersão dos dados em torno da média é a variância. Para uma população a variância é definida como:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

A variância para a amostra pode ser expressa como:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{x})^2}{n-1}$$

A variância é uma medida de variabilidade que considera todos os valores da amostra e mede a variabilidade das observações em torno da média. Como para sua obtenção os valores dos desvios em relação a média são elevados ao quadrado, o valor final da variância terá sua unidade elevada ao quadrado.

Para obtenção da variância de uma amostra utiliza-se o comando:

```
>s²=var(x)
227.77
```

Desvio padrão

Como a variância é expressa com sua unidade elevada ao quadrado, utiliza-se o desvio padrão para apresentar uma medida de variação na mesma unidade da média.
O desvio-padrão para a população (σ) pode ser expresso como:

$$\sigma = \sqrt{(\sigma^2)}$$

O desvio-padrão para a amostra (s) pode ser expresso como:

$$s = \sqrt{(s^2)}$$

Para obtenção do desvio padrão utiliza-se o comando:

```
>s=std(x)
15.09
```

Para a amostra utilizada foi obtida uma média de 41,82 cm e um desvio padrão de 15,09 cm, o que significa que os valores apresentam uma dispersão, em média, de 15,09 cm em torno do centro.

Coeficiente de variação

Muitas vezes o Engenheiro Florestal avalia variáveis que não são comuns em seu dia a dia de trabalho, fato este que torna difícil a interpretação do desvio padrão. Outra dificuldade encontrada é comparar a variabilidade de variáveis que têm unidades de medida distintas, tais como o diâmetro à altura do peito (em cm) e o volume do fuste (em m³) das árvores. Uma alternativa é o uso do coeficiente de variação, que é uma medida relativa da variabilidade. Esta estatística além de facilitar a interpretação permite a comparação da variabilidade entre diferentes variáveis. Isto deve-se ao fato de o coeficiente de variação ser uma estatística adimensional. O coeficiente de variação é calculado por:

$$CV = \frac{\sigma}{\mu} 100, \text{para a população}$$

$$cv = \frac{s}{\bar{x}} 100, \text{para a amostra}$$

Para obtenção do coeficiente de variação utiliza-se o comando:

```
>cv=s/Média*100
36.09
```

A variância e o desvio-padrão medem a variação dos valores em torno da média. Valores dos dados próximos um dos outros dão origem a variância e desvio padrão baixos, enquanto que para valores muito diferentes um dos outros, variância e desvio-padrão altos. Para demonstrar isto, na Figura 5.1 estão apresentados os resultados de três simulações:

1. Conjunto de dados com 10.000 valores, com média 20 (u.m.), variância 0 (u.m.²) e desvio-padrão 0 (u.m.). Estes valores estão representados na linha azul.
2. Conjunto de dados com 10.000 valores, com média 20 (u.m.), variância 15,65 (u.m.²) e desvio-padrão 3,96 (u.m.). Estes valores estão representados na linha laranja.
3. Conjunto de dados com 10.000 valores, com média 20 (u.m.), variância 64,51 (u.m.²) e desvio-padrão 8,03 (u.m.). Estes valores estão representados na linha verde.

Arquivo: 6 – Estatísticas descritivas

```
>using Distributions, PyPlot

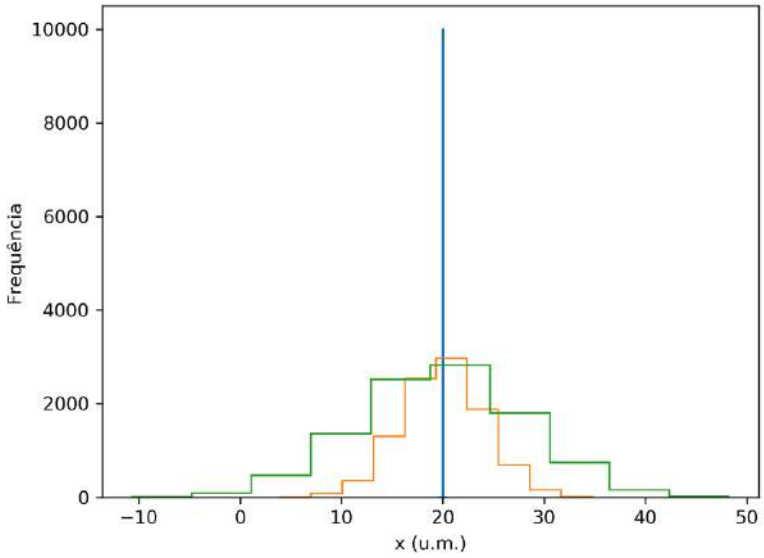
#Simulação 1
>d=Normal(20,0)
Normal{Float64}(μ=20.0, σ=0.0)
>x = rand(d, 10000);
>PyPlot.plt.hist(x,10,histtype = "step")
>ylabel("Frequência")
>xlabel("x (u.m.)")
>mean(x)
```

```
20.0
>var(x)
0.0
>std(x)
0.0

#Simulação 2
>d=Normal(20,4)
Normal{Float64}(μ=20.0, σ=4.0)
>x = rand(d, 10000);
>PyPlot.plt.hist(x,10,histtype = "step")
>mean(x)
20.03
>var(x)
15.65
>std(x)
3.96
```

```
#Simulação 3
>d=Normal(20,8)
>Normal{Float64}(μ=20.0, σ=8.0)
x = rand(d, 10000);
>PyPlot.plt.hist(x,10,histtype = "step")
>mean(x)
20.00
>var(x)
64.51
>std(x)
8.03
```

FIGURA 5.1. Resultado das simulações.



Em que:
Linha azul $\overline{x}$ = 20 (u.m.) e s = 0 (u.m.)
Linha laranja $\overline{x}$ = 20 (u.m.) e s = 3,96 (u.m.)
Linha verde $\overline{x}$ = 20 (u.m.) e s = 8,03 (u.m.)

Na primeira simulação, todos os valores são iguais a 20 (u.m.), portanto, não existe variabilidade entre os dados, ou seja, é uma constante. Nas três simulações a média é igual a 20 (u.m.), entretanto, nas simulações 2 e 3 os dados se dispersam com o aumento da variância e do desvio padrão.

5.1.3 ERRO PADRÃO DA MÉDIA E INTERVALO DE CONFIANÇA PARA A MÉDIA

Quando o conjunto de dados é oriundo do censo, os valores são paramétricos para a média, moda, mediana, variância, desvio-padrão e coeficiente de variação. São valores únicos. Quando os dados são oriundos de amostras estas estatísticas podem variar.

DESVIO PADRÃO DA MÉDIA OU ERRO PADRÃO

Normalmente, o Engenheiro Florestal trabalha com dados obtidos por meio de amostragem, assim, está ciente que em sucessivas amostragens em uma mesma floresta, por exemplo, pode resultar em diferentes valores para a média de uma variável qualquer. Se fossem conduzidas várias amostragens em uma mesma floresta, várias médias seriam obtidas e normalmente seriam diferentes. A variação existente entre médias de amostras, de mesmo tamanho, obtidas da mesma população pode ser avaliada por meio do desvio padrão das médias também conhecido como erro padrão. É possível, no entanto, calcular a variação entre médias de diferentes amostras pelo cálculo do desvio padrão das médias, ou erro padrão. Esta estatística pode ser calculada a partir de apenas uma amostra, por meio do cálculo do erro padrão. Esta estatística é dada por:

$$S_{\bar{x}} = \frac{S}{\sqrt{n}}$$

Para obtenção do desvio padrão da média, ou erro padrão, utiliza-se o comando:

```
>sx=s/sqrt(n)
5.34
```

A utilização do desvio padrão da média (erro padrão) é utilizada pela mesma razão da utilização do desvio padrão de uma amostra, ou seja, é uma estatística com a mesma unidade de medida da média.

Em resumo, pode-se dizer que o desvio padrão de uma amostra mede a dispersão dos valores da amostra em relação ao centro (média). O erro padrão mede a dispersão de diferentes médias obtidas de diferentes amostras em relação ao centro (média das médias).

No exercício de simulação apresentado na sequência, foram obtidos 62 valores de diâmetros à altura do peito da espécie acácia-negra (Figura 5.2). Estes valores foram considerados como se fossem uma população. Desta população foi retirada uma amostra de tamanho n = 62 e calculadas as estatísticas $\bar{x}$ = 12,64 cm, s = 3,3763 cm e $s_{\bar{x}}$ = 0,4289 cm. Para esta mesma população foram efetuadas outras amostras, tomadas aleatoriamente num total de 1.000.000 amostras (N) de tamanho 62 (n) e para cada amostra foi calculada a média. A média das médias das amostras foi de $\bar{x}$ = 12,64 cm e o desvio-padrão das médias, calculado por meio de $s_{\bar{x}} = \sqrt{(s_{\bar{x}}^2)}$, foi de $s_{\bar{x}}$ = 0,4255 cm. Nessa simulação pode ser observado que as duas estatísticas são próximas indicando que a estimativa do erro padrão obtida a partir de uma amostra se assemelha bastante ao valor obtido quando foram realizadas 1.000.000 de amostras.

FIGURA 5.2. Diâmetro à altura do peito de acácia-negra.



ARQUIVO: 6 – ESTATÍSTICAS DESCRITIVAS

```
>using DataFrames, CSV, Distributions
>X=CSV.read("Z:/Alexandre/Julia/Exemplos/descritiva.csv", DataFrame)
```

julia

```
julia> X=CSV.read("Z:/Alexandre/Julia/Exemplos/descritiva.csv", DataFrame)
62×8 DataFrame
```

Row	Idade	d	h	cc	b	me	v	ff
	Float64	Float64	Float64	Float64	Float64	Float64	Float64	Float64
1	10.75	19.74	20.2	7.7	206.09	0.65	0.28	0.45
2	10.75	15.12	19.1	4.9	93.17	0.5	0.16	0.48
3	10.75	12.89	17.3	3.0	65.88	0.57	0.11	0.47
4	10.75	11.62	16.9	7.2	49.3	0.52	0.09	0.5
5	10.75	16.55	20.5	11.8	142.0	0.56	0.22	0.5
⋮								
57	10.75	10.03	15.1	3.7	56.07	0.59	0.06	0.48
58	10.75	12.73	17.6	7.4	89.64	0.54	0.12	0.54
59	10.75	14.64	17.7	12.1	104.02	0.61	0.16	0.53
60	10.75	12.73	18.2	9.2	85.58	0.58	0.12	0.53
61	10.75	7.32	11.7	7.5	26.23	0.51	0.03	0.59
62	10.75	18.3	19.5	7.6	179.41	0.54	0.24	0.47

```
>Media=mean(X.d)
12.6414
>Desviopadrao=std(X.d)
3.3763
```

```
>ErroPadrao=std(X.d)/sqrt(length(X.d))
0.4289
>n, N = 62, 1000000;
>Reamostragem = [mean(rand(X.d, n)) for i in 1:N]
1000000-element Vector{Float64}:
 12.404838709677415
 11.884193548387097
 12.710645161290323
 12.767096774193549
 12.42806451612903
 12.251129032258067
 13.393387096774193
 12.479677419354843
...
 13.350161290322582
 12.543870967741933
 12.87241935483871
 12.615806451612904
 13.208709677419353
 13.077419354838709
 12.032903225806448

>MediaReamostragem=mean(Reamostragem)
12.6416
>DesvioPadraoReamostragem=std(Reamostragem)
0.4255
```

Como observado pelo exercício de simulação, a estatística erro-padrão revela a exatidão da média, consequentemente, permite calcular o intervalo de confiança para a média.

Intervalo de confiança para a média

O intervalo de confiança descreve os limites para o valor paramétrico da média populacional é esperado ocorrer, para um determinado nível de probabilidade. É uma estatística fundamental, pois a descrição de um conjunto de dados geralmente é dada pela média com uma probabilidade a ela associada. O intervalo de confiança para a média pode ser expresso como:

$$IC[\ \bar{x} - t_{(\alpha,n-1)}\, s_{\bar{x}} \leq \mu \leq \bar{x} + t_{(\alpha,n-1)}\, s_{\bar{x}}\ ] = 1 - \alpha$$

Para obtenção do intervalo de confiança para a média da amostra, utiliza-se o comando:

Arquivo: 6 – Estatísticas descritivas

```
>alpha = 0.05
0.05
>tVal = quantile(TDist(n-1),1-alpha/2)
2.2281
>ICInferior = (Média- tVal * sx)
```

```
31.6791
>ICSuperior = (Média+ tVal * sx)
51.9572
```

O intervalo de confiança é expresso graficamente na Figura 5.3, desenvolvida conforme o algoritmo apresentado na sequência.

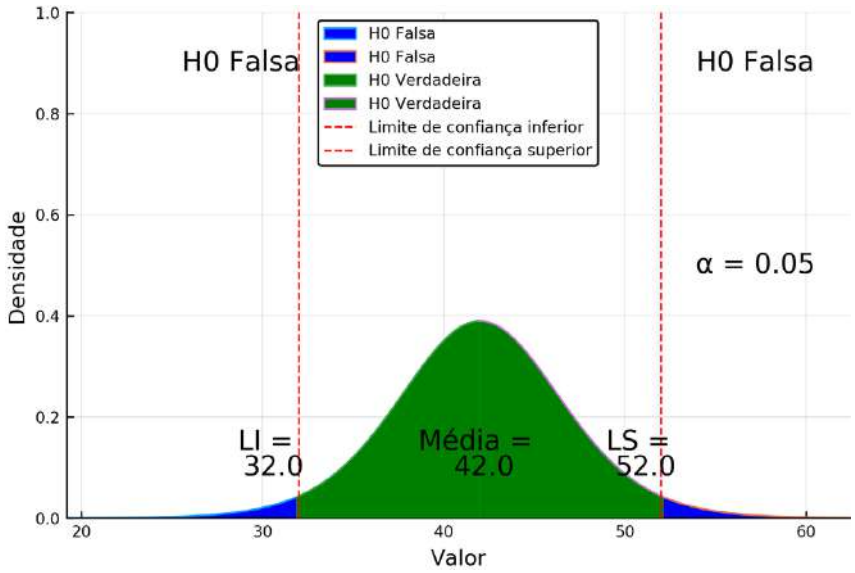
Arquivo: 6 – Estatísticas descritivas

```
>using Distributions, Plots; pyplot()

>function IC(x)
    Alpha=0.05
    n= length(x)
    xbar=sum(x)/n
    xbar=round(xbar)
    s²=sum((x.-xbar).^2)/(n-1)
    s=sqrt(s²)
    sx=s/sqrt(n)
    tVal = quantile(TDist(n-1),1-Alpha/2)
    ICInferior = (xbar- tVal * sx)
    ICInferior=round(ICInferior)
    ICSuperior = (xbar+ tVal * sx)
    ICSuperior=round(ICSuperior)
    xGridInfH0=(Alpha/2):0.0001:0.5
    xGridSupH0=0.5:0.0001:(1-(Alpha/2))
    xGridInfH1=0:0.0001: (Alpha/2)
    xGridSupH1=(1-(Alpha/2)):0.0001:1
    H0Inf = quantile.(TDist(n-1),xGridInfH0)
    H0Sup= quantile.(TDist(n-1),xGridSupH0)
    H1Inf = quantile.(TDist(n-1),xGridInfH1)
    H1Sup= quantile.(TDist(n-1),xGridSupH1)
    DensidadeH0Inf= pdf.(TDist(n),H0Inf)
    DensidadeH0Sup= pdf.(TDist(n),H0Sup)
    DensidadeH1Inf= pdf.(TDist(n),H1Inf)
    DensidadeH1Sup= pdf.(TDist(n),H1Sup)
    XH1Inf = xbar.+(H1Inf.*sx)
    XH1Sup = xbar.+( H1Sup.*sx)
    XH0Inf = xbar.+(H0Inf.*sx)
    XH0Sup = xbar.+( H0Sup.*sx)
    plot!( XH1Inf, DensidadeH1Inf,fill=(0, :blue), xlims=((ICInferior*0.6),
    (ICSuperior*1.2)), ylims=(0,1), label="H0 Falsa",dpi=300)
    plot!( XH1Sup, DensidadeH1Sup,fill=(0, :blue), label="H0 Falsa")
    plot!( XH0Inf, DensidadeH0Inf,fill=(0, :green), label="H0 Verdadeira")
    plot!( XH0Sup, DensidadeH0Sup,fill=(0, :green), label="H0 Verdadeira")
    plot!([ ICInferior, ICInferior],[0,1],c=:red, ls=:dash,
    label="Limite de confiança inferior", xlabel="Valor", ylabel="Densidade")
    plot!([ ICSuperior, ICSuperior],[0,1],c=:red, ls=:dash,
    label="Limite de confiança superior")
    A=(ICInferior+ICSuperior)/2
```

```
            B=ICInferior*0.90
            C=ICSuperior*1.1
            D=(ICInferior+ICSuperior)/2
            E= ICSuperior*0.98
            F=ICInferior*0.95
            G= ICSuperior*1.10
          annotate!([(A, 0.9, text("H0 Verdadeira")),(B, 0.9, text("H0 Falsa")),(C, 0.9, text("H0
Falsa")),(D, 0.15, text("Média = ")), ((D+0.2), 0.1, xbar), (F, 0.15, text("LI = ")), ((F+0.2),
0.1, ICInferior), (E, 0.15, text("LS = ")), ((E+0.2),0.1, ICSuperior), (G, 0.5, text("α =
0.05"))])
       end
       >IC(x)
```

FIGURA 5.3. Intervalo de confiança para a média da amostra.



ASPECTOS GERAIS SOBRE O INTERVALO DE CONFIANÇA

Dada uma única amostra, qual é a exatidão da estimativa da média? Considere que o parâmetro a ser estimado é μ (média da população de onde foi tirada a amostra), então, um intervalo de confinça inferior (I) e superior (S) pode ser dado por: $P\ (I \leq \mu \leq S) = 1 - \alpha$. O valor de $1 - \alpha$ indica o nível de confiança da estimativa da média. Assim, conhecendo-se os valores de $[I,S]$, interpretações podem ser feitas sobre o nível de certeza em relação ao valor desconhecido de μ. Dessa forma, existe a chance de $1 - \alpha$ que o parâmetro μ esteja no intervalo $[I,S]$, ou seja, se a amostragem for repetida N vezes, em média $(1 - \alpha)100$ vezes o parâmetro μ está dentro do intervalo $[I,S]$.

No exercício de simulação, apresentado na sequência, foi obtida uma amostra de tamanho n ($n = 1000$) e a amostragem foi repetida N vezes ($N = 100$). Para cada amostra aleatória obtida foi calculado o seu respectivo intervalo de confiança. O resultado obtido comprova que somente $\alpha / 100$ vezes o intervalo de confiança não inclui o parâmetro μ em questão (Figura 5.4). Portanto, existe uma chance de $1 - \alpha$ de que o parâmetro μ esteja no intervalo de confiança.

Para obtenção dos intervalos de confiança para as 100 amostras tomadas, utiliza-se o comando:

Arquivo: 6 – Estatísticas descritivas

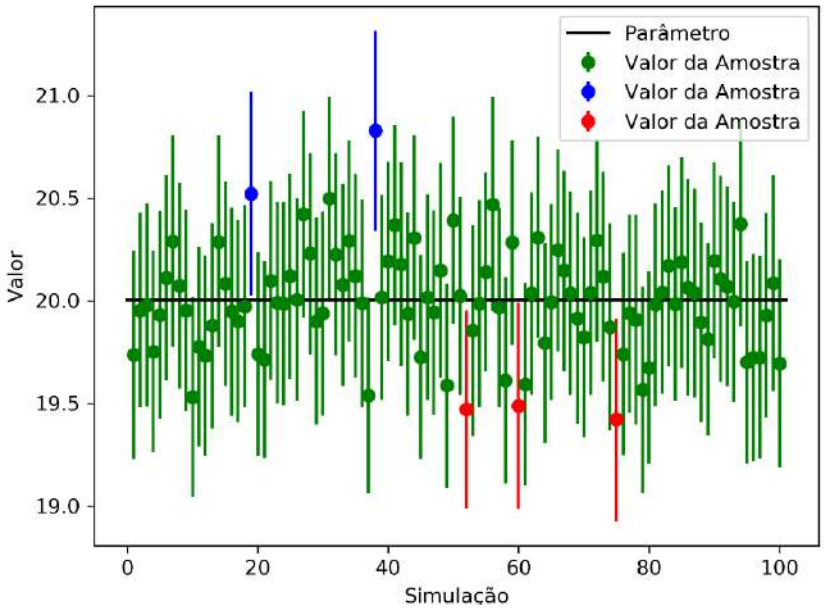
```
>using DataFrames, Random, Distributions, PyPlot

>d=Normal(20,8)
Normal{Float64}(μ=20.0, σ=8.0)
>X = rand(d, 100000000);

>function IC(X)
    mu=mean(X)
    n=1000
    N=100
    Amostra = [rand(X, n) for i in 1:N]
    Media=mean.(Amostra)
    Desvio=std.(Amostra)
    ErroPadrao=Desvio/sqrt(n)
    alpha = 0.05
    tVal = quantile(TDist(n-1),1-alpha/2)
    Delta= tVal*ErroPadrao
    LI=Media-Delta
    LS=Media+Delta
    df = DataFrame()
    df.NSimulacao = 1:N
    df.Media=Media
    df.Delta=Delta
    df.LI=LI
    df.LS=LS
    Dentro=df[(df.LI .< mu) .& (df.LS .>mu), :]
    ForaAcima=df[(df.LI .> mu) .& (df.LS .>mu), :]
    ForaAbaixo=df[(df.LI .< mu) .& (df.LS .<mu), :]
errorbar(Dentro.NSimulacao,Dentro.Media,Dentro.Delta,fmt="o",c=:green, label="Valor da
Amostra")
 errorbar(ForaAcima.NSimulacao,ForaAcima.Media,ForaAcima.Delta,fmt="o",
c=:blue,label="Valor da Amostra")
errorbar(ForaAbaixo.NSimulacao,ForaAbaixo.Media,ForaAbaixo.Delta,fmt="o",
c=:red, label="Valor da Amostra")
    plot([0,N+1],[mu,mu], c=:black, label="Parâmetro")
    xlabel("Simulação")
    ylabel("Valor")
    legend(loc="best")
end

>IC(X)
```

FIGURA 5.4. Intervalo de confiança para cada uma das 100 amostras tomadas



Na Figura 5.4 os intervalos destacados em vermelho indicam as amostras cujas médias ficaram abaixo da média paramétrica e os intervalos destacados em azul as que ficaram acima da média paramétrica.

EXERCÍCIO PROPOSTO 4: ESTATÍSTICA DESCRITIVA – VARIÁVEIS DENDROMÉTRICAS DE ACÁCIA-NEGRA

Considere o conjunto de dados em um arquivo com extensão CSV denominado descritiva. Os dados são oriundos de uma amostragem aleatória de árvores da espécie acácia-negra, cultivada em povoamentos no estado do Rio Grande do Sul com 10,75 anos. As variáveis mensuradas foram:

d = diâmetro à altura do peito, em cm
h = altura total, em m
cc = comprimento de copa, em m
b = biomassa, em kg
me = massa específica da madeira, em $g.cm^{-3}$
v = volume do fuste com casca, em m^3
f = fator de forma, adimensional

Apresente uma análise exploratória por meio das estatísticas descritivas calculadas para todas as variáveis do conjunto de dados. Interprete os resultados.

Para obtenção dos resultados para a variável *d* utiliza-se os comandos:

Arquivo: 7 - Função resumo para estatística descritiva

```
>using DataFrames, CSV, Statistics, Distributions

#Arquivo com os dados das variáveis dendrométricas
>X=CSV.read("Z:/Alexandre/Julia/Exemplos/descritiva.csv", DataFrame)
```

julia

```
julia> X=CSV.read("Z:/Alexandre/Julia/Exemplos/descritiva.csv", DataFrame)
62×8 DataFrame
```

Row	Idade	d	h	cc	b	me	v	ff
	Float64	Float64	Float64	Float64	Float64	Float64	Float64	Float64
1	10.75	19.74	20.2	7.7	206.09	0.65	0.28	0.45
2	10.75	15.12	19.1	4.9	93.17	0.5	0.16	0.48
3	10.75	12.89	17.3	3.0	65.88	0.57	0.11	0.47
4	10.75	11.62	16.9	7.2	49.3	0.52	0.09	0.5
5	10.75	16.55	20.5	11.8	142.0	0.56	0.22	0.5
⋮								
57	10.75	10.03	15.1	3.7	56.07	0.59	0.06	0.48
58	10.75	12.73	17.6	7.4	89.64	0.54	0.12	0.54
59	10.75	14.64	17.7	12.1	104.02	0.61	0.16	0.53
60	10.75	12.73	18.2	9.2	85.58	0.58	0.12	0.53
61	10.75	7.32	11.7	7.5	26.23	0.51	0.03	0.59
62	10.75	18.3	19.5	7.6	179.41	0.54	0.24	0.47

```
>function RESUMO(variavel)
    n= length(variavel)
    xbar=sum(variavel)/n
    md=median(variavel)
    mo=mode(variavel)
    max=maximum(variavel)
    min=minimum(variavel)
    at=maximum(variavel)-minimum(variavel)
    s2=sum((variavel.-xbar).^2)/(n-1)
    s=sqrt(s2)
    sx=s/sqrt(n)
    cv=s/xbar*100
    alpha = 0.05
    tVal = quantile(TDist(n-1),1-alpha/2)
    ICInferior = (xbar- tVal * sx)
    ICSuperior = (xbar+ tVal * sx)
    Confiança=(1-alpha)*100
    println("Média: ", xbar)
    println("Mediana: ", md)
    println("Moda: ", mo)
    println("Mínimo: ", min)
    println("Máximo: ", max)
    println("Amplitude: ", at)
    println("Variância: ", s2)
```

```
    println("Desvio padrão: ", s)
    println("Erro padrão da média: ", sx)
    println("Coeficiente de variação (%): ", cv)
    println("Limite de confiança inferior: ", ICInferior)
    println("Limite de confiança superior: ", ICSuperior)
    println("Confiança (%): ", Confiança)
end

>RESUMO(X.d)
```

```
julia> RESUMO(X.d)
Média: 12.641451612903225
Mediana: 12.65
Moda: 12.73
Mínimo: 3.82
Máximo: 20.53
Amplitude: 16.71
Variância: 11.399189661554733
Desvio padrão: 3.3762686003271027
Erro padrão da média: 0.42878654102829744
Coeficiente de variação (%): 26.707918550120617
Limite de confiança inferior: 11.784039932534641
Limite de confiança superior: 13.498863293271809
Confiança (%): 95.0
```

5.2 SIMETRIA DE UM CONJUNTO DE DADOS

A avaliação dos valores da média, mediana e moda podem revelar, de forma simples, a característica da simetria de uma distribuição dos dados. Uma distribuição é simétrica quando a metade direita do histograma é semelhante a metade esquerda, caso contrário, a distribuição é assimétrica. Considere o exemplo apresentado na Figura 5.5, para uma variável *X* (em unidades de medidas, u.m.) geradas utilizando a função de distribuição de probabilidade beta.

Quando a distribuição é assimétrica negativa a média e a mediana resulta em valores menores que a moda, portanto estão à esquerda da moda (Gráfico na cor laranja). Quando a distribuição é assimétrica positiva a média e a mediana resultam em valores maiores que a moda, portanto, estão à direita da moda (Gráfico na cor azul). Quando a distribuição é simétrica a média, mediana e moda resultam aproximadamente o mesmo valor (Gráfico na cor verde).

ARQUIVO: 8 - SIMETRIA DE DADOS

```
>using Distributions, PyPlot

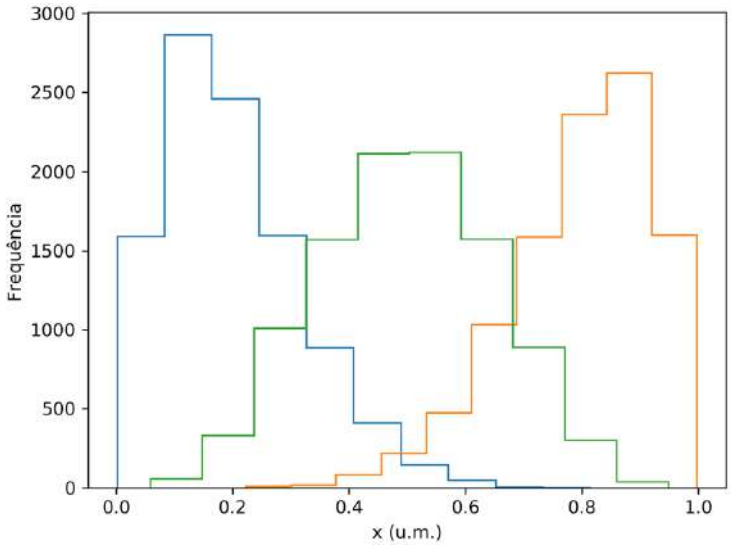
#Assimetria positiva
>d=Beta(2,8);
>x = rand(d, 10000);
>PyPlot.plt.hist(x,10,histtype = "step")
```

```
>ylabel("Frequência")
>xlabel("x (u.m.)")

#Assimetria negativa
>d=Beta(8,2)
>x = rand(d, 10000)
>plt.hist(x, 10, facecolor="orange", histtype = "step")

#Simétrica
>d=Beta(5,5)
>x = rand(d, 10000)
>plt.hist(x, 10, facecolor="green", histtype = "step")
```

FIGURA 5.5. Simetria e assimetria na distribuição dos dados.

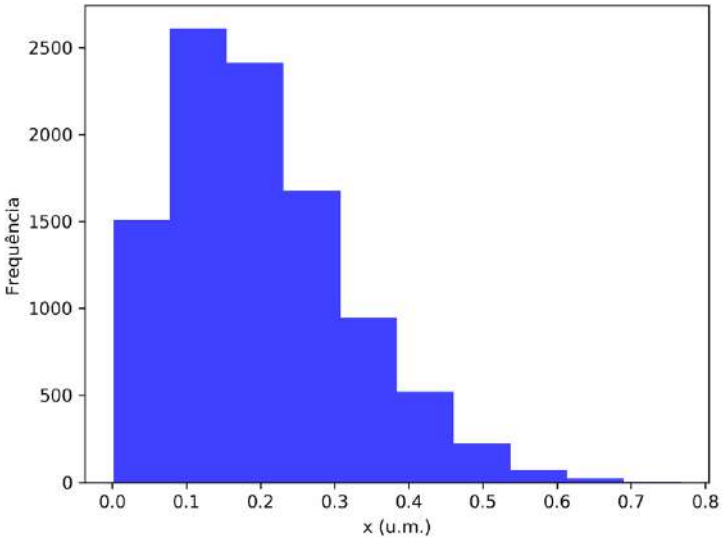


Para obtenção dos resultados utiliza-se os comandos:

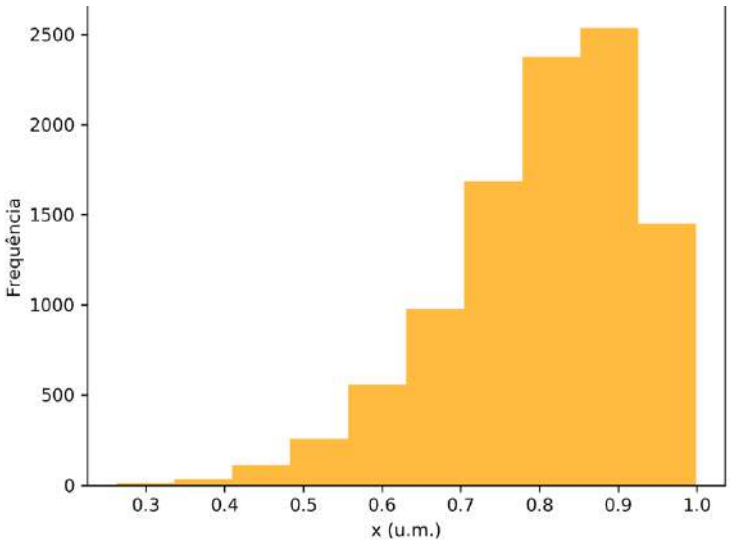
ARQUIVO: 8 - SIMETRIA DE DADOS

```
>using Distributions, PyPlot

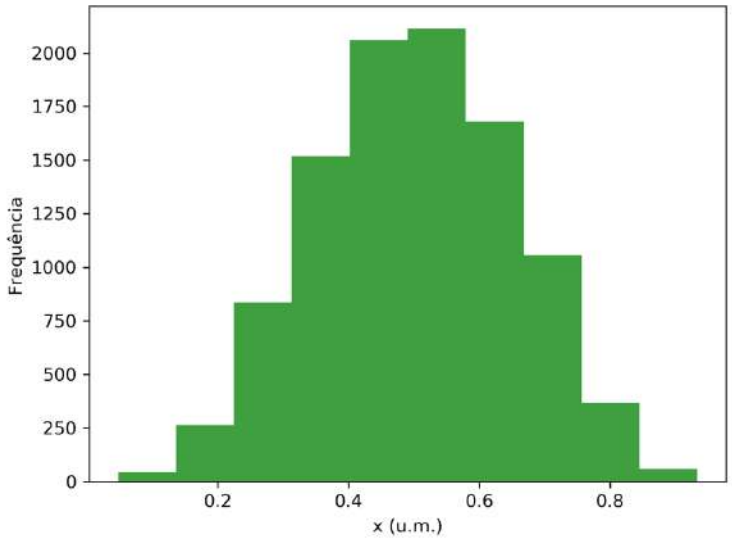
#Assimétrica positiva
>d=Beta(2,8)
>x = rand(d, 10000)
>plt.hist(x, 10, facecolor="blue", alpha=0.75)
>ylabel("Frequência")
>xlabel("x (u.m.)")
>mean(d)
0.20
>mode(d)
0.13
>median(d)
0.18
```




```
#Assimétrica negativa
>d=Beta(8,2)
>x = rand(d, 10000)
>plt.hist(x, 10, facecolor="orange", alpha=0.75)
>ylabel("Frequência")
>xlabel("x (u.m.)")
>mean(d)
0.80
>mode(d)
0.88
>median(d)
0.82
```



```
#Simétrica
>d=Beta(5,5)
>x = rand(d, 10000)
>plt.hist(x, 10, facecolor="green", alpha=0.75)
>ylabel("Frequência")
>xlabel("x (u.m.)")
>mean(d)
0.50
>mode(d)
0.50
>median(d)
0.50
```



5.3 SIMETRIA DAS MÉDIAS AMOSTRAIS

Uma característica importante da distribuição das médias amostrais é sua aproximação à normalidade. Se em uma população várias amostras do mesmo tamanho forem obtidas, independente da distribuição da variável, as médias estimadas para as amostras se aproximarão da distribuição Normal, o que é visível por meio da simulação realizada.

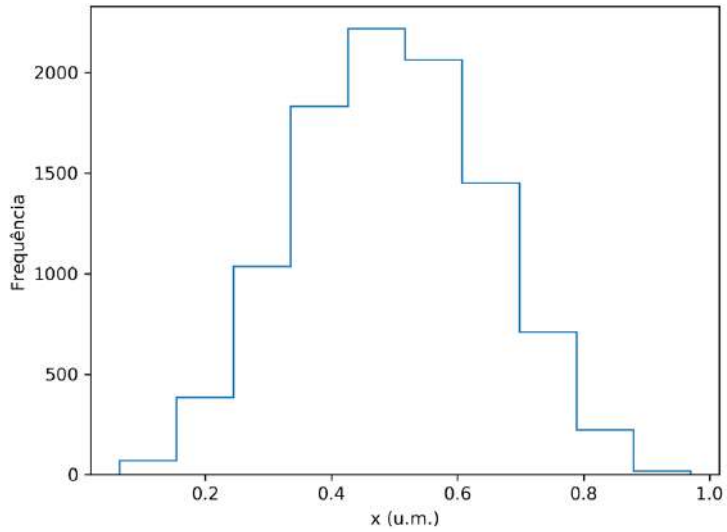
Considere os dados obtidos de variáveis com distribuição assimétrica, simétrica e retangular. Amostras de tamanho *n* foram obtidas e a amostragem foi repetida N vezes. As médias estimadas para as *N* vezes foram plotadas em um histograma.

Para obtenção dos resultados utiliza-se os comandos:

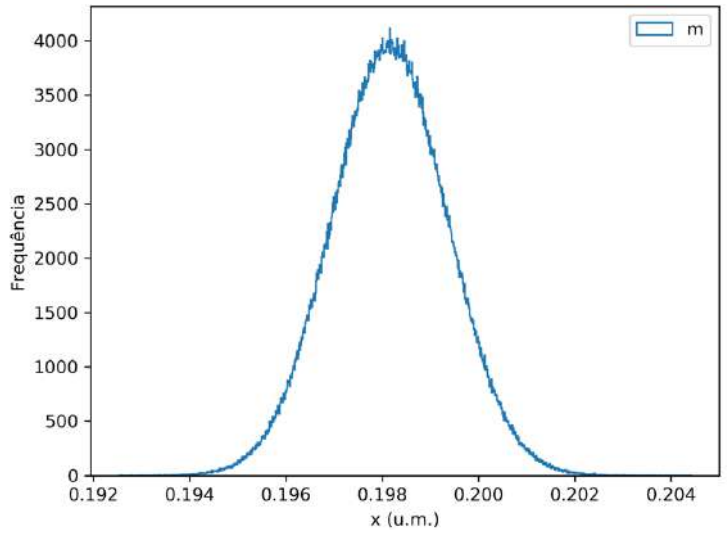
ARQUIVO: 8 - SIMETRIA DE DADOS

```
#Distribuição das médias amostrais
>using Distributions, PyPlot, Random
```

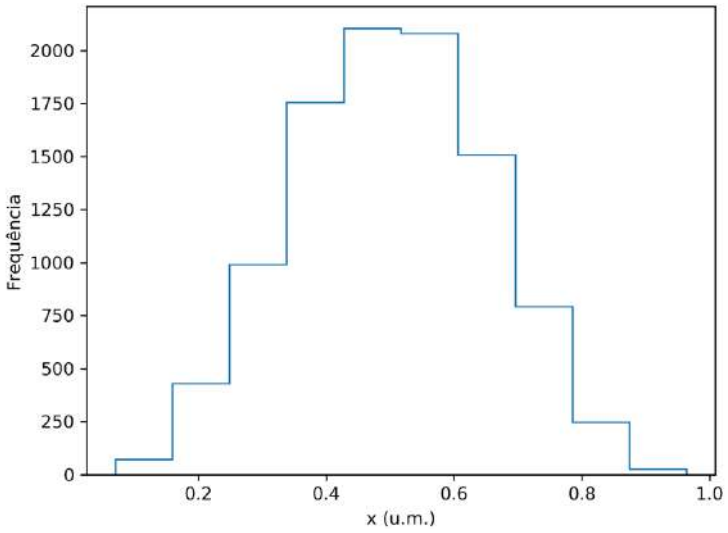
```
#Assimétrica positiva
>d=Beta(2,8);
>x = rand(d, 10000);
>PyPlot.plt.hist(x,10,histtype = "step")
>ylabel("Frequência")
>xlabel("x (u.m.)")
```



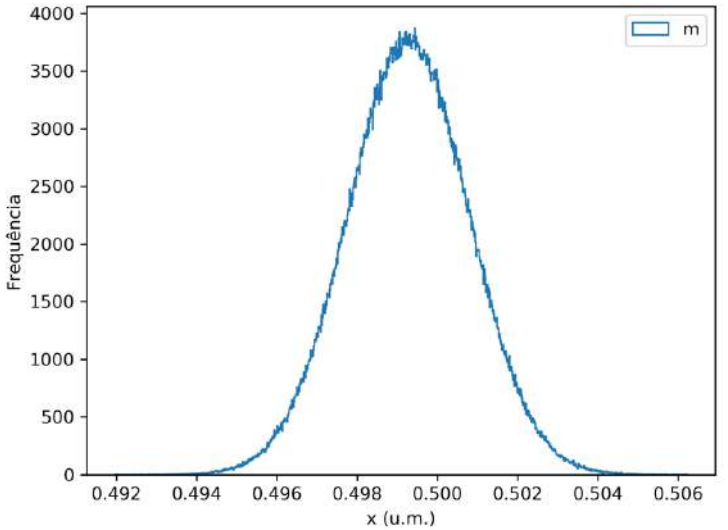

```
>n, N = length(x), 10^6
>Médias = [mean(rand(x, n)) for i in 1:N]
>PyPlot.plt.hist(Médias,1000,histtype = "step")
>ylabel("Frequência")
>xlabel("x (u.m.)")
>legend("média da amostra")
```



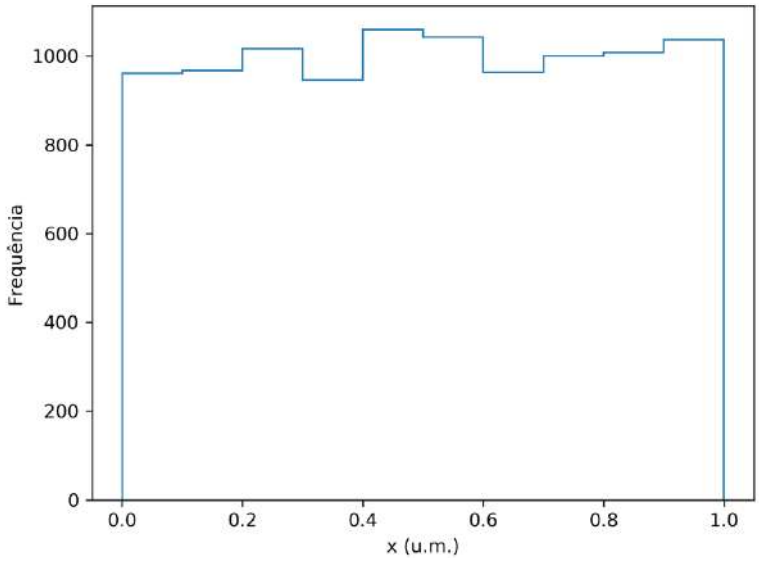

```
#Simétrica
>d=Beta(5,5);
>x = rand(d, 10000);
>PyPlot.plt.hist(x,10,histtype = "step")
>ylabel("Frequência")
>xlabel("x (u.m.)")
```



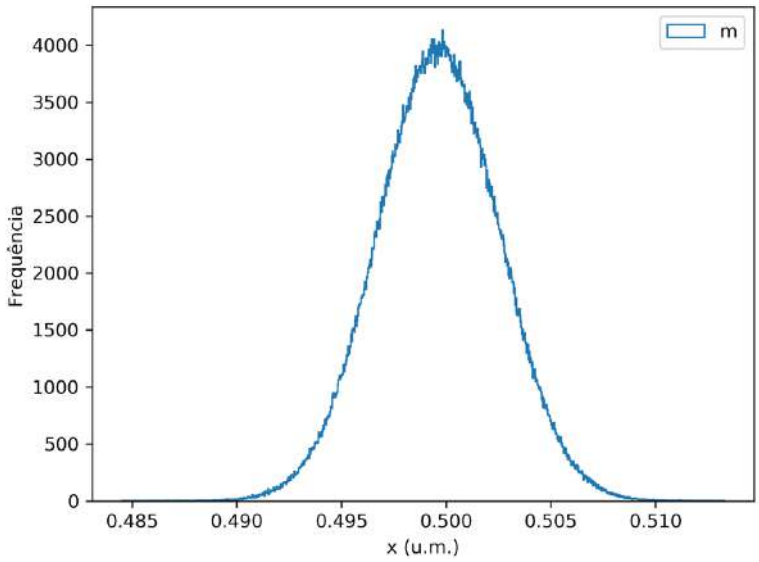

```
>n, N = length(x), 10^6
>Médias = [mean(rand(x, n)) for i in 1:N]
>PyPlot.plt.hist(Médias,1000,histtype = "step")
>ylabel("Frequência")
>xlabel("x (u.m.)")
>legend("média da amostra")
```




```
#Uniforme
>d=Uniform(0,1);
>x = rand(d, 10000);
>PyPlot.plt.hist(x,10,histtype = "step")
>ylabel("Frequência")
>xlabel("x (u.m.)")
```




```
>n, N = length(x), 10^6
>Médias = [mean(rand(x, n)) for i in 1:N]
>PyPlot.plt.hist(Médias,1000,histtype = "step")
>ylabel("Frequência")
>xlabel("x (u.m.)")
>legend("média da amostra")
```



Percebe-se que para todas as distribuições, simétrica, assimétrica e retangular, as distribuições das médias resultantes aproximam-se da normalidade, como se pode observar nos gráficos obtidos.

5.4 RESUMO GRÁFICO DE DADOS

Para a descrição estatística de um conjunto de dados é útil que, inicialmente, se as consiga para a amostra, bem como a obtenção de gráficos que podem contribuir para revelar informações. Assim, os Engenheiros Florestais podem combinar as técnicas estatísticas e gráficas para resumir os dados. No método gráfico, informações gerais podem ser extraídas de maneira rápida, enquanto informações específicas e de quantidade podem ser obtidas por meio das estatísticas.

O resumo numérico por meio de gráficos é útil e eficiente, porque permite obter informações de maneira geral e, também, de valores específicos presentes nos dados. Além disso, o baixo custo computacional atual dos computadores e uma grande gama de *softwares* livres para tal fim, têm contribuído decisivamente para que os métodos gráficos sejam utilizados para descrição de dados. Existem vários programas que permitem obtenção de gráficos especializados para determinados tipos de dados, os quais podem ser estudados em bibliografia específica. Na linguagem Julia existem centenas de gráficos que podem ser construídos utilizando as bibliotecas gráficas PyPlot, Gadfly, StatsPlots, entre outras. As diferentes bibliotecas podem ser exploradas para se construir de maneira fácil os gráficos com objetivo de entender os dados por meio visual.

5.4.1 DISTRIBUIÇÃO DE FREQUÊNCIA E HISTOGRAMAS

A distribuição de frequências é um resumo dos dados na qual as frequências das variáveis são descritivas em intervalos de classes. Ela permite obter uma impressão visual da forma da distribuição das medidas, bem como fornece uma ideia da tendência central (média) e da dispersão dos dados (desvio-padrão). Embora não seja a maneira mais apropriada, fornece uma visão acerca da possível distribuição de probabilidade da variável. Isso permite identificar aspectos simétricos ou assimétricos, aspectos unimodais ou bimodais, entre outras informações. Também é possível transformar o histograma em um gráfico de frequências acumuladas, o qual auxilia na interpretação dos dados.

Os histogramas de frequência simples e de frequência acumulada geralmente são utilizados para descrever variáveis quantitativas, enquanto os gráficos de barras são utilizados para variáveis qualitativas, usualmente na escala nominal.

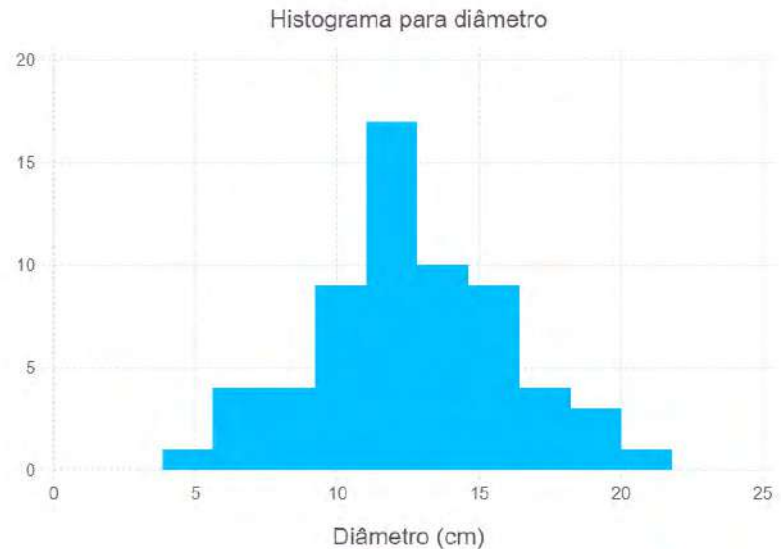
Na Figura 5.6 está apresentado um exemplo típico de histograma observado em florestas plantadas para a variável diâmetro à altura do peito (d), e o histograma das frequências acumuladas (Figura 5.7).

Para obtenção dos gráficos utiliza-se os comandos:

ARQUIVO: 9 - RESUMO GRÁFICO DE DADOS

```
#Histograma
>using DataFrames, CSV, Gadfly
>X=CSV.read("Z:/Alexandre/Julia/Exemplos/descritiva.csv", DataFrame)
>plot(X, x=:d, Geom.histogram(bincount=10), Guide.xlabel("Diâmetro (cm)"),
Guide.ylabel("Frequência"), Guide.title("Histograma para diâmetro"))
>Fig1 = SVG("Z:/Alexandre/Julia/Exemplos/g1.svg", 14cm, 8cm)
```

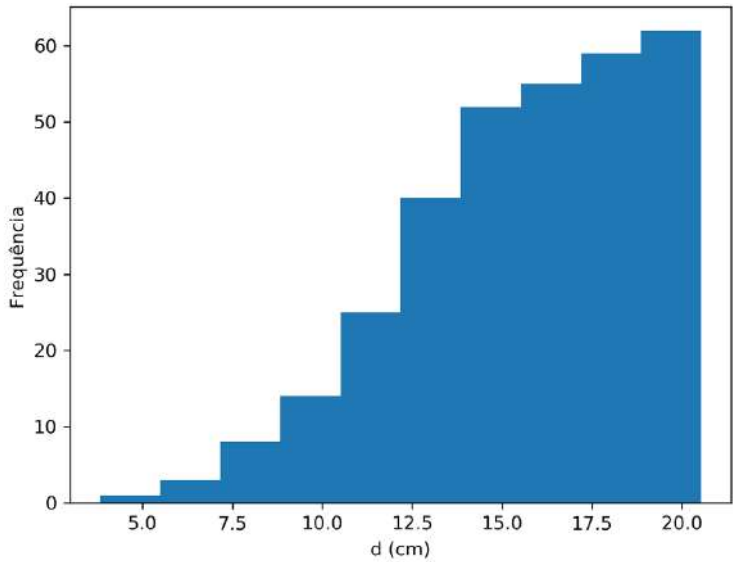
FIGURA 5.6. Histograma de frequências para a variável diâmetro à altura do peito obtidas utilizando-se o pacote Gadfly.



Para obtenção dos gráficos utiliza-se os comandos:

```
#Histograma de frequências acumuladas
>using DataFrames, CSV, PyPlot
>X=CSV.read("Z:/Alexandre/Julia/Exemplos/descritiva.csv", DataFrame)
>plt[:hist](X.d, 10, cumulative="true")
>xlabel("d (cm)")
>ylabel("Frequência")
>savefig("Z:/Alexandre/Julia/Exemplos/g2s.svg")
>savefig("Z:/Alexandre/Julia/Exemplos/gsp.png")
>savefig("Z:/Alexandre/Julia/Exemplos/g2.png", dpi=300)
```

FIGURA 5.7. Histograma de frequências acumuladas para a variável diâmetro à altura do peito utilizando o pacote PyPlot.



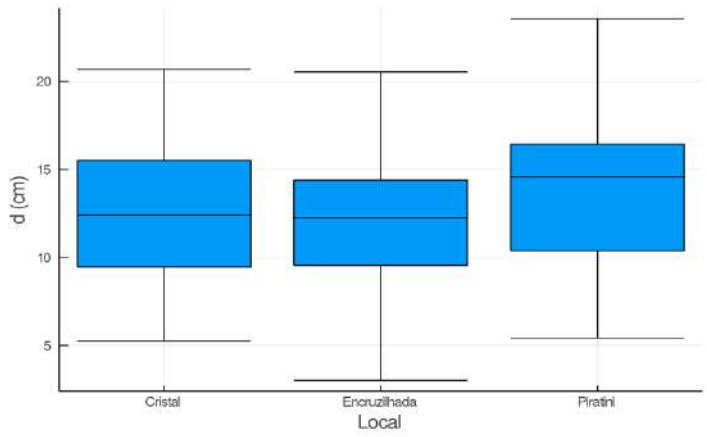
5.4.2 DIAGRAMAS DE CAIXA (BOX PLOT)

O diagrama de caixa descreve várias características dos dados: o centro, a dispersão, a simetria, os *outliers* (valores extremos) e os quartis. Esse gráfico é útil em comparações entre conjuntos de dados para a mesma variável. Na Figura 5.8 está apresentado um diagrama de caixa para a variável diâmetro à altura do peito de florestas de acácia-negra, com a mesma idade, cultivadas em três municípios do estado do Rio Grande do Sul.

Para obtenção dos gráficos utiliza-se os comandos:

```
#Box-plot
>using DataFrames, CSV, StatsPlots
>X=CSV.read("Z:/Alexandre/Julia/Exemplos/daplocais.csv", DataFrame)
@df X boxplot(:Local,:d, xlabel = "Local", ylabel = "d (cm)", legend = false,
dpi=300)
>savefig("Z:/Alexandre/Julia/Exemplos/g3.png")
```

FIGURA 5.8. Diagrama de caixa para a variável diâmetro à altura do peito de florestas de acácia-negra com mesma idade, utilizando o pacote StatsPlots.



Três caixas foram criadas, uma para cada local. Nas caixas estão indicadas a mediana (linha horizontal dentro da caixa) e os quartis, o primeiro representado pela linha mais inferior até a linha inferior da caixa, o segundo entre a linha inferior da caixa e a mediana, o terceiro entre a mediana e a linha superior da caixa e o quarto entre a linha superior da caixa e a linha horizontal superior. Valores extremos (*outliers*) aparecerão como pontos abaixo da linha inferior ou acima da linha superior.

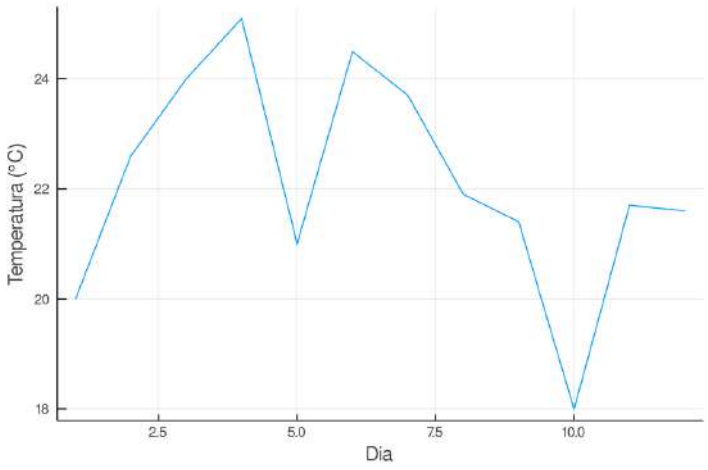
5.4.4 GRÁFICO DE LINHA

O gráfico de linha revela informações do comportamento de uma variável ao longo do eixo x, o que facilita a observações de tendências e de valores extremos. Na Figura 5.9 está representado o comportamento da temperatura média do ar dentro de uma estufa de produção de mudas ao longo de 10 dias.

Para obtenção dos gráficos utiliza-se os comandos:

```
#Linha
>using DataFrames, CSV, Plots
>X=CSV.read("Z:/Alexandre/Julia/Exemplos/temperatura.csv", DataFrame)
>plot(X.Dia, X.T, xlabel = "Dia", ylabel = "Temperatura (°C)",legend =false,
dpi=300)
>savefig("Z:/Alexandre/Julia/Exemplos/g4.png")
```

FIGURA 5.9. Comportamento da temperatura média do ar dentro de uma estufa de produção de mudas ao longo de 10 dias, utilizando o pacote Plots.

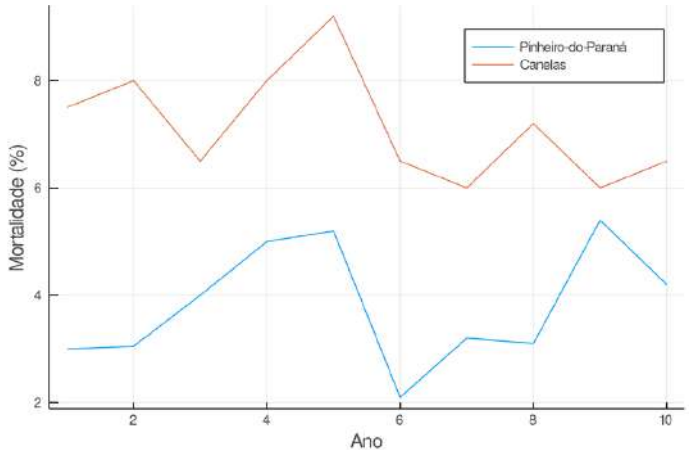


Na Figura 5.10 está apresentado o comportamento da mortalidade do Pinheiro-do-Paraná e de várias espécies de canelas encontradas, em uma Floresta Ombrófila Mista Montana, ao longo de uma década de monitoramento.

Para obtenção dos gráficos utiliza-se os comandos:

```
>using DataFrames, CSV, Plots
>X=CSV.read("Z:/Alexandre/Julia/Exemplos/mortalidade.csv", DataFrame)
>plot(X.Ano, [X.Pinheiro, X.Canelas],xlabel = "Ano", ylabel = "Mortalidade (%)",
label=["Pinheiro-do-Paraná","Canelas"],legend =true, dpi=300)
>savefig("Z:/Alexandre/Julia/Exemplos/g5.png")
```

FIGURA 5.10. Mortalidade do Pinheiro-do-Paraná e de várias espécies de canelas observada em uma Floresta Ombrófila Mista Montana ao longo da década de monitoramento, utilizando o pacote Plots.



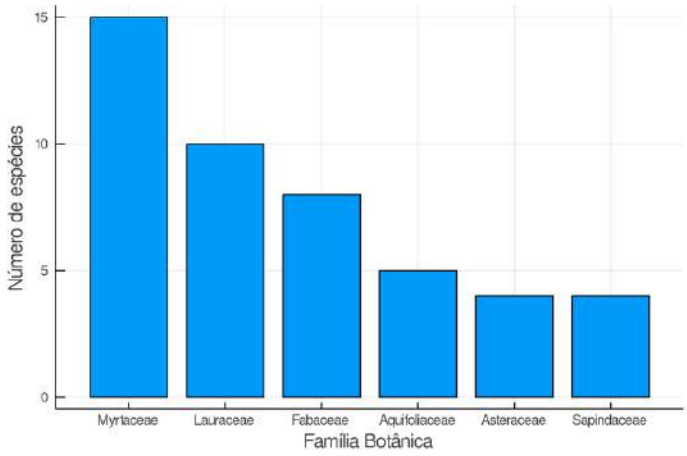
5.4.5 GRÁFICO DE BARRAS

O gráfico de barras revela informações da frequência de uma variável por categoria. Na Figura 5.11 está apresentado o número de espécies, por família botânica, observadas em uma Floresta Ombrófila Mista Montana, ao longo de uma década de monitoramento. Estas famílias representam 47% das espécies encontradas.

Para obtenção dos gráficos utiliza-se os comandos:

```
#Barras
>using DataFrames, CSV, Plots
>X=CSV.read("Z:/Alexandre/Julia/Exemplos/familias.csv", DataFrame)
>bar(X.Familia,X.NE,xlabel = "Família Botânica", ylabel = "Número de espécies",
legend =false, dpi=300)
>savefig("Z:/Alexandre/Julia/Exemplos/g6.png")
```

FIGURA 5.11. Número de espécie por família botânica observada em uma Floresta Ombrófila Mista Montana ao longo da década de monitoramento, utilizando o pacote Plots.



Os gráficos de barra são utilizados para variáveis nominais e se diferenciam dos histogramas por terem um espaço entre as barras.

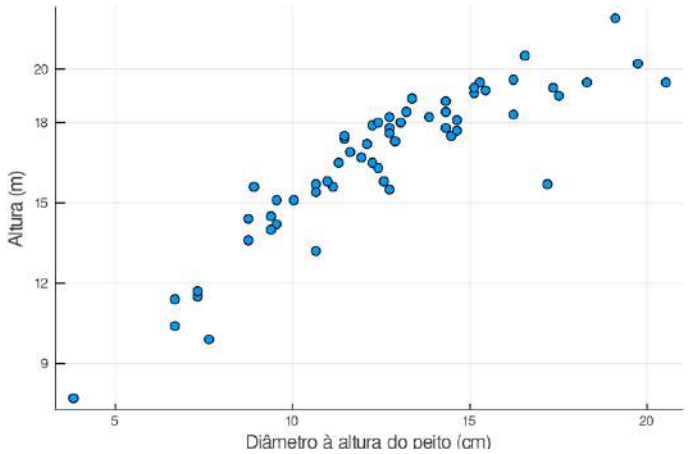
5.4.6 GRÁFICO DE DISPERSÃO

O gráfico de dispersão é utilizado para avaliar se existe um padrão no comportamento dos dados entre duas variáveis. Na Figura 5.12 está apresentada a relação entre diâmetro à altura do peito e altura das árvores de uma floresta de acácia-negra com 10 anos.

Para obtenção dos gráficos utiliza-se os comandos:

```
#Dispersão
>using DataFrames, CSV, Plots
>X=CSV.read("Z:/Alexandre/Julia/Exemplos/descritiva.csv", DataFrame)
>scatter(X.d, X.h,xlabel = "Diâmetro à altura do peito (cm)", ylabel = "Altura (m)",legend
=false, dpi=300)
>savefig("Z:/Alexandre/Julia/Exemplos/g7.png")
```

FIGURA 5.12. Relação entre diâmetro à altura do peito e altura das árvores de uma floresta de acácia-negra com 10 anos, utilizando o pacote Plots.



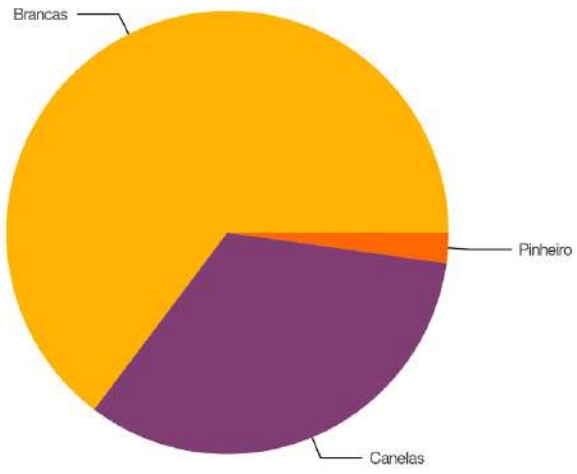
5.4.7 GRÁFICO DE SETORES

O gráfico de setores é utilizado para avaliar a proporção existente entre grupos. Na Figura 5.13 está apresentado o número de árvores ingressas entre os grupos ecológicos Araucária, Canelas e Brancas, observadas em uma área de 3 hectares de floresta Ombrófila Mista, durante uma década de monitoramento. O ingresso é obtido pela contagem de árvores que na primeira ocasião não se encontravam em determinada classe, mas que entraram em uma segunda ocasião de mensuração e no presente caso é a quantidade de árvores que atingiu o porte de 10 cm de diâmetro à altura do peito.

Para obtenção dos gráficos utiliza-se os comandos:

```
#Setores
>using DataFrames, CSV, Plots
>X=CSV.read("Z:/Alexandre/Julia/Exemplos/ingresso.csv", DataFrame)
>pie(X.Grupo,X.Ingresso)
>savefig("Z:/Alexandre/Julia/Exemplos/g8.png")
```

FIGURA 5.13. Proporção de árvores ingressas entre os grupos ecológicos Araucária, Canelas e Brancas em uma área de 3 hectares de Floresta Ombrófila Mista durante uma década, utilizando o pacote Plots.



Os gráficos de setores devem ser utilizados com muito cuidado. Eles são eficientes para um número pequeno de classes (menor do que 5), pois, caso contrário, se tornam de difícil interpretação, requerendo a leitura das frequências nas classes, o que seria mais eficiente se fosse efetuada por meio de um gráfico de barras. Outra limitação é o uso das cores, pois quanto maior o número de classes, maior o número de cores utilizadas, o que dificulta sua interpretação com uma simples observação do gráfico. Gráficos são feitos para ser vistos e não lidos.

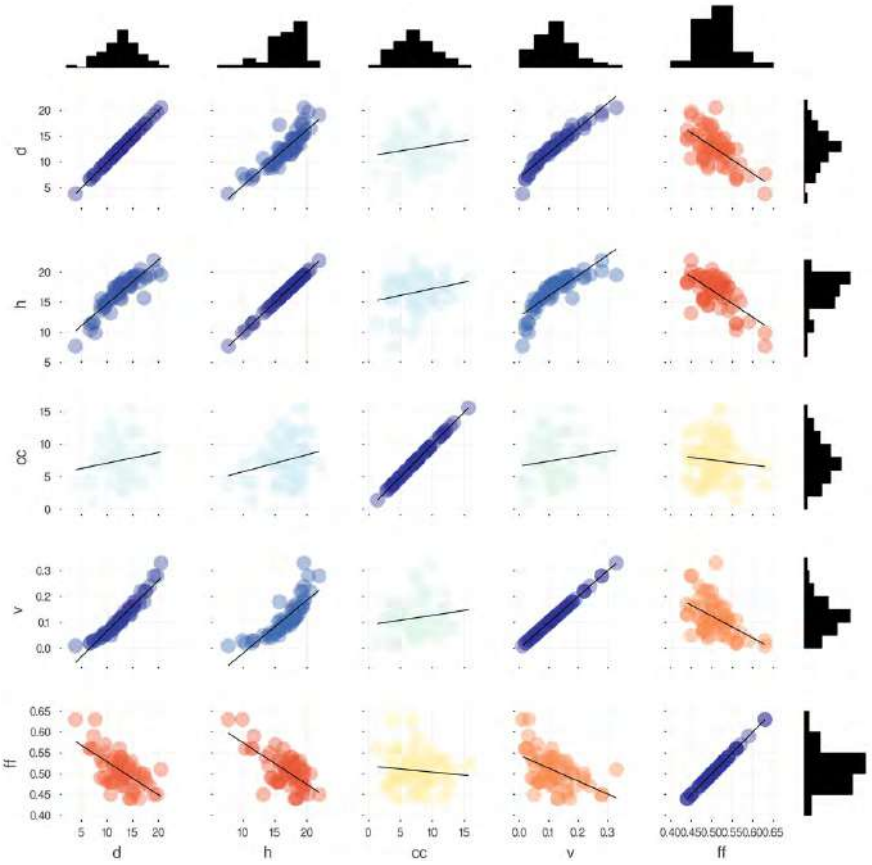
5.4.8 GRÁFICOS BIVARIADOS

Os gráficos bivariados são úteis para se entender a relação entre variáveis com a visualização das possíveis correlações existentes. Na Figura 5.14 está apresentada a relação entre as variáveis diâmetro à altura do peito, altura, comprimento da copa, volume do fuste e fator de forma para árvores de acácia-negra. A cor azul indica correlação positiva, vermelha negativa e amarela ausência de correlação.

Para obtenção dos gráficos utiliza-se os comandos:

```
#Bivariados
>using DataFrames, CSV, StatsPlots
>X=CSV.read("Z:/Alexandre/Julia/Exemplos/descritiva.csv", DataFrame)
>gr(size = (900, 900))
>@df X cornerplot([:d :h :cc :v :ff], grid = false, dpi=500)
>savefig("Z:/Alexandre/Julia/Exemplos/g9.png")
```

FIGURA 5.14. Correlação entre as variáveis diâmetro à altura do peito, altura, comprimento da copa, volume do fuste e fator de forma para árvores de acácia-negra com 10 anos, utilizando o pacote StatsPlots.



EXERCÍCIO PROPOSTO 5: ANÁLISE EXPLORATÓRIA DE DADOS

Neste exercício podem ser utilizados os dados do exercício sobre a avaliação das mudas conduzidas por um Engenheiro Florestal, apresentado na seção 3. A tarefa proposta é explorar e comparar os dados dos diferentes clones utilizando-se estatísticas descritivas, tabelas e gráficos.

6 TESTES DE HIPÓTESES

Os Engenheiros Florestais formulam muitas hipóteses ao longo de sua vida profissional. O que seria uma hipótese? De maneira simples e objetiva, hipótese é uma pergunta para qual se deseja a resposta, com base em informações científicas. Para tanto, procedimentos estatísticos são usados para tomar decisões sobre a população com base nas amostras, ou seja, valendo-se do raciocínio indutivo.

Os procedimentos estatísticos para comparar duas ou mais médias de tratamentos são efetuados por meio de um teste de hipótese. Os tratamentos representam as diferentes situações para as quais se deseja fazer inferências. Dentro do método científico, parte-se do princípio que não existe diferenças entre as situações testadas. Isto configura o que é chamado de hipótese de nulidade (H_0). A hipótese H_0 é sempre, inicialmente, considerada como verdadeira, podendo os resultados confirmarem ou negarem este pressuposto. Isto indica que as possíveis diferenças observadas entre as médias ocorrem devido ao acaso.

Entretanto, se na análise a hipótese da nulidade é rejeitada conclui-se que pelo menos um dos tratamentos testados tem efeito significante em relação aos demais, então é possível dizer que as médias são significantemente diferentes. As diferenças podem ser altamente siginificantes, dependendo do nível de probabilidade utilizado.

O processo estatístico usado para decidir se a hipótese da nulidade deve ser aceita ou rejeitada é denominado teste de hipóteses. Estes testes podem ser paramétricos ou não paramétricos, univariados ou multivariados. Estes são compostos por vários testes estatísticos, os quais são úteis para a análise de dados de experimentos efetuados pelos Engenheiros Florestais.

Os testes de hipótese sobre populações, cujas variáveis apresentaram distribuição normal, são usualmente efetuados sobre os parâmetros μ e σ (parâmetros da população). Na sequência é apresentado um exemplo de teste de hipótese.

EXERCÍCIO 6

O Engenheiro Florestal, responsável por uma serraria, sabe que a densidade média da população de tábuas de araucária é de 0,49 g.cm^{-3} (Figura 6.1). Para um novo lote de toras, oriundo de um povoamento manejado com 40 anos, foram produzidas tábuas e foi coletada uma amostra de dados de densidade da madeira com as seguintes características: n = 100, $\overline{x}$ = 0,53 g.cm^{-3} e s = 0,012 g.cm^{-3}, com distribuição normal dos dados. Assim, deseja-se saber se os dados amostrais com $\overline{x}$ = 0,53 g.cm^{-3} diferem ou não do valor paramétrico (da população) μ = 0,49 g.cm^{-3} ?

Se for admitido que μ = 0,49 g.cm^{-3} e aplicando o teorema do limite central – já evidenciado na seção 5, sabe-se que as médias amostrais tendem a se distribuir normalmente com $\mu_{\overline{x}} = \mu$ = 0,49 g.cm^{-3} e $s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{0{,}012}{\sqrt{100}} = 0{,}0012\ g.cm^{-3}$.

Se μ é realmente 0,49 g.cm^{-3}, então supõem-se que 95% das médias amostrais estariam compreendidas entre o limite inferior e superior (área verde – Figura 6.2), conforme cálculos apresentados. Neste caso, supõem-se que 95% das médias amostrais deveriam estar a menos de ± 1,98 desvios-padrões de μ, representado graficamente na Figura 6.2, construída utilizando-se a função *Hipótese*.

Note que a estatística do teste, denotada por z da amostra, é maior que 1,98 desvios-padrões. Observa-se que $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$.

FIGURA 6.1. Amostra para coleta de dados de densidade média de tábuas de Araucária.



Para obtenção dos resultados utiliza-se os comandos:

Arquivo: 10 - Teste de hipóteses

```
#Teste de hipóteses
>using Distributions

>media=0.53;
>mu=0.49;
>s=0.012;
>n=100;
>sx=s/sqrt(n);
>alpha = 0.05;
>tVal = quantile(TDist(n-1),1-alpha/2)
1.9842
```

```
>LimiteInferior = (mu- tVal * sx)
0.4876

>LimiteSuperior = (mu+ tVal * sx)
0.4924

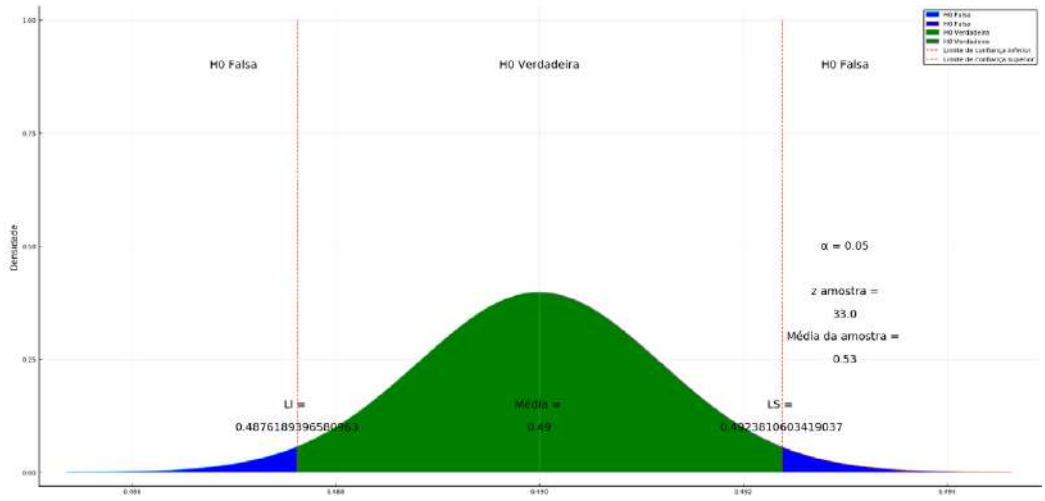
#Função HIPOTESE
>using Distributions, Plots; pyplot()

>function Hipotese(n,DesvioPadrao,media,mu)
    Alpha=0.05
    media=media
    n= n
    xbar=mu
    s=DesvioPadrao
    sx=s/sqrt(n)
    tVal = quantile(TDist(n-1),1-Alpha/2)
    z=(media-mu)/sx
    z=round(z)
    ICInferior = (xbar- tVal * sx)
    ICSuperior = (xbar+ tVal * sx)
    xGridInfH0=(Alpha/2):0.0001:0.5
    xGridSupH0=0.5:0.0001:(1-(Alpha/2))
    xGridInfH1=0:0.0001: (Alpha/2)
    xGridSupH1=(1-(Alpha/2)):0.0001:1
    H0Inf = quantile.(TDist(n-1),xGridInfH0)
    H0Sup= quantile.(TDist(n-1),xGridSupH0)
    H1Inf = quantile.(TDist(n-1),xGridInfH1)
    H1Sup= quantile.(TDist(n-1),xGridSupH1)
    DensidadeH0Inf= pdf.(TDist(n),H0Inf)
    DensidadeH0Sup= pdf.(TDist(n),H0Sup)
    DensidadeH1Inf= pdf.(TDist(n),H1Inf)
    DensidadeH1Sup= pdf.(TDist(n),H1Sup)
    XH1Inf = xbar.+(H1Inf.*sx)
    XH1Sup = xbar.+( H1Sup.*sx)
    XH0Inf = xbar.+(H0Inf.*sx)
    XH0Sup = xbar.+( H0Sup.*sx)
    plot!( XH1Inf, DensidadeH1Inf,fill=(0, :blue), label="H0 Falsa", dpi=300)
    plot!( XH1Sup, DensidadeH1Sup,fill=(0, :blue), label="H0 Falsa")
    plot!( XH0Inf, DensidadeH0Inf,fill=(0, :green), label="H0 Verdadeira")
    plot!( XH0Sup, DensidadeH0Sup,fill=(0, :green), label="H0 Verdadeira")
    plot!([ ICInferior, ICInferior],[0,1],c=:red, ls=:dash,
label="Limite de confiança inferior", xlabel="Valor", ylabel="Densidade")
    plot!([ ICSuperior, ICSuperior],[0,1],c=:red, ls=:dash,
label="Limite de confiança superior")
    A=(ICInferior+ICSuperior)/2
    B=ICInferior
    C=ICSuperior
    D=(ICInferior+ICSuperior)/2
    E= ICSuperior
```

```
    F=ICInferior
    annotate!([(A, 0.9, text("H0 Verdadeira")),(0.493, 0.9, text("H0 Falsa")),
(0.487, 0.9, text("H0 Falsa")),(D, 0.15, text("Média = ")), (D, 0.1, xbar),
(F, 0.15, text("LI = ")), (F, 0.1, ICInferior), (E, 0.15, text("LS = ")),
(E, 0.1, ICSuperior), (0.493, 0.5, text("α = 0.05")),(0.493, 0.4,
text("z amostra =")),(0.493, 0.35, z), (0.493, 0.3,
text("Média da amostra = ")),(0.493, 0.25, media)])
end

#Hipotese(n,DesvioPadrao,media,mu)
>Hipotese(100,0.012,0.53,0.49)
```

FIGURA 6.2. Distribuição de médias amostrais de densidade de tábuas (g.cm³) de Araucária.



Interpretação do resultado

- A média da amostra, encontrada pelo Engenheiro Florestal para o novo lote de tábuas, foi de $\overline{x}$ = 0,53 g.cm^{-3}.
- A média paramétrica da densidade básica é conhecida μ = 0,49 g.cm^{-3}.
- Avaliando-se a distribuição de médias amostrais obtidas para o tamanho da amostra medido resultou num intervalo de confiança de 0,4876 g.cm^{-3} a 0,4924 g.cm^{-3}. A média amostral é 0,53 g.cm^{-3} e maior do que o limite superior do intervalo de confiança. Isto significa que existe 5% de chance (área azul no gráfico) de este resultado ter ocorrido por chance, se realmente μ = 0,49 g.cm^{-3}.
- Motivos para a obtenção da média amostral 0,53 g.cm^{-3}: *i)* normalmente as tábuas de araucárias são produzidas a partir de toras oriundas de povoamentos com idades inferiores a do novo lote; *ii)* pode ter acontecido um evento raro na serraria, se as araucárias foram cortadas com elevada densidade da madeira; *iii)* a média não é realmente 0,49 g.cm^{-3} para a amostra obtida.
- Como a probabilidade de se obter uma média amostral $\overline{x}$ = 0,53 g.cm^{-3} é bem pequena, o que pode ser evidenciado na Figura anterior, é razoável admitir que o valor da média não é realmente 0,49 g.cm^{-3} e isto sugere que deve ser rejeitada a suposição de que μ = 0,49 g.cm^{-3}.

