

Ciência com R

Perguntas e respostas para pesquisadores e analistas de dados

Arthur de Sá Ferreira

Ciência com R

Perguntas e respostas para pesquisadores e analistas de dados

2025

Sumário

Sumário	iii
Lista de Figuras	ix
Lista de Tabelas	xi
Dedicatória	xiii
Agradecimentos	xv
Sobre o autor	xvii
Prefácio	xix
<hr/>	
PARTE 1: PENSAMENTO ESTATÍSTICO E METODOLÓGICO	1
Pense como um cientista	1
1 Pensamento probabilístico	3
1.1 Experimento	3
1.2 Espaço amostral e eventos discretos	3
1.3 Espaço amostral e eventos contínuos	3
1.4 Probabilidade	5
1.5 Independência e probabilidade	6
1.6 Leis dos números anômalos	7
1.7 Leis dos pequenos números	7
1.8 Leis dos grandes números	8
1.9 Teorema central do limite	8
1.10 Regressão para a média	12
2 Pensamento estatístico	15
2.1 População e Amostra	15
2.2 Unidade de análise	15
2.3 Amostragem	16
2.4 Reamostragem	16
2.5 Subamostragem e superamostragem	17
3 Pensamento metodológico	19
3.1 Metodologia da pesquisa	19
3.2 Relação Estatística-Metodologia	19
3.3 Reprodutibilidade	19
3.4 Robustez	19
3.5 Replicabilidade	21
3.6 Generalização	21
4 Pensamento computacional	23
4.1 Programas de computador	23

4.2 Scripts computacionais	25
4.3 Pacotes	25
4.4 Aplicativos Shiny	26
4.5 Manuscritos reproduzíveis	26
4.6 Compartilhamento	28
5 Paradoxos e falácia	31
5.1 Paradoxos estatísticos	31
5.2 Falácia estatísticas	33
6 Vieses metodológicos	35
6.1 Vieses metodológicos	35
6.2 Tipos de vieses metodológicos	35
6.3 Efeitos relacionados aos vieses metodológicos	35
<hr/>	
PARTE 2: DADOS – COLETA E PREPARAÇÃO	37
Organização e fundamentos de dados	37
7 Medidas e instrumentos	39
7.1 Escalas	39
7.2 Medição e Medidas	39
7.3 Erros de medida	42
7.4 Instrumentos	43
7.5 Acurácia e precisão	43
7.6 Viés e variabilidade	43
8 Dados, <i>big data</i> e metadados	45
8.1 Dados	45
8.2 <i>Big data</i>	45
8.3 Metadados	45
9 Tabulação de dados	47
9.1 Planilhas eletrônicas	47
10 Variáveis e fatores	51
10.1 Variáveis	51
10.2 Transformação de variáveis	52
10.3 Categorização de variáveis contínuas	53
10.4 Dicotomização de variáveis contínuas	53
10.5 Fatores	55
11 Dados perdidos e imputados	57
11.1 Dados perdidos	57
11.2 Dados imputados	58
12 Dados anonimizados e sintéticos	61
12.1 Dados anonimizados	61
12.2 Dados sintéticos	61
<hr/>	
PARTE 3: ANÁLISE EXPLORATÓRIA E DESCRIPTIVA	63
Analisando padrões	63
13 Descrição	65
13.1 Análise de descrição	65
13.2 Estimação	66

14 Análise inicial de dados	67
14.1 Análise inicial de dados	67
15 Distribuições e parâmetros	69
15.1 Distribuições de probabilidade	69
15.2 Parâmetros	72
15.3 Tendência central	73
15.4 Dispersão	74
15.5 Proporção	75
15.6 Distribuição	76
15.7 Extremos	76
15.8 Valores discrepantes	76
16 Análise exploratória de dados	79
16.1 Análise exploratória de dados	79
17 Análise descritiva	83
17.1 Análise descritiva	83
17.2 Apresentação de resultados numéricos	83
17.3 Tabelas	84
17.4 Tabela 1	85
17.5 Tabela 2	86
17.6 Gráficos	86
18 Análise robusta	89
18.1 Raciocínio inferencial robusto	89
PARTE 4: INFERÊNCIA E TESTES ESTATÍSTICOS	91
De amostras para populações	91
19 Análise inferencial	93
19.1 Raciocínio inferencial	93
19.2 Hipóteses científicas	93
19.3 Hipóteses estatísticas	94
19.4 Testes de hipóteses	94
19.5 Poder do teste	96
19.6 Inferência visual	98
19.7 Interpretação de análise inferencial	99
19.8 Erros de inferência	99
20 Tamanho do efeito e P-valor	103
20.1 Tamanho do efeito	103
20.2 Efeito fixo	104
20.3 Efeito aleatório	104
20.4 Efeito principal	104
20.5 Efeito de modificação	104
20.6 Efeito de interação	104
20.7 Efeito de mediação	106
20.8 Efeitos brutos e padronizados	106
20.9 P-valor	107
20.10 P-hacking	108
21 Seleção de testes	109
21.1 Multiverso de análises estatísticas	109
21.2 Escolha de testes para análise inferencial	109
22 Testes estatísticos	111
22.1 Testes de Qui-quadrado (χ^2)	111

22.2 Teste exato de Fisher	113
23 Comparação	115
23.1 Análise inferencial de comparação	115
24 Associação	117
24.1 Análise inferencial de associação	117
24.2 Associação bivariada	117
24.3 Associação multivariada	118
25 Correlação	119
25.1 Análise inferencial de correlação	119
25.2 Coeficientes de correlação	121
25.3 Colinearidade	125
26 Regressão	127
26.1 Análise de regressão	127
26.2 Preparação de variáveis para regressão	128
26.3 Multicolinearidade	131
26.4 Redução de dimensionalidade	133
<hr/>	
PARTE 5: MODELAGEM ESTATÍSTICA	135
Ferramentas preditivas e causais	135
27 Modelos	137
27.1 Modelos estatísticos	137
27.2 Suposições dos modelos	137
27.3 Avaliação de modelos	137
27.4 Modelos estocásticos	138
27.5 Comparação de modelos	138
28 Redes	139
28.1 Análise de redes	139
29 Aprendizado de máquina	141
29.1 Aprendizado de máquina	141
29.2 Aprendizado supervisionada	141
29.3 Aprendizado não-supervisionada	141
29.4 Aprendizado por reforço	141
29.5 Aprendizado profunda	141
30 Árvore de decisão	143
30.1 Árvore de decisão	143
31 Análise preditiva	145
31.1 Predição via modelagem	145
32 Análise causal	147
32.1 Causalidade	147
<hr/>	
PARTE 6: DELINEAMENTO DE PESQUISA	149
Delineamento antes da análise	149
33 Delineamento de estudos	151
33.1 Critérios de delineamento	151
33.2 Alocação	151
33.3 Cegamento	151

33.4 Pareamento	151
33.5 Aleatorização	151
33.6 Taxonomia de estudos	152
34 Tamanho da amostra	155
34.1 Tamanho da amostra	155
34.2 Cálculo do tamanho da amostra	156
34.3 Perdas de amostra	157
34.4 Ajustes no tamanho da amostra	158
34.5 Justificativa do tamanho da amostra	158
34.6 SPARKing	158
35 Estudos observacionais	159
35.1 Características	159
35.2 Diretrizes para redação	159
36 Propriedades psicométricas	161
36.1 Características	161
36.2 Análise fatorial exploratória	161
36.3 Análise fatorial confirmatória	162
36.4 Validade de conteúdo	162
36.5 Validade de face	162
36.6 Validade do construto	162
36.7 Validade fatorial	162
36.8 Validade convergente	162
36.9 Validade discriminante	162
36.10 Validade de critério	163
36.11 Validade concorrente	163
36.12 Responsividade	163
36.13 Concordância	163
36.14 Confiabilidade	165
36.15 Diretrizes para redação	165
37 Desempenho diagnóstico	167
37.1 Características	167
37.2 Tabelas 2x2	167
37.3 Gráficos <i>crosshair</i>	169
37.4 Curvas ROC	169
37.5 Interpretação da validade de um teste	170
37.6 Diretrizes para redação	170
38 Ensaios quase-experimentais	171
38.1 Características	171
38.2 Diretrizes para redação	171
39 Ensaios experimentais	173
39.1 Ensaio clínico aleatorizado	173
39.2 Modelos de análise de comparação	174
39.3 Comparação na linha de base	174
39.4 Comparação intragrupos	175
39.5 Comparação entre grupos	175
39.6 Comparação de subgrupos	176
39.7 Efeito de interação	176
39.8 Ajuste de covariáveis	177
39.9 Imputação de dados perdidos	177
39.10 Diretrizes para redação	178
40 Meta-análise	179
40.1 Características	179

40.2 Interpretação de efeitos em meta-análise	179
40.3 Diretrizes para redação	180
41 Simulação computacional	181
41.1 Características	181
41.2 Método de Monte Carlo	181
41.3 Diretrizes para redação	182
<hr/>	
<i>PARTE 7: APLICAÇÕES E COMUNICAÇÃO</i>	183
Da análise ao impacto	183
42 Plano de análise	185
42.1 Plano de análise estatística	185
43 Redação de resultados	187
43.1 Resultados da análise estatística	187
43.2 Diretrizes e Listas	187
<hr/>	
<i>PARTE 8: RECURSOS</i>	189
Material complementar	189
44 Shiny Apps	191
Aplicativos por delineamento de estudo	191
45 Fontes externas	193
45.1 Fontes de informação externas	193
46 Diretrizes e Listas	195
46.1 Diretrizes	195
46.2 Listas de verificação	195
Referências	197

Lista de Figuras

1.1	Exemplos de espaço amostral discreto. Superior: Todas as faces de uma moeda. Inferior: Todas as faces de um dado.	4
1.2	Exemplos de evento de experimento. Superior: 1 lançamento de 1 moeda. Inferior: 1 lançamento de 1 dado.	4
1.3	Espaço de eventos: União dos eventos face = 3 e face = 4 de um dado.	5
1.4	Superior: Eventos independentes. Inferior: Eventos dependentes.	6
1.5	Esquerda: Evento (face = 4). Direita: Experimentos de 1 lançamento de 1 dado (superior), 3 lançamentos de 1 dado (central), 10 lançamentos de 1 dado (inferior).	7
1.6	Esquerda: Histogramas de uma variável aleatória com distribuição uniforme (N = 100). Direita: Histogramas da média de 100 amostras de tamanhos 5, 50, 500 e 5000 tomadas da população com reposição e igual probabilidade.	9
1.7	Esquerda: Histogramas de lançamento de 1 dado com distribuição uniforme (N = 100). Direita: Histogramas da média de 100 amostras de tamanhos 5, 50, 500 e 5000 tomadas da população com reposição e igual probabilidade.	10
1.8	Esquerda: Histogramas de lançamento de 1 moeda com distribuição uniforme (N = 100). Direita: Histogramas da média de 100 amostras de tamanhos 5, 50, 500 e 5000 tomadas da população com reposição e igual probabilidade.	11
1.9	Representação gráfica da regressão para a média em medidas repetidas. A segunda medida (dado 2) é mais próxima da média (valor real) do que a primeira medida (dado 1).	13
3.1	Mapa mental da relação entre o pensamento estatístico e o pensamento metodológico.	20
4.1	Interface do RStudio. Fonte: https://docs.posit.co/ide/user/	24
5.1	Paradoxo de Simpson representado com dados simulados. Os pontos no gráfico representam observações individuais e as linhas de tendência representam as regressões lineares ajustadas para os dados desagregados da população e agregados por subpopulação.	32
7.1	Acurácia e precisão como propriedades de uma medida.	43
7.2	Viés e variabilidade de uma medida.	44
15.1	Distribuições e funções de probabilidade	72
15.2	Parâmetros de tendência central em distribuições assimétricas e normais.	74
15.3	Parâmetros de dispersão em distribuições normais.	75
17.1	Gráficos de barras representando médias, barras de erro e dados individuais.	88
19.1	Representação gráfica de um teste de hipótese (unicaudal).	95
19.2	Representação gráfica de um teste de hipótese (bicaudal).	95
19.3	Representação gráfica dos erros tipo I e tipo II em um teste de hipótese (bicaudal).	100
19.4	Representação gráfica do erro tipo S (sinal) em um teste de hipótese (bicaudal).	101
19.5	Representação gráfica do erro tipo M (magnitude) em um teste de hipótese (bicaudal).	102
20.1	Análise de efeito de interação (direta) entre grupos e tempo. Retas paralelas sugerem ausência de efeito de interação.	105
20.2	Análise de efeito de interação (inversa) entre grupos e tempo. Retas paralelas sugerem ausência de efeito de interação.	105

25.1 Exemplo de diferentes forças e direção de correlação entre duas variáveis X e Y.	120
25.2 Gráfico de dispersão do Quarteto de Anscombe para representação gráfica de conjuntos de dados bivariados com parâmetros quase idênticos e relações muito distintas.	122
26.1 Regressão linear.	129
26.2 Regressão não-linear.	129
26.3 Regressão polinomial.	130
26.4 Regressão ridge.	130
26.5 Regressão logística.	131
26.6 Multicolinearidade entre variáveis candidatas em modelos de regressão multivariável.	132
37.1 Árvore de frequência do desempenho diagnóstico de uma tabela de confusão 2x2 representando um método novo (dicotômico) comparado ao método padrão-ouro ou referência (dicotômico).	168
44.1 RCTapp: Shiny app para análise de ensaios clínicos aleatorizados.	191

Lista de Tabelas

7.1	Tabela de dados brutos com medidas únicas.	40
7.2	Tabela de dados brutos com medidas repetidas.	40
7.3	Tabela de dados brutos com medidas repetidas agregadas.	41
7.4	Tabela de dados brutos com medidas seriadas não agregadas.	41
7.5	Tabela de dados brutos com medidas seriadas não agregadas.	42
7.6	Tabela de dados brutos com medidas múltiplas.	42
9.1	Estrutura básica de uma tabela de dados.	47
9.2	Formatação recomendada para tabela de dados.	48
9.3	Formatação não recomendada para tabela de dados.	48
17.1	Quantidade de casas decimais e dígitos significativos.	84
17.2	Valores originais, arredondamentos e erros de arredondamento por casas decimais.	84
19.1	Tabela de erros tipos I e II de inferência estatística.	100
19.2	Tabela de erro tipo S de inferência estatística.	100
19.3	Tabela de erro tipo M de inferência estatística.	101
22.1	Teste Qui-quadrado (com correção de Yates)	112
22.2	Teste Qui-quadrado (sem correção de Yates)	113
22.3	Teste exato de Fisher	114
25.1	Quarteto de Anscombe.	121
25.2	Análise descritiva do Quarteto de Anscombe demonstrando os conjuntos de dados bivariados com parâmetros quase idênticos.	121
36.1	Tabela de confusão sobre propriedades psicométricas de instrumentos.	161
36.2	Tabela de confusão 2x2 para análise de concordância de testes e variáveis dicotômicas.	163
36.3	Tabela de confusão 3x3 para análise de concordância de testes e variáveis dicotômicas.	164
37.1	Tabela de confusão 2x2 para análise de desempenho diagnóstico de testes e variáveis dicotômicas.	167
37.2	Probabilidades calculados a partir da tabela de confusão 2x2 para análise de desempenho diagnóstico de testes e variáveis dicotômicas.	169

Copyright © 2023-2025 Arthur de Sá Ferreira

Todos os direitos reservados. Nenhuma parte deste livro pode ser reproduzida ou usada de qualquer maneira sem a permissão prévia por escrito do proprietário dos direitos autorais, exceto para o uso de breves citações em uma resenha do livro.

Para solicitar permissões, entre em contato com cienciacomr@gmail.com

Capa dura: ISBN

Brochura: ISBN

E-book: ISBN

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

Dedicatória

Esta obra é dedicada a todos que, em princípio, buscam conhecimento para melhorar a qualidade da pesquisa científica - seja a sua própria, a de colegas ou a de desconhecidos - mas, em última análise, desejam mesmo prover melhores condições de saúde e desenvolvimento da sociedade.

Dedico também ao leitor eventual que chegou aqui por acaso.

RASCUNHO

Agradecimentos

Este trabalho não seria possível sem o apoio e suporte da minha esposa Daniele, minha irmã Mônica, meu pai José Victorino, minha mãe Angela (*in memoriam*) e meus filhos Giovanna, Victor e Lucas.

RASCUNHO

Sobre o autor



Arthur de Sá Ferreira

Obtive minha Graduação em Fisioterapia pela Universidade Federal do Rio de Janeiro (UFRJ, 1999), Formação em Acupuntura pela Academia Brasileira de Arte e Ciência Oriental (ABACO, 2001), Mestrado em Engenharia Biomédica pela Universidade Federal do Rio de Janeiro (UFRJ, 2002) e Doutorado em Engenharia Biomédica pela Universidade Federal do Rio de Janeiro (UFRJ, 2006).

