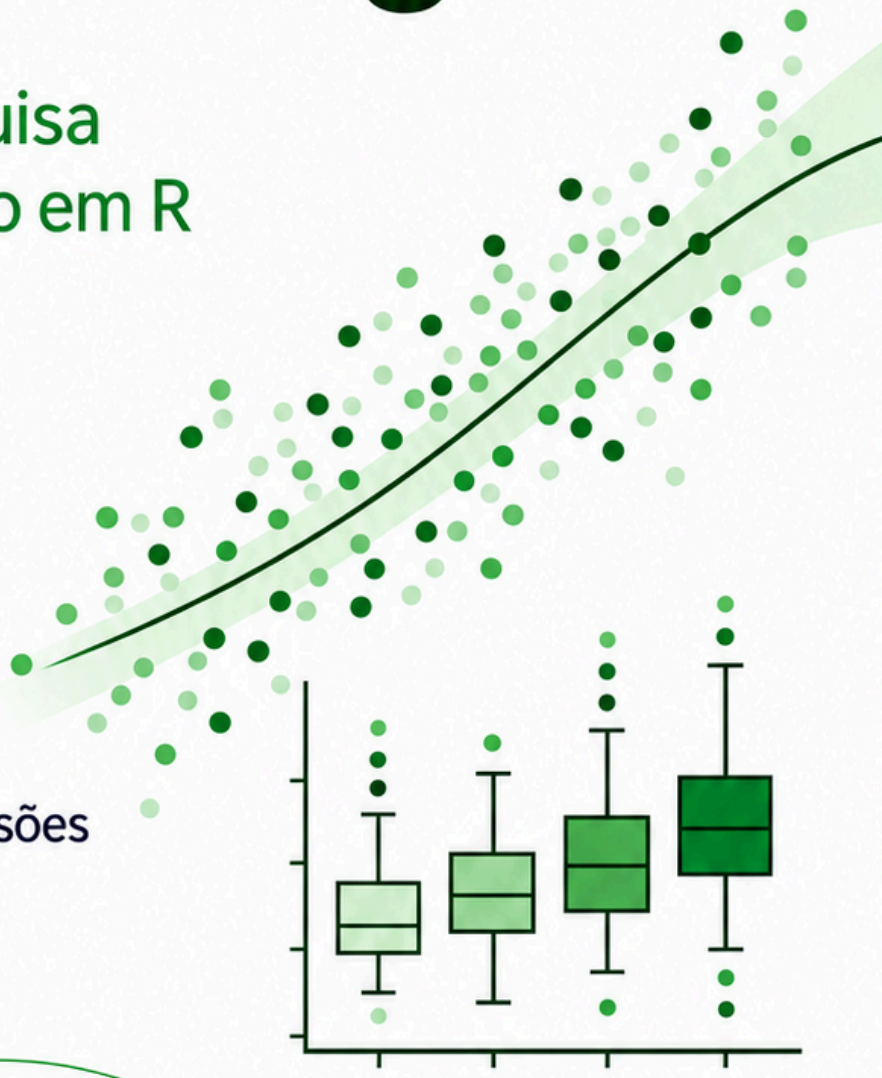
Análise e Visualização de Dados em Fitopatologia

Um guia para pesquisa e tomada de decisão em R

- Explorar dados
- Visualizar padrões
- Modelar relações
- Tomar melhores decisões



Emerson M. Del Ponte

Análise e Visualização de Dados em Fitopatologia

© 2026 Emerson M. Del Ponte

Esta obra está licenciada sob a licença Creative Commons Atribuição-NãoComercial 4.0 Internacional (CC BY-NC 4.0). É permitido copiar, distribuir, imprimir e adaptar o conteúdo para fins não comerciais, desde que seja atribuído o devido crédito ao autor. Texto da licença: creativecommons.org/licenses/by-nc/4.0.

Os códigos em R apresentados no livro podem ser reutilizados livremente.

Edição	1. ed., Viçosa, MG: Edição do autor, 2026
ISBN	978-65-02-48740-2
DOI	10.5281/zenodo.23889987
Versão online	https://emdelponte.github.io/analise-dados-fitopatologia/
Código-fonte	github.com/emdelponte/analise-dados-fitopatologia

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa

D636a 2026

Del Ponte, Emerson Medeiros, 1973-
Análise e visualização de dados em fitopatologia : um guia para pesquisa e tomada de decisão em R / Emerson Medeiros Del Ponte. – Viçosa, MG : O Autor, 2026.
1 livro eletrônico (298 p.) : il.

Bibliografia: p. 297-298.
ISBN: 978-65-02-48740-2

1. Fitopatologia – Métodos estatísticos. 2. Análise de dados. 3. R (Linguagem de programação). I. Titulo.

CDD 22. ed. 632.3

Bibliotecário responsável: Fábio Jaderson Miguel Reis CRB-6/3030

Como citar

DEL PONTE, E. M. **Análise e Visualização de Dados em Fitopatologia**. 1. ed. Viçosa, MG: Edição do autor, 2026. Disponível em: https://emdelponte.github.io/analise-dados-fitopatologia/. DOI: 10.5281/zenodo.23889987.

Sumário

Boas-vindas **7**

1 Por que pesquisar em Fitopatologia? **9**

1.1 O que é Fitopatologia? 9
1.2 Que tipos de pesquisa são realizados em Fitopatologia? 9
1.3 Onde a pesquisa em Fitopatologia é realizada? 9
1.4 Estudos representativos 10
1.5 Que tipos de dados os fitopatologistas coletam? 15
1.6 Como esses dados devem ser analisados? 16
1.7 Como este livro está organizado 17
1.8 Resumo do capítulo 19
1.9 Conceitos-chave 19
1.10 Exercícios 20
1.11 Reflexão 21
1.12 Próximos passos 21

2 Como delinear um estudo? **23**

2.1 Objetivos de aprendizagem 23
2.2 Por que o delineamento importa 23
2.3 Da questão biológica ao estudo 25
2.4 Estudos observacionais e experimentais 28
2.5 Os componentes básicos de um estudo 29
2.6 Unidades experimentais e unidades amostrais 31
2.7 Repetição 32
2.8 Casualização 34
2.9 Bloqueamento 35
2.10 Delineamentos experimentais comuns em Fitopatologia 36
2.11 Problemas comuns de delineamento em Fitopatologia 37
2.12 Pesquisa na prática: revisitando um experimento de manejo da doença 39
2.13 Planejando um estudo antes de coletar dados 39
2.14 Resumo do capítulo 40
2.15 Conceitos-chave 40
2.16 Exercícios 41
2.17 Reflexão 42
2.18 Próximos passos 42

3 Como os dados científicos devem ser organizados? **43**

3.1 Objetivos de aprendizagem 43
3.2 Por que a organização dos dados importa 43
3.3 Das observações ao conjunto de dados 44
3.4 Princípios de dados organizados (*tidy data*) 46
3.5 Estruturas de dados comuns em Fitopatologia 48
3.6 Tipos de variáveis 50
3.7 Metadados 52
3.8 Convenções de nomes 53

3.9 Problemas comuns na digitação de dados . . . 54
3.10 Validação dos dados . . . 56
3.11 Dados brutos versus dados processados . . . 57
3.12 Organizando projetos de pesquisa . . . 58
3.13 Pesquisa na prática . . . 59
3.14 Lista de verificação da organização dos dados . . . 61
3.15 Resumo do capítulo . . . 61
3.16 Conceitos-chave . . . 61
3.17 Exercícios . . . 62
3.18 Reflexão . . . 63
3.19 Próximos passos . . . 63

4 O que os dados revelam? . . . 65
4.1 Objetivos de aprendizagem . . . 65
4.2 Por que explorar os dados? . . . 65
4.3 Análise exploratória versus confirmatória . . . 66
4.4 Conhecendo o seu conjunto de dados . . . 67
4.5 Explorando uma variável . . . 70
4.6 Explorando relações . . . 73
4.7 Padrões que merecem investigação . . . 76
4.8 Interpretação biológica . . . 78
4.9 Pesquisa na prática . . . 79
4.10 Erros comuns . . . 81
4.11 Um fluxo de trabalho para a análise exploratória . . . 81
4.12 Resumo do capítulo . . . 82
4.13 Conceitos-chave . . . 83
4.14 Exercícios . . . 83
4.15 Reflexão . . . 84
4.16 Próximos passos . . . 84

5 Como visualizar dados de forma eficaz? . . . 85
5.1 Objetivos de aprendizagem . . . 85
5.2 Por que visualizar dados? . . . 85
5.3 A visualização como parte do raciocínio científico . . . 86
5.4 A anatomia de um gráfico estatístico . . . 88
5.5 Associando variáveis a propriedades visuais . . . 89
5.6 Escolhendo um gráfico . . . 91
5.7 Mostre os dados sempre que possível . . . 93
5.8 Representando a incerteza . . . 95
5.9 Cor em gráficos científicos . . . 96
5.10 Rótulos, escalas e eixos . . . 98
5.11 Múltiplas variáveis e múltiplos painéis . . . 100
5.12 Construindo gráficos com o ggplot2 . . . 101
5.13 Problemas comuns em figuras científicas . . . 107
5.14 Visualizações enganosas . . . 109
5.15 Elaborando figuras para publicações científicas . . . 110
5.16 Pesquisa na prática: aprimorando uma figura científica . . . 111
5.17 Um fluxo de trabalho para a visualização científica . . . 113
5.18 Lista de verificação da figura . . . 113
5.19 Resumo do capítulo . . . 114
5.20 Conceitos-chave . . . 114
5.21 Exercícios . . . 114

5.22 Reflexão 115
5.23 Próximos passos 116
6 Como tiramos conclusões científicas? 117
6.1 Objetivos de aprendizagem 117
6.2 Das observações à inferência 117
6.3 O que queremos estimar? 118
6.4 População e amostra 119
6.5 Por que as estimativas variam? 120
6.6 Estimação e incerteza 122
6.7 Tamanho de efeito 124
6.8 Hipóteses científicas e hipóteses estatísticas 125
6.9 Teste de hipóteses 126
6.10 Significância estatística versus importância biológica 127
6.11 Erros nas decisões estatísticas 129
6.12 Poder estatístico 129
6.13 Estimação ou teste? 130
6.14 Comparações múltiplas e testes múltiplos 131
6.15 Os pressupostos importam 132
6.16 Associação, predição e causalidade 133
6.17 Pesquisa na prática: interpretando o efeito de um tratamento 134
6.18 Dos resultados estatísticos às conclusões científicas 135
6.19 Um fluxo de trabalho para o raciocínio estatístico 137
6.20 Lista de verificação da inferência estatística 137
6.21 Resumo do capítulo 138
6.22 Conceitos-chave 138
6.23 Exercícios 139
6.24 Reflexão 140
6.25 Próximos passos 140
7 Como comparamos grupos? 141
7.1 Objetivos de aprendizagem 141
7.2 Por que comparar grupos? 141
7.3 Da questão biológica à comparação estatística 142
7.4 Olhando os dados antes de comparar médias 143
7.5 Comparando dois grupos independentes 145
7.6 Observações independentes versus pareadas 146
7.7 Comparando mais de dois grupos 148
7.8 A ANOVA como modelo linear 148
7.9 O que o teste global nos diz? 150
7.10 Contrastes: fazendo perguntas específicas 150
7.11 Comparações múltiplas 151
7.12 Das letras às estimativas 152
7.13 Experimentos fatoriais 153
7.14 Blocos e comparações entre grupos 156
7.15 Pressupostos do modelo 157
7.16 Diagnóstico do modelo 158
7.17 E se os pressupostos não forem atendidos? 160
7.18 Uma nota sobre os testes de normalidade 161
7.19 Médias marginais estimadas 162
7.20 Pesquisa na prática: avaliando tratamentos de manejo da doença 164
7.21 Um fluxo de trabalho prático para comparar grupos 165

7.22 Relatando comparações entre grupos . . . 165
7.23 Erros comuns . . . 166
7.24 Lista de verificação para comparações entre grupos . . . 167
7.25 Resumo do capítulo . . . 167
7.26 Conceitos-chave . . . 168
7.27 Exercícios . . . 168
7.28 Reflexão . . . 170
7.29 Próximos passos . . . 170

8 Como estudamos relações entre variáveis? . . . 171
8.1 Objetivos de aprendizagem . . . 171
8.2 Por que estudar relações? . . . 171
8.3 Da questão biológica à relação . . . 172
8.4 Olhando os dados antes de ajustar uma reta . . . 173
8.5 Correlação . . . 176
8.6 Regressão linear simples . . . 178
8.7 Incerteza nas estimativas da regressão . . . 181
8.8 Quanto da variação é explicado? . . . 182
8.9 Pressupostos e diagnóstico do modelo . . . 183
8.10 Quando a relação não é linear . . . 186
8.11 Extrapolação . . . 187
8.12 Comparando relações entre grupos . . . 187
8.13 Regressão linear múltipla . . . 189
8.14 Dados de vários experimentos ou locais . . . 191
8.15 Regressão ou correlação? . . . 191
8.16 Pesquisa na prática: estimando a perda de produtividade causada pela doença . . . 192
8.17 Um fluxo de trabalho prático para estudar relações . . . 193
8.18 Relatando relações . . . 193
8.19 Erros comuns . . . 194
8.20 Lista de verificação para a análise de relações . . . 195
8.21 Resumo do capítulo . . . 195
8.22 Conceitos-chave . . . 196
8.23 Exercícios . . . 196
8.24 Reflexão . . . 198
8.25 Próximos passos . . . 198

9 Como analisamos respostas não lineares? . . . 199
9.1 Objetivos de aprendizagem . . . 199
9.2 Por que respostas não lineares? . . . 199
9.3 Da questão biológica ao parâmetro da curva . . . 200
9.4 Por que não uma reta ou um polinômio? . . . 201
9.5 A anatomia de um modelo não linear . . . 201
9.6 Formas comuns de curvas . . . 202
9.7 Análise de dose–resposta . . . 203
9.8 Comparando curvas de dose–resposta . . . 206
9.9 Respostas à temperatura e a outras variáveis ambientais . . . 209
9.10 Como os modelos não lineares são ajustados . . . 211
9.11 Incerteza nos modelos não lineares . . . 212
9.12 Pressupostos e diagnóstico do modelo . . . 213
9.13 Escolhendo entre modelos . . . 214
9.14 Delineando experimentos para respostas não lineares . . . 215
9.15 Pesquisa na prática: monitorando a sensibilidade a fungicidas . . . 216

9.16 Um fluxo de trabalho prático para respostas não lineares 217
9.17 Relatando análises não lineares 217
9.18 Erros comuns 218
9.19 Lista de verificação para análises não lineares 219
9.20 Resumo do capítulo 219
9.21 Conceitos-chave 220
9.22 Exercícios 220
9.23 Reflexão 222
9.24 Próximos passos 222

10 Como analisamos dados categóricos, de contagem e de proporção? 223
10.1 Objetivos de aprendizagem 223
10.2 Por que esses dados são diferentes? 223
10.3 Da questão biológica ao estimando 224
10.4 Por que não usar o modelo linear usual? 225
10.5 Descrevendo dados categóricos 226
10.6 Testes de qui-quadrado e testes exatos 227
10.7 Modelos lineares generalizados 229
10.8 Respostas binárias e binomiais: regressão logística 230
10.9 Respostas de contagem: modelos do tipo Poisson 233
10.10 Superdispersão 234
10.11 Considerando o delineamento experimental 237
10.12 Proporções contínuas: severidade em porcentagem 237
10.13 Respostas ordinais: escalas de notas de doença 238
10.14 Escolhendo um modelo 239
10.15 Pesquisa na prática: avaliando um tratamento de sementes 239
10.16 Um fluxo de trabalho prático para dados categóricos, de contagem e de proporção 241
10.17 Relatando análises de dados categóricos, de contagem e de proporção 242
10.18 Erros comuns 243
10.19 Lista de verificação para dados categóricos, de contagem e de proporção 244
10.20 Resumo do capítulo 244
10.21 Conceitos-chave 244
10.22 Exercícios 245
10.23 Reflexão 246
10.24 Próximos passos 247

11 Como exploramos dados multivariados? 249
11.1 Objetivos de aprendizagem 249
11.2 Por que dados multivariados? 249
11.3 Da questão biológica à exploração multivariada 250
11.4 Organizando dados multivariados 251
11.5 Preparando os dados 252
11.6 Distância e similaridade 252
11.7 Análise de componentes principais 253
11.8 Agrupamento 258
11.9 Mapas de calor 260
11.10 Relações entre variáveis 262
11.11 Da exploração à inferência 263
11.12 Questões de delineamento em estudos multivariados 264
11.13 Pesquisa na prática: expressão de genes de defesa em cultivares resistentes e suscetíveis 265
11.14 Um fluxo de trabalho prático para dados multivariados 266

11.15 Relatando análises multivariadas 266
11.16 Erros comuns 267
11.17 Lista de verificação para análises multivariadas 268
11.18 Resumo do capítulo 269
11.19 Conceitos-chave 269
11.20 Exercícios 270
11.21 Reflexão 271
11.22 Próximos passos 271
12 Como os resultados científicos devem ser comunicados? 273
12.1 Objetivos de aprendizagem 273
12.2 Por que a comunicação importa 273
12.3 Do resultado estatístico à mensagem científica 274
12.4 Relatando os métodos estatísticos 274
12.5 Relatando os resultados 275
12.6 Tabelas, figuras ou texto? 276
12.7 Figuras de resultados 278
12.8 Escrevendo a interpretação 278
12.9 Relatórios reprodutíveis 279
12.10 Pesquisa na prática: redigindo os resultados de um ensaio de fungicidas 279
12.11 Erros comuns 280
12.12 Um fluxo de trabalho para comunicar resultados 281
12.13 Lista de verificação para a comunicação de resultados 281
12.14 Resumo do capítulo 282
12.15 Conceitos-chave 282
12.16 Exercícios 282
12.17 Reflexão 283
12.18 Próximos passos 283
13 Como tornar a pesquisa totalmente reprodutível? 285
13.1 Objetivos de aprendizagem 285
13.2 Por que a reprodutibilidade importa 285
13.3 Reprodutibilidade, replicabilidade e robustez 286
13.4 Organizando o projeto 286
13.5 Scripts em vez de cliques 287
13.6 Documentos dinâmicos com o Quarto 289
13.7 O ambiente computacional 290
13.8 Controle de versões com Git 291
13.9 Compartilhando dados e código 292
13.10 “Bom o suficiente” é melhor do que perfeito 292
13.11 Pesquisa na prática: tornando reprodutível um ensaio de fungicidas 293
13.12 Erros comuns 293
13.13 Um fluxo de trabalho para a pesquisa reprodutível 294
13.14 Lista de verificação da reprodutibilidade 294
13.15 Resumo do capítulo 294
13.16 Conceitos-chave 295
13.17 Exercícios 295
13.18 Reflexão 296
13.19 Considerações finais 296
Referências 297

Boas-vindas

A Fitopatologia é uma ciência cada vez mais quantitativa. Avanços nas técnicas experimentais, na biologia molecular, no monitoramento ambiental, no processamento de imagens e na ciência de dados transformaram a maneira como as doenças de plantas são investigadas, gerando conjuntos de dados maiores, mais diversos e mais complexos do que nunca. Ao mesmo tempo, espera-se que os pesquisadores não apenas analisem esses dados corretamente, mas também comuniquem seus resultados por meio de fluxos de trabalho científicos transparentes e reprodutíveis.

Este livro nasceu de notas de aula elaboradas para a disciplina de pós-graduação *Análise e Visualização de Dados em Fitopatologia*, oferecida no Programa de Pós-Graduação em Fitopatologia da Universidade Federal de Viçosa. Ao longo dos anos, a disciplina evoluiu a partir de discussões com estudantes de pós-graduação e pesquisadores de diversas áreas da Fitopatologia, que revelaram uma necessidade comum: um material que conectasse as questões biológicas ao raciocínio quantitativo.

Em vez de apresentar os métodos estatísticos como técnicas isoladas, este livro aborda a análise de dados como parte integrante do processo científico. A ênfase não está na escolha de um teste estatístico, mas na compreensão de como as questões biológicas orientam o delineamento do estudo, a coleta de dados, a visualização, a análise, a interpretação e a comunicação das evidências científicas. Ao longo do livro, análises reprodutíveis implementadas em R são usadas para ilustrar esses conceitos no contexto de problemas reais de pesquisa.

Embora os exemplos tenham como foco a Fitopatologia, os princípios aqui apresentados aplicam-se de modo geral à pesquisa experimental nas ciências biológicas. Minha esperança é que este livro ajude estudantes e pesquisadores a desenvolver não apenas habilidades analíticas práticas, mas também uma compreensão mais profunda do raciocínio científico que fundamenta uma pesquisa rigorosa e reprodutível.

CAPÍTULO 1

Por que pesquisar em Fitopatologia?

1.1 O que é Fitopatologia?

A Fitopatologia é a disciplina científica dedicada a compreender as doenças de plantas, incluindo suas causas, os mecanismos subjacentes às interações entre hospedeiros, patógenos e ambiente, os fatores que influenciam o desenvolvimento das doenças e os princípios que fundamentam o seu manejo eficaz (AGRIOS 2005). Como as doenças de plantas afetam a produção agrícola, a segurança alimentar, os ecossistemas naturais e o comércio internacional, a Fitopatologia integra conceitos e métodos de diversas áreas, como microbiologia, genética, biologia molecular, ecologia, agronomia e ciência de dados (Oliver 2024).

1.2 Que tipos de pesquisa são realizados em Fitopatologia?

A pesquisa em Fitopatologia abrange um amplo espectro de atividades científicas. As investigações podem descrever novas doenças ou patógenos, caracterizar mecanismos biológicos ou moleculares, avaliar estratégias de manejo, quantificar perdas de produção, investigar epidemias, desenvolver modelos de previsão ou sintetizar evidências de vários estudos. Esses objetivos diversos geram formas igualmente diversas de dados científicos e exigem abordagens analíticas adequadas tanto à questão de pesquisa quanto à natureza das observações coletadas.

Embora esses estudos difiram substancialmente em seus objetivos, todos compartilham uma meta comum: responder a questões biológicas por meio da coleta, análise e interpretação sistemáticas de evidências científicas. Seja o objetivo identificar um patógeno até então desconhecido, avaliar uma estratégia de manejo, compreender os fatores ambientais que impulsionam as epidemias ou investigar os mecanismos moleculares das interações patógeno–hospedeiro, o progresso científico depende de transformar observações em evidências confiáveis.

1.3 Onde a pesquisa em Fitopatologia é realizada?

A diversidade da pesquisa em Fitopatologia também se reflete nos ambientes em que os estudos são conduzidos. As investigações podem ser realizadas *in silico*, por meio de modelos computacionais e simulações; *in vitro*, sob condições controladas de laboratório; *in planta*, em câmaras de crescimento ou casas de vegetação; no campo, por meio de experimentos conduzidos sob condições agrícolas naturais; ou em escalas espaciais e temporais mais amplas,

incluindo estudos em nível de paisagem e de região que investigam a dinâmica das doenças em vários campos, ambientes ou safras. Outros estudos concentram-se nas interações patógeno–hospedeiro nos níveis celular, fisiológico ou molecular, enquanto o monitoramento de longo prazo fornece informações sobre o desenvolvimento das doenças ao longo do tempo (Figura 1.1). Cada um desses ambientes de pesquisa responde a questões biológicas distintas, opera em diferentes escalas espaciais e temporais e gera dados com características próprias. Em conjunto, eles oferecem perspectivas complementares que contribuem para uma compreensão mais abrangente das doenças de plantas.



Figura 1.1. Principais ambientes de pesquisa em Fitopatologia. Cada ambiente responde a questões de pesquisa distintas e gera formas complementares de evidência científica.

1.4 Estudos representativos

A amplitude da Fitopatologia talvez seja mais bem apreciada por meio de estudos representativos. Embora as investigações individuais difiram em seus objetivos, delineamentos e métodos analíticos, todas seguem o mesmo processo científico fundamental: uma questão biológica motiva o estudo, evidências são coletadas sistematicamente e conclusões são extraídas dos dados resultantes. Os quatro exemplos apresentados a seguir ilustram tipos comuns de pesquisa em Fitopatologia e a diversidade de questões científicas abordadas pelos pesquisadores. Juntos, eles cobrem boa parte da disciplina — da descoberta de novos patógenos e da avaliação de estratégias de manejo à compreensão das influências ambientais sobre o desenvolvimento das doenças e à investigação de mecanismos moleculares. Ao longo

deste livro, esses exemplos serão revisitados para mostrar como diferentes questões científicas levam naturalmente a diferentes delineamentos, tipos de dados e abordagens analíticas.

Pesquisa na prática 1: Descobrindo uma nova doença de planta

Questão científica

O que está causando uma doença até então não relatada que afeta lavouras de coentro no Brasil?

Contexto

Em 2021, produtores de coentro na Bahia observaram tombamento, podridão do colo e podridão de raízes generalizados, que causaram perdas severas e tornaram a produção comercial economicamente inviável nas áreas afetadas. Como a doença ainda não havia sido associada a esse patógeno no Brasil, pesquisadores conduziram um estudo para identificar o agente causal e determinar se se tratava de um novo relato da doença (Silva et al. 2024).

Ambiente do estudo

- 🌱 **No campo** — observação dos sintomas e coleta de plantas doentes
- 🫧 **In vitro** — isolamento do patógeno e caracterização morfológica
- 🧬 **Laboratório de biologia molecular** — sequenciamento de DNA e análises filogenéticas
- 🌿 **In planta** — testes de patogenicidade e confirmação dos postulados de Koch

Tipos de dados coletados

- Observações de campo
- Sintomas da doença
- Morfologia das colônias
- Imagens de microscopia
- Medidas morfológicas
- Sequências de DNA
- Árvores filogenéticas
- Resultados de testes de patogenicidade

Evidência científica

Os pesquisadores integraram várias fontes independentes de evidência para identificar o patógeno:

- observação dos sintomas da doença no campo;
- isolamento do patógeno suspeito a partir de tecidos doentes;
- caracterização morfológica das culturas e das estruturas reprodutivas;
- sequenciamento das regiões ITS e *cox2*, seguido de análises filogenéticas;
- testes de patogenicidade mostrando que as plantas inoculadas desenvolveram os mesmos sintomas, seguidos do reisolamento bem-sucedido do patógeno.

Conclusão principal

As evidências morfológicas, moleculares e de patogenicidade, em conjunto, demonstraram que *Pythium aphanidermatum* era o agente causal do tombamento, da podridão do colo e da podridão de raízes do coentro, constituindo o **primeiro relato dessa doença no Brasil**.

Reflita sobre este estudo

Antes de ler os próximos capítulos, considere as questões abaixo.

1. Qual era a questão biológica principal?
2. Quais ambientes de pesquisa estiveram envolvidos?
3. Quais observações eram descritivas?
4. Quais observações eram quantitativas?
5. Quais dados exigiram análise estatística?
6. Que tipos de dados foram gerados?
7. A conclusão poderia ter sido alcançada com uma única fonte de evidência?

Mensagem principal

Descobertas científicas nem sempre dependem de testes de hipóteses ou de significância estatística. Em muitas áreas da Fitopatologia, conclusões robustas emergem da integração de múltiplas linhas independentes de evidência científica.

Pesquisa na prática 2: Avaliando uma estratégia de manejo da doença

Questão científica

Aplicações de fungicidas reduzem de forma eficaz o míldio e protegem o manjericão cultivado em condições de campo?

Contexto

O míldio, causado por *Peronospora belbahrii*, é uma das doenças mais destrutivas do manjericão no mundo. Como as folhas comercializáveis precisam estar livres de sintomas visíveis, o controle eficaz da doença é essencial para a produção comercial. Pesquisadores conduziram uma série de experimentos de campo para avaliar a eficácia de diversos fungicidas sob pressão natural da doença (McGrath 2020).

Ambiente do estudo

🚜 Experimentos de campo conduzidos em condições de produção comercial.

Delineamento do estudo

- Cultivar de manjericão suscetível
- Tratamentos com fungicidas e uma testemunha sem tratamento
- Delineamento em blocos casualizados
- Quatro parcelas (repetições) por tratamento
- Doença avaliada periodicamente durante o ciclo da cultura

Tipos de dados coletados

- Incidência da doença
- Área abaixo da curva de progresso da doença (AACPD)
- Identificação do tratamento
- Bloco experimental

Evidência científica

A doença foi monitorada repetidamente ao longo da epidemia, o que permitiu aos pesquisadores comparar o desenvolvimento da doença entre os tratamentos com fungicidas e a testemunha. As médias dos tratamentos foram comparadas estatisticamente para determinar se as diferenças observadas refletiam efeitos reais dos tratamentos, e não a variabilidade experimental natural.

Principais resultados

As aplicações preventivas de fungicidas reduziram substancialmente a doença em comparação com as plantas não tratadas. Alguns programas de fungicidas proporcionaram, de forma consistente, excelente controle, enquanto outros foram consideravelmente menos eficazes. O estudo também demonstrou que manter intervalos regulares de aplicação e boa cobertura da pulverização foi fundamental para o sucesso do manejo.

Reflita sobre este estudo

1. Qual é a questão biológica abordada?
2. Qual é o fator de tratamento?
3. Qual é a unidade experimental?
4. Por que uma testemunha sem tratamento é necessária?
5. Qual variável resposta representa melhor a intensidade da doença?
6. Por que foram incluídas várias parcelas por tratamento?
7. Quais capítulos deste livro ajudarão a analisar esses dados?

Mensagem principal

Estudos de manejo de doenças estão entre as investigações experimentais mais comuns em Fitopatologia. Eles comparam tratamentos sob condições experimentais controladas para determinar se uma intervenção reduz a doença. Ao longo deste livro, experimentos semelhantes serão revisitados para ilustrar conceitos que vão do delineamento experimental e da organização dos dados à análise estatística, à visualização e ao relato reprodutível.

Pesquisa na prática 3: Compreendendo como as condições ambientais influenciam o desenvolvimento da doença

Questão científica

Como a temperatura e a duração do molhamento foliar influenciam a infecção de folhas de videira pelo agente da antracnose, Elsinoë ampelina? (Baseado em (Carisse et al. 2020))

Contexto

O sucesso da infecção por muitos patógenos de plantas depende das condições ambientais. Compreender como a doença responde a variações de temperatura e umidade ajuda a explicar o desenvolvimento das epidemias e fornece a base biológica para a previsão e o manejo de doenças.

Delineamento do estudo

Os pesquisadores inocularam plantas de videira com *E. ampelina* em ambiente controlado e as expuseram a combinações de seis temperaturas (5–30 °C) e seis durações de molhamento foliar (0–24 h). A severidade da doença foi avaliada cerca de uma semana após a inoculação.

Dados coletados

- Temperatura (contínua)
- Duração do molhamento foliar (contínua)
- Severidade da doença (% da área foliar doente)

Principais resultados

Não houve doença na ausência de molhamento foliar nem a 5 °C. A severidade aumentou com a duração do molhamento, com a resposta mais intensa em torno de 25 °C. Esses resultados demonstraram que o desenvolvimento da doença depende da interação entre as condições ambientais, e não de um único fator.

Reflita sobre este estudo

- Qual é a variável resposta?
- Quais variáveis são explicativas?
- Por que a temperatura e o molhamento foliar são medidos em escalas contínuas?
- Que tipo de gráfico resumiria melhor essas relações?
- Um modelo linear descreveria adequadamente a resposta biológica?

Pesquisa na prática 4: Investigando respostas fisiológicas e moleculares à infecção por patógenos

Compreender como as plantas respondem à infecção por patógenos frequentemente exige medir processos que não podem ser observados diretamente. Em vez de se concentrarem apenas na severidade da doença, os pesquisadores podem investigar alterações na fisiologia, no metabolismo e na expressão gênica da planta para compreender os mecanismos subjacentes à suscetibilidade ou à resistência.

Um exemplo é o estudo de Silva et al. (Silva et al. 2023), que examinou como a infecção por *Septoria lycopersici* afetava a fisiologia de folhas de tomateiro. Plantas de tomate foram inoculadas com o patógeno ou mantidas sadias como controle, sob condições ambientais controladas. Durante o desenvolvimento da doença, os pesquisadores mediram diversas variáveis fisiológicas, bioquímicas e moleculares para determinar como a infecção alterava o metabolismo da planta.

Questão de pesquisa

Como a infecção por *Septoria lycopersici* altera as respostas fisiológicas e moleculares do tomateiro?

Delineamento do estudo

- Experimento em casa de vegetação, em condições controladas
- Dois tratamentos:
  - Plantas sadias
  - Plantas inoculadas
- Medições realizadas repetidamente durante o desenvolvimento da doença

Exemplos de dados coletados

- Taxa fotossintética líquida
- Fluorescência da clorofila
- Concentração de pigmentos fotossintéticos
- Concentração de peróxido de hidrogênio
- Peroxidação de lipídios
- Expressão relativa de genes relacionados à defesa e à fotorrespiração

O que o estudo concluiu?

A infecção comprometeu substancialmente a fotossíntese, alterou o metabolismo oxidativo e modificou a expressão de diversos genes associados à fotorrespiração e às respostas ao

estresse. Em conjunto, essas medidas mostraram que a infecção afeta múltiplos processos fisiológicos além dos sintomas visíveis da doença.

Por que este estudo é importante para este livro?

Este exemplo mostra que uma única questão biológica pode exigir a coleta de muitos tipos diferentes de dados quantitativos. Embora cada variável seja medida de forma diferente, todas contribuem com evidências complementares para responder à mesma questão de pesquisa. Capítulos posteriores revisitarão conjuntos de dados semelhantes para explorar medidas repetidas, visualização, análise multivariada e a interpretação de múltiplas variáveis resposta.

Os quatro exemplos apresentados acima ilustram a notável diversidade da pesquisa científica em Fitopatologia. Embora cada estudo aborde uma questão biológica diferente e empregue abordagens experimentais distintas, todos seguem a mesma lógica científica: uma questão de pesquisa motiva o estudo, evidências são coletadas sistematicamente e conclusões são extraídas dos dados resultantes. Mais importante, esses exemplos mostram que diferentes objetivos de pesquisa geram naturalmente diferentes tipos de dados.

Um estudo que descreve um novo patógeno baseia-se principalmente em observações qualitativas, microscopia, sequências de DNA e testes de patogenicidade, ao passo que um ensaio de manejo gera comparações entre tratamentos e avaliações de doença. Estudos sobre efeitos ambientais frequentemente combinam medidas de doença com variáveis ambientais contínuas, enquanto investigações fisiológicas e moleculares podem quantificar simultaneamente respostas bioquímicas, fisiológicas e de expressão gênica. Reconhecer essas diferenças é o primeiro passo para escolher métodos adequados de organização, visualização, análise estatística e interpretação dos dados. Os exemplos resumidos na Tabela 1.1 servirão como estudos de caso recorrentes ao longo do livro, à medida que novos conceitos analíticos forem apresentados.

Tabela 1.1. Exemplos representativos de pesquisa em Fitopatologia, ilustrando como diferentes questões científicas originam diferentes tipos de dados.

Exemplo	Objetivo da pesquisa	Exemplos de dados coletados
1	Descrever uma nova doença ou patógeno	Sintomas, morfologia, sequências de DNA, testes de patogenicidade
2	Avaliar uma estratégia de manejo	Incidência, severidade, produtividade
3	Compreender efeitos ambientais sobre o desenvolvimento da doença	Severidade, temperatura, duração do molhamento foliar
4	Compreender respostas fisiológicas e moleculares da planta	Fotossíntese, pigmentos, atividade enzimática, expressão gênica

1.5 Que tipos de dados os fitopatologistas coletam?

A diversidade de questões científicas encontradas na Fitopatologia dá origem, naturalmente, a uma coleção igualmente diversa de dados. Dependendo do estudo, os pesquisadores podem coletar observações de campo, notas de doença, medidas de incidência e severidade, contagens de lesões, estimativas de produtividade, registros ambientais, imagens de microscopia, sequências de DNA, medidas de expressão gênica, perfis de microbioma, estimativas de

populações de patógenos, dados metabolômicos e muitas outras formas de informação biológica. Alguns conjuntos de dados são pequenos e simples; outros são grandes, complexos e de alta dimensionalidade.

Apesar dessa diversidade, todo conjunto de dados tem o mesmo propósito final: fornecer evidências para responder a uma questão científica. Antes que conclusões significativas possam ser tiradas, no entanto, os dados precisam ser cuidadosamente organizados, explorados, visualizados, analisados, interpretados e comunicados.

1.6 Como esses dados devem ser analisados?

Coletar dados é apenas uma etapa do processo científico. Independentemente da questão de pesquisa ou do tipo de estudo, observações científicas não se tornam evidências simplesmente por terem sido medidas. Seu valor depende de como são organizadas, exploradas, analisadas, interpretadas e comunicadas. Conjuntos de dados mal organizados, métodos estatísticos inadequados, visualizações enganosas ou interpretações sem fundamento podem comprometer estudos que, de outra forma, seriam bem delineados. Por outro lado, uma análise cuidadosa permite distinguir padrões biológicos relevantes da variabilidade natural e tirar conclusões que sejam ao mesmo tempo estatisticamente sólidas e biologicamente relevantes.

A análise de dados deve, portanto, ser vista como um fluxo de trabalho completo, e não como a aplicação de técnicas estatísticas ao final de um experimento. Ela começa muito antes do primeiro teste estatístico, com decisões sobre como os dados são registrados, documentados e organizados.

Análises exploratórias e representações gráficas ajudam os pesquisadores a compreender a estrutura e a qualidade de seus dados, identificar possíveis problemas e gerar novas hipóteses. As análises estatísticas formais fornecem, então, ferramentas quantitativas para avaliar a incerteza, comparar tratamentos, estimar relações entre variáveis e testar hipóteses científicas. Por fim, os resultados devem ser comunicados com clareza por meio de figuras, tabelas e interpretação escrita, para que outros pesquisadores possam compreendê-los, reproduzi-los e avançar a partir deles. Essa perspectiva reflete a filosofia central deste livro.

> **! Importante**
>
> Os métodos estatísticos são apresentados não como procedimentos matemáticos isolados, mas como ferramentas para responder a questões biológicas.

A escolha de um método analítico deve sempre ser orientada pela questão de pesquisa, pelo delineamento do estudo e pelo tipo de dado coletado. Consequentemente, compreender os dados costuma ser mais importante do que conhecer um grande número de técnicas estatísticas.

Ao longo deste livro, o leitor aprenderá a abordar a análise de dados como parte integrante do raciocínio científico. Partindo de uma questão biológica, desenvolverá progressivamente as habilidades necessárias para organizar dados, explorar padrões, visualizar resultados, realizar análises estatísticas adequadas e comunicar evidências científicas por meio de fluxos de trabalho reprodutíveis implementados em R.

Os exemplos apresentados neste capítulo mostram que a pesquisa em Fitopatologia abrange questões científicas, delineamentos e tipos de dados diversos. Ainda assim, independentemente de sua complexidade, todos os estudos enfrentam o mesmo desafio fundamental: transformar

observações em evidências científicas confiáveis. Esse desafio é a motivação para o restante deste livro.

1.7 Como este livro está organizado

Este livro segue a progressão natural de um projeto de pesquisa científica. Em vez de apresentar os métodos estatísticos como uma coleção isolada de técnicas, cada capítulo se apoia no anterior, conduzindo o leitor da formulação de uma questão de pesquisa à comunicação de evidências científicas reprodutíveis (Figura 1.2).



Figura 1.2. Organização do livro. Cada etapa de um projeto de pesquisa corresponde a um ou mais capítulos.

Capítulo 1 – Por que fazemos pesquisa científica?

O livro começa discutindo a Fitopatologia como disciplina científica, o papel da investigação científica e como as questões de pesquisa dão origem à coleta e à análise de dados.

Capítulo 2 – Como delinear um estudo?

Conclusões científicas confiáveis dependem de estudos bem delineados. Este capítulo apresenta os princípios da pesquisa observacional e experimental, o delineamento de estudos, unidades experimentais, variáveis resposta, repetição e as armadilhas mais comuns na pesquisa em Fitopatologia.

Capítulo 3 – Como os dados devem ser organizados?

Antes de qualquer análise, os dados precisam ser cuidadosamente estruturados, documentados e gerenciados. Este capítulo apresenta os princípios de dados organizados (*tidy data*), metadados, qualidade e gestão de dados para a pesquisa reprodutível.

Capítulo 4 – O que os dados revelam?

A análise exploratória de dados é a primeira oportunidade de compreender as características de um conjunto de dados. O leitor aprende a resumir variáveis, identificar observações incomuns e explorar padrões antes das análises estatísticas formais.

Capítulo 5 – Como visualizar dados de forma eficaz?

Figuras científicas são ferramentas essenciais para compreender e comunicar resultados. Este capítulo apresenta os princípios da visualização eficaz de dados e mostra como gráficos adequados revelam padrões que os resumos numéricos sozinhos podem não mostrar.

Capítulo 6 – Como tiramos conclusões científicas?

Este capítulo apresenta os fundamentos da inferência estatística, incluindo hipóteses científicas e estatísticas, incerteza, intervalos de confiança, tamanhos de efeito, testes de significância e os princípios que orientam a escolha de métodos estatísticos adequados.

Capítulos 7–11 – Como analisar diferentes tipos de dados?

Em vez de organizar os métodos estatísticos por complexidade matemática, estes capítulos são organizados de acordo com o tipo de questão biológica e a natureza da variável resposta mais comuns na pesquisa em Fitopatologia. A sequência vai das comparações entre grupos às relações entre variáveis, às respostas não lineares, aos dados não gaussianos e, por fim, aos dados complexos e multivariados:

- **Capítulo 7 – Como comparamos grupos?** Respostas contínuas comparadas entre tratamentos ou grupos: comparação de dois grupos, testes *t*, análise de variância, comparações múltiplas e tamanhos de efeito.
- **Capítulo 8 – Como estudamos relações entre variáveis?** Correlação, regressão linear simples e múltipla e a interpretação biológica das inclinações.
- **Capítulo 9 – Como analisamos respostas não lineares?** Curvas de dose–resposta, regressão não linear, CE_{50}/DE_{50} e parâmetros de curvas com significado biológico.

- **Capítulo 10 – Como analisamos dados categóricos, de contagem e de proporção?** Testes de qui-quadrado, respostas binárias, proporções, contagens e modelos logísticos e do tipo Poisson.
- **Capítulo 11 – Como exploramos dados multivariados?** Análise de componentes principais, agrupamento e mapas de calor, aplicados a dados de expressão gênica e outros conjuntos de dados de alta dimensionalidade.

Essa organização reflete a maneira como os pesquisadores costumam abordar a análise de dados: considerando primeiro a questão biológica e o tipo de dado, antes de escolher um método analítico adequado.

Capítulo 12 – Como os resultados científicos devem ser comunicados?

Análises científicas só cumprem seu propósito quando seus resultados são comunicados com clareza. Este capítulo discute a preparação de tabelas, figuras, resumos estatísticos e interpretações escritas adequadas para publicações científicas.

Capítulo 13 – Como tornar a pesquisa totalmente reprodutível?

O capítulo final integra todos os conceitos anteriores em um fluxo de trabalho de pesquisa reprodutível. O leitor aprende a organizar projetos, documentar análises, gerar relatórios dinâmicos com o Quarto e compartilhar dados e códigos de acordo com os princípios atuais da ciência aberta e reprodutível.

Em conjunto, esses capítulos refletem uma filosofia central deste livro: a análise de dados não é uma etapa isolada da pesquisa científica, mas um processo que começa com uma questão de pesquisa bem formulada e termina com evidências científicas transparentes, reprodutíveis e comunicadas de forma eficaz.

1.8 Resumo do capítulo

Neste capítulo, você aprendeu que:

- A Fitopatologia aborda uma ampla diversidade de questões biológicas.
- A pesquisa pode ser conduzida em múltiplos ambientes.
- Diferentes questões geram diferentes tipos de dados.
- A análise de dados é parte integrante da pesquisa científica.
- Os métodos científicos devem ser escolhidos de acordo com a questão de pesquisa e a natureza dos dados.

1.9 Conceitos-chave

Ao concluir este capítulo, você deverá ser capaz de:

- Descrever o escopo da Fitopatologia como disciplina científica.
- Reconhecer diferentes tipos de pesquisa realizados em Fitopatologia.
- Identificar ambientes comuns de pesquisa, do laboratório ao campo e aos estudos computacionais.

- Explicar como as questões científicas determinam o tipo de evidência coletada.
- Reconhecer a diversidade de dados gerados pela pesquisa em Fitopatologia.
- Explicar por que a análise de dados é um componente essencial da investigação científica.

1.10 Exercícios

1.10.1 Exercício 1. Explorando um estudo publicado

Escolha um dos quatro exemplos de pesquisa apresentados neste capítulo e responda às questões a seguir.

1. Que questão biológica os pesquisadores buscaram responder?
2. Por que essa questão é importante?
3. Em que ambiente(s) de pesquisa o estudo foi conduzido?
4. O que os pesquisadores fizeram para responder à questão?
5. Que tipos de evidência coletaram?
6. Qual foi a conclusão principal?

1.10.2 Exercício 2. Comparando estudos científicos

Compare os quatro estudos representativos apresentados neste capítulo.

1. Qual estudo teve como foco a descoberta de um novo conhecimento?
2. Qual estudo avaliou uma estratégia de manejo da doença?
3. Qual estudo investigou a influência das condições ambientais?
4. Qual estudo buscou compreender mecanismos biológicos?
5. Em que esses estudos diferem quanto aos objetivos?
6. O que eles têm em comum?

1.10.3 Exercício 3. Lendo um artigo científico

Em pequenos grupos, analisem um artigo científico fornecido pelo professor e preparem uma apresentação de cinco minutos abordando as questões abaixo.

- Que questão científica motivou o estudo?
- Por que a questão é importante?
- Onde o estudo foi conduzido?
- O que os pesquisadores fizeram?
- Que evidências foram coletadas?
- A que conclusões chegaram?

Não descrevam as análises estatísticas. Em vez disso, concentrem-se em compreender a lógica científica do estudo.

1.10.4 Exercício 4. Pensando como um cientista

Escolha um dos estudos discutidos neste capítulo.

1. Que nova questão de pesquisa esse estudo levanta?
2. Se você fosse continuar essa pesquisa, o que investigaria no próximo estudo?
3. Que evidências adicionais fortaleceriam as conclusões?

1.11 Reflexão

A pesquisa científica muitas vezes começa com a curiosidade, e não com a metodologia. Ao longo deste capítulo, você viu exemplos de estudos que diferem muito em seus objetivos e abordagens, mas que contribuem todos para ampliar nossa compreensão das doenças de plantas.

Reflita sobre a questão a seguir e discuta-a com seus colegas.

> **O que torna uma evidência científica convincente?**

É a sofisticação dos métodos, a quantidade de dados coletados, a análise estatística realizada ou a qualidade da questão científica e do delineamento do estudo?

Não existe uma única resposta correta. O objetivo desta discussão é reconhecer que a ciência rigorosa depende da integração de todos esses elementos, e não de um único componente isolado.

1.12 Próximos passos

A diversidade de estudos apresentada neste capítulo mostra que a pesquisa científica em Fitopatologia não pode ser reduzida a uma única metodologia. Os pesquisadores escolhem delineamentos, medidas e abordagens analíticas de acordo com a questão biológica que buscam responder. No próximo capítulo, examinaremos como os estudos científicos são delineados e por que as decisões tomadas antes da coleta de dados costumam ter a maior influência sobre a confiabilidade das conclusões que podem ser tiradas.

CAPÍTULO 2

Como delinear um estudo?

2.1 Objetivos de aprendizagem

! Importante

Ao concluir este capítulo, o leitor deverá ser capaz de:

- explicar por que o delineamento do estudo determina a força da evidência científica;
- distinguir estudos observacionais de estudos experimentais;
- traduzir uma questão biológica em uma hipótese científica e em predições testáveis;
- identificar fatores, tratamentos, variáveis resposta, unidades experimentais e unidades amostrais;
- explicar os papéis da repetição, da casualização e do bloqueamento;
- distinguir a repetição dentro de um experimento da repetição do experimento;
- reconhecer pseudorrepetição, confundimento, ausência de testemunhas e outros problemas comuns de delineamento;
- avaliar criticamente o delineamento de estudos em Fitopatologia.

2.2 Por que o delineamento importa

A qualidade da evidência científica é determinada muito antes do início da análise de dados. Ela é determinada quando a questão biológica é formulada, quando os tratamentos ou grupos são definidos, quando as unidades experimentais são escolhidas, quando a repetição e a casualização são planejadas e quando as medidas são selecionadas. A análise estatística é essencial para quantificar padrões e incertezas, mas não consegue criar evidência confiável se o estudo não foi delineado para produzi-la.

Esse ponto é particularmente importante na Fitopatologia, porque muitos estudos envolvem sistemas biológicos complexos. O desenvolvimento da doença depende do hospedeiro, do patógeno, do ambiente e das práticas de manejo. Esses fatores variam entre plantas, parcelas, lavouras, casas de vegetação, laboratórios, safras e regiões. Se essa variação não for considerada no delineamento, pode se tornar impossível determinar se as diferenças observadas são causadas pelo fator de interesse ou por fontes de variação não controladas.

Por exemplo, suponha que um pesquisador queira avaliar se um fungicida reduz a severidade da doença. Se as parcelas tratadas forem colocadas em uma parte da lavoura e as não tratadas em outra, o efeito do tratamento pode ficar confundido com um gradiente de solo, diferenças de umidade, pressão de inóculo ou desenvolvimento do dossel. Mesmo que a análise produza um

valor-P pequeno, a conclusão continua fraca, porque o delineamento não permite separar o efeito do fungicida de outras diferenças no campo.

Da mesma forma, em um experimento em casa de vegetação, se todas as plantas de um tratamento forem colocadas em uma bancada e todas as de outro tratamento em outra bancada, os efeitos dos tratamentos podem ficar confundidos com diferenças de luz, temperatura, umidade ou circulação de ar entre as bancadas. Em um experimento em câmara de crescimento, se cada temperatura for representada por uma única câmara, o efeito da temperatura não pode ser separado do efeito da câmara. Em um experimento de laboratório, se muitos esporos ou colônias forem medidos na mesma placa de Petri, essas medidas não representam necessariamente repetições independentes, se a placa for a unidade à qual o tratamento foi aplicado.

Esses exemplos ilustram um princípio central: a análise estatística não compensa um estudo mal delineado. Um modelo mais complexo, uma transformação diferente ou um teste mais sofisticado não conseguem reparar completamente problemas como confundimento, ausência de repetição verdadeira, unidades experimentais mal definidas, alocação tendenciosa dos tratamentos ou medidas inadequadas. Nesses casos, a fragilidade não é apenas analítica; ela está embutida na forma como a evidência foi produzida.

> **! Importante**
>
> A análise estatística pode quantificar a incerteza, mas não pode salvar um estudo sem delineamento adequado, sem unidades experimentais válidas ou sem repetição verdadeira.

2.3 Da questão biológica ao estudo

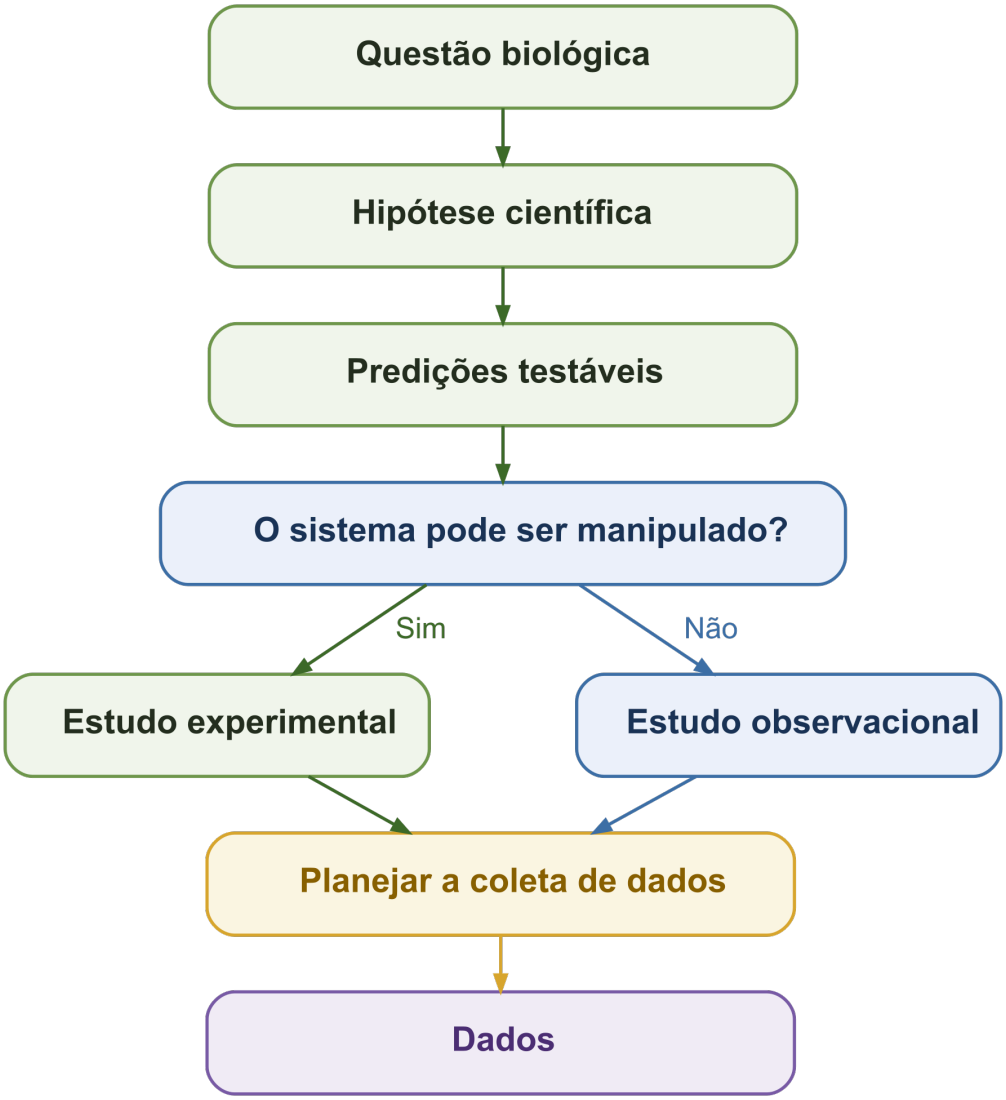


Figura 2.1. Da questão biológica à evidência científica.

2.3.1 Questões biológicas

Uma questão biológica expressa o que o pesquisador quer compreender sobre um patossistema. Ela pode surgir de observações de campo, de estudos anteriores, de resultados experimentais inesperados ou de problemas práticos de manejo. Nessa etapa, a questão deve ser formulada em termos biológicos, e não estatísticos.

Por exemplo, um pesquisador pode perguntar:

- Por que a doença causa mais danos em algumas lavouras do que em outras?
- Um fungicida reduz a severidade da doença?
- Uma prática cultural reduz o inóculo inicial?
- Algumas cultivares são menos suscetíveis do que outras?
- A temperatura influencia a infecção?
- A infecção pelo patógeno altera a fisiologia da planta?

Essas questões diferem em escopo, mas todas identificam um fenômeno biológico que exige explicação ou avaliação. Algumas tratam de causas; outras, de comparações, relações,

mecanismos ou consequências. Uma boa questão biológica deve ser ampla o suficiente para ter relevância científica, mas específica o suficiente para orientar o estudo.

Por exemplo, a questão: **O manejo da doença funciona?** é ampla demais. Uma versão mais útil seria: **A aplicação de fungicida reduz a incidência de brusone do trigo em condições de campo?** A questão revisada identifica a intervenção, a doença, a resposta de interesse e o contexto do estudo. Ela também começa a sugerir o tipo de estudo necessário.

Da mesma forma, a questão: **O clima afeta a doença?** é biologicamente relevante, mas genérica demais. Uma questão mais precisa seria: **Como a temperatura e a duração do molhamento foliar influenciam a infecção por um patógeno foliar?** Essa versão identifica duas variáveis ambientais e as relaciona a um processo biológico específico: a infecção.

> **Dica**
>
> Uma questão biológica útil deixa claro qual sistema está sendo estudado, qual resposta interessa e qual comparação, relação ou mecanismo motiva o estudo.

2.3.2 Hipóteses científicas

Uma hipótese científica é uma expectativa ou explicação com fundamento biológico. Ela propõe uma resposta possível à questão biológica e fornece o raciocínio que liga a questão às observações esperadas.

Questão biológica	Hipótese científica
Um fungicida reduz a severidade da doença?	Plantas tratadas com fungicida desenvolverão menos doença porque o produto suprime a infecção ou a colonização pelo patógeno.
As cultivares diferem em resistência?	As cultivares diferem em resistência porque variam em características que limitam o estabelecimento do patógeno, a colonização ou o desenvolvimento de sintomas.
A temperatura influencia a infecção?	A eficiência de infecção aumenta dentro de uma faixa favorável de temperatura e diminui quando a temperatura está baixa ou alta demais para o desenvolvimento do patógeno.
A infecção altera a fisiologia da planta?	Plantas infectadas apresentam menor atividade fotossintética porque a colonização pelo patógeno danifica o tecido foliar e perturba o metabolismo do hospedeiro.

Uma hipótese científica é diferente de uma hipótese estatística. A hipótese científica é expressa em linguagem biológica e está ligada a mecanismos, processos ou efeitos esperados. Uma hipótese estatística, como a hipótese nula, é um enunciado analítico formal usado mais tarde para avaliar a evidência dos dados.

Por exemplo, a hipótese científica: **A aplicação de fungicida reduz a severidade da doença.** pode ser representada estatisticamente como uma comparação entre a severidade média das parcelas tratadas e não tratadas. No entanto, a hipótese estatística não é a ideia científica em si. É apenas uma forma de avaliar a evidência relacionada a essa ideia.

Vale notar que as hipóteses científicas pertencem ao raciocínio biológico do estudo, enquanto as hipóteses estatísticas pertencem ao arcabouço analítico usado para avaliar a evidência. Essa distinção é importante porque um estudo pode produzir evidência biológica útil mesmo quando o teste estatístico não é o componente central da investigação. No primeiro relato de uma doença, por exemplo, a hipótese diz respeito à identidade e à patogenicidade do organismo causal. A evidência pode vir de sintomas, isolamento, morfologia, sequências de DNA, análise filogenética,

testes de patogenicidade e reisolamento. A lógica é biológica, mesmo quando alguns componentes são quantitativos.

Falseabilidade

Uma característica importante de uma boa hipótese é a falseabilidade. Uma hipótese é falseável quando faz predições que, em princípio, poderiam ser contrariadas pelas evidências. Por exemplo, a hipótese de que "um fungicida reduz a severidade da doença em condições de campo" pode ser testada comparando parcelas tratadas e não tratadas. Se as parcelas tratadas não apresentarem menor severidade em condições experimentais adequadas, a hipótese não é sustentada. Falseabilidade não significa que a hipótese seja falsa; significa que a hipótese é científica porque se expõe à possível contradição pelos dados. Nesse sentido, hipóteses não são meras declarações de crença. São expectativas estruturadas que ajudam os pesquisadores a delinear estudos, coletar dados relevantes e avaliar se as observações sustentam ou contestam uma explicação biológica.

2.3.3 Predições testáveis

Uma hipótese torna-se útil para delinear um estudo quando é traduzida em predições: enunciados sobre o que deveria ser observado, sob condições especificadas, se a hipótese estiver correta. As predições ligam a explicação geral aos dados específicos que o estudo vai produzir.

Por exemplo, a hipótese de que um fungicida reduz a doença porque suprime a infecção leva a predições como:

- parcelas pulverizadas com o fungicida terão menor severidade do que parcelas não tratadas nas mesmas condições;
- a diferença será maior quando a pressão da doença for alta do que quando for baixa;
- o efeito será mais forte quando o fungicida for aplicado antes da infecção do que após o aparecimento dos sintomas.

Cada predição especifica uma comparação e as condições em que ela deve se verificar. A primeira sugere um ensaio de campo com parcelas tratadas e não tratadas; a segunda, repetir o ensaio em ambientes com diferentes pressões de doença; a terceira, comparar épocas de aplicação. Assim, as predições orientam a escolha de tratamentos, testemunhas, condições e medidas.

Uma predição útil também especifica o que contaria como evidência contra a hipótese. Se parcelas tratadas e não tratadas apresentarem severidade semelhante em uma safra de alta pressão de doença e com aplicação adequada, a hipótese não é sustentada. Pensar com antecedência nos resultados que contradiriam a hipótese ajuda a garantir que o estudo seja capaz de produzir evidência informativa em qualquer direção.

2.3.4 Definições operacionais

Muitos conceitos em Fitopatologia são abstratos: resistência, agressividade, controle da doença, infecção, tolerância ou suscetibilidade. Para serem estudados, precisam ser traduzidos em grandezas mensuráveis. Uma definição operacional estabelece exatamente como um conceito será observado ou medido em um estudo específico.

Tabela 2.1. Exemplos de definições operacionais de conceitos comuns em Fitopatologia.

Conceito	Definições operacionais possíveis
Controle da doença	Redução da severidade (% da área foliar afetada) em relação à testemunha na última avaliação
Resistência	Número de lesões por cm² de folha, período latente ou severidade final após inoculação padronizada
Agressividade	Taxa de expansão das lesões (mm/dia) em uma cultivar suscetível
Infecção	Presença de sintomas 14 dias após a inoculação, ou DNA do patógeno detectado por qPCR
Perda de produtividade	Diferença de produtividade comercial (kg/ha) entre parcelas protegidas e não protegidas

Diferentes definições operacionais do mesmo conceito podem levar a conclusões diferentes. Uma cultivar pode ser resistente segundo o número de lesões, mas não segundo o tamanho das lesões; um fungicida pode reduzir a severidade sem aumentar a produtividade. As definições operacionais devem, portanto, ser escolhidas para representar a questão biológica da forma mais fiel possível e ser declaradas explicitamente, para que os leitores possam avaliá-las e reproduzi--las.

2.4 Estudos observacionais e experimentais

2.4.1 Estudos observacionais

Em um estudo observacional, o pesquisador observa e registra as condições como elas ocorrem, sem atribuí-las deliberadamente às unidades estudadas. O pesquisador pode escolher onde, quando e o que observar, mas não controla quais unidades são expostas a quais condições.

Estudos observacionais são comuns e valiosos em Fitopatologia. Exemplos incluem levantamentos que registram a prevalência de uma doença em lavouras e regiões, programas de monitoramento regional que acompanham a doença ao longo de vários anos, estudos que relacionam variáveis meteorológicas à intensidade da doença, levantamentos de populações de patógenos e de sua distribuição, questionários sobre práticas de manejo dos produtores e análises retrospectivas de bases de dados acumuladas por serviços de extensão ou programas de melhoramento.

Esses estudos descrevem o mundo como ele é, muitas vezes em grandes áreas, ao longo de longos períodos e em condições reais de produção. Sua principal limitação é que as condições de interesse não são atribuídas pelo pesquisador. Lavouras pulverizadas com mais frequência também podem diferir em cultivar, data de semeadura, intensidade de manejo e localização, de modo que as diferenças na doença não podem ser atribuídas com segurança apenas às pulverizações.

2.4.2 Estudos experimentais

Em um estudo experimental, o pesquisador impõe deliberadamente condições, chamadas tratamentos, às unidades experimentais e observa a resposta. Idealmente, os tratamentos são atribuídos por sorteio, de modo que as unidades que recebem tratamentos diferentes sejam comparáveis em todos os outros aspectos.

Estudos experimentais em Fitopatologia incluem ensaios de eficácia de fungicidas, comparações de cultivares sob inoculação natural ou artificial, inoculações controladas com diferentes isolados

ou concentrações de inóculo, experimentos em câmara de crescimento manipulando temperatura e molhamento foliar e estudos em casa de vegetação testando o efeito de nutrição, irrigação ou agentes de controle biológico sobre a doença.

Como o pesquisador controla os tratamentos, os experimentos permitem separar o efeito do tratamento de outras fontes de variação. Sua principal limitação é que as condições controladas podem não representar a diversidade de condições da produção comercial. Um ranqueamento de cultivares obtido em câmara de crescimento com um único isolado pode não se confirmar no campo, sob inóculo natural.

2.4.3 O que cada tipo de estudo permite concluir

Estudos observacionais são adequados para descrever padrões, estimar a frequência de uma doença, identificar associações, gerar hipóteses e estudar fenômenos que não podem ser manipulados, como o clima regional ou tendências de longo prazo. Associações encontradas em estudos observacionais podem refletir efeitos causais, mas também podem refletir variáveis de confundimento que influenciam tanto a suposta causa quanto a resposta.

Estudos experimentais são mais adequados para avaliar efeitos causais de intervenções, porque a alocação aleatória dos tratamentos torna improvável que outras diferenças entre os grupos expliquem os resultados. Suas conclusões, porém, aplicam-se mais diretamente às condições em que o experimento foi conduzido. Generalizar para outros ambientes, safras ou sistemas de produção exige repetir o experimento nessas condições ou combinar evidências experimentais e observacionais.

Na prática, as duas abordagens se complementam. Um levantamento pode revelar que a doença é mais severa em lavouras semeadas mais cedo; um experimento pode então testar se a data de semeadura afeta a doença em comparações controladas; e novos levantamentos podem avaliar se a mudança recomendada na data de semeadura reduz a doença em uma região.

Tabela 2.2. Comparação geral entre estudos observacionais e experimentais.

Característica	Estudo observacional	Estudo experimental
O pesquisador manipula as condições	Não	Sim
Os tratamentos são impostos deliberadamente	Não	Geralmente
Principal ponto forte	Descreve padrões e associações	Avalia efeitos em comparações controladas
Exemplos comuns em Fitopatologia	Levantamentos, monitoramento, bases de dados regionais	Ensaios de campo, ensaios em casa de vegetação, experimentos de laboratório
Principal limitação	Confundimento e inferência causal limitada	Realismo ou capacidade de generalização podem ser limitados

2.5 Os componentes básicos de um estudo

2.5.1 Fatores e tratamentos

Um **fator** é uma variável cujo efeito sobre a resposta está sendo estudado. Cada fator tem **níveis**, que são os valores ou categorias específicos comparados. Em um ensaio de fungicidas, o fator é

o programa de fungicida, e os níveis podem ser a testemunha e três produtos. Em um estudo em câmara de crescimento, a temperatura é um fator com níveis como 15, 20, 25 e 30 °C.

Um **tratamento** é a condição específica aplicada a uma unidade experimental. Quando o estudo tem um único fator, cada nível é um tratamento. Quando tem vários fatores, cada tratamento é uma combinação de níveis: um experimento com três cultivares e dois programas de fungicida tem seis tratamentos. Exemplos de fatores em Fitopatologia incluem fungicidas e suas doses ou épocas de aplicação, cultivares, inoculação (inoculado ou não inoculado), isolados, concentração de inóculo, temperatura, duração do molhamento e regimes de nutrição ou irrigação.

Os tratamentos devem ser definidos com precisão: produto, dose, época, método de aplicação e número de aplicações, no caso de um fungicida; isolado, concentração de inóculo e método, no caso de uma inoculação. Tratamentos definidos de forma imprecisa não podem ser reproduzidos e tornam os resultados difíceis de interpretar.

2.5.2 Variáveis resposta

Uma variável resposta é a grandeza medida para avaliar o efeito dos fatores. Incidência, severidade, número de lesões, tamanho das lesões, período latente, esporulação, biomassa do patógeno, produtividade e variáveis fisiológicas como a taxa fotossintética são respostas comuns.

A **resposta principal** é a variável que responde mais diretamente à questão biológica e na qual se baseiam as conclusões principais. As **respostas secundárias** fornecem informações complementares. Em um ensaio de fungicidas cujo objetivo é proteger a produtividade, a produtividade pode ser a resposta principal, e a severidade e a incidência, respostas secundárias que ajudam a explicar os resultados de produtividade.

A resposta deve representar a questão biológica. Um estudo que pretende reduzir perdas de produtividade, mas mede apenas a severidade no início da safra, não consegue mostrar se a produtividade foi protegida. Um estudo de resistência que mede apenas a incidência pode não detectar diferenças na velocidade de desenvolvimento das lesões. A escolha da resposta também envolve considerações práticas: precisão, custo, tempo necessário e possibilidade de que diferentes avaliadores façam a medida de forma consistente.

2.5.3 Variáveis explicativas e covariáveis

Variáveis explicativas são as usadas para explicar a variação na resposta. Em experimentos, as principais variáveis explicativas são os fatores manipulados deliberadamente pelo pesquisador. Outras variáveis explicativas são medidas, e não manipuladas: doença inicial, altura das plantas, umidade do solo ou temperatura registrada durante o experimento.

Variáveis medidas que não são de interesse principal, mas podem influenciar a resposta, são chamadas **covariáveis** ou variáveis de perturbação. Registrá-las permite que análises posteriores levem em conta parte da variação que causam, melhorando a precisão e ajudando a interpretar resultados inesperados. Por exemplo, registrar o nível inicial de doença em cada parcela antes da primeira aplicação de fungicida permite levar em conta diferenças entre parcelas que já existiam antes dos tratamentos. Como essas variáveis são usadas na análise é discutido em capítulos posteriores; na etapa de delineamento, a decisão-chave é identificá-las e registrá-las.

2.5.4 Testemunhas e condições de referência

Uma testemunha (ou controle) é uma condição de referência com a qual os tratamentos são comparados. Sem uma testemunha adequada, não há como saber o que teria acontecido sem o tratamento. Vários tipos de testemunha são comuns em Fitopatologia:

- **testemunhas sem tratamento**, que não recebem fungicida nem prática de manejo e mostram o nível de doença na ausência da intervenção;
- **testemunhas não inoculadas**, que mostram se ocorrem sintomas sem o patógeno e ajudam a detectar contaminação ou infecção natural;
- **padrões suscetíveis e resistentes**, cultivares de resposta conhecida que confirmam se a pressão da doença foi adequada e servem de referência para novos genótipos;
- **controles positivos**, tratamentos sabidamente eficazes, como um fungicida-padrão, com os quais novos produtos são comparados;
- **controles negativos**, como água ou amostras em branco em ensaios de laboratório, que detectam resultados falso-positivos;
- **controles de procedimento**, que recebem todas as manipulações exceto o componente ativo, como plantas pulverizadas com água e espalhante, ou plantas inoculadas apenas com o veículo (*mock*).

A escolha da testemunha depende da questão. Para saber se um novo fungicida controla a doença, é necessária uma testemunha sem tratamento. Para saber se ele é melhor que a prática atual, também é necessário um fungicida-padrão. Para saber se os sintomas são causados pelo patógeno inoculado, testemunhas não inoculadas ou inoculadas apenas com o veículo são essenciais.

2.6 Unidades experimentais e unidades amostrais

2.6.1 O que é uma unidade experimental?

A unidade experimental é a menor unidade à qual um tratamento é atribuído de forma independente. É a unidade que poderia, em princípio, ter recebido qualquer um dos tratamentos, e é o número dessas unidades que determina a repetição verdadeira do experimento.

A unidade experimental depende de como o experimento é conduzido, e não do que é medido. Se cada placa de Petri recebe uma concentração de fungicida, a placa é a unidade experimental. Se cada vaso recebe uma concentração de inóculo, o vaso é a unidade experimental, mesmo que contenha cinco plantas. Se cada parcela é pulverizada com um fungicida, a parcela é a unidade experimental, mesmo que centenas de plantas sejam avaliadas. Se uma lavoura inteira recebe uma prática de manejo, a lavoura é a unidade experimental. Se uma temperatura é aplicada a uma câmara de crescimento inteira, a câmara é a unidade experimental para a temperatura, não importa quantas plantas contenha.

2.6.2 O que é uma unidade amostral?

A unidade amostral é a unidade na qual as medidas são feitas. Ela pode coincidir com a unidade experimental, mas muitas vezes é menor: folhas dentro de uma planta, plantas dentro de uma parcela, lesões dentro de uma folha, colônias dentro de uma placa. Medir várias unidades amostrais dentro de cada unidade experimental chama-se **subamostragem**.

A subamostragem é útil. Ela fornece uma estimativa mais precisa da resposta de cada unidade experimental e informação sobre a variabilidade dentro das unidades. No entanto, as subamostras não são repetições independentes do tratamento, porque compartilham as

condições da mesma unidade experimental. Dez folhas de uma parcela pulverizada fornecem informação sobre aquela parcela, e não dez testes independentes do fungicida.

2.6.3 Hierarquias de observação

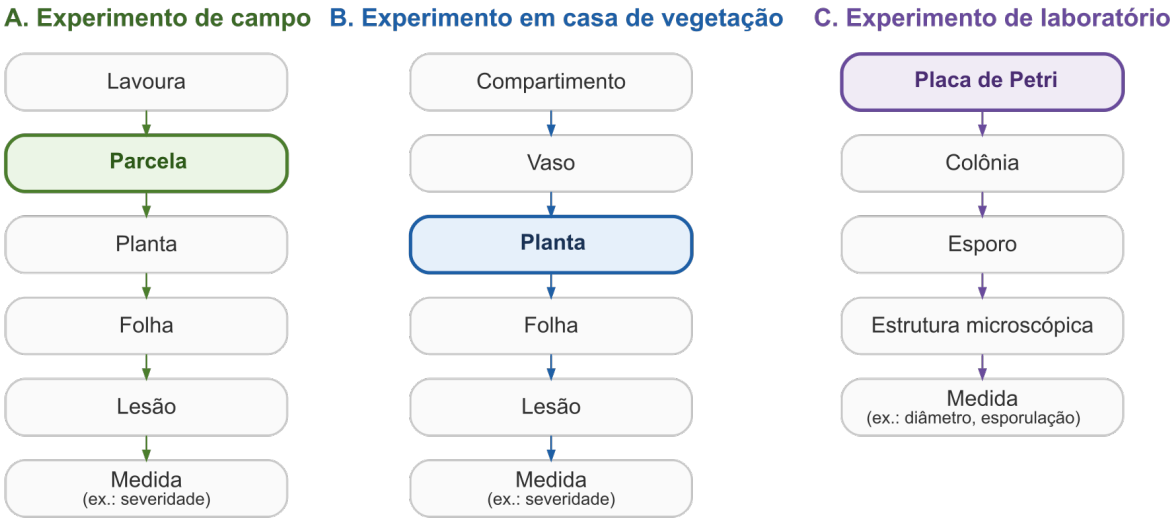


Figura 2.2. Exemplos de organização hierárquica em estudos de Fitopatologia. A unidade experimental varia com o delineamento e é determinada pelo nível em que os tratamentos são atribuídos de forma independente; os níveis destacados indicam a unidade experimental em cada exemplo. Medidas obtidas em níveis hierárquicos inferiores (por exemplo, plantas, folhas, lesões ou colônias) representam unidades amostrais e não devem ser confundidas com repetição verdadeira.

As observações em Fitopatologia costumam estar organizadas em hierarquias: lavouras contêm parcelas, parcelas contêm plantas, plantas contêm folhas e folhas contêm lesões. Da mesma forma, câmaras de crescimento contêm bandejas, bandejas contêm vasos e vasos contêm plantas; incubadoras contêm placas de Petri, e placas contêm colônias.

Identificar a hierarquia é a chave para identificar a unidade experimental. A regra é encontrar o nível em que os tratamentos foram atribuídos de forma independente. Os níveis acima dele costumam ser blocos ou ambientes; os níveis abaixo dele são unidades amostrais. A mesma hierarquia deve estar refletida na organização dos dados (Capítulo 3) e na análise estatística.

> ⚠ **Fonte comum de confusão**
>
> Várias folhas medidas na mesma planta, várias lesões medidas na mesma folha ou várias colônias medidas na mesma placa de Petri não constituem, automaticamente, repetições independentes.

2.7 Repetição

2.7.1 Por que repetir?

Unidades biológicas diferem entre si. Mesmo sob o mesmo tratamento, plantas diferem em tamanho e vigor, parcelas diferem em solo e microclima, e inoculações diferem no número de esporos que chegam e germinam. Por causa dessa variabilidade natural, a resposta de uma única unidade não permite saber se uma diferença entre tratamentos é causada pelo tratamento ou por diferenças casuais entre unidades.

A repetição — a aplicação de cada tratamento a várias unidades experimentais independentes — resolve esse problema. Ela permite estimar a variação entre unidades que recebem o mesmo tratamento, chamada erro experimental. Essa estimativa é a base para julgar se as diferenças entre tratamentos são maiores do que o esperado pela variabilidade natural. A repetição também aumenta a precisão, porque a média de várias unidades estima o efeito do tratamento com mais confiabilidade do que uma única unidade, e aumenta a robustez, porque os resultados não dependem das peculiaridades de uma unidade.

2.7.2 Repetição verdadeira e subamostragem

Repetição verdadeira significa que cada repetição é uma unidade experimental independente que recebeu o tratamento separadamente. Subamostragem significa fazer várias medidas na mesma unidade experimental. Somente a repetição verdadeira fornece uma estimativa do erro experimental adequada para comparar tratamentos.

Considere um experimento em casa de vegetação comparando dois isolados, com um vaso por isolado e dez plantas por vaso. O experimento tem dez medidas por isolado, mas apenas uma unidade experimental por isolado. Qualquer diferença entre os dois vasos — posição na bancada, irrigação ou inoculação — é indistinguível da diferença entre os isolados. Com cinco vasos por isolado e duas plantas por vaso, o mesmo número de plantas forneceria cinco repetições verdadeiras por isolado.

Tabela 2.3. Unidades experimentais, unidades amostrais e repetição verdadeira em estudos típicos.

Situação	Unidade experimental	Unidade amostral	Repetições verdadeiras por tratamento
4 placas por dose de fungicida, 3 colônias medidas por placa	Placa de Petri	Colônia	4
6 vasos por isolado, 5 plantas por vaso	Vaso	Planta	6
4 parcelas por fungicida, 20 folhas avaliadas por parcela	Parcela	Folha	4
1 câmara por temperatura, 30 plantas por câmara	Câmara	Planta	1

2.7.3 Repetição dentro de um experimento

Dentro de um único experimento, a repetição consiste em atribuir cada tratamento a várias parcelas, plantas, vasos ou placas. O número de repetições necessário depende da variabilidade da resposta, do tamanho da diferença biologicamente importante e de restrições práticas. Ensaios de campo em Fitopatologia costumam usar de quatro a seis parcelas por tratamento; experimentos em casa de vegetação e laboratório frequentemente usam mais unidades, porque cada uma é mais barata. A relação entre o número de repetições e a capacidade de detectar efeitos é discutida no Capítulo 6.

2.7.4 Repetindo um experimento

A repetição dentro de um experimento leva em conta a variabilidade entre unidades nas condições daquele experimento. Ela não leva em conta a variação entre execuções, safras, locais ou anos. Um fungicida com bom desempenho em uma safra pode ter desempenho diferente em uma safra com outro clima ou outra pressão de doença; um experimento em câmara de

crescimento pode dar resultados diferentes com um novo lote de inóculo ou outro conjunto de plantas.

Por isso, os experimentos são frequentemente repetidos de forma independente: em diferentes safras ou anos, em diferentes locais ou como execuções separadas, com novas plantas e novo inóculo. Muitas revistas de Fitopatologia exigem que experimentos em ambiente controlado sejam conduzidos pelo menos duas vezes. A repetição do experimento mostra se os resultados são consistentes e amplia a gama de condições às quais as conclusões se aplicam.

2.7.5 Combinando experimentos repetidos

Quando um experimento é repetido, os resultados das diferentes execuções podem ser analisados em conjunto. Uma análise conjunta pode estimar o efeito médio do tratamento entre as execuções e avaliar se o efeito varia entre elas. Isso é apropriado quando os experimentos compartilham os mesmos tratamentos, a mesma resposta medida da mesma forma e delineamentos comparáveis. A variação entre execuções passa então a fazer parte da informação sobre o grau de generalidade dos resultados. Os métodos para combinar experimentos são apresentados em capítulos posteriores; na etapa de delineamento, o importante é planejar os experimentos repetidos de modo que possam ser combinados.

2.8 Casualização

2.8.1 Por que a casualização é essencial

A casualização (ou aleatorização) é a atribuição dos tratamentos às unidades experimentais por um processo de sorteio, como um gerador de números aleatórios. Ela garante que nenhum fator sistemático, conhecido ou desconhecido, determine quais unidades recebem quais tratamentos. Sem casualização, as escolhas do pesquisador, mesmo inconscientes, podem associar tratamentos a unidades específicas: as plantas mais sadias, as parcelas mais próximas da estrada, as bancadas com melhor iluminação.

A casualização protege contra esse tipo de viés sistemático. Ela não torna os grupos idênticos, mas faz com que as diferenças entre eles resultem do acaso, e não de uma causa oculta — e o acaso é justamente o que os métodos estatísticos são feitos para considerar. Nesse sentido, a casualização é o que torna válida a inferência estatística sobre os efeitos dos tratamentos.

2.8.2 Alocação aleatória dos tratamentos

Em ensaios de campo, os tratamentos são sorteados entre as parcelas, geralmente dentro de blocos. Em casas de vegetação, os vasos são sorteados entre tratamentos e posições na bancada. Em laboratórios, placas ou frascos são sorteados entre tratamentos e posições na incubadora. Em câmaras de crescimento, a atribuição dos tratamentos às câmaras, e a posição das unidades dentro delas, também deve ser casualizada.

A casualização deve ser feita com um método documentado, como uma tabela de números aleatórios ou um programa de computador, e o croqui resultante deve ser registrado. A escolha “ao acaso” feita à mão não é casualização verdadeira, porque as pessoas tendem a fazer escolhas sistemáticas sem perceber.

2.8.3 Ordem aleatória das medições

A casualização também se aplica à ordem em que as unidades são processadas e medidas. Instrumentos podem sofrer deriva durante uma sessão, avaliadores podem ficar cansados ou

mais experientes à medida que trabalham, e amostras processadas mais tarde podem ter ficado armazenadas por mais tempo. Se todas as unidades de um tratamento forem medidas primeiro e as de outro por último, esses efeitos ficam confundidos com o tratamento.

Medir as unidades em ordem aleatória, ou em uma ordem equilibrada entre tratamentos, evita esses problemas. O mesmo princípio se aplica à posição das amostras em placas de qPCR e microplacas, onde efeitos de borda são comuns, e à ordem de processamento no laboratório. Sempre que possível, os avaliadores não devem saber qual tratamento cada unidade recebeu (avaliação cega), especialmente em estimativas visuais de doença.

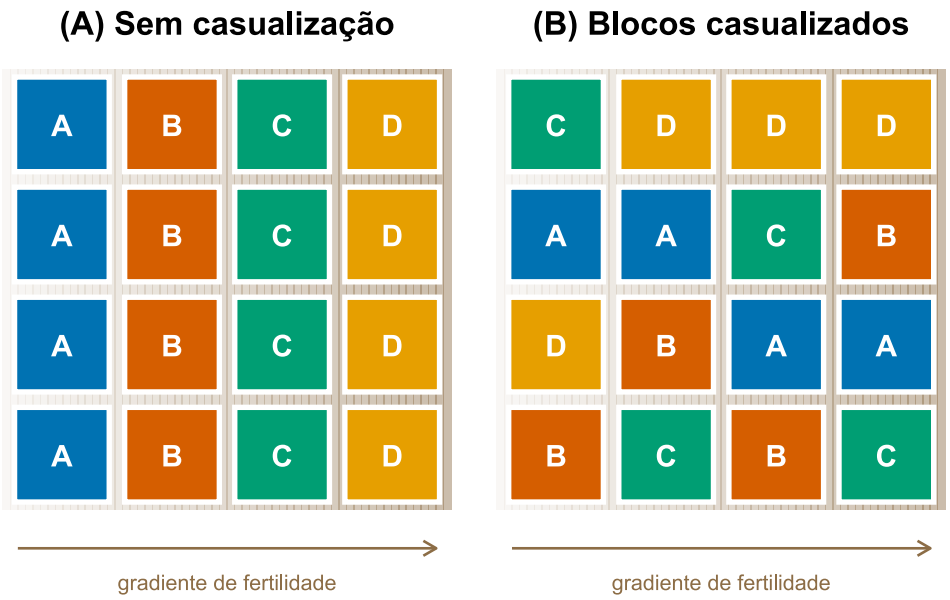


Figura 2.3. Exemplo de alocação de quatro tratamentos (A–D) com quatro repetições em uma área com gradiente de fertilidade (fundo sombreado). (A) Sem casualização: cada tratamento ocupa uma coluna, e o efeito do tratamento fica confundido com o gradiente. (B) Blocos casualizados: cada coluna é um bloco perpendicular ao gradiente, e os tratamentos são sorteados dentro de cada bloco.

2.9 Bloqueamento

2.9.1 Por que usar blocos?

As unidades experimentais raramente são uniformes. Uma lavoura pode ter um gradiente de fertilidade ou de drenagem do solo; uma casa de vegetação pode ser mais quente junto às paredes; plantas inoculadas em dias diferentes podem receber inóculo de qualidade diferente. Quando essas fontes de variação são conhecidas ou suspeitas, podem ser controladas por meio do bloqueamento.

Um bloco é um grupo de unidades experimentais que se espera serem semelhantes entre si. Cada bloco contém todos os tratamentos, sorteados dentro dele. Como os tratamentos são comparados dentro dos blocos, as diferenças entre blocos não afetam a comparação dos tratamentos, e a variação que causam pode ser separada do erro experimental. O resultado é uma comparação mais precisa dos tratamentos.

2.9.2 Blocos em experimentos de campo

Em ensaios de campo, os blocos geralmente são grupos de parcelas adjacentes dispostos perpendicularmente a um gradiente conhecido ou suspeito, como declividade, tipo de solo,

fertilidade, drenagem, distância de um quebra-vento ou distância de uma fonte de inóculo. As parcelas dentro de um bloco devem ser tão homogêneas quanto possível, enquanto os blocos podem diferir entre si. Se a direção do gradiente for desconhecida, blocos compactos e aproximadamente quadrados geralmente são preferíveis a blocos longos e estreitos.

2.9.3 Blocos em casa de vegetação e laboratório

Em casas de vegetação e laboratórios, os blocos frequentemente correspondem a bancadas, prateleiras, bandejas, posições na incubadora, placas, dias de inoculação, avaliadores ou execuções do experimento. Se nem todas as unidades puderem ser inoculadas ou avaliadas no mesmo dia, cada dia pode se tornar um bloco contendo todos os tratamentos. Isso impede que diferenças entre dias sejam confundidas com diferenças entre tratamentos.

2.9.4 Bloco não é repetição

Blocos e repetições são frequentemente confundidos porque, no delineamento em blocos casualizados, cada tratamento aparece uma vez em cada bloco, e o número de blocos é igual ao número de repetições. Conceitualmente, porém, são coisas diferentes. A repetição refere-se ao número de unidades experimentais independentes que recebem cada tratamento. O bloqueamento refere-se ao agrupamento das unidades para controlar fontes conhecidas de variação. Um bloco que contém várias unidades do mesmo tratamento continua fornecendo repetições, e um experimento pode ter repetições sem ter blocos. O bloqueamento também não substitui a casualização: os tratamentos ainda precisam ser sorteados dentro de cada bloco.

2.10 Delineamentos experimentais comuns em Fitopatologia

2.10.1 Delineamento inteiramente casualizado

No delineamento inteiramente casualizado (DIC), os tratamentos são atribuídos às unidades experimentais inteiramente ao acaso, sem blocos. É o delineamento mais simples e é adequado quando as unidades experimentais são homogêneas, como placas de Petri em uma única incubadora ou vasos em uma câmara de crescimento com condições uniformes. Seus pontos fortes são a simplicidade e a flexibilidade, inclusive para números desiguais de repetições. Sua limitação é que qualquer heterogeneidade entre unidades aumenta o erro experimental, dificultando a detecção de diferenças entre tratamentos.

2.10.2 Delineamento em blocos casualizados

No delineamento em blocos casualizados (DBC), as unidades são agrupadas em blocos, cada bloco contém todos os tratamentos uma vez, e os tratamentos são sorteados independentemente dentro de cada bloco. É o delineamento mais comum em ensaios de campo e também é muito usado em casas de vegetação. Seu ponto forte é controlar a variação associada aos blocos, melhorando a precisão quando os blocos realmente diferem. Suas limitações são que os blocos precisam ser grandes o suficiente para conter todos os tratamentos, o que se torna difícil com muitos tratamentos, e que o delineamento é menos eficaz se a variação dentro dos blocos for grande.

2.10.3 Experimentos fatoriais

Experimentos fatoriais estudam dois ou mais fatores simultaneamente, com tratamentos formados pela combinação dos níveis de cada fator. Exemplos incluem cultivar × fungicida, temperatura × duração do molhamento, isolado × cultivar e inoculação × nutrição.

Experimentos fatoriais são eficientes, porque cada unidade fornece informação sobre todos os fatores. Sobretudo, permitem estudar **interações**, situações em que o efeito de um fator depende do nível de outro. Um fungicida pode reduzir muito a doença em uma cultivar suscetível e ter pouco efeito em uma resistente; a infecção pode aumentar com a duração do molhamento a 25 °C, mas não a 10 °C. As interações estão frequentemente entre os resultados biologicamente mais informativos de um experimento e não podem ser detectadas estudando um fator de cada vez. Os tratamentos fatoriais podem ser dispostos em DIC ou em DBC.

2.10.4 Parcelas subdivididas e delineamentos hierárquicos

Em alguns experimentos, os fatores são aplicados a unidades de tamanhos diferentes. A irrigação ou o preparo do solo, por exemplo, podem ser praticáveis apenas em parcelas grandes, enquanto cultivares ou fungicidas podem ser aplicados a subparcelas dentro delas. Em um experimento em câmara de crescimento, a temperatura é aplicada a câmaras inteiras, enquanto os isolados podem ser atribuídos a vasos dentro de cada câmara. Esses delineamentos são chamados de parcelas subdivididas e, de modo mais geral, delineamentos hierárquicos ou aninhados.

Esses delineamentos são práticos e comuns, mas têm duas unidades experimentais de tamanhos diferentes, cada uma com sua própria repetição e seu próprio erro experimental. As comparações do fator aplicado às unidades maiores costumam ser menos precisas, porque há menos delas. Reconhecer uma estrutura de parcelas subdivididas na etapa de delineamento é essencial, porque analisá-la como se todos os fatores tivessem sido aplicados às menores unidades superestima a repetição do fator da parcela principal.

2.10.5 Medidas repetidas no tempo

A doença é frequentemente avaliada várias vezes nas mesmas parcelas ou plantas para acompanhar seu desenvolvimento durante a safra ou após a inoculação. Essas avaliações repetidas não acrescentam unidades experimentais: são medidas sucessivas da mesma unidade, e avaliações da mesma unidade tendem a ser mais parecidas entre si do que avaliações de unidades diferentes.

As medidas repetidas são valiosas porque descrevem como a doença se desenvolve, e não apenas seu nível final. Na análise, elas podem ser resumidas em um único valor por unidade ou modeladas levando em conta a dependência entre as avaliações da mesma unidade. Na etapa de delineamento, os pontos essenciais são definir as épocas de avaliação com antecedência, avaliar todas as unidades em cada época e registrar a qual unidade pertence cada avaliação.

2.11 Problemas comuns de delineamento em Fitopatologia

2.11.1 Pseudorrepetição

A pseudorrepetição ocorre quando subamostras ou medidas repetidas são tratadas como se fossem repetições independentes (Hurlbert 1984). Exemplos típicos incluem várias colônias medidas em uma única placa de Petri por tratamento, folhas da mesma planta analisadas como repetições separadas, todas as plantas de um tratamento cultivadas em uma única bandeja e leituras repetidas da mesma amostra tratadas como observações independentes. A pseudorrepetição infla o tamanho aparente da amostra e subestima o erro experimental, fazendo as diferenças entre tratamentos parecerem mais certas do que são.