6.1 TESTE DE HIPÓTESES ESTATÍSTICAS E NÍVEL DE SIGNIFICÂNCIA

Em um experimento no qual tratamentos são comparados é necessário estabelecer hipóteses estatísticas para auxiliar o pesquisador a tomar decisão sobre os seus efeitos. Duas hipóteses são estabelecidas:

1 H_0: HIPÓTESE DA NULIDADE
Na qual se declara não existir diferenças entre os tratamentos que estão sendo testados, isto é, que os efeitos dos tratamentos são nulos.

2 H_1: HIPÓTESE ALTERATIVA
Na qual se declara que os tratamentos diferem entre si, ou seja, existem evidências reais dos efeitos dos tratamentos.

Ao testar a hipótese nula, por meio de um teste estatístico, está-se sujeito a uma margem de erro que julga-se aceitável na conclusão do teste, ou seja, o nível de significância. As conclusões ora podem ser corretas e ora podem ser incorretas, originando dois tipos de erros em um teste de hipótese:

- **Erro tipo I ou α:** rejeita-se a hipótese da nulidade quando na verdade ela deveria ser aceita.
- **Erro tipo II ou β:** aceita-se a hipótese da nulidade quando na verdade ela deveria ser rejeitada.

Ressalta-se que a ocorrência desses erros nas conclusões pode não ser devida aos cálculos malfeitos, ou experimentos malconduzidos, ou seja, é um erro que pode ocorrer devido ao acaso.

A probabilidade de ocorrer o erro tipo I na decisão dos testes de hipóteses em um experimento está sob o controle do pesquisador e é determinada pelo nível de significância especificado, comumente 0,01(1%) ou 0,05 (5%). Ao nível de 0,05 (5/100) de probabilidade é esperado rejeitar a hipótese da nulidade não mais que cinco em cada cem vezes, quando de fato a hipótese é verdadeira. Portanto, o nível de significância corresponde a probabilidade máxima de ocorrer um erro tipo I.

A frequência em que o erro tipo II ocorre depende de vários fatores e, entre eles, pode-se apontar: a precisão com que as estatísticas são calculadas, o número de repetições, o delineamento experimental adotado e a qualidade do experimento conduzido, ou seja, depende do planejamento e condução do experimento. Dependendo do pesquisador, os testes de hipóteses podem ser unilaterais ou bilaterais, uma vez que depende do critério estabelecido, o qual deve declarar as hipóteses que deseja testar a priori, em termos estatísticos. Isto vai determinar se o teste é bilateral ou unilateral, conforme especificado adiante.

Em um teste de hipótese sempre é testada H_0, a qual pressupõe não existir diferenças reais dos efeitos médios dos tratamentos, ou seja, parte-se do princípio da nulidade de efeitos. O teste estatístico a ser escolhido para julgar se a hipótese em questão deve ser aceita ou rejeitada depende da quantidade de tratamentos testados, do tipo de tratamento que é testado, do tipo da variável aleatória mensurada no experimento e, ainda, das exigências dos testes estatísticos a serem aplicados. Daí decorre a necessidade de o Engenheiro Florestal conhecer os tipos de variáveis, tipos de testes estatísticos e as condições requeridas para sua aplicação.

A estatística de um teste é baseada nos dados amostrais. A maioria dos testes estatísticos aplicados por Engenheiros Florestais é paramétrica, em função de suas exigências. Testes não paramétricos podem, algumas vezes, ser aplicados. Estes testes exigem, entretanto, procedimentos específicos para sua aplicação.

Os testes paramétricos estão baseados em variáveis métricas (de intervalo ou de razão), e as principais exigências destes testes referem-se à distribuição normal da variável aleatória, variância homogênea nos grupos testados e observações independentes.

Ao longo deste texto alguns testes paramétricos estão apresentados, limitados à técnicas univariadas e divididos em: teste para uma amostra, teste para duas amostras independentes, testes para duas amostras relacionadas (pareadas) e testes para três ou mais médias. Independentemente do tipo de teste estatístico a ser utilizado, um teste de hipótese é testado por meio da seguinte sequência:

1. Especificação do problema, que determina a forma de condução do experimento.
2. Estabelecimento da hipótese nula, bem como do nível de significância.
3. Escolha do teste estatístico a ser utilizado.
4. Obtenção dos dados experimentais e avaliação das condições para aplicação do teste estatístico.
5. Obtenção do valor do teste estatístico e conclusão da aceitação ou rejeição da hipótese da nulidade.
6. Conclusão a respeito do problema estabelecido (pergunta formulada).

6.1.1 CONCLUSÕES NOS TESTES DE HIPÓTESES

Como já evidenciado, o processo final em um teste de hipótese é tomar decisões sobre a hipótese nula, ou seja, fazer uma escolha entre as diferentes possibilidades. A forma de se escrever as conclusões são as seguintes:

i) Não rejeitar a hipótese nula H_0

"à evidência amostral não é suficientemente forte para rejeitar a hipótese de não haver diferenças...."
"não há evidência suficiente para garantir a rejeição de que...."
"A probabilidade de H_0 ser verdadeira é representada pela cor verde nas próximas três figuras, ocupando a área $1-\alpha = 0,95$".

ii) Rejeitar a hipótese nula H_0

"à evidência amostral é suficientemente forte para rejeitar a hipótese de não haver diferenças...."
"há evidência suficiente para garantir a rejeição de que...."
"A probabilidade de Ho ser falsa é representada pela cor azul nas próximas três figuras, para um nível $\alpha = 0,05$".

A estatística de um teste amostral é utilizada para tomar decisões sobre a rejeição da hipótese nula. Esta será representada por uma barra laranja nas próximas três figuras.

A região crítica em um teste de hipótese, em uma determinada distribuição de probabilidades, refere-se a um conjunto de valores da estatística do teste que direcionam para a rejeição da hipótese nula. Esta será representada pela cor azul nas próximas três figuras.

O valor crítico separa as regiões de aceitação ou rejeição da hipótese. Este valor será representado por uma linha vermelha nas próximas três figuras.

Para obtenção das três figuras citadas no texto utiliza-se os seguintes comandos:

Arquivo: 11 - Função thipotesebi para teste t bilateral

```
#Função THIPOTESEBI para teste t bilateral

>using Distributions, Plots; pyplot()

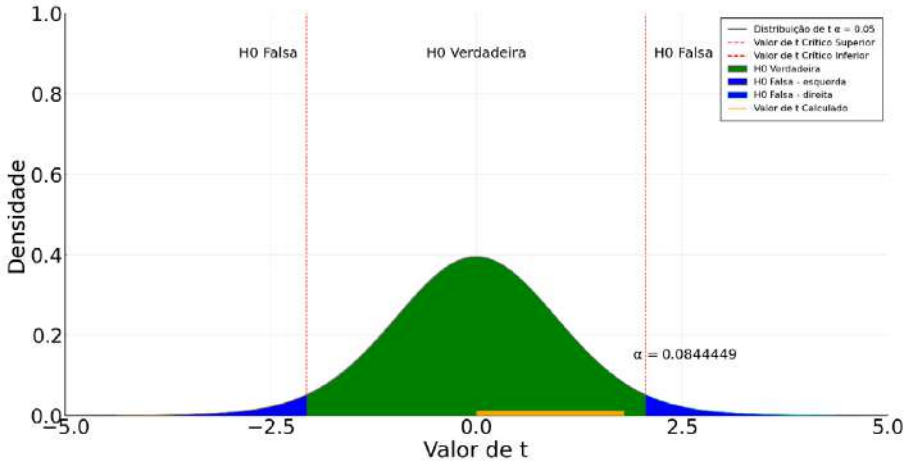
>function tHipoteseBi(n,tCalculado)
    xGrid = -5:0.01:5
    Alpha=0.05
    plot!(xGrid, pdf.(TDist(n),xGrid), c=:black, xlims=(-5, 5), ylims=(0,1),
label="Distribuição de t α = 0.05")
    ValorCriticoS = quantile(TDist(n-1),1-Alpha/2)
    ValorCriticoI =(quantile(TDist(n-1),1-Alpha/2))*-1
    plot!([ ValorCriticoS, ValorCriticoS],[0,1],c=:red, ls=:dash,
label="Valor de t Crítico Superior", xlabel="Valor de t", ylabel="Densidade")
    plot!([ ValorCriticoI, ValorCriticoI],[0,1],c=:red, ls=:dash,
label="Valor de t Crítico Inferior")
    alphaH0=(Alpha/2):0.0001:(1-(Alpha/2))
    H0 = quantile(TDist(n-1),alphaH0)
    DensidadeH0= pdf.(TDist(n),H0)
    plot!(H0, DensidadeH0,fill=(0, :green), label="H0 Verdadeira")
    alphaH1I=0:0.0001:(Alpha/2)
    alphaH1S=(1-(Alpha/2)):0.0001:1
    H1I = quantile(TDist(n-1),alphaH1I)
    H1S = quantile(TDist(n-1),alphaH1S)
    DensidadeH1I= pdf.(TDist(n),H1I)
    DensidadeH1S= pdf.(TDist(n),H1S)
    plot!(H1I, DensidadeH1I,fill=(0, :blue), label="H0 Falsa - esquerda")
    plot!(H1S, DensidadeH1S,fill=(0, :blue), label="H0 Falsa - direita")
    plot!([0, tCalculado],[0,0],c=:orange, lw=12, label="Valor de t Calculado")
    α = 2*ccdf(TDist(n-1), abs(tCalculado))
    A=(ValorCriticoI+ValorCriticoS)/2
    B=ValorCriticoS*1.15
    C=ValorCriticoI*1.15
    D=tCalculado*1.15
    annotate!([(A, 0.9, text("H0 Verdadeira")),(B, 0.9, text("H0 Falsa")),
(C, 0.9, text("H0 Falsa")),(D, 0.15, text("α =")), ((D+0.2), 0.1, α)])
end

#tHipoteseBi(n,tCalculado)
>tHipoteseBi(25,1.8)
```

Quando a hipótese é bilateral isto significa que a média da amostra pode ser menor do que o limite inferior ou maior do que o limite superior (áreas azuis). A probabilidade associada a estas áreas é $\frac{\alpha}{2}$, ou seja, a metade do erro tipo I, pois a distribuição é simétrica.

No exemplo da Figura 6.3 a linha em cor laranja representa o resultado obtido a partir da amostra. Como ele encontra-se dentro da área verde, aceita-se a hipótese da nulidade.

FIGURA 6.3. Teste de hipótese t bilateral.



ARQUIVO: 12 - FUNÇÃO THIPOTESEDI PARA TESTE T LATERAL À DIREITA

```
using Distributions, Plots; pyplot()

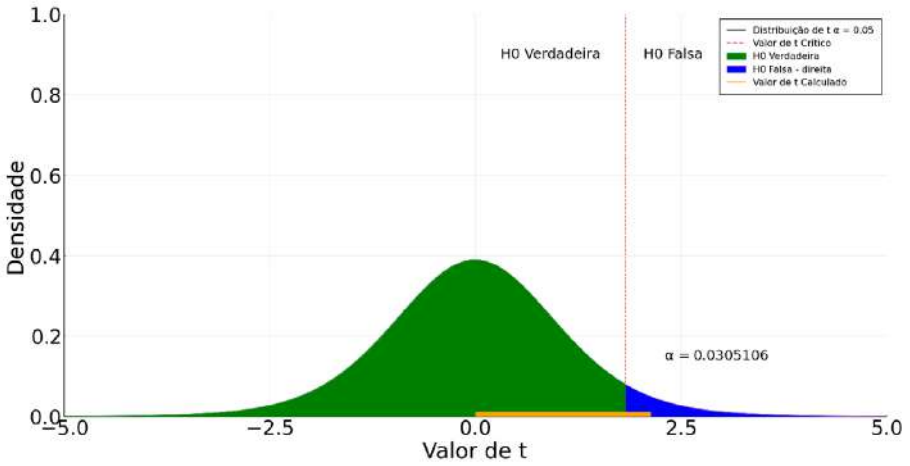
function tHipoteseDi(n,tCalculado)
    xGrid = -5:0.01:5
    Alpha=0.05
         plot!(xGrid, pdf.(TDist(n),xGrid), c=:black, xlims=(-5, 5), ylims=(0,1),
label="Distribuição de t α = 0.05",
              tickfontsize=26, guidefontsize=28, legendfontsize=12)
    ValorCritico = quantile(TDist(n-1),1-Alpha)
      plot!([ ValorCritico, ValorCritico],[0,1],c=:red, ls=:dash, label="Valor de t Crítico",
xlabel="Valor de t", ylabel="Densidade")
    alphaH0=0:0.0001:(1-Alpha)
    H0 = quantile(TDist(n-1),alphaH0)
    DensidadeH0= pdf.(TDist(n),H0)
    plot!(H0, DensidadeH0,fill=(0, :green), label="H0 Verdadeira")
    alphaH1=(1-Alpha):0.0001:1
    H1 = quantile(TDist(n-1),alphaH1)
    DensidadeH1= pdf.(TDist(n),H1)
    plot!(H1, DensidadeH1,fill=(0, :blue), label="H0 Falsa - direita")
    plot!([0, tCalculado],[0,0],c=:orange, lw=12, label="Valor de t Calculado")
    α = round(ccdf(TDist(n-1), abs(tCalculado)), digits = 7)
    A=ValorCritico/2
    B=ValorCritico*1.15
    C=tCalculado*1.15
      annotate!([(A, 0.9, text("H0 Verdadeira", 18)),((B+0.3), 0.9, text("H0 Falsa",
18)),(C, 0.15, text("α =", 18)), ((C+0.65), 0.15, text(α, 18))])
end

#tHipoteseDi(n,tCalculado)
tHipoteseDi(10,2.14)
savefig("ttd.png")
```

No caso de uma hipótese unilateral à direita, isto significa que a média da amostra deve ser maior do que o limite superior (área azul). A probabilidade associada a esta área é α, ou seja, erro tipo I.

No exemplo da Figura 6.4 a linha em cor laranja representa o resultado obtido a partir da amostra. Como ele encontra-se dentro da área azul, rejeita-se a hipótese da nulidade.

FIGURA 6.4. Teste de hipótese t unilateral direita.



ARQUIVO: 13 - FUNÇÃO THIPOTESEES PARA TESTE T LATERAL À ESQUERDA

```
#Função THIPOTESEES para teste t lateral à esquerda

>using Distributions, Plots; pyplot()

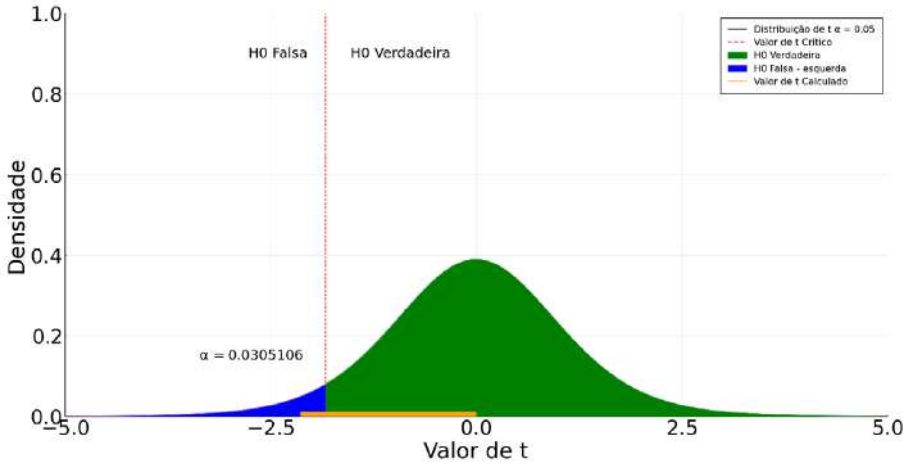
>function tHipoteseEs(n,tCalculado)
    xGrid = -5:0.01:5
    Alpha=0.05
    plot!(xGrid, pdf.(TDist(n),xGrid), c=:black, xlims=(-5, 5), ylims=(0,1),
label="Distribuição de t α = 0.05")
    ValorCritico = quantile(TDist(n-1),1-Alpha)*-1
    plot!([ ValorCritico, ValorCritico],[0,1],c=:red, ls=:dash,
label="Valor de t Crítico", xlabel="Valor de t", ylabel="Densidade")
    alphaH0=Alpha:0.0001:1
    H0 = quantile(TDist(n-1),alphaH0)
    DensidadeH0= pdf.(TDist(n),H0)
    plot!(H0, DensidadeH0,fill=(0, :green), label="H0 Verdadeira")
    alphaH1=0:0.0001:Alpha
    H1 = quantile(TDist(n-1),alphaH1)
    DensidadeH1= pdf.(TDist(n),H1)
    plot!(H1, DensidadeH1,fill=(0, :blue), label="H0 Falsa - esquerda")
    plot!([0, tCalculado],[0,0],c=:orange, lw=12, label="Valor de t Calculado")
    α = ccdf(TDist(n-1), abs(tCalculado))
    A=ValorCritico/2
```

```
    B=ValorCritico*1.15
    C=tCalculado*1.15
    annotate!([(A, 0.9, text("H0 Verdadeira")),(B, 0.9, text("H0 Falsa")),
(C, 0.15, text("α =")), ((C+0.2), 0.1, α)])
end

#tHipoteseEs(n,tCalculado)
>tHipoteseEs(10,-2.14)
```

Para uma hipótese unilateral à esquerda, isto significa que a média da amostra deve ser menor do que o limite inferior (área azul). A probabilidade associada a esta área é α, ou seja, erro tipo I. No exemplo da Figura 6.5 a linha em cor laranja representa o resultado obtido a partir da amostra. Como ele encontra-se dentro da área azul, rejeita-se a hipótese da nulidade.

FIGURA 6.5. Teste de hipótese t unilateral esquerda.



Com base nesses aspectos, utiliza-se o seguinte critério de decisão sobre a hipótese H_0:

i. rejeitar a hipótese de nulidade se a estatística do teste estiver na região crítica (áreas azuis)
ii. aceitar a hipótese de nulidade se a estatística do teste não estiver na região crítica (área verde)

Note que nas figuras é impresso o valor de α, o qual em muitos *softwares* é denominado de *p-valor*. Este valor corresponde à probabilidade associada à estatística do teste, assim, pode-se utilizar os seguintes critérios de decisão sobre a hipótese H_0:

i. se α é menor que 0,01, o teste é dito altamente significante. Existe evidência muito forte para rejeição da hipótese nula.
ii. se α variar de 0,01 a 0,05, o teste é dito significante. Existe evidência contra a aceitação da hipótese nula.
iii. se α for maior 0,05, o teste é dito não significante. Não existe evidência para rejeição da hipótese nula.
iv. ressalta-se que normalmente os níveis de significância nos testes de hipóteses são $\alpha = 0,01$ e $\alpha = 0,05$, e são escolhidos pelo pesquisador antes de proceder o teste estatístico.

Um valor de α muito pequeno sugere evidências para rejeitar a hipótese da nulidade. Quanto menor este valor, conclui-se que os resultados amostrais são extremamente incomuns, e provavelmente não ocorrem ao acaso, mas sim devido à presença de efeitos diferentes entre os tratamentos. Se o valor de α for alto não se rejeita a hipótese da nulidade, pois os resultados amostrais têm alta probabilidade de ocorrência (0,95).

Nas três figuras utilizadas para representar os testes de hipótese, as caudas da distribuição são as regiões extremas delimitadas pelos valores críticos (regiões em azul). Nelas estão representados os testes: bilateral, unilateral à direita e unilateral à esquerda, formulados para o teste de hipótese *t*, o qual será discutido posteriormente. Assim:

Teste bilateral

- Quando a hipótese aceita é a alternativa (H_1), a média de um tratamento testado é diferente da média do outro.
- A forma de se expressar as hipóteses testadas é:

$$\begin{cases} H_0: \mu_{Trat1} = \mu_{Trat2} \\ H_1: \mu_{Trat1} \neq \mu_{Trat2} \end{cases}$$
$$\alpha = 0{,}05$$

- A região crítica (cor azul) está situada nas duas regiões extremas (caudas) da distribuição.
- O nível de significância α é dividido igualmente entre as duas caudas, motivo pelo qual, para determinação dos valores críticos, existe a seguinte expressão na função denominada *tHipoteseBi*:

```
>ValorCriticoS = quantile(TDist(n-1),1-Alpha/2)
>ValorCriticoI =(quantile(TDist(n-1),1-Alpha/2))*-1
```

Nessa expressão foi determinado um valor crítico para a direita (ValorCriticoS) e um valor para a esquerda (ValorCriticoI) da média e α foi divido por 2.

- Note que o valor de α é multiplacado por 2, pois engloba as duas áreas extremas, delimitada pela estatística do teste.

```
>α = 2*ccdf(TDist(n-1), abs(tCalculado))
```

Unilateral à esquerda

- Quando a hipótese alternativa é aceita, a média de um tratamento é menor do que a do outro tratamento, e é expressa por:

$$\begin{cases} H_0: \mu_{Trat1} = \mu_{Trat2} \\ H_1: \mu_{Trat1} < \mu_{Trat2} \end{cases}$$
$$\alpha = 0{,}05$$

- A área da região crítica é α e está localiza na extremidade à esquerda do gráfico.

Unilateral à direita

- Quando a hipótese alternativa é aceita, a média de um tratamento é maior do que a do outro tratamento, e é expressa por:

$$\begin{cases} H_0: \mu_{Trat1} = \mu_{Trat2} \\ H_1: \mu_{Trat1} > \mu_{Trat2} \end{cases}$$
$$\alpha = 0{,}05$$

- A área da região crítica é α e está localiza na extremidade à direita do gráfico.

As três figuras anteriores geradas pelos algoritmos *tHipoteseBi* para teste bilateral, *tHipoteseDi* para unilateral à direita e *tHipoteseEs* para unilateral à esquerda demostram como o sinal de desigualdade em H_1 direciona a região crítica.

A área delimitada pela cor verde nas Figuras pode ser convertida em intervalos de confiança, conforme efetuado no exercício anterior, referente à densidade das tábuas de araucária. Isto pode ser denominado de estimativa intervalar.

Uma estimativa intervalar é um espaço que contém todos os valores prováveis daquela média, o que pode ser realizado para um teste bilateral. Naquele exercício, o valor médio amostral $\bar{x} = 0{,}53$ g.cm^{-3} estava fora do intervalo de confiança para 0,95, o que permitiu alegar que a média das tábuas de araucária era diferente de $\mu = 0{,}49$ g.cm^{-3}.

Para um teste unilateral, a correspondência entre um intervalo de confiança e o teste de hipótese deve considerar a probabilidade de erro 2α, assim os limites de confiança passam a ser $1 - 2\alpha$. Um teste de hipótese unilateral com nível de significância $\alpha = 0{,}05$, corresponde a um intervalo de confiança de 90%, duas vezes α, porque a área fica restrita a um dos lados da distribuição. Entender essas variações nos intervalos de confiança auxiliam na identificação de diferenças significantes nos testes de hipóteses.

6.2 TESTES PARAMÉTRICOS PARA A MÉDIA DE UMA AMOSTRA

Este teste pode ser utilizado pelo Engenheiro Florestal quando apenas uma amostra é obtida e se quer comparar a média dessa amostra contra um valor pré-determinado. Apesar de existirem outros testes, a maioria das aplicações está baseada no teste *t*.

Para tanto, uma amostra aleatória de tamanho *n* em uma determinada população é selecionada de maneira a obter a média e o desvio padrão. A média desta amostra é comparada estatisticamente com um determinado valor fixo. As hipóteses H_0 e H_1 são estabelecidas e o teste *t* é aplicado para decidir se a hipótese da nulidade deve ser aceita ou rejeitada.

As condições para aplicação do teste são: i) a amostra é pequena ($n \leq 30$); σ é desconhecido e a variável tem distribuição normal.

A estatística do teste *t* é dada por:

$$t = \frac{\bar{x} - valor}{\frac{s}{\sqrt{n}}}$$

Em que:
$\bar{x}$ é a média da amostra
s é o desvio-padrão da amostra
n é o tamanho da amostra
valor é um valor padrão, que definirá a hipótese da nulidade

Para obtenção dos valores críticos, utiliza-se o grau de liberdade definido por $GL = n - 1$.

EXERCÍCIO 7

Deseja-se saber se a média das alturas de mudas de acácia-negra encontra-se dentro dos limites estabelecidos pelo viveirista, que é de 200 mm. Mudas com 200 mm de altura apresentam um excelente padrão para o plantio para esta espécie, daí a importância de averiguar se as mudas produzidas atingem o padrão especificado (Figura 6.6). Para esta avaliação, foram selecionadas de maneira aleatória 13 mudas de acácia-negra e a altura foi mensurada utilizando-se uma régua graduada em milímetros. Os dados experimentais estão apresentados na (Tabela 6.1). Note que a média da amostra das 13 mudas é de 197,77 mm, o que permite a formulação da seguinte questão: Existe evidência para afirmar que a média da amostra não difere de 200 mm? Para o teste, considere o nível de significância de 0,05.

FIGURA 6.6. Medição de altura de mudas de acácia-negra.

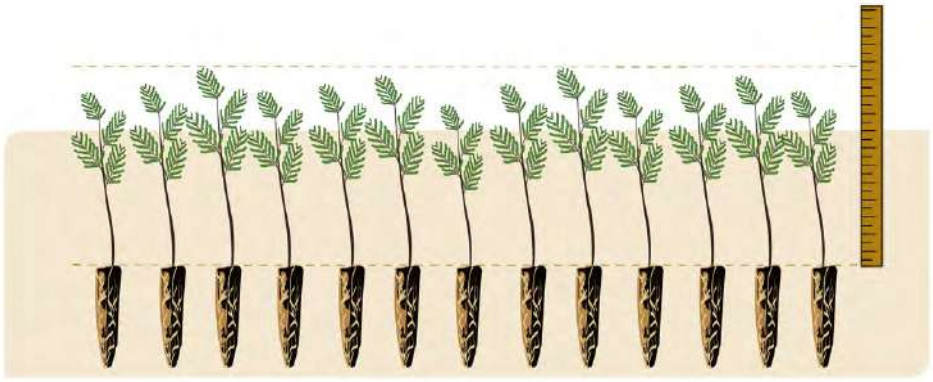


TABELA 6.1. Dados amostrais de altura de mudas (em mm) de acácia-negra.

N	Altura (mm)
1	198
2	185
3	194
4	189
5	180
6	204
7	219
8	190
9	185
10	214
11	198
12	214
13	201

Foram estabelecidas as hipóteses:

$$\begin{cases} H_0: \mu_{AN} = \mu_{200} \\ H_1: \mu_{AN} \neq \mu_{200} \\ \alpha = 0{,}05 \end{cases}$$

Os comandos utilizados para testar estas hipóteses são:

ARQUIVO: 14 - FUNÇÃO TUMAAMOSTRA PARA TESTE T DA MÉDIA DE UMA AMOSTRA

```
>using DataFrames, CSV, Statistics, Distributions
>X=CSV.read("Z:/Alexandre/Julia/Exemplos/umaamostra.csv", DataFrame)
```

julia

```
julia> X=CSV.read("Z:/Alexandre/Julia/Exemplos/amostrasindependentes.csv", DataFrame)
11×3 DataFrame
| Row | n     | x1    | x2       |
|     | Int64 | Int64 | Int64?   |
|-----|-------|-------|----------|
| 1   | 1     | 201   | 158      |
| 2   | 2     | 204   | 164      |
| 3   | 3     | 232   | 172      |
| 4   | 4     | 214   | 170      |
| 5   | 5     | 219   | 165      |
| 6   | 6     | 242   | 160      |
| 7   | 7     | 233   | 178      |
| 8   | 8     | 225   | 160      |
| 9   | 9     | 219   | 174      |
| 10  | 10    | 218   | missing  |
| 11  | 11    | 213   | missing  |
```

```
>function TUMAAMOSTRA(xamostra,xpadrao)
    n= length(xamostra)
    xbar=sum(xamostra)/n
    s2=sum((xamostra.-xbar).^2)/(n-1)
    s=sqrt(s2)
    sx = s/sqrt(n)
    tcalculado = (xbar- xpadrao)/(s/sqrt(n))
    Alpha = 2*ccdf(TDist(n-1), abs(tcalculado))
    println("Média: ", xbar)
    println("Variância: ", s2)
    println("Desvio-padrão: ", s)
    println("Erro padrão: ", sx)
    println("tcalculado: ", tcalculado)
    println("Alpha: ", Alpha)
end

#TUMAAMOSTRA(xamostra,xpadrao)
>TUMAAMOSTRA(X.h,200)
Média: 197.77
Variância: 151.69
Desvio-padrão: 12.32
Erro padrão: 3.42
tcalculado: -0.65
Alpha: 0.53
```

O valor de $\alpha = 0{,}526$ foi maior que o nível de significância de 0,05, revelando que não há evidência suficiente para rejeitar a hipótese da nulidade. Logo, a média da amostra de 197,77 mm não é diferente do valor padrão de 200 mm, ao nível de 0,05 de significância. Portanto, as mudas produzidas enquadram-se no padrão especificado e podem ser enviadas para campo, quando considerada esta variável.

Ressalta-se que, quando o tamanho da amostra aumenta a distribuição de t se aproxima da distribuição normal, segundo o teorema do limite central, conforme pode ser visualizado na seção 3. Especialmente, quando o tamanho da amostra for maior que 30, as diferenças entre os valores da distribuição de t e os da distribuição Normal são muitos pequenas. Nestes casos, poderiam ser utilizados os valores críticos da distribuição normal padronizada (z), sendo a estatística do teste dada por:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Se σ (desvio-padrão da população) é desconhecido, pode-se estimá-lo utilizando o desvio padrão da amostra (s).

6.3 TESTES T PARA DUAS AMOSTRAS INDEPENDENTES

Um problema frequente encontrado por Engenheiros Florestais em seus experimentos é comparar médias de duas amostras (ou tratamentos). As amostras são definidas como independentes quando não existe nada que as relacione. Normalmente, os valores amostrais são obtidos em unidades amostrais distintas. Os métodos para testar se as duas médias amostrais não são diferentes estão apresentados na sequência.

6.3.1 TESTE T PARA DUAS AMOSTRAS INDEPENDENTES PEQUENAS E COM MESMA VARIÂNCIA

As condições para aplicação deste tipo de teste são:

- As amostras são independentes, ou seja, tomadas em amostras exclusivas
- Pelo menos uma amostra é considerada pequena ($n \leq 30$)
- As amostras são extraídas de populações normalmente distribuídas
- As variâncias das duas populações são homogêneas. Pode-se aplicar um teste de hipótese para avaliar essa condição (H_0: $\sigma_1^2 = \sigma_2^2$).
- As variâncias das duas populações são desconhecidas, mas podem ser estimadas por meio das duas amostras.

Quando essas condições são satisfeitas, a estatística do t para a média das duas amostras, que podem ser compostas pelo mesmo ou diferente número de observações, é dada por:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Em que:

$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$ é a média ponderada das variâncias pelo tamanho das duas amostras

$\bar{x}_1$ é a média da amostra da população 1
$\bar{x}_2$ é a média da amostra da população 2
n_1 é o tamanho da amostra da população 1
n_2 é o tamanho da amostra da população 2
$\mu_1 - \mu_2$ representa a hipótese testada e normalmente atribui-se um valor zero, indicando que não existem diferenças entre as médias

Para obtenção dos valores críticos da distribuição de *t*, os graus de liberdade são calculados por $GL = n_1 + n_2 - 2$.

Pode-se construir um intervalo de confiança como segue:

$$(\bar{x}_1 - \bar{x}_2) - t_{\frac{\alpha}{2}}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \leq (\mu_1 - \mu_2) \leq (\bar{x}_1 - \bar{x}_2) + t_{\frac{\alpha}{2}}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

EXERCÍCIO 8

Um Engenheiro Florestal deseja saber se a produção de madeira ($m^3.ha^{-1}$) de duas espécies, eucalipto e acácia-negra, é a mesma 7 anos após o plantio (Figuras 6.7 – 6.8). Nos povoamentos de eucalipto e acácia-negra foram instaladas parcelas de 400 m². Em cada uma das parcelas todas as árvores foram derrubadas e cubadas para obtenção do volume por árvore, em m³. A partir desses dados, foi obtido o estoque de volume de madeira por parcela e por hectare. Os dados estão apresentados na Tabela 6.2.

FIGURA 6.7. Parcela instalada no povoamento de eucalipto.

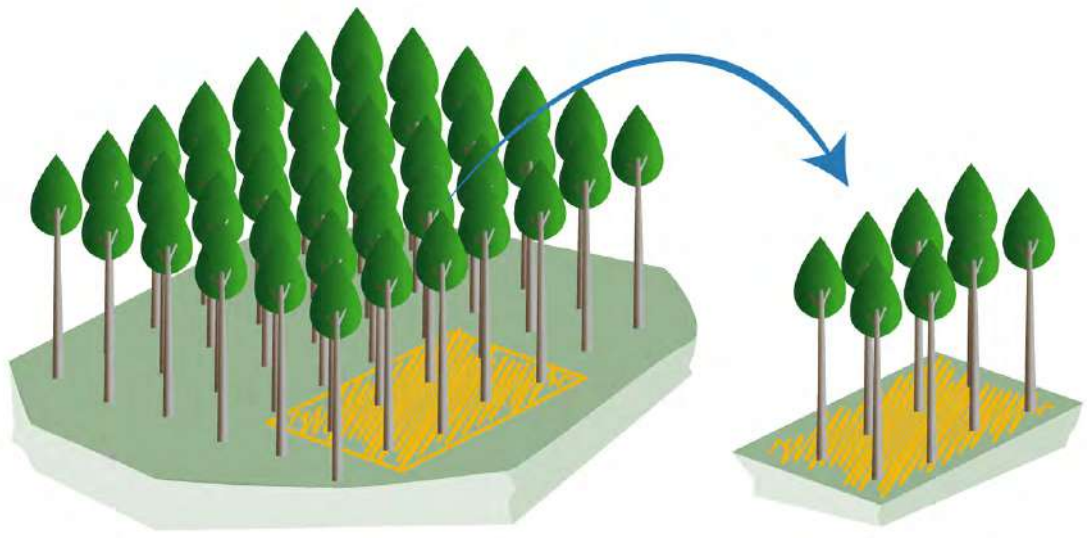


FIGURA 6.8. Parcela instalada no povoamento de acácia-negra.

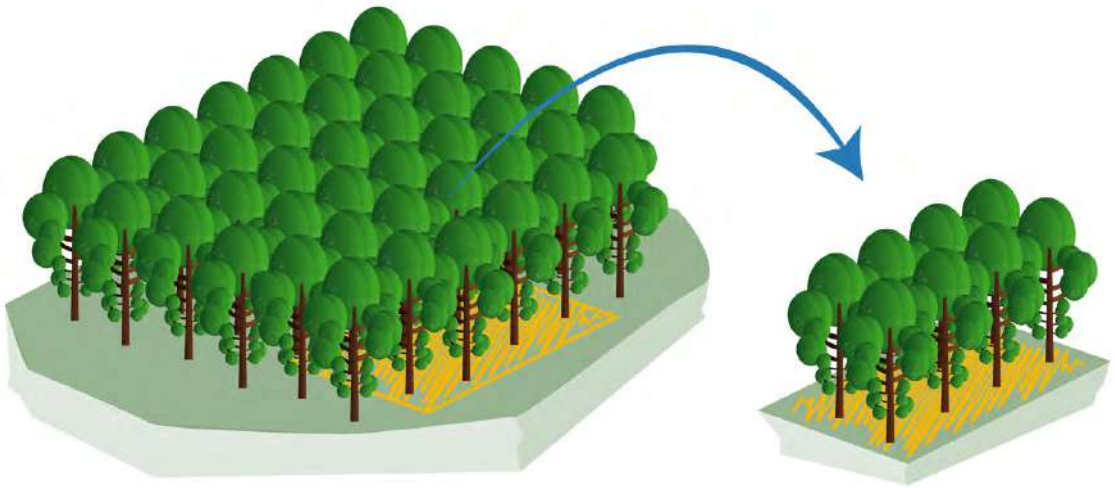


TABELA 6.2. Dados experimentais de volume de madeira ($m^3.ha^{-1}$) de eucalipto (X1) e acácia-negra (X2).

n	X1	X2
1	201	158
2	204	164
3	232	172
4	214	170
5	219	165
6	142	160
7	233	178
8	225	160
9	219	174
10	218	
11	213	

X1 é o estoque de madeira ($m^3.ha^{-1}$) para a espécie eucalipto e X2 é o estoque de madeira ($m^3.ha^{-1}$) para a espécie acácia-negra.

Foram estabelecidas as seguintes hipóteses:

$$\begin{cases} H_0: \mu_{X1} = \mu_{X2} \\ H_1: \mu_{X1} \neq \mu_{X2} \end{cases}$$
$$\alpha = 0{,}05$$

Os comandos utilizados para testar as médias de amostras independentes são:

Arquivo: 15 - Função amostrasindependentes para teste t de amostras independentes

```
#Teste t para duas amostras independentes pequenas e com mesma variância

>using DataFrames, CSV, Statistics, Distributions
>X=CSV.read("Z:/Alexandre/Julia/Exemplos/amostrasindependentes.csv", DataFrame)
```



```
julia> X=CSV.read("Z:/Alexandre/Julia/Exemplos/amostraspareadas.csv", DataFrame)
15×3 DataFrame
```

Row	n	x1	x2
	Int64	Float64	Float64
1	1	20.6	19.8
2	2	22.4	22.8
3	3	25.6	25.6
4	4	27.8	26.4
5	5	23.4	22.9
6	6	25.6	26.5
7	7	23.2	22.9
8	8	24.5	25.6
9	9	27.6	26.1
10	10	25.4	25.6
11	11	26.4	26.0
12	12	24.3	24.3
13	13	19.2	20.2
14	14	20.3	21.1
15	15	22.4	22.3

```
>#
```

Ressalta-se que o tamanho das amostras dos dois grupos (X1 e X2) pode ser diferente. Para salvar um conjunto de dados com duas variáveis com tamanhos diferentes, pode-se optar em deixar as células vazias para a variável com menor tamanho (nos valores faltantes), no arquivo de dados em .CSV. Logicamente, os valores faltantes (missing) não serão considerados no processamento da análise. X1 são os valores do grupo (tratamento) 1. X2 são os valores do grupo (tratamento) 2.

```
>function AMOSTRASINDEPENDENTES(X1,X2)
    dfX1 = DataFrame([X1])
    X1=dropmissing(dfX1)
    X1= X1.x1
    dfX2 = DataFrame([X2])
    X2=dropmissing(dfX2)
    X2= X2.x1
    n1= length(X1)
    xbar1=sum(X1)/n1
    s²1=sum((X1.-xbar1).^2)/(n1-1)
    s1=sqrt(s²1)
    n2= length(X2)
    xbar2=sum(X2)/n2
    s²2=sum((X2.-xbar2).^2)/(n2-1)
    s2=sqrt(s²2)
    s²p=(s²1*(n1-1)+s²2*(n2-1))/(n1+n2-2)
    tcalculado = ( xbar1-xbar2) /  sqrt(s²p*(1/n1+1/n2))
    Alpha = 2*ccdf(TDist(n1+n2 -2), abs(tcalculado))
    println("Média de x1: ", xbar1)
    println("Variância de x1: ", s²1)
    println("Desvio-padrão de x1: ", s1)
```

```
    println("Média de x2: ", xbar2)
    println("Variância de x2: ", s²2)
    println("Desvio-padrão de x2: ", s2)
    println("Variância ponderada: ", s²p)
    println("tcalculado: ", tcalculado)
    println("Alpha: ", Alpha)
end

#AMOSTRASINDEPENDENTES(X1,X2)
>AMOSTRASINDEPENDENTES(X.x1,X.x2)
Média de x1: 220.00
Variância de x1: 153.00
Desvio-padrão de x1: 12.37
Média de x2: 166.78
Variância de x2: 49.44
Desvio-padrão de x2: 7.03
Variância ponderada: 106.98
tcalculado: 11.45
Alpha: 1.07e-9
```

A média da amostra do eucalipto foi de 220 $m^3.ha^{-1}$ e da acácia-negra de 166,78 $m^3.ha^{-1}$, o valor encontrado de $\alpha = 1{,}07\ e^{-9}$ se aproxima de zero, portanto, existe evidência para rejeitar a hipótese da nulidade. A floresta de eucalipto foi mais produtiva que a floresta de acácia-negra para a presente condição, dado que as médias diferem estatisticamente ao nível de 0,01 de significância.

Destaca-se que é raro as variâncias serem conhecidas nos experimentos conduzidos por Engenheiros Florestais, mas, se eventualmente o são, neste caso a estatística do teste pode ser calculada como:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Note que a estatística do teste (z) refere-se à distribuição normal padronizada, logo os valores críticos devem ser obtidos desta distribuição (distribuição normal padrão).

Pode-se construir um intervalo de confiança como segue:

$$(\bar{x}_1 - \bar{x}_2) - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq (\mu_1 - \mu_2) \leq (\bar{x}_1 - \bar{x}_2) + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

6.3.2 TESTE T PARA DUAS AMOSTRAS INDEPENDENTES PEQUENAS E COM DIFERENTES VARIÂNCIAS

As condições para aplicação desse teste são:

- As amostras são independentes.
- Pelo menos uma amostra é considerada pequena ($n \leq 30$).
- As amostras são extraídas de populações normalmente distribuídas.
- As variâncias das duas populações não são homogêneas, ou seja, existe evidências para rejeitar a hipótese de que H_0: $\sigma_1^2 = \sigma_2^2$.
- As variâncias das duas populações são desconhecidas, mas podem ser estimadas por meio das duas amostras.

Note, que no presente caso, o teste é aplicado quando não existem evidências para assumir a igualdade das variâncias das duas populações. Logo, o teste t para comparar se as diferenças entre as médias é igual a zero (não existem diferenças) é dado por:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Em que:
$\bar{x}_1$ é a média da amostra da população 1
$\bar{x}_2$ é a média da amostra da população 2
n_1 é o tamanho da amostra da população 1
n_2 é o tamanho da amostra da população 2
μ_1 - μ_2 representa a hipótese testada e normalmente atribui-se um valor zero, indicando que não existem diferenças entre as médias

Para obtenção dos valores críticos da distribuição de *t*, utiliza-se o menor grau de liberdade, ou seja, o menor entre $GL_1 = n_1 - 1$ e $GL_2 = n_2 - 1$.

Pode-se construir um intervalo de confiança como segue:

$$(\bar{x}_1 - \bar{x}_2) - t_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq (\mu_1 - \mu_2) \leq (\bar{x}_1 - \bar{x}_2) + t_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Com grau de liberdade sendo o menor entre $GL_1 = n_1 - 1$ e $GL_2 = n_2 - 1$.

6.3.3 TESTE T PARA DUAS AMOSTRAS INDEPENDENTES GRANDES

Amostras são consideradas grandes quando o tamanho para ambos os grupos supera 30 observações, ou seja, $n > 30$. Nesta condição utiliza-se a distribuição Normal padrão ao invés da distribuição de *t*.

As condições para aplicação desse teste são:
- As amostras são independentes.
- As duas amostras contém mais do que 30 observações, ou seja, $n_1 > 30$ e $n_2 > 30$.
- As amostras são extraídas de populações normalmente distribuídas.

Já foi discutido que as médias amostrais, geralmente, seguem a distribuição normal. As diferenças entre médias amostrais $\bar{x}_1 - \bar{x}_2$ também se distribuem dessa forma. Logo, a estatística do teste é dada por:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Em que:

$\bar{x}_1$ é a média da amostra da população 1

$\bar{x}_2$ é a média da amostra da população 2

n_1 é o tamanho da amostra da população 1

n_2 é o tamanho da amostra da população 2

$\mu_1 - \mu_2$ representa a hipótese testada e normalmente atribui-se o valor zero, indicando que não existem diferenças entre as médias

σ_1^2 é a variância da população 1

σ_2^2 é a variância da população 2

Quando as variâncias da população são desconhecidas, pode-se estimá-las por meio das variâncias amostrais s_1^2 e s_2^2.

Os valores críticos para verificar se a estatística do teste (z) está na área de aceitação ou rejeição da hipótese da nulidade são obtidos na distribuição normal padronizada.

Pode-se construir um intervalo de confiança como segue:

$$(\bar{x}_1 - \bar{x}_2) - z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq (\mu_1 - \mu_2) \leq (\bar{x}_1 - \bar{x}_2) + z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Note que no intervalo de confiança as variâncias das duas amostras são consideradas homogêneas, podendo-se, então, obter a média delas.

6.4 TESTE T PARA DUAS AMOSTRAS PAREADAS

Outra aplicação interessante nos experimentos conduzidos por Engenheiros Florestais é quando se tem duas amostras pareadas. Isto quer dizer que os valores das amostras são obtidos na mesma unidade experimental. Portanto, existe relacionamento entre as duas amostras, pois, os valores da amostra um e da amostra dois são obtidos nas mesmas unidades experimentais. Assim, cada elemento da amostra II é tomado com o seu próprio controle (amostra I), o que é denominado de pré- e pós-tratamento, o que caracteriza as chamadas amostras pareadas.

Assim, obtém-se a seguinte configuração em relação à obtenção dos valores nas amostras:

- Dois valores para cada elemento da amostra, pois são efetuadas leituras em uma unidade experimental antes e após a aplicação de algum tratamento. Por exemplo, teor de clorofila nas mudas antes e após adubação nitrogenada, neste caso, as folhas e as mudas serão as mesmas.

- Quando são realizadas duas leituras na mesma amostra, mas com elementos distintos. Por exemplo, análise do teor de matéria orgânica de uma amostra por dois laboratoristas, nesse caso, a pequena porção utilizada por cada um dos laboratoristas não será a mesma, mas será da mesma amostra e terá a mesma característica.

Como resultado disso, o tamanho das duas amostras é sempre igual para o caso, pois são tomadas aos pares, sendo então testada a diferença entre as médias das duas amostras por meio do teste *t*. O teste é utilizado para evidenciar se a média das diferenças é zero, sendo a estatística do teste formalizada como segue:

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

Em que:
$\bar{d}$ é a média das diferenças para os dados pareados, $\bar{d} = \frac{\sum d}{n}$.
d é a diferença entre dois dados pareados (antes – depois ou depois - antes)
s_d é o desvio-padrão das diferenças
n é o tamanho da amostra, ou seja, o número de valores pareados
μ_d representa a hipótese testada e normalmente atribui-se um valor zero, indicando que não existem diferenças entre as médias antes e depois.

Para obtenção dos valores críticos, utiliza-se o grau de liberdade definido por $GL = n - 1$.

EXERCÍCIO 9

Em uma floresta de Pinus foram mensuradas a altura de 15 árvores utilizando dois métodos: medição com hipsômetro (testemunha) e medição com LiDAR (Figura 6.9). Os dados obtidos estão apresentados na Tabela 6.3. Assim, o Engenheiro Florestal deseja avaliar se os valores das alturas mensuradas por esses dois métodos não diferem entre si. Se não há diferença, espera-se que a média dos valores das diferenças não seja diferente de zero ($\mu_d = 0$).

TABELA 6.3. Dados experimentais de altura (m) mensurada com hipsômetro (X_1) e por LiDAR (X_2) em uma floresta de Pinus.

n	X_1	X_2
1	20,6	19,8
2	22,4	22,8
3	25,6	25,6
4	27,8	26,4
5	23,4	22,9
6	25,6	26,5
7	23,2	22,9
8	24,5	25,6
9	27,6	26,1
10	25,4	25,6
11	26,4	26,0
12	24,3	24,3
13	19,2	20,2
14	20,3	21,1
15	22,4	22,3

FIGURA 6.9. Medição de altura total com hipsômetro e LiDAR.



Foram estabelecidas as seguintes hipóteses:

$$\begin{cases} H_0: \mu_d = 0 \\ H_1: \mu_d \neq 0 \end{cases}$$
$$\alpha = 0{,}05$$

Os comandos utilizados para o teste *t* pareado são:

ARQUIVO: 16 - FUNÇÃO AMOSTRASPAREADAS PARA TESTE T DE AMOSTRAS PAREADAS

```
>using DataFrames, CSV, Statistics, Distributions

>X=CSV.read("Z:/Alexandre/Julia/Exemplos/amostraspareadas.csv", DataFrame)
```

```
julia> X=CSV.read("Z:/Alexandre/Julia/Exemplos/amostraspareadas.csv", DataFrame)
15×3 DataFrame
```

Row	n	x1	x2
	Int64	Float64	Float64
1	1	20.6	19.8
2	2	22.4	22.8
3	3	25.6	25.6
4	4	27.8	26.4
5	5	23.4	22.9
6	6	25.6	26.5
7	7	23.2	22.9
8	8	24.5	25.6
9	9	27.6	26.1
10	10	25.4	25.6
11	11	26.4	26.0
12	12	24.3	24.3
13	13	19.2	20.2
14	14	20.3	21.1
15	15	22.4	22.3

```
>function AMOSTRASPAREADAS(x1,x2)
    n= length(x1)
    x1bar=mean(x1)
    x2bar=mean(x2)
    d=x1-x2
    xbard=sum(d)/n
    s²=sum((d.-xbard).^2)/(n-1)
    s=sqrt(s²)
    tcalculado = xbard/(s/sqrt(n))
    Alpha = 2*ccdf(TDist(n-1), abs(tcalculado))
    println("Média de x1: ", x1bar)
    println("Média de x2: ", x2bar)
    println("Média da diferença: ", xbard)
    println("Variância da diferença: ", s²)
    println("Desvio-padrão da difrença: ", s)
    println("tcalculado: ", tcalculado)
    println("Alpha: ", Alpha)
end
#AMOSTRASPAREADAS(x1,x2)
>AMOSTRASPAREADAS(X.x1,X.x2)
Média de x1: 23.91
Média de x2: 23.87
Média da diferença: 0.040
Variância da diferença: 0.66
Desvio-padrão da difrença: 0.81
tcalculado: 0.19
Alpha: 0.85
```

A média da altura de Pinus mensurada com hipsômetro ($\overline{x}_1$) foi de 23,91 m e por LiDAR ($\overline{x}_2$) de 23,87 m. A média da diferença da altura por meio dos dois métodos foi de 0,039 m. O valor de $\alpha = 0,85$ foi maior do que o nível de significância de 0,05, revelando que não existe evidência para rejeitar a hipótese da nulidade, ou seja, o teste estatístico indicou não haver diferenças entre as médias dos métodos de medição das alturas, ao nível de 5% de significância. Portanto, a obtenção dos valores médios das alturas com tecnologia LiDAR pode ser substituída pelo método empregado pela empresa, considerando somente a obtenção dos valores de altura por hipsômetro.

Adicionalmente ao teste, pode-se construir um intervalo de confiança para a diferença da média populacional (μ_d). O intervalo é dado por:

$$\left[\bar{d} - t_{\frac{\alpha}{2}}\frac{s_d}{\sqrt{n}} \leq \mu_d \leq \bar{d} - t_{\frac{\alpha}{2}}\frac{s_d}{\sqrt{n}}\right] = 1 - \alpha$$

Ressalta-se que o número de graus de liberdade é $GL = n - 1$.

Para esse exemplo foram utilizados os comandos:

Arquivo: 16 - Função amostraspareadas para teste t de amostras pareadas

```
#Função ICPAREADO para intervalo de confiança para a diferença da média
populacional

>function ICPAREADO(x1,x2)
    alpha=0.05
    n= length(x1)
    x1bar=mean(x1)
    x2bar=mean(x2)
    d=x1-x2
    xbard=sum(d)/n
    s²=sum((d.-xbard).^2)/(n-1)
    s=sqrt(s²)
    sx=s/sqrt(n)
    tVal = quantile(TDist(n-1),1-alpha/2)
    ICInferior = xbard- tVal * sx
    ICSuperior = xbard+ tVal * sx
    Confiança=(1-alpha)*100
    println("Média da diferença: ", xbard)
    println("Limite de confiança inferior: ", ICInferior)
    println("Limite de confiança superior: ", ICSuperior)
    println("Confiança (%): ", Confiança)
end

#ICPAREADO(x1,x2)
>ICPAREADO(X.x1,X.x2)
Média da diferença: 0.0399
Limite de confiança inferior: -0.41
Limite de confiança superior: 0.49
Confiança (%): 95.0
```

O intervalo de confiança obtido foi de -0,41 $m \leq \mu_d \leq$ 0,49 m = 0,95. Note que no intervalo de confiança está incluso o valor de zero, o que indica que a média da diferença de 0,0399 m não é considerada diferente de zero. Quando o intervalo de confiança não contiver o valor zero, então a diferença da média é considerada estatisticamente diferente de zero, o que evidencia condições para rejeitar a hipótese da nulidade.

Ressalta-se que para pequenas amostras é utilizado o teste t, enquanto que para amostras grandes ($n > 30$) pode-se utilizar os valores críticos da tabela z (distribuição normal padronizada).

7 TESTE SOBRE UMA PROPORÇAO

Os testes *ts*, evidenciados na seção anterior, referiam-se às afirmações sobre médias populacionais. Em muitas situações práticas, o Engenheiro Florestal está interessado em uma resposta expressa por uma proporção, percentagem ou probabilidade de um determinado evento.

Como exemplos, pode-se considerar que um viveirista comprou sementes para as quais havia garantia de germinação de 95%. A estatística do teste sobre a afirmação de uma proporção é obtida por:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p\,q}{n}}}$$

Em que:
$\hat{p} = \frac{\bar{x}}{n}$, ou seja, é a proporção amostral
p é a proporção populacional, declarada na hipótese da nulidade
$q = 1 - p$
n é o número de provas

Destaca-se que para aplicação desta expressão deve-se assegurar as condições $np \geq 5$ e $nq \geq 5$, para que a distribuição binomial das proporções amostrais se aproximem de uma distribuição normal com média $\mu = np$ e $\sigma = \sqrt{npq}$.

Assim, a estatística do teste pode ser comparada aos valores críticos da distribuição normal padronizada.

EXERCÍCIO 10

Considere o caso de germinação de sementes da espécie louro-pardo (Figura 7.1). Um Engenheiro Florestal, responsável pela produção e vendas de sementes de espécies nativas, garante que a taxa de sementes não germinadas é inferior a 10% para a espécie louro-pardo. Um viveiro adquiriu 3 quilogramas das sementes e tomou uma amostra de 1.000 sementes para avaliar a porcentagem de germinação. A amostra evidenciou que 879 germinaram. Teste ao nível de 0,01 de significância que a porcentagem de sementes não germinadas não difere estatisticamente da garantia fornecida pelo Engenheiro Florestal, ou seja, é inferior a 10%.

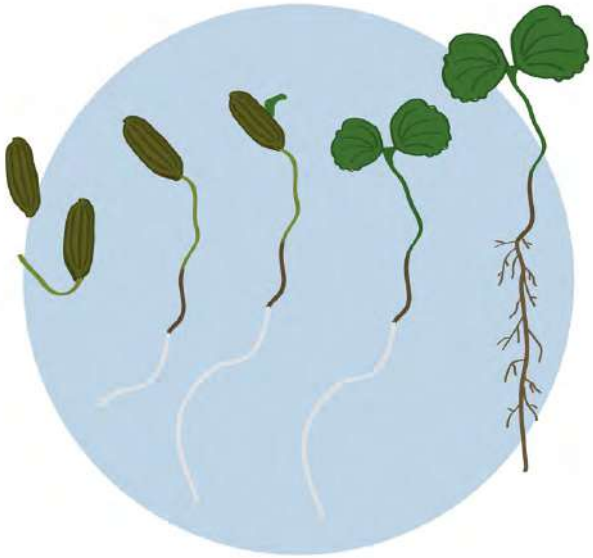
Informações:

- Observa-se que a garantia é que a taxa para não germinação das sementes é inferior a 10%, logo $p < 0{,}1$. Assim, a taxa de sementes que germinarão é $p \geq 0{,}1$.
- Um teste unilateral à esquerda foi estabelecido com as seguintes hipóteses testadas:

$$\begin{cases} H_0: p \geq 0{,}1 \\ H_1: p < 0{,}1 \end{cases}$$
$$\alpha = 0{,}01$$

O valor de $\hat{p} = \frac{\bar{x}}{n} = 0{,}121$

FIGURA 7.1. Processo de germinação da semente de louro-pardo.



Os comandos utilizados para testar proporções estão apresentados abaixo. Note que as exigências para o teste foram satisfeitas.

ARQUIVO: 17 - FUNÇÃO PROPORÇÃO PARA TESTE SOBRE UMA PROPORÇÃO

```
>phat=(1000-879)/1000
0.121
>n=1000;
>p=0.01;
>q=1-p
0.99
>n*p
10
>n*q
990

using Distributions

function proporcao(n,x,p)
    Alpha=0.01
    phat=x/n
    q=1-p
    z=(phat-p)/sqrt((p*q)/n)
    zVal = quantile(Normal(),Alpha)
    Alpha =(ccdf(Normal(),abs(z)))
    println("zcalculado: ", z)
    println("z Crítico: ", zVal)
    println("Alpha: ", Alpha)
end

#proporcao(n,x,p)
>proporcao(1000,(1000-879),0.1)
zcalculado: 2.22
z Crítico: -2.33
Alpha: 0.013
```

Conforme o resultado obtido, a hipótese da nulidade não é rejeitada, uma vez que $\alpha = 0{,}013$ é maior que o nível de significância de 0,01. Logo, não há evidência suficiente de que a taxa de sementes germinadas (12,1%) difere daquela garantida pelo Engenheiro Florestal (10%), ao nível de 0,01 de significância.

Ressalta-se que, quando o teste for bilateral, a estatística Alpha pode ser obtida por:

```
>Alpha =(ccdf(Normal(),abs(z))*2)
```

8 TESTE DE QUI-QUADRADO

O teste de Qui-quadrado (χ^2) permite analisar dados categóricos, ou seja, aqueles provenientes de variáveis nominais e ordinais, mais especificamente em dados de frequências de diferentes categorias.

É um teste apropriado quando os elementos das amostras são divididos em categorias e tem objetivo de comparar as frequências observadas e esperadas. A estatística Qui-quadrado é utilizada para evidenciar se as frequências observadas estão próximas das esperadas. Quando isto acontece a hipótese da nulidade é aceita, caso contrário ela é rejeitada.

Os dados de frequência (contagem) são organizados em uma tabela, denominada tabela de contingência.

A distribuição de χ^2 tem as seguintes características:

- Não é simétrica.
- Os valores são ≥ 0, ou seja, nunca são negativos.
- Assim como na distribuição de t e F, cada grau de liberdade resulta em uma diferente distribuição de χ^2.

A estatística do teste de Qui-quadrado χ^2 é dada por:

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

Em que:
f_o é a frequência observada
f_e é a frequência esperada

O teste de χ^2 mede a discrepância entre as frequências observadas e esperadas e revela se as diferenças entre os valores observados e os valores teoricamente esperados são estatisticamente significantes. As hipóteses testadas são:

$$\begin{cases} H_0: F_o = F_e \\ H_1: F_o \neq F_e \end{cases}$$
$$\alpha = 0{,}05$$

8.1 TESTE DE QUI-QUADRADO PARA AJUSTE

O teste de χ^2 pode ser utilizado para verificar o ajuste ou aderência. Isto implica em testar a hipótese da nulidade de que a distribuição de frequências observadas se ajusta (tem aderência) a uma determinada distribuição teórica.

O teste de Qui-quadrado para ajuste é aplicado quando as proporções são definidas a priori, ou seja, as frequências esperadas são conhecidas. Se valores pequenos da estatística do teste χ^2 forem obtidos, então, existe uma pequena diferença entre as frequências observadas e esperadas e aceita-se H_0; se os valores de χ^2 obtidos forem grandes, existe discrepância entre as frequências, indicando que H_0 é rejeitada, ou seja, não existe aderência.

Alguns cuidados devem ser tomados para aplicação do teste:

- Qualquer valor na tabela de contingência para a frequência esperada não deve ser menor do que 5. Não existe nenhuma suposição em relação ao número mínimo para frequência observada.
- Os valores de frequência devem ser amostrados aleatoriamente.
- Não se pode ter frequências negativas e valores maiores do que o total da frequência observada para todas as categorias.
- O teste de ajuste é efetuado por meio de um teste unilateral à direita.

As hipóteses testadas são:

1. **H_0:** a distribuição de frequências observadas não difere da distribuição teórica de frequência (esperada).
2. **H_1:** a distribuição de frequências observadas difere da distribuição teórica de frequência (esperada).

A estatística do teste de χ^2 é aplicada para testar a hipótese da nulidade (H_0) sobre frequências observadas e esperadas. O valor calculado para χ^2 pode ser comparado com um valor tabelado da distribuição de Qui-quadrado com graus de liberdade $GL = (h - 1)\ (k - 1)$, em que h refere-se ao número de linhas e k o número de colunas na tabela de contingência.

Uma importante aplicação do teste de χ^2 é efetuada para avaliar se os dados de uma amostra, oriundos de uma população, são normalmente distribuídos. Neste caso, os valores esperados são obtidos por meio do modelo da função densidade Normal.

EXERCÍCIO 11

Um Engenheiro Florestal efetuou uma pesquisa de recuperação de área degradada. Para tanto, em uma área de 50 ha foram plantadas mudas de 10 espécies florestais e, 5 anos após o plantio, foi realizado um inventário da área (Figura 8.1). O Engenheiro Florestal tem interesse em avaliar se o número de indivíduos para as espécies plantadas é o mesmo 5 anos após o plantio. Foram efetuadas duas amostragens independentes, cada qual com área amostral de 0,5 ha. Os dados estão apresentados na Tabela 8.1, sendo considerada a frequência esperada aquela obtida no momento do plantio e a observada 5 anos após o plantio.

FIGURA 8.1. Espécies florestais para recuperação de área degradada.



1. Pinheiro-do-Paraná
2. Louro-pardo
3. Farinha-seca
4. Angico-vermelho
5. Grápia
6. Alecrim
7. Ariticum
8. Laranjeira-do-mato
9. Aroeira-vermelha
10. Açoita-cavalo

TABELA 8.1. Número de indivíduos por espécie observados em uma área de 5.000 m² no estado do Paraná.

Espécie	FO	FE
Pinheiro-do-Paraná	36	26
Louro-pardo	18	16
Farinha-seca	60	55
Angico-vermelho	15	14
Grápia	20	17
Alecrim	16	14
Araticum	19	19
Laranjeira-do-mato	26	20
Aroeira-vermelha	19	19
Açoita-cavalo	16	15

Em que: FE é o número de indivíduos no momento do plantio e FO o número de indivíduos 5 anos após o plantio.

Foi estabelecido o seguinte teste de hipótese:

$$\begin{cases} H_0: FE = FO \\ H_1: FE \neq FO \end{cases}$$
$$\alpha = 0{,}05$$

Os comandos para realização do teste são:

ARQUIVO: 18 - FUNÇÃO QUIAJUSTE PARA TESTE DE QUI-QUADRADO

```
>using DataFrames, CSV, Statistics, Distributions

>X=CSV.read("Z:/Alexandre/Julia/Exemplos/quiajuste.csv", DataFrame)
```

julia

```
julia> X=CSV.read("Z:/Alexandre/Julia/Exemplos/quiajuste.csv", DataFrame)
10×3 DataFrame
| Row | Especie              | Plantio | CincoAnos |
|     | String               | Int64   | Int64     |
|-----|----------------------|---------|-----------|
| 1   | Pinheiro-do-Parana   | 36      | 26        |
| 2   | Louro-pardo          | 18      | 16        |
| 3   | Farinha-seca         | 60      | 55        |
| 4   | Angico-vermelho      | 15      | 14        |
| 5   | Grapia               | 20      | 17        |
| 6   | Alecrim              | 16      | 14        |
| 7   | Ariticum             | 19      | 19        |
| 8   | Laranjeira-do-mato   | 26      | 20        |
| 9   | Aroeira-vermelha     | 19      | 19        |
| 10  | Acoita-cavalo        | 16      | 15        |
```

```
>function QUIAJUSTE(fo,fe)
    Quiquadradocalculado = sum((fe-fo).^2 ./fe)
    Gl = (length(fo)-1)
    Alpha = ccdf(Chisq(Gl), Quiquadradocalculado)
    println("Qui-quadrado: ", Quiquadradocalculado)
    println("Grau de liberdade: ", Gl)
    println("Alpha: ", Alpha)
end

#QUIAJUSTE(fo,fe)
>QUIAJUSTE(X.CincoAnos,X.Plantio)
Qui-quadrado: 5.63
Grau de liberdade: 9
Alpha: 0.78
```

O teste de Qui-quadrado resultou em uma estatística $\chi^2 = 5{,}63$ com um valor correspondente de $\alpha = 0{,}78$. Como este valor é maior que o valor de $\alpha = 0{,}05$, o resultado do teste atesta que H_0 é verdadeira, portanto, indica não haver diferenças na avaliação das frequências das espécies no momento do plantio e 5 anos após, ao nível de 0,05 de significância. Desta forma, as espécies escolhidas para a recuperação da área demonstraram-se adaptadas ao local concernente à sua sobrevivência.

8.2 TESTE DE QUI-QUADRADO PARA ASSOCIAÇÃO

O teste de Qui-quadrado para associação é aplicado quando se tem somente as proporções observadas. Neste caso o que está sendo testado é se as variáveis são associadas.

O teste de χ^2 para associação ou de independência para uma tabela de contingência de dupla entrada é efetuado de tal modo que variáveis categóricas caracterizam as colunas e as linhas. O teste de independência verifica se as duas variáveis categóricas (uma na linha e outra na coluna) não estão associadas. Isto implica em assumir que as frequências esperadas em todas as células da tabela de contingência sejam iguais, ou proporcionais aos valores observados.

Alguns cuidados devem ser tomados para aplicação do teste:

- Qualquer valor na tabela de contingência para a frequência esperada não deve ser menor do que 5.
- Os valores de frequência observada devem ser tomados aleatoriamente.
- Não se pode ter frequências negativas e valores maiores do que o total de frequência observada para uma determinada linha ou coluna.
- O teste de independência é efetuado por meio de um teste unilateral à direita.

Na tabela de contingência somente os valores de frequência observada são conhecidos, requerendo que os valores de frequência esperada sejam calculados a partir dos dados obtidos. A forma de cálculo das frequências esperadas pode ser obtida por:

$$f_e = \frac{\sum_{i=1}^{n} l \sum_{i=1}^{n} c}{\sum}$$

Em que:
$\sum_{i=1}^{n} l$ é o total da linha
$\sum_{i=1}^{n} c$ é o total da coluna
$\sum$ é o total geral $(\sum_{i=1}^{n} l + \sum_{i=1}^{n} c)$

As hipóteses testadas são:

1. **H_0:** As variáveis de linha e coluna são independentes (não associadas).
2. **H_1:** As variáveis de linha e coluna são dependentes (associadas).

A estatística do teste de χ^2 é aplicada para testar a hipótese da nulidade (H_0). Valores pequenos de χ^2 indicam pequena diferença entre os valores observados e esperados, enquanto grandes valores (à direita da distribuição) revelam diferença significante entre as frequências observadas e esperadas. A estatística do teste de χ^2 pode ser comparada com um valor da distribuição de Qui-quadrado com graus de liberdade $GL = (h - 1)\ (k - 1)$, em que h refere-se ao número de linhas e k o número de colunas na tabela de contingência.

EXERCÍCIO 12

Um Engenheiro Florestal efetuou uma pesquisa para avaliar se o número de acidentes em um viveiro florestal está associado ao turno de trabalho e idade dos colaboradores. A variável número de acidentes foi avaliada em 100 viveiros florestais durante um período de 10 anos. Os dados obtidos estão apresentados na Tabela 8.2.

TABELA 8.2. Número de acidentes observado em 100 viveiros florestais durante um período de 10 anos em função do turno de trabalho e idade dos colaboradores.

Turno de trabalho	I_1	I_2	I_3	I_4
Manhã	19	12	8	9
Tarde	25	16	11	8

Em que I_1 = 18 -| 25, I_2 = 25 -| 35, I_3 = 35 -| 45, $I_4 \geq 45$ anos.

Foram estabelecidas as seguintes hipóteses:

$$\begin{cases} H_0: \textit{acidentes em viveiros florestais não estão associados ao turno de trabalho e idade dos colaboradores} \\ H_1: \textit{acidentes em viveiros florestais estão associados ao turno de trabalho e idade dos colaboradores} \end{cases}$$

$$\alpha = 0{,}05$$

Os comandos usados no teste de Qui-quadrado para associação são dados por:

ARQUIVO: 19 - FUNÇÃO QUIASSOCIACAO PARA TESTE DE QUI-QUADRADO PARA ASSOCIAÇÃO

```
>using DataFrames, CSV, Statistics, Distributions

>X=CSV.read("Z:/Alexandre/Julia/Exemplos/quiassociacao.csv", DataFrame)
```

julia

```
julia> X=CSV.read("Z/Alexandre/Julia/Exemplos/quiassociacao.csv", DataFrame)
2×5 DataFrame
 Row │ Turno    I1     I2     I3     I4
     │ String7  Int64  Int64  Int64  Int64
─────┼──────────────────────────────────────
   1 │ Manha      19     12      8      9
   2 │ Tarde      25     16     11      8
```

```
>function QUIASSOCIACAO(X)
    FO=X[:,2:end]
    L=size(FO,1)
    C=size(FO,2)
    TotalLinha = [sum(FO[i,:]) for i in 1:L]
    TotalColuna = [sum(FO[:,i]) for i in 1:C]
    n = sum(TotalColuna)
    PropLinha = TotalLinha/n
    PropColuna = TotalColuna/n
```

```
    FE = [PropColuna[c]*PropLinha[l]*n for l in 1:L, c in 1:C]
    Quiquadradocalculado = sum([(FO[l,c]-
FE[l,c])^2 / FE[l,c] for l in 1:L, c in 1:C])
    Gl = (L-1)*(C-1)
    Alpha = ccdf(Chisq(Gl), Quiquadradocalculado)
    println("Frequência Esperada: ", FE)
    println("Qui-quadrado: ", Quiquadradocalculado)
    println("Grau de liberdade: ", Gl)
    println("Alpha: ", Alpha)
end
#QUIASSOCIACAO(X)
>QUIASSOCIACAO(X)
Frequência Esperada: [19.56 12.44 8.44 7.56; 24.44 15.56 10.56 9.44]
Qui-quadrado: 0.60
Grau de liberdade: 3
Alpha: 0.90
```

O teste de Qui-quadrado resultou em um valor de $\chi^2 = 0{,}60$ e um valor de $\alpha = 0{,}90$. Como este valor é maior que o valor de $\alpha = 0{,}05$, o valor encontrado revela que H_0 é verdadeira, portanto, indica não haver associação entre a idade e o turno de trabalho para o número de acidentes em viveiros florestais, ao nível de 0,05 de significância. Desta forma, o programa de treinamento nos viveiros florestais para prevenção de acidentes pode ser direcionado a todos os funcionários, sem a necessidade de considerar grupos de idade e turno de trabalho.

8.3 TESTE PARA A GRANDEZA DO DESVIO-PADRÃO

O procedimento para testar a grandeza do desvio-padrão de uma população (σ) pode ser útil para o controle de qualidade em várias operações florestais. Se a variável apresentar distribuição normal, então um teste de qui-quadrado pode ser aplicado, utilizando-se a seguinte estatística:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

Em que:

n = tamanho da amostra

s^2 = variância da amostra

σ^2 = variância da população, que configura a hipótese da nulidade, representada por um valor fixo para este parâmetro.

A estatística tem como graus de liberdade $GL = n - 1$.

Ressalta-se que o teste de Qui-quadrado pode ser:

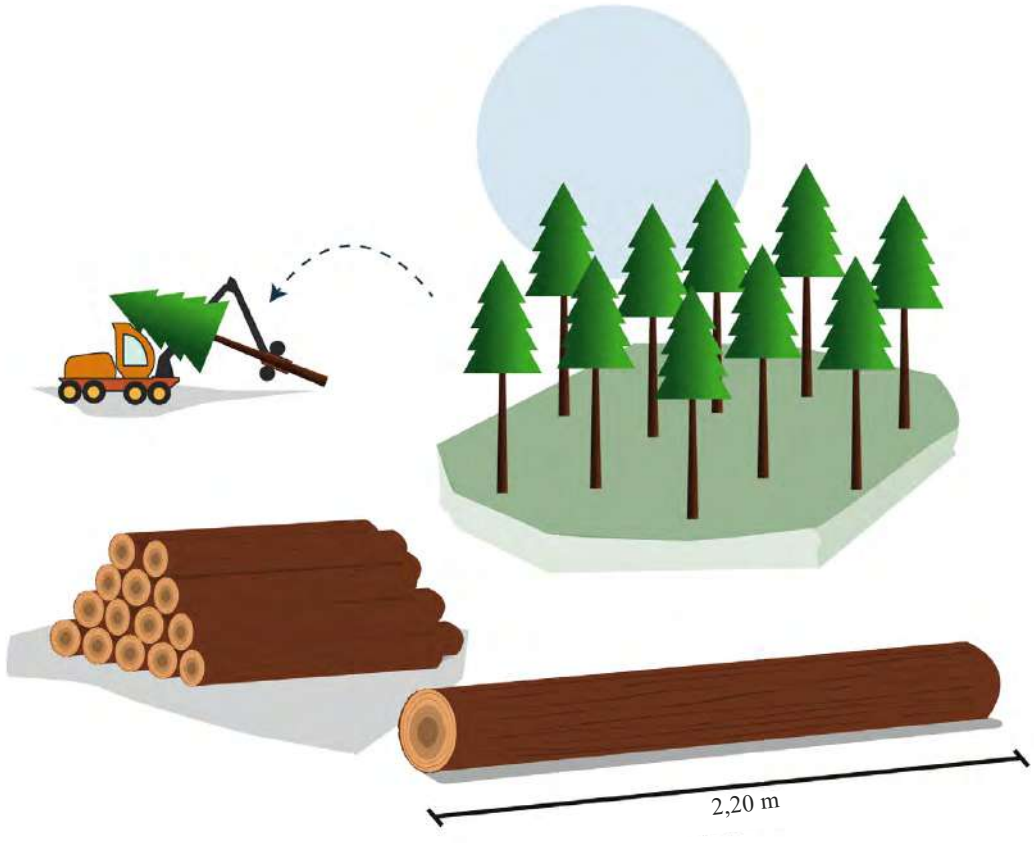
- **Unilateral direito:** localizado à direita da distribuição de Qui-quadrado, para um nível de significância α.
- **Unilateral esquerdo:** localizado à esquerda da distribuição de Qui-quadrado, para um nível de significância α. O valor crítico em uma tabela é encontrado observado o valor para o nível $(1 - \alpha)$.
- **Bilateral:** localizado nos dois extremos da distribuição de Qui-quadrado, para um nível de significância $\alpha/2$. O valor crítico superior em uma tabela é encontrado observando o valor para o nível $(1 - \alpha)$ e o valor crítico inferior observando o valor para o nível α.

EXERCÍCIO 13

Um Engenheiro Florestal que coordena a atividade de colheita florestal com processador florestal quer garantir que as toras processadas tenham um comprimento estabelecido de 2,20 m e, principalmente, que elas tenham um comprimento consistente. A consistência é obtida reduzindo-se a variação do tamanho do comprimento das toras processadas pela máquina (Figura 8.2).

Desta forma, foi estabelecido pelo Engenheiro Florestal o padrão de tamanho das toras com média 2,20 m e desvio-padrão de 0,2 m. Uma amostra de 100 toras foi obtida, com média 2,20 m e desvio-padrão de 0,3 m. Ao nível de significância de 0,05, teste a hipótese que as toras amostradas não têm desvio-padrão diferente do valor estabelecido como padrão (0,2 m).

FIGURA 8.2. Colheita florestal com harvester.



Foi estabelecido o seguinte teste de hipótese:

$$\begin{cases} H_0: \sigma = 0{,}2\ m \\ H_1: \sigma \neq 0{,}2\ m \end{cases}$$
$$\alpha = 0{,}05$$

A hipótese H_0 assume que o desvio-padrão da amostra não é diferente de 0,2 m. Observe que a hipótese formulada como um teste bicaudal, logo α (também denominado de *p-valor*) terá uma área dobrada à direita da estatística do teste, similar ao que foi discutido no teste de hipótese *t*.

Comandos para teste do desvio-padrão:

Arquivo: 20 - Função quidesvio para teste do desvio-padrão

```
>using DataFrames, Statistics, Distributions

>function quidesvio(s,σ,n)
    Quiquadrado=((n-1)*s^2)/(σ^2)
    GL=n-1
    Alpha = ccdf(Chisq(GL), Quiquadrado)*2
    println("Qui-quadrado: ", Quiquadrado)
    println("Grau de liberdade: ", GL)
    println("Alpha: ", Alpha)
end

#quidesvio(s,σ,n)
>quidesvio(0.3,0.2,100)
Qui-quadrado: 222.75
Grau de liberdade: 99
Alpha: 3.21e-11
```

Como o valor de α é inferior ao nível de significância de 0,05, rejeita-se a hipótese da nulidade. Portanto, há evidência suficiente de que o desvio-padrão é diferente de 0,2 m. Assim, seria aconselhável realizar uma nova regulagem no cabeçote da colheitadeira, com objetivo de produzir toras com comprimentos menos variáveis.

9 COMPARAÇÃO DE DUAS VARIÂNCIAS – A DISTRIBUIÇÃO F

Se um experimento for repetido, no qual várias amostras são selecionadas aleatoriamente de duas populações normalmente distribuídas e com a mesma variância, a distribuição da razão da variância amostral da população 1 pela variância amostral da população 2, $\frac{s_1^2}{s_2^2}$, origina a distribuição F.

Características da distribuição F:

- A distribuição é assimétrica e não possui valores negativos.
- A distribuição depende de dois graus de liberdade, pois ela é obtida pela razão de duas variâncias. Existe uma distribuição F diferente para cada par de graus de liberdade.
- O numerador da razão F deve ser sempre a maior variância entre as duas.