Tenho experiência em docência no ensino superior, atuei com professor da graduação em cursos de Fisioterapia, Enfermagem e Odontologia, entre outros (2001-2018); pós-graduação *lato sensu* em Fisioterapia (2001-atual) e *stricto sensu* níveis mestrado e doutorado (2010-atual).

Como pesquisador, sou Professor Adjunto do Centro Universitário Augusto Motta (UNISUAM), atuando nos Programas de Pós-graduação em Ciências da Reabilitação (PPGCR; 2009-atual) e Desenvolvimento Local (PPGDL; 2018-atual). Também sou pesquisador do Instituto D'Or de Pesquisa e Ensino (IDOR; 2024-atual). Fundei o Laboratório de Simulação Computacional e Modelagem em Reabilitação (LSCMR) em 2012, onde desenvolvo projetos de pesquisa principalmente nos seguintes temas: Bioestatística, Modelagem e simulação computacional, Processamento de sinais biomédicos, Movimento funcional humano, Medicina tradicional (chinesa), Distúrbios musculoesqueléticos, Doenças cardiovasculares e Doenças respiratórias.

Dentre os editais públicos que obtive financimento, destaco o Jovem Cientista do Nossa Estado (JCNE; 2012-2015; 2015-2017) e Cientista do Nossa Estado (2021-atual) pela Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ; e Bolsista Produtividade em Pesquisa nível PQ2 pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq; 2021-atual).

Como gestor, estou na Coordenação do Programa de Pós-Graduação *stricto sensu* em Ciências da Reabilitação (PPGCR; 2016-atual). Atuei como coordenador do Comitê de Ética em Pesquisa (CEP) do Centro Universitário Augusto Motta (UNISUAM; 2020-2024) e como Coordenador do Curso de Graduação em Fisioterapia da Universidade Salgado de Oliveira (UNIVERSO; 2004-2009). Atualmente coordeno o Curso Superior de Tecnologia em Radiologia da Faculdade IDOR de Ciências Médicas (IDOR; 2024-atual).

Sou membro efetivo da Associação Brasileira de Pesquisa e Pós-Graduação em Fisioterapia (ABRAPG-FT) (2007-atual), Consórcio Acadêmico Brasileiro de Saúde Integrativa (CABSIN) (2019-atual), Committee on Publication Ethics (COPE) (2018-atual) e Royal Statistical Society (RSS) (2021-atual).

Componho o corpo editorial e de revisores de periódicos nacionais e internacionais como *Scientific Reports*, *Frontiers in Rehabilitation Sciences*, *The Journal of Clinical Hypertension*, *Chinese Journal of Integrative Medicine*, *Journal of Integrative Medicine*, *Brazilian Journal of Physical Therapy*, Fisioterapia e Pesquisa.

Curículos externos

5432142731317894

0000-0001-7014-2002

F-6831-2012

RASCUNHO

Prefácio

No âmbito da análise estatística de dados, os processos envolvidos são marcados por uma série de escolhas críticas. Estas decisões abrangem considerações metodológicas e ações operacionais que moldam toda a jornada analítica. Deve-se selecionar, cuidadosamente, um delineamento de estudo para enfrentar os desafios únicos colocados por um projeto de pesquisa. Além disso, a escolha de métodos estatísticos adequados para lidar com os dados gerados pelo delineamento escolhido tem um peso importante. Estas decisões necessitam de uma base construída sobre as evidências mais convincentes da literatura existente e na adesão a práticas sólidas de investigação.

Interpretar os resultados destas análises não é uma tarefa simples. Confiar apenas na formação educacional convencional, no bom senso e na intuição para decifrar tabelas e gráficos pode revelar-se inadequado. Interpretações errôneas podem gerar consequências indesejáveis, incluindo a utilização de testes diagnósticos imprecisos ou o endosso de tratamentos ineficazes.

Este livro emerge do reconhecimento desses desafios.

A proposta gira em torno da organização de um compêndio abrangente de métodos e técnicas de ponta, para análise estatística de dados em pesquisa científica, apresentados em formato de perguntas e respostas. Esse formato promove um diálogo direto e objetivo com o leitor, respondendo a dúvidas comumente colocadas por alunos de graduação, pós-graduação, mestrado e doutorado, bem como por pesquisadores.

O objetivo geral de cada capítulo é elucidar as questões metodológicas fundamentais: “*O que é?*”, “*Por que usar?*”, “*Quando usar?*”, “*Quando não usar?*” e “*Como fazer?*”. Em cada capítulo, diversas questões específicas são propostas e respondidas sistematicamente, permitindo ao leitor uma melhor elaboração do conteúdo e resultado do seu trabalho.

Os capítulos foram organizados para seguir uma progressão de conceitos e aplicações. Embora sejam fragmentados para maior clareza instrucional, as referências cruzadas ajudam a mitigar a fragmentação do conteúdo e reforçar a interconexão dos tópicos.

O público-alvo comprehende pesquisadores, professores, analistas de dados, profissionais e estudantes que regularmente lidam com a tomada de decisões em pesquisa. Os estudantes de pós-graduação encontrarão aqui uma obra repleta de exemplos para adaptar na análise dos dados de seus projetos de pesquisa. Professores de graduação e pós-graduação terão acesso a uma obra didática de referência, direcionada para seus alunos. Pesquisadores e analistas de dados iniciantes descobrirão um valioso acervo de informações e referências para a construção de projetos e manuscritos. Pesquisadores e os cientistas mais experientes podem recorrer às referências e esclarecimentos mais atuais sobre vieses, paradoxos, mitos e mal práticas em pesquisa. E mesmo os leitores não familiarizados ainda com as técnicas de análise de dados em pesquisa terão a oportunidade de apreciar o papel fundamental de colocar e responder suas perguntas na busca do conhecimento científico.

Arthur de Sá Ferreira

RASCUNHO

PARTE 1: PENSAMENTO ESTATÍSTICO E METODOLÓGICO

Pense como um cientista

RASCUNHO

Capítulo 1

Pensamento probabilístico

1.1 Experimento

1.1.1 O que é um experimento?

- Um experimento é um processo de simulação ou medição cujo resultado é chamado de desfecho.¹
- Tentativa se refere a uma repetição de um experimento.¹

1.1.2 O que é um experimento aleatório?

- Em um experimento aleatório, o desfecho de cada tentativa é imprevisível.¹

1.2 Espaço amostral e eventos discretos

1.2.1 O que é espaço amostral discreto?

- O espaço amostral S de um experimento aleatório é definido como o conjunto de todos os desfechos possíveis de um experimento.¹
- Em probabilidade discreta, o espaço amostral S pode ser enumerado e contado.¹

1.2.2 O que é evento discreto?

- Um evento E é um único desfecho ou uma coleção de desfechos.¹
- Um evento E é um subconjunto do espaço amostral S de um experimento.¹

1.2.3 O que é espaço de eventos discretos?

- Um espaço de eventos E_s também é um subconjunto do espaço amostral S de um experimento.¹
- A união de dois eventos $E_1 \cup E_2$ é o conjunto de todos os desfechos que estão em ambos.¹
- A intersecção de dois eventos $E_1 \cap E_2$, ou evento conjunto, é o conjunto de todos os desfechos que estão em ambos os eventos.¹
- O complemento de um evento E^C consiste em todos os desfechos que não estão incluídos no evento E .¹

1.3 Espaço amostral e eventos contínuos

1.3.1 O que é espaço amostral contínuo?

- ?

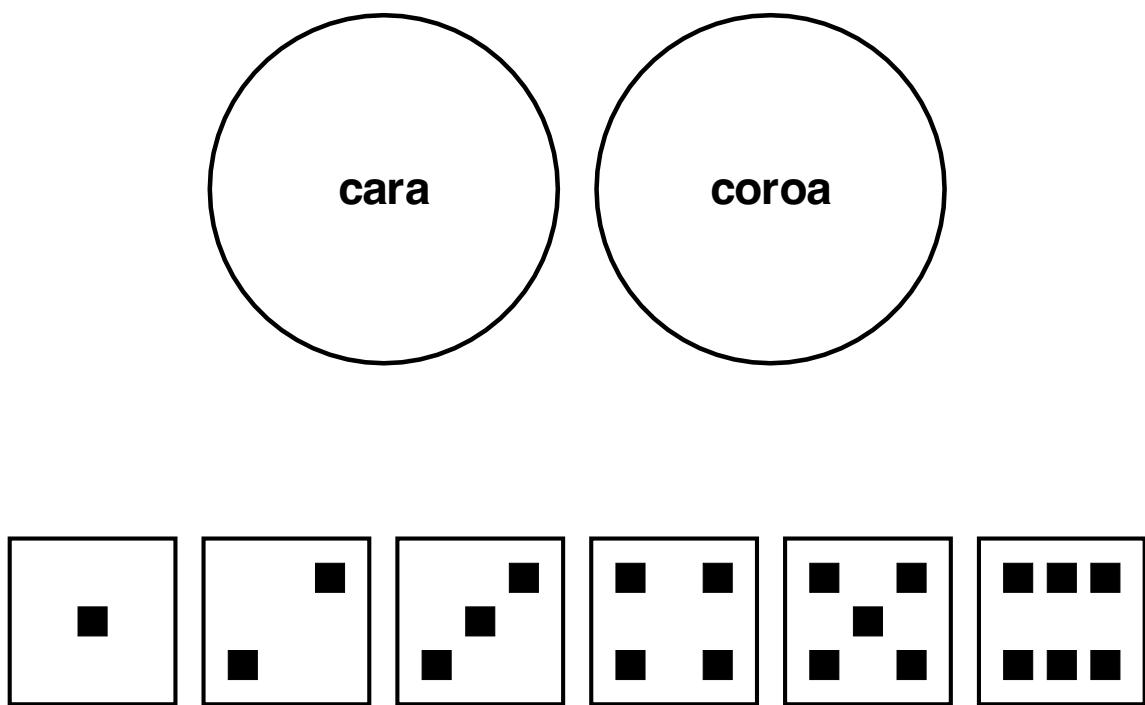


Figura 1.1: Exemplos de espaço amostral discreto. Superior: Todas as faces de uma moeda. Inferior: Todas as faces de um dado.

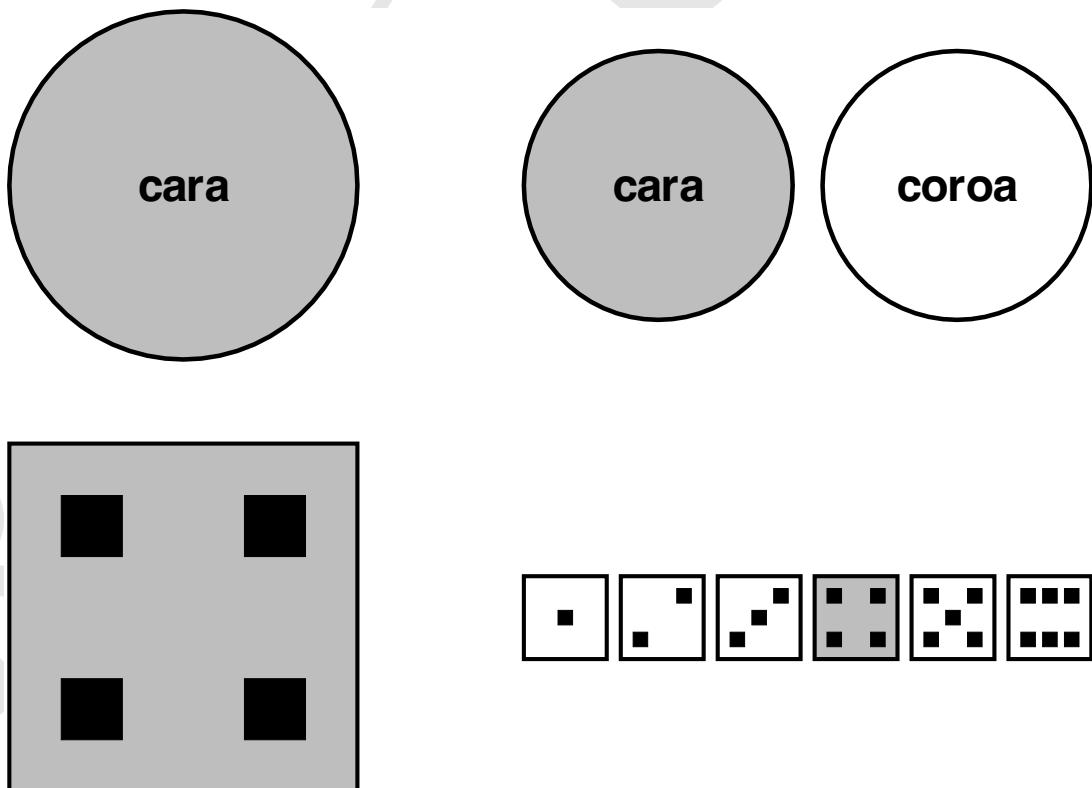


Figura 1.2: Exemplos de evento de experimento. Superior: 1 lançamento de 1 moeda. Inferior: 1 lançamento de 1 dado.

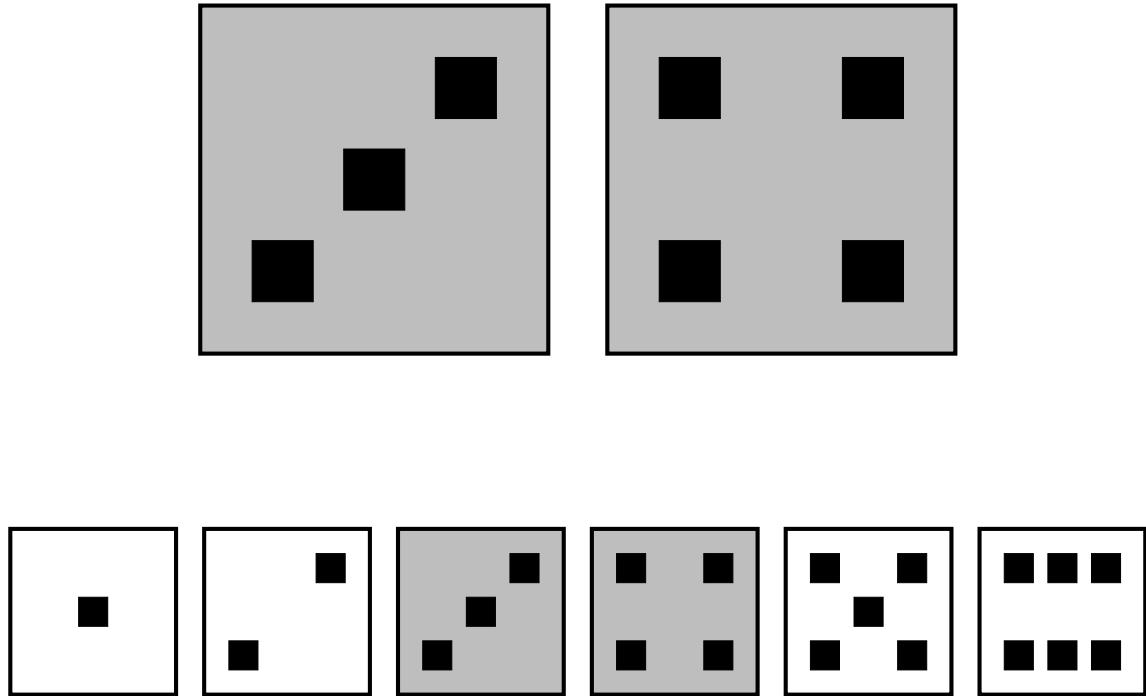


Figura 1.3: Espaço de eventos: União dos eventos face = 3 e face = 4 de um dado.

1.3.2 O que é evento contínuo?

- ?

1.3.3 O que é espaço de eventos contínuo?

- ?

1.4 Probabilidade

1.4.1 O que é probabilidade?

- Com um espaço amostral S finito e não vazio de desfechos igualmente prováveis, a probabilidade P de um evento E é a razão entre o número de desfechos no evento E e o número de desfechos no espaço amostral S .¹
- Um evento E impossível não contém um desfecho e, portanto, nunca ocorre: $P(E) = 0$.^{1,2}
- Um evento E é certo consiste em qualquer um dos desfechos possíveis e, portanto, sempre ocorre: $P(E) = 1$.¹

1.4.2 Quais são os axiomas da probabilidade?

- A probabilidade de um evento é um número real que satisfaz os seguintes axiomas descritos por Andrei Nikolaevich Kolmogorov em 1950:^{1,2}
 - Axioma I. Probabilidades de um evento E são números não-negativos: $P(E) \geq 0$.
 - Axioma II. Probabilidade de todos os eventos do espaço amostral A ocorrerem é 100%: $P(S) = 1$.
 - Axioma III. A probabilidade de um conjunto k de eventos mutuamente exclusivos é igual a soma da probabilidade de cada evento: $P(E_1 \cup E_2 \cup \dots \cup E_k) = P(E_1) + P(E_2) + \dots + P(E_k)$.
- Os axiomas possuem as seguintes consequências:¹
 - A soma da probabilidade de dois eventos que dividem o espaço amostral é 100%: $P(E) + P(E^C) = 1$.