2.11.2 Confundimento

O confundimento ocorre quando o efeito de um tratamento não pode ser separado do efeito de outra variável que difere entre os grupos de tratamento. Exemplos clássicos incluem uma câmara de crescimento por temperatura, de modo que temperatura e câmara ficam confundidas; cada cultivar cultivada em um único local, de modo que cultivar e local ficam confundidos; tratamentos aplicados ou inoculados em datas diferentes; e equipamentos específicos para cada tratamento, como um pulverizador diferente para cada produto. O confundimento não pode ser removido pela análise; deve ser evitado pelo delineamento, por meio de casualização, repetição das condições envolvidas e bloqueamento.

2.11.3 Ausência de testemunha adequada

Sem uma testemunha adequada, o efeito de um tratamento não pode ser avaliado. Um estudo mostrando que um agente de controle biológico reduziu a doença para 10% de severidade não é informativo sem saber a severidade das plantas não tratadas nas mesmas condições. Uma testemunha com pouca doença revela que a pressão da doença foi baixa demais para avaliar os tratamentos. Um teste de patogenicidade sem plantas não inoculadas não pode excluir a possibilidade de que os sintomas tenham sido causados por outra coisa.

2.11.4 Unidades experimentais inadequadas ou pouco claras

Muitos artigos publicados não informam explicitamente qual foi a unidade experimental, quantas unidades receberam cada tratamento ou como as subamostras foram tratadas. O leitor não consegue, então, determinar se a repetição foi adequada ou se a análise foi apropriada. Todo estudo deve informar claramente a unidade experimental, o número de repetições e as unidades amostrais medidas dentro de cada uma.

2.11.5 Tratar repetições técnicas como repetições biológicas

Em estudos moleculares e fisiológicos, distingue-se entre **repetições biológicas**, amostras independentes de plantas ou unidades diferentes, e **repetições técnicas**, medidas repetidas da mesma amostra, como vários poços de qPCR carregados com a mesma extração de RNA ou leituras repetidas da mesma folha com um instrumento. As repetições técnicas medem a precisão do procedimento de medida; somente as repetições biológicas medem a variabilidade biológica e podem sustentar conclusões sobre tratamentos.

2.11.6 Implementação desigual dos tratamentos

Os tratamentos devem diferir apenas no fator estudado. Surgem problemas quando os tratamentos são implementados com lotes de inóculo diferentes, por operadores diferentes, em datas diferentes, em câmaras diferentes ou com equipamentos de pulverização diferentes. Qualquer diferença desse tipo fica confundida com o tratamento. Os procedimentos devem ser padronizados e, quando não puderem ser idênticos, equilibrados entre os tratamentos — por exemplo, fazendo cada operador trabalhar com todos os tratamentos.

2.11.7 Poucos experimentos independentes

Um único experimento, mesmo bem repetido, reflete as condições de uma safra, um local, um lote de plantas ou uma execução. Seus resultados podem não se manter em outras condições. Afirmações amplas, como recomendar um programa de fungicidas ou classificar uma cultivar como resistente, exigem evidências de vários experimentos independentes. Por isso, muitas

revistas exigem a repetição de experimentos em ambiente controlado, e recomendações de campo costumam se basear em ensaios conduzidos em vários anos ou locais.

2.11.8 Medir o que é conveniente, e não o que responde à questão

Às vezes, as respostas são escolhidas por serem fáceis de medir, e não por responderem à questão. A incidência é mais rápida de avaliar do que a severidade, e uma única avaliação final é mais barata do que várias, mas nenhuma delas pode capturar o efeito de interesse. Um estudo sobre proteção da produtividade que não mede a produtividade, ou um estudo de resistência que mede apenas se as plantas estão infectadas, pode produzir dados precisos que não respondem à questão biológica.

2.12 Pesquisa na prática: revisitando um experimento de manejo da doença

O estudo do míldio do manjericão apresentado no Capítulo 1 avaliou programas de fungicidas em experimentos de campo sob pressão natural da doença (McGrath 2020). Em vez de repetir sua descrição, vamos agora examiná-lo à luz dos conceitos apresentados neste capítulo. Responder às questões abaixo para qualquer estudo publicado é uma forma eficaz de avaliar a força de sua evidência.

Pesquisa revisitada: Eficácia de fungicidas em manjericão

Questão biológica

Aplicações de fungicidas reduzem o míldio do manjericão em condições de campo?

Questões sobre o delineamento

1. Quais tratamentos foram comparados?
2. Qual foi a testemunha sem tratamento?
3. Qual foi a unidade experimental?
4. Por que foram usados blocos?
5. O que foi repetido?
6. O que foi medido repetidamente?
7. Por que os experimentos foram conduzidos em vários anos?
8. Quais conclusões se aplicam a um único experimento e quais exigem evidências de vários experimentos?

2.13 Planejando um estudo antes de coletar dados

2.13.1 Um roteiro prático de planejamento

As questões abaixo resumem as decisões discutidas neste capítulo. Respondê-las por escrito antes da coleta de dados, de preferência em um breve protocolo do estudo, ajuda a detectar problemas de delineamento enquanto ainda podem ser corrigidos.

1. Que questão biológica o estudo vai responder?
2. Que hipótese científica está sendo avaliada?
3. Que observações sustentariam ou contradiriam essa hipótese?
4. O estudo é observacional ou experimental?

5. Quais são os fatores, os tratamentos e as testemunhas?
6. Qual é a unidade experimental ou observacional?
7. Qual é a unidade amostral?
8. Quais variáveis resposta serão medidas?
9. Como as unidades serão repetidas?
10. Como os tratamentos serão casualizados?
11. É necessário usar blocos?
12. O experimento será repetido de forma independente?
13. Que fontes de viés ou confundimento são plausíveis?
14. Como os dados serão registrados e documentados?

2.14 Resumo do capítulo

Uma análise confiável começa com um estudo delineado para responder a uma questão biológica claramente definida. Um bom delineamento traduz a questão em uma hipótese falseável, em predições testáveis e em definições operacionais dos conceitos envolvidos. Estudos observacionais e experimentais fornecem formas diferentes de evidência: os observacionais descrevem padrões e associações em condições reais, enquanto os experimentos avaliam os efeitos de tratamentos impostos deliberadamente. Todo estudo é construído a partir de fatores, tratamentos, testemunhas e variáveis resposta, e a unidade experimental — a menor unidade à qual um tratamento é atribuído de forma independente — determina a repetição verdadeira. A repetição estima o erro experimental, a casualização protege contra o viés sistemático e o bloqueamento controla fontes conhecidas de variação. Subamostras, repetições técnicas e medidas repetidas não substituem a repetição verdadeira, e repetir o experimento de forma independente costuma ser necessário antes que conclusões mais amplas possam ser sustentadas. Problemas comuns como pseudorrepetição, confundimento e ausência de testemunhas adequadas não podem ser corrigidos pela análise estatística e devem ser evitados na etapa de delineamento.

2.15 Conceitos-chave

Ao concluir este capítulo, você deverá ser capaz de explicar:

- questão biológica;
- hipótese científica;
- predição;
- definição operacional;
- estudo observacional;
- estudo experimental;
- fator;
- tratamento;
- testemunha;
- variável resposta;
- variável explicativa;
- covariável;
- unidade experimental;
- unidade amostral;
- repetição;

- subamostragem;
- repetição do experimento;
- casualização;
- bloqueamento;
- interação;
- pseudorrepetição;
- confundimento.

2.16 Exercícios

2.16.1 Exercício 1. Identificando os componentes de um estudo

Escolha um artigo fornecido pelo professor e responda:

1. Que questão biológica motivou o estudo?
2. O estudo foi observacional ou experimental?
3. Que condições, tratamentos ou grupos foram comparados?
4. Qual foi a unidade experimental ou observacional?
5. O que foi medido?
6. Havia uma testemunha adequada?
7. Como a repetição foi obtida?
8. O estudo foi repetido de forma independente?

2.16.2 Exercício 2. Unidade experimental ou unidade amostral?

Para cada cenário, identifique a unidade experimental, a unidade amostral e o número de repetições verdadeiras por tratamento.

1. Quatro concentrações de fungicida são testadas *in vitro*. Cada concentração é adicionada a cinco placas de Petri, e o diâmetro de três colônias é medido em cada placa.
2. Duas concentrações de inóculo são comparadas em casa de vegetação. Cada concentração é aplicada a seis vasos com quatro plantas cada, e a severidade é avaliada em todas as plantas.
3. Três programas de fungicida e uma testemunha são comparados em um ensaio de campo com quatro blocos. Em cada parcela, a severidade é avaliada em 20 folhas.
4. A expressão de um gene de defesa é comparada entre plantas inoculadas e plantas inoculadas apenas com o veículo (*mock*). Três plantas por tratamento são amostradas, o RNA é extraído de cada planta e cada extrato é analisado em três poços de qPCR.
5. Três temperaturas são comparadas usando três câmaras de crescimento, uma por temperatura, cada uma com 30 plantas inoculadas.

2.16.3 Exercício 3. Detectando problemas de delineamento

Leia as breves descrições de estudos fornecidas e identifique problemas envolvendo:

- pseudorrepetição;
- confundimento;
- ausência de testemunhas;
- repetição inadequada;
- atribuição dos tratamentos pouco clara;
- conclusões inadequadas.

2.16.4 Exercício 4. Delineando um estudo

Escolha uma questão:

- Um fungicida reduz a severidade da doença?
- A resistência da cultivar reduz a infecção?
- A temperatura influencia o desenvolvimento das lesões?
- A inoculação altera as respostas fisiológicas da planta?

Prepare um breve delineamento especificando:

1. a questão biológica;
2. a hipótese;
3. os tratamentos ou grupos;
4. a testemunha;
5. a unidade experimental;
6. a variável resposta;
7. a repetição;
8. a casualização;
9. o bloqueamento, se necessário;
10. se e como o estudo deve ser repetido.

2.17 Reflexão

> **Nenhum método estatístico compensa um estudo mal delineado, mas um estudo bem delineado muitas vezes pode ser analisado com métodos relativamente simples.**

Discuta essa afirmação usando um exemplo da Fitopatologia.

2.18 Próximos passos

Um estudo bem delineado produz observações que precisam ser registradas de forma consistente, transparente e analisável. O próximo capítulo examina como os dados científicos devem ser estruturados, documentados, verificados e preparados para a exploração e a análise.

CAPÍTULO 3

Como os dados científicos devem ser organizados?

3.1 Objetivos de aprendizagem

! Importante

Ao concluir este capítulo, o leitor deverá ser capaz de:

- explicar por que a organização dos dados é essencial para a pesquisa reprodutível;
- distinguir dados brutos de dados processados;
- construir conjuntos de dados organizados (*tidy*) adequados para a análise;
- identificar observações, variáveis e identificadores;
- organizar metadados e informações de apoio;
- reconhecer erros comuns de digitação de dados;
- aplicar boas práticas de documentação e armazenamento de dados de pesquisa.

3.2 Por que a organização dos dados importa

Um estudo bem delineado produz observações, mas as observações só se tornam evidência depois de registradas, organizadas, verificadas e documentadas. Entre o campo, a casa de vegetação ou o laboratório e a análise final existe uma etapa frequentemente tratada como burocrática: a construção do conjunto de dados. Na prática, essa etapa determina se a análise será simples ou penosa, transparente ou opaca, correta ou silenciosamente errada.

Dados mal organizados causam problemas que nenhum método estatístico consegue corrigir. Um valor de severidade digitado na linha errada atribui doença ao tratamento errado. Um valor perdido registrado como zero reduz a média de um tratamento. Uma planilha em que os tratamentos são identificados apenas pela cor das células não pode ser lida por programas de análise. Um arquivo chamado `dados_final_v2_NOVO_corrigido.xlsx` não indica se contém as observações brutas ou uma versão da qual se removeram valores discrepantes. Cada um desses problemas é comum, e cada um pode mudar as conclusões de um estudo sem que ninguém perceba.

A organização dos dados também determina se a pesquisa pode ser reproduzida. Um colaborador, um revisor ou o próprio pesquisador dois anos depois deve conseguir abrir o conjunto de dados e entender o que cada linha representa, o que cada coluna contém, em que unidades os valores estão expressos e como o conjunto foi obtido a partir dos registros originais.

Quando essa informação existe apenas na memória de quem coletou os dados, os resultados não podem ser verificados, reutilizados ou combinados com outros estudos.

Por fim, uma boa organização economiza tempo. Os pesquisadores frequentemente gastam mais tempo preparando os dados para a análise do que analisando-os. Boa parte desse esforço decorre de decisões tomadas na digitação: nomes inconsistentes, células mescladas, várias tabelas na mesma planilha ou informação codificada em formatação. Organizar os dados com cuidado desde o início é muito mais barato do que consertá-los depois.

> **! Importante**
>
> Dados bem organizados são a base da pesquisa científica reprodutível. A estatística não compensa dados mal organizados.

3.3 Das observações ao conjunto de dados

3.3.1 O que é um conjunto de dados?

Um conjunto de dados é uma coleção estruturada de observações registradas segundo regras consistentes. Na maioria dos estudos em Fitopatologia, ele tem a forma de uma tabela retangular: as linhas representam o que foi observado, e as colunas representam as características registradas para cada observação. Essa forma retangular não é mera convenção. É a estrutura esperada pelos programas estatísticos, e ela obriga o pesquisador a declarar explicitamente o que foi observado e o que foi medido.

Um conjunto de dados não é o mesmo que um arquivo de planilha. Um único arquivo pode conter vários conjuntos de dados (por exemplo, uma tabela com informações das parcelas e outra com as avaliações de doença), e um único conjunto de dados pode estar guardado em vários arquivos (por exemplo, avaliações de anos diferentes). O que importa é que cada tabela tenha uma estrutura clara e que as relações entre as tabelas estejam documentadas.

3.3.2 Observações

Uma observação é o registro único das características de uma unidade em um momento. Em um ensaio de campo, uma observação pode ser a severidade da doença em uma parcela em uma data de avaliação. Em um ensaio de laboratório, pode ser o diâmetro da colônia em uma placa de Petri após determinado período de incubação. Em um levantamento, pode ser a informação registrada para uma lavoura visitada.

Definir a observação é a decisão mais importante na organização dos dados, e ela decorre diretamente do delineamento do estudo discutido no Capítulo 2. Se a severidade foi avaliada em dez folhas por parcela, o pesquisador precisa decidir se cada linha representará uma folha ou uma parcela. As duas escolhas podem ser válidas, mas levam a conjuntos de dados diferentes e devem ser declaradas claramente. Manter o nível mais detalhado efetivamente registrado (aqui, a folha) costuma ser preferível, porque resumos em níveis superiores (médias por parcela) sempre podem ser calculados depois, ao passo que a informação detalhada não pode ser recuperada a partir dos resumos.

3.3.3 Variáveis

Uma variável é uma característica registrada para todas as observações e que pode variar entre elas. Tratamento, bloco, cultivar, data de avaliação, incidência, severidade, temperatura e

produtividade são variáveis. Cada variável é registrada em uma coluna, e todos os valores dessa coluna devem se referir à mesma característica, medida da mesma forma e expressa nas mesmas unidades.

Algumas variáveis descrevem o delineamento do estudo (tratamento, bloco, repetição, execução), outras descrevem o contexto (local, data, cultivar, avaliador) e outras são as respostas que respondem à questão biológica (severidade, incidência, número de lesões, produtividade). Todas pertencem ao conjunto de dados. A informação sobre o delineamento é particularmente fácil de omitir, porque parece óbvia para quem coletou os dados, mas é indispensável para a análise.

3.3.4 Identificadores

Um identificador é uma variável, ou combinação de variáveis, que distingue de forma única cada unidade. Números de parcela, etiquetas de plantas, códigos de isolados, códigos de barras de amostras e identificadores de lavouras são exemplos típicos. Os identificadores ligam as observações às unidades físicas de onde foram obtidas, conectam tabelas diferentes (por exemplo, informações das parcelas e avaliações de doença) e permitem rastrear qualquer valor até sua origem.

Na terminologia de bancos de dados, uma variável ou combinação de variáveis que identifica de forma única cada linha de uma tabela é chamada chave primária. Em uma tabela de características das parcelas, o número da parcela pode bastar. Em uma tabela de avaliações repetidas, é necessária a combinação de parcela e data, porque cada parcela aparece uma vez para cada avaliação. Se duas linhas têm a mesma chave, ou os dados contêm duplicatas ou a chave está incompleta.

3.3.5 Unidades de observação

A unidade de observação é a entidade descrita por uma linha da tabela. Ela está relacionada, mas nem sempre é igual, à unidade experimental e à unidade amostral apresentadas no Capítulo 2. Em um ensaio de campo em que os tratamentos são aplicados às parcelas e a severidade é avaliada em folhas individuais, a parcela é a unidade experimental, a folha é a unidade amostral, e o conjunto de dados poderia ser organizado tendo a folha ou a parcela como unidade de observação.

O importante é que a unidade de observação seja explícita e constante dentro de uma tabela. Uma tabela em que algumas linhas representam folhas individuais e outras representam médias de parcelas não pode ser analisada corretamente. Quando há informação em mais de um nível, geralmente é melhor guardá-la em tabelas separadas, uma para cada nível, ligadas por identificadores.

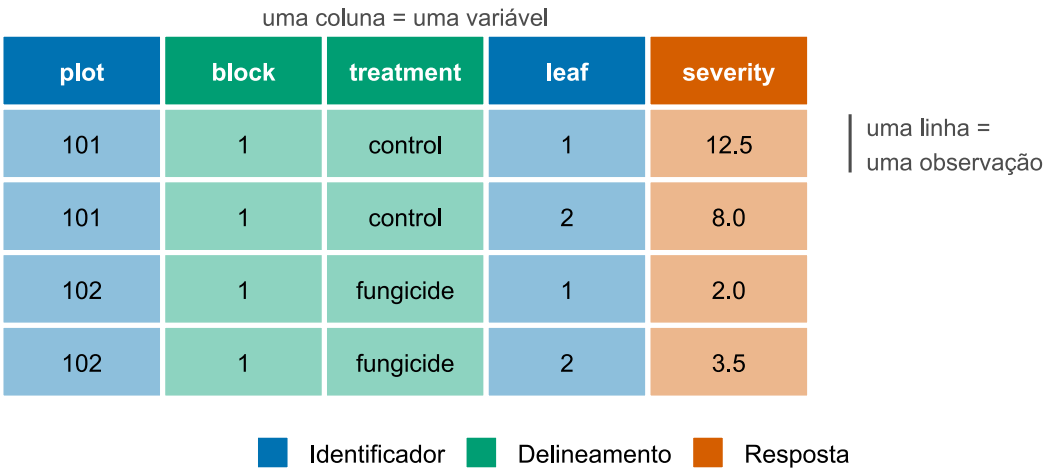


Figura 3.1. Estrutura de um conjunto de dados científicos. Cada linha é uma observação (aqui, uma folha) e cada coluna é uma variável. As cores indicam o papel das variáveis: identificadores, variáveis do delineamento e resposta.

Tabela 3.1. Um pequeno conjunto de dados em que cada linha é uma folha (a unidade de observação), cada coluna é uma variável, e a combinação de `plot` e `leaf` identifica cada linha de forma única.

plot	block	treatment	leaf	severity
101	1	control	1	12.5
101	1	control	2	8.0
102	1	fungicide	1	2.0
102	1	fungicide	2	3.5

ℹ Nota

Nos exemplos deste livro, nomes de variáveis e códigos de categorias são mantidos em inglês (`plot`, `block`, `treatment`, `severity`), como é comum em códigos de análise e em dados compartilhados internacionalmente. O mais importante é escolher uma convenção e aplicá-la de forma consistente.

3.4 Princípios de dados organizados (*tidy data*)

A estrutura descrita acima foi formalizada como **dados organizados** (*tidy data*) (Wickham 2014). Um conjunto de dados é organizado quando sua estrutura corresponde ao seu significado: a forma da tabela reflete diretamente o que foi observado e o que foi medido. Dados organizados seguem três princípios.

3.4.1 Uma linha por observação

Cada linha contém todos os valores registrados para uma única observação e nada mais. Se a severidade foi avaliada em quatro datas em cada parcela, um conjunto organizado contém quatro linhas por parcela, uma para cada avaliação. Uma linha nunca deve conter observações de mais de uma unidade ou de mais de um momento.

3.4.2 Uma coluna por variável

Cada coluna contém uma única variável. A informação que varia entre observações, como a data de avaliação, pertence à sua própria coluna, e não aos nomes das colunas. Uma tabela com colunas chamadas `sev_jun10`, `sev_jun24` e `sev_jul08` guarda a variável *data* nos cabeçalhos. A versão organizada tem uma coluna chamada `date` e uma coluna chamada `severity`.

O princípio também proíbe guardar duas variáveis em uma coluna. Uma coluna com valores como `R1-control` combina repetição e tratamento, e uma coluna com `25 C/12 h` combina temperatura e duração do molhamento. Cada uma deve ser dividida em colunas separadas.

3.4.3 Uma tabela por unidade de observação

Cada tipo de unidade de observação fica em sua própria tabela. As características das parcelas (bloco, tratamento, cultivar, data de semeadura) são registradas uma vez por parcela, em uma tabela de parcelas. As avaliações repetidas ficam em uma tabela de avaliações, com uma linha por parcela e data. Registros meteorológicos, com uma linha por hora ou por dia, formam uma terceira tabela. Essas tabelas são ligadas por identificadores compartilhados, como o número da parcela ou a data, em vez de repetir toda a informação em uma única tabela grande.

3.4.4 Por que dados organizados simplificam as análises

Dados organizados não são questão de estilo. A maioria dos programas estatísticos e gráficos, incluindo as ferramentas do R usadas neste livro, espera que cada variável seja uma coluna e cada observação uma linha. Quando os dados estão organizados, o mesmo pequeno conjunto de operações (filtrar linhas, selecionar colunas, agrupar, resumir e unir tabelas) pode ser aplicado a praticamente qualquer conjunto de dados. Quando não estão, toda análise começa com uma reorganização diferente e improvisada, demorada e sujeita a erros.

A estrutura organizada também torna visível o delineamento do estudo. Olhando uma tabela organizada, o leitor vê os tratamentos, os blocos, as unidades e as medidas repetidas, e pode verificar se o conjunto de dados é coerente com o estudo descrito.

Não organizado (largo)

plot	treatment	sev_jun10	sev_jun24	sev_jul08
101	control	2	10	35
102	fungicide	0	2	6

Organizado (longo)

plot	treatment	date	severity
101	control	2025-06-10	2
101	control	2025-06-24	10
101	control	2025-07-08	35
102	fungicide	2025-06-10	0
102	fungicide	2025-06-24	2
102	fungicide	2025-07-08	6

Figura 3.2. Os mesmos dados em formato não organizado (largo), com as datas de avaliação guardadas nos nomes das colunas, e em formato organizado (longo), em que data e severidade são variáveis em colunas próprias.

Tabela 3.2. Uma tabela não organizada (formato largo): as datas de avaliação estão nos nomes das colunas.

plot	treatment	sev_jun10	sev_jun24	sev_jul08
101	control	2	10	35
102	fungicide	0	2	6

Tabela 3.3. Os mesmos dados em formato organizado (longo): uma linha por parcela e data, com data e severidade como variáveis separadas.

plot	treatment	date	severity
101	control	2025-06-10	2
101	control	2025-06-24	10
101	control	2025-07-08	35
102	fungicide	2025-06-10	0
102	fungicide	2025-06-24	2
102	fungicide	2025-07-08	6

ℹ Nota

Tabelas largas costumam ser práticas para anotar dados no campo, e não há nada de errado em usá-las na ficha de campo. O que importa é que o conjunto de dados usado na análise esteja organizado e que a conversão de um formato para o outro seja documentada e reprodutível.

3.5 Estruturas de dados comuns em Fitopatologia

Embora os estudos em Fitopatologia variem muito, a maioria dos conjuntos de dados se enquadra em algumas estruturas recorrentes. Reconhecê-las ajuda a decidir como organizar os dados antes da coleta.

3.5.1 Experimentos de avaliação de doença

Em muitos experimentos, a doença é avaliada em várias unidades amostrais dentro de cada unidade experimental: folhas em uma planta, plantas em um vaso ou lesões em uma folha. O conjunto de dados deve registrar a unidade experimental, a unidade amostral e a resposta.

Tabela 3.4. Experimento de avaliação com folhas aninhadas em plantas e plantas aninhadas em vasos.

pot	isolate	plant	leaf	lesions
1	A12	1	1	4
1	A12	1	2	7
1	A12	2	1	3

3.5.2 Ensaios de campo

Ensaios de campo normalmente incluem tratamentos dispostos em blocos, muitas vezes com várias variáveis resposta medidas por parcela. Cada linha é uma parcela, e cada resposta é uma coluna.

Tabela 3.5. Ensaio de campo com uma linha por parcela.

plot	block	treatment	incidence	severity	yield_kg_ha
101	1	control	85	22.4	2850
102	1	fung_A	30	4.1	3620

3.5.3 Experimentos em casa de vegetação

Experimentos em casa de vegetação e câmara de crescimento acrescentam fontes de estrutura que precisam ser registradas: bancada, câmara, bandeja, posição e execução do experimento. Essas variáveis podem ser usadas depois como blocos ou efeitos aleatórios, e omiti-las impossibilita considerá-las.

Tabela 3.6. Experimento em casa de vegetação registrando execução, câmara e bancada.

run	chamber	bench	pot	inoculum	severity
1	C1	2	14	10^5	18
1	C1	2	15	10^6	31

3.5.4 Dados de progresso da doença

Quando a doença é avaliada repetidamente nas mesmas unidades, cada avaliação se torna uma linha separada, identificada pela unidade e pela data ou pelo tempo desde um evento de referência (semeadura, inoculação ou primeiros sintomas). Tanto a data do calendário quanto o tempo decorrido costumam ser úteis.

Tabela 3.7. Avaliações repetidas ao longo do tempo na mesma parcela.

plot	treatment	date	days_after_planting	severity
101	control	2025-06-10	35	2
101	control	2025-06-24	49	10

3.5.5 Dados meteorológicos

Registros meteorológicos geralmente são coletados em intervalos regulares por estações automáticas. Cada linha é um instante, e cada coluna é uma variável meteorológica. As tabelas de clima normalmente se ligam às tabelas de doença pela data e pelo local.

Tabela 3.8. Registros meteorológicos horários.

station	datetime	temp_c	rh_pct	rain_mm	leaf_wet_min
ST01	2025-06-10 06:00	14.2	96	0.0	60
ST01	2025-06-10 07:00	15.8	91	0.2	45

3.5.6 Dados moleculares

Estudos moleculares frequentemente produzem muitas variáveis por amostra, como os níveis de expressão de muitos genes. Normalmente são necessárias duas tabelas: uma tabela de amostras, que descreve cada amostra biológica (tratamento, tempo, repetição), e uma tabela de medidas. Quando o número de variáveis é grande, a tabela de medidas pode ser guardada em formato longo, com uma linha por amostra e gene, ou no formato de matriz exigido por programas especializados, mas a informação das amostras deve sempre ficar em uma tabela separada e organizada.

Tabela 3.9. Dados de expressão gênica em formato longo; `hpi` = horas após a inoculação.

sample	treatment	hpi	bio_rep	gene	ct
S01	mock	24	1	PR1	27.3
S01	mock	24	1	EF1a	18.9
S02	inoculated	24	1	PR1	22.1

3.5.7 Dados de levantamento

Levantamentos registram informações sobre lavouras, propriedades ou regiões, muitas vezes combinando localização, práticas de manejo e observações de doença. Coordenadas, datas e o procedimento de amostragem devem ser registrados para que os resultados possam ser mapeados, relacionados ao clima ou comparados entre anos.

Tabela 3.10. Dados de levantamento com uma linha por lavoura visitada.

field	lat	lon	date	cultivar	tillage	incidence
F001	−20.7542	−42.8821	2025-03-12	BRS 123	no-till	15
F002	−20.8123	−42.9010	2025-03-12	BRS 456	conv	40

3.6 Tipos de variáveis

O tipo de uma variável determina como ela deve ser registrada, resumida, representada graficamente e analisada. Identificar os tipos de variáveis na organização dos dados prepara o terreno para as escolhas feitas nos capítulos seguintes.

3.6.1 Variáveis contínuas

Variáveis contínuas podem assumir qualquer valor dentro de um intervalo, limitadas apenas pela precisão da medida: comprimento da lesão, diâmetro da colônia, temperatura, produtividade ou taxa fotossintética. Devem ser registradas como números, com tanta precisão quanto a medida permitir, e em uma única unidade para toda a coluna.

A severidade expressa em porcentagem é frequentemente tratada como contínua, mas é limitada entre 0 e 100. Quando a severidade é estimada visualmente, os valores podem se concentrar em certos números (5, 10, 25), refletindo o método de avaliação e não a biologia. Essas características importam mais tarde, na escolha de gráficos e modelos.

3.6.2 Contagens

Contagens são números inteiros não negativos: lesões por folha, pústulas por cm², esporos por armadilha ou plantas infectadas por parcela. Devem ser registradas como inteiros. Quando uma contagem é convertida em proporção (plantas infectadas divididas por plantas avaliadas), tanto o numerador quanto o denominador devem ser mantidos, porque o denominador carrega informação sobre a precisão.

3.6.3 Categorias

Variáveis categóricas classificam as observações em grupos: tratamento, cultivar, isolado, bloco ou tipo de solo. As categorias devem ser registradas com rótulos consistentes. Algumas variáveis categóricas têm uma ordem natural, como as classes de severidade em uma escala ordinal (0 a 5) ou os níveis de resistência (resistente, moderadamente resistente, suscetível). Categorias

ordenadas são registradas como categorias, e não como medidas contínuas, mesmo quando codificadas com números.

Números usados como rótulos merecem atenção especial. Um bloco codificado como 1, 2, 3, 4 ou um tratamento codificado de 1 a 6 é categórico, embora pareça numérico. Os programas tratarão essas colunas como números se não forem instruídos do contrário, o que pode levar a análises sem sentido, como calcular a média dos números dos blocos.

3.6.4 Variáveis binárias

Variáveis binárias têm apenas dois valores possíveis: infectado ou não infectado, presente ou ausente, germinado ou não germinado. O melhor é registrá-las de forma consistente como 0 e 1 ou com dois rótulos claramente definidos (por exemplo, `yes` e `no`). Misturar códigos como `S`, `sim`, `1` e `+` na mesma coluna gera erros.

3.6.5 Datas e horários

Datas e horários são essenciais na Fitopatologia, porque a doença se desenvolve ao longo do tempo e depende do clima. Devem ser registrados em um formato sem ambiguidade. O padrão internacional ISO 8601 (`AAAA-MM-DD`, por exemplo `2025-06-10`) evita a confusão entre as convenções dia-mês e mês-dia e ordena corretamente como texto. Os horários devem incluir o fuso horário ou ser claramente documentados como hora local.

Programas de planilha frequentemente convertem datas, sem aviso, em números internos ou as reinterpretam de acordo com as configurações regionais. Conferir as datas após a importação é, portanto, uma etapa importante da validação.

3.6.6 Cadeias de caracteres

Texto livre, como anotações de campo, comentários do avaliador ou descrições de sintomas, é registrado como cadeia de caracteres. O texto é um contexto valioso, mas difícil de analisar. A informação que será usada na análise deve ser registrada em variáveis próprias, codificadas, enquanto o texto livre deve ficar em uma coluna separada de `notes` ou `comments`, e não misturado com valores numéricos.

3.6.7 Valores perdidos

Valores perdidos indicam que uma observação foi planejada, mas não obtida: uma planta morreu, uma parcela foi danificada, uma amostra se perdeu ou uma avaliação não foi feita. Os valores perdidos devem ser registrados de forma consistente, de preferência como células vazias ou como `NA`, e nunca como zero. Zero significa que a grandeza foi medida e é igual a zero; um valor perdido significa que nada se sabe.

O motivo da perda também deve ser documentado sempre que possível. Valores perdidos porque as plantas morreram da doença carregam informação diferente de valores perdidos porque uma etiqueta se extraviou. Essa distinção se torna importante em análises posteriores.

> ⚠ **Aviso**
>
> Nunca use zero, um espaço em branco, um traço ou um número implausível como −99 ou 999 para indicar dado perdido, a menos que o código esteja documentado e seja tratado de forma consistente na análise. Zero é uma medida válida, não um valor perdido.

Tabela 3.11. Tipos comuns de variáveis e como registrá-las.

Tipo de variável	Exemplos em Fitopatologia	Como registrar
Contínua	Comprimento da lesão, produtividade, temperatura	Número, uma única unidade por coluna
Contagem	Lesões por folha, esporos por armadilha	Inteiro, manter denominadores
Categórica (nominal)	Tratamento, cultivar, isolado	Rótulos de texto consistentes
Categórica (ordinal)	Classes de severidade, níveis de resistência	Rótulos ou códigos ordenados e documentados
Binária	Infectado ou não, presente ou ausente	0/1 ou dois rótulos definidos
Data e hora	Data de avaliação, horário da inoculação	ISO 8601 (`AAAA-MM-DD`)
Perdido	Planta morta, amostra perdida	Célula vazia ou `NA`, nunca 0

3.7 Metadados

3.7.1 O que são metadados?

Metadados são dados sobre os dados. Descrevem o que o conjunto de dados contém, como, quando, onde e por quem foi coletado e como deve ser interpretado. Sem metadados, um conjunto de dados é uma tabela de números cujo significado depende da memória de alguém.

Os metadados operam em vários níveis. Os metadados do projeto descrevem o estudo: objetivos, local, datas, delineamento e responsáveis. Os metadados do conjunto de dados descrevem cada arquivo: conteúdo, unidade de observação, número de linhas e relação com outros arquivos. Os metadados das variáveis descrevem cada coluna: nome, definição, unidades, valores permitidos e codificação.

Bons metadados também são centrais para os princípios que tornam os dados de pesquisa localizáveis, acessíveis, interoperáveis e reutilizáveis, conhecidos como princípios FAIR (Wilkinson et al. 2016). Dados bem documentados podem ser reutilizados em novas análises, combinados em meta-análises e verificados por outros.

3.7.2 Dicionários de dados

Um dicionário de dados é uma tabela que documenta todas as variáveis de um conjunto de dados. É a forma mais prática de metadados das variáveis e deve ser criado no início do estudo, e não no fim. Um dicionário simples inclui o nome da variável, uma descrição, o tipo, as unidades e os valores ou a faixa permitidos.

variável	descrição	tipo	unidades	valores permitidos
plot	Identificador da parcela	categórica	—	101–416
block	Bloco experimental	categórica	—	1, 2, 3, 4
treatment	Programa de fungicida	categórica	—	control, fung_A, fung_B
date	Data da avaliação	data	—	AAAA-MM-DD
severity	Porcentagem da área foliar com sintomas (média de 10 folhas)	contínua	%	0–100

variável	descrição	tipo	unidades	valores permitidos
yield_kg_ha	Produtividade comercial	contínua	kg/ha	≥ 0

Os nomes das variáveis devem ser curtos, informativos e utilizáveis pelos programas de análise. Bons nomes evitam espaços, acentos, caracteres especiais e números no início, e usam um estilo consistente. Uma convenção muito usada é o `snake_case`, em que as palavras são escritas em minúsculas e separadas por sublinhado: `disease_severity`, `days_after_planting`, `yield_kg_ha`.

Nomes como `Sev (%)`, `Severidade - 10 jun` ou `2ª avaliação` podem ser legíveis em uma planilha, mas causam problemas nos programas e no código. Descrições detalhadas pertencem ao dicionário de dados, e não ao nome da coluna.

3.7.3 Unidades de medida

Toda variável quantitativa deve ter uma unidade documentada, e cada coluna deve usar uma única unidade do início ao fim. Incluir a unidade no nome da variável (`temp_c`, `yield_kg_ha`, `lesion_mm`) ajuda a evitar erros, especialmente quando os conjuntos de dados são compartilhados ou combinados. As unidades nunca devem ser digitadas dentro das células (`25 mm`), porque isso transforma números em texto.

3.7.4 Sistemas de codificação

Códigos são rótulos curtos usados para representar categorias, como `fung_A` para um programa de fungicida ou `R`, `MR` e `S` para classes de resistência. Os códigos economizam espaço e reduzem erros de digitação, mas precisam ser documentados. Todo código usado no conjunto de dados deve aparecer no dicionário com seu significado completo. As escalas de notas devem ser documentadas com uma descrição ou referência para cada classe.

3.7.5 Vocabulários controlados

Um vocabulário controlado é uma lista predefinida de valores permitidos para uma variável. Usar vocabulários controlados para cultivares, tratamentos, espécies de patógenos ou estádios de desenvolvimento impede que a mesma categoria seja escrita de várias formas. Os programas de planilha permitem listas suspensas (validação de dados) que restringem as entradas aos valores permitidos. Para algumas variáveis já existem padrões estabelecidos pela comunidade, como os nomes científicos dos organismos ou as escalas padronizadas de estádios fenológicos das culturas, e estes devem ser preferidos a rótulos improvisados.

3.8 Convenções de nomes

Nomes consistentes para variáveis, arquivos e pastas tornam os dados mais fáceis de encontrar, entender e processar. A convenção específica importa menos do que aplicá-la de forma consistente em todo o projeto.

3.8.1 Variáveis

Os nomes das variáveis seguem as orientações apresentadas acima: minúsculas, sem espaços, palavras separadas por sublinhado, unidades incluídas quando relevante e o mesmo nome para a mesma variável em todas as tabelas e arquivos. Se as parcelas se chamam `plot` em uma tabela, não devem se chamar `Plot_ID` em outra.

3.8.2 Arquivos

Os nomes de arquivos devem descrever o conteúdo, ser legíveis por pessoas e por computadores e ordenar-se de forma significativa. Regras úteis incluem:

- sem espaços nem caracteres especiais;
- datas escritas como `AAAA-MM-DD` no início, quando a ordem cronológica importa;
- palavras descritivas separadas por sublinhado ou hífen;
- números de versão explícitos, em vez de palavras como `final` ou `novo`.

Tabela 3.12. Exemplos de nomes de arquivos ruins e melhores.

Nome de arquivo ruim	Nome de arquivo melhor
`dados final NOVO.xlsx`	`2025_basil_trial_severity_raw.csv`
`ensaio(2)-corrigido.xls`	`2025_basil_trial_severity_clean_v02.csv`
`clima junho.xlsx`	`2025-06_station01_weather_hourly.csv`
`Analise.R`	`02_fit_severity_models.R`

Formatos de texto simples, como CSV (valores separados por vírgula), são preferíveis para guardar os dados usados na análise. Podem ser lidos por qualquer programa, não escondem informação em formatação e continuam legíveis à medida que os programas mudam.

3.8.3 Pastas

As pastas agrupam arquivos relacionados e separam os dados em diferentes etapas de processamento. Uma estrutura típica separa dados brutos, dados processados, scripts e resultados, como descrito mais adiante neste capítulo. Os nomes das pastas seguem as mesmas convenções dos nomes de arquivos.

3.8.4 Projetos

Cada projeto de pesquisa — seja um capítulo de tese, um experimento ou um artigo — deve ter sua própria pasta autocontida, com tudo o que é necessário para reproduzir a análise. Um breve arquivo `README` no nível superior explica o objetivo do projeto, o conteúdo de cada pasta e como executar a análise. Números de versão ou datas devem ser usados quando conjuntos de dados ou análises forem atualizados, para que os resultados possam sempre ser rastreados até a versão exata dos dados utilizada.

3.9 Problemas comuns na digitação de dados

Muitos erros são introduzidos quando as observações são transferidas de fichas de campo, instrumentos ou cadernos para arquivos digitais. Conhecer os problemas mais comuns ajuda a preveni-los e facilita detectá-los.

3.9.1 Erros de digitação

Erros de digitação incluem dígitos trocados (`52` em vez de `25`), vírgulas decimais deslocadas (`345` em vez de `3,45`) e valores digitados na linha ou coluna errada. Esses erros são especialmente perigosos porque produzem números plausíveis. A dupla digitação, em que duas pessoas digitam os mesmos dados de forma independente e os arquivos são comparados, é uma forma eficaz de detectá-los em conjuntos de dados críticos.

3.9.2 Codificação inconsistente

A mesma categoria é frequentemente escrita de formas diferentes: `Controle`, `controle`, `CTRL`, `testemunha` e `controle` (com um espaço no final). Os programas tratam cada versão como uma categoria diferente, e a análise pode relatar cinco tratamentos em vez de um. Vocabulários controlados e listas suspensas evitam esse problema.

3.9.3 Valores perdidos ocultos

Valores perdidos podem estar ocultos como zeros, espaços em branco, traços, `n/a`, `-`, `?` ou números como `-99`. Cada um pode ser interpretado de forma diferente pelos programas, e os zeros, em particular, são indistinguíveis de medidas verdadeiras. Os valores perdidos devem ser codificados de uma única forma documentada.

3.9.4 Unidades misturadas

Valores na mesma coluna podem estar em unidades diferentes: comprimento de lesão em milímetros e centímetros, produtividade em kg/ha e t/ha, temperatura em °C e °F. Unidades misturadas são comuns quando conjuntos de dados de diferentes fontes ou avaliadores são combinados. Elas produzem valores que parecem razoáveis individualmente, mas são inconsistentes em conjunto.

3.9.5 Observações duplicadas

A mesma observação pode ser digitada duas vezes, especialmente quando os dados são copiados entre arquivos ou combinados a partir de diferentes planilhas. As duplicatas inflam o tamanho aparente da amostra e dão mais peso a algumas unidades. Verificar se os identificadores são únicos detecta a maioria das duplicatas.

3.9.6 Valores impossíveis

Valores impossíveis estão fora da faixa que a variável pode assumir: severidade de 120%, contagem de lesões negativa, umidade relativa acima de 100% ou data de avaliação anterior à semeadura. Esses valores indicam erros de digitação, de medida ou de cálculo e devem ser investigados, e não removidos silenciosamente.

3.9.7 Problemas de formatação em planilhas

As planilhas incentivam práticas que dificultam a análise dos dados (Broman e Woo 2018):

- células mescladas, que deixam valores apenas na primeira célula do intervalo mesclado;
- cores ou negrito usados para codificar informação, como destacar plantas doentes;
- várias tabelas na mesma planilha, separadas por linhas em branco;
- linhas de resumo (médias, totais) inseridas no meio dos dados;
- comentários, unidades ou símbolos digitados dentro de células numéricas (`12*`, `25 mm`, `<0,1`);
- conversão automática de valores em datas (o nome de um gene ou isolado, como `MARCH1` ou `1-2`, virando data).

Uma planilha usada para guardar dados deve conter uma única tabela retangular por aba, com uma única linha de cabeçalho, sem células mescladas, sem formatação que carregue significado e sem cálculos misturados aos dados.

⚠ Aviso

> Muitos erros surgem na digitação dos dados, e não na análise estatística.

3.10 Validação dos dados

A validação dos dados é o processo de verificar se um conjunto de dados está completo, consistente e plausível antes de ser analisado. A validação não garante que os dados estejam corretos, mas detecta muitos erros enquanto ainda é possível rastreá-los até a origem.

3.10.1 Verificação de faixas

A verificação de faixas confere se cada valor está dentro da faixa possível ou esperada para sua variável. A severidade deve estar entre 0 e 100%, as contagens devem ser inteiros não negativos e as datas devem estar dentro do período do estudo. A faixa permitida de cada variável deve constar do dicionário de dados, o que torna a verificação de faixas simples.

3.10.2 Verificação de consistência

A verificação de consistência confere se os valores concordam entre si e com o delineamento do estudo. Cada parcela deve pertencer a um único tratamento e a um único bloco, cada tratamento deve aparecer uma vez em cada bloco de um DBC, e o número de linhas deve corresponder ao número de observações planejadas. Relações lógicas também podem ser verificadas: a incidência não pode ser zero quando a severidade é maior que zero, e a data dos primeiros sintomas não pode ser anterior à inoculação.

3.10.3 Conferência com os registros originais

Conferir a digitação significa comparar o conjunto de dados digital com uma fonte independente: as fichas de campo originais, fotografias das fichas, arquivos de saída dos instrumentos ou uma segunda digitação independente dos mesmos dados. Mesmo conferir uma amostra aleatória de registros com as fichas originais fornece uma estimativa útil da taxa de erro.

3.10.4 Resumos exploratórios

Resumos simples revelam muitos problemas rapidamente. Contar o número de observações por tratamento, bloco e data revela registros faltantes ou duplicados. Listar os valores distintos de cada variável categórica revela codificação inconsistente. Calcular os valores mínimo e máximo revela valores impossíveis. Representar graficamente cada variável revela valores discrepantes e padrões inesperados. Esses resumos são os primeiros passos da análise exploratória de dados apresentada no Capítulo 4.

Tabela 3.13. Verificações comuns na validação dos dados.

Verificação	Pergunta feita	Exemplo de problema detectado
Faixa	Cada valor é possível?	Severidade de 150%
Consistência	Os valores concordam entre si e com o delineamento?	Parcela atribuída a dois tratamentos
Unicidade	Cada identificador é único?	Parcela 203 digitada duas vezes na mesma data
Completude	Todas as observações planejadas estão presentes?	Um bloco faltando na última data de avaliação
Codificação	As categorias estão escritas de forma consistente?	`control`, `Control` e `CTRL`
Conferência	O arquivo corresponde aos registros originais?	Dígitos trocados na produtividade

Quando um erro é encontrado, a correção deve ser documentada: o que foi alterado, por quê e com base em quê. Valores não devem ser apagados ou alterados apenas porque parecem incomuns. Um valor extremo pode ser uma observação real, e removê-lo sem justificativa é uma forma de manipulação de dados.

3.11 Dados brutos versus dados processados

3.11.1 Nunca sobrescreva os dados brutos

Dados brutos são os registros originais das observações, tal como foram coletados: fichas de campo preenchidas, arquivos de saída de instrumentos, imagens e a primeira transcrição digital das observações. Os dados brutos devem ser guardados com segurança, protegidos contra modificações acidentais e nunca sobrescritos. Todo o trabalho posterior deve ser feito em cópias ou, de preferência, por meio de scripts que leem os arquivos brutos e gravam novos arquivos.

Manter os dados brutos intactos permite corrigir erros cometidos no processamento, aplicar decisões de processamento diferentes e demonstrar exatamente como os dados analisados foram obtidos.

> **! Importante**
>
> Trate os dados brutos como somente leitura. Toda alteração deve ser feita em um novo arquivo, e toda alteração deve ser documentada.

3.11.2 Variáveis derivadas

Variáveis derivadas são calculadas a partir das variáveis registradas: a incidência calculada a partir do número de plantas doentes e avaliadas, a severidade média por parcela calculada a partir das avaliações por folha, os dias após a semeadura calculados a partir das datas, ou a produtividade relativa calculada a partir das parcelas tratadas e não tratadas. As variáveis derivadas devem ser calculadas por procedimentos documentados, de preferência em um script, e não digitadas à mão. Isso garante que o cálculo seja consistente para todas as observações e possa ser refeito se os dados brutos mudarem.

3.11.3 Limpeza dos dados

A limpeza dos dados é o processo de transformar os dados brutos em um conjunto pronto para a análise. Inclui corrigir erros documentados, harmonizar códigos e unidades, recodificar valores perdidos, reorganizar tabelas no formato organizado e unir tabelas relacionadas. Cada decisão de limpeza deve ser registrada. Quando a limpeza é feita com um script, o próprio script é o registro: qualquer pessoa pode lê-lo para ver exatamente o que foi feito, e executá-lo novamente reproduz o conjunto processado a partir dos dados brutos.

3.11.4 Controle de versões

Conjuntos de dados e análises mudam durante um projeto, à medida que novos dados são coletados e erros são corrigidos. Controle de versões significa acompanhar essas mudanças para que seja sempre possível saber qual versão dos dados produziu qual resultado. Na forma mais simples, consiste em salvar novas versões com números de versão e datas explícitos e manter um registro das alterações. Ferramentas mais poderosas, como o Git, registram automaticamente todas as mudanças em scripts e arquivos de texto e são amplamente usadas na pesquisa reprodutível.

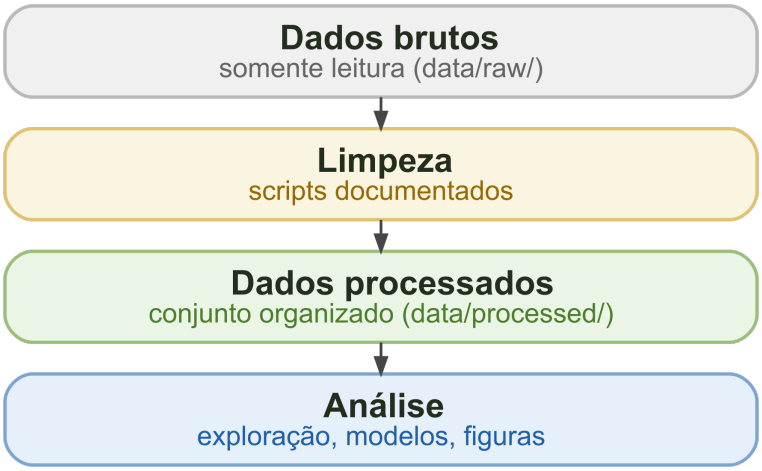


Figura 3.3. Fluxo de trabalho recomendado para os dados. Os dados brutos nunca são alterados; a limpeza, feita por scripts documentados, gera os dados processados usados na análise.

3.12 Organizando projetos de pesquisa

Uma estrutura de projeto consistente facilita encontrar arquivos, separar os dados em diferentes etapas e reproduzir as análises. A estrutura exata varia com o projeto, mas um modelo simples funciona bem para a maioria dos estudos em Fitopatologia.

```
basil_fungicide_trial/
├── README.md
├── data/
│   ├── raw/
│   └── processed/
├── scripts/
├── figures/
├── results/
└── manuscript/
```

3.12.1 Pastas do projeto

Cada projeto tem uma única pasta principal contendo tudo o que é necessário para entender e reproduzir o trabalho. Um arquivo `README` no nível superior explica o objetivo do projeto, as fontes dos dados, o conteúdo de cada pasta e a ordem em que os scripts devem ser executados.

3.12.2 Dados

A pasta `data` contém os conjuntos de dados, separados nas subpastas `raw` (brutos) e `processed` (processados). Os arquivos brutos nunca são modificados. Os arquivos processados são gerados a partir dos brutos por scripts e podem ser apagados e recriados a qualquer momento. O dicionário de dados e outros metadados ficam junto com os dados.

3.12.3 Scripts

A pasta `scripts` contém o código usado para limpar, analisar e visualizar os dados. Numerar os scripts na ordem em que devem ser executados (`01_clean_data.R`, `02_explore.R`, `03_fit_models.R`) torna o fluxo de trabalho fácil de acompanhar.

3.12.4 Figuras

A pasta `figures` contém os gráficos produzidos pelos scripts. Como as figuras são geradas por código, podem ser atualizadas automaticamente quando os dados ou as análises mudam, em vez de serem editadas à mão.

3.12.5 Resultados

A pasta `results` contém tabelas, saídas de modelos e outros produtos da análise. Assim como as figuras, os resultados devem ser gerados por scripts, para que todo número relatado possa ser rastreado até os dados e o código que o produziram.

3.12.6 Manuscritos

A pasta `manuscript` contém versões de relatórios, teses ou artigos. Ferramentas como o Quarto, usado para escrever este livro, permitem combinar texto, código, resultados e figuras em um único documento, de modo que o manuscrito seja atualizado automaticamente quando a análise muda.

3.13 Pesquisa na prática

Pesquisa revisitada: Organizando os dados do ensaio de fungicidas em manjericão

Contexto

O Capítulo 1 apresentou experimentos de campo que avaliaram fungicidas para o controle do míldio do manjericão, e o Capítulo 2 examinou seu delineamento: tratamentos com fungicidas e uma testemunha dispostos em blocos casualizados com quatro parcelas por tratamento, com a doença avaliada periodicamente durante a safra (McGrath 2020). Aqui, acompanhamos como os dados de um experimento desse tipo passam da ficha de campo a um conjunto de dados organizado. Os valores abaixo são ilustrativos e não reproduzem os dados publicados.

Etapa 1: a planilha bruta

Uma primeira transcrição típica das anotações de campo poderia ser assim:

Tabela 3.14. Planilha bruta, como digitada a partir das anotações de campo (ilustrativa).

Bloco	Parcela	Trat	10-jun	24-jun	8-jul	Obs
I	101	Controle	2%	10	35	
I	102	Fung A	0	2	6	
I	103	fung A	-	1	4	parcela alagada
II	201	CTRL	3	12	40	
	Média				**21,3**	

Essa planilha contém vários problemas: as datas de avaliação estão nos nomes das colunas; a testemunha está escrita de três formas diferentes, e o fungicida A, com maiúsculas e espaços inconsistentes; um sinal de porcentagem foi digitado em uma célula; um traço foi usado para um valor perdido; a parcela 103 parece repetir o tratamento da parcela 102 no mesmo bloco, o que precisa ser conferido no croqui; e uma linha de resumo com uma média foi inserida no meio dos dados.

Etapa 2: a planilha limpa

As seguintes alterações foram feitas e registradas em um registro de limpeza:

1. Os rótulos dos tratamentos foram harmonizados segundo um vocabulário controlado (`control` , `fung_A` , `fung_B` , ...).
2. O sinal de porcentagem foi removido para que a severidade fique registrada como número; a unidade (%) foi documentada no dicionário de dados.
3. O traço foi substituído por `NA` , e o motivo (parcela alagada antes da primeira avaliação) foi mantido na coluna `notes` .
4. A parcela 103 foi conferida no croqui e na ficha original, e seu tratamento foi corrigido para `fung_B` .
5. A linha de resumo foi apagada; as médias serão calculadas na análise.
6. Os números dos blocos foram convertidos de algarismos romanos para inteiros e documentados como variável categórica.
7. Os nomes das colunas foram convertidos para `snake_case` em minúsculas.

Etapa 3: o conjunto de dados organizado

Por fim, a tabela foi reorganizada de modo que cada linha represente uma parcela em uma data de avaliação:

Tabela 3.15. Conjunto de dados organizado: uma linha por parcela e data de avaliação (ilustrativo).

block	plot	treatment	date	severity	notes
1	101	control	2025-06-10	2	
1	101	control	2025-06-24	10	
1	101	control	2025-07-08	35	
1	102	fung_A	2025-06-10	0	
1	103	fung_B	2025-06-10	NA	parcela alagada
...	...	...	...	...	...

A planilha bruta foi preservada sem alterações em `data/raw/` , e o conjunto organizado foi salvo como um novo arquivo CSV em `data/processed/` , junto com o dicionário de dados e o registro de limpeza.

Questões

1. Qual é a unidade de observação no conjunto de dados organizado?
2. Que combinação de variáveis identifica cada linha de forma única?
3. Por que a linha de resumo deve ser removida, em vez de mantida no conjunto de dados?
4. Por que é importante manter o motivo do valor perdido?
5. Como os rótulos inconsistentes dos tratamentos poderiam ter sido evitados na digitação?

3.14 Lista de verificação da organização dos dados

Antes de iniciar qualquer análise, verifique:

1. As observações estão claramente definidas?
2. Cada linha é uma observação?
3. Cada coluna é uma variável?
4. Os identificadores são únicos?
5. As unidades estão documentadas?
6. Os valores perdidos estão codificados de forma consistente?
7. Os dados brutos estão preservados?
8. Os metadados estão completos?
9. Os nomes dos arquivos são informativos?
10. O conjunto de dados está pronto para a análise?

3.15 Resumo do capítulo

A organização dos dados liga o delineamento do estudo à análise. Um conjunto de dados é uma coleção estruturada de observações em que a unidade de observação é explícita, cada linha representa uma observação, cada coluna representa uma variável e identificadores ligam cada registro à unidade da qual foi obtido. Esses princípios de dados organizados simplificam a análise e tornam visível a estrutura do estudo. Reconhecer os tipos de variáveis, documentá-las em um dicionário de dados e usar nomes consistentes para variáveis, arquivos e pastas tornam os conjuntos de dados compreensíveis para outras pessoas e para os próprios pesquisadores. Muitos erros surgem na digitação, e verificações simples de validação detectam a maioria deles antes que afetem os resultados. Os dados brutos devem ser preservados sem alterações, e cada etapa entre os dados brutos e o conjunto analisado deve ser documentada, de preferência em scripts, para que a análise possa ser reproduzida.

3.16 Conceitos-chave

Ao concluir este capítulo, você deverá ser capaz de explicar:

- conjunto de dados;
- observação;
- variável;
- identificador;
- unidade de observação;
- dados organizados (*tidy data*);

- metadados;
- dicionário de dados;
- vocabulário controlado;
- valores perdidos;
- validação dos dados;
- dados brutos;
- dados processados;
- variável derivada;
- reprodutibilidade.

3.17 Exercícios

3.17.1 Exercício 1. Identificando problemas em conjuntos de dados mal organizados

Examine as três planilhas fornecidas pelo professor. Para cada uma, liste todos os problemas encontrados, incluindo estrutura não organizada, codificação inconsistente, valores perdidos ocultos, unidades misturadas, observações duplicadas, valores impossíveis e problemas de formatação. Explique como cada problema poderia afetar a análise.

3.17.2 Exercício 2. Convertendo um conjunto de dados para o formato organizado

A tabela abaixo registra o número de lesões em folhas de três cultivares avaliadas em duas épocas.

cultivar	rep	lesions_7dai	lesions_14dai
A	1	3	12
A	2	5	15
B	1	0	2
B	2	1	4

1. Que informação está guardada nos nomes das colunas?
2. Reescreva a tabela no formato organizado.
3. Quantas linhas tem a tabela organizada?
4. Quais variáveis identificam cada linha de forma única?

3.17.3 Exercício 3. Criando uma tabela de metadados

Escolha um conjunto de dados da sua própria pesquisa ou um fornecido pelo professor e construa um dicionário de dados contendo, para cada variável, o nome, a descrição, o tipo, as unidades e os valores ou a faixa permitidos.

3.17.4 Exercício 4. Revisando os dados suplementares de um artigo publicado

Baixe os dados suplementares de um artigo publicado em Fitopatologia. Você conseguiria reproduzir a análise principal usando apenas esses arquivos? Que informação está faltando? O conjunto de dados está organizado? As variáveis, unidades e códigos estão documentados?

3.17.5 Exercício 5. Planejando uma estrutura de pastas

Planeje uma estrutura de pastas para um projeto de pesquisa de sua escolha, como um capítulo de tese ou um ensaio de campo de vários anos. Escreva um breve arquivo `README` descrevendo o conteúdo de cada pasta e as convenções de nomes dos arquivos.

3.18 Reflexão

> **Boas análises começam com dados bem organizados.**

Discuta por que essa afirmação é verdadeira, usando um exemplo da sua própria experiência ou de um estudo publicado.

3.19 Próximos passos

Uma vez organizados os dados, o passo seguinte é entender o que eles contêm antes de aplicar métodos estatísticos. O Capítulo 4 apresenta a análise exploratória de dados e a exploração gráfica.

CAPÍTULO 4

O que os dados revelam?

4.1 Objetivos de aprendizagem

! Importante

Ao concluir este capítulo, o leitor deverá ser capaz de:

- explicar o papel da análise exploratória de dados (AED) na pesquisa científica;
- distinguir a exploração da inferência estatística;
- resumir variáveis individuais numérica e graficamente;
- identificar observações incomuns, valores perdidos e possíveis erros;
- explorar relações entre variáveis;
- reconhecer padrões que orientam as análises estatísticas subsequentes;
- formular questões biológicas que surgem das análises exploratórias.

4.2 Por que explorar os dados?

Depois que os dados foram coletados e organizados, é tentador passar diretamente ao teste estatístico que responde à questão do projeto de pesquisa. Pesquisadores experientes raramente fazem isso. Antes de estimar efeitos ou testar hipóteses, eles olham os dados: quantas observações existem, como os valores se distribuem, se algo está faltando ou é implausível e que padrões aparecem entre as variáveis. Essa etapa é chamada **análise exploratória de dados** (AED).

O termo foi popularizado pelo estatístico John Tukey, que defendia que olhar os dados com a mente aberta é tão importante para a ciência quanto confirmar expectativas, e que os números devem ser examinados graficamente e com resumos simples e robustos antes da aplicação de modelos formais (Tukey 1977). Tukey comparava a análise exploratória ao trabalho de um detetive: ela procura pistas, enquanto a análise confirmatória avalia a força das provas no tribunal.

A exploração serve a vários propósitos. Ela revela erros que escaparam à validação dos dados (Capítulo 3), como um valor de severidade de 250% ou um tratamento que aparece em apenas três dos quatro blocos. Mostra se os dados têm as características pressupostas pelos métodos planejados para a análise, como distribuições aproximadamente simétricas ou variabilidade semelhante entre grupos. Mostra o tamanho e a direção das diferenças antes de qualquer teste, o que mantém a atenção nos efeitos biológicos, e não nos valores-P. E frequentemente revela

padrões não previstos, como um gradiente entre blocos ou a resposta incomum de uma cultivar, que podem se tornar os resultados mais interessantes do estudo.

> **! Importante**
>
> A análise exploratória de dados não é feita para confirmar hipóteses, mas para compreender os dados antes das análises estatísticas formais.

4.3 Análise exploratória versus confirmatória

4.3.1 O que é análise exploratória de dados?

A análise exploratória de dados é um exame aberto dos dados, baseado principalmente em gráficos e resumos simples, cujo objetivo é compreender o que os dados contêm. Ela faz perguntas como: Que valores cada variável assume? Quão variáveis são as observações? Há observações incomuns? Quais variáveis parecem estar relacionadas? Os grupos diferem, e em quanto?

A exploração é flexível. O pesquisador pode experimentar diferentes gráficos, olhar subconjuntos dos dados e seguir pistas que surgem pelo caminho. Seus produtos são descrições, perguntas e hipóteses, e não conclusões formais.

4.3.2 O que é análise confirmatória?

A análise confirmatória avalia a evidência para questões e hipóteses definidas com antecedência, usando métodos escolhidos de acordo com o delineamento do estudo. Ela estima efeitos com medidas de incerteza, como intervalos de confiança, e pode testar hipóteses estatísticas (Capítulo 6). A análise confirmatória segue regras fixadas antes de os resultados serem vistos. Essas regras protegem contra encontrar padrões que são simplesmente fruto do acaso.

4.3.3 Por que ambas são necessárias

Exploração e confirmação desempenham papéis complementares. Sem exploração, análises confirmatórias podem ser aplicadas a dados com erros, a variáveis cujas distribuições violam os pressupostos do método escolhido ou a questões que os dados não conseguem responder. Sem confirmação, os padrões exploratórios continuam sendo impressões: com gráficos e subconjuntos suficientes, algum padrão marcante sempre aparecerá, mesmo em dados sem nenhum efeito real.

A distinção importa sobretudo quando a exploração sugere uma nova hipótese. Um padrão descoberto ao explorar um conjunto de dados não pode ser testado de forma justa com os mesmos dados, porque eles foram usados para encontrá-lo. Esses padrões devem ser relatados como achados exploratórios e, sempre que possível, confirmados com novos dados de um estudo independente.

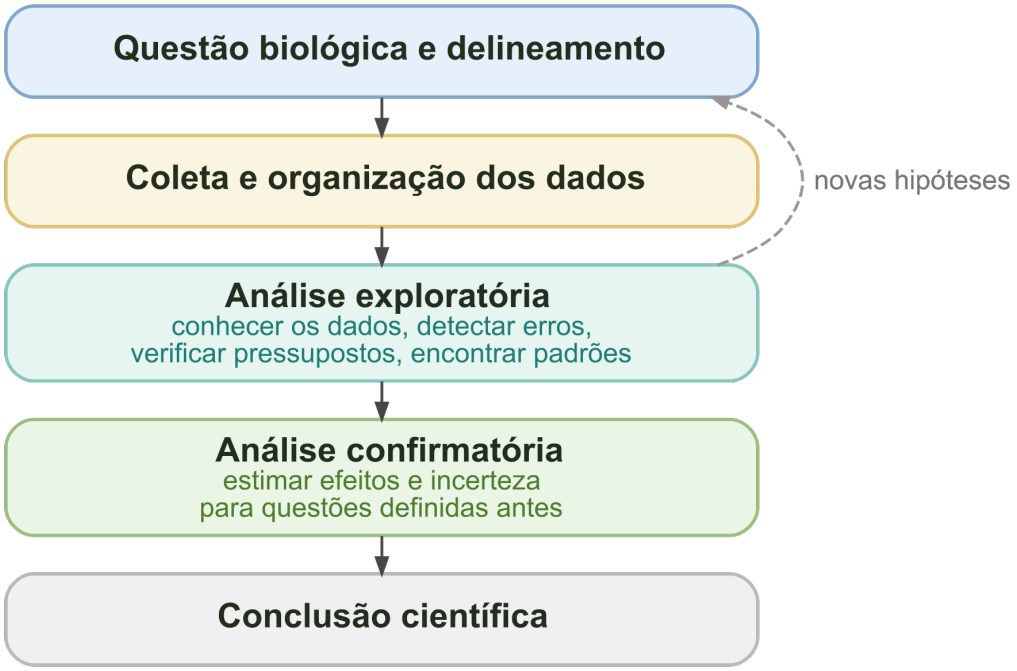


Figura 4.1. O papel da análise exploratória de dados no fluxo de trabalho científico. A exploração precede a análise confirmatória e frequentemente gera novas questões e hipóteses, que exigem novos estudos para serem confirmadas.

Tabela 4.1. Comparação entre análise exploratória e confirmatória.

Característica	Análise exploratória	Análise confirmatória
Objetivo principal	Compreender os dados, encontrar padrões	Avaliar a evidência para questões predefinidas
Questões	Abertas, podem mudar durante a análise	Definidas antes da análise
Principais ferramentas	Gráficos, resumos simples	Estimativas, intervalos, modelos, testes
Flexibilidade	Alta	Baixa; o plano de análise é fixo
Produto típico	Descrições, perguntas, hipóteses	Conclusões com incerteza quantificada
Principal risco	Superinterpretar padrões casuais	Aplicar métodos a dados que não se ajustam a eles

4.4 Conhecendo o seu conjunto de dados

Toda exploração começa com perguntas simples sobre o conjunto de dados como um todo. Os exemplos deste capítulo usam o R e funções do tidyverse, que são apresentados com mais detalhe em capítulos posteriores. O código é mostrado para ilustrar as etapas; as mesmas etapas podem ser feitas em qualquer programa estatístico.

```
library(tidyverse)

data ← read_csv("data/processed/field_trial.csv")
```

Para que os exemplos possam ser executados, este capítulo usa um conjunto de dados simulado com a mesma estrutura: um ensaio em blocos casualizados com seis tratamentos, quatro blocos

e quatro datas de avaliação. Como em dados reais, ele contém alguns problemas para a exploração encontrar.

4.4.1 Dimensões do conjunto de dados

A primeira verificação é o número de linhas e colunas. O número de linhas deve corresponder ao número de observações esperado pelo delineamento. Um DBC com seis tratamentos, quatro blocos e quatro datas de avaliação deve produzir 6 × 4 × 4 = 96 linhas. Se o conjunto de dados tiver 95 ou 100 linhas, algo está faltando ou duplicado, e o motivo precisa ser encontrado antes de prosseguir.

```
dim(data)    # número de linhas e colunas
```

```
[1] 96  6
```

```
nrow(data)
```

```
[1] 96
```

4.4.2 Nomes das variáveis

Em seguida, liste os nomes das variáveis e confira se correspondem ao dicionário de dados. Colunas inesperadas (como `...7`, criada quando a planilha tem uma coluna vazia com formatação) ou colunas ausentes indicam problemas de importação.

```
names(data)
```

```
[1] "plot"      "block"     "treatment" "date"      "severity"  "yield"
```

```
head(data)   # primeiras linhas do conjunto de dados
```

```
# A tibble: 6 × 6
   plot block treatment  date       severity yield
  <dbl> <fct> <fct>      <date>        <dbl> <dbl>
1   101 1     Testemunha 2025-06-10      2.3  2723
2   101 1     Testemunha 2025-06-24      5.8  2723
3   101 1     Testemunha 2025-07-08     15.3  2723
4   101 1     Testemunha 2025-07-22     42.2  2723
5   102 1     Fung. A    2025-06-10      1.6  3983
6   102 1     Fung. A    2025-06-24      3.8  3983
```

4.4.3 Classes das variáveis

Os programas atribuem uma classe a cada variável: numérica, inteira, caractere, fator, lógica ou data. Conferir as classes revela muitos problemas de uma vez. Uma coluna de severidade lida como caractere geralmente contém, em algum lugar, uma entrada não numérica, como `12*`, `n/a` ou `25%`. Uma coluna de bloco lida como numérica precisa ser convertida em variável categórica (fator) antes da análise. Uma data lida como caractere precisa ser convertida em data.

```
glimpse(data)  # nomes, classes e primeiros valores de cada variável
```

```
Rows: 96
Columns: 6
$ plot      <dbl> 101, 101, 101, 101, 102, 102, 102, 102, 103, 103, 103, 103, …
$ block     <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
$ treatment <fct> Testemunha, Testemunha, Testemunha, Testemunha, Fung. A, Fun…
$ date      <date> 2025-06-10, 2025-06-24, 2025-07-08, 2025-07-22, 2025-06-10,…
$ severity  <dbl> 2.3, 5.8, 15.3, 42.2, 1.6, 3.8, 4.0, 7.9, 1.9, 4.5, 5.1, 7.3…
$ yield     <dbl> 2723, 2723, 2723, 2723, 3983, 3983, 3983, 3983, 3965, 3965, …
```

4.4.4 Valores perdidos

Conte os valores perdidos em cada variável e verifique onde ocorrem. Poucos valores perdidos espalhados têm implicações diferentes de muitos valores perdidos concentrados em um tratamento, bloco ou data. Se as plantas morreram no tratamento mais suscetível, a perda está relacionada à própria resposta e pode enviesar as comparações.

```
colSums(is.na(data))           # valores perdidos por coluna
```

```
     plot     block treatment      date  severity     yield 
        0         0         0         0         2         0 
```

```
data |>
  filter(is.na(severity)) |>
  count(treatment, date)       # onde eles estão?
```

```
# A tibble: 2 × 3
  treatment date           n
  <fct>     <date>     <int>
1 Fung. C   2025-06-10     1
2 Fung. E   2025-07-08     1
```

4.4.5 Primeiros resumos

Por fim, calcule resumos simples de todas as variáveis e conte as observações em cada combinação dos fatores do delineamento. Esses resumos revelam valores impossíveis (mínimo ou máximo fora da faixa permitida), rótulos de categorias inconsistentes e delineamentos desbalanceados.

```
summary(data)
```

```
      plot       block       treatment       date                severity     
 Min.   :101.0   1:24   Testemunha:16   Min.   :2025-06-10   Min.   : 1.300  
 1st Qu.:177.2   2:24   Fung. A   :16   1st Qu.:2025-06-20   1st Qu.: 2.850  
 Median :253.5   3:24   Fung. B   :16   Median :2025-07-01   Median : 5.300  
 Mean   :253.5   4:24   Fung. C   :16   Mean   :2025-07-01   Mean   : 9.528  
 3rd Qu.:329.8          Fung. D   :16   3rd Qu.:2025-07-11   3rd Qu.:10.425  
 Max.   :406.0          Fung. E   :16   Max.   :2025-07-22   Max.   :97.200  
                                                             NA's   :2       
```

```
     yield
 Min.   :2670
 1st Qu.:3542
 Median :3864
 Mean   :3740
 3rd Qu.:4022
 Max.   :4392
```

```
data ▷ count(treatment)           # observações por tratamento
```

```
# A tibble: 6 × 2
  treatment      n
  <fct>      <int>
1 Testemunha    16
2 Fung. A       16
3 Fung. B       16
4 Fung. C       16
5 Fung. D       16
6 Fung. E       16
```

```
data ▷ count(treatment, block)     # o delineamento está completo?
```

```
# A tibble: 24 × 3
   treatment  block     n
   <fct>      <fct> <int>
 1 Testemunha 1         4
 2 Testemunha 2         4
 3 Testemunha 3         4
 4 Testemunha 4         4
 5 Fung. A    1         4
 6 Fung. A    2         4
 7 Fung. A    3         4
 8 Fung. A    4         4
 9 Fung. B    1         4
10 Fung. B    2         4
# ℹ 14 more rows
```

Dica

Compare o conjunto de dados com o delineamento do estudo: número de tratamentos, de blocos ou repetições, de datas e de unidades experimentais. Discrepâncias são quase sempre erros de digitação ou de importação.

4.5 Explorando uma variável

Depois da visão geral, cada variável importante é examinada individualmente. O objetivo é descrever seu centro, sua dispersão, a forma da sua distribuição e eventuais valores incomuns.

4.5.1 Resumos numéricos

Um pequeno conjunto de resumos descreve uma variável quantitativa:

- **média**: a média aritmética, a soma dos valores dividida pelo seu número;
- **mediana**: o valor central quando as observações são ordenadas; metade das observações está abaixo dela e metade acima;
- **desvio padrão**: uma medida da distância típica entre as observações e sua média, expressa nas unidades da variável;
- **quartis**: os valores abaixo dos quais estão 25% (primeiro quartil) e 75% (terceiro quartil) das observações; a distância entre eles é a amplitude interquartil (AIQ), que contém a metade central dos dados;
- **mínimo** e **máximo**: o menor e o maior valor, que definem a amplitude.

A média e o desvio padrão são sensíveis a valores extremos, enquanto a mediana e os quartis são resistentes a eles. Quando média e mediana diferem muito, a distribuição é assimétrica ou contém observações extremas. Por isso, é útil relatar ambas.

```
data ▷
  summarise(
    n      = sum(!is.na(severity)),
    mean   = mean(severity, na.rm = TRUE),
    median = median(severity, na.rm = TRUE),
    sd     = sd(severity, na.rm = TRUE),
    q1     = quantile(severity, 0.25, na.rm = TRUE),
    q3     = quantile(severity, 0.75, na.rm = TRUE),
    min    = min(severity, na.rm = TRUE),
    max    = max(severity, na.rm = TRUE)
  )
```

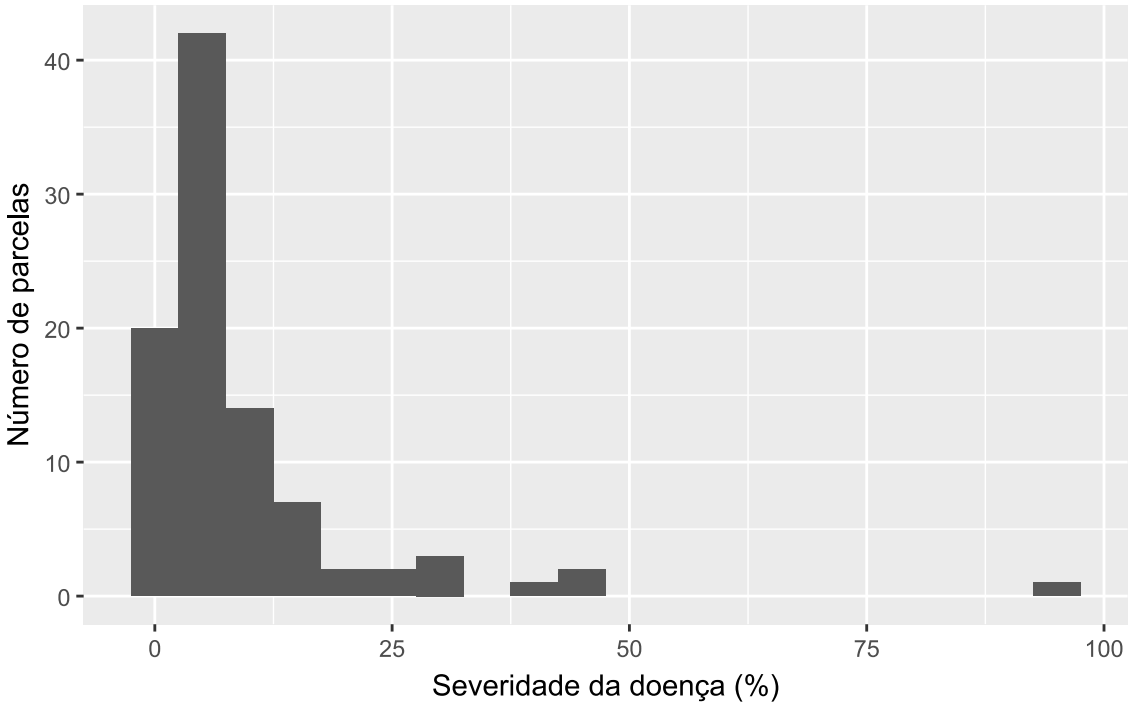
```
# A tibble: 1 × 8
      n  mean median    sd    q1    q3   min   max
  <int> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1    94  9.53    5.3  13.0  2.85  10.4   1.3  97.2
```

Para variáveis categóricas, os resumos relevantes são as contagens e proporções em cada categoria.

4.5.2 Distribuição

A distribuição de uma variável descreve quais valores ocorrem e com que frequência. Os resumos numéricos condensam a distribuição em poucos números; os gráficos a mostram diretamente. Histogramas, gráficos de densidade, gráficos de pontos e diagramas de caixa (*boxplots*) são as ferramentas usuais, e sua construção é discutida no Capítulo 5.

```
ggplot(data, aes(x = severity)) +
  geom_histogram(binwidth = 5) +
  labs(x = "Severidade da doença (%)", y = "Número de parcelas")
```



Três características da distribuição merecem atenção:

- **simetria**: se os valores se distribuem de modo semelhante dos dois lados do centro;
- **assimetria**: uma distribuição com uma cauda mais longa de um lado. Dados de doença frequentemente têm assimetria à direita: a maioria das plantas ou parcelas tem pouca doença, e algumas têm muita. Contagens de lesões, de esporos e quantidades de DNA do patógeno costumam se comportar assim;
- **dispersão**: quão amplamente os valores variam. Na Fitopatologia, a variabilidade frequentemente aumenta com a média, de modo que tratamentos com mais doença também são mais variáveis.

Outras características também são informativas. Uma distribuição com dois picos (bimodal) pode indicar que dois grupos diferentes foram misturados, como plantas resistentes e suscetíveis ou duas datas de avaliação. Muitos zeros podem indicar escape à infecção ou um limite de detecção. Valores acumulados em 0% ou 100% revelam os limites de uma escala percentual. Valores concentrados em 5, 10, 25 e 50% refletem hábitos de arredondamento da avaliação visual, e não a biologia.

Essas características antecipam decisões tomadas mais tarde. A assimetria e a variabilidade que muda com a média podem exigir transformações ou modelos próprios para contagens e proporções (Capítulos 7 a 10).

4.5.3 Valores discrepantes

Um valor discrepante (*outlier*) é uma observação excepcionalmente distante do restante dos dados. Uma regra comum classifica como possíveis valores discrepantes os valores situados mais de 1,5 vez a amplitude interquartil abaixo do primeiro quartil ou acima do terceiro, que é a regra usada nos diagramas de caixa. Essa regra é uma ferramenta de triagem, não uma definição: um valor sinalizado pela regra pode ser perfeitamente válido, e um erro pode estar dentro da faixa normal.

Um valor discrepante pode ter três origens. Pode ser um **erro** de medida, registro ou digitação, como uma vírgula decimal deslocada. Pode ser uma **observação real, mas incomum**, como uma parcela localizada perto de uma fonte de inóculo. Ou pode indicar que a observação pertence a uma **população diferente**, como uma planta infectada por outro patógeno. Cada origem exige uma ação diferente: corrigir o erro a partir dos registros originais, manter o valor e examinar sua influência nos resultados, ou reconsiderar o escopo do estudo.

> ⚠ **Aviso**
>
> Valores discrepantes nunca devem ser removidos apenas porque são inconvenientes ou porque sua remoção produz um resultado significativo. Toda exclusão deve ser justificada com base documentada e relatada.

4.5.4 Valores perdidos

Os valores perdidos também afetam as variáveis individualmente. A proporção de valores perdidos, e se eles estão relacionados a outras variáveis, determinam o quanto podem afetar as conclusões. Resumos calculados com os valores perdidos removidos silenciosamente (como com `na.rm = TRUE`) baseiam-se em menos observações do que o planejado, e isso deve ser relatado.

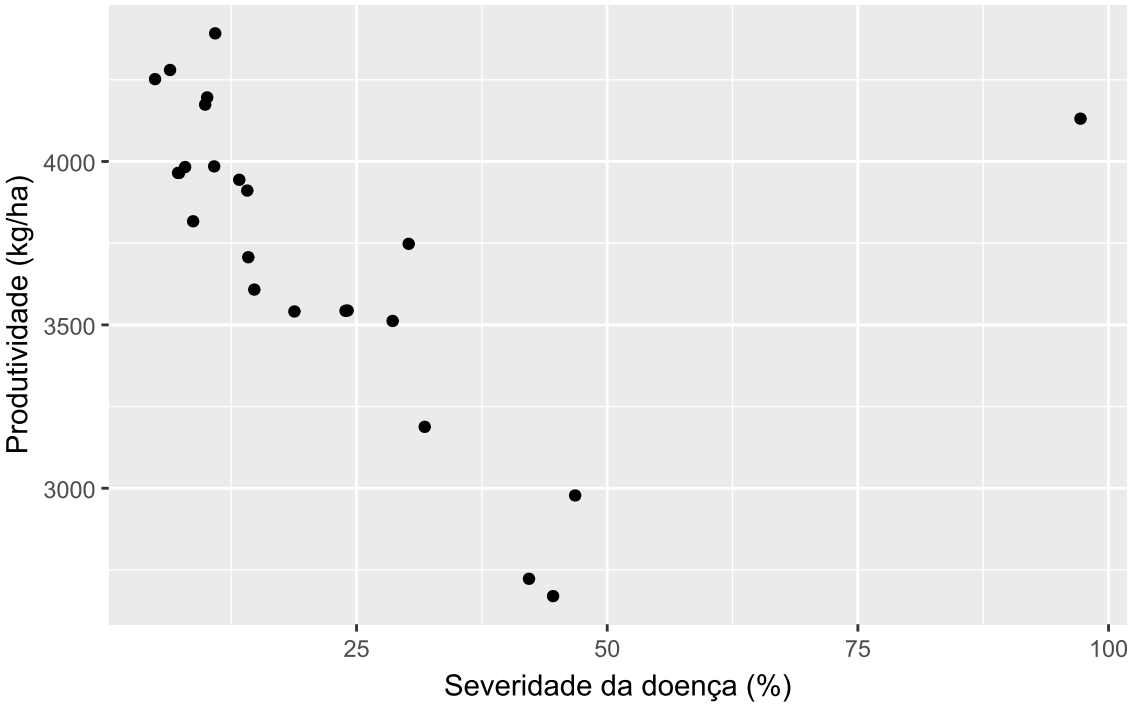
4.6 Explorando relações

A maioria das questões em Fitopatologia envolve relações: entre doença e produtividade, entre tratamento e severidade, entre clima e infecção. A exploração das relações depende dos tipos de variáveis envolvidas.

4.6.1 Contínua versus contínua

A relação entre duas variáveis quantitativas, como severidade e produtividade, ou temperatura e tamanho da lesão, é explorada com um gráfico de dispersão. O gráfico mostra a **direção** da relação (crescente, decrescente, nenhuma), sua **forma** (linear, curva, com platô ou limiar), sua **intensidade** (quão de perto os pontos seguem um padrão) e as **observações incomuns** que não seguem o padrão geral.

```
# uma observação por parcela: a severidade na última avaliação
final ← filter(data, date == max(date))

ggplot(final, aes(x = severity, y = yield)) +
  geom_point() +
  labs(x = "Severidade da doença (%)", y = "Produtividade (kg/ha)")
```



Na etapa exploratória, o objetivo é descrever o padrão, e não ajustar um modelo. Correlação e regressão são apresentadas no Capítulo 8.

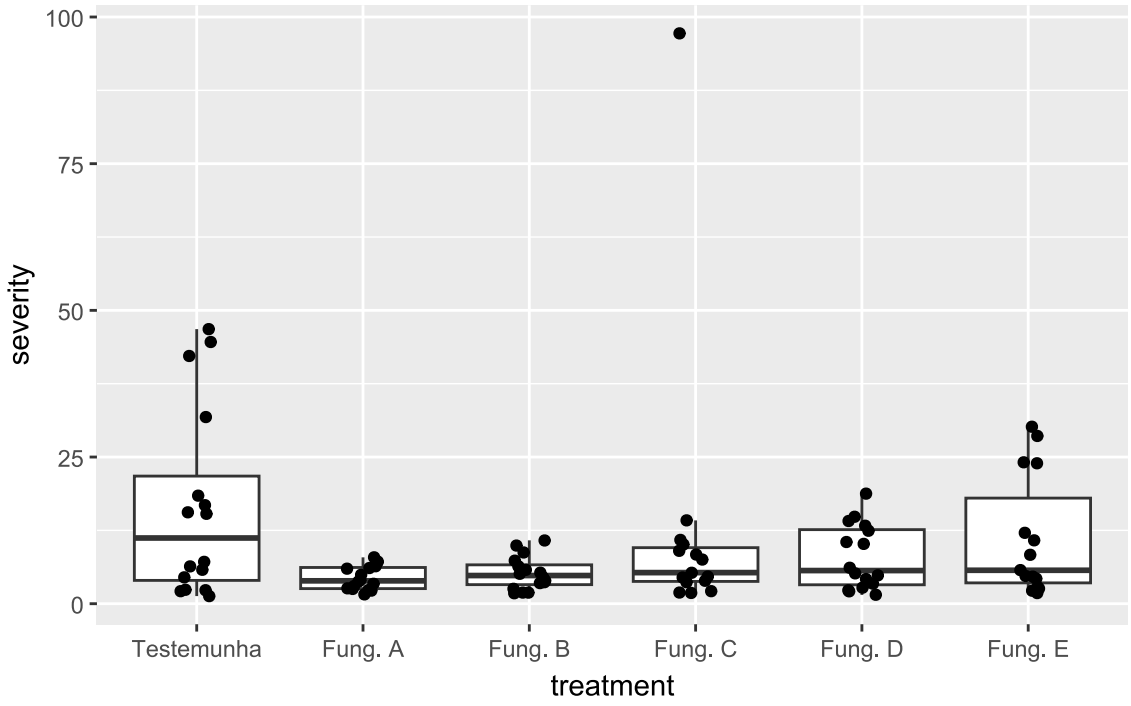
4.6.2 Contínua versus categórica

Comparar uma resposta quantitativa entre grupos, como a severidade entre tratamentos com fungicidas ou o diâmetro das lesões entre isolados, é a situação mais comum em experimentos. Resumos numéricos por grupo (contagens, médias, medianas, desvios padrão) são complementados por gráficos que mostram as observações individuais, como gráficos de pontos ou diagramas de caixa com pontos.

```
data ▷
  group_by(treatment) ▷
  summarise(n = n(), mean = mean(severity), sd = sd(severity))
```

```
# A tibble: 6 × 4
  treatment      n  mean    sd
  <fct>      <int> <dbl> <dbl>
1 Testemunha    16 16.5  16.1
2 Fung. A       16  4.44  2.07
3 Fung. B       16  5.19  2.83
4 Fung. C       16 NA    NA
5 Fung. D       16  7.9   5.50
6 Fung. E       16 NA    NA
```

```
ggplot(data, aes(x = treatment, y = severity)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.1)
```



Além das diferenças de centro, procure diferenças de dispersão entre grupos, assimetria dentro dos grupos, grupos com poucas observações e observações incomuns dentro do seu grupo, mesmo que não o sejam no conjunto.

4.6.3 Categórica versus categórica

A relação entre duas variáveis categóricas, como cultivar e presença de doença, ou região e espécie do patógeno, é explorada com tabelas de contingência que mostram contagens e proporções em cada combinação de categorias. As proporções devem ser calculadas na direção que responde à questão: por exemplo, a proporção de lavouras infectadas em cada sistema de preparo do solo, e não a proporção de sistemas de preparo entre as lavouras infectadas.

```
levantamento ▷
  count(tillage, disease_present) ▷
  group_by(tillage) ▷
  mutate(proportion = n / sum(n))
```

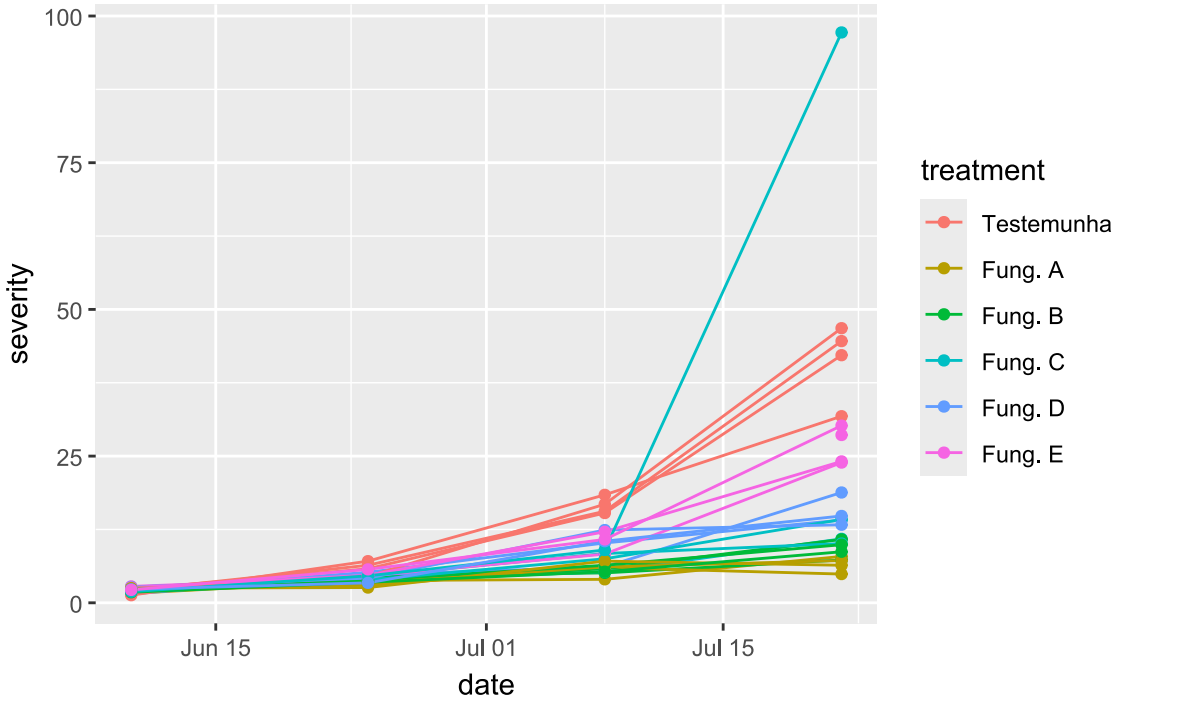
```
# A tibble: 4 × 4
# Groups:   tillage [2]
  tillage        disease_present     n proportion
  <chr>          <chr>           <int>      <dbl>
1 Convencional   não                49      0.721
2 Convencional   sim                19      0.279
3 Plantio direto não                17      0.327
4 Plantio direto sim                35      0.673
```

4.6.4 O tempo como variável

Quando as observações se repetem ao longo do tempo, o tempo torna-se uma variável central. O progresso da doença é explorado representando a resposta em função do tempo, com as observações da mesma unidade ligadas por linhas. Isso revela a forma da epidemia, o momento

de início, diferenças entre tratamentos que aparecem apenas no fim da safra e unidades que se comportam de modo diferente das demais.

```
ggplot(data, aes(x = date, y = severity, group = plot, color = treatment)) +
  geom_line() +
  geom_point()
```



Séries meteorológicas e outras séries temporais são exploradas da mesma forma. Observações próximas no tempo tendem a ser semelhantes, o que significa que não são independentes — uma característica que precisa ser considerada em análises posteriores.