Para testar a hipótese que duas variâncias não diferem entre si, utiliza-se a estatística F:

$$F = \frac{s^2_{\text{maior}}}{s^2_{\text{menor}}}$$

Em que:
s^2_{maior} é a maior variância das duas populações amostradas
s^2_{menor} é a menor variância das duas populações amostradas

Os valores críticos, para avaliação da aceitação ou rejeição da hipótese de nulidade são obtidos na distribuição de F, observando-se:

- Nível de significância.
- Graus de liberdade do numerador $GLN = n - 1$ da população com maior variância.
- Graus de liberdade do denominador $GLD = n - 1$ da população com menor variância.

Quando as variâncias das duas populações não diferem entre si, o valor da estatística F se aproxima de 1. Um valor grande de F revela evidências que as variâncias são diferentes.

As condições para aplicação do teste são:

- As duas populações são independentes (a amostra de uma população não tem relação com a amostra obtida na outra população).
- As duas populações são normalmente distribuídas.

EXERCÍCIO 14

No exemplo mencionado na seção do teste t para amostras pequenas independentes, um Engenheiro Florestal deseja saber se a produção de madeira ($m^3.ha^{-1}$) de duas espécies, eucalipto e acácia-negra, seriam a mesma após 7 anos de plantio. A partir dos dados deste exercício, teste a hipótese de que as duas variâncias estimadas são oriundas de uma mesma e população.

Foi estabelecido o seguinte teste de hipótese:

$$\begin{cases} H_0: \sigma_1^2 = \sigma_2^2 \\ H_1: \sigma_1^2 \neq \sigma_2^2 \end{cases}$$
$$\alpha = 0{,}05$$

Comandos utilizados para testar duas variâncias:

ARQUIVO: 21 - FUNÇÃO DUASVARIANCIAS PARA TESTE DE DUAS VARIÂNCIAS

```
>using DataFrames, CSV, Statistics, Distributions

>X=CSV.read("Z:/Alexandre/Julia/Exemplos/amostrasindependentes.csv", DataFrame);

>function DUASVARIANCIAS(X1,X2)
    dfX1 = DataFrame([X1])
    X1=dropmissing(dfX1)
    X1= X1.x1
    dfX2 = DataFrame([X2])
    X2=dropmissing(dfX2)
    X2= X2.x1
    n1= length(X1)
    xbar1=sum(X1)/n1
    s²1=sum((X1.-xbar1).^2)/(n1-1)
    s1=sqrt(s²1)
    gl1=n1-1
    n2= length(X2)
    xbar2=sum(X2)/n2
    s²2=sum((X2.-xbar2).^2)/(n2-1)
    s2=sqrt(s²2)
    gl2=n2-1
    df = DataFrame()
    df.GrupoOriginal =["x1", "x2"]
    df.Media=[ xbar1, xbar2]
    df.Variancia=[ s²1, s²2]
    df.GrauDeLiberdade=[ n1, n2]
    df
    AT=sort(df, :Variancia, rev=true)
    F= AT.Variancia[1,1]/AT.Variancia[2,1]
    α= ccdf(FDist(gl1, gl2),abs(F))
    df = DataFrame()
    df.GrupoOriginal =["x1", "x2"]
    df.Media=[AT.Media[1,1], AT.Media[2,1]]
```

```
    df.Variancia=[AT.Variancia[1,1], AT.Variancia[2,1]]
    df. GrauDeLiberdade =[AT.GrauDeLiberdade[1,1], AT.GrauDeLiberdade[2,1]]
    df.F=[F,-]
    df.α=[α,-]
    return df
end

#DUASVARIANCIAS(X1,X2)
>DUASVARIANCIAS(X.x1,X.x2)
```

julia

```
julia> DUASVARIANCIAS(X.x1,X.x2)
2×6 DataFrame
```

Row	GrupoOriginal	Media	Variancia	GrauDeLiberdade	F	α
	String	Float64	Float64	Int64	Any	Any
1	x1	220.0	153.0	11	3.09438	0.0613937
2	x2	166.778	49.4444	9	-	-

O valor da estatística F foi de 3,09 para um valor de $\alpha = 0{,}06$. Como este valor é maior que o do nível de significância do teste de hipótese ($\alpha = 0{,}05$), assume-se que as variâncias são homogênas.

10 MEDIDAS DE ASSOCIAÇÃO

Existem situações no dia a dia do Engenheiro Florestal em que é necessário estabelecer associações entre variáveis. A altura da árvore aumenta quando aumenta o tamanho da área foliar da árvore? Para árvores com uma mesma idade, a densidade básica da madeira é maior para árvores com diâmetros menores? O preço do metro cúbico de madeira de pinus aumenta quando diminui o preço do metro cúbico de eucalipto? O poder calorífico da biomassa das árvores varia com a sua idade?

Observe que essas perguntas podem ser respondidas utilizando técnicas de regressão. É aconselhável antes de proceder a modelagem, avaliar como as duas variáveis se comportam. Os conjuntos de dados amostrais são sempre bivariados (2 variáveis) e um resumo dos dados resulta em duas médias, duas variâncias e uma covariância.

As variáveis estão associadas? Quanto e como as duas variáveis variam? Isto pode ser avaliado por meio das estatísticas covariância e correlação (linear). Estas duas estatísticas também são consideradas como parte da análise exploratória de dados.

Correlação, semelhante à covariância, é uma medida do grau de associação de duas variáveis e expressa pela média da intensidade de sua associação.

A estatística para avaliar o quanto as duas variáveis variam conjuntamente é denotada pela covariância. Seu valor é influenciado pelas unidades de medidas das variáveis e pela grandeza dos valores. A correlação padroniza o valor da covariância, com objetivo de mantê-la em uma escala que varia de -1 a 1, tornando as interpretações mais fáceis e mais claras.

10.1 COVARIÂNCIA

A covariância é uma estatística utilizada para medir a relação linear entre duas variáveis. Entretanto, ela não é apropriada para comparar o grau de relação entre variáveis, porque ela é influenciada pelas unidades de medida de cada variável e pela sua grandeza. O sinal e a grandeza (tamanho do valor da covariância) indicam o sentido e o grau da relação linear entre as variáveis. O resultado da covariância é de difícil interpretação. A covariância (s_{xy}) é obtida pela seguinte expressão:

$$s_{xy} = \frac{\sum_{i=1}^{n} xy - \frac{\sum_{i=1}^{n} x \sum_{i=1}^{n} y}{n}}{n - 1}$$

Em que:
x e y são as variáveis aleatórias.
n é o tamanho da amostra.

10.2 CORRELAÇÃO

A estatística coeficiente de correlação linear (r), ou coeficiente de correlação linear de Pearson, é obtida por:

$$r = \frac{S_{xy}}{S_x S_y}$$

Em que:
s_{xy} é o valor da covariância entre as variáveis x e y
s_x é o valor do desvio-padrão da variável x
s_y é o valor do desvio-padrão da variável y

O valor do coeficiente de correlação linear varia no intervalo de $-1 \leq r \leq 1$:

- Quando o valor do coeficiente é positivo, ocorre um aumento (ou diminuição) nas duas variáveis.
- Quando o valor do coeficiente é negativo, ocorre o aumento em uma variável e diminuição na outra.
- Quanto mais o valor se aproximar de 1 (positivo ou negativo) mais forte é correlação entre as variáveis e quando se obtém o valor 1 a correlação é perfeita.
- Quanto mais próximo de zero for este valor, não se tem correlação.

EXERCÍCIO 15

Um Engenheiro Florestal coletou dados de produção de biomassa em árvores (b, em $kg.m^{-2}$) em uma floresta e a radiação fotossinteticamente ativa interceptada por estas árvores (rfai, em $MJ.m^{-2}$), estando os dados apresentados na Tabela 10.1. Calcule e interprete os valores de covariância e correlação.

TABELA 10.1. Dados de produção de biomassa de árvores e radiação fotossinteticamente ativa interceptada por elas.

rfai	1.965,0	3.985,0	9.500,0	5.400,0	6.135,0	7.235,0	3.140,0	8.102,0	2.563,0	4.367,0
b	5,0	10,2	21,01	13,2	13,1	17,02	10,5	19,01	7,2	14,2

Comandos usados para cálculo de covariância e coeficiente de correlação.

Arquivo: 22 - Função covecor para cálculo de covariância e coeficiente de correlação

```
>using DataFrames, CSV, Statistics, Distributions

>X=CSV.read("Z:/Alexandre/Julia/Exemplos/covariancia.csv", DataFrame)
```



```
julia> X=CSV.read("Z:/Alexandre/Julia/Exemplos/covariancia.csv", DataFrame)
10×2 DataFrame
| Row | rfai    | b       |
|     | Float64 | Float64 |
|-----|---------|---------|
| 1   | 1965.0  | 5.0     |
| 2   | 3985.0  | 10.2    |
| 3   | 9500.0  | 21.01   |
| 4   | 5400.0  | 13.5    |
| 5   | 6135.0  | 16.1    |
| 6   | 7235.0  | 17.02   |
| 7   | 3140.0  | 10.5    |
| 8   | 8102.0  | 19.01   |
| 9   | 2563.0  | 7.2     |
| 10  | 4367.0  | 14.2    |

>function COVECOR(x,y)
    n= length(x)
    cov=(sum(x.*y)-((sum(x)*sum(y))/n))/(n-1)
    xbar=sum(x)/n
    s²x=sum((x.-xbar).^2)/(n-1)
    sx=sqrt(s²x)
    ybar=sum(y)/n
    s²y=sum((y.-ybar).^2)/(n-1)
    sy=sqrt(s²y)
    cor=cov/(sx*sy)
    println("Covariância: ", cov)
    println("Correlação: ", cor)
end

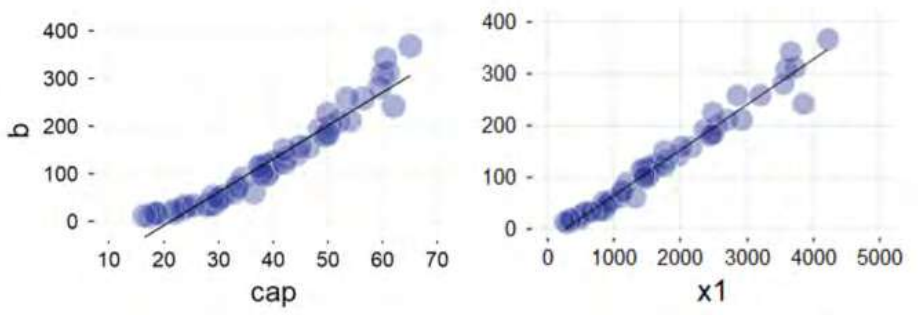
#COVECOR(x,y)
>COVECOR(X.rfai,X.b)
Covariância: 12404.85
Correlação: 0.97
```

O valor da covariância foi de 12.404,85 (MJ/m²)/(kg/m²). Como o valor encontrado foi positivo, conclui-se que quando os valores de *rfai* aumentam os valores de b também aumentam. Entretanto, não se sabe qual é o grau de associação entre estas duas variáveis, pois elas não se encontram na mesma escala, o que torna difícil a interpretação do valor acima apresentado. Para contornar esta dificuldade, calcula-se o coeficiente de correlação entre as variáveis.

O valor do coeficiente de correlação linear foi de 0,97. O valor encontrado varia no intervalo $-1 \leq r \leq 1$. Conclui-se que existem evidências que essas duas variáveis estão associadas e com forte dependência. O valor do coeficiente de correlação é adimensional, uma vez que a unidade de medida da covariância (MJ/m² Kg/m²) se anula ao dividí-lo pelo produto dos desvios-padrões de *rfai* (MJ/m²) e *b* (Kg/m²).

Para a aplicação do coeficiente de correlação linear, assume-se que existe uma relação linear entre as duas variáveis de interesse. Dada esta condição, a mudança de uma variável corresponde à mudança constante no valor médio da outra variável. Na Figura 10.1, observa-se a relação entre as variáveis *cap* (circunferência à altura do peito, cm) e *b* (biomassa, kg). A relação entre estas duas variáveis pode ser representada por uma equação polinomial de segundo grau. Quando se utiliza a variável transformada x_1 $(x_1 = cap^2)$, a relação passa a ser representada pela equação de uma reta. Em muitas situações, embora não seja o mais correto, admite-se uma relação linear entre as variáveis *b* e *cap*, ao menos para se ter uma ideia do grau desta associação.

FIGURA 10.1. Relação entre as variáveis biomassa e circunferência à altura do peito para árvores de acácia-negra. com 10 anos.



Considerações sobre análise de correlação

- O valor do coeficiente de correlação é afetado pela presença de valores estranhos (*outliers*), principalmente em pequenas amostras, o que pode comprometer a análise. Para amostras grandes a presença de valores estranhos afeta pouco o coeficiente. Existem testes estatísticos próprios para a identificação de *outliers*, a própria análise gráfica já permite detectar sua ocorrência.
- A independência das observações é requerida, ou seja, a ocorrência de uma observação (x_1, y_1) não influencia a ocorrência de outra observação (x_2, y_2).
- O coeficiente de correlação linear de Pearson deve ser aplicado quando ambas as variáveis estão na escala de intervalo ou razão.
- As variáveis devem ter uma distribuição normal conjunta bivariada (para cada valor de X, a variável Y é normalmente distribuída). Esta condição é necessária para realização de inferências estatísticas, tais como teste de hipótese da correlação e construção de intervalos de confiança. Muitas vezes, dado que ambas as variáveis apresentam distribuição normal univariadas, espera-se que, então, se obtenha uma distribuição normal conjunta bivariada. Quando o objetivo é obter apenas o valor da correlação amostral, a condição de normalidade não é requerida.
- Em algumas situações a distribuição bivariada não é normal. Uma solução para isso é a transformação dos dados destas variáveis para aproximá-la da distribuição normal conjunta bivariada, que podem ser transformações lineares ou não-lineares.

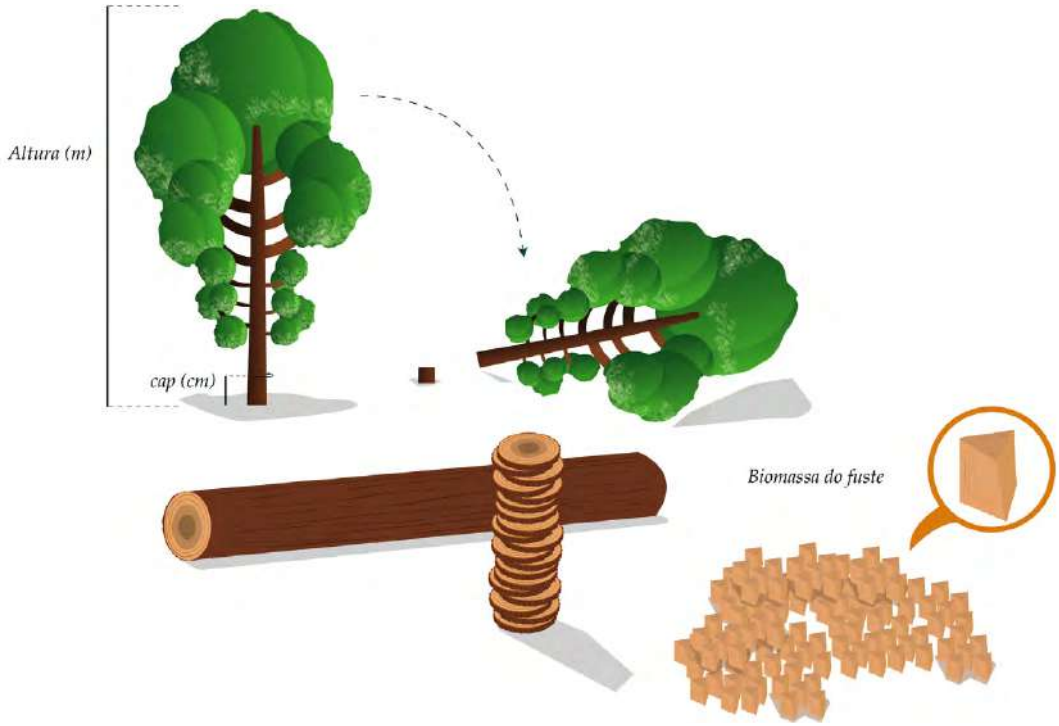
- Transformações lineares resultam em apenas mudança da escala ou de origem da variável, sendo pouco efetivo para aproximação da normalidade. Se uma variável não é normal, uma transformação linear também não será normal. Assim, as transformações mais importantes são as não lineares (logarítmica, quadrado, raiz quadrada, entre outras), em que um incremento na escala original não corresponde ao mesmo incremento para a variável transformada (que está em uma nova escala). Ao utilizar transformações não lineares, é importante preservar a ordenação das observações. É lógico esperar que transformações lineares não afetarão a correlação entre duas variáveis e que transformações não lineares podem afetar os valores de correlação.

A construção de gráficos bivariados para representar a correlação entre as variáveis é muito útil. Além de permitir avaliar condicionantes acerca da aplicação da análise de correlação, é uma maneira interessante, intuitiva e visual para se observar a existência de associações entre as variáveis. Assim, a representação numérica da correlação em uma tabela e, complementarmente, a adição de gráficos *cornerplot*, por exemplo, revela se é apropriado avaliar a associação entre as variáveis por meio do coeficiente de correlação linear.

EXERCÍCIO 16

Um Engenheiro Florestal interessou-se em saber se as variáveis circunferência à altura do peito (cap, em cm), altura (h, em m), biomassa do fuste (b, em kg), teor de carbono do fuste (c, em %) e poder calorífico do fuste (pcs, em kcal kg^{-1}) estão associadas. Assim, uma amostra de 60 árvores foi coletada aleatoriamente em uma floresta de acácia-negra com 7 anos e mensuradas as variáveis citadas (Figura 10.2). Os dados estão disponibilizados em um arquivo denominado *energia*.

FIGURA 10.2. Coleta de variáveis altura, cap e biomassa do fuste da espécie florestal acácia-negra.



Pede-se:

- Construa uma matriz de covariância. Interprete os resultados.
- Construa uma matriz de correlação. Interprete os resultados.
- Construa um gráfico cornerplot. Interprete os resultados.

Comandos para obtenção das matrizes de covariância e correlação e do gráfico:

Arquivo: 23 - Função SR para matriz de covariância e correlação

```
>using DataFrames, Statistics, Distributions, LinearAlgebra

>X=CSV.read("Z:/Alexandre/Julia/Exemplos/energia.csv", DataFrame)
```

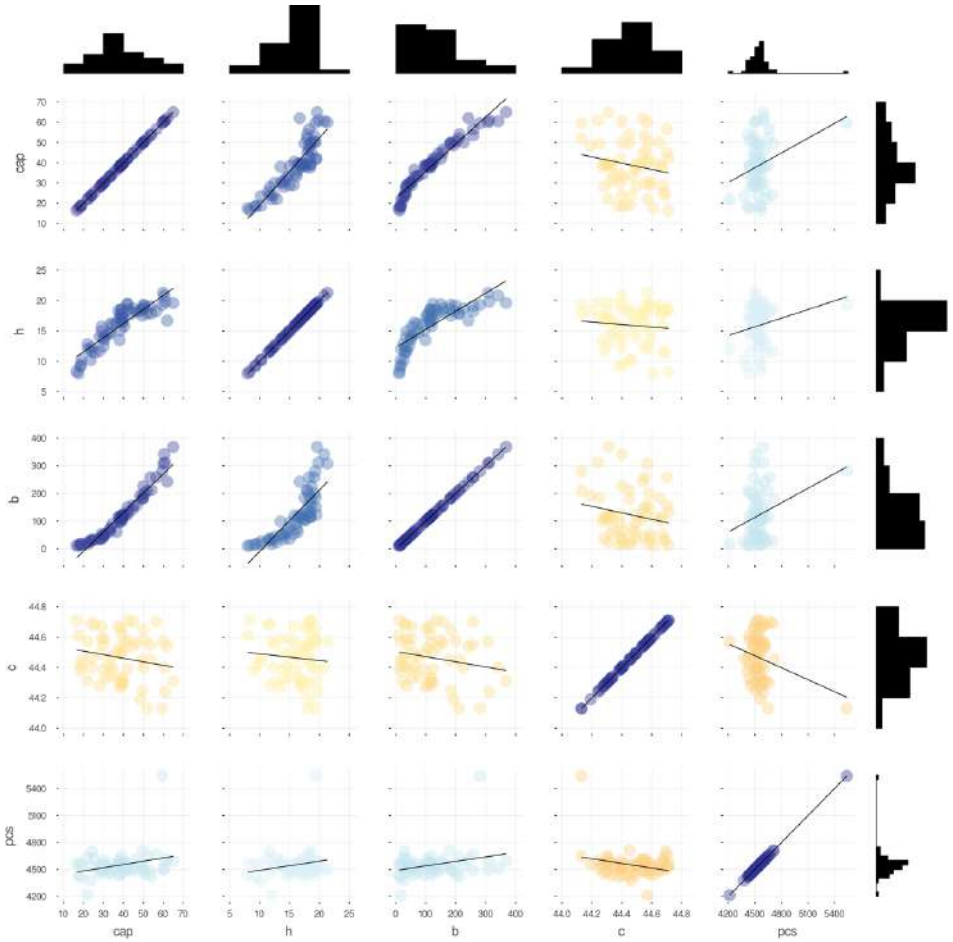
julia

```
julia> X=CSV.read("Z:/Alexandre/Julia/Exemplos/energia.csv", DataFrame)
60×5 DataFrame
| Row | cap     | h       | b       | c       | pcs   |
|     | Float64 | Float64 | Float64 | Float64 | Int64 |
|-----|---------|---------|---------|---------|-------|
| 1   | 39.5    | 16.45   | 115.75  | 44.33   | 4558  |
| 2   | 23.5    | 10.2    | 26.43   | 44.46   | 4378  |
| 3   | 33.5    | 14.8    | 75.04   | 44.34   | 4562  |
| 4   | 17.5    | 8.0     | 10.4    | 44.61   | 4435  |
| 5   | 30.0    | 12.5    | 43.65   | 44.55   | 4560  |
⋮
| 55  | 42.5    | 18.5    | 133.92  | 44.71   | 4529  |
| 56  | 42.0    | 19.5    | 123.21  | 44.39   | 4503  |
| 57  | 42.0    | 19.4    | 150.2   | 44.58   | 4578  |
| 58  | 37.5    | 18.4    | 114.22  | 44.67   | 4579  |
| 59  | 38.0    | 13.5    | 98.7    | 44.27   | 4555  |
| 60  | 45.0    | 18.7    | 157.27  | 44.7    | 4513  |
```

```
>function SR(X)
    n=size(X,1)
    p=size(X,2)
    A = [[X[:,i].-mean(X[:,i])]' for i in 1:p]
    S=A*A'/(n-1)
    s²=Diagonal(S)
    s=sqrt(s²)
    INVS=inv(s)
    R=INVS*S*INVS
    println("S: ", S)
    println("R: ", R)
end
```

```
#SR(X)
>SR(X)
S:
5×5 Array{Float64,2}:
157.53          36.63          1099.85-0.37          569.50
36.63           11.02          244.63         -0.05          111.41
1099.85 244.63         8222.52-2.81          4113.24
-0.37           -0.05          -2.81          0.02           -6.20
569.50          111.41         4113.24-6.20          23405.40
R:
5×5 Array{Float64,2}:
1.0000          0.8793         0.9664         -0.19240.2966
0.8793          1.0000         0.8129         -0.09670.2194
0.9664          0.8129         1.0000         -0.20070.2965
-0.1924 -0.0967-0.20071.0000         -0.2628
0.2966          0.2194         0.2965         -0.26281.0000

>using StatsPlots
gr(size = (1200, 1200))
@df X cornerplot([:cap :h :b :c :pcs], grid = false, dpi=900)
```



Observação: A cor azul indica correlação positiva, vermelha negativa e amarela ausência de correlação.

Com base na análise gráfica e numérica (por meio da aplicação do coeficiente de correlação linear), pergunta-se:

1. Todos os pares de variáveis apresentam relação linear?
2. Existe suspeita de algum *outlier*? Se existir evidências de *outlier*, qual será o tratamento apropriado para os respectivos pares de valores?
3. Seria interessante transformar as variáveis para se obter o coeficiente de correlação linear?
4. Entre as variáveis: circunferência à altura do peito (cap, em cm), altura (h, em m), biomassa do fuste (b, em kg), teor de carbono do fuste (c, em %) e poder calorífico do fuste (pcs, em kcal. kg^{-1}), quais estão relacionadas?

10.3 INTERVALO DE CONFIANÇA E TESTE DE HIPÓTESE PARA A CORRELAÇÃO

O cálculo do intervalo de confiança para a correlação de Pearson (ρ – *valor* da correlação para a população) apresenta uma característica única: a distribuição amostral de *r* não é normalmente distribuída. A solução para construir um intervalo de confiança é a utilização da transformação Z' de Fisher, uma vez que os valores de Z' se aproximam de uma distribuição normal, permitindo construir os intervalos de confiança e a realização de testes de hipóteses.

Os passos para calcular um intervalo de confiança para ρ são:

- Converter *r* em *Z'*.
- Calcular um intervalo de confiança em função de *Z'*.
- Reconverter o intervalo de confiança novamente para *r*.

Observa-se que Z' correponde aos valores da população e z' aos valores da amostra (o qual será utilizado).

O intervalo de confiança para a correlação pode ser obtido por:

$$z' = 0{,}5 \ln \frac{1+r}{1-r} \text{ (média)}$$

$$\sigma_{z'} = \sqrt{\frac{1}{n-3}} \text{ (desvio)}$$

$$IC(\mu_{z'}) = z' \pm z_{0{,}95}\, \sigma_{z'}, \text{ para 95\% de confiança}$$

$$r = \frac{\left(e^{2\,z'} - 1\right)}{\left(e^{2\,z'} + 1\right)}, \text{ reconversão do valor de } z' \text{ para } r$$

O valor de *z* para 0,95 de probabilidade ($z_{0,95}$) é 1,96, o qual é obtido por meio da distribuição normal.

Considere o resultado da correlação *r* = 0,97 entre as variáveis *rfai* e *b*, mencionadas no exemplo anterior, para uma amostra de tamanho *n* = 10. Obtenha o intervalo de confiança para a correlação ρ, para 0,95 de probabilidade, com base no valor *r* = 0,97.

Comandos para a obtenção do intervalo de confiança para correlação ρ:

Arquivo: 24 – Teste de hipótese para a correlação

```
#intervalo de confiança para correlação
>using DataFrames, CSV, Statistics, Distributions, LinearAlgebra

>X=CSV.read("Z:/Alexandre/Julia/Exemplos/energia.csv", DataFrame)
>n=10;
>r=0.97;
>Z=0.5*log((1+r)/(1-r))
2.09
>σZ=1/sqrt(n-3)
0.38
>ICInferiorZ=Z-1.96*σZ
1.35
>ICSuperiorZ=Z+1.96*σZ
2.83
>r_ICInferior=((exp(2*ICInferiorZ))-1)/((exp(2*ICInferiorZ))+1)
0.87
>r_ICSuperior=((exp(2*ICSuperiorZ))-1)/((exp(2*ICSuperiorZ))+1)
0.99
```

O intervalo de confiança a 0,95 de probabilidade para ρ é:

$$IC = [0{,}87 \leq \rho \leq 0{,}99\ ] = 0{,}95$$

Como já mencionado, a distribuição de ρ é assimétrica, especialmente para valores altos (negativos ou positivos). Entretanto, quando os valores de ρ se aproximam de zero a distribuição se aproxima de uma forma simétrica, o que torna possível testar a hipótese H_0: $\rho = 0$ por meio de um teste t de Student.

Para testar se a correlação é significante, ou seja, testar a hipótese da nulidade de que $\rho = 0$, aplica-se um teste t:

$$t_r = \frac{r}{\sqrt{\frac{(1-r^2)}{n-2}}}$$

A estatística do teste (t_r) tem distribuição t de Student com n - 2 graus de liberdade.

O valor do teste elevado ao quadrado corresponde ao valor de F na análise de variância da regressão linear. A distinção entre significância e não-significância não é um problema sério com amostras pequenas, mas pode ser com amostras grandes, nas quais pequenos valores de r podem ser significantes.

No teste de hipótese da correlação têm-se:

H_0: $\rho = 0$, ou seja, não existe correlação entre as variáveis
H_1: $\rho \neq 0$, ou seja, existe correlação entre as variáveis

Teste a hipótese que $\rho = 0$ para as variáveis *rfai* e *b*, mencionadas no primeiro exemplo.

Comandos para a obtenção do teste *t* para correlação ρ.

Arquivo: 24 – Teste de hipótese para a correlação

```
#Teste t para a correlação

>using Distributions
>tr=r/sqrt((1-r^2)/(n-2))
11.29
>Alpha=0.05;
>tVal = quantile(TDist(n-2),1-Alpha/2)
2.31
>Alpha = 2*ccdf(TDist(n-2), abs(tr))
3.42e-6
```

O valor de t_r de 11,29 foi maior que o valor crítico 2,31, sendo o valor de Alpha pequeno (3,42 e^{-6}), revelando que a correlação linear r = 0,97 difere significantemente de zero. Portanto, existe evidência de correlação positiva entre as variáveis *rfai* e *b*.

10.4 TESTE DOS COEFICIENTES DE CORRELAÇÃO

Em algumas situações busca-se testar a diferença dos valores de ρ entre duas populações.

Para testar a hipótese sobre um ou dois valores de r, esses são convertidos para Z'. A retransformação para obtenção de r não é requerida, pois isto é somente feito quando se constrói o intervalo de confiança. As hipóteses testadas são:

$$\begin{cases} H_0: \rho_1 = \rho_2 \\ H_1: \rho_1 \neq \rho_2 \end{cases}$$
$$\alpha = 0{,}05$$

Para obtenção do valor, usa-se a seguinte transformação da correlação:

$$z' = 0{,}5 \ln \frac{1+r}{1-r}$$

A retranformação, se for de interesse, é obtida por:

$$r_w = \frac{\left(e^{2\,z'} - 1\right)}{\left(e^{2\,z'} + 1\right)}$$

Uma vez que os valores de correlação foram convertidos para valores de z', a distribuição normal pode ser utilizada para avaliar o teste de hipótese.

$$z_c = \frac{z'_1 - z'_2}{\sigma_{z'_1 - z'_2}}$$

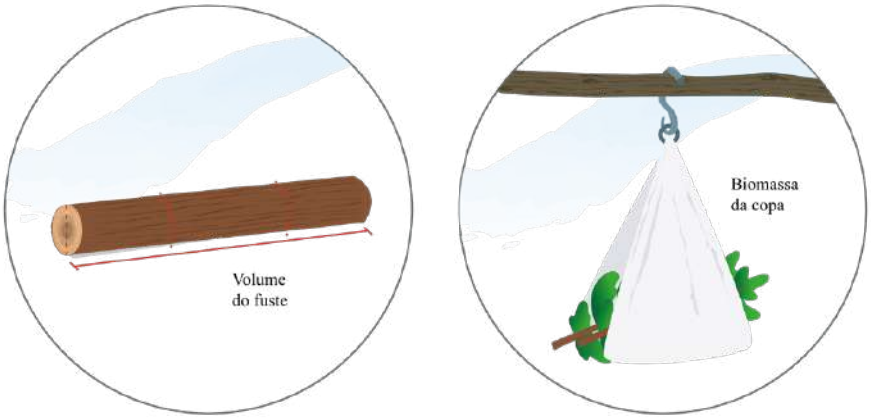
O desvio-padrão das diferenças é obtido por:

$$\sigma_{z'_1 - z'_2} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}$$

EXERCÍCIO 17

Um Engenheiro Florestal obteve a biomassa da copa e volume do fuste de árvores mortas (em pé) e árvores vivas em povoamentos de acácia-negra (Figura 10.3). O valor da correlação entre a biomassa da copa e o volume do fuste para as árvores mortas foi de $r_1 = 0{,}77$, para uma amostra de tamanho $n_1 = 23$. O valor da correlação entre a biomassa da copa e o volume do fuste para as árvores vivas foi de $r_2 = 0{,}92$, para uma amostra de tamanho $n_2 = 53$. Teste a hipótese que os valores das duas correlações não são diferentes.

FIGURA 10.3. Obtenção da biomassa de copa e volume do fuste de árvores de acácia-negra.



Comandos para a obtenção do teste de dois coeficientes de correlação:

ARQUIVO: 24 – TESTE DE HIPÓTESE PARA A CORRELAÇÃO

```
#Teste para dois coeficientes de correlação
>n1=23;
>r1=0.77;
>n2=53;
>r2=0.92;
>Z1=0.5*log((1+r1)/(1-r1))
1.02
>Z2=0.5*log((1+r2)/(1-r2))
1.59
>σZ1Z2=1/sqrt((n1-3)+(n2-3))
0.12
```

```
>Zc=(Z1-Z2)/σZ1Z2
-4.76
```

O valor crítico de *z* para 0,95 de probabilidade ($z_{0,95}$) é 1,96, o qual é obtido por meio da distribuição normal. Assim, o valor de $Zc = -4{,}76$ é maior que o valor crítico, revelando evidências que as duas correlações diferem estatisticamente. Portanto, a correlação entre a biomassa da copa e o volume do fuste, para árvores vivas e mortas, é diferente.

Em outras situações busca-se testar a diferença dos valores de ρ entre várias populações. A razão para isto é a possibilidade de obter um único valor de coeficiente de correlação ao invés de vários.

Por exemplo, a construção de modelos alométricos implica na coleta de dados em vários locais, logo uma análise de correlação é exigida previamente. Assim, amostras são obtidas em povoamentos situados em diferentes locais e, às vezes, é mais interessante analisar estes locais como um todo e não considerando uma correlação por local. Entretanto, pode ser detectado que as variâncias, devido aos locais, não sejam homogêneas e, neste caso, a obtenção de um valor de correlação simples (comum aos diferentes locais) não é válida. Para avaliar a homogeneidade das correlações um teste estatístico pode ser efetuado.

As hipóteses testadas são:

$$\begin{cases} H_0: \rho_1=\rho_2=\rho_3=\ldots=\rho_I \\ H_1: \rho_i \neq \rho_{i'}, \text{ em que } i \neq i' \end{cases}$$
$$\alpha=0{,}05$$

Uma vez que os valores de correlação foram convertidos para valores de *z*', um teste de Qui-quadrado é utilizado para avaliar a hipótese da nulidade.

$$\chi_c^2 = \sum_{i=1}^{k}\left(\frac{z_i' - \bar{z}_w'}{1/\sqrt{n_i - 3}}\right)^2 = \sum_{i=1}^{k}(n_i - 3)(z_i' - \bar{z}_w')^2$$

Em que:

$$\bar{z}_w' = \frac{\sum_{i=1}^{k}(n_i - 3)z_i'}{\sum_{i=1}^{k}(n_i - 3)}$$

O valor de χ_c^2 tem distribuição de χ^2 com $k - 1$ graus de liberdade (k é o número de correlações).

EXEMPLO 18

Um Engenheiro Florestal efetuou uma avaliação da biomassa da copa e volume do fuste de árvores vivas de acácia-negra em três locais. O valor da correlação para o local 1 foi de $r_1 = 0{,}94$, para uma amostra de tamanho $n_1 = 57$, para o local 2 foi de $r_2 = 0{,}91$, para uma amostra de tamanho $n_2 = 61$ e para o local 3 foi de $r_3 = 0{,}88$, para uma amostra de tamanho $n_3 = 51$. Teste a hipótese que os valores das três correlações não são diferentes.

Comandos para a obtenção do teste de várias correlações.

Arquivo: 24 – Teste de hipótese para a correlação

```
#Teste para várias correlações
using Distributions
>n1=57;
>n2=61;
>n3=50;
>r1=0.94;
>r2=0.91;
>r3=0.88;
>Z1=0.5*log((1+r1)/(1-r1))
1.74
>Z2=0.5*log((1+r2)/(1-r2))
1.53
>Z3=0.5*log((1+r3)/(1-r3))
1.38
>Zw=((n1-3)*Z1+(n2-3)*Z2+(n3-3)*Z3)/((n1-3)+(n2-3)+(n3-3))
1.55
>QuiQc=(((n1-3)*(Z1-Zw)^2)+(n2-3)*(Z2-Zw)^2)+((n3-3)*(Z3-Zw)^2)
3.37
>QuiQuadCrit = quantile(Chisq(3-1),0.95)
5.99
```

O valor de Qui-quadrado calculado foi de 3,37 e o valor de Qui-quadrado crítico, obtido de tabela, é de 5,99, revelando existir evidências que os coeficientes de correlação são homogêneos.

Dada a homogeneidade dos coeficientes de correlação, pode-se obter um coeficiente de correlação ponderado (r_w), de modo a retranformar o valor de $\bar{z}'_w$:

$$r_w = \frac{\left(e^{2\,z'_w} - 1\right)}{\left(e^{2\,z'_w} + 1\right)}$$

Comando para a obtenção do valor do coeficiente de correlação ponderado.

```
>rw=((exp(2*Zw))-1)/((exp(2*Zw))+1)
0.91
```

A retransformação do valor de $\bar{z}'_w$ para o valor do coeficiente de correlação ponderado (comum as três populações amostradas) resultou no valor de $r_w = 0{,}91$.

Um intervalo de confiança para a correlação ponderada pode ser obtido por:

$$\sigma_{\bar{z}'_w} = \frac{1}{\sqrt{\sum_{i=1}^{k}(n_i - 3)}} \text{ (desvio-padrão)}$$

$$IC\left(\mu_{\bar{z}'_w}\right) = \bar{z}'_w \pm z_{0,95}\,\sigma_{\bar{z}'_w}, \text{ para 95\% de confiança}$$

O valor de z para 0,95 de confiança ($z_{0,95}$) é 1,96, obtido por meio da distribuição normal. O valor do intervalo de confiança pode ser retransformado para os limites de r_w.

Comandos para a obtenção do intervalo de confiança:

Arquivo: 24 – Teste de hipótese para a correlação

```
#Intervalo de confiança para várias correlações
>Zw
1.55
>σZw=1/sqrt((n1-3)+(n2-3)+(n3-3))
0.08
>ICInferiorZw=Zw-1.96*σZw
1.40
>ICSuperiorZw=Zw+1.96*σZw
1.71
>rw_ICInferior=((exp(2*ICInferiorZw))-1)/((exp(2*ICInferiorZw))+1)
0.89
>rw_ICSuperior=((exp(2*ICSuperiorZw))-1)/((exp(2*ICSuperiorZw))+1)
0.94
```

O intervalo de confiança com 0,95 de probabilidade para ρ, comum aos três locais, é:

$$IC = [0{,}89 \leq \rho \leq 0{,}94\ ] = 0{,}95$$

11 TERMOS E CONCEITOS EM DELINEAMENTOS EXPERIMENTAIS

O termo delineamento experimental diz respeito à estrutura básica em que o experimento é instalado. Ela é a forma de como os tratamentos provenientes de um fator são atribuídos às unidades experimentais, tendo como principal finalidade reduzir o erro experimental.

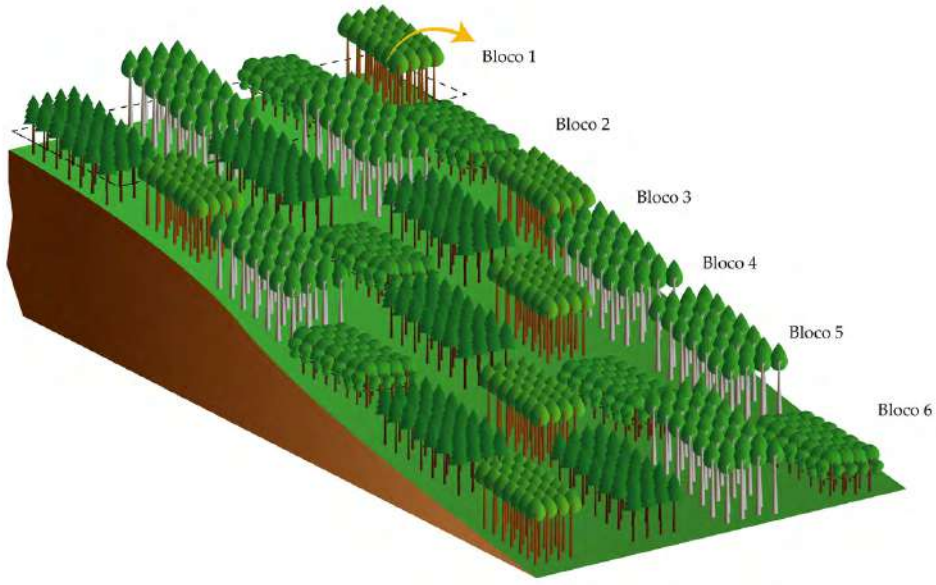
Existem vários arranjos de um delineamento experimental e os principais que são utilizados em pesquisas na Engenharia Florestal estão apresentados na sequência.

11.1 TERMOS E CONCEITOS

Em pesquisa, alguns dos termos e conceitos importantes na estruturação dos experimentos devem ser conhecidos e descritos detalhadamente.

Por exemplo, na Figura 11.1 está apresentado um experimento que foi conduzido em campo para avaliar o crescimento de espécies florestais. Quais são os termos que você usaria para descrever a estrutura (desenho) do experimento?

FIGURA 11.1. Experimento para avaliar a produção em volume de madeira do fuste oito anos após a instalação de diferentes espécies florestais com delineamento em blocos ao acaso.



Tratamentos

São os diferentes procedimentos que se quer comparar, por exemplo, tratamentos podem ser tipos ou quantidades de fertilizantes, espaçamentos de cultivos, tipos de substratos e clones de uma espécie.

Unidades experimentais

É o local onde serão aplicados os tratamentos. Unidades experimentais podem ser uma área de terra com árvores, as quais receberão tipos ou quantidades de fertilizantes, espaçamentos de cultivos, tubetes com plântulas e substratos, tubetes com substrato para os clones de uma espécie.

É o espaço ou objetos nos quais as respostas são obtidas. As unidades respostas podem diferir das unidades experimentais. Por exemplo, considere o efeito de diferentes fertilizantes no teor de nitrogênio de mudas de Cedro. Diferentes parcelas de campo são as unidades experimentais, mas as unidades de medida podem ser um subconjunto das plantas de Cedro na parcela de campo, ou uma amostra de folhas. Em outros casos, a unidade experimental pode ser as árvores medidas nas unidades experimentais, excluindo-se as plantas de bordadura.

Respostas

É o resultado que se observa após a aplicação de um tratamento a uma unidade experimental. As respostas são obtidas por meio de mensuração de variáveis, as quais fornecem informações para avaliar o que aconteceu no experimento. Comumente são avaliadas várias variáveis respostas, como, por exemplo, altura, diâmetro à altura do peito, biomassa e volume.

Aleatorização

Consiste em alocar aleatoriamente os tratamentos nas unidades experimentais. Vários métodos podem ser utilizados, o mais comum é o sorteio dos tratamentos nas unidades experimentais previamente numeradas. Outros aspectos de um experimento também podem ser aleatorizados: ordem em que os blocos serão instalados ou a ordem em que as unidades serão avaliadas.

Erro experimental

É a variação aleatória presente nos experimentos. Um mesmo tratamento aplicado a diferentes unidades experimentais (mesmo que estas sejam homogêneas), ou mesmo um tratamento aplicado na mesma unidade experimental em sucessivos experimentos, resultará em respostas diferentes. Erro experimental não se refere à condução do experimento de maneira errada, mas sim as variações que ocorrem entre unidades experimentais tratadas da mesma forma e que não se tem controle.

Avaliação às cegas

A avaliação às cegas consiste em conduzir o experimento no qual os avaliadores não sabem quais foram os tratamentos aplicados às unidades experimentais. Isto pode evitar a ocorrência de tendenciosidade na avaliação, especialmente quando se têm expectativas para certos tratamentos.

Controle ou testemunha

Um tratamento de controle ou testemunha é denominado padrão, ou seja, é usado como uma linha de base ou base de comparação para os outros tratamentos. Geralmente o controle é um tratamento de uso comum (que já é utilizado) ou pode ser um tratamento nulo (sem nenhum tratamento). Por exemplo, em um estudo de doses de fertilização em mudas de um viveiro, a testemunha poderia ser a ausência de adubação. Em um experimento para avaliar o espaçamento de cultivo de eucalipto, poderia ser o espaçamento 3x2 m, que é o mais comum nos plantios da espécie.

Um tratamento é chamado de nulo quando a sua aplicação pode resultar em alguma interferência. Por exemplo, um experimento para avaliar a eficácia da aplicação de herbicidas com pulverizador em um plantio de acácia-negra e seu efeito sobre o crescimento da espécie. A aplicação dos herbicidas é efetuada com pulverizadores acoplados a um trator agrícola, que pode compactar (minimamente) o solo nas unidades experimentais que os receberão, além de uma mínima quantidade adicional de água. Assim, a pulverização será também efetuada sobre a testemunha, porém apenas com água, o que dá origem ao tratamento nulo.

Fatores

Fatores podem ser únicos ou combinados e são utilizados para formar os tratamentos. Por exemplo, os cultivos florestais são conduzidos em um determinado espaçamento sobre um tipo de adubação. Os tratamentos podem ser a combinação dos espaçamentos e dos tipos de adubação, testados ao mesmo tempo. Assim, tem-se um fator de espaçamento e um fator de tipo de adubação. Configurações individuais para cada fator são chamadas de níveis do fator. Por exemplo, para os espaçamentos os níveis seriam 2x1m, 3x1m, 3x2m e para os tipos de adubação, a orgânica, a química e a testemunha.

12 PRINCÍPIOS BÁSICOS DA EXPERIMENTAÇÃO

Existem três princípios básicos indispensáveis para validar as conclusões dos experimentos florestais, que são:

1. **Repetição:** repetir cada tratamento visando possibilitar a avaliação de seu efeito e propiciar a obtenção de estimativa do erro experimental.
2. **Casualização:** aplicar os tratamentos às unidades experimentais de maneira aleatória visando evitar a tendenciosidade em seus efeitos e, também, no erro experimental.
3. **Controle local (controle das unidades experimentais):** consiste em agrupar unidades experimentais homogêneas a partir de um conjunto de unidades heterogêneas. O objetivo, neste caso, é reduzir o erro experimental, o que influência na escolha do tipo de delineamento a ser utilizado.

13 BASES PARA TESTE DE HIPÓTESES E ANÁLISE DE VARIÂNCIA

O modelo mais simples de um delineamento experimental é dado por:

$$y_{ij} = \mu + \alpha_i + e_{ij} \quad (1)$$

A análise de variância, conhecida como ANOVA, é um procedimento aritmético de particionamento da variabilidade total dos dados em componentes atribuídos aos tratamentos (α_i) e uma parte remanescente atribuída ao erro e_{ij}, que não é controlada no experimento, chamada de erro experimental.

Uma maneira de construir critérios para comparar modelos estatisticamente, é testar a hipótese da nulidade de que um modelo único para a média é verdadeiro versus a hipótese alternativa de que vários modelos separados são verdadeiros. Utilizando a notação de médias por grupos $\mu_i = \mu + \alpha_i$, a hipótese da nulidade para a média única pode ser expressa como segue:

H_0: $\alpha_i = 0$, (para todos os tratamentos)
H_1: $\alpha_i \neq 0$, (para algum ou mais tratamentos)

Os cálculos da ANOVA são sumarizados em uma tabela onde existe uma linha para cada fonte de variação do modelo linear do delineamento utilizado. Por exemplo, no modelo 1 as fontes de variação são tratamentos e erro, que podem ser também denominadas de entre grupos (*Between*) e dentro dos grupos (*Within*). Na maioria das vezes, a variação total dos dados também é apresentada.

A tabela 13.1 é típica para sumarização dos cálculos de ANOVA.

TABELA 13.1. Tabela resumo da análise de variância.

Fontes de variação	Grau de Liberdade	Soma de Quadrado	Médias Quadraticas ou Quadrados Médios	F
Tratamentos	$t - 1$	SQ_e	$MQ_e = SQ_e / (t - 1)$	MQ_e / MQ_d
Erro	$tr - t$	SQ_d	$MQ_d = SQ_d / (tr - t)$	
Total	$tr - 1$	SQ_t		

Em que t é o número de tratamento e r é o número de repetições.

A estatística F (razão de F) na tabela de ANOVA é utilizada para testar a hipótese da nulidade que as médias dos tratamentos não diferem entre si. Quando a hipótese da nulidade é verdadeira os valores de F se aproximam de 1 e quando é rejeitada o valor de F é maior que 1.

Adicionalmente, é computado a estatística *p-valor* ou nível de significância do teste (α). Pequenos valores de *p-valor* indicam evidências para rejeição da hipótese da nulidade. Geralmente, p-valores menores que 0,05 são denominados estatisticamente significantes, e quando menores que 0,01 altamente significantes.

Em alguns casos, um pequeno p-valor pode ser obtido com um valor relativo pequeno de α_i, se a amostra não for grande o suficiente, ou se a variância σ^2 não for pequena o suficiente. Diferenças grandes entre tratamentos podem não ser reveladas devido a estatística F não ser significante, devido a amostragem ser pequena e a variância do erro ser grande.

A expectativa do quadrado médio do tratamento é uma variável aleatória e é dada por:

$$E(QM_e) = \sigma^2 + \sum_{i=1}^{I} n_i \alpha_i^2 \Big/ (I-1)$$

Quando os valores α_i são iguais a zero, o quadrado médio do tratamento estima σ^2. Quando algum efeito α_i difere de zero, o quadrado médio de tratamento é maior do que σ^2. Quando a hipótese da nulidade é verdadeira, QM_e e QM_d variam em torno de σ^2, resultando em um valor de F próximo de 1. Quando a hipótese da nulidade é rejeitada, QM_e tende ser maior do que σ^2 e a estatística F é maior do que 1. Assim, a hipótese de nulidade é rejeitada para valores da estatística F suficientemente grandes.

Em resumo, a estatística F refere-se à razão entre duas variâncias, podendo ser calculada de várias maneiras:

$$F = \frac{s_e^2}{s_d^2} = \frac{\left[\frac{\sum_{i=1}^{t} n_i (\bar{x}_i - \bar{x}_g)^2}{I-1}\right]}{\left[\frac{\sum_{i=1}^{t} (n_i - 1) s_i^2}{\sum_{i=1}^{t} (n_i - 1)}\right]}$$

ou

$$F = \frac{\left[\frac{n_1(\bar{x}_1 - \bar{x}_g)^2 + n_2(\bar{x}_2 - \bar{x}_g)^2 + \cdots + n_I(\bar{x}_I - \bar{x}_g)^2}{I-1}\right]}{\left[\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \cdots + (n_I-1)s_I^2}{N-I}\right]}$$

ou

$$F = \frac{\frac{SQ_{tratamento}}{GL_{tramento}}}{\frac{SQ_{erro}}{GL_{erro}}} = \frac{\frac{\left[\frac{\sum_i Y_i^2}{J} - \frac{(\sum_i \sum_j Y_{ij})^2}{IJ}\right]}{I-1}}{\frac{\left[\sum_i \sum_j Y_{ij}^2 - \frac{(\sum_i \sum_j Y_{ij})^2}{IJ}\right] - \left[\frac{\sum_i Y_i^2}{J} - \frac{(\sum_i \sum_j Y_{ij})^2}{IJ}\right]}{I(J-1)}}$$

ou

$$F = \frac{QM_{tratamento}}{QM_{erro}}$$

Em que:

s_e^2 é a variância entre as médias dos tratamentos (variação entre)

s_d^2 é a variância entre os valores de cada tratamento em relação a sua média (variação dentro)

n_i é o número de repetições para o *i*-ésimo tratamento

$\bar{x}_i$ é a média dos valores para o *i*-ésimo tratamento

$\bar{x}_g$ é a média geral de todos os valores obtidos no experimento

I é o número de tratamentos testados

J é o número de repetições

N é o número de valores

Y é o valor de uma determinada variável aletória mensurada no experimento

s_i^2 é a variância dos valores para o *i*-ésimo tratamento

A variação de s_e^2 depende das médias dos tratamentos, enquanto s_d^2 depende das variâncias dentro dos tratamentos (erro experimental), uma vez que não é afetada pelo efeito dos tratamentos.

SIMULAÇÃO PARA A ESTATÍSTICA F

Na função desenvolvida *FHipotese*, para representação do teste de hipótese de uma análise de variância, quando a barra na cor laranja sobrepõe a área verde a hipótese da nulidade é verdadeira; quando sobrepõe a área azul a hipótese da nulidade é falsa. O valor de α na Figura 13.1 refere-se à área ocupada pela estatística F para os tratamentos.

Comandos para obtenção da representação gráfica da ANOVA:

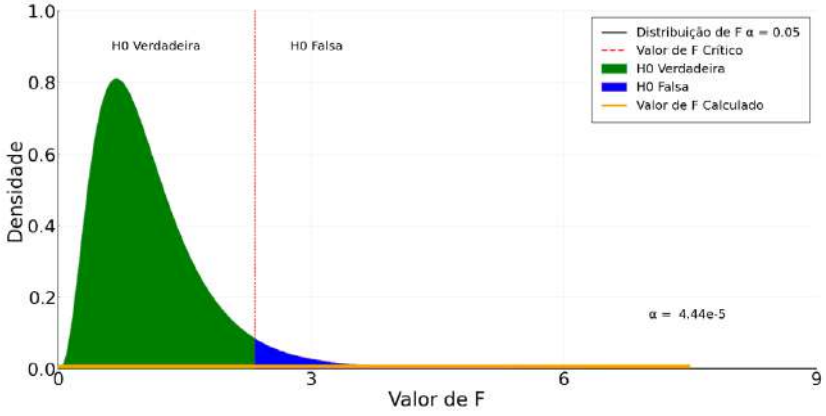
Arquivo: 25 – Função fhipotese para teste F na anova

```
>using Distributions, Plots; pyplot()
>QMTrat=15
>QMRes=2
>FTrat=QMTrat/QMRes
>GLTrat=8
>GLErro =25

>function FHipotese(FTrat,GLTrat,GLErro)
    xGrid = 0:0.01:(FTrat*1.2)
    Alpha=0.05
    plot!(xGrid, pdf.(FDist(GLTrat, GLErro),xGrid),c=:black,
xlims=(0, (FTrat*1.2)), ylims=(0,1),
label="Distribuição de F α = 0.05", dpi=300)
    ValorCritico = quantile(FDist(GLTrat, GLErro),(1-Alpha))
    plot!([ ValorCritico, ValorCritico],[0,1],c=:red, ls=:dash,
label="Valor de F Crítico", xlabel="Valor de F", ylabel="Densidade")
    alphaH0=0:0.0001:(1-Alpha)
    H0 = quantile(FDist(GLTrat, GLErro),alphaH0)
    DensidadeH0=pdf.(FDist(GLTrat, GLErro),H0)
    plot!(H0, DensidadeH0,fill=(0, :green), label="H0 Verdadeira")
    alphaH1=(1-Alpha):0.0001:1
    H1 = quantile(FDist(GLTrat, GLErro),alphaH1)
    DensidadeH1=pdf.(FDist(GLTrat, GLErro),H1)
    plot!(H1, DensidadeH1,fill=(0, :blue), label="H0 Falsa")
    plot!([0, FTrat],[0,0],c=:orange, lw=12, label="Valor de F Calculado")
    α = ccdf(FDist(GLTrat, GLErro),abs(FTrat))
    A=ValorCritico/2
    B=ValorCritico*1.1
    C=FTrat*1.1
    annotate!([(A, 0.9, text("H0 Verdadeira")),(B, 0.9, text("H0 Falsa")),
(C, 0.15, text("α =")), ((C+0.2), 0.1, α)])
end

#FHipotese(FTrat,GLTrat,GLErro)
>FHipotese(FTrat,GLTrat,GLErro)
```

FIGURA 13.1. Área ocupada pela estatística F para testar as médias dos tratamentos.



13.1 CONDICIONANTES PARA A ANÁLISE DE VARIÂNCIA

A aplicação do teste F em uma análise de variância requer o atendimento de algumas condições (condicionantes da ANOVA). O motivo para isso é assegurar a veracidade do teste, e/ou evitar a obtenção de conclusões no experimento que não se justificam, ou ainda, deixar de obter conclusões importantes.

Na análise de variância, quatro condições são observadas para aplicação do teste F:

1. Os efeitos principais devem ser aditivos.
2. Os erros de observações devem ser independentes.
3. Os erros de observações devem ser normalmente distribuídos.
4. As variâncias dos tratamentos devem ser homogêneas.

Será tratado aqui apenas a condição de homogeneidade das variâncias.

13.2 AVALIAÇÃO DA HOMOGENEIDADE DAS VARIÂNCIAS DOS TRATAMENTOS

As variâncias dos tratamentos não devem ser muito diferentes umas das outras. Quando as variâncias dos tratamentos resultam em valores numéricos próximos diz-se que a variância dos tratamentos é homogênea (possui homoscedasticidade). Para se avaliar isto estatisticamente pode-se aplicar o teste de Bartlett para testar as seguintes hipóteses:

$$\begin{cases} H_0: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_I^2 \\ H_1: \sigma_i^2 \neq \sigma_{i'}^2 \end{cases}$$
$$\alpha = 0,05$$

Em que i é diferente de i'.

Pela hipótese da nulidade parte-se do princípio de que as variâncias dos tratamentos são homogêneas, enquanto na hipótese alternativa não são homogêneas. Se as variâncias dos tratamentos forem homogêneas a análise de variância pode ser efetuada, considerando este critério. Caso contrário, transformações nos dados podem ser efetuadas com objetivo de assegurar variâncias de tratamentos homogêneas.

A estatística do teste de Bartlett é dada por:

$$\chi^2_{corr} = \frac{1}{C} \cdot \left\{ 2{,}3026 \left[\left(log\bar{S}^2 \cdot \sum GL_i \right) - \sum \left(GL \cdot logS_i^2 \right) \right] \right\}$$

$$C = 1 + \frac{1}{3(n-1)} \left(\sum \frac{1}{GL} \frac{1}{\sum GL} \right)$$

n = número de variâncias que estão sendo comparadas.

A hipótese de que as variâncias são homogêneas (H_0) é rejeitada quando o valor de χ^2 do teste de Bartlett for maior que a estatística $\chi^2_{\alpha,n-1}$ tabelada. Outra maneira de avaliar a significância da estatística χ^2 do teste de Bartlett é utilizar o valor de α obtido no teste, o qual é comparado com o valor de α estabelecido para o teste de hipótese.

Considere um experimento em que foram testados 3 tipos de substratos (denominados de A, B e C) na produção de mudas de *Eucalyptus grandis*. O experimento foi conduzindo utilizando-se o delineamento inteiramente casualizado e cada tratamento foi repetido 4 vezes. A variável aleatória mensurada foi a altura das mudas (em cm), tomada 90 dias após a emergência. Os dados estão disponibilizados no arquivo *dic*.

Comandos da função Bartlett para obtenção do teste de Bartlett:

Arquivo: 26 - Função bartlett para obtenção do teste de bartlett

```
>using DataFrames, CSV, Statistics, Distributions

>X=CSV.read("Z:/Alexandre/Julia/Exemplos/dic1.csv", DataFrame)
```

julia

```
julia> X=CSV.read("Z:/Alexandre/Julia/Exemplos/dic1.csv", DataFrame)
12×3 DataFrame
| Row | Trat   | Rep   | h     |
|     | String | Int64 | Int64 |
|-----|--------|-------|-------|
| 1   | A      | 1     | 25    |
| 2   | A      | 2     | 30    |
| 3   | A      | 3     | 27    |
| 4   | A      | 4     | 28    |
| 5   | B      | 1     | 40    |
| 6   | B      | 2     | 35    |
| 7   | B      | 3     | 37    |
| 8   | B      | 4     | 35    |
| 9   | C      | 1     | 16    |
| 10  | C      | 2     | 19    |
| 11  | C      | 3     | 18    |
| 12  | C      | 4     | 15    |
```

```
>function Bartlett(Tratamento,Repeticao,Variavel,Alpha)
    X=DataFrame()
    X.Trat=Tratamento
    X.Rep=Repeticao
    X.Y=Variavel
    X
    I=length(unique(X.Trat))
    J=length(unique(X.Rep))
    AT= by(X, :Trat) do df
    (m = mean(df.Y),ni=length(df.Y),gl=length(df.Y)-1, s² = var(df.Y))
    end
    GLs²=sum(AT.gl.*(log10.(AT.s²)))
    logs²bar=log10(mean(AT.s²))
    C=1+(1/(3*(I-1)))*((sum(AT.gl.^-1))-(1/sum(AT.gl)))
    X²c=(1/C)*2.3026*((logs²bar*sum(AT.gl))-GLs²)
    X²Crit = quantile(Chisq(I-1),1-Alpha)
    Alpha = ccdf(Chisq(I-1), X²c)
    a=DataFrame(Tratamento=AT.Trat, Média=AT.m, Variância=AT.s²)
    println("Tabela de média e variância: ", a)
    println("X²c: ", X²c)
    println("Alpha: ", Alpha)
    println("X² crítico: ", X²Crit)
end
#Bartlett(Tratamento,Repeticao,Variavel,Alpha)
>Bartlett(X.Trat,X.Rep,X.h,0.05)
```

```
julia> Bartlett(X.Trat,X.Rep,X.h,0.05)
Tabela de média e variância: 3×3 DataFrame
| Row | Tratamento | Média   | Variância |
|     | String     | Float64 | Float64   |
|-----|------------|---------|-----------|
| 1   | A          | 27.5    | 4.33333   |
| 2   | B          | 36.75   | 5.58333   |
| 3   | C          | 17.0    | 3.33333   |
X²c: 0.17260937018633532
Alpha: 0.9173146965235116
X² crítico: 5.99146454710798
```

O valor da estatística do teste de Bartlett (χ^2_{corr}) foi de 0,1726, enquanto o valor de χ^2 crítico é de 5,99 e o valor de α correspondente é de 0,92, revelando que há evidências para aceitar a hipótese da nulidade. Assim, pode-se concluir que as variâncias dos tratamentos são homegêneas, atendendo uma das condicionantes exigidas para a aplicação da análise de variância.

14 DELINEAMENTO INTEIRAMENTE CASUALIZADO

O delineamento inteiramente casualizado (DIC) é o mais simples dos delineamentos experimentais. A única fonte de variação controlada neste delineamento é a dos tratamentos, o que o caracteriza como experimentos de simples classificação.

No DIC apenas os princípios básicos de repetição e casualização estão presentes, pois este delineamento é apropriado para implantação quando as unidades experimentais são homogêneas. Isso implica considerar o controle local das unidades experimentais como implícito, ao mesmo tempo em que maximiza os graus de liberdade do resíduo (o que é desejável).

Nenhum tipo de restrição é imposto à casualização dos tratamentos nas parcelas. Seu arranjo é estruturado por meio da atribuição dos tratamentos às unidades experimentais de forma inteiramente casualizados (sorteados), o que origina a denominação do delineamento.

A função apresentada na sequência pode ser utilizada para que os tratamentos sejam aleatórios nas unidades experimentais (denominadas de "Row"). No exemplo, o experimento é definido por 3 tratamentos (A, B e C) e três repetições.

Comandos para casualizar os tratamentos em um DIC:

ARQUIVO: 27 - CASUALIZAR OS TRATAMENTOS EM DIC

```
>using DataFrames, CSV
>I=3 #Número de tratamentos
>J=3 #Número de repetições
#Obtenção de uma tabela com os tratamentos e repetições
>XDIC = DataFrame(Tratamento=repeat(Symbol.('A':'C'), inner=J),
Repeticao=repeat(1:J, outer=I))
```

julia

```
julia> XDIC = DataFrame(Tratamento=repeat(Symbol.('A':'C'), inner=J),Repeticao=repeat(1:J, outer=I))
9×2 DataFrame
```

Row	Tratamento	Repeticao
	Symbol	Int64
1	A	1
2	A	2
3	A	3
4	B	1
5	B	2
6	B	3
7	C	1
8	C	2
9	C	3

```
#Salvar a tabela de dados
>XDIC|> CSV.write("Z://Alexandre//Julia//Exemplos//ALEATORIZAR.csv")

#Casualização dos tratamentos
>using Random
>AleatDIC(X) = X[shuffle(1:end), :]
>AleatDIC(XDIC)
```

```
julia> AleatDIC(XDIC)
9×2 DataFrame
 Row | Tratamento | Repeticao
     | Symbol     | Int64
-----|------------|----------
   1 | B          |        1
   2 | B          |        3
   3 | A          |        3
   4 | B          |        2
   5 | A          |        1
   6 | C          |        2
   7 | A          |        2
   8 | C          |        1
   9 | C          |        3
```

Algumas características do DIC são:

» É um delineamento flexível, uma vez que o número de tratamentos e de repetições depende apenas do número de parcelas homogêneas disponíveis;

» O número de repetições pode ser diferente de um tratamento para outro, embora não seja o ideal; a análise estatística é simples mesmo nesta situação;

» O número de graus de liberdade para o resíduo é o maior dentre todos os delineamentos;

» O delineamento exige homogeneidade total das condições experimentais e quando isto não acontece pode conduzir a uma estimativa de variância residual grande, pois todas as variações, exceto as devidas aos tratamentos, são consideradas como variações estritamente ao acaso (erro experimental).

Na área florestal este tipo de delineamento é frequentemente utilizado em ensaios efetuados em laboratórios, estufas e viveiros. Na maioria dos experimentos conduzidos nestes locais, devido à característica dos experimentos, as unidades experimentais são homogêneas.

O modelo linear aditivo do delineamento inteiramente casualizado é dado por:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

Em que:

y_{ij} é a observação do i-ésimo tratamento na j-ésima unidade experimental;

μ é o efeito constante (usualmente corresponde à média geral) – constante inerente a todas as observações;

τ_i é o efeito do i-ésimo tratamento – efeito propiciado pelo tratamento i (desvio em relação a μ, decorrente da ação);

ε_{ij} é o erro associado ao i-ésimo tratamento na j-ésima unidade experimental - efeito aleatório na unidade experimental (erro experimental).

O esquema de análise de variância para o delineamento inteiramente casualizado está apresentado na Tabela 14.1.

TABELA 14.1. Esquema para análise de variância para o delineamento inteiramente casualizado.

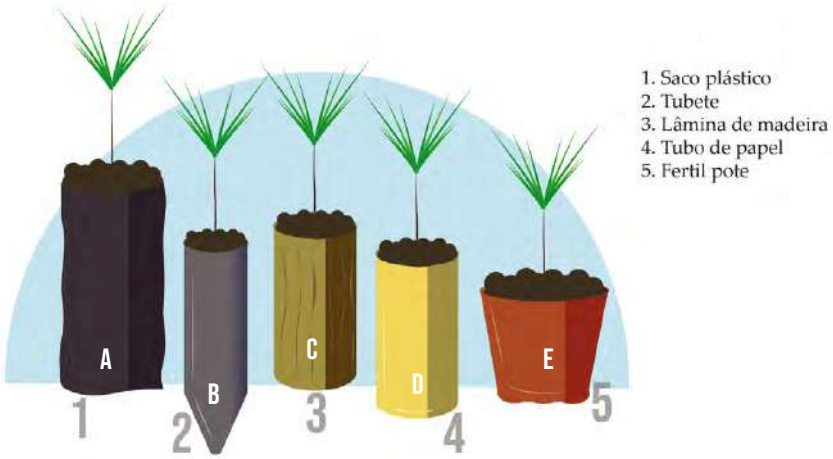
Fonte de variação	Graus de Liberdade	Soma de Quadrados	Quadrados Médio	F
Tratamentos (t)	$I - 1$	$SQ_t = \frac{\sum_i Y_i^2}{J} - FC$	$QM_t = \frac{SQ_t}{GL_t}$	$\frac{QM_t}{QM_e}$
Resíduo (e)	$I \cdot (J - 1)$	$SQ_e = SQ_{Total} - SQ_t$	$QM_e = \frac{SQ_e}{GL_e}$	
Total	$I \cdot J - 1$	$SQ_{total} = \sum_i \sum_j Y_{ij}^2 - FC$		

$FC = \frac{\left(\sum_i \sum_j Y_{ij}\right)^2}{IJ}$, conhecido como fator de correção para a média.

I é o número de tratamentos.
J é o número de repetições.

Um experimento foi conduzido para avaliar o efeito de recipientes no crescimento de mudas de *Pinus taeda*. Assim, foram testados os seguintes tratamentos: saco plástico (1-A), tubete (2-B), lâmina de madeira (3-C), tubo de papel (4-D) e fértil pote (5-E) (Figura 14.1).

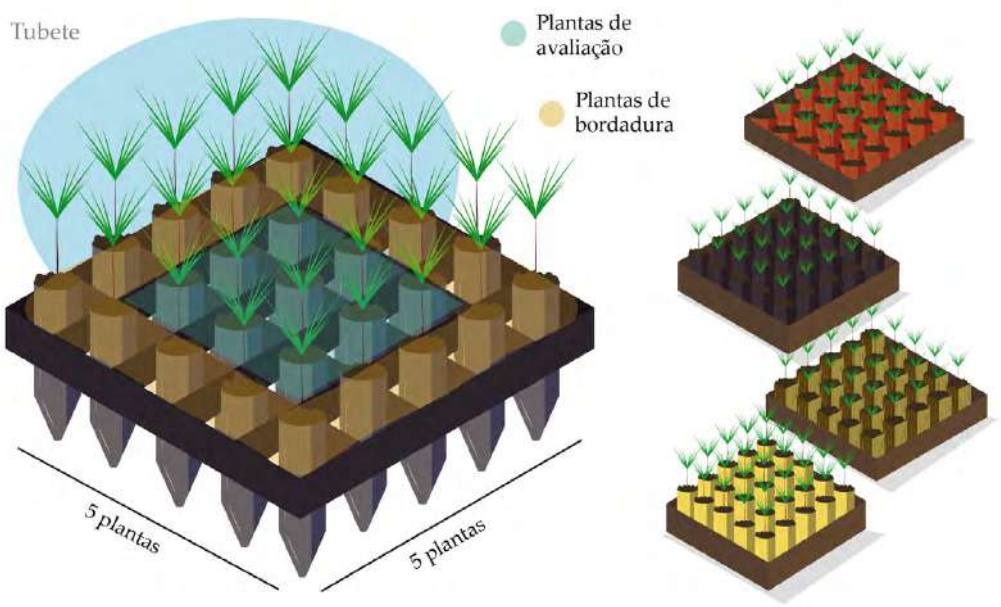
FIGURA 14.1. Recipientes utilizados como tratamento para a avaliação do crescimento em altura de mudas de *Pinus taeda*.



O experimento foi conduzido em casa de vegetação. A unidade experimental foi definida como: mudas de *Pinus taeda* com aproximadamente 3 cm de altura e substrato recomendado para a espécie oriundo do mesmo lote. Tendo em vista que as unidades experimentais são homogêneas e que o ambiente onde elas seriam cultivadas também pode ser considerado como homogêneo (casa de vegetação), foi adotado o DIC.

A unidade experimental foi composta por 25 plantas (Figura 14.2). Uma linha de bordadura foi definida totalizando, desta maneira, apenas 9 plantas para avaliação.

FIGURA 14.2. Unidade experimental com os tratamentos testados.



No experimento foram utilizadas 6 repetições. A alocação dos tratamentos nas unidades experimentais foi efetuada de maneira inteiramente casualizado como está apresentado na Figura 14.3.

Aos 120 dias após a instalação do experimento (Figura 14.4), foi mensurada a altura das mudas (em cm) com uma régua. Os dados obtidos estão apresentados na Tabela 14.2.

FIGURA 14.3. Experimento para avaliar o efeito de diferentes tipos de recipientes no crescimento em altura de mudas de *Pinus taeda* em delineamento inteiramente casualizado.

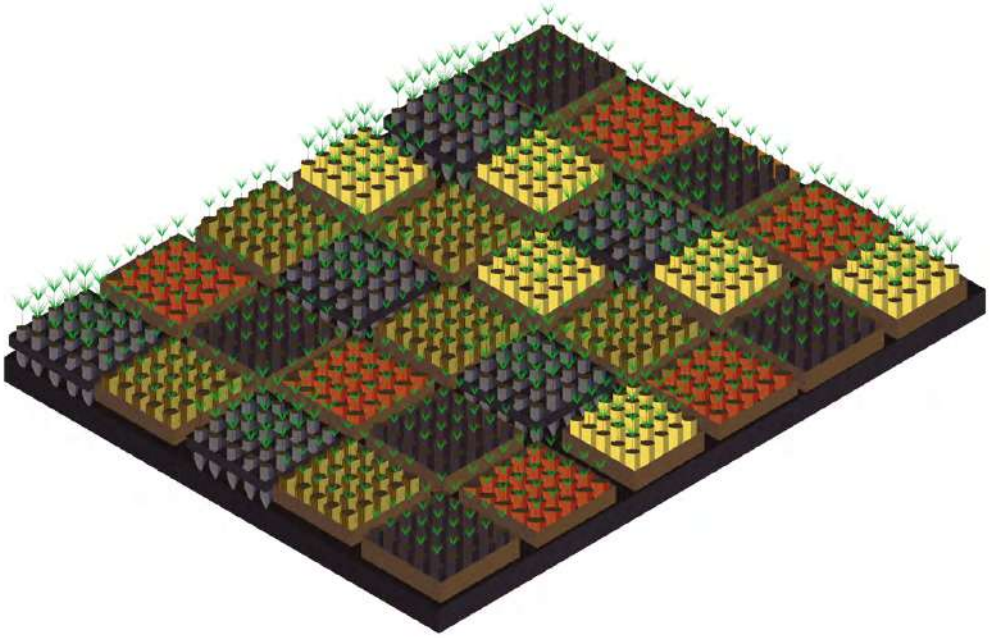


FIGURA 14.4 Experimento para avaliar o efeito de tipos de recipientes no crescimento em altura de mudas de *Pinus taeda* em delineamento inteiramente casualizado, aos 120 dias após o plantio.

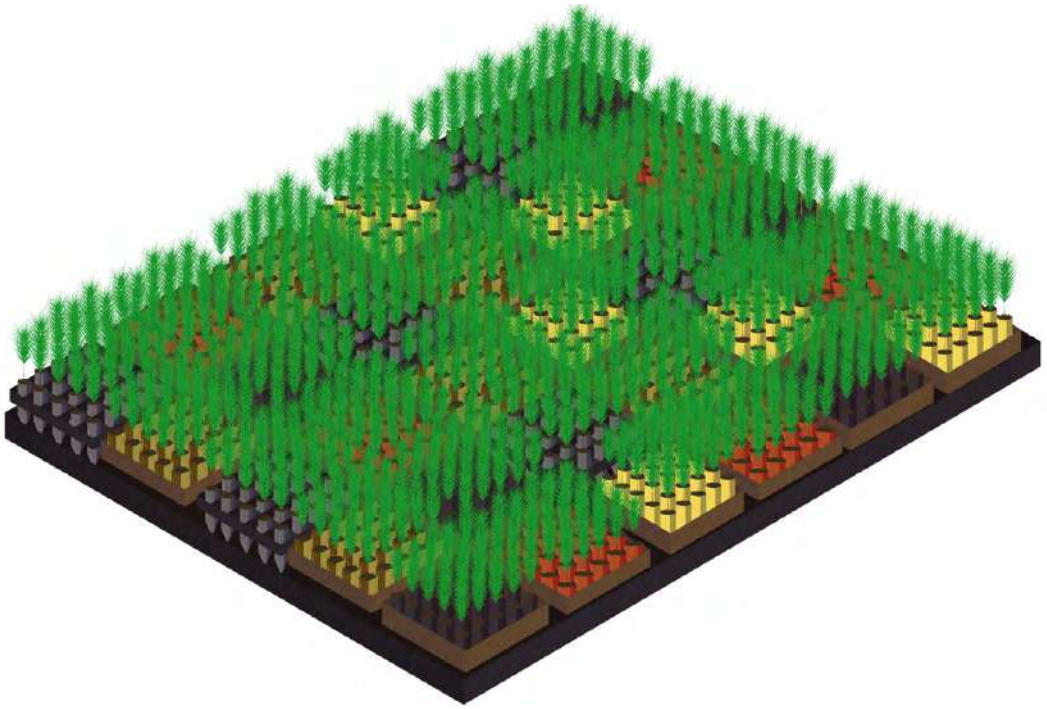


TABELA 14.2. Dados de altura média de mudas de *Pinus taeda* (em cm) obtidas em um experimento para avaliar o efeito de recipientes no seu crescimento em delineamento inteiramente casualizado (continua).

	Tratamento	Repetição	Média das alturas
A	Saco plástico	1	9,0
	Saco plástico	2	9,3
	Saco plástico	3	9,2
	Saco plástico	4	9,1
	Saco plástico	5	8,9
	Saco plástico	6	9,3
B	Tubete	1	10,4
	Tubete	2	10,1
	Tubete	3	10,4
	Tubete	4	10,7
	Tubete	5	10,6
	Tubete	6	10,2
C	Lâmina de madeira	1	9,3
	Lâmina de madeira	2	8,4
	Lâmina de madeira	3	8,9
	Lâmina de madeira	4	9,1
	Lâmina de madeira	5	9,4
	Lâmina de madeira	6	9,4
D	Tubo de papel	1	9,2
	Tubo de papel	2	9,1
	Tubo de papel	3	9,6

TABELA 14.2. Dados de altura média de mudas de *Pinus taeda* (em cm) obtidas em um experimento para avaliar o efeito de recipientes no seu crescimento em delineamento inteiramente casualizado (continuação).

	Tratamento	Repetição	Média das alturas
D	Tubo de papel	4	9,4
	Tubo de papel	5	9,6
	Tubo de papel	6	9,1
E	Fertil pote	1	10,2
	Fertil pote	2	9,8
	Fertil pote	3	10,1
	Fertil pote	4	10,4
	Fertil pote	5	10,1
	Fertil pote	6	10,3

Os dados foram submetidos ao teste de homogeneidade de variâncias de Bartlett. O valor de χ^2 corrigido foi de 4,08, sendo não significante ao nível de 0,05. Assim, foi evidenciado que as variâncias dos tratamentos são homogêneas.

Os dados experimentais, na escala original, foram submetidos à análise de variância e os resultados obtidos estão apresentados na Tabela 14.3.

TABELA 14.3. Análise de variância aplicada para avaliar os efeitos de tipos de recipiente sobre o crescimento da altura (em cm) de mudas de *Pinus taeda*.