- O valor máximo de probabilidade de um evento é 100%: $P(S) \leq 1$.
- A probabilidade é uma função não decrescente do número de desfechos de um evento.

1.5 Independência e probabilidade

1.5.1 O que é independência em estatística?

- Em experimentos aleatórios, é comum assumir que os eventos de tentativas separadas são independentes devido a independência física de eventos e experimentos.¹
- Se a ocorrência do evento E_1 não tiver efeito na ocorrência do evento E_2 , os eventos E_1 e E_2 são considerados estatisticamente independentes.¹
- Eventos são mutuamente exclusivos, ou disjuntos, se a ocorrência de um exclui a ocorrência dos outros.¹
- Se dois eventos E_1 e E_2 são mutuamente exclusivos, então os eventos E_1 e E_2 não podem ocorrer ao mesmo tempo e, portanto, são eventos independentes.¹

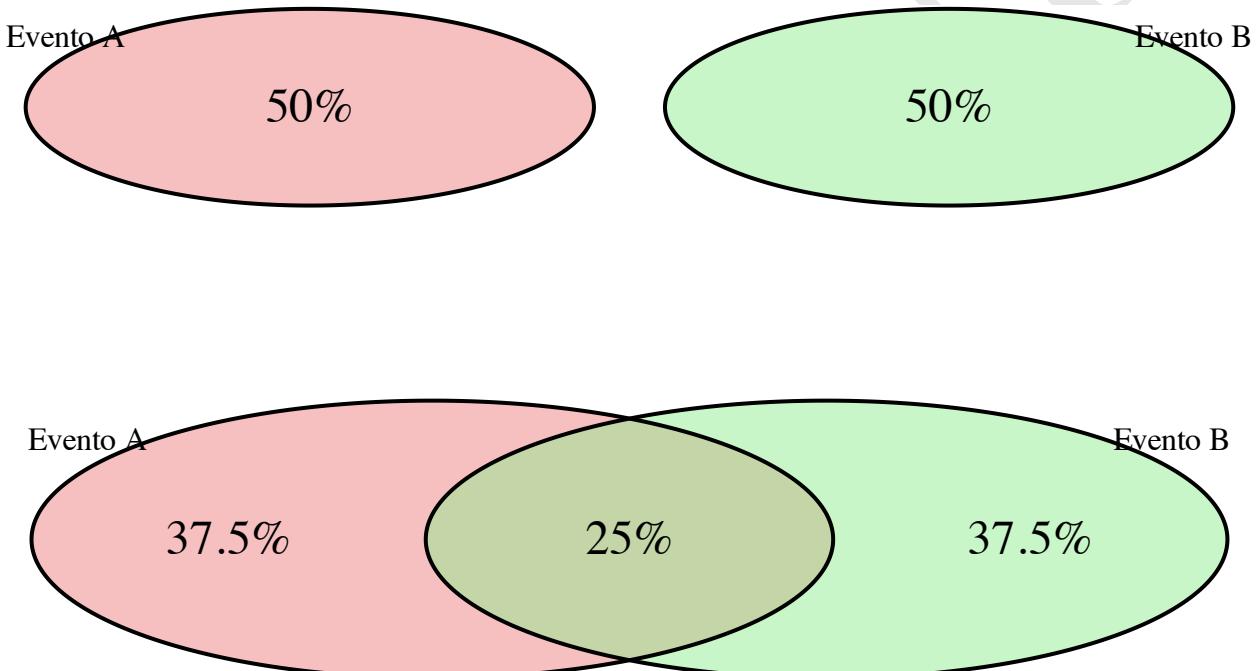


Figura 1.4: Superior: Eventos independentes. Inferior: Eventos dependentes.

- Em experimentos independentes, o desfecho de uma tentativa é independente dos desfechos de outras tentativas, passadas e/ou futuras. Uma tentativa em um experimento aleatório é independente se a probabilidade de cada desfecho possível não mudar de tentativa para tentativa.¹

1.5.2 O que é probabilidade marginal?

- Probabilidade marginal é a probabilidade de ocorrência de um evento E independentemente da(s) probabilidade(s) de outro(s) evento(s).¹

1.5.3 O que é probabilidade conjunta?

- Probabilidade conjunta é a probabilidade de ocorrência de dois ou mais eventos independentes E_1, E_2, \dots, E_k , independentemente da(s) probabilidade(s) de outro(s) evento(s).¹
- Se a probabilidade conjunta dos eventos é nula ($E_1 \cup E_2 = \emptyset$), esses dois eventos E_1 e E_2 são mutuamente exclusivos ou disjuntos.¹

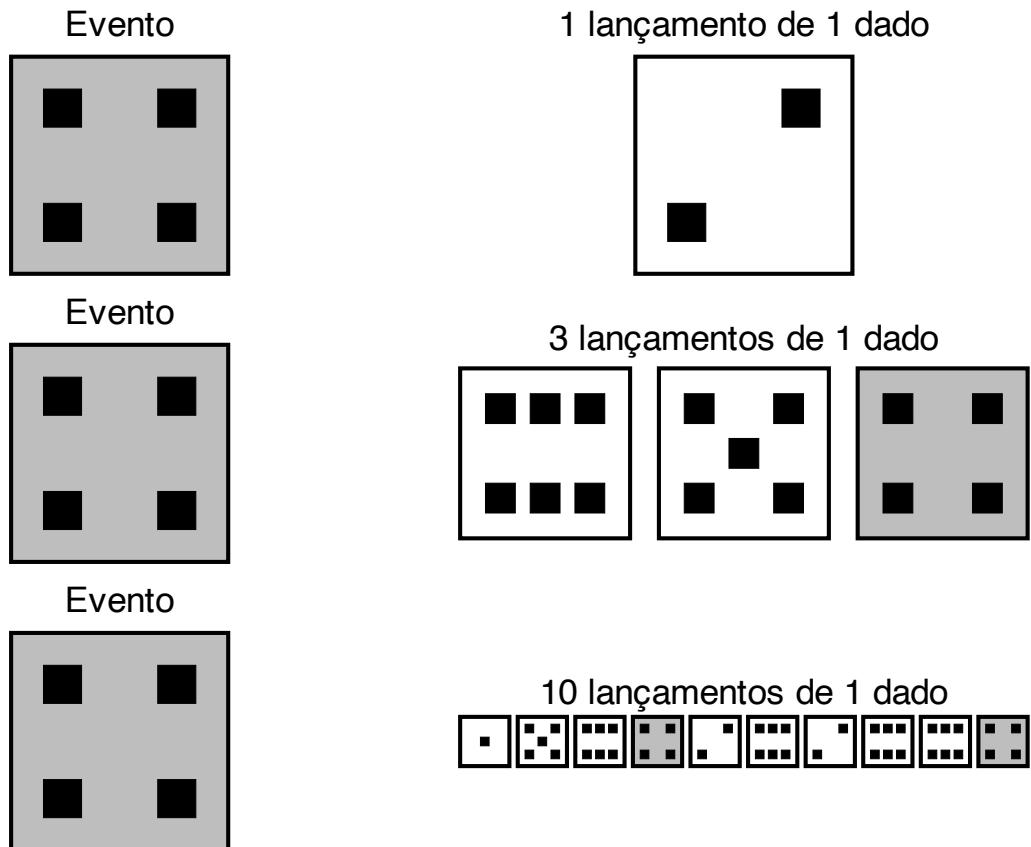


Figura 1.5: Esquerda: Evento (face = 4). Direita: Experimentos de 1 lançamento de 1 dado (superior), 3 lançamentos de 1 dado (central), 10 lançamentos de 1 dado (inferior).

1.5.4 O que é probabilidade condicional?

- Probabilidade condicional é a probabilidade de ocorrência do evento E_2 quando se sabe que o evento E_1 já ocorreu $P(E_2|E_1)$.¹
- A probabilidade condicional $P(E_2|E_1)$ representa que a ocorrência do evento E_1 fornece informação sobre a ocorrência do evento E_2 .¹
- Se a ocorrência do evento E_1 tiver alguma influência na ocorrência do evento E_2 , então a probabilidade condicional do evento E_2 dado o evento E_1 pode ser maior ou menor do que a probabilidade marginal.¹

1.6 Leis dos números anômalos

1.6.1 O que é a lei dos números anômalos?

- A lei dos números anômalos - lei de Benford - é uma distribuição de probabilidade que descreve a frequência de ocorrência do primeiro dígito em muitos conjuntos de dados do mundo real.³

1.7 Leis dos pequenos números

1.7.1 O que é a lei dos pequenos números?

- A crença exagerada na probabilidade de replicar com sucesso os achados de um estudo, pela tendência de se considerar uma amostra como representativa da população.⁴
- A crença na lei dos pequenos números se refere à tendência de superestimar a estabilidade das estimativas provenientes de estudos com amostras pequenas.⁵
- Quando se percebe um padrão, pode não ser possível identificar se tal padrão é real.⁶

1.7.2 Quais são as versões da lei dos pequenos números?

- 1a Lei Forte dos Pequenos Números: “Não há pequenos números suficientes para atender às muitas demandas que lhes são feitas”.⁶
- 2a Lei Forte dos Pequenos Números: “Quando dois números parecem iguais, não são necessariamente assim”.⁷

1.8 Leis dos grandes números

1.8.1 O que é a lei dos grandes números?

- A lei dos grandes números descreve que, ao realizar o mesmo experimento E um grande número de vezes (n), a média μ dos resultados obtidos tende a se aproximar do valor esperado $E[\bar{X}]$ à medida que mais experimentos forem realizados ($n \rightarrow \infty$).⁸
- De acordo com a lei dos grandes números, a média amostral converge para a média populacional à medida que o tamanho da amostra aumenta.⁹

1.8.2 Quais são as versões da lei dos grandes números?

- Lei Fraca dos Grandes Números (de Poisson): ““.”?
- Lei Fraca dos Grandes Números (de Bernoulli): ““.”?
- Lei Forte dos Grandes Números: ““.”?

1.9 Teorema central do limite

1.9.1 O que é teorema central do limite?

- O teorema central do limite - equação (1.1) - afirma que, para uma amostra aleatória de tamanho n de uma população com valor esperado igual à média $E[\bar{X}_i] = \mu$ e variância $Var[\bar{X}_i] = \sigma^2$, a distribuição amostral da média de uma variável \bar{X} se aproxima de uma distribuição normal N com média μ e variância σ^2/n à medida que n aumenta ($n \rightarrow \infty$):⁸

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{n \rightarrow \infty} N(0, \sigma^2) \quad (1.1)$$

- O teorema central do limite demonstra que se o tamanho da amostra n for suficientemente grande, a distribuição amostral das médias obtidas utilizando reamostragem com substituição será aproximadamente normal, com média μ e variância σ^2/n , independentemente da distribuição da população.⁸
- No exemplo abaixo, uma variável aleatória numérica com distribuição uniforme no espaço amostral $S = [18; 65]$ tem média $\mu = 38.53$ e variância $\sigma^2 = 172.433$. A distribuição amostral da média de 100 amostras de tamanho 5, 50, 500 e 5000 tomadas da população com reposição e igual probabilidade se aproxima de uma distribuição normal com média $\mu = 38.493$ e variância $\sigma^2 = 0.038$, independentemente da distribuição da população:
- Em outro exemplo, o lançamento de um dado com distribuição uniforme no espaço amostral $S = \{1, 2, 3, 4, 5, 6\}$ tem média $\mu = 3.77$ e variância $\sigma^2 = 3.169$. A distribuição amostral da média de 100 amostras de tamanho 5, 50, 500 e 5000 tomadas da população com reposição e igual probabilidade se aproxima de uma distribuição normal com média $\mu = 3.767$ e variância $\sigma^2 = 0.001$, independentemente da distribuição da população:
- Mais um exemplo, o lançamento de uma moeda com distribuição uniforme no espaço amostral $S = \{0, 1\}$ — codificado para *sucesso* = 1 e *insucesso* = 0 — tem média $\mu = 0.48$ e variância $\sigma^2 = 0.252$. A distribuição amostral da média de 100 amostras de tamanho 5, 50, 500 e 5000 tomadas da população com reposição e igual probabilidade se aproxima de uma distribuição normal com média $\mu = 0.48$ e variância $\sigma^2 = 0$, independentemente da distribuição da população:

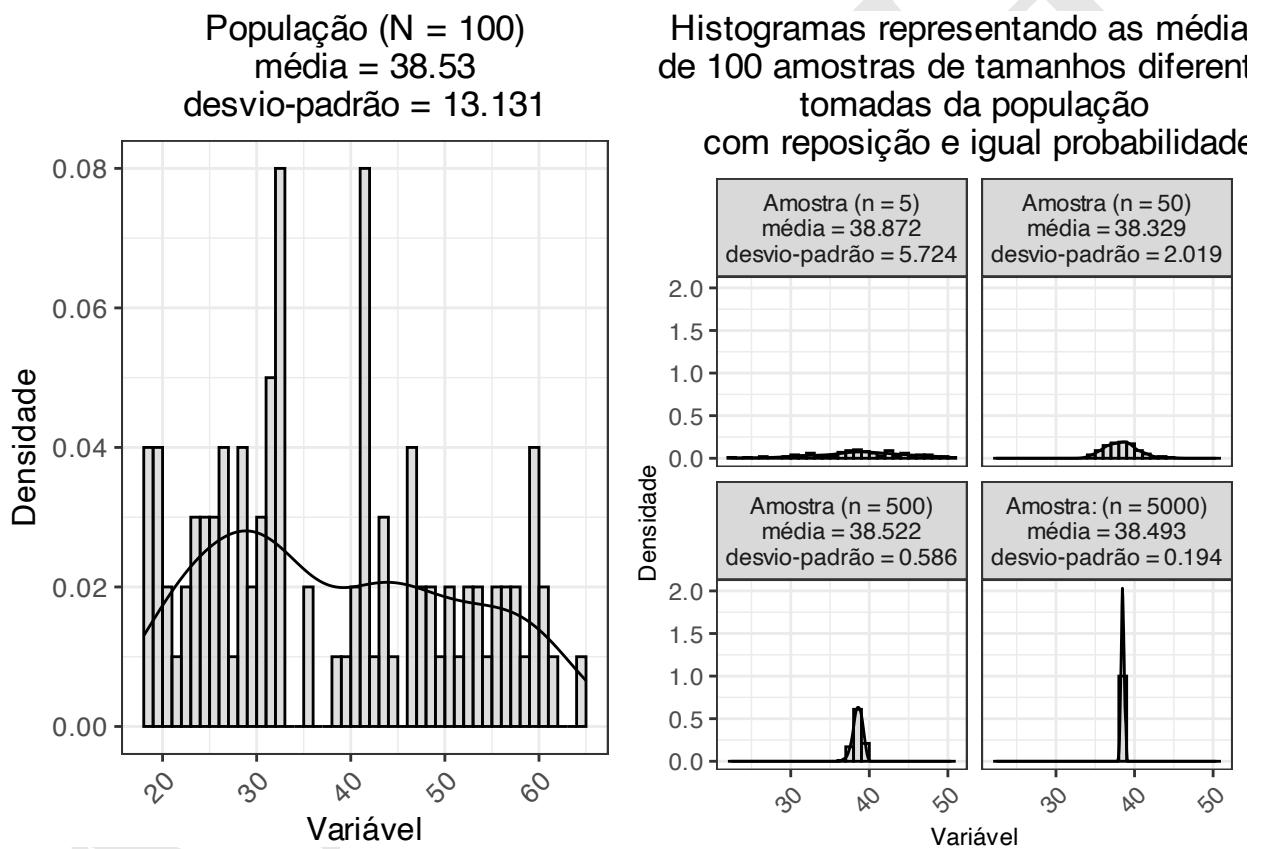


Figura 1.6: Esquerda: Histogramas de uma variável aleatória com distribuição uniforme ($N = 100$). Direita: Histogramas da média de 100 amostras de tamanhos 5, 50, 500 e 5000 tomadas da população com reposição e igual probabilidade.