Tabela 4.2. Ferramentas exploratórias por tipo de variável.

Variáveis envolvidas	Resumos numéricos	Gráficos
Uma quantitativa	Média, mediana, DP, quartis, amplitude	Histograma, densidade, diagrama de caixa, gráfico de pontos
Uma categórica	Contagens, proporções	Gráfico de barras de contagens
Quantitativa vs quantitativa	Descrição da direção, forma, intensidade	Gráfico de dispersão
Quantitativa vs categórica	Resumos por grupo	Gráfico de pontos, diagrama de caixa com pontos
Categórica vs categórica	Tabela de contingência, proporções	Barras agrupadas ou empilhadas, gráfico de mosaico
Quantitativa ao longo do tempo	Resumos por data	Gráfico de linhas ligando cada unidade

4.7 Padrões que merecem investigação

A exploração é uma busca por padrões. Alguns confirmam as expectativas; outros são surpreendentes e merecem atenção.

4.7.1 Tendências

Uma tendência é um aumento ou uma diminuição sistemática de uma resposta com outra variável: a doença aumentando com o tempo, a produtividade diminuindo com a severidade, a infecção aumentando com a duração do molhamento. As tendências podem ser lineares ou curvas, podem se estabilizar em um platô ou mudar de direção em um ótimo, como ocorre com a resposta de muitos patógenos à temperatura.

4.7.2 Agrupamentos

Agrupamentos são grupos de observações semelhantes entre si e diferentes das demais. Em um gráfico de dispersão de severidade versus produtividade, pontos de locais ou anos diferentes frequentemente formam nuvens separadas. Os agrupamentos sugerem uma variável de agrupamento não considerada, e uma relação aparente no conjunto pode desaparecer, ou se inverter, dentro dos grupos.

4.7.3 Gradientes

Um gradiente é uma mudança sistemática da resposta no espaço ou ao longo da ordem em que as unidades foram manuseadas. A doença pode aumentar de um lado da lavoura para o outro, seguindo a drenagem ou o vento predominante a partir de uma fonte de inóculo, ou pode diferir entre bancadas de uma casa de vegetação. Representar as respostas em função do bloco, da linha, da coluna, da bancada ou da ordem de processamento revela gradientes. Eles justificam o uso de blocos (Capítulo 2) e, quando inesperados, podem apontar problemas na condução do experimento.

4.7.4 Interações

Uma interação ocorre quando o efeito de um fator depende do nível de outro. Um fungicida pode reduzir muito a doença em uma cultivar suscetível e ter pouco efeito em uma resistente; a infecção pode aumentar com a duração do molhamento a 25 °C, mas não a 10 °C. Nos gráficos exploratórios, as interações aparecem como linhas não paralelas quando as médias dos grupos são representadas em função de um fator, com linhas separadas para os níveis do outro. As interações são frequentemente importantes do ponto de vista biológico e determinam como os dados deverão ser modelados.

4.7.5 Valores discrepantes

Os valores discrepantes, discutidos antes para variáveis individuais, também podem aparecer apenas nas relações: uma parcela com severidade e produtividade médias pode ainda assim ser incomum se sua produtividade for muito maior do que o esperado para sua severidade. Essas observações só são visíveis em gráficos de dispersão ou em gráficos por grupo.

4.7.6 Erros de digitação

A exploração é a última oportunidade de detectar erros antes que afetem as conclusões. Sinais típicos são valores fora da faixa possível, valores exatamente dez ou cem vezes maiores que o esperado, linhas idênticas, saltos bruscos na série temporal de uma unidade e rótulos de categoria que aparecem uma única vez.

4.7.7 Observações inesperadas

Algumas observações não se encaixam em nenhum padrão esperado: uma parcela testemunha sem doença em uma safra de alta pressão, uma cultivar resistente com sintomas severos, um único isolado muito mais agressivo do que todos os outros. Essas observações podem ser erros,

mas também podem ser o primeiro sinal de algo novo, como uma nova raça do patógeno superando a resistência.

> **Dica**
>
> Descobertas interessantes muitas vezes começam como observações inesperadas durante a análise exploratória.

4.8 Interpretação biológica

4.8.1 Padrões estatísticos versus significado biológico

Um padrão nos dados ainda não é um achado biológico. Uma diferença entre tratamentos, uma tendência ao longo do tempo ou um agrupamento de pontos só ganha significado quando interpretado à luz da biologia do patossistema e do delineamento do estudo. O mesmo padrão pode ter explicações muito diferentes. Uma severidade maior em um bloco pode refletir um gradiente ambiental real, um avaliador diferente ou uma data de avaliação posterior. Uma associação negativa entre severidade e produtividade em um levantamento pode refletir o dano causado pela doença, ou o fato de que lavouras com manejo ruim têm mais doença e menor produtividade por outros motivos.

Interpretar um padrão exige, portanto, perguntar: O padrão é grande o suficiente para importar biologicamente? É consistente entre blocos, experimentos ou anos? Poderia ser explicado pelo delineamento, pelo processo de coleta ou por um erro? É coerente com o que se sabe sobre o hospedeiro, o patógeno e o ambiente?

4.8.2 Fazendo novas perguntas

Uma boa análise exploratória costuma terminar com mais perguntas do que começou. Por que uma cultivar respondeu de modo diferente ao fungicida? Por que o progresso da doença parou depois da terceira avaliação? Por que algumas lavouras estão livres da doença apesar do clima favorável? Essas perguntas podem ser respondidas examinando outras variáveis do conjunto de dados, consultando as anotações de campo ou delineando um novo estudo.

4.8.3 Gerando hipóteses

A exploração é uma fonte legítima e valiosa de hipóteses científicas. Um padrão observado em um levantamento, como mais doença em lavouras semeadas mais cedo, pode levar à hipótese de que a semeadura antecipada expõe as plantas a condições mais favoráveis à infecção. Essa hipótese pode então ser testada em um experimento planejado, fechando o ciclo entre observação, hipótese e teste descrito no Capítulo 2.

A regra essencial é ser transparente sobre como a hipótese foi gerada. Uma hipótese sugerida pela exploração dos dados não é confirmada pelos mesmos dados. Relatar os achados exploratórios como exploratórios, e os confirmatórios como confirmatórios, é uma questão de integridade científica.

> **Importante**
>
> A exploração gera hipóteses; a confirmação exige dados que não foram usados para gerá-las.

4.9 Pesquisa na prática

Pesquisa revisitada: Explorando os efeitos da temperatura e do molhamento na antracnose da videira

Contexto

O Capítulo 1 apresentou um estudo em ambiente controlado sobre como a temperatura e a duração do molhamento foliar influenciam a infecção de folhas de videira por *Elsinoë ampelina*, agente causal da antracnose da videira (Carisse et al. 2020). As plantas foram inoculadas e expostas a combinações de seis temperaturas (5 a 30 °C) e seis durações de molhamento (0 a 24 h), e a severidade da doença foi avaliada cerca de uma semana depois. Aqui perguntamos: antes de ajustar qualquer modelo, o que esses dados sugerem? Os valores abaixo são ilustrativos e não reproduzem os dados publicados.

Os dados brutos

O conjunto de dados organizado tem uma linha por planta inoculada:

Tabela 4.3. Primeiras linhas do conjunto de dados da antracnose da videira (ilustrativo).

temp_c	wetness_h	plant	severity
5	0	1	0.0
5	24	1	0.0
15	12	1	4.2
25	12	1	11.8
25	24	1	24.5
30	24	1	13.1
…	…	…	…

Conhecendo o conjunto de dados

Com seis temperaturas, seis durações de molhamento e um número fixo de plantas por combinação, o número esperado de linhas é conhecido de antemão e pode ser comparado com o número real. As duas variáveis explicativas são numéricas e controladas; a severidade é uma porcentagem limitada entre 0 e 100.

Resumos exploratórios

A severidade média em cada combinação de temperatura e duração do molhamento pode ser organizada em uma tabela de dupla entrada:

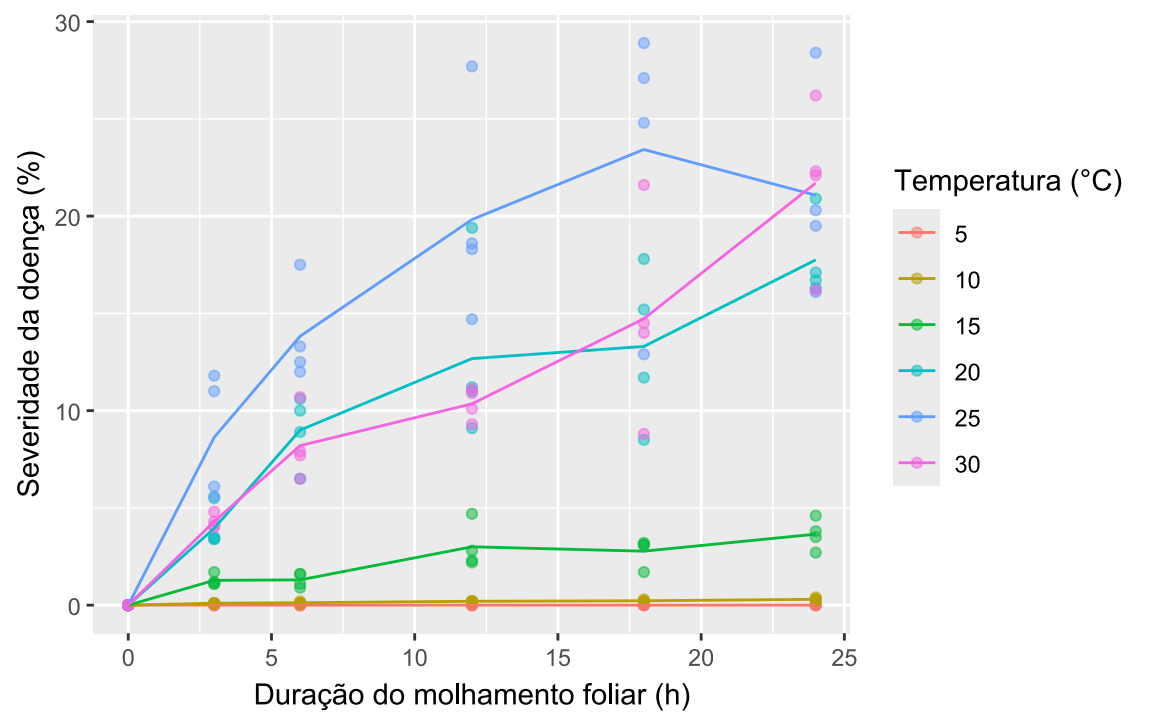
Tabela 4.4. Severidade média da doença (%) por temperatura e duração do molhamento (subconjunto ilustrativo).

Temperatura	0 h	6 h	12 h	24 h
5 °C	0,0	0,0	0,0	0,0
15 °C	0,0	1,1	4,0	9,5
25 °C	0,0	4,8	12,3	23,9
30 °C	0,0	2,2	6,5	12,7

A distribuição da severidade é fortemente assimétrica à direita: muitas plantas não têm doença alguma, e poucas têm severidade alta.

Gráficos simples

```
ggplot(grape, aes(x = wetness_h, y = severity, color = factor(temp_c))) +
  geom_point(alpha = 0.5) +
  stat_summary(fun = mean, geom = "line") +
  labs(x = "Duração do molhamento foliar (h)", y = "Severidade da doença (%)",
       color = "Temperatura (°C)")
```



O que esses dados sugerem?

- Não houve doença sem molhamento foliar, qualquer que fosse a temperatura, nem a 5 °C, qualquer que fosse a duração do molhamento. Os zeros são estruturais, e não aleatórios: marcam condições em que a infecção não ocorreu.
- A severidade aumentou com a duração do molhamento em todas as temperaturas em que houve infecção.
- O aumento foi mais acentuado perto de 25 °C e menos pronunciado a 15 e 30 °C, indicando uma temperatura ótima.
- Como o efeito do molhamento depende da temperatura, as linhas não são paralelas: os dois fatores interagem.
- A variabilidade entre plantas aumenta com a severidade média.

O que a exploração não nos diz

Essas observações descrevem padrões, mas não quantificam sua incerteza, não testam se a interação poderia ter surgido por acaso e não fornecem um modelo para prever a infecção em novas condições. Elas indicam, porém, que um modelo de linha reta com efeitos aditivos seria inadequado, que os zeros e a variabilidade crescente precisam ser considerados e que a análise deve permitir que o efeito do molhamento varie com a temperatura. São decisões que seriam difíceis de tomar sem a exploração.

Questões

1. Por que o número esperado de linhas é conhecido antes de olhar os dados?
2. O que os zeros a 0 h de molhamento e a 5 °C representam biologicamente?
3. Como a tabela de médias revela uma interação?
4. Por que uma única severidade média geral seria um resumo enganoso desses dados?
5. Que hipóteses para estudos futuros surgem dessa exploração?

4.10 Erros comuns

4.10.1 Ignorar os valores perdidos

Análises que removem silenciosamente os valores perdidos podem comparar grupos com números muito diferentes de observações, ou podem ser enviesadas se os valores faltarem por motivos relacionados à resposta, como plantas que morreram da doença. O número e o padrão dos valores perdidos devem sempre ser examinados e relatados.

4.10.2 Ignorar os valores discrepantes

Valores extremos podem dominar médias, desvios padrão e modelos ajustados. Ignorá-los pode levar a conclusões determinadas por uma única observação. Todo valor discrepante deve ser conferido com os registros originais, e sua influência nos resultados, examinada.

4.10.3 Tratar erros como biologia

O erro oposto é interpretar um erro como achado biológico. Uma produtividade dez vezes maior do que todas as outras ou um valor de severidade acima de 100% é quase certamente um erro. Explicações biológicas elaboradas não devem ser construídas sobre valores não verificados.

4.10.4 Ir direto aos testes de hipóteses

Executar testes estatísticos sem olhar os dados significa aplicar métodos sem saber se seus pressupostos se verificam, se os dados contêm erros ou se a questão testada é a que os dados conseguem responder. A exploração deve sempre preceder a inferência.

4.10.5 Olhar apenas os valores-P

Um valor-P resume um único aspecto da evidência e nada diz sobre o tamanho de um efeito, a variabilidade das observações ou a presença de padrões incomuns (Capítulo 6). Os resultados devem ser interpretados junto com os gráficos e resumos que mostram os próprios dados.

> ⚠ **Aviso**
>
> As análises exploratórias ajudam a identificar problemas que os métodos estatísticos formais muitas vezes não conseguem detectar.

4.11 Um fluxo de trabalho para a análise exploratória

A análise exploratória é iterativa, mas segue uma sequência reconhecível de etapas.

1. **Importar**: ler o conjunto de dados processado no programa de análise.
2. **Inspecionar**: conferir dimensões, nomes, classes e valores perdidos em relação ao delineamento e ao dicionário de dados.

3. **Limpar**: corrigir erros documentados, converter classes e harmonizar códigos, voltando aos dados brutos quando necessário (Capítulo 3).
4. **Resumir**: calcular resumos numéricos de cada variável e dos grupos definidos pelo delineamento.
5. **Visualizar**: representar graficamente distribuições, comparações entre grupos, relações e mudanças ao longo do tempo.
6. **Interpretar**: descrever os padrões em termos biológicos e anotar observações inesperadas e novas perguntas.
7. **Decidir**: escolher as análises adequadas aos dados e às questões.

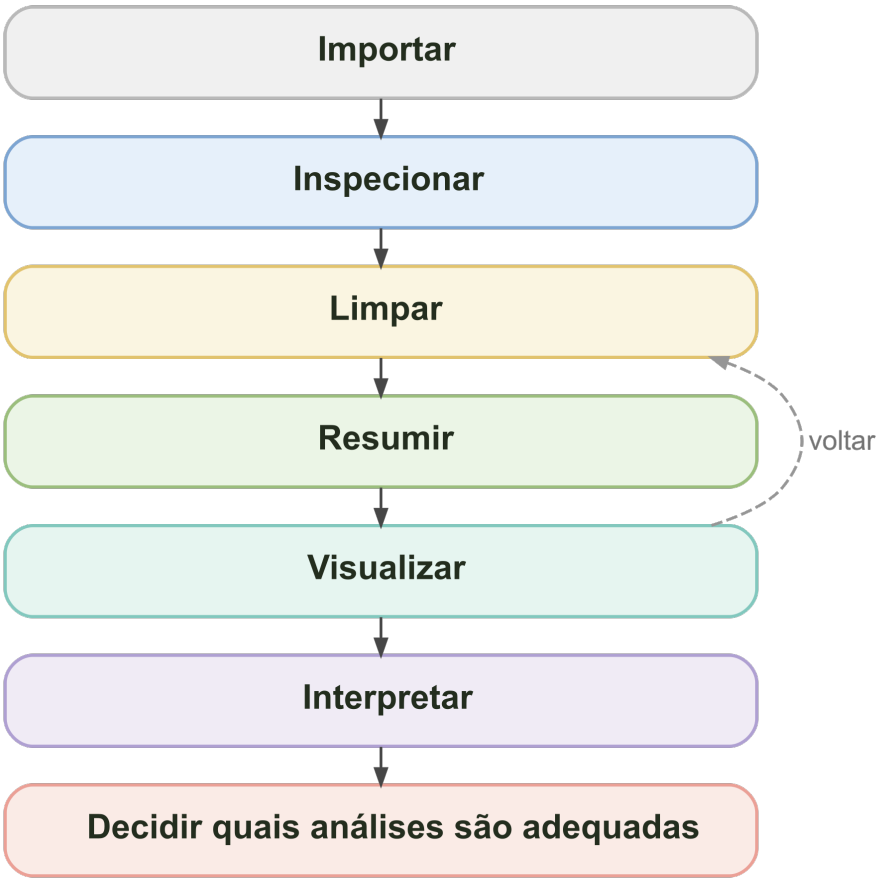


Figura 4.2. Fluxo de trabalho sugerido para a AED. A seta tracejada indica que descobertas nas etapas finais frequentemente levam de volta à limpeza dos dados.

Descobertas em qualquer etapa podem levar a análise de volta a uma etapa anterior: um valor estranho encontrado na visualização leva de volta à limpeza; um agrupamento inesperado leva de volta a resumir por uma nova variável de agrupamento.

4.12 Resumo do capítulo

A análise exploratória de dados é a etapa em que os pesquisadores passam a compreender seus dados antes da inferência formal. Ela começa conferindo o conjunto de dados em relação ao delineamento do estudo (dimensões, variáveis, classes e valores perdidos) e prossegue descrevendo cada variável com resumos numéricos e gráficos que revelam seu centro, dispersão, forma e valores incomuns. As relações entre variáveis são exploradas com ferramentas adequadas aos tipos de variáveis envolvidas, e padrões como tendências,

agrupamentos, gradientes e interações são identificados. Esses padrões só ganham significado por meio da interpretação biológica e frequentemente geram novas perguntas e hipóteses. A exploração complementa, mas não substitui, a análise confirmatória: ela revela erros, orienta a escolha dos métodos e mantém a atenção nos dados, enquanto as hipóteses sugeridas pela exploração exigem dados independentes para serem confirmadas.

4.13 Conceitos-chave

Ao concluir este capítulo, você deverá ser capaz de explicar:

- análise exploratória de dados;
- análise confirmatória;
- distribuição;
- centro e dispersão;
- assimetria;
- variabilidade;
- valor discrepante;
- valor perdido;
- padrão;
- associação;
- interação;
- interpretação biológica;
- geração de hipóteses.

4.14 Exercícios

4.14.1 Exercício 1. Conhecendo um conjunto de dados

Explore o conjunto de dados fornecido pelo professor e descreva:

- as variáveis e seus tipos;
- o número de observações e se ele corresponde ao delineamento do estudo;
- o número e a localização dos valores perdidos;
- eventuais observações incomuns.

4.14.2 Exercício 2. Da exploração às perguntas

Com base na exploração do Exercício 1, liste três questões biológicas que surgem dos dados. Para cada uma, indique se poderia ser respondida com o conjunto de dados atual ou se exigiria um novo estudo.

4.14.3 Exercício 3. Detectando erros de digitação

Examine os resumos e gráficos do conjunto de dados e identifique possíveis erros de digitação. Para cada um, explique que evidência sugere um erro e como você o verificaria.

4.14.4 Exercício 4. Comparando a variabilidade

Compare dois conjuntos de dados fornecidos pelo professor, por exemplo a severidade em duas safras ou o tamanho das lesões de dois isolados. Qual deles parece ter maior variabilidade? Fundamente sua resposta com resumos numéricos e um gráfico, e sugira razões biológicas para a diferença.

4.14.5 Exercício 5. Interpretando uma figura exploratória

Escolha uma figura de um artigo publicado em Fitopatologia. Sem realizar nenhum teste estatístico, descreva o que a figura mostra: a distribuição dos dados, diferenças entre grupos, tendências, interações e observações incomuns. Quais conclusões podem ser tiradas apenas da figura, e quais exigiriam análise formal?

4.15 Reflexão

> **A análise exploratória de dados é, muitas vezes, a etapa em que os pesquisadores mais aprendem sobre seus dados.**

Discuta essa afirmação.

4.16 Próximos passos

As análises exploratórias dependem frequentemente de representações gráficas. O próximo capítulo apresenta os princípios da visualização científica eficaz e mostra como figuras bem elaboradas revelam padrões que os resumos numéricos sozinhos podem não mostrar.

CAPÍTULO 5

Como visualizar dados de forma eficaz?

5.1 Objetivos de aprendizagem

! Importante

Ao concluir este capítulo, o leitor deverá ser capaz de:

- explicar o papel da visualização na análise e na comunicação científica;
- distinguir gráficos exploratórios de gráficos explicativos;
- identificar os principais componentes de um gráfico estatístico;
- associar adequadamente variáveis a propriedades gráficas;
- escolher representações gráficas de acordo com a questão biológica e a estrutura dos dados;
- reconhecer visualizações enganosas ou ineficazes;
- construir figuras científicas claras e informativas usando o R;
- aprimorar figuras por meio do uso adequado de escalas, rótulos, símbolos e ênfase gráfica.

5.2 Por que visualizar dados?

Conjuntos de dados científicos frequentemente contêm padrões difíceis de reconhecer apenas por tabelas ou resumos numéricos. A visualização estabelece uma conexão direta entre os dados e a percepção humana, permitindo identificar distribuições, relações, variação, observações incomuns e padrões biológicos.

O Capítulo 4 mostrou que a análise exploratória depende fortemente de gráficos. Histogramas revelam assimetria e valores discrepantes, gráficos de dispersão revelam curvaturas e agrupamentos, e gráficos de linhas revelam como a doença se desenvolve ao longo do tempo. Os gráficos também são o principal meio de comunicar resultados: em artigos, teses, relatórios, apresentações e materiais de extensão, os leitores muitas vezes olham as figuras antes — e às vezes em vez — de ler o texto. Uma figura serve, portanto, a dois públicos: o pesquisador que tenta compreender os dados e o leitor que tenta avaliar a evidência.

Os dois usos exigem escolhas. Quais variáveis vão em qual eixo, se mostrar as observações individuais ou apenas resumos, como representar a incerteza e como usar as cores — tudo isso afeta o que o leitor vê e conclui. Este capítulo apresenta princípios para fazer essas escolhas de forma deliberada, para que as figuras revelem a evidência em vez de obscurecê-la ou distorcê-la.

> **! Importante**
>
> Uma figura científica deve ajudar o leitor a ver a evidência, e não apenas enfeitar os resultados.

5.3 A visualização como parte do raciocínio científico

5.3.1 Vendo padrões nos dados

Os resumos numéricos condensam os dados em poucos números e, ao fazer isso, inevitavelmente descartam informação. Conjuntos de dados muito diferentes podem ter as mesmas médias, desvios padrão e correlações. A demonstração clássica é o quarteto de Anscombe: quatro pequenos conjuntos de dados com médias, variâncias, correlação e reta de regressão praticamente idênticas, dos quais um mostra uma relação linear, outro uma curva, outro uma reta perfeita distorcida por um único valor discrepante e outro nenhuma relação, exceto por um ponto extremo (Anscombe 1973).

O mesmo fenômeno ocorre na Fitopatologia. Imagine quatro ensaios de campo em que a severidade média da testemunha é 20% e seu desvio padrão é 8%. No primeiro ensaio, a severidade varia suavemente entre as parcelas. No segundo, a maioria das parcelas tem cerca de 15% de severidade e uma parcela tem 50%, talvez por estar ao lado de uma fonte de inóculo. No terceiro, as parcelas se dividem em dois grupos, 10% e 30%, correspondentes a dois tipos de solo. No quarto, a severidade aumenta continuamente de uma extremidade da lavoura à outra. Os resumos são idênticos; as histórias biológicas são completamente diferentes. Somente um gráfico das observações individuais revela qual história se aplica.

O quarteto de Anscombe está disponível no R base (objeto `anscombe`), e basta reorganizá-lo para vê-lo (Figura 5.1).

```
library(ggplot2)

quarteto ← data.frame(
  conjunto = rep(paste("Conjunto", 1:4), each = 11),
  x = c(anscombe$x1, anscombe$x2, anscombe$x3, anscombe$x4),
  y = c(anscombe$y1, anscombe$y2, anscombe$y3, anscombe$y4)
)

ggplot(quarteto, aes(x, y)) +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE,
              colour = "grey60", linewidth = 0.8) +
  geom_point(size = 2.5, colour = "#0072B2") +
  facet_wrap(~ conjunto) +
  theme_bw(base_size = 13)
```

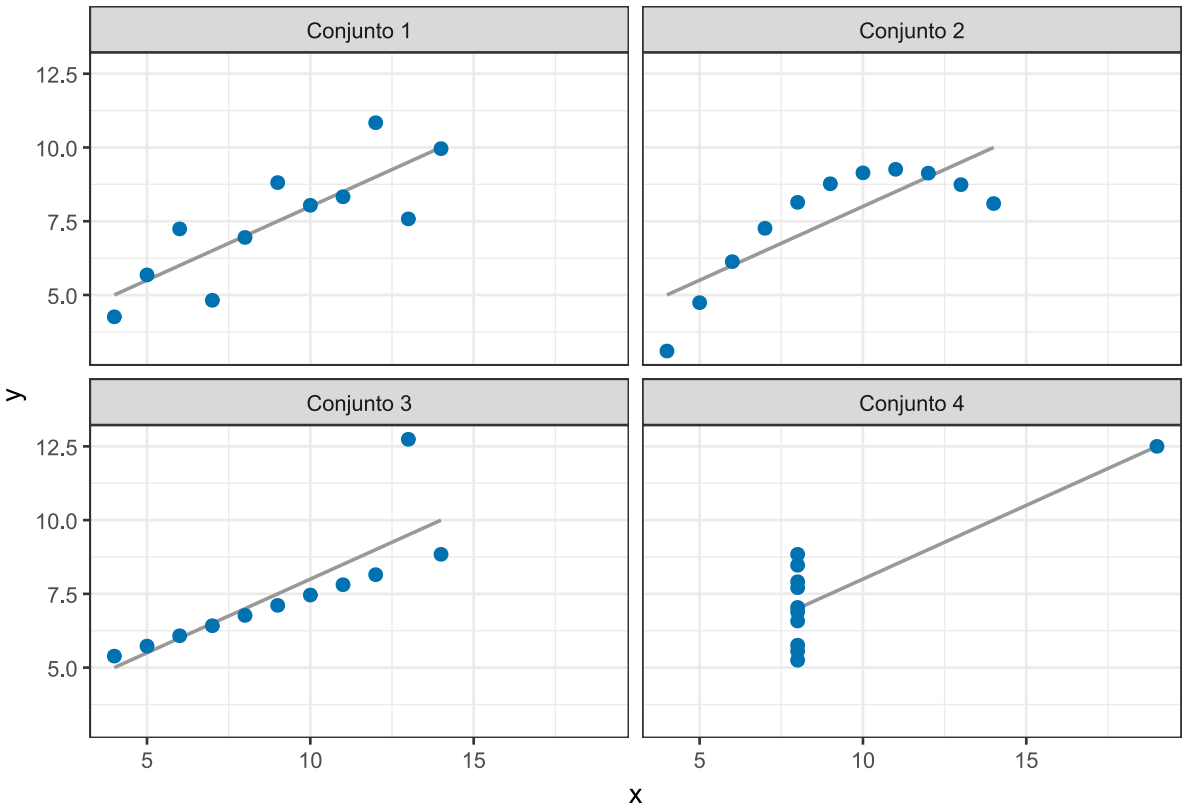


Figura 5.1. O quarteto de Anscombe: quatro conjuntos de dados com médias, variâncias, correlação (r ≈ 0,82) e reta de regressão praticamente idênticas, mas com padrões completamente diferentes.

5.3.2 Exploração e comunicação

A visualização serve a dois propósitos distintos. **Gráficos exploratórios** são feitos pelos pesquisadores para si mesmos, para compreender os dados: pesquisador → dados → compreensão. São produzidos rapidamente, em grande número, e descartados livremente. Seu objetivo é revelar tudo o que possa existir nos dados, inclusive erros e padrões inesperados.

Gráficos explicativos são feitos para outras pessoas, para comunicar um resultado: pesquisador → figura → leitor. São produzidos com cuidado e em pequeno número, e cada um é planejado para transmitir uma mensagem específica de forma clara e honesta. Seu objetivo é ajudar os leitores a ver e avaliar a evidência que sustenta uma conclusão.

Os dois tipos diferem no grau de acabamento, mas não na honestidade. Uma figura explicativa seleciona e enfatiza, mas não deve esconder características dos dados que sejam relevantes para a conclusão.

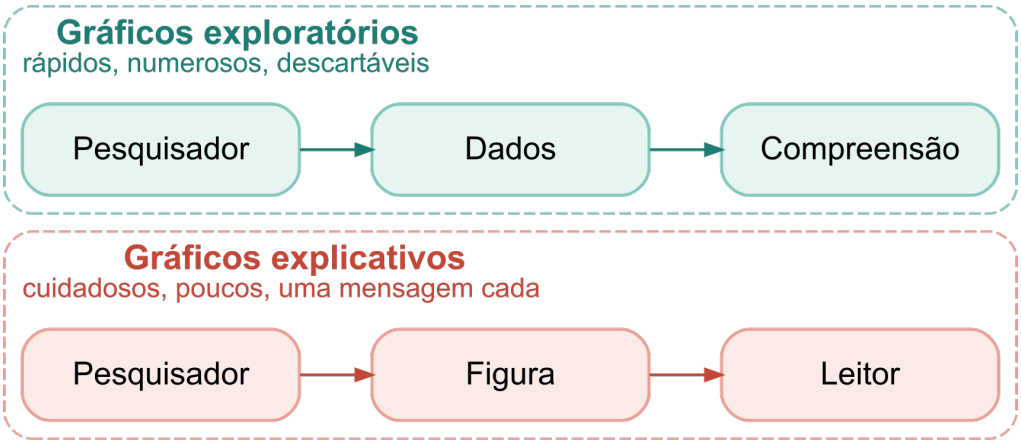


Figura 5.2. Dois papéis complementares da visualização científica: explorar os dados e comunicar a evidência.

5.3.3 Dos dados à evidência gráfica

Uma figura é uma representação dos dados, e não os dados em si. Toda figura resulta de decisões: quais observações incluir, se mostrar valores individuais ou resumos, quais comparações colocar lado a lado, que escala usar e o que enfatizar com cor ou posição. Cada decisão torna algumas características dos dados mais fáceis de ver e outras mais difíceis.

Não existe figura neutra. Por isso, a pergunta a fazer não é se uma figura envolve escolhas, mas se essas escolhas são orientadas pela questão científica e se representam a evidência de forma justa. O restante deste capítulo discute como fazer bem essas escolhas.

5.4 A anatomia de um gráfico estatístico

Gráficos estatísticos podem parecer muito diferentes entre si, mas são construídos a partir dos mesmos componentes. Essa ideia, formalizada como **gramática dos gráficos** (Wickham 2010; Wilkinson 2005), fornece uma forma sistemática de descrever, planejar e construir qualquer gráfico.

5.4.1 Dados

Todo gráfico começa com um conjunto de dados. A organização descrita no Capítulo 3, com uma linha por observação e uma coluna por variável, é a entrada natural para os gráficos: cada linha se torna um elemento gráfico (um ponto, uma barra, um segmento de linha), e cada variável pode ser associada a uma propriedade visual.

5.4.2 Mapeamentos gráficos

Um mapeamento, também chamado mapeamento estético, associa uma variável a uma propriedade visual: severidade à posição vertical, tratamento à posição horizontal, cultivar à cor, ano ao formato. Os mapeamentos são as decisões centrais no planejamento de um gráfico, porque determinam quais comparações o leitor consegue fazer com facilidade.

5.4.3 Representações geométricas

As geometrias são as marcas usadas para representar as observações: pontos, linhas, barras, caixas, áreas ou texto. O mesmo mapeamento pode ser desenhado com geometrias diferentes. A severidade por tratamento pode ser mostrada com pontos, diagramas de caixa ou barras, e cada opção enfatiza características diferentes dos dados.

5.4.4 Escalas

As escalas traduzem os valores dos dados em valores visuais: severidade de 0 a 100% em posições verticais, cultivares em cores, número de esporos em tamanhos de pontos. As escalas também produzem os guias que permitem ao leitor fazer a tradução inversa: eixos e legendas. As escolhas sobre escalas incluem seus limites, marcações e transformações, como escalas logarítmicas.

5.4.5 Sistemas de coordenadas

O sistema de coordenadas define o espaço em que o gráfico é desenhado. A maioria dos gráficos científicos usa coordenadas cartesianas, com eixos horizontal e vertical perpendiculares. Outros sistemas incluem as coordenadas polares (usadas em gráficos de setores e gráficos circulares) e as projeções geográficas (usadas em mapas).

5.4.6 Rótulos e anotações

Rótulos e anotações tornam o gráfico interpretável: títulos dos eixos com unidades, títulos das legendas, rótulos dos painéis, linhas de referência (como um limiar econômico) e texto destacando observações específicas. Eles ligam a representação gráfica de volta ao significado biológico.

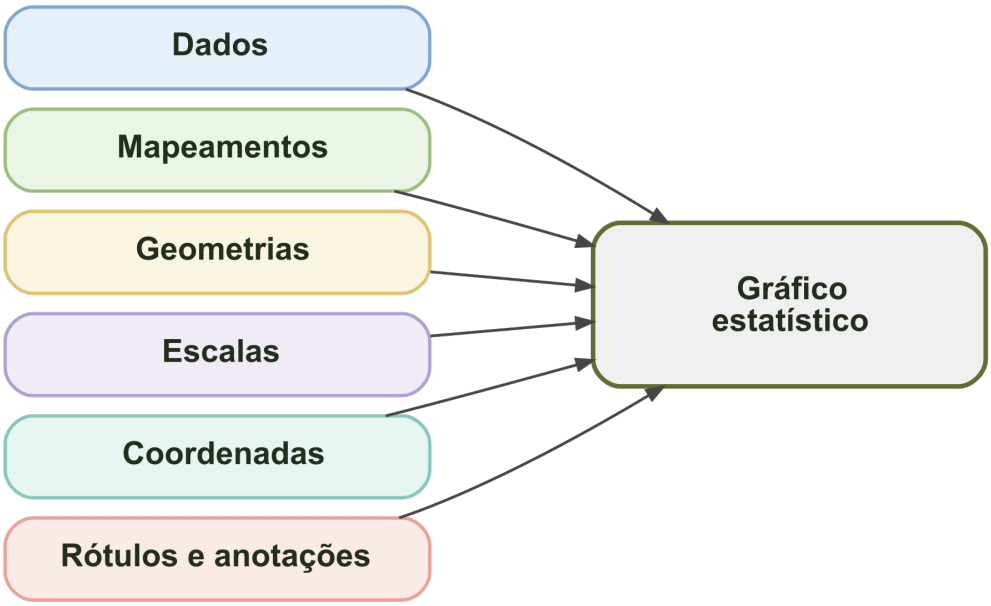


Figura 5.3. Principais componentes de um gráfico estatístico, segundo a gramática dos gráficos.

5.5 Associando variáveis a propriedades visuais

As propriedades visuais não são igualmente eficazes para transmitir informação quantitativa. Pesquisas sobre percepção gráfica mostraram que as pessoas julgam algumas propriedades visuais com mais precisão do que outras (Cleveland e McGill 1984). A escolha de qual variável associar a qual propriedade deve, portanto, refletir quais comparações são mais importantes.

5.5.1 Posição

A posição ao longo de uma escala comum é a propriedade visual percebida com mais precisão. Os leitores conseguem comparar com grande precisão as alturas de pontos alinhados no mesmo

eixo. Por isso, a comparação mais importante de uma figura deve ser representada pela posição, geralmente no eixo vertical.

5.5.2 Comprimento

O comprimento, como nos gráficos de barras, também é julgado com precisão, desde que as barras partam de uma linha de base comum, geralmente zero. Comprimentos que não começam em zero, ou que não estão alinhados, são comparados com muito menos precisão.

5.5.3 Tamanho

O tamanho, como a área dos pontos, é percebido com menos precisão do que a posição ou o comprimento. Os leitores tendem a subestimar diferenças de área. O tamanho é útil para representar uma variável secundária, como o número de observações por trás de cada ponto, mas não para comparações precisas.

5.5.4 Formato

O formato é adequado para distinguir um pequeno número de categorias, como anos ou locais. Ele não transmite quantidade, e mais de quatro ou cinco formatos tornam-se difíceis de distinguir.

5.5.5 Cor

A cor pode representar categorias (matiz) ou quantidades (luminosidade e saturação). O matiz é eficaz para distinguir poucos grupos; a intensidade da cor transmite apenas magnitudes aproximadas. A cor é discutida em detalhe mais adiante neste capítulo.

5.5.6 Transparência

A transparência é usada principalmente para lidar com observações sobrepostas. Quando muitos pontos caem uns sobre os outros, pontos semitransparentes revelam onde as observações se concentram. A transparência também pode atenuar informações de fundo, como as observações individuais por trás das médias dos grupos.

Tabela 5.1. Propriedades visuais e seus usos típicos.

Propriedade visual	Melhor uso	Precisão para comparação quantitativa
Posição em escala comum	Comparação quantitativa principal	Máxima
Comprimento a partir de base comum	Contagens, totais, quantidades	Alta
Tamanho (área)	Quantidades secundárias	Baixa
Intensidade da cor	Magnitudes aproximadas, mapas de calor	Baixa
Matiz da cor	Poucas categorias	Não quantitativa
Formato	Poucas categorias	Não quantitativa
Transparência	Pontos sobrepostos, atenuação	Não quantitativa

A Figura 5.4 mostra os mesmos cinco valores representados por quatro propriedades visuais diferentes. Na posição, as diferenças são lidas com facilidade; na área e na intensidade da cor, mesmo diferenças grandes parecem pequenas e imprecisas.

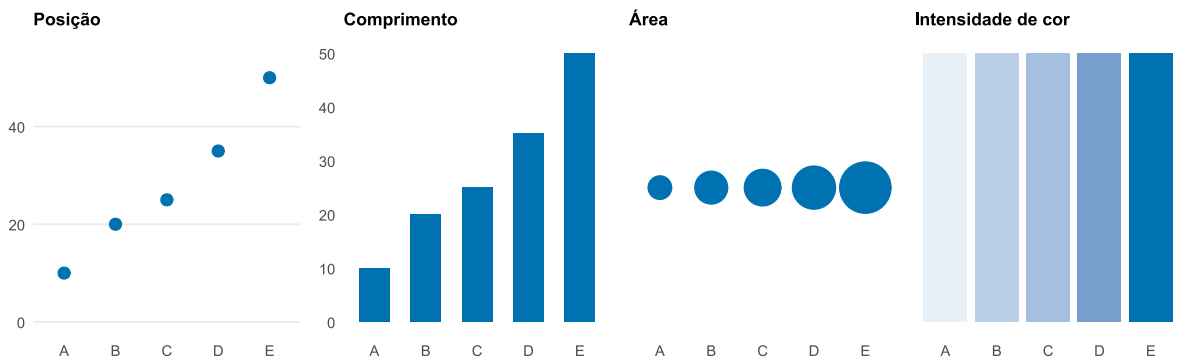


Figura 5.4. Os mesmos cinco valores (severidade de 10, 20, 25, 35 e 50%) representados por posição, comprimento, área e intensidade de cor. A precisão com que o leitor compara os valores diminui da esquerda para a direita.

Dica

Sempre que a comparação quantitativa precisa for importante, prefira propriedades visuais que os leitores interpretam com precisão, especialmente a posição ao longo de uma escala comum.

5.6 Escolhendo um gráfico

A representação gráfica adequada depende principalmente da questão biológica e das variáveis representadas.

5.6.1 Mostrando distribuições

A distribuição de uma única variável quantitativa é mostrada com histogramas, gráficos de densidade, diagramas de caixa, gráficos de violino ou gráficos das observações brutas. Os histogramas dividem a amplitude em intervalos e contam as observações em cada um; sua aparência depende da largura dos intervalos, que deve ser escolhida com cuidado. Os gráficos de densidade mostram uma versão suavizada da distribuição. Os diagramas de caixa resumem a distribuição com a mediana, os quartis e os possíveis valores discrepantes, e são compactos o suficiente para comparar muitos grupos. Os gráficos de violino combinam a compacidade do diagrama de caixa com a forma do gráfico de densidade. Quando o número de observações é pequeno, como na maioria dos experimentos planejados, as próprias observações, mostradas em um gráfico de pontos, costumam ser a representação mais informativa.

5.6.2 Comparando grupos

Comparar uma resposta quantitativa entre tratamentos, cultivares ou isolados é a tarefa mais comum nas figuras de Fitopatologia. Gráficos de pontos mostrando todas as observações, diagramas de caixa e gráficos de médias com intervalos são todos eficazes. Gráficos de barras de médias são comuns, mas muitas vezes inadequados para respostas contínuas: mostram um único número por grupo, escondem a distribuição e o número de observações e sugerem que valores abaixo do topo da barra foram observados. A comparação entre grupos deve estar em um eixo comum, e os grupos devem estar ordenados de forma significativa.

5.6.3 Mostrando relações

A relação entre duas variáveis quantitativas, como severidade e produtividade ou temperatura e crescimento da lesão, é mostrada com um gráfico de dispersão. Variáveis adicionais podem ser mostradas por cor, formato ou painéis separados. Linhas ou curvas ajustadas podem ser acrescentadas, mas na etapa exploratória os próprios pontos são mais importantes do que qualquer linha-resumo.

5.6.4 Mostrando mudanças ao longo do tempo

Mudanças ao longo do tempo são mostradas com gráficos de linhas, em que o tempo fica no eixo horizontal e as observações da mesma unidade são ligadas. Curvas de progresso da doença, dinâmica de populações de patógenos e séries ambientais como temperatura, chuva e molhamento foliar são exemplos típicos. As linhas devem ligar observações apenas quando pertencem à mesma unidade ou grupo e quando os valores intermediários têm significado.

5.6.5 Mostrando proporções e frequências

Contagens e proporções de desfechos categóricos, como o número de lavouras em que um patógeno foi detectado ou a proporção de isolados resistentes a um fungicida, são mais bem mostradas com gráficos de barras, em que o comprimento da barra a partir do zero representa a contagem ou a proporção. Barras empilhadas mostram a composição dos grupos, mas dificultam a comparação de todos os segmentos exceto o de baixo. Gráficos de setores (“pizza”) representam proporções como ângulos e áreas, que são julgados com menos precisão do que comprimentos; raramente são a melhor escolha em figuras científicas.

5.6.6 Mostrando a incerteza

Estimativas como médias ou efeitos de tratamentos devem ser acompanhadas de uma representação de sua incerteza, normalmente um intervalo de confiança desenhado como uma linha em torno de um ponto. Sempre que possível, a incerteza deve ser mostrada junto com as observações, para que os leitores vejam tanto a variabilidade dos dados quanto a precisão das estimativas. Essas representações são discutidas em mais detalhe adiante.

Tabela 5.2. Relacionando questões científicas a abordagens gráficas comuns.

Questão científica	Abordagem gráfica típica
Como uma variável está distribuída?	Histograma, gráfico de densidade, diagrama de caixa
Como os grupos diferem?	Gráfico de pontos, diagrama de caixa, gráfico de intervalos
Duas variáveis quantitativas estão relacionadas?	Gráfico de dispersão
Como uma resposta muda ao longo do tempo?	Gráfico de linhas
Qual a frequência de diferentes categorias?	Gráfico de barras
Quão incerta é uma estimativa?	Gráfico de pontos e intervalos

A Figura 5.5 reúne exemplos dessas representações, todos construídos com o mesmo conjunto de dados simulados de um ensaio de fungicidas. O código de cada painel segue o padrão apresentado na seção sobre o `ggplot2`, mais adiante.

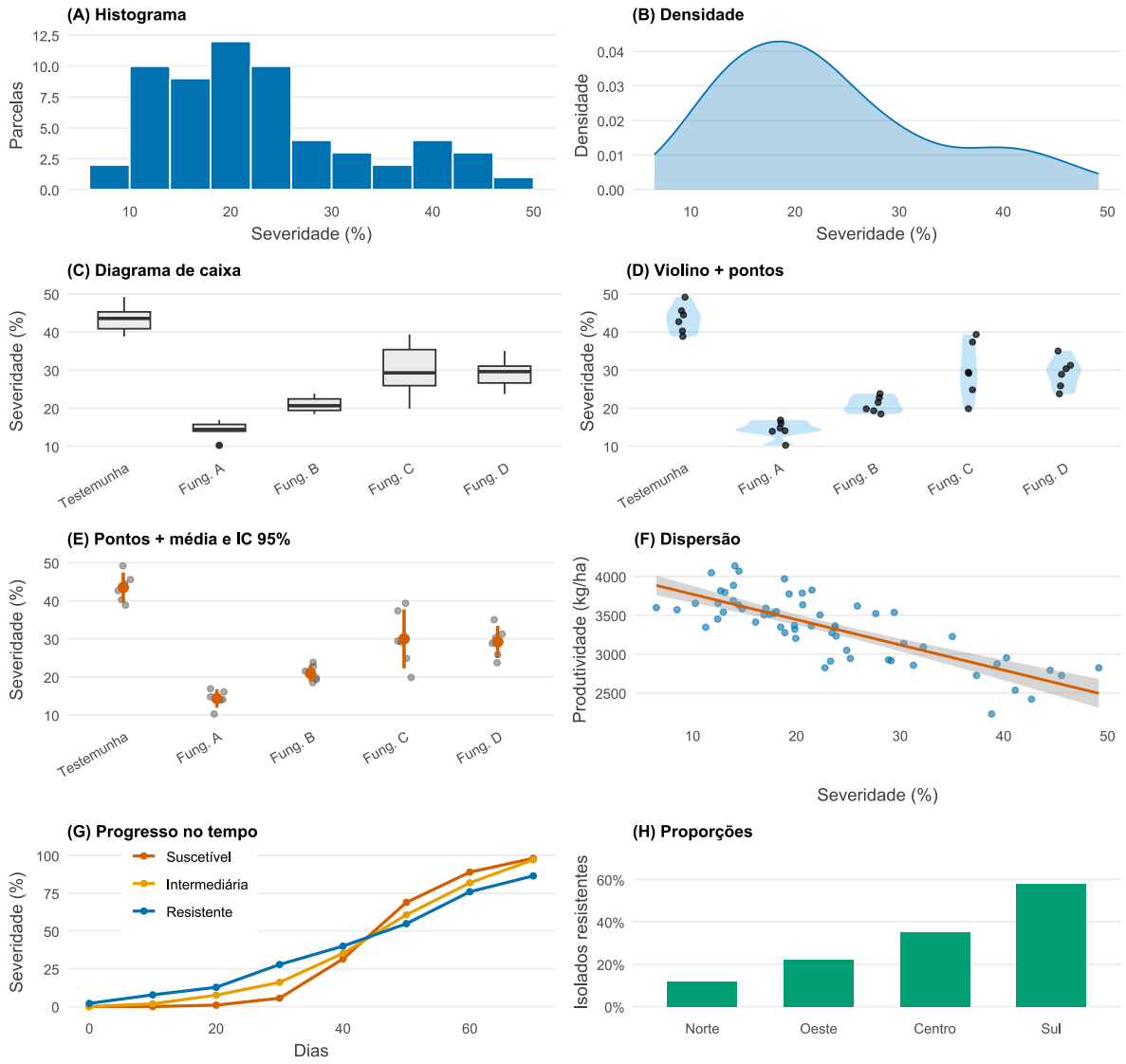


Figura 5.5. Galeria de gráficos comuns em Fitopatologia (dados simulados). Distribuições: (A) histograma, (B) densidade, (C) diagrama de caixa e (D) violino com pontos. Comparação de grupos: (E) observações com média e IC 95%. Relações: (F) dispersão com reta ajustada. Tempo: (G) curvas de progresso. Proporções: (H) barras de proporção.

5.7 Mostre os dados sempre que possível

5.7.1 Observações brutas

A maioria dos experimentos em Fitopatologia tem um número modesto de unidades experimentais por tratamento, muitas vezes entre quatro e vinte. Com tão poucas observações, não há motivo para escondê-las. Mostrar cada observação permite ao leitor ver o número de repetições, a dispersão dos dados, valores incomuns e se as estatísticas-resumo representam bem os dados. Quando as observações se sobrepõem, podem ser espalhadas horizontalmente com um pequeno deslocamento aleatório (*jitter*) ou tornadas semitransparentes.

5.7.2 Estatísticas-resumo

Estatísticas-resumo, como médias ou medianas, ajudam o leitor a se concentrar na comparação de interesse. São mais bem mostradas sobre as observações, com um estilo visual distinto, como um ponto maior ou uma linha horizontal, para que os dois níveis de informação fiquem visíveis.

5.7.3 Distribuição e variação

Um tipo comum de figura mostra as médias dos tratamentos como barras com barras de erro padrão. Essa representação esconde quase tudo sobre os dados: o número de observações, sua distribuição, valores discrepantes e se a distribuição é assimétrica ou bimodal. Conjuntos de dados muito diferentes podem produzir barras idênticas (Weissgerber et al. 2015). Uma alternativa mais informativa mostra as observações individuais, as médias dos tratamentos e os intervalos de confiança. Com o mesmo espaço e a mesma tinta, ela revela os dados e a incerteza das estimativas.

```
library(ggplot2)

# Convencional: barras e erros padrão
ggplot(data, aes(x = treatment, y = severity)) +
  stat_summary(fun = mean, geom = "col") +
  stat_summary(fun.data = mean_se, geom = "errorbar", width = 0.2)

# Melhor: observações, médias e intervalos de confiança de 95%
ggplot(data, aes(x = treatment, y = severity)) +
  geom_jitter(width = 0.1, alpha = 0.5) +
  stat_summary(fun.data = mean_cl_normal, color = "red")
```

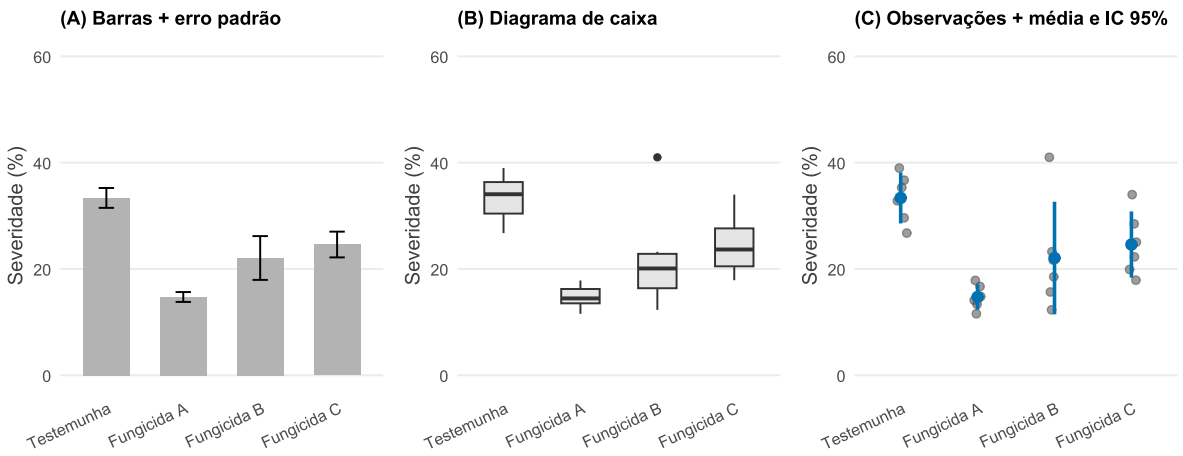


Figura 5.6. Diferentes representações gráficas dos mesmos dados experimentais (dados simulados; seis parcelas por tratamento). (A) Barras com erro padrão mostram apenas a média e sua precisão. (B) Diagramas de caixa mostram a distribuição, mas com poucas observações escondem quantas são. (C) Observações individuais com a média e o intervalo de confiança de 95% revelam os dados e a incerteza, incluindo a parcela discrepante do Fungicida B e a maior variabilidade da testemunha.

! Importante

As estatísticas-resumo não devem esconder desnecessariamente as observações a partir das quais foram calculadas.

5.8 Representando a incerteza

5.8.1 Variabilidade versus incerteza

Dois conceitos diferentes são frequentemente confundidos nas figuras. A **variabilidade** descreve o quanto as observações individuais diferem entre si; é uma propriedade dos dados e é resumida pelo desvio padrão ou pela amplitude. A **incerteza** descreve com que precisão uma grandeza, como a média de um tratamento, foi estimada; é resumida pelo erro padrão ou por um intervalo de confiança. A variabilidade não diminui à medida que mais observações são coletadas; a incerteza, sim. Os fundamentos estatísticos desses conceitos são apresentados no Capítulo 6. Para os gráficos, o ponto-chave é que o leitor precisa saber qual dos dois está sendo mostrado.

5.8.2 Barras de erro

As barras de erro podem representar o desvio padrão, o erro padrão, um intervalo de confiança ou a amplitude. Essas alternativas podem diferir muito em comprimento: o erro padrão é menor que o desvio padrão por um fator igual à raiz quadrada do tamanho da amostra. Barras de erro sem a indicação do que representam são ininterpretáveis. Toda figura com barras de erro deve informar na legenda o que elas mostram e em quantas observações cada uma se baseia.

5.8.3 Intervalos de confiança

Um intervalo de confiança mostra uma faixa de valores para a grandeza estimada que são compatíveis com os dados, dado o modelo. Os intervalos de confiança são, em geral, a representação mais útil da incerteza para estimativas como médias, diferenças entre tratamentos ou inclinações, porque sua largura reflete diretamente a precisão da estimativa nas unidades da resposta.

5.8.4 Intervalos de predição

Um intervalo de predição mostra a faixa em que se espera que uma nova observação individual caia. Ele é sempre mais largo do que o intervalo de confiança da média, porque inclui tanto a incerteza da estimativa quanto a variabilidade das observações individuais. Os intervalos de predição são relevantes quando o objetivo é prever, por exemplo, o nível de doença esperado em uma nova lavoura sob determinadas condições (Capítulo 8).

5.8.5 Quando a incerteza deve ser mostrada

A incerteza deve ser mostrada sempre que uma figura sustenta uma conclusão sobre estimativas, como diferenças entre tratamentos, efeitos ou relações ajustadas. Pode ser omitida em gráficos puramente descritivos dos dados brutos, em que as próprias observações mostram a variabilidade. Em todos os casos, a representação deve corresponder à comparação de interesse: se a conclusão é sobre a diferença entre dois tratamentos, a incerteza dessa diferença é mais relevante do que a incerteza de cada média separadamente.

Tabela 5.3. Representações comuns de variabilidade e incerteza.

Representação	O que descreve	Diminui com mais observações?
Desvio padrão	Variabilidade entre observações	Não
Erro padrão	Precisão de uma média estimada	Sim
Intervalo de confiança	Faixa de valores compatíveis com os dados para uma estimativa	Sim
Intervalo de predição	Faixa esperada para uma nova observação	Só em parte

5.9 Cor em gráficos científicos

5.9.1 A cor como informação

A cor é poderosa: atrai a atenção e distingue grupos instantaneamente. Pelo mesmo motivo, deve ser usada de forma deliberada. Em gráficos científicos, a cor deve codificar uma variável ou direcionar a atenção para a mensagem principal. O tipo de paleta deve corresponder ao tipo de variável.

5.9.2 Paletas sequenciais

Paletas sequenciais variam gradualmente do claro ao escuro (ou de baixa a alta saturação) e representam quantidades ordenadas que vão de baixo a alto: severidade da doença em um mapa, concentração de DNA do patógeno, chuva. Cores mais claras costumam representar valores menores, e cores mais escuras, valores maiores.

5.9.3 Paletas divergentes

Paletas divergentes usam dois matizes contrastantes que divergem a partir de um ponto central neutro. Representam quantidades com um centro significativo, como mudança zero, uma cultivar de referência ou correlação zero. Exemplos incluem mudanças na doença em relação a uma testemunha, anomalias de temperatura e matrizes de correlação.

5.9.4 Paletas qualitativas

Paletas qualitativas usam matizes distintos de luminosidade semelhante para representar categorias sem ordem, como cultivares, isolados ou locais. Não devem sugerir uma ordem inexistente e funcionam bem apenas para um número limitado de categorias, geralmente não mais que seis a oito.

5.9.5 Acessibilidade das cores

Cerca de 8% dos homens e 0,5% das mulheres têm alguma forma de deficiência na visão de cores, mais comumente dificuldade em distinguir vermelho e verde. Paletas vermelho–verde, infelizmente comuns em figuras de doença e sanidade, são portanto inacessíveis para uma parte considerável do público. Devem ser preferidas paletas projetadas para daltonismo, como a família viridis ou a paleta de Okabe–Ito. Distinções essenciais não devem depender exclusivamente da cor: combinar cor com formato, tipo de linha, rótulos diretos ou posição garante que as figuras continuem interpretáveis em impressão em tons de cinza e para todos os leitores.



Figura 5.7. Os três tipos de paleta. Sequencial (viridis): quantidades de baixo a alto. Divergente: desvios em torno de um centro com significado, como zero. Qualitativa (Okabe–Ito): categorias sem ordem; é segura para pessoas com deficiência na visão de cores.

O código abaixo aplica uma paleta acessível às curvas de progresso de três cultivares e combina cor com formato dos pontos, para que as curvas continuem distinguíveis mesmo sem cor.

```
ggplot(progresso, aes(x = dias, y = severity,
                      color = cultivar, shape = cultivar)) +
  geom_line(linewidth = 0.9) +
  geom_point(size = 2.5) +
  scale_color_viridis_d(end = 0.85) +
  labs(x = "Dias após a semeadura", y = "Severidade (%)",
       color = "Cultivar", shape = "Cultivar") +
  theme_bw(base_size = 13)
```

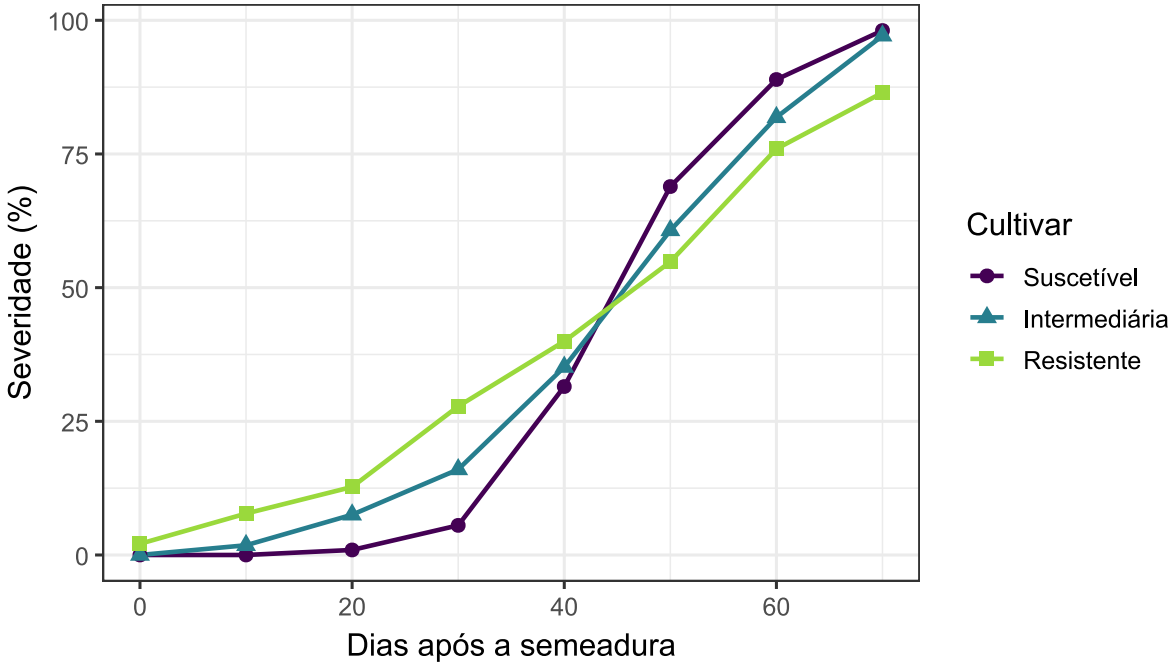


Figura 5.8. Curvas de progresso da doença (dados simulados) com a paleta viridis e formatos diferentes para cada cultivar.

5.9.6 Quando não usar cor

A cor é desnecessária quando duplica informação já mostrada pela posição, como quando cada barra de um gráfico de barras recebe uma cor diferente embora as categorias já estejam

rotuladas no eixo. A cor decorativa distrai e pode sugerir agrupamentos inexistentes. Usar uma única cor neutra para a maioria dos elementos e uma cor contrastante apenas para os elementos que carregam a mensagem é, muitas vezes, a estratégia mais eficaz.

> **Dica**
>
> A cor deve codificar informação ou direcionar a atenção. Se remover uma cor não muda a interpretação da figura, ela talvez não seja necessária.

5.10 Rótulos, escalas e eixos

5.10.1 Rótulos informativos nos eixos

Os rótulos dos eixos devem descrever a variável com palavras que o leitor entende, e não com as abreviações usadas no conjunto de dados. `sev_pct` e `dap` fazem sentido para quem analisa; "Severidade da doença (%)" e "Dias após a semeadura" fazem sentido para o leitor.

5.10.2 Unidades de medida

Todo eixo que representa uma grandeza medida deve incluir sua unidade, geralmente entre parênteses: comprimento da lesão (mm), produtividade (kg/ha), temperatura (°C). As unidades de grandezas derivadas, como taxas ou razões, merecem atenção especial.

5.10.3 Limites dos eixos

Os limites dos eixos determinam o tamanho aparente das diferenças. Em gráficos de barras, o eixo deve começar em zero, porque o comprimento da barra representa a magnitude. Em gráficos de pontos, de linhas e de dispersão, o eixo não precisa incluir o zero, e começar em outro valor pode revelar variação relevante. Truncar um eixo, porém, pode exagerar diferenças pequenas, e a escolha deve refletir qual tamanho de diferença é biologicamente relevante. Para variáveis limitadas, como porcentagens, mostrar a faixa completa de 0 a 100% pode ajudar o leitor a julgar a magnitude das diferenças (Figura 5.9).

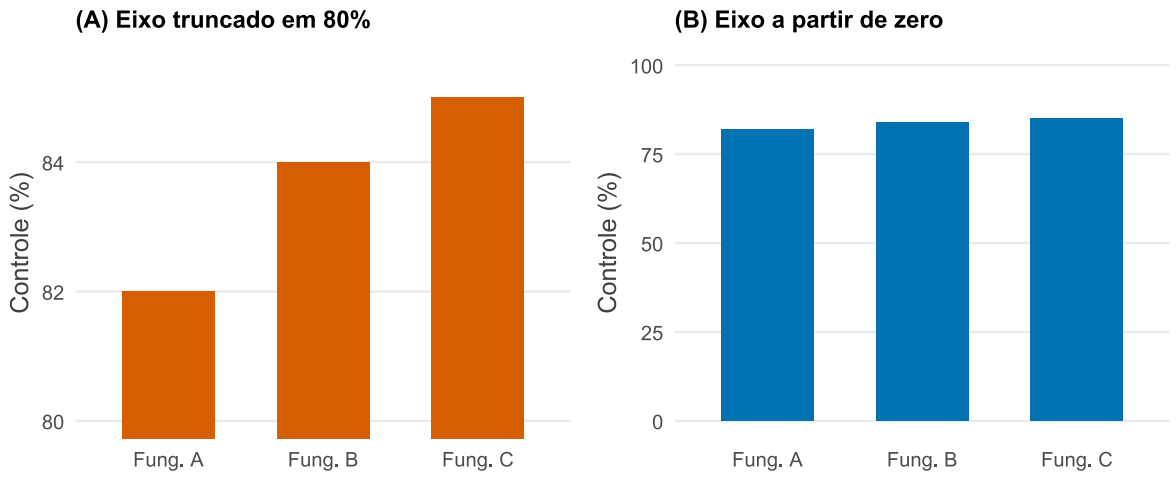


Figura 5.9. O efeito de truncar o eixo de um gráfico de barras. Os dados são os mesmos nos dois painéis (controle de 82, 84 e 85% por três fungicidas). Com o eixo começando em 80% (A), uma diferença de 3 pontos percentuais parece enorme; com o eixo começando em zero (B), ela aparece na proporção correta.

5.10.4 Transformações

Quando os valores abrangem várias ordens de grandeza, como contagens de esporos, quantidades de DNA do patógeno ou valores de CE_{50} de fungicidas, uma escala logarítmica espalha os dados e torna visíveis as diferenças relativas. Os rótulos do eixo devem então mostrar os valores na escala original (1, 10, 100, 1000), e não os logaritmos, para que o leitor possa interpretá-los diretamente.

```
ggplot(isolados, aes(x = reorder(isolate, ec50), y = ec50)) +
  geom_point(size = 2.5, colour = "#0072B2") +
  scale_y_log10() +
  labs(x = "Isolado", y = expression(CE[50]~(mu*g/mL))) +
  theme_bw(base_size = 13) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

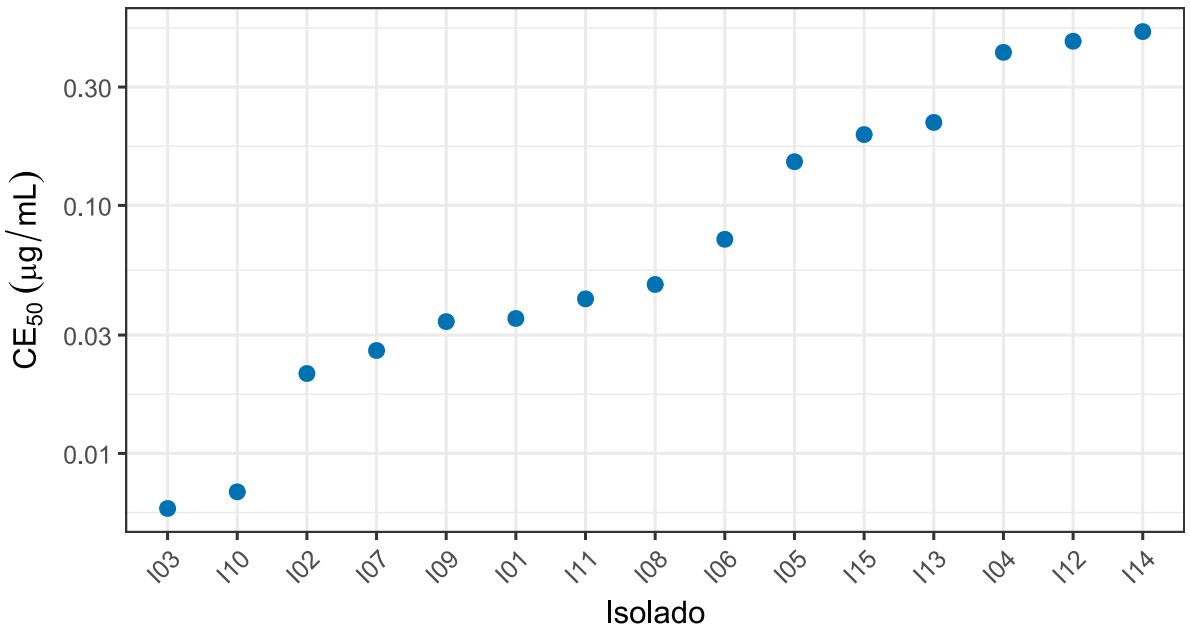


Figura 5.10. Valores de CE_{50} de 15 isolados (dados simulados) em escala logarítmica. Os rótulos do eixo mostram os valores originais, e as diferenças relativas entre isolados sensíveis e menos sensíveis ficam visíveis.

5.10.5 Ordenando categorias

Os programas costumam ordenar as categorias alfabeticamente, o que raramente corresponde a algo significativo. As categorias devem ser ordenadas de forma que ajude a comparação: tratamentos em uma sequência lógica (testemunha primeiro, depois doses crescentes), cultivares segundo sua resposta (da menos à mais doente), meses em ordem cronológica. Ordenar pela resposta torna os ranqueamentos imediatamente visíveis.

```
pA ← ggplot(notas_cv, aes(x = cultivar, y = severity)) +
  geom_point(colour = "grey40") +
  labs(title = "(A) Ordem alfabética", x = NULL, y = "Severidade (%)")

pB ← ggplot(notas_cv, aes(x = reorder(cultivar, severity), y = severity)) +
  geom_point(colour = "#0072B2") +
  labs(title = "(B) Ordenado pela resposta", x = NULL, y = "Severidade (%)")
```

```
pA + pB & theme_bw(base_size = 12)
```

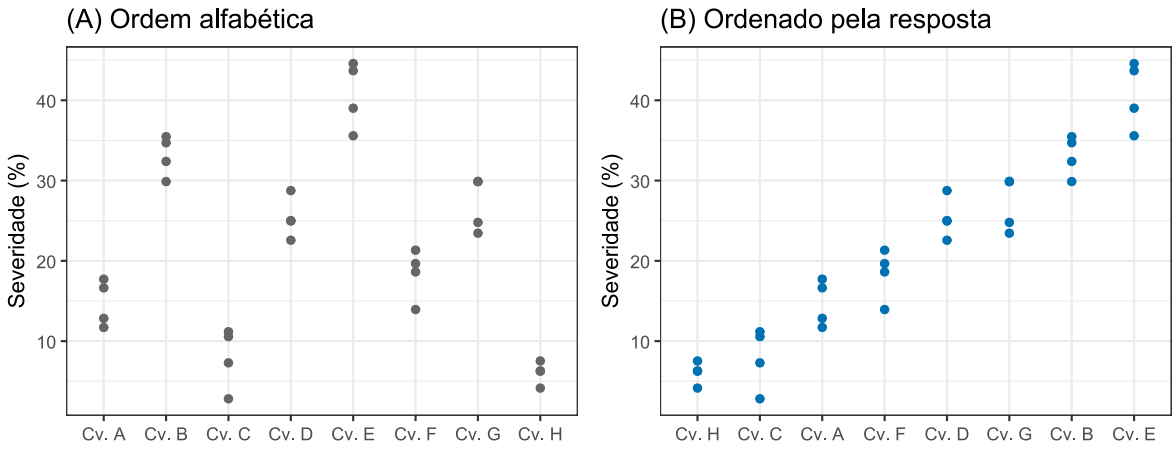


Figura 5.11. Severidade em oito cultivares (dados simulados). (A) Ordem alfabética, o padrão do R. (B) Cultivares ordenadas pela severidade média com `reorder()`: o ranqueamento fica imediatamente visível.

5.10.6 Legendas

As legendas obrigam o leitor a olhar alternadamente para a legenda e para os dados. Quando possível, rotular os grupos diretamente no gráfico, no fim das linhas ou ao lado das nuvens de pontos, facilita a leitura. Quando uma legenda é necessária, ela deve ter um título informativo, listar as categorias na mesma ordem em que aparecem no gráfico e ficar perto dos dados.

5.11 Múltiplas variáveis e múltiplos painéis

5.11.1 Variáveis de agrupamento

Muitas questões envolvem mais de duas variáveis. Uma terceira variável pode ser acrescentada associando-a à cor, ao formato ou ao tipo de linha, de modo que os grupos sejam comparados no mesmo painel. Isso funciona bem quando há poucos grupos e quando a comparação direta entre eles é o objetivo principal, como ao comparar as curvas de progresso da doença de três cultivares.

5.11.2 Facetas

As facetas dividem os dados em subconjuntos e desenham o mesmo gráfico para cada um em painéis separados. São preferíveis ao agrupamento quando há muitos grupos, quando os grupos se sobrepõem muito ou quando cada grupo deve ser examinado isoladamente. Os painéis devem compartilhar as mesmas escalas nos eixos, a menos que haja um bom motivo para não fazê-lo, para que os valores possam ser comparados entre painéis.

5.11.3 Pequenos múltiplos

Uma série de painéis com estrutura idêntica, cada um mostrando um subconjunto diferente, é chamada de conjunto de pequenos múltiplos (*small multiples*) (Tufte 2001). Como o leitor aprende a ler um painel, todos os outros podem ser lidos imediatamente, e a atenção pode se concentrar nas diferenças entre eles. Os pequenos múltiplos estão entre as formas mais eficazes de mostrar conjuntos de dados complexos.

5.11.4 Comparando ambientes

Os dados de Fitopatologia frequentemente vêm de vários locais, safras ou condições ambientais. Mostrar cada ambiente em seu próprio painel, como curvas de progresso da doença para cada local ou relações severidade–produtividade para cada ano, revela se os padrões são consistentes entre ambientes ou mudam com eles, o que muitas vezes é a questão central.

5.11.5 Comparando experimentos

Quando um experimento é repetido, os resultados de cada execução podem ser mostrados em painéis separados ou lado a lado no mesmo painel. Essas figuras mostram se os efeitos dos tratamentos são consistentes entre experimentos, o que é essencial para avaliar a generalidade das conclusões (Capítulo 2). Exemplos típicos incluem respostas cultivar × doença entre experimentos, comparações tratamento × ano em ensaios de fungicidas e curvas de progresso da doença × ambiente.

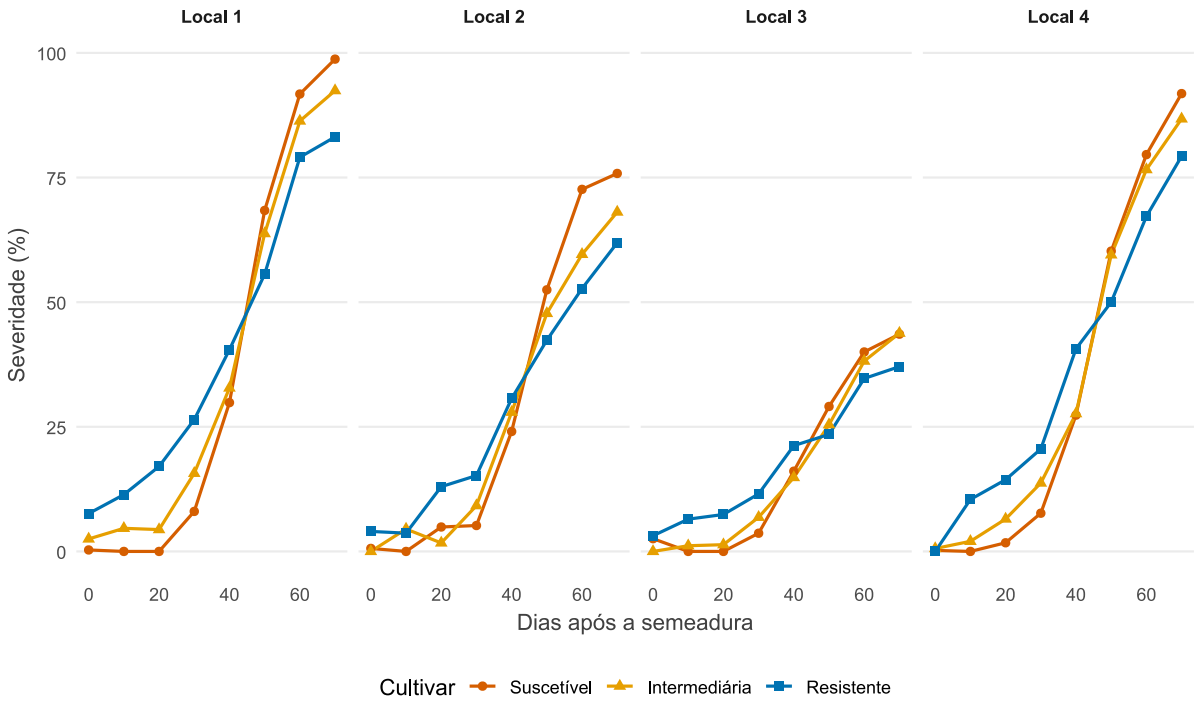


Figura 5.12. Pequenos múltiplos permitem comparar conjuntos de dados complexos usando estruturas gráficas repetidas. Curvas de progresso da doença (dados simulados) de três cultivares em quatro locais: todos os painéis compartilham as mesmas escalas, de modo que diferenças entre locais e entre cultivares podem ser comparadas diretamente.

5.12 Construindo gráficos com o ggplot2

5.12.1 A gramática dos gráficos

O pacote `ggplot2` do R implementa a gramática dos gráficos descrita no início deste capítulo (Wickham 2016). Um gráfico é construído especificando os dados, associando variáveis às estéticas e acrescentando camadas de geometrias, escalas, facetas, rótulos e temas. Todo gráfico segue o mesmo modelo básico:

```
library(ggplot2)

ggplot(data, aes(x = x_variable, y = y_variable)) +
  geom_point()
```

Os exemplos a seguir usam o conjunto de dados simulado `ensaio`: um ensaio de fungicidas com uma testemunha e quatro programas de fungicida, seis blocos e dois ambientes. Ele está organizado no formato descrito no Capítulo 3, com uma linha por parcela:

```
head(ensaio)
```

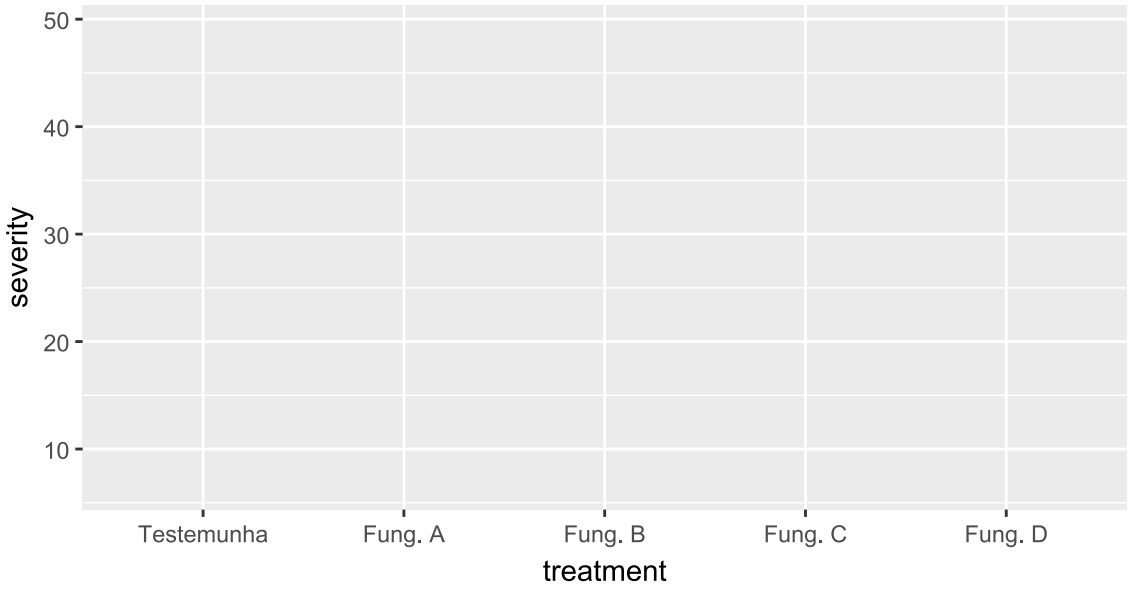
```
  block  treatment environment severity yield
1     1 Testemunha  Ambiente 1 44.46945  2792
2     2 Testemunha  Ambiente 1 40.25662  2951
3     3 Testemunha  Ambiente 1 45.56671  2725
4     4 Testemunha  Ambiente 1 49.16192  2823
5     5 Testemunha  Ambiente 1 42.67102  2420
6     6 Testemunha  Ambiente 1 38.82745  2228
```

Cada passo abaixo acrescenta um componente ao gráfico, e o resultado aparece logo após o código.

5.12.2 Dados e estéticas

A função `ggplot()` recebe o conjunto de dados e os mapeamentos estéticos, definidos com `aes()`. O código abaixo associa o tratamento ao eixo horizontal e a severidade ao eixo vertical. Sozinho, ele produz um painel vazio com eixos, porque nenhuma geometria foi especificada ainda.

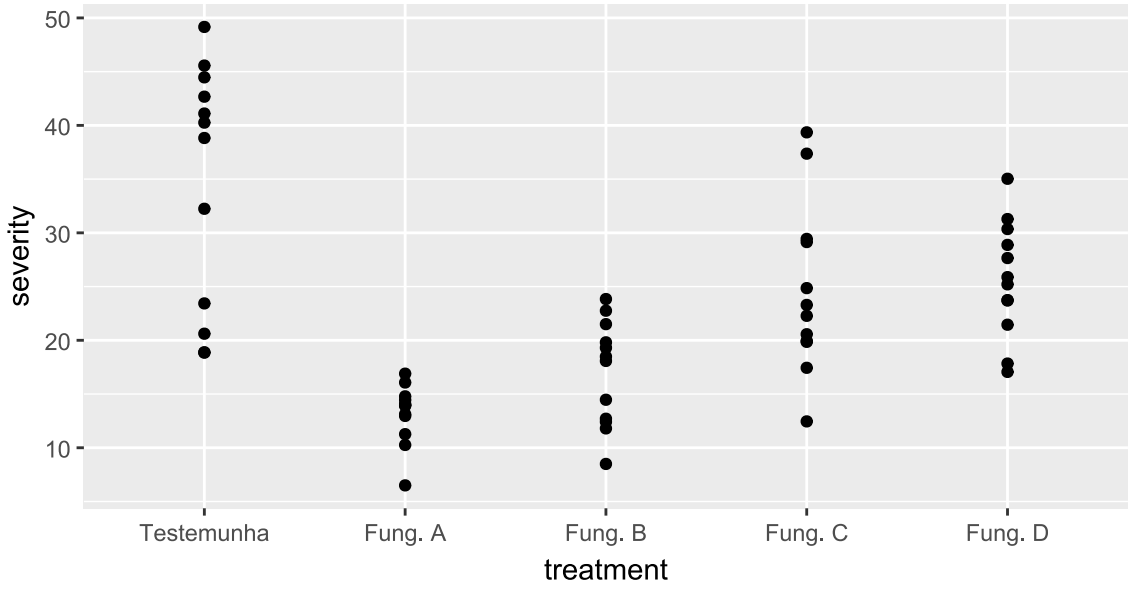
```
ggplot(ensaio, aes(x = treatment, y = severity))
```



5.12.3 Geometrias

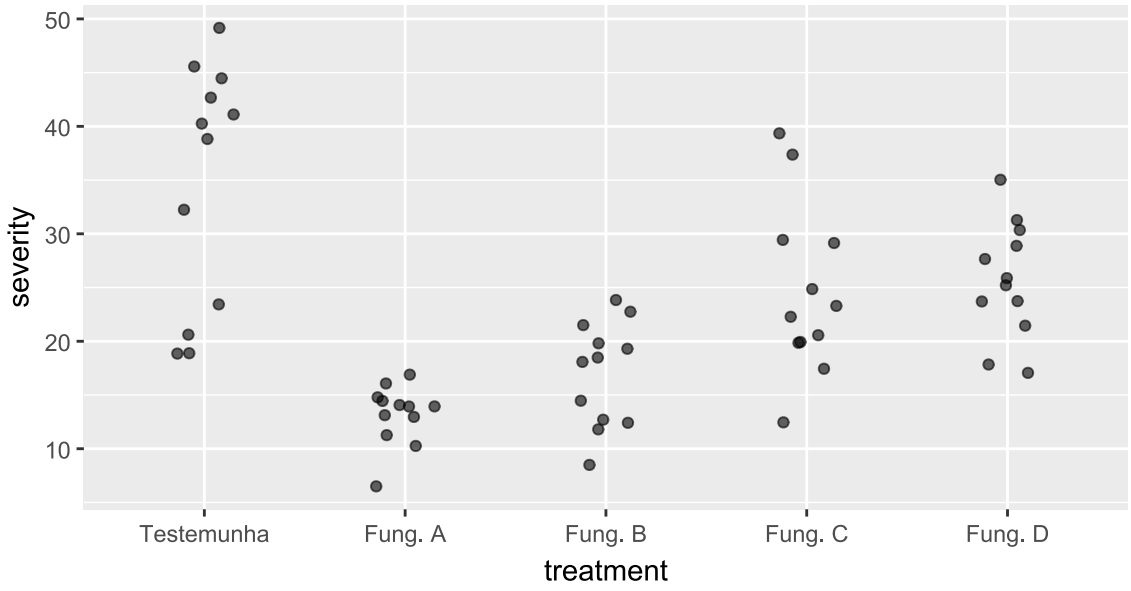
As geometrias são acrescentadas com funções que começam com `geom_`: `geom_point()` para pontos, `geom_line()` para linhas, `geom_boxplot()` para diagramas de caixa, `geom_histogram()` para histogramas e `geom_col()` para barras. Trocar uma geometria por outra muda a representação, mantendo o mesmo mapeamento.