Fonte de variação	Graus de Liberdade	Soma de Quadrado	Quadrado Médio	F calculado	F tabelado (α = 5%)	P-valor (%)
Tratamento	4	8,978	2,2445	34,4248**	2,76	0,0000001
Erro Exp.	25	1,63	0,0652			
Total	29	10,608				

Em que: Erro Exp. é o erro experimental

Existe uma notação convencional para representar os resultados obtidos em qualquer teste de hipótese. Se a hipótese da nulidade for aceita, esta é representada por "ns", indicando que o resultado foi não significante. Se a hipótese da nulidade for rejeitada ao nível de 0,05, esta é representada por "*" e se for rejeitada ao nível de 0,01 por "**".

Os cálculos foram efetuados em Julia e estão apresentados abaixo. Os tratamentos foram denominados de A, B, C, D e E, respectivamente, conforme está apresentado na Tabela 14.2. A variável tratamento foi denominada de "Trat", a repetição de "Rep" e a altura das mudas de "Y". O arquivo de dados utilizado foi denominado de *DIC2*.

Comandos para obtenção da análise de variância de um DIC:

Arquivo: 28 - Função dic para anova em dic

```
>using DataFrames, CSV, Statistics, Distributions

>X=CSV.read("Z:/Alexandre/Julia/Exemplos/dic2.csv", DataFrame)
```

```
julia> X=CSV.read("Z:/Alexandre/Julia/Exemplos/dic2.csv", DataFrame)
30×3 DataFrame
| Row | Trat   | Rep   | Y       |
|     | String | Int64 | Float64 |
|-----|--------|-------|---------|
| 1   | A      | 1     | 9.0     |
| 2   | A      | 2     | 9.3     |
| 3   | A      | 3     | 9.2     |
| 4   | A      | 4     | 9.1     |
⋮
| 26  | E      | 2     | 9.8     |
| 27  | E      | 3     | 10.1    |
| 28  | E      | 4     | 10.4    |
| 29  | E      | 5     | 10.1    |
| 30  | E      | 6     | 10.3    |
```

```
>function DIC(Tratamento, Repeticao,Variavel)
    X=DataFrame()
    X.Trat=Tratamento
    X.Rep=Repeticao
    X.Y=Variavel
    X
    I=length(unique(X.Trat))
    J=length(unique(X.Rep))
    GLTrat=I-1
    GLResiduo=I*(J-1)
    GLTotal=I*J-1
    FC=((sum(X.Y))^2)/(I*J)
    AT= by(X, :Trat) do df
    (m = mean(df.Y), s² = var(df.Y), total= sum(df.Y), ss=sum((df.Y).^2))
    end
    SQTrat=((sum(AT.total.^2))/J)-FC
    SQTotal=(sum(X.Y.^2))-FC
    SQResiduo=SQTotal-SQTrat
    QMTrat=SQTrat/GLTrat
    QMResiduo=SQResiduo/GLResiduo
    FTrat=QMTrat/QMResiduo
    AlphaTrat= ccdf(FDist(GLTrat, GLResiduo),abs(FTrat))
    R²=SQTrat/SQTotal*100
    CV=sqrt(QMResiduo)/mean(X.Y)*100
    println("GLTrat: ", GLTrat)
    println("GLResíduo: ", GLResiduo)
    println("GLTotal: ", GLTotal)
    println("SQTrat: ", SQTrat)
    println("SQResíduo: ", SQResiduo)
    println("SQTotal: ", SQTotal)
    println("QMTrat: ", QMTrat)
    println("QMResíduo: ", QMResiduo)
    println("FTrat: ", FTrat)
    println("Alpha: ", AlphaTrat)
```

```
    println("R²(%): ", R²)
    println("CV(%): ", CV)
end

#DIC(Tratamento, Repeticao,Variavel)
>DIC(X.Trat,X.Rep,X.Y)
GLTrat: 4
GLResíduo: 25
GLTotal: 29
SQTrat: 8.978
SQResíduo: 1.630
SQTotal: 10.608
QMTrat: 2.2445
QMResíduo: 0.065
FTrat: 34.423
Alpha: 7.863e-10
R²(%): 84.63
CV(%): 2.65
```

O teste de F da análise de variância foi significante (**), indicando haver diferenças significantes em pelo menos uma média entre os tratamentos. O teste de Tukey foi aplicado para a comparação das médias entre os tratamentos (que será detalhado oportunamente) e está apresentado na Tabela 14.4. Este teste foi efetuado utilizando-se a função "Tukey", apresentada na seção 19 para um nível de significância de 0,05 (α). A diferença mínima para atestar significância foi de 0,43 cm. O teste de Tukey revelou que as médias dos tratamentos B e E não diferem entre si e são superiores às demais médias de tratamentos testadas.

Comandos para obtenção do teste de Tukey para comparação de médias:

Arquivo: 28 - Função DIC para ANOVA em DIC

```
>using DataFrames, Statistics, PyPlot, Distributions, StatsFuns
#Tukey(Tratamento,Repeticao,Variavel,GLResiduo,QMResiduo,Alpha)
>Tukey(X.Trat,X.Rep,X.Y,25,0.0652,0.05)
```

julia

```
julia> Tukey(X.Trat,X.Rep,X.Y,25,0.0652,0.05)
Tabela de média: 5×3 DataFrame
| Row | Tratamento | Média   | Variância |
|     | String     | Float64 | Float64   |
|-----|------------|---------|-----------|
| 1   | B          | 10.4    | 0.052     |
| 2   | E          | 10.15   | 0.043     |
| 3   | D          | 9.33333 | 0.0546667 |
| 4   | A          | 9.13333 | 0.0266667 |
| 5   | C          | 9.08333 | 0.149667  |
```

```
Diferença Mínimca Significativa: 0.43296029654054147
Tabela DMS: 5×6 DataFrame
| Row | Tratamento | B         | E         | D         | A         | C         |
|     | String     | Float64   | Float64   | Float64   | Float64   | Float64   |
|-----|------------|-----------|-----------|-----------|-----------|-----------|
| 1   | B          | 0.0       | 0.25      | 1.06667   | 1.26667   | 1.31667   |
| 2   | E          | -0.25     | 0.0       | 0.816667  | 1.01667   | 1.06667   |
| 3   | D          | -1.06667  | -0.816667 | 0.0       | 0.2       | 0.25      |
| 4   | A          | -1.26667  | -1.01667  | -0.2      | 0.0       | 0.05      |
| 5   | C          | -1.31667  | -1.06667  | -0.25     | -0.05     | 0.0       |
Tabela de hipóteses DMS: 5×6 DataFrame
| Row | Tratamento | B      | E      | D      | A      | C      |
|     | String     | String | String | String | String | String |
|-----|------------|--------|--------|--------|--------|--------|
| 1   | B          | ns     | ns     | *      | *      | *      |
| 2   | E          | ns     | ns     | *      | *      | *      |
| 3   | D          | *      | *      | ns     | ns     | ns     |
| 4   | A          | *      | *      | ns     | ns     | ns     |
| 5   | C          | *      | *      | ns     | ns     | ns     |
PyObject Text(0, 0.5, 'Altura (cm)')
```

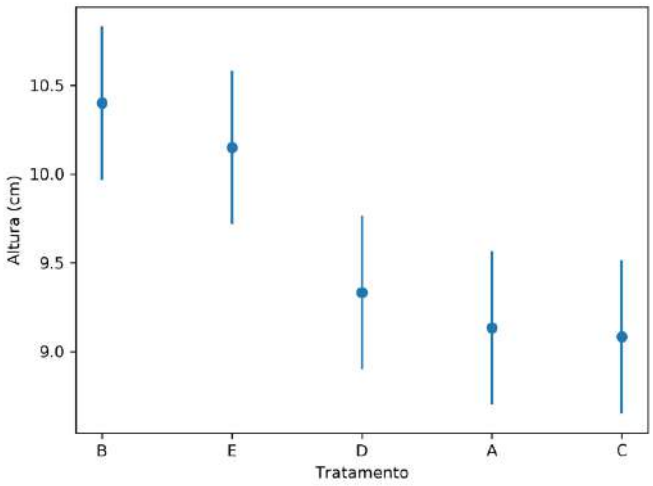


TABELA 14.4. Sumarização dos resultados obtidos pelo teste de Tukey, na comparação de médias dos diferentes tipos de recipientes sobre o crescimento em altura (em cm) de mudas de *Pinus taeda*.

Tratamento	Média	Grupo
Tubete	10,40	a
Fertil pote	10,15	a
Tubo de papel	9,33	b
Saco plástico	9,13	b
Lâmina de madeira	9,08	b

Assim, por meio do teste de Tukey foi identificado que os melhores tratamentos, concernente ao efeito sobre o crescimento em altura das mudas de *Pinus taeda*, foram o tubete e o fértil pote. O viveirista pode escolher entre um deles.

Um resumo geral sobre o delineamento inteiramente casualizado está apresentado na Figura 14.5.

FIGURA 14.5. Resumo do delineamento inteiramente casualizado.

DELINEAMENTO INTEIRAMENTE CASUALIZADO

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

O Delineamento inteiramente casualizado é o mais simples dos delineamentos experimentais.

Somente um fator varia enquanto todos os demais são mantidos constantes. A atribuição dos tratamento às unidades experimentais ocorre de forma totalmente aleatória.

Os princípios experimentais são: a repetição e a casualização, nenhum tipo de restrição é imposta à casualização das parcelas no tratamento.

VANTAGENS

- É um delineamento flexível
- Tem análise estatística simples
- O número de graus de liberdade para os resíduos é sempre o maior possível

DESVANTAGENS

- Exige a homogeneidade total das condições experimentais
- Pode conduzir a uma estimativa de variância residual muito alta
- As variações ocorrem ao acaso

julia

```
function DIC (Tratamento, Repeticao,Variavel)
```

15 DELINEAMENTO EM BLOCOS AO ACASO

O delineamento em blocos ao acaso (DBA) é o mais utilizado em pesquisas na Engenharia Florestal. Neste delineamento as repetições são denominadas de blocos e um bloco consiste em um agrupamento de unidades experimentais o mais homogêneo que for possível, de tal modo a minimizar as diferenças entre elas. A variabilidade entre blocos é menos importante. Quando as unidades experimentais não são homogêneas, devido a um fator previamente conhecido, por exemplo, tipo de solo, umidade, fertilidade, gradiente de luz e temperatura, a solução para a redução do erro experimental é aglutinar as unidades em grupos homogêneos, comumente chamados de blocos.

Um bloco conterá todos os tratamentos do experimento, portanto, cada novo bloco instalado significa repetir todos os tratamentos do experimento e estes representam o número de repetições de cada tratamento na estrutura experimental.

A função apresentada na sequência pode ser utilizada para aleatorizar (sortear) os tratamentos em um delineamento de blocos ao acaso. No exemplo, o experimento é definido por 3 tratamentos (A, B e C) e três repetições (blocos).

Comandos para casualizar os blocos:

ARQUIVO: 29 - CASUALIZAR OS TRATAMENTOS EM DBA

```
>using DataFrames, CSV
>I=3; #Número de tratamentos
>J=3; #Número de blocos
#Obtenção de uma tabela com os tratamentos e repetições
>XDBA = DataFrame(Tratamento=repeat(Symbol.('A':'C'), inner=J),Bloco=repeat(1:J,
outer=I))
#Salvar a tabela de dados
>XDBA|> CSV.write("Z://Alexandre//Julia//Exemplos//ALEATDBA.csv")
```

julia

```
julia> XDBA = DataFrame(Tratamento=repeat(Symbol.('A':'C'), inner=J),Bloco=repeat(1:J, outer=I))
9×2 DataFrame
 Row │ Tratamento  Bloco
     │ Symbol      Int64
─────┼──────────────────
   1 │ A               1
   2 │ A               2
   3 │ A               3
   4 │ B               1
   5 │ B               2
   6 │ B               3
   7 │ C               1
   8 │ C               2
   9 │ C               3
```

```
#Casualização dos tratamentos
>using DataFrames, StatsBase

>function AleatDBA(Tratamento,Bloco)
    df = DataFrame()
    df.Trat = Tratamento
    df.Bloco=Bloco
    df
    n = size(df, 1)
    ord = sample(1:n, n, replace = false)
    A=df[ord,:]
    X=sort(A, :Bloco)
end
#AleatDBA(Tratamento,Bloco)
>AleatDBA(XDBA.Tratamento,XDBA.Bloco)
```

```
julia> AleatDBA(XDBA.Tratamento,XDBA.Bloco)
9×2 DataFrame
| Row | Trat   | Bloco |
|     | Symbol | Int64 |
|-----|--------|-------|
| 1   | C      | 1     |
| 2   | B      | 1     |
| 3   | A      | 1     |
| 4   | C      | 2     |
| 5   | B      | 2     |
| 6   | A      | 2     |
| 7   | A      | 3     |
| 8   | C      | 3     |
| 9   | B      | 3     |
```

Outra forma para aleatorizar os tratamentos em cada bloco pode ser efetuada conforme apresentado na sequência.

Comandos para casualizar os tratamentos nos blocos.

Arquivo: 29 - Casualizar os tratamentos em DBA

```
>using Random
>NTratamentos=5 #Número de tratamentos
>NBlocos=4 #Número de blocos
>Tratamentos=collect(1: NTratamentos)
>function adba(x)
    Desenho=shuffle(x)
end

>desenho = for i=1: NBlocos
    println(a(Tratamentos))
end
[4, 5, 2, 1, 3]
[4, 5, 3, 2, 1]
[3, 5, 2, 4, 1]
[5, 3, 1, 2, 4]
```

Em cada vetor (cada linha) está à disposição dos tratamentos em um bloco. Caso prefira codificar os tratamentos por letra utilize:

```
>Tratamentos=Symbol.('A':'E')
```

As principais características do DBA são:

- Controlar as diferenças que ocorrem nas condições experimentais, por meio da utilização de blocos;
- Minimizar as diferenças entre as unidades experimentais que constituem um bloco;
- As fontes de variação controladas correspondem aos tratamentos e uma fonte de variação nas unidades experimentais (blocos); as demais fontes de variação são consideradas como variações aleatórias;
- Devido ao controle das unidades experimentais (controle local), ocorre redução do número de graus de liberdade;
- Especialmente em experimentos florestais, devido as unidades experimentais serem grandes, a exigência de homogeneidade das parcelas dentro de cada bloco limita o número de tratamentos, não podendo, portanto, ser muito elevado.
- Os 3 princípios da experimentação constituem o delineamento: repetição, casualização e controle local.
- A atribuição dos tratamentos às unidades experimentais é efetuada de forma aleatória dentro de cada bloco.

O modelo linear aditivo do delineamento DBA é dado por:

$$Y_{ij}=\mu+\tau_i+\beta_j+\epsilon_{ij}$$

Em que:

Y_{ij} é o valor de uma observação ao *i*-ésimo tratamento do *j*-ésimo bloco;

μ é o efeito constante (usualmente corresponde à média geral) – inerente a todas as observações;

τ_i é o efeito do *i*-ésimo tratamento – efeito propiciado pelo tratamento *i* (desvio em relação a μ, decorrente de seus efeitos);

β_j é o efeito do *j*-ésimo bloco;

ϵ_{ij} é o erro experimental do *i*-ésimo tratamento associado no do *j*-ésimo bloco – efeito aleatório na unidade experimental (erro).

O esquema da análise de variância para DBA está apresentado na Tabela 15.1.

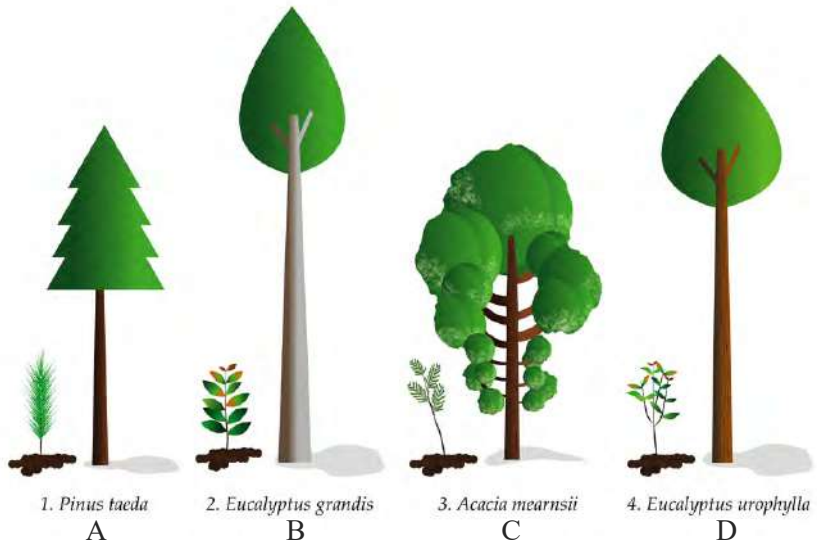
TABELA 15.1. Esquema para análise de variância para o delineamento em blocos ao acaso.

Causas de variação	Graus de liberdade	Soma de quadrados	Quadrados médios	F calculado
Blocos (b)	$J - 1$	$SQ_b = \frac{\sum_j Y_j^2}{I} - FC$	$QM_b = \frac{SQ_b}{GL_b}$	$\frac{QM_b}{QM_e}$
Tratamentos (t)	$I - 1$	$SQ_t = \frac{\sum_i Y_i^2}{J} - FC$	$QM_t = \frac{SQ_t}{GL_t}$	$\frac{QM_t}{QM_e}$
Resíduo (e)	$(I - 1) \cdot (J - 1)$	$SQ_e = SQ_{Total} - SQ_b - SQ_t$	$QM_e = \frac{SQ_e}{GL_e}$	
Total	$I \cdot J - 1$	$SQ_{total} = \sum_i \sum_j Y_{ij}^2 - FC$		

$FC = \frac{\left(\sum_i \sum_j Y_{ij}\right)^2}{IJ}$, fator de correção para a média.

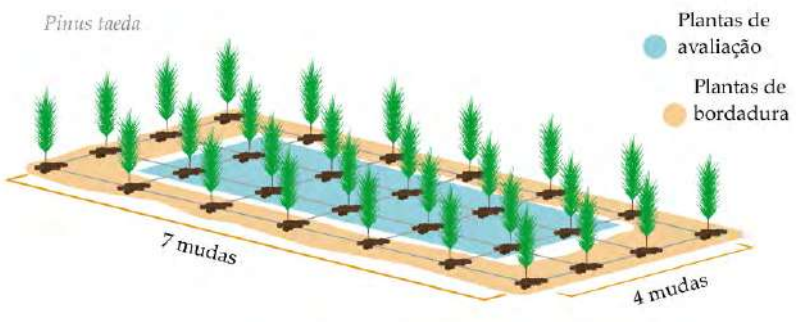
Um experimento foi conduzido para avaliar a produção em volume de madeira do fuste de diferentes espécies florestais. As espécies testadas foram: (1-A) *Pinus taeda*, (2-B) *Eucalyptus grandis*, (3-C) *Acacia mearnsii* e (4-D) *Eucalyptus urophylla* (Figura 15.1).

FIGURA 15.1. Espécies (mudas e árvores com 8 anos) utilizadas como tratamento para avaliar a produção em volume de madeira do fuste.



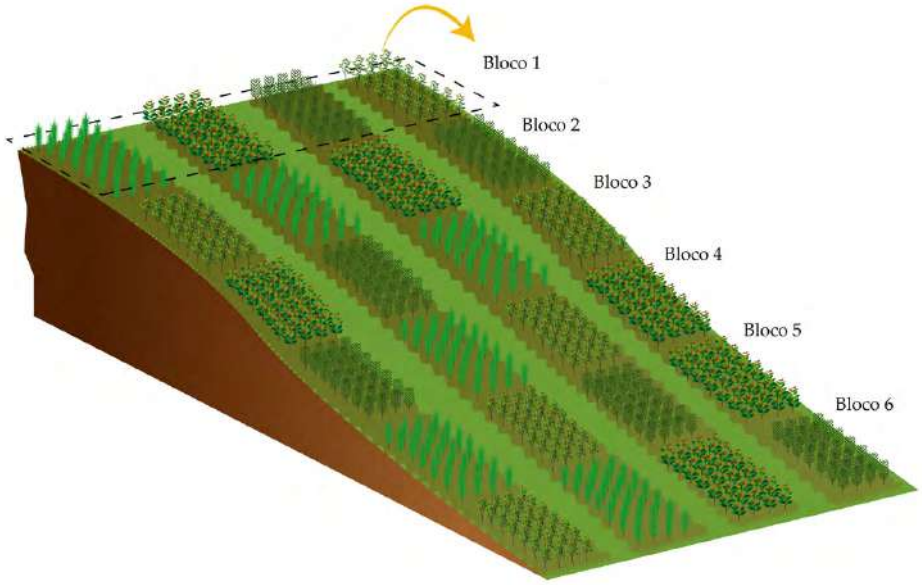
A unidade experimental foi definida com 28 plantas. O plantio das mudas das espécies (com aproximadamente 30 cm) foi efetuado em espaçamento de 3m x 2m (Figura 15.2). Uma linha de bordadura foi definida.

FIGURA 15.2. Unidade experimental das diferentes espécies utilizadas para avaliar a produção em volume de madeira do fuste



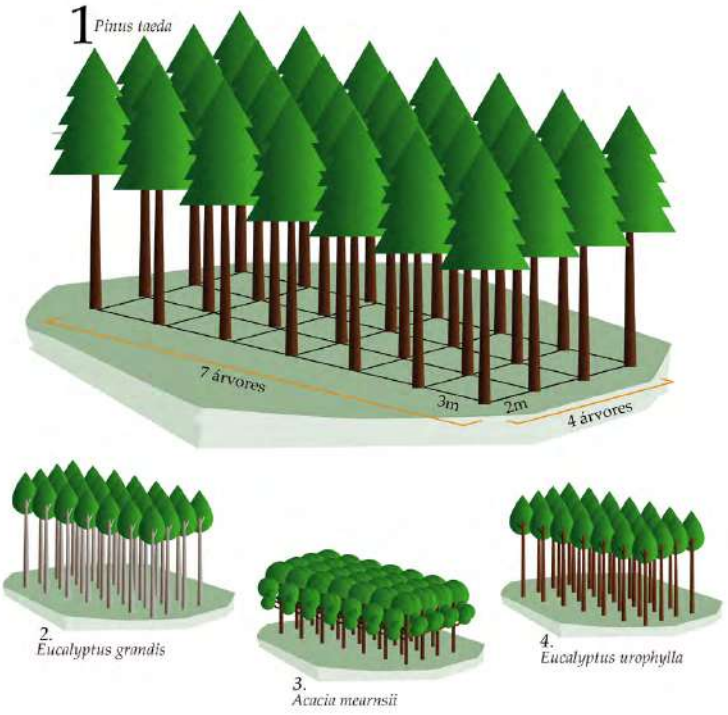
O experimento foi instalado em campo em uma área com declividade de aproximadamente 30%. As unidades experimentais, após uma avaliação, não seriam alocadas em condições homogêneas devido às variações nos diferentes patamares de relevo no solo. Assim, para controlar esta fonte de variação, foi utilizado o delineamento em blocos ao acaso. Seis blocos foram instalados (Figura 15.3).

FIGURA 15.3. Experimento para avaliar a produção em volume de madeira do fuste, de diferentes espécies florestais, em delineamento com blocos ao acaso.



Oito anos após a instalação do experimento (excluindo-se as árvores de bordadura) foi mensurado o volume de madeira (Figuras 15.4 - 15.5). Todas as árvores na parcela foram cortadas e cubadas para obtenção do volume. Estes valores foram somados, obtendo-se o estoque por unidade de área, o qual foi extrapolado para hectare.

FIGURA 15.4. Unidade experimental das diferentes espécies utilizadas para avaliar a produção em volume de madeira do fuste oito anos após a instalação do experimento.

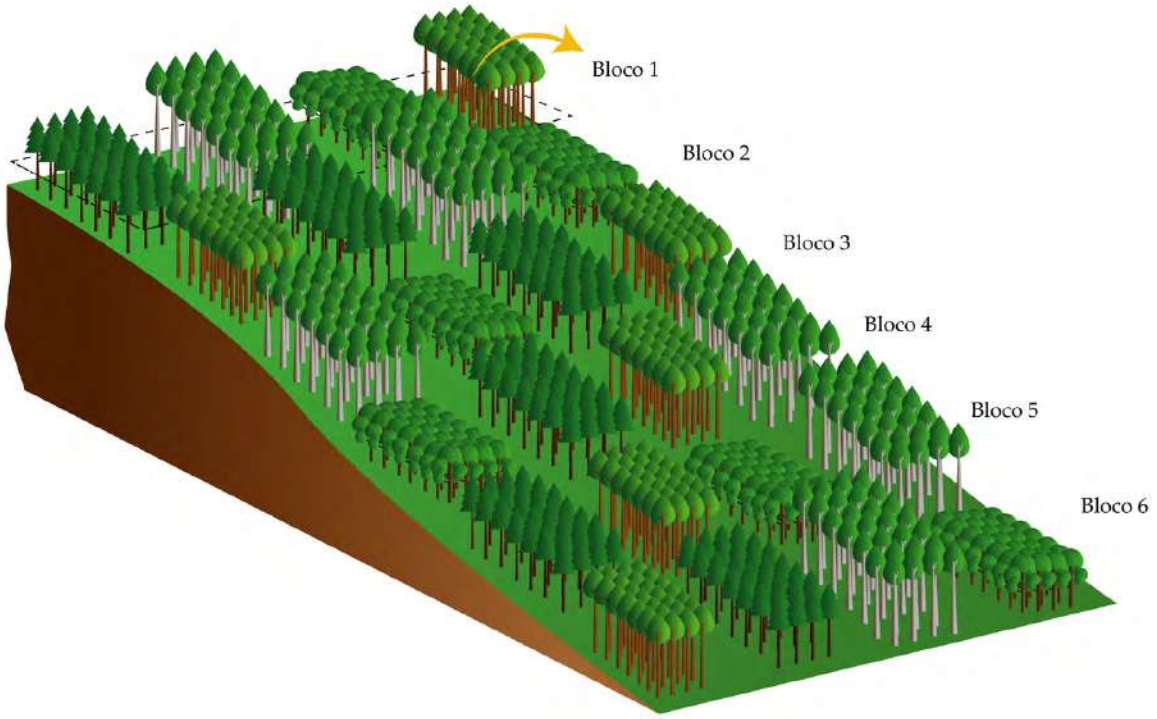


A variável resposta mensurada foi a produção de volume de madeira em $m^3.ha^{-1}$. Os dados obtidos estão apresentados na Tabela 15.2.

TABELA 15.2. Dados de volume de madeira (em $m^3.ha^{-1}$) obtidos em um experimento para avaliar a produção de espécies florestais em delineamento com blocos ao acaso.

Tratamento		Bloco	Volume ($m^3.ha^{-1}$)
A	*Eucalyptus grandis*	1	260
	Eucalyptus grandis	2	265
	Eucalyptus grandis	3	273
	Eucalyptus grandis	4	275
	Eucalyptus grandis	5	280
	Eucalyptus grandis	6	280
B	*Pinus taeda*	1	239
	Pinus taeda	2	245
	Pinus taeda	3	247
	Pinus taeda	4	256
	Pinus taeda	5	260
	Pinus taeda	6	272
C	*Acacia mearnsii*	1	140
	Acacia mearnsii	2	143
	Acacia mearnsii	3	148
	Acacia mearnsii	4	150
	Acacia mearnsii	5	154
	Acacia mearnsii	6	162
D	*Eucalyptus urophylla*	1	230
	Eucalyptus urophylla	2	238
	Eucalyptus urophylla	3	245
	Eucalyptus urophylla	4	254
	Eucalyptus urophylla	5	257
	Eucalyptus urophylla	6	262

FIGURA 15.5. Experimento em campo para avaliar a produção em volume de madeira do fuste de diferentes espécies florestais em delineamento com blocos ao acaso oito anos após a instalação.



Os dados foram submetidos ao teste de Bartlett para avaliar a homogeneidade de variâncias. O valor de χ^2 corrigido foi de 1,52, sendo não significante ao nível de 0,05 de significância. Assim, pode-se concluir que as variâncias são homogêneas.

Os dados experimentais, na escala original, foram submetidos à análise de variância, cujos resultados estão resumidos na Tabela 15.3. Os cálculos foram efetuados utilizando a função *DBA*. Os tratamentos foram denominados de A, B, C e D, respectivamente, conforme apresentados na Tabela 15.2. A variável tratamento foi denominada de "Trat", bloco de "Bloco" e a produção em volume de madeira de "Y".

Comandos utilizados para obtenção da ANOVA de um delineamento em blocos ao acaso:

Arquivo: 30 - Função dba para anova em dba

```
>using DataFrames, CSV, Statistics, Distributions

>X=CSV.read("Z:/Alexandre/Julia/Exemplos/dba2.csv", DataFrame)
```

```
julia> X=CSV.read("Z:/Alexandre/Julia/Exemplos/dba2.csv", DataFrame)
24×3 DataFrame
| Row | Trat   | Bloco | Volume |
|     | String | Int64 | Int64  |
|-----|--------|-------|--------|
| 1   | A      | 1     | 260    |
| 2   | A      | 2     | 265    |
| 3   | A      | 3     | 273    |
| 4   | A      | 4     | 275    |
| 5   | A      | 5     | 280    |
⋮
| 19  | D      | 1     | 230    |
| 20  | D      | 2     | 238    |
| 21  | D      | 3     | 245    |
| 22  | D      | 4     | 254    |
| 23  | D      | 5     | 257    |
| 24  | D      | 6     | 262    |
```

```
>function DBA(Tratamento,Bloco,Variavel)
    X=DataFrame()
    X.Trat=Tratamento
    X.Bloco=Bloco
    X.Y=Variavel
    X
    I=length(unique(X.Trat))
    J=length(unique(X.Bloco))
    GLTrat=I-1
    GLBlocos=J-1
    GLResiduo=(I-1)*(J-1)
    GLTotal=I*J-1
    FC=((sum(X.Y))^2)/(I*J)
    AB= by(X, :Bloco) do df
    (m = mean(df.Y), s² = var(df.Y), total= sum(df.Y), ss=sum((df.Y).^2))
    end
    SQBlocos=((sum(AB.total.^2))/I)-FC
    AA= by(X, :Trat) do df
    (m = mean(df.Y), s² = var(df.Y), total= sum(df.Y), ss=sum((df.Y).^2))
    end
    SQTrat=((sum(AA.total.^2))/J)-FC
    SQTotal=(sum(X.Y.^2))-FC
    SQResiduo=SQTotal-SQBlocos-SQTrat
    QMBlocos=SQBlocos/GLBlocos
    QMTrat=SQTrat/GLTrat
    QMResiduo=SQResiduo/GLResiduo
    FBlocos=QMBlocos/QMResiduo
    FTrat=QMTrat/QMResiduo
    AlphaBlocos= ccdf(FDist(GLBlocos, GLResiduo),abs(FBlocos))
    AlphaTrat= ccdf(FDist(GLTrat, GLResiduo),abs(FTrat))
    R²=SQTrat/SQTotal*100
    CV=sqrt(QMResiduo)/mean(X.Y)*100
    println("GLBlocos: ", GLBlocos)
    println("GLTrat: ", GLTrat)
    println("GLResíduo: ", GLResiduo)
    println("GLTotal: ", GLTotal)
    println("SQBlocos: ", SQBlocos)
```

```
    println("SQTrat: ", SQTrat)
    println("SQResíduo: ", SQResiduo)
    println("SQTotal: ", SQTotal)
    println("QMBlocos: ", QMBlocos)
    println("QMTrat: ", QMTrat)
    println("QMResíduo: ", QMResiduo)
    println("FBlocos: ", FBlocos)
    println("FTrat: ", FTrat)
    println("AlphaBlocos: ", AlphaBlocos)
    println("AlphaTrat: ", AlphaTrat)
    println("R²(%): ", R²)
    println("CV(%): ", CV)
end

#DBA(Tratamento,Bloco,Variavel)
>DBA(X.Trat,X.Bloco,X.Volume)
GLBlocos: 5
GLTrat: 3
GLResíduo: 15
GLTotal: 23
SQBlocos: 1943.88
SQTrat: 54633.13
SQResíduo: 158.63
SQTotal: 56735.63
QMBlocos: 388.78
QMTrat: 18211.04
QMResíduo: 10.58
FBlocos: 36.76
FTrat: 1722.08
AlphaBlocos: 6.70e-8
AlphaTrat: 2.97e-19
R²(%): 96.29
CV(%): 1.41
```

TABELA 15.3. Resultados da análise de variância aplicada para avaliar a produção em volume de madeira (em $m^3.ha^{-1}$) do fuste de diferentes espécies florestais em delineamento em blocos ao acaso.

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	F calculado	F tabelado ($\alpha = 5\%$)	P-valor (%)
Blocos	5	1.943,88	388,78	36,76**	2,90	<0,001%
Tratamentos	3	54.633,13	18.211,04	1.722,08**	3,29	<0,001%
Erro Exp	15	158,63	10,58			
Total	23	56.735,63				

Em que: ** significante ao nível de $\alpha = 0{,}01$.

O teste de F para avaliar os efeitos dos blocos foi significante, indicando que estes foram efetivos para o controle local. O teste F na análise de variância para avaliar o efeito de tratamentos foi significante, indicando que pelo menos uma média de tratamentos é diferente.

O teste de Tukey foi aplicado para comparação das médias dos tratamentos, utilizando-se a função *Tukey*, descrita na seção 19, para um nível de significância de 0,05 (α). A diferença mínima de significância foi de 5,41 $m^3.ha^{-1}$. A tabela das significâncias encontradas e o gráfico ilustrativo dos resultados revelam que as médias de todos os tratamentos são diferentes. Um resumo dos resultados está apresentado na Tabela 15.4.

Arquivo: 30 - Função DBA para ANOVA em DBA

```
>using DataFrames, Statistics, PyPlot, Distributions, StatsFuns
#Tukey(Tratamento,Repeticao,Variavel,GLResiduo,QMResiduo,Alpha)
>Tukey(X.Trat,X.Bloco,X.Volume,15,10.575,0.05)
```

julia

```
julia> Tukey(X.Trat,X.Bloco,X.Volume,15,10.575,0.05)
Tabela de média: 4×3 DataFrame
| Row | Tratamento | Média    | Variância |
|     | String     | Float64  | Float64   |
|-----|------------|----------|-----------|
| 1   | A          | 272.167  | 66.1667   |
| 2   | B          | 253.167  | 142.967   |
| 3   | D          | 247.667  | 149.067   |
| 4   | C          | 149.5    | 62.3      |
Diferença Mínimca Significativa: 5.411229310213345
Tabela DMS: 4×5 DataFrame
| Row | Tratamento | A         | B         | D         | C         |
|     | String     | Float64   | Float64   | Float64   | Float64   |
|-----|------------|-----------|-----------|-----------|-----------|
| 1   | A          | 0.0       | 19.0      | 24.5      | 122.667   |
| 2   | B          | -19.0     | 0.0       | 5.5       | 103.667   |
| 3   | D          | -24.5     | -5.5      | 0.0       | 98.1667   |
| 4   | C          | -122.667  | -103.667  | -98.1667  | 0.0       |
Tabela de hipóteses DMS: 4×5 DataFrame
| Row | Tratamento | A      | B      | D      | C      |
|     | String     | String | String | String | String |
|-----|------------|--------|--------|--------|--------|
| 1   | A          | ns     | *      | *      | *      |
| 2   | B          | *      | ns     | *      | *      |
| 3   | D          | *      | *      | ns     | *      |
| 4   | C          | *      | *      | *      | ns     |
PyObject Text(0, 0.5, 'Volume (m³/ha)')
```

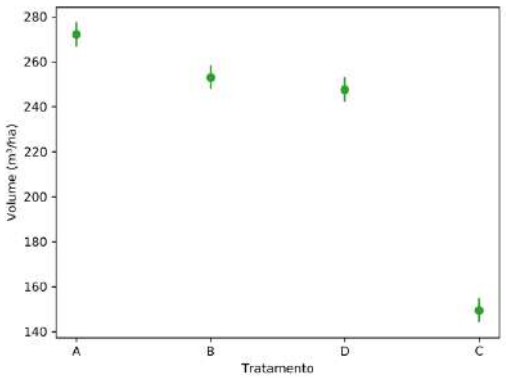


TABELA 15.4. Resultado do teste de Tukey para comparação das médias das diferentes espécies, na produção em volume de madeira (em m^3 ha^{-1}) do fuste de diferentes espécies florestais.

Tratamento	Média	Grupo
Eucalyptus grandis	272,17	A
Pinus taeda	253,17	B
Eucalyptus urophylla	247,67	C
Acacia mearnsii	149,50	D

Médias seguidas pela mesma letra não diferem significantemente pelo teste de Tukey.

Por meio do teste de Tukey foi identificado que a espécie com maior produção em volume foi o *Eucalyptus grandis*, que diferiu estatisticamente das demais. Assim, se o objetivo da silvicultura é maximizar a produção de volume de madeira em um hectare, aos 8 anos, sugere-se o plantio desta espécie.

Um resumo geral sobre o delineamento em blocos ao acaso está apresentado na Figura 15.6.

FIGURA 15.6. Resumo do Delineamento em blocos ao acaso.

DELINEAMENTO EM BLOCOS AO ACASO

$$Y_{ij}=\mu+\tau_i+\beta_j+\epsilon_{ij}$$

É o delineamento mais utilizado em pesquisas na Engenharia Florestal. Tem como objetivo manter a variabilidade entre as unidades experimentais o menor possível e maximizar a diferença entre blocos.

Neste delineamento cada bloco deverá ser tão uniforme quanto possível, porém os blocos podem diferir uns dos outros.

Os princípios experimentais são: a repetição, a casualização e o controle local.

VANTAGENS

- Controla as diferenças que ocorrem nas condições experimentais;
- Permite utilizar qualquer número de tratamentos e blocos (dentro de certo limites);
- Conduz a uma estimativa mais exata para a variância residual.

DESVANTAGENS

- Redução do número de graus de liberdade.

julia function AleatDBA(Tratamento,Bloco, Variável)

16 DELINEAMENTO EM QUADRADO LATINO

O delineamento em quadrado latino (DQL) é uma extensão do delineamento em blocos ao acaso, apropriado para ambientes experimentais heterogêneos, nos quais as fontes de variação a serem controladas nas unidades experimentais são duas, determinando, assim, um duplo bloqueamento.

A casualização dos tratamentos é restrita, pois estes devem ser sistematicamente ordenados em linhas e colunas. O número de tratamentos determina o número de linhas e colunas do delineamento, assim como o número de repetições do experimento. As linhas constituem um tipo de bloco para controlar uma fonte de variação e as colunas outro tipo de bloco para controlar uma segunda fonte de variação. Assim como no DBA, os blocos no DQL (linhas e colunas) podem representar uma distribuição espacial física das unidades experimentais, ou outros fatores que se deseje controlar.

A função apresentada na sequência pode ser utilizada para aleatorizar os tratamentos em um DQL. A partir de uma matriz com os tratamentos distribuídos sistematicamente, de modo que cada um deles é repetido apenas uma vez em cada linha e em cada coluna, é necessário para atender ao princípio básico de casualização, sortear as linhas e, a partir do resultado obtido sortear as colunas. No exemplo apresentado na sequência, foi utilizado um experimento composto por 5 tratamentos (A, B, C, D e E), os quais foram sistematicamente distribuídos em linhas e colunas. Na sequência, as linhas e colunas foram aleatorizadas.

Comandos para obter a casualização de um DQL:

ARQUIVO: 31 - CASUALIZAR OS TRATAMENTOS EM DQL

```
>using DataFrames, CSV, Random
>X=CSV.read("Z:/Alexandre/Julia/Exemplos/aleatdql.csv", DataFrame)
```

julia

```
julia> X=CSV.read("Z:/Alexandre/Julia/Exemplos/aleatdql.csv", DataFrame)
5×6 DataFrame
```

Row	LC	C1	C2	C3	C4	C5
	String	String	String	String	String	String
1	L1	A	B	C	D	E
2	L2	E	A	B	C	D
3	L3	D	E	A	B	C
4	L4	C	D	E	A	B
5	L5	B	C	D	E	A

```
>function AleatDQL(X)
    Linha(X) = X[shuffle(1:end), :]
    Linhas=Linha(X)
    Coluna(X) = X[:, shuffle(2:end)]
    Colunas=Coluna(X)
end

#AleatDQL(X)
>AleatDQL(X)
julia> AleatDQL(X)
```

5×5 DataFrame

Row	C4	C5	C2	C1	C3
	String	String	String	String	String
1	D	E	B	A	C
2	C	D	A	E	B
3	B	C	E	D	A
4	A	B	D	C	E
5	E	A	C	B	D

As principais características do DQL são:

- Controlar duas fontes de variação existentes nas unidades experimentais, perpendiculares entre si;
- O número de tratamentos é restrito, variando entre 5 e 8. De modo geral, mais do que 8 tratamentos compõem um experimento grande, enquanto poucos tratamentos forçam a redução no número de graus de liberdade dos resíduos (o que não é desejável);
- O número de graus de liberdade para o erro experimental é reduzido devido às fontes de variação controladas pelas linhas e colunas;
- Os princípios da experimentação que o caracteriza são casualização e duplo controle local;
- Os tratamentos são aleatorizados nas unidades experimentais de forma a participar somente uma vez em cada linha e cada coluna;
- As fontes de variação controladas correspondem às linhas, colunas e aos tratamentos.
- Aplicações florestais com este delineamento são recomendadas para possibilitar o controle da fertilidade do solo em dois gradientes ou experimentos em que são necessários controlar outros fatores (genéticos e idade, por exemplo).

O modelo linear aditivo do DQL é dado por:

$$Y_{ijk}=\mu+L_i+C_j+\tau_k+\epsilon_{ijk}$$

Em que:

Y_{ijk} é o valor de uma observação correspondente a *i*-ésima linha na *j*-ésima coluna na *k*-ésima parcela;

μ é o efeito constante (usualmente corresponde à média geral) – inerente à todas as observações;

L_i é o efeito do bloco da i-ésima linha;

C_j é o efeito do bloco da j-ésima coluna;

τ_k é o efeito do k-ésimo tratamento – decorrente de sua ação (desvio em relação a μ);

ϵ_{ijk} é o erro experimental associado a i-ésima linha na j-ésima coluna na k-ésima parcela.

O esquema de análise de variância para o delineamento em quadrado latino está apresentado na Tabela 16.1.

TABELA 16.1. Esquema para análise de variância para o delineamento em quadrado latino.

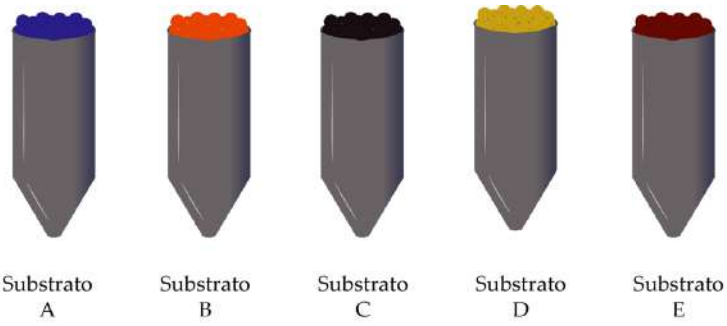
Causas de variação	Graus de Liberdade	Soma de Quadrados	Quadrados Médios	F calculado
Linhas (L)	$t - 1$	$SQ_L = \frac{\sum_j Y_j^2}{t} - FC$	$QM_L = \frac{SQ_L}{GL_L}$	$\frac{QM_L}{QM_e}$
Colunas (C)	$t - 1$	$SQ_C = \frac{\sum_k Y_k^2}{t} - FC$	$QM_C = \frac{SQ_C}{GL_C}$	$\frac{QM_C}{QM_e}$
Tratamentos (t)	$t - 1$	$SQ_t = \frac{\sum_i Y_i^2}{t} - FC$	$QM_t = \frac{SQ_t}{GL_t}$	$\frac{QM_t}{QM_e}$
Resíduo (e)	$(t - 1) \cdot (t - 2)$	$SQ_e = SQ_{Total} - SQ_t - SQ_L - SQ_C$	$QM_e = \frac{SQ_e}{GL_e}$	
Total	$t^2 - 1$	$SQ_{total} = \sum_i \sum_j Y_{ijk}^2 - FC$		

$FC = \frac{\left(\sum_j \sum_k Y_{jk}\right)^2}{I^2}$, fator de correção para a média e $t = I$, correspondendo ao número de tratamentos.

Um experimento foi conduzido para avaliar o crescimento de mudas de *Araucaria angustifolia* sobre efeito de 5 tipos de substratos: (A) Casca de pinus decomposta (70%) e terra de subsolo (30%); (B) Composto orgânico (70%), casca de arroz carbonizada (20%) e terra de subsolo (10%); (C) Composto orgânico (60%), casca de arroz carbonizada (20%) e areia (20%); (D) Casca de pinus (50%), vermiculita (30%) e cama de aviário (20%) e (E) Terra de subsolo (50%), cama de aviário (30%), vermiculita (10%) e areia (10%).

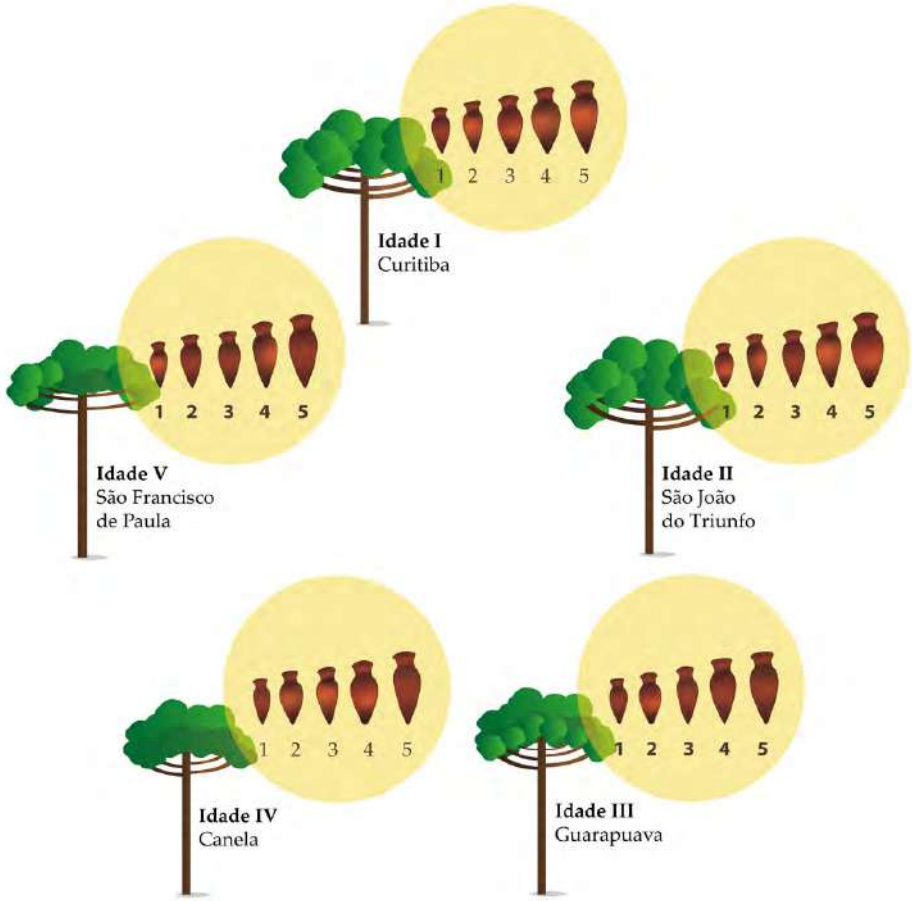
Os tratamentos estão representados na Figura 16.1.

FIGURA 16.1. Substratos utilizados como tratamentos para avaliar o crescimento de mudas de *Araucaria angustifolia*.



A unidade experimental foi composta por 12 tubetes, cada um com uma semente. As sementes pertencem a 5 idades e 5 tamanhos diferentes (Figura 16.2).

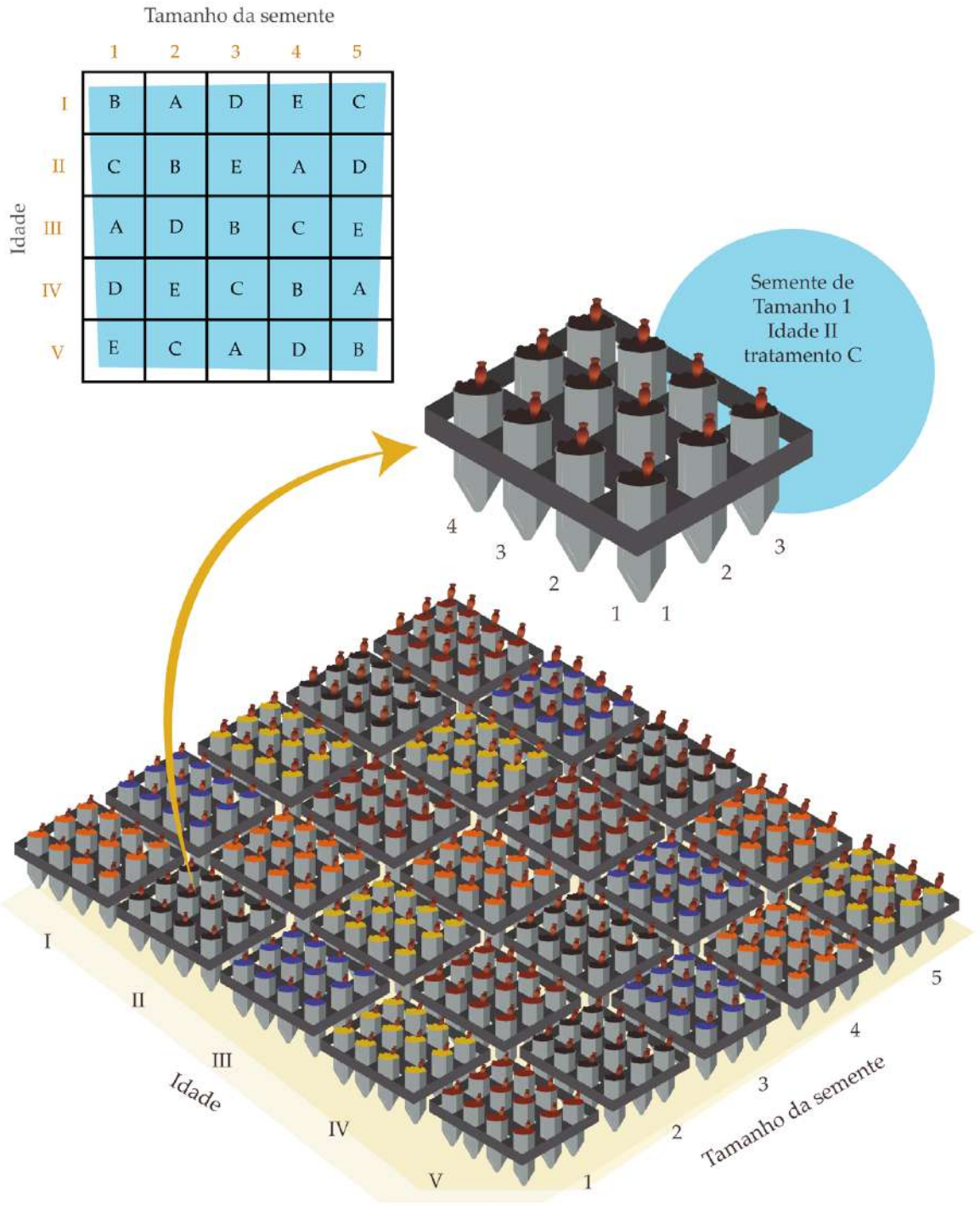
FIGURA 16.2. Árvores de *Araucaria angustifolia* para seleção de sementes a serem submetidas à avaliação de efeito de tipos de substratos no crescimento das mudas, para 5 idades e tamanhos de sementes.



Para controlar as possíveis diferenças no crescimento das mudas em função da idade das árvores e do tamanho das sementes foi utilizado o delineamento em quadrado latino. Nas colunas deste delineamento foi controlado o tamanho das sementes e nas linhas foram controladas as diferentes idades.

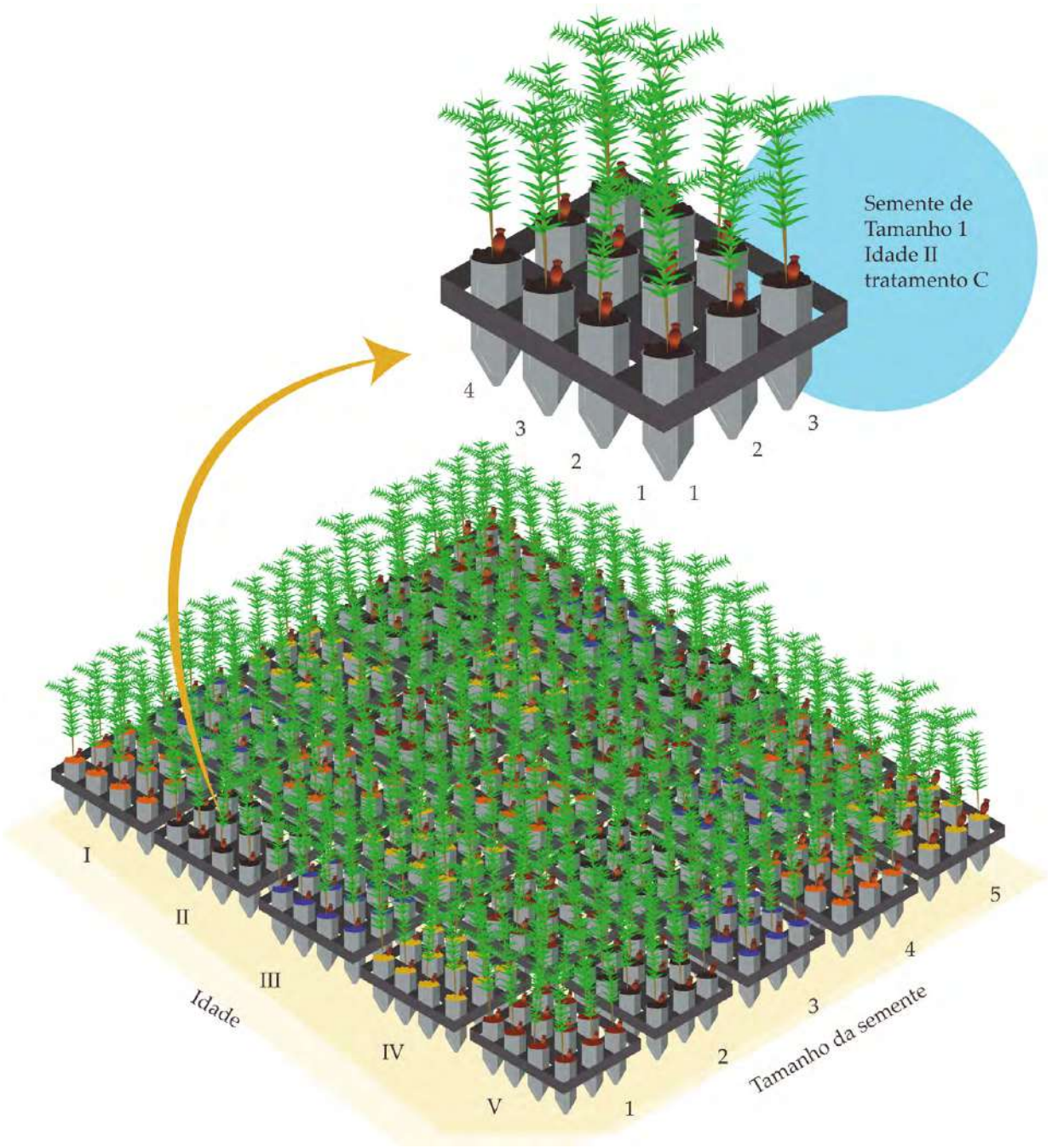
A unidade experimental e a atribuição dos tratamentos às subunidades dentro dela estão apresentadas na Figura 16.3. Plantas de bordadura não foram adicionadas a este experimento.

FIGURA 16.3. Experimento para avaliar o efeito de tipos de substratos no crescimento de mudas de *Araucaria angustifolia* com delineamento em quadrado latino.



O experimento foi conduzido em casa de vegetação, sendo efetuadas irrigações diárias. Um ano após a semeadura foi avaliada a altura das mudas (em cm) e o diâmetro de colo (em mm) (Figura 16.4).

FIGURA 16.4. Experimento para avaliar o efeito de tipos de substratos no crescimento de mudas de *Araucaria angustifolia* com delineamento em quadrado latino um ano após a instalação.



Todas as mudas foram avaliadas nas unidades experimentais e o valor médio observado em cada uma delas está apresentado na Tabela 16.2.

TABELA 16.2. Resultados experimentais obtidos para a altura (em cm) e diâmetro de colo (em mm) da espécie *Araucaria angustifolia*, sob efeito de cinco diferentes substratos.

Linha	Coluna	Tratamento	Altura	Diâmetro de colo
I	1	B	36	2,2
I	2	A	27	1,1
I	3	D	28	0,8
I	4	E	29	0,8
I	5	C	28	0,8
II	1	C	26	2,2
II	2	B	34	2,5
II	3	E	20	1,9
II	4	A	20	1,7
II	5	D	26	1,5
III	1	A	35	0,4
III	2	D	33	0,8
III	3	B	37	1,8
III	4	C	32	1,2
III	5	E	28	1,1
IV	1	D	29	1,2
IV	2	E	32	0,6
IV	3	C	37	0,8
IV	4	B	39	1,8
IV	5	A	30	0,9
V	1	E	43	2,7
V	2	C	46	2,5
V	3	A	47	2,0
V	4	D	44	2,0
V	5	B	57	3,1

Os dados foram submetidos ao teste de Bartlett para avaliar a homogeneidade das variâncias. O valor de χ^2 corrigido para a variável altura foi de 0,512 e para a variável diâmetro o valor foi de 1,67, sendo ambos não significantes ao nível de 0,05 de significância. Assim, as variâncias dos tratamentos para as duas variáveis são homogêneas.

Os dados na escala original foram submetidos à análise de variância, cujos resultados obtidos estão apresentados na Tabela 16.3.

TABELA 16.3. Resultados da análise de variância realizada para testar o efeito de cinco substratos no crescimento em altura (em cm) de mudas de *Araucaria angustifolia*.

Fonte de variação	Graus de Liberdade	Soma de Quadrado	Quadrado Médio	F calculado	F tabelado (α = 5%)	P-valor (%)
Linhas (L)	4	1386,64	346,66	40,50 **	3,26	0,00007
Colunas (C)	4	6,64	1,66	0,19 ns	3,26	93,90
Tratamentos (t)	4	325,04	81,26	9,49 **	3,26	0,1067
Erro Exp.	12	102,72	8,56			
Total	24	1821,04				

ns: não significante, ** significante a α = 0,01.

A variável tratamento foi denominada de “Trat” e a altura das mudas de “Y”.

Os comandos utilizados para obtenção da análise de variância foram:

julia

```
julia> X=CSV.read("Z:/Alexandre/Julia/Exemplos/dql2.csv", DataFrame)
25×4 DataFrame
| Row | Linha  | Coluna | Trat   | Y     |
|     | String | Int64  | String | Int64 |
|-----|--------|--------|--------|-------|
| 1   | I      | 1      | B      | 36    |
| 2   | I      | 2      | A      | 27    |
| 3   | I      | 3      | D      | 28    |
| 4   | I      | 4      | E      | 29    |
| 5   | I      | 5      | C      | 28    |
⋮
| 20  | IV     | 5      | A      | 30    |
| 21  | V      | 1      | E      | 43    |
| 22  | V      | 2      | C      | 46    |
| 23  | V      | 3      | A      | 47    |
| 24  | V      | 4      | D      | 44    |
| 25  | V      | 5      | B      | 57    |

>function DQL(Tratamento,Linha,Coluna,Variavel)
    X=DataFrame()
    X.Trat=Tratamento
    X.Linha=Linha
    X.Coluna=Coluna
    X.Y=Variavel
    X
    I=length(unique(X.Linha))
    J= length(unique(X.Coluna))
    K= length(unique(X.Trat))
    GLLinhas=I-1
    GLColunas=J-1
    GLTrat=K-1
    GLResiduo=(K-1)*(K-2)
    GLTotal=K^2-1
    FC=((sum(X.Y))^2)/(K^2)
```

```
    AL= by(X, :Linha) do df
    (m = mean(df.Y), s² = var(df.Y), total= sum(df.Y), ss=sum((df.Y).^2))
    end
    SQLinhas=((sum(AL.total.^2))/K)-FC
    AC= by(X, :Coluna) do df
    (m = mean(df.Y), s² = var(df.Y), total= sum(df.Y), ss=sum((df.Y).^2))
    end
    SQColunas=((sum(AC.total.^2))/K)-FC
    AT= by(X, :Trat) do df
    (m = mean(df.Y), s² = var(df.Y), total= sum(df.Y), ss=sum((df.Y).^2))
    end
    SQTrat=((sum(AT.total.^2))/K)-FC
    SQTotal=(sum(X.Y.^2))-FC
    SQResiduo=SQTotal-SQLinhas-SQColunas-SQTrat
    QMLinhas=SQLinhas/GLLinhas
    QMColunas=SQColunas/GLColunas
    QMTrat=SQTrat/GLTrat
    QMResiduo=SQResiduo/GLResiduo
    FLinhas=QMLinhas/QMResiduo
    FColunas=QMColunas/QMResiduo
    FTrat=QMTrat/QMResiduo
    AlphaLinhas= ccdf(FDist(GLLinhas, GLResiduo),abs(FLinhas))
    AlphaColunas= ccdf(FDist(GLColunas, GLResiduo),abs(FColunas))
    AlphaTrat= ccdf(FDist(GLTrat, GLResiduo),abs(FTrat))
    R²=SQTrat/SQTotal*100
    CV=sqrt(QMResiduo)/mean(X.Y)*100
    println("GLLinhas: ", GLLinhas)
    println("GLColunas: ", GLColunas)
    println("GLTrat: ", GLTrat)
    println("GLResíduo: ", GLResiduo)
    println("GLTotal: ", GLTotal)
    println("SQLinhas: ", SQLinhas)
    println("SQColunas: ", SQColunas)
    println("SQTrat: ", SQTrat)
    println("SQResíduo: ", SQResiduo)
    println("SQTotal: ", SQTotal)
    println("QMLinhas: ", QMLinhas)
    println("QMColunas: ", QMColunas)
    println("QMTrat: ", QMTrat)
    println("QMResíduo: ", QMResiduo)
    println("FLinhas: ", FLinhas)
    println("FColunas: ", FColunas)
    println("FTrat: ", FTrat)
    println("AlphaLinhas: ", AlphaLinhas)
    println("AlphaColunas: ", AlphaColunas)
    println("AlphaTrat: ", AlphaTrat)
    println("R²(%): ", R²)
    println("CV(%): ", CV)
end
```

```
#DQL(Tratamento,Linha,Coluna,Variavel)
>DQL(X.Trat,X.Linha,X.Coluna,X.Y)
GLLinhas: 4
GLColunas: 4
GLTrat: 4
GLResíduo: 12
GLTotal: 24
SQLinhas: 1386.64
SQColunas: 6.64
SQTrat: 325.04
SQResíduo: 102.72
SQTotal: 1821.04
QMLinhas: 346.66
QMColunas: 1.66
QMTrat: 81.26
QMResíduo: 8.56
FLinhas: 40.50
FColunas: 0.19
FTrat: 9.49
AlphaLinhas: 7.09e-7
AlphaColunas: 0.94
AlphaTrat: 0.001
R²(%): 17.85
CV(%): 8.68
```

O teste F da análise de variância, para esta variável, revelou que somente as linhas (idades) foram significantes, indicando haver diferenças entre elas, o que não foi observado para o tamanho das sementes (colunas). O teste F na análise de variância para tratamentos foi significante, indicando que pelo menos uma média de tratamentos é diferente. Foi aplicado o teste de Tukey para a comparação das médias entre tratamentos.

O teste de Tukey para comparação de médias foi aplicado utilizando-se a função Tukey, descrita na seção 19, para um nível 0,05 de significância (α). A diferença mínima de significância foi de 5,90 cm, a tabela contendo as significâncias detectadas e o gráfico revelam que a média do tratamento B difere das demais médias. Um resumo dos resultados está apresentado na Tabela 16.4.

Comandos utilizados para a realização do teste de Tukey:

Arquivo: 32 - Função dql para anova em dql

```
>using DataFrames, Statistics, PyPlot, Distributions, StatsFuns
#Tukey(Tratamento,Repeticao,Variavel,GLResiduo,QMResiduo,Alpha)
>Tukey(X.Trat,X.Trat,X.Y,12,8.56,0.05)
```

```
julia> Tukey(X.Trat,X.Trat,X.Y,12,8.56,0.05)
Tabela de média: 5×3 DataFrame
| Row | Tratamento | Média   | Variância |
|     | String     | Float64 | Float64   |
|-----|------------|---------|-----------|
| 1   | B          | 40.6    | 87.3      |
| 2   | C          | 33.8    | 64.2      |
| 3   | D          | 32.0    | 51.5      |
| 4   | A          | 31.8    | 101.7     |
| 5   | E          | 30.4    | 69.3      |
Diferença Mínimca Significativa: 5.898041726560932
Tabela DMS: 5×6 DataFrame
| Row | Tratamento | B       | C       | D       | A       | E       |
|     | String     | Float64 | Float64 | Float64 | Float64 | Float64 |
|-----|------------|---------|---------|---------|---------|---------|
| 1   | B          | 0.0     | 6.8     | 8.6     | 8.8     | 10.2    |
| 2   | C          | -6.8    | 0.0     | 1.8     | 2.0     | 3.4     |
| 3   | D          | -8.6    | -1.8    | 0.0     | 0.2     | 1.6     |
| 4   | A          | -8.8    | -2.0    | -0.2    | 0.0     | 1.4     |
| 5   | E          | -10.2   | -3.4    | -1.6    | -1.4    | 0.0     |
Tabela de hipóteses DMS: 5×6 DataFrame
| Row | Tratamento | B      | C      | D      | A      | E      |
|     | String     | String | String | String | String | String |
|-----|------------|--------|--------|--------|--------|--------|
| 1   | B          | ns     | *      | *      | *      | *      |
| 2   | C          | *      | ns     | ns     | ns     | ns     |
| 3   | D          | *      | ns     | ns     | ns     | ns     |
| 4   | A          | *      | ns     | ns     | ns     | ns     |
| 5   | E          | *      | ns     | ns     | ns     | ns     |
PyObject Text(0, 0.5, 'Altura (cm)')
```

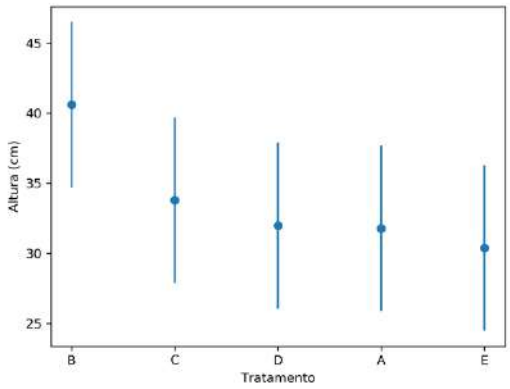


TABELA 16.4. Resultados do teste de Tukey para comparação das médias dos efeitos do substratos no crescimento em altura (em cm) de mudas de *Araucaria angustifolia*.

Tratamento	Média	Grupo
A	31,8	b
B	40,6	a
C	33,8	b
D	32,0	b
E	30,4	b

Médias seguidas pela mesma letra não diferem pelo teste de Tukey.

O teste de Tukey revelou que o tratamento B (Composto orgânico - 70%, casca de arroz carbonizada - 20% e terra de subsolo - 10%), teve a melhor resposta para a variável altura. Assim, na produção de mudas de Araucária sugere-se a utilização deste substrato.

Um resumo geral sobre o delineamento em blocos ao acaso está apresentado na Figura 16.5.

FIGURA 16.5. Resumo do Delineamento em quadrado latino.

DELINEAMENTO EM QUADRADO LATINO

julia

function DQL(Tratamento,Linha,Coluna,Variavel)

$$Y_{ijk}=\mu+L_i+C_j+\tau_k+\epsilon_{ijk}$$

O DQL é utilizado quando se tem duas causas de variação perpendiculares que são conhecidas e podem ser controladas, sendo apropriado para ambientes experimentais heterogêneos.

A casualização dos tratamentos é mais restrita, cada tratamento deve estar presente somente uma vez em cada linha e coluna.

Em experimentos com menos de 5 tratamentos há uma redução nos graus de liberdade do resíduo. Quando mais que 8 tratamentos há elevação no número de repetições, o que torna o experimento grande.

Vantagens

» Controla duas fontes de variação ao mesmo tempo;
» Reduz o erro experimental: remove, pelo menos, mais uma fonte de variação que estaria incluída no erro;

Desvantagens

» O número de tratamentos fica restrito, variando entre 5 e 8;
» Reduz o número de graus de liberdade para o erro experimental.

17 DELINEAMENTO EM ENSAIOS FATORIAIS

Experimentos com arranjos fatoriais dos tratamentos são ensaios em que dois ou mais fatores, em seus respectivos níveis, são testados simultaneamente. Os tratamentos correspondem à combinação dos níveis dos fatores estudados. A ideia neste tipo de arranjo é testar os fatores principais, assim como sua interação. A interação permite avaliar a dependência dos fatores e isso é uma característica positiva nos experimentos florestais, pois os fatores testados, em muitos experimentos, influenciam as variáveis de modo dependente.

Ensaios com arranjo fatorial dos tratamentos as vezes são considerados como sendo um novo tipo de delineamento. No presente contexto os fatoriais são considerados como uma especialização dos delineamentos experimentais básicos.

As principais características são:

- Avaliação de mais de um fator no mesmo ensaio;
- Avaliação da interação dos efeitos dos fatores;
- Na ausência de interação o número de repetições é aumentado no teste dos efeitos principais (repetição escondida);
- Propicia a utilização mais racional dos recursos experimentais;
- Uma vez que os tratamentos são constituídos por todas as combinações possíveis entre os níveis dos diversos fatores, o número de tratamentos aumenta e o experimento se torna grande.
- Ensaios com mais que três fatores são difíceis de interpretar. Devido a isso, ensaios com 2 ou 3 fatores são os mais apropriados. Às vezes é melhor fixar um fator do que considerá-lo na análise, pois isto facilita de maneira significante a interpretação dos resultados.

A combinação dos níveis do Fator A com os níveis do Fator B, que geram os tratamentos, pode ser obtida conforme apresentado abaixo.

Comandos utilizados para obtenção das combinações que constituem os tratamentos testados:

Arquivo: 33 - Casualizar os tratamentos em fatorial

```
>using DataFrames
>NFA=3 #Número de níveis do Fator A
>NFB=2 #Número de níveis do Fator B
>df = DataFrame(FatorA=repeat(1:NFA, inner=NFB),FatorB=repeat(1:NFB, outer=NFA),
Tratamento=1:(NFA*NFB))
```

```
julia> df = DataFrame(FatorA=repeat(1:NFA, inner=NFB),FatorB=repeat(1:NFB, outer=NFA),Tratamento=1:(NFA*NFB))
6×3 DataFrame
| Row | FatorA | FatorB | Tratamento |
|     | Int64  | Int64  | Int64      |
|-----|--------|--------|------------|
| 1   | 1      | 1      | 1          |
| 2   | 1      | 2      | 2          |
| 3   | 2      | 1      | 3          |
| 4   | 2      | 2      | 4          |
| 5   | 3      | 1      | 5          |
| 6   | 3      | 2      | 6          |
```

O resultado da função gerou uma tabela de tratamentos (*df*) com a combinação de todos os níveis do Fator A com todos os níveis do Fator B, o que resulta na composição dos tratamentos.

A função abaixo pode ser utilizada para aleatorizar os tratamentos, dependendo do tipo de delineamento a ser utilizado. Normalmente são utilizados o delineamento inteiramente casualizado ou em blocos ao acaso.

Comandos para casualização dos tratamentos.

Arquivo: 33 - Casualizar os tratamentos em fatorial

```
#casualização dos tratamentos
>using Random
>function AleatFat(X)
    Desenho=X[shuffle(1:end), :]
end
```

Para obter a estrutura de um experimento com arranjo fatorial dos tratamentos com delineamento de blocos, as funções abaixo podem ser utilizadas. Para vários blocos, a função anterior, que gera a tabela das combinações dos tratamentos (*df*) pode ser repetida. Por exemplo, se o experimento for composto por três blocos, o resultado será uma tabela (DataFrame) para cada bloco. Em cada tabela a coluna Row indica a parcela dentro do bloco que receberá o tratamento.

Comandos para casualização dos tratamentos em DBA:

Arquivo: 33 - Casualizar os tratamentos em fatorial

```
>NBlocos=3#Númeroo de blocos
>desenho= for i=1: NBlocos
    println(AleatFat(df))
end
```

```
6×3 DataFrame
 Row │ FatorA  │ FatorB  │ Tratamento
     │ Int64   │ Int64   │ Int64
─────┼─────────┼─────────┼────────────
   1 │ 3       │ 1       │ 5
   2 │ 1       │ 2       │ 2
   3 │ 1       │ 1       │ 1
   4 │ 3       │ 2       │ 6
   5 │ 2       │ 2       │ 4
   6 │ 2       │ 1       │ 3
6×3 DataFrame
 Row │ FatorA  │ FatorB  │ Tratamento
     │ Int64   │ Int64   │ Int64
─────┼─────────┼─────────┼────────────
   1 │ 2       │ 2       │ 4
   2 │ 1       │ 2       │ 2
   3 │ 1       │ 1       │ 1
   4 │ 3       │ 1       │ 5
   5 │ 2       │ 1       │ 3
   6 │ 3       │ 2       │ 6
6×3 DataFrame
 Row │ FatorA  │ FatorB  │ Tratamento
     │ Int64   │ Int64   │ Int64
─────┼─────────┼─────────┼────────────
   1 │ 3       │ 1       │ 5
   2 │ 1       │ 2       │ 2
   3 │ 3       │ 2       │ 6
   4 │ 2       │ 1       │ 3
   5 │ 1       │ 1       │ 1
   6 │ 2       │ 2       │ 4
```

Para um experimento fatorial em delineamento inteiramente casualizado, as seguintes funções podem ser utilizadas para casualizar os tratamentos. A coluna *Row* indica a parcela que receberá o tratamento.

Comandos para casualização dos tratamentos em DIC:

ARQUIVO: 33 - CASUALIZAR OS TRATAMENTOS EM FATORIAL

```
>NRep=3 #Número de repetições
>X=repeat(df, inner = NRep)
>AleatFat(X)
```

```
julia> AleatFat(X)
18×3 DataFrame
 Row │ FatorA  │ FatorB  │ Tratamento
     │ Int64   │ Int64   │ Int64
─────┼─────────┼─────────┼────────────
   1 │ 1       │ 1       │ 1
   2 │ 1       │ 1       │ 1
   3 │ 3       │ 1       │ 5
   4 │ 2       │ 1       │ 3
   5 │ 3       │ 1       │ 5
   6 │ 1       │ 1       │ 1
   7 │ 2       │ 2       │ 4
   8 │ 2       │ 1       │ 3
   9 │ 2       │ 1       │ 3
  10 │ 2       │ 2       │ 4
  11 │ 3       │ 2       │ 6
  12 │ 1       │ 2       │ 2
  13 │ 3       │ 2       │ 6
  14 │ 3       │ 2       │ 6
  15 │ 2       │ 2       │ 4
  16 │ 1       │ 2       │ 2
  17 │ 3       │ 1       │ 5
  18 │ 1       │ 2       │ 2
```

O modelo linear aditivo de um ensaio fatorial conduzido em DIC é dado por:

$$Y_{ijk}=\mu+\tau_i+\beta_j+(\tau\beta)_{ij}+\epsilon_{ijk}$$

Em que:

Y_{ijk} é o valor de uma observação correspondente ao *i*-ésimo nível do fator A, *j*-ésimo nível do fator B, na *k*-ésima unidade experimental;

μ é o efeito constante (usualmente corresponde à média geral) – constante inerente a todas as observações;

τ_i é o efeito do *i*-ésimo nível do fator A;

β_j é o efeito da *j*-ésimo nível do fator B;

$(\tau\beta)_{ij}$ é o efeito da interação entre τ_i e β_j;

ϵ_{ijk} é o erro experimental associado ao *i*-ésimo nível do fator A, *j*-ésimo nível do fator B, na *k*-ésima unidade experimental.

O esquema da análise de variância para o delineamento inteiramente casualizado em ensaios fatoriais está apresentado na Tabela 17.1.

TABELA 17.1. Esquema para análise de variância um experimento fatorial a·b conduzido em delineamento em delineamento inteiramento casualizado.

Causas de variação	Graus de Liberdade	Soma de Quadrados	Quadrados Médios	F calculado
Tratamento (t)	$(I \cdot J - 1) = t - 1$	$SQ_t = \frac{\sum_i \sum_j Y_{ij}^2}{K} - FC$	$QM_t = \frac{SQ_t}{GL_t}$	$\frac{QM_t}{QM_e}$
Fator A	$I - 1$	$SQ_A = \frac{\sum_i Y_i^2}{J.K} - FC$	$QM_a = \frac{SQ_a}{GL_a}$	$\frac{QM_a}{QM_e}$
Fator B	$J - 1$	$SQ_B = \frac{\sum_j Y_j^2}{I.J} - FC$	$QM_b = \frac{SQ_b}{GL_b}$	$\frac{QM_B}{QM_e}$
Interação A.B	$(I - 1) \cdot (J - 2)$	$SQ_{A.B} = SQ_t - SQ_A - SQ_B$	$QM_{A.B} = \frac{SQ_{A.B}}{GL_{A.B}}$	$\frac{QM_{A.B}}{QM_e}$
Resíduo (e)	$I \cdot I \cdot (K - 1)$	$SQ_e = SQ_{Total} - SQ_t$	$QM_e = \frac{SQ_e}{GL_e}$	
Total	$I \cdot I \cdot K - 1$	$SQ_{total} = \sum_i \sum_j \sum_k Y_{ijk}^2 - FC$		

$FC = \frac{\left(\sum_i \sum_j \sum_k Y_{ijk}\right)^2}{I\,J\,K}$, fator de correção para a média.