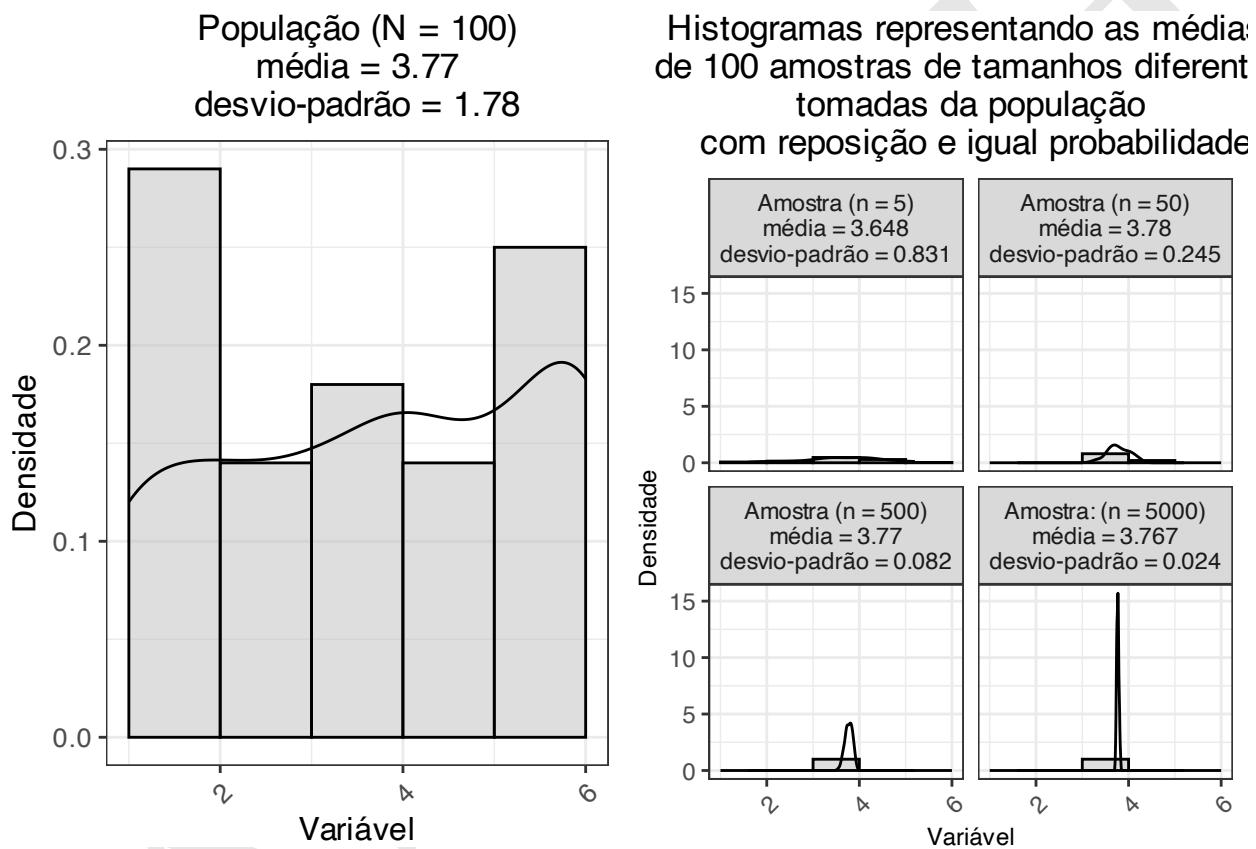


Figura 1.7: Esquerda: Histogramas de lançamento de 1 dado com distribuição uniforme (N = 100). Direita: Histogramas da média de 100 amostras de tamanhos 5, 50, 500 e 5000 tomadas da população com reposição e igual probabilidade.

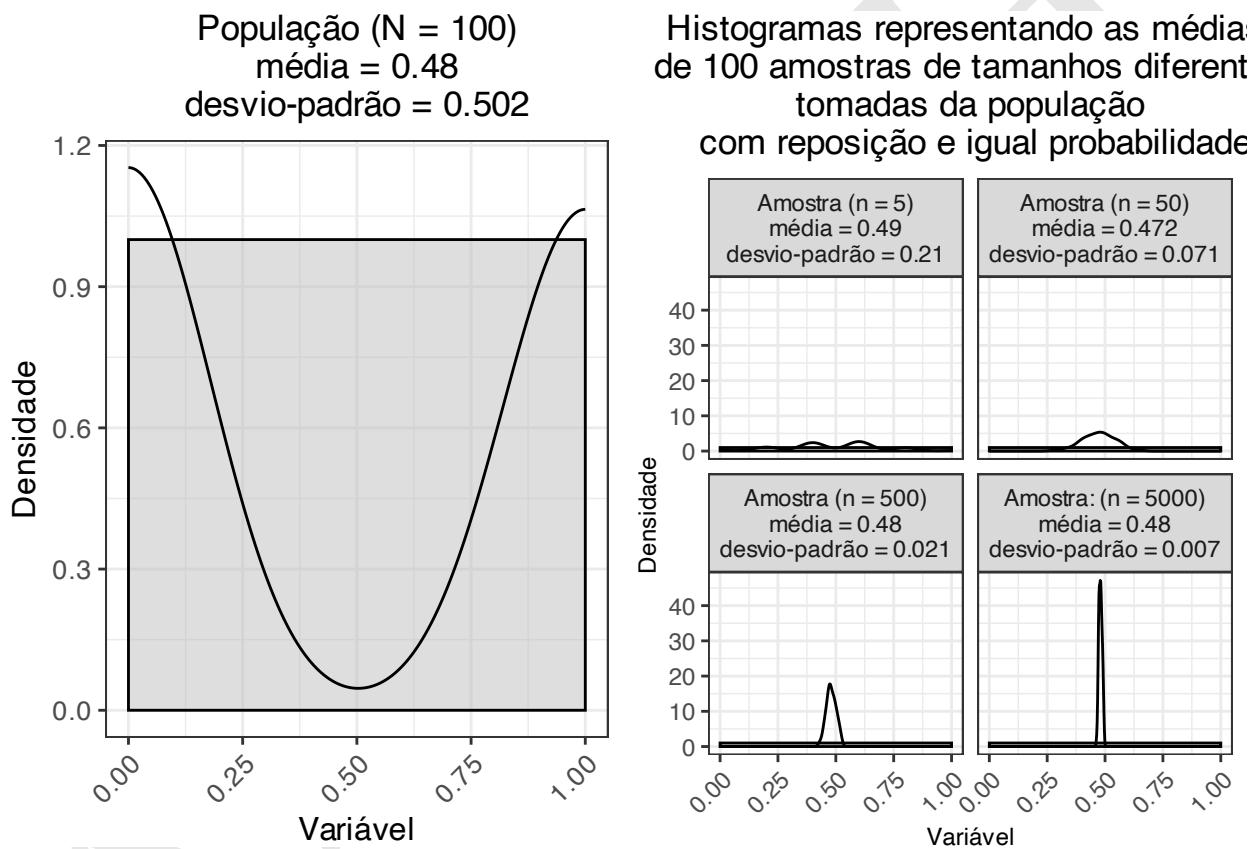


Figura 1.8: Esquerda: Histogramas de lançamento de 1 moeda com distribuição uniforme ($N = 100$). Direita: Histogramas da média de 100 amostras de tamanhos 5, 50, 500 e 5000 tomadas da população com reposição e igual probabilidade.

1.9.2 Quais as condições de validade do teorema central do limite?

- As condições de validade do teorema central do limite são:⁸
 - As variáveis aleatórias devem ser independentes e identicamente distribuídas (*independent and identically distributed* ou i.i.d.);
 - As variáveis aleatórias devem ter média μ e variância σ^2 finitas;
 - O tamanho da amostra deve ser suficientemente grande (geralmente, $n \geq 30$).

1.9.3 Qual a relação entre a lei dos grandes números e o teorema central do limite?

- A lei dos grandes números é um precursor do teorema central do limite, pois estabelece que a média da amostra se torna cada vez mais próxima da média populacional (isto é, mais representativa) à medida que o tamanho da amostra aumenta, e o teorema central do limite demonstra que a distribuição da soma das variáveis aleatórias se aproxima de uma distribuição normal também à medida que o tamanho da amostra aumenta.⁹

1.9.4 Qual a relevância do teorema central do limite para a análise estatística?

- O teorema central do limite explica porque os testes paramétricos têm maior poder estatístico do que os testes não paramétricos, os quais não requerem suposições de distribuição de probabilidade.⁸
- O teorema central do limite implica que os métodos estatísticos que se aplicam a distribuições normais podem ser aplicados a outras distribuições quando suas suposições são satisfeitas.⁸
- Como o teorema central do limite determina a distribuição amostral Z - equação (1.2) - das médias com tamanho amostral suficientemente grande, a média pode ser padronizada para uma distribuição normal com média 0 e variância 1, $N(0, 1)$:⁸

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad (1.2)$$

- Para amostras com $n \geq 30$, a distribuição amostral Student- t se aproxima da distribuição normal padrão Z e, portanto, as suposições sobre a distribuição populacional não são mais necessárias de acordo com o teorema central do limite. Neste cenário, a suposição de distribuição normal pode ser usada para a distribuição de probabilidade.⁸

1.10 Regressão para a média

1.10.1 O que é regressão para a média?

- Regressão para a média⁹ é um fenômeno estatístico que ocorre quando uma variável aleatória X é medida na mesma unidade de análise em dois ou mais momentos diferentes, X_1, X_2, \dots, X_t e X_t é mais próximo da média populacional do que X_1 , ou seja, $E(X_t)$ é mais próxima de $E(X)$ do que $E(X_1)$ é de $E(X)$.¹⁰
- O valor real - sem erros aleatório ou sistemático - em geral não é conhecido, mas pode ser estimado pela média de várias observações.¹⁰
- Regressão para a média pode ocorrer em qualquer pesquisa cujo delineamento envolva medidas repetidas.¹¹
- Em medidas repetidas, a média de várias observações é mais próxima da média verdadeira do que qualquer observação individual, pois o erro aleatório é reduzido pela média.¹⁰
- Valores extremos - em direção ao mínimo ou máximo - em uma medição inicial tendem a ser seguidos por valores mais próximos da média (valor real) na medição subsequente.¹⁰
- No exemplo abaixo, a 2a medida (dado 2 = 121) é mais próxima da média (valor real = 120) do que a 1a medida (dado 1 = 118):

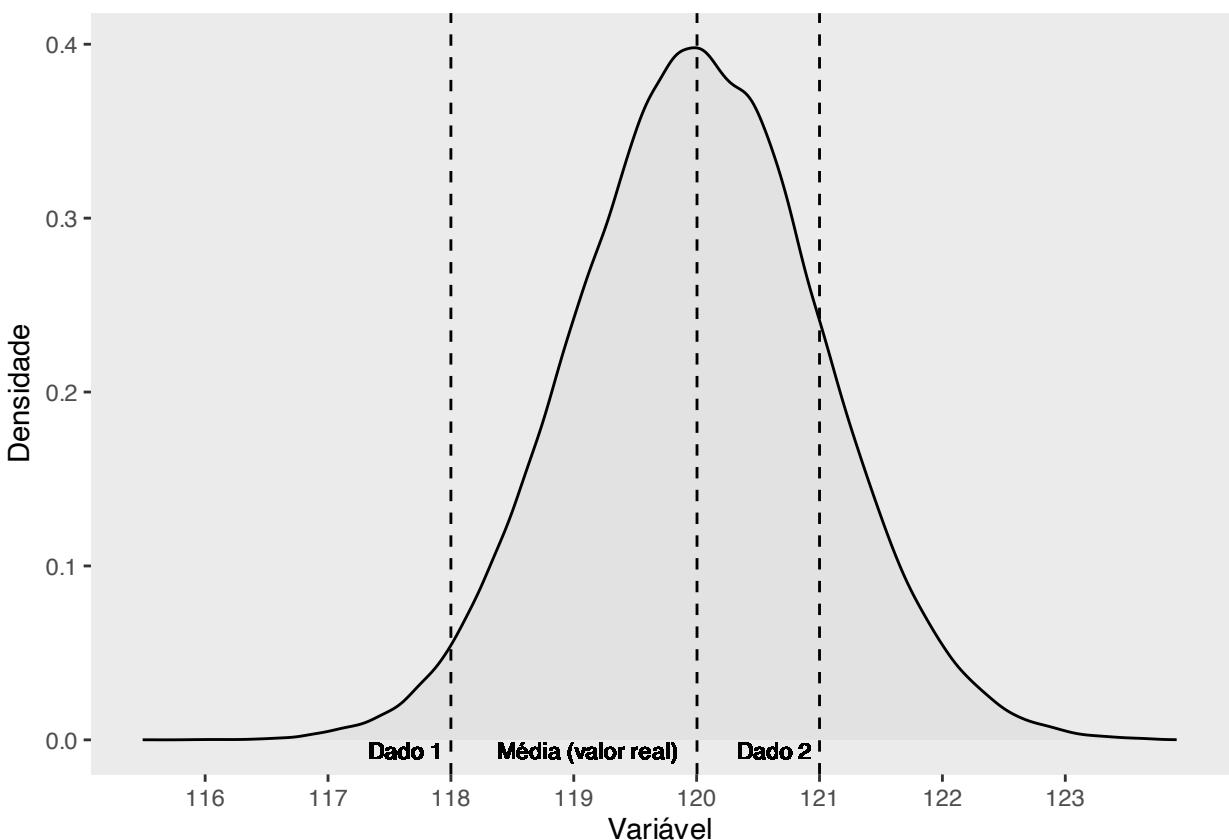


Figura 1.9: Representação gráfica da regressão para a média em medidas repetidas. A segunda medida (dado 2) é mais próxima da média (valor real) do que a primeira medida (dado 1).

1.10.2 Qual a causa da regressão para a média?

- A regressão para a média pode ser atribuída ao erro aleatório, que é a variação não sistemática nos valores observados em torno de uma média verdadeira (por exemplo, erro de medição aleatório ou variações aleatórias em um participante).¹⁰
- Regressão para a média é uma consequência da observação de que dados extremos não se repetem com frequência.¹¹
- Deve-se assumir que a regressão para a média ocorreu até que os dados mostrem o contrário.¹⁰

1.10.3 Por que detectar o fenômeno de regressão para a média?

- A regressão para a média pode levar a conclusões errôneas sobre a eficácia de uma intervenção, pois a mudança observada pode ser devida ao erro aleatório e não ao tratamento.¹¹

1.10.4 Com detectar o fenômeno de regressão para a média?

- O fenômeno de regressão para a média pode ser detectado por meio de gráfico de dispersão da diferença (estudos transversais) ou mudança (estudos longitudinais) versus os valores da 1a medida.¹⁰

R

O pacote *regtomean*¹² fornece as funções *cordata*^a para calcular a correlação entre medidas tipo antes-e-depois e *meechua_reg*^b para ajustar modelos lineares de regressão.

^a<https://www.rdocumentation.org/packages/regtomean/versions/1.1/topics/cordata>

^bhttps://www.rdocumentation.org/packages/regtomean/versions/1.1/topics/meechua_reg

1.10.5 Como o fenômeno de regressão para a média pode ser evitado?

- Aloque os participantes de modo aleatório nos grupos de tratamento e controle pode reduzir o fenômeno de regressão para a média.¹⁰
- Selecione participantes com base em medidas repetidas ao invés de medidas únicas.¹⁰

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

Capítulo 2

Pensamento estatístico

2.1 População e Amostra

2.1.1 O que é população?

- População — ou população-alvo — refere-se ao conjunto completo sobre o qual se pretende obter informações.¹³
- População é metodologicamente delimitada pelos critérios de inclusão e exclusão do estudo.¹³
- Em estudos observacionais, inicialmente as características geográficas e/ou demográficas, por exemplo, definem a população a ser estudada.¹³
- Em estudos analíticos, a população é inicialmente definida pelos objetivos da pesquisa e, posteriormente, as observações são realizadas na amostra.¹³

2.1.2 O que é amostra?

- Amostra é uma parte finita da população do estudo.¹³
- Em pesquisa científica, utilizam-se dados de uma amostra de participantes (ou outras unidades de análise) para realizar inferências sobre a população.¹⁴

2.1.3 Por que usar dados de amostras?

- Dados de uma amostra de tamanho suficiente e características representativas podem ser utilizados para inferência sobre uma população.⁸
- Em geral, amostras de tamanhos maiores possuem médias mais próximas da média populacional e menores variâncias.⁸

2.2 Unidade de análise

2.2.1 O que é unidade de análise?

- A unidade de análise (ou unidade experimental) de pesquisas na área de saúde geralmente é o indivíduo.¹⁵
- A unidade de análise também pode ser a instituição em estudos multicêntricos (ex.: hospitais, clínicas) ou um estudo publicado em meta-análise (ex.: ensaios clínicos).¹⁵

2.2.2 Por que identificar a unidade de análise de um estudo?

- É fundamental identificar corretamente a unidade de análise para evitar inflação do tamanho da amostra (ex.: medidas bilaterais resultando em o dobro de participantes), violações de suposições dos testes de hipótese (ex.: independência entre medidas e/ou unidade de análise) e resultados espúrios em testes de hipótese (ex.: P-valores menores que aqueles observados se a amostra não estivesse inflada).^{15,16}

2.2.3 Que medidas podem ser obtidas da unidade de análise de um estudo?

- Da unidade de análise podem ser coletadas informações em medidas únicas, repetidas, seriadas ou múltiplas.