```
ggplot(ensaio, aes(x = treatment, y = severity)) +
  geom_point()
```



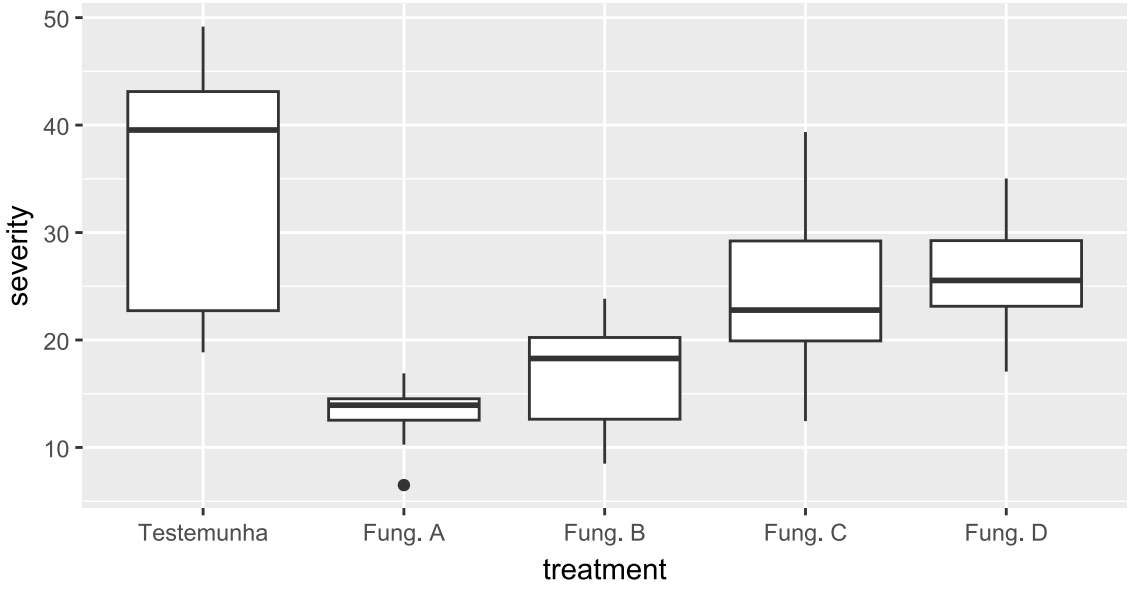
Com 12 parcelas por tratamento, vários pontos se sobrepõem. Trocar `geom_point()` por `geom_jitter()` espalha os pontos horizontalmente, e a transparência (`alpha`) mostra onde eles se acumulam:

```
ggplot(ensaio, aes(x = treatment, y = severity)) +
  geom_jitter(width = 0.15, alpha = 0.6)
```



Com o mesmo mapeamento, `geom_boxplot()` resume a distribuição de cada tratamento:

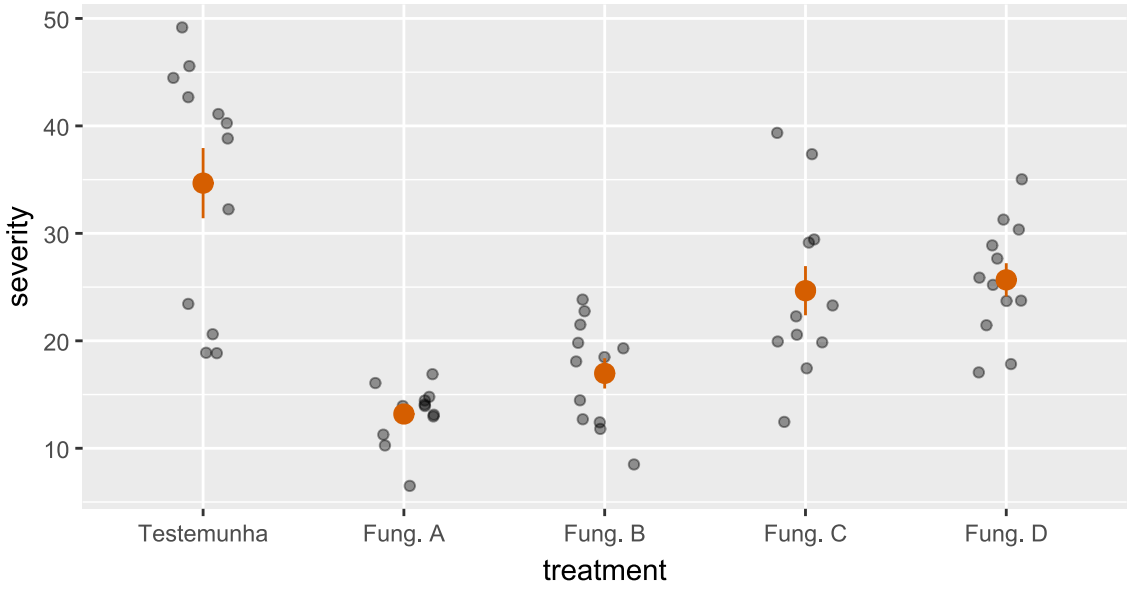
```
ggplot(ensaio, aes(x = treatment, y = severity)) +
  geom_boxplot()
```



5.12.4 Acrescentando camadas

Os gráficos são construídos acrescentando camadas com o operador `+`. Cada camada pode mostrar um aspecto diferente dos dados. Aqui, uma camada de resumos é acrescentada sobre as observações. Por padrão, `stat_summary()` mostra a média e seu erro padrão; os argumentos `colour` e `size` a distinguem visualmente das observações.

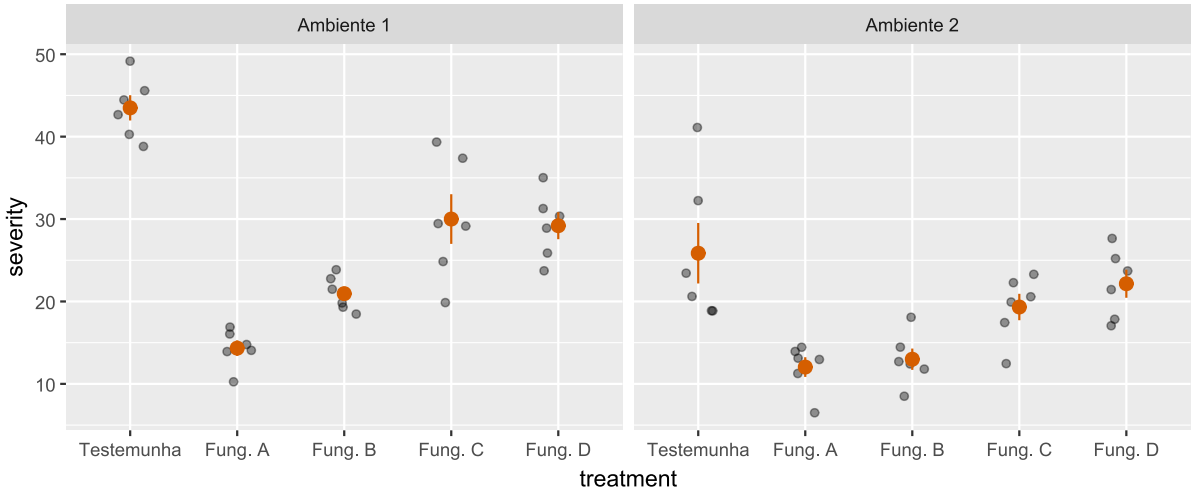
```
ggplot(ensaio, aes(x = treatment, y = severity)) +
  geom_jitter(width = 0.15, alpha = 0.4) +
  stat_summary(colour = "#D55E00", size = 0.6)
```



5.12.5 Facetas

As facetas são acrescentadas com `facet_wrap()` ou `facet_grid()`, que dividem os dados por uma ou duas variáveis e desenham um painel para cada subconjunto. Aqui, cada ambiente aparece em seu próprio painel, com a mesma escala vertical:

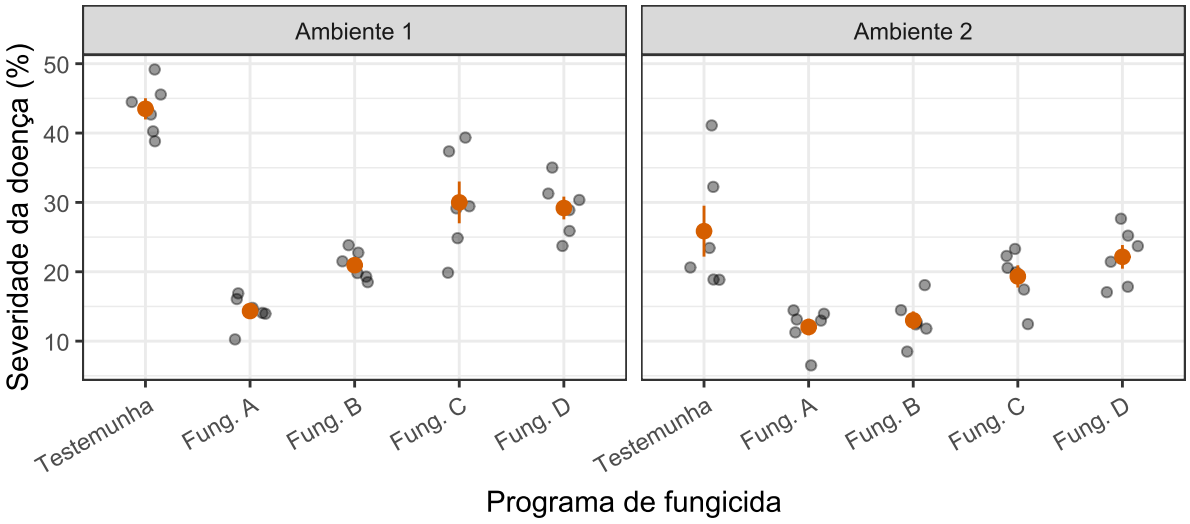
```
ggplot(ensaio, aes(x = treatment, y = severity)) +
  geom_jitter(width = 0.15, alpha = 0.4) +
  stat_summary(colour = "#D55E00", size = 0.5) +
  facet_wrap(~ environment)
```



5.12.6 Rótulos e temas

Os títulos dos eixos, das legendas e as legendas da figura são definidos com `labs()`. Os temas controlam os elementos do gráfico que não são dados, como o fundo, as linhas de grade e o tamanho da fonte. Um tema limpo, como `theme_bw()` ou `theme_classic()`, e um tamanho de fonte maior costumam bastar para figuras científicas. Os temas são secundários em relação às decisões sobre dados, mapeamentos e geometrias: um gráfico bem planejado com o tema padrão comunica melhor do que um gráfico mal planejado com um tema elaborado.

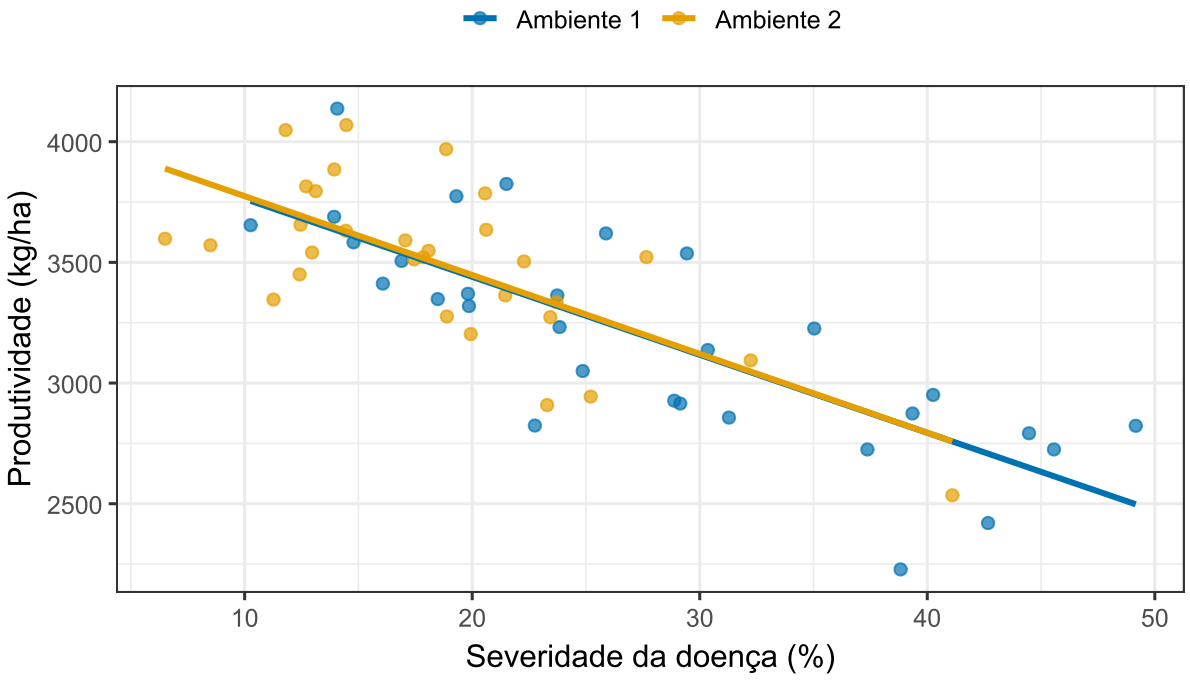
```
ggplot(ensaio, aes(x = treatment, y = severity)) +
  geom_jitter(width = 0.15, alpha = 0.4) +
  stat_summary(colour = "#D55E00", size = 0.5) +
  facet_wrap(~ environment) +
  labs(x = "Programa de fungicida", y = "Severidade da doença (%)") +
  theme_bw(base_size = 14) +
  theme(axis.text.x = element_text(angle = 30, hjust = 1))
```



5.12.7 Juntando tudo

Os mesmos componentes servem para qualquer tipo de gráfico. O exemplo abaixo relaciona a produtividade à severidade da doença, com uma cor por ambiente, uma reta ajustada para cada um e rótulos completos:

```
ggplot(ensaio, aes(x = severity, y = yield, colour = environment)) +
  geom_point(alpha = 0.7, size = 2) +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE) +
  scale_colour_manual(values = c("#0072B2", "#E69F00")) +
  labs(x = "Severidade da doença (%)", y = "Produtividade (kg/ha)",
       colour = NULL) +
  theme_bw(base_size = 13) +
  theme(legend.position = "top")
```



> **Nota**
>
> O objetivo não é memorizar funções do `ggplot2`. A gramática dos gráficos fornece uma forma sistemática de traduzir variáveis e questões científicas em representações gráficas.

5.13 Problemas comuns em figuras científicas

Muitas fragilidades das figuras científicas não são erros nos dados, mas más escolhas gráficas. Os problemas a seguir são frequentes em publicações de Fitopatologia.

5.13.1 Gráficos de barras que escondem os dados

Barras com erros padrão resumem cada grupo com dois números e escondem as observações. São adequadas para contagens e proporções, mas, para respostas contínuas como severidade, tamanho de lesão ou produtividade, gráficos de pontos ou diagramas de caixa com as observações são quase sempre mais informativos.

5.13.2 Uso inadequado de gráficos tridimensionais

Gráficos de barras e de setores tridimensionais acrescentam profundidade sem acrescentar informação. A perspectiva distorce o tamanho aparente dos elementos, tornando os valores mais difíceis de ler e comparar. A menos que a terceira dimensão represente uma variável real e possa ser lida com precisão, os gráficos devem ser bidimensionais.

5.13.3 Decoração excessiva

Linhas de grade pesadas, fundos sombreados, bordas, sombras, degradês e ilustrações competem com os dados pela atenção. Cada elemento de uma figura deve ter uma função. Remover elementos que não carregam informação geralmente faz os dados se destacarem.

5.13.4 Cores demais

Figuras com muitas cores são difíceis de decodificar, especialmente quando a legenda precisa ser consultada repetidamente. Quando é preciso distinguir mais de seis a oito categorias, facetas, rótulos diretos ou o destaque de alguns grupos de interesse são opções melhores.

5.13.5 Escalas mal escolhidas

Escalas lineares para dados que abrangem várias ordens de grandeza comprimem a maioria das observações em um canto do gráfico. Escalas que se estendem muito além dos dados desperdiçam espaço e achatam as diferenças. As escalas devem ser escolhidas para mostrar claramente a amplitude e a variação dos dados.

5.13.6 Escalas inconsistentes entre painéis

Quando os painéis usam faixas diferentes nos eixos, os leitores tendem a compará-los como se as escalas fossem iguais, e as diferenças de magnitude são mal interpretadas. Painéis que devem ser comparados devem compartilhar a mesma escala. Se escalas diferentes forem necessárias, isso deve ser informado com clareza.

5.13.7 Observações sobrepostas

Quando muitas observações têm valores iguais ou semelhantes, como é comum em avaliações visuais de severidade, os pontos se sobrepõem e o gráfico sugere menos observações do que

realmente existem. Deslocamento aleatório (*jitter*), transparência ou gráficos que mostram contagens ajudam a revelar todas as observações (Figura 5.13).

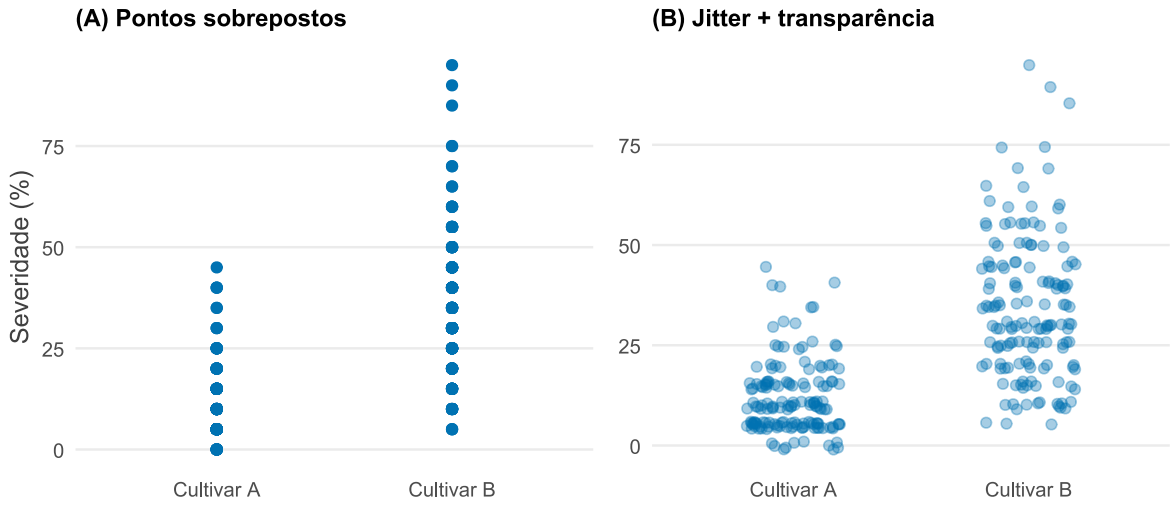


Figura 5.13. Observações sobrepostas (dados simulados). Notas visuais de severidade arredondadas para múltiplos de 5% em 300 plantas de duas cultivares. (A) Com pontos simples, cada valor parece ter sido observado uma única vez. (B) Com deslocamento aleatório e transparência, a concentração das observações fica visível.

5.13.8 Legendas desnecessárias

Uma legenda que repete informação já mostrada em um eixo, ou uma legenda para um único grupo, acrescenta poluição visual sem informação. As legendas devem aparecer apenas quando necessárias para decodificar o gráfico.

5.13.9 Textos pequenos e rótulos ilegíveis

As figuras costumam ser reduzidas para caber em uma coluna de revista ou projetadas em uma sala grande. Um texto legível na tela do computador pode se tornar ilegível. O tamanho das fontes deve ser escolhido para o tamanho final da figura, e os rótulos dos eixos não devem se sobrepor.

5.13.10 Ligar observações que não deveriam ser ligadas

Linhas sugerem continuidade e conexão. Ligar por linhas as médias de diferentes cultivares ou tratamentos sugere uma tendência entre categorias que não têm ordem natural. As linhas devem ligar observações apenas ao longo de uma dimensão contínua, como tempo, dose ou temperatura, ou medidas repetidas da mesma unidade.

5.13.11 Gráficos que sugerem relações sem fundamento

Linhas ajustadas, curvas suavizadas e extrapolações sugerem relações que podem não ser sustentadas pelos dados ou pelo delineamento. Uma reta de regressão traçada através de dois grupos de pontos de locais diferentes, ou uma curva estendida além da faixa de temperaturas observadas, sugere mais do que a evidência sustenta.

5.14 Visualizações enganosas

Algumas figuras criam impressões falsas sobre a evidência. Às vezes isso é deliberado, mas com muito mais frequência resulta de escolhas padrão não examinadas. Em qualquer caso, o efeito sobre o leitor é o mesmo.

5.14.1 Eixos truncados

Um gráfico de barras cujo eixo vertical começa em 80 em vez de 0 faz uma diferença entre 82% e 85% de controle parecer um grande efeito. Como o comprimento da barra representa a magnitude, eixos truncados em gráficos de barras são sempre enganosos. Para pontos e linhas, o truncamento é aceitável, mas o leitor deve conseguir julgar o tamanho das diferenças em relação a valores biologicamente relevantes.

5.14.2 Distorções de área e volume

Quando os símbolos são dimensionados pela altura ou largura, e não pela área, um valor duas vezes maior é representado por um símbolo quatro vezes maior. Pictogramas, ícones escalonados e bolhas cujo diâmetro, e não a área, é proporcional ao valor exageram as diferenças.

5.14.3 Intervalos gráficos desiguais

Eixos em que distâncias iguais representam intervalos desiguais, como datas de avaliação colocadas em espaçamento igual embora estivessem separadas por 7, 14 e 30 dias, distorcem taxas e formas. O tempo e outras variáveis contínuas devem ser representados em uma escala contínua adequada.

5.14.4 Apresentação seletiva dos dados

Mostrar apenas os experimentos, anos, locais ou tratamentos que sustentam uma conclusão, ou recortar uma série temporal para começar ou terminar em pontos convenientes, dá uma visão distorcida da evidência. As figuras devem representar todos os dados relevantes, ou informar claramente o que foi selecionado e por quê.

5.14.5 Suavização e agregação

A suavização e as médias simplificam dados complexos, mas também podem esconder variabilidade, valores discrepantes e padrões. Médias entre locais ou anos podem esconder efeitos opostos; uma curva suavizada pode sugerir uma tendência determinada por poucos pontos. Os resumos devem ser acompanhados, quando possível, pelos dados em que se baseiam.

> ⚠ **Aviso**
>
> Um gráfico pode estar numericamente correto e ainda assim criar uma impressão enganosa da evidência.

5.15 Elaborando figuras para publicações científicas

5.15.1 Uma figura, uma mensagem principal

Cada figura de uma publicação deve ter uma mensagem principal clara, que geralmente pode ser enunciada em uma frase: “Os programas de fungicida A e B reduziram a severidade em comparação com a testemunha nos dois anos.” Planejar a figura para tornar essa mensagem visível — colocando a comparação-chave em um eixo comum e dispondo os grupos de modo a facilitá-la — é mais importante do que qualquer escolha de estilo.

5.15.2 Painéis

Figuras com vários painéis combinam gráficos relacionados, como severidade e produtividade, ou resultados de vários experimentos. Os painéis devem ser identificados (A, B, C), dispostos em uma ordem lógica de leitura, alinhados em eixos comuns quando possível e referidos na legenda. Gráficos não relacionados devem ser figuras separadas.

5.15.3 Linguagem visual consistente

Dentro de uma figura, e em todas as figuras de uma publicação, o mesmo tratamento ou cultivar deve ter a mesma cor e o mesmo símbolo, a mesma variável deve ter o mesmo rótulo de eixo e o mesmo estilo deve ser usado para observações, médias e intervalos. A consistência permite ao leitor transferir o que aprendeu de uma figura para a seguinte.

5.15.4 Legendas das figuras

A legenda deve permitir que a figura seja compreendida sem a leitura do texto principal. Deve informar o que é mostrado, o delineamento experimental, o número de observações ou repetições, o que os pontos, linhas e barras de erro representam e quaisquer abreviações usadas. Uma boa legenda muitas vezes começa com uma frase que enuncia a mensagem principal da figura.

5.15.5 Resolução e dimensões

As revistas especificam as dimensões das figuras (frequentemente a largura de uma ou duas colunas) e a resolução mínima para imagens rasterizadas, tipicamente 300 dpi ou mais para fotografias e mais ainda para desenhos de linhas. Produzir as figuras nas dimensões finais, em vez de redimensioná-las depois, garante que textos e símbolos apareçam no tamanho pretendido.

5.15.6 Formatos vetoriais e rasterizados

Formatos vetoriais, como PDF, SVG e EPS, guardam as figuras como formas e texto e podem ser ampliados sem perda de qualidade; são ideais para gráficos. Formatos rasterizados, como PNG, TIFF e JPEG, guardam as figuras como grades de pixels, e a qualidade depende da resolução; são necessários para fotografias. PNG e TIFF preservam linhas e textos nítidos, enquanto a compressão do JPEG pode borrá-los. As exigências variam entre revistas e devem ser conferidas antes da submissão.

```
ggsave("figures/fig1_severity.pdf", width = 17, height = 10, units = "cm")
ggsave("figures/fig1_severity.png", width = 17, height = 10, units = "cm", dpi = 600)
```

5.15.7 Planejando para o formato final

A mesma figura pode precisar ser adaptada a diferentes formatos. Na tela e em artigos de revista, o leitor pode examinar os detalhes no seu próprio ritmo. Em apresentações, o público tem segundos para captar a mensagem, de modo que as figuras devem ser mais simples, com textos maiores e menos painéis. Pôsteres são vistos à distância e devem enfatizar uma mensagem clara por figura. Planejar para o formato final significa considerar o tamanho, a distância de visualização e o tempo disponível para o leitor.

5.16 Pesquisa na prática: aprimorando uma figura científica

Um ensaio de campo comparou a severidade de uma doença foliar em uma testemunha e quatro programas de fungicida, em blocos casualizados com seis blocos. O exemplo é ilustrativo, mas a sequência de melhorias se aplica à maioria dos experimentos que comparam tratamentos. Partindo de uma figura convencional, porém fraca, a figura é aprimorada passo a passo.

1. **O ponto de partida.** Barras mostram as médias dos tratamentos em ordem alfabética, com barras de erro padrão, fundo cinza, uma cor diferente para cada barra e o rótulo de eixo `sev`. A figura mostra cinco números e seus erros padrão, e nada mais.
2. **Mostrar as observações brutas.** Substituir as barras pelos seis valores das parcelas de cada tratamento revela que um programa de fungicida tem uma parcela com severidade excepcionalmente alta e que a testemunha é muito mais variável que as parcelas tratadas.
3. **Mostrar a distribuição.** Com seis observações por tratamento, os próprios pontos mostram a distribuição. Com mais observações, um diagrama de caixa ou de violino poderia ser acrescentado por trás deles.
4. **Acrescentar os resumos dos tratamentos.** As médias são acrescentadas como pontos maiores, distinguíveis das observações, para que a comparação de interesse seja fácil de ver.
5. **Representar a incerteza.** São acrescentados intervalos de confiança para as médias dos tratamentos, e a legenda informa o que eles representam.
6. **Ordenar de forma significativa.** A testemunha é colocada primeiro, seguida dos programas de fungicida em ordem crescente de eficácia, para que o ranqueamento fique imediatamente visível.
7. **Rotular e finalizar.** O rótulo do eixo passa a ser “Severidade da doença (%)”, as cores redundantes e a legenda são removidas, uma única cor de destaque é usada para as médias, o tema é simplificado e o tamanho da fonte é aumentado para a largura final da figura. A legenda informa o delineamento, o número de parcelas e o que os pontos e intervalos representam.

A Figura 5.14 mostra o ponto de partida e o resultado final, usando as parcelas do Ambiente 1 do conjunto `ensaio`. O código abaixo produz as duas versões.

```
trial ← subset(ensaio, environment == "Ambiente 1")

# (A) Ponto de partida
antes ← ggplot(trial, aes(x = as.character(treatment), y = severity,
                          fill = as.character(treatment))) +
  stat_summary(fun = mean, geom = "col") +
  stat_summary(fun.data = mean_se, geom = "errorbar", width = 0.3) +
```

```
  labs(title = "(A) Antes", x = "trat", y = "sev", fill = "trat")

# (B) Versão final: testemunha primeiro, depois em ordem de eficácia
ordem ← c("Testemunha",
          names(sort(tapply(trial$severity, trial$treatment, mean)[-1])))
trial$treatment ← factor(trial$treatment, levels = ordem)

depois ← ggplot(trial, aes(x = treatment, y = severity)) +
  geom_jitter(width = 0.08, alpha = 0.4, size = 2) +
  stat_summary(fun.data = media_ic95, colour = "#0072B2",
               size = 0.6, linewidth = 1.1) +
  scale_y_continuous(limits = c(0, NA)) +
  labs(title = "(B) Depois", x = NULL, y = "Severidade da doença (%)") +
  theme_classic(base_size = 14) +
  theme(axis.text.x = element_text(angle = 30, hjust = 1))

antes + depois
```

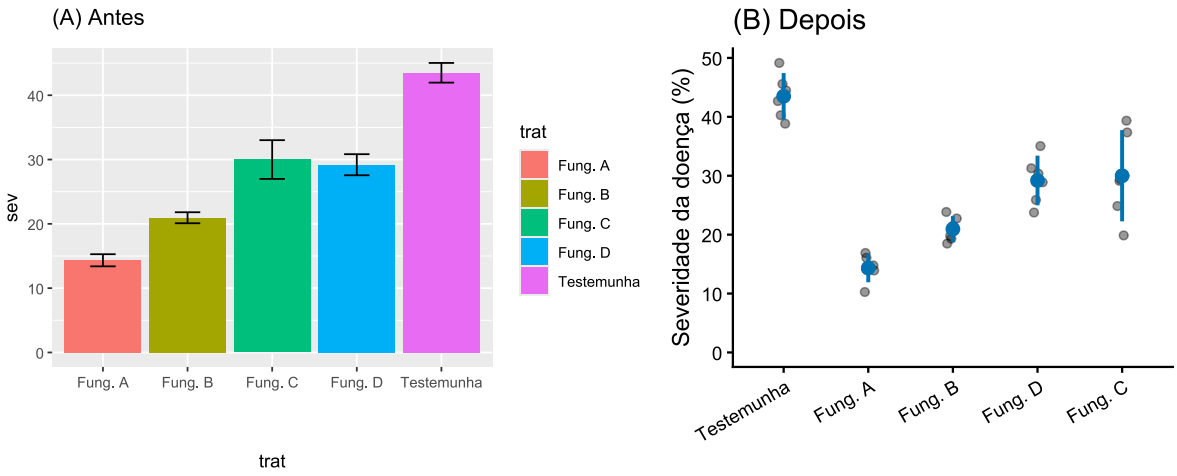


Figura 5.14. Aprimorando uma figura (dados simulados; seis parcelas por tratamento). (A) Ponto de partida: barras com erro padrão em ordem alfabética, cores redundantes, legenda desnecessária e rótulo abreviado. (B) Versão final: observações individuais, média e intervalo de confiança de 95% em uma única cor de destaque, testemunha primeiro e tratamentos em ordem de eficácia, rótulos completos e tema limpo.

Pesquisa revisitada: Dos dados à figura científica

Considere um experimento que compara a severidade da doença entre tratamentos.

Questão

Qual representação gráfica comunica com mais clareza a evidência experimental?

Compare várias alternativas:

1. tabela de médias dos tratamentos;
2. gráfico de barras das médias;
3. diagramas de caixa;
4. observações brutas;
5. observações combinadas com os efeitos estimados dos tratamentos e sua incerteza.

Para cada representação, pergunte:

- Que informação está visível?
- Que informação está escondida?
- Que comparações são fáceis de fazer?
- Que conclusão biológica a figura enfatiza?

5.17 Um fluxo de trabalho para a visualização científica

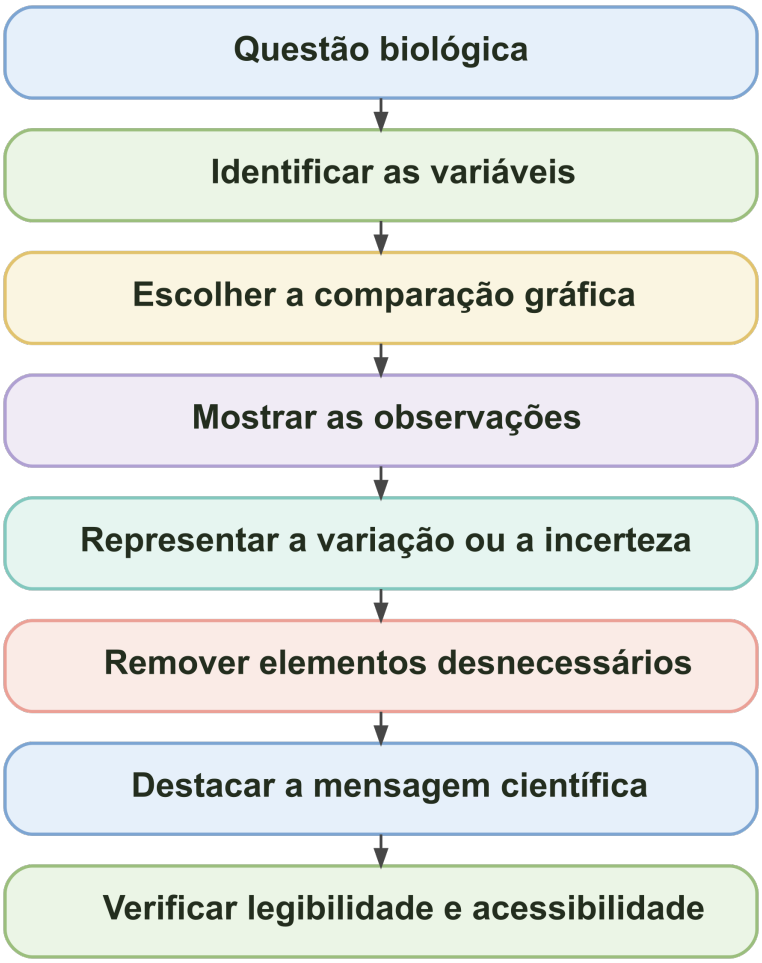


Figura 5.15. Um fluxo de trabalho prático para construir gráficos científicos.

5.18 Lista de verificação da figura

Antes de finalizar uma figura científica, pergunte:

1. Que questão científica a figura aborda?
2. O que o leitor deve notar primeiro?
3. As observações subjacentes estão visíveis, quando apropriado?
4. As comparações são fáceis de fazer?
5. Os eixos e as unidades estão claramente identificados?
6. A incerteza está representada de forma adequada?
7. A cor codifica informação útil?
8. A figura pode ser compreendida sem depender exclusivamente da cor?
9. Todos os elementos gráficos são necessários?

10. Algum aspecto da figura poderia enganar o leitor?
11. A figura é legível no tamanho final de publicação?
12. A legenda permite interpretar a figura de forma independente?

5.19 Resumo do capítulo

A visualização científica é ao mesmo tempo uma ferramenta analítica e de comunicação. Figuras eficazes revelam padrões, variação, relações e incerteza, direcionando a atenção para a evidência biológica relevante para a questão de pesquisa. As escolhas gráficas devem, portanto, ser orientadas pela estrutura dos dados e pelas comparações de interesse científico, e não por convenções ou decoração. A gramática dos gráficos fornece um arcabouço sistemático para traduzir essas decisões em visualizações reprodutíveis no R.

5.20 Conceitos-chave

Ao concluir este capítulo, você deverá ser capaz de explicar:

- visualização científica;
- gráfico exploratório;
- gráfico explicativo;
- mapeamento gráfico;
- geometria;
- escala;
- posição;
- distribuição;
- variabilidade versus incerteza;
- incerteza;
- pequenos múltiplos;
- facetas;
- acessibilidade gráfica;
- gramática dos gráficos.

5.21 Exercícios

5.21.1 Exercício 1. O que o gráfico revela?

Examine uma figura fornecida pelo professor sem ler sua legenda nem o texto que a acompanha.

Identifique:

1. as variáveis representadas;
2. a comparação biológica aparente;
3. o padrão principal;
4. a variação entre as observações;
5. qualquer informação difícil de interpretar.

5.21.2 Exercício 2. Escolhendo uma representação gráfica

Para cada questão de pesquisa, proponha uma representação gráfica adequada:

1. A severidade da doença difere entre cinco cultivares?

2. A severidade da doença está associada à temperatura?
3. Como a severidade da doença muda durante uma epidemia?
4. Quão variável é a eficácia de um fungicida entre experimentos?
5. A relação entre doença e produtividade difere entre ambientes?

Justifique cada escolha.

5.21.3 Exercício 3. Aprimorando uma figura

Você recebe uma figura científica mal elaborada.

Identifique problemas envolvendo:

- a representação gráfica;
- as escalas;
- os rótulos;
- as cores;
- as legendas;
- elementos desnecessários;
- a representação da variabilidade.

Refaça a figura.

5.21.4 Exercício 4. Dos dados à figura

Usando um conjunto de dados fornecido pelo professor:

1. identifique a questão biológica;
2. selecione as variáveis necessárias para abordá-la;
3. construa uma visualização exploratória;
4. construa uma figura adequada para comunicar o resultado principal;
5. explique por que as duas figuras podem diferir.

5.21.5 Exercício 5. Avaliando criticamente figuras publicadas

Escolha uma figura de um artigo recente de Fitopatologia.

Avalie:

1. Que questão científica a figura aborda?
2. A representação gráfica escolhida é adequada?
3. As observações individuais estão visíveis?
4. A variabilidade ou a incerteza está representada?
5. As comparações são fáceis de fazer?
6. A figura poderia ser simplificada?
7. Você a elaboraria de outra forma?

5.22 Reflexão

> Uma boa figura científica não apenas mostra os dados; ela torna a evidência mais fácil de compreender.

Discuta essa afirmação usando um exemplo da Fitopatologia.

5.23 Próximos passos

A visualização permite reconhecer padrões e comunicar evidências, mas diferenças visuais, por si só, não estabelecem se um efeito observado é compatível com a variação aleatória nem com que precisão um efeito foi estimado. O próximo capítulo apresenta a inferência estatística e o raciocínio usado para passar dos dados observados às conclusões científicas.

CAPÍTULO 6

Como tiramos conclusões científicas?

6.1 Objetivos de aprendizagem

! Importante

Ao concluir este capítulo, o leitor deverá ser capaz de:

- explicar a distinção entre população, amostra, parâmetro e estimativa;
- definir a grandeza de interesse científico antes de escolher um método estatístico;
- explicar a variabilidade amostral e a incerteza estatística;
- distinguir estimação de teste de hipóteses;
- interpretar tamanhos de efeito e intervalos de confiança;
- formular hipóteses científicas e estatísticas;
- explicar o significado e as limitações do valor-P;
- distinguir significância estatística de importância biológica;
- reconhecer os erros dos tipos I e II;
- compreender os papéis do poder estatístico e do tamanho da amostra;
- avaliar conclusões científicas com base na magnitude, na incerteza e na relevância biológica dos efeitos estimados.

6.2 Das observações à inferência

Os capítulos anteriores trataram do delineamento de estudos, da organização dos dados, da exploração das observações e da visualização de padrões. Agora enfrentamos uma questão diferente: **que conclusões podem ser razoavelmente tiradas dos dados?**

Suponha que, em um ensaio de campo, as parcelas pulverizadas com um fungicida tenham tido severidade média de 40%, enquanto as parcelas não tratadas tiveram 60%. A diferença observada é de 20 pontos percentuais. Mas o pesquisador raramente se interessa apenas por essas parcelas específicas nessa safra específica. A verdadeira pergunta é sobre o fungicida: ele reduz a doença, e em quanto, em condições como as do ensaio? Se o ensaio tivesse sido conduzido em outras parcelas da mesma lavoura, com outras plantas, a diferença observada não teria sido exatamente 20 pontos percentuais. Poderia ter sido 14 ou 27.

A diferença observada, portanto, não é o mesmo que o efeito subjacente do fungicida. É uma estimativa desse efeito, influenciada pela variabilidade natural das parcelas, das plantas e da doença. A **inferência estatística** é o raciocínio que liga o que foi observado ao que queremos saber. Ela responde a perguntas como: Qual é a nossa melhor estimativa do efeito? Com que

precisão ele foi estimado? Quais tamanhos de efeito são compatíveis com os dados? Uma diferença desse tamanho poderia surgir plausivelmente se o tratamento não tivesse efeito?

Este capítulo apresenta os conceitos em que esse raciocínio se baseia. Eles são comuns a todos os métodos estatísticos apresentados nos capítulos seguintes, das comparações entre grupos aos modelos não lineares.

> **! Importante**
>
> A inferência estatística é o processo de usar os dados observados, junto com pressupostos sobre como esses dados foram gerados, para quantificar o que pode ser razoavelmente concluído além das próprias observações.

6.3 O que queremos estimar?

6.3.1 A questão científica vem primeiro

Um ponto de partida comum na análise de dados é perguntar: "Que teste estatístico devo usar?" Essa pergunta vem cedo demais. Antes de escolher um método, o pesquisador precisa decidir que grandeza responderia à questão biológica.

Considere a questão: Um fungicida reduz a doença? Várias grandezas poderiam respondê-la:

- a diferença na severidade média entre parcelas tratadas e não tratadas na última avaliação;
- a redução relativa da severidade, expressa como porcentagem da testemunha;
- a diferença na incidência da doença em um estádio fenológico específico;
- a diferença na taxa de progresso da doença durante a safra;
- a diferença de produtividade entre parcelas tratadas e não tratadas.

Cada uma dessas grandezas responde a uma pergunta ligeiramente diferente, e cada uma pode levar a uma conclusão diferente. Um fungicida pode reduzir muito a severidade final e ter pouco efeito sobre a produtividade, ou pode atrasar a epidemia sem mudar seu nível final. Escolher a grandeza de interesse é uma decisão científica, e não estatística.

6.3.2 Da questão biológica ao estimando

A grandeza que um estudo pretende conhecer é chamada **estimando**. Defini-lo com precisão é a ponte entre a questão biológica e a análise estatística. Exemplos de estimandos em Fitopatologia incluem:

- a severidade média de uma cultivar em determinadas condições;
- a diferença entre a severidade média de parcelas tratadas e não tratadas;
- a redução relativa da doença causada por um tratamento;
- a inclinação que relaciona a produtividade à severidade da doença;
- a chance de infecção de plantas expostas a determinada duração de molhamento em relação a plantas expostas a uma duração menor;
- a temperatura em que a infecção é máxima.

Um estimando bem definido especifica o que é comparado, em que unidades, em que momento e em que condições. "O efeito do fungicida" é vago; "a diferença na severidade média (pontos percentuais) entre parcelas pulverizadas com o fungicida A e parcelas não tratadas na última avaliação, sob infecção natural na região-alvo" é um estimando.

> **Nota**
>
> Antes de perguntar **"Que teste estatístico devo usar?"**, pergunte **"Que grandeza eu quero estimar?"**

6.3.3 Parâmetro e estimativa

O valor verdadeiro do estimando, que só seria conhecido se pudéssemos observar todas as unidades em todas as condições relevantes, é chamado **parâmetro**. Ele é fixo, mas desconhecido: por exemplo, a redução média verdadeira da severidade que o fungicida produziria em todas as parcelas que poderiam ter sido usadas no ensaio. O valor calculado a partir dos dados é uma **estimativa** do parâmetro: por exemplo, a diferença observada de 20 pontos percentuais.

A distinção é central para a inferência. O parâmetro descreve a realidade biológica que queremos compreender; a estimativa é nossa melhor aproximação, obtida a partir de dados limitados. As estimativas variam de estudo para estudo; o parâmetro, não. Os métodos estatísticos quantificam o quanto a estimativa provavelmente se afasta do parâmetro.

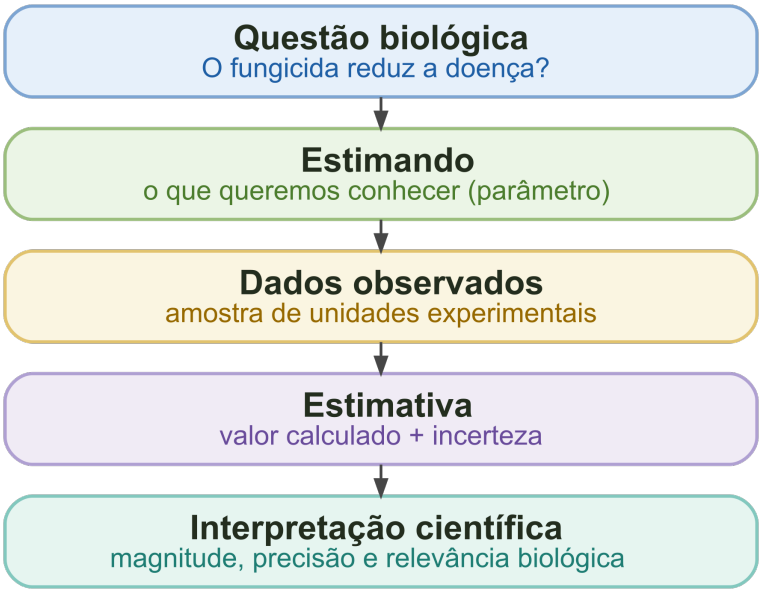


Figura 6.1. Da questão biológica ao estimando, aos dados observados, à estimativa e à interpretação científica.

6.4 População e amostra

6.4.1 Qual é a população de interesse?

Em estatística, uma **população** é o conjunto de todas as unidades, ou de todas as observações possíveis, sobre as quais se quer tirar conclusões. Em Fitopatologia, a população raramente significa todas as plantas do mundo. Pode ser:

- plantas de uma cultivar cultivadas em condições definidas de casa de vegetação;
- parcelas em lavouras de uma região produtora-alvo;
- o conjunto de ambientes (safras e locais) em que um fungicida seria usado;
- isolados de uma população do patógeno em um país;
- lavouras comerciais de uma cultura em um estado em determinado ano.

Definir a população faz parte de definir o estimando. Os mesmos dados podem sustentar conclusões sobre uma população e não sobre outra.

6.4.2 O que é uma amostra?

Uma **amostra** é o conjunto de unidades efetivamente observadas: as parcelas do ensaio, as plantas inoculadas, os isolados coletados, as lavouras visitadas. A inferência é o uso da amostra para aprender sobre a população. Para que seja válida, a amostra precisa representar a população de forma conhecida. Em levantamentos, isso é obtido por amostragem aleatória ou sistemática das unidades da população. Em experimentos, a alocação aleatória dos tratamentos às unidades garante que os grupos de tratamento sejam comparáveis, de modo que as diferenças entre grupos possam ser atribuídas aos tratamentos.

6.4.3 Para onde podemos generalizar?

A população para a qual as conclusões podem ser estendidas é determinada pelo delineamento do estudo (Capítulo 2). Um experimento bem casualizado conduzido em uma lavoura em uma safra sustenta conclusões sobre o efeito do tratamento naquela lavoura e naquela safra. Estendê-las a outras lavouras, cultivares ou safras exige repetir o experimento nessas condições ou supor que o efeito é semelhante — um pressuposto biológico, e não estatístico.

Da mesma forma, um experimento em casa de vegetação com um único isolado sustenta conclusões sobre aquele isolado em condições de casa de vegetação. Um levantamento de lavouras escolhidas por conveniência à beira das estradas pode não representar todas as lavouras de uma região. A casualização dentro de um experimento garante comparações válidas entre tratamentos, mas não garante, por si só, que os resultados se apliquem a todos os ambientes.

> ⚠ **Aviso**
>
> O escopo da inferência é determinado principalmente por como o estudo foi delineado e por como as unidades experimentais ou observacionais foram obtidas — e não apenas pelo modelo estatístico.

6.5 Por que as estimativas variam?

6.5.1 Variação amostral

Imagine que o ensaio de fungicida pudesse ser repetido muitas vezes, cada vez com um novo conjunto de parcelas da mesma lavoura e da mesma safra. Mesmo que o fungicida tivesse exatamente o mesmo efeito todas as vezes, a diferença observada entre parcelas tratadas e não tratadas variaria: 20 pontos em um ensaio, 15 em outro, 26 em um terceiro. Essa variação surge porque cada ensaio usa parcelas diferentes, que diferem em solo, microclima e inóculo, e plantas diferentes, que diferem em suscetibilidade.

Essa variação entre as estimativas que seriam obtidas a partir de amostras diferentes é chamada **variação amostral**. Ela afeta toda grandeza estimada: médias, diferenças, proporções, inclinações e correlações. É o motivo pelo qual um único estudo raramente fornece uma resposta definitiva e pelo qual as estimativas sempre devem vir acompanhadas de uma medida de sua incerteza.

6.5.2 Distribuições amostrais

Se pudéssemos registrar as estimativas de muitas repetições hipotéticas de um estudo, sua distribuição seria chamada **distribuição amostral** da estimativa. Ela descreve o quanto a estimativa varia de amostra para amostra e se tende a estar centrada no valor verdadeiro.

Na prática, um estudo é conduzido uma única vez, e a distribuição amostral não pode ser observada diretamente. A teoria estatística, porém, permite inferir sua forma e sua dispersão a partir de uma única amostra, sob pressupostos sobre como os dados foram gerados. Dois fatos gerais são úteis. Primeiro, estimativas baseadas em mais observações independentes variam menos. Segundo, estimativas baseadas em observações mais variáveis variam mais. Esses dois fatos fundamentam os conceitos de erro padrão, intervalos de confiança e poder estatístico.

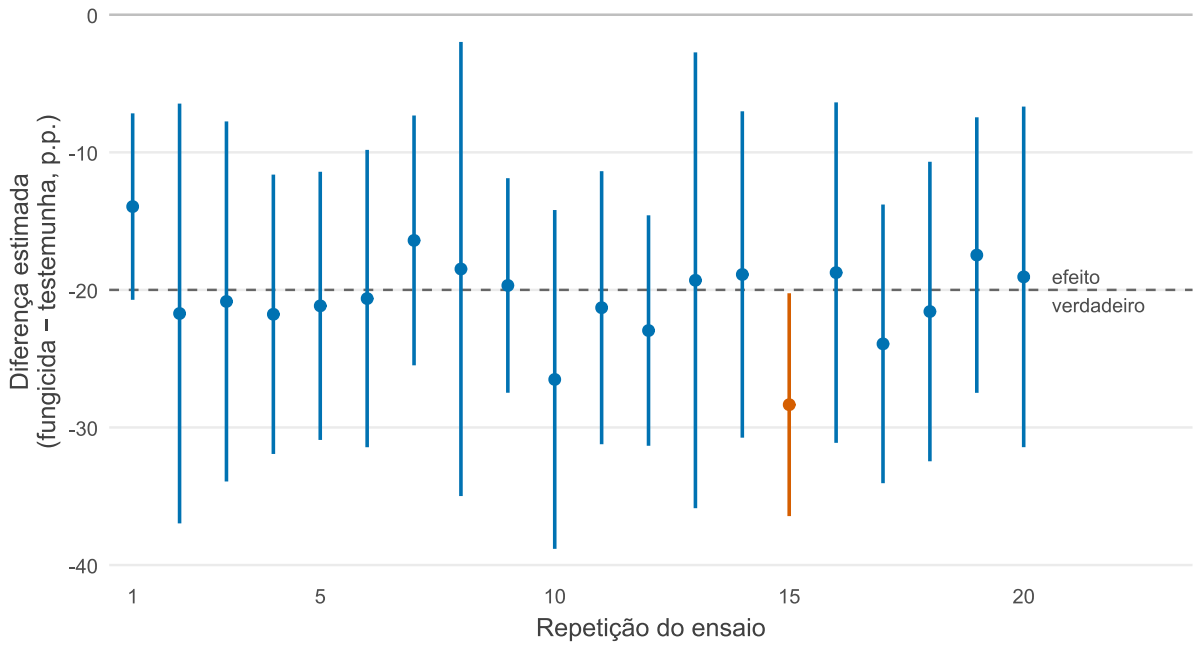


Figura 6.2. Amostras repetidas produzem estimativas diferentes da mesma grandeza subjacente. Vinte repetições simuladas do mesmo ensaio (seis parcelas por tratamento), em que o efeito verdadeiro do fungicida é uma redução de 20 pontos percentuais (linha tracejada). Cada ponto é a diferença estimada em um ensaio, com seu intervalo de confiança de 95%; os intervalos em laranja não contêm o valor verdadeiro.

6.5.3 Erro padrão

O **erro padrão** (EP) é o desvio padrão da distribuição amostral de uma estimativa. Em outras palavras, mede o quanto a estimativa variaria tipicamente de uma repetição do estudo para outra. É a medida básica da incerteza de uma estimativa.

Para uma média calculada a partir de n observações independentes com desvio padrão s, o erro padrão é aproximadamente:

$$EP = \frac{s}{\sqrt{n}}$$

O erro padrão diminui à medida que aumenta o número de observações independentes, mas devagar: para reduzi-lo à metade, são necessárias quatro vezes mais observações. Toda estimativa, inclusive diferenças, inclinações e proporções, tem seu próprio erro padrão.

6.5.4 Desvio padrão versus erro padrão

O desvio padrão e o erro padrão são frequentemente confundidos, em parte porque ambos são usados para desenhar barras de erro (Capítulo 5). O **desvio padrão** descreve a variação entre observações individuais: o quanto as parcelas diferem umas das outras. Ele não diminui com mais observações, porque reflete a variabilidade natural do sistema. O **erro padrão** descreve a incerteza de uma estimativa: com que precisão uma média ou diferença é conhecida. Ele diminui à medida que mais observações independentes são coletadas.

A escolha entre os dois depende da mensagem. Para descrever o quanto parcelas ou plantas são variáveis, relate o desvio padrão. Para descrever com que precisão a média ou o efeito de um tratamento foi estimado, relate o erro padrão ou, de preferência, um intervalo de confiança.

Tabela 6.1. Distinção entre desvio padrão e erro padrão.

Grandeza	O que descreve?
Desvio padrão	Variação entre observações
Erro padrão	Incerteza de uma grandeza estimada

6.6 Estimação e incerteza

6.6.1 Estimativas pontuais

Uma **estimativa pontual** é um único valor calculado a partir dos dados como a melhor estimativa do parâmetro: uma severidade média, uma diferença entre médias de tratamentos, uma redução relativa, uma inclinação, uma proporção de plantas infectadas ou uma razão de riscos. As estimativas pontuais respondem à pergunta “quanto?”, mas, sozinhas, nada dizem sobre a confiabilidade da resposta.

6.6.2 Estimativas intervalares

Uma **estimativa intervalar** fornece uma faixa de valores plausíveis para o parâmetro, junto com a estimativa pontual. Ela comunica tanto a magnitude do efeito quanto a incerteza sobre ele. A estimativa intervalar mais comum é o intervalo de confiança.

6.6.3 Intervalos de confiança

Um intervalo de confiança de 95% é construído por um procedimento que, se o estudo fosse repetido muitas vezes, produziria intervalos contendo o parâmetro verdadeiro em cerca de 95% das repetições. Para um único intervalo calculado a partir dos dados disponíveis, o parâmetro está dentro ou fora; não sabemos qual. Por isso, não é correto dizer que há 95% de probabilidade de o valor verdadeiro estar dentro deste intervalo específico.

Uma interpretação prática é que o intervalo de confiança contém os valores do parâmetro razoavelmente compatíveis com os dados, dado o modelo estatístico. Valores próximos ao centro do intervalo são os mais compatíveis; valores próximos aos limites, menos; valores muito fora do intervalo são pouco compatíveis com os dados. Por exemplo, se um fungicida reduziu a severidade em 20 pontos percentuais, com intervalo de confiança de 95% de 10 a 30 pontos, então reduções entre cerca de 10 e 30 pontos são compatíveis com os dados, enquanto uma redução zero, ou de 40 pontos, não é.

! Importante

Um intervalo de confiança deve ser interpretado junto com o efeito estimado. Ele comunica a faixa de tamanhos de efeito razoavelmente compatíveis com os dados e com o modelo estatístico.

6.6.4 Precisão

A largura de um intervalo de confiança reflete a **precisão** da estimativa. Um intervalo estreito indica uma estimativa precisa; um intervalo largo indica que os dados são compatíveis com uma ampla faixa de efeitos. A precisão aumenta com o número de unidades experimentais independentes e diminui com a variabilidade entre elas. Um bom delineamento aumenta a precisão sem aumentar o número de unidades: o bloqueamento, por exemplo, remove fontes conhecidas de variação da comparação entre tratamentos (Capítulo 2).

Precisão não é o mesmo que exatidão. Uma estimativa pode ser precisa, mas enviesada, como quando as parcelas não tratadas são sistematicamente colocadas na parte mais úmida da lavoura. Os métodos estatísticos quantificam a precisão; somente um bom delineamento protege contra o viés.

6.6.5 Visualizando estimativas e incerteza

Estimativas e seus intervalos de confiança são mostrados de forma eficaz como pontos com linhas horizontais ou verticais representando os intervalos (Capítulo 5). Quando a conclusão principal diz respeito a diferenças entre tratamentos, as próprias diferenças, com seus intervalos, podem ser representadas em relação a uma linha de referência no zero. Esses gráficos tornam fácil ver de uma vez a direção, a magnitude e a precisão de cada efeito.

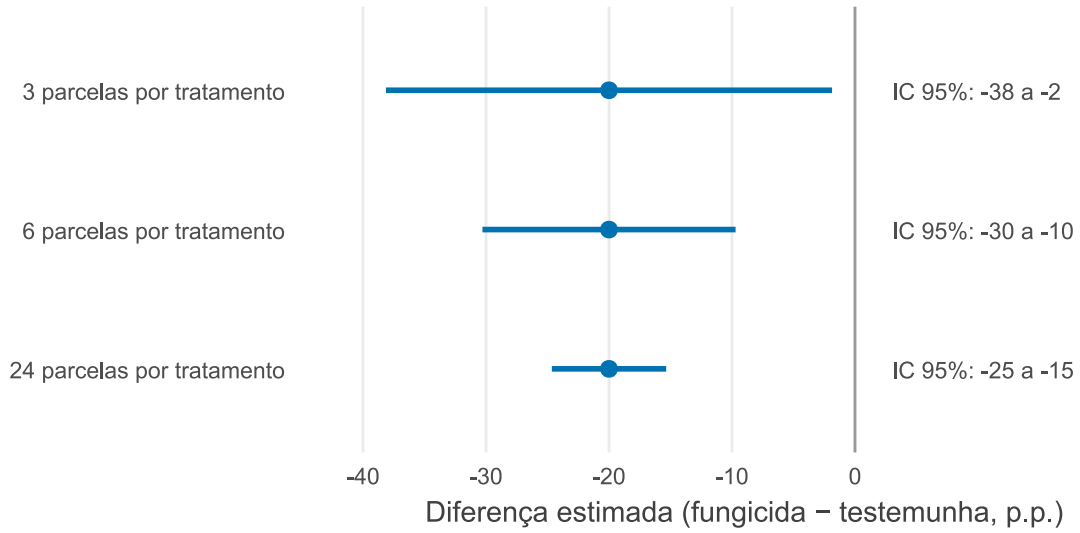


Figura 6.3. O mesmo efeito estimado pode ter diferentes níveis de incerteza. A redução estimada é de 20 pontos percentuais nos três ensaios hipotéticos, mas a precisão aumenta com o número de parcelas por tratamento (desvio padrão de 8 pontos percentuais entre parcelas). Com três parcelas, os dados são compatíveis desde efeitos desprezíveis até reduções muito grandes.

6.7 Tamanho de efeito

6.7.1 Qual é o tamanho do efeito?

Um **tamanho de efeito** é uma medida quantitativa da magnitude de um efeito: o quanto um tratamento reduz a doença, o quanto a produtividade diminui por unidade de severidade, com que intensidade a infecção depende da temperatura. Os tamanhos de efeito respondem à pergunta que mais interessa aos pesquisadores e aos usuários da pesquisa: qual é o tamanho do efeito?

Grande parte da comunicação científica tem se concentrado, em vez disso, em saber se um efeito é "estatisticamente significativo". Essa mudança de ênfase — da existência de um efeito para sua magnitude e precisão — é uma das mensagens centrais deste capítulo.

6.7.2 Efeitos absolutos

Um efeito absoluto é a diferença entre grupos expressa nas unidades da resposta. Se as parcelas não tratadas tiveram 40% de severidade e as tratadas 25%, o efeito absoluto é uma redução de 15 pontos percentuais. Se a produtividade foi de 3200 kg/ha nas parcelas tratadas e de 2800 kg/ha nas não tratadas, o efeito absoluto é de 400 kg/ha. Efeitos absolutos são diretamente interpretáveis e costumam ser o que importa para decisões econômicas.

Para variáveis expressas em porcentagem, a diferença deve ser expressa em **pontos percentuais**, e não como porcentagem, para evitar confusão: uma mudança de 40% para 25% de severidade é uma redução de 15 pontos percentuais.

6.7.3 Efeitos relativos

Um efeito relativo expressa a diferença como proporção de um valor de referência, geralmente a testemunha. A redução de 40% para 25% de severidade corresponde a uma redução relativa de 15/40 = 37,5%. Na Fitopatologia, os efeitos relativos são muito usados como medidas de eficácia de fungicidas ou de controle da doença.

Os efeitos relativos são úteis para comparar efeitos entre situações com diferentes níveis de doença, mas podem ser enganosos quando o valor de referência é pequeno. Uma redução de 2% para 1% de severidade é uma redução relativa de 50%, mas de apenas um ponto percentual em termos absolutos, o que pode não ter nenhuma importância prática. Relatar efeitos absolutos e relativos fornece o quadro mais completo.

6.7.4 Efeitos padronizados

Efeitos padronizados expressam uma diferença em relação à variabilidade das observações, por exemplo a diferença entre médias dividida pelo desvio padrão. Não têm unidades e são usados principalmente para comparar efeitos medidos em escalas diferentes, como em meta-análises ou cálculos de poder. Como dependem da variabilidade de um estudo específico, os efeitos padronizados são menos diretamente interpretáveis em termos biológicos, e efeitos absolutos ou relativos em unidades naturais são geralmente preferidos ao relatar resultados.

6.7.5 Relevância biológica

Se um efeito é importante depende de suas consequências, que só podem ser julgadas com conhecimento biológico e prático. Um fungicida que reduz a severidade em 5 pontos percentuais pode ser irrelevante se a produtividade não for afetada, ou valioso se a cultura tiver alto valor de mercado e a doença causar perdas diretas de qualidade. Um aumento de um dia no período

latente pode ser trivial em um único ciclo, mas substancial ao longo de uma epidemia policíclica. Uma diferença de resistência entre cultivares pode importar para o melhoramento, mesmo que seja pequena demais para importar para o manejo.

Os pesquisadores devem definir, idealmente antes de coletar os dados, o menor efeito que seria biológica ou economicamente relevante. Esse limiar fornece o contexto para interpretar as estimativas e seus intervalos de confiança.

> **! Importante**
>
> A evidência estatística trata da incerteza. A importância biológica diz respeito à magnitude e às consequências de um efeito. São perguntas relacionadas, mas não são a mesma pergunta.

6.8 Hipóteses científicas e hipóteses estatísticas

6.8.1 Hipóteses científicas

Como discutido no Capítulo 2, uma hipótese científica é uma proposição biológica que oferece uma explicação ou expectativa: um fungicida reduz a doença porque inibe a germinação dos esporos; uma cultivar tem menos doença porque possui um gene de resistência; a infecção aumenta com a duração do molhamento porque o patógeno precisa de água livre para germinar e penetrar. As hipóteses científicas são avaliadas por meio das predições que geram.

6.8.2 Hipóteses estatísticas

Uma hipótese estatística é um enunciado preciso sobre o valor de um parâmetro em um modelo estatístico. Ela traduz uma predição da hipótese científica em uma forma que pode ser avaliada com dados. Por exemplo, a predição de que parcelas tratadas terão menos doença que parcelas não tratadas pode ser traduzida em um enunciado sobre a diferença entre as severidades médias dos dois grupos.

6.8.3 Hipótese nula

A **hipótese nula**, H_0, geralmente afirma que não há efeito ou diferença: a severidade média das parcelas tratadas é igual à das não tratadas, ou a inclinação que relaciona doença e produtividade é zero. A hipótese nula não é necessariamente aquilo em que o pesquisador acredita; é um modelo de referência com o qual os dados são comparados.

6.8.4 Hipótese alternativa

A **hipótese alternativa**, H_A, afirma o que se considera caso a hipótese nula não seja sustentada, geralmente que existe uma diferença ou um efeito. Ela pode ser bilateral (as médias diferem, em qualquer direção) ou unilateral (a média tratada é menor).

Para uma comparação entre duas médias populacionais:

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_A : \mu_1 - \mu_2 \neq 0$$

Essa é uma representação estatística de uma comparação científica, e não a própria hipótese científica. Rejeitar H_0 indica que os dados não são compatíveis com uma diferença zero sob o

modelo; não demonstra o mecanismo biológico proposto pela hipótese científica, nem diz qual é o tamanho da diferença.

6.9 Teste de hipóteses

6.9.1 A lógica de um teste estatístico

Um teste estatístico segue uma lógica indireta. Começa supondo que a hipótese nula é verdadeira, junto com os demais pressupostos do modelo estatístico. Em seguida, pergunta quão compatíveis os dados observados são com esse modelo. Se dados como os observados seriam muito incomuns sob o modelo nulo, isso é tomado como evidência contra ele. Se seriam comuns, os dados fornecem pouca evidência contra o modelo nulo — o que não significa que a hipótese nula seja verdadeira.

6.9.2 Estatísticas de teste

Uma estatística de teste é um número calculado a partir dos dados que resume o quanto as observações se afastam do que a hipótese nula prevê, em relação à incerteza da estimativa. Muitas estatísticas de teste têm a forma geral:

$$\text{estatística de teste} = \frac{\text{estimativa} - \text{valor sob } H_0}{\text{erro padrão da estimativa}}$$

A estatística t usada para comparar duas médias, por exemplo, é a diferença observada dividida pelo seu erro padrão. Valores grandes, positivos ou negativos, indicam que a estimativa está longe do valor nulo em relação à sua incerteza.

6.9.3 O valor-P

O **valor-P** é a probabilidade, calculada supondo que o modelo nulo seja verdadeiro, de obter uma estatística de teste pelo menos tão extrema quanto a observada. Valores-P pequenos indicam que os dados observados seriam incomuns se o modelo nulo fosse verdadeiro; valores-P grandes indicam que não seriam incomuns.

O valor-P depende tanto do tamanho do efeito quanto da sua precisão. Um efeito pequeno estimado com muita precisão, a partir de muitas observações, pode produzir um valor-P muito pequeno; um efeito grande estimado com pouca precisão pode produzir um valor-P grande.

> **! Importante**
>
> Um valor-P mede a compatibilidade dos dados observados, ou de dados mais extremos, com um modelo estatístico especificado que inclui a hipótese nula. Ele não é a probabilidade de a hipótese nula ser verdadeira.

6.9.4 O que um valor-P não nos diz

Um valor-P **não** nos diz diretamente:

- a probabilidade de a hipótese ser verdadeira;
- a probabilidade de os resultados terem ocorrido “por acaso”;
- a magnitude do efeito biológico;
- se o efeito é cientificamente importante;
- se o experimento foi bem delineado;
- se os resultados serão reprodutíveis.

Essas limitações foram enfatizadas em uma declaração da Associação Americana de Estatística (ASA) sobre valores-P (Wasserstein e Lazar 2016).

6.9.5 Significância estatística

Na prática convencional, um resultado é chamado "estatisticamente significativo" quando o valor-P é menor que um limiar preestabelecido, chamado nível de significância, α. O limiar mais comum é $\alpha = 0,05$. Essa convenção foi popularizada no início do século XX como um guia aproximado de quais resultados mereciam mais atenção, e não como uma regra para decidir o que é verdadeiro. Com o tempo, porém, o limiar passou a ser tratado como uma fronteira entre resultados que "funcionam" e resultados que "não funcionam".

6.9.6 O problema dos limiares

A evidência não muda bruscamente em um valor arbitrário. Considere dois experimentos:

```
P = 0,049  → "significativo"
P = 0,051  → "não significativo"
```

Os dois resultados fornecem evidências praticamente idênticas, mas a dicotomia leva a conclusões opostas. Tratar os limiares como decisivos incentiva a exagerar resultados logo abaixo de 0,05, descartar resultados logo acima e escolher análises que produzam valores-P abaixo dele. Por isso, muitos estatísticos recomendam relatar os valores-P exatos, interpretá-los como uma medida contínua de compatibilidade com o modelo nulo e concentrar-se nos tamanhos de efeito e em seus intervalos de confiança (Amrhein et al. 2019; Wasserstein e Lazar 2016).

> ⚠ **Aviso**
>
> As conclusões científicas não devem mudar bruscamente porque um valor-P ficou logo acima ou logo abaixo de um limiar arbitrário.

6.10 Significância estatística versus importância biológica

Essa distinção é particularmente importante nas ciências biológicas aplicadas.

A significância estatística diz respeito a se os dados são compatíveis com a ausência de efeito. A importância biológica diz respeito a se o efeito é grande o suficiente para importar. São dimensões independentes: um efeito pode ser estatisticamente detectável, mas biologicamente trivial, ou biologicamente grande, mas estimado com imprecisão demais para conclusões firmes. Considerar juntos a estimativa, seu intervalo de confiança e o menor efeito biologicamente relevante leva a quatro situações típicas.

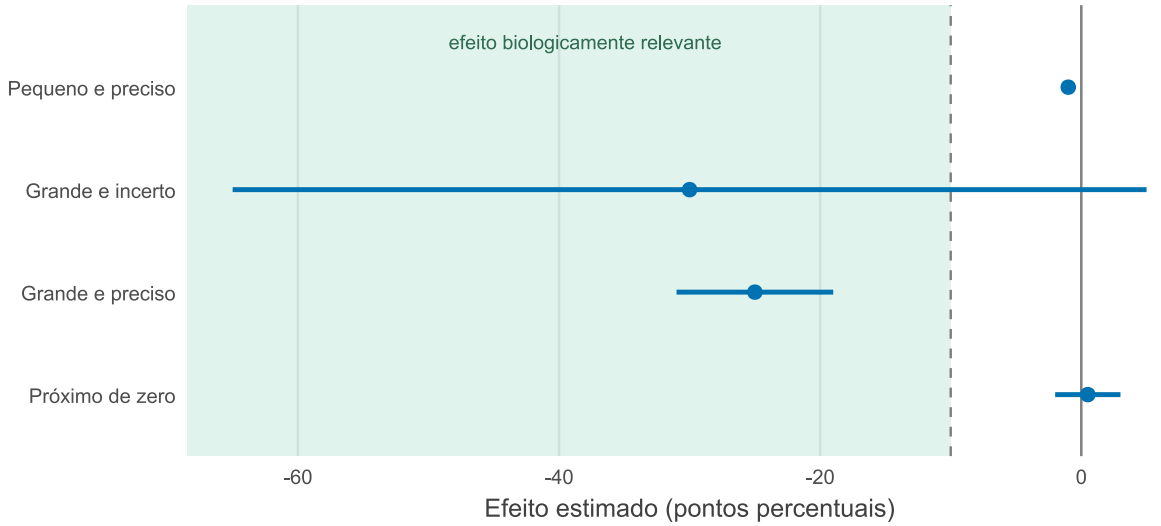


Figura 6.4. A evidência estatística e a importância biológica representam dimensões diferentes da interpretação científica. Quatro situações descritas no texto: cada ponto é um efeito estimado com seu intervalo de confiança de 95%. A faixa sombreada indica reduções maiores que 10 pontos percentuais, consideradas aqui biologicamente relevantes.

6.10.1 Efeitos pequenos, mas estimados com precisão

Um levantamento de milhares de lavouras constata que um sistema de preparo do solo está associado a 1 ponto percentual a menos de doença, com intervalo de confiança de 0,6 a 1,4 ponto e P < 0,001. O efeito é claramente diferente de zero, mas todo o intervalo está abaixo de qualquer efeito relevante para o manejo da doença. A evidência é forte; o efeito é pequeno.

6.10.2 Efeitos grandes, mas incertos

Um pequeno experimento em casa de vegetação constata que um agente de controle biológico reduziu a severidade em 30 pontos percentuais, com intervalo de confiança de −5 a 65 pontos e P = 0,09. O efeito estimado é grande, mas os dados são compatíveis com qualquer coisa, desde nenhum efeito (ou um leve aumento) até uma redução muito grande. A conclusão adequada não é que o agente “não funciona”, mas que a evidência é inconclusiva e que é necessário um experimento maior ou repetido.

6.10.3 Efeitos grandes e estimados com precisão

Um ensaio de campo bem repetido constata que um fungicida reduziu a severidade em 25 pontos percentuais, com intervalo de confiança de 19 a 31 pontos. Todo o intervalo está na faixa de efeitos biologicamente relevantes. Essa é a situação mais forte: o efeito está claramente presente e é claramente importante.

6.10.4 Efeitos próximos de zero

Um ensaio bem repetido encontra uma diferença de 0,5 ponto percentual entre dois fungicidas, com intervalo de confiança de −2 a 3 pontos. Os dados são compatíveis com ausência de diferença, e também com pequenas diferenças em qualquer direção, mas não com nenhuma diferença grande o suficiente para importar. Aqui, um resultado “não significativo” fornece evidência útil de que os dois produtos têm desempenho semelhante, porque o intervalo é estreito. Essa conclusão só é possível porque a estimativa é precisa.

6.11 Erros nas decisões estatísticas

Quando um teste é usado para tomar uma decisão — rejeitar ou não rejeitar a hipótese nula —, dois tipos de erro são possíveis.

6.11.1 Erro do tipo I

O **erro do tipo I** ocorre quando a hipótese nula é rejeitada embora seja verdadeira: concluir que um fungicida reduz a doença quando, na verdade, ele não tem efeito. A probabilidade de um erro do tipo I, quando a hipótese nula é verdadeira, é igual ao nível de significância α. Com $\alpha = 0,05$, cerca de 1 em cada 20 testes de hipóteses nulas verdadeiras produzirá um resultado "significativo".

6.11.2 Erro do tipo II

O **erro do tipo II** ocorre quando a hipótese nula não é rejeitada embora exista um efeito relevante: concluir que não há evidência de efeito do fungicida quando ele, na verdade, reduz a doença. A probabilidade de um erro do tipo II, denotada β, depende do tamanho do efeito verdadeiro, da variabilidade dos dados e do número de unidades experimentais independentes. Experimentos pequenos e com poucas repetições têm alta probabilidade de erros do tipo II.

Tabela 6.2. Resultados possíveis de um teste convencional de significância da hipótese nula.

Realidade	Rejeitar H_0	Não rejeitar H_0
H_0 verdadeira	Erro do tipo I	Decisão correta
H_0 falsa	Decisão correta	Erro do tipo II

6.11.3 Por que "não significativo" não significa "sem efeito"

Um valor-P maior que 0,05 significa que os dados são compatíveis com ausência de efeito. Não significa que os dados mostram que não há efeito. Os mesmos dados podem ser igualmente compatíveis com um efeito substancial. Essa é a diferença entre **ausência de evidência** e **evidência de ausência**.

O intervalo de confiança revela qual situação se aplica. Se for estreito e contiver apenas efeitos biologicamente desprezíveis, os dados fornecem evidência de que qualquer efeito é pequeno. Se for largo e incluir tanto o zero quanto efeitos grandes, o estudo é simplesmente inconclusivo. Concluir que "os tratamentos não diferiram" apenas com base em um valor-P acima de 0,05 é um dos erros mais comuns na literatura científica (Amrhein et al. 2019).

6.12 Poder estatístico

6.12.1 O que é poder?

O **poder estatístico** é a probabilidade de um estudo detectar um efeito de determinado tamanho — isto é, produzir um resultado estatisticamente significativo — se esse efeito realmente existir. É igual a $1 - \beta$. Um poder de 0,80 significa que, se o efeito verdadeiro for tão grande quanto o suposto, 80% das repetições do estudo o detectariam, e 20% não.

O poder é uma propriedade do delineamento do estudo para um tamanho de efeito especificado, e não do resultado. Ele é mais útil no planejamento, para determinar se a repetição planejada é suficiente para detectar efeitos de interesse prático.

6.12.2 O que determina o poder?

Conceitualmente, o poder estatístico depende:

- da magnitude do efeito;
- da variabilidade entre as observações;
- do tamanho da amostra;
- do delineamento estatístico;
- do critério usado para a evidência estatística.

Efeitos maiores, menos variabilidade, mais unidades experimentais independentes, delineamentos que removem fontes conhecidas de variação (como o bloqueamento) e limiares menos rigorosos aumentam o poder.

6.12.3 Tamanho da amostra

No contexto do poder, tamanho da amostra significa o número de **unidades experimentais independentes**, e não o número de medidas. Avaliar 50 folhas em cada uma de quatro parcelas não fornece o poder de 200 parcelas, porque as folhas são subamostras (Capítulo 2). As subamostras melhoram a precisão do valor de cada parcela, mas somente mais parcelas, vasos, placas ou execuções aumentam o número de unidades independentes. Uma amostra maior também não torna um estudo bom por si só: um estudo grande, mas enviesado, produz estimativas precisas da grandeza errada.

> ⚠ **Aviso**
>
> Aumentar o número de subamostras não compensa ter poucas unidades experimentais independentes.

6.12.4 Planejando o tamanho da amostra

Planejar o tamanho da amostra exige três informações: o menor efeito que seria biologicamente relevante, uma estimativa da variabilidade entre unidades experimentais (de estudos anteriores ou experimentos-piloto) e o poder ou a precisão desejados. Com elas, o número de repetições necessário pode ser calculado ou estimado por simulação. Alternativamente, o tamanho da amostra pode ser planejado para obter um intervalo de confiança com a largura desejada.

Esse planejamento frequentemente revela que experimentos típicos, com três ou quatro repetições, só conseguem detectar efeitos grandes. Isso não é motivo para evitar experimentos, mas é motivo para interpretar com cautela resultados não significativos de experimentos pequenos e para repetir os experimentos quando efeitos de tamanho moderado importam.

6.13 Estimação ou teste?

Teste de hipóteses e estimação são complementares, e não mutuamente exclusivos. Um teste responde a uma pergunta restrita: os dados são compatíveis com o modelo nulo? A estimação responde a perguntas mais amplas: qual é o tamanho do efeito, e com que precisão ele é conhecido? Na maioria dos estudos de Fitopatologia, o segundo conjunto de perguntas está mais próximo do que os pesquisadores e os usuários da pesquisa precisam saber. Um produtor que escolhe um fungicida quer saber quanto de doença ele controla, e não apenas se controla alguma doença.

Por isso, a estimação deve geralmente ser a abordagem principal, com os testes fornecendo informação complementar. Um intervalo de confiança contém a informação de um teste — já que um intervalo que exclui o zero corresponde a um resultado significativo no nível correspondente — e muito mais.

6.13.1 Raciocínio com a estimação em primeiro lugar

Uma sequência útil é:

```
Qual é o efeito estimado?
  ↓
Quão incerta é a estimativa?
  ↓
Que tamanhos de efeito são compatíveis com os dados?
  ↓
A magnitude é biologicamente relevante?
  ↓
O que o teste estatístico acrescenta a essa interpretação?
```

> **Dica**
>
> Sempre que possível, relate o efeito estimado e sua incerteza, em vez de apresentar apenas um valor-P.

6.14 Comparações múltiplas e testes múltiplos

6.14.1 Por que testes múltiplos importam

Cada teste conduzido com $\alpha = 0,05$ tem 5% de probabilidade de produzir um resultado significativo quando não existe efeito. Quando muitos testes são realizados, os falsos positivos se acumulam. Um experimento comparando dez programas de fungicida envolve 45 comparações dois a dois; se nenhum programa diferisse de nenhum outro, ainda assim se esperaria que cerca de duas dessas comparações parecessem significativas. Estudos que medem muitas respostas (diversas variáveis de doença, componentes da produção ou milhares de genes) enfrentam o mesmo problema em escala maior.

6.14.2 Comparações planejadas versus exploratórias

Comparações planejadas antes de os dados serem vistos, porque correspondem diretamente às questões que motivaram o estudo — como cada fungicida versus a testemunha —, têm um status diferente das comparações escolhidas depois de olhar os dados. As primeiras são testes de hipóteses predefinidas; as segundas são exploratórias e devem ser interpretadas como geradoras de hipóteses, e não como confirmação (Capítulo 4). Um pequeno número de comparações planejadas é preferível a todas as comparações possíveis.

6.14.3 Ajuste para multiplicidade

Vários procedimentos ajustam os testes ou os intervalos de confiança para controlar a taxa global de falsos positivos quando muitas comparações são feitas. Exemplos comuns incluem o procedimento de Tukey para todas as comparações dois a dois entre médias, o procedimento de Dunnett para comparações de vários tratamentos com uma testemunha, a correção de Bonferroni para um conjunto de testes e procedimentos que controlam a taxa de falsas descobertas (FDR),

muito usados em estudos genômicos. Esses métodos são aplicados em capítulos posteriores. Aqui, o ponto essencial é que o número de testes afeta a forma como os resultados devem ser interpretados.

6.14.4 Evitando a busca por significância

A análise de dados envolve muitas escolhas: quais observações incluir, qual resposta analisar, que transformação aplicar, que covariáveis incluir, que modelo usar. Quando essas escolhas são feitas depois de ver quais opções produzem resultados significativos, os valores-P relatados perdem o sentido, mesmo que nenhuma escolha isolada pareça irrazoável. Essa prática é chamada *P-hacking*, e a flexibilidade que a permite é às vezes chamada de graus de liberdade do pesquisador (Simmons et al. 2011).

Os remédios são simples: definir as questões, respostas e análises principais antes de coletar os dados; relatar todas as análises realizadas, e não apenas as que produziram resultados significativos; e distinguir claramente as análises confirmatórias das exploratórias.

6.15 Os pressupostos importam

6.15.1 Todo método estatístico tem pressupostos

Todo método estatístico se baseia em um modelo de como os dados foram gerados, e todo modelo tem pressupostos. Pressupostos comuns incluem a independência das observações, uma distribuição específica da resposta ou dos resíduos (como a distribuição normal), variância constante entre grupos ou ao longo da faixa de um preditor e uma forma específica de relação (como a linearidade). Erros padrão, intervalos de confiança e valores-P só são válidos na medida em que esses pressupostos sejam razoáveis.

6.15.2 A independência começa no delineamento

De todos os pressupostos, a independência é o mais importante e o menos passível de correção posterior. As observações são independentes quando conhecer o valor de uma nada diz sobre o valor de outra, além do que o modelo já considera. Folhas da mesma planta, plantas do mesmo vaso e avaliações repetidas da mesma parcela não são independentes. Tratá-las como independentes infla o tamanho aparente da amostra, produz erros padrão pequenos demais e valores-P pequenos demais. A independência é determinada pelo delineamento do estudo e pela unidade experimental (Capítulo 2), e a análise deve respeitá-los.

6.15.3 Os pressupostos do modelo devem ser verificados

Outros pressupostos, como a distribuição dos resíduos e a constância da variância, podem e devem ser verificados com gráficos e diagnósticos depois do ajuste do modelo. Dados de doença frequentemente os violam: contagens são assimétricas, proporções são limitadas entre 0 e 1 e a variabilidade muitas vezes aumenta com a média. Quando os pressupostos são violados, as alternativas incluem transformar a resposta ou, de preferência, usar modelos próprios para o tipo de dado, como os modelos lineares generalizados para contagens e proporções. Essas ferramentas são apresentadas nos Capítulos 7 a 10.

> **! Importante**
>
> Os pressupostos estatísticos descrevem como os dados são representados por um modelo. Devem ser avaliados, e não aceitos automaticamente.

6.16 Associação, predição e causalidade

6.16.1 Associação

Duas variáveis estão associadas quando seus valores tendem a variar juntos: lavouras com mais chuva têm mais doença, plantas com mais lesões têm menores taxas fotossintéticas. As associações podem ser descritas por coeficientes de correlação, inclinações de regressão ou diferenças entre grupos. Uma associação, por mais forte que seja, não indica por si só por que as variáveis variam juntas.

6.16.2 Predição

A predição usa valores conhecidos de algumas variáveis para antecipar os valores de outra: prever a doença a partir do clima, estimar a perda de produtividade a partir da severidade, prever o risco de uma epidemia a partir de observações no início da safra. Um bom modelo preditivo precisa de uma associação estável, não necessariamente causal. As predições só são confiáveis dentro da faixa de condições representada nos dados usados para construir o modelo.

6.16.3 Efeitos causais

Um efeito causal é a mudança em uma resposta que resultaria de intervir em uma variável: o que aconteceria com a doença se uma lavoura fosse pulverizada em vez de não ser. As questões causais são o cerne da maioria dos experimentos em Fitopatologia.

Associações podem surgir sem causalidade. Lavouras manejadas por produtores mais experientes podem ter menos doença e maior produtividade, de modo que doença e produtividade estão associadas em parte porque ambas dependem do manejo. Variáveis que influenciam tanto a suposta causa quanto a resposta são chamadas **variáveis de confundimento**. Em um experimento casualizado, a alocação aleatória dos tratamentos torna improvável que as variáveis de confundimento difiram sistematicamente entre os grupos, e é por isso que os experimentos sustentam conclusões causais, sob os pressupostos do delineamento. Dados observacionais também podem contribuir para a compreensão causal, mas exigem muito mais cuidado: conhecimento das possíveis variáveis de confundimento, sua medida e uma análise adequada.

6.16.4 O que o delineamento do estudo permite concluir?

O tipo de conclusão que um estudo pode sustentar depende principalmente do seu delineamento, e não do método estatístico. O mesmo modelo de regressão aplicado a dados de um experimento casualizado e a dados de um levantamento estima um efeito causal no primeiro caso e uma associação no segundo. As conclusões devem corresponder ao delineamento: experimentos justificam afirmações sobre efeitos nas condições estudadas; estudos observacionais justificam afirmações sobre associações e padrões e só sustentam afirmações causais com evidências e pressupostos adicionais (Capítulo 2).

Tabela 6.3. Três tipos de conclusão e os delineamentos que as sustentam.

Tipo de conclusão	Pergunta respondida	Delineamento típico
Associação	As variáveis variam juntas?	Observacional ou experimental
Predição	Uma variável permite antecipar outra?	Qualquer um, se a associação for estável
Efeito causal	O que acontece se intervirmos?	Experimento casualizado, ou observacional com pressupostos fortes

6.17 Pesquisa na prática: interpretando o efeito de um tratamento

Considere um experimento de campo como os ensaios de fungicidas apresentados no Capítulo 1, com uma testemunha e um programa de fungicida em blocos casualizados com seis blocos. Os valores a seguir são hipotéticos, mas típicos:

Tabela 6.4. Resultados hipotéticos de um ensaio de fungicida.

Grandeza	Valor
Severidade média, testemunha	60%
Severidade média, fungicida	40%
Diferença estimada (fungicida − testemunha)	−20 pontos percentuais
Intervalo de confiança de 95% para a diferença	−30 a −10 pontos percentuais
Redução relativa	33%
Valor-P	0,002

Cada grandeza tem um significado específico. As **médias** descrevem a severidade média observada em cada tratamento. A **diferença estimada** é a melhor estimativa do efeito do fungicida: uma redução de 20 pontos percentuais. O **intervalo de confiança** indica que reduções entre cerca de 10 e 30 pontos são compatíveis com os dados; os dados não são compatíveis com ausência de efeito, nem com reduções tão grandes quanto 40 pontos. A **redução relativa** expressa o efeito como um terço da doença nas parcelas não tratadas. O **valor-P** indica que uma diferença desse tamanho seria muito incomum se o fungicida não tivesse efeito.

Se o efeito é importante depende do que se sabe sobre o sistema. Se as perdas de produtividade começam com severidades acima de 30%, os dois tratamentos continuam acima desse nível, e uma redução de 20 pontos pode não ser suficiente para proteger a produtividade, mesmo sendo estatisticamente clara. Se a meta for manter a doença abaixo de 50%, o fungicida pode ser suficiente. Os resultados estatísticos são os mesmos nos dois casos; a conclusão biológica é diferente.

Pesquisa revisitada: O tratamento reduz a doença?

Considere um experimento que compara a severidade da doença entre uma testemunha e um tratamento de manejo da doença.

Suponha que a análise forneça:

- a severidade média de cada tratamento;
- a diferença estimada entre os tratamentos;
- um intervalo de confiança para essa diferença;

- um valor-P.

Faça quatro perguntas separadas:

1. **Direção:** a doença foi maior ou menor com o tratamento?
2. **Magnitude:** qual foi o tamanho do efeito estimado do tratamento?
3. **Precisão:** quão incerta foi essa estimativa?
4. **Relevância biológica:** a redução estimada é grande o suficiente para importar?

Somente depois de responder a essas perguntas o valor-P deve ser considerado, como informação adicional sobre a compatibilidade estatística com o modelo nulo.

6.18 Dos resultados estatísticos às conclusões científicas

Uma análise estatística produz estimativas, intervalos e valores-P. Uma conclusão científica é algo mais: uma afirmação sobre o sistema biológico, sustentada por evidências e qualificada por suas limitações. Construir uma conclusão exige reunir todos os elementos discutidos neste livro até aqui.

Uma conclusão científica deve integrar:

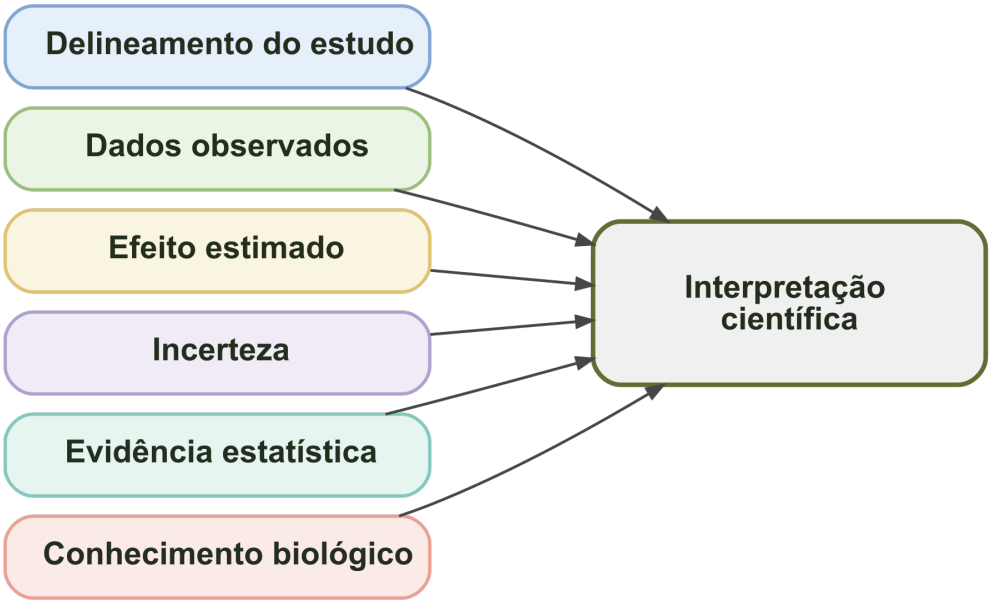


Figura 6.5. Componentes que contribuem para uma conclusão científica.

O delineamento determina que tipo de conclusão é possível e a que população ela se aplica. Os dados observados, explorados e visualizados, mostram o que aconteceu. O efeito estimado e sua incerteza quantificam o que foi aprendido. A evidência estatística indica quão compatíveis os dados são com modelos alternativos. O conhecimento biológico determina se o efeito importa e como ele se encaixa no que já se sabe. Uma conclusão que omite qualquer desses elementos está incompleta.

Boas conclusões também são adequadamente modestas. Um único experimento raramente estabelece um fato geral. Informar claramente as condições em que os resultados foram obtidos, a precisão das estimativas e o que permanece incerto ajuda o leitor a julgar até onde as conclusões podem ser estendidas. ## Erros comuns na inferência estatística

6.18.1 Igualar P < 0,05 a importância científica

Um valor-P pequeno indica que os dados não são compatíveis com o modelo nulo; não indica que o efeito seja grande ou importante. Com observações suficientes, efeitos triviais tornam-se estatisticamente significativos.

6.18.2 Interpretar P > 0,05 como prova de ausência de efeito

Um resultado não significativo pode refletir um efeito pequeno, alta variabilidade ou repetição insuficiente. Sem examinar o intervalo de confiança, ele não pode ser interpretado como evidência de que os tratamentos são equivalentes.

6.18.3 Relatar valores-P sem tamanhos de efeito

Uma tabela de valores-P, ou letras indicando diferenças significativas, sem os efeitos estimados e sua incerteza, esconde a informação de que os leitores mais precisam: qual é o tamanho dos efeitos e com que precisão foram estimados.

6.18.4 Ignorar a incerteza

Estimativas pontuais relatadas sem erros padrão ou intervalos de confiança dão uma falsa impressão de precisão. Toda estimativa da qual depende uma conclusão deve vir acompanhada de uma medida de sua incerteza.

6.18.5 Confundir desvio padrão e erro padrão

Usar o erro padrão para descrever a variabilidade entre observações faz os dados parecerem menos variáveis do que são; usar o desvio padrão para descrever a precisão de uma média faz as estimativas parecerem menos precisas. Figuras e tabelas devem informar qual dos dois está sendo mostrado.

6.18.6 Ignorar a unidade experimental

Analisar subamostras, como folhas, lesões ou repetições técnicas, como se fossem unidades experimentais independentes infla o tamanho da amostra e produz conclusões excessivamente confiantes (pseudorrepetição; Capítulo 2).

6.18.7 Tratar associação como causalidade

Associações de dados observacionais, ou de relações entre variáveis medidas dentro de experimentos, não estabelecem por si sós que uma variável causa mudanças em outra.

6.18.8 Generalizar além da população estudada

Resultados de uma cultivar, isolado, local, safra ou conjunto de condições controladas não se aplicam automaticamente a outros. As conclusões devem informar as condições em que foram obtidas.

6.18.9 Escolher as análises depois de procurar significância

Experimentar várias análises, transformações ou subconjuntos de dados e relatar apenas os que produzem resultados significativos invalida os valores-P e leva a achados que não se replicam.

6.18.10 Relatar precisão numérica excessiva

Relatar uma severidade média de 23,4719% sugere uma precisão inexistente quando as avaliações visuais variam em vários pontos percentuais. As estimativas devem ser relatadas com

um número de dígitos coerente com sua incerteza, tipicamente dois ou três algarismos significativos.

> ⚠ **Aviso**
>
> A análise estatística não consegue determinar a importância biológica de um resultado. Esse julgamento exige conhecimento do sistema biológico e da questão de pesquisa.

6.19 Um fluxo de trabalho para o raciocínio estatístico

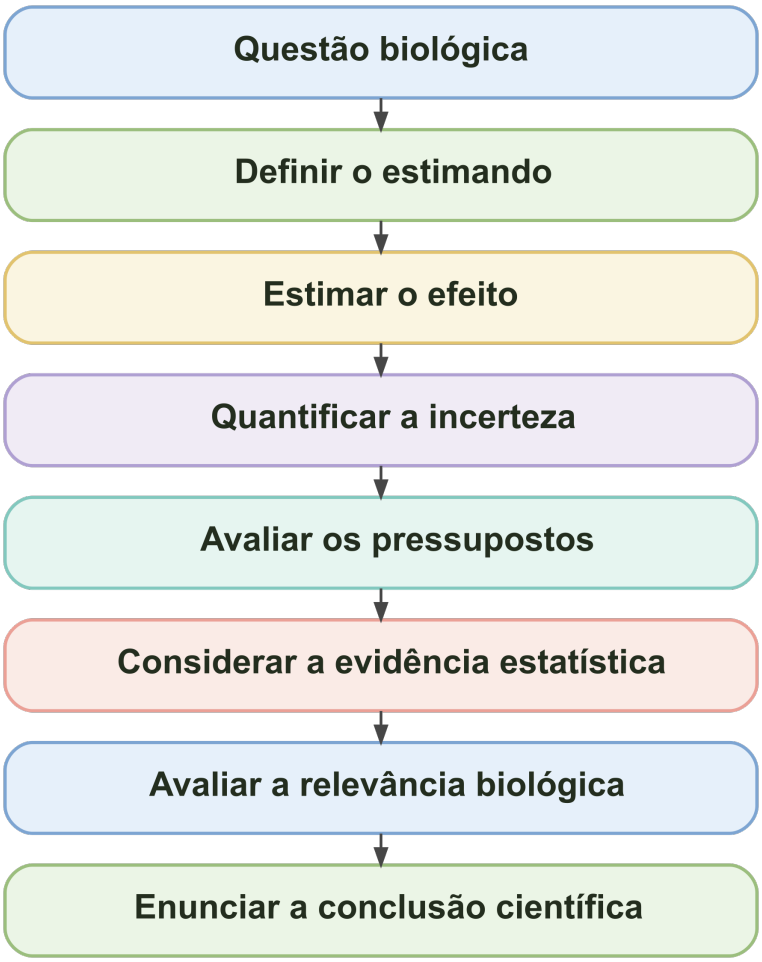


Figura 6.6. Um fluxo de trabalho prático para passar dos dados à inferência científica.

6.20 Lista de verificação da inferência estatística

Antes de interpretar uma análise, pergunte:

1. Que questão biológica está sendo respondida?
2. Que grandeza foi estimada?
3. Qual é a unidade experimental ou observacional?
4. Que população o estudo representa?
5. Qual é o efeito estimado?
6. Quão precisa é essa estimativa?
7. Que pressupostos foram necessários?

8. Esses pressupostos são razoáveis?
9. O que o valor-P acrescenta à interpretação?
10. O efeito é biologicamente relevante?
11. O delineamento sustenta uma conclusão causal?
12. As conclusões são mais amplas do que a evidência sustenta?

6.21 Resumo do capítulo

A inferência estatística fornece um arcabouço para raciocinar a partir dos dados observados reconhecendo explicitamente a incerteza. A interpretação científica deve, portanto, concentrar-se não apenas em saber se um efeito é estatisticamente detectável, mas também em sua direção, magnitude, precisão e relevância biológica. Intervalos de confiança e tamanhos de efeito fornecem informações essenciais que não podem ser substituídas apenas por valores-P. Sobretudo, a força e o escopo de qualquer conclusão estatística permanecem limitados pelo delineamento do estudo e pelo processo que gerou os dados.

6.22 Conceitos-chave

Ao concluir este capítulo, você deverá ser capaz de explicar:

- população;
- amostra;
- parâmetro;
- estimativa;
- estimando;
- variação amostral;
- distribuição amostral;
- desvio padrão;
- erro padrão;
- estimativa pontual;
- intervalo de confiança;
- precisão;
- tamanho de efeito;
- hipótese científica;
- hipótese estatística;
- hipótese nula;
- hipótese alternativa;
- valor-P;
- significância estatística;
- relevância biológica;
- erro do tipo I;
- erro do tipo II;
- poder estatístico;
- comparações múltiplas;
- associação;
- predição;
- causalidade.

6.23 Exercícios

6.23.1 Exercício 1. O que está sendo estimado?

Para cada estudo fornecido pelo professor:

1. identifique a questão biológica;
2. identifique a população de interesse;
3. identifique a unidade experimental ou observacional;
4. enuncie a grandeza que deve ser estimada;
5. descreva como essa grandeza responde à questão biológica.

6.23.2 Exercício 2. Tamanho de efeito ou valor-P?

Considere vários resultados experimentais hipotéticos com diferentes combinações de:

- magnitude do efeito;
- largura do intervalo de confiança;
- valor-P.

Ordene os resultados segundo a força da evidência biológica e justifique seu raciocínio.

6.23.3 Exercício 3. Interprete um intervalo de confiança

Para cada efeito estimado de tratamento e seu intervalo de confiança:

1. descreva a direção do efeito;
2. descreva sua magnitude;
3. avalie sua precisão;
4. identifique os tamanhos de efeito compatíveis com os dados;
5. discuta se o efeito poderia ser biologicamente relevante.

6.23.4 Exercício 4. O que há de errado com esta conclusão?

Avalie criticamente afirmações como:

> "Não houve diferença entre os tratamentos porque P = 0,08."

> "O tratamento A foi altamente eficaz porque P < 0,001."

> "As variáveis estavam fortemente associadas; portanto, uma causa a outra."

> "O experimento tinha 100 observações e, portanto, repetição adequada."

Reescreva cada conclusão de forma mais adequada.

6.23.5 Exercício 5. Significância estatística versus biológica

Considere um experimento de manejo em que um tratamento reduz a severidade média da doença de 50% para 48%, com um intervalo de confiança muito estreito e P < 0,001.

Discuta:

1. Há evidência estatística de diferença?

2. Qual é o efeito estimado?
3. O efeito é necessariamente importante do ponto de vista biológico?
4. Que informação adicional seria útil?

6.23.6 Exercício 6. Lendo um artigo publicado

Escolha um artigo recente de Fitopatologia e identifique:

1. a questão biológica principal;
2. o principal efeito estimado;
3. como a incerteza foi relatada;
4. como os valores-P foram usados;
5. se os tamanhos de efeito foram enfatizados;
6. se as conclusões são sustentadas pelo delineamento do estudo e pela evidência estatística.

6.24 Reflexão

> **Significância estatística não é uma conclusão científica.**

Discuta essa afirmação usando um exemplo da Fitopatologia.

6.25 Próximos passos

Os princípios apresentados neste capítulo se aplicam qualquer que seja o método estatístico utilizado. O modelo analítico adequado, porém, depende da questão biológica, do delineamento do estudo e da natureza da variável resposta. Os capítulos seguintes aplicam esses princípios aos principais tipos de dados encontrados na pesquisa em Fitopatologia.

CAPÍTULO 7

Como comparamos grupos?

7.1 Objetivos de aprendizagem

! Importante

Ao concluir este capítulo, o leitor deverá ser capaz de:

- identificar questões de pesquisa que envolvem comparações entre grupos ou tratamentos;
- distinguir comparações independentes de comparações pareadas;
- estimar e interpretar diferenças entre médias de grupos;
- usar intervalos de confiança para quantificar a incerteza dos efeitos dos tratamentos;
- explicar a relação entre o teste *t*, a análise de variância e os modelos lineares;
- analisar experimentos com dois ou mais grupos;
- interpretar experimentos fatoriais e interações;
- distinguir comparações planejadas de comparações múltiplas *a posteriori*;
- avaliar os pressupostos dos modelos lineares;
- reconhecer quando transformações ou modelos alternativos podem ser necessários;
- comunicar comparações entre grupos usando estimativas, incerteza e gráficos adequados.

7.2 Por que comparar grupos?

Muitos experimentos em Fitopatologia são construídos em torno de comparações.

Questões típicas incluem:

- Um fungicida reduz a severidade da doença em relação a uma testemunha sem tratamento?
- As cultivares diferem em resistência?
- A inoculação altera uma resposta fisiológica?
- Diferentes tratamentos nutricionais afetam o desenvolvimento das lesões?
- A eficácia do tratamento depende da cultivar?

Embora essas questões difiram biologicamente, elas compartilham uma estrutura estatística comum: uma resposta quantitativa é comparada entre grupos definidos por um ou mais fatores experimentais.

! Importante

O objetivo principal de uma comparação entre grupos não é determinar se os grupos são "significativamente diferentes". É estimar **o quanto eles diferem e quão incerta é essa diferença**.

7.3 Da questão biológica à comparação estatística

Considere um experimento simples de manejo da doença.

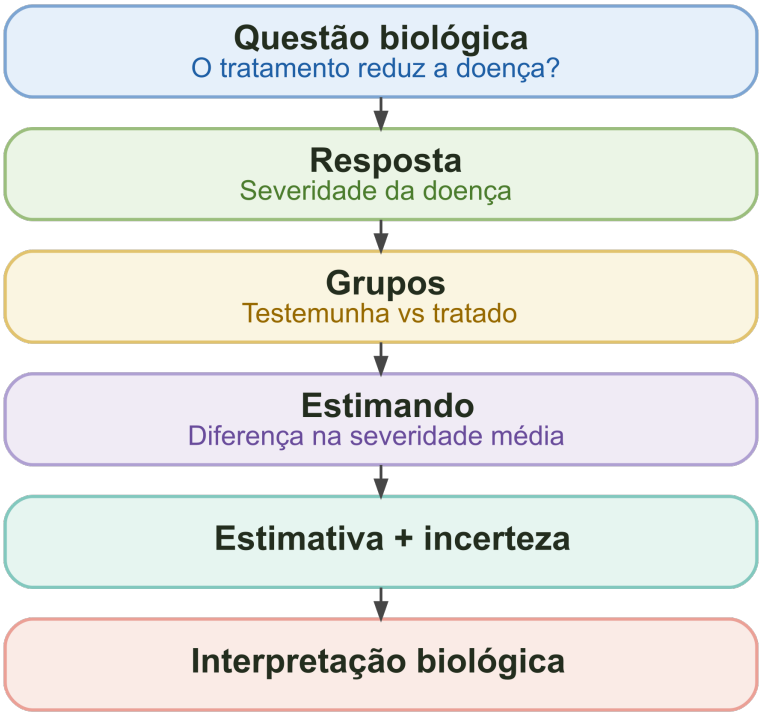


Figura 7.1. Da questão biológica à estimação de uma diferença entre grupos.

7.3.1 Qual é o estimando?

Suponha que a severidade média da doença seja μ_C na testemunha e μ_T no tratamento.

Um estimando natural é a diferença:

$$\Delta = \mu_T - \mu_C$$

A estimativa correspondente obtida no experimento é:

$$\hat{\Delta} = \bar{y}_T - \bar{y}_C$$

A direção da subtração define o sinal do resultado. Com a convenção "tratamento menos testemunha", valores negativos indicam redução da doença, e valores positivos, aumento. A convenção oposta é igualmente válida, desde que declarada. O importante é definir o contraste explicitamente, no texto e nas figuras, para que o leitor não precise adivinhá-lo.

Dica

Sempre defina a direção de uma comparação. "O tratamento diferiu da testemunha em 15%" é ambíguo, a menos que a direção da subtração esteja clara.

7.4 Olhando os dados antes de comparar médias

Antes de ajustar um modelo estatístico, examine as observações. Os princípios da análise exploratória (Capítulo 4) e da visualização (Capítulo 5) aplicam-se diretamente às comparações entre grupos.

7.4.1 Mostre as observações individuais

Com o número de repetições típico dos experimentos em Fitopatologia, cada observação pode ser mostrada. Um gráfico de pontos por tratamento revela imediatamente o número de unidades, a dispersão e os valores incomuns.

7.4.2 Examine as distribuições dos grupos

Verifique se as observações de cada grupo são aproximadamente simétricas ou assimétricas, se há acúmulo de valores em zero ou em 100% e se algum grupo parece conter subgrupos distintos.

7.4.3 Examine a variabilidade entre os grupos

Compare a dispersão dos grupos. Em dados de doença, a variabilidade frequentemente aumenta com a média: a testemunha, com mais doença, costuma ser mais variável que os tratamentos eficazes. Essa característica afeta a escolha do método.

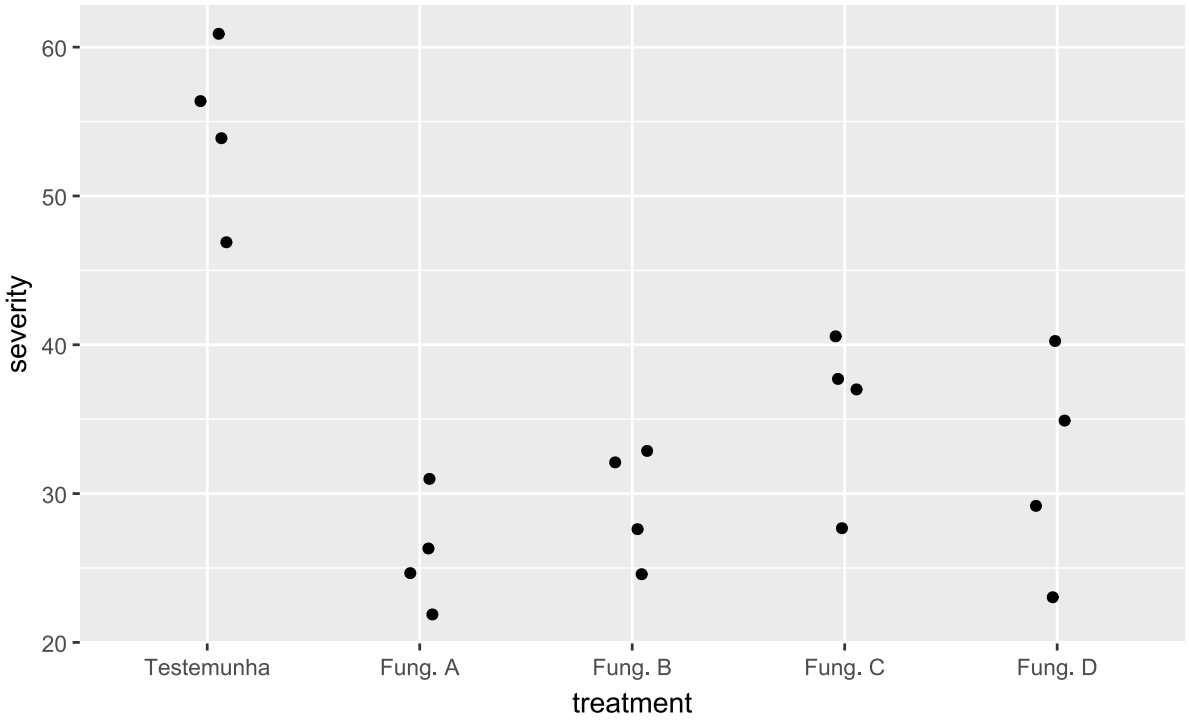
7.4.4 Identifique observações incomuns

Uma parcela com valor muito diferente das demais do mesmo tratamento deve ser conferida com os registros originais antes da análise.

7.4.5 Verifique a estrutura experimental

Confirme o número de unidades experimentais por tratamento, a presença de blocos e a existência de subamostras. A análise deve refletir essa estrutura (Capítulo 2).