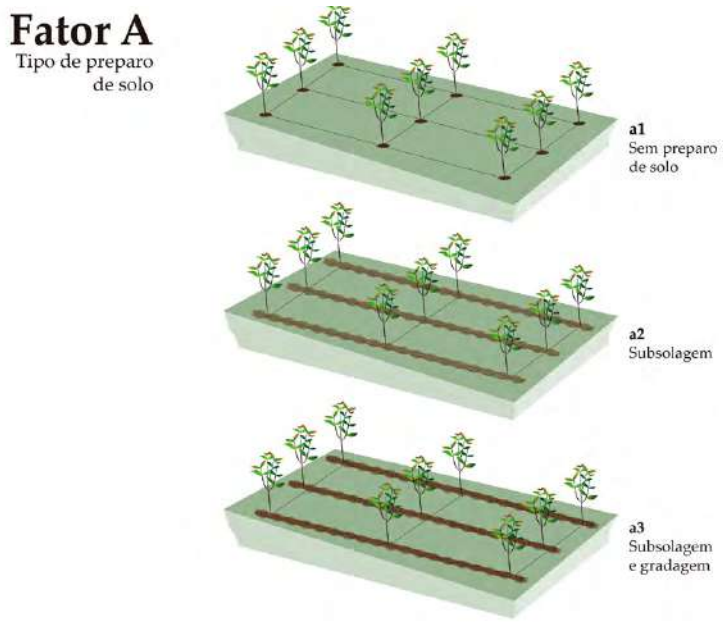
I é o número de níveis do fator A.
J é o número de níveis do fator B.
K é o número de repetições.

Se o experimento fatorial for conduzido utilizando o modelo de blocos ao acaso, então o modelo do ensaio é dado por: $Y_{ijk}=\mu+\tau_i+\beta_j+(\tau\beta)_{ij}+\Psi_k+\epsilon_{ijk}$, onde Ψ_k é o efeito do *k*-ésimo bloco, como consequência a tabela da ANOVA deve seguir a composição deste modelo.

Um experimento foi conduzido para avaliar o crescimento inicial de *Eucalyptus urophylla* sob o efeito de três diferentes tipos de preparo de solo e dois métodos de adubação. Os fatores e seus respectivos níveis são:

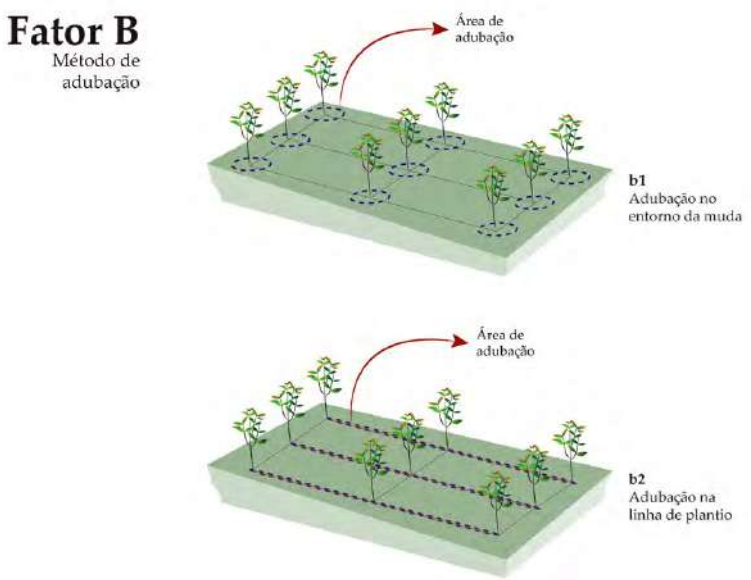
- **Fator A:** Tipo de preparo de solo (Figura 17.1)
  Níveis:
  (a1) Sem preparo
  (a2) Subsolagem
  (a3) Subsolagem e gradagem

FIGURA 17.1. Níveis testados para o preparo de solo (fator a) para avaliar o crescimento inicial de *Eucalyptus urophylla*.



- **Fator B:** Método de adubação (Figura 17.2)
  Níveis:
  (b1) Adubação no entorno da muda
  (b2) Adubação na linha de plantio

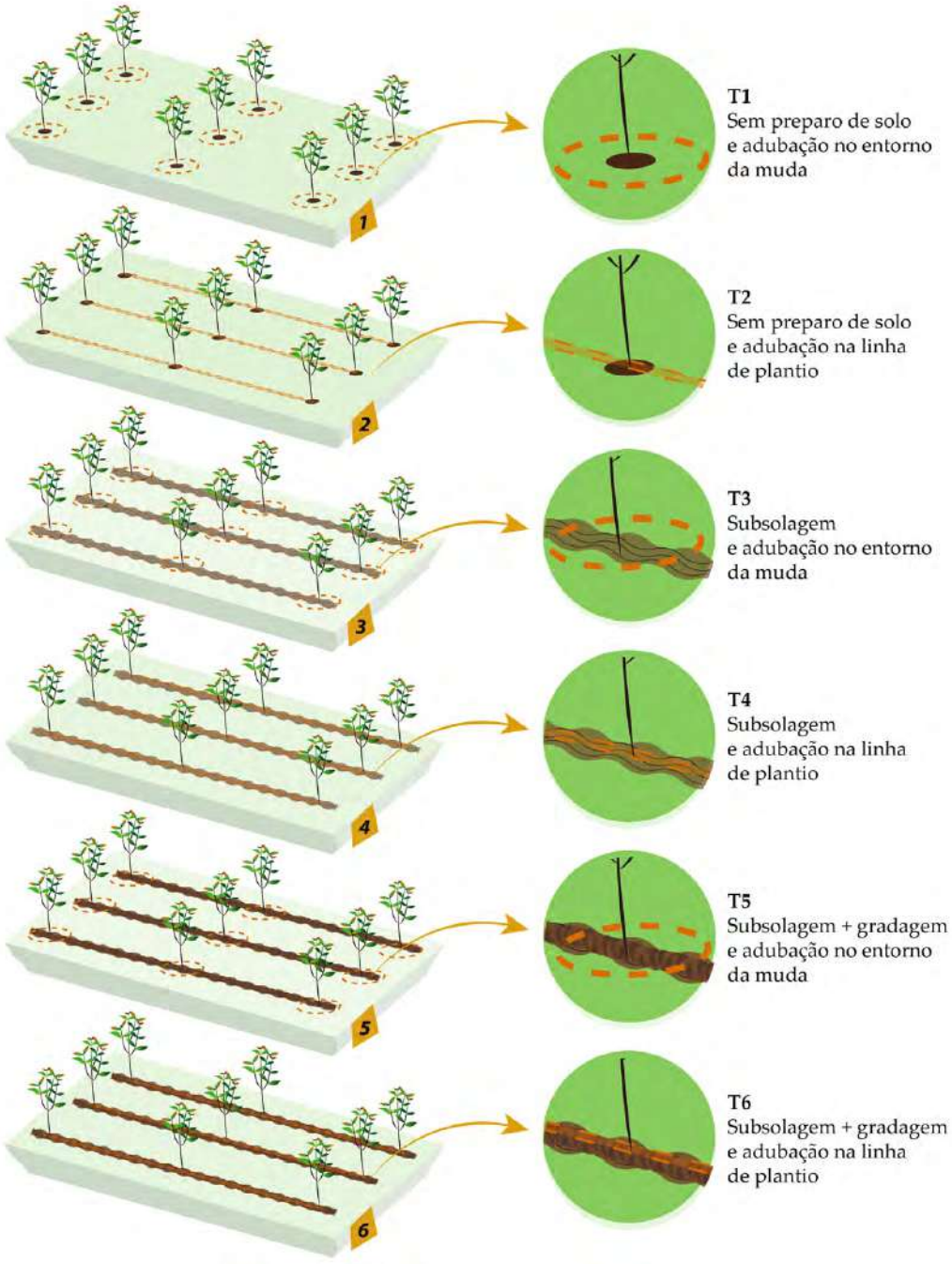
FIGURA 17.2. Níveis testados para os métodos de adubação (fator b) para avaliar o crescimento inicial de E*ucalyptus urophylla*.



Os tratamentos resultantes da combinação de todos os níveis preparo do solo e métodos de adubação foram (Figura 17.3):

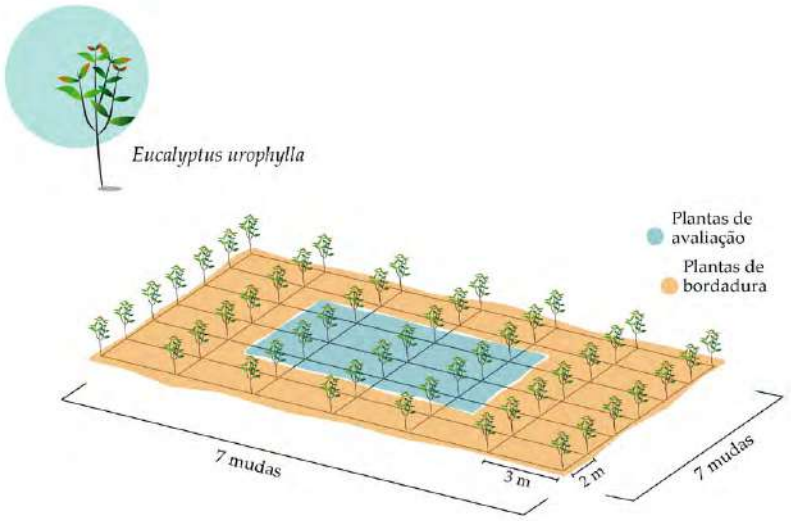
- Tratamento 1: a1 + b1
- Tratamento 2: a1 + b2
- Tratamento 3: a2 + b1
- Tratamento 4: a2 + b2
- Tratamento 5: a3 + b1
- Tratamento 6: a3 + b2

FIGURA 17.3. Tratamentos resultantes da combinação de diferentes níveis de preparo do solo e métodos de adubação testados para avaliar o crescimento inicial de *Eucalyptus urophylla*.



A unidade experimental foi composta com 49 plantas. O plantio das mudas de *Eucalyptus urophylla* (com aproximadamente 35 cm) foi efetuado em espaçamento de 3m x 2m (Figura 17.4). Duas linhas de bordadura foram definidas, totalizando, ao final do experimento, 9 plantas para avaliação em cada unidade experimental.

FIGURA 17.4. Unidade experimental com mudas de *Eucalyptus urophylla* utilizadas para avaliar o crescimento inicial sob efeito de diferentes tipos de preparo de solo e métodos de adubação.



O experimento foi instalado em campo em uma área levemente ondulada. Devido à vegetação presente no local pressupôs-se existir um gradiente de variação na fertilidade do solo. Amostras de solo foram coletadas e analisadas quanto ao teor de macro- e micronutrientes, o que possibilitou identificar a necessidade de controle das unidades experimentais. Portanto, o experimento foi conduzido com delineamento em blocos ao acaso com 5 repetições (blocos) (Figura 17.5). Adicionalmente, cada bloco está representado nas Figuras 17.6 a 17.10.

FIGURA 17.5. Experimento para avaliar o efeito de diferentes tipos de preparo de solo e métodos de adubação no crescimento inicial de mudas de *Eucalyptus urophylla*.

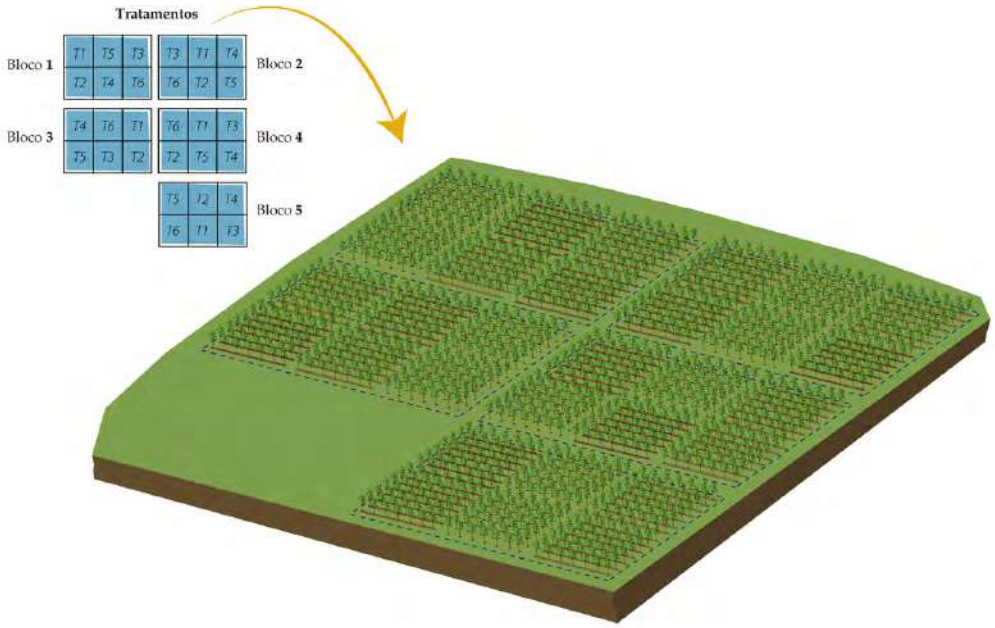


FIGURA 17.6. Experimento para avaliar o efeito de diferentes tipos de preparo de solo e métodos de adubação no crescimento inicial de mudas de *Eucalyptus urophylla* - bloco 1.

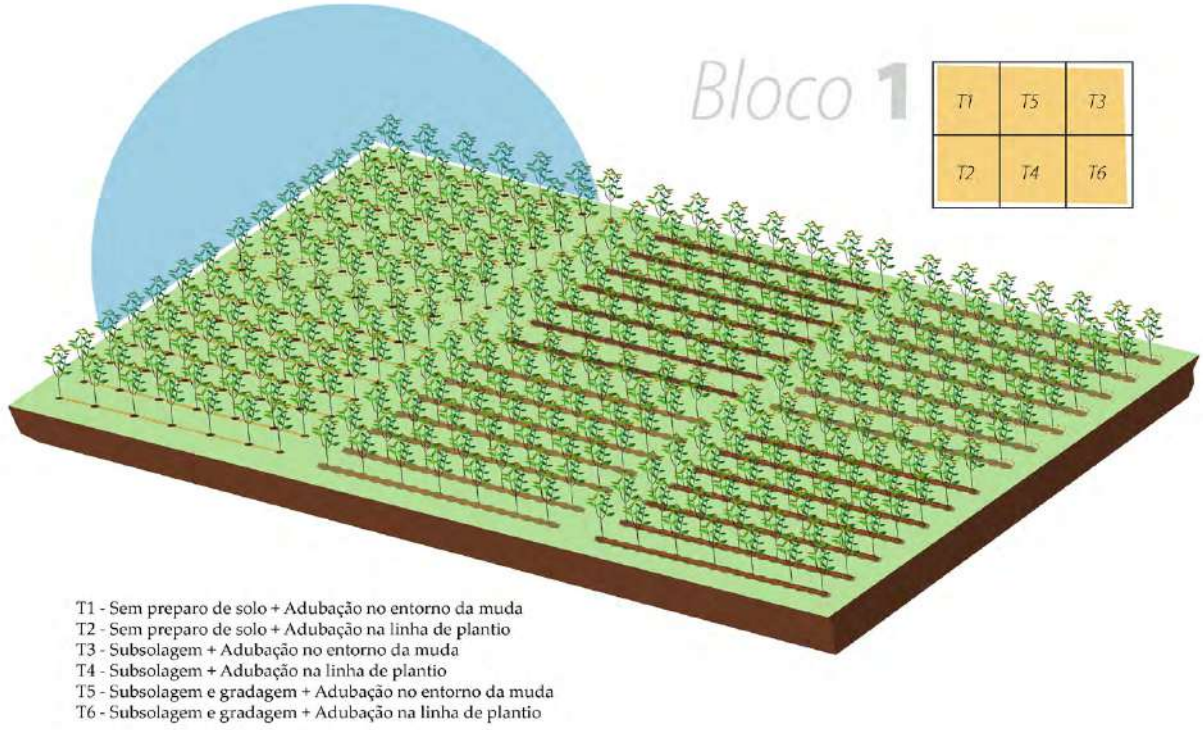


FIGURA 17.7. Experimento para avaliar o efeito de diferentes tipos de preparo de solo e métodos de adubação no crescimento inicial de mudas de *Eucalyptus urophylla* - bloco 2.

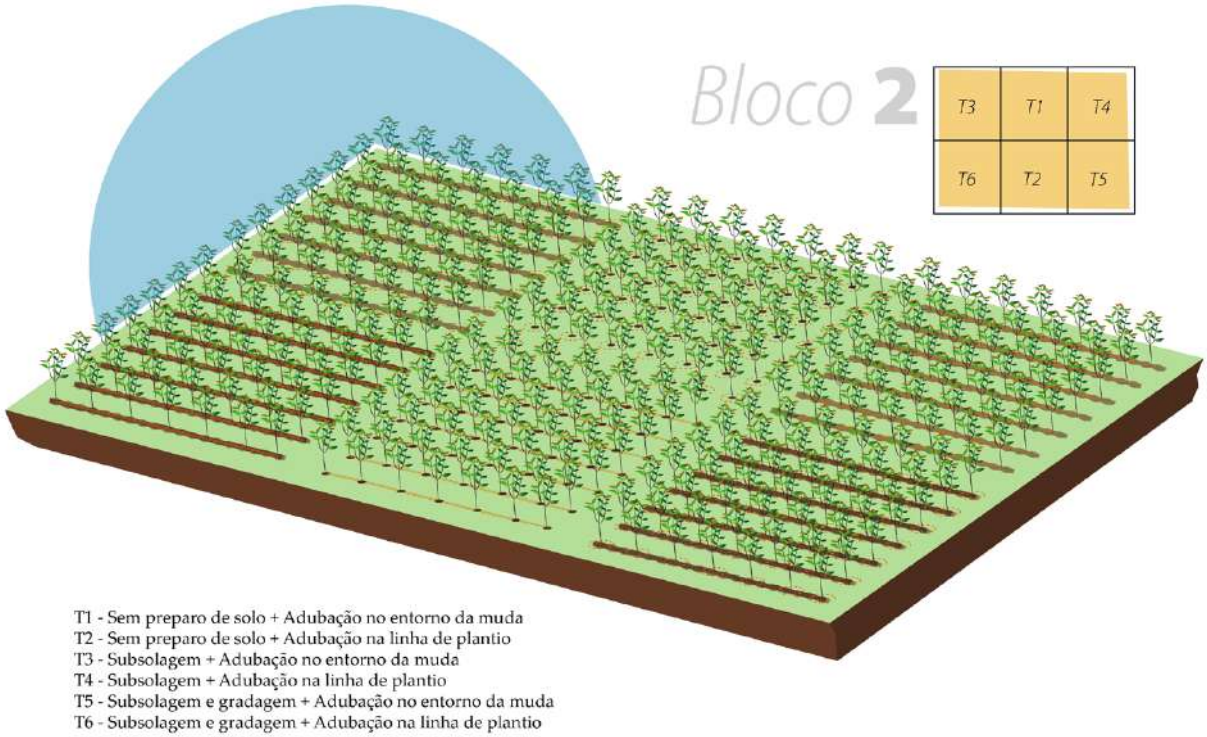
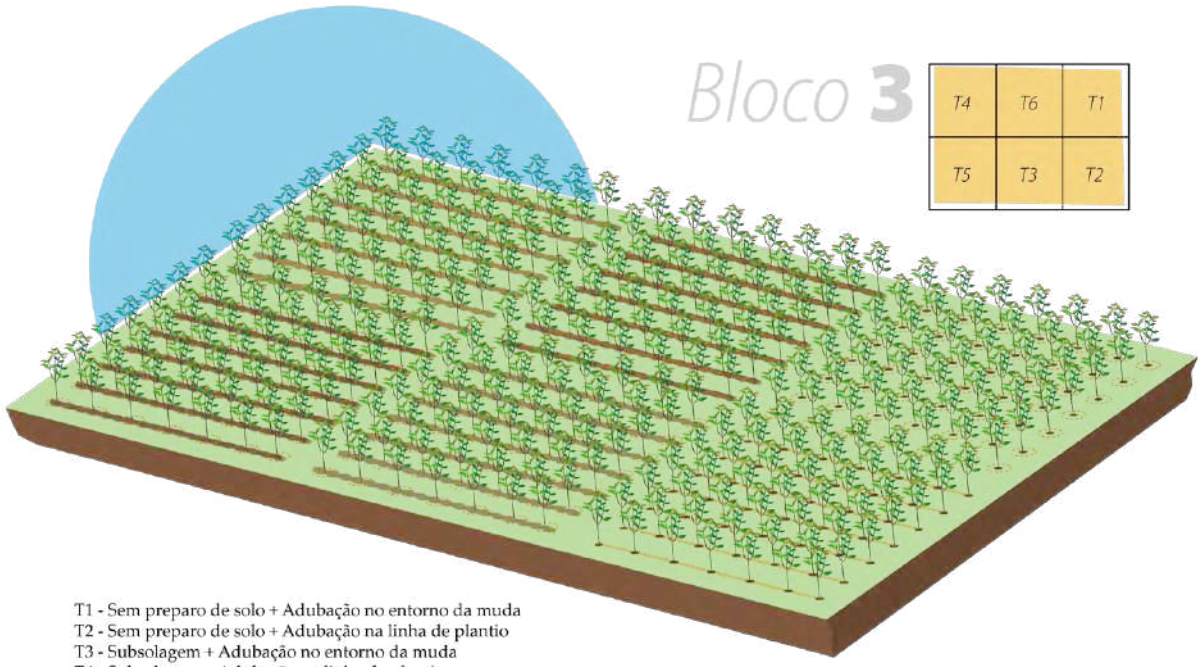
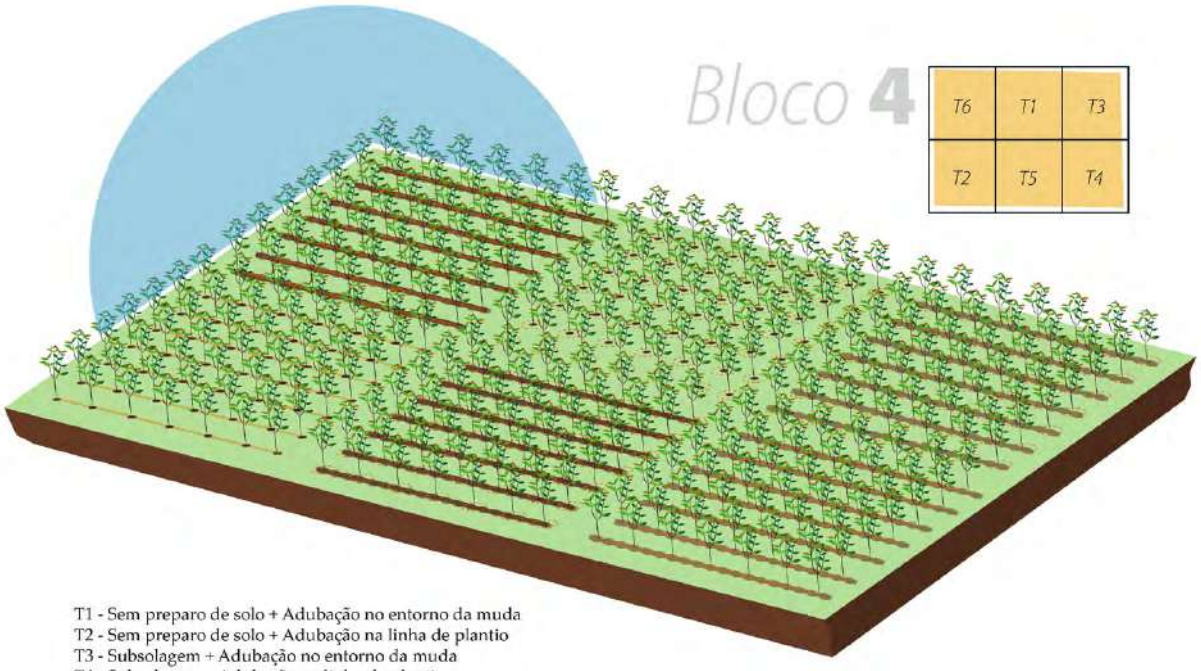


FIGURA 17.8. Experimento para avaliar o efeito de diferentes tipos de preparo de solo e métodos de adubação no crescimento inicial de mudas de *Eucalyptus urophylla* - bloco 3.



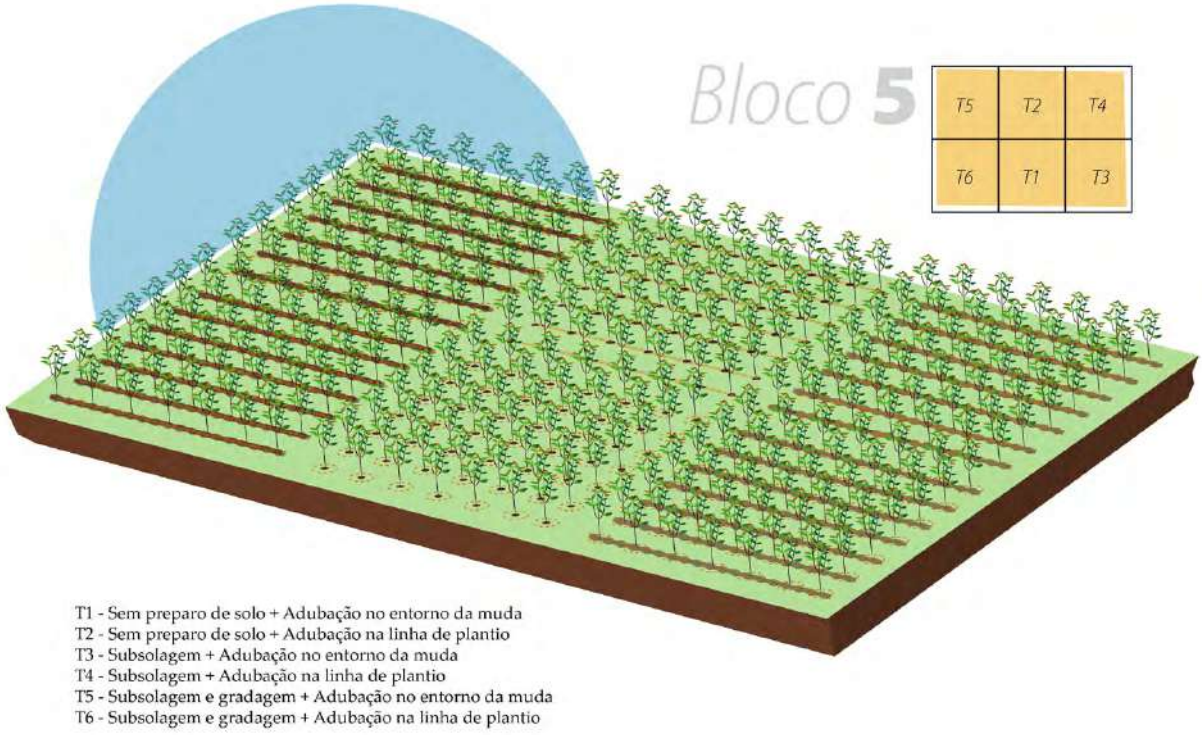
T1 - Sem preparo de solo + Adubação no entorno da muda
T2 - Sem preparo de solo + Adubação na linha de plantio
T3 - Subsolagem + Adubação no entorno da muda
T4 - Subsolagem + Adubação na linha de plantio
T5 - Subsolagem e gradagem + Adubação no entorno da muda
T6 - Subsolagem e gradagem + Adubação na linha de plantio

FIGURA 17.9. Experimento para avaliar o efeito de diferentes tipos de preparo de solo e métodos de adubação no crescimento inicial de mudas de *Eucalyptus urophylla* - bloco 4.



T1 - Sem preparo de solo + Adubação no entorno da muda
T2 - Sem preparo de solo + Adubação na linha de plantio
T3 - Subsolagem + Adubação no entorno da muda
T4 - Subsolagem + Adubação na linha de plantio
T5 - Subsolagem e gradagem + Adubação no entorno da muda
T6 - Subsolagem e gradagem + Adubação na linha de plantio

FIGURA 17.10. Experimento para avaliar o efeito de diferentes tipos de preparo de solo e métodos de adubação no crescimento inicial de mudas de *Eucalyptus urophylla* - bloco 5.



Um ano após a instalação do experimento foi medida a altura das plantas com uma régua telescópica, o diâmetro à altura do peito com uma suta e diâmetro de copa com uma trena (Figuras 17.11 - 17.12). Os dados obtidos estão apresentados na Tabela 17.2.

Os dados da variável altura foram submetidos ao teste de Bartlett para avaliar a homogeneidade das variâncias. O valor de χ^2 corrigido foi de 2,24, sendo não significante ao nível de 0,05 de significância. Assim, as variâncias são homogêneas.

FIGURA 17.11. Unidade experimental com mudas de *Eucalyptus urophylla* utilizadas para avaliar o crescimento inicial sob efeito de diferentes tipos de preparo de solo e métodos de adubação após a instalação do experimento.

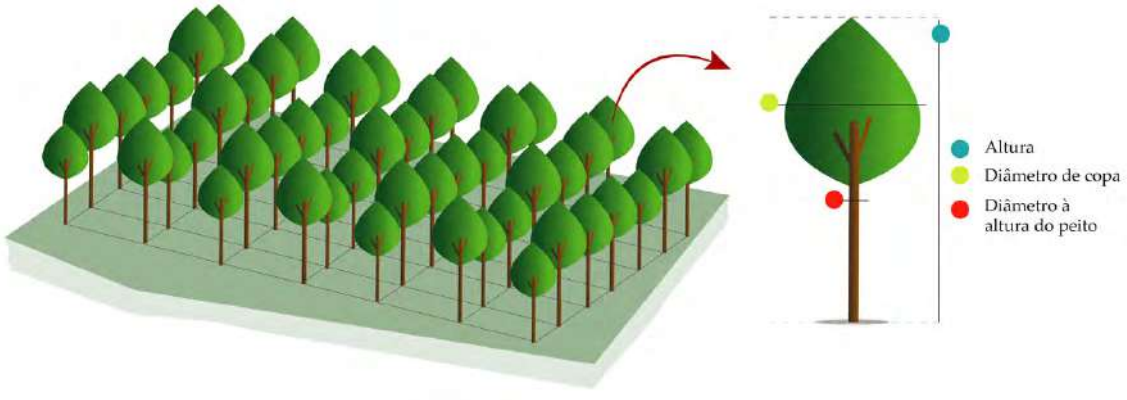


FIGURA 17.12. Experimento para avaliar o efeito de diferentes tipos de preparo de solo e métodos de adubação no crescimento inicial de mudas de *Eucalyptus urophylla* um ano após a instalação com delineamento de blocos ao acaso com arranjo fatorial

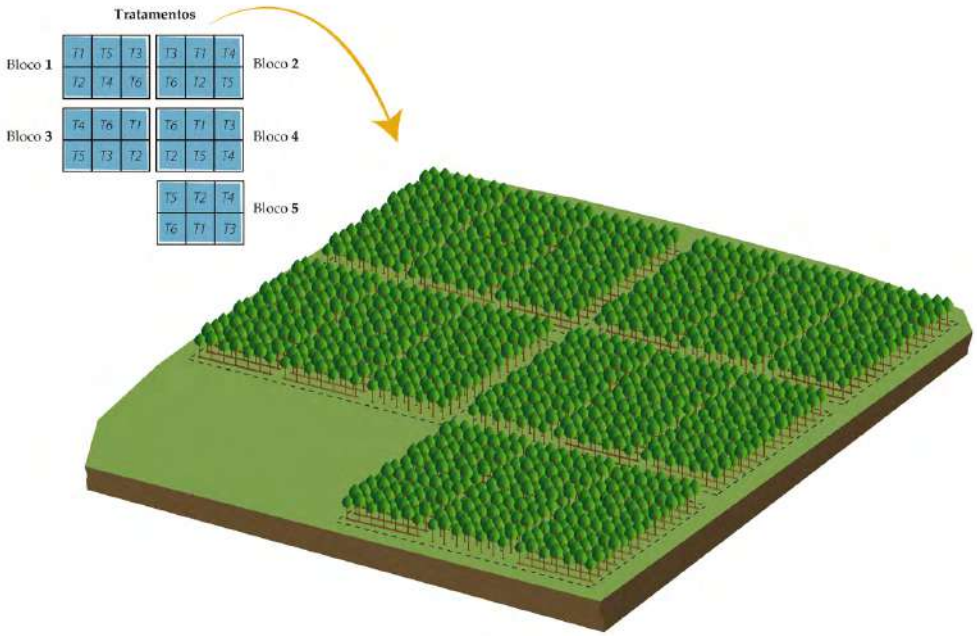


TABELA 17.2. Resultados obtidos para a altura (em m), diâmetro de copa (em m) e diâmetro à altura do peito (em cm) da espécie *Eucalyptus urophylla* sob efeito de diferentes tipos de preparo de solo e métodos de adubação (continua).

Fator A	Fator B	Bloco	Altura	Diâmetro de copa	Diâmetro à altura do peito
a1	b1	1	2,40	2,05	1,70
a1	b1	2	2,60	2,40	1,50
a1	b1	3	2,70	1,85	1,60
a1	b1	4	2,80	1,97	1,80
a1	b1	5	3,20	2,03	1,70
a1	b2	1	2,10	2,15	1,90
a1	b2	2	2,20	1,77	1,40
a1	b2	3	2,60	2,23	1,50
a1	b2	4	2,60	1,63	1,40
a1	b2	5	2,90	1,77	1,60
a2	b1	1	3,00	2,14	1,70
a2	b1	2	3,10	2,04	1,80
a2	b1	3	3,70	2,00	1,60
a2	b1	4	3,80	1,67	1,50
a2	b1	5	3,90	1,69	1,60
a2	b2	1	3,10	1,98	1,70
a2	b2	2	3,20	1,84	1,70
a2	b2	3	3,60	1,89	1,80
a2	b2	4	3,80	1,93	1,60
a2	b2	5	4,10	2,04	1,30

TABELA 17.2. Resultados obtidos para a altura (em m), diâmetro de copa (em m) e diâmetro à altura do peito (em cm) da esécie *Eucalyptus urophylla* sob efeito de diferentes tipos de preparo de solo e métodos de adubação (continuação).

Fator A	Fator B	Bloco	Altura	Diâmetro de Copa	Diâmetro à altura do peito
a3	b1	1	5,60	2,15	1,70
a3	b1	2	5,30	2,16	1,40
a3	b1	3	5,50	1,86	1,50
a3	b1	4	5,60	1,83	1,60
a3	b1	5	4,80	1,78	1,50
a3	b2	1	3,50	1,98	1,60
a3	b2	2	4,20	1,76	1,40
a3	b2	3	4,30	1,98	1,30
a3	b2	4	4,60	1,99	1,70
a3	b2	5	4,70	1,89	1,80

Os dados experimentais, na escala original, foram submetidos à análise de variância e os resultados obtidos estão apresentados na Tabela 17.3.

Comandos utilizados para obtenção da ANOVA em delineamento em blocos ao acaso com arranjo fatorial dos tratamentos:

ARQUIVO: 34 - FUNÇÃO FATORIAL PARA ANOVA EM FATORIAL COM DBA

```
>using DataFrames, CSV, Statistics, Distributions

>X=CSV.read("Z:/Alexandre/Julia/Exemplos/fatorial2.csv", DataFrame)
```



```
julia> X=CSV.read("Z:/Alexandre/Julia/Exemplos/fatorial2.csv", DataFrame)
30×6 DataFrame
| Row | FatorA | FatorB | Bloco | h       | dc      | dap     |
|     | String | String | Int64 | Float64 | Float64 | Float64 |
|-----|--------|--------|-------|---------|---------|---------|
| 1   | a1     | b1     | 1     | 2.4     | 2.05    | 1.7     |
| 2   | a1     | b1     | 2     | 2.6     | 2.4     | 1.5     |
| 3   | a1     | b1     | 3     | 2.7     | 1.85    | 1.6     |
| 4   | a1     | b1     | 4     | 2.8     | 1.97    | 1.8     |
| 5   | a1     | b1     | 5     | 3.2     | 2.03    | 1.7     |
⋮
| 25  | a3     | b1     | 5     | 4.8     | 1.78    | 1.5     |
| 26  | a3     | b2     | 1     | 3.5     | 1.98    | 1.6     |
| 27  | a3     | b2     | 2     | 4.2     | 1.76    | 1.4     |
| 28  | a3     | b2     | 3     | 4.3     | 1.98    | 1.3     |
| 29  | a3     | b2     | 4     | 4.6     | 1.99    | 1.7     |
| 30  | a3     | b2     | 5     | 4.7     | 1.89    | 1.8     |
```

```
>function FATORIAL2F(FatorA,FatorB,Bloco,Variavel)
    X=DataFrame()
    X.FatorA=FatorA
    X.FatorB=FatorB
    X.Bloco=Bloco
```

```
X.Y=Variavel
X
I= length(unique(X.FatorA))
J= length(unique(X.FatorB))
K= length(unique(X.Bloco))
GLBlocos=K-1
GLTrat=I*J-1
GLTotal=I*J*K-1
GLResiduo=(I*J-1)*(K-1)
FC=((sum(X.Y))^2)/(I*J*K)
AB= by(X, :Bloco) do df
(m = mean(df.Y), s² = var(df.Y), total= sum(df.Y), ss=sum((df.Y).^2))
end
SQBlocos=((sum(AB.total.^2))/(I*J))-FC
AT= by(X, [:FatorA, :FatorB]) do df
(m = mean(df.Y), s² = var(df.Y), total= sum(df.Y), ss=sum((df.Y).^2))
end
SQTrat=((sum(AT.total.^2))/(K))-FC
SQTotal=(sum(X.Y.^2))-FC
SQResiduo=SQTotal-SQBlocos-SQTrat
QMBlocos=SQBlocos/GLBlocos
QMTrat=SQTrat/GLTrat
QMResiduo=SQResiduo/GLResiduo
FBlocos=QMBlocos/QMResiduo
FTrat=QMTrat/QMResiduo
AlphaBlocos= ccdf(FDist(GLBlocos, GLResiduo),abs(FBlocos))
AlphaTrat= ccdf(FDist(GLTrat, GLResiduo),abs(FTrat))
R²=SQTrat/SQTotal*100
CV=sqrt(QMResiduo)/mean(X.Y)*100
GLFatorA=I-1
GLFatorB=J-1
GLFatorAFatorB=(I-1)*(J-1)
AFA= by(X, :FatorA) do df
(m = mean(df.Y), s² = var(df.Y), total= sum(df.Y), ss=sum((df.Y).^2))
end
SQFatorA=((sum(AFA.total.^2))/(J*K))-FC
AFB= by(X, :FatorB) do df
(m = mean(df.Y), s² = var(df.Y), total= sum(df.Y), ss=sum((df.Y).^2))
end
SQFatorB=((sum(AFB.total.^2))/(I*K))-FC
SQFatorAFatorB=SQTrat-SQFatorA-SQFatorB
QMFatorA=SQFatorA/GLFatorA
QMFatorB=SQFatorB/GLFatorB
QMFatorAFatorB=SQFatorAFatorB/GLFatorAFatorB
FFatorA=QMFatorA/QMResiduo
FFatorB=QMFatorB/QMResiduo
FFatorAFatorB=QMFatorAFatorB/QMResiduo
AlphaFatorAFatorB=ccdf(FDist(GLFatorAFatorB, GLResiduo),abs(FFatorAFatorB))
AlphaFatorA= ccdf(FDist(GLFatorA, GLResiduo),abs(FFatorA))
AlphaFatorB= ccdf(FDist(GLFatorB, GLResiduo),abs(FFatorB))
```

```
    println("GLBlocos: ", GLBlocos)
    println("GLTrat: ", GLTrat)
    println("GLResíduo: ", GLResiduo)
    println("GLTotal: ", GLTotal)
    println("SQBlocos: ", SQBlocos)
    println("SQTrat: ", SQTrat)
    println("SQResíduo: ", SQResiduo)
    println("SQTotal: ", SQTotal)
    println("QMBlocos: ", QMBlocos)
    println("QMTrat: ", QMTrat)
    println("QMResíduo: ", QMResiduo)
    println("FBlocos: ", FBlocos)
    println("FTrat: ", FTrat)
    println("AlphaBlocos: ", AlphaBlocos)
    println("AlphaTrat: ", AlphaTrat)
    println("GLFatorA: ", GLFatorA)
    println("GLFatorB: ", GLFatorB)
    println("GLFatorAFatorB: ", GLFatorAFatorB)
    println("SQFatorA: ", SQFatorA)
    println("SQFatorB: ", SQFatorB)
    println("SQFatorAFatorB: ", SQFatorAFatorB)
    println("QMFatorA: ", QMFatorA)
    println("QMFatorB: ", QMFatorB)
    println("QMFatorAFatorB: ", QMFatorAFatorB)
    println("FFatorAFatorB: ", FFatorAFatorB)
    println("FFatorA: ", FFatorA)
    println("FFatorB: ", FFatorB)
    println("AlphaFatorAFatorB: ", AlphaFatorAFatorB)
    println("AlphaFatorA: ", AlphaFatorA)
    println("AlphaFatorB: ", AlphaFatorB)
end

#FATORIAL2F(FatorA,FatorB,Bloco,Variavel)
>FATORIAL2F(X.FatorA,X.FatorB,X.Bloco,X.h)
GLBlocos: 4
GLTrat: 5
GLResíduo: 20
GLTotal: 29
SQBlocos: 1.89
SQTrat: 27.62
SQResíduo: 1.62
SQTotal: 31.14
QMBlocos: 0.47
QMTrat: 5.52
QMResíduo: 0.081
FBlocos: 5.83
FTrat: 68.08
AlphaBlocos: 0.0028
AlphaTrat: 7.27e-12
GLFatorA: 2
```

```
GLFatorB: 1
GLFatorAFatorB: 2
SQFatorA: 24.42
SQFatorB: 1.41
SQFatorAFatorB: 1.79
QMFatorA: 12.21
QMFatorB: 1.41
QMFatorAFatorB: 0.90
FFatorAFatorB: 11.06
FFatorA: 150.47
FFatorB: 17.36
AlphaFatorAFatorB: 0.00058
AlphaFatorA: 8.83e-13
AlphaFatorB: 0.000477
```

Para a altura das mudas, o teste F para avaliar os efeitos dos blocos foi significante, indicando que eles foram efetivos para o controle das diferenças entre as unidades experimentais. O teste F na análise de variância para a interação foi significante, indicando que os fatores testados são dependentes. Ambos os fatores principais também foram significantes. A interpretação de um delineamento com arranjo fatorial dos tratamentos sempre deve ser iniciada pela avaliação da interação. O resumo da análise de variância está apresentado na Tabela 17.3.

TABELA 17.3. Resultado da análise de variância para avaliar o crescimento em altura (em m) da espécie florestal *Eucalyptus urophylla* sob o efeito de diferentes tipos de preparo de solo e métodos de adubação.

Fonte de variação	Graus de Liberdade	Soma de Quadrado	Quadrado Médio	F calculado	F tabelado (α = 5%)	Alpha (%)
Blocos	4	1,893	0,473	5,83 **	2,87	0,28
Fator A	2	24,416	12,208	150,47 **	3,49	< 0,01
Fator B	1	1,408	1,408	17,36 **	4,35	0,05
A x B	2	1,795	0,897	11,06 **	3,49	0,06
Erro Exp.	20	1,623	0,081			
Total	29	31,135				

** significante ao nível de α= 0,01 e A x B = Interação entre fatores

A comparação de médias depende do resultado da avaliação da interação entre os fatores. Se esta for não significante, somente as médias dos fatores principais serão comparadas. Por outro lado, se a interação for significante as médias de cada fator dentro dos níveis do outro fator devem ser comparadas separadamente.

Como a análise de variância revelou que a interação foi significante, o teste de Tukey foi aplicado para comparação das médias entre os níveis de cada fator. Para a aplicação do teste de comparação de médias, inicialmente os dados foram organizados para comparar as médias dos níveis do fator A (tipos de preparo do solo) dentro de cada nível do fator B (métodos de adubação), bem como o seu inverso, ou seja, as médias dos níveis do fator B dentro de cada nível do fator A.

Comandos utilizados para obtenção do teste de Tukey entre os níveis dos fatores:

ARQUIVO: 34 - FUNÇÃO FATORIAL PARA ANOVA EM FATORIAL COM DBA

```
>Dados=sort(by(X, [:FatorB,:FatorA], :h => mean));
>nFA=length(levels(X.FatorA));
>nFB=length(levels(X.FatorB));
#XA: Tabela de médias dos níveis do Fator A para os níveis de B.
>Medias=[Dados[1:nFA,3] Dados[(nFA+1):end,3]];
>Tabela=convert(DataFrame, Medias);
>TratFB=levels(Dados.FatorB);
>Tabela=names!(Tabela,[Symbol(name) for name in TratFB]);
>TratFA=levels(Dados.FatorA);
>XA=insertcols!(Tabela, 1, :FatorA => TratFA)
```

```
julia> XA=insertcols!(Tabela, 1, :FatorA => TratFA)
3×3 DataFrame
| Row | FatorA | b1      | b2      |
|     | String | Float64 | Float64 |
|-----|--------|---------|---------|
| 1   | a1     | 2.74    | 2.48    |
| 2   | a2     | 3.5     | 3.56    |
| 3   | a3     | 5.36    | 4.26    |
```

```
#XB: Tabela de médias dos níveis do Fator B para os níveis de A.
>Medias=[Dados[1:nFA,3] Dados[(nFA+1):end,3]]';
>Tabela=convert(DataFrame, Medias);
>Tabela=names!(Tabela,[Symbol(name) for name in TratFA]);
>XB=insertcols!(Tabela, 1, :FatorB => TratFB)
```

```
julia> XB=insertcols!(Tabela, 1, :FatorB => TratFB)
2×4 DataFrame
| Row | FatorB | a1      | a2      | a3      |
|     | String | Float64 | Float64 | Float64 |
|-----|--------|---------|---------|---------|
| 1   | b1     | 2.74    | 3.5     | 5.36    |
| 2   | b2     | 2.48    | 3.56    | 4.26    |
```

```
#Teste de médias de Tukey
>using DataFrames, Statistics, PyPlot, Distributions, StatsFuns

>function TukeyFat(NumeroRepeticao,Tratamento,Medias,GLResiduo,QMResiduo,Alpha)
    I= length(Medias)
    J=NumeroRepeticao
    q=srdistinvccdf(I,GLResiduo,Alpha)
    DMS=q*sqrt(QMResiduo / 2 * (2 / J))
    A=Medias
    B=A'
    DIF = [A[l]-B[c] for l in 1:I, c in 1:I]
    DIFA=[if abs(DIF[i,c])>=DMS diferença="*" elseif abs(DIF[i,c])<DMS
diferença="ns" end for i in 1:I, c in 1:I]
    Trat=levels(Tratamento)
    NomeTratamentos=convert(Array,Tratamento)
```

```
    TabelaHipótese=convert(DataFrame, DIFA)
    TabelaHipótese=names!(TabelaHipótese,[Symbol(name) for name in
NomeTratamentos])
    Tratamentos=names(TabelaHipótese)
    TabelaSignificânciaDMS=insertcols!(TabelaHipótese, 1, :Tratamento =>
Tratamentos)
    TabelaDMS=convert(DataFrame, DIF)
    TabelaDMS=names!(TabelaDMS,[Symbol(name) for name in NomeTratamentos])
    TabelaDMS=insertcols!(TabelaDMS, 1, :Tratamento => Tratamentos)
    println("DMS: ", DMS)
    println("Tabela DMS: ", TabelaDMS)
    println("Tabela de hipóteses DMS: ", TabelaSignificânciaDMS)
end

#Teste para as médias: diferença entre os níveis a1, a2 e a3 dentro do nível b1.
>XA.b1
2.74
3.5
5.36
>TukeyFat(5,XA.FatorA,XA.b1,20,0.081,0.05)
```

```
julia> TukeyFat(5,XA.FatorA,XA.b1,20,0.081,0.05)
DMS: 0.4553967249429451
Tabela DMS: 3×4 DataFrame
| Row | Tratamento | a1      | a2      | a3      |
|     | String     | Float64 | Float64 | Float64 |
|-----|------------|---------|---------|---------|
| 1   | a1         | 0.0     | -0.76   | -2.62   |
| 2   | a2         | 0.76    | 0.0     | -1.86   |
| 3   | a3         | 2.62    | 1.86    | 0.0     |
Tabela de hipóteses DMS: 3×4 DataFrame
| Row | Tratamento | a1     | a2     | a3     |
|     | String     | String | String | String |
|-----|------------|--------|--------|--------|
| 1   | a1         | ns     | *      | *      |
| 2   | a2         | *      | ns     | *      |
| 3   | a3         | *      | *      | ns     |
```

```
#Teste para as médias: diferença entre os níveis a1, a2 e a3 dentro do nível b2.
>XA.b2
2.48
3.56
4.26
>TukeyFat(5,XA.FatorA,XA.b2,20,0.081,0.05)
```

```
julia> TukeyFat(5,XA.FatorA,XA.b2,20,0.081,0.05)
DMS: 0.4553967249429451
Tabela DMS: 3×4 DataFrame
| Row | Tratamento | a1      | a2      | a3      |
|     | String     | Float64 | Float64 | Float64 |
|-----|------------|---------|---------|---------|
| 1   | a1         | 0.0     | -1.08   | -1.78   |
| 2   | a2         | 1.08    | 0.0     | -0.7    |
| 3   | a3         | 1.78    | 0.7     | 0.0     |
Tabela de hipóteses DMS: 3×4 DataFrame
| Row | Tratamento | a1     | a2     | a3     |
|     | String     | String | String | String |
|-----|------------|--------|--------|--------|
| 1   | a1         | ns     | *      | *      |
| 2   | a2         | *      | ns     | *      |
| 3   | a3         | *      | *      | ns     |
```

```
#Teste para as médias: diferença entre os níveis b1 e b2 dentro do nível a1.
>XB.a1
2.74
2.48
>TukeyFat(5,XB.FatorB,XB.a1,20,0.081,0.05)
```

```
julia> TukeyFat(5,XB.FatorB,XB.a1,20,0.081,0.05)
DMS: 0.37547340093132586
Tabela DMS: 2×3 DataFrame
| Row | Tratamento | b1      | b2      |
|     | String     | Float64 | Float64 |
|-----|------------|---------|---------|
| 1   | b1         | 0.0     | 0.26    |
| 2   | b2         | -0.26   | 0.0     |
Tabela de hipóteses DMS: 2×3 DataFrame
| Row | Tratamento | b1     | b2     |
|     | String     | String | String |
|-----|------------|--------|--------|
| 1   | b1         | ns     | ns     |
| 2   | b2         | ns     | ns     |
```

```
#Teste para as médias: diferença entre os níveis b1 e b2 dentro do nível a2.
>XB.a2
3.50
3.56
>TukeyFat(5,XB.FatorB,XB.a2,20,0.081,0.05)
```

```
julia> TukeyFat(5,XB.FatorB,XB.a2,20,0.081,0.05)
DMS: 0.37547340093132586
Tabela DMS: 2×3 DataFrame
| Row | Tratamento | b1      | b2      |
|     | String     | Float64 | Float64 |
|-----|------------|---------|---------|
| 1   | b1         | 0.0     | -0.06   |
| 2   | b2         | 0.06    | 0.0     |
Tabela de hipóteses DMS: 2×3 DataFrame
| Row | Tratamento | b1     | b2     |
|     | String     | String | String |
|-----|------------|--------|--------|
| 1   | b1         | ns     | ns     |
| 2   | b2         | ns     | ns     |
```

```
#Teste para as médias: diferença entre os níveis b1 e b2 dentro do nível a3.
>XB.a3
5.36
4.26
>TukeyFat(5,XB.FatorB,XB.a3,20,0.081,0.05)
```

```
julia> TukeyFat(5,XB.FatorB,XB.a3,20,0.081,0.05)
DMS: 0.37547340093132586
Tabela DMS: 2×3 DataFrame
| Row | Tratamento | b1      | b2      |
|     | String     | Float64 | Float64 |
|-----|------------|---------|---------|
| 1   | b1         | 0.0     | 1.1     |
| 2   | b2         | -1.1    | 0.0     |
Tabela de hipóteses DMS: 2×3 DataFrame
| Row | Tratamento | b1     | b2     |
|     | String     | String | String |
|-----|------------|--------|--------|
| 1   | b1         | ns     | *      |
| 2   | b2         | *      | ns     |
```

```
#Construção de um gráfico
#Níveis a1, a2 e a3 dentro do nível b1.
>DMS=0.46

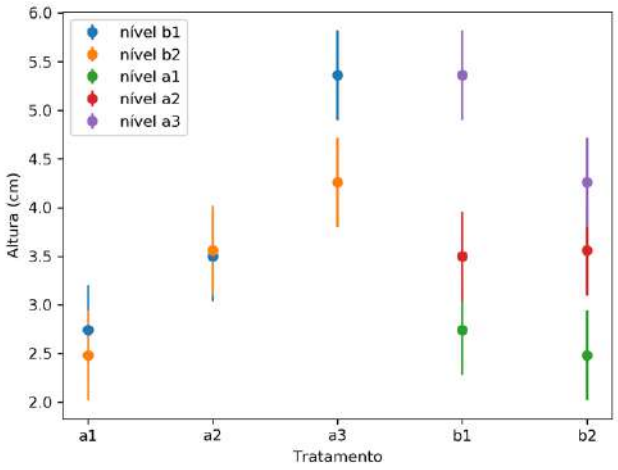
>errorbar(XA.FatorA,XA.b1,DMS,fmt="o",label="nível b1")
>xlabel("Tratamento")
>ylabel("Altura (cm)")
>legend(loc="best")

#Níveis a1, a2 e a3 dentro do nível b2
>errorbar(XA.FatorA,XA.b2,DMS,fmt="o",label="nível b2")
>legend(loc="best")

#Níveis b1 e b2 dentro do nível a1.
>DMS: 0.38
>errorbar(XB.FatorB,XB.a1,DMS,fmt="o",label="nível a1")
>legend(loc="best")

#Níveis b1 e b2 dentro do nível a2
>errorbar(XB.FatorB,XB.a2,DMS,fmt="o",label="nível a2")
>legend(loc="best")

#Níveis b1 e b2 dentro do nível a3
>errorbar(XB.FatorB,XB.a3,DMS,fmt="o",label="nível a3")
>legend(loc="best")
```



Um resumo dos resultados obtidos no teste de Tukey para comparação de médias está apresentado na Tabela 17.4.

TABELA 17.4. Resultados do teste de Tukey na comparação de médias para avaliar o crescimento inicial *Eucalyptus urophylla* sob o efeito de diferentes tipos de preparo de solo e métodos de adubação.

Tipo de preparo de solo	Método de adubação	
	No entorno da muda (b1)	Na linha de plantio (b2)
Sem preparo (a1)	2,74 cA	2,48 cA
Subsolagem (a2)	3,50 bA	3,56 bA
Subsolagem e gradagem (a3)	5,36 aA	4,26 aB

Médias seguidas pelas mesmas letras minúsculas nas colunas e maiúsculas nas linhas, não diferem estatisticamente ao nível de 0,05 de significância pelo teste de Tukey.

Na avaliação de cada método de adubação dentro de cada preparo do solo, os dois métodos de adubação não apresentaram diferenças para a ausência de preparo do solo e subsolagem. Para o solo preparado com subsolagem e gradagem, a adubação em torno da muda mostrou-se superior a adubação na linha.

Na avaliação de cada preparo do solo dentro de cada tipo de adubação, o tipo de preparo com subsolagem e gradagem foi superior e diferente dos demais tipos testados, para ambos os métodos de adubação.

Considerando todos os resultados, foi observado que o preparo do solo por meio da subsolagem e gradagem resultou em um maior crescimento em altura das mudas, quando combinada com adubação no entorno da muda. Assim, diante dos resultados experimentais observados para a variável altura sugere-se a aplicação deste tratamento.

Para as demais variáveis, sugere-se a efetuação dos cálculos e interpretação dos resultados pelos leitores.

CONCLUSÃO DA ANÁLISE DA VARIÂNCIA

» **Sempre se inicia pela interação**

Se o valor de F for significante para a interação, então, os fatores são dependentes e compara-se as médias de um fator para cada nível do outro fator.

Quando os fatores são independentes, ou seja, a interação não é significante compara-se as médias dos efeitos principais.

Comandos para se obter as médias dos efeitos principais:

ARQUIVO: 34 - FUNÇÃO FATORIAL PARA ANOVA EM FATORIAL COM DBA

```
>MédiasFatorA=by(X, :FatorA, :h => mean)
```

```
julia> MédiasFatorA=by(X, :FatorA, :h => mean)
3×2 DataFrame
| Row | FatorA | h_mean   |
|     | String | Float64  |
|-----|--------|----------|
| 1   | a1     | 2.61     |
| 2   | a2     | 3.53     |
| 3   | a3     | 4.81     |
```

```
>MédiasFatorB=by(X, :FatorB, :h => mean)
```

```
julia> MédiasFatorB=by(X, :FatorB, :h => mean)
2×2 DataFrame
| Row | FatorB | h_mean   |
|     | String | Float64  |
|-----|--------|----------|
| 1   | b1     | 3.86667  |
| 2   | b2     | 3.43333  |
```

Se a análise de variância indicou que os fatores são independentes, aplica-se o teste de comparação de médias apenas dentro de cada fator.

Um resumo geral sobre o delineamento em ensaios fatoriais está apresentado na Figura 17.13.

FIGURA 17.13. Resumo do Delineamento em ensaios fatoriais.

DELINEAMENTO EM ENSAIOS FATORIAIS

function FATORIAL2F(FatorA,FatorB,Bloco,Variavel)
Fatorial em DBA

$$Y_{ijk}=\mu+\tau_i+\beta_j+(\tau\beta)_{ij}+\epsilon_{ijk} \text{ (DIC)}$$

$$Y_{ijk}=\mu+\tau_i+\beta_j+(\tau\beta)_{ij}+ \Psi_k + \epsilon_{ijk} \text{ (DBA)}$$

Experimentos com arranjos fatoriais dos tratamentos são aqueles onde o efeito de dois ou mais fatores são investigados ao mesmo tempo.

Estes representam todas as combinações possíveis de vários conjuntos de tratamentos ou fatores.

As respostas são mais interessantes do que nos experimentos simples com apenas um fator, porque pode-se avaliar a dependência dos fatores.

São utilizados em qualquer tipo de delineamento experimental.

VANTAGENS

- Avaliação de mais de um fator no mesmo ensaio
- Avaliação da interação dos efeitos dos fatores
- Utilização mais racional dos recursos disponíveis

DESVANTAGENS

- Conforme aumenta o número de fatores, aumenta-se o número de tratamentos, o que pode dificultar a condução do experimento;
- Quando o número de fatores é elevado as interpretações práticas podem ser difíceis.

18 DELINEAMENTO EM PARCELAS SUBDIVIDIDAS

O arranjo dos tratamentos em parcelas subdivididas é uma variação do arranjo fatorial, com a finalidade de facilitar o manuseio dos tratamentos. Estes experimentos são utilizados quando em um mesmo ensaio deseja-se testar dois ou mais fatores, porém, em condições experimentais aninhadas.

Nas parcelas, por exemplo, são alocados os níveis do Fator A. Cada parcela é subdividida, por exemplo, para receber os níveis do Fator B. Logo, as parcelas para os níveis do fator A são maiores que as parcelas para os níveis do fator B.

Comandos para casualização de experimentos em parcelas subdivididas:

ARQUIVO: 35 - CASUALIZAR OS TRATAMENTOS EM PARCELA SUBDIVIDIDA

```
>using Random, DataFrames, CSV
>X=CSV.read("Z:/Alexandre/Julia/Exemplos/aleatPS.csv", DataFrame)
```

julia

```
julia> X=CSV.read("Z:/Alexandre/Julia/Exemplos/aleatPS.csv", DataFrame)
3×4 DataFrame
| Row | FatorA | B1     | B2     | B3     |
|     | String | String | String | String |
|-----|--------|--------|--------|--------|
| 1   | A1     | b1     | b2     | b3     |
| 2   | A2     | b1     | b2     | b3     |
| 3   | A3     | b1     | b2     | b3     |
```

```
>function PS(X)
    B=X.FatorA
    Linha=B[shuffle(1:end), :]
    I=length(unique(X.FatorA))
    A=convert(Matrix, X[:,2:end])
    Coluna=A
    for i=1:I
    Coluna[i,:]=A[i,shuffle(1:end)]
    end
    Coluna
    Desenho=[Linha Coluna]
    Desenho=convert(DataFrame, Desenho)
end

>PS(X)
```

```
julia> PS(X)
3×4 DataFrame
| Row | x1     | x2     | x3     | x4     |
|     | String | String | String | String |
|-----|--------|--------|--------|--------|
| 1   | A2     | b3     | b1     | b2     |
| 2   | A3     | b1     | b3     | b2     |
| 3   | A1     | b2     | b3     | b1     |
```

O resultado da função gerou a estrutura para um delineamento em blocos ao acaso. A coluna x1 refere-se aos tratamentos distribuídos nas parcelas (níveis do Fator A) e nas colunas x2, x3 e x4 os tratamentos distribuídos nas sub-parcelas (níveis do Fator B). Para vários blocos, a função anterior pode ser repetida. Por exemplo, se o experimento for composto por quatro 4 blocos, pode-se utilizar as seguintes funções para obter a estrutura, a qual resultará um *dataframe* para cada bloco.

Comandos para casualização nos blocos em delineamento com parcelas subdivididas:

```
>NBlocos=4 #Número de blocos
>desenho = for i=1: NBlocos
    println(PS(X))
end
```

```
3×4 DataFrame
| Row | x1     | x2     | x3     | x4     |
|     | String | String | String | String |
|-----|--------|--------|--------|--------|
| 1   | A3     | b3     | b1     | b2     |
| 2   | A2     | b3     | b1     | b2     |
| 3   | A1     | b2     | b1     | b3     |
3×4 DataFrame
| Row | x1     | x2     | x3     | x4     |
|     | String | String | String | String |
|-----|--------|--------|--------|--------|
| 1   | A3     | b2     | b3     | b1     |
| 2   | A2     | b2     | b3     | b1     |
| 3   | A1     | b3     | b1     | b2     |
3×4 DataFrame
| Row | x1     | x2     | x3     | x4     |
|     | String | String | String | String |
|-----|--------|--------|--------|--------|
| 1   | A2     | b2     | b3     | b1     |
| 2   | A3     | b3     | b1     | b2     |
| 3   | A1     | b3     | b1     | b2     |
3×4 DataFrame
| Row | x1     | x2     | x3     | x4     |
|     | String | String | String | String |
|-----|--------|--------|--------|--------|
| 1   | A2     | b3     | b2     | b1     |
| 2   | A3     | b1     | b2     | b3     |
| 3   | A1     | b3     | b1     | b2     |
```

Devido à variação nos tamanhos das parcelas e, também, no número de graus de liberdade resultantes, os fatores A e B são comparados com distintos níveis de acurácia, determinando um erro experimental para as parcelas e outro para as subparcelas. Em alguns casos o erro experimental associado às parcelas é maior que o erro experimental associado às subparcelas, devido o quadrado médio do resíduo ser menor que o quadrado médio do resíduo da parcela. Assim, os níveis do fator que requerem maior detalhamento no experimento são alocados na subparcela, devido ao controle local ser mais rigoroso. Se isso não for efetuado, o fator que é atribuído às parcelas ou subparcelas é escolhido para facilitar a instalação e condução do experimento.

Experimentos com parcelas subdivididas são frequentes na área florestal, especialmente nos ensaios de espaçamento x adubação, genótipos x espaçamento e preparo do solo x adubação, por exemplo.

Assim como para os ensaios com arranjo fatorial dos tratamentos, o modelo estatístico para parcelas subdividas dependerá dos princípios da experimentação para o qual o experimento foi estruturado, que pode, por exemplo, ser em DIC, DBA e DQL.

O modelo linear aditivo de um delineamento em parcelas subdivididas, conduzidos em DBA, é dado por:

$$Y_{ijk}=\mu+\tau_i+\omega_k+\delta_{ik}+\beta_j+(\tau\beta)_{ij}+\epsilon_{ijk}$$

Em que:

Y_{ijk} é o valor observado de uma variável aleatória na *k*-ésima repetição da combinação do *i*-ésimo nível do fator A com o *j*-ésimo nível do fator B.

μ usualmente corresponde à média geral.

τ_i é o efeito o *i*-ésimo nível do fator A.

δ_{ik} é o efeito residual das parcelas, caracterizado como componente do erro a.

β_j é o efeito o *j*-ésimo nível do fator B.

$(\tau\beta)_{ij}$ é o efeito da interação do *i*-ésimo nível do fator A com o *j*-ésimo nível do fator B.

ω_k é o efeito do *k*-ésimo bloco

ϵ_{ijk} são os efeitos residuais das subparcelas, caracterizados como componentes do erro b.

O esquema de análise de variância para um delineamento em blocos ao acaso com parcelas subdivididas é apresentado na Tabela 18.1.

TABELA 18.1. Esquema para análise de variância um experimento conduzido em delineamento em blocos ao acaso com parcelas subdivididas.

Causas de variação	Graus de Liberdade	Soma de Quadrados	Quadrados Médios	F calculado
Blocos (r)	$K-1$	$SQ_r=\frac{\sum_k Y_k^2}{I.J}-FC$	$QM_r=\frac{SQ_r}{GL_r}$	$\frac{QM_r}{QM_a}$
Fator A	$I-1$	$SQ_A=\frac{\sum_i Y_i^2}{J.K}-FC$	$QM_A=\frac{SQ_A}{GL_A}$	$\frac{QM_A}{QM_a}$
Erro exp. A	$(I-1)\cdot(K-1)$	$SQ_a=SQ_P-SQ_A-SQ_r$	$QM_a=\frac{SQ_a}{GL_a}$	
Parcelas (P)	$IK-1$	$SQ_P=\frac{\sum_i\sum_k Y_{ik}^2}{J}-FC$		
Fator B	$J-1$	$SQ_B=\frac{\sum_i Y_i^2}{J.K}-FC$	$QM_B=\frac{SQ_B}{GL_B}$	$\frac{QM_B}{QM_b}$
Interação A x B	$(I-1)\cdot(J-1)$	$SQ_A=SQ_t-SQ_A-SQ_B$	$QM_{A.B}=\frac{SQ_{A.B}}{GL_{A.B}}$	$\frac{QM_{A.B}}{QM_b}$
Erro exp. B	$I\cdot(J-1)\cdot(K-1)$	$SQ_b=SQ_{Total}-SQ_P-SQ_B-SQ_{A.B}$	$QM_b=\frac{SQ_b}{GL_b}$	
Total	$I\cdot I\cdot K-1$	$SQ_{total}=\sum_i\sum_j\sum_k Y_{ijk}^2-FC$		

$FC=\frac{\left(\sum_i\sum_j\sum_k Y_{ijk}\right)^2}{IJK}$, fator de correção para a média.

K é o número de blocos.
I é o número de níveis do fator A.
J é o número de níveis do fator B.

Principais características

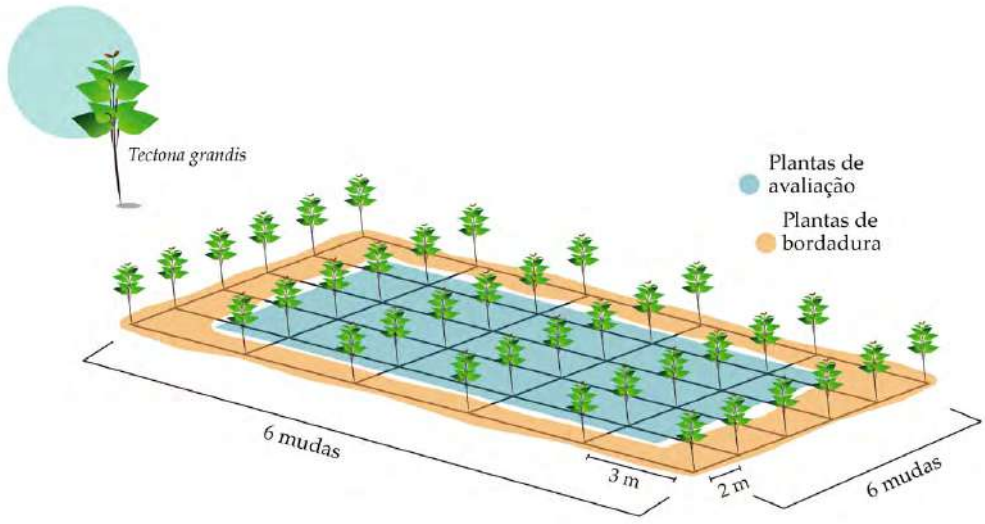
Como características principais dos delineamentos com arranjo dos tratamentos em parcelas subdivididas pode-se citar:

- » o fator aplicado à subparcela tem mais ênfase do que o fator aplicado à parcela devido este ter mais repetições.
- » o erro experimental associado às parcelas é usualmente maior do que o erro associado às subparcelas.
- » como existem dois erros experimentais, existirão dois coeficientes de variação, um para cada erro experimental.
- » o arranjo em parcelas subdivididas propicia o uso mais racional dos recursos experimentais, pois permite uma instalação mais racional e mais eficiente dos tratamentos a nível de campo.
- » as características descritas nos delineamentos com arranjo fatorial dos tratamentos também se aplicam aos delineamentos com arranjo em parcelas subdivididas.

O exemplo abaixo consiste no mesmo conjunto de tratamentos referentes ao experimento utilizado no delineamento em ensaios fatoriais, porém, neste caso a espécie avaliada foi a *Tectona grandis* em uma parcela de 6x6 plantas.

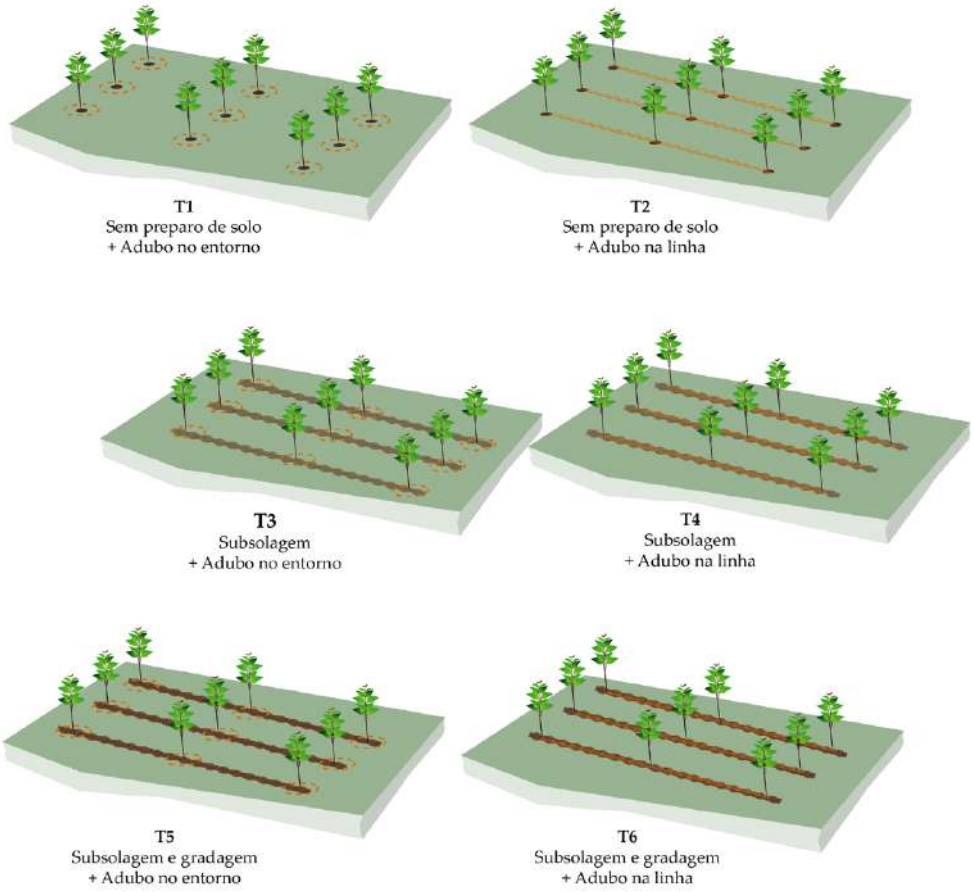
A unidade amostral foi composta com 36 plantas em um espaçamento de 3m x 2m (Figura 18.1). Uma linha de bordadura foi plantada, totalizando 12 plantas em cada unidade amostral.

FIGURA 18.1. Unidade amostral com mudas de *Tectona grandis* utilizadas para avaliar o crescimento inicial sob efeito de diferentes tipos de preparo de solo e métodos de adubação.



Do mesmo modo que no exemplo anterior, em ensaios fatoriais, os tratamentos foram definidos conforme a Figura 18.2.

FIGURA 18.2. Tratamentos testados para avaliar o crescimento inicial da espécie florestal *Tectona grandis* sob o efeito de diferentes tipos de preparo de solo e métodos de adubação.



Visando facilitar a distribuição dos tratamentos nas unidades experimentais, optou-se por utilizar parcelas subdivididas (Figura 18.3), sendo utilizados 6 blocos. Os tipos de preparo de solo foram aplicados às parcelas e os métodos de adubação às subparcelas. O experimento instalado está representado na Figura 18.4. Adicionalmente, cada bloco foi representado nas Figuras 18.5 a 18.10.

Dois anos após a instalação do experimento (Figura 18.11) a altura das mudas foi mensurada com uma régua telescópica, o diâmetro à altura do peito (DAP) com uma suta e o diâmetro de copa com uma trena. Os dados obtidos estão apresentados na Tabela 18.2.

TABELA 18.2. Resultados experimentais obtidos para a altura (em m), diâmetro de copa (em m) e diâmetro à altura do peito (em cm) da espécie *Tectona grandis* submetidas a diferentes tipos de preparo de solo e métodos de adubação.

Fator A	Fator B	Bloco	Altura	Diâmetro de copa	Diâmetro à altura do peito
a1	b1	1	7,20	1,60	3,40
a1	b1	2	7,00	1,56	4,00
a1	b1	3	7,10	1,74	4,20
a1	b1	4	6,70	1,66	4,40
a1	b1	5	6,40	1,59	4,70
a1	b1	6	6,60	1,61	4,50
a1	b2	1	6,70	1,50	4,00
a1	b2	2	6,80	1,71	4,10
a1	b2	3	6,20	1,62	4,60
a1	b2	4	6,90	1,77	4,60
a1	b2	5	7,00	1,78	4,70
a1	b2	6	6,80	1,68	4,00
a2	b1	1	7,40	2,01	5,10
a2	b1	2	6,30	1,78	5,10
a2	b1	3	6,80	1,69	5,30
a2	b1	4	6,80	1,67	5,50
a2	b1	5	7,20	1,91	5,60
a2	b1	6	6,50	1,87	5,40
a2	b2	1	6,30	1,84	4,70
a2	b2	2	6,30	1,69	4,80
a2	b2	3	6,70	1,67	4,90
a2	b2	4	6,90	1,78	5,30
a2	b2	5	7,00	1,75	5,80
a2	b2	6	6,40	1,69	5,50
a3	b1	1	7,00	1,67	6,70
a3	b1	2	7,40	1,66	6,80
a3	b1	3	6,60	1,74	6,90
a3	b1	4	6,80	1,73	7,30
a3	b1	5	6,50	1,93	7,40
a3	b1	6	6,70	1,88	7,20
a3	b2	1	6,70	1,92	6,90
a3	b2	2	7,20	1,87	6,80
a3	b2	3	7,40	2,02	6,90
a3	b2	4	6,90	1,70	7,20
a3	b2	5	6,50	1,59	7,20
a3	b2	6	7,10	1,68	7,00

Os dados da variável DAP foram testados quanto à homogeneidade de variâncias dos tratamentos pelo teste de Bartlett. O valor de χ^2 corrigido foi de 6,72, sendo não significante ao nível de 0,05 de significância. Assim, pode-se concluir que as variâncias dos tratamentos são homogêneas.

FIGURA 18.3. Tratamentos testados para avaliar o crescimento inicial da espécie florestal *Tectona grandis* sob o efeito de diferentes tipos de preparo de solo e métodos de adubação.

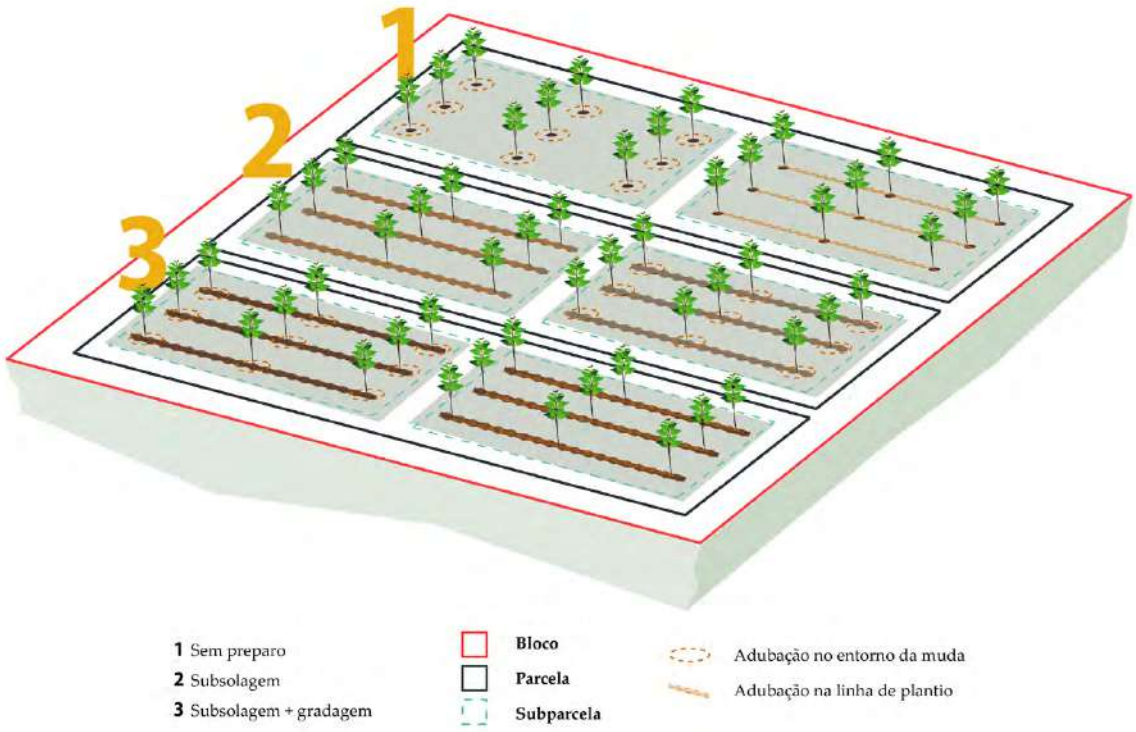


FIGURA 18.4. Experimento instalado em campo para avaliar o crescimento inicial da espécie florestal *Tectona grandis* sob o efeito de diferentes tipos de preparo de solo e métodos de adubação com delineamento de blocos ao acaso em esquema de parcelas subdivididas.