2.3 Amostragem

2.3.1 O que é amostragem?

- ?

2.3.2 Quais métodos de amostragem são usados para obter uma amostra da população?

- O método de amostragem é geralmente definido pelas condições de viabilidade do estudo, no que diz respeito a acesso aos participantes, ao tempo de execução e aos custos envolvidos, entre outras.¹³
- Não-probabilísticas ou intencionais:¹³
 - Bola de neve.
 - Conveniência.
 - Participantes encaminhados.
- Probabilísticas:¹³
 - Simples.
 - Sistemática.
 - Multiestágio.
 - Estratificada.
 - Agregada.

2.3.3 O que é erro de amostragem?

- ?

2.4 Reamostragem

2.4.1 O que é reamostragem?

- Reamostragem é um procedimento que cria vários conjuntos de dados sorteados a partir de um conjunto de dados real - a amostra da população - sem a necessidade de fazer suposições sobre os dados e suas distribuições.¹⁴
- O procedimento é repetido várias vezes para usar a variabilidade dos resultados para obter um intervalo de confiança do parâmetro no nível de significância α pré-estabelecido.¹⁴

2.4.2 Por que utilizar reamostragem?

- Quando se dispõe de dados de apenas 1 amostra, as diversas suposições que são feitas podem não ser atingidas.¹⁴
- Procedimentos de reamostragem produzem um conjunto de observações escolhidas aleatoriamente da amostra, igualmente representativo da população original.¹⁴
- Procedimentos de reamostragem permitem estimar o erro-padrão e intervalos de confiança sem a necessidade de tais suposições, sendo, portanto, um conjunto de procedimentos não-paramétricos.¹⁴

2.4.3 Quais procedimentos de reamostragem podem ser realizados?

- *Bootstrap*: Cada iteração gera uma amostra *bootstrap* do mesmo tamanho do conjunto de dados original escolhendo aleatoriamente observações reais, uma de cada vez. Cada observação tem chances iguais de ser escolhida a cada vez, portanto, algumas observações serão escolhidas mais de uma vez e outras nem serão escolhidas.¹⁴

2.5 Subamostragem e superamostragem

2.5.1 O que é subamostragem?

- . ?

2.5.2 O que é superamostragem?

- . ?

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

RAASCUNHO

Capítulo 3

Pensamento metodológico

3.1 Metodologia da pesquisa

3.1.1 O que é metodologia da pesquisa?

- A utilização de um vocabulário próprio — incluindo termos frequentemente usados em metodologia, epidemiologia e estatística — facilita a discussão na comunidade científica e melhora a compreensão das publicações.^{17,18}

3.2 Relação Estatística-Metodologia

3.2.1 Qual a relação entre estatística e metodologia da pesquisa?

- ¹⁹

3.3 Reproduzibilidade

3.3.1 O que é reproduzibilidade?

- Reproduzibilidade é a habilidade de se obter resultados iguais ou similares quando uma análise ou teste estatístico é repetido.²⁰⁻²²

3.3.2 Por que reproduzibilidade é importante?

- Analisar a reproduzibilidade pode fornecer evidências a respeito da objetividade e confiabilidade dos achados, em detrimento de terem sido obtidos devido a vieses ou ao acaso.²⁰
- A reproduzibilidade não é apenas uma questão metodológica, mas também ética, uma vez que pode envolver mal práticas científicas como fabricação e/ou falsificação de dados.²⁰
- Reproduzibilidade pode ser considerada um padrão mínimo em pesquisa científica.²¹

3.4 Robustez

3.4.1 O que é robustez?

- [?]

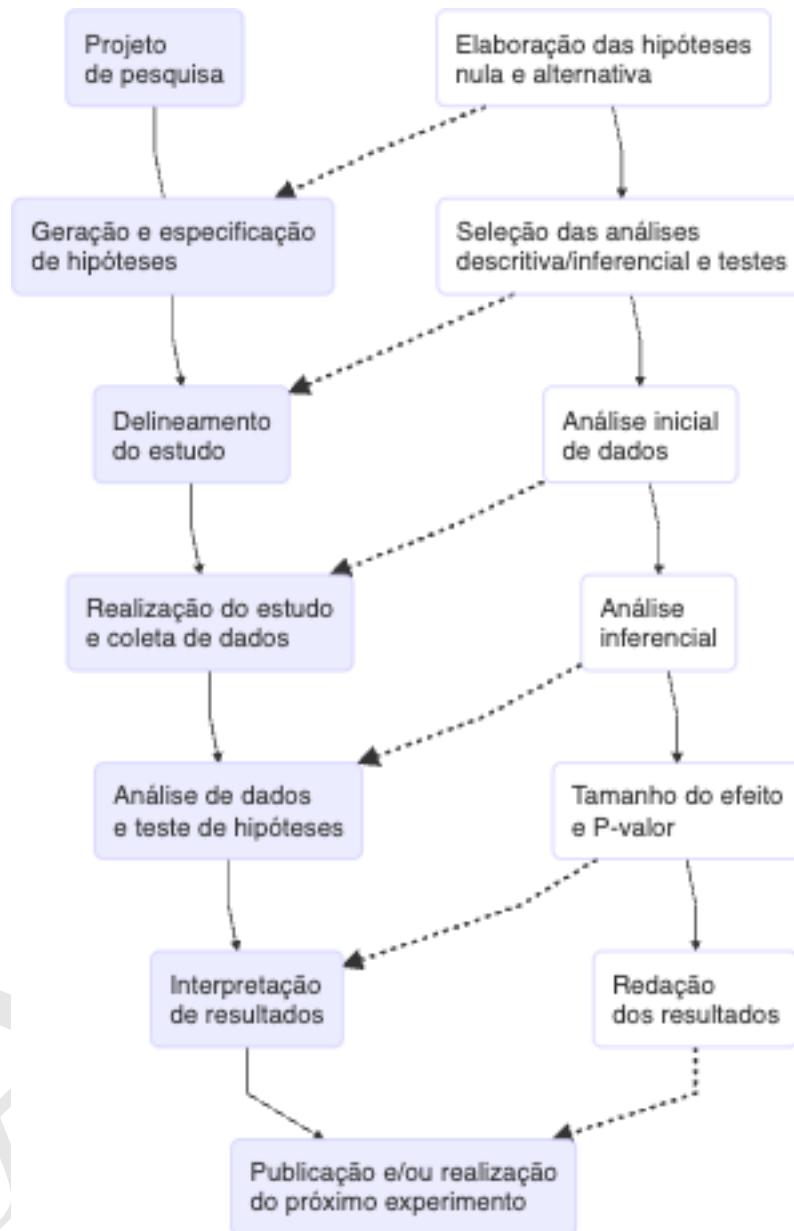


Figura 3.1: Mapa mental da relação entre o pensamento estatístico e o pensamento metodológico.

3.5 Replicabilidade

3.5.1 O que é replicabilidade?

- Replicabilidade é a habilidade de se obter conclusões iguais ou similares quando um experimento é repetido.^{21,22}

3.6 Generalização

3.6.1 O que é generalização?

- Generalização refere-se à extração das conclusões do estudo, observados na amostra, para a população.¹³

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

RASCUNHO

Capítulo 4

Pensamento computacional

4.1 Programas de computador

4.1.1 O que é R?

- R é um programa de computador de código aberto com linguagem computacional direcionada para análise estatística.^{23,24}
- R version 4.5.1 (2025-06-13) está disponível gratuitamente em *Comprehensive R Archive Network* (CRAN).²⁵

4.1.2 Por que usar R?

- R é o software de maior abrangência de métodos estatísticos, possui sintaxe que permite análises estatísticas reproduzíveis e está disponível gratuitamente no *Comprehensive R Archive Network* (CRAN).^{22,25}

4.1.3 O que é RStudio?

- RStudio é um ambiente de desenvolvimento integrado (*integrated development environment*, IDE) desenvolvido visando a reprodutibilidade e a simplicidade para a criação e disseminação de conhecimento.^{24,26}
- O ambiente do RStudio é dividido em painéis:
 - *Source/Script editor*: para edição de R scripts.²⁴
 - *Console*: para execução de códigos simples.²⁴
 - *Environments*: para visualização de objetos criados durante a sessão de trabalho.²⁴
 - *Output*: para visualização de gráficos criados durante a sessão de trabalho.²⁴
- As principais características do RStudio incluem um ambiente de edição com abas para acesso rápido a arquivos, comandos e resultados; histórico de comandos previamente utilizados; ferramentas para visualização de bancos de dados e elaboração de scripts e gráficos e tabelas.^{24,26}
- RStudio está disponível gratuitamente em Posit¹.

 O pacote *learnr*²⁷ fornece tutoriais interativos para RStudio.

4.1.4 Que programas de computador podem ser usados para análise estatística com R?

- JASP²⁸

¹<https://posit.co/download/rstudio-desktop/>

²<https://jasp-stats.org>

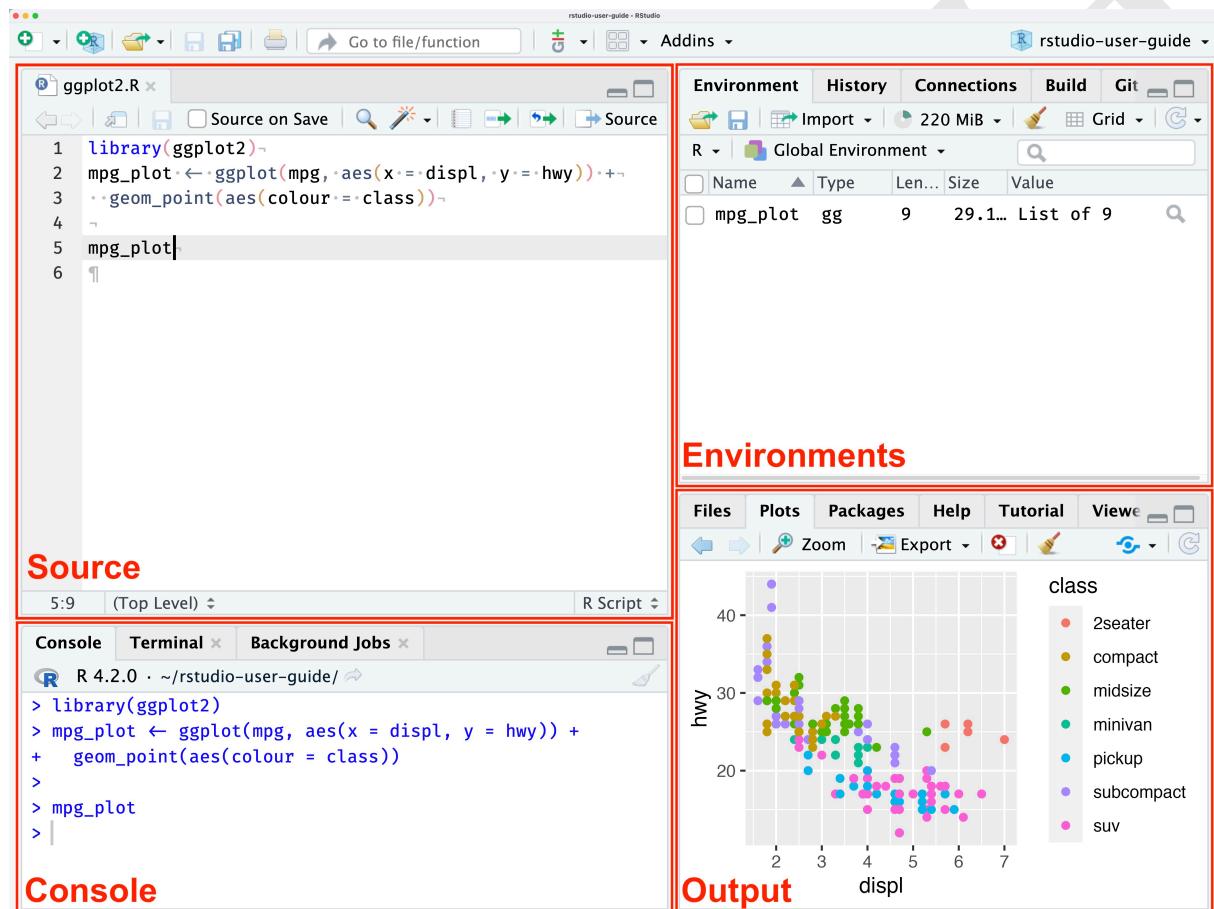


Figura 4.1: Interface do RStudio. Fonte: <https://docs.posit.co/ide/user/>

- jamovi³.²⁹
- BlueSky⁴.



Os pacotes *jmv*³⁰ e *jmvconnect*³¹ fornecem funções para análise descritiva e inferencial com interface com jamovi.

4.2 Scripts computacionais

4.2.1 O que são R scripts?

- “Scripts são dados”.³²
- Scripts permitem ao usuário se concentrar nas tarefas mais importantes da computação e utilizar pacotes ou bibliotecas para executar as funções mais básicas com maior eficiência.³²
- Um script é um arquivo de texto contendo (quase) os mesmos comandos que você digitaria na linha de comando do R. O “quase” refere-se ao fato de que se você estiver usando *sink()* para enviar a saída para um arquivo, você terá que incluir alguns comandos em *print()* para obter a mesma saída da linha de comando.

4.3 Pacotes

4.3.1 O que são pacotes?

- Pacotes são conjuntos de scripts programados pela comunidade e compartilhados para uso público.²⁴
- Os pacotes ficam armazenados no *Comprehensive R Archive Network* (CRAN) e podem ser instalados diretamente no RStudio.^{24,25}
- Na mais recente atualização deste livro, o [Comprehensive R Archive Network (CRAN) possui 381395 pacotes disponíveis.^{24,25}
- Os pacotes disponíveis podem ser encontrados em *R PACKAGES DOCUMENTATION*.³³



O pacote *utils*³⁴ fornece a função *install.packages*^a para instalar os pacotes no computador.

^a<https://www.rdocumentation.org/packages/utils/versions/3.6.2/topics/install.packages>



O pacote *utils*³⁴ fornece a função *library*^a para carregar os pacotes instalados no computador.

^a<https://www.rdocumentation.org/packages/utils/versions/3.6.2/topics/library>



O pacote *utils*³⁴ fornece a função *require*^a para indicar se o pacote requisitado está disponível.

^a<https://www.rdocumentation.org/packages/utils/versions/3.6.2/topics/require>



O pacote *utils*³⁴ fornece a função *installed.packages*^a para listar os pacotes instalados no computador.

^a<https://www.rdocumentation.org/packages/utils/versions/3.6.2/topics/installed.packages>

³<https://www.jamovi.org>

⁴<https://www.blueskystatistics.com>

 O pacote *utils*³⁴ fornece a função *update.packages*^a para atualizar os pacotes instalados no computador.

^a<https://www.rdocumentation.org/packages/utils/versions/3.6.2/topics/update.packages>

4.3.2 Quais práticas são recomendadas na redação de scripts?

- Use nomes consistentes para as variáveis.³⁵
- Defina os tipos de variáveis adequadamente no banco de dados.³⁵
- Defina constantes - isto é, variáveis de valor fixo - ao invés de digitar valores.³⁵
- Use e cite os pacotes disponíveis para suas análises.³⁵
- Controle as versões do script.^{35,36}
- Teste o script antes de sua utilização.³⁵
- Conduza revisão por pares do código durante a redação (digitação em dupla).³⁵

 O pacote *formatR*³⁷ fornece a função *tidy_source*^a para formatar um R script.

^ahttps://www.rdocumentation.org/packages/formatR/versions/1.14/topics/tidy_source

 O pacote *styler*³⁸ fornece a função *style_file*^a para formatar um R script.

^ahttps://www.rdocumentation.org/packages/styler/versions/1.10.1/topics/style_file

 O pacote *lintr*³⁹ fornece a função *lint*^a para verificar a adesão de um script a um determinado estilo, identificando erros de sintaxe e possíveis problemas semânticos.