```
library(ggplot2)

ggplot(data, aes(x = treatment, y = severity)) +
  geom_point(position = position_jitter(width = 0.1))
```



7.4.6 Por que as médias sozinhas são insuficientes

Dois tratamentos podem ter médias semelhantes, mas diferir muito em:

- variabilidade;
- distribuição;
- tamanho da amostra;
- observações incomuns.

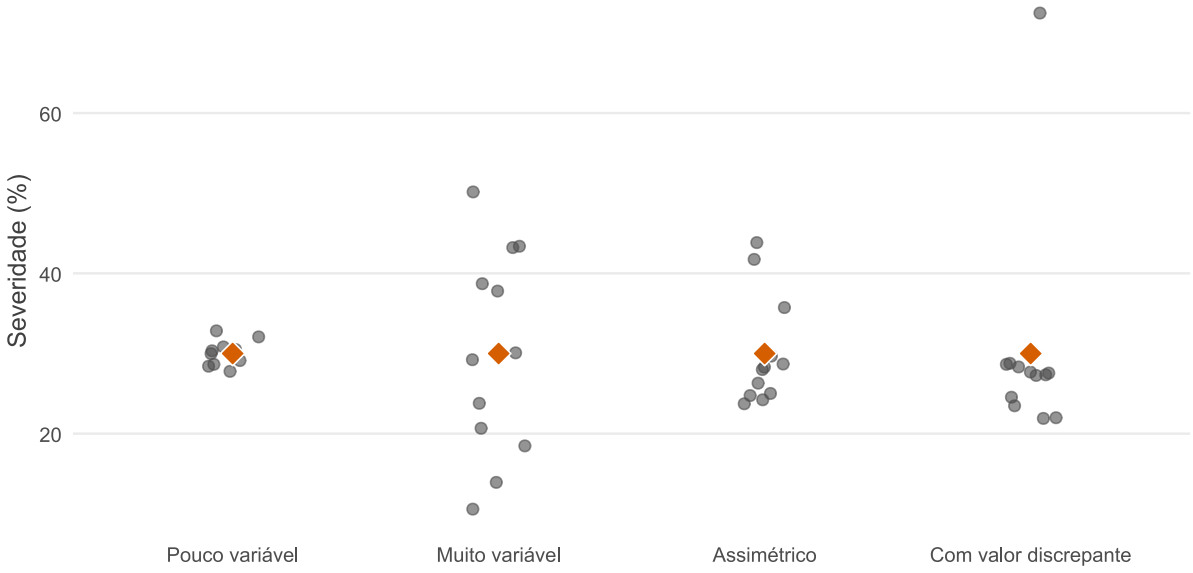


Figura 7.2. As mesmas médias de tratamento podem surgir de distribuições de observações muito diferentes. Os quatro grupos simulados têm média de 30% de severidade (losango), mas diferem na variabilidade, na simetria e na presença de valores discrepantes.

7.5 Comparando dois grupos independentes

7.5.1 Um exemplo simples

Considere um ensaio de campo em que seis parcelas receberam um fungicida e seis parcelas foram mantidas sem tratamento, alocadas ao acaso. A severidade média foi de 38% na testemunha e de 24% nas parcelas tratadas. Os dados simulados usados neste capítulo (objeto `dois`) reproduzem esse exemplo.

7.5.2 Diferença entre médias

A grandeza central é:

$$\hat{\Delta} = \bar{y}_1 - \bar{y}_2$$

No exemplo, com a convenção tratamento menos testemunha, $\hat{\Delta} = 24 - 38 = -14$ pontos percentuais.

7.5.3 Incerteza da diferença

A diferença estimada tem seu próprio erro padrão, que depende da variabilidade dentro dos grupos e do número de unidades em cada um. A partir dele, calcula-se um intervalo de confiança para a diferença (Capítulo 6). No exemplo, o intervalo de confiança de 95% vai de −21,1 a −6,9 pontos percentuais (ver a saída do teste *t* abaixo): reduções entre cerca de 7 e 21 pontos são compatíveis com os dados.

7.5.4 O teste t para amostras independentes

O teste *t* avalia uma comparação entre duas médias populacionais.

Conceitualmente:

$$t = \frac{\text{diferença estimada}}{\text{erro padrão da diferença}}$$

Quanto maior a diferença em relação à sua incerteza, maior o valor absoluto de t e menor o valor-P. Os detalhes do cálculo são feitos pelo programa; o importante é compreender que o teste compara o tamanho da diferença com a precisão com que ela foi estimada.

7.5.5 O teste t de Welch

O teste *t* clássico pressupõe que os dois grupos têm a mesma variância. A versão de Welch não faz esse pressuposto e ajusta o erro padrão e os graus de liberdade quando as variâncias diferem. Como variâncias desiguais são comuns em dados de doença, e como o teste de Welch perde muito pouco quando as variâncias são de fato iguais, ele é geralmente preferível a assumir automaticamente variâncias iguais. No R, `t.test()` usa a versão de Welch por padrão.

```
t.test(severity ~ treatment, data = dois)
```

```
    Welch Two Sample t-test

data:  severity by treatment
t = 4.4094, df = 10, p-value = 0.001316
alternative hypothesis: true difference in means between group Testemunha and group
Fungicida is not equal to 0
```

```
95 percent confidence interval:
  6.933797 21.099536
sample estimates:
mean in group Testemunha  mean in group Fungicida
                38.00000                 23.98333
```

7.5.6 Interpretando o resultado

Não pare em:

> P = 0,001

Em vez disso, relate:

> A severidade média da doença foi 14,0 pontos percentuais menor no grupo tratado do que na testemunha (IC 95%: 6,9 a 21,1 pontos percentuais).

! Importante

A diferença estimada e seu intervalo de confiança geralmente fornecem mais informação biológica do que o valor-P sozinho.

7.6 Observações independentes versus pareadas

Nem todas as comparações entre dois grupos envolvem observações independentes.

7.6.1 Observações independentes

Exemplos:

- parcelas diferentes recebendo tratamentos diferentes;
- plantas diferentes atribuídas independentemente aos tratamentos;
- isolados diferentes atribuídos às condições experimentais.

7.6.2 Observações pareadas

Exemplos:

- medidas antes e depois do tratamento na mesma unidade experimental;
- duas medidas obtidas em unidades experimentais emparelhadas;
- dois tratamentos avaliados em material naturalmente pareado, como as duas metades da mesma folha.

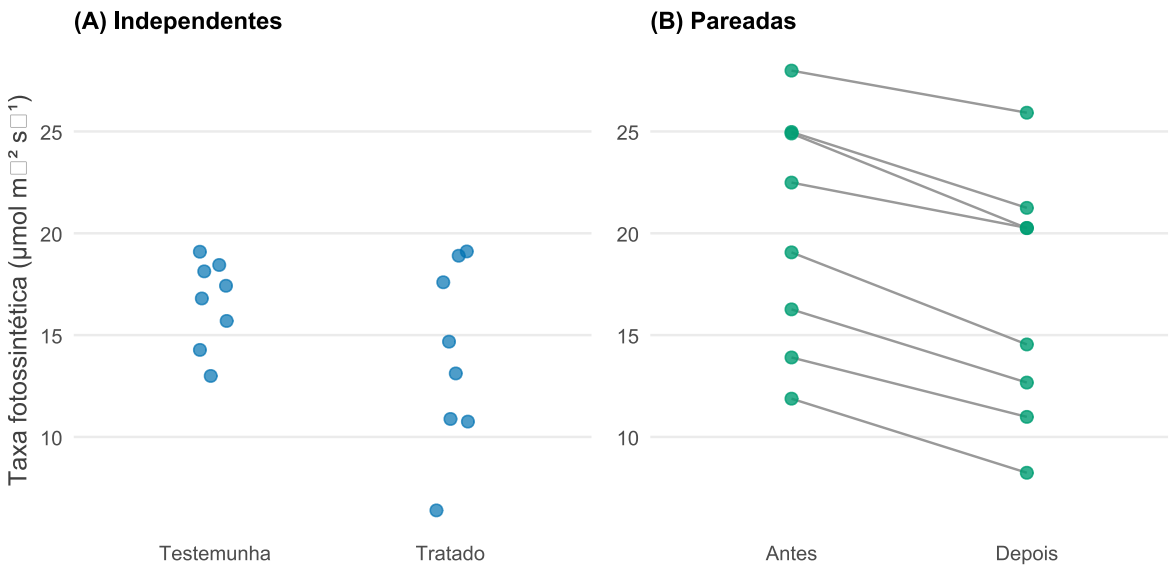


Figura 7.3. Estruturas experimentais independentes e pareadas (dados simulados). (A) Plantas diferentes em cada tratamento: as observações dos dois grupos não têm relação entre si. (B) A mesma planta medida antes e depois do tratamento: cada linha liga as duas medidas de uma unidade, e a análise se baseia nas diferenças dentro de cada par, o que remove a variação entre plantas.

7.6.3 O teste t pareado

Para observações pareadas, a análise se concentra nas diferenças dentro de cada par:

$$d_i = y_{i1} - y_{i2}$$

e pergunta sobre a diferença média, μ_d. Como cada par compartilha as condições de uma mesma unidade, a variação entre unidades é removida da comparação, o que geralmente aumenta a precisão.

```
t.test(before, after, paired = TRUE)
```

```
    Paired t-test

data:  before and after
t = 9.7412, df = 7, p-value = 2.542e-05
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 2.745048 4.504952
sample estimates:
mean difference
          3.625
```

> **⚠ Aviso**
>
> O pareamento é uma propriedade do delineamento do estudo. Ele não deve ser introduzido na análise apenas porque dois grupos têm o mesmo número de observações.

7.7 Comparando mais de dois grupos

Experimentos em Fitopatologia frequentemente têm vários tratamentos.

Exemplos incluem:

- vários fungicidas;
- várias cultivares;
- diferentes concentrações de inóculo;
- várias temperaturas;
- vários tratamentos nutricionais.

7.7.1 Por que não fazer muitos testes t?

Suponha que quatro tratamentos sejam comparados. Há seis comparações dois a dois possíveis. Se cada uma for testada com $\alpha = 0,05$, a probabilidade de que pelo menos uma pareça significativa por acaso, mesmo sem nenhuma diferença real, é bem maior que 5% (Capítulo 6). Além disso, cada teste *t* usa apenas os dados de dois grupos para estimar a variabilidade, desperdiçando a informação dos demais. Um modelo que analisa todos os grupos juntos resolve os dois problemas.

7.7.2 Análise de variância

A análise de variância (ANOVA) fornece um arcabouço para avaliar simultaneamente a variação entre vários grupos.

7.7.3 Variação dentro e entre grupos

Conceitualmente, a ANOVA separa a variação em:

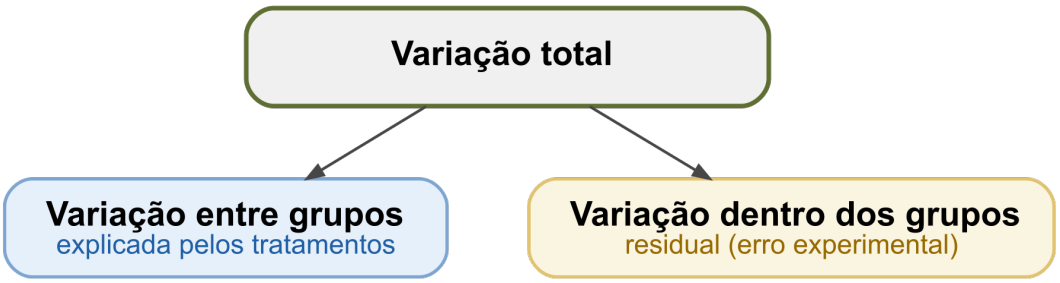


Figura 7.4. Partição conceitual da variabilidade em uma comparação entre grupos.

7.7.4 O teste F

Conceitualmente:

$$F = \frac{\text{variação explicada pelos grupos}}{\text{variação não explicada}}$$

Se os tratamentos não tivessem efeito, as médias dos grupos difeririam apenas pela variação casual, e as duas fontes de variação seriam de magnitude semelhante, com F próximo de 1. Quando as médias diferem mais do que a variação dentro dos grupos faria esperar, F aumenta e o valor-P diminui.

7.8 A ANOVA como modelo linear

Esta é uma transição conceitual importante.

Uma ANOVA com um fator pode ser escrita como:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \varepsilon_i$$

em que as variáveis x indicam a qual grupo pertence cada observação. O mesmo arcabouço de modelo linear usado para comparar grupos será usado depois para estudar relações entre variáveis quantitativas.

> **! Importante**
>
> O teste *t*, a ANOVA e a regressão linear não são procedimentos estatísticos sem relação entre si. São casos particulares do mesmo arcabouço geral do modelo linear.

7.8.1 Ajustando o modelo no R

```
model ← lm(severity ~ treatment, data = data)

anova(model)
```

```
Analysis of Variance Table

Response: severity
          Df  Sum Sq Mean Sq F value    Pr(>F)
treatment  4 2021.37  505.34  16.947 1.948e-05 ***
Residuals 15  447.28   29.82
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

7.8.2 Examinando as estimativas do modelo

```
summary(model)
```

```


Call:
lm(formula = severity ~ treatment, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-8.725 -3.050  0.850  3.100  8.475

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       54.575      2.730  19.989 3.18e-12 ***
treatmentFung. A  -28.550      3.861  -7.394 2.24e-06 ***
treatmentFung. B  -25.250      3.861  -6.539 9.37e-06 ***
treatmentFung. C  -18.800      3.861  -4.869 0.000204 ***
treatmentFung. D  -22.750      3.861  -5.892 2.96e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.461 on 15 degrees of freedom
Multiple R-squared:  0.8188,    Adjusted R-squared:  0.7705
F-statistic: 16.95 on 4 and 15 DF,  p-value: 1.948e-05
```

Na parametrização padrão do R, o intercepto é a média estimada do grupo de referência (o primeiro nível do fator, geralmente em ordem alfabética), e cada coeficiente de tratamento é a diferença entre a média daquele tratamento e a do grupo de referência. Se a testemunha for definida como nível de referência, os coeficientes passam a ser diretamente os efeitos dos tratamentos em relação a ela. Por isso, vale a pena definir o nível de referência de forma deliberada, por exemplo com `relevel()`.

7.9 O que o teste global nos diz?

Para vários tratamentos, a hipótese nula é:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

A alternativa é que pelo menos uma média populacional difere.

> ⚠ **Aviso**
> Uma ANOVA global significativa não nos diz quais tratamentos diferem nem qual é o tamanho dessas diferenças.

7.10 Contrastes: fazendo perguntas específicas

Muitas vezes, a questão biológica é mais específica do que:

> Algum tratamento é diferente?

Exemplos:

- Os fungicidas reduzem a doença em relação à testemunha?
- As cultivares resistentes diferem de um padrão suscetível?
- O aumento da dose produz uma resposta sistemática?
- Uma classe de tratamentos é mais eficaz que outra?

7.10.1 Contrastes planejados

Um contraste representa uma comparação cientificamente definida entre médias de tratamentos. Formalmente, é uma combinação das médias com coeficientes que somam zero. A diferença entre dois tratamentos é o contraste mais simples (coeficientes +1 e −1); a comparação da média de dois fungicidas com a testemunha é outro (coeficientes +½, +½ e −1).

7.10.2 Tratamento versus testemunha

Quando o objetivo é comparar cada tratamento com uma testemunha comum, o número de comparações é igual ao número de tratamentos, e não ao de todos os pares possíveis. O procedimento de Dunnett ajusta esses contrastes para a multiplicidade, mantendo a atenção nas comparações que interessam.

7.10.3 Grupos de tratamentos

Às vezes, a questão envolve classes de tratamentos: fungicidas de sítio específico versus multissítios, cultivares resistentes versus suscetíveis, produtos biológicos versus químicos.

Contrastes que comparam a média de uma classe com a de outra respondem diretamente a essas questões.

> **Dica**
>
> Quando as comparações decorrem diretamente da questão científica, defina-as antes de examinar os resultados, sempre que possível.

7.11 Comparações múltiplas

7.11.1 Por que a multiplicidade importa

Quanto mais comparações são feitas, maior a oportunidade de surgirem diferenças aparentemente incomuns apenas por acaso.

7.11.2 Comparações dois a dois

Comparar todos os pares de tratamentos é útil quando não há uma estrutura prévia entre os tratamentos, como em uma avaliação de muitas cultivares. Com k tratamentos, há $k(k-1)/2$ pares, e o número cresce rapidamente.

7.11.3 Comparações de Tukey

O procedimento de Tukey (diferença honestamente significativa, HSD) ajusta todos os intervalos de confiança e testes dois a dois para que a probabilidade de pelo menos um falso positivo em todo o conjunto de comparações permaneça no nível escolhido.

7.11.4 Comparações com uma testemunha

O procedimento de Dunnett faz o mesmo para o conjunto, menor, de comparações de cada tratamento com a testemunha. Como há menos comparações, ele é mais poderoso do que o de Tukey para essa finalidade.

7.11.5 Métodos de ajuste

Os principais métodos incluem:

- Tukey, para todas as comparações dois a dois;
- Dunnett, para comparações com uma testemunha;
- Bonferroni, simples e geral, porém conservador quando há muitas comparações;
- taxa de falsas descobertas (FDR), que controla a proporção esperada de falsos positivos entre os resultados declarados significativos, útil quando há muitos testes.

Um tratamento matemático detalhado não é necessário aqui.

7.11.6 Apresentação com letras

Artigos de Fitopatologia frequentemente relatam resultados como:

```
Tratamento A    a
Tratamento B    ab
Tratamento C    b
Tratamento D    c
```

7.11.7 O que as letras dizem — e o que não dizem

Nessa apresentação, tratamentos que compartilham uma letra não diferiram significativamente segundo o procedimento usado, e tratamentos sem letras em comum diferiram. As letras resumem um conjunto de decisões dicotômicas. Elas não informam o tamanho das diferenças, sua incerteza nem se as diferenças são biologicamente relevantes. Duas médias com letras diferentes podem diferir por uma quantidade trivial, se estimadas com muita precisão, e duas médias com a mesma letra podem diferir muito, se estimadas com pouca precisão. Além disso, letras são frequentemente lidas como se definissem grupos homogêneos de tratamentos, o que não é verdade: no exemplo acima, A não difere de B, B não difere de C, mas A difere de C.

> ⚠ **Aviso**
>
> A apresentação com letras resume decisões estatísticas, mas esconde a magnitude e a incerteza das diferenças entre tratamentos. Sempre que possível, interprete-a junto com os efeitos estimados e os intervalos de confiança.

7.12 Das letras às estimativas

Esta seção resume a proposta central do capítulo: substituir, ou pelo menos complementar, a apresentação com letras por estimativas dos efeitos com sua incerteza.

Compare:

7.12.1 Apresentação convencional

Os números abaixo vêm do ensaio simulado deste capítulo (testemunha e quatro fungicidas em quatro blocos), analisado mais adiante com o `emmeans`.

```
Tratamento   Média (%)
Fung. A      26,0 a
Fung. B      29,3 ab
Fung. D      31,8 ab
Fung. C      35,8 b
Testemunha   54,6 c
```

com uma figura baseada em estimação, que mostra o efeito de cada tratamento em relação à testemunha e seu intervalo de confiança (ajustado por Dunnett):

```
Efeito em relação à testemunha (pontos percentuais)

Fung. A  ●──────          -28,6  (-34,4 a -22,7)
Fung. B    ●──────        -25,2  (-31,1 a -19,4)
Fung. D      ●──────      -22,8  (-28,6 a -16,9)
Fung. C        ●──────    -18,8  (-24,6 a -13,0)
         |         |        |
        -40       -20       0
```

A figura mostra diretamente que todos os tratamentos reduziram a doença, que as reduções vão de cerca de 19 a 29 pontos percentuais, que a incerteza de cada estimativa é de aproximadamente ±6 pontos e que os intervalos se sobrepõem bastante, de modo que a diferença entre os tratamentos é muito menor, e mais incerta, do que a diferença entre qualquer tratamento e a testemunha.

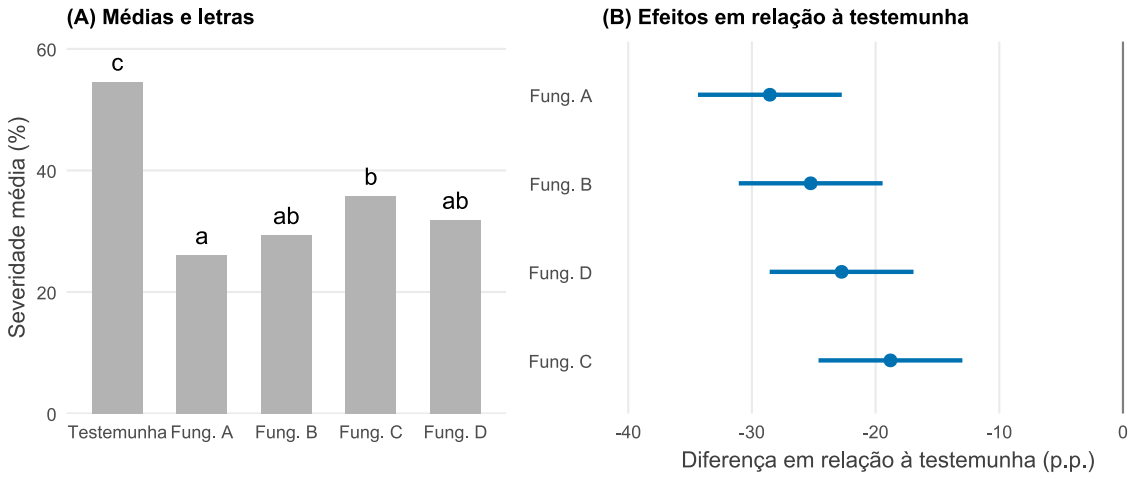


Figura 7.5. Formas alternativas de apresentar as mesmas comparações entre tratamentos (ensaio simulado deste capítulo). (A) Médias com letras (Tukey): informam decisões estatísticas, mas não o tamanho nem a incerteza das diferenças. (B) Efeito de cada tratamento em relação à testemunha, com intervalo de confiança de 95% ajustado por Dunnett: mostra diretamente quanto cada tratamento reduziu a doença e com que precisão.

Discuta:

- Qual tratamento tem o melhor desempenho?
- Por quanto?
- Quão incerta é a diferença?
- A diferença é biologicamente relevante?

7.13 Experimentos fatoriais

Muitos experimentos em Fitopatologia envolvem mais de um fator.

Exemplos:

```
Cultivar × Fungicida
Temperatura × Molhamento foliar
Cultivar × Inoculação
Silício × Patógeno
Fungicida × Época de aplicação
```

7.13.1 Efeitos principais

Um efeito principal descreve o efeito médio de um fator ao longo dos níveis de outro fator.

7.13.2 Interações

Uma interação ocorre quando o efeito de um fator depende do nível de outro.

> **! Importante**
>
> Quando uma interação é biologicamente relevante, a interpretação deve se concentrar nos efeitos condicionais, e não nos efeitos principais médios ao longo do fator com o qual ele interage.

7.13.3 Visualizando interações

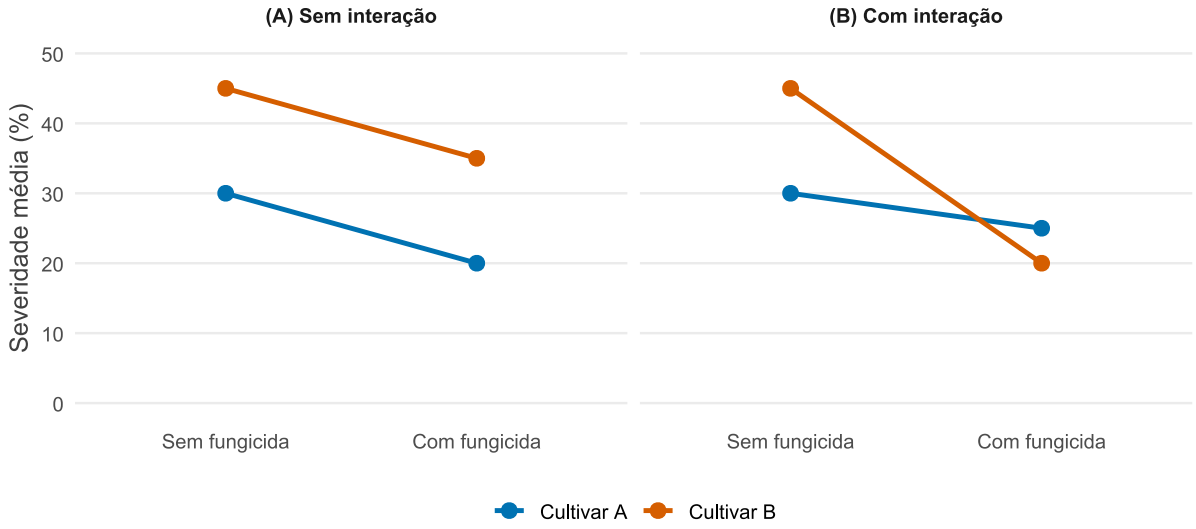
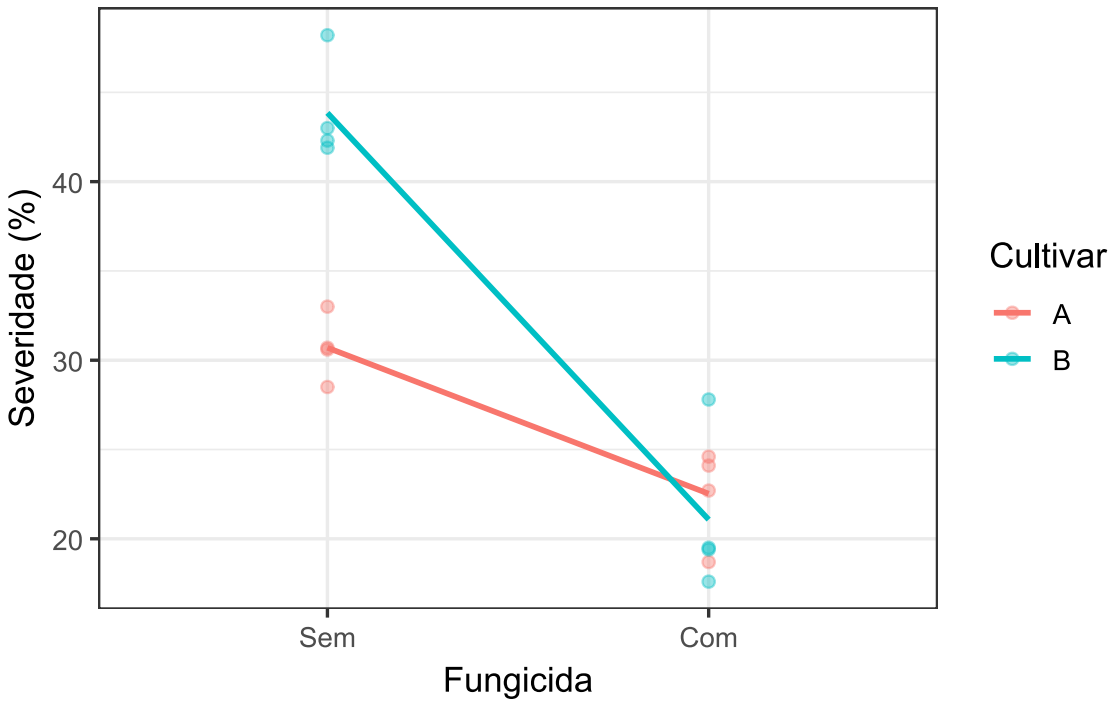


Figura 7.6. Exemplos conceituais de ausência e presença de interação entre dois fatores experimentais. (A) Sem interação: o fungicida reduz a severidade em 10 pontos percentuais nas duas cultivares, e as linhas são paralelas. (B) Com interação: o efeito do fungicida é de 5 pontos na cultivar A e de 25 pontos na cultivar B, e as linhas não são paralelas.

Com os dados simulados do experimento fatorial (duas cultivares, com e sem fungicida, quatro repetições), o gráfico de interação mostra linhas claramente não paralelas:

```
ggplot(fatorial,
       aes(x = fungicide,
           y = severity,
           group = cultivar,
           colour = cultivar)) +
  geom_point(alpha = 0.4) +
  stat_summary(fun = mean, geom = "line", linewidth = 1) +
  labs(x = "Fungicida", y = "Severidade (%)", colour = "Cultivar") +
  theme_bw(base_size = 13)
```



7.13.4 Ajustando um modelo fatorial

```
model_fat ← lm(
  severity ~ cultivar * fungicide,
  data = fatorial
)

anova(model_fat)
```

```
Analysis of Variance Table

Response: severity
                  Df Sum Sq Mean Sq F value    Pr(>F)
cultivar           1 136.89  136.89  13.686 0.0030387 **
fungicide          1 957.90  957.90  95.770 4.521e-07 ***
cultivar:fungicide  1 213.16  213.16  21.312 0.0005938 ***
Residuals          12 120.03   10.00
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A linha `cultivar:fungicide` testa a interação. Os efeitos do fungicida dentro de cada cultivar, que são a informação relevante quando há interação, são obtidos com o `emmeans` (ver adiante):

```
library(emmeans)

emm_fat ← emmeans(model_fat, ~ fungicide | cultivar)
contrast(emm_fat, "revpairwise")
```

```
cultivar = A:
 contrast  estimate   SE df t.ratio p.value
 Com - Sem    -8.18 2.24 12  -3.656  0.0033
```

```
cultivar = B:
 contrast  estimate   SE df t.ratio p.value
 Com - Sem   -22.77 2.24 12 -10.184 <0.0001
```

Nos dados simulados, o fungicida reduziu a severidade em 8,2 pontos percentuais na cultivar A e em 22,8 pontos na cultivar B: o mesmo padrão descrito a seguir.

A notação `*` inclui:

```
cultivar
+
fungicide
+
cultivar × fungicide
```

7.13.5 Interpretando uma interação

Suponha que um fungicida reduza a severidade em 5 pontos percentuais na cultivar A e em 25 pontos percentuais na cultivar B. O efeito principal do fungicida, a média dos dois efeitos, seria uma redução de 15 pontos — um valor que não descreve nenhuma das duas cultivares. A informação biologicamente relevante é que o benefício do fungicida depende da cultivar: é pequeno na cultivar A, provavelmente mais resistente, e grande na cultivar B. Nesse caso, os efeitos do fungicida devem ser relatados separadamente para cada cultivar.

7.14 Blocos e comparações entre grupos

Retome o delineamento em blocos casualizados apresentado no Capítulo 2.

7.14.1 Por que os blocos pertencem ao modelo

Se os tratamentos foram casualizados dentro dos blocos, a análise deve refletir essa estrutura.

Um modelo simples pode ser escrito como:

$$y_{ij} = \mu + \tau_i + b_j + \varepsilon_{ij}$$

em que:

- μ é a média geral;
- τ_i é o efeito do tratamento;
- b_j representa a variação entre blocos;
- ε_{ij} representa a variação residual.

```
model ← lm(
  severity ~ block + treatment,
  data = data
)

anova(model)
```

```
Analysis of Variance Table

Response: severity
```

```
          Df  Sum Sq Mean Sq F value    Pr(>F)
block      3  345.97  115.32  13.659 0.0003555 ***
treatment  4 2021.37  505.34  59.856 7.936e-08 ***
Residuals 12  101.31    8.44
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Incluir o bloco no modelo remove a variação entre blocos do erro experimental, tornando as comparações entre tratamentos mais precisas. Em delineamentos mais complexos, como parcelas subdivididas ou experimentos repetidos em vários locais, os blocos e outras fontes de variação são frequentemente tratados como efeitos aleatórios em modelos mistos, um tema que vai além deste capítulo.

> **ℹ Nota**
>
> O modelo estatístico deve refletir a forma como o experimento foi delineado. Os blocos não são algo acrescentado na análise apenas para melhorar a significância estatística.

7.15 Pressupostos do modelo

A validade da inferência a partir de um modelo linear depende de pressupostos sobre a variação residual.

7.15.1 Independência

A independência vem principalmente do delineamento do estudo: casualização, unidades experimentais bem definidas e ausência de pseudorrepetição.

7.15.2 Linearidade

Nas comparações entre grupos, o modelo deve representar adequadamente a estrutura sistemática dos tratamentos, incluindo blocos e interações relevantes.

7.15.3 Variância constante

A variabilidade residual deve ser razoavelmente semelhante entre os valores ajustados ou entre os grupos.

7.15.4 Distribuição dos resíduos

O pressuposto de normalidade diz respeito aos resíduos do modelo — e não necessariamente à variável resposta bruta.

> **! Importante**
>
> Os pressupostos do modelo aplicam-se principalmente aos erros, ou resíduos, gerados pelo modelo, e não simplesmente à variável resposta observada.

7.16 Diagnóstico do modelo

7.16.1 Resíduos versus valores ajustados

Este gráfico deve mostrar uma nuvem de pontos sem padrão. Um formato de funil, com resíduos mais dispersos para valores ajustados maiores, indica variância que aumenta com a média, muito comum em dados de doença.

7.16.2 Gráficos quantil–quantil

O gráfico quantil–quantil compara os resíduos com o que seria esperado sob uma distribuição normal. Pontos próximos à linha indicam resíduos aproximadamente normais; curvaturas sistemáticas indicam assimetria ou caudas pesadas.

7.16.3 Observações incomuns

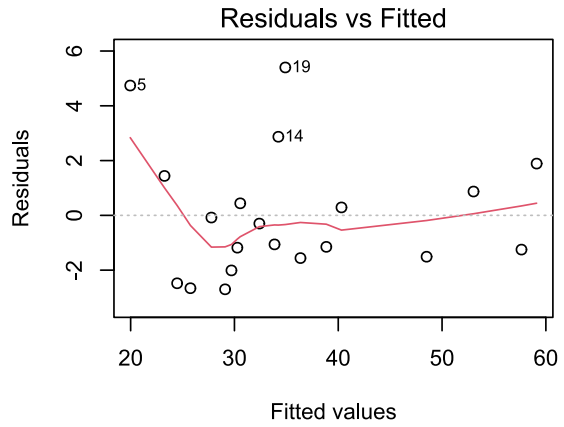
Resíduos muito grandes identificam observações que o modelo descreve mal. Elas devem ser conferidas com os registros originais.

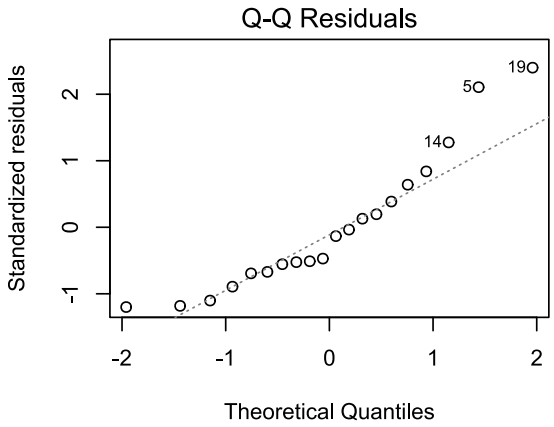
7.16.4 Influência

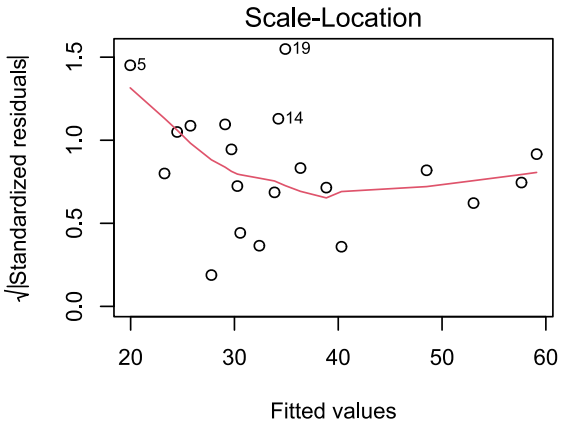
Algumas observações têm influência desproporcional sobre as estimativas. Medidas como a distância de Cook ajudam a identificá-las. Uma observação influente não deve ser removida automaticamente, mas convém verificar se as conclusões mudam sem ela.

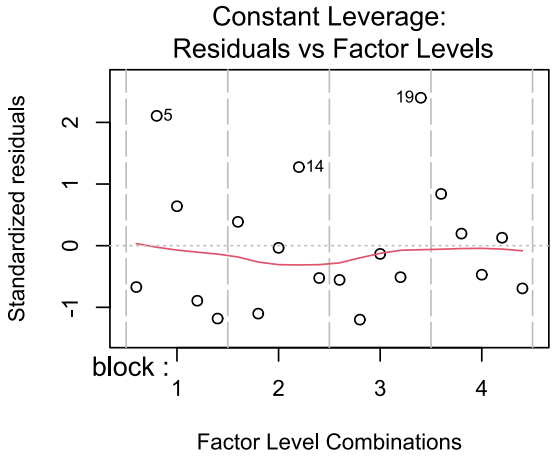
Aplicados ao modelo com blocos ajustado aos dados simulados, os quatro gráficos de diagnóstico do R base são:

```
par(mfrow = c(2, 2))
plot(model)
```


Residuals vs Fitted
Residuals
Fitted values
5
19
14
-2
0
2
4
6
20
30
40
50
60


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
19
5
14
-1
0
1
2
-2
-1
0
1
2


Scale-Location
√|Standardized residuals|
Fitted values
5
19
14
0.0
0.5
1.0
1.5
20
30
40
50
60


Constant Leverage:
Residuals vs Factor Levels
Standardized residuals
Factor Level Combinations
block :
19
5
14
-1
0
1
2
1
2
3
4

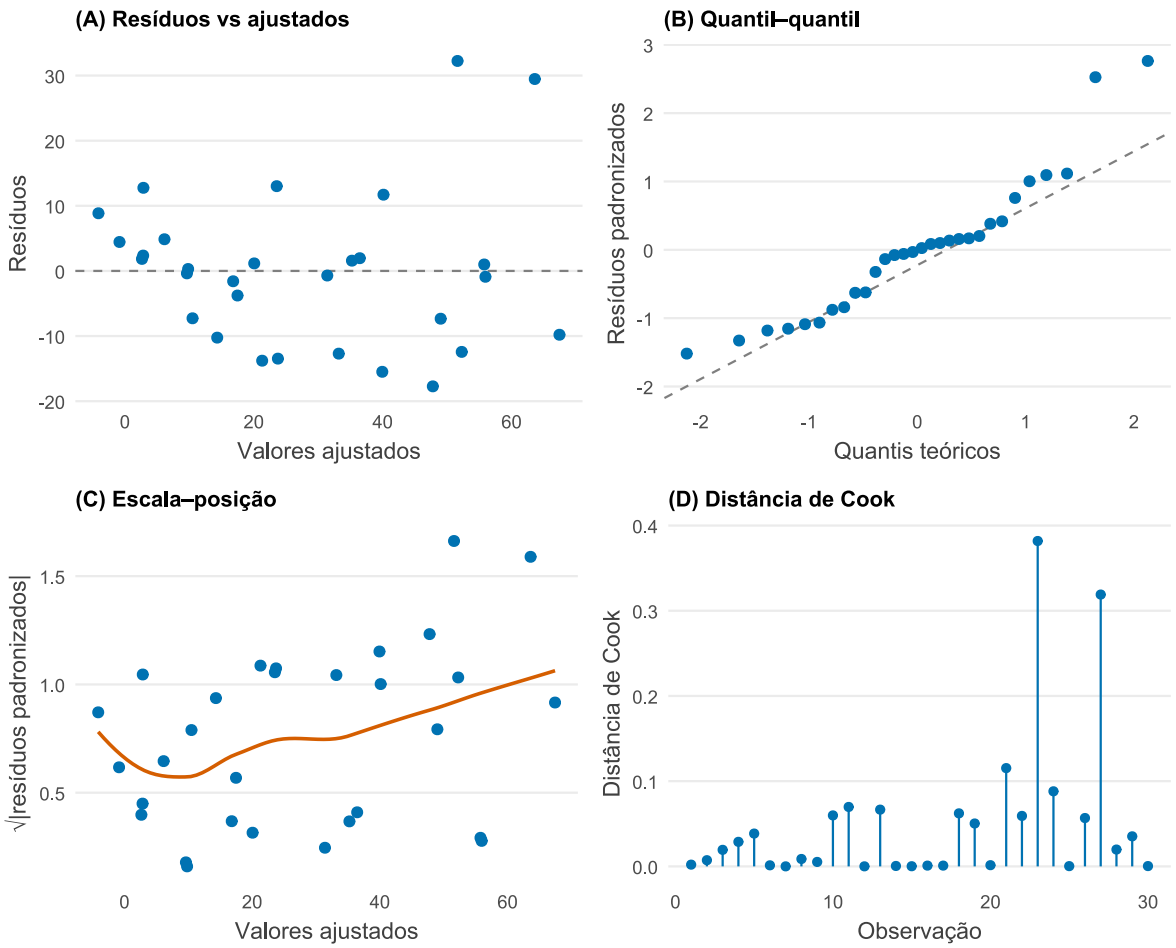


Figura 7.7. Gráficos de diagnóstico comuns para um modelo linear, a partir de um ensaio simulado com cinco tratamentos e seis blocos em que a variabilidade aumenta com a média. (A) Resíduos versus valores ajustados: o formato de funil indica variância não constante. (B) Gráfico quantil–quantil dos resíduos. (C) Escala–posição: confirma o aumento da dispersão com a média. (D) Distância de Cook: identifica observações com influência desproporcional sobre as estimativas.

7.17 E se os pressupostos não forem atendidos?

7.17.1 Primeiro: confira os dados

Causas possíveis incluem:

- erros de digitação;
- unidades incorretas;
- unidades experimentais incomuns;
- fatores do delineamento omitidos;
- definição inadequada da resposta.

7.17.2 Transformações

Transformações comuns incluem $\log(y)$ e $\sqrt{y}$. A transformação logarítmica é útil quando os efeitos são multiplicativos e a variabilidade cresce proporcionalmente à média, como em contagens de esporos ou quantidades de DNA. A raiz quadrada é às vezes usada para contagens. Uma transformação, porém, muda a escala em que os efeitos são estimados e interpretados: uma diferença na escala logarítmica corresponde a uma razão na escala original. A

escolha deve considerar o significado biológico dos efeitos, e não apenas o comportamento dos resíduos.

7.17.3 Os dados em porcentagem devem sempre ser transformados?

Não.

Durante décadas, recomendou-se transformar rotineiramente dados percentuais, por exemplo com a transformação arco-seno da raiz quadrada. Essa recomendação não deve ser aplicada automaticamente. Quando as porcentagens estão longe de 0 e 100% e os resíduos se comportam razoavelmente, a análise na escala original é adequada e tem a vantagem de produzir efeitos diretamente interpretáveis em pontos percentuais. Quando muitos valores estão próximos dos limites, ou quando a porcentagem é uma proporção de contagens (plantas doentes em plantas avaliadas), modelos próprios para proporções costumam ser mais apropriados do que transformações (Capítulo 10).

7.17.4 Quando outro modelo é mais apropriado

Algumas variáveis resposta exigem naturalmente outros modelos estatísticos:

- desfechos binários;
- proporções baseadas em contagens;
- contagens;
- notas ordinais.

Esses casos são tratados no Capítulo 10.

> ⚠ **Aviso**
>
> A transformação não deve ser uma resposta automática a um teste de normalidade que falhou. A primeira pergunta é se o modelo estatístico representa adequadamente o processo que gerou os dados.

7.18 Uma nota sobre os testes de normalidade

Testes formais como o de Shapiro–Wilk são frequentemente usados como regras de decisão automáticas:

```
P > 0,05 → usar ANOVA
P < 0,05 → transformar os dados
```

Essa abordagem é problemática. Com amostras grandes, os testes detectam desvios da normalidade triviais e irrelevantes para a inferência. Com as amostras pequenas típicas dos experimentos, têm pouco poder e deixam passar desvios importantes. Além disso, a ANOVA é razoavelmente robusta a desvios moderados da normalidade, especialmente com delineamentos balanceados, ao passo que a heterogeneidade de variâncias e a falta de independência costumam ser problemas mais sérios. Os gráficos de diagnóstico, interpretados com conhecimento do tipo de dado, são mais informativos do que um valor-P de um teste de normalidade.

> ⚠ **Aviso**

Não use um teste de normalidade como uma barreira mecânica para decidir se a ANOVA pode ser feita.

7.19 Médias marginais estimadas

Em delineamentos mais complexos, as médias aritméticas simples podem não representar as comparações implicadas pelo modelo estatístico.

7.19.1 Médias ajustadas pelo modelo

As médias marginais estimadas (também chamadas médias de mínimos quadrados) são as médias dos tratamentos previstas pelo modelo, calculadas em média sobre os demais fatores de forma equilibrada. Em experimentos desbalanceados — por exemplo, com parcelas perdidas em alguns blocos —, elas corrigem as distorções que as médias aritméticas sofreriam. Em experimentos fatoriais, permitem obter médias e contrastes de um fator dentro de cada nível do outro.

7.19.2 Contrastes dois a dois

```
library(emmeans)

emm ← emmeans(model, ~ treatment)

pairs(emm)
```

```
 contrast             estimate   SE df t.ratio p.value
 Testemunha - Fung. A    28.55 2.05 12  13.896 <0.0001
 Testemunha - Fung. B    25.25 2.05 12  12.290 <0.0001
 Testemunha - Fung. C    18.80 2.05 12   9.150 <0.0001
 Testemunha - Fung. D    22.75 2.05 12  11.073 <0.0001
 Fung. A - Fung. B       -3.30 2.05 12  -1.606  0.5207
 Fung. A - Fung. C       -9.75 2.05 12  -4.745  0.0035
 Fung. A - Fung. D       -5.80 2.05 12  -2.823  0.0921
 Fung. B - Fung. C       -6.45 2.05 12  -3.139  0.0543
 Fung. B - Fung. D       -2.50 2.05 12  -1.217  0.7425
 Fung. C - Fung. D        3.95 2.05 12   1.923  0.3569

Results are averaged over the levels of: block
P value adjustment: tukey method for comparing a family of 5 estimates
```

7.19.3 Intervalos de confiança

```
confint(emm)
```

```
 treatment  emmean   SE df lower.CL upper.CL
 Testemunha   54.6 1.45 12     51.4     57.7
 Fung. A      26.0 1.45 12     22.9     29.2
 Fung. B      29.3 1.45 12     26.2     32.5
 Fung. C      35.8 1.45 12     32.6     38.9
 Fung. D      31.8 1.45 12     28.7     35.0
```

```
Results are averaged over the levels of: block
Confidence level used: 0.95
```

7.19.4 Comparações com a testemunha

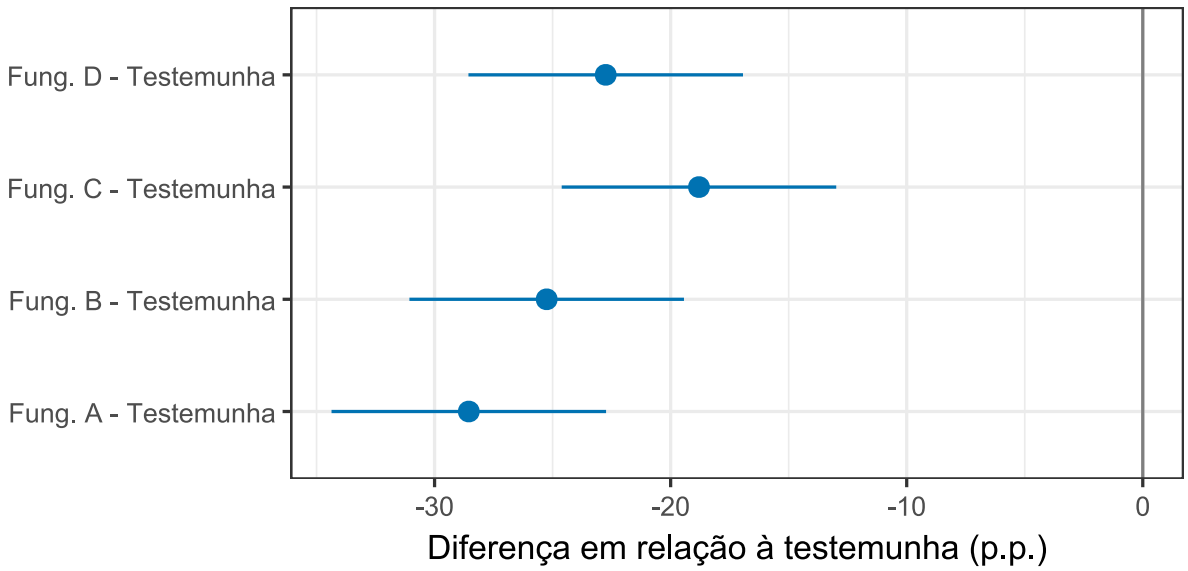
Quando a questão é o efeito de cada tratamento em relação à testemunha, o contraste `"trt.vs.ctrl"` aplica o ajuste de Dunnett e fornece diretamente as diferenças com seus intervalos de confiança:

```
efeitos ← contrast(emm, "trt.vs.ctrl")
confint(efeitos)
```

```
 contrast             estimate   SE df lower.CL upper.CL
 Fung. A - Testemunha    -28.6 2.05 12    -34.4    -22.7
 Fung. B - Testemunha    -25.2 2.05 12    -31.1    -19.4
 Fung. C - Testemunha    -18.8 2.05 12    -24.6    -13.0
 Fung. D - Testemunha    -22.8 2.05 12    -28.6    -16.9

Results are averaged over the levels of: block
Confidence level used: 0.95
Conf-level adjustment: dunnettx method for 4 estimates
```

Esses contrastes podem ser representados no gráfico de efeitos discutido na seção “Das letras às estimativas”:

```
efeitos_df ← as.data.frame(confint(efeitos))

ggplot(efeitos_df, aes(x = estimate, y = contrast,
                       xmin = lower.CL, xmax = upper.CL)) +
  geom_vline(xintercept = 0, colour = "grey50") +
  geom_pointrange(colour = "#0072B2") +
  labs(x = "Diferença em relação à testemunha (p.p.)", y = NULL) +
  theme_bw(base_size = 13)
```



O pacote `emmeans` também permite obter efeitos condicionais em modelos com interação (`emmeans(model, ~ fungicide | cultivar)`, como feito acima para o experimento fatorial) e ajustes para multiplicidade. Esse fluxo de trabalho — ajustar o modelo, obter as médias marginais e estimar os contrastes de interesse com intervalos de confiança — será usado nos capítulos seguintes.

7.20 Pesquisa na prática: avaliando tratamentos de manejo da doença

Retome os ensaios de fungicidas apresentados no Capítulo 1, como o do míldio do manjericão (McGrath 2020). Um experimento desse tipo — vários programas de fungicida e uma testemunha, em blocos casualizados — reúne quase todos os conceitos deste capítulo: blocos no modelo, contrastes com a testemunha, comparações entre tratamentos, multiplicidade e interpretação biológica.

Pesquisa revisitada: Quais tratamentos reduzem a doença?

Considere um experimento de campo que compara vários tratamentos de manejo da doença com uma testemunha sem tratamento.

A questão científica não é simplesmente:

> Os tratamentos são significativamente diferentes?

Em vez disso, pergunte:

1. Quanto cada tratamento reduz a doença em relação à testemunha?
2. Quão incerta é cada redução estimada?
3. Quais comparações foram planejadas cientificamente?
4. As diferenças entre os tratamentos são grandes o suficiente para importar biologicamente?
5. O desempenho dos tratamentos varia entre blocos ou condições experimentais?

A análise deve, portanto, enfatizar **os efeitos estimados dos tratamentos e sua incerteza**, com os testes estatísticos fornecendo evidência complementar.

7.21 Um fluxo de trabalho prático para comparar grupos

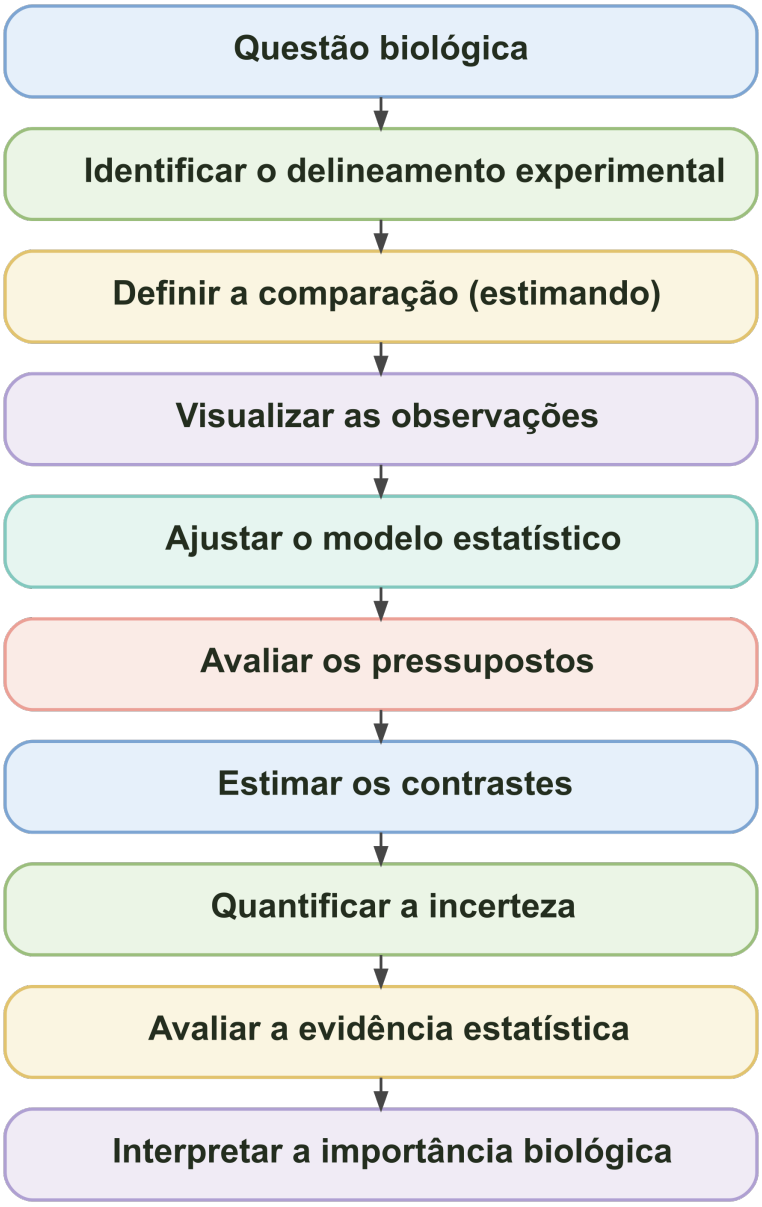


Figura 7.8. Fluxo de trabalho para analisar experimentos que envolvem comparações entre grupos.

7.22 Relatando comparações entre grupos

Um bom enunciado de resultados geralmente inclui:

7.22.1 O que foi comparado?

Os tratamentos ou grupos e a resposta, com o contraste definido explicitamente.

7.22.2 Direção do efeito

Se a resposta aumentou ou diminuiu.

7.22.3 Magnitude do efeito

O tamanho da diferença, em unidades da resposta e, se útil, em termos relativos.

7.22.4 Incerteza

O intervalo de confiança da diferença.

7.22.5 Evidência estatística

O valor-P ou o resultado do teste, como informação complementar.

7.22.6 Interpretação biológica

O que a diferença significa para o sistema estudado.

Em vez de:

> Os tratamentos com fungicidas diferiram significativamente ($P < 0,05$).

prefira algo como:

> O Fungicida A reduziu a severidade média da doença em 28,6 pontos percentuais em relação à testemunha (IC 95% ajustado por Dunnett: 22,7 a 34,4 pontos percentuais), uma redução substancial nas condições do experimento.

7.23 Erros comuns

7.23.1 Relatar apenas valores-P

O valor-P não informa o tamanho nem a precisão do efeito.

7.23.2 Relatar apenas as letras dos tratamentos

As letras escondem a magnitude e a incerteza das diferenças.

7.23.3 Comparar médias sem considerar o delineamento experimental

Ignorar blocos, parcelas subdivididas ou medidas repetidas leva a erros padrão incorretos.

7.23.4 Tratar subamostras como repetições

Folhas, lesões ou leituras repetidas não são unidades experimentais independentes (Capítulo 2).

7.23.5 Fazer vários testes t em vez de modelar todos os grupos

Isso aumenta os falsos positivos e desperdiça informação sobre a variabilidade.

7.23.6 Interpretar uma ANOVA significativa como evidência de que todos os tratamentos diferem

O teste global indica apenas que pelo menos uma média difere.

7.23.7 Ignorar interações

Quando os efeitos de um fator dependem de outro, efeitos médios podem ser enganosos.

7.23.8 Interpretar efeitos principais apesar de interações fortes

Na presença de interação forte, o efeito médio de um fator pode não descrever nenhuma das condições estudadas.

7.23.9 Transformar automaticamente dados em porcentagem

A transformação deve ser justificada pelo comportamento dos dados e pelo objetivo da análise.

7.23.10 Usar testes de normalidade como regras de decisão

Os gráficos de diagnóstico são mais informativos.

7.23.11 Relatar médias sem incerteza

Toda média ou diferença relatada deve vir acompanhada de uma medida de sua precisão.

7.23.12 Igualar significância estatística a importância biológica

Uma diferença significativa pode ser biologicamente trivial, e uma diferença não significativa pode ser importante, mas incerta (Capítulo 6).

7.24 Lista de verificação para comparações entre grupos

Antes de interpretar uma comparação entre grupos, pergunte:

1. Que questão biológica está sendo abordada?
2. Qual é a unidade experimental?
3. Que fator define os grupos?
4. Que variável resposta está sendo comparada?
5. Qual é o estimando principal?
6. As observações são independentes ou pareadas?
7. O modelo reflete os blocos ou outras características do delineamento?
8. As interações são relevantes?
9. Os pressupostos do modelo são razoáveis?
10. Qual é o efeito estimado?
11. Quão incerta é a estimativa?
12. Quais comparações foram planejadas?
13. A multiplicidade foi considerada?
14. O efeito é biologicamente relevante?
15. A conclusão permanece dentro do escopo do experimento?

7.25 Resumo do capítulo

As comparações entre grupos estão entre as análises mais comuns em Fitopatologia. Embora os testes *t* e a análise de variância sejam ferramentas estatísticas familiares, seu propósito vai além de determinar se um valor-P ultrapassa um limiar de significância. O objetivo central é estimar diferenças biologicamente relevantes entre tratamentos ou grupos e quantificar a incerteza associada a essas estimativas.

O teste *t*, a ANOVA, a ANOVA fatorial e as comparações relacionadas podem ser compreendidos dentro do arcabouço geral do modelo linear. Essa perspectiva oferece uma base coerente para analisar delineamentos experimentais simples e complexos, enfatizando contrastes que respondem diretamente à questão biológica. Uma interpretação confiável também exige atenção ao delineamento experimental, aos pressupostos do modelo, às interações, à multiplicidade, à magnitude dos efeitos e à incerteza.

7.26 Conceitos-chave

Ao concluir este capítulo, você deverá ser capaz de explicar:

- comparação entre grupos;
- observações independentes;
- observações pareadas;
- diferença entre médias;
- teste *t*;
- teste *t* de Welch;
- análise de variância;
- teste F;
- modelo linear;
- fator;
- efeito principal;
- interação;
- contraste;
- comparação planejada;
- comparação dois a dois;
- comparações múltiplas;
- média marginal estimada;
- resíduo;
- homogeneidade de variâncias;
- diagnóstico do modelo;
- tamanho de efeito;
- intervalo de confiança.

7.27 Exercícios

7.27.1 Exercício 1. Qual é a comparação?

Para cada questão biológica, identifique:

1. a variável resposta;
2. os grupos comparados;
3. a unidade experimental;
4. o estimando adequado;
5. a direção em que o contraste deve ser definido.

Use questões como:

- Um fungicida reduz a severidade da doença?
- Uma cultivar resistente desenvolve menos doença que uma suscetível?
- A inoculação reduz a taxa fotossintética?

7.27.2 Exercício 2. Dois grupos

Usando um conjunto de dados fornecido pelo professor:

1. visualize as observações individuais;
2. calcule a resposta média de cada grupo;
3. estime a diferença entre as médias;
4. calcule um intervalo de confiança;

5. faça o teste *t* adequado;
6. interprete o efeito biologicamente.

7.27.3 Exercício 3. Vários tratamentos

Analise um experimento que contenha uma testemunha e vários tratamentos.

1. visualize os dados;
2. ajuste um modelo linear adequado;
3. faça o teste global;
4. estime os efeitos dos tratamentos em relação à testemunha;
5. calcule os intervalos de confiança;
6. compare esses resultados com uma apresentação convencional de comparações múltiplas.

7.27.4 Exercício 4. Apresentação com letras

Você recebe uma tabela de médias de tratamentos e letras de um procedimento de comparações múltiplas.

Discuta:

1. Quais comparações podem ser inferidas a partir das letras?
2. Que informação sobre a magnitude dos efeitos está faltando?
3. Que informação sobre a incerteza está faltando?
4. Como os resultados poderiam ser apresentados de forma mais informativa?

7.27.5 Exercício 5. Interação

Considere um experimento que avalia dois fungicidas em duas cultivares.

Determine se os dados hipotéticos sugerem:

1. ausência de interação;
2. interação quantitativa;
3. respostas fortemente dependentes do tratamento.

Desenhe um gráfico de interação e descreva a interpretação biológica.

7.27.6 Exercício 6. Diagnóstico do modelo

Ajuste um modelo linear ao conjunto de dados fornecido.

Examine:

- os resíduos versus os valores ajustados;
- o gráfico quantil–quantil;
- a variabilidade residual;
- as observações incomuns.

Discuta se o modelo parece adequado e que investigação adicional pode ser necessária.

7.27.7 Exercício 7. Avaliação de um artigo publicado

Escolha um artigo recente de Fitopatologia que compare tratamentos experimentais.

Identifique:

1. o delineamento experimental;
2. a unidade experimental;
3. a variável resposta;
4. o modelo estatístico;

5. as comparações realizadas;
6. se os tamanhos de efeito foram relatados;
7. se a incerteza foi relatada;
8. se foram usadas letras para os tratamentos;
9. se a interpretação biológica é sustentada pelos efeitos estimados.

7.28 Reflexão

> **A questão não é simplesmente se os tratamentos diferem, mas o quanto diferem e se essa diferença importa biologicamente.**

Discuta essa afirmação usando um exemplo da Fitopatologia.

7.29 Próximos passos

Muitas questões biológicas envolvem não comparações entre grupos discretos, mas relações entre variáveis quantitativas. A doença pode aumentar com a temperatura, a produtividade pode diminuir com o aumento da intensidade da doença, ou o crescimento do patógeno pode mudar com as condições ambientais. O Capítulo 8 estende o arcabouço do modelo linear apresentado aqui à análise de relações entre variáveis quantitativas.

CAPÍTULO 8

Como estudamos relações entre variáveis?

8.1 Objetivos de aprendizagem

! Importante

Ao concluir este capítulo, o leitor deverá ser capaz de:

- identificar questões de pesquisa que envolvem relações entre variáveis quantitativas;
- explorar relações graficamente antes de ajustar um modelo;
- distinguir associação de causalidade;
- estimar e interpretar coeficientes de correlação e reconhecer suas limitações;
- distinguir correlação de concordância entre dois métodos de medida;
- formular um modelo de regressão linear simples e interpretar seu intercepto e sua inclinação em unidades biológicas;
- quantificar a incerteza das estimativas da regressão e distinguir intervalos de confiança de intervalos de predição;
- interpretar o coeficiente de determinação sem superestimá-lo;
- diagnosticar modelos de regressão, incluindo observações influentes;
- comparar relações entre grupos combinando preditores quantitativos e categóricos;
- ajustar e interpretar modelos de regressão linear múltipla;
- reconhecer colinearidade, extrapolação e outras ameaças à interpretação;
- comunicar resultados de regressão usando estimativas, incerteza e gráficos adequados.

8.2 Por que estudar relações?

O Capítulo 7 tratou de questões em que a variável explicativa define grupos: tratado versus não tratado, resistente versus suscetível, inoculado versus não inoculado. Muitas questões em Fitopatologia, porém, envolvem variáveis explicativas que são, elas próprias, quantitativas.

Questões típicas incluem:

- Quanto de produtividade se perde para cada unidade de aumento na severidade da doença?
- O tamanho da lesão aumenta com a temperatura?
- Como o número de esporos capturados se relaciona com a incidência posterior da doença?
- A infecção aumenta com a duração do molhamento foliar?

- As estimativas de severidade obtidas por análise de imagens correspondem às estimativas visuais?
- A biomassa do patógeno no tecido está associada à severidade dos sintomas?

Essas questões compartilham uma estrutura comum: queremos descrever **como uma resposta muda à medida que outra variável quantitativa muda**, e com que confiança podemos descrever essa mudança.

> **! Importante**
>
> A grandeza central na maioria dos estudos de relações não é um valor-P nem um coeficiente de correlação, mas a **taxa de variação** da resposta em relação à variável explicativa, expressa em unidades biológicas e acompanhada de sua incerteza.

8.3 Da questão biológica à relação

Considere um estudo de campo que relaciona a severidade da doença em um estádio-chave da cultura à produtividade final.

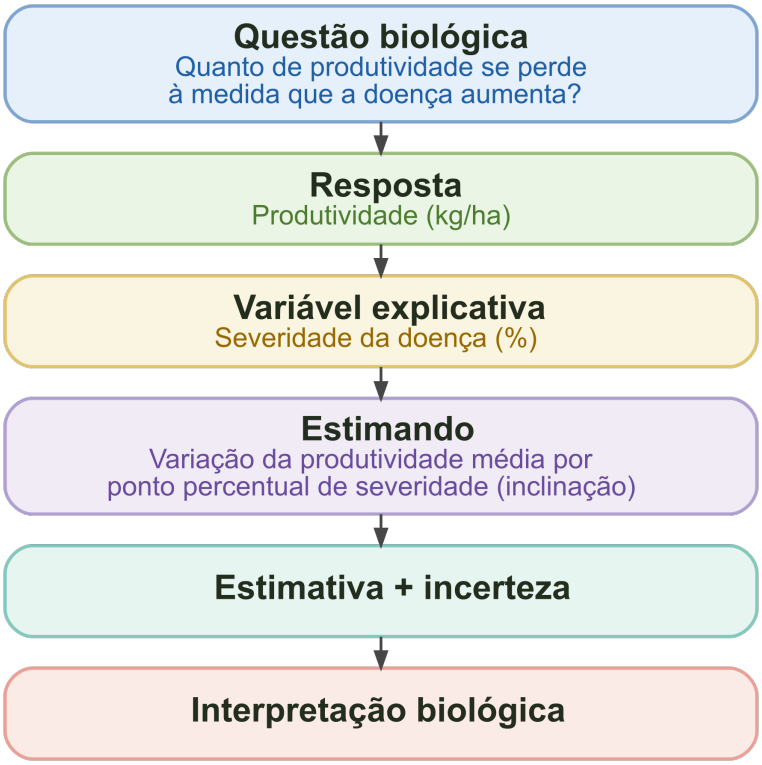


Figura 8.1. Da questão biológica à estimação de uma taxa de variação (inclinação).

8.3.1 Qual é o estimando?

Se a produtividade média varia linearmente com a severidade, podemos escrever a produtividade esperada na severidade x como:

$$E(Y \mid x) = \beta_0 + \beta_1 x$$

O estimando de interesse principal geralmente é a inclinação:

$$\beta_1 = \text{variação da produtividade média por unidade de aumento na severidade}$$

A estimativa correspondente obtida dos dados é $\hat{\beta}_1$.

> **Dica**
>
> Uma inclinação sempre tem unidades: unidades da resposta por unidade da variável explicativa. “A inclinação foi −42” é incompleto; “a produtividade diminuiu 42 kg/ha para cada ponto percentual de severidade” é interpretável.

8.3.2 Qual variável é a resposta?

Em um modelo de regressão, a escolha da resposta e da variável explicativa não é arbitrária. Ela decorre da questão biológica:

- a **resposta** é a variável cujo comportamento queremos explicar ou prever;
- a **variável explicativa** é a variável que usamos para explicá-la ou prevê-la.

A produtividade é modelada em função da doença, e não o contrário, porque a questão é sobre a consequência da doença para a produtividade. Fazer a regressão de y em x e a de x em y geralmente produz retas diferentes e responde a perguntas diferentes.

8.4 Olhando os dados antes de ajustar uma reta

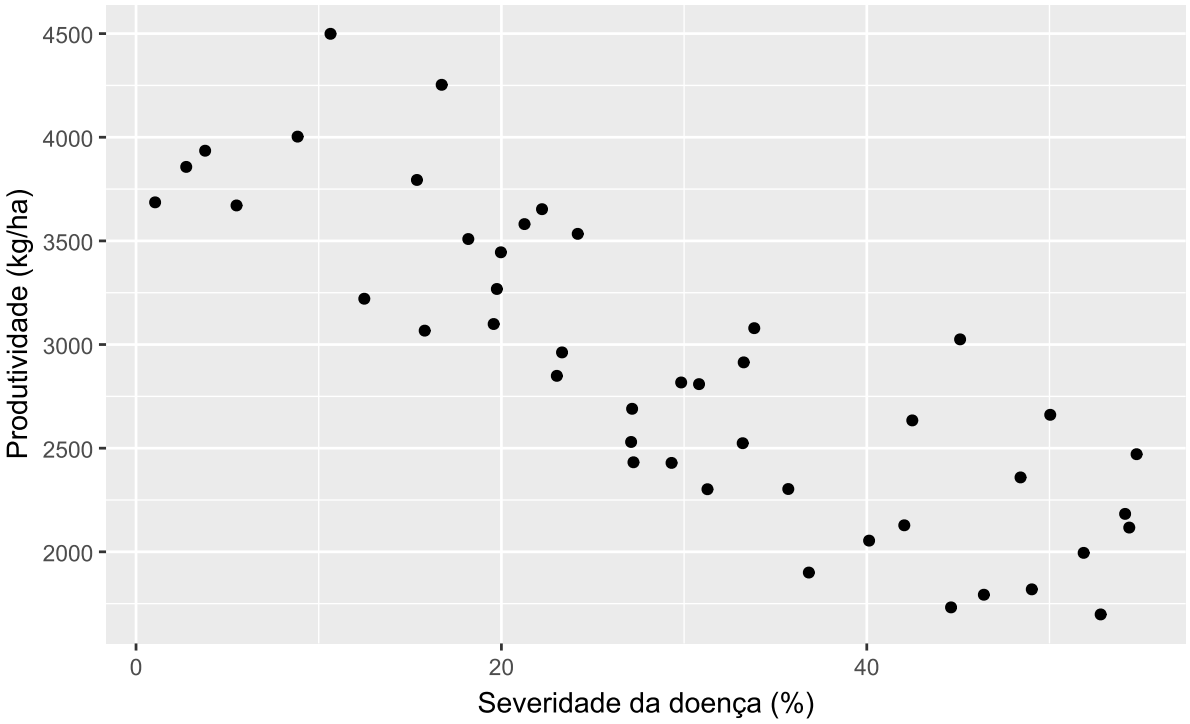
Como nos capítulos anteriores, a análise começa pelas observações, e não pelo modelo (ver Capítulos 4 e 5).

8.4.1 O gráfico de dispersão

O gráfico de dispersão é o gráfico básico para a relação entre duas variáveis quantitativas. Antes de ajustar qualquer modelo, examine nele:

- a **direção** da relação (crescente, decrescente, nenhuma);
- a **forma** da relação (linear, curva, um platô, um limiar);
- a **intensidade** da relação (quão próximos os pontos estão de uma tendência);
- a **variabilidade** que muda ao longo da variável explicativa;
- **observações incomuns** e possíveis erros de digitação;
- **grupos ou agrupamentos** que podem corresponder a locais, anos, cultivares ou experimentos.

```
library(ggplot2)

ggplot(data, aes(x = severity, y = yield)) +
  geom_point() +
  labs(x = "Severidade da doença (%)",
       y = "Produtividade (kg/ha)")
```



8.4.2 Por que os resumos sozinhos são insuficientes

Relações muito diferentes podem produzir as mesmas médias, desvios padrão, coeficiente de correlação e reta de regressão ajustada. A ilustração clássica é o quarteto de Anscombe: quatro conjuntos de dados com estatísticas-resumo praticamente idênticas, dos quais apenas um é bem descrito por uma reta.

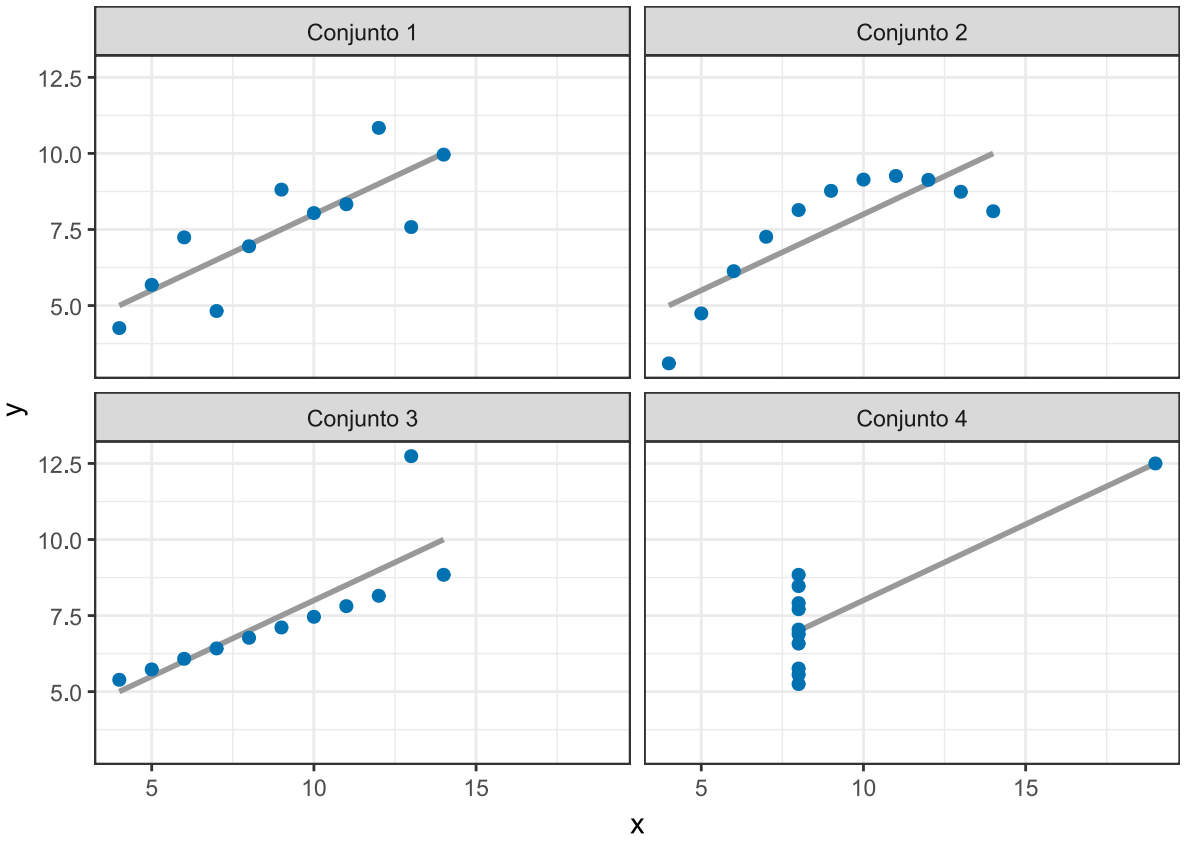
```
# O quarteto de Anscombe está disponível no R base
sapply(1:4, function(i) {
  x ← anscombe[[paste0("x", i)]]; y ← anscombe[[paste0("y", i)]]
  round(c(media_y = mean(y), dp_y = sd(y), r = cor(x, y),
          inclinacao = unname(coef(lm(y ~ x))[2])), 2)
})
```

```
           [,1] [,2] [,3] [,4]
media_y    7.50 7.50 7.50 7.50
dp_y       2.03 2.03 2.03 2.03
r          0.82 0.82 0.82 0.82
inclinacao 0.50 0.50 0.50 0.50
```

Os quatro conjuntos têm praticamente as mesmas médias, desvios padrão, correlações e inclinações, mas o gráfico mostra relações completamente diferentes:

```
quarteto ← data.frame(
  conjunto = rep(paste("Conjunto", 1:4), each = 11),
  x = unlist(anscombe[, 1:4]), y = unlist(anscombe[, 5:8])
)

ggplot(quarteto, aes(x, y)) +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE, colour = "grey60") +
  geom_point(colour = "#0072B2", size = 2) +
```

```
  facet_wrap(~ conjunto) +
  theme_bw(base_size = 12)
```

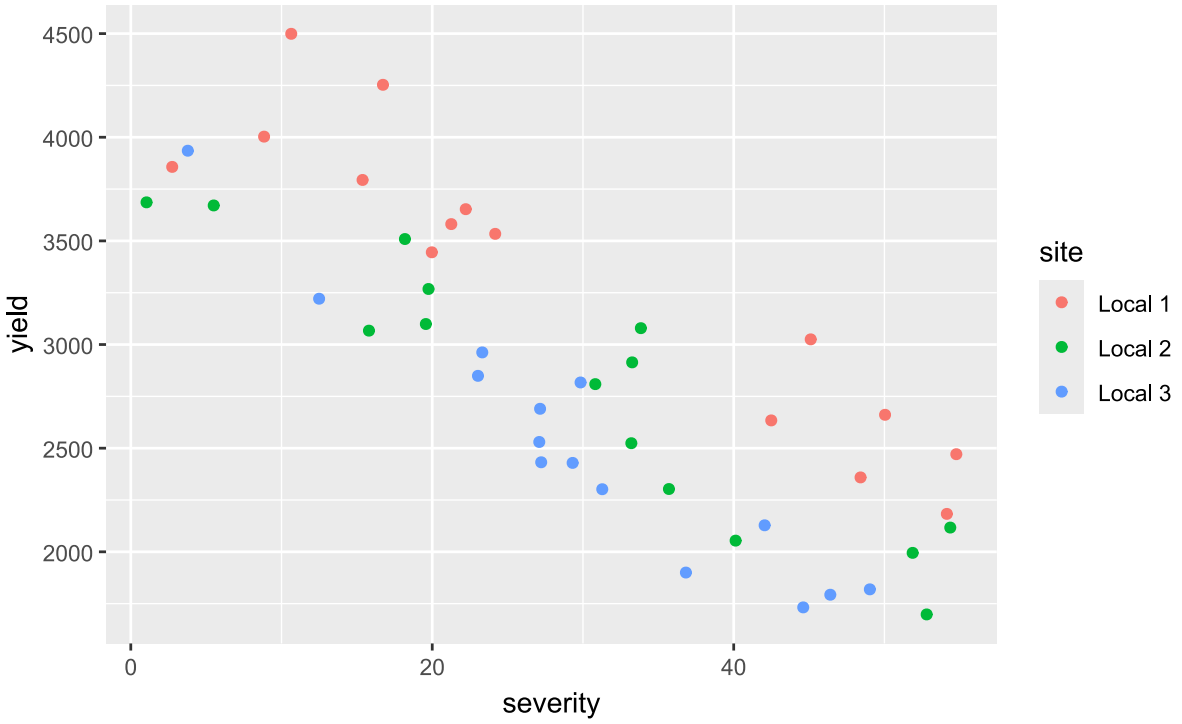


> **⚠ Aviso**
>
> Nunca ajuste nem relate uma correlação ou regressão sem olhar o gráfico de dispersão. Um número-resumo não revela curvaturas, valores discrepantes ou agrupamentos.