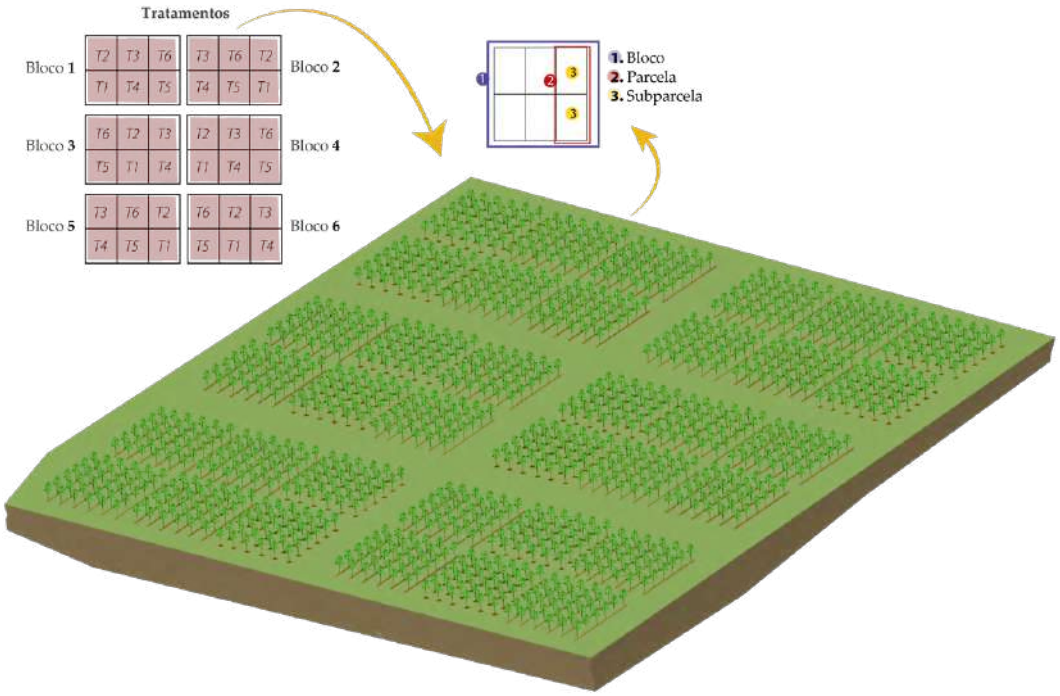


FIGURA 18.5. Experimento instalado em campo para avaliar o crescimento inicial da espécie florestal *Tectona grandis* sob o efeito de diferentes tipos de preparo de solo e métodos de adubação com delineamento de blocos ao acaso em esquema de parcelas subdivididas - bloco 1.

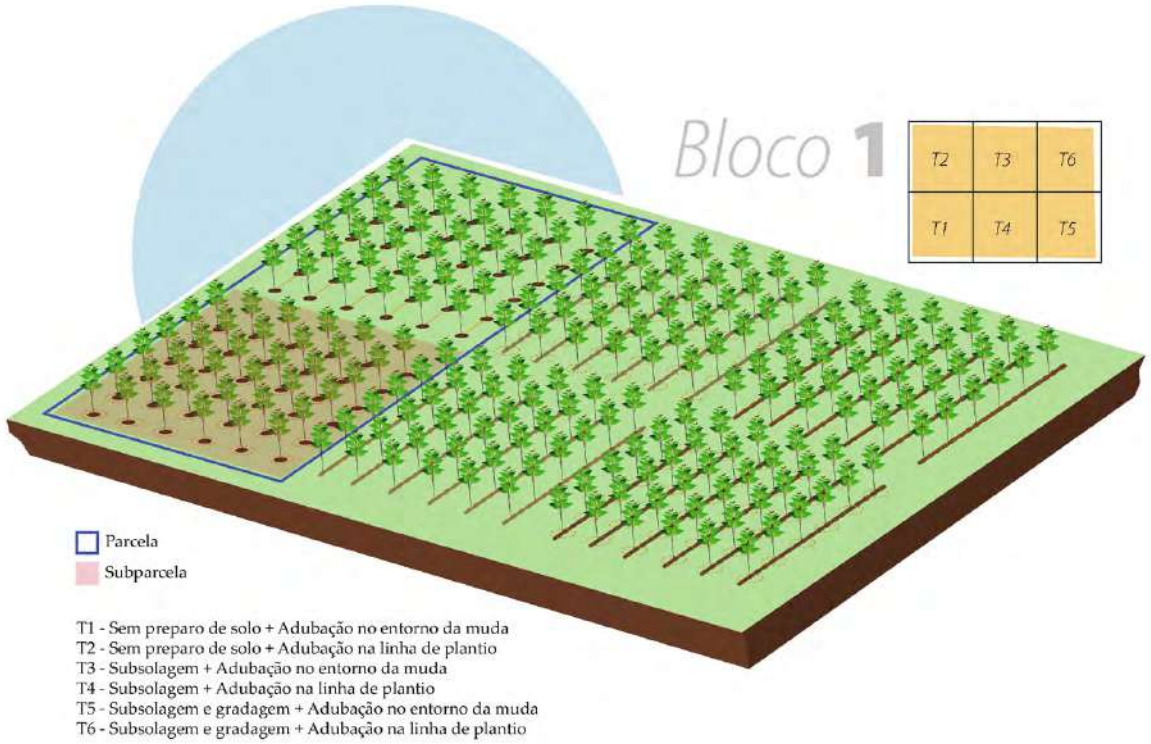


FIGURA 18.6. Experimento instalado em campo para avaliar o crescimento inicial da espécie florestal *Tectona grandis* sob o efeito de diferentes tipos de preparo de solo e métodos de adubação com delineamento de blocos ao acaso em esquema de parcelas subdivididas - bloco 2.

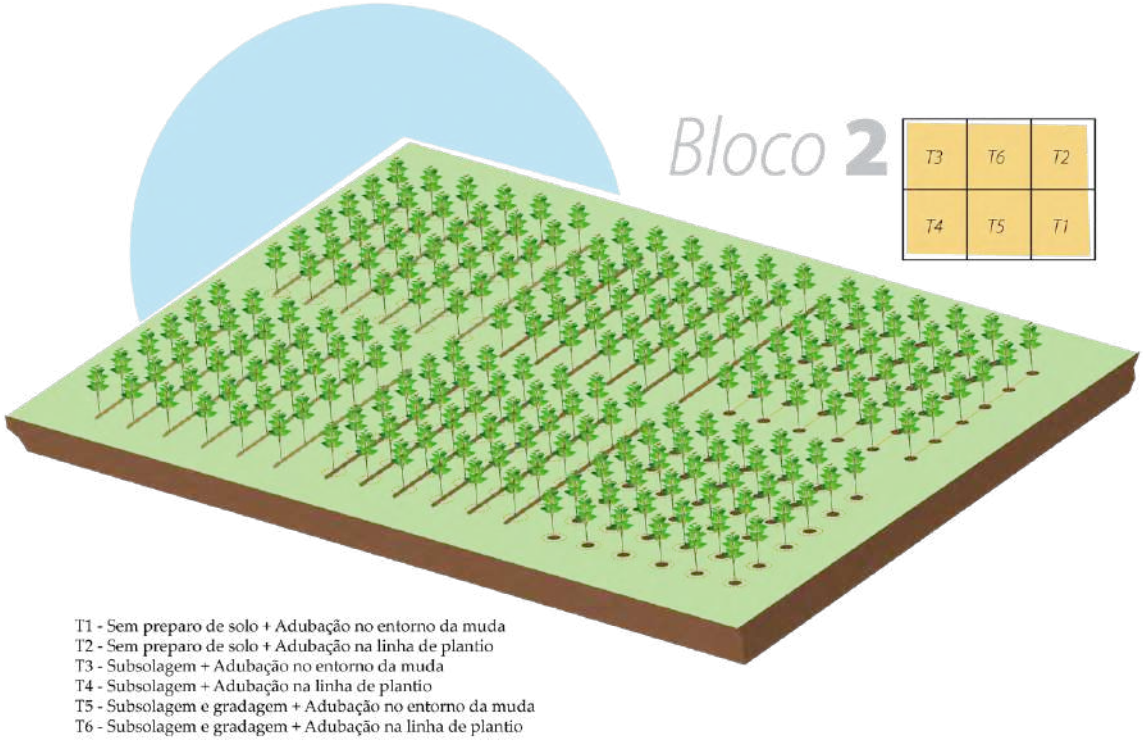
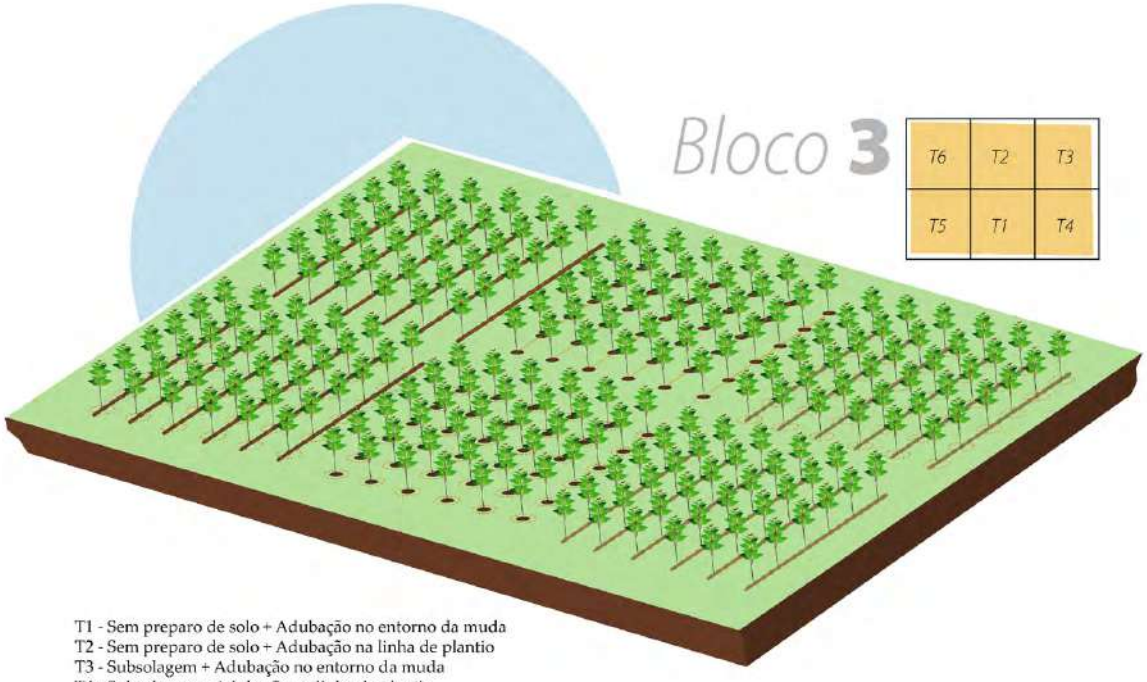
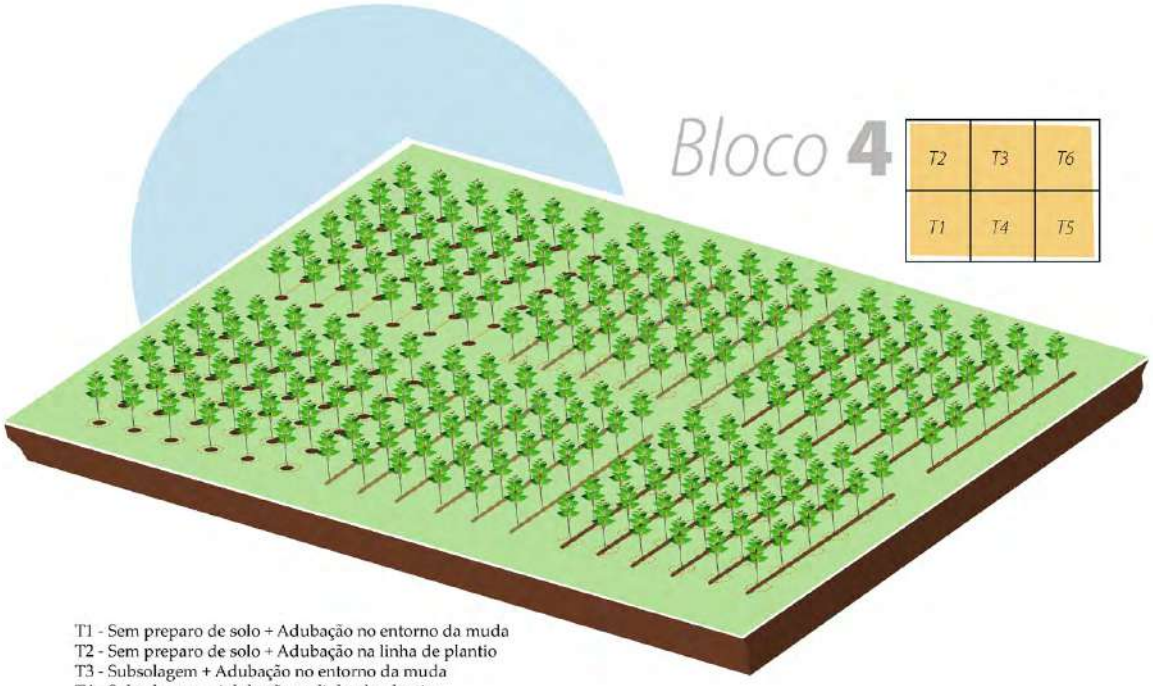


FIGURA 18.7. Experimento instalado em campo para avaliar o crescimento inicial da espécie florestal *Tectona grandis* sob o efeito de diferentes tipos de preparo de solo e métodos de adubação com delineamento de blocos ao acaso em esquema de parcelas subdivididas - bloco 3.



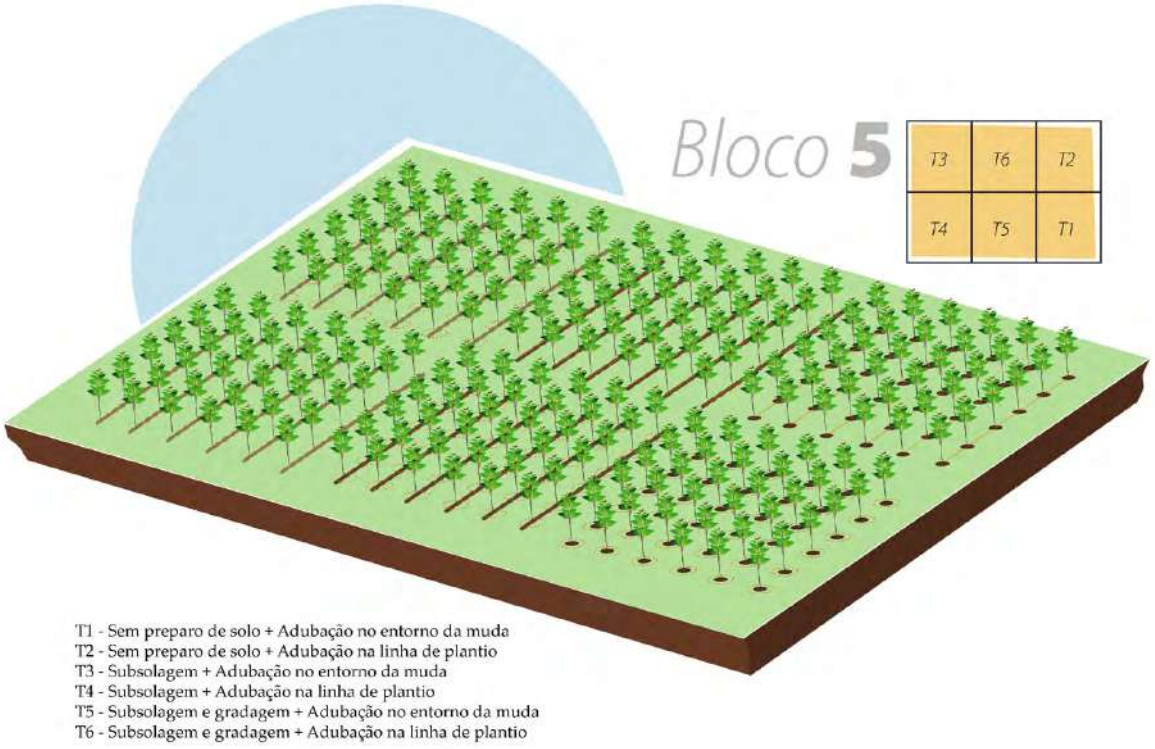
T1 - Sem preparo de solo + Adubação no entorno da muda
T2 - Sem preparo de solo + Adubação na linha de plantio
T3 - Subsolagem + Adubação no entorno da muda
T4 - Subsolagem + Adubação na linha de plantio
T5 - Subsolagem e gradagem + Adubação no entorno da muda
T6 - Subsolagem e gradagem + Adubação na linha de plantio

FIGURA 18.8. Experimento instalado em campo para avaliar o crescimento inicial da espécie florestal *Tectona grandis* sob o efeito de diferentes tipos de preparo de solo e métodos de adubação com delineamento de blocos ao acaso em esquema de parcelas subdivididas - bloco 4.



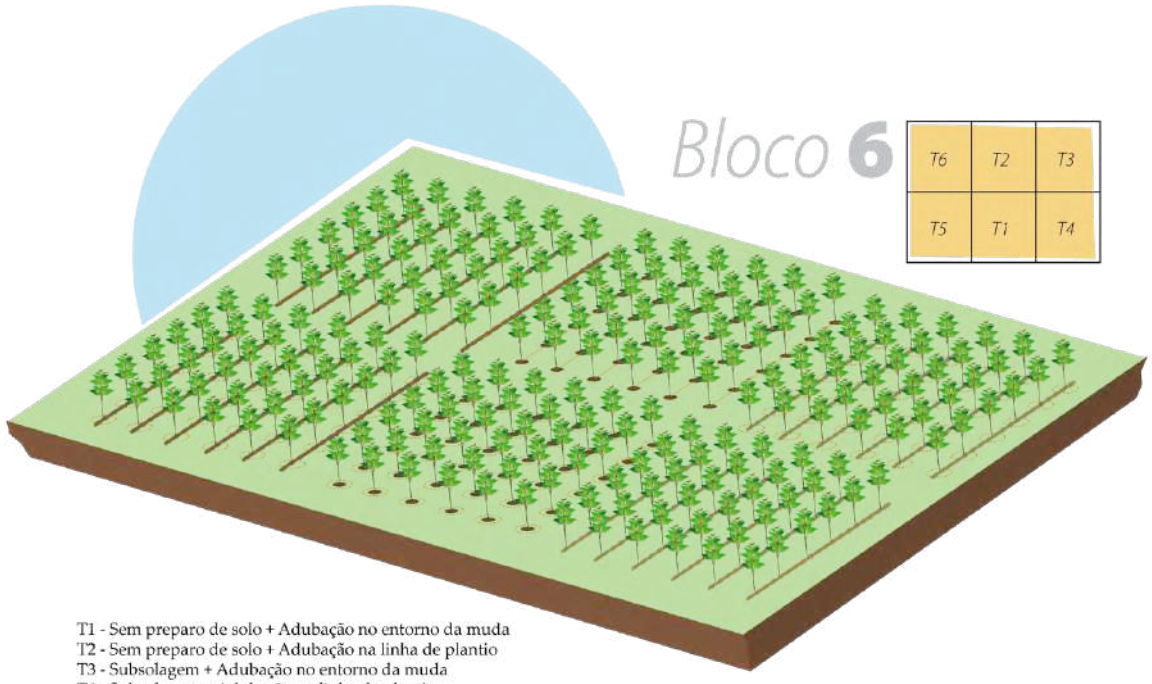
T1 - Sem preparo de solo + Adubação no entorno da muda
T2 - Sem preparo de solo + Adubação na linha de plantio
T3 - Subsolagem + Adubação no entorno da muda
T4 - Subsolagem + Adubação na linha de plantio
T5 - Subsolagem e gradagem + Adubação no entorno da muda
T6 - Subsolagem e gradagem + Adubação na linha de plantio

FIGURA 18.9. Experimento instalado em campo para avaliar o crescimento inicial da espécie florestal *Tectona grandis* sob o efeito de diferentes tipos de preparo de solo e métodos de adubação com delineamento de blocos ao acaso em esquema de parcelas subdivididas - bloco 5.



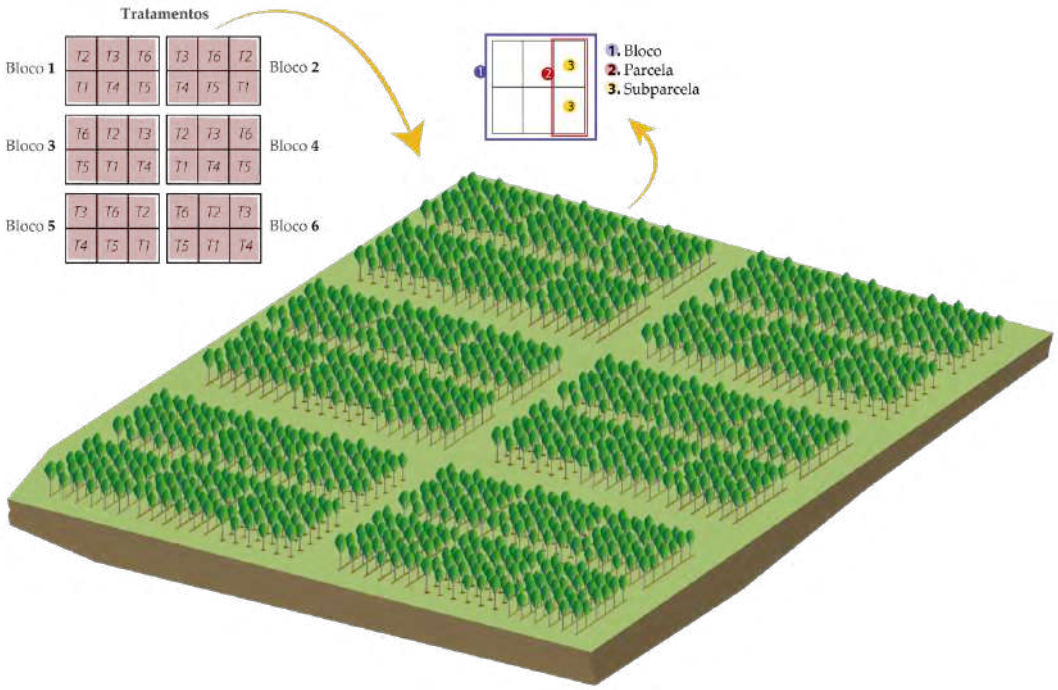
T1 - Sem preparo de solo + Adubação no entorno da muda
T2 - Sem preparo de solo + Adubação na linha de plantio
T3 - Subsolagem + Adubação no entorno da muda
T4 - Subsolagem + Adubação na linha de plantio
T5 - Subsolagem e gradagem + Adubação no entorno da muda
T6 - Subsolagem e gradagem + Adubação na linha de plantio

FIGURA 18.10. Experimento instalado em campo para avaliar o crescimento inicial da espécie florestal *Tectona grandis* sob o efeito de diferentes tipos de preparo de solo e métodos de adubação com delineamento de blocos ao acaso em esquema de parcelas subdivididas - bloco 6.



T1 - Sem preparo de solo + Adubação no entorno da muda
T2 - Sem preparo de solo + Adubação na linha de plantio
T3 - Subsolagem + Adubação no entorno da muda
T4 - Subsolagem + Adubação na linha de plantio
T5 - Subsolagem e gradagem + Adubação no entorno da muda
T6 - Subsolagem e gradagem + Adubação na linha de plantio

FIGURA 18.11. Experimento instalado em campo para avaliar o crescimento inicial da espécie florestal *Tectona grandis* sob o efeito de diferentes tipos de preparo de solo e métodos de adubação com delineamento de blocos ao acaso em esquema de parcelas subdivididas 2 anos após a instalação.



Os dados, na escala original, foram submetidos à análise de variância e os resultados obtidos estão apresentados na Tabela 18.3.

Comandos utilizados para obtenção da análise de variância de um experimento em blocos ao caso com parcelas subdivididas:

ARQUIVO: 36 - FUNÇÃO SUBDIVIDIDA PARA ANOVA EM PARCELA SUBDIVIDIDA COM DBA

```
>using DataFrames, CSV, Statistics, Distributions
>X=CSV.read("Z:/Alexandre/Julia/Exemplos/PS2.csv", DataFrame)
```

julia

```
julia> X=CSV.read("Z:/Alexandre/Julia/Exemplos/PS2.csv", DataFrame)
36×6 DataFrame
| Row | FatorA | FatorB | Bloco | h       | dc      | dap     |
|     | String | String | Int64 | Float64 | Float64 | Float64 |
|-----|--------|--------|-------|---------|---------|---------|
| 1   | a1     | b1     | 1     | 7.2     | 1.6     | 3.4     |
| 2   | a1     | b1     | 2     | 7.0     | 1.56    | 4.0     |
| 3   | a1     | b1     | 3     | 7.1     | 1.74    | 4.2     |
| 4   | a1     | b1     | 4     | 6.7     | 1.66    | 4.4     |
| 5   | a1     | b1     | 5     | 6.4     | 1.59    | 4.7     |
⋮
| 31  | a3     | b2     | 1     | 6.7     | 1.92    | 6.9     |
| 32  | a3     | b2     | 2     | 7.2     | 1.87    | 6.8     |
| 33  | a3     | b2     | 3     | 7.4     | 2.02    | 6.9     |
| 34  | a3     | b2     | 4     | 6.9     | 1.7     | 7.2     |
| 35  | a3     | b2     | 5     | 6.5     | 1.59    | 7.2     |
| 36  | a3     | b2     | 6     | 7.1     | 1.68    | 7.0     |
```

```
>function SUBDIVIDIDA(FatorA, FatorB, Bloco, Variavel)
    X=DataFrame()
    X.FatorA=FatorA
    X.FatorB=FatorB
    X.Bloco=Bloco
    X.Y=Variavel
    I= length(unique(X.FatorA))
    J= length(unique(X.FatorB))
    K= length(unique(X.Bloco))
    GLBlocos=K-1
    GLTrat=I*J-1
    GLErroA=(I-1)*(K-1)
    GLParcela=I*K-1
    GLTotal=I*J*K-1
    FC=((sum(X.Y))^2)/(I*J*K)
    AB= by(X, :Bloco) do df
    (m = mean(df.Y), s² = var(df.Y), total= sum(df.Y), ss=sum((df.Y).^2))
    end
    SQBlocos=((sum(AB.total.^2))/(I*J))-FC
    AT= by(X, [:FatorA,:FatorB]) do df
    (m = mean(df.Y), s² = var(df.Y), total= sum(df.Y), ss=sum((df.Y).^2))
    end
    SQTrat=((sum(AT.total.^2))/(K))-FC
    AP= by(X, [:Bloco,:FatorA]) do df
    (m = mean(df.Y), s² = var(df.Y), total= sum(df.Y), ss=sum((df.Y).^2))
    end
    SQParcela=((sum(AP.total.^2))/(J))-FC
    AFA= by(X, :FatorA) do df
    (m = mean(df.Y), s² = var(df.Y), total= sum(df.Y), ss=sum((df.Y).^2))
    end
    SQFatorA=((sum(AFA.total.^2))/(J*K))-FC
    SQErroA=SQParcela-SQFatorA-SQBlocos
    QMBlocos=SQBlocos/GLBlocos
    QMTrat=SQTrat/GLTrat
    QMErroA=SQErroA/GLErroA
    FBlocos=QMBlocos/QMErroA
    FTrat=QMTrat/QMErroA
    AlphaBlocos= ccdf(FDist(GLBlocos, GLErroA),abs(FBlocos))
    AlphaTrat= ccdf(FDist(GLTrat, GLErroA),abs(FTrat))
    GLFatorA=I-1
    GLFatorB=J-1
    GLFatorAFatorB=(I-1)*(J-1)
    GLErroB=I*(J-1)*(K-1)
    AFB= by(X, :FatorB) do df
    (m = mean(df.Y), s² = var(df.Y), total= sum(df.Y), ss=sum((df.Y).^2))
    end
    SQFatorB=((sum(AFB.total.^2))/(I*K))-FC
    SQTotal=(sum(X.Y.^2))-FC
    SQFatorAFatorB=SQTrat-SQFatorA-SQFatorB
    SQErroB=SQTotal-SQParcela-SQFatorB-SQFatorAFatorB
```

```
    QMFatorA=SQFatorA/GLFatorA
    QMFatorB=SQFatorB/GLFatorB
    QMFatorAFatorB=SQFatorAFatorB/GLFatorAFatorB
    QMErroB=SQErroB/GLErroB
    FFatorA=QMFatorA/QMErroA
    FFatorB=QMFatorB/QMErroB
    FFatorAFatorB=QMFatorAFatorB/QMErroB
    AlphaFatorAFatorB=ccdf(FDist(GLFatorAFatorB, GLErroB),abs(FFatorAFatorB))
    AlphaFatorA= ccdf(FDist(GLFatorA, GLErroA),abs(FFatorA))
    AlphaFatorB= ccdf(FDist(GLFatorB, GLErroB),abs(FFatorB))
    println("GLBlocos: ", GLBlocos)
    println("GLTrat: ", GLTrat)
    println("GLParcela: ", GLParcela)
    println("GLErroA: ", GLErroA)
    println("GLTotal: ", GLTotal)
    println("SQBlocos: ", SQBlocos)
    println("SQTrat: ", SQTrat)
    println("SQParcela: ", SQParcela)
    println("SQErroA: ", SQErroA)
    println("SQTotal: ", SQTotal)
    println("QMBlocos: ", QMBlocos)
    println("QMTrat: ", QMTrat)
    println("QMErroA: ", QMErroA)
    println("FBlocos: ", FBlocos)
    println("FTrat: ", FTrat)
    println("AlphaBlocos: ", AlphaBlocos)
    println("AlphaTrat: ", AlphaTrat)
    println("GLFatorA: ", GLFatorA)
    println("GLFatorB: ", GLFatorB)
    println("GLFatorAFatorB: ", GLFatorAFatorB)
    println("GLErroB: ", GLErroB)
    println("SQFatorA: ", SQFatorA)
    println("SQFatorB: ", SQFatorB)
    println("SQFatorAFatorB: ", SQFatorAFatorB)
    println("SQErroB: ", SQErroB)
    println("QMFatorA: ", QMFatorA)
    println("QMFatorB: ", QMFatorB)
    println("QMFatorAFatorB: ", QMFatorAFatorB)
    println("QMErroB ", QMErroB)
    println("FFatorAFatorB: ", FFatorAFatorB)
    println("FFatorA: ", FFatorA)
    println("FFatorB: ", FFatorB)
    println("AlphaFatorAFatorB: ", AlphaFatorAFatorB)
    println("AlphaFatorA: ", AlphaFatorA)
    println("AlphaFatorB: ", AlphaFatorB)
end
```

```
#SUBDIVIDIDA(FatorA, FatorB, Bloco, Variavel)

>SUBDIVIDIDA(X.FatorA, X.FatorB, X.Bloco, X.dap)
GLBlocos: 5
GLTrat: 5
GLParcela: 17
GLErroA: 10
GLTotal: 35
SQBlocos: 2.44
SQTrat: 47.05
SQParcela: 49.66
SQErroA: 0.319
SQTotal: 50.38
QMBlocos: 0.487
QMTrat: 9.41
QMErroA: 0.032
FBlocos: 15.24
FTrat: 294.56
AlphaBlocos: 0.0002
AlphaTrat: 1.623e-10
GLFatorA: 2
GLFatorB: 1
GLFatorAFatorB: 2
GLErroB: 15
SQFatorA: 46.90
SQFatorB: 0.0069
SQFatorAFatorB: 0.137
SQErroB: 0.581
QMFatorA: 23.45
QMFatorB: 0.00694
QMFatorAFatorB: 0.0686
QMErroB 0.0387
FFatorAFatorB: 1.77
FFatorA: 734.15
FFatorB: 0.179
AlphaFatorAFatorB: 0.204
AlphaFatorA: 1.42e-11
AlphaFatorB: 0.68
```

TABELA 18.3. Resultados da análise de variância para avaliar o crescimento em diâmetro à altura do peito (em cm) da espécie florestal *Tectona grandis* sob o efeito de diferentes tipos de preparo de solo e métodos de adubação.

Fonte de Variação	Grau de Liberdade	Soma de Quadrado	Quadrado Médio	F calculado	F tabelado ($\alpha = 0,05$)	P- valor (%)
Blocos	5	2,43	0,49	15,24 **	3,32	0,02
Fator A	2	46,90	23,45	734,15 **	4,10	<0,001
Erro Exp a	10	0,32	0,03			
Parcelas	17	49,66				
Fator B	1	0,01	0,01	0,179 ns	4,54	67,80
A x B	2	0,14	0,07	1,772 ns	3,68	20,38
Erro Exp b	15	0,58	0,04			
Total	35	50,38				

ns = não significante, ** = significativo ao nível de $\alpha = 0,01$ e A x B= Interação entre fatores.

Para a variável DAP, o teste F para avaliar os efeitos dos blocos foi significante, indicando que os blocos foram efetivos para o uso do controle local. O teste F da análise de variância indicou que os fatores são independentes, pois a interação foi não significante. Diferença significante foi observada apenas para os tipos de preparo do solo. Assim, foi aplicado o teste de médias de Tukey para sua diferenciação entre os níveis para este fator (Tabela 18.4). A função TukeyFat foi utilizada para obtenção dos resultados. Quando se deseja comparar as médias do fator aplicado às parcelas utiliza-se o erro experimental das parcelas (erro a) na fórmula do teste. Para as demais comparações, ou seja, para o fator aplicado à subparcela e interação de efeitos utiliza-se o erro experimental da subparcela (erro b).

Comandos para obtenção do teste de Tukey:

>ARQUIVO: 36 - FUNÇÃO SUBDIVIDIDA PARA ANOVA EM PARCELA SUBDIVIDIDA COM DBA

```
>using DataFrames, Statistics, PyPlot, Distributions, StatsFuns
```

julia

```
julia> MédiasFatorA=by(X, :FatorA, :dap => mean)
3×2 DataFrame
| Row | FatorA | dap_mean |
|     | String | Float64  |
|-----|--------|----------|
| 1   | a1     | 4.26667  |
| 2   | a2     | 5.25     |
| 3   | a3     | 7.025    |

#Teste para as médias entre os níveis do FatorA
>MédiasFatorA=by(X, :FatorA, :dap => mean)
```

```
julia> TukeyFat(6,MédiasFatorA.FatorA,MédiasFatorA.dap_mean,10,0.0319,0.05)
DMS: 0.2826770227137929
Tabela DMS: 3×4 DataFrame
| Row | Tratamento | a1        | a2        | a3        |
|     | String     | Float64   | Float64   | Float64   |
|-----|------------|-----------|-----------|-----------|
| 1   | a1         | 0.0       | -0.983333 | -2.75833  |
| 2   | a2         | 0.983333  | 0.0       | -1.775    |
| 3   | a3         | 2.75833   | 1.775     | 0.0       |
Tabela de hipóteses DMS: 3×4 DataFrame
| Row | Tratamento | a1     | a2     | a3     |
|     | String     | String | String | String |
|-----|------------|--------|--------|--------|
| 1   | a1         | ns     | *      | *      |
| 2   | a2         | *      | ns     | *      |
| 3   | a3         | *      | *      | ns     |
```

```
>DMS=0.283
errorbar(MédiasFatorA.FatorA, MédiasFatorA.dap_mean,DMS,fmt="o")
>xlabel("Tratamento")
>ylabel("dap (cm)")
```

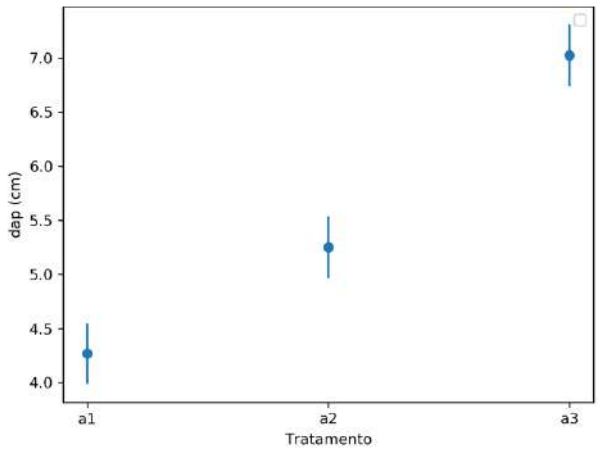


TABELA 18.4. Resultados do teste de Tukey para comparação das médias dos diferentes tipos de preparo do solo (fator A) para o crescimento em diâmetro à altura do peito (em cm) da espécie florestal *Tectona grandis*.

Tipo de preparo de solo	Média	Grupo
Sem preparo	4,27	c
Subsolagem	5,25	b
Subsolagem e gradagem	7,02	a

Médias seguidas pela mesma letra não diferem pelo teste de Tukey ao nível de $\alpha = 0{,}05$ de significância.

O teste de Tukey revelou, para a variável diâmetro à altura do peito, que o tratamento subsolagem e gradagem gerou os melhores resultados, dado que são superiores e diferentes dos obtidos nos demais tratamentos testados. As análises das demais variáveis não foram apresentadas, pois sugere-se que sejam efetuadas pelos leitores.

Experimento 2

Foi conduzido um experimento para testar o efeito do tipo de desbaste e da sua intensidade sobre a produção em área basal de *Pinus caribaea*. O desbaste foi conduzido sete anos depois e a avaliação de produção 14 anos após o plantio das mudas.

Os seguintes tipos e intensidades de desbastes foram testados:

1. **Sistemático:**
   20%: 5ª linha
   25%: 4ª linha
   33%: 3ª linha

2. **Seletivo:**
   20%: remoção de 20% das árvores menores remanescentes após o desbaste sistemático
   25%: remoção de 25% das árvores menores remanescentes após o desbaste sistemático
   30%: remoção de 30% das árvores menores remanescentes após o desbaste sistemático

Para facilitar a instalação do experimento no campo o desbaste sistemático (fator A) foi alocado às parcelas e o desbaste seletivo (fator B) nas subparcelas. Assim, os tratamentos foram os seguintes:

- **T1:** Desbaste sistemático 20% + seletivo 20%
- **T2:** Desbaste sistemático 20% + seletivo 25%
- **T3:** Desbaste sistemático 20% + seletivo 33%
- **T4:** Desbaste sistemático 25% + seletivo 20%
- **T5:** Desbaste sistemático 25% + seletivo 25%
- **T6:** Desbaste sistemático 25% + seletivo 33%
- **T7:** Desbaste sistemático 33% + seletivo 20%
- **T8:** Desbaste sistemático 33% + seletivo 25%
- **T9:** Desbaste sistemático 33% + seletivo 33%

A unidade experimental para o desbaste sistemático 20% e 33% foi composta com 234 plantas (18 linhas x 13 plantas) e a subparcela composta com 48 plantas (16 linhas x 3 plantas). Para o tratamento de desbaste sistemático 25%, a unidade experimental foi composta com 247 plantas (19 linhas x 13 plantas) e a subparcela composta com 51 plantas (17 linhas x 3 plantas), a fim de viabilizar o estabelecimento da linha de bordadura no entorno da parcela e, também, uma linha entre as subparcelas. As plantas foram cultivadas em espaçamento 3m x 2m e o experimento foi conduzido com 6 repetições.

Observe que a condução do experimento ocorre para um período de no mínimo 14 anos. Uma alternativa para reduzir este tempo e obter respostas mais rápidas pode ser alcançada por meio da instalação do experimento em um povoamento já estabelecido com 7 anos, reduzindo o tempo de condução do experimento pela metade.

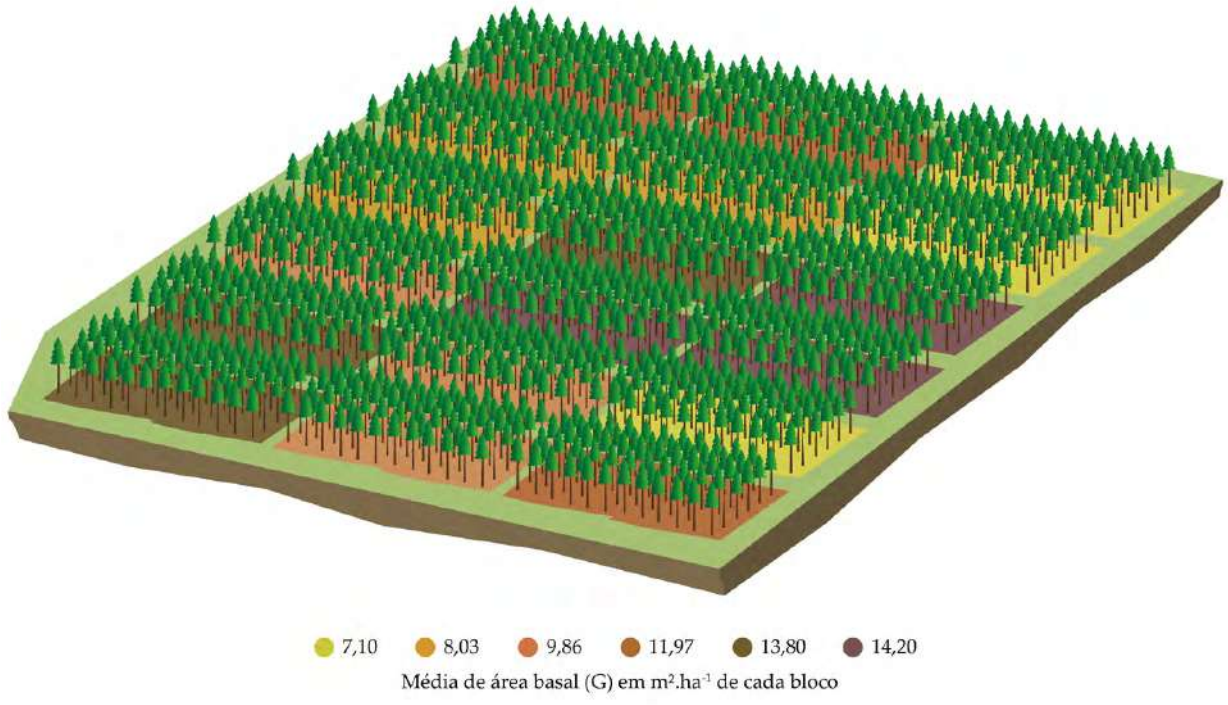
Dessa maneira, foi selecionada uma área em um povoamento para a instalação de 18 parcelas. Nestas, previamente demarcadas em campo, foi conduzido um inventário florestal quando foi mensurado o DAP, visando obter a área basal da parcela ($m^2.ha^{-1}$). Os resultados do inventário florestal estão apresentados na Tabela 18.5.

TABELA 18.5. Resultados obtidos para a área basal ($m^2\ ha^{-1}$) da espécie *Pinus caribaea* aos sete anos de idade, obtidos nas parcelas, nas quais foram efetuados os tratamentos para testar o efeito e a intensidade do desbaste sobre a produção em área basal.

Área basal ($m^2\ ha^{-1}$)		
12,0	11,6	7,0
8,0	8,1	7,5
8,0	13,5	14,3
10,0	14,1	14,2
14,0	9,5	6,9
13,9	10,1	12,3

A partir destes resultados iniciais, as unidades experimentais foram agrupadas em blocos em função dos valores observados para a área basal, definindo-se, dessa maneira, a estrutura experimental em blocos ao acaso (Figura 18.12). Os tipos de desbaste foram aplicados às parcelas e as intensidades às subparcelas.

FIGURA 18.12. Parcelas selecionadas para avaliar o efeito do tipo e intensidade de desbaste sobre a produção da espécie *Pinus caribaea*.



Após o estabelecimento das parcelas, cada uma foi dividida em três subparcelas para receber os tratamentos de intensidade dos desbastes, conforme as Figuras 18.13 a 18.16.

FIGURA 18.13. Parcela instalada em campo para avaliar o efeito do tipo e intensidade de desbaste sobre a produção em área basal da espécie *Pinus caribaea* – fator a: desbaste sistemático 20%.

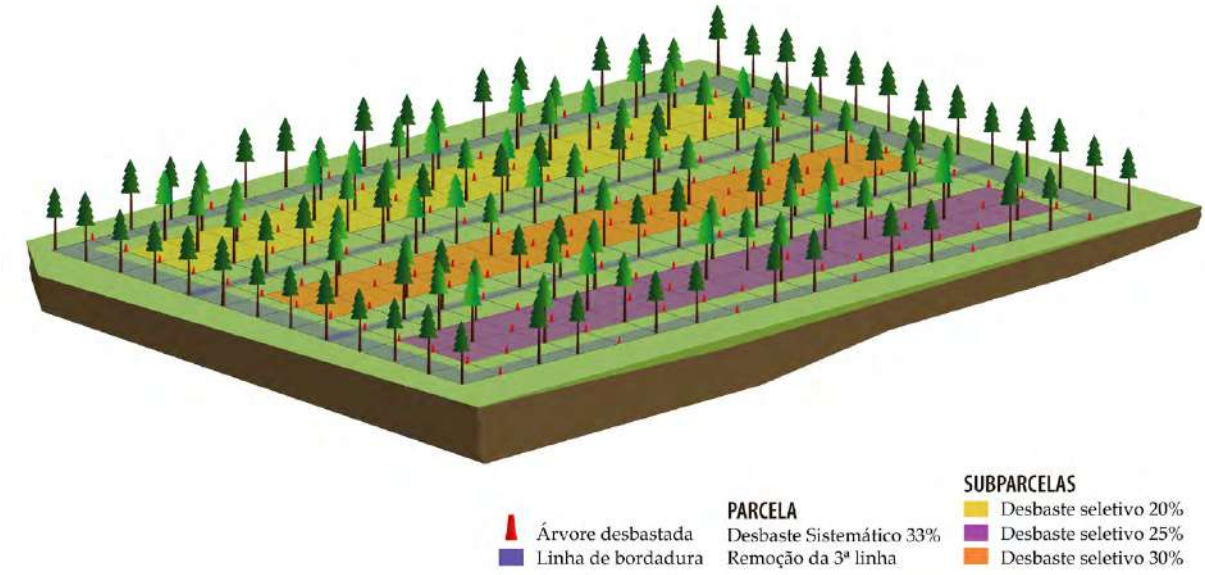


FIGURA 18.14. Parcela instalada em campo para avaliar o efeito do tipo e intensidade de desbaste sobre a produção em área basal da espécie *Pinus caribaea* – fator a: desbaste sistemático 25%.

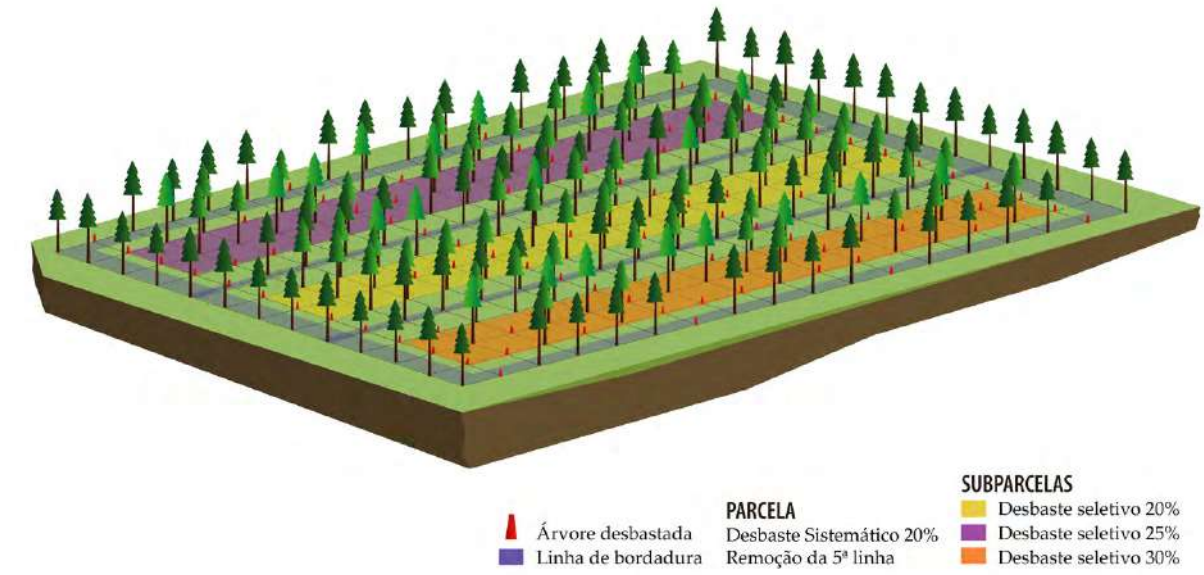


FIGURA 18.15. Parcela instalada em campo para avaliar o efeito do tipo e intensidade de desbaste sobre a produção em área basal da espécie *Pinus caribaea* – fator a: desbaste sistemático 33%.

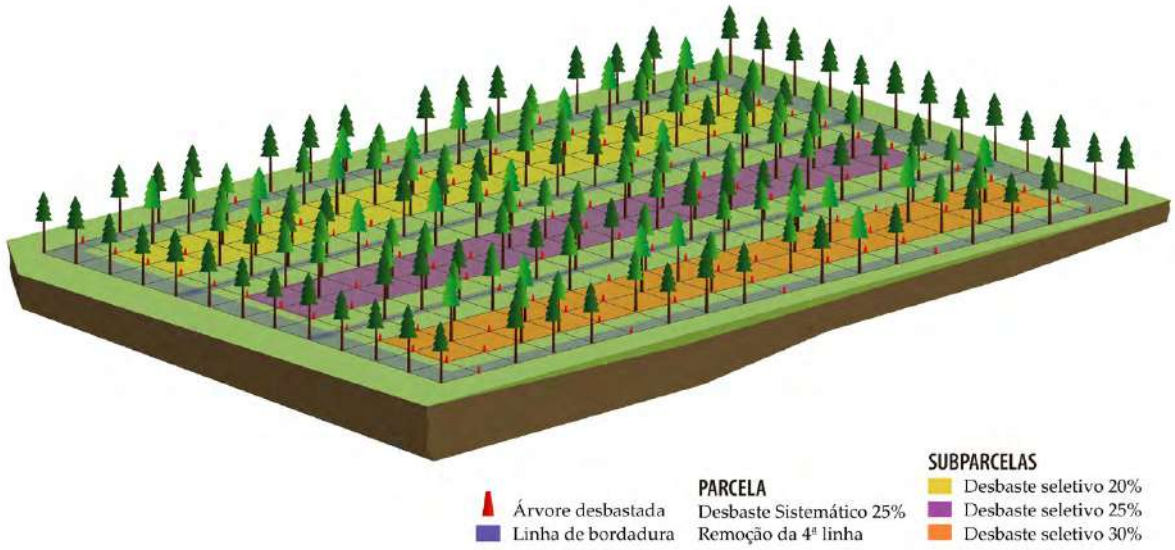
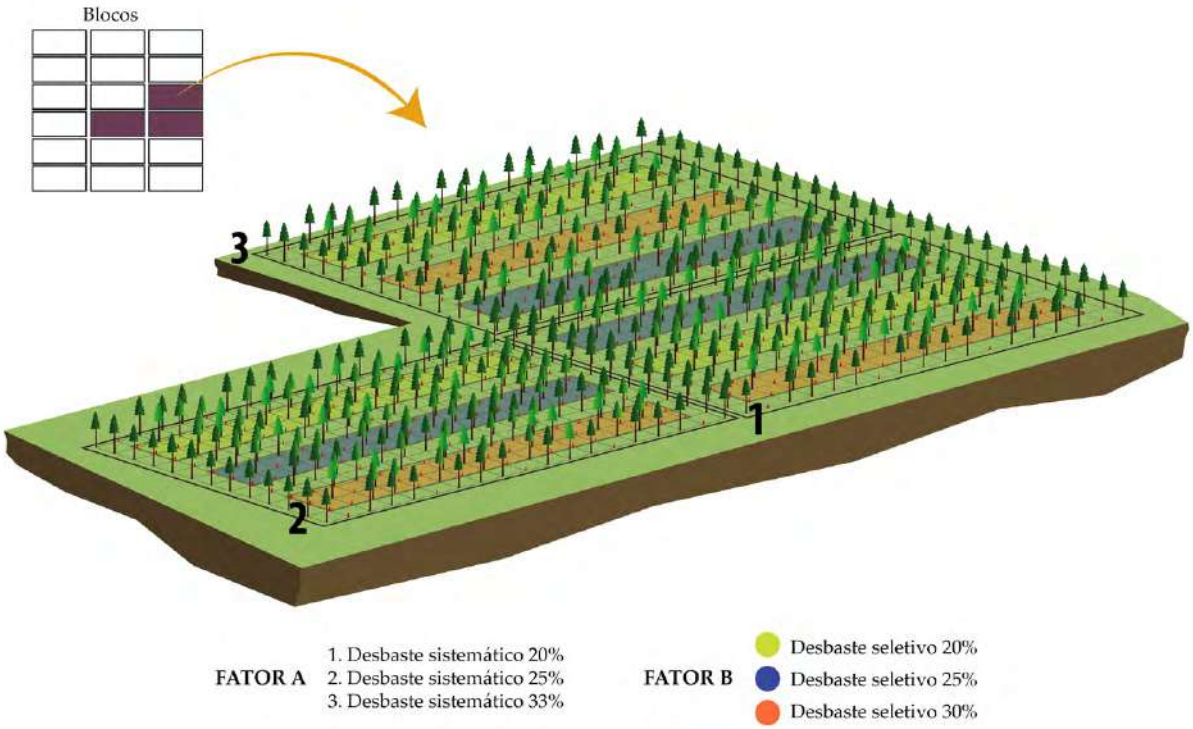


FIGURA 18.16. Bloco instalado em campo para avaliar o efeito do tipo e intensidade de desbaste sobre a produção em área basal da espécie *Pinus caribaea* – bloco de área basal média de 14,20 $m^2.ha^{-1}$ com os 9 tratamentos testados.



Sete anos após a instalação do experimento foi realizada a avaliação da área basal nas unidades experimentais, estando os dados apresentados na Tabela 18.6.

TABELA 18.6. Resultados obtidos para a área basal (G, em $m^2.ha^{-1}$) da espécie *Pinus caribaea* aos 14 anos de idade para testar o efeito do tipo e intensidade do desbaste sete anos após a instalação do experimento (continua).

Fator A	Fator B	Bloco	G ($m^2.ha^{-1}$)
Sistemático 20%	Seletivo 20%	1	40
Sistemático 20%	Seletivo 20%	2	38
Sistemático 20%	Seletivo 20%	3	37
Sistemático 20%	Seletivo 20%	4	36
Sistemático 20%	Seletivo 20%	5	39
Sistemático 20%	Seletivo 20%	6	37
Sistemático 20%	Seletivo 25%	1	34
Sistemático 20%	Seletivo 25%	2	36
Sistemático 20%	Seletivo 25%	3	33
Sistemático 20%	Seletivo 25%	4	35
Sistemático 20%	Seletivo 25%	5	32
Sistemático 20%	Seletivo 25%	6	36
Sistemático 20%	Seletivo 30%	1	30
Sistemático 20%	Seletivo 30%	2	32
Sistemático 20%	Seletivo 30%	3	33
Sistemático 20%	Seletivo 30%	4	30
Sistemático 20%	Seletivo 30%	5	32
Sistemático 20%	Seletivo 30%	6	30
Sistemático 25%	Seletivo 20%	1	28
Sistemático 25%	Seletivo 20%	2	28
Sistemático 25%	Seletivo 20%	3	29
Sistemático 25%	Seletivo 20%	4	29
Sistemático 25%	Seletivo 20%	5	30
Sistemático 25%	Seletivo 20%	6	28
Sistemático 25%	Seletivo 25%	1	28
Sistemático 25%	Seletivo 25%	2	26
Sistemático 25%	Seletivo 25%	3	28
Sistemático 25%	Seletivo 25%	4	25
Sistemático 25%	Seletivo 25%	5	26
Sistemático 25%	Seletivo 25%	6	27
Sistemático 25%	Seletivo 30%	1	28
Sistemático 25%	Seletivo 30%	2	29
Sistemático 25%	Seletivo 30%	3	25
Sistemático 25%	Seletivo 30%	4	24
Sistemático 25%	Seletivo 30%	5	25
Sistemático 25%	Seletivo 30%	6	26
Sistemático 33%	Seletivo 20%	1	26
Sistemático 33%	Seletivo 20%	2	27
Sistemático 33%	Seletivo 20%	3	30

TABELA 18.6. Resultados obtidos para a área basal (G, em $m^2.ha^{-1}$) da espécie *Pinus caribaea* aos 14 anos de idade para testar o efeito do tipo e intensidade do desbaste sete anos após a instalação do experimento (continuação).

Fator A	Fator B	Bloco	G ($m^2.ha^{-1}$)
Sistemático 33%	Seletivo 20%	4	28
Sistemático 33%	Seletivo 20%	5	25
Sistemático 33%	Seletivo 20%	6	24
Sistemático 33%	Seletivo 25%	1	27
Sistemático 33%	Seletivo 25%	2	26
Sistemático 33%	Seletivo 25%	3	25
Sistemático 33%	Seletivo 25%	4	24
Sistemático 33%	Seletivo 25%	5	28
Sistemático 33%	Seletivo 25%	6	26
Sistemático 33%	Seletivo 30%	1	22
Sistemático 33%	Seletivo 30%	2	23
Sistemático 33%	Seletivo 30%	3	25
Sistemático 33%	Seletivo 30%	4	24
Sistemático 33%	Seletivo 30%	5	21
Sistemático 33%	Seletivo 30%	6	20

A análise dos dados deste experimento não será apresentada, pois este é um exercício proposto para resolução.

Um resumo geral sobre o delineamento em parcelas subdivididas está apresentado na Figura 18.17.

FIGURA 18.17. Resumo do Delineamento em parcelas subdivididas.

DELINEAMENTO EM PARCELAS SUBDIVIDIDAS

julia

function
SUBDIVIDIDA(FatorA, FatorB, Bloco, Variavel)

$$Y_{ijk}=\mu+\tau_i+\omega_k+\delta_{ik}+\beta_j+(\tau\beta)_{ij}+\epsilon_{ijk}$$

quando conduzido em DBA

A característica principal deste tipo de arranjo é a utilização de dois tamanhos de parcelas, um para cada fator de estudo. São mais práticos de instalar que o ensaio fatorial.

Casualização em duas etapas:

- Casualiza-se os níveis de um fator na parcela
- Casualiza-se os níveis do outro fator na subparcela

A utilização de dois tamanhos de parcela resulta em dois níveis de precisão. Normalmente, o erro experimental associado às parcelas é maior que o erro experimental associado às subparcelas.

VANTAGENS

» Economia e uso racional na aplicação dos recursos
» Análise de mais de uma fator em um mesmo ensaio, análise da dependência dos fatores
» Facilidade de instalação do experimento em campo

DESVANTAGENS

» Dois erros experimentais, um para a parcela e outro para a subparcela (entretanto isso pode ser uma vantagem)
» Todos os efeitos são avaliados com menor precisão que nos experimentos fatoriais.

19 TESTES DE COMPARAÇÕES MÚLTIPLAS DE MÉDIAS

Em todos os experimentos é formulada uma hipótese estatística inicial, a qual será testada por meio do teste de F na análise de variância. Por meio dela é testado se as médias dos tratamentos não diferem entre si ou se pelo menos uma média é diferente das demais. Se o valor de F é não significante, então as médias dos tratamentos não diferem estatisticamente entre si. Quando o valor de F é significante, pelo menos uma média difere estatisticamente das demais, exigindo, desta forma, a verificação de quais médias são iguais e quais são diferentes. Isto é obtido por meio de um teste de comparação de médias.

Neste sentido, destacam-se duas prioridades especificadas pelos pesquisadores para a análise das médias:

1. Em algumas pesquisas busca-se agrupá-las, visando a identificar quais tratamentos são diferentes, confirmados pelo teste de F. Aplicadas para tratamentos qualitativos (com exceções), busca-se agrupá-las baseada em uma diferença mínima de significância entre as médias.
2. Quando os tratamentos testados são quantitativos é possível complementar a análise por meio do ajuste de uma equação de regressão para determinar a tendência dos resultados em função dos tratamentos aplicados. Uma vez identificada a equação que define a tendência observada, é possível determinar os pontos de máximos e de mínimos. A formação de grupos de tratamentos não é prioridade, mas sim a análise do comportamento do gradiente da função ajustada para os níveis dos tratamentos.

Aqui será abordado somente o primeiro item, cujo objetivo final é agrupar os tratamentos, já ciente de que as médias dentro dos grupos não difiram estatisticamente entre si, mas que essas difiram entre os grupos. Para tanto, um teste de comparação de médias deve ser aplicado.

19.1 TESTES DE COMPARAÇÃO DE MÉDIAS

No experimento, quando 3 ou mais tratamentos são avaliados, o teste F é aplicado para testar se as médias dos tratamentos não diferem entre si. Quando o teste F é significante, sabe-se que pelo menos uma média é diferente das demais. A pergunta que pode ser feita é: quais são os tratamentos cujas médias são estatisticamente diferentes das demais? Esta é uma questão que enseja a aplicação dos testes de comparação de médias.

Os testes de comparação de médias implicam na comparação de todas as médias, duas a duas, contra um valor definido pelo tipo de teste a ser aplicado. Desta forma, quatro passos serão efetuados:

- cálculo do comparador
- ordenamento das médias
- comparação das médias duas a duas contra o comparador
- sumarização dos resultados

19.2 CONTRASTES PLANEJADOS: A PRIORI

Os contrastes planejados são comparações entre médias de tratamentos específicos, ou grupos de tratamentos. Isto é utilizado quando se deseja testar hipóteses específicas sobre os grupos de tratamentos, devido a um interesse prático, ou porque não seria lógico comparar certos tratamentos. As comparações são efetuadas de duas a duas. Por exemplo, na análise de variância em que são avaliados 5 tratamentos, as comparações de interesse podem ser $\bar{x}_1$ com $\bar{x}_3$ e $\bar{x}_2$, $\bar{x}_3$ e $\bar{x}_2$ com $\bar{x}_4$ e $\bar{x}_1$ com $\bar{x}_5$. Isto é denominado de comparação simples. Observando-se as condições de todos os tratamentos pode não ser interessante compará-los em todas as suas possibilidades combinatórias.

Outras comparações podem se referir a grupos de tratamentos, ou seja, quando se deseja comparar as médias dos grupos. Por exemplo, comparar a média do grupo $\bar{x}_1$ e $\bar{x}_3$ com a média do grupo $\bar{x}_2$, $\bar{x}_3$ e $\bar{x}_5$. Em outras situações, a média dos tratamentos $\bar{x}_2$, $\bar{x}_3$, $\bar{x}_4$ e $\bar{x}_5$ contra a média de $\bar{x}_1$, que pode ser uma testemunha. Observe que uma comparação planejada pressupõe combinar os tratamentos a serem comparados antecipadamente e não necessariamente aguardar os seus resultados para, por exemplo, definir qual tratamento está associado às melhores respostas.

19.3 CONTRASTES A POSTERIORI (POST HOC)

A análise *post hoc* (em latim *depois disso*) é aplicada para efetuar comparações entre as médias dos tratamentos quando não se tem expectativas prévias sobre como os tratamentos devem ser comparados, ou seja, as comparações não são planejadas, quando o interesse for comparar todas as combinações de médias. A efetuação de comparações não planejadas é denominada de efeitos de testes revelados pelas respostas obtidas nos dados.

Existem vários testes de comparação de médias e os mais comuns são: t ou DMS (diferença mínima significativa), Tukey, Duncan, Student-Newman-Keuls (SNK), Dunnet, Scheffé e Scott-Knott. Estes podem ser classificados conforme o critério da abrangência das comparações, de forma a ser efetuado:

- Comparações entre todos os pares de tratamentos: testes DMS, Tukey, Duncan e SNK;
- Comparações entre grupos de tratamentos: testes DMS, Scheffé e Scott-Knott;
- Comparação de cada tratamento contra uma testemunha: teste de Dunnett.

Entre os vários testes *post hoc*, a maioria deles utiliza basicamente a estatística *t*, como segue:

$$t_{(n-1,\alpha)} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{2\,MSE}{n}}}$$

A maior diferença entre os testes de comparação de médias está focada no uso dos graus de liberdade e, também, nos níveis α.

Os testes de comparação de médias são aplicados somente quando o valor de F na análise de variância é significante, caso contrário não é apropriado.

Previamente à execução dos cálculos da análise de variância, foi evidenciada a importância da avaliação da homogeneidade das variâncias dos tratamentos. Esta premissa, de se ter variâncias homogêneas também é exigida pela maioria dos testes *post hoc*. Quando as variâncias não são homogêneas, o teste *post hoc* atestará diferença significante, quando ela não existe.

19.4 TESTE DMS

O teste DMS (diferença mínima significante) é equivalente *Post Hoc* de um contraste linear, utilizado para comparar médias duas a duas. O nível α para todas as comparações, por exemplo, corresponde a α = 0,05 de significância, com grau de liberdade do erro experimental.

É recomendado para realizar comparações previamente planejadas e não para efetuar inferências evidenciadas nos dados, embora possa, também, ser aplicado com restrições. Neste caso, o teste é utilizado para comparar pares de médias de no máximo cinco tratamentos.

O teste DMS é dado por:

$$DMS = t_{(\alpha,GLE)}\sqrt{\frac{2QME}{r}}$$

Em que:

$t_{(\alpha,GLE)}$ é o valor da distribuição de t para o nível de significância α e número de graus de liberdade do erro experimental, obtido da análise de variância.

QME é o quadrado médio do erro experimental, obtido da análise de variância.

r é o número de repetições.

EXERCÍCIO 19 : TRATAMENTO DE MADEIRA

Em determinadas situações, devido ao uso da madeira em ambiente natural, esta está sujeita à ação de agentes biológicos e, também, do próprio ambiente. Uma alternativa para aumentar a durabilidade da madeira é a aplicação de preservativos industriais ou alternativos, para diminuir sua deterioração em comparação com a madeira que não é tratada. Neste sentido, experimentos são realizados com objetivo de revelar o que acontece com a madeira que foi tratada comparada com a que não foi tratada. Esta condição motiva o estabelecimento de um teste de comparação de médias dos tratamentos definidos à priori, nos quais estas, para a madeira tratada com preservativos, são comparadas com a média da madeira que não é tratada (testemunha).

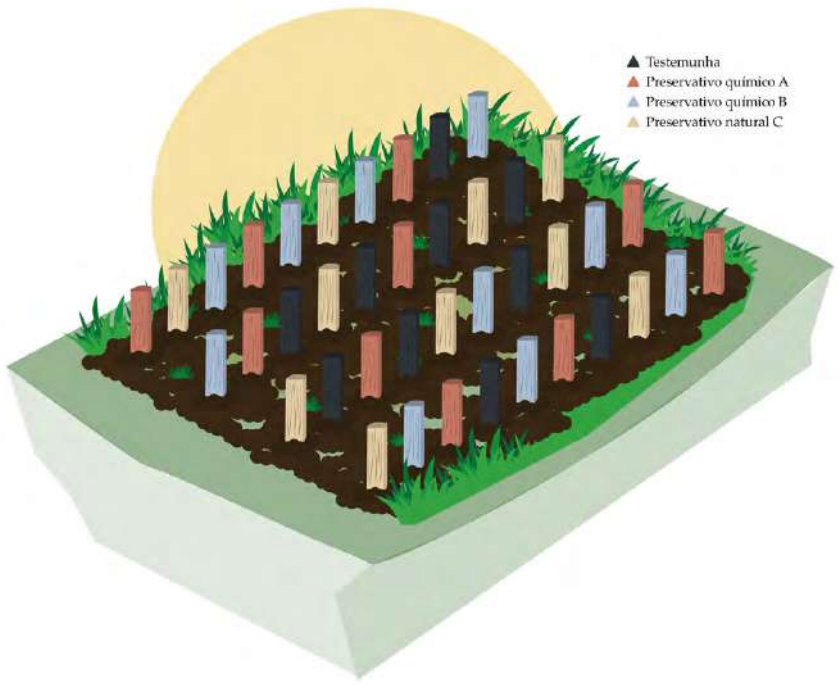
Um Engenheiro Florestal conduziu um experimento, no qual foram testados 4 tratamentos:

1. Testemunha (sem aplicação de preservativos)
2. Preservativo químico A
3. Preservativo químico B
4. Preservativo natural C

Os tratamentos foram aplicados em corpos de prova da madeira da espécie canafístula (Figura 19.1), proveniente de uma árvore com 20 anos, repetidos 15 vezes. Os corpos de prova tinham 50 cm de comprimento e 5 cm de largura.

Após a aplicação dos tratamentos nos corpos de prova, as amostras foram levadas para campo e enterradas parcialmente em solo (25 cm). Como estes eram semelhantes entre si, assim como o solo era homogêneo no local da instalação, o delineamento inteiramente casualizado foi adotado. Antes e após 12 meses da instalação do experimento foi efetuada a pesagem dos corpos de prova, logo após as amostras serem secas em estufa. Com estes dados, foi calculada a perda de massa de madeira (em gramas). Estes dados estão armazenados no arquivo *Tratamentodamadeira*.

FIGURA 19.1. Aplicação dos tratamentos nos corpos de prova.



Comandos para a leitura dos dados:

ARQUIVO: 37 - EXERCÍCIO TRATAMENTO DE MADEIRA

```
>using DataFrames, CSV
>X=CSV.read("Z:/Alexandre/Julia/Exemplos/TratMad.csv", DataFrame)
```

```
julia> X=CSV.read("Z:/Alexandre/Julia/Exemplos/TratMad.csv", DataFrame)
60×3 DataFrame
| Row | Tratamento | Rep   | PM        |
|     | String     | Int64 | Float64   |
|-----|------------|-------|-----------|
| 1   | Testemunha | 1     | 10.3255   |
| 2   | Testemunha | 2     | 11.2635   |
| 3   | Testemunha | 3     | 13.0984   |
| 4   | Testemunha | 4     | 11.0387   |
| 5   | Testemunha | 5     | 11.0654   |
⋮
| 55  | PQC        | 10    | 2.61043   |
| 56  | PQC        | 11    | 1.89132   |
| 57  | PQC        | 12    | 0.924287  |
| 58  | PQC        | 13    | 0.0163914 |
| 59  | PQC        | 14    | 3.75367   |
| 60  | PQC        | 15    | 2.61767   |
```

Utilizando-se a função dic, foi calculada a estatística F por meio da análise de variância. O valor da estatística F foi de 182,83 e o α = 6,79 e^{-29}, revelando evidências que pelo menos uma média dos tratamentos difere das demais.

Comandos para obtenção da análise de variância de um DIC:

ARQUIVO: 37 - EXERCÍCIO TRATAMENTO DE MADEIRA

```
>using DataFrames, Statistics, Distributions

#DIC(Tratamento, Repeticao,Variavel)
>DIC(X.Tratamento,X.Rep,X.PM)
GLTrat: 3
GLResíduo: 56
GLTotal: 59
SQTrat: 848.72
SQResíduo: 86.65
SQTotal: 935.37
QMTrat: 282.91
QMResíduo: 1.55
FTrat: 182.83
Alpha: 6.79e-29
R²(%): 90.74
CV(%): 23.34
```

A comparação de médias a ser efetuada foi definida à priori à condução do experimento. O teste DMS foi aplicado para comparar as médias dos tratamentos com uso do preservativo em relação à testemunha, utilizando-se a função *DMS*.

Comandos utilizados para a aplicação do teste DMS:

Arquivo: 38 - Função dms para teste de médias de dms

```
>using Statistics, Distributions, StatsFuns

>function DMS(Tratamento,Repeticao,Variavel,GLResiduo,QMResiduo,Alpha)
    X=DataFrame()
    X.Trat=Tratamento
    X.Rep=Repeticao
    X.Y=Variavel
    X
    I=length(unique(X.Trat))
    J=length(unique(X.Rep))
    AT= by(X, :Trat) do df
    (m = mean(df.Y), s² = var(df.Y))
    end
    t = quantile(TDist(GLResiduo),1-Alpha/2)
    DMS=t*sqrt(2*QMResiduo/J)
    AT=sort(AT, :m, rev=true)
    A=AT.m
    B=A'
    DIF = [A[l]-B[c] for l in 1:I, c in 1:I]
    DIFA=[if abs(DIF[i,c])>=DMS diferença="*" elseif abs(DIF[i,c])<DMS
diferença="ns" end for i in 1:I, c in 1:I]
    DIFA=convert(DataFrame, DIFA)
    Trat=levels(AT.Trat)
    NomeTratamentos=convert(Array,AT.Trat)
    TabelaHipótese=convert(DataFrame, DIFA)
    TabelaHipótese=names!(TabelaHipótese,[Symbol(name) for name in
NomeTratamentos])
    Tratamentos=names(TabelaHipótese)
    TabelaSignificânciaDMS=insertcols!(TabelaHipótese, 1, :Tratamento =>
Tratamentos)
    println("Tabela de média: ", AT)
    println("DMS: ", DMS)
    println("Tabela de Significância: ", DIFA)
end

#DMS(Tratamento,Repeticao,Variavel,GLResiduo,QMResiduo,Alpha)
>DMS(X.Tratamento,X.Rep,X.PM,56,2.23,0.05)
```

julia

```
julia> DMS(X.Tratamento,X.Rep,X.PM,56,2.23,0.05)
Tabela de média: 4×3 DataFrame
```

Row	Trat	m	s²
	String	Float64	Float64
1	Testemunha	11.5281	1.95407
2	PQA	5.11817	1.30775
3	PQB	2.55956	1.83145
4	PQC	2.09816	1.09632

```
DMS: 1.0923327107409362
Tabela de Significância: 4×5 DataFrame
```

Row	Tratamento	Testemunha	PQA	PQB	PQC
	String	String	String	String	String
1	Testemunha	ns	*	*	*
2	PQA	*	ns	*	*
3	PQB	*	*	ns	ns
4	PQC	*	*	ns	ns

A função foi construída para efetuar a comparação entre todas as médias. Entretanto, apenas a primeira coluna da tabela de significância interessa para o presente caso (observe que a posição da Testemunha está em *Row* 1 na tabela de médias), na qual foram expressas as diferenças entre médias pelo teste DMS dos tratamentos definidos pela aplicação do preservativo em relação à testemunha. O DMS foi de 1,09 e os resultados revelam que houve diferença significante de todos os tratamentos com o uso do preservativo em relação ao tratamento testemunha, que por sua vez, apresentou a maior média de perda de massa. Logo, a aplicação dos preservativos na madeira propiciou a redução da perda de massa, quando comparado com as madeiras que não foram tratadas.

19.5 TESTE DE BONFERONI

No teste DMS a determinação da diferença mínima de significância foi calculada utilizando os graus de liberdade, o quadrado médio do erro experimental e o valor da distribuição de *t*.

No teste de Bonferoni a amplitude também é calculada utilizando os graus de liberdade, o quadrado médio do erro experimental e o valor da distribuição de *t*, porém utilizando o nível de significância de $\frac{\alpha}{m}$ para as comparações. Se as comparações forem efetuadas duas a duas, então, o número de comparações m é dado por:

$$m = \binom{I}{2} = \frac{I(I-2)}{2}$$

Em que I é o número de tratamentos.

O pesquisador determina o nível de significância α ($\alpha = 0{,}05$, por exemplo) para o experimento. No teste de Bonferoni o nível de significância utilizado em cada comparação corresponde a $\frac{\alpha}{m}$.

A diferença mínima de significância para o teste de Bonferoni é dada por:

$$DMS = t_{\left(GLE,\frac{\alpha}{m}\right)}\sqrt{2\frac{QME}{J}}$$

Em que:

$t_{(GLE\,\frac{\alpha}{m})}$ é o valor da distribuição de t para o nível de significância de $\frac{\alpha}{m}$ e do número de graus de liberdade do erro experimental, obtido da análise de variância.