^a<https://www.rdocumentation.org/packages/lintr/versions/3.1.0/topics/lint>

4.4 Aplicativos Shiny

4.4.1 O que são Shiny Apps?

- Shiny Apps são aplicativos web interativos que permitem a criação de interfaces gráficas para visualização e análise de dados em tempo real, utilizando o R como backend.²

4.5 Manuscritos reproduzíveis

4.5.1 O que são manuscritos reproduzíveis?

- Manuscritos reproduzíveis - manuscritos executáveis ou relatórios dinâmicos - permitem a produção de um manuscrito completo a partir da integração do banco de dados da(s) amostra(s), do(s) script(s) de análise estatística (incluindo comentários para sua interpretação), dos pacotes ou bibliotecas utilizados, das fontes e referências bibliográficas citadas, além dos demais elementos textuais (tabelas, gráficos) - todos gerados dinamicamente.³²

4.5.2 Por que usar manuscritos reproduzíveis?

- No processo tradicional de redação científica há muitas etapas de copiar e colar não reproduzíveis envolvidas. Documentos dinâmicos combinam uma ferramenta de processamento de texto com o R script que produz o texto/tabela/figura a ser incorporado no manuscrito.²²

- Ao trabalhar com relatórios dinâmicos, é possível extrair o mesmo script usado para análise estatística. Os documentos podem ser compilados em vários formatos de saída e salvos como DOCX, PPTX e PDF.²²
- Muitos erros de análise poderiam ser evitados com a adoção de boas práticas de programação em manuscritos reproduzíveis.⁴⁰

R

O pacote *rmarkdown*⁴¹ fornece as funções *render*^a para criar manuscritos reproduzíveis a partir de arquivos .Rmd.

^a<https://www.rdocumentation.org/packages/rmarkdown/versions/2.24/topics/render>

R

O pacote *officedown*⁴² fornece as funções *rdocx_document*^a e *rpptx_document*^b para criar arquivos DOCX e PPTX, respectivamente, com o conteúdo criado no manuscrito reproduzível.

^ahttps://www.rdocumentation.org/packages/officedown/versions/0.3.0/topics/rdocx_document

^bhttps://www.rdocumentation.org/packages/officedown/versions/0.3.0/topics/rpptx_document

R

O pacote *bookdown*⁴³ fornece as funções *gitbook*^a, *pdf_book*^b, *epub_book*^c e *html_document2*^d para criar documentos reproduzíveis em diversos formatos (Git, PDF, EPUB e HTML, respectivamente).

^a<https://www.rdocumentation.org/packages/bookdown/versions/0.35/topics/gitbook>

^bhttps://www.rdocumentation.org/packages/bookdown/versions/0.35/topics/pdf_book

^chttps://www.rdocumentation.org/packages/bookdown/versions/0.35/topics/epub_book

^dhttps://www.rdocumentation.org/packages/bookdown/versions/0.35/topics/html_document2

4.5.3 O que é RMarkdown?

- RMarkdown⁴¹ é uma ferramenta que permite a integração de texto, código e saída em um único documento. O RMarkdown é uma extensão do Markdown, que é uma linguagem de marcação simples e fácil de aprender, que é usada para formatar texto. O RMarkdown permite a inclusão de blocos de código R, Python, SQL, C++, entre outros, e a saída desses blocos de código é incorporada ao documento final. O RMarkdown é uma ferramenta poderosa para a criação de relatórios dinâmicos, que podem ser facilmente atualizados com novos dados ou análises. O RMarkdown é amplamente utilizado na comunidade científica para a criação de relatórios de pesquisa, artigos científicos, apresentações, livros, entre outros.
- O trabalho com RMarkdown⁴¹ permite um fluxo de dados totalmente transparente, desde o conjunto de dados coletados até o manuscrito finalizado. Todos os aspectos do fluxo de dados podem ser incorporados em blocos de R script (*chunk*), exibindo tanto o R script quanto o respectivo texto, tabelas e figuras formatadas no estilo científico de interesse.⁴⁴
- O RMarkdown⁴¹ foi projetado especificamente para relatórios dinâmicos onde a análise é realizada em R e oferece uma flexibilidade incrível por meio de uma linguagem de marcação.²²

4.5.4 Como manuscritos reproduzíveis contribuem para a ciência?

- O compartilhamento de bancos de dados e seus scripts de análise estatística permitem a adoção de práticas reproduzíveis, tais como a reanálise dos dados.⁴⁵

R

O pacote *projects*⁴⁶ fornece a função *setup_projects*^a para criar um projeto com arquivos organizados em diretórios.

^ahttps://www.rdocumentation.org/packages/projects/versions/2.1.3/topics/setup_projects

4.5.5 Como contribuir para a reproduzibilidade?

- Disponibilize publicamente os bancos de dados, respeitando as considerações éticas vigentes (ex.: autorização dos participantes e do Comitê de Ética em Pesquisa) e internacionalmente.²²

- Produza manuscritos reprodutíveis - manuscritos executáveis ou relatórios dinâmicos - que permitem a integração do banco de dados da(s) amostra(s), do(s) script(s) de análise estatística (incluindo comentários para sua interpretação), dos pacotes ou bibliotecas utilizados, das fontes e referências bibliográficas citadas, além dos demais elementos textuais (tabelas, gráficos) - todos gerados dinamicamente.³²

R O pacote *rmarkdown*⁴¹ fornece a função *render*^a para criar manuscritos reprodutíveis a partir de arquivos .Rmd.

^a<https://www.rdocumentation.org/packages/rmarkdown/versions/2.24/topics/render>

R O pacote *bookdown*⁴³ fornece as funções *gitbook*^a, *pdf_book*^b, *epub_book*^c e *html_document2*^d para criar documentos reprodutíveis em diversos formatos (Git, PDF, EPUB e HTML, respectivamente).

^a<https://www.rdocumentation.org/packages/bookdown/versions/0.35/topics/gitbook>

^bhttps://www.rdocumentation.org/packages/bookdown/versions/0.35/topics/pdf_book

^chttps://www.rdocumentation.org/packages/bookdown/versions/0.35/topics/epub_book

^dhttps://www.rdocumentation.org/packages/bookdown/versions/0.35/topics/html_document2

4.6 Compartilhamento

4.6.1 Por que compartilhar scripts?

- Compartilhar o script — principalmente junto aos dados — pode facilitar a replicação direta do estudo, a detecção de eventuais erros de análise, a detecção de pesquisas fraudulentas.⁴⁷

4.6.2 O que pode ser compartilhado?

- Idealmente, todos os scripts, pacotes/bibliotecas e dados necessários para outros reproduzirem seus dados.³⁶
- Minimamente, partes importantes incluindo implementações de novos algoritmos e dados que permitam reproduzir um resultado importante.³⁶

4.6.3 Como preparar dados para compartilhamento?

- ?

4.6.4 Como preparar scripts para compartilhamento?

- Providencie a documentação sobre seu script (ex.: arquivo README).³⁶
- Inclua a versão dos pacotes usados no seu script por meio de um script inicial para instalação de pacotes (ex.: ‘instalar.R’).⁴⁰
- Documente em um arquivo README os arquivos disponíveis e os pré-requisitos necessários para executar o código (ex.: pacotes e respectivas versões). Uma lista de configurações (hardware e software) que foram usadas para rodar o código pode ajudar na reprodução dos resultados.²¹
- Use endereços de arquivos relativos.⁴⁰
- Crie links persistentes para versões do seu script.³⁶
- Defina uma semente para o gerador de números aleatórios em scripts com métodos computacionais que dependem da geração de números pseudoaleatórios.²¹

R O pacote *base*⁴⁸ fornece a função *set.seed*^a para especificar uma semente para reprodutibilidade de computações que envolvem números aleatórios.

^a<https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/Random>

- Escolha uma licença apropriada para garantir os direitos de criação e como outros poderão usar seus scripts.³⁶
- Teste o script em uma nova sessão antes de compartilhar.⁴⁰
- Cite todos os pacotes relacionados à sua análise.⁴⁹

R

O pacote *utils*³⁴ fornece a função *citation*^a para citar o programa R e os pacotes da sessão atual.

^a<https://www.rdocumentation.org/packages/utils/versions/3.6.2/topics/citation>

R

O pacote *grateful*⁵⁰ fornece a função *cite_packages*^a para citar os pacotes utilizados em um projeto R.

^ahttps://www.rdocumentation.org/packages/grateful/versions/0.2.0/topics/cite_packages

R

- Inclua a informação da sessão em que os scripts foram rodados.⁴⁰

^a<https://www.rdocumentation.org/packages/utils/versions/3.6.2/topics/sessionInfo>

4.6.5 O que incluir no arquivo README?

- Título do trabalho.²¹
- Autores do trabalho.²¹
- Principais responsáveis pela escrita do script e quaisquer outras pessoas que fizeram contribuições substanciais para o desenvolvimento do script.²¹
- Endereço de e-mail do autor ou contribuidor a quem devem ser direcionadas dúvidas, comentários, sugestões e bugs sobre o script.²¹
- Lista de configurações nas quais o script foi testado, tais com nome e versão do programa, pacotes e plataforma.²¹

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

RAASCUNHO

Capítulo 5

Paradoxos e falácia

5.1 Paradoxos estatísticos

5.1.1 O que são paradoxos estatísticos?

- Paradoxos podem originar da incompreensão ou mal informação da nossa intuição a respeito do fenômeno.⁵¹

5.1.2 O que é o paradoxo de Abelson?

- ⁵²

5.1.3 O que é o paradoxo de Berkson?

- ⁵³

5.1.4 O que é o paradoxo de *Big Data*?

- “Quanto maior a quantidade de dados, maior a certeza de que vamos nos enganar”.⁵¹

5.1.5 O que é o paradoxo de Ellsberg?

- ⁵⁴

5.1.6 O que é o paradoxo de Freedman?

- ^{55,56}

5.1.7 O que é o paradoxo de Hand?

- ⁵⁷

5.1.8 O que é o paradoxo de Lindley?

- ⁵⁸

5.1.9 O que é o paradoxo de Lord?

- ^{59,60}

5.1.10 O que é o paradoxo de Proebsting?

- [?]

5.1.11 O que é o paradoxo de Simpson?

- O paradoxo de Simpson ocorre quando a associação entre duas variáveis X e Y desaparece ou mesmo reverte sua direção quando condicionadas em uma terceira variável Z .^{61,62}
- Para decisão do paradoxo de Simpson pode-se utilizar o conceito de ‘back-door’, o qual considera os ‘caminhos’ (isto é, associações) no gráfico acíclico direcionado e assegura que todos as associações espúrias do tratamento X para o desfecho Y nesse diagrama causal sejam interceptados pela variável Z .⁶³
- Dependendo do contexto em que os dados foram obtidos — delineamento do estudo, escolha dos instrumentos e dos tipos de variáveis — a melhor escolha para a análise pode variar entre a análise da população agregada ou da subpopulação desagregada.⁶³
- É possível que em alguns contextos nem a análise agregada ou a desagregada podem oferecer a resposta correta, sendo necessário o uso de outras (mais) covariáveis.⁶³

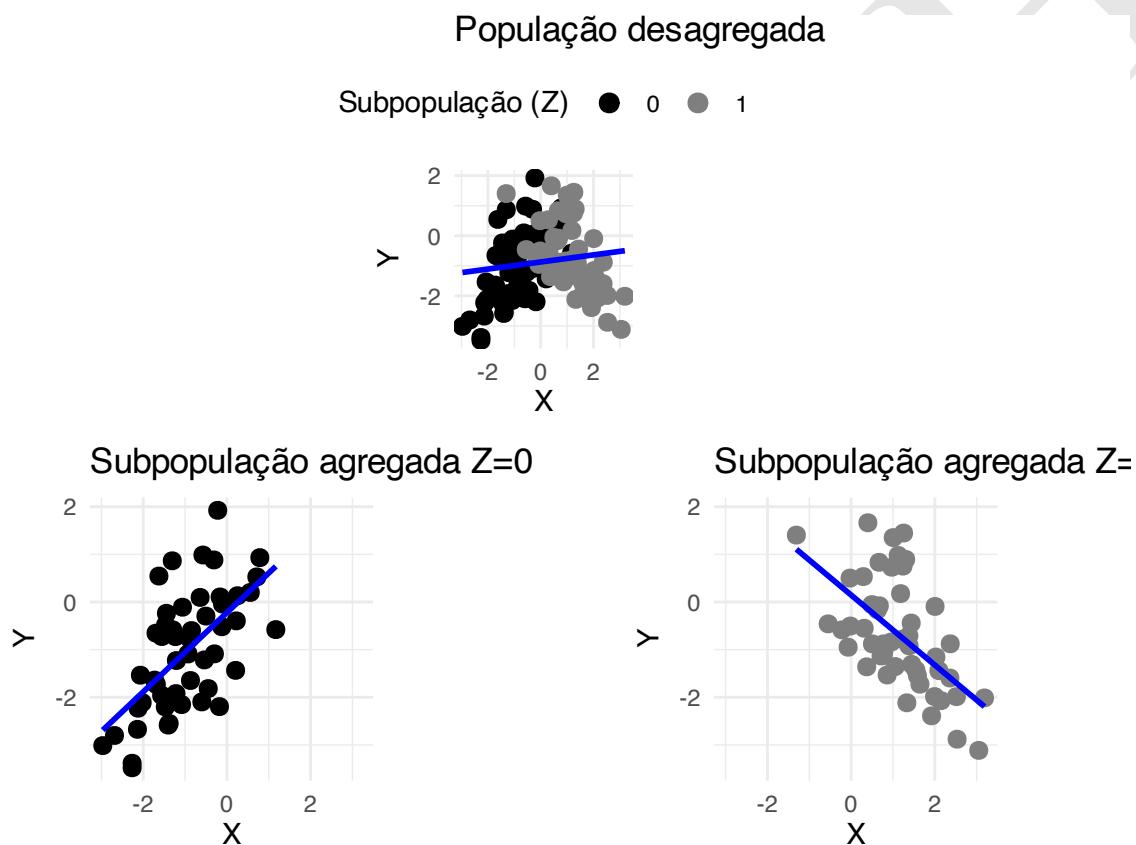


Figura 5.1: Paradoxo de Simpson representado com dados simulados. Os pontos no gráfico representam observações individuais e as linhas de tendência representam as regressões lineares ajustadas para os dados desagregados da população e agregados por subpopulação.

5.1.12 O que é o paradoxo de Stein?

- ⁶⁴

5.1.13 O que é o paradoxo de Okie?

- ?

5.1.14 O que é o paradoxo da acurácia?

- ?

5.1.15 O que é o paradoxo do falso positivo?

- [?]

5.1.16 O que é o paradoxo da caixa de Bertrand?

- [?]

5.1.17 O que é o paradoxo do elevador?

- ⁶⁵

5.1.18 O que é o paradoxo da amizade?

- ⁶⁶

5.1.19 O que é o paradoxo do menino ou menina?

- ⁶⁵

5.1.20 O que é o paradoxo do teste surpresa?

- [?]

5.1.21 O que é o paradoxo do nó da gravata?

- [?]

5.1.22 O que é o paradoxo da Bela Adormecida?

- [?]

5.2 Falácia estatísticas

5.2.1 O que são falácia estatísticas?

- Falácia estatísticas são erros de raciocínio que ocorrem em situações que envolvem dados e estatísticas. Elas podem ocorrer em qualquer etapa do processo de análise de dados, desde a coleta até a interpretação dos resultados. Elas podem ser intencionais ou não intencionais, e podem ser usadas para manipular, enganar ou confundir as pessoas.⁶⁸
- As falácia estatísticas podem ser difíceis de detectar, pois muitas vezes são sutis e podem parecer plausíveis à primeira vista. No entanto, é importante estar ciente delas e saber como identificá-las para evitar erros de interpretação e tomada de decisão.⁶⁹

5.2.2 O que é a falácia do jogador?

- A falácia do jogador é a crença de que eventos independentes têm uma influência sobre eventos futuros. Por exemplo, se uma moeda é lançada várias vezes e cai cara em todas as vezes, a falácia do jogador sugere que a próxima jogada será coroa, pois a moeda “deve” se equilibrar. No entanto, cada lançamento da moeda é independente e não afeta o resultado do próximo lançamento.⁷⁰

5.2.3 O que é a falácia da mão quente?

- A falácia da mão quente é a crença de que um jogador que teve sucesso em um jogo de azar terá mais chances de sucesso no futuro. Por exemplo, se uma moeda é lançada várias vezes e cai cara em todas as vezes, a falácia da mão quente sugere que a próxima jogada será cara, pois o jogador está “quente”. No entanto, cada lançamento da moeda é independente e não afeta o resultado do próximo lançamento.⁷¹

Citar como: Ferreira, Arthur de Sá. Ciéncia com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

RAASCUNHO

Capítulo 6

Vieses metodológicos

6.1 Vieses metodológicos

6.1.1 O que são vieses metodológicos?

- ?

6.2 Tipos de vieses metodológicos

6.2.1 Quais são os tipos de vieses metodológicos?

6.3 Efeitos relacionados aos vieses metodológicos

6.3.1 Quais são os efeitos relacionados aos vieses metodológicos?

6.3.2 O que é efeito placebo?

- ?

6.3.3 O que é efeito nocebo?

- ?

6.3.4 O que é efeito Hawthorne?

- ?

6.3.5 O que é efeito Rosenthal?

- ?