8.4.3 Mostre a estrutura dos dados

Os pontos de um gráfico de dispersão frequentemente não são independentes. As observações podem vir de várias lavouras, anos, cultivares ou de medidas repetidas nas mesmas plantas. Colorir ou separar em painéis os pontos segundo essas fontes frequentemente revela que uma relação aparente no conjunto é determinada por diferenças entre grupos, e não por uma relação dentro deles.

```
ggplot(data, aes(x = severity, y = yield, color = site)) +
  geom_point()
```



8.5 Correlação

8.5.1 O que um coeficiente de correlação mede?

O coeficiente de correlação de Pearson, r, mede a intensidade e a direção de uma associação **linear** entre duas variáveis quantitativas. Ele varia de −1 a 1:

- r próximo de 1: os pontos se agrupam em torno de uma reta crescente;
- r próximo de −1: os pontos se agrupam em torno de uma reta decrescente;
- r próximo de 0: ausência de associação linear (mas possivelmente uma forte associação não linear).

A correlação é simétrica: a correlação entre x e y é igual à correlação entre y e x. Ela não tem unidades e não distingue resposta de variável explicativa.

```
cor.test(data$severity, data$yield)
```

```
    Pearson's product-moment correlation

data:  data$severity and data$yield
t = -9.9782, df = 43, p-value = 9.224e-13
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9067872 -0.7183787
sample estimates:
       cor
-0.8356916
```

8.5.2 Correlação por postos

Quando a relação é monotônica, mas não linear, ou quando alguns valores extremos dominam, a correlação por postos de Spearman descreve a associação entre as posições (postos) das observações, e não entre seus valores.

```
cor.test(data$severity, data$yield, method = "spearman")
```

```
    Spearman's rank correlation rho

data:  data$severity and data$yield
S = 28162, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.8552042
```

8.5.3 O que a correlação não nos diz

Um coeficiente de correlação não nos diz:

- **quanto** a resposta muda por unidade da variável explicativa;
- se a relação é **causal**;
- se a relação é **linear** (uma relação curva forte pode ter r próximo de 0);
- se um método de medida **concorda** com outro.

O valor de r também depende da amplitude da variável explicativa estudada. Amostrar apenas uma faixa estreita de severidade (por exemplo, um único local com pouca doença) produzirá uma correlação mais fraca do que amostrar uma faixa ampla, mesmo que a relação subjacente seja idêntica.

8.5.4 Correlação e causalidade

Estudos observacionais (Capítulo 2) frequentemente revelam associações: lavouras com mais inóculo têm mais doença, plantas com mais lesões produzem menos, regiões com mais chuva têm mais epidemias. Essas associações podem refletir relações causais, mas também podem surgir de **variáveis de confundimento** que afetam ambas as grandezas, como data de semeadura, cultivar, manejo ou solo.

> **! Importante**
>
> Uma regressão ou correlação estimada a partir de dados observacionais descreve uma associação. A interpretação causal exige um delineamento experimental, ou uma argumentação cuidadosa de que as variáveis de confundimento foram medidas e consideradas.

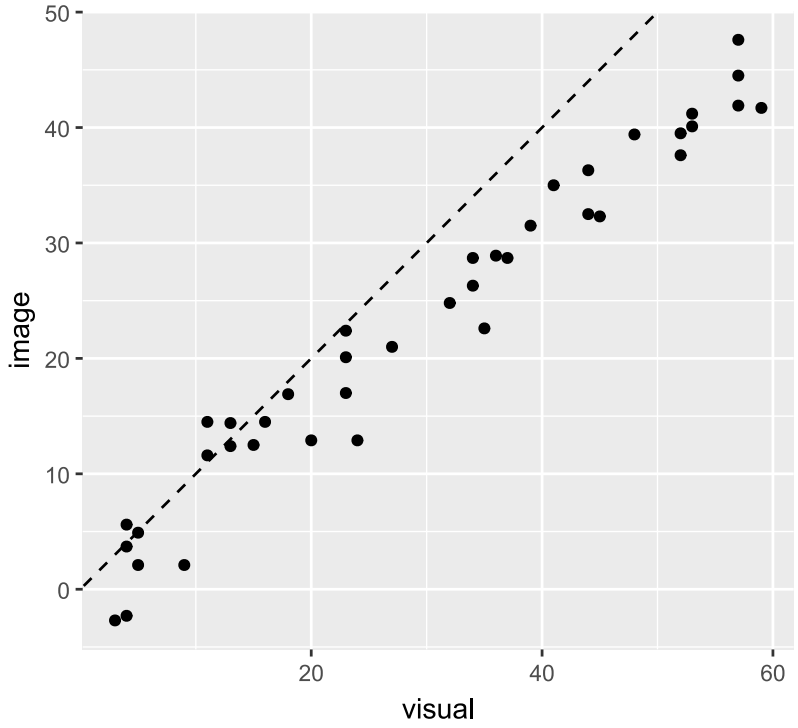
8.5.5 Correlação versus concordância

Um uso comum da correlação em Fitopatologia é comparar dois métodos de medir a mesma grandeza, como estimativas visuais de severidade e estimativas por análise de imagens, ou estimativas de dois avaliadores.

Uma correlação alta não significa que os métodos concordam. Se um avaliador estima sistematicamente o dobro da severidade estimada pelo outro, suas estimativas são perfeitamente correlacionadas, mas sistematicamente diferentes.

A concordância exige que as observações caiam sobre a **linha de igualdade** ($y = x$), e não simplesmente sobre alguma reta. Medidas como o coeficiente de correlação de concordância de Lin e ferramentas gráficas como os gráficos de Bland–Altman avaliam diretamente a concordância.

```
ggplot(aval, aes(x = visual, y = image)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, linetype = 2) +
  coord_equal()
```



> ⚠ **Aviso**
>
> Não use o coeficiente de correlação como evidência de que dois métodos de avaliação são intercambiáveis. Avalie o viés e a concordância em relação à linha de igualdade.

8.6 Regressão linear simples

8.6.1 O modelo

Um modelo de regressão linear simples descreve cada observação como:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

em que:

- y_i é a resposta da observação i;
- x_i é o valor da variável explicativa;

- β_0 é o intercepto, a resposta esperada quando $x = 0$;
- β_1 é a inclinação, a variação esperada da resposta para um aumento de uma unidade em x;
- ε_i é a variação residual não explicada pela reta.

É o mesmo arcabouço de modelo linear usado no Capítulo 7. A única diferença é que a variável explicativa agora é quantitativa, e não um conjunto de grupos.

> **! Importante**
>
> Comparar grupos e estudar relações não são mundos estatísticos separados. Ambos são modelos lineares; diferem apenas em se as variáveis explicativas são categóricas, quantitativas ou ambas.

8.6.2 Ajustando uma reta: mínimos quadrados

A reta ajustada é aquela que minimiza a soma dos quadrados das distâncias verticais entre as observações e a reta:

$$\sum_i (y_i - \hat{y}_i)^2$$

Essas distâncias verticais são os **resíduos**. Como as distâncias são medidas verticalmente, os mínimos quadrados tratam x como conhecido e atribuem toda a variação não explicada à resposta.

8.6.3 Um exemplo simples

O código a seguir simula um conjunto de dados de produtividade e severidade que o leitor pode reproduzir. Na prática, os dados viriam de um estudo de campo.

```
set.seed(606)

data ← data.frame(
  severity = runif(40, min = 0, max = 60)
)

data$yield ← 4200 - 35 * data$severity + rnorm(40, sd = 250)
```

```
model ← lm(yield ~ severity, data = data)

summary(model)
```

```

Call:
lm(formula = yield ~ severity, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-486.3 -190.1  -41.8  136.2  653.1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4183.98      87.82   47.64   <2e-16 ***
severity      -36.22       2.46  -14.72   <2e-16 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 275.8 on 38 degrees of freedom
Multiple R-squared:  0.8509,    Adjusted R-squared:  0.8469
F-statistic: 216.8 on 1 and 38 DF,  p-value: < 2.2e-16
```

```
confint(model)
```

```
                2.5 %     97.5 %
(Intercept) 4006.1948 4361.77245
severity     -41.2022  -31.24159
```

8.6.4 Interpretando o intercepto

O intercepto é a produtividade esperada quando a severidade é zero. Em um estudo de perda de produtividade, ele tem um significado claro: a **produtividade atingível** na ausência da doença, nas condições do estudo.

Em muitos outros casos, $x = 0$ está muito fora dos dados observados (uma temperatura de 0 °C em um estudo conduzido entre 15 e 30 °C, por exemplo), e o intercepto não tem significado biológico direto. Centralizar a variável explicativa torna o intercepto interpretável:

```
temp_data$temp_c ← temp_data$temperature - 20

lm(lesion_size ~ temp_c, data = temp_data)
```

```

Call:
lm(formula = lesion_size ~ temp_c, data = temp_data)

Coefficients:
(Intercept)       temp_c
      8.423        0.232
```

O intercepto agora representa o tamanho de lesão esperado a 20 °C; a inclinação não muda.

8.6.5 Interpretando a inclinação

No exemplo da produtividade, a inclinação estima quantos quilogramas por hectare de produtividade se perdem para cada ponto percentual adicional de severidade. Essa grandeza tem significado agronômico direto e pode ser usada para avaliar se o controle da doença se justifica economicamente.

Dividir a inclinação pelo intercepto expressa a perda em termos relativos:

$$\frac{\hat{\beta}_1}{\hat{\beta}_0} \times 100$$

a porcentagem da produtividade atingível perdida por ponto percentual de severidade. Inclinações relativas são úteis para comparar estudos conduzidos em diferentes níveis de produtividade.

Dica

> Relate as inclinações em unidades que o leitor possa usar e escolha a unidade da variável explicativa de forma sensata. Uma inclinação por 1 °C, por 10 esporos ou por 10 pontos percentuais de severidade pode ser mais informativa do que uma inclinação por unidade.

8.7 Incerteza nas estimativas da regressão

8.7.1 Erros padrão e intervalos de confiança

Como qualquer estimativa (Capítulo 6), a inclinação estimada seria diferente se o estudo fosse repetido. Seu erro padrão descreve essa variabilidade amostral, e seu intervalo de confiança descreve a faixa de inclinações compatíveis com os dados.

Um intervalo de confiança para a inclinação que exclui o zero indica evidência de associação linear. Mais importante, a **largura** do intervalo mostra com que precisão a taxa de variação foi estimada, e seus **limites** mostram se os valores plausíveis são todos biologicamente importantes, todos desprezíveis ou uma mistura.

8.7.2 Testando a inclinação

O teste relatado por `summary()` avalia:

$$H_0 : \beta_1 = 0$$

Essa hipótese nula raramente tem interesse científico por si só. Poucos fitopatologistas duvidam de que a produtividade diminui à medida que a doença severa aumenta; a pergunta é **com que rapidez**.

8.7.3 Intervalos de confiança versus intervalos de predição

Dois tipos de intervalo são comumente desenhados em torno de uma reta de regressão, e eles respondem a perguntas diferentes:

- um **intervalo de confiança** para a resposta média descreve a incerteza sobre a resposta *média* em determinado valor de x;
- um **intervalo de predição** descreve a faixa em que se espera que uma *nova observação individual* caia.

Os intervalos de predição são sempre mais largos, porque incluem tanto a incerteza da reta quanto a variabilidade das observações individuais em torno dela.

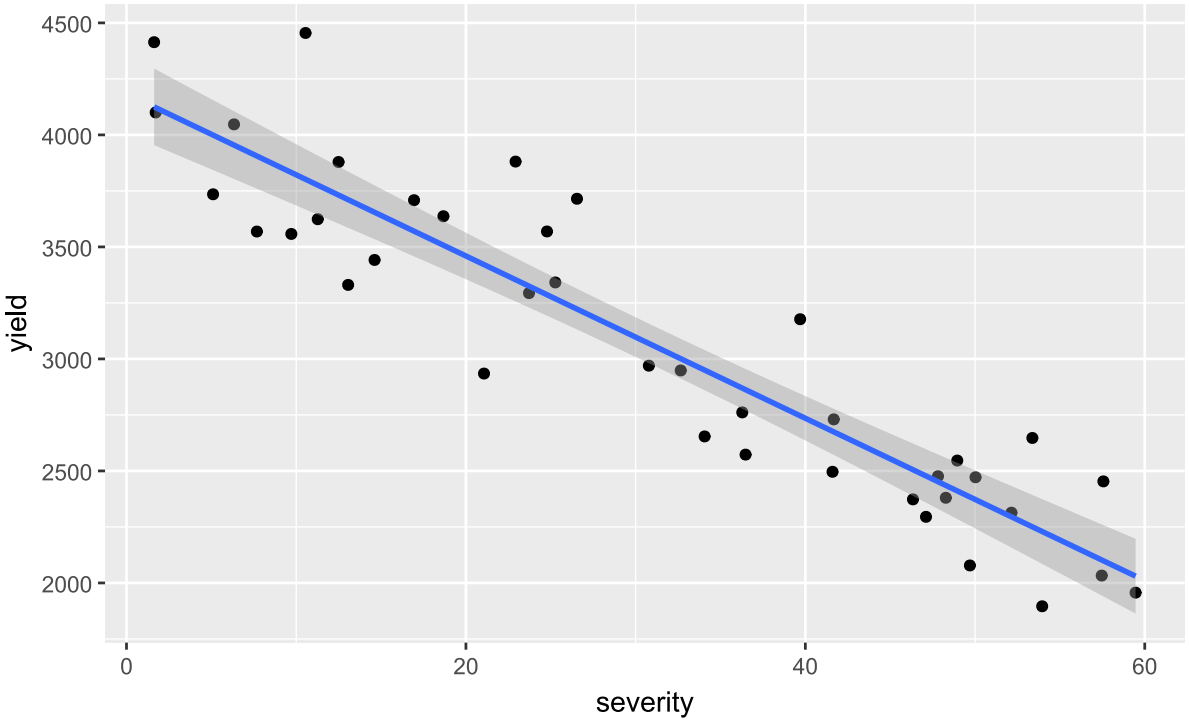
```
new ← data.frame(severity = c(10, 30, 50))

predict(model, newdata = new, interval = "confidence")
```

```
       fit      lwr      upr
1 3821.765 3684.953 3958.577
2 3097.327 3008.914 3185.739
3 2372.889 2243.433 2502.345
```

```
predict(model, newdata = new, interval = "prediction")
```

```
       fit      lwr      upr
1 3821.765 3246.940 4396.589
2 3097.327 2532.063 3662.590
3 2372.889 1799.770 2946.007
```

```
ggplot(data, aes(x = severity, y = yield)) +
  geom_point() +
  geom_smooth(method = "lm")   # faixa sombreada: intervalo de confiança da média
```



> ⚠ **Aviso**
>
> Se o objetivo é prever a produtividade de uma lavoura específica, use um intervalo de predição. A estreita faixa de confiança em torno de uma reta de regressão subestima muito a incerteza para lavouras individuais.

8.8 Quanto da variação é explicado?

8.8.1 O coeficiente de determinação

O coeficiente de determinação, R^2, é a proporção da variação na resposta explicada pela reta ajustada. Na regressão linear simples, ele é igual ao quadrado da correlação de Pearson.

8.8.2 O que o R^2 diz e o que não diz

O R^2 é frequentemente superinterpretado. Ele não nos diz:

- se a relação é linear;
- se a inclinação é grande ou biologicamente importante;
- se o modelo vai prever bem em outros ambientes;
- se o modelo está correto.

Uma inclinação acentuada e biologicamente importante pode coexistir com um R^2 modesto quando as observações individuais são variáveis, como é comum em dados de campo. Por outro lado, um R^2 alto pode surgir de uma relação curva forçada em uma reta, ou de alguns pontos extremos.

> **! Importante**
>
> Interprete primeiro a inclinação e sua incerteza. O R^2 é uma descrição secundária da dispersão em torno da reta, e não uma medida do valor científico de uma relação.

8.9 Pressupostos e diagnóstico do modelo

Os pressupostos do modelo linear foram apresentados no Capítulo 7. Na regressão, um deles ganha destaque.

8.9.1 Linearidade

O modelo pressupõe que a resposta média varia linearmente com x. Nas comparações entre grupos, esse pressuposto é essencialmente satisfeito por construção; na regressão, é uma afirmação biológica substantiva. Muitas respostas biológicas se curvam, saturam ou têm limiares, e forçar uma reta sobre elas produz inclinações enganosas.

8.9.2 Independência

As observações devem ser independentes, dada a variável explicativa. Medidas repetidas nas mesmas plantas, várias parcelas da mesma lavoura ou observações de datas consecutivas na mesma lavoura não são independentes, e tratá-las como tal superestima a precisão da inclinação.

8.9.3 Variância constante

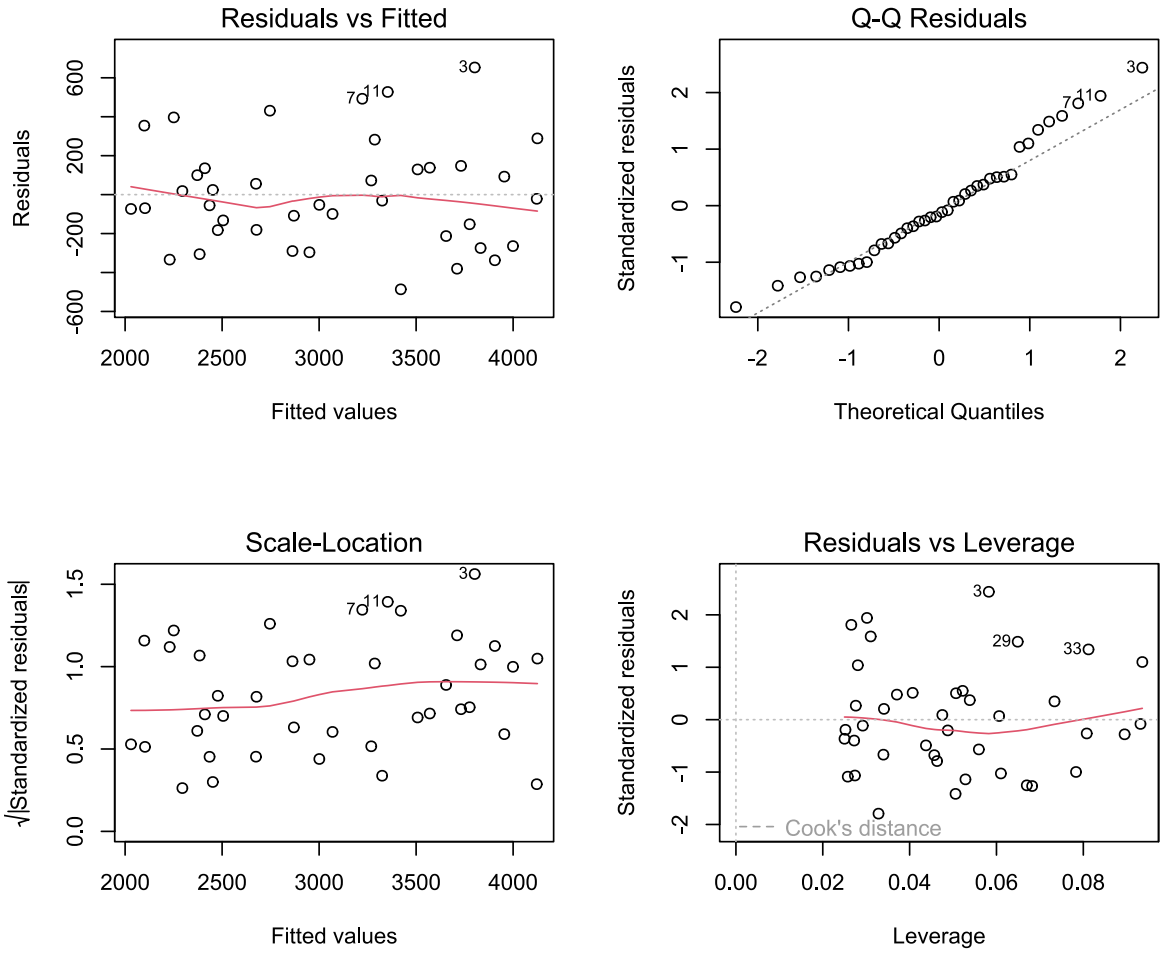
Em dados de Fitopatologia, a variabilidade frequentemente aumenta com a média, por exemplo quando as contagens de lesões ou de esporos aumentam. Um gráfico de resíduos em forma de funil indica variância não constante.

8.9.4 Distribuição dos resíduos

Como no Capítulo 7, o pressuposto distribucional diz respeito aos resíduos, e não à resposta bruta nem à variável explicativa. A variável explicativa não precisa ter distribuição normal.

8.9.5 Diagnosticando o modelo

```
par(mfrow = c(2, 2))
plot(model)
```



O gráfico de resíduos versus valores ajustados é o diagnóstico mais informativo na regressão: um padrão curvo indica que uma reta não descreve a relação.

8.9.6 Alavancagem e influência

Observações com valores incomuns de x têm alta **alavancagem**: têm o potencial de puxar a reta de regressão em sua direção. Uma observação é **influente** quando removê-la mudaria substancialmente a inclinação estimada.

Em estudos de perda de produtividade, uma única lavoura com severidade extremamente alta pode determinar a maior parte da inclinação estimada. Essas observações não devem ser removidas automaticamente, mas sua influência deve ser examinada e relatada.

```
round(cooks.distance(model), 3)
```

```
    1     2     3     4     5     6     7     8     9    10    11    12    13
0.000 0.036 0.184 0.011 0.005 0.004 0.045 0.001 0.006 0.004 0.059 0.062 0.002
   14    15    16    17    18    19    20    21    22    23    24    25    26
0.040 0.016 0.059 0.007 0.004 0.056 0.034 0.008 0.042 0.016 0.055 0.000 0.006
   27    28    29    30    31    32    33    34    35    36    37    38    39
0.016 0.008 0.077 0.010 0.053 0.001 0.079 0.015 0.003 0.000 0.000 0.001 0.002
```

```
   40
0.000
```

O exemplo abaixo acrescenta aos dados uma única lavoura com severidade muito alta e produtividade maior do que o esperado. Essa observação, de alta alavancagem, muda visivelmente a inclinação:

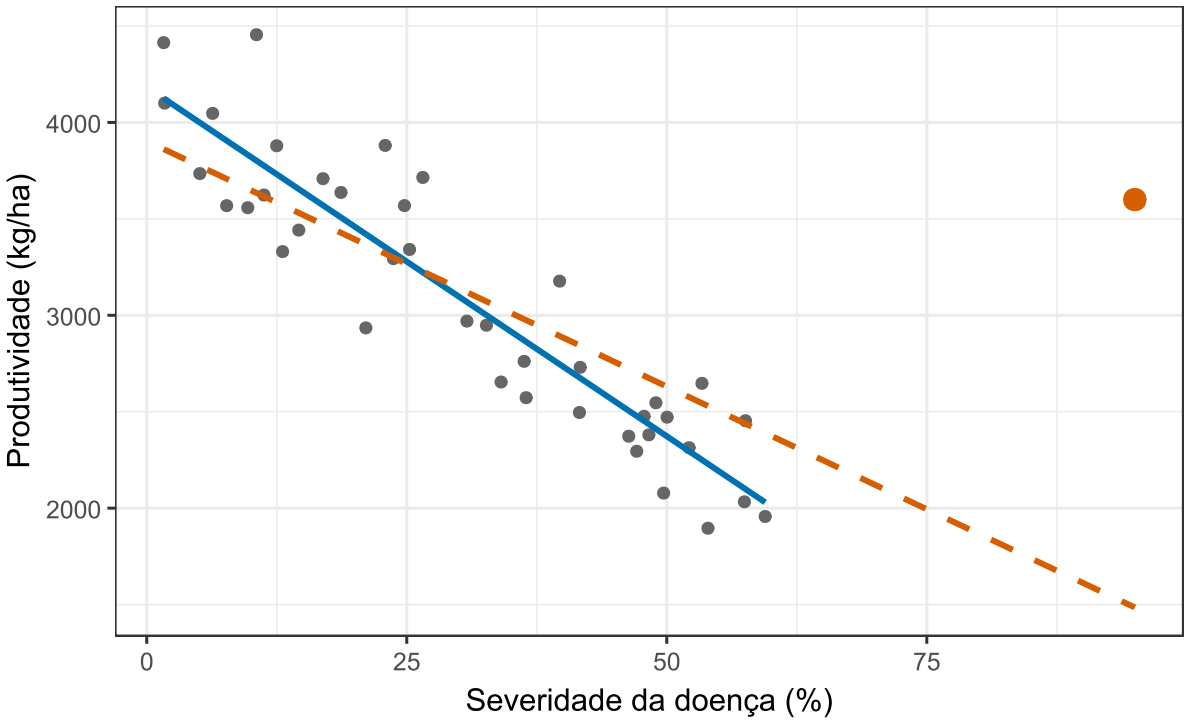
```
com_extremo ← rbind(data[, c("severity", "yield")],
                    data.frame(severity = 95, yield = 3600))

coef(lm(yield ~ severity, data = data))[2]          # sem a lavoura extrema
```

```
severity
-36.2219
```

```
coef(lm(yield ~ severity, data = com_extremo))[2]   # com a lavoura extrema
```

```
severity
-25.4502
```

```
ggplot(com_extremo, aes(severity, yield)) +
  geom_point(colour = "grey40") +
  geom_point(data = data.frame(severity = 95, yield = 3600),
             colour = "#D55E00", size = 3.5) +
  geom_smooth(data = data, method = "lm", formula = y ~ x, se = FALSE,
              colour = "#0072B2") +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE,
              colour = "#D55E00", linetype = 2) +
  labs(x = "Severidade da doença (%)", y = "Produtividade (kg/ha)") +
  theme_bw(base_size = 12)
```



A linha contínua azul é a reta sem a lavoura extrema; a tracejada laranja, com ela.

> ⚠ **Aviso**
>
> Não apague observações porque são inconvenientes. Verifique se são erros; se não forem, relate como as conclusões mudam com e sem elas.

8.10 Quando a relação não é linear

8.10.1 Transformando variáveis

Algumas relações curvas tornam-se aproximadamente lineares em uma escala transformada. Transformações logarítmicas são comuns para contagens de esporos, concentrações do patógeno e doses. Uma transformação muda a interpretação da inclinação: com uma variável explicativa transformada em log, a inclinação descreve a variação da resposta por mudança multiplicativa (por exemplo, de dez vezes) em x.

```
lm(severity ~ log10(spores), data = esporos)
```

```

Call:
lm(formula = severity ~ log10(spores), data = esporos)

Coefficients:
  (Intercept)  log10(spores)
      -0.9353         6.2078
```

8.10.2 Termos polinomiais

Acrescentar um termo quadrático permite uma curvatura simples:

```
lm(lesion_size ~ temperature + I(temperature^2), data = temp_data)
```

```

Call:
lm(formula = lesion_size ~ temperature + I(temperature^2), data = temp_data)

Coefficients:
     (Intercept)       temperature  I(temperature^2)
       -23.64829           2.99182          -0.06272
```

Polinômios podem descrever uma temperatura ótima dentro da faixa observada, mas seus parâmetros raramente têm significado biológico direto e eles se comportam mal fora dos dados.

8.10.3 Modelos não lineares com significado biológico

Quando a própria forma da resposta é o objeto de estudo, como uma curva de dose–resposta, uma curva de crescimento ou uma resposta à temperatura com um ótimo, são preferíveis modelos cujos parâmetros tenham significado biológico. Esses modelos são o tema do Capítulo 9.

8.11 Extrapolação

Uma reta de regressão descreve a relação dentro da faixa dos dados observados. Fora dessa faixa, não há evidência sobre se a relação continua.

Uma inclinação de perda de produtividade estimada em lavouras com severidade entre 0 e 20% nada diz sobre a produtividade com 60% de severidade. Da mesma forma, uma relação entre temperatura e lesão estimada entre 15 e 25 °C não pode ser estendida a 35 °C, temperatura em que o patógeno pode simplesmente parar de crescer.

> ⚠ **Aviso**
>
> Não use um modelo de regressão para prever fora da faixa das variáveis explicativas usadas para ajustá-lo, e não desenhe retas ajustadas além dos dados observados.

8.12 Comparando relações entre grupos

Muitas vezes, a questão biológica é se uma relação difere entre grupos: a produtividade diminui mais rapidamente com a doença em uma cultivar do que em outra? O tamanho da lesão responde de forma diferente à temperatura em dois isolados?

8.12.1 Combinando preditores quantitativos e categóricos

Um modelo com uma variável quantitativa e um fator permite que o intercepto difira entre grupos, mantendo uma inclinação comum:

```
lm(yield ~ severity + cultivar, data = cv_data)
```

```

Call:
```

```
lm(formula = yield ~ severity + cultivar, data = cv_data)

Coefficients:
      (Intercept)           severity  cultivarTolerante
          3863.15             -29.54             611.85
```

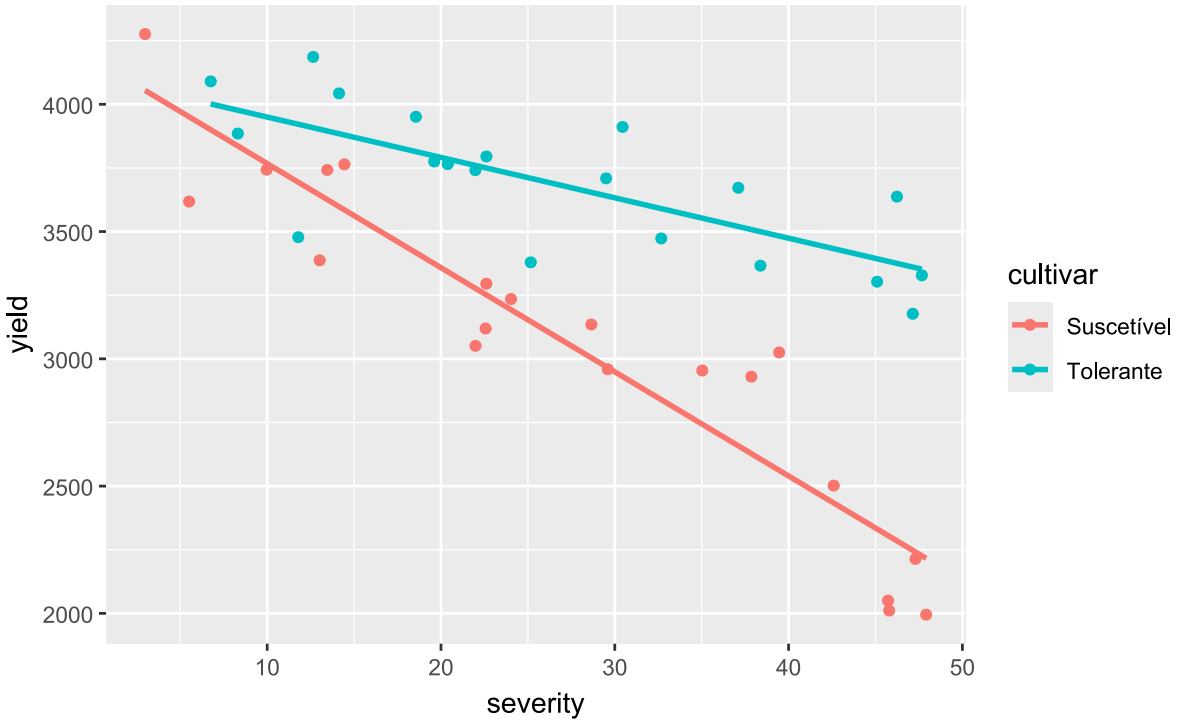
Acrescentar a interação permite que cada grupo tenha sua própria inclinação:

```
model_int ← lm(yield ~ severity * cultivar, data = cv_data)
```

8.12.2 Inclinações diferentes são uma interação

A interação entre uma variável quantitativa e um fator tem o mesmo significado das interações apresentadas no Capítulo 7: o efeito da severidade sobre a produtividade **depende** da cultivar.

```
ggplot(cv_data, aes(x = severity, y = yield, color = cultivar)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



8.12.3 Estimando e comparando inclinações

As inclinações estimadas para cada grupo, e as diferenças entre elas, podem ser obtidas diretamente:

```
library(emmeans)

slopes ← emtrends(model_int, ~ cultivar, var = "severity")

slopes
```

```
 cultivar    severity.trend   SE df lower.CL upper.CL
 Suscetível          -40.9 3.32 36    -47.6   -34.18
 Tolerante           -15.9 3.64 36    -23.2    -8.49

Confidence level used: 0.95
```

```
pairs(slopes)
```

```
 contrast               estimate   SE df t.ratio p.value
 Suscetível - Tolerante      -25 4.92 36  -5.086 <0.0001
```

É o mesmo fluxo de trabalho baseado em estimação usado para as médias de grupos no Capítulo 7, agora aplicado a taxas de variação.

> **Dica**
>
> Quando os grupos diferem em suas inclinações, comparar os grupos em um único valor de x depende de qual valor é escolhido. Compare as próprias inclinações, ou compare os grupos em valores de x biologicamente relevantes.

8.13 Regressão linear múltipla

8.13.1 Mais de uma variável explicativa

Muitas respostas biológicas dependem de várias variáveis quantitativas simultaneamente. A infecção, por exemplo, frequentemente depende tanto da temperatura quanto da duração do molhamento foliar (Carisse et al. 2020).

Um modelo de regressão linear múltipla com duas variáveis explicativas pode ser escrito como:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

```
model_clima ← lm(infection ~ temperature + wetness, data = clima)
summary(model_clima)
```

```

Call:
lm(formula = infection ~ temperature + wetness, data = clima)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0148 -1.9462 -0.2487  1.8679  7.9381

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.41522    1.70318  -6.115 9.31e-08 ***
temperature   0.87322    0.08826   9.893 5.57e-14 ***
wetness       1.11869    0.11945   9.365 3.93e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.78 on 57 degrees of freedom
```

```
Multiple R-squared:  0.8505,    Adjusted R-squared:  0.8452
F-statistic: 162.1 on 2 and 57 DF,  p-value: < 2.2e-16
```

8.13.2 Efeitos parciais

Em um modelo de regressão múltipla, cada coeficiente descreve a variação esperada da resposta para uma mudança de uma unidade naquela variável **mantendo constantes as demais variáveis do modelo**.

É por isso que um coeficiente pode mudar, às vezes substancialmente, quando outra variável é acrescentada ao modelo ou removida dele. O coeficiente da temperatura em um modelo que contém o molhamento responde a uma pergunta diferente do coeficiente da temperatura sozinha.

8.13.3 Interações entre variáveis quantitativas

O efeito de uma variável quantitativa pode depender do valor de outra. Por exemplo, o aumento da infecção por hora adicional de molhamento pode ser maior em temperaturas favoráveis:

```
lm(infection ~ temperature * wetness, data = clima)
```

```

Call:
lm(formula = infection ~ temperature * wetness, data = clima)

Coefficients:
        (Intercept)          temperature              wetness
          -9.265518             0.818670             0.997346
temperature:wetness
           0.005603
```

Com uma interação, o coeficiente do molhamento descreve seu efeito apenas quando a temperatura é igual a zero. Centralizar as variáveis explicativas torna os coeficientes dos efeitos principais interpretáveis em valores típicos.

8.13.4 Colinearidade

Quando as variáveis explicativas estão fortemente correlacionadas entre si, o modelo não consegue separar suas contribuições individuais. Os coeficientes tornam-se imprecisos e instáveis, e seus sinais podem até parecer biologicamente implausíveis.

A colinearidade é comum com variáveis meteorológicas: temperatura, umidade relativa, déficit de pressão de vapor e molhamento foliar frequentemente variam juntos.

```
cor(clima[, c("temperature", "humidity", "wetness")])
```

```
            temperature  humidity   wetness
temperature   1.0000000 0.3085412 0.4277422
humidity      0.3085412 1.0000000 0.7570720
wetness       0.4277422 0.7570720 1.0000000
```

A colinearidade é principalmente uma propriedade dos dados e do delineamento do estudo, e não algo que um teste estatístico possa corrigir. Experimentos que variam os fatores de forma independente a evitam; estudos observacionais precisam reconhecê-la.

8.13.5 Escolhendo as variáveis explicativas

As variáveis incluídas em um modelo devem decorrer da questão biológica e do conhecimento do sistema, especificadas antes da análise sempre que possível.

Procedimentos automáticos que acrescentam e removem variáveis com base em valores-P (seleção *stepwise*) tendem a produzir modelos que se ajustam bem ao conjunto de dados atual, mas generalizam mal, com precisão superestimada. São especialmente problemáticos quando há muitas variáveis meteorológicas candidatas e poucas observações.

> **! Importante**
>
> Modelos construídos para **explicar** um processo biológico e modelos construídos para **prever** um desfecho têm objetivos diferentes. Para explicar, escolha as variáveis com base no raciocínio biológico; para prever, avalie o desempenho em dados que não foram usados para ajustar o modelo.

8.14 Dados de vários experimentos ou locais

As relações são frequentemente estimadas combinando dados de várias lavouras, anos ou experimentos repetidos, cada um contribuindo com várias observações. Observações do mesmo experimento compartilham condições e não são independentes.

Reunir todas as observações em uma única regressão ignora essa estrutura. Isso pode superestimar a precisão e confundir a relação **dentro** dos experimentos com as diferenças **entre** experimentos.

Um primeiro passo útil é ajustar e representar graficamente a relação separadamente para cada experimento. Modelos que estimam uma relação geral levando em conta a variação entre experimentos, como os modelos mistos com interceptos e inclinações aleatórios, são o arcabouço adequado para esses dados.

8.15 Regressão ou correlação?

Tabela 8.1. Escolhendo entre correlação e regressão.

Questão	Ferramenta
Quão fortemente duas variáveis estão linearmente associadas?	Correlação
Quanto a resposta muda por unidade da variável explicativa?	Inclinação da regressão
Que resposta se espera em determinado valor de x?	Predição pela regressão
Dois métodos de medida concordam?	Análise de concordância (não correlação)
A relação difere entre grupos?	Regressão com interação
Quais são as contribuições separadas de várias variáveis?	Regressão múltipla

Na maioria dos estudos em Fitopatologia, a questão envolve uma taxa de variação, o que torna a regressão a ferramenta mais informativa.

8.16 Pesquisa na prática: estimando a perda de produtividade causada pela doença

Pesquisa revisitada: Quanto de produtividade a doença retira?

Considere um conjunto de ensaios de campo em que a severidade da doença foi medida em um estádio-chave da cultura e a produtividade foi medida na colheita, com a severidade variando entre parcelas por causa de diferentes programas de fungicida.

A questão científica não é simplesmente:

> A produtividade está correlacionada com a severidade da doença?

Em vez disso, pergunte:

1. Quantos quilogramas por hectare se perdem por ponto percentual de severidade?
2. Qual é a produtividade atingível na ausência da doença?
3. Qual é a perda relativa, expressa como porcentagem da produtividade atingível?
4. Quão incertas são essas estimativas?
5. A relação é linear ao longo da faixa de severidade observada?
6. A inclinação difere entre cultivares, locais ou anos?
7. Em que faixa de severidade a relação pode ser aplicada?

As respostas a essas perguntas são o que permite que uma relação de perda de produtividade oriente decisões econômicas sobre o manejo da doença.

8.17 Um fluxo de trabalho prático para estudar relações

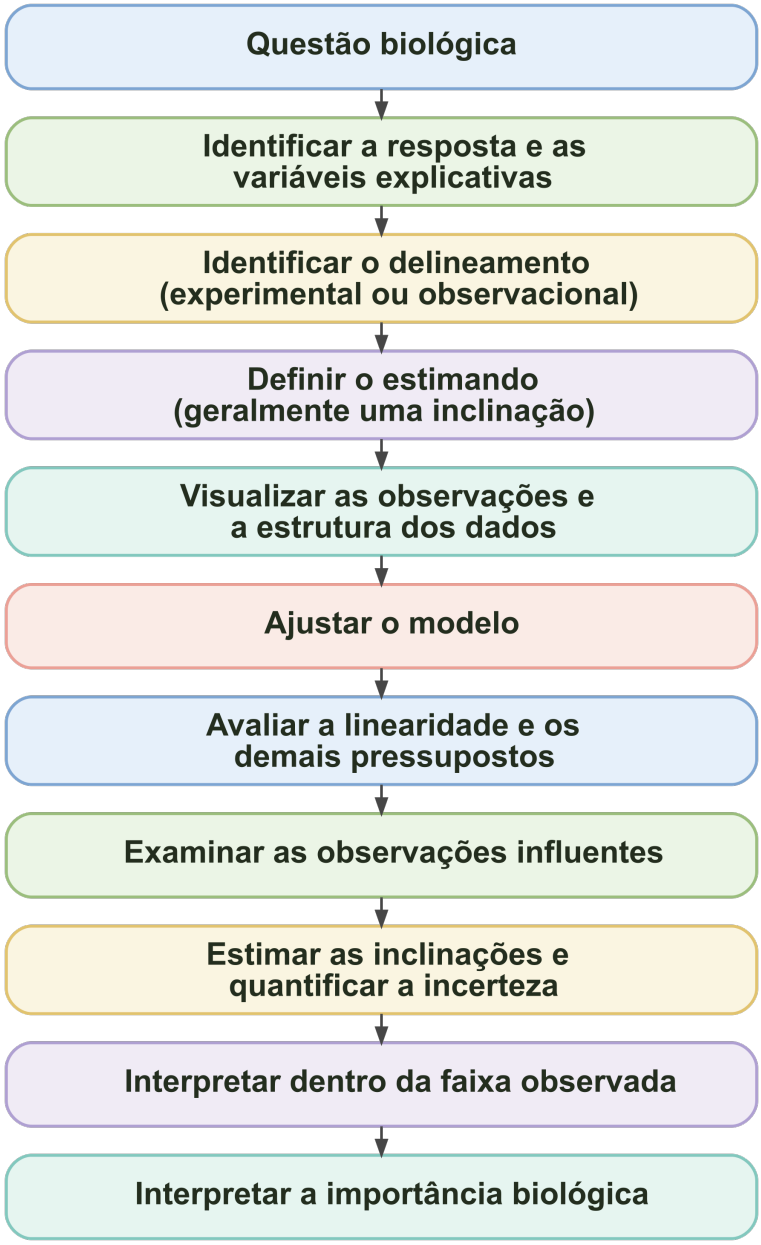


Figura 8.2. Fluxo de trabalho para estudar relações entre variáveis.

8.18 Relatando relações

Um bom enunciado de resultados geralmente inclui:

8.18.1 Que relação foi estudada?

Nomeie a resposta, a variável explicativa e suas unidades.

8.18.2 Direção e magnitude

Relate a inclinação estimada em unidades biológicas.

8.18.3 Incerteza

Relate um intervalo de confiança para a inclinação.

8.18.4 Amplitude dos dados

Informe a faixa da variável explicativa em que a relação foi estimada.

8.18.5 Ajuste e adequação

Relate o R^2 ou a variabilidade residual como descrição da dispersão, junto com uma afirmação sobre a adequação do modelo linear.

8.18.6 Interpretação biológica

Explique o que a taxa de variação estimada significa para o sistema.

Em vez de:

> A produtividade foi significativamente correlacionada com a severidade (r = −0,92, P < 0,001).

prefira algo como:

> A produtividade diminuiu 36,2 kg/ha (IC 95%: 31,2 a 41,2 kg/ha) para cada ponto percentual de severidade entre 0 e 60%, a partir de uma produtividade atingível estimada de 4.184 kg/ha, o que corresponde a uma perda de 0,87% da produtividade atingível por ponto percentual de severidade.

Esses números vêm do exemplo simulado da seção “Um exemplo simples” (saídas de `summary(model)` e `confint(model)`).

8.19 Erros comuns

8.19.1 Ajustar uma reta sem olhar o gráfico de dispersão

Curvaturas, valores discrepantes e agrupamentos são invisíveis nas estatísticas-resumo.

8.19.2 Relatar apenas r ou R^2

Nenhuma dessas grandezas diz ao leitor quanto a resposta muda.

8.19.3 Interpretar correlação como causalidade

Associações observacionais podem refletir variáveis de confundimento.

8.19.4 Usar a correlação para avaliar a concordância entre métodos

Métodos perfeitamente correlacionados podem discordar sistematicamente.

8.19.5 Relatar inclinações sem unidades

Uma inclinação só tem significado com as unidades das duas variáveis.

8.19.6 Interpretar um intercepto muito fora dos dados

O intercepto pode corresponder a condições impossíveis.

8.19.7 Confundir intervalos de confiança e de predição

A faixa de confiança da média é muito mais estreita do que a incerteza para observações individuais.

8.19.8 Extrapolar além dos dados observados

A relação fora da faixa observada é desconhecida.

8.19.9 Ignorar observações influentes

Um único ponto extremo pode determinar a inclinação.

8.19.10 Reunir dados de vários experimentos como observações independentes

Isso superestima a precisão e pode misturar relações dentro e entre experimentos.

8.19.11 Interpretar coeficientes de regressão múltipla como efeitos simples

Cada coeficiente é condicional às demais variáveis do modelo.

8.19.12 Selecionar variáveis automaticamente

Procedimentos *stepwise* produzem modelos otimistas e instáveis.

8.19.13 Forçar uma reta sobre uma relação curva

Muitas respostas biológicas são não lineares (Capítulo 9).

8.20 Lista de verificação para a análise de relações

Antes de interpretar uma relação entre variáveis, pergunte:

1. Que questão biológica está sendo abordada?
2. Qual variável é a resposta e qual é a explicativa?
3. A variável explicativa foi controlada experimentalmente ou observada?
4. Quais são as unidades das duas variáveis?
5. O que o gráfico de dispersão mostra?
6. A relação é aproximadamente linear na faixa observada?
7. As observações são independentes, ou vêm de grupos, locais ou medidas repetidas?
8. Há observações influentes?
9. Qual é a inclinação estimada e o que ela significa biologicamente?
10. Quão incerta é a inclinação?
11. O intercepto tem significado?
12. O objetivo é estimar uma resposta média ou prever novas observações?
13. A relação difere entre grupos?
14. As variáveis explicativas estão correlacionadas entre si?
15. A conclusão permanece dentro da faixa e do escopo dos dados?

8.21 Resumo do capítulo

Muitas questões em Fitopatologia dizem respeito a como uma variável quantitativa muda com outra: como a produtividade diminui com a doença, como as lesões crescem com a temperatura ou como a infecção aumenta com o molhamento foliar. A correlação descreve a intensidade de

uma associação linear, mas não quantifica a taxa de variação, não estabelece causalidade nem demonstra concordância entre métodos de medida.

A regressão linear estende o arcabouço do modelo linear do Capítulo 7 a variáveis explicativas quantitativas. Sua estimativa central, a inclinação, expressa a variação esperada da resposta por unidade da variável explicativa e deve ser relatada em unidades biológicas, com sua incerteza. Uma interpretação confiável exige examinar os dados, verificar a linearidade e as observações influentes, distinguir intervalos de confiança de intervalos de predição, respeitar a faixa dos dados e levar em conta a estrutura das observações. Combinar preditores quantitativos e categóricos permite comparar relações entre grupos, e a regressão múltipla permite considerar várias variáveis explicativas simultaneamente, desde que os coeficientes sejam interpretados como efeitos parciais e as variáveis sejam escolhidas com base no raciocínio biológico.

8.22 Conceitos-chave

Ao concluir este capítulo, você deverá ser capaz de explicar:

- variável resposta;
- variável explicativa;
- gráfico de dispersão;
- associação;
- confundimento;
- correlação de Pearson;
- correlação de Spearman;
- concordância;
- linha de igualdade;
- regressão linear simples;
- intercepto;
- inclinação;
- mínimos quadrados;
- resíduo;
- centralização;
- intervalo de confiança da resposta média;
- intervalo de predição;
- coeficiente de determinação;
- linearidade;
- alavancagem;
- observação influente;
- extrapolação;
- interação entre uma variável quantitativa e uma categórica;
- regressão linear múltipla;
- efeito parcial;
- colinearidade.

8.23 Exercícios

8.23.1 Exercício 1. Qual é a relação?

Para cada questão biológica, identifique:

1. a variável resposta e suas unidades;
2. a variável explicativa e suas unidades;
3. se a variável explicativa é controlada ou observada;
4. o estimando de interesse;
5. a faixa da variável explicativa em que as conclusões se aplicam.

Use questões como:

- Quanto a produtividade diminui à medida que a severidade da doença aumenta?
- O tamanho da lesão aumenta com a temperatura?
- A densidade de esporos no ar está associada à incidência posterior da doença?

8.23.2 Exercício 2. Olhe antes de ajustar

Usando o quarteto de Anscombe (`anscombe` no R):

1. calcule a média, o desvio padrão e a correlação de cada conjunto de dados;
2. ajuste uma regressão linear a cada conjunto;
3. represente cada conjunto com sua reta ajustada;
4. explique quais conjuntos são adequadamente descritos por uma reta.

8.23.3 Exercício 3. Perda de produtividade

Usando um conjunto de dados de produtividade e severidade fornecido pelo professor:

1. represente os dados graficamente;
2. ajuste uma regressão linear simples;
3. interprete o intercepto e a inclinação em unidades biológicas;
4. calcule a perda relativa de produtividade por ponto percentual de severidade;
5. calcule intervalos de confiança para a inclinação;
6. examine os gráficos de diagnóstico e identifique observações influentes.

8.23.4 Exercício 4. Confiança ou predição?

Usando o modelo do Exercício 3, calcule um intervalo de confiança e um intervalo de predição para a produtividade com 10% e com 40% de severidade. Explique qual intervalo um produtor deveria usar para antecipar a produtividade de uma lavoura específica, e por quê.

8.23.5 Exercício 5. Concordância entre métodos

Dois avaliadores estimaram a severidade da doença no mesmo conjunto de folhas.

1. calcule a correlação entre os avaliadores;
2. represente suas estimativas junto com a linha de igualdade;
3. determine se um avaliador superestima ou subestima sistematicamente em relação ao outro;
4. explique por que uma correlação alta não é evidência suficiente de concordância.

8.23.6 Exercício 6. Inclinações diferentes

A produtividade e a severidade foram medidas em duas cultivares.

1. represente os dados por cultivar;
2. ajuste um modelo que permita inclinações diferentes;
3. estime a inclinação de cada cultivar e a diferença entre elas;
4. interprete o resultado biologicamente. Uma cultivar é mais tolerante à doença?

8.23.7 Exercício 7. Avaliação de um artigo publicado

Escolha um artigo recente de Fitopatologia que relate uma correlação ou regressão.

Identifique:

1. a resposta e as variáveis explicativas;
2. se o estudo foi experimental ou observacional;
3. se os dados foram mostrados em um gráfico de dispersão;
4. se foi relatada uma inclinação com unidades;
5. se foi relatada a incerteza da inclinação;
6. se os autores se basearam principalmente em r, R^2 ou valores-P;
7. se a relação foi aplicada além da faixa dos dados;
8. se a linguagem causal era justificada pelo delineamento.

8.24 Reflexão

> **Uma correlação nos diz que duas variáveis se movem juntas. Uma inclinação nos diz em quanto.**

Discuta essa afirmação usando um exemplo da Fitopatologia.

8.25 Próximos passos

Retas são aproximações úteis, mas muitas respostas biológicas não são lineares. O crescimento de um fungo pode atingir o máximo em uma temperatura ótima, a inibição de um patógeno pode aumentar com a dose do fungicida segundo uma curva em forma de S, e a doença pode aumentar rapidamente antes de atingir um platô. O Capítulo 9 apresenta modelos não lineares cujos parâmetros, como a concentração efetiva que inibe o crescimento em 50% (CE_{50}), têm significado biológico direto.

CAPÍTULO 9

Como analisamos respostas não lineares?

9.1 Objetivos de aprendizagem

! Importante

Ao concluir este capítulo, o leitor deverá ser capaz de:

- reconhecer respostas biológicas intrinsecamente não lineares;
- explicar por que retas e polinômios costumam descrever mal essas respostas;
- descrever formas comuns de curvas não lineares e o significado biológico de seus parâmetros;
- ajustar e interpretar modelos de dose–resposta e estimar valores de CE_{50};
- distinguir concentrações efetivas absolutas e relativas;
- comparar curvas de dose–resposta e valores de CE_{50} entre isolados ou tratamentos;
- ajustar modelos que descrevem respostas à temperatura e a outras variáveis ambientais;
- compreender como os modelos não lineares são ajustados, incluindo o papel dos valores iniciais;
- quantificar a incerteza dos parâmetros do modelo e de grandezas derivadas;
- escolher entre modelos não lineares concorrentes usando critérios biológicos e estatísticos;
- delinear experimentos que permitam estimar com precisão os parâmetros não lineares;
- comunicar análises não lineares usando estimativas dos parâmetros, incerteza e curvas ajustadas.

9.2 Por que respostas não lineares?

O Capítulo 8 descreveu relações em que a resposta muda a uma taxa constante com a variável explicativa. Muitas respostas biológicas não se comportam assim. Elas sobem e depois se estabilizam, decrescem em direção a um piso, seguem uma curva em forma de S ou atingem um máximo em um valor intermediário.

Exemplos típicos em Fitopatologia incluem:

- **sensibilidade a fungicidas**: o crescimento micelial ou a germinação de esporos diminui com a concentração do fungicida, devagar no início, depois rapidamente, e então se aproxima de um mínimo;

- **respostas à temperatura**: o crescimento, a germinação ou a infecção aumentam com a temperatura até um ótimo e diminuem além dele;
- **densidade de inóculo**: a infecção aumenta com a concentração de inóculo, mas se satura quando a maioria dos sítios suscetíveis já está infectada;
- **duração do molhamento foliar**: a proporção de folhas infectadas aumenta com a duração do molhamento e se aproxima de um platô;
- **progresso da doença**: a doença aumenta devagar, depois rapidamente, e então se aproxima de um máximo ao longo da safra.

Em cada caso, a forma da resposta não é um incômodo a ser eliminado por uma transformação. **A forma é a biologia.** A curva nos diz em que concentração um patógeno é inibido, em que temperatura cresce melhor ou quanto inóculo é necessário para causar doença.

> **! Importante**
>
> O objetivo da modelagem não linear não é simplesmente desenhar uma curva sobre os dados. É estimar **parâmetros com significado biológico**, como uma concentração efetiva, uma temperatura ótima ou uma resposta máxima, e quantificar sua incerteza.

9.3 Da questão biológica ao parâmetro da curva

Considere um ensaio de laboratório em que o crescimento micelial de um isolado fúngico é medido em meio de cultura com concentrações crescentes de fungicida.

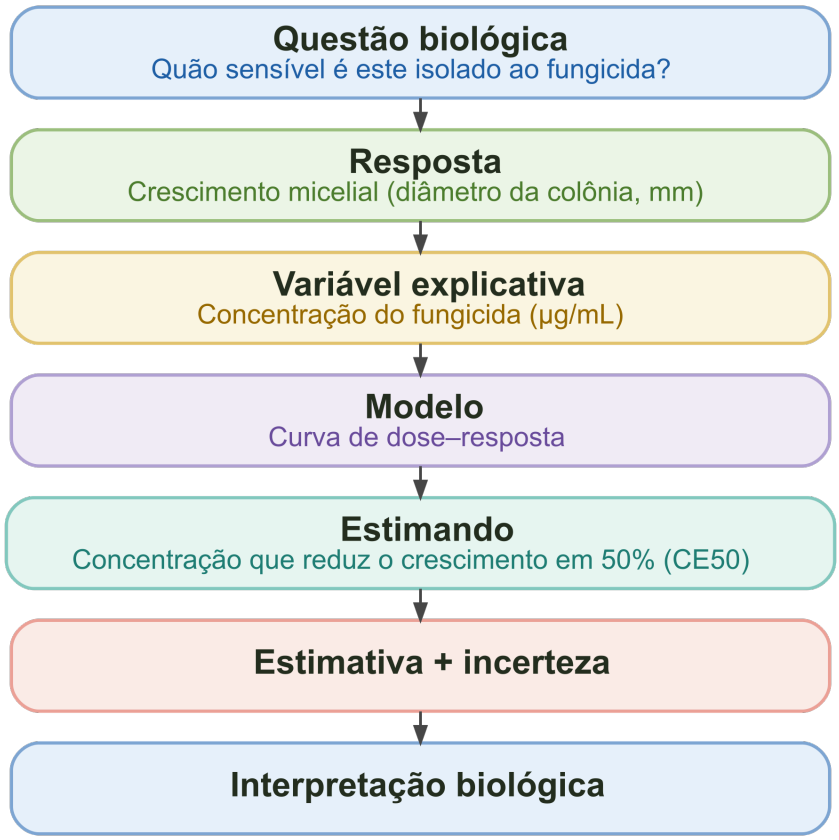


Figura 9.1. Da questão biológica à estimação de um parâmetro de curva (CE_{50}).

Aqui o estimando não é uma diferença entre médias (Capítulo 7) nem uma inclinação (Capítulo 8). É um **parâmetro de uma curva**: a concentração em que a resposta atinge a metade de sua amplitude.

9.4 Por que não uma reta ou um polinômio?

9.4.1 Retas

Uma reta ajustada a uma resposta curva a representa mal em todos os pontos: superestima a resposta em algumas regiões e a subestima em outras, e sua inclinação é a média de uma taxa de variação que não é constante. Os gráficos de resíduos revelam o problema como um padrão curvo sistemático (Capítulo 8).

9.4.2 Transformações

Algumas relações curvas podem ser retificadas. Representar a resposta em função do logaritmo da concentração, por exemplo, frequentemente torna aproximadamente linear a parte central de uma curva de dose–resposta. As análises clássicas de sensibilidade a fungicidas faziam a regressão da inibição transformada em probit ou logit sobre o logaritmo da concentração.

A linearização tem limitações. Geralmente descreve apenas parte da curva, altera a estrutura de erro dos dados e pode exigir o descarte de observações nos extremos, como respostas de 0% ou 100%.

9.4.3 Polinômios

Polinômios quadráticos ou cúbicos podem acompanhar a curvatura dentro da faixa observada, mas:

- seus parâmetros não têm interpretação biológica direta;
- não conseguem representar assíntotas e, por isso, preveem valores impossíveis (crescimento negativo, inibição acima de 100%) fora dos dados;
- frequentemente oscilam entre as observações.

> **Dica**
>
> Um polinômio pode ser uma descrição exploratória razoável. Quando a questão biológica diz respeito a uma dose efetiva, a um ótimo ou a um máximo, prefira um modelo em que essa grandeza seja um parâmetro.

9.5 A anatomia de um modelo não linear

Um modelo de regressão não linear pode ser escrito como:

$$y_i = f(x_i, \boldsymbol{\theta}) + \varepsilon_i$$

em que f é uma função escolhida para descrever a forma da resposta, $\boldsymbol{\theta}$ é um conjunto de parâmetros e ε_i representa a variação residual.

O modelo é *não linear* porque a resposta não depende linearmente dos parâmetros. Um polinômio quadrático é curvo, mas ainda é linear em seus parâmetros e pode ser ajustado com `lm()`; uma curva logística, não.

9.5.1 Parâmetros com significado biológico

A maioria das curvas usadas em Fitopatologia é construída a partir de um pequeno conjunto de características interpretáveis:

- **assíntota superior**: a resposta máxima (por exemplo, o crescimento na ausência de fungicida);
- **assíntota inferior**: a resposta mínima (por exemplo, o crescimento residual em concentrações muito altas);
- **posição**: onde, ao longo do eixo x, a resposta muda (por exemplo, a CE_{50} ou o ponto de inflexão);
- **inclinação**: quão abruptamente a resposta muda em torno dessa posição;
- **ótimo**: o valor de x em que a resposta é máxima;
- **limiares**: valores mínimo e máximo de x além dos quais não há resposta.

9.6 Formas comuns de curvas

9.6.1 Declínio exponencial

A resposta diminui a uma taxa proporcional ao seu valor atual, aproximando-se de zero. Exemplos incluem a sobrevivência de propágulos do patógeno ao longo do tempo ou o declínio de resíduos de fungicida.

$$y = a\, e^{-bx}$$

9.6.2 Curva monomolecular (de saturação)

A resposta aumenta rapidamente no início e depois se estabiliza em direção a um máximo, sem ponto de inflexão. Exemplos incluem a infecção em função da concentração de inóculo ou da duração do molhamento.

$$y = a\left(1 - e^{-bx}\right)$$

9.6.3 Curvas logística e de Gompertz

Curvas em forma de S, simétricas (logística) ou assimétricas (Gompertz), que sobem de uma assíntota inferior a uma superior. São muito usadas para descrever o progresso da doença ao longo do tempo e as respostas à duração do molhamento.

$$y = \frac{a}{1 + e^{-b(x-m)}}$$

O ajuste detalhado de modelos epidemiológicos está além do escopo deste livro.

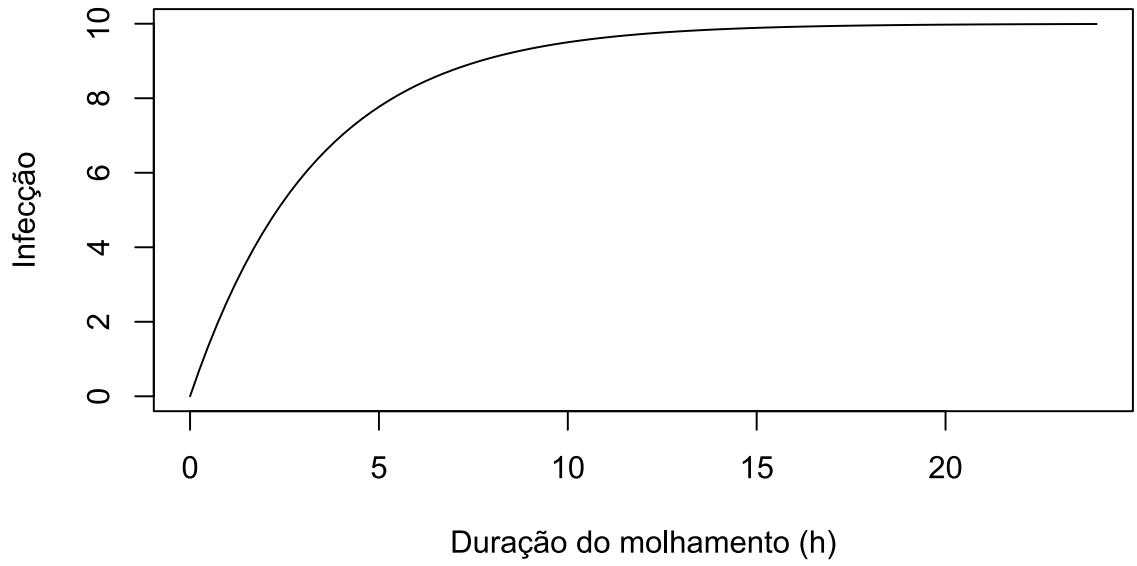
9.6.4 Curvas log-logísticas de dose–resposta

Curvas em forma de S no logaritmo da dose, decrescendo de uma assíntota superior a uma inferior. São os modelos-padrão para a sensibilidade a fungicidas e são descritas em detalhe a seguir.

9.6.5 Curvas com um ótimo

Curvas que sobem até um máximo e depois decrescem, usadas para descrever respostas à temperatura. A função beta e modelos relacionados são parametrizados diretamente em termos das temperaturas mínima, ótima e máxima.

```r
library(ggplot2)

# Explorar formas candidatas antes do ajuste
curve(10 * (1 - exp(-0.3 * x)), from = 0, to = 24,
      xlab = "Duração do molhamento (h)", ylab = "Infecção")
```



Dica

Escolha uma família de curvas com base no que se sabe sobre o processo biológico e no gráfico de dispersão, e não experimentando todas as funções disponíveis.

9.7 Análise de dose–resposta

9.7.1 Por que experimentos de dose–resposta?

Os experimentos de dose–resposta são centrais no monitoramento da resistência a fungicidas, na avaliação de novos ingredientes ativos, em estudos de agentes de controle biológico e em ensaios de extratos vegetais ou outros compostos antimicrobianos. Eles respondem a perguntas como:

- Quão sensível é uma população do patógeno a um fungicida?
- Alguns isolados são menos sensíveis que uma população de referência (linha de base)?
- A sensibilidade está mudando ao longo do tempo?
- Qual de dois compostos é mais potente?

9.7.2 O modelo log-logístico

O modelo log-logístico de quatro parâmetros é:

$$f(x) = c + \frac{d - c}{1 + \exp\{b\,[\log(x) - \log(e)]\}}$$

em que:

- d é a assíntota superior, a resposta na ausência do composto;
- c é a assíntota inferior, a resposta em concentrações muito altas;
- e é a concentração que produz uma resposta a meio caminho entre c e d, a **CE_{50}**;
- b descreve a inclinação da curva em torno da CE_{50}.

Quando se sabe que a resposta se aproxima de zero em altas concentrações, a assíntota inferior pode ser fixada em zero, resultando no modelo de três parâmetros.

9.7.3 Um exemplo simples

O código a seguir simula um ensaio de crescimento micelial com seis concentrações e quatro placas por concentração, além da testemunha.

```
set.seed(606)

conc ← rep(c(0, 0.01, 0.03, 0.1, 0.3, 1, 3), each = 4)

growth ← 2 + (60 - 2) / (1 + exp(1.5 * (log(conc) - log(0.15))))
growth[conc == 0] ← 60

data ← data.frame(
  conc = conc,
  growth = growth + rnorm(length(conc), sd = 3)
)
```

9.7.4 Ajustando o modelo no R

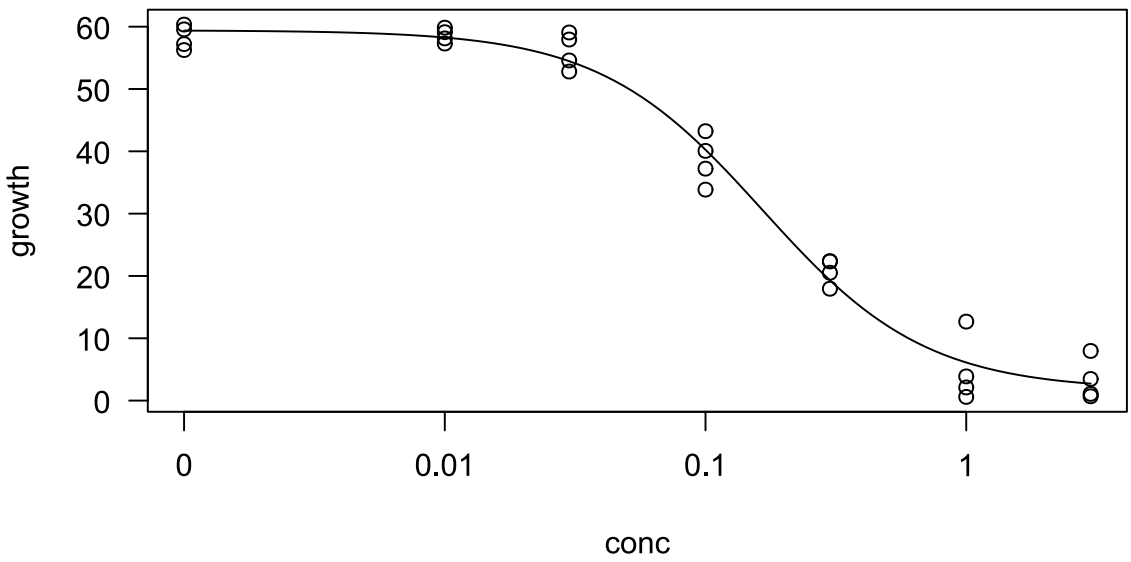
O pacote `drc` fornece modelos de dose–resposta e ferramentas para estimar doses efetivas.

```
library(drc)

model ← drm(growth ~ conc, data = data, fct = LL.4(
  names = c("b", "c", "d", "e")
))

summary(model)
```

```
Model fitted: Log-logistic (ED50 as parameter) (4 parms)

Parameter estimates:

              Estimate Std. Error t-value   p-value
b:(Intercept)  1.38519    0.15130  9.1554 2.675e-09 ***
c:(Intercept)  1.69492    1.76596  0.9598    0.3467
d:(Intercept) 59.41427    1.22573 48.4725 < 2.2e-16 ***
e:(Intercept)  0.16564    0.01608 10.3011 2.745e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error:
```

```
 3.33071 (24 degrees of freedom)
```

```
plot(model, type = "all", log = "x")
```



9.7.5 Estimando a CE_{50}

```
ED(model, 50, interval = "delta")
```

```
Estimated effective doses

       Estimate Std. Error   Lower   Upper
e:1:50  0.16564    0.01608 0.13245 0.19883
```

A CE_{50} é relatada nas unidades de concentração, junto com seu intervalo de confiança. Outras concentrações efetivas, como a CE_{90}, podem ser estimadas da mesma forma quando outro nível de inibição é relevante.

9.7.6 CE_{50} absoluta e relativa

Duas definições de CE_{50} são de uso comum:

- a CE_{50} **relativa** é a concentração que produz uma resposta a meio caminho entre as assíntotas superior e inferior estimadas pelo modelo, que é o parâmetro e acima;
- a CE_{50} **absoluta** é a concentração que reduz a resposta a 50% da testemunha.

As duas só coincidem quando a assíntota inferior é zero. Quando a curva se estabiliza acima de zero, como acontece quando algum crescimento persiste mesmo em altas concentrações, elas podem diferir substancialmente.

```
ED(model, 50, type = "relative")
```

```
Estimated effective doses

       Estimate Std. Error
e:1:50  0.16564    0.01608
```

```
ED(model, 50, type = "absolute")
```

```
Estimated effective doses

       Estimate Std. Error
e:1:50 0.050869   0.007294
```

! Importante

Sempre informe qual definição de CE_{50} foi usada. Valores calculados com definições diferentes não são comparáveis entre estudos.

9.7.7 Expressando a resposta em relação à testemunha

As respostas são frequentemente expressas como porcentagem de inibição em relação à testemunha antes da análise. Isso é conveniente, mas trata a média da testemunha como conhecida sem erro e descarta a informação sobre a variabilidade das testemunhas. Ajustar o modelo às medidas originais, com a assíntota superior estimada a partir dos dados, propaga essa incerteza corretamente. Os resultados ainda podem ser apresentados em escala relativa depois.

9.7.8 A dose zero

A testemunha tem concentração zero, que não pode ser mostrada em um eixo logarítmico. O modelo log-logístico trata corretamente as doses zero durante o ajuste; nas figuras, a testemunha geralmente é mostrada em uma posição pequena e arbitrária do eixo, com uma quebra, e isso deve ser indicado.

9.7.9 Respostas acima da testemunha

Alguns compostos estimulam o crescimento em baixas concentrações antes de inibi-lo em concentrações mais altas, um fenômeno chamado hormese. Os modelos log-logísticos-padrão não conseguem descrever esse padrão, e existem modelos específicos para hormese. Antes de concluir que há hormese, verifique se o aparente estímulo é maior do que a variabilidade entre as repetições da testemunha.

9.8 Comparando curvas de dose–resposta

9.8.1 Vários isolados ou compostos

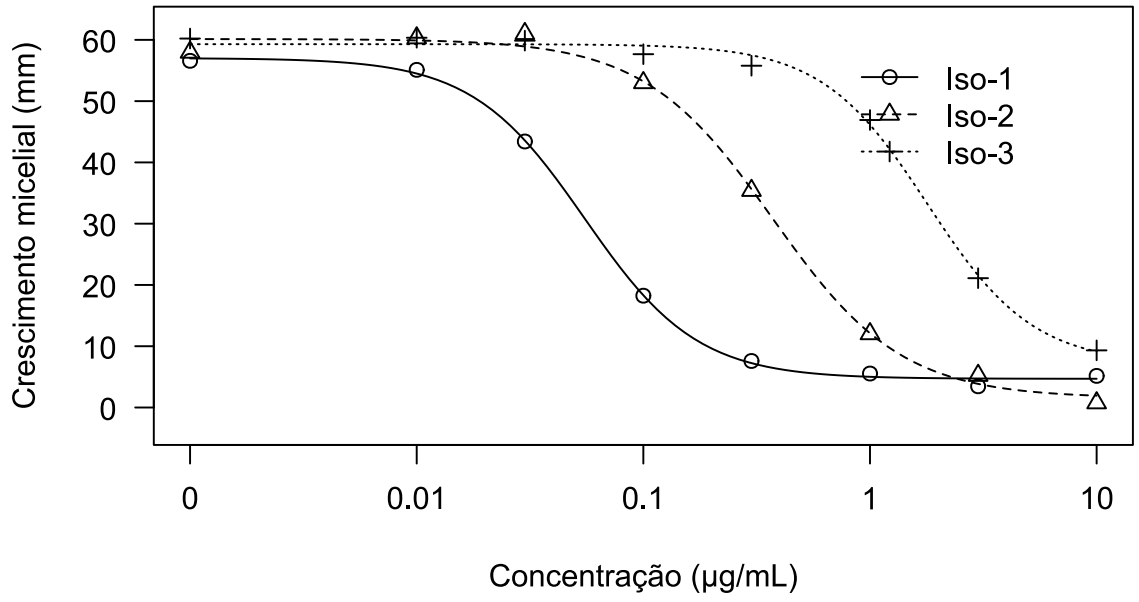
Estudos de sensibilidade geralmente envolvem muitos isolados, ou vários compostos testados no mesmo isolado. As curvas podem ser ajustadas a todos os grupos simultaneamente:

```
model_iso ← drm(
  growth ~ conc,
  curveid = isolate,
  data = data_iso,
  fct = LL.4()
)

ED(model_iso, 50, interval = "delta")
```

```
Estimated effective doses

            Estimate Std. Error     Lower     Upper
e:Iso-1:50 0.0550798  0.0042369 0.0466048 0.0635549
e:Iso-2:50 0.3720363  0.0284418 0.3151442 0.4289284
e:Iso-3:50 1.7942362  0.1538093 1.4865718 2.1019006
```

As curvas ajustadas dos três isolados mostram o deslocamento horizontal que corresponde às diferenças de sensibilidade:

```
plot(model_iso, log = "x", legendPos = c(5, 60),
     xlab = "Concentração (µg/mL)", ylab = "Crescimento micelial (mm)")
```



9.8.2 Comparando valores de CE_{50}

As diferenças de sensibilidade geralmente são expressas como razões, e não como diferenças, porque as concentrações efetivas variam em escala multiplicativa:

```
EDcomp(model_iso, c(50, 50))
```

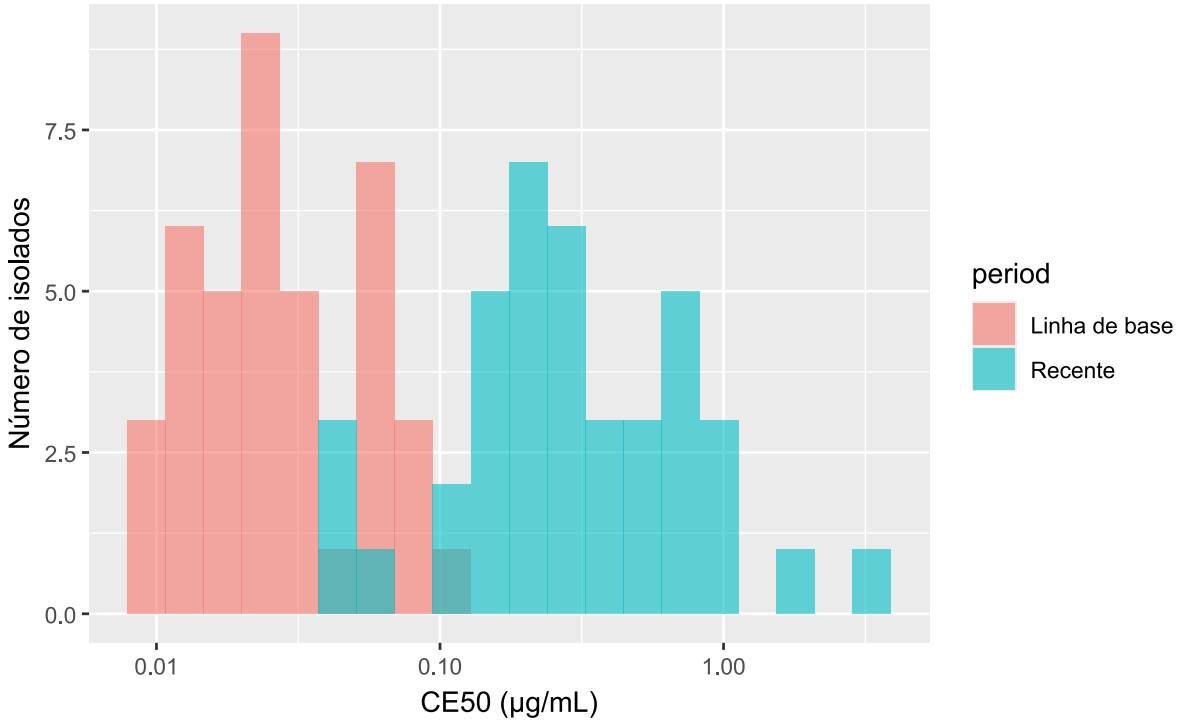
```
Estimated ratios of effect doses

                    Estimate  Std. Error     t-value     p-value
Iso-1/Iso-2:50/50  1.4805e-01  1.6056e-02 -5.3061e+01  1.9851e-52
Iso-1/Iso-3:50/50  3.0698e-02  3.5357e-03 -2.7415e+02  5.8277e-95
Iso-2/Iso-3:50/50  2.0735e-01  2.3817e-02 -3.3281e+01  1.1122e-40
```

O **fator de resistência** de um isolado é sua CE_{50} dividida pela CE_{50} de um isolado sensível de referência ou da população de linha de base. Um fator de resistência de 10 significa que é preciso dez vezes mais fungicida para obter a mesma inibição.

9.8.3 Distribuições de sensibilidade

Em estudos de monitoramento, a CE_{50} de cada isolado torna-se, ela própria, uma observação. A distribuição dos valores de CE_{50} em uma população, geralmente examinada em escala logarítmica, descreve sua sensibilidade. Comparar as distribuições entre uma linha de base coletada antes do uso do fungicida e coletas posteriores revela mudanças de sensibilidade.

```
ggplot(ec50_table, aes(x = ec50, fill = period)) +
  geom_histogram(bins = 20, position = "identity", alpha = 0.6) +
  scale_x_log10() +
  labs(x = "CE50 (µg/mL)", y = "Número de isolados")
```



A mudança de sensibilidade é estimada na escala logarítmica e expressa como razão entre as médias geométricas:

```
tt ← t.test(log10(ec50) ~ period, data = ec50_table)

# quantas vezes a CE50 da coleta recente é maior que a da linha de base
razao  ← 10^(tt$estimate[2] - tt$estimate[1])
```

```
razao_ic ← 10^(-rev(tt$conf.int))
round(c(razao = unname(razao), inferior = razao_ic[1], superior = razao_ic[2]), 1)
```

```
   razao inferior superior
    10.7      7.3     15.6
```

> **⚠ Aviso**
>
> Os valores de CE_{50} são estimativas, cada uma com sua própria incerteza. Quando são analisados depois como dados, os isolados cujas curvas foram mal estimadas devem ser identificados, em vez de tratados como conhecidos com precisão.

9.8.4 Doses discriminatórias

Quando grandes números de isolados precisam ser avaliados, uma curva de dose–resposta completa para cada um pode não ser viável. Uma alternativa comum é medir o crescimento relativo em uma única **dose discriminatória**, escolhida a partir das curvas de dose–resposta de um conjunto representativo de isolados. A escolha dessa dose depende da análise não linear descrita aqui.

9.9 Respostas à temperatura e a outras variáveis ambientais

9.9.1 Curvas com um ótimo

O crescimento do patógeno, a germinação de esporos e a infecção geralmente aumentam com a temperatura até um ótimo e depois diminuem. Polinômios são frequentemente usados para descrever esse padrão, mas modelos parametrizados em termos de temperaturas cardeais são mais informativos:

- T_{min}: a temperatura mínima para o processo;
- T_{opt}: a temperatura ótima;
- T_{max}: a temperatura máxima;
- y_{opt}: a resposta no ótimo.

Uma forma muito usada é a função beta:

$$y = y_{opt}\left(\frac{T - T_{min}}{T_{opt} - T_{min}}\right)^{a}\left(\frac{T_{max} - T}{T_{max} - T_{opt}}\right)^{a\frac{T_{max}-T_{opt}}{T_{opt}-T_{min}}}$$

para $T_{min} < T < T_{max}$, e zero fora desse intervalo.

9.9.2 Ajustando com nls()

Modelos não fornecidos por pacotes especializados podem ser ajustados com `nls()`, que exige a fórmula do modelo e valores iniciais para cada parâmetro:

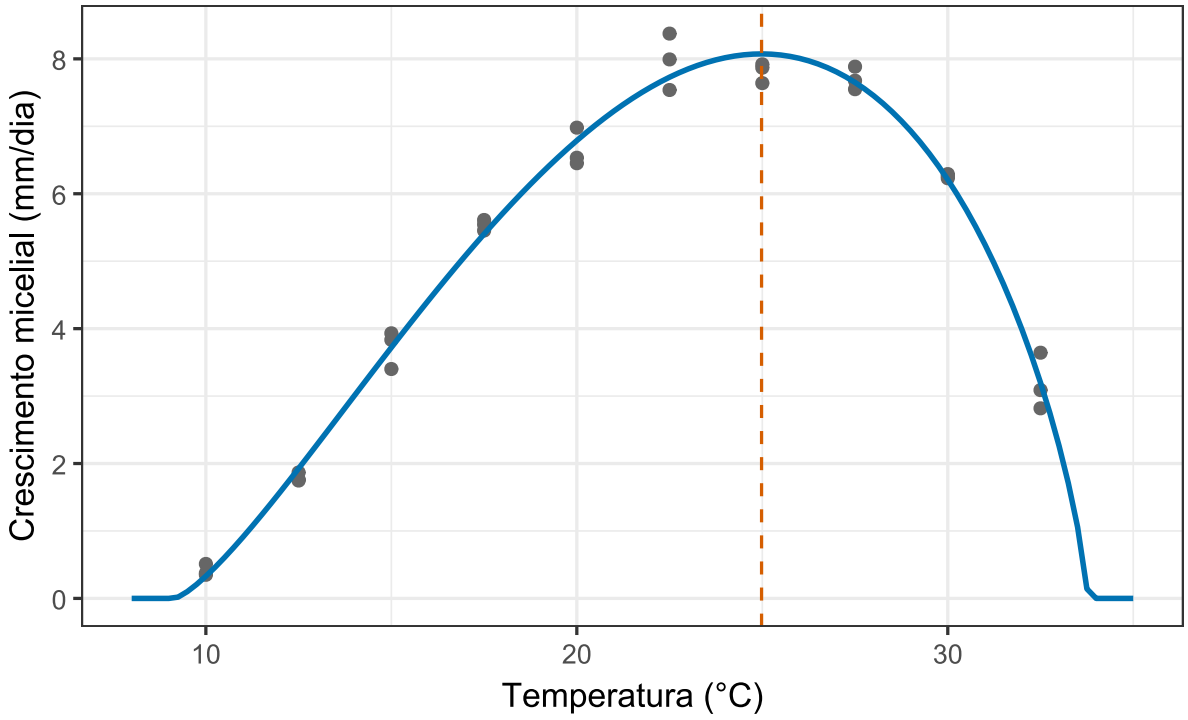
```
beta_fun ← function(T, yopt, Tmin, Topt, Tmax, a) {
  b ← a * (Tmax - Topt) / (Topt - Tmin)
  ifelse(T > Tmin & T < Tmax,
         yopt * ((T - Tmin) / (Topt - Tmin))^a *
                ((Tmax - T) / (Tmax - Topt))^b,
         0)
```

```
}

model_temp ← nls(
  growth ~ beta_fun(temperature, yopt, Tmin, Topt, Tmax, a),
  data = data_temp,
  start = list(yopt = 8, Tmin = 7, Topt = 24, Tmax = 35, a = 1.5)
)

summary(model_temp)
```

```

Formula: growth ~ beta_fun(temperature, yopt, Tmin, Topt, Tmax, a)

Parameters:
     Estimate Std. Error t value Pr(>|t|)
yopt   8.0735     0.0892  90.506  < 2e-16 ***
Tmin   9.1490     0.4655  19.653  < 2e-16 ***
Topt  24.9799     0.1813 137.810  < 2e-16 ***
Tmax  33.7680     0.3316 101.828  < 2e-16 ***
a      1.3493     0.1361   9.917 3.79e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.264 on 25 degrees of freedom

Number of iterations to convergence: 6
Achieved convergence tolerance: 5.396e-07
```

A curva ajustada pode ser desenhada sobre as observações a partir das previsões do modelo:

```
grade_t ← data.frame(temperature = seq(8, 35, by = 0.25))
grade_t$growth ← predict(model_temp, newdata = grade_t)

ggplot(data_temp, aes(temperature, growth)) +
  geom_point(colour = "grey40") +
  geom_line(data = grade_t, colour = "#0072B2", linewidth = 1) +
  geom_vline(xintercept = coef(model_temp)["Topt"], linetype = 2,
             colour = "#D55E00") +
  labs(x = "Temperatura (°C)", y = "Crescimento micelial (mm/dia)") +
  theme_bw(base_size = 13)
```



A linha tracejada marca a temperatura ótima estimada (T_{opt}).

9.9.3 Interpretando as temperaturas cardeais

T_{opt} costuma ser o parâmetro de maior interesse biológico, e também frequentemente o estimado com mais precisão, porque depende de observações ao longo de toda a curva. T_{min} e T_{max} são estimadas extrapolando a curva até zero e são muitas vezes muito incertas, a menos que temperaturas próximas a elas tenham sido incluídas no experimento.

> **⚠ Aviso**
>
> Temperaturas cardeais estimadas além da faixa de temperaturas testadas são extrapolações. Relate seus intervalos de confiança e não as interprete como limiares medidos.

9.10 Como os modelos não lineares são ajustados

9.10.1 Estimação iterativa

Como na regressão linear, a regressão não linear geralmente estima os parâmetros por mínimos quadrados. Diferentemente da regressão linear, não há uma fórmula direta para a solução. O algoritmo parte de valores iniciais e ajusta iterativamente os parâmetros até que a soma dos quadrados dos resíduos pare de diminuir.

9.10.2 Valores iniciais

Bons valores iniciais costumam ser lidos diretamente no gráfico de dispersão:

- a assíntota superior, a partir da resposta nas doses baixas ou na testemunha;
- a assíntota inferior, a partir da resposta nas doses mais altas;
- a CE_{50}, a partir da concentração em que a resposta está aproximadamente na metade;

- um ótimo, a partir da posição do máximo.

Pacotes como o `drc` e os modelos autoiniciáveis do R base (por exemplo, `SSlogis`) calculam os valores iniciais automaticamente.

9.10.3 Quando o ajuste falha

Falhas de convergência e estimativas implausíveis são comuns e informativas. Geralmente indicam que:

- os valores iniciais estavam longe de valores razoáveis;
- os dados não contêm informação suficiente sobre algum parâmetro, por exemplo, nenhuma dose além do ponto de inflexão;
- a curva escolhida é inadequada para os dados;
- o modelo tem mais parâmetros do que os dados conseguem sustentar.

> **Dica**
>
> Quando um modelo não converge, olhe os dados de novo antes de trocar o algoritmo. O problema geralmente está no delineamento ou no modelo, e não no programa.

9.11 Incerteza nos modelos não lineares

9.11.1 Erros padrão e intervalos de confiança

As estimativas dos parâmetros são relatadas com erros padrão e intervalos de confiança, como nos modelos lineares. Nos modelos não lineares, esses intervalos são aproximações. Os intervalos aproximados (de Wald) são simétricos em torno da estimativa; os intervalos de perfil de verossimilhança podem ser assimétricos e costumam ser mais exatos para parâmetros como a CE_{50}.

```
confint(model_temp)   # intervalos de perfil para modelos nls
```

```
          2.5%     97.5%
yopt  7.893417  8.256896
Tmin  7.902474  9.896287
Topt 24.602663 25.364128
Tmax 33.236020 34.694971
a     1.134537  1.711541
```

9.11.2 Grandezas derivadas

Algumas grandezas de interesse não são parâmetros do modelo, mas funções deles, como uma CE_{90}, um fator de resistência ou a temperatura em que o crescimento atinge a metade do máximo. Sua incerteza pode ser obtida pelo método delta ou por *bootstrap*.

```
ED(model, c(10, 50, 90), interval = "delta")
```

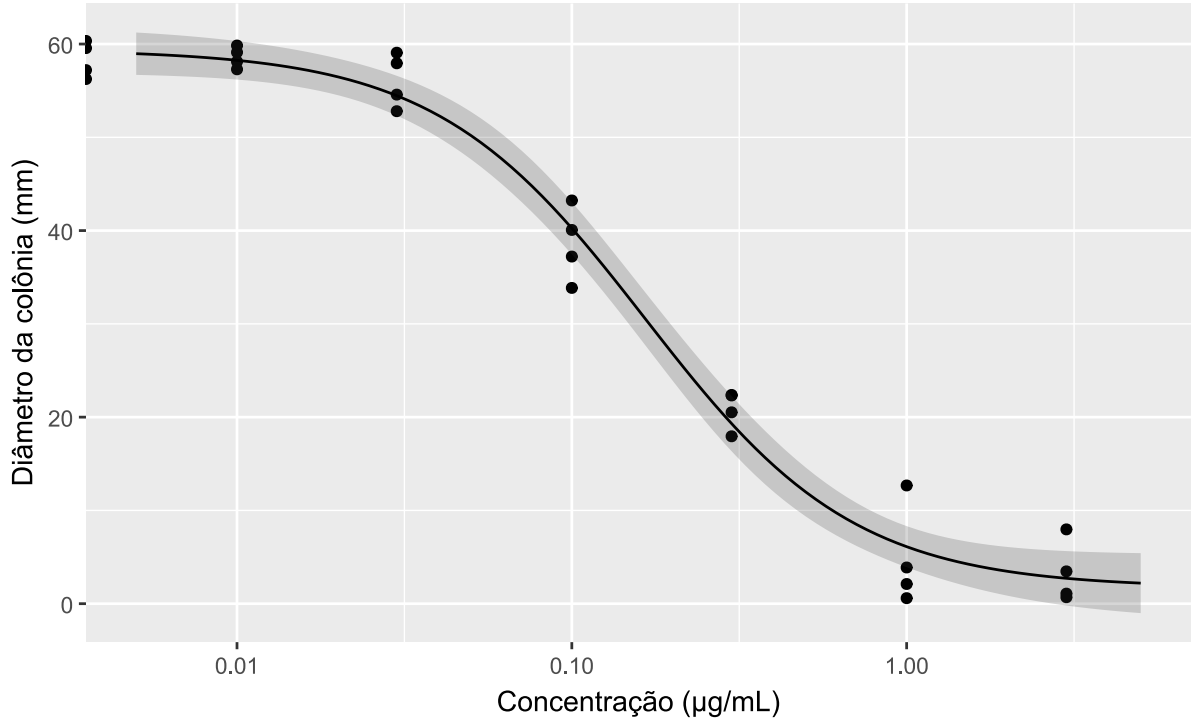
```
Estimated effective doses

        Estimate Std. Error     Lower     Upper
```

```
e:1:10 0.0339059  0.0060713 0.0213754 0.0464364
e:1:50 0.1656387  0.0160796 0.1324520 0.1988254
e:1:90 0.8091859  0.1750881 0.4478218 1.1705500
```

9.11.3 Faixas de confiança para a curva

Representar a curva ajustada junto com uma faixa de confiança e as observações individuais mostra tanto o modelo quanto o quão bem ele descreve os dados.