Comandos para obtenção do teste de Bonferoni:

Arquivo: 39 - Função bonferroni para teste de médias de bonferoni

```
using DataFrames, Statistics, PyPlot, Distributions

function Bonferoni(Tratamento,Repeticao,Variavel,GLResiduo,QMResiduo,Alpha)
    X=DataFrame()
    X.Trat=Tratamento
    X.Rep=Repeticao
    X.Y=Variavel
    X
    I=length(unique(X.Trat))
    J=length(unique(X.Rep))
    AT= by(X, :Trat) do df
    (m = mean(df.Y), s² = var(df.Y))
    end
    NMC=2 #comparação de médias duas a duas
    m=(I*(I-1))/NMC
    alphaC=Alpha/m
    t=quantile(TDist(GLResiduo),1-alphaC/2)
    DMS=t*sqrt(2*(QMResiduo/J))
    AT=sort(AT, :m, rev=true)
    A=AT.m
    B=A'
    DIF = [A[l]-B[c] for l in 1:I, c in 1:I]
    DIFA=[if abs(DIF[i,c])>=DMS diferença="*" elseif abs(DIF[i,c])<DMS
diferença="ns" end for i in 1:I, c in 1:I]
    Trat=levels(AT.Trat)
    NomeTratamentos=convert(Array,AT.Trat)
    TabelaHipótese=convert(DataFrame, DIFA)
    TabelaHipótese=names!(TabelaHipótese,[Symbol(name) for name in
NomeTratamentos])
    Tratamentos=names(TabelaHipótese)
    TabelaSignificânciaDMS=insertcols!(TabelaHipótese, 1, :Tratamento =>
Tratamentos)
    TabelaDMS=convert(DataFrame, DIF)
    TabelaDMS=names!(TabelaDMS,[Symbol(name) for name in NomeTratamentos])
    TabelaDMS=insertcols!(TabelaDMS, 1, :Tratamento => Tratamentos)
    TabelaMédia=DataFrame(Tratamento=AT.Trat,Média=AT.m,Variância=AT.s²)
```

```
    println("Tabela de média: ", TabelaMédia)
    println("Diferença Mínima Significativa: ", DMS)
    println("Tabela DMS: ", TabelaDMS)
    println("Tabela de hipóteses DMS: ", TabelaSignificânciaDMS)
    errorbar(AT.Trat,AT.m,DMS,fmt="o")
    xlabel("Tratamento")
    ylabel("Y")
end

#Bonferoni(Tratamento,Repeticao,Variavel,GLResiduo,QMResiduo,Alpha)
```

19.6 TESTE DE TUKEY

Para a obtenção do desvio mínimo de significância pelo teste de Tukey são utilizados os graus de liberdade, o quadrado médio do erro experimental e a amplitude estudentizada (q).

A diferença mínima de significância – DMS (também simbolizada por Δ ou denominada de diferença honestamente significante) é calculada por:

$$DMS = q_{\alpha(I;GLE)} \sqrt{\frac{QME}{J}}$$

Em que:

$q_{\alpha(I;GLE)}$ é a amplitude estudentizada (valor tabelar), ao nível de significância α para I tratamentos e GLE graus de liberdade do erro expeimental da ANOVA.

Assim, toda diferença entre duas médias de tratamentos, que seja maior que DMS, é considerada significante ao nível α.

Os comandos para obtenção do teste de Tukey são:

ARQUIVO: 40 - FUNÇÃO TUKEY PARA TESTE DE MÉDIAS DE TUKEY

```
using DataFrames, Statistics, PyPlot, Distributions, StatsFuns

function Tukey(Tratamento,Repeticao,Variavel,GLResiduo,QMResiduo,Alpha)
    X=DataFrame()
    X.Trat=Tratamento
    X.Rep=Repeticao
    X.Y=Variavel
    X
    I=length(unique(X.Trat))
    J=length(unique(X.Rep))
    AT= by(X, :Trat) do df
    (m = mean(df.Y), s² = var(df.Y))
```

```
    end
    q=srdistinvccdf(I,GLResiduo,Alpha)
    DMS=q*sqrt(QMResiduo / 2 * (2 / J))
    AT=sort(AT, :m, rev=true)
    A=AT.m
    B=A'
    DIF = [A[l]-B[c] for l in 1:I, c in 1:I]
    DIFA=[if abs(DIF[i,c])>=DMS diferença="*" elseif abs(DIF[i,c])<DMS
diferença="ns" end for i in 1:I, c in 1:I]
    Trat=levels(AT.Trat)
    NomeTratamentos=convert(Array,AT.Trat)
    TabelaHipótese=convert(DataFrame, DIFA)
    TabelaHipótese=names!(TabelaHipótese,[Symbol(name) for name in
NomeTratamentos])
    Tratamentos=names(TabelaHipótese)
    TabelaSignificânciaDMS=insertcols!(TabelaHipótese, 1, :Tratamento =>
Tratamentos)
    TabelaDMS=convert(DataFrame, DIF)
    TabelaDMS=names!(TabelaDMS,[Symbol(name) for name in NomeTratamentos])
    TabelaDMS=insertcols!(TabelaDMS, 1, :Tratamento => Tratamentos)
    TabelaMédia=DataFrame(Tratamento=AT.Trat,Média=AT.m,Variância=AT.s²)
    println("Tabela de média: ", TabelaMédia)
    println("Diferença Mínimca Significativa: ", DMS)
    println("Tabela DMS: ", TabelaDMS)
    println("Tabela de hipóteses DMS: ", TabelaSignificânciaDMS)
    errorbar(AT.Trat,AT.m,DMS,fmt="o")
    xlabel("Tratamento")
    ylabel("Y")
end

#Tukey(Tratamento,Repeticao,Variavel,GLResiduo,QMResiduo,Alpha)
```

19.7 TESTE DE DUNCAN

No teste de Duncan o cálculo da diferença mínima de significância também é calculada utilizando o grau de liberdade e o quadrado médio do erro experimental, mas com uma amplitude estudentizada padronizada (z).

As diferenças mínimas de significância no teste de Duncan dependem do número de médias (ordenadas) que são abrangidas na comparação. Como o número de médias abrangidas pelo contraste (diferença) varia em cada comparação, o teste é denominado de comparação múltipla.

A diferença mínima de significância para o teste de Duncan, para cada contraste, é dada por:

$$DMS_u = Z_u \sqrt{\frac{QME}{J}}$$

Em que:

u é o número de médias de tratamentos abrangidas no contraste (1, 3, 4, ..., I).

Z_u é o valor correspondente aos GL_e, sendo u o número de médias abrangidas em cada contraste a um nível α de significância, cujo valor é obtido na tabela de valores de Duncan.

O teste de Duncan é sequencial e toda diferença entre duas médias de tratamentos maior que DMS_u é considerada significante ao nível α. As médias são ordenadas decrescentemente, o primeiro contraste é o de maior amplitude (diferença entre a maior e a menor média), e os demais são sequenciais até os contrastes de menor amplitude. A sequência de comparação de contrastes é efetuada somente até se observar um contraste não significante.

Comandos para obtenção do teste de Duncan em comparação de médias:

Arquivo: 41 - Função duncan para teste de médias de duncan

```
using Statistics, DataFrames

function Duncan(Tratamento,Repeticao,Variavel,GLResiduo,QMResiduo,z)
    X=DataFrame()
    X.Trat=Tratamento
    X.Rep=Repeticao
    X.Y=Variavel
    X
    I=length(unique(X.Trat))
    J=length(unique(X.Rep))
    AT= by(X, :Trat) do df
    (m = mean(df.Y), s² = var(df.Y))
    end
    z=sort(z)
    DMS=z.*sqrt(QMResiduo/J)
    AT=sort(AT, :m, rev=true)
    A=AT.m
    B=A'
    DIF = [A[l]-B[c] for l in 1:I, c in 1:I]
    DMS=[0; sort(DMS)]
    DIFA=[if abs(DIF[i,c])>DMS[abs(i-c)+1] diferença="*" elseif abs(DIF[i,c])<=
DMS[abs(i-c)+1] diferença="ns" end for i in 1:I, c in 1:I]
    Trat=levels(AT.Trat)
    NomeTratamentos=convert(Array,AT.Trat)
    TabelaHipótese=convert(DataFrame, DIFA)
    TabelaHipótese=names!(TabelaHipótese,[Symbol(name) for name in
NomeTratamentos])
    Tratamentos=names(TabelaHipótese)
    TabelaSignificânciaDMS=insertcols!(TabelaHipótese, 1, :Tratamento =>
Tratamentos)
    TabelaDMS=convert(DataFrame, DIF)
    TabelaDMS=names!(TabelaDMS,[Symbol(name) for name in NomeTratamentos])
    TabelaDMS=insertcols!(TabelaDMS, 1, :Tratamento => Tratamentos)
    TabelaMédia=DataFrame(Tratamento=AT.Trat,Média=AT.m,Variância=AT.s²)
    println("Tabela de média: ", TabelaMédia)
```

```
    println("Diferença Mínimca Significativa: ", DMS)
    println("Tabela DMS: ", TabelaDMS)
    println("Tabela de hipóteses DMS: ", TabelaSignificânciaDMS)
end

#Duncan(Tratamento,Repeticao,Variavel,GLResiduo,QMResiduo,z)

z=[3.22,3.15,3.06,2.91] #Valores tabelares de Duncan. Atualizar para novas
comparações.
```

19.8 TESTE DE SCHEFFÉ

Nos testes de médias apresentados anteriormente é calculada a diferença mínima de significância utilizando os graus de liberdade, o quadrado médio do erro experimental e a distribuição apropriada. No teste de Scheffé a diferença mínima de significância também é calculada por meio dos graus de liberdade e do quadrado médio do erro experimental, entretanto, utiliza a distribuição de F.

Nos testes de médias anteriores as comparações entre os tratamentos eram feitas aos pares. No teste de Scheffé isso também pode ser efetuado (hipóteses estabelecidas a posteriori), incluindo a possibilidade de comparar contrastes que não são aos pares (hipóteses estabelecidas a priori), o que é recomendado.

No teste de Scheffé a diferença mínima de significância também é denominada de diferença completamente significante. A diferença mínima de significância, para comparações aos pares, é calculada por:

$$DMS = \sqrt{(I-1)F_{I-1,GLE,\alpha}}\sqrt{2\frac{QME}{J}}$$

Comandos para obtenção do teste de Scheffé:

ARQUIVO: 42 - FUNÇÃO SCHEFFE PARA TESTE DE MÉDIAS DE SCHEFFÉ

```
#Função SCHEFFE para teste de médias de Scheffé

using DataFrames, Statistics, PyPlot, Distributions

function Scheffe(Tratamento,Repeticao,Variavel,GLResiduo,QMResiduo,Alpha)
    X=DataFrame()
    X.Trat=Tratamento
    X.Rep=Repeticao
    X.Y=Variavel
    X
    I=length(unique(X.Trat))
    J=length(unique(X.Rep))
    AT= by(X, :Trat) do df
```

```
    (m = mean(df.Y), s² = var(df.Y))
    end
    F=quantile(FDist(I-1, GLResiduo),(1-Alpha))
    DMS=sqrt((I-1)*F)*sqrt(2*(QMResiduo/J))
    AT=sort(AT, :m, rev=true)
    A=AT.m
    B=A'
    DIF = [A[l]-B[c] for l in 1:I, c in 1:I]
    DIFA=[if abs(DIF[i,c])>=DMS diferença="*" elseif abs(DIF[i,c])<DMS
diferença="ns" end for i in 1:I, c in 1:I]
    Trat=levels(AT.Trat)
    NomeTratamentos=convert(Array,AT.Trat)
    TabelaHipótese=convert(DataFrame, DIFA)
    TabelaHipótese=names!(TabelaHipótese,[Symbol(name) for name in
NomeTratamentos])
    Tratamentos=names(TabelaHipótese)
    TabelaSignificânciaDMS=insertcols!(TabelaHipótese, 1, :Tratamento =>
Tratamentos)
    TabelaDMS=convert(DataFrame, DIF)
    TabelaDMS=names!(TabelaDMS,[Symbol(name) for name in NomeTratamentos])
    TabelaDMS=insertcols!(TabelaDMS, 1, :Tratamento => Tratamentos)
    TabelaMédia=DataFrame(Tratamento=AT.Trat,Média=AT.m,Variância=AT.s²)
    println("Tabela de média: ", TabelaMédia)
    println("Diferença Mínimca Significativa: ", DMS)
    println("Tabela DMS: ", TabelaDMS)
    println("Tabela de hipóteses DMS: ", TabelaSignificânciaDMS)
    errorbar(AT.Trat,AT.m,DMS,fmt="o")
    xlabel("Tratamento")
    ylabel("Y")
end

#Scheffe(Tratamento,Repeticao,Variavel,GLResiduo,QMResiduo,Alpha)
```

19.9 TESTE DE SCOTT-KNOTT

Os testes apresentados até o momento revelam em alguns casos respostas ambíguas, em que um tratamento pode pertencer a dois ou mais grupos estabelecidos para o teste de média. Uma alternativa para obter respostas exclusivas na formação dos grupos é a aplicação do teste de Scott-Knott, de modo a agrupar médias de tratamentos em grupos homogêneos, sendo este um procedimento com maior objetividade e clareza na formação dos grupos de tratamentos.

Observe que nos testes de médias anteriores a diferença mínima de significância utiliza os graus de liberdade, o quadrado médio do erro experimental e a distribuição correspondente. No teste de Scott-Knott também são utilizados os graus de liberdade e o quadrado médio do erro experimental, porém utilizando a distribuição de Qui-quadrado.

O objetivo do teste de Scott-Knott é minimizar a soma de quadrados dentro dos grupos e maximizar a soma de quadrados entre os grupos, sem sobreposição de médias. No teste, os tratamentos são divididos em dois grupos de forma a maximizar a soma de quadrados entre eles. Para tanto, utiliza-se o que se denomina de razão de verossimilhança. O teste é repetido K vezes, até determinar subgrupos com um tratamento em cada grupo ou grupos de médias de tratamentos homogêneas.

As seguintes hipóteses são testadas:

$$\begin{cases} H_0: G_1 = G_2 \\ H_1: G_1 \neq G_2 \\ \alpha = 0{,}05 \end{cases}$$

Em que: G_1 e G_2 são os totais dos dois grupos de médias com k_1 e k_2 tratamentos cada.

A estatística do teste é dada por:

$$\lambda = \frac{\pi}{2 \cdot (\pi - 2)} \cdot \frac{\beta_0}{\hat{\sigma}_0^2}$$

A estatística λ tem distribuição de Qui-quadrado com $k/(\pi\text{-}2)$ graus de liberdade e nível de significância α . k é o número de partições, nas quais têm-se que $k = I - 1$, sendo I o número de médias comparadas.

As médias dos tratamentos devem ser ordenadas decrescentemente. Define-se o número de partições e obtém-se a soma de quadrados entre os grupos de cada partição - β_0, dada por:

$$\beta_0 = \frac{T_1^2}{k_1} + \frac{T_2^2}{k_2} - \frac{(T_1+T_2)^2}{(k_1+k_2)}$$

Em que:
β_0 é a estimativa da soma de quadrados entre grupos.
k_1 é o número de tratamentos no grupo 1.
k_2 é o número de tratamentos no grupo 2.
T_1 é o total das médias dos tramentos do grupo 1.
T_2 é o total das médias dos tramentos do grupo 2.

Para cada partição, obtém-se o valor de β_0, sendo necessário considerar o valor máximo entre as partições dos grupos para o teste. Após isto, efetua-se o cálculo do estimador de máxima verossimilhança, dado por:

$$\hat{\sigma}_0^2 = \frac{\sum_{i=1}^{k} (\overline{y}_i - \overline{y})^2 + GLE \frac{QME}{J}}{(k + GLE)}$$

Em que:
$\overline{y}_i$ é a média do tratamento i
$\overline{y}$ é a média geral dos tratamentos I
QME é o quadrado médio do erro experimental, obtido da análise de variância
GLE é o grau de liberdade do erro experimental, obtido da análise de variância
J é o número de repetições

O cálculo da estatística λ é dado por:

$$\lambda = \frac{\pi}{2 \cdot (\pi - 2)} \cdot \frac{\beta_0}{\hat{\sigma}_0^{\,2}}$$

Quando $\lambda \leq \chi^2$ a hipótese da nulidade H_0 não é rejeitada e se infere que os grupos não diferem entre si. Quando $\lambda > \chi^2$ a hipótese da nulidade H_0 é rejeitada, revelando que os grupos diferem entre si. Quando isso acontecer os tratamentos dos dois subgrupos formados devem ser testados entre si, repetindo-se os cálculos. O teste encerra quando H_0 não for rejeitada ou quando sobrar apenas uma média nos subgrupos. No final do teste as médias ordenadas são classificadas em seus respectivos grupos.

Comandos para obtenção dos cálculos do teste Scott-Knott:

Arquivo: 43 - Função sk para teste de médias de scott-knott

```
using Statistics, Distributions, DataFrames, CSV

function sk(Dados::DataFrame, J, QME, GLE, Alpha; verbose::Bool=true)
    Dados = sort(Dados, :m, rev=true)
    grupo = Dict{String,Int}()
    proximo_grupo = Ref(1)

    function particionar!(sub::DataFrame)
        p = nrow(sub)

        # caso base: um único tratamento não pode mais ser dividido
        if p == 1
            grupo[sub.Trat[1]] = proximo_grupo[]
            proximo_grupo[] += 1
            return
        end

        x  = sub.m
        nt = sub.Trat

        B0i = Vector{Float64}(undef, p - 1)
        for i in 1:(p - 1)
            t1, t2 = sum(x[1:i]), sum(x[i+1:end])
            k1, k2 = i, p - i
            B0i[i] = (t1^2) / k1 + (t2^2) / k2 - ((t1 + t2)^2) / (k1 + k2)
        end
        k_best = argmax(B0i)
        B0     = B0i[k_best]

        MG  = mean(x)
        S20 = (sum((x .- MG) .^ 2) + (QME / J) * GLE) / (p + GLE)
        Lambda = (pi / (2 * (pi - 2))) * (B0 / S20)
        nu   = p / (pi - 2)
        crit = quantile(Chisq(nu), 1 - Alpha)
        sig  = Lambda > crit
```

```
        if verbose
            println("---- Partição de $p tratamentos: ", nt)
            println("  Melhor corte: ", nt[1:k_best], " | ", nt[k_best+1:end])
            println("  B0 = ", round(B0, digits=4),
                    "  S20 = ", round(S20, digits=4),
                    "  Lambda = ", round(Lambda, digits=4),
                    "  Qui² crítico = ", round(crit, digits=4),
                    "  => ", sig ? "significativo (*)" : "não significativo (ns)")
        end

        if sig
            particionar!(sub[1:k_best, :])
            particionar!(sub[k_best+1:end, :])
        else
            # sem diferença: todo o subconjunto forma um único grupo
            for t in nt
                grupo[t] = proximo_grupo[]
            end
            proximo_grupo[] += 1
        end
    end

    particionar!(Dados)

    Dados.GrupoNum = [grupo[t] for t in Dados.Trat]
    letras = collect('a':'z')
    Dados.Grupo = [string(letras[g]) for g in Dados.GrupoNum]
    select!(Dados, Not(:GrupoNum))

    return Dados
end

QME   = 10.08
GLE   = 15
Alpha = 0.05
J     = length(X.Bloco)

Dados = combine(groupby(X, :Trat)) do df
    (m = mean(df.Y),)
end

sk(Dados, J, QME, GLE, Alpha)
```

19.10 TESTE DE DUNNETT

Os testes de médias anteriores são aplicados para comparar todos os pares de médias dos tratamentos, não havendo entre eles uma testemunha (ou tratamento controle). Nestas condições, é importante aplicar os testes estatísticos para discriminar qual é o melhor ou os melhores tratamentos.

Em alguns experimentos, um dos tratamentos corresponde a uma testemunha (controle). Neste caso, o objetivo é comparar os demais tratamentos contra a testemunha e a aplicação do teste estatístico revela quais tratamentos não diferem da testemunha. Para tanto, o teste de Dunnett pode ser efetuado.

A diferença mínima significante do teste de Dunnett, quando se tem o mesmo número de repetições para todos os tratamento, é dada por:

$$DMS = d_{\alpha(I;GLE)} \sqrt{2\frac{QME}{J}}$$

Em que:

$d_{\alpha(I;GLE)}$ é um valor tabelar de Dunnet, ao nível de significância α para I tratamentos comparados com a testemunha e GLE graus de liberdade do erro experimental na ANOVA; J é o numero de repetições. É importante observar que a estatística utilizada no teste de Dunnet segue a distribuição de t multivariada. Se os graus de liberdade da variância do erro experimental forem grandes, então a distribuição se aproxima da distribuição normal multivariada.

Comandos para obtenção do teste de Dunnet:

ARQUIVO: 44 - FUNÇÃO DUNNET PARA TESTE DE MÉDIAS DE DUNNET

```
using DataFrames, Statistics

function Dunnet(Tratamento,Repeticao,Variavel,GLResiduo,QMResiduo,Testemunha,d)
    X=DataFrame()
    X.Trat=Tratamento
    X.Rep=Repeticao
    X.Y=Variavel
    X
    I=length(unique(X.Trat))
    J=length(unique(X.Rep))
    AT= by(X, :Trat) do df
    (m = mean(df.Y), s² = var(df.Y))
    end
    AT=sort(AT, :m, rev=true)
    c=AT[in.(AT.Trat, Ref([Testemunha])), :]
    DMS=d*sqrt(2*QMResiduo/J)
    A=AT.m
    DIF = c.m.-A
    DIFA=[if abs(DIF[i])>=DMS diferença="*" elseif abs(DIF[i])<DMS diferença="ns"
end for i in 1:I]
```

```
    Trat=levels(AT.Trat)
    NomeTratamentos=convert(Array,AT.Trat)
    TabelaHipótese=convert(DataFrame, DIFA)
    TabelaHipótese=names!(TabelaHipótese,[Symbol(name) for name in
NomeTratamentos])
    Tratamentos=names(TabelaHipótese)
    TabelaSignificânciaDMS=insertcols!(TabelaHipótese, 1, :Tratamento =>
Tratamentos)
    TabelaDMS=convert(DataFrame, DIF)
    TabelaDMS=names!(TabelaDMS,[Symbol(name) for name in NomeTratamentos])
    TabelaDMS=insertcols!(TabelaDMS, 1, :Tratamento => Tratamentos)
    TabelaMédia=DataFrame(Tratamento=AT.Trat,Média=AT.m,Variância=AT.s²)
    println("Tabela de média: ", TabelaMédia)
    println("Diferença Mínimca Significativa: ", DMS)
    println("Tabela DMS: ", TabelaDMS)
    println("Tabela de hipóteses DMS: ", TabelaSignificânciaDMS)
end

#Dunnet(Tratamento,Repeticao,Variavel,GLResiduo,QMResiduo,Testemunha,d)

d= 2.04                      #valor tabelar de Dunnet.
                              Atualizar para novas comparações.
Testemunha= Testem           #indicar o nome do tratamento testemunha
```

20 ANÁLISE DE REGRESSÃO

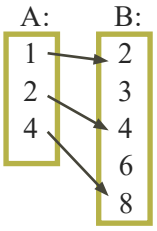
Regressão pode ser definida como a técnica que visa determinar a forma e a significância da relação funcional entre duas variáveis. A relação funcional diz respeito a função matemática a ser utilizada, e esta função não implica em uma relação de causa e efeito.

Desta forma, na Ciência Florestal a utilização de funções, além de desempenhar um papel fundamental, viabiliza o estabelecimento de importantes relações entre variáveis que expressam o comportamento funcional das florestas.

20.1 FUNÇÕES

Funções fazem parte de um campo da matemática no qual os modelos são formulados e a partir de dados experimentais equações são ajustadas. A formulação de modelos matemáticos permite compreender a relação entre variáveis para analisar um determinado fenômeno.

Uma função pode ser representada por tabelas, gráficos ou equações. Tomando-se um conjunto de dados A={1,2,4} e B={2,3,4,6,8}, nos quais em *B* é o dobro de *A*, utilizando-se a lei da correspondência para representar a relação entre A e B.



Todo elemento do conjunto *A* corresponde a um e somente um elemento no conjunto *B*. Quando estas condições não são satisfeitas não se tem uma função.

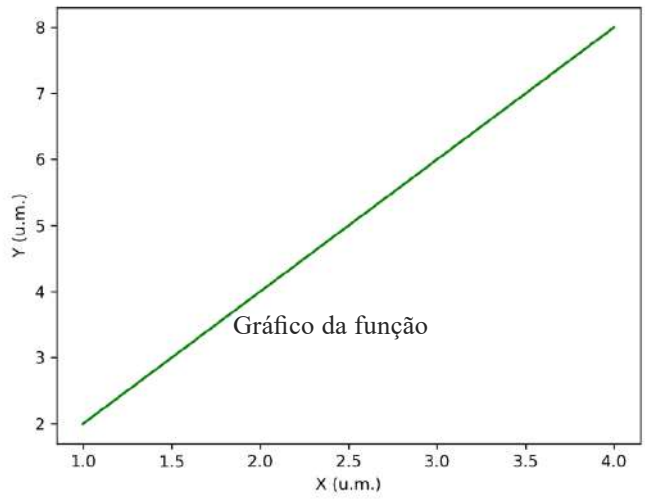
No exemplo, a função $f(x) = 2x$ foi representada por conjuntos. Uma função também pode ser representada por meio de um gráfico.

Comandos para obtenção do gráfico da função $f(x) = 2x$:

```
Arquivo: 45 - AR - funções

A=[1, 2, 4];    #Domínio
f(x)=2*x        #função
f(A)            #Imagem da função
2
4
8

using PyPlot
plot(A,f(A),"g")
xlabel("X (u.m.)")
ylabel("Y (u.m.)")
```

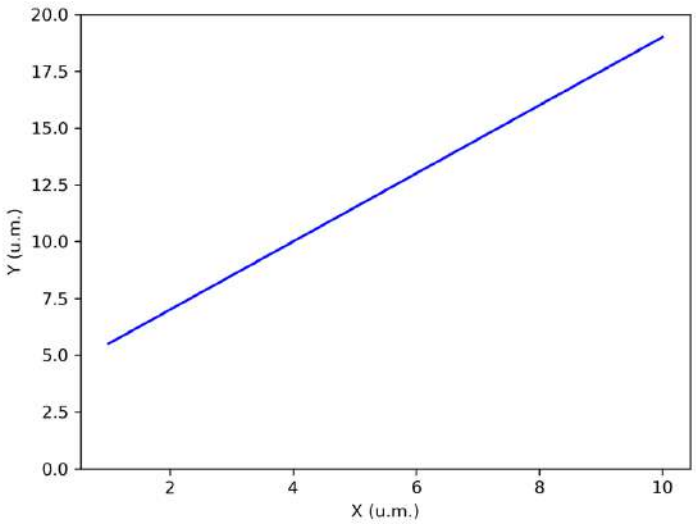


Considere um modelo que tem a forma Y=a+bX e sua equação $Y = 4 + 1{,}5\,X$.

Comandos para obtenção do gráfico da função $Y = 4 + 1{,}5\,X$:

Arquivo: 45 - AR - funções

```
#Gráfico da função Y=4+1,5*X
xGrid = 1:0.01:10;
a, b = 4, 1.5;
f(x) = a + b*x;
plot(xGrid,f.(xGrid),"b",label="função linear")
xlabel("X (u.m.)")
ylabel("Y (u.m.)")
ylim(0,20)
```



O gráfico obtido para a função $Y = 4 + 1{,}5\,X$ contém os seguintes elementos:

- A variável *Y* é denominada de variável dependente ou variável resposta.
- A variável *X* é denominada de variável independente ou variável explicativa.
- Nesta equação, quando $X = 0$, $Y = a$, ou seja, *a* é o valor onde o gráfico (linha azul) intercepta o eixo *Y*. O valor de *a* é denominado de intercepto.
- O intercepto pode ter o valor de zero e, neste caso, a linha do gráfico passa pela origem
- Uma unidade de mudança de *X*, resulta em *b* unidades de mudança em *Y*. Assim, *b* é uma medida de inclinação do gráfico (linha azul). Quando *b* é positivo ambas as variáveis (*X* e *Y*) aumentam, quando *b* é negativo uma variável aumenta e outra diminui.
- Para uma linha reta, quaisquer dois pontos determinam exclusivamente a posição da linha, ou seja, a inclinação e o intercepto. Isto é denominado de relação funcional, ou seja, existe uma fórmula matemática exata para explicar a relação entre as duas variáveis.
- A função gerada é uma função linear, mas existem outros modelos matemáticos importantes, tais como: polinomiais, racionais, exponenciais e logarítmicos, com ampla aplicação na Engenharia florestal.
- A técnica de regressão tem dois usos principais:
  - » Estimar uma variável de difícil obtenção em função de uma ou mais variáveis de fácil obtenção. Por exemplo, diâmetro à altura do peito *X* e o volume do fuste *Y*.
  - » Estabelecer qual a tendência da relação entre a variável *Y* e a(s) variável(eis) *X*(s). Por exemplo, dosagem de fertilizantes *X* e produção de biomassa *Y*.
- Engenheiros Florestais trabalham com modelos matemáticos ou modelos estatísticos.

20.2 REGRESSÃO LINEAR SIMPLES

A análise de regressão linear é aplicada ao estudo das relações de uma variável resposta (dependente) Y e uma ou mais variáveis explicativas (independentes) (*X1*, *X2*, …, *Xn*) por meio de uma função (modelo) linear, de tal forma que a variável resposta possa ser predita a partir das demais.

Um modelo de regressão pode ser linear ou não linear e simples ou múltiplo. Regressão linear simples é aquela que só tem uma variável independente. Regressão linear múltipla é aquela que tem mais do que uma variável independente. Um modelo de regressão é dito linear quando todos os coeficientes são elevados a potência unitária. Caso isso não ocorra o modelo é chamado de não linear.

O modelo de regressão linear simples é representado por:

$$Y_i = \beta_0 + \beta_1 \cdot X1_i + \varepsilon_i$$

enquanto o modelo de regressão linear múltiplo é representado por:

$$Y_i = \beta_0 + \beta_1 \cdot X1_i + \beta_2 \cdot X2_i + \beta_3 \cdot X3_i + \varepsilon_i$$

Engenheiros Florestais rotineiramente trabalham com observações bivariadas (pares de valores de X,Y), nas quais os dados não expressam uma relação perfeita. Por exemplo, árvores com o mesmo valor de diâmetro à altura do peito na floresta (*variável X*) podem ter diferentes alturas (*variável Y*).

Entretanto, pode ser que exista uma relação funcional implícita entre as variáveis, relação esta que não é perfeita. No exemplo, é apropriado pensar que árvores com diâmetros menores têm alturas menores, árvores com diâmetros maiores têm alturas maiores. Assim, ao invés de estabelecer uma relação matemática (ou seja exata), é apropriado estabelecer um modelo estatístico e ajustar uma função que minimize a soma dos quadrados dos desvios (diferença entre os valores observados e estimados por meio da função).

Considere o seguinte modelo estatístico:

$$Y = \alpha + \beta X + \varepsilon$$

Em que:
α e β descrevem o intercepto e a inclinação, respectivamente
ε é o termo de erro

A diferença principal entre o modelo matemático e o modelo estatístico é a inclusão de um termo de erro. Embora diferentes, ambos têm o mesmo objetivo: estabelecer uma função de correspondência para descrever a relação entre as variáveis.

Engenheiros Florestais costumam mensurar o diâmetro à altura do peito (d) e o volume do fuste. Uma vez coletados os dados, modelos estatísticos podem ser ajustados visando obter a melhor função que represente a relação entre d e o volume do fuste.

Considere o exemplo em que um Engenheiro Florestal está interessado em avaliar a relação que o diâmetro à altura do peito (*d*) tem com a altura (*h*) das árvores. A variável dependente é a altura das árvores que varia de acordo com a variável independente diâmetro à altura do peito (cm), pois esta é uma variável mais acessível e de menor custo para ser medida. No exemplo, em uma floresta de acácia-negra foi obtida uma amostra aleatória de 62 pares de valores dessas variáveis, apresentada no arquivo *descritiva*.

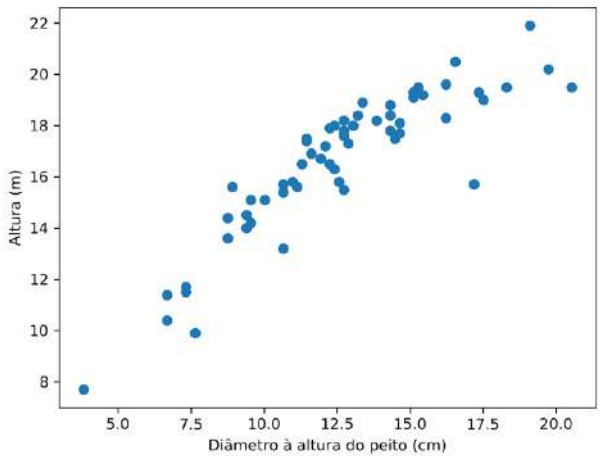
Comandos para obtenção do diagrama de dispersão entre altura e diâmetro:

Arquivo: 46 - Regressão linear simples

```
>using DataFrames, CSV, PyPlot
```

```
julia> X=CSV.read("Z:/Alexandre/Julia/Exemplos/descritiva.csv", DataFrame)
62×8 DataFrame
| Row | Idade   | d       | h       | cc      | b       | me      | v       | ff      |
|     | Float64 | Float64 | Float64 | Float64 | Float64 | Float64 | Float64 | Float64 |
|-----|---------|---------|---------|---------|---------|---------|---------|---------|
| 1   | 10.75   | 19.74   | 20.2    | 7.7     | 206.09  | 0.65    | 0.28    | 0.45    |
| 2   | 10.75   | 15.12   | 19.1    | 4.9     | 93.17   | 0.5     | 0.16    | 0.48    |
| 3   | 10.75   | 12.89   | 17.3    | 3.0     | 65.88   | 0.57    | 0.11    | 0.47    |
| 4   | 10.75   | 11.62   | 16.9    | 7.2     | 49.3    | 0.52    | 0.09    | 0.5     |
| 5   | 10.75   | 16.55   | 20.5    | 11.8    | 142.0   | 0.56    | 0.22    | 0.5     |
| 6   | 10.75   | 19.1    | 21.9    | 7.0     | 194.87  | 0.6     | 0.28    | 0.45    |
⋮
| 56  | 10.75   | 12.25   | 16.5    | 8.1     | 69.62   | 0.72    | 0.1     | 0.51    |
| 57  | 10.75   | 10.03   | 15.1    | 3.7     | 56.07   | 0.59    | 0.06    | 0.48    |
| 58  | 10.75   | 12.73   | 17.6    | 7.4     | 89.64   | 0.54    | 0.12    | 0.54    |
| 59  | 10.75   | 14.64   | 17.7    | 12.1    | 104.02  | 0.61    | 0.16    | 0.53    |
| 60  | 10.75   | 12.73   | 18.2    | 9.2     | 85.58   | 0.58    | 0.12    | 0.53    |
| 61  | 10.75   | 7.32    | 11.7    | 7.5     | 26.23   | 0.51    | 0.03    | 0.59    |
| 62  | 10.75   | 18.3    | 19.5    | 7.6     | 179.41  | 0.54    | 0.24    | 0.47    |
```

```
plot(X.d,X.h,"o",label="dados")
xlabel("Diâmetro à altura do peito (cm)")
ylabel("Altura (m)")
```



Um gráfico da forma $Y = \alpha + \beta X$ pode ser utilizado para expressar a relação funcional entre as variáveis h e d. Pode-se, então, estimar os valores de α e β com o mínimo de erro possível, pois trata-se de um modelo estatístico. Isto é obtido pelo método dos mínimos quadrados. Este método assegura a obtenção de equações com o mínimo de erro possível. Dada uma série de pares de valores: $(x_1, y_2),(x_2, y_2),\ldots,(x_n, x_n)$, o método de mínimos quadrados tem como objetivo ajustar a melhor função para representar a relação entre X e Y. O modelo ajustado pode ser representado por $\hat{y} = \hat{\alpha} + \hat{\beta}x$, em que $\hat{\alpha}$ e $\hat{\beta}$ são estimadores de α e β, obtidos a partir dos dados.

Para qualquer valor de α e β a soma dos quadrados do resíduos ou do erro é dada por:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 + \beta_1 x_i)^2$$

Em que:
$\sum_{i=1}^{n} e_i^2$ é a soma dos quadrados dos resíduos
y_i é o i-ésimo valor observado
$\hat{y}_i$ é o i-ésimo valor predito

Estimativas de α e β são representadas como $\hat{\alpha}$ *e* $\hat{\beta}$, $\hat{\beta}_0$ e $\hat{\beta}_1$ ou ainda a e b.

Assim, a tarefa consiste em estimar os valores de $\hat{\alpha}$ *e* $\hat{\beta}$ de forma a minimizar a soma de quadrados dos resíduos. A solução, para um modelo de regressão linear simples, é dada algebricamente por:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

Esta solução tem as seguintes propriedades:
O ponto $(\bar{y}, \bar{x})$ corresponde ao gráfico de regressão da amostra.
A soma dos desvios (resíduos) obtidos pela equação ajustada é zero $\sum (y_i - \hat{y}_i) = 0$.
A soma dos quadrados dos resíduos é mínima.

EXERCÍCIO 20: RELAÇÃO HIPSOMÉTRICA

Para os dados de altura e diâmetro à altura do peito, ajuste o modelo: $Y = \alpha + \beta X + \varepsilon$.

Comandos para o ajuste de um modelo de regressão linear simples algebricamente:

ARQUIVO: 46 - REGRESSÃO LINEAR SIMPLES

```
function rls(x,y)
    Σx = sum(x)
    Σy = sum(y)
    Σxy = sum(x.*y)
    Σx² = sum(x.^2)
    n = length(x)
    ybar=Σy/n
```

```
    xbar= Σx/n
    β1= (Σxy-(Σx*Σy)/n)/(Σx²- (Σx^2)/n)
    β0= ybar- β1* xbar
    Coeficientes=[ β0, β1]
end

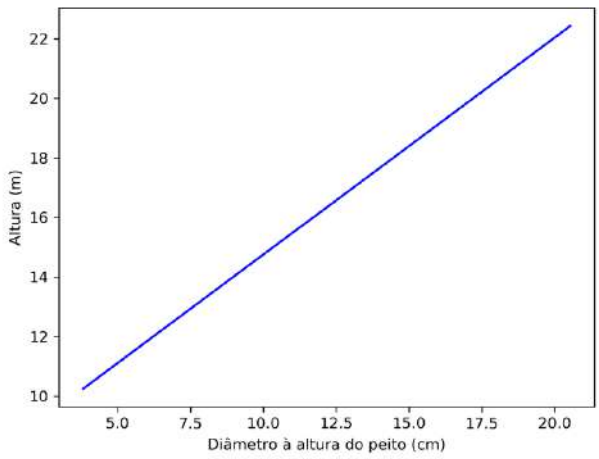
#rls(x,y)
rls(X.d,X.h)
7.4671
0.7287
```

O valor do coeficiente $\hat{\beta}_0$ encontrado foi de 7,4671 e de $\hat{\beta}_1$ 0,7287. A equação ajustada pode ser representada como $\hat{y} = 7{,}4671 + 0{,}7287\,x$ ou $\hat{h} = 7{,}4671 + 0{,}7287\,d$.

Comandos para representar o gráfico da função $\hat{y} = 7{,}4671 + 0{,}7287\,x$:

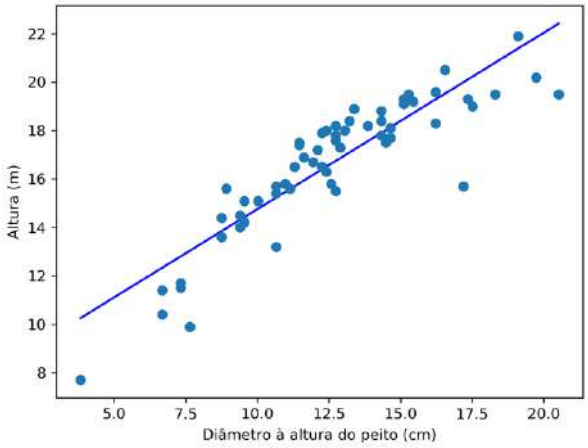
Arquivo: 46 - Regressão linear simples

```
using PyPlot
x = minimum(X.d):0.01: maximum(X.d)
f(x) =7.4671+0.7287*x
plot(x,f.(x),"b")
xlabel("Diâmetro à altura do peito (cm)")
ylabel("Altura (m)")
```



Comandos para representação dos valores observados utilizados no ajuste.

```
plot(X.d,X.h,"o",label="dados")
```



Comandos para predizer a altura das árvores ($\hat{y}$) para um novo conjunto de valores de d:

Arquivo: 46 - Regressão linear simples

```
d=[5.2,6.1,7.8,9.10,10,5,12.8,15.2,17.5,19.2];
ŷ=[]
for i in d
    yhat = 7.4671+0.7287*i
    append!(ŷ,yhat)
end
ŷ
10-element Array{Any,1}:
 11.26
 11.91
 13.15
 14.10
 14.75
 11.11
 16.79
 18.54
 20.22
 21.46

Comandos alternativos para estimar as alturas.
f(x) =7.4671.+0.7287*x;
f(d)
11.26
 11.91
 13.15
 14.10
 14.75
 11.11
 16.79
 18.54
 20.22
 21.46
```

```
f(x)=7.4671+0.7287*x;
f.(d)
11.26
 11.91
 13.15
 14.10
 14.75
 11.11
 16.79
 18.54
 20.22
 21.46
```

Dados os valores observados e preditos, questiona-se: A equação ajustada resultou em uma boa função de correspondência para os dados da amostra? Como avaliar a equação ajustada? Será que o modelo estatístico estabelecido é o melhor para representar a função intrínseca dos dados amostrais?

Existem estatísticas que auxiliam nas respostas a estas questões, bem como para auxiliar a escolha da melhor equação ajustada.

20.3 ESTATÍSTICAS DE AJUSTES DA EQUAÇÃO

Após o ajuste da equação de regressão sua avaliação é necessária. Para esse propósito existem várias estatísticas que podem ser aplicadas. A maneira de se realizar e obter estas estatísticas parte da elaboração de um quadro de análise de variância da equação ajustada, a partir do qual é possível encontrar estas estatísticas.

20.3.1 ANÁLISE DE VARIÂNCIA DA EQUAÇÃO AJUSTADA

Na análise de uma equação linear ajustada, existem três fontes de variação:

- Variação Total (variabilidade dos valores de Y)
- Variação devido à regressão
- Variação devido aos resíduos

Estas informações são resumidas em uma tabela de análise de variância. O quadro resumo da análise de variância contém os graus de liberdade, as somas de quadrados e os quadrados médios (variância) correspondentes a cada uma das fontes de variação citadas.

O teste de F também é calculado e representa a razão de dois quadrados médios (quadrado médio da regressão/quadrado médio do resíduo).

20.3.2 ESTATÍSTICA F

Por meio da estatística F é possível testar se o modelo ajustado representa ou não a relação existente entre a variável dependente e independente. As hipóteses testadas pela razão F são:

$$\begin{cases} H_0\text{: o modelo ajustado não representa a relação entre X e Y} \\ H_1\text{: o modelo ajustado representa a relação entre X e Y} \\ \alpha=0,05 \end{cases}$$

A estatística F é dada por:

$$F_c = \frac{QMR}{QME}$$

$$QMR = \frac{SQR}{GLR}$$

$$QME = \frac{SQE}{GLE}$$

$$SQE = \sum\left(y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i\right)^2 = \sum(y_i - \hat{y}_i)^2$$

$$SQR = \sum(\hat{y}_i - \bar{y})^2$$

$$SQT = \sum(y_i - \bar{y})^2$$

$$GLR = n_\beta - 1$$

$$GLE = GLT - GLR$$

$$GLT = n - 1$$

Em que:

QMR é o quadrado médio da regressão, *QME* é o quadrado médio do erro ou resíduo, *SQE* é a soma de quadrados do erro ou resíduo, *SQR* é a soma de quadrados da regressão, *SQT* é a soma de quadrados total, n_β é o número de coeficientes, n é o tamanho da amostra, *GLR* são os graus de liberdade da regressão, *GLE* são os graus de liberdade do erro ou resíduo e *GLT* são os grau de liberdade total.

Para o exercício *relação hipsométrica*, calcule a estatística F e faça a conclusão do teste de hipótese.

Comandos para a obtenção da análise de variância de um modelo linear simples:

Arquivo: 47 - AR - Estatísticas de ajustes da equação

```
>using Statistics, Distributions

>function Frl(x,y,yhat,nβ)
    n=length(x)
    GLR= nβ-1
    GLT=n-1
    GLE= GLT- GLR
    SSR= sum((yhat.- mean(y)).^2)
    SSE= sum((yhat- y).^2)
    SST = sum((y.- mean(y)).^2)
    MSR=SSR/GLR
    MSE=SSE/GLE
    Fc=MSR/MSE
    AlphaRegressao= ccdf(FDist(GLR, GLE),abs(Fc))
    println("Fc: ", Fc)
    println("α: ", AlphaRegressao)
end

>using DataFrames, CSV
>X=CSV.read("Z:/Alexandre/Julia/Exemplos/descritiva.csv", DataFrame);
f(x) =7.4671.+0.7287*x  #equação ajustada
yhat=f(X.d)             #obtenção dos valores estimados

#Frl(x,y,yhat,nβ)
Frl(X.d,X.h,yhat,2)
Fc: 225.27
α: 5.59e-22
```

O valor da estatística F para a regressão foi de 225,27 e o valor de α foi menor que 0,05, revelando evidências de que a hipótese da nulidade deva ser rejeitada. Portanto, a equação ajustada representa a relação entre as variáveis *h* e *d*.

A soma da decomposição das somas de quadrados é válida para regressão linear e permite o cálculo de estatísticas complementares, valendo-se dos valores obtidos na análise de variância. Essas estatísticas são frequentemente apresentadas na análise de regressão com o objetivo de avaliar a qualidade do ajuste da função aos dados amostrais.

20.3.3 DESVIO-PADRÃO DO ERRO (OU RESÍDUO)

Esta estatística também é conhecida como erro-padrão da estimativa e é uma medida da variabilidade das estimativas obtidas pela equação ajustada. Ela mede o erro médio cometido quando se utiliza a equação ajustada para obtenção de valores preditos. Quanto menor for o valor do erro padrão da estimativa, melhor o ajuste da equação.

$$S_{YX} = \sqrt{QME}$$

20.3.4 DESVIO-PADRÃO EM PORCENTAGEM

Esta estatística também é conhecida como erro-padrão da estimativa em percentagem ou coeficiente de variação (CV). Valores menores desta estatística são mais desejados.

$$CV = \frac{S_{YX}}{\bar{y}}$$

20.3.5 COEFICIENTE DE DETERMINAÇÃO

O coeficiente de determinação mede o quanto da relação existente entre as variáveis está representada na equação ajustada e é importante porque é de fácil interpretação.

$$R^2 = 1 - \frac{SQE}{SQT}$$

A estatística varia no intervalo entre 0 e 1. Quando R² se aproxima de 1, indica que a variação residual é pequena. Quando R^2 se aproxima de 0, indica que a variação residual é grande. Esta estatística revela o quanto a regressão sobre X explica a variação de Y, logo maiores valores de R^2 são mais bem avaliados. Se o valor de $R^2 = 1$ então a relação linear é perfeita, enquanto $R^2 = 0$ indica não existir relação linear entre as variáveis.

20.3.6 COEFICIENTE DE DETERMINAÇÃO AJUSTADO

$$\bar{R}^2 = 1 - \frac{QME}{QMT}$$

Em que:
QME é o quadrado médio do resíduo.
QMT é o quadrado médio total $\left(QMT = \frac{SQT}{n-1}\right)$

A estatística $\bar{R}^2$ (ou R^2_{aj}) é similar a R^2, porém em seu cálculo são utilizados os valores dos quadrados médios. O motivo para isso é a aplicação de uma correção devida aos graus de liberdade, especialmente em regressões com mais de uma variável independente. O valor ordinário de R^2 nunca diminuirá ao se acrescentar mais variáveis independentes em um modelo, mesmo que elas não contribuam para aumentar a explicação da variação de Y. O valor de $\bar{R}^2$ pode aumentar ou diminuir quando são acrescentadas mais variáveis independentes a um modelo.

20.3.7 COEFICIENTE DE CORRELAÇÃO LINEAR (R)

O coeficiente de correlação mede o grau de associação entre as duas variáveis (X, Y) e o sinal de r é o mesmo do coeficiente angular.

$$r = \pm\sqrt{R^2}$$

A estatística *r* varia de -1 a +1, ou seja, entre correlação negativa perfeita e correlação positiva perfeita. Quanto mais próximo r estiver de 1 ou -1, mais forte será a relação entre *X* e *Y*. Quando o valor de $r = 0$, significa não existir uma relação entre *X* e *Y*, logo *X* não é apropriado para prever *Y*.

Na avaliação do ajuste de várias equações é comum se apresentar um resumo com os valores numéricos obtidos para as estatísticas: β_0, β_1, *QME*, R^2, $\bar{R}^2$, S_{YX}, *CV* e *F*. Complementarmente, uma análise gráfica dos resíduos e testes de significância para os coeficientes podem ser obtidos.

Comandos para obtenção das estatísticas do ajuste de uma equação de regressão linear simples:

Arquivo: 47 - ar - Estatísticas de ajustes da equação

```
>using Statistics, Distributions

>function Resumo_rl(x,y,yhat,nβ)
    n=length(x)
    GLR= nβ-1
    GLT=n-1
    GLE= GLT- GLR
    SQR= sum((yhat.- mean(y)).^2)
    SQE= sum((yhat- y).^2)
    SQT = sum((y.- mean(y)).^2)
    QMR=SQR/GLR
    QME=SQE/GLE
    Fc=QMR/QME
    SYX=sqrt(QME)
    CV= SYX/mean(y)*100
    R²=1-(SQE/SQT)
    R²aj=1-(QME/(SQT/GLT))
    println("Fc: ", Fc)
    println("SYX: ", SYX)
    println("CV (%): ", CV)
    println("R²: ", R²)
    println("R²aj: ", R²aj)
end

f(x) =7.4671.+0.7287*x  #equação ajustada
yhat=f(X.d)             #obtenção dos valores estimados

#Resumo_rl(x,y,yhat,nβ)
Resumo_rl(X.d,X.h,yhat,2)
Fc: 225.27
SYX: 1.28
CV (%): 7.68
R²: 0.79
R²aj: 0.79
```

Conforme exposto na descrição de cada uma das estatísticas e por meio da avaliação dos valores obtidos, é possível concluir que a equação ajustada é apropriada para predizer a altura das árvores. Análises adicionais são requeridas, como testes de significância para os coeficientes e análise gráfica dos resíduos.

20.3.8 TESTES DE HIPÓTESES SOBRE OS COEFICIENTES

A razão F é um teste global para os coeficientes da equação ajustada. Uma análise complementar pode ser efetuada para identificar se todos os coeficientes na equação ajustada são significantes. Por exemplo, se o coeficiente que representa o intercepto ($\hat{\beta}_0$) não for significante, uma regressão que passa pela origem pode ser ajustada. Em outras situações, prioriza-se avaliar se os outros coeficientes da equação são significantes, com objetivo de melhor compreender a bondade das relações entre as variáveis que estão sendo estudadas.

Testes de hipóteses podem ser aplicados para avaliar:

$$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$$
$$\alpha = 0{,}05$$

Em que a hipótese de nulidade explicita não existir efeito de X sobre Y, enquanto a hipótese alternativa explicita que existe tal efeito.

O mesmo procedimento pode ser usado para testar se a equação passa pela origem:

$$\begin{cases} H_0: \beta_0 = 0 \\ H_1: \beta_0 \neq 0 \end{cases}$$
$$\alpha = 0{,}05$$

Para ambos, um teste t pode ser aplicado para avaliar a hipótese da nulidade:

$$t_i = \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}}$$

Se $i = 1$ então o teste se refere ao coeficiente angular e se $i = 0$ ao intercepto. A estatística t_i tem distribuição de t com $n - 2$ graus de liberdade.

Para uma amostra de tamanho n, um grau de liberdade é devido a estimativa da média, equivalendo a n-1 graus de liberdade para a variação de Y. Para ajustar os coeficientes sobre X mais um grau de liberdade é perdido, resultando em $n - 2$ graus de liberdade para a soma de quadrados do erro (SQE). O quadrado médio do erro ou resíduo (QME ou S^2_{YX}) é dado por $QME = SQE / (n - 2)$, para um modelo linear simples, o que também é conhecido como variância da equação.

A estimativa do erro padrão dos coeficientes ($S_{\hat{\beta}i}$) depedende do coeficiente que é testado.

Para $i = 1$: $S_{\hat{\beta}_1} = \sqrt{\frac{QME}{S_{xx}}}$

Para $i = 0$: $S_{\hat{\beta}_0} = \sqrt{QME\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x}\right)}$

Sendo, $SS_x = \sum(x_i - \bar{x})^2$

Comandos para testar as hipóteses de que $H_0: \beta_1 = 0$ e $H_0: \beta_0 = 0$:

Arquivo: 47 - ar - Estatísticas de ajustes da equação

#Teste das hipóteses $H_0: \beta_1 = 0$ e $H_0: \beta_0 = 0$

```
>using Statistics, Distributions
>function t_rls(x,y,yhat,β0,β1)
    n=length(x)
    SSE= sum((yhat- y).^2)
    MSE=SSE/(n-2)
    SSx=sum((x.- mean(x)).^2)
    S1=sqrt(MSE/SSx)
    t1= β1/S1
    xbar= mean(x)
    S0=sqrt(MSE*((1/n)+(xbar^2)/SSx))
    t0= β0/S0
    αβ0= 2*ccdf(TDist(n-2), abs(t0))
    αβ1= 2*ccdf(TDist(n-2), abs(t1))
    println("t0: ", t0)
    println("αβ0: ", αβ0)
    println("t1: ", t1)
    println("αβ1: ", αβ1)
end

#t_rls(x,y,yhat,β0,β1)
t_rls(X.d,X.h,yhat, 7.4671, 0.7287)
t0: 11.76
αβ0: 3.24e-17
t1: 15.01
αβ1: 5.59e-2
```

O valor da estatística *t* para o coeficiente angular foi de 15,01 e o valor de α correspondente foi muito pequeno, revelando que a estatística é altamente significante e, portanto, a hipótese da nulidade não é verdadeira, ou seja, esta deve ser rejeitada e, consequentemente, a altura das árvores está relacionada com o diâmetro à altura do peito.

Um intervalo de confiança para o valor paramétrico (desconhecido) de β_i pode ser estabelecido:

$$IC\left[\hat{\beta}_i - t_{\alpha,GLE}\ S_{\hat{\beta}_i} \leq \beta_i \leq \hat{\beta}_i + t_{\alpha,GLE}\ S_{\hat{\beta}_i}\right] = 1 - \alpha$$

Se $i = 1$ o teste refere-se ao coeficiente angular, se $i = 0$ ao intercepto.

Comandos para obtenção dos intervalos de confiança para β_1 e β_0:

Arquivo: 47 - ar - Estatísticas de ajustes da equação

```
#Intervalos de confiança para  β_1 e β_0

>using Statistics, Distributions
>function IC_rls(x,y,yhat,β0,β1,α)
    n=length(x)
    SQE= sum((yhat- y).^2)
    QME=SQE/(n-2)
    SSx=sum((x.- mean(x)).^2)
    S1=sqrt(QME/SSx)
    xbar= mean(x)
    S0=sqrt(QME*((1/n)+(xbar^2)/SSx))
    tα = quantile(TDist(n-2),1- α/2)
    ICInferior_β0 = β0 - tα * S0
    ICSuperior_β0 = β0 + tα * S0
    ICInferior_β1 = β1 - tα * S1
    ICSuperior_β1 = β1 + tα * S1
    println("ICInferior_β0: ", ICInferior_β0)
    println("ICSuperior_β0: ", ICSuperior_β0)
    println("ICInferior_ β1: ", ICInferior_β1)
    println("ICSuperior_ β1: ", ICSuperior_β1)
end

#IC_rls(x,y,yhat,β0,β1,α)
IC_rls(X.d,X.h,yhat, 7.4671, 0.7287,0.05)
ICInferior_β0: 6.20
ICSuperior_β0: 8.74
ICInferior_ β1: 0.63
ICSuperior_ β1: 0.83
```

$$IC\ [6{,}20 \leq \beta_0 \leq 8{,}74] = 0{,}95$$
$$IC\ [0{,}63 \leq \beta_1 \leq 0{,}83] = 0{,}95$$

Ressalta-se que o melhor estimador para β_1 corresponde a 0,7287. Entretanto, devido à variabilidade dos dados na amostra, o aumento de altura em função de uma unidade de diâmetro à altura do peito pode variar entre 0,63 e 0,83, com 95% de probabilidade.

Um intervalo de confiança foi estabelecido para β_0, mas isto tem pouca importância prática, porque seria equivalente a conhecer a altura para um diâmetro à altura do peito $X = 0$, ou seja, um modelo que passa pela origem. Essa é uma questão importante, dado que as alturas das árvores são mensuradas somente quando $h \geq 1{,}30$ m.

Em relação ao intervalo de confiança para o coeficiente angular, se o intervalo não incluir o valor 0, existem evidências de que X e Y estão relacionados. Se X e Y não estiverem relacionados, $\hat{\beta}_1$ será 0, e o intervalo de confiança indica que o modelo não é útil para obtenção de estimativas.

20.3.9 ANÁLISE DE RESÍDUOS

Para os resíduos, é importante observar as seguintes condições:

- Ter média zero
- Ter a mesma variância (σ^2), ou seja, ser homogêneo
- Não serem correlacionados
- Serem normalmente distribuídos

A análise gráfica dos resíduos é muito interessante, especialmente quando seus valores são plotados para os valores de X. Nesta análise avalia-se:

- Se os resíduos estão distribuídos de forma homogênea para a variável X.
- Se os resíduos estão distribuídos de forma aleatória para a variável X.
- Se os resíduos tendem a ter o mesmo sinal nas duas extremidades do gráfico e de sinal oposto no meio, o que pode evidenciar ser a função de correspondência não linear.
- Se a amplitude dos resíduos aumenta para X isto indica que não são homogêneos. Neste caso, existe evidências de heterogeneidade de variância, possivelmente devido à natureza das variáveis. Isto sugere que a regressão pode ser ajustada com a aplicação de uma ponderação.
- Valores considerados *outliers* ou extremos podem ser detectados previamente em uma análise exploratória. A plotagem dos resíduos pode auxiliar também nestas circunstâncias, ou seja, espera-se observar nenhum resíduo discrepante.
- Plotar os resíduos para outras variáveis envolvidas na modelagem também pode revelar alguma informação interessante no entendimento do problema em questão.
- Estas características são analisadas calculando-se os resíduos ordinários (e_i), isto é, $e_i = y_i - \hat{y}_i$.

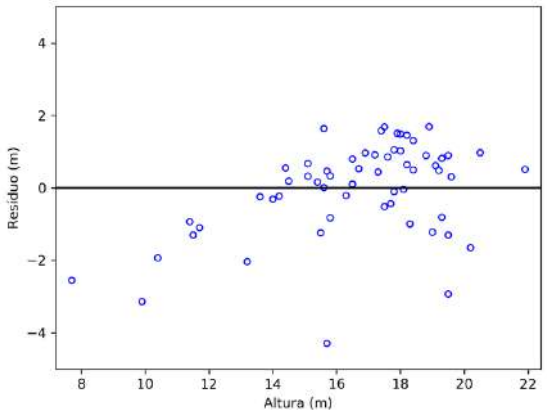
PEDE-SE:

Para o exercício *relação hipsométrica*, construa uma Figura para representar a diferença entre os valores observados e preditos, ou seja, os resíduos. Plote os valores dos resíduos em função das variáveis altura e diâmetro à altura do peito e faça uma interpretação das figuras.

Comandos para obtenção dos resíduos:

Arquivo: 47 - ar - Estatísticas de ajustes da equação

```
#Análise de Resíduos
>using PyPlot
f(x) =7.4671.+0.7287*x
f(X.d);
ei=X.h-f(X.d)
-1.6516
  0.6150
  0.4400
  0.9654
  0.9729
  0.5147
  0.4979
  0.8150
  0.5322
  0.8984
 -0.9939
  0.8150
 -2.9273
 ...
 -0.3096
  1.6820
 -0.9348
  0.6738
  0.3318
  0.1063
  0.3241
  0.8565
 -0.4353
  1.4565
 -1.1012
 -1.3023
scatter(X.h, ei, facecolor= "none", edgecolors="blue", s=20)
xlabel("Altura (m)")
ylabel("Resíduo (m)")
maximum(ei)
minimum(ei)
ylim(-5, 5)
xlim(minimum(X.h)-0.5, maximum(X.h)+0.5)
plot([0, maximum(X.h)+0.5],[0,0],c=:black)
```




```
scatter(X.d, ei, facecolor= "none", edgecolors="orange", s=20)
xlabel("Diâmetro à altura do peito (cm)")
ylabel("Resíduo (m)")
ylim(-5, 5)
xlim(minimum(X.d)-0.5, maximum(X.d)+0.5)
plot([0, maximum(X.d)+0.5],[0,0],c=:black)
```

Os resíduos estão distribuídos de forma aleatória para as variáveis *d* e *h*. A amplitude dos resíduos não aumenta para a variável *d*, logo existe evidência de serem homogêneos. Não foram observados valores considerados *outliers* ou extremos, uma vez que nenhum resíduo foi discrepante dos demais. Os resíduos têm o mesmo sinal nas duas extremidades do gráfico para a variável *d* e de sinal oposto no meio, o que evidencia ser a função de correspondência não linear. Uma alternativa para isto, para manter o modelo linear, é a transformação de dados das variáveis.

Assim, por meio da análise gráfica dos resíduos, sugere-se testar outros modelos para expressar a relação hipsométrica, tal como: $Y = \alpha + \beta X^2 + \varepsilon$. Logo, uma atividade comum realizada por Engenheiros Florestais é testar vários modelos visando escolher a melhor função de correspondência subjacentes aos dados amostrais para expressar a relação do diâmetro à altura do peito com a altura das árvores, do diâmetro à altura do peito e altura de forma combinada com o volume do fuste ou com a biomassa, entre várias outras aplicações.

Resíduo padronizado

Prioriza-se equações em que aproximadamente 95% dos resíduos padronizados se distribuam entre os valores de -2 e 2. Resíduos padronizados são aqueles que foram transformados para a escala de uma distribuição normal padrão. A padronização pode ser obtida por:

$$e_{is} = \frac{y_i - \hat{y}_i}{\sqrt{QME}}$$

Pede-se:

Para o exercício *relação hipsométrica*, construa uma Figura para representar os resíduos padronizados em função das variáveis altura e diâmetro à altura do peito.

20.3.10 INTERVALO DE CONFIANÇA PARA A MÉDIA E INTERVALO DE PREDIÇÃO

O principal uso da regressão é prever o valor de *Y* correspondente a um valor *X* específico, ou seja, dado que um modelo foi ajustado, este é usado para efetuar estimativas e predições da variável *Y*. A estimativa corresponde à aplicação da equação para obter valores médios da variável dependente em função de valores da variável independente que estão presentes no conjunto de dados. Predição corresponde a aplicação da equação para obter valores médios da variável dependente para qualquer valor da variável independente, que pertençam ao domínio da função ou estão próximos dele.

Entretanto, além dos valores estimados, pode ser interessante avaliar um intervalo de confiança para a média ou para um intervalo de predição. Estes intervalos também são denominados de bandas de confiança ou envelopes.

Um intervalo de confiança (IC) estabelece a delimitação dos limites inferior e superior para o parâmetro desconhecido da população. Isto quer dizer que se o processo de amostragem for repetido várias vezes, 95% dos intervalos de confiança conterão a média da população, para um $\alpha = 0{,}05$ de significância.

Um intervalo de predição (IP) estabelece a delimitação dos limites inferior e superior entre os quais uma observação futura existirá, com um certo nível de confiança, efetuado com as observações que já foram mensuradas. Novamente, se o processo de amostragem for repetido várias vezes, 95% dos intervalos de predição conterão o valor de uma nova observação, para um $\alpha = 0{,}05$ de significância.

Os intervalos de predição (IP) se assemelham aos intervalos de confiança (IC), mas seu espectro de dispersão é, por definição, maior que o espectro de dispersão do IC. Isto ocorre porque o intervalo de predição contém duas fontes de incerteza: a média estimada (assim como no intervalo de confiança) e a variação aleatória de novas observações (desvio da média).

Para calcular os intervalos de confiança com 95% de probabilidade para a verdadeira resposta média, tem-se:

$$CI = \hat{\beta}_0 + \hat{\beta}_1 X_0 \pm t_{\frac{\alpha}{2},GLE} \sqrt{QME \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{SS_x}\right]}$$

Os intervalos de confiança de predição podem ser calculadas por:

$$CIP = \hat{\beta}_0 + \hat{\beta}_1 X_0 \pm t_{\frac{\alpha}{2},GLE} \sqrt{QME \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{SS_x}\right]}$$

As bandas de confiança podem ser representadas para o intervalo superior (U) e inferior (L), para um nível de $(1 - \alpha)$ de confiança.

No exemplo *relação hipsométrica*, considere uma árvore com $d = 15{,}7$ cm, que não está presente no conjunto de dados utilizados para o ajuste.

Comandos utilizados para verificar a existência do valor $d = 15{,}7$ cm no conjunto de dados:

Arquivo: 47 - ar - Estatísticas de ajustes da equação

```
X[in.(X.d, Ref([15.7])), :]
0×8 DataFrame
```

Embora o valor de d = 15,7 cm não esteja inicialmente no conjunto de domínio, ele pertence ao espaço do domínio da função. Assim, para estimar o valor da altura aplica-se a equação ajustada, que revelará um valor predito de 18,91 m.

Comandos para obtenção da estimativa da altura de uma árvore com d = 15,7 cm.

```
f(x) =7.4671.+0.7287*x
yhat=f(15.7)
18.91
```

Essa predição é muito útil, mas também é de interesse prático obter os limites de incerteza em torno de *f(15,7)*, com objetivo de avaliar o quão real é a predição realizada, ou seja, obter faixas de incerteza. De acordo com a equação, a altura predita é de 18,91 m e é o valor médio esperado para árvores com diâmetro à altura do peito de 15,7 cm. Por outro lado, 18,91 m é o valor predito para altura de uma árvore em função de um valor arbitrário de diâmetro à altura do peito de 15,7 cm. No cálculo das faixas de incerteza, seus espectros de dispersão nas faixas diferem, dependendo se é para o valor esperado ou para o valor predito.

Para árvores com d = 15,7 cm, pode-se ter 95% de confiança de que a verdadeira altura média está entre 18,47 e 19,35 m. Este intervalo de confiança é estabelecido para um valor médio específico. Às vezes, objetiva-se obter uma estimativa de intervalo para uma resposta individual correspondente a um determinado valor X_i (em vez de uma estimativa para a resposta média). A melhor estimativa de uma resposta individual ainda é $\hat{Y}$, mas o intervalo de confiança é maior porque os valores individuais variam mais que a média, ou seja, é mais difícil prever um valor individual do que uma média. Para a árvore com d = 15,7 cm, tem-se 95% de confiança de que a altura real se encontra entre 16,31 e 21,51 m.

Pede-se:

Para o exercício *relação hipsométrica*, construa uma tabela para representar os valores da variável independente, os valores observados e preditos da variável dependente, intervalos de confiança para a média (CIL – Intervalo de confiança inferior, CIU – Intervalo de confiança superior) e intervalos de confiança de predição (CIPL – Intervalo de confiança de predição inferior, CIPU – Intervalo de confiança de predição superior).

Comandos para obtenção dos intervalos de confiança para a média e predição:

Arquivo: 47 - ar - Estatísticas de ajustes da equação

```
using PyPlot, Distributions

function bandas_rls(x,y,β0,β1,α)
    X0= x
    f(x) = β0+ β1*x
    n=length(x)
    yhat= f.(x)
    SSE= sum((yhat- y).^2)
    MSE=SSE/(n-2)
    SSx=sum((x.- mean(x)).^2)
    xbar= mean(x)
    tα = quantile(TDist(n-2),1- α/2)
```

```
    CIU(X0)= β0+ β1*X0+ tα*sqrt(MSE*((1/n)+(((X0- xbar)^2)/SSx)))
    CIL(X0)= β0+ β1*X0- tα*sqrt(MSE*((1/n)+(((X0- xbar)^2)/SSx)))
    CIPU(X0)= β0+ β1*X0+ tα*sqrt(MSE*(1+(1/n)+(((X0- xbar)^2)/SSx)))
    CIPL(X0)= β0+ β1*X0- tα*sqrt(MSE*(1+(1/n)+(((X0- xbar)^2)/SSx)))
    df=DataFrame(X = x, Y = y, ValorEstimado=yhat, CIL=CIL.(X0), CIU=CIU.(X0), CIPL=CIPL.
(X0),CIPU=CIPU.(X0))
end
```

```
julia> bandas_rls(X.d,X.h,7.4671, 0.7287,0.05)
62×7 DataFrame
```

Row	X	Y	ValorEstimado	CIL	CIU	CIPL	CIPU
	Float64	Float64	Float64	Float64	Float64	Float64	Float64
1	19.74	20.2	21.8516	21.0894	22.6139	19.1797	24.5236
2	15.12	19.1	18.485	18.0804	18.8897	15.8924	21.0777
3	12.89	17.3	16.86	16.5339	17.1862	14.2785	19.4416
4	11.62	16.9	15.9346	15.5946	16.2746	13.3512	18.518
5	16.55	20.5	19.5271	19.0272	20.0269	16.9179	22.1363
⋮							
57	10.03	15.1	14.776	14.3635	15.1884	12.1821	17.3699
58	12.73	17.6	16.7435	16.4181	17.0688	14.162	19.3249
59	14.64	17.7	18.1353	17.7565	18.514	15.5465	20.724
60	12.73	18.2	16.7435	16.4181	17.0688	14.162	19.3249
61	7.32	11.7	12.8012	12.1906	13.4118	10.1685	15.4339
62	18.3	19.5	20.8023	20.1637	21.4409	18.163	23.4416

Adicionalmente, construa uma Figura para representar os valores de altura observados e estimados, intervalos de confiança para os valores esperados e intervalos de confiança para os valores preditos.

Comandos para obtenção do gráfico dos intervalos de confiança:

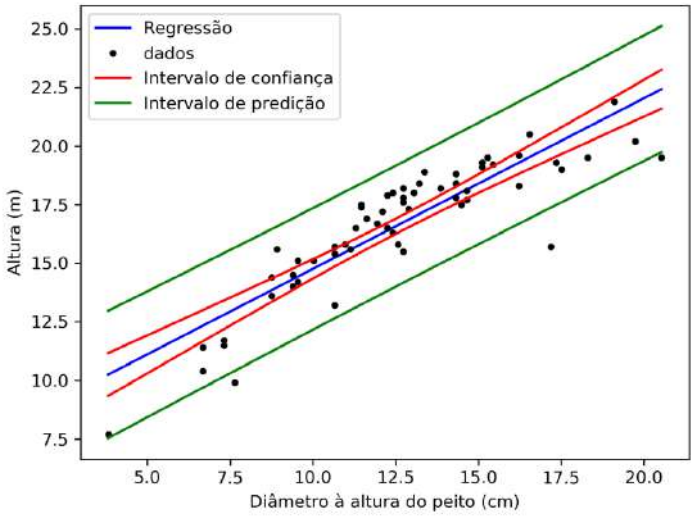
Arquivo: 47 - ar - Estatísticas de ajustes da equação

```
#Gráfico dos intervalos de confiança
using PyPlot, Distributions

function Gbandas_rls(x,y,β0,β1,α)
    X0= minimum(x):0.001: maximum(x)
    f(x) = β0+ β1*x
    plot(X0,f.(X0),"b", label="Regressão")
    xlabel("Diâmetro à altura do peito (cm)")
    ylabel("Altura (m)")
    plot(x,y,".k",label="dados")
    n=length(x)
    yhat= f.(x)
    SSE= sum((yhat- y).^2)
    MSE=SSE/(n-2)
    SSx=sum((x.- mean(x)).^2)
    xbar= mean(x)
    tα = quantile(TDist(n-2),1- α/2)
    CIU(X0)= β0+ β1*X0+ tα*sqrt(MSE*((1/n)+(((X0- xbar)^2)/SSx)))
    CIL(X0)= β0+ β1*X0- tα*sqrt(MSE*((1/n)+(((X0- xbar)^2)/SSx)))
    CIPU(X0)= β0+ β1*X0+ tα*sqrt(MSE*(1+(1/n)+(((X0- xbar)^2)/SSx)))
    CIPL(X0)= β0+ β1*X0- tα*sqrt(MSE*(1+(1/n)+(((X0- xbar)^2)/SSx)))
```

```
    plot(X0, CIU.(X0) ,"r",label="Intervalo de confiança")
    plot(X0, CIL.(X0) ,"r")
    plot(X0, CIPU.(X0) ,"g",label="Intervalo de predição")
    plot(X0, CIPL.(X0) ,"g")
    legend(loc="upper left")
end

#Gbandas_rls(x,y,β0,β1,α)
Gbandas_rls(X.d,X.h,7.4671, 0.7287,0.05)
```



20.4 REGRESSÃO PELA ORIGEM

Os modelos estatísticos tratados até aqui têm um intercepto. Em algumas situações práticas, a teoria envolve modelos apenas com o coeficiente angular. Desta forma, o intercepto é igual a zero e a regressão tem seu ponto inicial na origem.

A regressão que passa pela origem é representada pelo modelo: $Y = \beta X + \varepsilon$.

O coeficiente β pode ser estimado por:

$$\beta = \frac{\sum XY}{\sum X^2}$$

EXERCÍCIO 21: ÁREA FOLIAR E PESO DAS FOLHAS

Em alguns modelos aplicados na Engenharia Florestal, como por exemplo, para previsão de crescimento e produção, a área foliar é uma variável importante. Entretanto, mensurá-la é uma atividade onerosa e particularmente difícil. Uma das maneiras para contornar esse problema é tomar uma amostra da massa de folhas e seus respectivos valores de área foliar. Logo, conhecendo-se o peso das folhas (que é mais fácil de medir do que a área foliar), pode-se estimar a área foliar. Observe, que intuitivamente, a regressão da área foliar em função do peso das folhas deve passar pela origem, pois para uma área foliar igual a zero não existe peso de folhas correspondente.

Um Engenheiro Florestal efetuou uma mensuração da área foliar (cm²) e de sua massa seca (g) para uma amostra de tamanho 144. Esta amostra foi obtida em florestas de acácia-negra com idade variando de 1 a 7 anos.

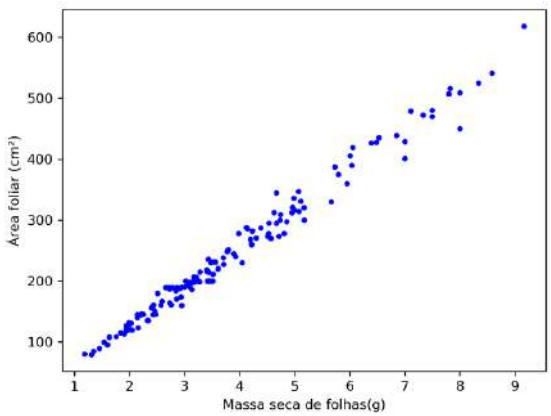
Comandos para construir um gráfico dos valores:

ARQUIVO: 48 - AR - REGRESSÃO PELA ORIGEM

```
>using DataFrames, CSV, PyPlot
>X=CSV.read("Z:/Alexandre/Julia/Exemplos/af.csv", DataFrame)
```

```
julia> X=CSV.read("Z:/Alexandre/Julia/Exemplos/af.csv", DataFrame)
144×2 DataFrame
| Row | Massa   | AreaFoliar |
|     | Float64 | Float64    |
|-----|---------|------------|
| 1   | 1.1809  | 80.0       |
| 2   | 1.9408  | 123.0      |
| 3   | 1.9854  | 132.0      |
| 4   | 2.0251  | 131.0      |
| 5   | 2.1408  | 140.0      |
⋮
| 139 | 4.2174  | 260.0      |
| 140 | 4.7142  | 274.0      |
| 141 | 3.5102  | 200.0      |
| 142 | 1.4497  | 89.0       |
| 143 | 2.8574  | 189.0      |
| 144 | 2.9496  | 160.0      |
```

```
plot(X.Massa,X.AreaFoliar,"b.",label="dados")
xlabel("Massa seca de folhas(g)")
ylabel("Área foliar (cm²)")
```



Comandos para ajuste da equação que passa pela origem.

```
>function rlso(x,y)
    Σx = sum(x)
    Σxy = sum(x.*y)
    Σx² = sum(x.^2)
    β= Σxy/ Σx²
end

#rlso(x,y)
rlso(X.Massa,X.AreaFoliar)
63.6461
```

A regressão ajustada foi $\hat{y}_i = 63{,}6461\ x_i$.
Observe que a soma dos desvios não equivale a zero.

Comandos para obtenção da soma dos desvios:

Arquivo: 48 - ar - Regressão pela origem

```
f(x) = 63.6461*x
SE=sum(X.AreaFoliar.-f.(X.Massa))
 -54.7809
```

Para o modelo ajustado, as somas dos quadrados devido à regressão, aos resíduos e total podem ser obtidas por:

$$SQR = \frac{(\sum xy)^2}{\sum x^2}$$

$$SQE = \sum (y_i - \hat{y}_i)^2$$

$$SQT = \sum y^2$$

Uma vez que os ajustes não foram efetuados em função da média, a soma de quadrados para o total tem *n* graus de liberdade e a soma de quadrados dos resíduos (n - 1) graus de liberdade.

Para este modelo, o coeficiente de determinação pode ser calculado como:

$$R^2 = \frac{(\sum xy)^2}{\sum x^2 \sum y^2}$$

As demais estatísticas de interesse (erro-padrão, coeficiente de variação e F) são calculadas de maneira análoga à que foi usada na regressão com intercepto.

Para o exercício *área foliar*, faça um resumo das estatísticas *QME, R^2, $\overline{R}^2$, S_{YX}, CV* e *F*.

Comandos para obtenção das estatísticas do modelo que passa pela origem:

Arquivo: 48 - ar - Regressão pela origem

```
>using Statistics

>function Resumo_rlso(x,y)
    Σx = sum(x)
    Σxy = sum(x.*y)
    Σx² = sum(x.^2)
    β= Σxy/ Σx²
    f(x)= β*x
    yhat=f.(x)
    n=length(x)
    GLT=n
    GLE= n-1
    GLR=GLT-GLE
    SQR=((Σxy)^2)/Σx²
    SQE= sum((yhat- y).^2)
    SQT = sum(y.^2)
    QMR=SQR/GLR
    QME=SQE/GLE
    Fc=QMR/QME
    SYX=sqrt(QME)
    CV= SYX/mean(y)*100
    Σy² = sum(y.^2)
    R²=(Σxy^2)/(Σx²*Σy²)
    println("QME: ", QME)
    println("Fc: ", Fc)
    println("SYX: ", SYX)
    println("CV (%): ", CV)
    println("R²: ", R²)
end

#Resumo_rlso(x,y)
Resumo_rlso(X.Massa,X.AreaFoliar)
QME: 214.3826
Fc: 48771.6234
SYX: 14.6418
CV (%): 5.98
R²: 0.9971
```

Se for necessária a realização de teste de hipóteses sobre o valor de $\hat{\beta}$ *($\beta = 0$)*, um teste *t* ou F pode ser aplicado. Neste caso, $F_c = QMR/QME$ e $t_c = \sqrt{F_c}$ com $(n-1)$ graus de liberdade.