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

RAASCUNHO

PARTE 2: DADOS – COLETA E PREPARAÇÃO

Organização e fundamentos de dados

RASCUNHO

Capítulo 7

Medidas e instrumentos

7.1 Escalas

7.1.1 O que são escalas?

- Uma escala de medição grosseira representa um construto de natureza contínua medido por itens tais que diferentes pontuações são agrupadas na mesma categoria no ato da coleta de dados.⁶⁸
- Em escalas grosseiras, erros são introduzidos porque as variações contínuas do constructo são colapsadas em uma mesma categorias ou separadas entre categorias próximas.⁶⁸
- Escalas tipo Likert com 5 categorias tipo “discordo totalmente”, “discordo parcialmente”, “nem concordo nem discordo”, “concordo parcialmente”, e “concordo totalmente” são escalas grosseira porque as diferenças entre as categorias não são iguais. Por exemplo, a diferença entre “discordo totalmente” e “discordo parcialmente” não é a mesma que a diferença entre “concordo parcialmente” e “concordo totalmente”.⁶⁸

R

O pacote *likert*⁶⁹ fornece a função *likert*^a para analisar respostas de instrumentos em escala Likert.

^a<https://www.rdocumentation.org/packages/likert/versions/1.3.5/topics/likert>

- O erros em escalas grosseiras é considerado sistemático mas não pode ser corrigido em nível da unidade de análise.⁶⁸

7.2 Medição e Medidas

7.2.1 O que é medição?

- Processo empírico, realizado por meio de um instrumento, que estabelece uma correspondência rigorosa e objetiva entre uma observação e uma categoria em um modelo da observação.⁷⁰
- Esse processo tem como objetivo distinguir de maneira substantiva a manifestação observada de outras possíveis manifestações que também possam ser diferenciadas.⁷⁰

7.2.2 O que são medidas diretas?

- ?

7.2.3 O que são medidas derivadas?

- ?

7.2.4 O que são medidas por teoria?

- ?

Tabela 7.1: Tabela de dados brutos com medidas únicas.

Unidade de análise	Pressão arterial, braço esquerdo (mmHg)
1	118
2	113
3	116
4	110
5	111
6	116
7	120
8	111
9	120
10	112

Tabela 7.2: Tabela de dados brutos com medidas repetidas.

Unidade de análise	Pressão arterial, braço esquerdo (mmHg) #1	Pressão arterial, braço esquerdo (mmHg) #2	Pressão arterial, braço esquerdo (mmHg) #3
1	114	112	112
2	115	120	113
3	115	110	120
4	117	116	114
5	110	118	116
6	110	120	113
7	118	114	117
8	111	112	119
9	120	112	117
10	110	115	115

7.2.5 O que são medidas únicas?

- A medida única da pressão arterial sistólica no braço esquerdo resulta em um valor pontual.⁷
- Medidas únicas obtidas de diferentes unidades de análise podem ser consideradas independentes se observadas outras condições na coleta de dados.⁷
- O valor pontual será considerado representativo da variável para a unidade de análise (ex.: **120 mmHg** para o participante #9).

7.2.6 O que são medidas repetidas?

- As medidas repetidas podem ser tabuladas separadamente, por exemplo para análise da confiabilidade de obtenção dessa medida.⁷
- A medida repetida da pressão arterial no braço esquerdo resulta em um conjunto de valores pontuais (ex.: **110 mmHg**, **118 mmHg** e **116 mmHg** para o participante #5).
- As medidas repetidas podem ser agregadas por algum parâmetro — ex.: média, mediana, máximo, mínimo, entre outros —, observando-se a relevância biológica, clínica e/ou metodológica desta escolha.⁷
- Medidas agregadas obtidas de diferentes unidades de análise podem ser consideradas independentes se observadas outras condições na coleta de dados.⁷
- O valor agregado será considerado representativo da variável para a unidade de análise (ex.: média = **115 mmHg** para o participante #5).

Tabela 7.3: Tabela de dados brutos com medidas repetidas agregadas.

Unidade de análise	Pressão arterial, braço esquerdo (mmHg) média
1	113
2	116
3	115
4	116
5	115
6	114
7	116
8	114
9	116
10	113

Tabela 7.4: Tabela de dados brutos com medidas seriadas não agregadas.

Unidade de análise	Tempo (min)	Pressão arterial, braço esquerdo (mmHg)
1	1	114
1	2	120
1	3	110
2	1	119
2	2	120
2	3	114
3	1	116
3	2	114
3	3	116
4	1	113

R

O pacote *stats*⁷¹ fornece a função *aggregate*^a para agrregar medidas repetidas utilizando uma função personalizada.

^a<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/aggregate>

7.2.7 O que são medidas seriadas?

- Medidas seriadas são possivelmente relacionadas e, portanto, dependentes na mesma unidade de análise.⁷
- Por exemplo, a medida seriada da pressão arterial no braço esquerdo, em intervalos tipicamente regulares (ex.: **114 mmHg**, **120 mmHg** e **110 mmHg** em **1 min**, **2 min** e **3 min**, respectivamente, para o participante #1).
- Medidas seriadas também agregadas por parâmetros — ex.: máximo, mínimo, amplitude — são consideradas representativas da variação temporal ou de uma característica de interesse (ex.: amplitude = **10 mmHg** para o participante #1).

R

O pacote *stats*⁷¹ fornece a função *aggregate*^a para agrregar medidas repetidas utilizando uma função personalizada.

^a<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/aggregate>

7.2.8 O que são medidas múltiplas?

- Medidas múltiplas também são possivelmente relacionadas e, portanto, são dependentes na mesma unidade de análise. Medidas múltiplas podem ser obtidas de modo repetido para análise agregada ou seriada.⁷

Tabela 7.5: Tabela de dados brutos com medidas seriadas não agregadas.

Unidade de análise	Pressão arterial, braço esquerdo (mmHg) amplitude
1	10
2	6
3	2
4	6
5	1
6	8
7	9
8	10
9	7
10	5

Tabela 7.6: Tabela de dados brutos com medidas múltiplas.

Unidade de análise	Pressão arterial, braço esquerdo (mmHg)	Pressão arterial, braço direito (mmHg)
1	117	115
2	120	118
3	112	118
4	112	112
5	116	112
6	112	118
7	115	113
8	114	118
9	119	114
10	112	116

- A medida de pressão arterial bilateral resulta em um conjunto de valores pontuais (ex.: braço esquerdo = **114 mmHg**, braço direito = **118 mmHg** para o participante #8). Neste caso, ambos os valores pontuais são considerados representativos daquela unidade de análise.

 O pacote *stats*⁷¹ fornece a função *aggregate*^a para agrregar medidas repetidas utilizando uma função personalizada.

^a<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/aggregate>

7.3 Erros de medida

7.3.1 O que são erros de medida?

- ?
- A natureza dos erros de medida são em geral atribuídos aos (1) instrumentos utilizados e variações no protocolo, na medida em que o seu tamanho médio pode ser reduzido por modificações e melhorias nesses instrumentos; e (2) variações genuínas medida em de curto prazo.⁷²

7.3.2 Quais fontes de variabilidade são comumente investigadas?

- Intra/Entre participantes (isto é, unidades de análise).⁷³
- Intra/Entre repetições.⁷³
- Intra/Entre observadores.⁷³

7.4 Instrumentos

7.4.1 O que são instrumentos?

- ?

7.5 Acurácia e precisão

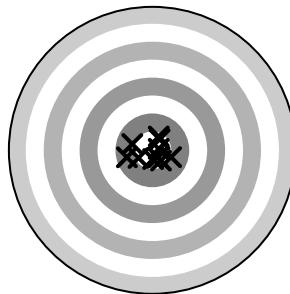
7.5.1 O que é acurácia?

- ?

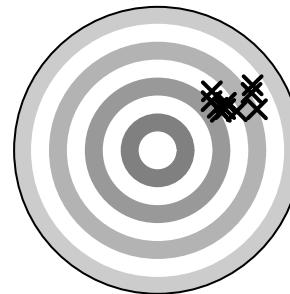
7.5.2 O que é precisão?

- ?

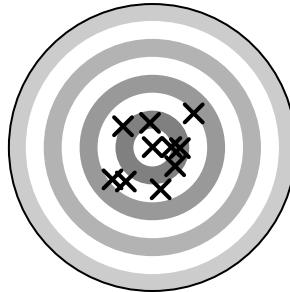
Acurácia alta, Precisão alta



Acurácia baixa, Precisão alta



Acurácia alta, Precisão baixa



Acurácia baixa, Precisão baixa

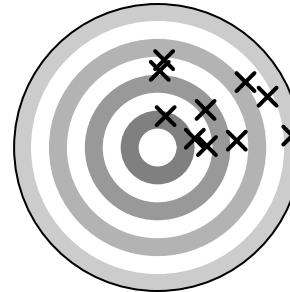


Figura 7.1: Acurácia e precisão como propriedades de uma medida.

7.6 Viés e variabilidade

7.6.1 Qual é a relação entre viés e variabilidade?

- ?

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

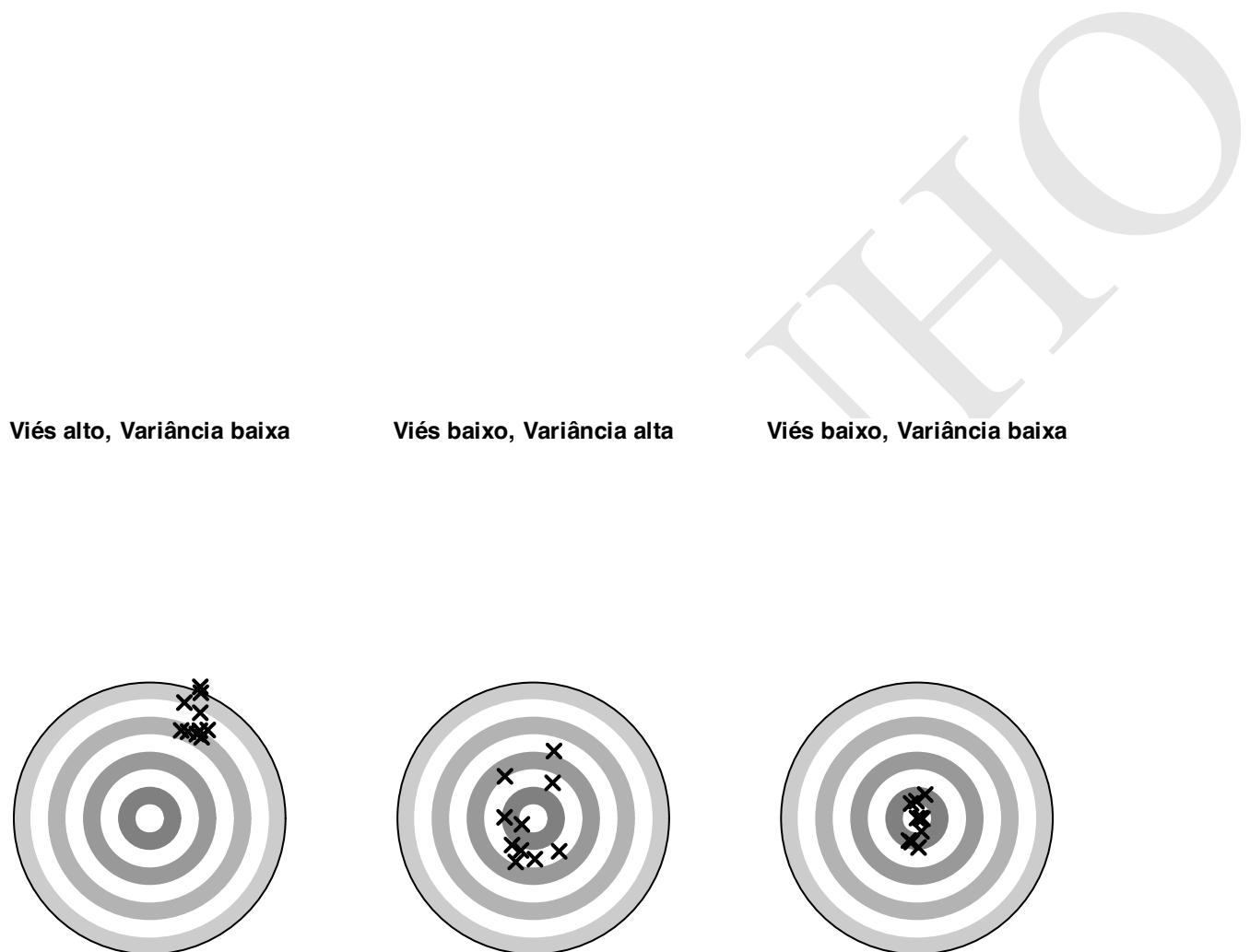


Figura 7.2: Viés e variabilidade de uma medida.

Capítulo 8

Dados, *big data* e metadados

8.1 Dados

8.1.1 O que são dados?

- “Tudo são dados”.⁷⁴
- Dados coletados em um estudo geralmente contêm erros de mensuração e/ou classificação, dados perdidos e são agrupados por alguma unidade de análise.⁷⁵

8.1.2 Quais são as fontes de dados?

- Experimentos.⁷⁶
- Mundo real.⁷⁷
- Simulação.⁷⁸

8.1.3 O que são dados primários e secundários?

- Dados primários são dados originais coletados intencionalmente para uma determinada análise exploratória ou inferencial planejada a priori.⁷⁹
- Dados secundários compreendem dados coletados inicialmente para análises de um estudo, e são subsequentemente utilizados para outras análises.⁸⁰

8.1.4 O que são dados quantitativos e qualitativos?

- ?

8.2 *Big data*

8.2.1 O que são *big data*?

- *Big data* refere-se a bancos de dados muito grandes com um mecanismo “R” — aleatório (*Random*), auto-reportado (*self-Reported*), reportado administrativamente (*administratively reported*), seletivamente respondido (*selectively respondend*) — descontrolado ou desconhecido.⁸¹

8.3 Metadados

8.3.1 O que são metadados?

- Metadados são informações técnicas relacionadas às variáveis do estudo, tais como rótulos, limites de valores plausíveis, códigos para dados perdidos e unidades de medida.⁷⁷

- Metadados também são informações relacionadas ao delineamento e/ou protocolo do estudo, recrutamento dos participantes, e métodos para realização das medidas.⁷⁷

8.3.2 Quais são as recomendações para os metadados de um banco de dados?

- Utilize rótulos padronizados para variáveis e fatores para facilitar o reuso (reprodutibilidade) do conjuntos de dados e scripts de análise.⁷⁸
- Crie rótulos de variáveis concisos, claros e mutuamente exclusivos.⁷⁸
- Evite muitas letras maiúsculas ou outros caracteres especiais que usam a *shift*.⁷⁸
- Na existência de versões de instrumentos publicadas em diferentes anos, use o ano de publicação das escalas no rótulo.⁷⁸
- Divida o rótulo da variável ou fator em partes e ordene-as do mais geral para o mais particular geral (ex.: experimento -> repetição -> escala -> item).⁷⁸

 O pacote *base*⁴⁸ fornece a função *names*^a para declarar o nome de uma variável.

^a<https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/names>

 O pacote *base*⁴⁸ fornece a função *labels*^a para declarar o rótulo de uma variável.

^a<https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/labels>

 O pacote *units*⁷⁹ fornece a função *units*^a para declarar as unidades de medida de uma variável.

^a<https://www.rdocumentation.org/packages/units/versions/0.8-3/topics/units>

 O pacote *units*⁷⁹ fornece a função *valid_udunits*^a para listar as opções de unidades de medida de uma variável.

^ahttps://www.rdocumentation.org/packages/units/versions/0.8-3/topics/valid_udunits

 O pacote *janitor*⁸⁰ fornece a função *clean_names*^a para formatar de modo padronizado o nome das variáveis utilizando apenas caracteres, números e o símbolo ‘_’.

^ahttps://www.rdocumentation.org/packages/janitor/versions/2.2.0/topics/clean_names

 O pacote *Hmisc*⁸¹ fornece a função *contents*^a para criar um objeto com os metadados (nomes, rótulos, unidades, quantidade e níveis das variáveis categóricas, e quantidade de dados perdidos) de um dataframe.

^a<https://www.rdocumentation.org/packages/Hmisc/versions/5.1-0/topics/contents>

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

Capítulo 9

Tabulação de dados

9.1 Planilhas eletrônicas

9.1.1 Qual a organização de uma tabela de dados?

- As informações podem ser organizadas em formato de dados retangulares (ex.: matrizes, tabelas, quadro de dados) ou não retangulares (ex.: listas).⁸²
- Cada variável possui sua própria coluna (vertical).⁸²
- Cada observação possui sua própria linha (horizontal).⁸²
- Cada valor possui sua própria célula especificada em um par (linha, coluna).⁸²
- Cada célula possui seu próprio dado.⁸²

R O pacote *DataEditR*⁸³ fornece a função *data_edit*^a para interativamente criar, editar e salvar a tabela de dados.