```
new ← data.frame(conc = exp(seq(log(0.005), log(5), length.out = 100)))

pred ← predict(model, newdata = new, interval = "confidence")

new ← cbind(new, pred)

ggplot(data, aes(x = conc, y = growth)) +
  geom_point() +
  geom_ribbon(data = new,
              aes(x = conc, ymin = Lower, ymax = Upper),
              inherit.aes = FALSE, alpha = 0.2) +
  geom_line(data = new, aes(x = conc, y = Prediction)) +
  scale_x_log10() +
  labs(x = "Concentração (µg/mL)", y = "Diâmetro da colônia (mm)")
```



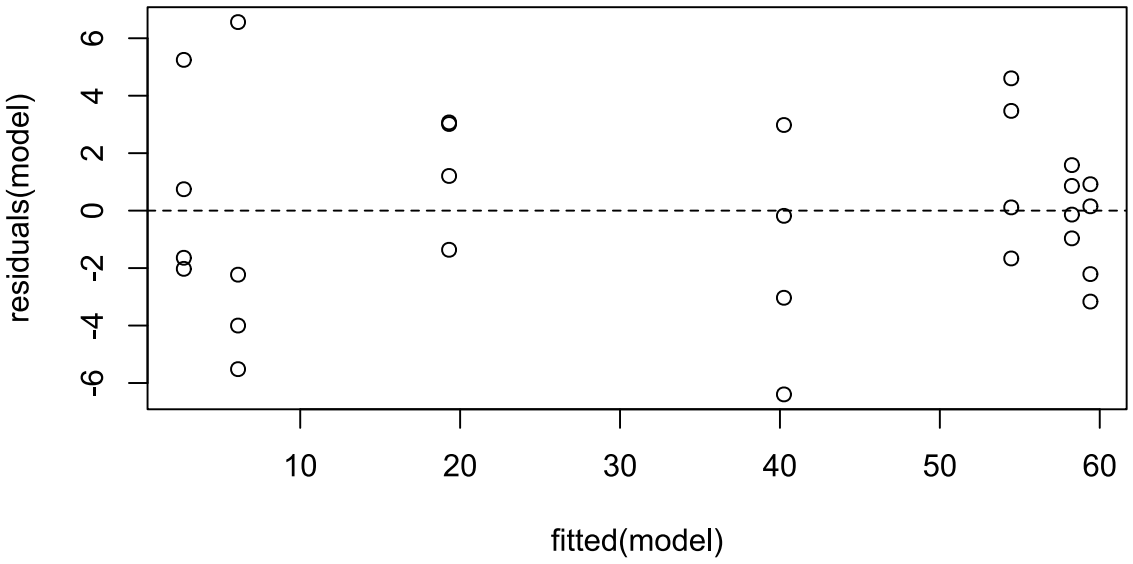
9.12 Pressupostos e diagnóstico do modelo

Os pressupostos da regressão não linear são paralelos aos do modelo linear (Capítulos 7 e 8): a curva descreve corretamente a resposta média, as observações são independentes, a variância residual é razoavelmente constante e os resíduos são aproximadamente simétricos em torno de zero.

9.12.1 Resíduos versus valores ajustados

Um padrão sistemático nos resíduos indica que a curva não acompanha os dados, por exemplo uma assíntota baixa demais ou uma transição abrupta demais.

```
plot(fitted(model), residuals(model))
abline(h = 0, lty = 2)
```



9.12.2 Variância não constante

Em dados de dose–resposta, a variabilidade frequentemente é maior nas concentrações baixas, onde o crescimento é alto, do que nas concentrações altas, onde o crescimento é próximo de zero. O ajuste ponderado ou uma transformação dos dois lados do modelo podem lidar com isso.

9.12.3 Quando a resposta não é contínua

Quando a resposta é uma contagem de esporos germinados em um total conhecido ou o número de folhas infectadas, a variabilidade é determinada pela distribuição binomial ou de Poisson, e modelos próprios para esses dados são mais adequados. Eles são apresentados no Capítulo 10; modelos de dose–resposta para esse tipo de resposta estão disponíveis nos mesmos pacotes.

9.13 Escolhendo entre modelos

Várias famílias de curvas podem descrever os mesmos dados. A escolha entre elas combina critérios biológicos e estatísticos:

1. O modelo representa a biologia conhecida (assíntotas, presença de um ótimo, simetria)?
2. Os parâmetros necessários para a questão estão presentes e são estimáveis?
3. Os resíduos mostram padrões sistemáticos?
4. Critérios de informação, como o AIC, favorecem algum modelo?

```
model_ll3 ← drm(growth ~ conc, data = data, fct = LL.3())
model_ll4 ← drm(growth ~ conc, data = data, fct = LL.4())
model_w14 ← drm(growth ~ conc, data = data, fct = W1.4())

mselect(model_ll4, list(LL.3(), W1.4(), W2.4()))
```

```
        logLik       IC Lack of fit  Res var
W2.4 -70.67105 151.3421   0.4018585 10.63559
LL.3 -71.74741 151.4948   0.3208891 11.02611
LL.4 -71.26136 152.5227   0.2792751 11.09363
W1.4 -72.47342 154.9468   0.1272162 12.09686
```

> ⚠ **Aviso**
>
> O R^2 não é um critério confiável para comparar modelos não lineares. Compare os modelos usando os padrões dos resíduos, critérios de informação e a plausibilidade biológica de seus parâmetros.

Quando modelos concorrentes dão ajustes semelhantes, mas estimativas diferentes de CE_{50}, a incerteza devida à escolha do modelo deve ser reconhecida.

9.14 Delineando experimentos para respostas não lineares

Na regressão linear, as observações nos extremos da variável explicativa são as mais informativas sobre a inclinação. Na regressão não linear, **a posição das observações determina quais parâmetros podem ser estimados**.

9.14.1 Escolhendo as doses

Para experimentos de dose–resposta:

- inclua uma testemunha sem tratamento;
- espace as doses em escala logarítmica (por exemplo, 0,01; 0,03; 0,1; 0,3; 1; 3 µg/mL);
- inclua doses que cubram toda a curva, de pouco efeito a efeito quase máximo;
- coloque várias doses perto da CE_{50} esperada;
- use experimentos preliminares de definição da faixa quando a sensibilidade for desconhecida.

9.14.2 Escolhendo as temperaturas

Em experimentos de resposta à temperatura, inclua temperaturas próximas da mínima e da máxima esperadas, além de temperaturas em torno do ótimo. Sem elas, as temperaturas cardeais não podem ser estimadas com confiabilidade.

9.14.3 Repetição e experimentos repetidos

Como no Capítulo 2, a repetição ocorre no nível da unidade experimental, como uma placa, folha ou planta. Medir a mesma colônia várias vezes não constitui repetição. Os ensaios de dose–resposta são comumente repetidos no tempo; as curvas de experimentos repetidos devem ser examinadas separadamente antes de serem combinadas, e modelos que levam em conta a variação entre experimentos, como os modelos mistos não lineares, são o arcabouço natural para combiná-las.

! Importante

Muitos problemas no ajuste de modelos não lineares são problemas de delineamento. Se as doses ou temperaturas testadas não cobrem a região em que a curva muda, nenhuma análise consegue recuperar a informação que falta.

9.15 Pesquisa na prática: monitorando a sensibilidade a fungicidas

Pesquisa revisitada: A população do patógeno está se tornando menos sensível?

Considere um estudo em que isolados de um fungo fitopatogênico, coletados antes e depois de vários anos de uso de um fungicida, foram testados em meio de cultura com uma série de concentrações do fungicida.

A questão científica não é simplesmente:

> O fungicida inibe o crescimento?

Em vez disso, pergunte:

1. Qual é a CE_{50} de cada isolado, e com que precisão ela foi estimada?
2. Qual definição de CE_{50} foi usada, e por quê?
3. Como os valores de CE_{50} se distribuem na população de linha de base e na população recente?
4. Qual é o tamanho da mudança de sensibilidade, expressa como razão?
5. Há isolados com fatores de resistência grandes o suficiente para importar na prática?
6. As doses testadas foram adequadas para estimar a CE_{50} tanto de isolados sensíveis quanto de menos sensíveis?
7. As estimativas de sensibilidade em laboratório se relacionam com o controle da doença no campo?

9.16 Um fluxo de trabalho prático para respostas não lineares

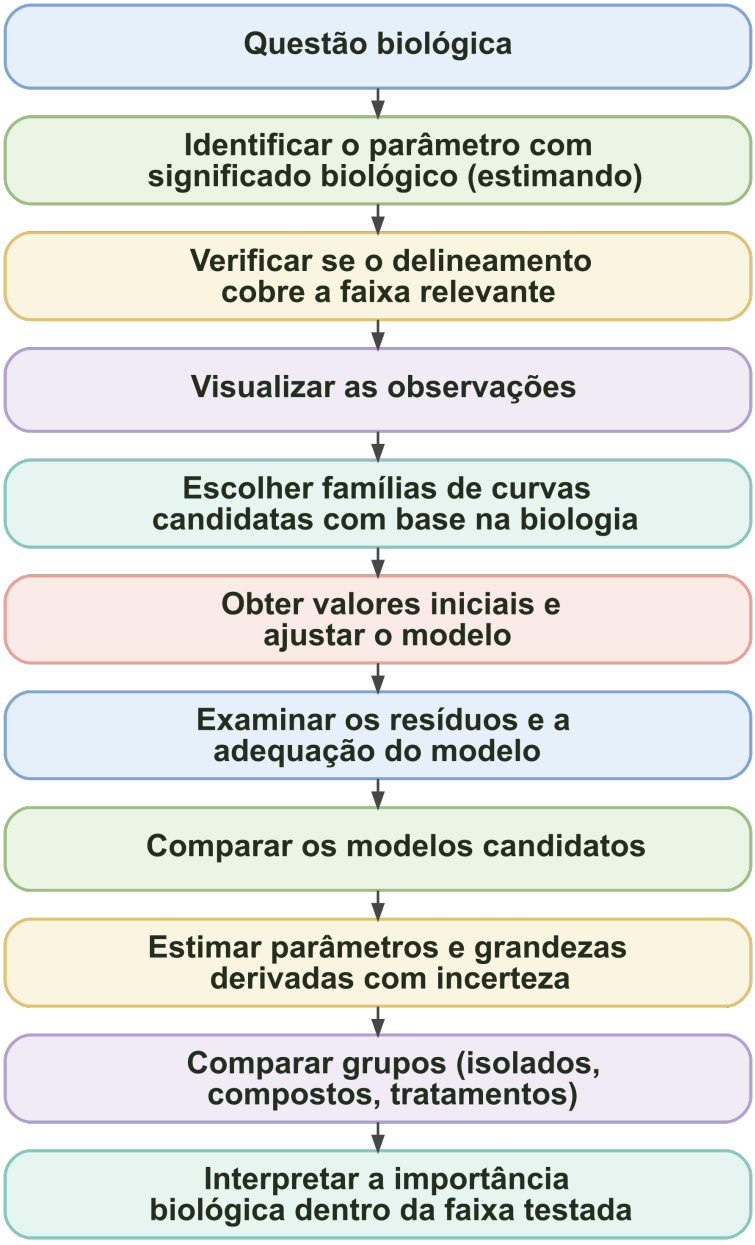


Figura 9.2. Fluxo de trabalho para analisar respostas não lineares.

9.17 Relatando análises não lineares

Um bom enunciado de resultados geralmente inclui:

9.17.1 Que modelo foi ajustado?

Nomeie o modelo, sua parametrização e o programa usado.

9.17.2 Estimativas dos parâmetros

Relate os parâmetros com significado biológico (CE_{50}, assíntotas, temperaturas cardeais) com suas unidades.

9.17.3 Incerteza

Relate os intervalos de confiança e informe como foram obtidos.

9.17.4 Definições

Informe se os valores de CE_{50} são absolutos ou relativos e se a resposta foi expressa em relação à testemunha.

9.17.5 Adequação do modelo

Descreva como o ajuste foi avaliado e, quando pertinente, como os modelos concorrentes foram comparados.

9.17.6 Interpretação biológica

Explique o que os parâmetros estimados significam para o organismo ou para a prática de manejo.

Em vez de:

> O fungicida reduziu significativamente o crescimento micelial (P < 0,001).

prefira algo como:

> O crescimento micelial seguiu uma resposta log-logística de quatro parâmetros à concentração. A CE_{50} relativa foi de 0,17 µg/mL (IC 95%: 0,13 a 0,20 µg/mL), e os isolados da coleta recente foram, em média, 10,7 vezes menos sensíveis que os isolados da linha de base (IC 95%: 7,3 a 15,6 vezes).

Esses números vêm das saídas de `ED(model, 50, interval = "delta")` e da comparação das distribuições de CE_{50} apresentadas neste capítulo.

9.18 Erros comuns

9.18.1 Ajustar retas a respostas curvas

A inclinação estimada é a média de uma taxa de variação que não é constante.

9.18.2 Usar polinômios quando os parâmetros importam

Os coeficientes de um polinômio não estimam doses efetivas nem ótimos.

9.18.3 Não informar a definição de CE_{50}

Valores de CE_{50} absolutos e relativos podem diferir substancialmente.

9.18.4 Normalizar em relação à testemunha e ignorar sua variabilidade

A média da testemunha também é uma estimativa.

9.18.5 Relatar a CE_{50} sem incerteza

Uma estimativa pontual sozinha não revela curvas mal estimadas.

9.18.6 Comparar valores de CE_{50} em escala linear

As diferenças de sensibilidade são multiplicativas e são mais bem expressas como razões.

9.18.7 Usar doses que não cobrem a curva

Parâmetros fora da faixa testada não podem ser estimados com confiabilidade.

9.18.8 Interpretar limiares extrapolados como medidas

As temperaturas mínima e máxima são frequentemente extrapolações.

9.18.9 Escolher modelos pelo R^2

O R^2 não discrimina entre modelos não lineares.

9.18.10 Ignorar avisos de convergência

Um modelo que não convergiu não produziu estimativas válidas.

9.18.11 Tratar medidas repetidas na mesma unidade como repetições

A precisão é superestimada quando subamostras são tratadas como independentes.

9.19 Lista de verificação para análises não lineares

Antes de interpretar uma análise não linear, pergunte:

1. Que questão biológica está sendo abordada?
2. Qual parâmetro da curva a responde?
3. A curva escolhida reflete a biologia conhecida?
4. As doses ou valores testados cobrem toda a resposta?
5. Qual é a unidade experimental?
6. Os valores iniciais foram razoáveis e o modelo convergiu?
7. Os resíduos mostram padrões sistemáticos?
8. Modelos alternativos foram considerados?
9. Qual definição de CE_{50} (ou de outro parâmetro) foi usada?
10. Quais são as estimativas dos parâmetros e seus intervalos de confiança?
11. As grandezas derivadas são relatadas com incerteza?
12. As comparações entre grupos são expressas em uma escala adequada?
13. Algum parâmetro é uma extrapolação além dos dados?
14. A resposta é contínua, ou um modelo para contagens ou proporções seria mais adequado?
15. Os efeitos estimados são biologicamente relevantes?

9.20 Resumo do capítulo

Muitas respostas biológicas em Fitopatologia são intrinsecamente não lineares: o crescimento do patógeno diminui de forma sigmoide com a concentração do fungicida, aumenta até um ótimo com a temperatura e se satura com a densidade de inóculo ou a duração do molhamento. Retas e polinômios descrevem mal essas respostas e não fornecem parâmetros com significado biológico.

Os modelos não lineares descrevem essas respostas com curvas cujos parâmetros correspondem a grandezas de interesse direto, como as assíntotas superior e inferior, a concentração efetiva que produz 50% de inibição e as temperaturas cardeais. Ajustar esses

modelos exige uma família de curvas adequada, valores iniciais razoáveis e um delineamento experimental que cubra a região em que a resposta muda. Uma interpretação confiável exige definições explícitas das grandezas estimadas, intervalos de confiança para os parâmetros e as grandezas derivadas, exame dos resíduos, consideração de modelos alternativos e cautela com parâmetros que são extrapolações além dos dados. As comparações de sensibilidade entre isolados ou compostos são naturalmente expressas como razões entre concentrações efetivas.

9.21 Conceitos-chave

Ao concluir este capítulo, você deverá ser capaz de explicar:

- resposta não linear;
- regressão não linear;
- assíntota;
- ponto de inflexão;
- modelo monomolecular;
- modelo logístico;
- modelo de Gompertz;
- modelo log-logístico;
- curva de dose–resposta;
- CE_{50};
- CE_{50} absoluta e relativa;
- hormese;
- fator de resistência;
- distribuição de sensibilidade;
- dose discriminatória;
- temperaturas cardeais;
- valores iniciais;
- convergência;
- intervalo de confiança de perfil;
- método delta;
- seleção de modelos;
- critérios de informação;
- delineamento experimental para modelos não lineares.

9.22 Exercícios

9.22.1 Exercício 1. Qual curva?

Para cada situação, esboce a forma esperada da resposta e identifique qual parâmetro responderia à questão biológica:

- crescimento micelial de um isolado em meio com concentrações crescentes de fungicida;
- germinação de esporos em temperaturas entre 5 e 35 °C;
- proporção de folhas infectadas após um número crescente de horas de molhamento foliar;
- número de lesões por folha após inoculação com concentrações crescentes de esporos.

9.22.2 Exercício 2. Reta, polinômio ou curva?

Usando os dados simulados de dose–resposta deste capítulo:

1. ajuste uma reta ao crescimento em função da concentração;
2. ajuste um polinômio quadrático;
3. ajuste um modelo log-logístico de quatro parâmetros;
4. compare os gráficos de resíduos;
5. explique qual modelo permite estimar a CE_{50}.

9.22.3 Exercício 3. Estimando a CE_{50}

Usando um conjunto de dados de dose–resposta fornecido pelo professor:

1. represente os dados com o eixo de concentração em escala logarítmica;
2. ajuste os modelos log-logísticos de três e de quatro parâmetros;
3. estime as CE_{50} relativa e absoluta com intervalos de confiança;
4. explique por que os dois valores diferem ou coincidem;
5. represente a curva ajustada com sua faixa de confiança e as observações.

9.22.4 Exercício 4. Comparando isolados

Usando dados de dose–resposta de vários isolados:

1. ajuste as curvas de todos os isolados simultaneamente;
2. estime a CE_{50} de cada isolado;
3. calcule os fatores de resistência em relação a um isolado de referência;
4. identifique os isolados cuja CE_{50} foi mal estimada e explique por quê.

9.22.5 Exercício 5. Temperatura ótima

Usando dados de crescimento medidos em várias temperaturas:

1. represente os dados;
2. ajuste um modelo com parâmetros de temperaturas cardeais;
3. relate T_{opt}, T_{min} e T_{max} com intervalos de confiança;
4. discuta quais parâmetros são bem sustentados pelas temperaturas testadas.

9.22.6 Exercício 6. Delineando um experimento de dose–resposta

Você precisa determinar a sensibilidade de um patógeno recém-descrito a um fungicida, sem nenhuma informação prévia. Proponha uma sequência de experimentos, incluindo as concentrações a testar, o número de repetições e como os resultados preliminares orientariam o delineamento final.

9.22.7 Exercício 7. Avaliação de um artigo publicado

Escolha um artigo recente de Fitopatologia que relate valores de CE_{50} ou outro modelo não linear.

Identifique:

1. o modelo ajustado;
2. as doses ou valores testados;
3. se a definição de CE_{50} foi informada;
4. se a incerteza dos parâmetros foi relatada;
5. se as curvas ajustadas foram mostradas junto com os dados;
6. se modelos alternativos foram considerados;
7. se algum parâmetro relatado era uma extrapolação;
8. se as comparações entre isolados ou compostos foram expressas de forma adequada.

9.23 Reflexão

> **Em uma resposta não linear, a forma da curva não é um incômodo estatístico. Ela é o resultado biológico.**

Discuta essa afirmação usando um exemplo da Fitopatologia.

9.24 Próximos passos

Os modelos dos Capítulos 7 a 9 pressupõem uma resposta contínua com variabilidade aproximadamente constante e simétrica em torno da média. Muitas respostas em Fitopatologia não se encaixam nessa descrição: uma planta está infectada ou não, uma folha tem determinado número de lesões, uma proporção das sementes germina ou uma lavoura é classificada em uma categoria de doença. O Capítulo 10 apresenta métodos para dados categóricos, de contagem e de proporção, incluindo modelos logísticos e do tipo Poisson, que respeitam a natureza dessas respostas.

CAPÍTULO 10

Como analisamos dados categóricos, de contagem e de proporção?

10.1 Objetivos de aprendizagem

! Importante

Ao concluir este capítulo, o leitor deverá ser capaz de:

- reconhecer respostas binárias, categóricas, ordinais, de contagem e de proporção;
- explicar por que essas respostas violam os pressupostos dos modelos usados nos Capítulos 7 a 9;
- resumir dados categóricos com frequências, proporções e sua incerteza;
- usar testes de qui-quadrado e testes exatos para aderência e para associação entre variáveis categóricas;
- descrever os componentes de um modelo linear generalizado;
- ajustar e interpretar a regressão logística para respostas binárias e binomiais usando razões de chances e probabilidades previstas;
- ajustar e interpretar modelos do tipo Poisson para contagens usando razões de taxas e *offsets*;
- detectar e tratar a superdispersão;
- distinguir proporções derivadas de contagens de proporções contínuas, como a severidade em porcentagem;
- reconhecer escalas ordinais de doença e os modelos adequados a elas;
- comunicar os resultados na escala da resposta original.

10.2 Por que esses dados são diferentes?

Os modelos dos Capítulos 7 a 9 descrevem uma resposta contínua cuja variabilidade em torno da média é aproximadamente constante e simétrica. Muitas das observações mais comuns em Fitopatologia não se encaixam nessa descrição.

Respostas típicas incluem:

- **desfechos binários**: uma planta está infectada ou não; uma semente carrega o patógeno ou não; um isolado é resistente ou sensível;
- **proporções a partir de contagens**: 37 de 50 plantas com sintomas; 120 de 200 esporos germinados; 8 de 30 sementes infectadas;

- **contagens**: número de lesões por folha, colônias por placa, esporos capturados por dia, insetos por armadilha adesiva;
- **categorias nominais**: tipo de acasalamento (*mating type*), espécie, haplótipo, tipo de sintoma;
- **categorias ordinais**: escalas de severidade com classes ordenadas, como de 0 a 5 ou a escala de Horsfall–Barratt;
- **proporções contínuas**: porcentagem da área foliar doente.

Essas respostas têm características que importam para a análise:

- são **limitadas**: uma proporção não pode ser menor que 0 nem maior que 1, e uma contagem não pode ser negativa;
- sua **variabilidade depende da média**: uma proporção próxima de 0,5 varia mais do que uma próxima de 0 ou 1, e contagens maiores variam mais do que contagens pequenas;
- são frequentemente **discretas**: as contagens assumem valores inteiros, e os desfechos binários, apenas dois.

> **! Importante**
>
> A questão não é como transformar esses dados até que pareçam normais. É qual modelo estatístico descreve como esses dados são gerados.

10.3 Da questão biológica ao estimando

Considere um ensaio de campo em que se registrou o número de plantas com sintomas em cada parcela.

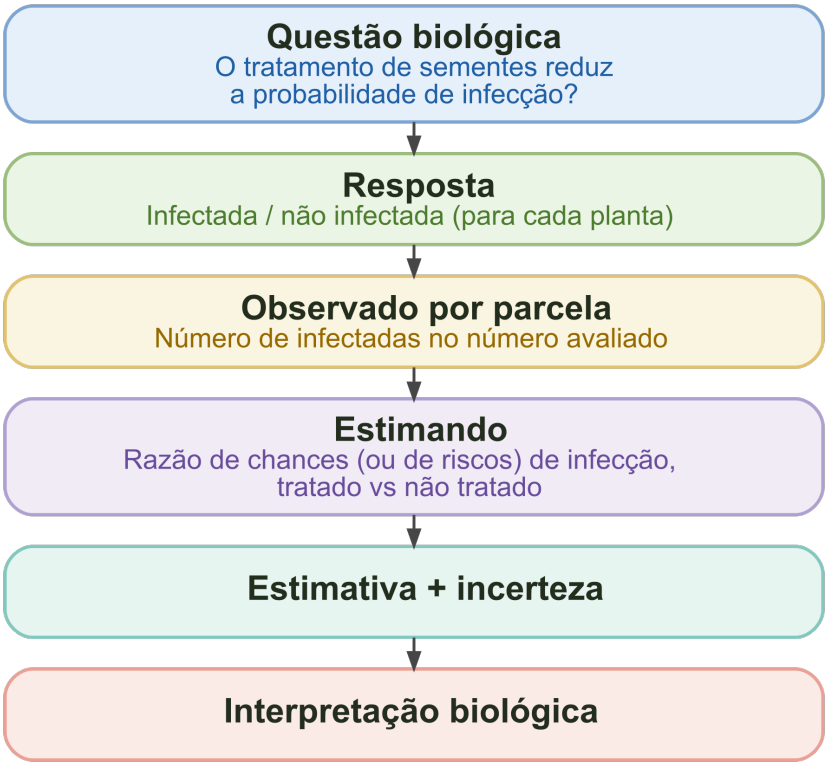


Figura 10.1. Da questão biológica à estimação de uma razão de chances.

10.3.1 Comparando proporções

Quando a resposta é uma probabilidade, o efeito de um tratamento pode ser expresso de diferentes formas. Se a probabilidade de infecção é p_C na testemunha e p_T no tratamento:

- a **diferença de riscos** é $p_T - p_C$;
- a **razão de riscos** é p_T / p_C;
- a **razão de chances** (*odds ratio*) é $\frac{p_T/(1-p_T)}{p_C/(1-p_C)}$.

A diferença de riscos é a mais intuitiva para decisões práticas; a razão de chances é a escala natural da regressão logística.

10.3.2 Comparando contagens

Quando a resposta é uma contagem, os efeitos geralmente são expressos como **razões de taxas**: o número de lesões por folha em um tratamento dividido pelo número em outro. Uma razão de taxas de 0,4 significa uma redução de 60%.

> **Dica**
>
> Informe em que escala o efeito é relatado. Uma redução de 10% para 5% de infecção é uma diferença de riscos de −5 pontos percentuais, uma razão de riscos de 0,5 e uma razão de chances de cerca de 0,47. Todas estão corretas; respondem a perguntas ligeiramente diferentes.

10.4 Por que não usar o modelo linear usual?

Ajustar `lm()` a proporções ou contagens pode produzir:

- proporções previstas abaixo de 0 ou acima de 1, e contagens previstas negativas;
- erros padrão e intervalos de confiança incorretos, porque a variabilidade não é constante;
- efeitos que parecem aditivos em uma escala em que os processos biológicos são multiplicativos.

10.4.1 E as transformações?

A solução tradicional era transformar a resposta — com o arco-seno da raiz quadrada para proporções ou com a raiz quadrada ou o logaritmo para contagens — e depois aplicar a ANOVA. Essas transformações estabilizam aproximadamente a variância, mas:

- os efeitos estimados ficam em uma escala difícil de interpretar;
- a transformação não leva em conta números diferentes de plantas avaliadas por parcela;
- os zeros exigem ajustes arbitrários antes de tomar logaritmos;
- a análise não reflete como os dados foram gerados.

Os modelos lineares generalizados resolvem esses problemas diretamente e hoje estão amplamente disponíveis. As transformações continuam sendo uma aproximação razoável em algumas situações, particularmente para contagens grandes e distantes de zero, mas não devem ser o padrão (ver também a seção “Os dados em porcentagem devem sempre ser transformados?” do Capítulo 7).

10.5 Descrevendo dados categóricos

10.5.1 Tabelas de frequência

O primeiro passo é tabular as observações.

```
table(plantas$treatment, plantas$infected)
```

```
            não sim
  Testemunha  30  30
  Tratado     55   5
```

```
prop.table(table(plantas$treatment, plantas$infected), margin = 1)
```

```
                   não        sim
  Testemunha 0.50000000 0.50000000
  Tratado    0.91666667 0.08333333
```

10.5.2 Proporções e sua incerteza

Uma proporção estimada a partir de n observações tem uma incerteza que depende tanto de n quanto da própria proporção. O intervalo simples $\hat{p} \pm 1,96\sqrt{\hat{p}(1-\hat{p})/n}$ funciona mal quando n é pequeno ou $\hat{p}$ está próximo de 0 ou 1, e pode incluir valores impossíveis. O intervalo de Wilson é uma escolha-padrão melhor.

```
prop.test(x = 8, n = 30)$conf.int     # intervalo de Wilson com correção de continuidade
```

```
[1] 0.1297544 0.4617278
attr(,"conf.level")
[1] 0.95
```

```
binom.test(x = 8, n = 30)$conf.int    # intervalo exato (Clopper-Pearson)
```

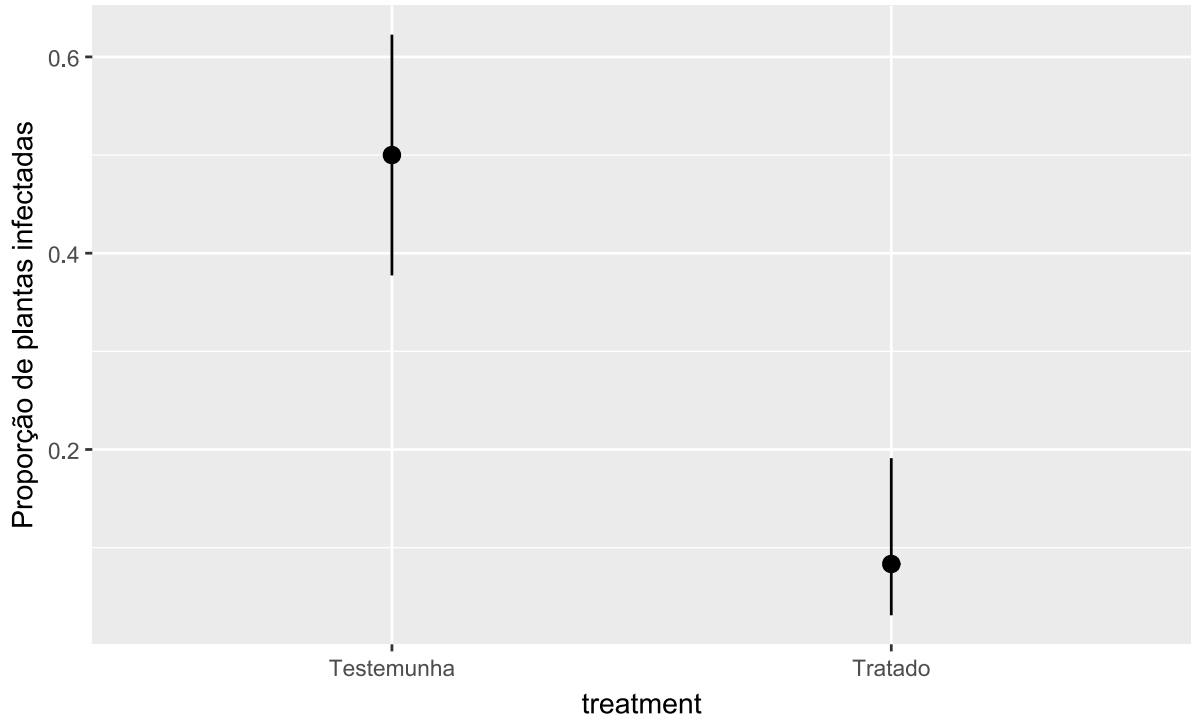
```
[1] 0.1227948 0.4588937
attr(,"conf.level")
[1] 0.95
```

> **⚠ Aviso**
>
> Nenhuma semente infectada em uma amostra de 30 não mostra que o lote está livre do patógeno. O limite superior do intervalo de confiança de 95% para a taxa de infecção fica em torno de 10% a 12%.

10.5.3 Visualizando dados categóricos

As proporções são mostradas como pontos com intervalos de confiança ou como barras empilhadas. Quando o número de observações difere entre os grupos, mostrar as contagens ou os tamanhos de amostra junto com as proporções ajuda o leitor a julgar a precisão.

```
library(ggplot2)

ggplot(summary_data, aes(x = treatment, y = prop)) +
  geom_pointrange(aes(ymin = lower, ymax = upper)) +
  labs(y = "Proporção de plantas infectadas")
```



10.6 Testes de qui-quadrado e testes exatos

10.6.1 Aderência

Um teste de aderência pergunta se as frequências observadas são compatíveis com proporções esperadas definidas por uma hipótese biológica. Exemplos clássicos em Fitopatologia incluem as razões de segregação de genes de resistência (por exemplo, 3 resistentes : 1 suscetível em uma população F_2) e a frequência dos tipos de acasalamento em uma população (1:1 sob acasalamento ao acaso).

```
observed ← c(resistant = 152, susceptible = 48)

chisq.test(observed, p = c(3/4, 1/4))
```

```
    Chi-squared test for given probabilities

data:  observed
X-squared = 0.10667, df = 1, p-value = 0.744
```

10.6.2 Associação entre duas variáveis categóricas

Um teste de independência pergunta se a distribuição de uma variável categórica difere entre os níveis de outra. Por exemplo: a frequência de isolados resistentes a um fungicida difere entre regiões?

```
tab ← table(isolados_reg$region, isolados_reg$phenotype)

chisq.test(tab)
```

```
    Pearson's Chi-squared test

data:  tab
X-squared = 17.76, df = 2, p-value = 0.0001391
```

10.6.3 Contagens esperadas e testes exatos

Os testes de qui-quadrado se baseiam em aproximações que funcionam mal quando as contagens esperadas são pequenas (uma regra comum é menos de 5 em algumas células). Nesse caso, o teste exato de Fisher é preferível.

```
fisher.test(tab)
```

```
    Fisher's Exact Test for Count Data

data:  tab
p-value = 4.056e-05
alternative hypothesis: two.sided
```

10.6.4 Dados categóricos pareados

Quando as mesmas amostras são classificadas duas vezes — por exemplo, as mesmas sementes testadas por dois métodos de detecção —, as observações são pareadas. O teste de McNemar compara as duas classificações concentrando-se nos pares discordantes.

```
mcnemar.test(table(deteccao$method_A, deteccao$method_B))
```

```
    McNemar's Chi-squared test with continuity correction

data:  table(deteccao$method_A, deteccao$method_B)
McNemar's chi-squared = 0.071429, df = 1, p-value = 0.7893
```

10.6.5 Acurácia diagnóstica

Quando um novo método de detecção é comparado com um método de referência, as questões de interesse geralmente não são respondidas por um teste de associação. São respondidas estimando:

- a **sensibilidade**: a proporção de amostras realmente infectadas que o método detecta;

- a **especificidade**: a proporção de amostras realmente sadias que o método classifica como sadias.

Cada uma é uma proporção e deve ser relatada com seu intervalo de confiança.

10.6.6 Limitações dos testes de qui-quadrado

Os testes de qui-quadrado respondem se as variáveis estão associadas, e não com que intensidade. Lidam com apenas um ou dois fatores, não podem incluir blocos ou covariáveis e não conseguem modelar a dependência de uma probabilidade em relação a uma variável quantitativa. Essas limitações são superadas pelos modelos lineares generalizados.

> **! Importante**
>
> Como a ANOVA no Capítulo 7, um teste de qui-quadrado significativo é um ponto de partida. O resultado informativo é a diferença ou a razão estimada entre as proporções, com sua incerteza.

10.7 Modelos lineares generalizados

10.7.1 Três componentes

Os modelos lineares generalizados (MLGs) estendem os modelos lineares dos Capítulos 7 e 8 a respostas não normais. Um MLG tem três componentes:

1. uma **distribuição de probabilidade** para a resposta, como a binomial para proporções a partir de contagens ou a de Poisson para contagens;
2. um **preditor linear**, $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$, exatamente como no modelo linear;
3. uma **função de ligação** que conecta a média da resposta ao preditor linear, como o logito para probabilidades ou o logaritmo para contagens.

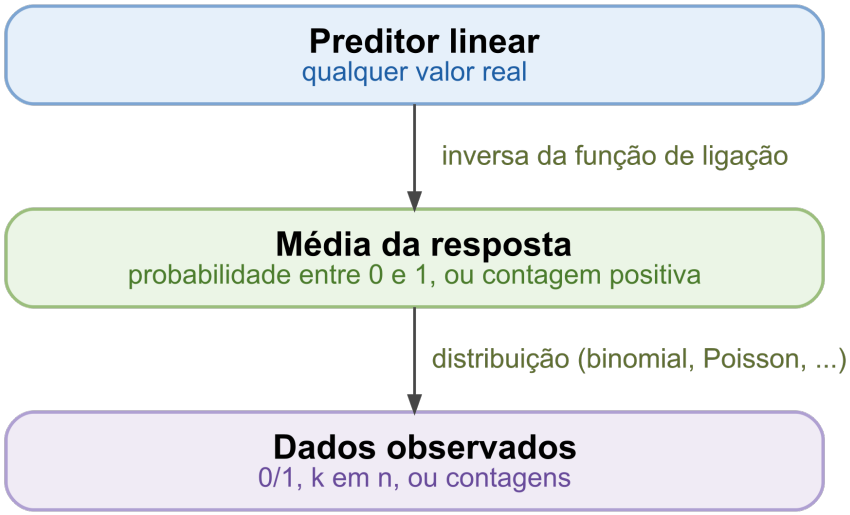


Figura 10.2. Os três componentes de um modelo linear generalizado.

A função de ligação mantém as previsões dentro de limites válidos, e a distribuição descreve como a variabilidade depende da média.

> **! Importante**

O teste *t*, a ANOVA, a regressão linear, a regressão logística e a regressão de Poisson pertencem à mesma família. A estrutura do modelo, com fatores, covariáveis, interações e blocos, é mantida sem alterações; mudam apenas a distribuição e a ligação.

10.7.2 Ajustando um MLG no R

```
model ← glm(response ~ treatment, family = binomial, data = data)   # modelo genérico
```

Todas as ferramentas dos capítulos anteriores, incluindo `summary()`, `anova()`, `emmeans()` e os contrastes, aplicam-se aos MLGs.

10.8 Respostas binárias e binomiais: regressão logística

10.8.1 O modelo

Para uma probabilidade de infecção p, a regressão logística modela o logaritmo da chance (*odds*):

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

A exponencial de um coeficiente fornece uma **razão de chances**: e^{β_1} é a mudança multiplicativa na chance de infecção para um aumento de uma unidade em x, ou entre um tratamento e o nível de referência.

10.8.2 Observações individuais ou contagens agrupadas

Os dados podem ser organizados com uma linha por planta e uma resposta 0/1, ou com uma linha por parcela com o número de plantas infectadas e sadias. A forma agrupada usa `cbind(sucessos, fracassos)`:

```
set.seed(606)

data ← data.frame(
  treatment = factor(rep(c("untreated", "treated"), each = 6),
                     levels = c("untreated", "treated")),
  block = factor(rep(1:6, times = 2)),
  plot = factor(1:12),
  assessed = 50
)

p ← ifelse(data$treatment == "untreated", 0.40, 0.15)
data$infected ← rbinom(nrow(data), size = data$assessed, prob = p)

model ← glm(
  cbind(infected, assessed - infected) ~ treatment,
  family = binomial,
  data = data
)

summary(model)
```

```
Call:
glm(formula = cbind(infected, assessed - infected) ~ treatment,
    family = binomial, data = data)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.3091     0.1169  -2.645  0.00816 **
treatmenttreated  -1.5919     0.2077  -7.665 1.78e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 79.945  on 11  degrees of freedom
Residual deviance: 12.851  on 10  degrees of freedom
AIC: 64.022

Number of Fisher Scoring iterations: 4
```

> ⚠ **Aviso**
>
> Não analise a proporção de infectadas como se fosse uma resposta contínua, ignorando quantas plantas foram avaliadas. Uma proporção de 0,5 baseada em 2 plantas carrega muito menos informação do que uma baseada em 200.

10.8.3 Dos coeficientes a grandezas interpretáveis

Os coeficientes na escala do logito são difíceis de interpretar. Os resultados são mais úteis quando convertidos em:

- **probabilidades previstas** para cada tratamento;
- **razões de chances** entre tratamentos;
- **diferenças de probabilidade**, quando as decisões práticas dependem delas.

```
library(emmeans)

emm ← emmeans(model, ~ treatment, type = "response")

emm            # probabilidades previstas com intervalos de confiança
```

```
 treatment  prob     SE  df asymp.LCL asymp.UCL
 untreated 0.423 0.0285 Inf    0.3686     0.480
 treated   0.130 0.0194 Inf    0.0964     0.173

Confidence level used: 0.95
Intervals are back-transformed from the logit scale
```

```
pairs(emm)     # razão de chances entre os tratamentos
```

```
 contrast            odds.ratio   SE  df null z.ratio p.value
 untreated / treated       4.91 1.02 Inf    1   7.665 <0.0001
```

```
Tests are performed on the log odds ratio scale
```

Com `reverse = TRUE`, a razão de chances passa a ser “tratado / não tratado”; com `regrid()`, o contraste é calculado na escala da probabilidade, como uma diferença:

```
confint(pairs(emm, reverse = TRUE))          # razão de chances com IC
```

```
 contrast            odds.ratio     SE  df asymp.LCL asymp.UCL
 treated / untreated      0.204 0.0423 Inf     0.135     0.306

Confidence level used: 0.95
Intervals are back-transformed from the log odds ratio scale
```

```
confint(pairs(regrid(emm), reverse = TRUE))   # diferença de probabilidade com IC
```

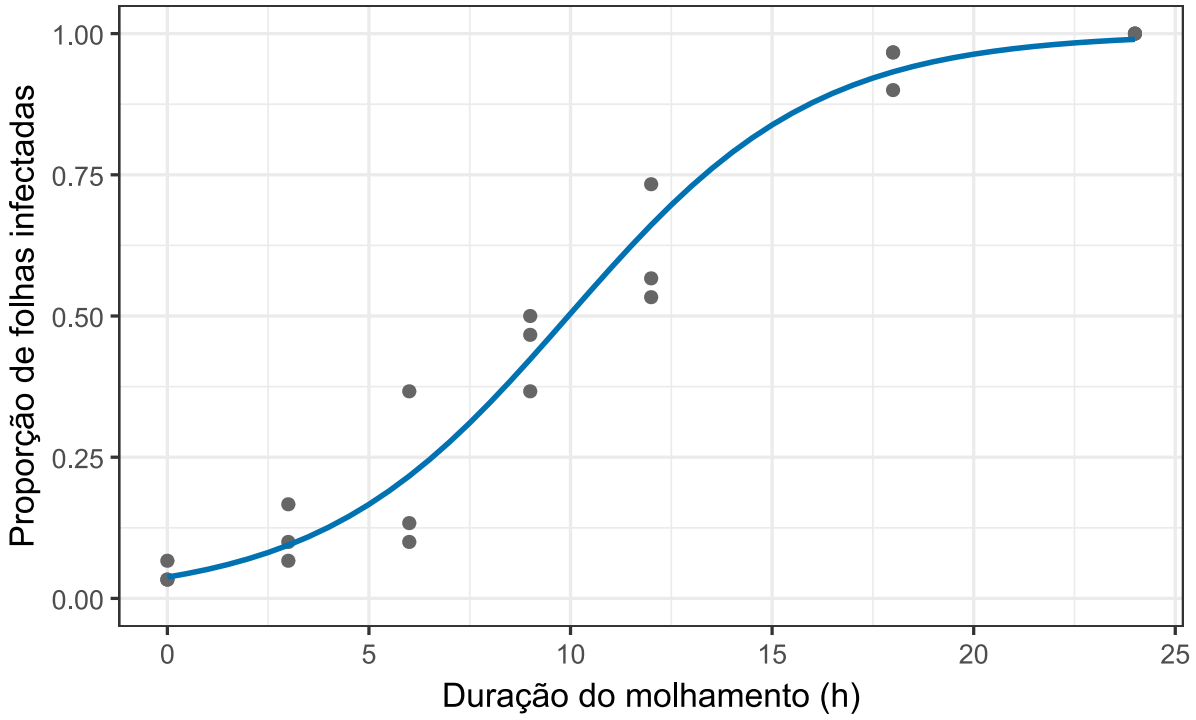
```
 contrast            estimate     SE  df asymp.LCL asymp.UCL
 treated - untreated   -0.293 0.0345 Inf    -0.361    -0.226

Confidence level used: 0.95
```

10.8.4 Variáveis explicativas quantitativas

A regressão logística também descreve como uma probabilidade muda com uma variável quantitativa, como a probabilidade de infecção à medida que aumenta a duração do molhamento foliar. A curva ajustada tem forma de S na escala da probabilidade e não pode ultrapassar 0 nem 1.

Isso se liga diretamente ao Capítulo 9: uma curva de dose–resposta para a proporção de esporos germinados é uma regressão logística sobre o logaritmo da concentração, e a CE_{50} pode ser estimada a partir dela.

```
model_wet ← glm(
  cbind(infected, total - infected) ~ wetness,
  family = binomial,
  data = data_wet
)

# curva ajustada na escala da probabilidade
grade_w ← data.frame(wetness = seq(0, 24, by = 0.5))
grade_w$p ← predict(model_wet, newdata = grade_w, type = "response")

ggplot(data_wet, aes(wetness, infected / total)) +
  geom_point(colour = "grey40") +
  geom_line(data = grade_w, aes(wetness, p), colour = "#0072B2", linewidth = 1) +
  scale_y_continuous(limits = c(0, 1)) +
  labs(x = "Duração do molhamento (h)", y = "Proporção de folhas infectadas") +
  theme_bw(base_size = 13)
```



10.9 Respostas de contagem: modelos do tipo Poisson

10.9.1 O modelo de Poisson

Para contagens, como lesões por folha ou colônias por placa, o MLG de Poisson modela o logaritmo da contagem esperada:

$$\log(\mu) = \beta_0 + \beta_1 x$$

A exponencial de um coeficiente fornece uma **razão de taxas**, a mudança multiplicativa na contagem esperada.

```
model ← glm(lesions ~ treatment, family = poisson, data = lesoes)

emmeans(model, ~ treatment, type = "response")
```

```
 treatment  rate    SE  df asymp.LCL asymp.UCL
 Testemunha 11.9 0.771 Inf     10.48     13.51
 Fung. A     3.2 0.400 Inf      2.50      4.09
 Fung. B     4.7 0.485 Inf      3.84      5.75

Confidence level used: 0.95
Intervals are back-transformed from the log scale
```

```
pairs(emmeans(model, ~ treatment), type = "response")   # razões de taxas
```

```
 contrast             ratio    SE  df null z.ratio p.value
 Testemunha / Fung. A 3.719 0.524 Inf    1   9.328 <0.0001
 Testemunha / Fung. B 2.532 0.308 Inf    1   7.626 <0.0001
```

```
 Fung. A / Fung. B    0.681 0.110 Inf    1  -2.372  0.0465

P value adjustment: tukey method for comparing a family of 3 estimates
Tests are performed on the log scale
```

10.9.2 Esforços de amostragem diferentes: *offsets*

As contagens são frequentemente obtidas com esforço desigual: lesões em folhas de tamanhos diferentes, esporos de armadilhas que funcionaram por períodos diferentes, colônias de volumes diferentes semeados. Um ***offset*** inclui o logaritmo do esforço no modelo, de modo que a análise estime taxas (por cm², por hora, por mL), e não contagens brutas.

```
model ← glm(
  lesions ~ treatment + offset(log(leaf_area)),
  family = poisson,
  data = lesoes
)

# razões de taxas (lesões por cm² de folha)
exp(coef(model))
```

```
     (Intercept) treatmentFung. A treatmentFung. B
       0.2776482        0.2650426        0.4546844
```

> **Dica**
>
> Dividir as contagens pelo esforço e analisar as taxas resultantes como resposta contínua descarta informação sobre a precisão. Modele as contagens e use um *offset*.

10.10 Superdispersão

10.10.1 O que é superdispersão?

As distribuições binomial e de Poisson pressupõem uma relação específica entre a média e a variância. Os dados de campo são frequentemente **mais variáveis** do que essas distribuições permitem, porque parcelas, plantas ou folhas diferem de formas não captadas pelo modelo: microclima, inóculo, vigor das plantas, agregação da doença.

Essa variabilidade extra é chamada **superdispersão**. Ignorá-la produz erros padrão pequenos demais e intervalos de confiança estreitos demais.

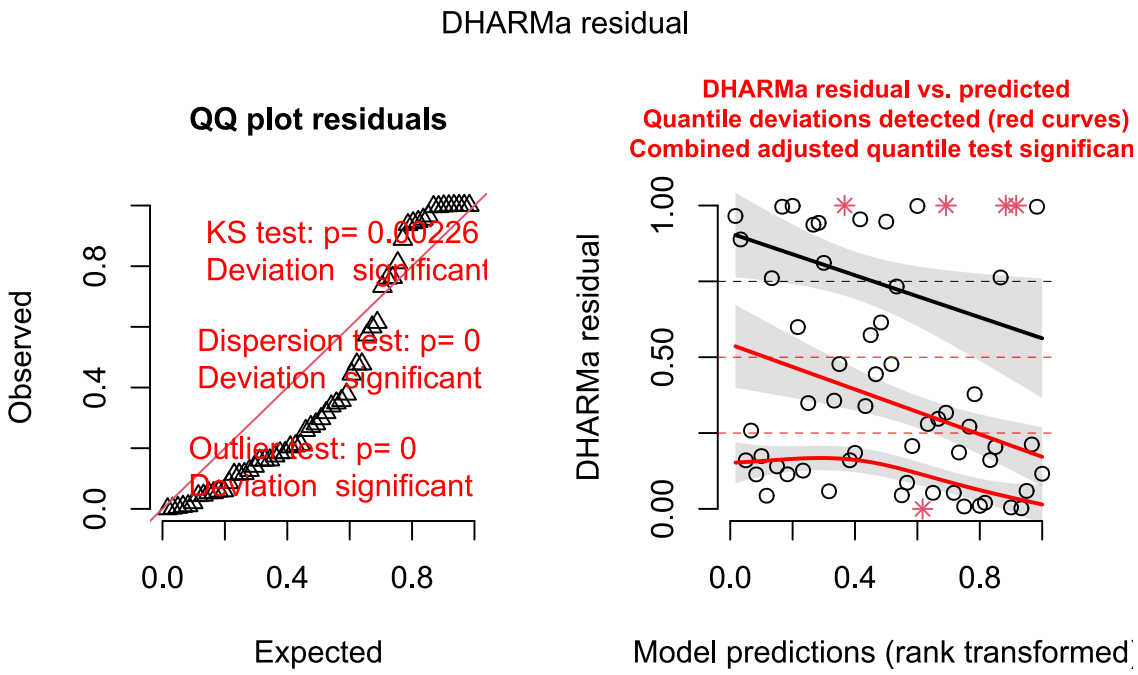
10.10.2 Detectando a superdispersão

Uma verificação simples compara o desvio residual (ou a soma dos quadrados dos resíduos de Pearson) com os graus de liberdade residuais; razões muito acima de 1 sugerem superdispersão. Resíduos baseados em simulação fornecem um diagnóstico mais completo.

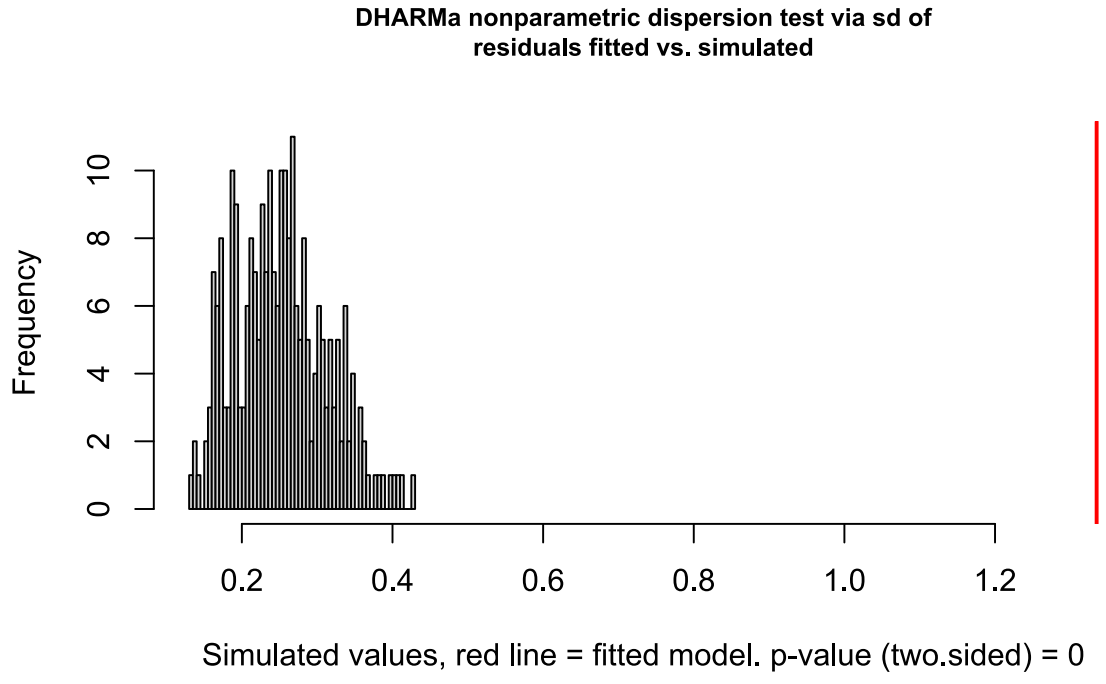
```
sum(residuals(model, type = "pearson")^2) / df.residual(model)
```

```
[1] 4.012932
```

```
library(DHARMa)

res ← simulateResiduals(model)
plot(res)
```




```
testDispersion(res)
```



```
    DHARMa nonparametric dispersion test via sd of residuals fitted vs.
    simulated

data:  simulationOutput
dispersion = 5.3134, p-value < 2.2e-16
alternative hypothesis: two.sided
```

10.10.3 Tratando a superdispersão

As opções comuns incluem:

- **modelos de quase-verossimilhança** (`quasibinomial`, `quasipoisson`), que inflacionam os erros padrão por um fator de dispersão estimado;
- **modelos binomiais negativos** para contagens, que permitem que a variância aumente mais rapidamente do que a média;
- **modelos beta-binomiais** para proporções a partir de contagens;
- **efeitos aleatórios** para as unidades experimentais (por exemplo, parcelas), que modelam explicitamente a variação extra.

```
library(MASS)

model_nb ← glm.nb(lesions ~ treatment, data = lesoes)
summary(model_nb)
```

```
Call:
glm.nb(formula = lesions ~ treatment, data = lesoes, init.theta = 2.624938252,
    link = log)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)        2.4765     0.1525  16.242  < 2e-16 ***
treatmentFung. A  -1.3134     0.2407  -5.457 4.84e-08 ***
treatmentFung. B  -0.9290     0.2301  -4.038 5.40e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(2.6249) family taken to be 1)

    Null deviance: 96.447  on 59  degrees of freedom
Residual deviance: 61.400  on 57  degrees of freedom
AIC: 332.38

Number of Fisher Scoring iterations: 1

              Theta:  2.625
          Std. Err.:  0.682

 2 x log-likelihood:  -324.377
```

! Importante

A superdispersão é a regra, e não a exceção, em dados de campo. Verifique-a sempre que um modelo binomial ou de Poisson for ajustado.

10.10.4 Excesso de zeros

Alguns dados de contagem contêm mais zeros do que uma distribuição de Poisson ou binomial negativa prevê, por exemplo quando muitas folhas escapam completamente da infecção e as demais têm muitas lesões. Modelos inflacionados de zeros e modelos *hurdle* foram criados para esses dados. Antes de usá-los, verifique se um modelo binomial negativo ou a inclusão de variáveis explicativas relevantes já explicam os zeros.

10.11 Considerando o delineamento experimental

Tudo o que foi dito nos Capítulos 2 e 7 sobre unidades experimentais, blocos e subamostragem se aplica aos MLGs.

Quando várias plantas são avaliadas em cada parcela, a **parcela** é a unidade experimental. Tratar cada planta como uma observação independente superestima a evidência. Os blocos são incluídos como no modelo linear, e a variação entre parcelas pode ser representada por efeitos aleatórios em um modelo linear generalizado misto.

```
library(lme4)

model_glmm ← glmer(
  cbind(infected, assessed - infected) ~ treatment + (1 | block) + (1 | plot),
  family = binomial,
  data = data
)

summary(model_glmm)$coefficients
```

```
                  Estimate Std. Error   z value     Pr(>|z|)
(Intercept)      -0.3102152  0.1266140 -2.450086 1.428223e-02
treatmenttreated -1.5958796  0.2190257 -7.286267 3.186603e-13
```

10.12 Proporções contínuas: severidade em porcentagem

Nem toda porcentagem é uma proporção de contagens. A **severidade**, a porcentagem da área do tecido com sintomas, é uma proporção contínua: não há um número de “sucessos” em um número conhecido de “tentativas”.

As opções para dados de severidade incluem:

- um modelo linear, quando os valores estão em sua maioria longe de 0% e 100% e os resíduos se comportam razoavelmente (Capítulo 7);
- a **regressão beta**, que modela diretamente proporções contínuas e respeita os limites;
- modelos que tratam separadamente os zeros ou uns exatos, quando muitas observações assumem esses valores.

```
library(glmmTMB)

sev_data$sev_prop ← sev_data$severity / 100

model_beta ← glmmTMB(sev_prop ~ treatment, family = beta_family(), data = sev_data)
summary(model_beta)
```

```
 Family: beta  ( logit )
Formula:          sev_prop ~ treatment
Data: sev_data

      AIC       BIC    logLik -2*log(L)  df.resid
    -48.7     -44.5      27.3     -54.7        27


Dispersion parameter for beta family (): 16.9

Conditional model:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -0.3946     0.1240  -3.183  0.00146 **
treatmentTratado  -1.1987     0.2003  -5.984 2.18e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A regressão beta exige valores estritamente entre 0 e 1, de modo que as parcelas sem doença precisam de tratamento à parte.

> ⚠ **Aviso**
>
> Não use um modelo binomial para a severidade em porcentagem tratando a porcentagem como um número de sucessos em 100. Não houve 100 tentativas independentes.

10.13 Respostas ordinais: escalas de notas de doença

A doença é frequentemente avaliada usando escalas com classes ordenadas, como de 0 (sem sintomas) a 5 (planta morta), ou escalas de intervalos como a de Horsfall–Barratt.

Os dados ordinais apresentam problemas específicos:

- as distâncias entre as classes não são necessariamente iguais, de modo que calcular a média dos números das classes pressupõe algo que pode não ser verdade;
- converter as classes em seus pontos médios e analisar o resultado como severidade contínua introduz erro, particularmente em escalas de intervalos com classes desiguais.

Os **modelos de ligação cumulativa** (modelos de chances proporcionais) tratam a resposta como categorias ordenadas e estimam o efeito dos tratamentos sobre a probabilidade de estar nas classes mais altas.

```
library(ordinal)

notas$score ← factor(notas$score, ordered = TRUE)

model_ord ← clm(score ~ treatment, data = notas)
```

```
summary(model_ord)
```

```
formula: score ~ treatment
data:    notas

 link  threshold nobs logLik AIC    niter max.grad cond.H
 logit flexible  75   -98.00 210.01 6(0)  6.92e-08 5.7e+01

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
treatmentCultivar B  -3.4431     0.7421  -4.640 3.49e-06 ***
treatmentCultivar C  -5.2843     0.8400  -6.291 3.15e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Threshold coefficients:
    Estimate Std. Error z value
0|1  -5.6434     0.8059  -7.002
1|2  -4.4803     0.7508  -5.968
2|3  -2.5578     0.6366  -4.018
3|4  -0.1830     0.3912  -0.468
4|5   1.4315     0.5009   2.858
```

10.14 Escolhendo um modelo

Tabela 10.1. Escolhendo um modelo de acordo com o tipo de resposta.

Resposta	Exemplo	Família de modelos	Escala do efeito
Binária	Infectada ou não	Binomial (logística)	Razão de chances, probabilidade
Proporção a partir de contagens	37 de 50 plantas infectadas	Binomial (logística)	Razão de chances, probabilidade
Contagem	Lesões por folha	Poisson ou binomial negativa	Razão de taxas
Contagem com esforço desigual	Esporos por armadilha, horas variáveis	Poisson ou binomial negativa com *offset*	Razão de taxas
Categorias nominais	Tipo de acasalamento por região	Qui-quadrado ou multinomial	Diferença de frequências
Categorias ordinais	Escala de doença 0–5	Modelo de ligação cumulativa	Razão de chances de classe mais alta
Proporção contínua	Severidade em porcentagem	Modelo linear ou regressão beta	Diferença ou razão de chances

10.15 Pesquisa na prática: avaliando um tratamento de sementes

Pesquisa revisitada: O tratamento de sementes reduz a infecção?

Considere um experimento de campo em que parcelas semeadas com sementes tratadas ou não tratadas foram dispostas em blocos casualizados. Em cada parcela, 50 plântulas foram examinadas e classificadas como infectadas ou sadias, e o número de lesões foi contado em uma amostra de folhas.

A questão científica não é simplesmente:

> Há diferença significativa de incidência entre os tratamentos?

Em vez disso, pergunte:

1. Qual é a unidade experimental, e quantas havia por tratamento?
2. Qual é a probabilidade estimada de infecção em cada tratamento, com seu intervalo de confiança?
3. Qual é o tamanho do efeito, expresso como diferença de probabilidade e como razão de chances?
4. Há mais variação entre parcelas do que um modelo binomial espera?
5. Como comparar contagens de lesões em folhas de tamanhos diferentes?
6. A redução da incidência é grande o suficiente para justificar o custo do tratamento?

10.16 Um fluxo de trabalho prático para dados categóricos, de contagem e de proporção

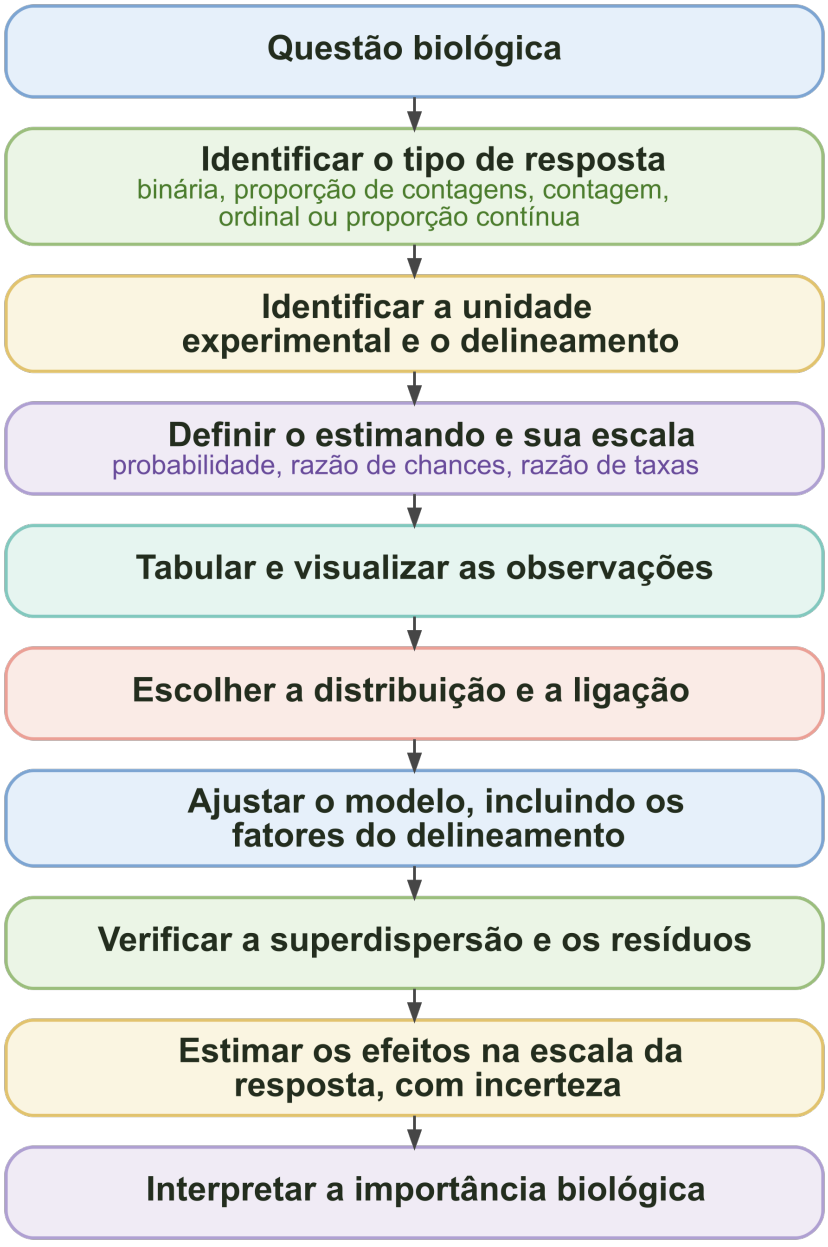


Figura 10.3. Fluxo de trabalho para dados categóricos, de contagem e de proporção.

A escolha da família de modelos decorre diretamente do tipo de resposta (Figura 10.4).

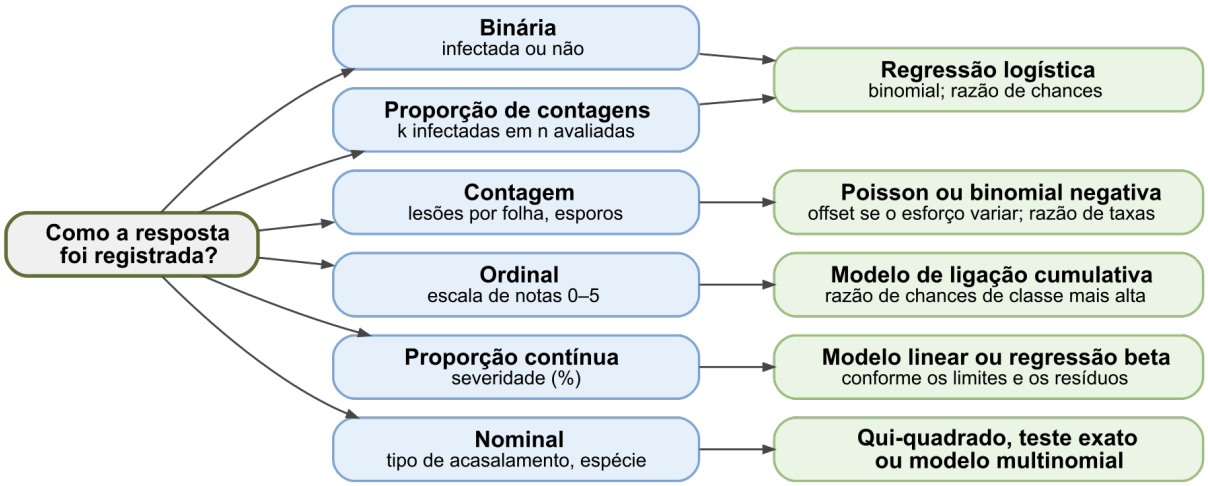


Figura 10.4. Do tipo de resposta à família de modelos.

10.17 Relatando análises de dados categóricos, de contagem e de proporção

Um bom enunciado de resultados geralmente inclui:

10.17.1 Que tipo de resposta foi analisado?

Descreva como a resposta foi registrada e qual foi a unidade experimental.

10.17.2 Que modelo foi ajustado?

Nomeie a distribuição, a ligação e como a superdispersão e os fatores do delineamento foram tratados.

10.17.3 Efeitos em uma escala interpretável

Relate probabilidades, razões de chances ou razões de taxas, e não coeficientes na escala da ligação.

10.17.4 Incerteza

Relate os intervalos de confiança dos efeitos estimados.

10.17.5 Tamanhos de amostra

Informe quantas unidades e quantos indivíduos por unidade contribuíram para cada estimativa.

10.17.6 Interpretação biológica

Explique o que o efeito estimado significa para o manejo da doença ou para a biologia.

Em vez de:

> A incidência diferiu significativamente entre os tratamentos ($\chi^2 = 63{,}0$, $P < 0{,}001$).

prefira algo como:

> A probabilidade estimada de infecção foi de 0,42 (IC 95%: 0,37 a 0,48) nas parcelas não tratadas e de 0,13 (IC 95%: 0,10 a 0,17) nas parcelas com sementes tratadas, uma redução de 29 pontos percentuais (IC 95%: 23 a 36 pontos; razão de chances: 0,20; IC 95%: 0,14 a 0,31).

Esses números vêm do exemplo simulado da regressão logística (saídas de `emmeans()` e dos contrastes mostrados naquela seção).

10.18 Erros comuns

10.18.1 Analisar proporções sem o número de indivíduos avaliados

A precisão depende do denominador.

10.18.2 Tratar plantas individuais como repetições

A parcela, e não a planta, geralmente é a unidade experimental.

10.18.3 Transformar automaticamente

As transformações arco-seno ou logarítmica não são mais necessárias como padrão.

10.18.4 Somar constantes arbitrárias antes de tomar logaritmos

Os resultados podem depender da constante escolhida.

10.18.5 Ignorar a superdispersão

Os intervalos de confiança ficam estreitos demais.

10.18.6 Relatar coeficientes na escala do logito ou do log

O leitor precisa de probabilidades, razões de chances ou razões de taxas.

10.18.7 Confundir razões de chances com razões de riscos

Elas só são semelhantes quando o desfecho é raro.

10.18.8 Usar testes de qui-quadrado com contagens esperadas pequenas

Testes exatos são mais adequados.

10.18.9 Tratar porcentagens de severidade como contagens binomiais

A severidade é uma proporção contínua.

10.18.10 Calcular médias das classes de escalas ordinais

Os números das classes são rótulos de categorias ordenadas, e não medidas.

10.18.11 Concluir ausência a partir de nenhuma detecção

Uma amostra pequena sem positivos ainda é compatível com uma taxa de infecção considerável.

10.19 Lista de verificação para dados categóricos, de contagem e de proporção

Antes de interpretar uma análise desses dados, pergunte:

1. Que questão biológica está sendo abordada?
2. Como a resposta foi registrada?
3. Ela é binária, uma proporção a partir de contagens, uma contagem, ordinal ou uma proporção contínua?
4. Qual é a unidade experimental?
5. Quantos indivíduos foram avaliados por unidade?
6. O esforço de amostragem difere entre as observações?
7. Que distribuição e que ligação foram usadas?
8. O modelo inclui blocos e outros fatores do delineamento?
9. Há superdispersão ou excesso de zeros?
10. Em que escala o efeito é relatado?
11. Quais são as probabilidades, razões de chances ou razões de taxas estimadas?
12. Quão incertas elas são?
13. Contagens esperadas pequenas ou células com zero foram tratadas adequadamente?
14. O efeito é biologicamente relevante?
15. A conclusão permanece dentro do escopo do estudo?

10.20 Resumo do capítulo

Muitas respostas em Fitopatologia são binárias, contagens, proporções a partir de contagens, categorias ou classes ordenadas. Esses dados são limitados, frequentemente discretos e têm variabilidade que depende da média, de modo que os modelos lineares dos capítulos anteriores, com ou sem transformações, os descrevem mal.

Tabelas de frequência, proporções com intervalos de confiança adequados e testes de qui-quadrado ou exatos fornecem uma primeira descrição dos dados categóricos. Os modelos lineares generalizados oferecem um arcabouço unificado que mantém a estrutura dos modelos lineares — fatores, covariáveis, interações e blocos —, especificando uma distribuição e uma função de ligação adequadas à resposta. A regressão logística descreve probabilidades e produz razões de chances; os modelos de Poisson e binomial negativo descrevem contagens e produzem razões de taxas, com *offsets* para esforços de amostragem desiguais. A superdispersão é comum em dados de campo e deve ser verificada e tratada. Proporções contínuas, como a severidade em porcentagem, e escalas ordinais de doença exigem modelos próprios. Em todos os casos, os resultados são mais úteis quando relatados na escala da resposta original, com sua incerteza e com a unidade experimental claramente identificada.

10.21 Conceitos-chave

Ao concluir este capítulo, você deverá ser capaz de explicar:

- resposta binária;
- proporção a partir de contagens;
- resposta de contagem;
- categorias nominais e ordinais;

- proporção contínua;
- diferença de riscos;
- razão de riscos;
- chance e razão de chances;
- razão de taxas;
- intervalo de confiança de Wilson;
- teste de aderência;
- teste de independência;
- teste exato de Fisher;
- teste de McNemar;
- sensibilidade e especificidade;
- modelo linear generalizado;
- função de ligação;
- regressão logística;
- regressão de Poisson;
- *offset*;
- superdispersão;
- modelo binomial negativo;
- inflação de zeros;
- regressão beta;
- modelo de ligação cumulativa.

10.22 Exercícios

10.22.1 Exercício 1. Que tipo de resposta?

Para cada observação, identifique o tipo de resposta, a unidade experimental e um modelo adequado:

- número de plantas infectadas em 40 por parcela;
- número de lesões por folha;
- porcentagem da área foliar com sintomas;
- nota de doença de 0 a 5 por planta;
- presença ou ausência do patógeno em uma amostra de sementes;
- tipo de acasalamento de isolados de três regiões.

10.22.2 Exercício 2. Uma proporção e sua incerteza

Um lote de sementes foi testado e 0 de 50 sementes estavam infectadas.

1. calcule a taxa de infecção estimada;
2. calcule um intervalo de confiança exato de 95%;
3. discuta se o lote pode ser declarado livre do patógeno;
4. determine quantas sementes precisariam ser testadas para que o limite superior ficasse abaixo de 1%.

10.22.3 Exercício 3. Razão de segregação

Em uma população F_2, 112 plantas foram resistentes e 48 suscetíveis. Teste se os dados são compatíveis com uma razão 3:1 e discuta o que o resultado mostra e o que não mostra sobre a genética da resistência.

10.22.4 Exercício 4. Regressão logística

Usando os dados simulados de tratamento de sementes deste capítulo:

1. ajuste um MLG binomial com o tratamento como variável explicativa;
2. acrescente os blocos ao modelo;
3. obtenha a probabilidade prevista de infecção para cada tratamento;
4. estime a razão de chances e a diferença de probabilidade;
5. verifique se há superdispersão.

10.22.5 Exercício 5. Contagens com esforço desigual

Armadilhas de esporos funcionaram por números diferentes de horas em vários locais.

1. explique por que as contagens brutas não podem ser comparadas diretamente;
2. ajuste um modelo de Poisson com um *offset* para o tempo de captura;
3. verifique a superdispersão e, se necessário, ajuste um modelo binomial negativo;
4. relate as razões de taxas entre os locais.

10.22.6 Exercício 6. Escala de doença

Plantas de quatro cultivares receberam notas em uma escala de 0 a 5.

1. tabule a frequência de cada classe por cultivar;
2. discuta os problemas de calcular a média das notas;
3. ajuste um modelo de ligação cumulativa;
4. interprete o efeito da cultivar.

10.22.7 Exercício 7. Avaliação de um artigo publicado

Escolha um artigo recente de Fitopatologia que analise incidência, contagens ou notas de doença.

Identifique:

1. o tipo de resposta;
2. a unidade experimental e o número de indivíduos por unidade;
3. se os dados foram transformados, e como;
4. se foi usado um MLG ou outro modelo adequado;
5. se a superdispersão foi considerada;
6. em que escala os efeitos foram relatados;
7. se a incerteza foi relatada;
8. se as conclusões são sustentadas pelos efeitos estimados.

10.23 Reflexão

> **Antes de escolher um teste estatístico, pergunte como os dados foram gerados.**

Discuta essa afirmação usando um exemplo de incidência, contagens ou notas de doença da Fitopatologia.

10.24 Próximos passos

Os Capítulos 7 a 10 analisaram uma resposta de cada vez. Muitos estudos em Fitopatologia medem dezenas a milhares de variáveis em cada amostra, como perfis de expressão gênica, metabólitos, características fisiológicas ou comunidades microbianas. O Capítulo 11 apresenta métodos para explorar dados multivariados, incluindo a análise de componentes principais, o agrupamento e os mapas de calor.

CAPÍTULO 11

Como exploramos dados multivariados?

11.1 Objetivos de aprendizagem

! Importante

Ao concluir este capítulo, o leitor deverá ser capaz de:

- reconhecer questões de pesquisa e conjuntos de dados em que muitas variáveis são medidas em cada amostra;
- organizar dados multivariados como uma matriz de amostras por variáveis;
- preparar dados multivariados para a análise, incluindo transformação, centralização e padronização;
- explicar como se definem distâncias e similaridades entre amostras;
- realizar e interpretar a análise de componentes principais, incluindo escores, cargas e variância explicada;
- realizar e interpretar o agrupamento hierárquico e o agrupamento por *k*-médias;
- construir e ler mapas de calor;
- distinguir a busca exploratória de padrões da inferência confirmatória;
- reconhecer os problemas de testes múltiplos e de efeitos de lote em dados de alta dimensionalidade;
- comunicar resultados multivariados sem exagerá-los.

11.2 Por que dados multivariados?

Os Capítulos 7 a 10 analisaram uma resposta de cada vez. Cada vez mais, os fitopatologistas medem muitas variáveis em cada amostra.

Exemplos típicos incluem:

- **expressão gênica**: a abundância de milhares de transcritos em plantas inoculadas e em plantas inoculadas apenas com o veículo (*mock*);
- **painéis moleculares direcionados**: a expressão de um conjunto de genes relacionados à defesa medida por qPCR;
- **metabolômica e proteômica**: centenas de compostos ou proteínas por amostra;
- **perfis fisiológicos**: taxa fotossintética, condutância estomática, fluorescência da clorofila e atividades enzimáticas medidas nas mesmas plantas;
- **fenótipos de isolados**: taxa de crescimento, esporulação, agressividade e sensibilidade a vários fungicidas para cada isolado;

- **comunidades microbianas**: a abundância relativa de centenas de táxons em cada amostra de solo ou de folha;
- **dados espectrais**: a reflectância em centenas de comprimentos de onda de folhas ou dosséis.

Nesses conjuntos de dados, as perguntas muitas vezes não dizem respeito a uma única variável, mas a **padrões entre muitas variáveis**:

- Quais amostras se parecem entre si?
- Os tratamentos separam as amostras no espaço definido por todas as variáveis?
- Quais variáveis mudam juntas?
- Há grupos de amostras ou de variáveis que não foram previstos?

> ! **Importante**
>
> Os métodos multivariados são, antes de tudo, ferramentas para **enxergar a estrutura** de dados complexos. Eles resumem e revelam padrões; não testam hipóteses por si sós nem estabelecem que os padrões são biologicamente reais.

11.3 Da questão biológica à exploração multivariada

Considere um experimento que mede a expressão de genes relacionados à defesa em cultivares resistentes e suscetíveis, inoculadas ou não.



Figura 11.1. Da questão biológica à exploração multivariada: o produto são padrões e hipóteses, e não uma única estimativa.

11.3.1 Exploração e confirmação

O Capítulo 4 distinguiu a análise exploratória da confirmatória. A distinção é especialmente importante aqui. Com muitas variáveis, alguns padrões aparecerão apenas por acaso, e os

métodos multivariados sempre produzirão uma imagem. Os padrões encontrados na exploração tornam-se hipóteses a serem avaliadas com delineamentos e análises adequados, e não conclusões.

11.4 Organizando dados multivariados

11.4.1 A matriz de dados

Os métodos multivariados operam sobre uma **matriz de dados** em que:

- cada **linha** é uma amostra (planta, parcela, isolado, folha);
- cada **coluna** é uma variável (gene, metabólito, característica, táxon);
- uma tabela separada de **informações das amostras** registra tratamentos, cultivares, blocos e outras variáveis do delineamento.

```
           gene1  gene2  gene3  ...  gene_p
sample1     5.2    0.8    3.1   ...
sample2     4.9    1.2    2.7   ...
...
sample_n
```

Esse formato largo difere do formato longo e organizado recomendado no Capítulo 3 para a maioria das análises. Ambos são úteis: o formato organizado é a melhor forma de guardar e documentar os dados, e a matriz larga é derivada dele para a análise multivariada.

```
library(tidyr)

mat ← expr_long ▷
  pivot_wider(names_from = gene, values_from = expression) ▷
  tibble::column_to_rownames("sample") ▷
  as.matrix()
```

11.4.2 Muitas variáveis, poucas amostras

Os conjuntos de dados multivariados em Fitopatologia frequentemente têm muito mais variáveis do que amostras: milhares de genes medidos em uma dúzia de plantas. Nessa situação:

- muitos padrões podem ser encontrados por acaso;
- métodos que estimam um parâmetro por variável tornam-se instáveis;
- a exploração visual e a redução de dimensionalidade são essenciais;
- o número de repetições biológicas continua determinando o quanto se pode concluir.

> ⚠ **Aviso**
>
> Medir milhares de variáveis não aumenta o número de repetições. Um experimento com três plantas por tratamento continua sendo um experimento com três repetições, não importa quantos genes sejam medidos em cada planta.

11.5 Preparando os dados

11.5.1 Valores perdidos e controle de qualidade

Antes de qualquer análise multivariada, verifique valores perdidos, amostras que falharam e variáveis sem variação. Muitos métodos multivariados não lidam diretamente com valores perdidos, e a forma como eles são removidos ou imputados pode afetar os resultados.

11.5.2 Transformação

Muitas variáveis multivariadas são fortemente assimétricas à direita: níveis de expressão, abundâncias de metabólitos e contagens abrangem várias ordens de grandeza. Uma transformação logarítmica geralmente torna sua variação mais comparável e impede que algumas variáveis muito abundantes dominem a análise.

```
mat_log ← log2(mat + 1)
```

Para contagens de sequenciamento, métodos especializados de normalização e transformação levam em conta as diferenças de profundidade de sequenciamento entre as amostras.

11.5.3 Centralização e padronização

Quando as variáveis são medidas em unidades diferentes, como taxa fotossintética (µmol m^{-2} s^{-1}), teor de clorofila (unidades SPAD) e tamanho de lesão (mm), suas variâncias não são comparáveis. Sem padronização, as variáveis com maior variância numérica dominam a análise.

Padronizar cada variável para média zero e desvio padrão um dá a todas as variáveis o mesmo peso:

```
mat_scaled ← scale(mat_log)
```

Padronizar ou não é uma decisão com consequências. Quando todas as variáveis estão na mesma escala e a magnitude de sua variação tem significado, como nos valores de expressão em log, a padronização pode ser desnecessária ou até indesejável, porque amplifica variáveis que quase não mudam.

> **Dica**
>
> Informe como os dados foram transformados, centralizados e padronizados. Escolhas diferentes podem produzir ordenações, agrupamentos e mapas de calor visivelmente diferentes a partir dos mesmos dados.

11.6 Distância e similaridade

Muitos métodos multivariados se baseiam em uma medida de quão diferentes duas amostras são.

11.6.1 Distância euclidiana

A distância em linha reta entre amostras no espaço definido pelas variáveis. É a distância natural para dados contínuos e adequadamente padronizados, e é a base da análise de componentes principais e do agrupamento por *k*-médias.

11.6.2 Distância baseada em correlação

Amostras ou variáveis são consideradas semelhantes se seus perfis sobem e descem juntos, independentemente de seus níveis absolutos. É comum no agrupamento de genes com padrões de expressão semelhantes.

11.6.3 Distâncias para dados de comunidades

Para abundâncias de táxons em comunidades microbianas, medidas como a dissimilaridade de Bray–Curtis são preferidas, porque não são influenciadas por ausências compartilhadas.

11.6.4 Tipos de variáveis misturados

Quando as variáveis incluem dados contínuos, categóricos e binários, como as características de isolados, a distância de Gower os combina em uma única medida.

```
d_euc ← dist(mat_scaled)                          # euclidiana
d_cor ← as.dist(1 - cor(t(mat_scaled)))           # baseada em correlação

library(vegan)
d_bc ← vegdist(community, method = "bray")        # Bray-Curtis
```

11.7 Análise de componentes principais

11.7.1 A ideia

Quando muitas variáveis estão correlacionadas, boa parte da variação entre as amostras pode ser descrita por poucas direções subjacentes. A análise de componentes principais (ACP, ou PCA na sigla em inglês) encontra essas direções:

- o **primeiro componente principal** (PC1) é a direção ao longo da qual as amostras mais variam;
- o **segundo componente principal** (PC2) é a direção de maior variação restante, perpendicular ao PC1;
- e assim por diante, até tantos componentes quantas forem as variáveis ou as amostras.

Cada componente principal é uma combinação ponderada das variáveis originais. Projetar as amostras nos dois ou três primeiros componentes produz uma imagem de baixa dimensionalidade dos padrões dominantes nos dados.

11.7.2 Um exemplo simples

O código a seguir simula a expressão de 50 genes em 12 amostras de um delineamento dois por dois.

```
set.seed(606)

samples ← expand.grid(
  rep = 1:3,
  inoculation = c("mock", "inoculated"),
  cultivar = c("susceptible", "resistant")
)

n ← nrow(samples)
p ← 50
```

```
mat ← matrix(rnorm(n * p, mean = 6, sd = 1), nrow = n,
             dimnames = list(paste0("s", 1:n), paste0("gene", 1:p)))

# os genes 1–10 respondem à inoculação, mais fortemente na cultivar resistente
effect ← ifelse(samples$inoculation == "inoculated",
                ifelse(samples$cultivar == "resistant", 3, 1.5), 0)

mat[, 1:10] ← mat[, 1:10] + effect
```

11.7.3 Ajustando uma ACP no R

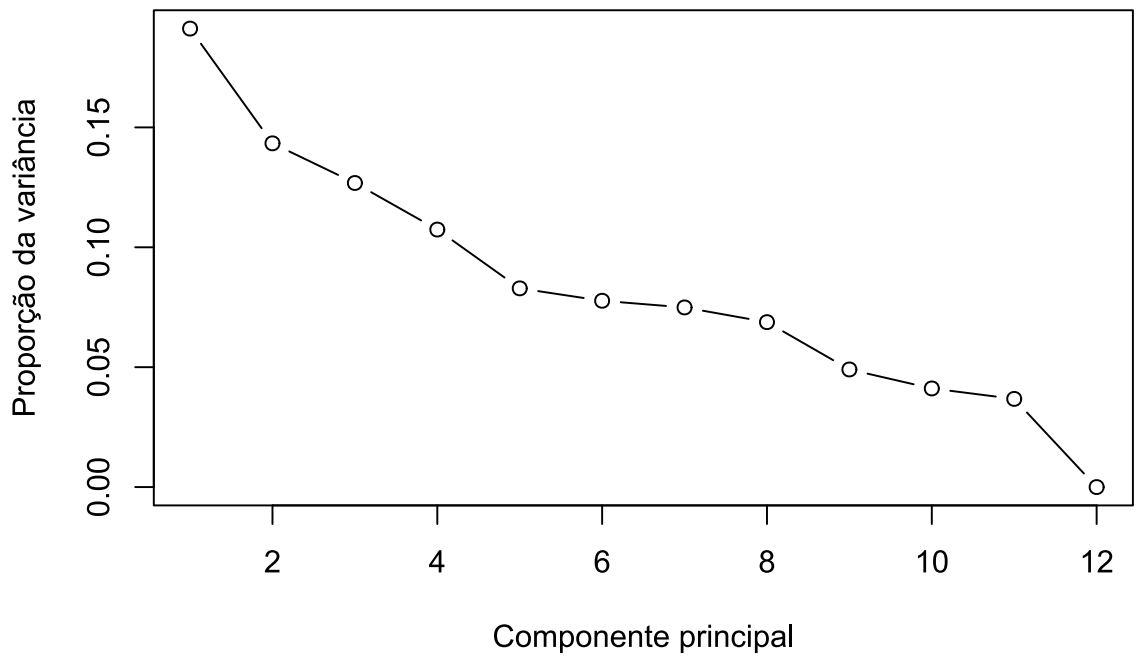
```
pca ← prcomp(mat, center = TRUE, scale. = TRUE)

summary(pca)
```

```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5    PC6     PC7
Standard deviation     3.0920 2.6772 2.5181 2.3173 2.03560 1.9710 1.93543
Proportion of Variance 0.1912 0.1434 0.1268 0.1074 0.08287 0.0777 0.07492
Cumulative Proportion  0.1912 0.3346 0.4614 0.5688 0.65164 0.7293 0.80426
                           PC8     PC9    PC10    PC11     PC12
Standard deviation     1.85453 1.56600 1.43408 1.35602 2.05e-15
Proportion of Variance 0.06879 0.04905 0.04113 0.03678 0.00e+00
Cumulative Proportion  0.87305 0.92209 0.96322 1.00000 1.00e+00
```

11.7.4 Variância explicada

A proporção da variância total explicada por cada componente indica quanto da estrutura dos dados cada componente captura. Um **gráfico de sedimentação** (*scree plot*) mostra essas proporções em ordem decrescente.

```
var_explained ← pca$sdev^2 / sum(pca$sdev^2)

plot(var_explained, type = "b",
     xlab = "Componente principal", ylab = "Proporção da variância")
```

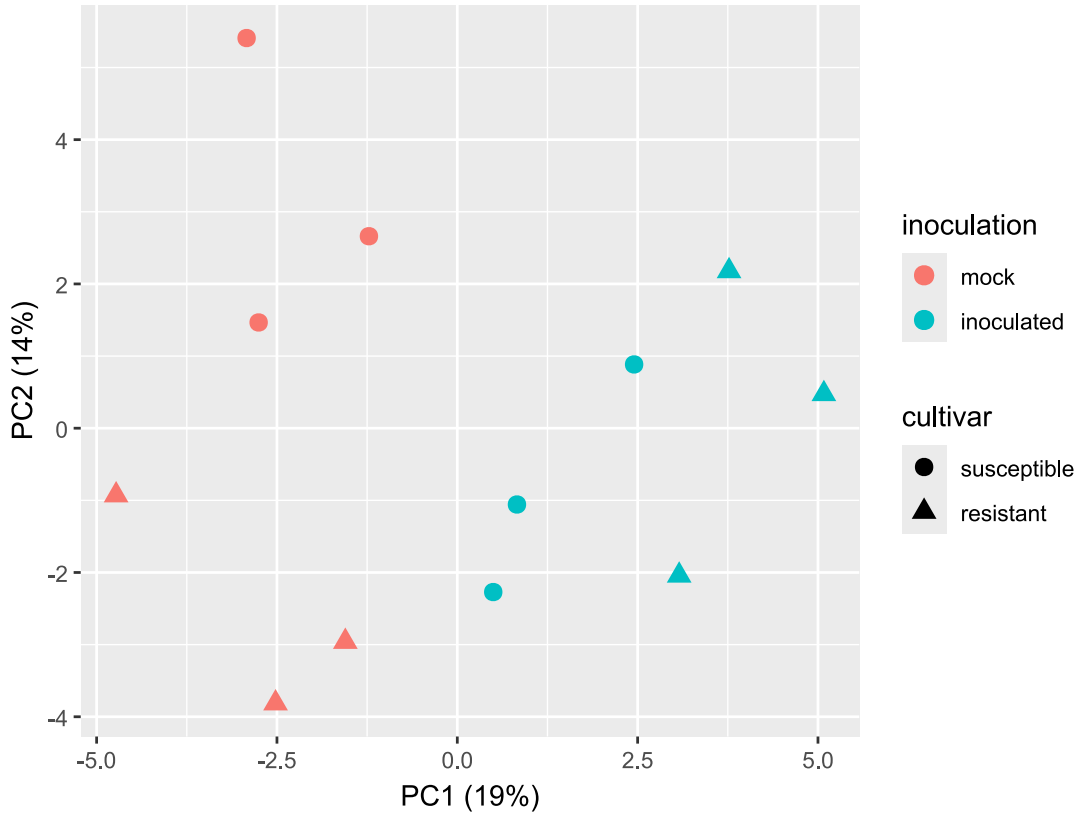


Em conjuntos de dados com muitas variáveis, é comum que PC1 e PC2 juntos expliquem uma parcela modesta da variância. Isso não significa que a análise falhou; significa que a variação está distribuída em muitas direções.