Comandos para testar a significância de $\hat{\beta}$:

Arquivo: 48 - ar - Regressão pela origem

```
>using Distributions

Fc= 48771.62339989079
gl=length(X.Massa)-1
tc=sqrt(Fc)
αβ= 2*ccdf(TDist(gl), abs(tc))
4.323e-183
```

Os resultados indicam que o coeficiente angular $\hat{\beta}$ foi significante ao nível de 0,01 de significância.

20.5 AJUSTE DE UMA REGRESSÃO LINEAR SIMPLES POR MATRIZES

A modelagem de regressão utilizando o método matricial é mais fácil para a obtenção dos cálculos por computador, principalmente quando o modelo compreende mais de duas variáveis independentes (regressão múltipla).

Considere o modelo estatístico para uma equação de regressão linear simples:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Em que:
β_0 e β_1 descrevem o intercepto e a inclinação, respectivamente.
ε é o termo de erro
i corresponde ao i-ésimo elemento da amostra, $i = 1,...,n$.

Este modelo pode ser reescrito na forma matricial como:

$$Y = X\beta + \varepsilon$$

Alternativamente,

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Em que:
Y é o vetor das observações da variável dependente
X é a matriz com os valores da variável independente. Os valores de 1 são adicionados para se obter a estimativa de β_0. Quando a regressão passa pela origem, a coluna com os valores um não é necessária

β é o vetor dos parâmetros da regressão a serem estimados.
ε é o vetor dos erros aleatórios, que é desconhecido.

Os valores do vetor β podem ser estimados por meio do método de mínimos quadrados, como segue:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Em que:

$$X'X = \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum {X_i}^2 \end{bmatrix}$$

$$(X'X)^{-1} = \frac{1}{n\sum {X_i}^2 - (\sum X_i)^2}\begin{bmatrix} \sum {X_i}^2 & -\sum X_i \\ -\sum X_i & n \end{bmatrix}$$

$$X'Y = \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix}$$

Se a matriz X' X é não singular (o resultado de $n\sum X_i^2 - (\sum X_i)^2$ deve ser diferente de zero), então existe sua inversa, e o valor do vetor β pode ser estimado como:

$$\hat{\beta} = \begin{bmatrix} \bar{Y} - \hat{\beta}_1 \bar{X} \\ \dfrac{\sum X_i Y_i - \dfrac{\sum X_i \sum Y_i}{n}}{\sum {X_i}^2 - \dfrac{(\sum X_i)^2}{n}} \end{bmatrix} \quad \text{ou} \quad \hat{\beta} = (X'X)^{-1}X'Y$$

EXERCÍCIO 22: RELAÇÃO HIPSOMÉTRICA

Para os dados do exercício *relação hipsométrica*, ajuste o modelo: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, em que a variável dependente é a altura e a variável independente o diâmetro à altura do peito. Estes dados estão inclusos no vetor e matrizes abaixo:

$$Y = \begin{bmatrix} 20{,}2 \\ 19{,}1 \\ \vdots \\ 19{,}5 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 19{,}74 \\ 1 & 15{,}12 \\ \vdots & \vdots \\ 1 & 18{,}30 \end{bmatrix} \beta = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

Comandos para o ajuste de uma equação de regressão linear simples pelo método matricial:

ARQUIVO: 49 - AR - REGRESSÃO LINEAR SIMPLES POR MATRIZES

```
>using DataFrames, CSV
>X=CSV.read("Z:/Alexandre/Julia/Exemplos/descritiva.csv", DataFrame);
```

```
function mrls(x,y)
    X = [ones(size(x,1)) x]
    β =inv(X'X)*X'*y
end

#mrls(x,y)
mrls(X.d,X.h)
7.4671
0.7287
```

O valor do coeficiente β_0 foi de 7,4671 e β_1 de 0,7287. A equação ajustada pode ser representada como: $\hat{y} = 7{,}4671 + 0{,}7287\, x$ ou $\hat{h} = 7{,}4671 + 0{,}7287\, d$.

De maneira análoga às seções anteriores, as somas de quadrados também podem ser obtidas por meio de operações matriciais. Assim, as somas de quadrados devidos à variação total (SQT), regressão (SQR) e resíduos (SQE) podem ser obtidas como:

$$SQR = \hat{\beta} X'Y - \frac{(\sum Y)^2}{n}, \text{ com } n_\beta - 1 \text{ graus de liberdade}$$

$$SQE = Y'Y - \hat{\beta} X'Y, \text{ com } n - 2 \text{ graus de liberdade}$$

$$SQT = Y'Y - \frac{(\sum y)^2}{n}, \text{ com } n - 1 \text{ graus de liberdade}$$

Em que:
n_β é o número de coeficientes.

Como para o método algébrico, às hipóteses $H_0: \beta_1 = 0$ e $H_0: \beta_0 = 0$, além de estabelecer um intervalo de confiança para o verdadeiro valor de β_i, também podem ser testadas por meio de operações matriciais. Para tanto, uma matriz de variância-covariância (variância dos coeficientes corresponde aos valores na diagonal da matriz e covariância dos coeficientes fora da diagonal da matriz), como segue:

$$V(\hat{\beta}) = \begin{bmatrix} \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum (X - \bar{X})^2} s^2_{Y\,X}\right) & \frac{-\bar{X} s^2_{Y\,X}}{\sum (X - \bar{X})^2} \\ \frac{-\bar{X} s^2_{Y\,X}}{\sum (X - \bar{X})^2} & \frac{s^2_{Y\,X}}{\sum (X - \bar{X})^2} \end{bmatrix} = (X'X)^{-1} s^2_{Y\,X}$$

Em que s^2_{YX} é a variância da regressão (QME).

O intervalo de confiança para a média matricialmente, pode ser calculado como:

$$IC\left[\hat{\mu}_y \pm t_{\frac{\alpha}{2}, GLE} s_{\hat{Y}}\right] = 1 - \alpha$$

Em que:

$s_{\hat{Y}} = \sqrt{s_{\hat{Y}}^2}$

$s_{\hat{Y}}^2 = X_0' V(\hat{\beta}) X_0$, variância da estimativa

$s_{\hat{Y}_{no}}^2 = [1 + X_0'(X'X)^{-1}X_0]s_{\hat{Y}\,X}^2$, variância da predição para uma nova observação

Pede-se:

Para o exercício relação hipsométrica, faça um resumo das estatísticas *QME, R², $\overline{R}^2$, S_{YX}, CV* e *F*, teste os coeficientes, obtenha os intervalos de confiança para os coeficientes e os intervalos de confiança para a média e para a predição. Considere as expressões anteriores para obtê-los matricialmente. Adicionalmente, apresente os gráficos com os intervalos de confiança e um gráfico de resíduos.

Comandos para ajuste e teste de uma equação de regressão linear simples por matrizes:

Arquivo: 50 - Função resumo_rlsm para regressão linear simples por matrizes

```
>using Statistics, Distributions, PyPlot

>function Resumo_rlsm(x,y,α)
    X = [ones(size(x,1)) x]
    β =inv(X'X)*X'*y
    yhat = X*β
    n=length(x)
    nβ=length(β)
    GLR= nβ-1
    GLT=n-1
    GLE= GLT- GLR
    Σy = sum(y)
    SQR= β'*X'*y-(Σy^2)/n
    SQE= y'*y-β'*X'*y
    SQT = y'*y-(Σy^2)/n
    QMR=SQR/GLR
    QME=SQE/GLE
    Fc=QMR/QME
    SYX=sqrt(QME)
    CV= SYX/mean(y)*100
    R²=1-(SQE/SQT)
    R²aj=1-(QME/(SQT/GLT))
    AlphaRegressao= ccdf(FDist(GLR, GLE),abs(Fc))
    AINV=inv(X'*X)
    Sβ0=sqrt(AINV[1,1]*QME)
    Sβ1=sqrt(AINV[2,2]* QME)
    t0 = β[1,1]/Sβ0
```

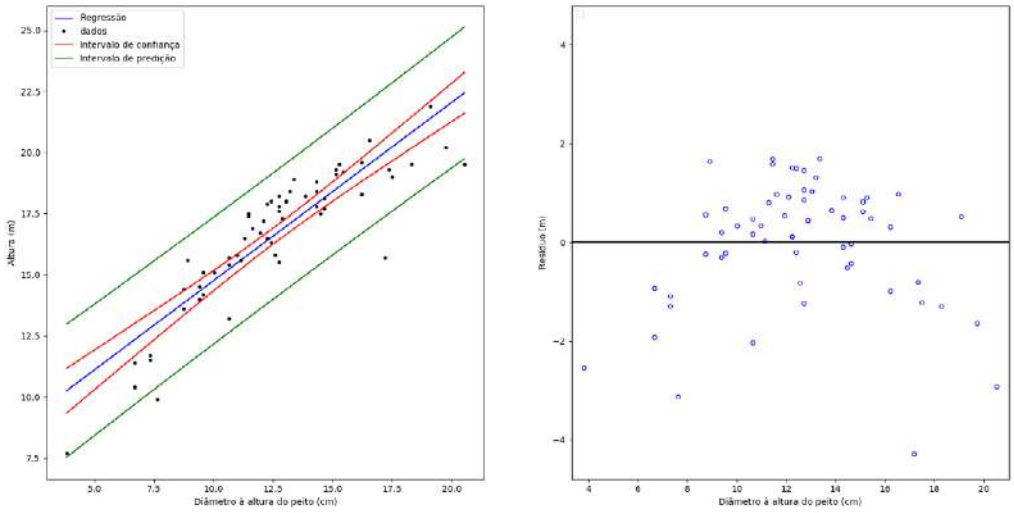
```
t1 = β[2,1]/Sβ1
αβ0= 2*ccdf(TDist(GLE), abs(t0))
αβ1= 2*ccdf(TDist(GLE), abs(t1))
tα = quantile(TDist(n-2),1- α/2)
ICInferior_β0 = β[1,1] - tα * Sβ0
ICSuperior_β0 = β[1,1] + tα * Sβ0
ICInferior_β1 = β[2,1] - tα * Sβ1
ICSuperior_β1 = β[2,1] + tα * Sβ1
x0= minimum(x):0.001: maximum(x)
nx0=length(x0)
X0 = [ones(size(x0,1)) x0]
yhatX0 = X0*β
tα = quantile(TDist(n-2),1- α/2)
Vβ= AINV*QME
Syhat_X0= sqrt.([X0[i,:]'*Vβ*X0[i,:] for i in 1: nx0])
CIL= yhatX0.-( tα.* Syhat_X0)
CIU= yhatX0.+( tα.* Syhat_X0)
Syhatno= sqrt.([(1+X0[i,:]'*AINV*X0[i,:])*QME for i in 1: nx0])
CIPL= yhatX0.-( tα.* Syhatno)
CIPU= yhatX0.+( tα.* Syhatno)
subplot(121)
plot(x0, yhatX0,"b", label="Regressão")
xlabel("Diâmetro à altura do peito (cm)")
ylabel("Altura (m)")
plot(x,y,".k",label="dados")
plot(x0, CIL,"r",label="Intervalo de confiança")
plot(x0, CIU,"r")
plot(x0, CIPU,"g",label="Intervalo de predição")
plot(x0, CIPL,"g")
legend(loc="upper left")
subplot(122)
ei= y.-yhat
scatter(x, ei, facecolor= "none", edgecolors="blue", s=20)
xlabel("Diâmetro à altura do peito (cm)")
ylabel("Resíduo (m)")
ylim(-maximum(abs.(ei))-0.5, maximum(abs.(ei))+0.5)
xlim(minimum(x)-0.5, maximum(x)+0.5)
plot([0, maximum(x)+0.5],[0,0],c=:black)
legend(loc="upper left")
println("Coeficientes_estimados: ", β)
println("GLR: ", GLR)
println("GLE: ", GLE)
println("GLT: ", GLT)
println("SQR: ", SQR)
println("SQE: ", SQE)
println("SQT: ", SQT)
println("QMR: ", QMR)
println("QME: ", QME)
println("Fc: ", Fc)
```

```
    println("αF: ", AlphaRegressao)
    println("SYX: ", SYX)
    println("CV (%): ", CV)
    println("R²: ", R²)
    println("R²aj: ", R²aj)
    println("Erro-padrão_β0: ", Sβ0)
    println("Erro-padrão_ β1: ", Sβ1)
    println("t0: ", t0)
    println("αβ0: ", αβ0)
    println("t1: ", t1)
    println("αβ1: ", αβ1)
    println("ICInferior_β0: ", ICInferior_β0)
    println("ICSuperior_β0: ", ICSuperior_β0)
    println("ICInferior_ β1: ", ICInferior_β1)
    println("ICSuperior_ β1: ", ICSuperior_β1)
end

>using DataFrames, CSV

>X=CSV.read("Z:/Alexandre/Julia/Exemplos/descritiva.csv", DataFrame)

#Resumo_rlsm(x,y,α)
Resumo_rlsm(X.d,X.h,0.05)
GLR: 1
GLE: 60
GLT: 61
SQR: 369.2394
SQE: 98.3433
SQT: 467.5827
QMR: 369.2394
QME: 1.6391
Fc: 225.27581
αF: 5.5856e-22
SYX: 1.2803
CV (%): 7.68
R²: 0.7897
R²aj: 0.7862
Erro-padrão_β0: 0.6349
Erro-padrão_ β1: 0.04856
t0: 11.7607
αβ0: 3.2376e-17
t1: 15.0092
αβ1: 5.5856e-22
ICInferior_β0: 6.1971
ICSuperior_β0: 8.7372
ICInferior_ β1: 0.6316
ICSuperior_ β1: 0.8258
```



Pequenas diferenças entre o método algébrico e o método matricial, devem-se ao arredondamento dos números.

20.6 REGRESSÃO LINEAR MÚLTIPLA

A regressão linear simples envolve a relação entre duas variáveis, uma dependente e outra independente. A Regressão linear múltipla refere-se à relação entre a variável dependente e duas ou mais variáveis independentes. O modelo para a regressão linear múltipla pode ser expresso como:

$$Y_i = \beta_0 + \beta_1 X1_i + \beta_2 X2_i + \ldots + \beta_k \; XK_i + \varepsilon_i$$

Em que:
K refere-se ao número de variáveis independentes e i ao i-ésimo elemento da amostra.

Este modelo pode ser reescrito na forma matricial como:

$$Y = X\beta + \varepsilon$$

Alternativamente,

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

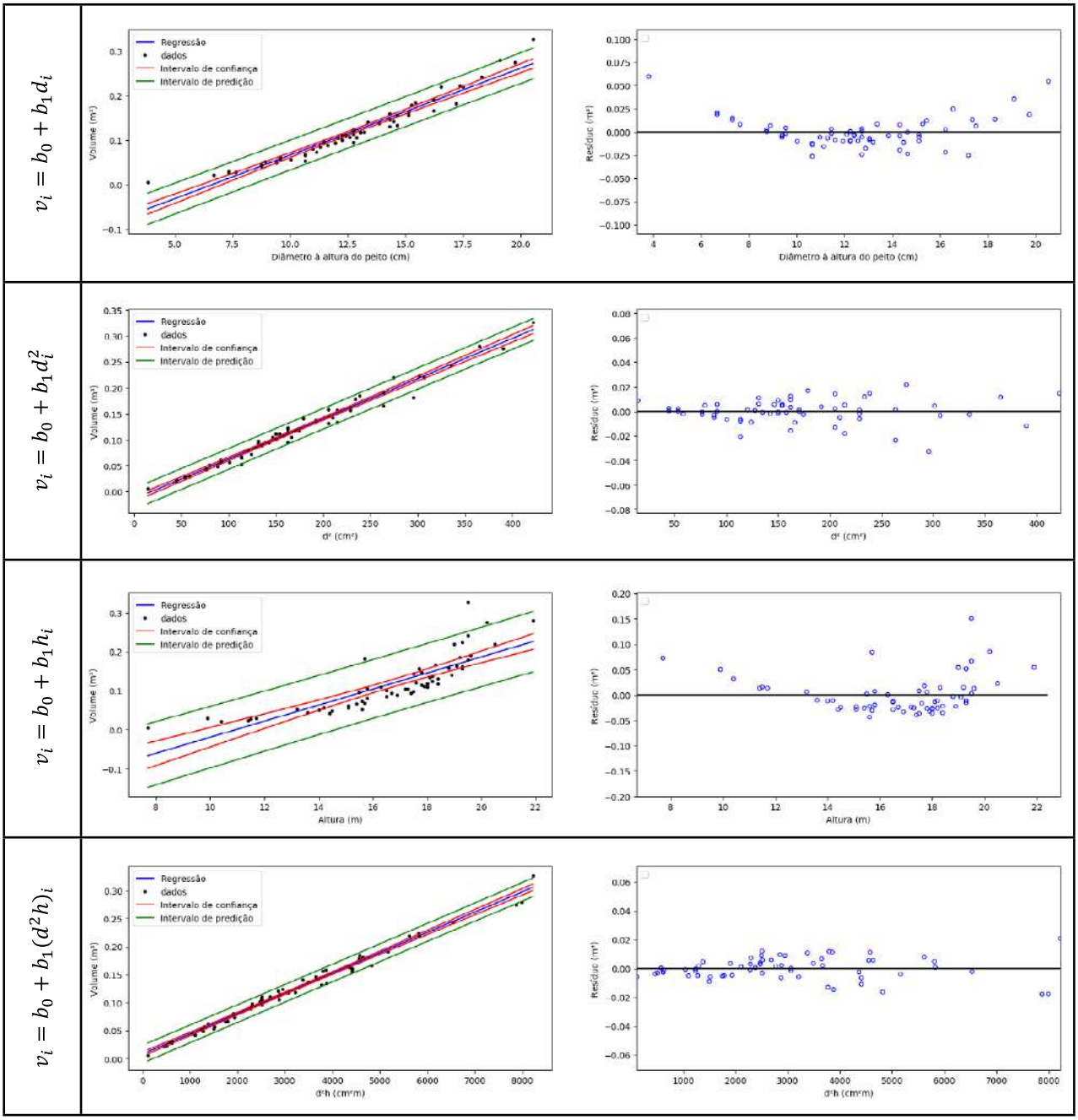
Os parâmetros podem ser estimados por meio do método de mínimos quadrados. O procedimento para estimativa dos coeficientes e das somas dos quadrados devido à regressão, resíduos e total são os mesmos aplicados à regressão linear simples.

EXERCÍCIO 23: VOLUME DO FUSTE

Considere o exemplo em que um Engenheiro Florestal está interessado em obter estimativas médias de volume em função das variáveis diâmetro à altura do peito (*d*) e altura (*h*) de árvores. A variável dependente é o volume do fuste das árvores (em m³), que depende do diâmetro à altura do peito (cm) e da altura (em m) das árvores, sendo estas variáveis independentes de fácil medição. No exemplo, em uma floresta de acácia-negra foi obtida uma amostra aleatória com 62 valores pareados destas variáveis, apresentados no arquivo *descritiva*.

Utilizando a função *Resumo_rlsm* foram ajustados os modelos apresentados na Figura 20.1. Os gráficos de valores preditos, intervalo de confiança para a média, intervalo de confiança para a predição e resíduos (em função da variável dependente) também estão apresentados na Figura 20.1.

FIGURA 20.1. Modelos volumétricos ajustados.



Na análise gráfica, observa-se que a equação ajustada para o modelo $v_i = b_0 + b_1 h_i$ foi a de pior resultado, revelando valores estimados de volume negativos – o que não é realista. A equação ajustada para o modelo $v_i = b_0 + b_1 d_i$ também não foi apropriada, uma vez que os resíduos têm o mesmo sinal (positivo) para valores de diâmetros pequenos e grandes, indicando uma relação não-linear entre as variáveis.

As equações ajustadas para os modelos $v_i = b_0 + b_1 d_i^2$ e $v_i = b_0 + b_1 (d^2 h)_i$ foram as que revelaram melhor os resultados para obtenção de estimativas de volume. Por meio das estatísticas para avaliação da qualidade do ajuste, observa-se que o modelo $v_i = b_0 + b_1 (d^2 h_i)$ revelou melhor os resultados para os dados amostrais, conforme apresentados na Tabela 20.1.

TABELA 20.1. Estatísticas para avaliação da qualidade do ajuste.

Estatística	$v_i = b_0 + b_1 d_i^2$	$v_i = b_0 + b_1 (d^2 h)_i$
MSE	0,000098	0,0000605
Fc	2825,453	4594,529
SXY	0,009882	0,007781
CV (%)	8,378886	6,597135
R^2	0,979206	0,987109
R^2aj	0,978859	0,986894
αβ0	2,58E-06	0,000301
αβ1	3,58E-52	2,10E-58

Embora o modelo com a variável independente altura não tenha sido apropriado, ao combinar esta variável com o diâmetro os resultados para as estimativas e predições foram melhores. Observe que são duas variáveis que foram combinadas em uma variável independente ($X1 = d^2 h$). A variável combinada também pode ser denominada de termo de produto ou interação. O termo de interação permite que a resposta de d^2 dependa do nível de h.

Uma análise adicional, com objetivo de melhorar ainda mais as estimativas e predições, seria testar o seguinte modelo: $v_i = b_0 + b_1 d_i^2 + b_2 h_i$. As variáveis são as mesmas, entretanto, agora o modelo linear é múltiplo e seu ajuste é realizado por meio de uma regressão linear múltipla.

O modelo $v_i = b_0 + b_1 d_i^2 + b_2 h_i + e_i$ também pode ser escrito como:

$$Y_i = \beta_0 + \beta_1 X1_i + \beta_2 X2_i + \varepsilon_i$$

Em que:
Y é o volume do fuste, em m^3
X1 é o quadrado do diâmetro à altura do peito (d^2, em cm^2)
X2 é a altura total, em m

Comandos para o ajuste de um modelo de regressão linear múltiplo, pelo método matricial:

ARQUIVO: 51 - FUNÇÃO RESUMO_RLM PARA REGRESSÃO LINEAR MÚLTIPLA

```
>using Statistics, Distributions,  LinearAlgebra
>function Resumo_rlm(x,y,α)
    X = [ones(size(x,1)) x]
    β =inv(X'X)*X'*y
    yhat = X*β
```

```
        n=size(x,1)
        nβ=length(β)
        GLR= nβ-1
        GLT=n-1
        GLE= GLT- GLR
        Σy = sum(y)
        SQR= β'*X'*y-(Σy^2)/n
        SQE= y'*y-β'*X'*y
        SQT = y'*y-(Σy^2)/n
        QMR=SQR/GLR
        QME=SQE/GLE
        Fc=QMR/QME
        SYX=sqrt(QME)
        CV= SYX/mean(y)*100
        R²=1-(SQE/SQT)
        R²aj=1-(QME/(SQT/GLT))
        AlphaRegressao= ccdf(FDist(GLR, GLE),abs(Fc))
        AINV=inv(X'*X)
        Vβ= AINV*QME
        tβ=β./sqrt(Diagonal(Vβ))
        αβ= 2*ccdf.(TDist(GLE), abs.(tβ))
        println("Coeficientes_estimados: ", β)
        println("GLR: ", GLR)
        println("GLE: ", GLE)
        println("GLT: ", GLT)
        println("SQR: ", SQR)
        println("SQE: ", SQE)
        println("SQT: ", SQT)
        println("QMR: ", QMR)
        println("QME: ", QME)
        println("Fc: ", Fc)
        println("αF: ", AlphaRegressao)
        println("SYX: ", SYX)
        println("CV (%): ", CV)
        println("R²: ", R²)
        println("R²aj: ", R²aj)
        println("Vβ: ", Vβ)
        println("tβ: ", tβ)
        println("αβ: ", αβ)
    end

    using DataFrames, CSV
    X=CSV.read("Z:/Alexandre/Julia/Exemplos/descritiva.csv", DataFrame)

    #Resumo_rlm(x,y,α)
    Resumo_rlm([(X.d.^2) X.h],X.v,0.05)
    Coeficientes_estimados:          [-0.03547501531190221,          0.0007324188873813576,
0.0016955455549279145]
```

```
GLR: 2
GLE: 59
GLT: 61
SQR: 0.27801117620278293
SQE: 0.005556565732701024
SQT: 0.28356774193548395
QMR: 0.13900558810139146
QME: 9.41790802152716e-5
Fc: 1475.970967051885
αF: 4.1534991537481714e-51
SYX: 0.009704590677368705
CV (%): 8.219735273181145
R²: 0.980404803117679
R²aj: 0.9797405591555663
Vβ:
  9.87008e-5   1.59866e-7   -7.4658e-6
  1.59866e-7   6.21081e-10  -1.59532e-8
 -7.4658e-6   -1.59532e-8    6.11196e-7
tβ:
-3.57077    -Inf      -Inf
 Inf        29.389     Inf
 Inf        Inf        2.1688
αβ:
 0.000715747  0.0         0.0
 0.0          6.2975e-37  0.0
 0.0          0.0         0.0341405
```

A equação ajustada foi $v_i = -0,03547 + 0,0007324\ d_i^2 + 0,00169\ h_i$. O valor de F na análise de variância foi significante, indicando que o modelo ajustado foi adequado. O teste t indicou que todos os coeficientes foram significantes. As estatísticas de ajuste: coeficiente de determinação ajustado, erro padrão e coeficiente de determinação revelaram valores confiáveis. Entretanto, ainda assim esta equação foi inferior quando comparada com a equação ajustada para o modelo $v_i = b_0 + b_1\ d^2\ h_i$.

20.7 AJUSTE DE MODELOS LINEARIZÁVEIS

Considere o modelo $y_i = \beta_0\ \ X1_i^{\beta_1}\ e_i$. O expoente da variável X_1 é um coeficiente (β_1), indicando que este não é um modelo linear. Entretanto, este modelo pode ser linearizado para se proceder o ajuste da equação por meio do método de mínimos quadrados. Assim, o modelo pode ser reescrito como $ln\ y_i = ln\ \beta_0 + \beta_1\ lnX1_i + ln\ e_i$.

Considere os dados do exercício de volume e ajuste o modelo $ln\ y_i = ln\ \beta_0 + \beta_1\ lnX1_i + ln\ e_i$, em que y é o volume do fuste e x o diâmetro à altura do peito.

Comandos para o ajuste de modelos linearizáveis:

Arquivo: 52 - Ajuste de modelos linearizáveis

```
>using DataFrames, CSV
>X=CSV.read("Z:/Alexandre/Julia/Exemplos/descritiva.csv", DataFrame);

x=log.(X.d)
y=log.(X.v)

>using Statistics, Distributions, LinearAlgebra
Resumo_rlm(x,y,0.05)
Coeficientes_estimados: [-8.05991, 2.29350]
QME: 0.011639
```

A equação ajustada foi $ln(\hat{v}_i) = -8{,}05991 + 2{,}29350\ ln(d_i)$, a qual estima valores médios de volume em m³, em escala logarítmica. A transformação utilizada, para linearizar o modelo, foi o logaritmo natural, então o valor do volume estimado em sua escala original (m³) pode ser obtido como $\hat{v}_i = e^{[-8{,}05991 + 2{,}29350 \ln(d_i)]}$.

Note que o termo que foi minimizado é a soma dos quadrados totais $\sum(Y - \hat{Y})^2 = mínimo$. Neste caso, Y e $\hat{Y}$ referem-se a $ln\ v_i$ e $ln\ \hat{v}_i$. Portanto, para que as estatísticas QME, coeficiente de determinação ajustado e erro padrão sejam comparáveis com os modelos testados anteriormente, elas devem ser recalculadas para as suas unidades originais.

Devido à transformação utilizada, uma correção nas estimativas da variável dependente (discrepância logarítmica) pode ser aplicada, como segue:

$$\hat{y}_i = e^{\left(\hat{\beta}_0' + \hat{\beta}_1 ln X_1 + \frac{QME}{2}\right)}$$

Em que:
e é o valor da constante de Euler.
QME/2 é o fator de correção devido à discrepância logarítmica.

Um gráfico dessa equação pode ser obtido. Nele observa-se um desvio sistemático devido à equação linearizada estimar os valores médios geométricos.

Comandos para obtenção do gráfico de uma equação linearizada:

Arquivo: 52 - Ajuste de modelos linearizáveis

```
>using DataFrames, CSV
>X=CSV.read("Z:/Alexandre/Julia/Exemplos/descritiva.csv", DataFrame)

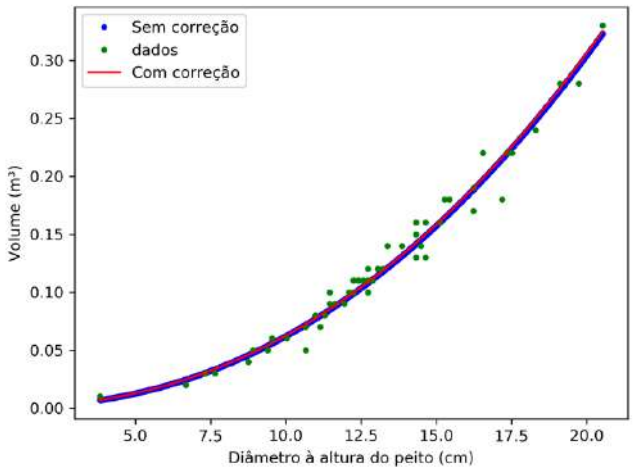
x=log.(X.d)
y=log.(X.v)

Resumo_rlm(x,y,0.05) #Conforme arquivo 51.

>using PyPlot
f(x) =-8.05991+2.29350*x
x = minimum(X.d):0.01: maximum(X.d)
```

```
logx=log.(x)
yhat=exp.(f.(logx))
plot(x,yhat,"b.",label="Sem correção")
xlabel("Diâmetro à altura do peito (cm)")
ylabel("Volume (m³)")
plot(X.d,X.v,"g.",label="dados")

#Correção
f(x) =-8.05991+2.29350*x+(0.011639/2)
x = minimum(X.d):0.01: maximum(X.d)
logx=log.(x)
yhatC=exp.(f.(logx))
plot(x,yhatC,"r",label="Com correção")
legend(loc="upper left")
```



A correlação do volume com o diâmetro à altura do peito é forte, logo, o valor do coeficiente de determinação é alto, enquanto o valor de QME é baixo, devido a uma diferença sistemática entre os valores corrigidos e não corrigidos, ou seja, uma discrepância logarítmica, que também é pequena.

Considere um novo conjunto de dados amostrais de biomassa de árvores nativas, apresentados no arquivo *BiomassaNativas*. Este conjunto refere-se a uma amostra de 100 pares de valores de biomassa de galhos vivos e diâmetro à altura do peito, mensurados em uma floresta da Mata Atlântica.

Pede-se:

Ajuste o modelo $ln\ y_i = ln\ \beta_0 + \beta_1\ lnXI_i + ln\ e_i$, em que y é a biomassa de galhos vivos e x é o diâmetro à altura do peito. Construa um gráfico com os valores observados, preditos e preditos com correção da discrepância logarítmica. O ajuste do modelo pode ser efetuado utilizando a função *Resumo_rlm*.

Comandos para ajuste de um modelo de regressão linear linearizado:

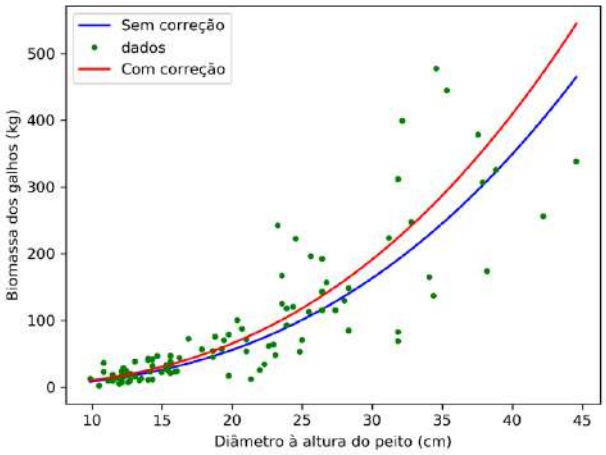
Arquivo: 52 - Ajuste de modelos linearizáveis

```
>using DataFrames, CSV
>X=CSV.read("Z:/Alexandre/Julia/Exemplos/BioNat.csv", DataFrame)
```

```
julia> X=CSV.read("Z:/Alexandre/Julia/Exemplos/BioNat.csv", DataFrame)
99×2 DataFrame
| Row | BGalhos | d       |
|     | Float64 | Float64 |
|-----|---------|---------|
| 1   | 120.45  | 24.35   |
| 2   | 44.31   | 16.23   |
| 3   | 8.03    | 12.1    |
| 4   | 53.44   | 24.83   |
| 5   | 13.95   | 11.46   |
| 6   | 307.28  | 37.88   |
⋮
| 93  | 23.37   | 10.82   |
| 94  | 125.34  | 23.55   |
| 95  | 78.86   | 19.74   |
| 96  | 112.79  | 25.46   |
| 97  | 10.23   | 11.14   |
| 98  | 82.89   | 31.83   |
| 99  | 48.27   | 23.08   |
```

```
x=log.(X.d)
y=log.(X.BGalhos)
Resumo_rlm(x,y,0.05)
Coeficientes_estimados: [-3.89732, 2.64394]
QME: 0.3164853999352024
using PyPlot
f(x) =-3.89732+2.64394*x
x = minimum(X.d):0.01: maximum(X.d)
logx=log.(x)
yhat=exp.(f.(logx))
plot(x,yhat,"b",label="Sem correção")
xlabel("Diâmetro à altura do peito (cm)")
ylabel("Biomassa dos galhos (kg)")
plot(X.d,X.BGalhos,"g.",label="dados")

#Correção
f(x) =-3.89732+2.64394*x+(0.3164853999352024/2)
x = minimum(X.d):0.01: maximum(X.d)
logx=log.(x)
yhatC=exp.(f.(logx))
plot(x,yhatC,"r",label="Com correção")
legend(loc="upper left")
```



O modelo ajustado resultou na equação $ln\ (\hat{B}Galhos_i) = -3{,}89732 + 2{,}64394\ ln(d_i)$. A correlação da biomassa dos galhos com o diâmetro à altura do peito é moderada, logo, o valor do coeficiente de determinação é moderado e, no presente caso, o valor de QME é relativamente alto. Observe que a diferença sistemática entre os valores corrigidos e não corrigidos, devido à discrepância logarítmica, é expressivo, daí a importância da aplicação da correção, especialmente quando o valor do QME é grande.

Ainda, concernente à seleção dos modelos do exercício de volume, uma última tentativa vale ser testada, o ajuste do modelo $v_i = \beta_0\, d_i^{\beta_1}\, h_i^{\beta_2}\, e_i$. Este é um modelo não linear e pode ser linearizado para que os coeficientes β_i possam ser estimados pelo método dos mínimos quadrados. Desta maneira, o modelo pode ser reescrito como segue:

$$ln\ v_i = ln\ \beta_0 + \beta_1\ ln\ d_i + \beta_2\ ln\ h_i + ln\ e_i$$

Em que:
ln é o logaritmo natural.

Pede-se:

Ajuste o modeloln $ln\ v_i = ln\ \beta_0 + \beta_1\ ln\ d_i + \beta_2\ ln\ h_i + ln\ e_i$ e obtenha os valores de QME, Coeficiente de determinação ajustado e coeficiente de variação corrigidos.

Comandos para ajuste de um modelo de regressão linear múltipla linearizável:

Arquivo: 53 - Função resumo_rlmc para regressão linear linearizável com correção

```
>using Statistics, Distributions,LinearAlgebra

>function Resumo_rlmC(x,y,α)
    X = [ones(size(x,1)) x]
    β =inv(X'X)*X'*y
    yhat = X*β
    n=n=size(x,1)
    nβ=length(β)
```

```
    GLR= nβ-1
    GLT=n-1
    GLE= GLT- GLR
    Σy = sum(y)
    SQR= β'*X'*y-(Σy^2)/n
    SQE= y'*y-β'*X'*y
    SQT = y'*y-(Σy^2)/n
    QMR=SQR/GLR
    QME=SQE/GLE
    FatorDeCorreção=QME/2
    yhatC = exp.(yhat.+(QME/2))
    yC=exp.(y)
    SSRC= sum((yhatC.- mean(yC)).^2)
    SSEC= sum((yhatC- yC).^2)
    SSTC = sum((yC.- mean(yC)).^2)
    MSEC=SSEC/GLE
    SYX=sqrt(MSEC)
    CV= SYX/mean(yC)*100
    R²=1-(SSEC/SSTC)
    R²aj=1-(MSEC/(SSTC/GLT))
    println("Coeficientes_estimados: ", β)
    println("QME: ", QME)
    println("SYX: ", SYX)
    println("CV (%): ", CV)
    println("R²: ", R²)
    println("R²aj: ", R²aj)
    println("FatorDeCorreção: ", FatorDeCorreção)
end

>using DataFrames, CSV
>X=CSV.read("Z:/Alexandre/Julia/Exemplos/descritiva.csv", DataFrame);
x=[log.(X.d) log.(X.h)]
y=log.(X.v)

Resumo_rlmC(x,y,0.05)
Coeficientes_estimados: [-8.95052, 1.895464, 0.67355]
QME: 0.00941
SYX: 0.00837
CV (%): 7.09
R²: 0.9854
R²aj: 0.9849
FatorDeCorreção: 0.004704
```

Os resultados obtidos por meio do ajuste do modelo $ln\ v_i = ln\ \beta_0 + \beta_1\ ln\ d_i + \beta_2\ ln\ h_i + ln\ e_i$ são um pouco melhores que o modelo $v_i = b_0 + b_1\ d^2\ h_i$. Os dois modelos são compostos pelas mesmas variáveis, mas suas diferentes estruturas podem resultar em estatísticas de ajuste diferentes.

Um gráfico de resíduos para cada modelo pode ser construído.

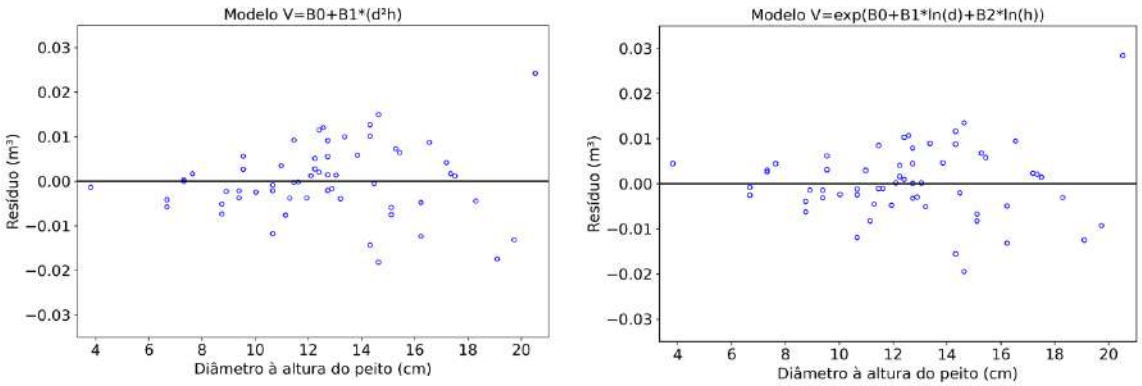
Comandos para obtenção do gráfico de resíduos para os modelos ajustados:

Arquivo: 54 - ar - Comparação de modelos

```
>using DataFrames, CSV, PyPlot
>X=CSV.read("Z:/Alexandre/Julia/Exemplos/descritiva.csv", DataFrame)

#Modelo V=B0+B1*(d²h)
f(x) =0.00732284+3.63139e-5*x
x=(X.d.^2).*X.h
yhat=f.(x)
ei=X.v.-yhat
scatter(X.d, ei, facecolor= "none", edgecolors="blue", s=20)
xlabel("Diâmetro à altura do peito (cm)")
ylabel("Resíduo (m³)")
title("Modelo V=B0+B1*(d²h)")
ylim(-0.035, 0.035)
xlim(minimum(X.d)-0.5, maximum(X.d)+0.5)
plot([0, maximum(X.d)+0.5],[0,0],c=:black)

#Modelo ln(v)=B0+B1*ln(d)+B2*ln(h)
f(x1,x2) =exp((-9.55579+1.87725*x1+0.903083*x2+0.0018508996914900855))
x1=log.(X.d)
x2=log.(X.h)
yhat=f.(x1,x2)
ei=X.v.-yhat
scatter(X.d, ei, facecolor= "none", edgecolors="blue", s=20)
xlabel("Diâmetro à altura do peito (cm)")
ylabel("Resíduo (m³)")
ylim(-0.035, 0.035)
xlim(minimum(X.d)-0.5, maximum(X.d)+0.5)
title("Modelo ln(v)=B0+B1*ln(d)+B2*ln(h)")
plot([0, maximum(X.d)+0.5],[0,0],c=:black)
```



Os resíduos da equação ajustada para o modelo $v_i = exp[b_0 + b_1 ln(d) + b_2 ln(h)]$ estão um pouco menos dispersos quando comparado com a equação ajustada para o modelo $v_i = b_0 + b_1 d^2 h_i$. O modelo linear múltiplo linearizado apresentou estatísticas de ajustes um pouco melhores para os dados amostrais, indicando que deve ser utilizado. Quando se utiliza modelos linearizados o mais interessante é avaliar os resíduos na escala em que a variável *y* foi ajustada, o que não foi aplicado para o presente caso.

Observe que, conforme aumentam os valores de diâmetro à altura do peito, também aumenta a dispersão dos resíduos. Isto é denominado de resíduos heterogêneos, notavelmente a variância dos resíduos, que é proporcional ao valor da variável independente. Neste caso, uma regressão ponderada é recomendada, especialmente se o objetivo é avaliar a significância das estatísticas F e t, bem como a construção dos intervalos de confiança para a média e para a predição (as fórmulas para o cálculo dessas estatísticas devem considerar um peso). Se apenas o valor médio a ser estimado é o objetivo, então a regressão ponderada nesse caso não é explicitamente importante.

No caso de regressão linear múltipla (regressão com várias variáveis independentes), os intervalos de confiança e para uma nova predição podem ser calculados da forma matricial utilizada para modelos lineares simples. A função *Bandas_rlsno* pode ser utilizada para o cálculo dos intervalos para um valor, em que x_0 é uma matriz que contém os valores das variáveis independentes com as novas observações. Observe, que as condicionantes da análise de regressão são requeridas para a apresentação destas bandas.

Comandos para obtenção do intervalo de confiança e predição para um novo valor:

Arquivo: 55 - Função bandas-rlsno para intervalo de confiança e predição para um novo valor

```
>using Statistics, Distributions

>function Bandas_rlsno(x,y,α,x0)
    X = [ones(size(x,1)) x]
    β =inv(X'X)*X'*y
    yhat = X*β
    n=size(x,1)
    nβ=length(β)
    GLR= nβ-1
    GLT=n-1
    GLE= GLT- GLR
    SQE= y'*y-β'*X'*y
    QME=SQE/GLE
    AINV=inv(X'*X)
    nx0= size(x0,1)
    X0 = [ones(size(x0,1)) x0]
    yhatX0 = X0*β
    tα = quantile(TDist(n-2),1- α/2)
    Vβ= AINV*QME
    Syhat_X0= sqrt.([X0[i,:]'*Vβ*X0[i,:] for i in 1: nx0])
    CIL= yhatX0.-( tα.* Syhat_X0)
    CIU= yhatX0.+( tα.* Syhat_X0)
    Syhatno= sqrt.([(1+X0[i,:]'*AINV*X0[i,:])*QME for i in 1: nx0])
    CIPL= yhatX0.-( tα.* Syhatno)
    CIPU= yhatX0.+( tα.* Syhatno)
    df=DataFrame(ValorEstimado=yhatX0, CIL=CIL, CIU=CIU, CIPL=CIPL, CIPU=CIPU)
end
```

```
>using DataFrames, CSV, PyPlot
>X=CSV.read("Z:/Alexandre/Julia/Exemplos/descritiva.csv", DataFrame);

#Bandas_rlsno(x,y,α,x0)
Bandas_rlsno([(X.d.^2) X.h],X.v,0.05,[18 15]);
```

20.8 REGRESSÃO PONDERADA

Se qualquer uma das condições associadas ao modelo for violada, os intervalos de confiança e os intervalos de predição também podem ser inválidos. Por esta razão que é tão importante averiguar as condições apresentadas pelos resíduos e por outros elementos elucidativos da qualidade de ajuste de um modelo.

Se na modelagem de variáveis os resíduos estão distribuídos de maneira heterogênea em relação às variáveis aleatórias utilizadas nas funções, para a própria variável dependente, isso indica que a variância não é constante. Para aplicação de mínimos quadrados a condição $E[\varepsilon_i \, \varepsilon_j] = \sigma_{ij} \, I$ (resíduos homogêneos) deve ser assegurada.

Existem várias formas para se detectar a homogeneidade das variâncias dos resíduos, destacando-se duas principais:

1. construir um gráfico de resíduos ε_i para a variáveis X_i e Y_i visando detectar a ocorrrência de algum padrão sistemático na sua distribuição dos resíduos, o que pode sugerir a não homogeneidade dos erros;
2. aplicar testes estatísticos.

No exercício da modelagem do volume do fuste, foi observado que ambos os modelos $ln \, v_i = ln \, \beta_0 + \beta_1 \, ln \, d_i + \beta_2 \, ln \, h_i + ln \, e_i$ (transformando para a escala original, ou seja, $\hat{v}_i = exp^{[\hat{\beta}_0 \cdot d_i^{\hat{\beta}_1} \cdot h_i^{\hat{\beta}_2}]}$) e $v_i = b_0 + b_1 \, d^2 \, h_i$ resultaram em resíduos com variâncias heterogêneas, observando-se um gráfico. Isto não é um problema do modelo, é uma condição observada nos dados. Árvores com menores valores de diâmetro à altura do peito e/ou altura apresentam menor variação do volume do que em árvores maiores (valores grandes de diâmetro à altura do peito e/ou altura). Logo, a aplicação de regressão ponderada torna-se apropriada para acomodar esta relação de resíduos não homogêneos.

Um estimador de mínimos quadrados ponderados é dado por:

$$\beta = (X' \, W^{-1} \, X)^{-1} \, X' \, W^{-1} \, Y$$

Em que:
W é uma matriz diagonal de pesos.

Os pesos podem ser obtidos utilizando-se as relações: $1/X$ ou $1/X^2$, por exemplo. Alternativamente, eles podem ser calculados. Ajusta-se, então, um modelo para estimar os resíduos: a variável dependente é o quadrado do resíduo e as variáveis independentes são as mesmas usadas no modelo que foi inicialmente ajustado para modelar a variável de interesse.

Os coeficientes angulares obtidos servirão para obter um peso para cada observação. Ao realizar a modelagem dos resíduos a estrutura de suas variâncias é verificada, cujo modelo é representado por:

$$\sigma^2 = \beta_0 X_1^{\beta_1} X_2^{\beta_2} + \cdots + X_k^{\beta_k}$$

Preferencialmente, as variáveis X_k devem ser as mesmas do modelo que foi inicialmente ajustado para modelar a variável de interesse. Os coeficientes β_i podem ser estimados via mínimos quadrados, logo, o modelo anterior deve ser linearizado.

$$\ln(\sigma^2) = \beta_0' + \beta_1 ln(X_1) + \beta_2 ln(X_2) + \cdots + \beta_3 ln(X_k)$$

Dado ao ajuste, os pesos para cada observação podem ser calculados como segue:

$$W_i = X_1^{\beta_1} X_2^{\beta_2} + \cdots + X_k^{\beta_k}$$

Os pesos obtidos são dispostos em uma matriz diagonal, a qual é denominada de matriz dos pesos:

$$W = \begin{bmatrix} W_1 & 0 & 0 & 0 \\ 0 & W_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & W_n \end{bmatrix}$$

Pede-se:

Para os dados do exemplo do volume, faça o ajuste ponderado do modelo $v_i = b_0 + b_1 d^2 h_i$. Apresente o gráfico dos resíduos ponderados.

Comandos para obtenção do ajuste de uma regressão linear ponderada:

Arquivo: 55 - Função bandas-rlsno para intervalo de confiança e predição para um novo valor

```
#Função COEF_RLMP para regressão linear ponderada

>using Statistics, Distributions,LinearAlgebra

>function Coef_rlmP(x,y,α)
    X = [ones(size(x,1)) x]
    β =inv(X'X)*X'*y
    yhat = X*β
    ei=y.-yhat
    lnei²=log.(ei.^2)
    lnx=log.(x)
    XR=[ones(size(x,1)) lnx]
    βR =inv(XR'XR)*XR'*lnei²
    βRd=Diagonal(βR[2:end])
    Wi=x.^βRd
    W=Diagonal(Wi)
    βP=inv(X'*inv(W)*X)*X'*inv(W)*y
end
```

```
using DataFrames, CSV,  PyPlot
X=CSV.read("Z:/Alexandre/Julia/Exemplos/descritiva.csv", DataFrame)

x=(X.d.^2).*X.h
y=X.v

Coef_rlmP(x,y,0.05)
Coeficientes_Não_Ponderados:
0.0073
 3.6313e-5
Coeficientes_Ponderados:
0.0073
 3.6313e-5
Coeficientes_FunçãoDePeso: 0.9355
```

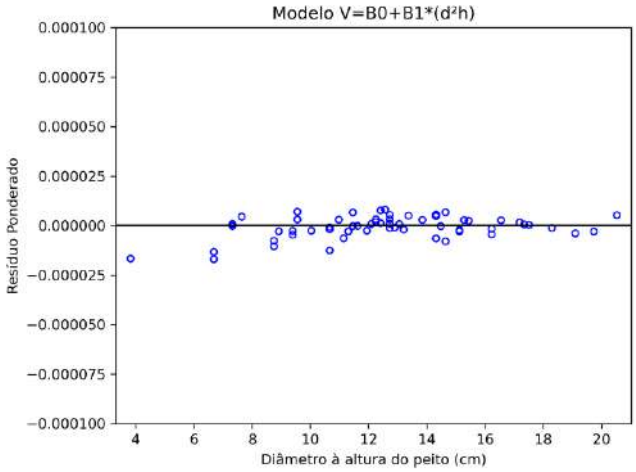
Note que os coeficientes estimados sem ponderação e com ponderação resultam em valores próximos. Isto revela que mesmo com resíduos heterogêneos a média é apropriadamente estimada, embora as estimativas intervalares estejam comprometidas. A estrutura da variância é modelada pela função $\sigma^2 = (d^2\ h)^{0,9355}$ e os pesos correspondem ao inverso da variável independente, ou seja, $W_i = 1/(d^2\ h)_i^{0,935519}$.

O Resíduo ponderado é dado por:

$$e_{iW} = (v_i - \hat{v}_i)\ W_i$$

Comandos para obtenção do gráfico dos resíduos de uma regressão linear ponderada:

```
#Modelo com ponderação V=B0+B1*(d²h)
f(x) =0.0073+3.6314e-5*x
x=(X.d.^2).*X.h
yhat=f.(x)
ei=X.v.-yhat
Wi=x.^-0.9355
eiP=ei.*Wi
scatter(X.d, eiP, facecolor= "none", edgecolors="blue", s=20)
xlabel("Diâmetro à altura do peito (cm)")
ylabel("Resíduo Ponderado")
title("Modelo V=B0+B1*(d²h)")
xlim(minimum(X.d)-0.5, maximum(X.d)+0.5)
ylim(-0.0001,0.0001)
plot([0, maximum(X.d)+0.5],[0,0],c=:black)
```



A ponderação da equação, por meio da modelagem da estrutura da variância dos resíduos, tornou os resíduos homogêneos. Ressalta-se que, para a árvore com menor valor, o seu resíduo se afastou dos demais.

Conduza o mesmo processo de modelagem para o modelo $ln\ v_i = ln\ \beta_0 + \beta_1\ ln\ d_i + \beta_2\ ln\ h_i + ln\ e_i$. Conclua sobre qual dos dois modelos concorrentes é indicado para estimar o volume no presente caso.

Observa-se que os valores dos estimadores: coeficiente de determinação ajustado, erro-padrão e coeficiente de variação podem ser calculados sem levar em consideração os pesos. Entretanto, se o interesse é avaliar as estatísticas F, *t* e construir os intervalos de confiança para a média e para a predição, os pesos devem ser considerados nas estruturas dos cálculos.

20.9 DETERMINAÇÃO DE EQUAÇÕES DE TENDÊNCIA

Existem situações em que os tratamentos testados em um experimento se constituem em variáveis quantitativas. Nestas situações é possível determinar qual a tendência da variável resposta em função dos tratamentos aplicados. Para tanto, deve-se ajustar uma equação de regressão para representar a resposta em função dos tratamentos, bem como a elaboração de um gráfico desta relação. O objetivo primordial é determinar que tipo de relação existe entre os tratamentos e as variáveis resposta mensuradas. O principal produto deste procedimento é determinar o tipo de relação existente, não havendo ênfase nas estatísticas da equação ajustada. Os pontos de máximo e mínimo podem ser determinados auxiliando a interpretação dos resultados.

Se o objetivo é obter a tendência da relação somente com os valores das médias dos tratamentos, então, a aplicação de regressão por meio de contrastes ortogonais torna-se apropriado.

21 ASPECTOS NA CONDUÇÃO DE EXPERIMENTOS - PLANO EXPERIMENTAL

Em geral nos experimentos objetiva-se o teste de uma determinada hipótese. Cada delineamento pode ser representado por meio de um modelo linear geral $y_{ij} = f(x_j) + e_{ij}$, que já foi discutido anteriormente. Assim, na sequência são destacados aspectos que podem ser úteis para a escolha do delineamento ou sua estrutura experimental, bem como revelar a importância do experimento, suas características e a apresentação de um relatório experimental.

1. Você realmente precisa conduzir um experimento?

- A informação pode ser encontrada de outra maneira?
- O tempo e os recursos financeiros necessários para executá-los são suficientes?
- Você tem restrições quanto às condições experimentais ou equipamentos específicos?

2. Indique as hipóteses a serem testadas.

Estabelecer uma hipótese é fundamental para qualquer pesquisa. A estrutura experimental é organizada para se fazer inferências sobre as hipóteses estabelecidas.

Nos ensaios experimentais, uma hipótese deve ser proposta para examinar a relação entre uma variável independente (tratamentos) e uma variável dependente (variável mensurada). O efeito sobre a variável dependente depende ou é determinado pelo que acontece quando se altera a variável independente (níveis do fator que está sendo alterado). Portanto, refere-se aos motivos que determinaram a execução de um experimento para maximizar ou minimizar os valores da variável dependente em função da variável independente.

Por exemplo, o teor de carbono da madeira (variável independente) aumenta o poder calorífico da madeira (variável dependente). Se for alterada a variável independente (aumentar o teor de carbono da madeira), ela mudará a variável dependente (poder calorífico da madeira). Provada a hipótese que o teor de carbono determina o aumento do poder calorífico, outra questão interessante da pesquisa aflora:

Como aumentar o teor de carbono da madeira visando propiciá-la a atingir um melhor poder calorífico?

O desenvolvimento de uma hipótese testável obriga o profissional que está planejando o experimento a refletir sobre os resultados esperados de um ensaio experimental, permitindo-lhe compreender como responder a questão por meio das variáveis (dependente e independente) envolvidas no experimento. Assim, o estabelecimento da hipótese auxilia-o a estabelecer previsões com base em pesquisas anteriores.

A construção de uma hipótese, de modo simplificado, pode ser considerada como uma pergunta para a qual deseja-se obter a resposta segundo critérios científicos. Para o exemplo anterior, pode-se imaginar se os teores de carbono podem ser alterados por meio de adubação, sugerindo a seguinte questão:

A aplicação de adubação altera os teores de carbono? Para teste desta hipótese deve-se:

▪ Realizar uma pesquisa preliminar

A resposta inicial à pergunta deve ser baseada no que já foi pesquisado sobre o assunto. A consulta a pesquisas anteriores é importante para formular suposições fundamentadas sobre os possíveis resultados experimentais. Quais são as informações da literatura que revelaram efeitos de adubação para o aumento do teor de carbono?

▪ Construção da hipótese

Dadas as evidências encontradas na literatura, novas suposições a respeito dos resultados que se esperam observar no futuro experimento podem ser realizadas. Escreva a pergunta inicial a ser respondida, e os resultados esperados como resposta de modo claro e conciso. Por exemplo, a resposta esperada para a hipótese formulada é que níveis altos de adubação fosfatada propiciam maiores valores do teor de carbono para as árvores.

▪ Refine e avalie a hipótese

A hipótese proposta deve ser específica, testável estatisticamente (para os experimentos florestais), e com alguns componentes importantes:

- Quais variáveis dependentes devem possibilitar a obtenção de respostas relevantes sobre o teor de carbono?
- Quais tratamentos devem ser estudados e que são representados pela variável independente (doses de adubação fosfatada)?
- Quais sãos os resultados esperados do experimento ou análise? Ou seja, o resultado da função de correspondência da variável dependente com a independente. (Níveis alto de adubação fosfatada implicam maiores valores do teor de carbono).

Para identificar os componentes da hipótese, pode-se utilizar a forma simples de raciocínio que implica em estabelecer a sequência representada por *se...então...por quê*. A primeira parte da frase indica a variável independente (tratamentos), a segunda parte indica a variável dependente (variável que se quer maximizar ou minimizar os valores) e a última justifica sua utilização. Assim, para o caso anterior, a seguinte hipótese poderia ser elaborada:

» **Se** as árvores forem manejadas com altas concentrações de adubações fosfatadas, **então** seus valores de teor de carbono serão maiores **porque** o fósforo proporciona maiores teores de carbono.

Essa hipótese pode ser reformulada em termos de efeito da variável independente em função da variável dependente, na qual se pressupõe uma relação possível entre as variáveis, tal como:

» Altas concentrações de adubação fosfatada propiciarão um efeito positivo nos teores de carbono da madeira.

Essas hipóteses especificadas que motivam a condução de experimentos são testadas por meio de testes estatísticos. Então, hipóteses estatísticas de efeito nulo (nulidade - H_0) e de efeito diferencial (alterantiva - H_1) são propostas:

$$\begin{cases} H_0: \text{não existe efeito de adubação sobre o teor de carbono} \\ H_1: \text{existe efeito de adubação sobre o teor de carbono} \\ \alpha = 0,05 \end{cases}$$

A hipótese testada por meio de qualquer teste estatístico é a hipótese da nulidade, sendo que quando da sua rejeição conclui-se pela hipótese H_1, sujeita a um erro α.

Note que a hipótese para a realização do experimento determina todo o seu planejamento experimental. A avaliação das hipóteses formuladas deve contemplar:

- Se a hipótese a ser testada contempla o objetivo da pesquisa.
- Se a hipótese inclui uma variável independente (VI) e uma variável dependente (VD). Se as variáveis a serem controladas (VI) e mensuradas (VD) em um experimento são viáveis.
- Se a hipótese formulada pode ser testada por meio dos experimentos florestais e por meio de análise estatística dos dados experimentais.
- Se a hipótese formulada engloba o que se tem como expectativa durante o experimento.

A hipótese formulada especifica as previsões que se espera de sua pesquisa. Ela é uma resposta provisória à sua pergunta de pesquisa e que ainda não foi testada, mas que, sendo programada por meio de um experimento, poderá revelar evidências experimentais que a afirmam ou a refutam. Para alguns experimentos, várias hipóteses podem ser formuladas para abordar os diferentes aspectos inerentes ao problema pesquisado. O estabelecimento de hipóteses devem ser baseados em teorias e conhecimentos existentes.

3. Definir os objetivos do experimento

A solução de um problema implica na sua prévia definição. Se o problema não for definido ele não pode ser resolvido. Se o objetivo de um experimento não pode ser definido, o experimento não pode ser efetuado. Isto implica em:

- Quais as perguntas sobre o problema você acha mais interessantes de serem respondidas?
- Os objetivos gerais foram estabelecidos?
- Quão abrangente a pesquisa precisa ser?

4. Escolha as variáveis independentes (tratamentos) que você deseja avaliar. Quais são as variáveis dependentes que você quer estimar para testar a hipótese formulada?

- Quais são as variáveis controladas ou independentes? Quais são seus níveis? São de interesse prático?
- Quais são as variáveis aleatórias dependentes que representam os efeitos avaliados?

5. Identificar as variáveis dependentes importantes

- Quais são as variáveis dependentes que realmente necessitam ser avaliadas?
- A exatidão da mensuração é suficiente para distinguir os efeitos dos tratamentos?

6. Estruture o experimento

- Como alcançar os objetivos experimentais da maneira mais simples, com o número mínimo de medidas e o menor custo?

FIGURA 21.1. Desenhando o experimento.



7. Pontos chaves no planejamento experimental

Construa um quadro resumo sobre o planejamento experimental.

FIGURA 21.2. Fluxograma do planejamento experimental.

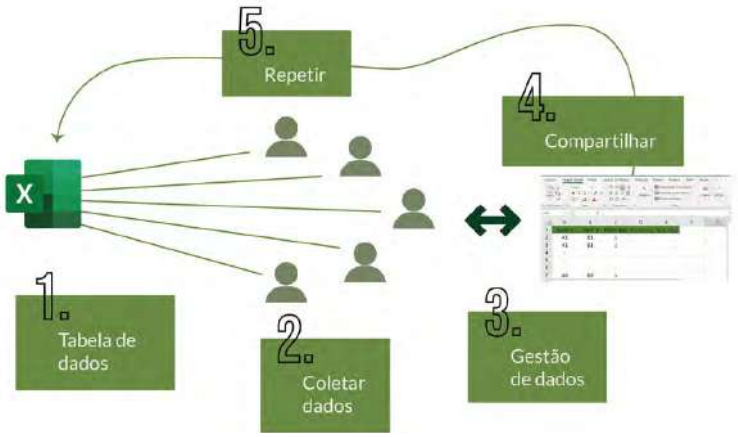


8. Coleta e organização de dados

A coleta e organização de dados são uma parte integral e crítica da pesquisa. A coleta de dados e sua análise ajudam os pesquisadores a responderem as questões da pesquisa formuladas nas hipóteses.

O armazenamento de dados em tabelas organizadas também é uma parte importante do experimento.

FIGURA 21.3. Passos da coleta e organização dos dados.



9. Análise e apresentação de dados

A análise de dados é o processo que leva a obtenção das respostas às perguntas formuladas e declaradas nas hipóteses, por meio do exame e interpretação dos dados.

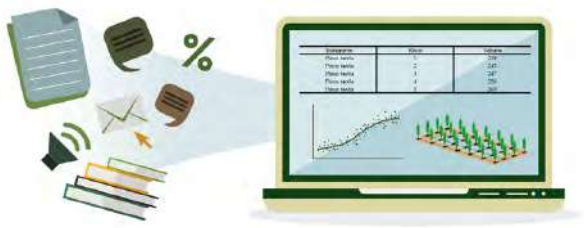
Análise de dados é o processo pelo qual o pesquisador aplica uma combinação de duas ou mais técnicas estatísticas para descrever, avaliar ou ilustrar um conjunto de dados obtidos de um experimento, permitindo a obtenção de conclusões, que é a parte mais importante de uma pesquisa.

A maioria dos leitores interessa-se pela interpretação dos resultados e não pelos dados em si. Sempre que possível é aconselhável apresentar um resumo dos resultados obtidos para os dados coletados por meio da apresentação das estatísticas descritivas básicas inerentes aos dados coletados. Estas informações podem ser importantes aos leitores quando do planejamento de experimentos similares aos que estão sendo reportados.

O uso de dados de pesquisas científicas relevantes é muito importante. Sempre que necessário cite essas referências.

Uma prática recomendável é a de realizar as análises pertinentes tão logo os dados sejam coletados. Este procedimento possibilita a detecção de erros prematuramente, bem como permite que o pesquisador fique familiar com as técnicas de análises sendo utilizadas. Este conhecimento permite que eventuais alterações necessárias possam ser feitas nos estágios iniciais da pesquisa.

FIGURA 21.4. Análise dos dados.



10. Apresentação dos resultados

Considere o modelo linear aditivo de um experimento instalado segundo um delineamento inteiramente casualizado (DIC):

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

Em que:

y_{ij} é a observação do *i*-ésimo tratamento na *j*-ésima unidade experimental;

μ é o efeito constante (usualmente corresponde à média geral) – constante inerente a todas as observações;

τ_i é o efeito do *i*-ésimo tratamento – efeito propiciado pelo tratamento *i* (desvio em relação a μ, decorrente do efeito);

ε_{ij} é o erro associado ao *i*-ésimo tratamento na *j*-ésima unidade experimental - efeito aleatório na unidade experimental (erro).

Considere, adicionalmente, que y_{ij} é uma variável dependente (VD) e τ_i é uma variável independente (VI).

O experimento é planejado para maximizar (ou minimizar) os valores do vetor $\bar{y}$ em função dos tratamentos τ_i. Se nas unidades experimentais o controle local for efetuado, então, o modelo pode ser expandido para os delineamentos DBA e DQL. Ainda, se no ensaio as respostas esperadas para maximizar (ou minimizar) os valores do vetor $\bar{y}$ dependem de mais de um fator, ensaios fatoriais ou em parcelas-subdivididas podem ser projetados.

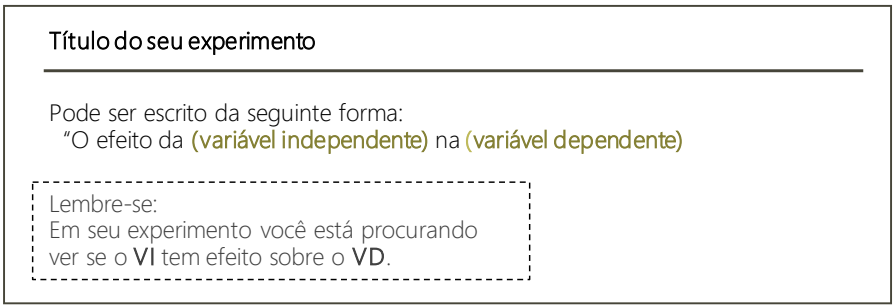
A apresentação dos resultados da pesquisa deve se a ter ao modelo estatístico com o qual o experimento foi conduzido. Considere um ensaio experimental simples em DIC, então um relatório pode ser escrito contemplando os seguintes componentes:

Título

O título do trabalho deve se referir ao efeito da variável independente (VI) sobre a variável dependente (VD). Lembre-se que, em um experimento, a hipótese da nulidade é testada por meio de um teste estatístico e implica que a variável independente não afeta a variável dependente.

Um título tal como efeito da VI sobre a VD seria considerado pobre nas atuais formas de redação científica. Entretanto, ainda remanesce a importância do modelo estatístico utilizado no ensaio experimental, para determinar o título do relatório.

FIGURA 21.5. Composição do título do experimento.



Resumo

- Definir o problema. Por que estudar o efeito da VI sobre a VD?
- Sumarizar como o problema foi abordado. Como foi estudado o efeito da VI sobre a VD?
- Indicar os resultados importantes que foram encontrados. Quais foram os efeitos de VI sobre a VD?
- Recomendar um tratamento (τ_i), que conforme objetivos maximiza ou minimiza os valores de $\overline{y}$.

Introdução

Definir o problema. Revelar porque pesquisar níveis τ_i (VI) e seus efeitos sobre $\overline{y}$ (VD) são um problema importante que merece ser estudado.

- Estabelecer a hipótese do experimento.
- Destacar o objetivo do experimento. Os objetivos referem-se à avaliação dos efeitos de VI sobre VD.

Materiais e métodos

Envolve descrever os procedimentos utilizados passo-a-passo, sobre como um ensaio experimental foi realizado.

- Local e período em que o experimento foi conduzido.
- Qual foi o modelo estatístico utilizado para o ensaio experimental (DIC, DBA, DQL)?
- Quais foram os níveis de VI testados? Se o ensaio foi unifatorial (τ_i) ou foram testados mais que um fator.
- Quais foram as variáveis respostas mensuradas (VD)?
- Como o experimento foi instalado? Quantas repetições J foram utilizadas? Como foi dimensionada a unidade experimental? Se foi necessário efetuar o controle das unidades experimentais por meio de bloqueamento (DBA, DQL).
- Como as variáveis respostas VD foram mensuradas? Qual é o tipo destas variáveis e quais são suas escalas?
- Como foi realizada a análise estatística dos dados para testar a hipótese da nulidade de que VI não afeta VD? Quais testes estatísticos foram aplicados? Quais condições ou condicionantes foram exigidas pelos testes?

Lembre-se da importância de apresentar um resumo dos dados (estatísticas descritivas) para as variáveis VD, que pode ser geral ou para cada nível de τ_i.

Destacar as fontes geradoras de erro (ε_{ij}) e as limitações do experimento.

Resultados

- Apresentar um quadro prospectivo para serem incluídas as estatísticas descritivas para as variáveis VD.
- Quais foram os resultados dos testes estatísticos aplicados para testar a hipótese da nulidade de que VI não afeta VD? Se a hipótese foi rejeitada, quais foram os grupos de tratamentos (τ_i) que minimizam ou maximizam VD?

Discussão dos resultados

Os resultados devem ser discutidos detalhadamente. Por que foram encontrados estes resultados? Elencar os motivos de VI não afetar VD ou de VI afetar VD. Adicionar as razões sobre a recomendação dos grupos de tratamentos (τ_i) que minimizam e maximizam VD.
Concluir a respeito das hipóteses do experimento. Observe que existe uma diferença entre a hipótese estatística e a hipótese para o experimento.

Conclusão

Apresentar sinteticamente as descobertas sobre os efeitos de VI sobre VD evidenciadas com a condução do experimento.
Apresentar suas recomendações sobre os níveis de τ_i (VI) que minimizarão ou maximizarão $\overline{y}$ (VD).

Referências bibliográficas

As referências bibliográficas devem ser selecionadas apropriadamente para contribuir sobre a formulação e conclusões concernentes às hipóteses testadas pelo experimento e, também, para construir um caminho coerente para a análise estatística dos dados. Para tanto, é interessante organizar uma base de dados de pesquisas científicas relevantes sobre o assunto a ser tratado no experimento.

11. Ciclo da pesquisa

Na Figura 21.6 são sumarizadas todas as etapas efetuadas na realização de qualquer experimento.

FIGURA 21.6. Ciclo de pesquisa.



A Figura apresentada representa todas as etapas envolvidas na execução de um experimento. O objetivo do presente fluxograma foi abordar alguns aspectos envolvidos na experimentação florestal.

REFERÊNCIAS.

ACKOFF, R. L.; GUPTA, S. K.; SAYER MINAS, J. Scientific method: optimizing applied research decisions. New York: John Wiley & Sons, 1962. 464 p.

ANDERSON, V. L.; McLEAN, R. A. Design of experiments: a realistic approach. New York: Marcel Dekker, 1974. 418 p.

ANDREW, I. Simple experimental design for forestry field trials. Revised ed. Rotorua: Forest Research Institute, 1986. 45 p.

BARBIN, D. Planejamento e análise estatística de experimentos agronômicos. Arapongas: Editora Midas, 2003. 208 p.

BEZANSON, J.; EDELMAN, A.; KARPINSKI, S.; SHAH, V. B. Julia: a fresh approach to numerical computing. SIAM Review, v. 59, n. 1, p. 65-98, 2017. Disponível em: https://doi.org/10.1137/141000671. Acesso em: 31 ago. 2026.

BOX, G. E. P.; HUNTER, W. G.; HUNTER, J. S. Statistics for experimenters: an introduction to design, data analysis, and model building. New York: Wiley, 1978. 653 p.

BOX, J. F. R. A. Fisher: the life of a scientist. New York: John Wiley & Sons, 1978. 512 p.

BUSSAB, W. O.; MORETTIN, P. A. Estatística básica. 6. ed. São Paulo: Saraiva, 2009. 535 p.

CHARNET, R.; FREIRE, C. A. L.; CHARNET, E. M. R.; BONVINO, H. Análise de modelos de regressão linear com aplicações. Campinas: Editora da UNICAMP, 1999. 356 p.

COCHRAN, W. G. Sampling techniques. 3. ed. New York: Wiley, 1977. 428 p.

COCHRAN, W. G.; COX, G. M. Experimental designs. 2. ed. New York: John Wiley & Sons; London: Chapman & Hall, 1957. 611 p.

CORDEIRO, G. M.; DEMÉTRIO, C. G. B.; MORAL, R. A. Modelos lineares generalizados e aplicações. São Paulo: Blucher, 2024. 256 p.

COX, D. R. Planning of experiments. New York: John Wiley & Sons, 1958. 308 p.

CUNHA, T. A.; FIGUEIREDO FILHO, A. Métodos estatísticos aplicados à ciência florestal: análise de regressão. Rio Branco: Stricto Sensu Editora, 2024. 335 p.

DRAPER, N. R.; SMITH, H. Applied regression analysis. New York: John Wiley & Sons, 1966. 407 p.

FERREIRA, P. V. Estatística experimental aplicada às ciências agrárias. Viçosa: Editora UFV, 2018. 588 p.

FERREIRA, R. S. Matemática aplicada às ciências agrárias: análise de dados e modelos. Viçosa: Editora UFV, 1999. 333 p.

FREESE, F. Linear regression methods for forest research. Fort Collins: U.S. Department of Agriculture, Forest Service, Rocky Mountain Forest and Range Experiment Station, 1964. 136 p.

HICKS, C. R. Fundamental concepts in the design of experiments. 2. ed. New York: Holt, Rinehart and Winston, 1973. 349 p.

HOEL, P. G. Estatística elementar. São Paulo: Atlas, 1977. 430 p.

KEMPTHORNE, O. The design and analysis of experiments. New York: John Wiley & Sons, 1952. 631 p.

KEPPEL, G. Design and analysis: a researcher's handbook. Englewood Cliffs, N.J.: Prentice-Hall, 1973. 658 p.

KOEHLER, H. S. Estatística experimental. Curitiba: UFPR/DFF, 1994. 123 p. (Apostila).

MEDEIROS, R. A.; PAIVA, H. N.; NOGUEIRA, G. S.; LEITE, H. G. Planejamento de experimentos com espécies florestais. 1. ed. Viçosa: Editora UFV, 2022. 184 p.

MONTGOMERY, D. C. Design and analysis of experiments. 8. ed. Hoboken: John Wiley & Sons, 2013. 724 p.

MONTGOMERY, D. C.; PECK, E. A. Introduction to linear regression analysis. 2. ed. New York: Wiley, 1992. 527 p.

MORETTIN, P. A.; BUSSAB, W. de O. Estatística básica. 5. ed. São Paulo: Saraiva, 2002. 526 p.

NAZARATHY, Y.; KLOK, H. Statistics with Julia: fundamentals for data science, machine learning and artificial intelligence. Cham: Springer, 2021. 527 p.

OEHLERT, G. W. A first course in design and analysis of experiments. New York: W. H. Freeman, 2000. 659 p.

OSTLE, B.; MENSING, R. W. Statistics in research: basic concepts and techniques for research workers. 3. ed. Ames: Iowa State University Press, 1975. 596 p.

PEARSON, F. A.; BENNETT, K. R. Statistical methods applied to agricultural economics. New York: John Wiley & Sons, 1942. 443 p.

PIMENTEL-GOMES, F. Curso de estatística experimental. 5. ed. São Paulo: Nobel, 1973. 430 p.

PIMENTEL-GOMES, F.; GARCIA, C. H. Estatística aplicada a experimentos agronômicos e florestais: exposição com exemplos e orientações para uso de aplicativos. Piracicaba: FEALQ, 2002. 309 p.

PRODAN, M. Forest biometrics. Oxford: Pergamon Press, 1968. 447 p.

QUEIROZ, W. T. de. Uma introdução à experimentação florestal. Belém: Universidade Federal Rural da Amazônia, 2023. 457 p.

SCHEFFÉ, H. The analysis of variance. New York: John Wiley & Sons, 1959. 477 p.

SCHNEIDER, P. R. Análise de regressão aplicada à Engenharia Florestal. 2. ed. Santa Maria: UFSM/CEPEF, 1998. 236 p.

SEBER, G. A. F. Linear regression analysis. New York: John Wiley & Sons, 1977.

SELTMAN, H. J. Experimental design and analysis. Pittsburgh: Carnegie Mellon University, 2018. 428 p.

SILVA, J. A. A.; SILVA, I. P. Estatística experimental aplicada à ciência florestal. Recife: Universidade Federal Rural de Pernambuco, 1982. 269 p.

SNEDECOR, G. W.; COCHRAN, W. G. Statistical methods. 6. ed. Ames: Iowa State University Press, 1967. 593 p.

SOKAL, R. R.; ROHLF, F. J. Biometry: the principles and practice of statistics in biological research. San Francisco: W. H. Freeman and Company, 1969. 776 p.

SOUZA, D. V.; OLIVEIRA, T. W. G. de; PIVA, L. R. de O.; SANTOS, J. X. dos; SANQUETTA, C. R.; CORTE, A. P. D. Introdução ao R: aplicações florestais. Curitiba: Ed. Autor, 2018. 135 p.

SPIEGEL, M. R. Estatística. Tradução de Pedro Cosentino. Rio de Janeiro: Ao Livro Técnico, 1969. 580 p.

SPROULL, N. L. Handbook of research methods: a guide for practitioners and students in the social sciences. 2. ed. Metuchen, N.J.: Scarecrow Press, 1995. 430 p.

STEEL, R. G. D.; TORRIE, J. H. Principles and procedures of statistics: with special reference to the biological sciences. New York: McGraw-Hill, 1960. 481 p.

STEEL, R. G. D.; TORRIE, J. H.; DICKEY, D. A. Principles and procedures of statistics: a biometrical approach. 3. ed. New York: McGraw-Hill, 1997. 666 p.

TRIOLA, M. F. Introdução à estatística. 7. ed. Rio de Janeiro: LTC, 1999. 410 p. Tradução de Alfredo Alves de Farias.

WINER, B. J. Statistical principles in experimental design. 2. ed. New York: McGraw-Hill, 1971.



JULIA E EXPERIMENTOS FLORESTAIS
PRIMEIRO CURSO

CURITIBA
2026