11.7.5 Escores: onde estão as amostras?

Os **escores** são as coordenadas das amostras nos componentes principais. Representá-los graficamente, coloridos pelas variáveis do delineamento, mostra se as amostras se agrupam por tratamento, por cultivar ou por outra coisa.

```
library(ggplot2)

scores ← cbind(samples, as.data.frame(pca$x[, 1:2]))

ggplot(scores, aes(PC1, PC2, color = inoculation, shape = cultivar)) +
  geom_point(size = 3) +
  labs(
    x = sprintf("PC1 (%.0f%%)", 100 * var_explained[1]),
    y = sprintf("PC2 (%.0f%%)", 100 * var_explained[2])
  ) +
  coord_equal()
```



11.7.6 Cargas: quais variáveis importam?

As **cargas** (*loadings*) são os pesos de cada variável original em cada componente. Variáveis com grandes cargas positivas ou negativas no PC1 são as que mais contribuem para a separação das amostras ao longo do PC1.

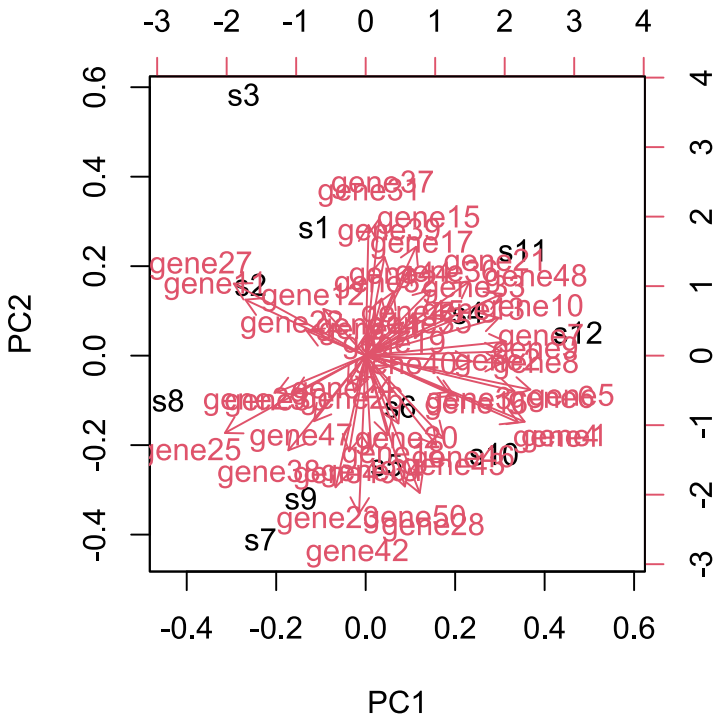
```
loadings ← pca$rotation[, 1:2]

head(loadings[order(-abs(loadings[, "PC1"])), ], 10)
```

```
             PC1          PC2
gene5   0.2764629 -0.065783837
gene1   0.2669092 -0.128121604
gene4   0.2558371 -0.129158483
gene6   0.2497328 -0.068759312
gene25 -0.2362822 -0.149913113
gene7   0.2355345  0.027272263
gene8   0.2287761 -0.014442768
gene48  0.2287524  0.119776907
gene9   0.2278760  0.007877981
gene10  0.2235339  0.073279846
```

11.7.7 Biplots

Um ***biplot*** mostra juntos os escores das amostras e as cargas das variáveis, com as variáveis desenhadas como setas. Ele mostra quais variáveis estão associadas a quais grupos de amostras. Os *biplots* tornam-se ilegíveis com muitas variáveis; funcionam melhor com dezenas de variáveis, como características fisiológicas, do que com milhares de genes.

```
biplot(pca)
```



11.7.8 Interpretando uma ACP

Ao interpretar uma ACP, lembre-se de que:

- a ACP é **não supervisionada**: ela não sabe que tratamento cada amostra recebeu. Se os tratamentos se separam ao longo de um componente, o efeito do tratamento é uma das principais fontes de variação nos dados;
- a ausência de separação não significa que os tratamentos não têm efeito sobre nenhuma variável; significa que seu efeito não está entre as fontes dominantes de variação;
- o sinal de um componente é arbitrário; apenas as posições relativas importam;
- os eixos devem ser desenhados com a mesma escala, para que as distâncias não sejam distorcidas;
- a ACP descreve combinações lineares das variáveis e pode não captar estruturas não lineares.

> **! Importante**
>
> Os primeiros componentes frequentemente refletem as maiores fontes de variação do experimento, e elas nem sempre são biológicas. Representar os escores da ACP coloridos por data de processamento, lote de extração ou corrida de sequenciamento é uma das formas mais eficazes de detectar problemas técnicos.

11.7.9 Outros métodos de ordenação

A ACP se baseia em distâncias euclidianas. Quando outra distância é mais adequada, outros métodos de ordenação representam as amostras em um espaço de baixa dimensionalidade:

- a **análise de coordenadas principais** (PCoA) funciona com qualquer distância, como a de Bray–Curtis para comunidades;
- o **escalonamento multidimensional não métrico** (NMDS) preserva a ordem das distâncias, e não seus valores.

Esses métodos são comuns em estudos de comunidades microbianas.

11.8 Agrupamento

11.8.1 A ideia

Os métodos de agrupamento reúnem amostras (ou variáveis) de modo que os membros de um mesmo grupo sejam mais semelhantes entre si do que aos membros de outros grupos. Diferentemente da ACP, que mostra variação contínua, o agrupamento impõe grupos discretos.

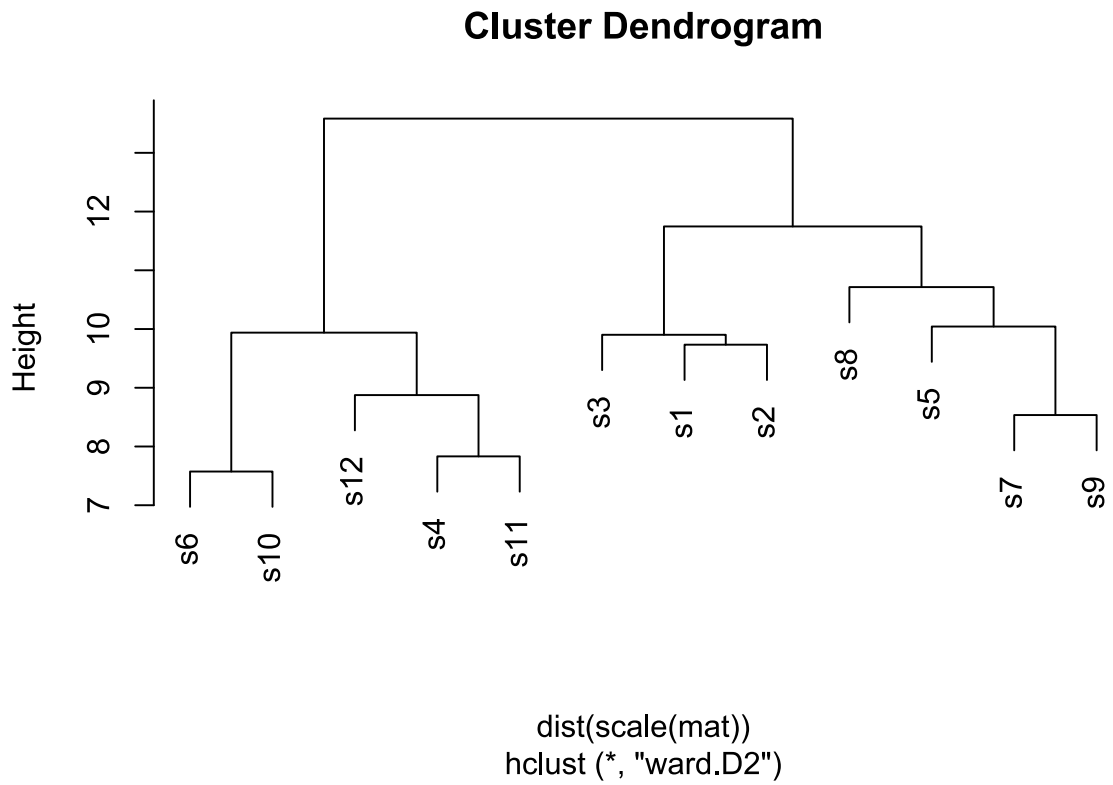
11.8.2 Agrupamento hierárquico

O agrupamento hierárquico começa com cada amostra como um grupo e funde sucessivamente os grupos mais semelhantes. O resultado é um **dendrograma** que mostra a ordem e a distância em que os grupos foram fundidos.

A forma como se definem as distâncias entre grupos, o **método de ligação**, afeta o resultado:

- a **ligação completa** usa a maior distância entre os membros;
- a **ligação média** usa a distância média;
- o **método de Ward** funde os grupos de modo a minimizar o aumento da variância dentro dos grupos e tende a produzir grupos compactos.

```
hc ← hclust(dist(scale(mat)), method = "ward.D2")

plot(hc)
```




```
groups ← cutree(hc, k = 2)
table(groups, samples$inoculation)
```

```
groups mock inoculated
     1    6          1
     2    0          5
```

11.8.3 Agrupamento por *k*-médias

O método *k*-médias divide as amostras em um número especificado de grupos, k, atribuindo cada amostra ao grupo com o centro mais próximo. O resultado depende de k e das posições iniciais aleatórias, de modo que o algoritmo deve ser executado a partir de vários inícios.

```
km ← kmeans(scale(mat), centers = 2, nstart = 25)

table(km$cluster, samples$inoculation)
```

```
   mock inoculated
1    0          6
2    6          0
```

11.8.4 Quantos grupos?

Raramente existe um único número correto de grupos. Critérios como a soma de quadrados dentro dos grupos (o gráfico do "cotovelo") e as larguras de silhueta podem orientar a escolha, mas o conhecimento biológico e o objetivo da análise importam mais.

```
library(cluster)

sil ← silhouette(km$cluster, dist(scale(mat)))
summary(sil)
```

```
Silhouette of 12 units in 2 clusters from silhouette.default(x = km$cluster, dist =
dist(scale(mat))) :
 Cluster sizes and average silhouette widths:
         6          6
0.12918363 0.02333586
Individual silhouette widths:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.01024  0.01962  0.05549  0.07626  0.13146  0.17607
```

11.8.5 O que os agrupamentos mostram e o que não mostram

Os algoritmos de agrupamento sempre produzem grupos, mesmo a partir de dados sem nenhuma estrutura de grupos. Um dendrograma, portanto, não demonstra que existem grupos distintos.

Perguntas úteis incluem:

- Os grupos são estáveis quando os dados são reamostrados ou o método é ligeiramente alterado?
- Eles correspondem a variáveis biológicas ou do delineamento conhecidas?
- São determinados por poucas variáveis ou por muitas?
- Poderiam refletir fatores técnicos, como o lote?

> ⚠ **Aviso**
>
> Grupos encontrados em um conjunto de dados não são evidência independente de grupos biológicos. Não teste diferenças entre os grupos usando as mesmas variáveis usadas para defini-los.

11.9 Mapas de calor

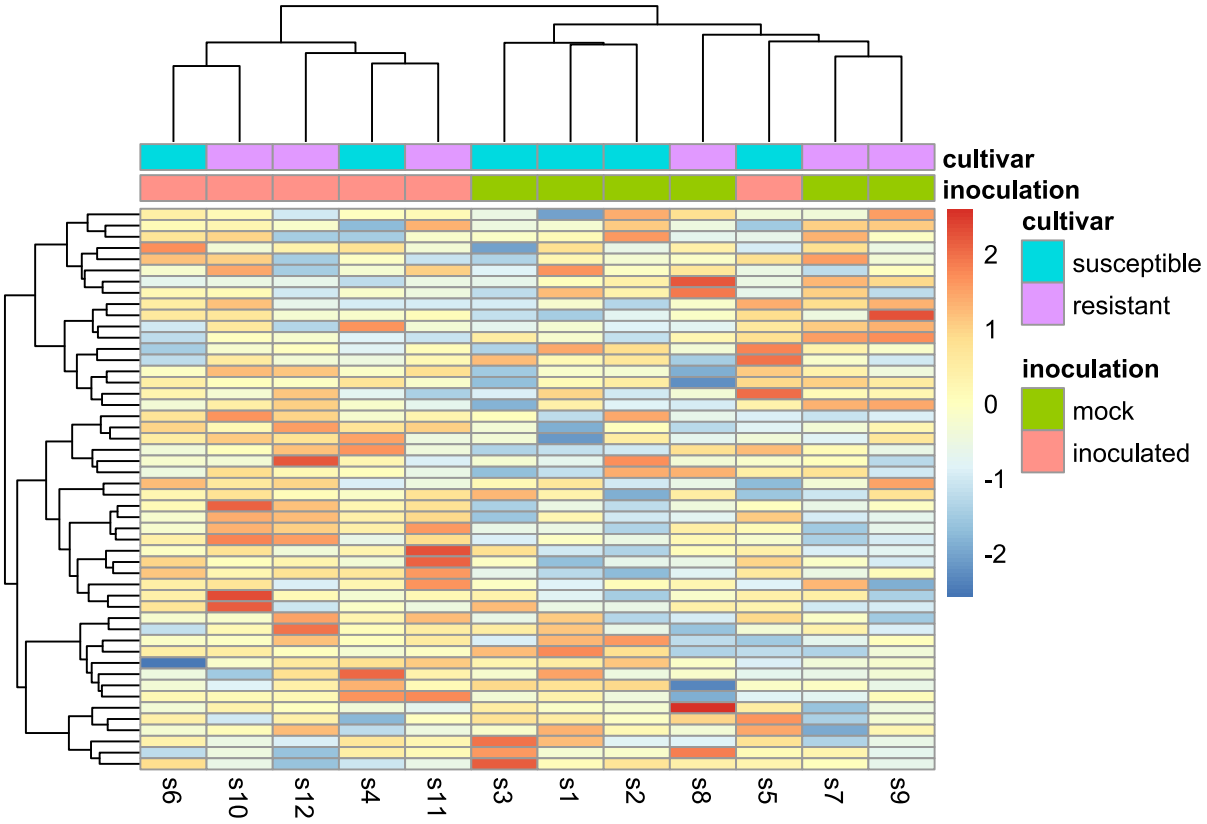
11.9.1 Lendo um mapa de calor

Um mapa de calor (*heatmap*) mostra a matriz de dados como uma grade de células coloridas, com as amostras em uma dimensão e as variáveis na outra. As linhas e colunas geralmente são reordenadas por agrupamento hierárquico, e dendrogramas são desenhados nas margens. Barras de anotação mostram variáveis do delineamento, como tratamento e cultivar.

Os mapas de calor estão entre as figuras mais comuns na Fitopatologia molecular, porque mostram tanto as amostras quanto as variáveis que as distinguem.

11.9.2 Construindo um mapa de calor

```
library(pheatmap)

annotation ← data.frame(
  inoculation = samples$inoculation,
  cultivar = samples$cultivar,
  row.names = rownames(mat)
)

pheatmap(
  t(mat),
  scale = "row",
  annotation_col = annotation,
  clustering_method = "ward.D2",
  show_rownames = FALSE
)
```



11.9.3 Padronizando as linhas

Quando cada linha é um gene, os valores frequentemente são padronizados dentro das linhas (escores z), para que as cores mostrem se cada gene está relativamente alto ou baixo em cada amostra. A padronização das linhas remove as diferenças de nível geral de expressão entre os genes; um gene expresso em níveis muito baixos pode parecer tão “induzido” quanto um gene muito expresso.

11.9.4 Cor

Os princípios do Capítulo 5 se aplicam:

- use uma paleta **divergente**, centrada em zero, para escores z ou *log fold changes*;

- use uma paleta **sequencial** para valores que apenas aumentam, como abundâncias;
- evite paletas vermelho–verde, que muitos leitores não conseguem distinguir;
- inclua uma legenda de cores com as unidades ou a padronização usada.

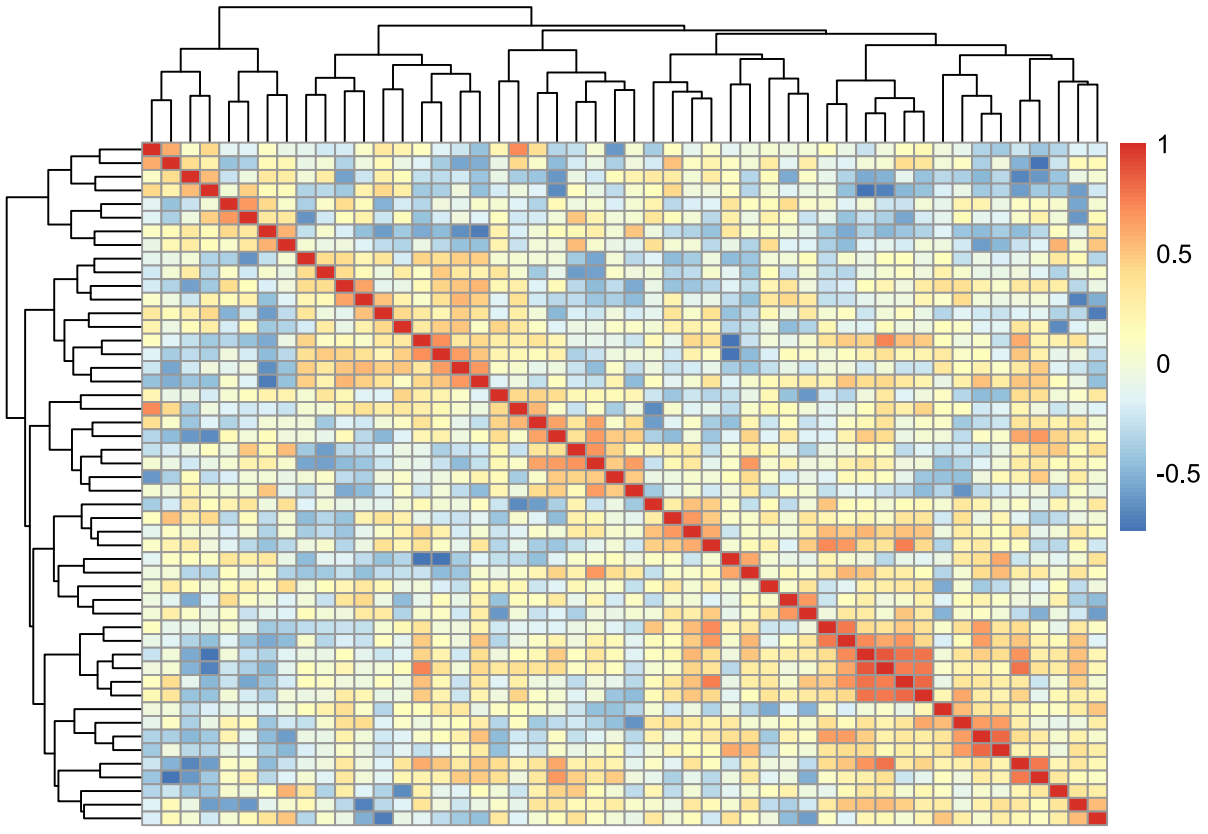
11.9.5 O que os mapas de calor podem esconder

Os mapas de calor enfatizam padrões, mas escondem magnitudes, incerteza e o número de repetições. O agrupamento produzirá blocos visualmente marcantes mesmo em dados aleatórios. Por isso, os mapas de calor são mais bem usados junto com as análises que quantificam as diferenças nas variáveis de interesse.

11.10 Relações entre variáveis

As matrizes de correlação resumem como as variáveis mudam juntas, estendendo as ideias do Capítulo 8 a muitas variáveis ao mesmo tempo. Elas podem revelar grupos de genes ou características que respondem juntos e redundâncias entre as variáveis medidas.

```
cor_mat ← cor(mat_scaled)

pheatmap(cor_mat, show_rownames = FALSE, show_colnames = FALSE)
```



Com muitas variáveis e poucas amostras, muitas correlações parecerão fortes apenas por acaso, e as matrizes de correlação devem ser lidas como exploratórias.

11.11 Da exploração à inferência

11.11.1 Testando muitas variáveis

Um passo comum depois da exploração é testar cada variável separadamente, por exemplo para identificar genes cuja expressão difere entre plantas inoculadas e plantas *mock*. Cada teste é uma comparação entre grupos como as dos Capítulos 7 e 10.

Com milhares de testes, muitos valores-P pequenos ocorrem por acaso. Se 5.000 genes forem testados com nível de significância de 5% e nenhum responder de fato, cerca de 250 ainda parecerão significativos.

11.11.2 Controlando as falsas descobertas

A **taxa de falsas descobertas** (FDR) é a proporção esperada de falsos positivos entre as variáveis declaradas significativas. Ajustar os valores-P para controlar a FDR é prática-padrão em genômica.

```
# um teste t por gene: inoculadas x mock
p_values ← apply(mat, 2, function(g) t.test(g ~ samples$inoculation)$p.value)
sum(p_values < 0.05)
```

```
[1] 10
```

```
p_adj ← p.adjust(p_values, method = "BH")

sum(p_adj < 0.05)
```

```
[1] 2
```

```
# genes significativos e seus log2 fold changes (inoculadas - mock)
lfc ← colMeans(mat[samples$inoculation == "inoculated", ]) -
  colMeans(mat[samples$inoculation == "mock", ])
round(lfc[p_adj < 0.05], 2)
```

```
gene5 gene6
 2.51  2.58
```

Métodos especializados para dados de expressão gênica melhoram o teste de cada gene isoladamente ao compartilhar informação entre os genes, o que é particularmente valioso com poucas repetições.

> **! Importante**
>
> Como nos capítulos anteriores, o número de variáveis significativas é menos informativo do que o tamanho das mudanças. Relate os tamanhos de efeito, como os *log fold changes*, junto com os valores-P ajustados.

11.11.3 Testando diferenças entre grupos no espaço multivariado

A análise de variância multivariada por permutação (PERMANOVA) testa se grupos de amostras diferem em sua posição no espaço multivariado, usando qualquer medida de distância. É muito usada em estudos de comunidades microbianas.

```
library(vegan)

adonis2(dist(mat_scaled) ~ cultivar * inoculation, data = samples)
```

```
Permutation test for adonis under reduced model
Permutation: free
Number of permutations: 999

adonis2(formula = dist(mat_scaled) ~ cultivar * inoculation, data = samples)
          Df SumOfSqs      R2      F Pr(>F)
Model      3   196.86 0.35793 1.4865  0.002 **
Residual   8   353.14 0.64207
Total     11   550.00 1.00000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A PERMANOVA é sensível tanto a diferenças de dispersão quanto a diferenças de posição. Um resultado significativo pode significar que os grupos diferem em sua variabilidade, e não em sua composição média, o que deve ser verificado separadamente.

11.11.4 Métodos supervisionados

Métodos como a análise discriminante e a análise discriminante por mínimos quadrados parciais (PLS-DA) encontram combinações de variáveis que melhor separam grupos predefinidos. Com muitas variáveis e poucas amostras, eles separam quase qualquer grupo, inclusive grupos aleatórios. Seus resultados só têm significado quando validados em amostras não usadas para ajustá-los.

> ⚠ **Aviso**
>
> Um gráfico em que um método supervisionado separa perfeitamente os tratamentos não é evidência de efeito do tratamento. Pode mostrar apenas que o método tinha variáveis suficientes para separar qualquer coisa.

11.12 Questões de delineamento em estudos multivariados

11.12.1 Repetição biológica e técnica

Os princípios do Capítulo 2 se aplicam integralmente. As repetições biológicas (plantas ou parcelas independentes) fornecem informação sobre a população; as repetições técnicas (extrações ou corridas repetidas da mesma amostra) fornecem informação sobre o erro de medida. Tratar repetições técnicas como biológicas infla a evidência aparente.

11.12.2 Efeitos de lote

As amostras são frequentemente processadas em lotes: dias de extração, placas, canaletas de sequenciamento. Os lotes podem introduzir diferenças sistemáticas maiores do que os efeitos biológicos de interesse.

Quando os lotes coincidem com os tratamentos — por exemplo, todas as amostras inoculadas extraídas em um dia e todas as amostras *mock* em outro —, os efeitos do tratamento e do lote não podem ser separados por nenhuma análise.

! Importante

Casualize as amostras entre os lotes e registre a informação de lote. Efeitos de lote equilibrados entre os tratamentos podem ser considerados na análise; efeitos de lote confundidos com os tratamentos, não.

11.13 Pesquisa na prática: expressão de genes de defesa em cultivares resistentes e suscetíveis

Pesquisa revisitada: Plantas resistentes respondem de forma diferente à infecção?

Considere um experimento em que folhas de uma cultivar resistente e de uma suscetível foram inoculadas com um patógeno ou inoculadas apenas com o veículo, com várias repetições biológicas por combinação, e a expressão de um painel de genes relacionados à defesa foi medida.

A questão científica não é simplesmente:

> As amostras se agrupam por tratamento?

Em vez disso, pergunte:

1. Qual é a unidade experimental, e quantas repetições biológicas havia?
2. Como os valores de expressão foram normalizados, transformados e padronizados?
3. Quais são as fontes dominantes de variação em uma ACP, e elas correspondem à cultivar, à inoculação ou ao lote de processamento?
4. Quais genes mais contribuem para a separação das amostras?
5. A resposta à inoculação é maior em uma cultivar do que na outra, e em quais genes?
6. Qual é o tamanho das mudanças de expressão, e quantos genes continuam significativos depois de controlar a taxa de falsas descobertas?
7. Quais padrões são robustos o suficiente para orientar experimentos direcionados de acompanhamento?

11.14 Um fluxo de trabalho prático para dados multivariados



Figura 11.2. Fluxo de trabalho para dados multivariados.

11.15 Relatando análises multivariadas

Uma boa descrição geralmente inclui:

11.15.1 O que foi medido?

O número de amostras, o número de variáveis e o delineamento.

11.15.2 Como os dados foram preparados?

Normalização, transformação, tratamento dos valores perdidos, centralização e padronização.

11.15.3 Quais métodos foram usados?

Medida de distância, método de ordenação, algoritmo de agrupamento e método de ligação.

11.15.4 Quanto da variação foi resumido?

A proporção da variância explicada pelos componentes mostrados.

11.15.5 Quão robustos são os padrões?

Estabilidade dos agrupamentos e verificação de efeitos de lote.

11.15.6 O que foi testado?

Número de testes, método de ajuste e tamanhos de efeito.

11.15.7 Interpretação biológica

Quais padrões são interpretados biologicamente e quais são tratados como hipóteses.

Em vez de:

> A ACP separou claramente os tratamentos, demonstrando uma forte resposta de defesa.

prefira algo como:

> O PC1 (19% da variância) separou as amostras inoculadas das amostras *mock*, com as maiores cargas para os genes 1 a 10. A separação foi maior na cultivar resistente. Dos 50 genes testados (inoculadas *versus mock*, três repetições biológicas por combinação), 10 tiveram $P < 0,05$, mas apenas 2 (genes 5 e 6) permaneceram significativos após o ajuste pela FDR, ambos com log_2 *fold change* de cerca de 2,5.

Esses números vêm das saídas da ACP e dos testes por gene apresentados neste capítulo. Eles também ilustram um ponto central: com apenas três repetições por grupo, o controle da taxa de falsas descobertas deixa poucos genes significativos, embora dez genes tenham sido simulados com respostas reais à inoculação.

11.16 Erros comuns

11.16.1 Tratar muitas variáveis como muitas repetições

A repetição depende do número de amostras independentes.

11.16.2 Não padronizar variáveis medidas em unidades diferentes

As variáveis com maior variância numérica dominam.

11.16.3 Não informar as escolhas de transformação e padronização

Escolhas diferentes produzem imagens diferentes.

11.16.4 Interpretar a separação na ACP como um teste estatístico

A ACP descreve a variação; não a testa.

11.16.5 Ignorar a variância explicada pelos componentes mostrados

Dois componentes podem resumir apenas uma pequena parte dos dados.

11.16.6 Tratar agrupamentos como grupos comprovados

O agrupamento sempre encontra grupos.

11.16.7 Testar os grupos com as variáveis usadas para criá-los

O teste é circular.

11.16.8 Ler as cores do mapa de calor como magnitudes após padronizar as linhas

As cores após a padronização das linhas mostram níveis relativos, e não absolutos.

11.16.9 Usar escalas de cor vermelho–verde

Muitos leitores não conseguem distingui-las.

11.16.10 Ignorar os testes múltiplos

Muitas variáveis produzem muitos falsos positivos.

11.16.11 Confundir lote com tratamento

Nenhuma análise consegue separá-los depois.

11.16.12 Confiar na separação perfeita obtida por métodos supervisionados

Com muitas variáveis, qualquer agrupamento pode ser separado.

11.17 Lista de verificação para análises multivariadas

Antes de interpretar uma análise multivariada, pergunte:

1. Que questão biológica está sendo abordada?
2. Quais são as amostras e quais são as variáveis?
3. Quantas repetições biológicas existem?
4. As amostras foram processadas em lotes, e os lotes estão confundidos com os tratamentos?
5. Como os valores perdidos foram tratados?
6. Como os dados foram normalizados, transformados e padronizados?
7. Qual medida de distância é adequada?
8. Quanto da variância os componentes mostrados explicam?
9. Quais variáveis determinam os padrões?
10. Os padrões correspondem ao delineamento, à biologia ou a fatores técnicos?
11. Os agrupamentos são estáveis?
12. As linhas ou colunas do mapa de calor foram padronizadas, e como?
13. Os testes múltiplos foram considerados?
14. Os tamanhos de efeito são relatados junto com a significância?
15. Quais conclusões são confirmadas e quais são hipóteses?

11.18 Resumo do capítulo

Muitos estudos em Fitopatologia medem numerosas variáveis em cada amostra, da expressão gênica e dos perfis de metabólitos às características fisiológicas e às comunidades microbianas. Os métodos multivariados resumem esses dados e revelam estruturas que não podem ser vistas uma variável de cada vez. Seus resultados dependem de como os dados são organizados, transformados, centralizados e padronizados, e da medida de distância escolhida.

A análise de componentes principais mostra as direções dominantes de variação entre as amostras por meio de escores, cargas e variância explicada. O agrupamento reúne amostras ou variáveis semelhantes, e os mapas de calor mostram toda a matriz de dados junto com esses agrupamentos. Esses métodos são exploratórios: sempre produzem uma imagem, e os padrões podem refletir o acaso, artefatos técnicos ou efeitos de lote, além da biologia. Passar da exploração à inferência exige testes direcionados com controle da taxa de falsas descobertas, atenção aos tamanhos de efeito e, sobretudo, delineamentos com repetição biológica adequada e casualização das amostras entre os lotes de processamento.

11.19 Conceitos-chave

Ao concluir este capítulo, você deverá ser capaz de explicar:

- dados multivariados;
- matriz de dados;
- dados de alta dimensionalidade;
- transformação;
- centralização;
- padronização;
- distância e similaridade;
- distância euclidiana;
- dissimilaridade de Bray–Curtis;
- análise de componentes principais;
- componente principal;
- variância explicada;
- gráfico de sedimentação;
- escores;
- cargas;
- *biplot*;
- ordenação;
- agrupamento hierárquico;
- método de ligação;
- dendrograma;
- agrupamento por *k*-médias;
- silhueta;
- mapa de calor;
- padronização das linhas;
- taxa de falsas descobertas;
- PERMANOVA;
- métodos supervisionados e não supervisionados;
- efeito de lote;

- repetição biológica e técnica.

11.20 Exercícios

11.20.1 Exercício 1. Organizando um conjunto de dados multivariados

Você recebe uma planilha em que cada linha contém um identificador de amostra, o nome de um gene e um valor de expressão.

1. descreva como convertê-la em uma matriz de amostras por genes;
2. descreva a tabela separada de informações das amostras de que você precisaria;
3. explique por que os dois formatos são úteis.

11.20.2 Exercício 2. Padronizar ou não padronizar?

Usando um conjunto de dados de características fisiológicas medidas em unidades diferentes:

1. faça uma ACP sem padronização;
2. faça uma ACP com padronização;
3. compare a variância explicada e as cargas;
4. explique qual análise é mais adequada e por quê.

11.20.3 Exercício 3. Interpretando uma ACP

Usando os dados simulados de expressão gênica deste capítulo:

1. faça uma ACP;
2. represente os escores coloridos pela inoculação e com formatos pela cultivar;
3. relate a variância explicada pelo PC1 e pelo PC2;
4. identifique os genes com as maiores cargas no PC1;
5. descreva o padrão biológico sugerido pelo gráfico.

11.20.4 Exercício 4. Grupos a partir de ruído

Gere uma matriz de valores aleatórios sem nenhuma estrutura.

1. faça um agrupamento hierárquico e desenhe o dendrograma;
2. desenhe um mapa de calor com agrupamento;
3. descreva o que você vê;
4. explique o que isso implica para a interpretação de agrupamentos em dados reais.

11.20.5 Exercício 5. Elaborando um mapa de calor

Usando os dados simulados, construa um mapa de calor com anotações das amostras. Compare versões com e sem padronização das linhas e com duas paletas de cores diferentes, e explique qual versão comunica melhor os resultados.

11.20.6 Exercício 6. Testes múltiplos

Teste cada gene do conjunto de dados simulado quanto à diferença entre amostras inoculadas e *mock*.

1. conte os genes com $P < 0,05$;
2. ajuste os valores-P para controlar a taxa de falsas descobertas;
3. conte os genes que continuam significativos;
4. relate os efeitos estimados dos genes significativos.

11.20.7 Exercício 7. Avaliação de um artigo publicado

Escolha um artigo recente de Fitopatologia que use ACP, agrupamento ou mapas de calor.

Identifique:

1. as amostras e as variáveis;
2. o número de repetições biológicas;
3. como os dados foram transformados e padronizados;
4. se a variância explicada foi relatada;
5. se os efeitos de lote foram considerados;
6. se os agrupamentos foram tratados como hipóteses ou como conclusões;
7. se os testes múltiplos foram considerados;
8. se os tamanhos de efeito foram relatados.

11.21 Reflexão

> **Os métodos multivariados sempre produzem uma imagem. A tarefa científica é decidir o que a imagem significa.**

Discuta essa afirmação usando um exemplo da Fitopatologia molecular.

11.22 Próximos passos

Os Capítulos 7 a 11 trataram da análise de diferentes tipos de dados. Uma análise só cumpre seu propósito quando seus resultados são comunicados com clareza a outras pessoas. O Capítulo 12 discute como preparar figuras, tabelas, resumos estatísticos e interpretações escritas que apresentem a evidência científica de forma exata e eficaz.

CAPÍTULO 12

Como os resultados científicos devem ser comunicados?

12.1 Objetivos de aprendizagem

! Importante

Ao concluir este capítulo, o leitor deverá ser capaz de:

- transformar resultados estatísticos em mensagens científicas claras;
- descrever os métodos estatísticos com detalhe suficiente para que possam ser avaliados e reproduzidos;
- relatar estimativas com unidades, incerteza e precisão numérica adequada;
- usar valores-P de forma complementar, sem substituí-los por afirmações dicotômicas;
- decidir quando usar tabelas, figuras ou texto;
- construir tabelas e figuras de resultados diretamente a partir do R;
- redigir interpretações que distingam evidência estatística de relevância biológica;
- reconhecer e evitar erros comuns na comunicação de resultados.

12.2 Por que a comunicação importa

Uma análise só cumpre seu propósito quando seus resultados chegam a outras pessoas de forma correta e compreensível. Revisores, editores, outros pesquisadores, extensionistas, produtores e gestores tomam decisões com base no que leem em artigos, relatórios e apresentações. Uma análise impecável comunicada de forma ambígua, incompleta ou exagerada produz conclusões erradas tanto quanto uma análise malfeita.

Os capítulos anteriores mostraram como delinear estudos, organizar e explorar dados, construir figuras, raciocinar sobre incerteza e analisar diferentes tipos de dados. Este capítulo trata da etapa final: como transformar tudo isso em texto, tabelas e figuras que permitam ao leitor compreender o que foi feito, o que foi encontrado e quão confiáveis são as conclusões.

Três princípios orientam o capítulo. Primeiro, **a comunicação deve servir à questão biológica**, e não aos procedimentos estatísticos. Segundo, **a evidência deve ser mostrada, e não apenas declarada**: estimativas, incerteza e dados, em vez de adjetivos e asteriscos. Terceiro, **o relato deve ser reprodutível**: outra pessoa deve conseguir entender exatamente como os resultados foram obtidos.

> **! Importante**
> Comunicar resultados não é decorar a análise; é permitir que o leitor avalie a evidência e chegue às próprias conclusões.

12.3 Do resultado estatístico à mensagem científica

12.3.1 Qual é a mensagem principal?

Antes de escrever, identifique a mensagem principal de cada resultado em uma frase, em linguagem biológica. Compare:

- "O efeito do tratamento foi significativo ($P < 0,001$)."
- "Todos os programas de fungicida reduziram a severidade em relação à testemunha, e o programa A foi o mais eficaz, com uma redução de cerca de 31 pontos percentuais."

A primeira frase descreve um procedimento; a segunda descreve um achado. Os números do procedimento continuam importantes, mas como apoio à mensagem, e não como a mensagem.

12.3.2 Público e formato

A mesma análise pode ser comunicada de formas diferentes conforme o público. Em um artigo científico, os leitores esperam detalhes metodológicos completos, estimativas com incerteza e uma discussão das limitações. Em um relatório técnico ou boletim de extensão, a ênfase recai sobre a magnitude prática dos efeitos e as recomendações. Em uma apresentação oral, uma ou duas figuras bem planejadas valem mais do que tabelas extensas. Em todos os casos, porém, as estimativas e sua incerteza devem ser corretas e coerentes entre si.

12.4 Relatando os métodos estatísticos

A seção de métodos deve permitir que um leitor experiente avalie se a análise é adequada e, idealmente, a reproduza. Ela deve informar:

- o **delineamento** do estudo e a **unidade experimental**, incluindo o número de repetições e de subamostras (Capítulo 2);
- a **variável resposta** e como foi medida, com unidades;
- o **modelo estatístico**: tipo de modelo (linear, linear generalizado, não linear, misto), efeitos incluídos (tratamentos, blocos, interações, efeitos aleatórios) e, para MLGs, a distribuição e a função de ligação;
- como os **pressupostos** foram avaliados e o que foi feito quando não foram atendidos;
- as **comparações** de interesse (contrastes planejados, comparações com a testemunha, todas as comparações dois a dois) e o **ajuste para multiplicidade** usado;
- como foram tratados **valores perdidos** e **observações excluídas**, com justificativa;
- os **programas e pacotes** usados, com suas versões.

Um parágrafo de métodos para o ensaio usado neste capítulo poderia ser:

> A severidade da doença (% da área foliar com sintomas) foi analisada com um modelo linear com efeitos fixos de bloco e de tratamento. Os pressupostos de normalidade e homogeneidade das variâncias foram avaliados por gráficos de resíduos. Os efeitos de

> cada programa de fungicida em relação à testemunha foram estimados como diferenças entre médias marginais, com intervalos de confiança de 95% ajustados pelo método de Dunnett. As análises foram feitas no R (versão 4.5.2) com o pacote emmeans (versão 2.0.1).

Note que as versões no parágrafo acima não foram digitadas: foram inseridas automaticamente pelo código R no momento em que este capítulo foi gerado. Esse recurso é discutido adiante.

As versões de todos os pacotes carregados podem ser obtidas com `sessionInfo()`, cuja saída é um bom candidato a material suplementar.

> **Dica**
>
> Descreva a análise que foi feita, e não a análise que se pretendia fazer. Se o modelo foi modificado depois de examinar os resíduos (por exemplo, trocando um modelo linear por um binomial negativo), relate isso e explique por quê.

12.5 Relatando os resultados

12.5.1 Estimativas, incerteza e unidades

O núcleo de um resultado é uma **estimativa com unidades e com sua incerteza** (Capítulo 6). Para o ensaio simulado deste capítulo, os efeitos dos fungicidas em relação à testemunha são:

```
modelo ← lm(severity ~ block + treatment, data = ensaio)
emm ← emmeans(modelo, ~ treatment)
efeitos ← confint(contrast(emm, "trt.vs.ctrl"))
efeitos
```

```
 contrast             estimate   SE df lower.CL upper.CL
 Fung. A - Testemunha    -31.0 3.35 12    -40.4   -21.51
 Fung. B - Testemunha    -26.3 3.35 12    -35.7   -16.81
 Fung. C - Testemunha    -15.1 3.35 12    -24.6    -5.66
 Fung. D - Testemunha    -22.8 3.35 12    -32.3   -13.33

Results are averaged over the levels of: block
Confidence level used: 0.95
Conf-level adjustment: dunnettx method for 4 estimates
```

A partir dessa saída, uma frase de resultados informativa seria: “O programa A reduziu a severidade em 31 pontos percentuais em relação à testemunha (IC 95%: 21 a 40)”. Ela informa a direção (redução), a magnitude (31 pontos), a unidade (pontos percentuais), a comparação (em relação à testemunha) e a incerteza (o intervalo).

12.5.2 Valores-P: como e quando

Valores-P podem complementar as estimativas, mas não substituí-las. Quando relatados:

- informe o **valor exato** (P = 0,03), e não apenas “P < 0,05” ou “n.s.”; para valores muito pequenos, “P < 0,001” é suficiente;
- não use o valor-P como medida do tamanho do efeito;
- não interprete P > 0,05 como evidência de ausência de efeito (Capítulo 6);

- evite asteriscos como única forma de comunicar a evidência.

12.5.3 Precisão numérica e arredondamento

O número de casas decimais deve refletir a precisão da medida e da estimativa, e não a capacidade da calculadora. Uma severidade média de 23,4719% sugere uma precisão inexistente quando as avaliações visuais variam em vários pontos percentuais. Regras práticas úteis:

- relate médias e efeitos com **dois ou três algarismos significativos**, coerentes com o erro padrão;
- relate estatísticas de teste com uma ou duas casas decimais e valores-P com dois algarismos significativos;
- use o **mesmo número de casas decimais** para grandezas comparáveis em uma tabela;
- arredonde apenas no final, nunca em etapas intermediárias dos cálculos.

12.5.4 Linguagem

A forma de escrever influencia o que o leitor conclui. Alguns cuidados:

- **"significativo"** tem um sentido estatístico e um sentido comum ("importante"). Prefira "estatisticamente significativo" quando for esse o sentido, ou, melhor ainda, descreva a magnitude e a incerteza;
- evite **"tendência à significância"** ou "marginalmente significativo" para valores-P próximos de 0,05; descreva a estimativa e seu intervalo;
- use **linguagem causal** ("o fungicida reduziu") apenas quando o delineamento a justificar; em estudos observacionais, prefira "esteve associado a" (Capítulo 6);
- não confunda **"não houve diferença"** com "não houve evidência de diferença".

Tabela 12.1. Exemplos de redação de resultados.

Em vez de	Prefira
O tratamento foi significativo (P < 0,05).	O programa A reduziu a severidade em 31 p.p. (IC 95%: 21 a 40) em relação à testemunha.
Não houve diferença entre B e C (P = 0,08).	A diferença entre B e C foi de 9 p.p., mas o intervalo (−1 a 19) é compatível tanto com nenhuma diferença quanto com uma diferença considerável.
A severidade foi de 18,3472%.	A severidade média foi de 18% (IC 95%: 15 a 22).
A temperatura causou o aumento da doença (levantamento).	A doença foi maior em lavouras com temperaturas mais altas.
Os resultados mostraram uma tendência à significância (P = 0,06).	A estimativa foi de −8 p.p. (IC 95%: −16 a 0,4); os dados são compatíveis com efeitos desde nulos até moderados.

12.6 Tabelas, figuras ou texto?

12.6.1 Quando usar cada um

Cada formato tem sua função:

- **texto**: para um ou dois números importantes, que fazem parte de uma frase;
- **tabelas**: quando os valores exatos importam e precisam ser consultados, ou quando há muitas grandezas diferentes (médias, intervalos, tamanhos de amostra) para vários grupos;
- **figuras**: quando o padrão, a comparação ou a forma de uma relação importam mais do que os valores exatos (Capítulo 5).

Os mesmos resultados não devem ser repetidos em tabela e em figura no corpo do texto; escolha o formato que comunica melhor e, se necessário, coloque o outro no material suplementar.

12.6.2 Anatomia de uma boa tabela

Uma boa tabela de resultados:

- tem um **título** que permite entendê-la sem ler o texto, informando o que é mostrado, a unidade e o tamanho da amostra;
- organiza as linhas em uma **ordem lógica** (testemunha primeiro, depois os tratamentos em ordem de dose ou de eficácia);
- coloca na mesma coluna grandezas **comparáveis**, com o mesmo número de casas decimais;
- mostra **estimativas e intervalos**, e não apenas médias com letras;
- explica em nota de rodapé as **abreviações** e o método de ajuste para multiplicidade;
- evita linhas verticais e bordas desnecessárias.

12.6.3 Tabelas a partir do R

Digitar números de uma saída do R em uma planilha ou em um editor de texto é uma fonte frequente de erros. Uma alternativa é construir a tabela diretamente a partir dos objetos da análise. O exemplo abaixo combina as médias marginais e os efeitos em relação à testemunha em uma única tabela:

```
medias ← as.data.frame(confint(emm))
efe ← as.data.frame(efeitos)

tabela ← data.frame(
  Tratamento = medias$treatment,
  `Severidade média (%)` = sprintf("%.1f (%.1f a %.1f)",
                                   medias$emmean, medias$lower.CL, medias$upper.CL),
  `Efeito vs. testemunha (p.p.)` = c("—", sprintf("%.1f (%.1f a %.1f)",
                                                  efe$estimate, efe$lower.CL,
efe$upper.CL)),
  check.names = FALSE
)
```

```
knitr::kable(tabela, align = "lcc")
```

Tabela 12.2. Severidade média da doença e efeito de cada programa de fungicida em relação à testemunha (ensaio simulado; quatro blocos). Valores entre parênteses: intervalos de confiança de 95%; os intervalos dos efeitos foram ajustados pelo método de Dunnett.

Tratamento	Severidade média (%)	Efeito vs. testemunha (p.p.)
Testemunha	48.2 (43.0 a 53.3)	—
Fung. A	17.2 (12.0 a 22.3)	−31.0 (-40.4 a −21.5)
Fung. B	21.9 (16.7 a 27.0)	−26.3 (-35.7 a −16.8)
Fung. C	33.0 (27.9 a 38.2)	−15.1 (-24.6 a −5.7)
Fung. D	25.3 (20.2 a 30.5)	−22.8 (-32.3 a −13.3)

Como a tabela é gerada pelo código, qualquer mudança nos dados ou no modelo é refletida automaticamente, sem risco de erro de transcrição. Pacotes como `gt` e `flextable` oferecem recursos adicionais de formatação e exportação para Word.

12.7 Figuras de resultados

Os princípios do Capítulo 5 aplicam-se integralmente às figuras de resultados: mostrar os dados sempre que possível, representar a incerteza, ordenar as categorias de forma significativa e escrever legendas autoexplicativas. Para comparações com uma testemunha, um gráfico de efeitos comunica diretamente a magnitude e a precisão de cada efeito (Figura 12.1).

```
efe$tratamento ← sub(" - Testemunha", "", efe$contrast)
efe$tratamento ← reorder(efe$tratamento, -efe$estimate)

ggplot(efe, aes(x = estimate, y = tratamento, xmin = lower.CL, xmax = upper.CL)) +
  geom_vline(xintercept = 0, colour = "grey50") +
  geom_pointrange(colour = "#0072B2", size = 0.6, linewidth = 1.1) +
  labs(x = "Diferença na severidade em relação à testemunha (pontos percentuais)",
       y = NULL) +
  theme_classic(base_size = 13)
```

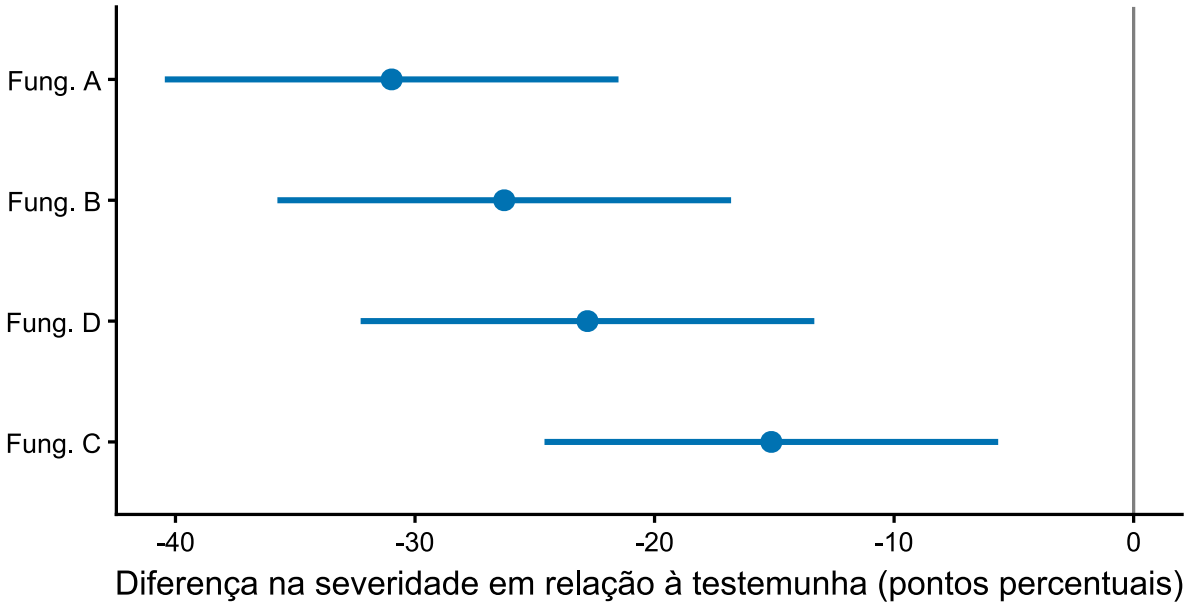


Figura 12.1. Efeito de cada programa de fungicida sobre a severidade da doença em relação à testemunha (ensaio simulado; quatro blocos). Pontos: diferenças estimadas entre médias marginais; barras: intervalos de confiança de 95% ajustados pelo método de Dunnett. A linha vertical marca a ausência de efeito.

Para submissão, a figura deve ser salva nas dimensões e no formato exigidos pela revista (Capítulo 5):

```
ggsave("figures/fig2_efeitos.pdf", width = 17, height = 9, units = "cm")
```

12.8 Escrevendo a interpretação

A discussão de um trabalho conecta os resultados à questão biológica. Uma boa interpretação:

- começa pela **resposta à questão** que motivou o estudo;
- distingue **evidência estatística** de **relevância biológica**: um efeito pode ser claro, mas pequeno demais para importar, ou grande, mas estimado com pouca precisão (Capítulo 6);
- compara os resultados com **estudos anteriores**, explicando concordâncias e discordâncias;

- reconhece as **limitações** do delineamento e da análise: uma única safra, um único local, um isolado, poucas repetições;
- indica a **população** à qual as conclusões se aplicam e evita generalizações além dela;
- propõe **hipóteses** para estudos futuros quando os resultados forem exploratórios.

> ⚠ **Aviso**
>
> A discussão não é o lugar para transformar resultados não significativos em "tendências" nem para apresentar como confirmados padrões descobertos durante a exploração dos dados.

12.9 Relatórios reprodutíveis

12.9.1 Números gerados pelo código

Uma das fontes mais comuns de erro em artigos científicos é a transcrição manual de números da saída da análise para o texto. Em documentos Quarto, como este livro, é possível inserir resultados diretamente no texto com **código R em linha**. Por exemplo, o texto a seguir não contém nenhum número digitado à mão:

> O programa A reduziu a severidade em 31,0 pontos percentuais em relação à testemunha (IC 95%: 21,5 a 40,4).

No arquivo-fonte, essa frase é escrita assim:

```
O programa A reduziu a severidade em `r fmt(efe_A$estimate)`
pontos percentuais em relação à testemunha (IC 95%: ...).
```

Se os dados forem corrigidos ou o modelo for alterado, o texto é atualizado automaticamente na próxima renderização. O Capítulo 13 aprofunda esse e outros aspectos da pesquisa reprodutível.

12.9.2 Material suplementar, dados e código

Cada vez mais revistas exigem ou incentivam a disponibilização dos dados e do código usados nas análises. Mesmo quando não é obrigatório, publicar os dados organizados (Capítulo 3), o dicionário de dados e os scripts de análise em um repositório permanente (como Zenodo, Figshare, OSF ou o repositório institucional) permite que outros verifiquem, reutilizem e ampliem o trabalho, e aumenta a confiança nos resultados (Wilkinson et al. 2016).

12.10 Pesquisa na prática: redigindo os resultados de um ensaio de fungicidas

> **Pesquisa revisitada: Comunicando a eficácia de fungicidas**
>
> Retome um ensaio de campo como o do míldio do manjericão apresentado no Capítulo 1 (McGrath 2020): vários programas de fungicida e uma testemunha, em blocos casualizados, com a severidade avaliada no fim da safra. Usando as análises deste capítulo como modelo, prepare:

1. um **parágrafo de métodos** que informe o delineamento, a unidade experimental, o modelo, as comparações e o ajuste para multiplicidade;
2. uma **tabela** com as médias e os efeitos em relação à testemunha, com intervalos de confiança;
3. uma **figura** que mostre os efeitos e sua incerteza;
4. um **parágrafo de resultados** em que cada afirmação venha acompanhada de uma estimativa com unidades e incerteza;
5. um **parágrafo de discussão** que diga se as reduções observadas são relevantes para a produção comercial e quais limitações impedem generalizar os resultados.

Em seguida, troque o texto com um colega e verifique: é possível entender o que foi feito e o que foi encontrado sem ver a saída do R?

12.11 Erros comuns

12.11.1 Relatar apenas valores-P ou letras

O leitor não consegue avaliar a magnitude nem a precisão dos efeitos.

12.11.2 Omitir a unidade experimental e o número de repetições

Sem essa informação, não é possível avaliar a adequação da análise.

12.11.3 Não informar o que as barras de erro representam

Desvio padrão, erro padrão e intervalo de confiança têm significados muito diferentes (Capítulo 5).

12.11.4 Excesso de casas decimais

Números como 23,4719% sugerem uma precisão inexistente.

12.11.5 Repetir os mesmos resultados em tabela, figura e texto

A redundância ocupa espaço e dificulta a leitura.

12.11.6 Transcrever números à mão

Cópias manuais introduzem erros difíceis de detectar.

12.11.7 Linguagem causal sem delineamento que a justifique

Associações observacionais não são efeitos causais.

12.11.8 Confundir “não significativo” com “sem efeito”

A ausência de evidência não é evidência de ausência.

12.11.9 Descrever a análise planejada em vez da análise feita

Mudanças no modelo devem ser relatadas e justificadas.

12.11.10 Generalizar além das condições estudadas

Uma safra, um local ou um isolado não representam todas as condições.

12.12 Um fluxo de trabalho para comunicar resultados

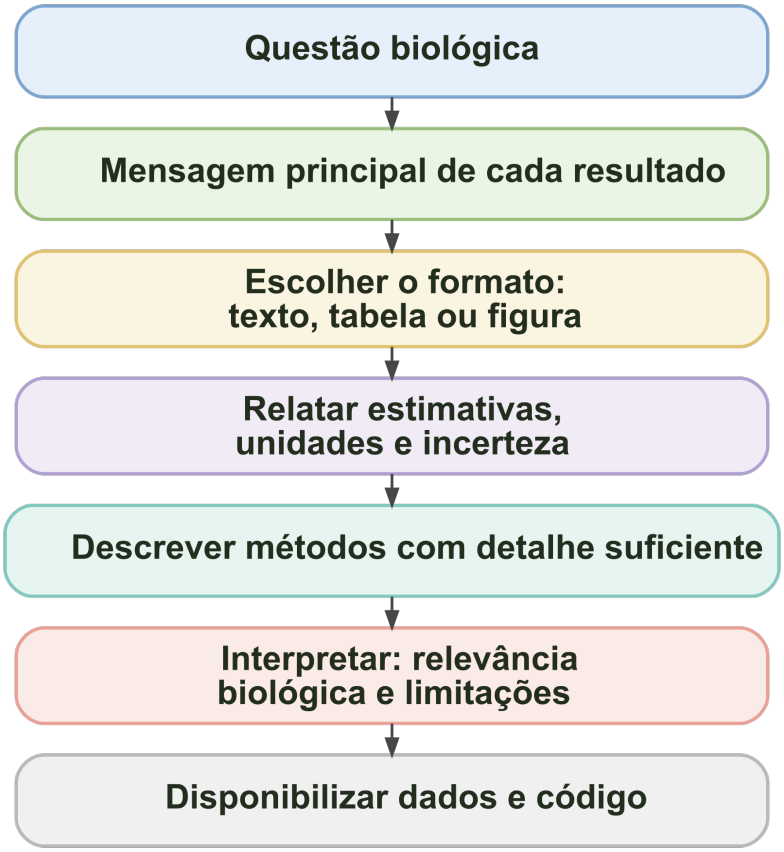


Figura 12.2. Um fluxo de trabalho para comunicar resultados científicos.

12.13 Lista de verificação para a comunicação de resultados

Antes de submeter um manuscrito ou relatório, pergunte:

1. A questão biológica e a mensagem principal estão claras?
2. O delineamento, a unidade experimental e o número de repetições estão descritos?
3. O modelo estatístico, as comparações e o ajuste para multiplicidade estão descritos?
4. Os programas e pacotes estão citados, com versões?
5. Cada resultado principal tem uma estimativa com unidades e incerteza?
6. Os valores-P, quando relatados, são exatos e complementares às estimativas?
7. O número de casas decimais é coerente com a precisão?
8. As figuras mostram os dados e informam o que as barras representam?
9. As tabelas têm títulos autoexplicativos e notas de rodapé?
10. O mesmo resultado não aparece repetido em tabela e figura?
11. A linguagem causal é justificada pelo delineamento?
12. A discussão distingue evidência estatística de relevância biológica?
13. As limitações e o escopo das conclusões estão explícitos?
14. Os números do texto foram gerados pelo código, ou conferidos com a saída da análise?
15. Os dados e o código estão disponíveis?

12.14 Resumo do capítulo

Comunicar resultados é a etapa em que uma análise se torna conhecimento compartilhado. Uma boa comunicação parte da questão biológica e transforma resultados estatísticos em mensagens científicas, sustentadas por estimativas com unidades e incerteza. Os métodos devem ser descritos com detalhe suficiente para serem avaliados e reproduzidos; os valores-P, quando usados, complementam as estimativas; e a precisão numérica deve refletir a precisão real das medidas. Tabelas, figuras e texto têm funções diferentes e devem ser escolhidos de acordo com o que precisa ser comunicado. Gerar tabelas, figuras e números diretamente a partir do código elimina erros de transcrição e torna o relato reprodutível. Por fim, a interpretação deve distinguir evidência estatística de relevância biológica e deixar claros as limitações e o escopo das conclusões.

12.15 Conceitos-chave

Ao concluir este capítulo, você deverá ser capaz de explicar:

- mensagem principal;
- relato de métodos estatísticos;
- estimativa com incerteza;
- precisão numérica;
- linguagem causal;
- tabela de resultados;
- gráfico de efeitos;
- código em linha;
- relatório reprodutível;
- material suplementar;
- dados abertos.

12.16 Exercícios

12.16.1 Exercício 1. Reescrevendo resultados

Reescreva as frases abaixo de forma mais informativa, inventando valores plausíveis quando necessário:

> “O fungicida foi significativamente melhor que a testemunha (P < 0,05).”

> “Não houve efeito da cultivar sobre a severidade (P = 0,12).”

> “A severidade média foi de 34,58219%.”

12.16.2 Exercício 2. Parágrafo de métodos

Escreva o parágrafo de métodos estatísticos de um experimento da sua área, informando o delineamento, a unidade experimental, o modelo, as comparações, o ajuste para multiplicidade e os programas usados.

12.16.3 Exercício 3. Da saída do R à tabela

Usando um conjunto de dados fornecido pelo professor, ajuste um modelo, obtenha as médias marginais e os contrastes de interesse com o `emmeans` e construa uma tabela de resultados diretamente no R, com título e nota de rodapé.

12.16.4 Exercício 4. Avaliação de um artigo publicado

Escolha um artigo recente de Fitopatologia e avalie:

1. se o delineamento e a unidade experimental estão descritos;
2. se o modelo estatístico e o ajuste para multiplicidade estão descritos;
3. se os resultados são relatados como estimativas com incerteza;
4. se as figuras informam o que as barras de erro representam;
5. se a linguagem causal é justificada;
6. se os dados e o código estão disponíveis.

12.16.5 Exercício 5. Texto reprodutível

Em um documento Quarto, ajuste um modelo simples e escreva uma frase de resultados em que todos os números sejam inseridos com código R em linha. Altere um valor dos dados e renderize novamente para verificar que o texto se atualiza.

12.17 Reflexão

> **Uma análise só está completa quando outra pessoa consegue entender o que foi feito, o que foi encontrado e quão confiáveis são as conclusões.**

Discuta essa afirmação usando um exemplo da Fitopatologia.

12.18 Próximos passos

Comunicar bem os resultados depende de saber exatamente como foram obtidos. O Capítulo 13 integra os conceitos deste livro em um fluxo de trabalho de pesquisa reprodutível: organização de projetos, documentação das análises, relatórios dinâmicos com o Quarto e compartilhamento de dados e código.

CAPÍTULO 13

Como tornar a pesquisa totalmente reprodutível?

13.1 Objetivos de aprendizagem

! Importante

Ao concluir este capítulo, o leitor deverá ser capaz de:

- distinguir reprodutibilidade, replicabilidade e robustez;
- organizar um projeto de pesquisa de modo que qualquer pessoa consiga reproduzir suas análises;
- preferir scripts a operações manuais e escrever código legível e documentado;
- usar documentos Quarto para integrar texto, código e resultados;
- registrar o ambiente computacional e controlar a aleatoriedade;
- compreender os princípios do controle de versões com Git;
- preparar dados e código para compartilhamento em repositórios permanentes;
- avaliar a reprodutibilidade de um estudo publicado.

13.2 Por que a reprodutibilidade importa

Ao longo deste livro, cada etapa da pesquisa foi tratada como parte de um mesmo fluxo: a questão biológica orienta o delineamento, o delineamento determina a estrutura dos dados, os dados organizados permitem a exploração, a exploração orienta a análise, e a análise sustenta as conclusões comunicadas. Este capítulo trata do que mantém esse fluxo íntegro ao longo do tempo: a capacidade de refazer cada etapa e obter os mesmos resultados.

Uma análise é **reprodutível** quando outra pessoa, ou o próprio pesquisador meses depois, consegue partir dos mesmos dados brutos, executar o mesmo código e chegar exatamente aos mesmos números, tabelas e figuras. Parece uma exigência mínima, mas na prática muitas análises não passam nesse teste: os dados foram editados à mão sem registro, os passos foram feitos clicando em menus, os arquivos intermediários se perderam, a versão de um pacote mudou ou ninguém lembra qual dos vários arquivos chamados `analise_final` produziu a figura do artigo.

A reprodutibilidade beneficia, antes de todos, o próprio pesquisador. Quando um revisor pede uma nova análise, quando um erro é descoberto nos dados ou quando um novo ano de experimento é acrescentado, um projeto reprodutível é atualizado em minutos; um projeto não

reprodutível exige refazer tudo à mão, com risco de introduzir novos erros. Ela também aumenta a confiança nos resultados, facilita a colaboração e permite que outros reutilizem dados e métodos.

> **! Importante**
>
> A pessoa que mais se beneficia de uma análise reprodutível é, quase sempre, o próprio autor alguns meses depois.

13.3 Reprodutibilidade, replicabilidade e robustez

Esses termos são usados de formas diferentes em diferentes áreas. Neste livro, adotamos as seguintes definições:

Tabela 13.1. Reprodutibilidade, robustez e replicabilidade.

Termo	Mesmos dados?	Mesma análise?	Pergunta respondida
Reprodutibilidade	Sim	Sim	Os resultados podem ser recalculados a partir dos dados e do código?
Robustez	Sim	Não	As conclusões se mantêm com escolhas analíticas razoáveis diferentes?
Replicabilidade	Não	Sim	Um novo experimento, com novos dados, leva às mesmas conclusões?

A reprodutibilidade é o requisito mais básico: se os resultados não podem nem ser recalculados a partir dos mesmos dados, não há como avaliar sua robustez nem compará-los com replicações. A replicabilidade, discutida no Capítulo 2 como a repetição de experimentos, depende do delineamento e da biologia; a reprodutibilidade depende sobretudo de como os dados e as análises são organizados — e está inteiramente sob o controle do pesquisador.

13.4 Organizando o projeto

13.4.1 Uma pasta por projeto

O ponto de partida é concentrar tudo o que um projeto precisa em uma única pasta: dados, scripts, resultados, figuras, manuscrito e documentação. A estrutura recomendada no Capítulo 3 é um bom modelo:

```
ensaio_fungicidas_manjericao/
├── README.md               # o que é o projeto e como reproduzi-lo
├── ensaio_fungicidas.Rproj # projeto do RStudio
├── data/
│   ├── raw/                # dados brutos: somente leitura
│   └── processed/          # dados gerados pelos scripts
├── R/                      # funções reutilizáveis
├── scripts/                # 01_limpeza.R, 02_analise.R, ...
├── figures/                # figuras geradas pelo código
├── results/                # tabelas e saídas de modelos
└── manuscript/             # relatório ou artigo em Quarto
```

13.4.2 Projetos do RStudio e caminhos relativos

Um **projeto do RStudio** (arquivo `.Rproj`) define a pasta do projeto como diretório de trabalho sempre que é aberto. Com isso, os scripts podem usar **caminhos relativos**, como `data/raw/severidade_2025.csv`, em vez de caminhos absolutos, como `C:/Users/maria/Documents/doutorado/dados/severidade_2025.csv`, que só funcionam no computador de quem os escreveu.

O pacote `here` torna os caminhos relativos ainda mais robustos, pois sempre parte da raiz do projeto, mesmo quando o código é executado a partir de uma subpasta (como a pasta do manuscrito):

```
library(here)

dados ← read.csv(here("data", "raw", "severidade_2025.csv"))
```

> ⚠ **Aviso**
>
> Evite `setwd()` com caminhos absolutos no início dos scripts. Esse é um dos motivos mais comuns pelos quais um código que "funcionava" deixa de funcionar no computador de um colaborador.

13.4.3 O arquivo README

O `README` é a porta de entrada do projeto. Ele deve explicar, em poucas linhas, o objetivo do estudo, a origem dos dados, o que há em cada pasta, a ordem em que os scripts devem ser executados e os pacotes necessários. Um bom teste é perguntar: um colega que nunca viu o projeto conseguiria reproduzir os resultados apenas lendo o `README`?

13.5 Scripts em vez de cliques

Toda operação feita com o mouse — filtrar linhas em uma planilha, apagar um valor discrepante, copiar uma média para o texto, ajustar um eixo em um programa gráfico — é uma operação sem registro. Ela não pode ser revista, corrigida ou repetida automaticamente. Scripts, ao contrário, são o registro exato de cada passo.

13.5.1 Uma cadeia de scripts

Uma organização simples e eficaz divide a análise em scripts numerados, cada um com uma função clara, que leem arquivos de uma etapa e gravam arquivos para a seguinte:

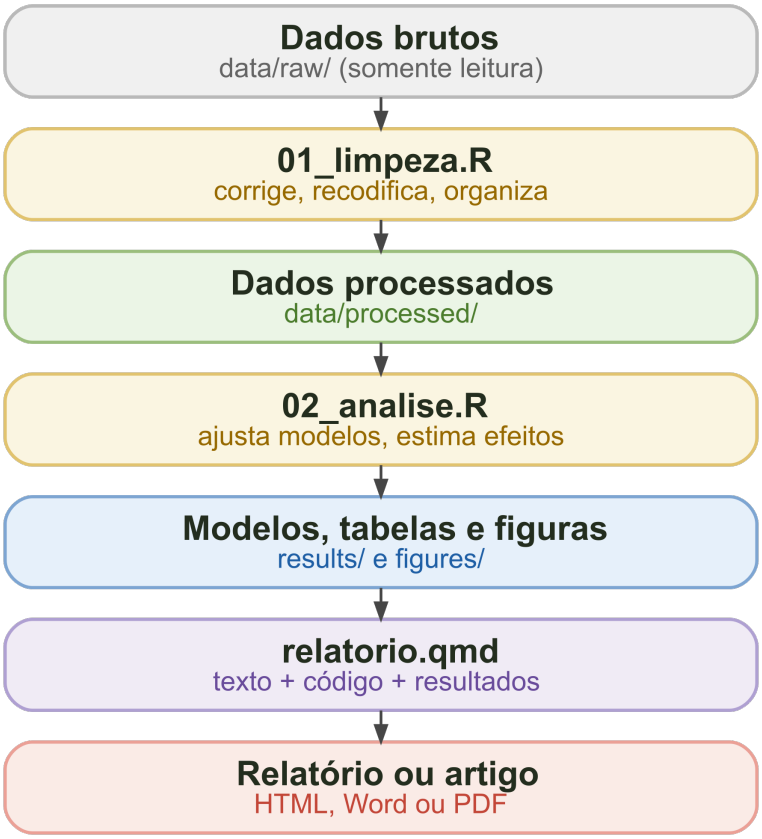


Figura 13.1. Uma cadeia de análise reprodutível. Os dados brutos nunca são alterados; cada script lê os produtos da etapa anterior e grava os seus, e o relatório final é gerado a partir deles.

Com essa estrutura, refazer toda a análise após uma correção nos dados brutos significa simplesmente executar os scripts de novo, na ordem.

13.5.2 Código legível

Código reprodutível também precisa ser compreensível. Algumas práticas simples fazem grande diferença:

- nomes de objetos e variáveis **descritivos** (`severidade_media`, e não `x2`);
- **comentários** que explicam o *porquê* de cada decisão, e não apenas o *quê*;
- um estilo consistente (espaços, indentação, `snake_case`);
- **funções** para operações repetidas, em vez de copiar e colar o mesmo bloco de código.

O exemplo abaixo transforma em função um cálculo usado várias vezes neste livro, a eficácia de controle em relação à testemunha:

```
# Eficácia de controle (%) de um tratamento em relação à testemunha.
# Usa as médias de severidade; valores negativos indicam mais doença que a testemunha.
eficacia ← function(sev_tratamento, sev_testemunha) {
  100 * (sev_testemunha - sev_tratamento) / sev_testemunha
}

eficacia(sev_tratamento = c(18, 22, 31), sev_testemunha = 48)
```

```
[1] 62.50000 54.16667 35.41667
```

Uma função documentada pode ser guardada na pasta `R/` e reutilizada em todos os scripts do projeto, garantindo que o cálculo seja sempre o mesmo.

13.6 Documentos dinâmicos com o Quarto

13.6.1 Texto, código e resultados no mesmo documento

O Quarto, usado para escrever este livro, permite combinar texto, código e resultados em um único documento. Ao renderizar o arquivo `.qmd`, o código é executado e suas saídas — tabelas, figuras, números — são inseridas automaticamente no documento final, que pode ser gerado em HTML, Word ou PDF.

Um documento Quarto mínimo tem três partes: um cabeçalho com metadados, blocos de código e texto:

````
---
title: "Ensaio de fungicidas - safra 2025"
format: html
---

```{r}
library(emmeans)
dados ← read.csv("data/processed/severidade.csv")
modelo ← lm(severity ~ block + treatment, data = dados)
```

A severidade média na testemunha foi de `r round(mean(dados$severity[dados$treatment ==
'Testemunha']), 1)`%.
````

O Capítulo 12 mostrou como o código em linha elimina a transcrição manual de números. Quando os dados mudam, basta renderizar o documento novamente para que texto, tabelas e figuras sejam atualizados juntos.

13.6.2 Opções dos blocos de código

As opções no início de cada bloco controlam o que aparece no documento final. As mais usadas são:

Tabela 13.2. Opções comuns dos blocos de código no Quarto.

Opção	Efeito
`#| echo: false`	executa o código, mas não o mostra
`#| eval: false`	mostra o código, mas não o executa
`#| include: false`	executa o código sem mostrar código nem saída
`#| label: fig-...`	rotula uma figura para referência cruzada
`#| fig-cap: "..."`	define a legenda da figura
`#| warning: false`	oculta avisos

13.6.3 Parâmetros

Documentos Quarto podem receber **parâmetros**, o que permite gerar o mesmo relatório para diferentes safras, locais ou experimentos sem duplicar o código. Por exemplo, um relatório com o parâmetro `safra` pode ser renderizado uma vez para cada ano:

```
quarto render relatorio.qmd -P safra:2024
quarto render relatorio.qmd -P safra:2025
```

13.7 O ambiente computacional

13.7.1 Registrando as versões

Os resultados de uma análise podem depender da versão do R e dos pacotes usados: funções mudam, padrões são alterados e erros são corrigidos. Registrar o ambiente é, portanto, parte da reprodutibilidade. A função `sessionInfo()` informa a versão do R, o sistema operacional e as versões de todos os pacotes carregados. Por exemplo, os pacotes usados para gerar este capítulo foram:

```
pacotes ← c("ggplot2", "emmeans", "patchwork", "DiagrammeR")
sapply(pacotes, function(p) as.character(packageVersion(p)))
```

```
   ggplot2    emmeans  patchwork DiagrammeR
   "4.0.3"    "2.0.1"    "1.3.2"   "1.0.12"
```

```
R.version.string
```

```
[1] "R version 4.5.2 (2025-10-31 ucrt)"
```

13.7.2 Ambientes isolados com renv

O pacote `renv` vai além de registrar as versões: ele cria uma biblioteca de pacotes própria para o projeto e grava as versões exatas em um arquivo (`renv.lock`). Outra pessoa, ou o mesmo pesquisador em outro computador, pode então restaurar exatamente o mesmo ambiente:

```
renv::init()       # cria o ambiente do projeto
renv::snapshot()   # registra as versões atuais em renv.lock
renv::restore()    # em outro computador: instala as mesmas versões
```

13.7.3 Aleatoriedade controlada

Simulações, reamostragens (*bootstrap*), permutações e alguns algoritmos (como o *k*-médias) usam números aleatórios. Sem controle, cada execução dá um resultado ligeiramente diferente. A função `set.seed()` fixa o ponto de partida do gerador de números aleatórios, tornando os resultados reprodutíveis:

```
set.seed(2025)
amostra_1 ← sample(1:100, 5)

set.seed(2025)
amostra_2 ← sample(1:100, 5)

amostra_1
```

```
[1] 13 76 36 26 65
```

```
identical(amostra_1, amostra_2)
```

```
[1] TRUE
```

Todos os dados simulados deste livro foram gerados dessa forma, e por isso qualquer leitor obtém exatamente os mesmos números e figuras.

13.8 Controle de versões com Git

13.8.1 Por que controlar versões

Projetos mudam: dados são corrigidos, análises são refeitas, textos são revisados por coautores. Sem controle de versões, o histórico vive em nomes de arquivos (`analise_v3_final_revisado.R`) e na memória das pessoas. O **Git** é um sistema que registra cada alteração nos arquivos do projeto, com data, autor e uma mensagem descritiva, e permite voltar a qualquer versão anterior.

13.8.2 Conceitos básicos

- **repositório**: a pasta do projeto sob controle do Git, com todo o seu histórico;
- **commit**: um registro do estado dos arquivos em um momento, acompanhado de uma mensagem (“corrige unidade da produtividade no bloco 3”);
- **histórico**: a sequência de commits, que mostra o que mudou, quando e por quê;
- **repositório remoto**: uma cópia do repositório em um serviço como GitHub ou GitLab, que serve de cópia de segurança e facilita a colaboração;
- `.gitignore`: lista de arquivos que não devem ser versionados, como arquivos temporários e saídas que podem ser regeneradas.

O RStudio tem uma interface gráfica para o Git, e os pacotes `usethis` e `gert` permitem configurá-lo a partir do R:

```
usethis::use_git()      # transforma o projeto em um repositório Git
usethis::use_github()   # cria um repositório remoto no GitHub
```

> **Dica**
>
> Faça commits pequenos e frequentes, cada um com uma mensagem que explique o que mudou e por quê. Um histórico bem escrito é uma documentação do projeto.

13.8.3 O que versionar

Scripts, documentos Quarto, o `README`, o dicionário de dados e, em geral, os dados brutos pequenos devem ser versionados. Arquivos grandes, dados sensíveis e produtos que podem ser regenerados a partir do código (figuras, tabelas, documentos renderizados) geralmente ficam fora do controle de versões ou são guardados em outro lugar.

13.9 Compartilhando dados e código

13.9.1 Repositórios permanentes

Disponibilizar dados e código permite que outras pessoas verifiquem, reutilizem e ampliem o trabalho. Para isso, eles devem estar em um **repositório permanente**, que atribui um identificador persistente (DOI) e garante que os arquivos continuem acessíveis. Exemplos incluem o Zenodo, o Figshare, o Open Science Framework (OSF), o Dryad e os repositórios institucionais. Um repositório no GitHub é ótimo para o desenvolvimento, mas não é, por si só, um arquivo permanente; o GitHub pode ser integrado ao Zenodo para gerar um DOI a cada versão publicada.

13.9.2 Princípios FAIR

Os dados compartilhados devem ser, na medida do possível, **localizáveis, acessíveis, interoperáveis e reutilizáveis** (princípios FAIR) (Wilkinson et al. 2016). Na prática, isso significa: um identificador persistente; formatos abertos (CSV em vez de formatos proprietários); um dicionário de dados que descreva cada variável e suas unidades (Capítulo 3); e uma **licença** que diga como os dados e o código podem ser reutilizados (por exemplo, CC BY para dados e MIT ou GPL para código).

13.9.3 O que compartilhar

Um pacote de reprodutibilidade completo inclui:

- os dados brutos (ou, quando não puderem ser compartilhados, os dados processados e uma explicação);
- o dicionário de dados;
- todos os scripts, na ordem em que devem ser executados;
- o documento Quarto do relatório ou artigo, quando houver;
- o registro do ambiente computacional (`sessionInfo()` ou `renv.lock`);
- um `README` com instruções.

13.10 "Bom o suficiente" é melhor do que perfeito

A reprodutibilidade não exige dominar todas as ferramentas deste capítulo de uma vez. Práticas simples, adotadas de forma consistente, já resolvem a maior parte dos problemas (Wilson et al. 2017). Uma progressão razoável é:

1. uma pasta por projeto, com dados brutos protegidos e um `README`;
2. scripts para todas as etapas, sem edições manuais nos dados;
3. relatórios em Quarto, com tabelas, figuras e números gerados pelo código;
4. registro do ambiente e uso de `set.seed()`;
5. controle de versões com Git;
6. compartilhamento em repositório permanente, com DOI e licença.

Cada degrau já representa um ganho real; não é preciso esperar dominar todos para começar.

13.11 Pesquisa na prática: tornando reprodutível um ensaio de fungicidas

Pesquisa revisitada: Do caderno de campo ao pacote reprodutível

Retome um ensaio de fungicidas como o do míldio do manjericão discutido ao longo do livro (McGrath 2020). Os dados chegaram em uma planilha preenchida no campo, com abas por data de avaliação, células mescladas e algumas correções feitas à mão (Capítulo 3).

Planeje como transformar esse material em um projeto reprodutível:

1. Que estrutura de pastas você criaria, e o que iria em cada uma?
2. Onde ficaria a planilha original, e como você garantiria que ela não fosse alterada?
3. Quais scripts seriam necessários, e o que cada um faria?
4. Como as correções feitas à mão na planilha seriam registradas?
5. Como você geraria a tabela e a figura de resultados do Capítulo 12 sem copiar números?
6. O que você incluiria no `README` e no repositório público?
7. O que aconteceria se um revisor pedisse para excluir um bloco da análise? Quanto tempo levaria para refazer tudo?

13.12 Erros comuns

13.12.1 Editar os dados brutos à mão

Alterações sem registro não podem ser revistas nem repetidas.

13.12.2 Usar caminhos absolutos e `setwd()`

O código deixa de funcionar em outro computador.

13.12.3 Copiar números da saída para o texto

Erros de transcrição são frequentes e difíceis de detectar.

13.12.4 Guardar versões em nomes de arquivos

`analise_final_v3_revisado.R` não informa o que mudou nem por quê.

13.12.5 Não registrar as versões dos pacotes

Mudanças nos pacotes podem alterar resultados sem aviso.

13.12.6 Esquecer de fixar a semente aleatória

Simulações e reamostragens dão resultados diferentes a cada execução.

13.12.7 Compartilhar apenas o resultado final

Sem dados brutos, scripts e instruções, a análise não pode ser verificada.

13.12.8 Depender de um único computador

Sem cópia de segurança e controle de versões, um disco danificado pode significar a perda de todo o projeto.

13.13 Um fluxo de trabalho para a pesquisa reprodutível

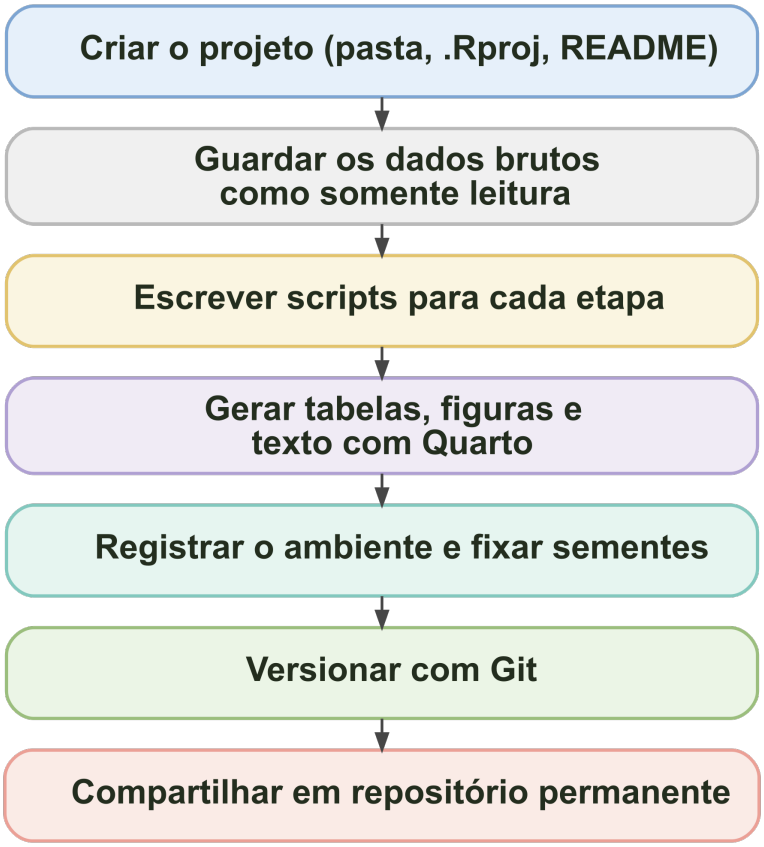


Figura 13.2. Um fluxo de trabalho para a pesquisa reprodutível, do início do projeto ao compartilhamento.

13.14 Lista de verificação da reprodutibilidade

Antes de considerar uma análise concluída, pergunte:

1. Todo o projeto está em uma única pasta, com um `README`?
2. Os dados brutos estão preservados e nunca são editados à mão?
3. Todas as etapas, da limpeza às figuras, estão em scripts?
4. Os scripts usam caminhos relativos?
5. Os scripts podem ser executados do início ao fim, em uma sessão limpa do R, sem erros?
6. Tabelas, figuras e números do texto são gerados pelo código?
7. As versões do R e dos pacotes estão registradas?
8. As sementes aleatórias estão fixadas?
9. O projeto está sob controle de versões?
10. Há um dicionário de dados?
11. Dados e código estão disponíveis em um repositório permanente, com licença?
12. Um colega conseguiria reproduzir os resultados sem ajuda?

13.15 Resumo do capítulo

A reprodutibilidade — a capacidade de recalcular exatamente os mesmos resultados a partir dos mesmos dados e do mesmo código — é a base sobre a qual a robustez e a replicabilidade podem ser avaliadas, e está inteiramente sob o controle do pesquisador. Ela começa com a organização

do projeto em uma única pasta, com dados brutos protegidos, caminhos relativos e um `README`. Scripts substituem operações manuais e registram cada decisão; funções e comentários tornam o código compreensível. Documentos Quarto integram texto, código e resultados, de modo que tabelas, figuras e números sejam atualizados automaticamente. O registro do ambiente computacional, o uso de sementes aleatórias e o controle de versões com Git garantem que os resultados possam ser recuperados ao longo do tempo. Por fim, o compartilhamento de dados e código em repositórios permanentes, segundo os princípios FAIR, permite que outros verifiquem e reutilizem o trabalho. Práticas simples, adotadas de forma consistente, já trazem a maior parte desses benefícios.

13.16 Conceitos-chave

Ao concluir este capítulo, você deverá ser capaz de explicar:

- reprodutibilidade;
- robustez;
- replicabilidade;
- projeto do RStudio;
- caminho relativo;
- README;
- cadeia de scripts;
- documento dinâmico;
- código em linha;
- parâmetros de um relatório;
- ambiente computacional;
- semente aleatória;
- controle de versões;
- commit;
- repositório permanente;
- DOI;
- princípios FAIR;
- licença de uso.

13.17 Exercícios

13.17.1 Exercício 1. Diagnóstico de reprodutibilidade

Escolha um projeto seu (um capítulo de tese, um experimento ou um trabalho de disciplina) e aplique a lista de verificação deste capítulo. Quais itens já são atendidos? Quais seriam mais fáceis de implementar primeiro?

13.17.2 Exercício 2. Reorganizando um projeto

Crie um projeto do RStudio com a estrutura de pastas apresentada neste capítulo, coloque nele um conjunto de dados e escreva dois scripts: um de limpeza, que grave os dados processados, e um de análise, que gere uma tabela e uma figura. Escreva um `README` com as instruções para reproduzir tudo.

13.17.3 Exercício 3. Relatório dinâmico

Transforme a análise do Exercício 2 em um documento Quarto em que todos os números do texto sejam gerados por código em linha. Altere um valor dos dados brutos, renderize novamente e verifique o que muda.

13.17.4 Exercício 4. Sementes e simulações

Simule um ensaio com cinco tratamentos e quatro blocos, ajuste um modelo e estime os efeitos dos tratamentos. Execute o código duas vezes sem `set.seed()` e duas vezes com `set.seed()`. Compare os resultados e explique as diferenças.

13.17.5 Exercício 5. Avaliação de um artigo publicado

Escolha um artigo recente de Fitopatologia e verifique:

1. se os dados estão disponíveis, onde e em que formato;
2. se o código está disponível;
3. se as versões dos programas e pacotes são informadas;
4. se há um identificador persistente (DOI) para os dados;
5. se haveria como reproduzir uma das figuras do artigo.

13.18 Reflexão

> **Uma análise que não pode ser reproduzida é uma afirmação; uma análise reprodutível é uma evidência que pode ser verificada.**

Discuta essa afirmação usando um exemplo da sua própria experiência.

13.19 Considerações finais

Este livro começou com uma questão biológica e terminou com um projeto reprodutível. Entre um e outro, procuramos mostrar que a análise de dados não é uma etapa isolada, aplicada no fim de um experimento, mas uma forma de raciocinar que acompanha toda a pesquisa: delinear estudos capazes de responder à questão, organizar os dados com cuidado, explorá-los antes de modelá-los, visualizar a evidência com honestidade, estimar efeitos com sua incerteza, escolher modelos adequados ao tipo de dado, comunicar os resultados com clareza e deixar cada passo registrado para que outros possam verificá-lo.

Os métodos estatísticos continuarão evoluindo, e novos pacotes substituirão os que usamos aqui. Os princípios, porém, permanecem: começar pela biologia, mostrar os dados, quantificar a incerteza, distinguir evidência estatística de relevância biológica e trabalhar de forma transparente. São eles que transformam observações em conhecimento confiável sobre as doenças de plantas.

Referências

AGRIOS, G. N. 2005. INTRODUCTION. Em Elsevier, p. 3–75.. https://doi.org/10.1016/b978-0-08-047378-9.50007-5.

Amrhein, V., Greenland, S., e McShane, B. 2019. Scientists rise up against statistical significance. Nature 567:305–307. https://doi.org/10.1038/d41586-019-00857-9.

Anscombe, F. J. 1973. Graphs in statistical analysis. The American Statistician 27:17–21. https://doi.org/10.1080/00031305.1973.10478966.

Broman, K. W., e Woo, K. H. 2018. Data organization in spreadsheets. The American Statistician 72:2–10. https://doi.org/10.1080/00031305.2017.1375989.

Carisse, O., Levasseur, A., e Provost, C. 2020. Influence of Leaf Wetness Duration and Temperature on Infection of Grape Leaves by Elsinoë ampelina under Controlled and Vineyard Conditions. Plant Disease 104:2817–2822. https://doi.org/10.1094/PDIS-02-20-0262-RE.

Cleveland, W. S., e McGill, R. 1984. Graphical perception: theory, experimentation, and application to the development of graphical methods. Journal of the American Statistical Association 79:531–554. https://doi.org/10.1080/01621459.1984.10478080.

Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. Ecological Monographs 54:187–211. https://doi.org/10.2307/1942661.

McGrath, M. T. 2020. Efficacy of Conventional Fungicides for Downy Mildew in Field-Grown Sweet Basil in the United States. Plant Disease 104:2967–2972. https://doi.org/10.1094/PDIS-11-19-2382-RE.

Oliver, R. P. 2024. Chapter 1 - The history and development of plant pathology org. Richard P. Oliver. Agrios' Plant Pathology (Sixth Edition) 3–61. https://doi.org/https://doi.org/10.1016/B978-0-12-822429-8.00001-7.

Silva, B. N., Picanço, B. B. M., Martins, S. C. V., e Rodrigues, F. A. 2023. Impairment of the photorespiratory pathway on tomato leaves during the infection process of Septoria lycopersici. Physiological and Molecular Plant Pathology 125:102020. https://doi.org/10.1016/j.pmpp.2023.102020.

Silva, E. M. d., Brito, N. D., Mesquita, N. L. S., Soares, P. P. d. S., Oliveira, R. J. V. d., Santos, A., e Novaes, Q. S. d. 2024. First report of Pythium aphanidermatum causing damping off, collar and root rot of coriander in Brazil. Journal of General Plant Pathology 90:168–172. https://doi.org/10.1007/s10327-024--01171-7.

Simmons, J. P., Nelson, L. D., e Simonsohn, U. 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science 22:1359–1366. https://doi.org/10.1177/0956797611417632.

Tufte, E. R. 2001. *The Visual Display of Quantitative Information*. 2º ed. Cheshire, CT: Graphics Press.

Tukey, J. W. 1977. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Wasserstein, R. L., e Lazar, N. A. 2016. The ASA's statement on p-values: context, process, and purpose. The American Statistician 70:129–133. https://doi.org/10.1080/00031305.2016.1154108.

Weissgerber, T. L., Milic, N. M., Winham, S. J., e Garovic, V. D. 2015. Beyond bar and line graphs: time for a new data presentation paradigm. PLoS Biology 13:e1002128. https://doi.org/10.1371/journal.pbio.1002128.

Wickham, H. 2010. A layered grammar of graphics. Journal of Computational and Graphical Statistics 19:3–28. https://doi.org/10.1198/jcgs.2009.07098.

Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. 2º ed. New York: Springer.

Wickham, H. 2014. Tidy data. Journal of Statistical Software 59:1–23. https://doi.org/10.18637/jss.v059.i10.

Wilkinson, L. 2005. *The Grammar of Graphics*. 2º ed. New York: Springer.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., e others. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3:160018. https://doi.org/10.1038/sdata.2016.18.

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., e Teal, T. K. 2017. Good enough practices in scientific computing. PLoS Computational Biology 13:e1005510. https://doi.org/10.1371/journal.pcbi.1005510.