^a<https://www.rdocumentation.org/packages/DataEditR/versions/0.1.5/topics/dataInput>

9.1.2 Qual a estrutura básica de uma tabela para análise estatística?

- Use apenas 1 (uma) planilha eletrônica para conter todas as informações coletadas. Evite múltiplas abas no mesmo arquivo, assim como múltiplos arquivos quando possível.⁸⁴
- Use apenas 1 (uma) linha de cabeçalho para nomear os fatores e variáveis do seu estudo.⁸⁴
- Tipicamente, cada linha representa um participante e cada coluna representa uma variável ou fator do estudo. Estudos com medidas repetidas dos participantes podem conter múltiplas linhas para o mesmo participante (repetindo os dados na mesma coluna, conhecido como *formato curto*) ou só uma linha para o participante (repetindo os dados em colunas separadas, conhecido como *formato longo*).⁸⁵

Tabela 9.1: Estrutura básica de uma tabela de dados.

V1	V2	V3	V4
$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$
$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$
$x_{3,1}$	$x_{3,2}$	$x_{3,3}$	$x_{3,4}$
$x_{4,1}$	$x_{4,2}$	$x_{4,3}$	$x_{4,4}$
$x_{5,1}$	$x_{5,2}$	$x_{5,3}$	$x_{5,4}$

Tabela 9.2: Formatação recomendada para tabela de dados.

ID	Data.Coleta	Estado.Civil	Numero.Filhos
1	21-07-2025	casado	NA
2	22-07-2025	casado	1
3	23-07-2025	casado	NA
4	24-07-2025	solteiro	NA
5	25-07-2025	casado	NA
6	26-07-2025	solteiro	0
7	27-07-2025	solteiro	NA
8	28-07-2025	solteiro	NA
9	29-07-2025	casado	NA
10	30-07-2025	solteiro	NA

Tabela 9.3: Formatação não recomendada para tabela de dados.

ID	Data de Coleta	Estado Civil	Número de Filhos
1	21-07-2025	casado	NA
2	22-07-2025	Casado	1
3	23-07-2025	casado	NaN
4	24-07-2025	Solteiro	N/A
5	25-07-2025	Casado	N.A.
6	26-07-2025	solteiro	0
7	27-07-2025	solteiro	
8	28-07-2025	Solteiro	na
9	29-07-2025	casado	n.a.
10	30-07-2025	Solteiro	999

9.1.3 O que usar para organizar tabelas para análise computadorizada?

- Seja consistente em: códigos para as variáveis categóricas; códigos para dados perdidos; nomes das variáveis; identificadores de participantes; nome dos arquivos; formato de datas; uso de caracteres de espaço.^{84,85}
- Crie um dicionário de dados (metadados) em um arquivo separado contendo: nome da variável, descrição da variável, unidades de medida e valores extremos possíveis.⁸⁴
- Use recursos para validação de dados antes e durante a digitação de dados.^{84,85}

 O pacote *data.table*⁸⁶ fornece a função *melt.data.table*^a para reorganizar a tabela em diferentes formatos.

^a<https://www.rdocumentation.org/packages/data.table/versions/1.14.8/topics/melt.data.table>

9.1.4 O que não usar para organizar tabelas para análise computadorizada?

- Não deixe células em branco: substitua dados perdidos por um código sistemático (ex.: NA [*not available*]).⁸⁴
- Não inclua análises estatísticas ou gráficos nas tabelas de dados brutos.⁸⁴
- Não utilize cores como informação. Se necessário, crie colunas adicionais - variáveis instrumentais ou auxiliares - para identificar a informação de modo que possa ser analisada.⁸⁴
- Não use células mescladas.
- Delete linhas e/ou colunas totalmente em branco (sem unidades de análise e/ou sem variáveis).

9.1.5 O que é recomendado e o que deve ser evitado na organização das tabelas para análise?

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

RAASCUNHO

Capítulo 10

Variáveis e fatores

10.1 Variáveis

10.1.1 O que são variáveis?

- Variáveis são informações que podem variar entre medidas em diferentes indivíduos e/ou repetições.⁸⁷
- Variáveis definem características de uma amostra extraída da população, tipicamente observados por aplicação de métodos de amostragem (isto é, seleção) da população de interesse.⁷⁶

10.1.2 Como são classificadas as variáveis?

- Quanto à informação:^{76,88–90}
 - Quantitativa
 - Qualitativa
- Quanto ao conteúdo:^{76,88–91}
 - Contínua: representam ordem e magnitude entre valores.
 - * Contínua (números inteiros) vs. Discreta (números racionais).
 - * Intervalo (valor ‘0’ é arbitrário) vs. Razão (valor ‘0’ verdadeiro).
 - Categórica ordinal (numérica ou nominal): representam ordem, mas não magnitude entre valores.
 - Categórica nominal (multinomial ou dicotômica): não representam ordem ou magnitude, apenas categorias.
- Quanto à interpretação:^{76,88–90}
 - Dependente (desfecho)
 - Independente (preditora, covariável, confundidora, controle)
 - Mediadora
 - Moderadora
 - Auxiliar
 - Indicadora

 O pacote *base*⁴⁸ fornece a função *class*^a para identificar qual é o tipo do objeto.

^a<https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/class>

R O pacote *base*⁴⁸ fornece as funções *as.numeric*^a e *as.character*^b para criar objetos numéricos e categóricos, respectivamente.

^a<https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/numeric>
^b<https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/character>

R O pacote *base*⁴⁸ fornece as funções *as.Date*^a e *as.logical*^b para criar objetos em formato de data e lógicos (VERDADEIRO, FALSO), respectivamente.

^a<https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/as.Date>
^b<https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/logical>

10.1.3 Por que é importante classificar as variáveis?

- Identificar corretamente os tipos de variáveis da pesquisa é uma das etapas da escolha dos métodos estatísticos adequados para as análises e representações no texto, tabelas e gráficos.⁸⁹

10.2 Transformação de variáveis

10.2.1 O que é transformação de variáveis?

- Transformação significa aplicar uma função matemática à variável medida em sua unidade original.⁹²
- A transformação visa atender aos pressupostos dos modelos estatísticos quanto à distribuição da variável, em geral a distribuição gaussiana.^{76,92}
- A dicotomização pode ser interpretada como um caso particular de agrupamento.⁹³

10.2.2 Por que transformar variáveis?

- Muitos procedimentos estatísticos supõem que as variáveis - ou seus termos de erro, mais especificamente - são normalmente distribuídas. A violação dessa suposição pode aumentar suas chances de cometer um erro do tipo I ou II.⁹⁴
- Mesmo quando se está usando análises consideradas robustas para violações dessas suposições ou testes não paramétricos (que não assumem explicitamente termos de erro normalmente distribuídos), atender a essas questões pode melhorar os resultados das análises (por exemplo, Zimmerman, 1995).⁹⁴

10.2.3 Quais transformações podem ser aplicadas?

- Distribuições com assimetria à direita:⁹⁴
 - Raiz quadrada
 - Logaritmo natural
 - Logaritmo base 10
 - Transformação inversa
- Distribuições com assimetria à esquerda:⁹⁴
 - Reflexão e raiz quadrada
 - Reflexão e logaritmo natural
 - Reflexão e logaritmo base 10
 - Reflexão e transformação inversa
- Transformação arco-seno.⁹⁴
- Transformação de Box-Cox.⁹⁵

- Transformação de escore padrão (Z-score ou padronização).
- Escala Mínimo-Máximo (0,1).
- Normalização (normas L1, L2).
- Diferenciação.
- Categorização.
- Dicotomização.

R

O pacote *MASS*⁹⁶ fornece a função *boxcox*^a para executar a transformação de Box-Cox.⁹⁵

^a<https://www.rdocumentation.org/packages/MASS/versions/7.3-58.3/topics/boxcox>

10.3 Categorização de variáveis contínuas

10.3.1 O que é categorização de uma variável?

- ?

10.3.2 Por que não é recomendado categorizar variáveis contínuas?

- Nenhum dos argumentos usados para defender a categorização de variáveis se sustenta sob uma análise técnica rigorosa.⁹⁷
- Categorizar variáveis não é necessário para conduzir análises estatísticas. Ao invés de categorizar, priorize as variáveis contínuas.⁹⁸⁻¹⁰⁰
- Em geral, não existe uma justificativa racional (plausibilidade biológica) para assumir que as categorias artificiais subjacentes existam.⁹⁸⁻¹⁰⁰
- Caso exista um ponto de corte ou limiar verdadeiro que discrimine três ou mais grupos independentes, identificar tal ponto de corte ainda é um desafio.¹⁰¹
- Categorização de variáveis contínuas aumenta a quantidade de testes de hipótese para comparações pareadas entre os quantis, inflando, portanto, o erro tipo I.¹⁰²
- Categorização de variáveis contínuas requer uma função teórica que pressupõe a homogeneidade da variável dentro dos grupos, levando tanto a uma perda de poder como a uma estimativa imprecisa.¹⁰²
- Categorização de variáveis contínuas pode dificultar a comparação de resultados entre estudos devido aos pontos de corte baseados em dados de um banco usados para definir as categorias.¹⁰²

R

O pacote *questionr*¹⁰³ fornece a função *irec*^a para executar uma interface interativa para codificação de variáveis categóricas.

^a<https://www.rdocumentation.org/packages/questionr/versions/0.7.8/topics/irec>

10.3.3 Quais são as alternativas à categorização de variáveis contínuas?

- Análise com os dados das variáveis na escala de medida original.⁹⁷
- Análise com modelos de regressão com pesos locais (*lowess*) tais como *splines* e polinômios fracionais.⁹⁷

10.4 Dicotomização de variáveis contínuas

10.4.1 O que são variáveis dicotômicas?

- Variáveis dicotômicas (ou binárias) podem representar categorias naturais tipo “presente/ausente”, “sim/não”?

- Variáveis dicotômicas podem representar categorias fictícias, criadas a partir de variáveis multinominais, em que cada nível é convertido em uma variável dicotômica indicatoda (*dummy*).⁹³
- Dicotomização é considerado um artefato da análise de dados, uma vez que é realizada após a coleta de dados.⁶⁸
- Geralmente são representadas por “1” (presente, sucesso) e “0” (ausente, falha).⁹⁴

10.4.2 Quais argumentos são usados para defender a categorização ou dicotomização de variáveis contínuas?

- O argumento principal para dicotomização de variáveis é que tal procedimento facilita e simplifica a apresentação dos resultados, principalmente para o público em geral.⁹⁵
- Os pesquisadores não conhecem as consequências estatísticas da dicotomização.⁹⁶
- Os pesquisadores não conhecem os métodos adequados de análise não-paramétrica, não-linear e robusta.⁹⁷
- As categorias representam características existentes dos participantes da pesquisa, de modo que as análises devam ser feitas por grupos e não por indivíduos.⁹⁷
- A confiabilidade da(s) variável(eis) medida(s) é baixa e, portanto, categorizar os participantes resultaria em uma medida mais confiável.⁹⁷

10.4.3 Por que não é recomendado dicotomizar variáveis contínuas?

- Nenhum dos argumentos usados para defender a dicotomização de variáveis se sustenta sob uma análise técnica rigorosa.⁹⁷
- Dicotomizar variáveis não é necessário para conduzir análises estatísticas. Ao invés de dicotomizar, priorize as variáveis contínuas.⁹⁸⁻¹⁰⁰
- Em geral, não existe uma justificativa racional (plausibilidade biológica) para assumir que as categorias artificiais subjacentes existam.⁹⁸⁻¹⁰⁰
- Dicotomização causa perda de informação e consequentemente perda de poder estatístico para detectar efeitos.^{97,98}
- Dicotomização também classifica indivíduos com valores próximos na variável contínua como indivíduos em pontos opostos e extremos, artificialmente sugerindo que são muito diferentes.⁹⁸
- Dicotomização pode diminuir a variabilidade das variáveis.⁹⁸
- Dicotomização pode ocultar não-linearidades presentes na variável contínua.^{97,98}
- A média ou a mediana, embora amplamente utilizadas, não são bons parâmetros para dicotomizar variáveis.^{93,98}
- Caso exista um ponto de corte ou limiar verdadeiro que discrimine dois grupos independentes, identificar tal ponto de corte ainda é um desafio.¹⁰¹

10.4.4 Quais cenários legitimam a dicotomização das variáveis contínuas?

- Quando existem dados e/ou análises que suportem a existência - não apenas a suposição ou teorização - de categorias com um ponto de corte claro e com significado entre elas.⁹⁷
- Quando a distribuição da variável contínua é muito assimétrica, de modo que uma grande quantidade de observações está em um dos extremos da escala.⁹⁷

10.4.5 Quais métodos são usados para dicotomizar variáveis contínuas?

- Em termos de tabelas de contingência 2x2, os seguintes métodos permitem¹⁰¹ a identificação do limiar verdadeiro:
 - Youden.¹⁰⁴
 - Gini Index.¹⁰⁵

- Estatística qui-quadrado (χ^2).¹⁰⁶
- Risco relativo (RR).¹⁰⁷
- Kappa (κ).¹⁰⁸

10.5 Fatores

10.5.1 O que são fatores?

- Fator é um sinônimo de variável categórica.⁹
- Na modelagem, fator é sinônimo de variável preditora, em particular quando se refere à modelagem de efeitos fixos e aleatórios – os fatores (variáveis) são fatores fixos ou fatores aleatórios.⁹
- Fatores são variáveis controladas pelos pesquisadores em um experimento para determinar seu efeito na(s) variável(ies) de resposta. Um fator pode assumir apenas um pequeno número de valores, conhecidos como níveis. Os fatores podem ser uma variável categórica ou baseados em uma variável contínua, mas usam apenas um número limitado de valores escolhidos pelos experimentadores.⁹

R O pacote *base*⁴⁸ fornece a função *as.factor*^a para converter uma variável em fator.

^a<https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/factor>

10.5.2 O que são níveis de um fator?

- Níveis de um fator são as possíveis categorias que descrevem um fator.⁹

R O pacote *base*⁴⁸ fornece as funções *levels*^a e *nlevels*^b para listar os níveis e a quantidade deles em um fator.

^a<https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/levels>

^b<https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/nlevels>

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

RASCUNHO

Capítulo 11

Dados perdidos e imputados

11.1 Dados perdidos

11.1.1 O que são dados perdidos?

- Dados perdidos são dados não coletados de um ou mais participantes, para uma ou mais variáveis.¹⁰⁹

 O pacote *base*⁴⁸ fornece a função *is.na*^a para identificar que elementos de um objeto são dados perdidos.

^a<https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/na>

11.1.2 Qual o problema de um estudo ter dados perdidos?

- Uma grande quantidade de dados perdidos pode comprometer a integridade científica do estudo, considerando-se que o tamanho da amostra foi estimado para observar um determinado tamanho de efeito mínimo.¹⁰⁹
- Perda de participantes no estudo por dados perdidos pode reduzir o poder estatístico (erro tipo II).¹⁰⁹
- Não existe solução globalmente satisfatória para o problema de dados perdidos.¹⁰⁹

11.1.3 Quais os mecanismos geradores de dados perdidos?

- Dados perdidos completamente ao acaso (*missing completely at random*, MCAR), em que os dados perdidos estão distribuídos aleatoriamente nos dados da amostra.^{110,111}
- Dados perdidos ao acaso (*missing at random*, MAR), em que a probabilidade de ocorrência de dados perdidos é relacionada a outras variáveis medidas.^{110,111}
- Dados perdidos não ao acaso (*missing not at random*, MNAR), em que a probabilidade da ocorrência de dados perdidos é relacionada com a própria variável.^{110,111}

11.1.4 Como identificar o mecanismo gerador de dados perdidos em um banco de dados?

- Por definição, não é possível avaliar se os dados foram perdidos ao acaso (MAR) ou não (MNAR).¹¹⁰
- Testes t e regressões logísticas podem ser aplicados para identificar relações entre variáveis com e sem dados perdidos, criando um fator de análise ('dado perdido' = 1, 'dado observado' = 0).¹¹⁰

 O pacote *misty*¹¹² fornece a função *na.test*^a para executar o Little's Missing Completely at Random (MCAR) test¹¹³.

^a<https://www.rdocumentation.org/packages/misty/versions/0.5.0/topics/na.test>

 O pacote *naniar*¹¹⁴ fornece a função *mcar_test*^a para executar o Little's Missing Completely at Random (MCAR) test¹¹³.

^ahttps://www.rdocumentation.org/packages/naniar/versions/1.0.0/topics/mcar_test

11.1.5 Que estratégias podem ser utilizadas na coleta de dados quando há expectativa de perda amostral?

- Na expectativa de ocorrência de perda amostral, com consequente ocorrência de dados perdidos, recomenda-se ampliar o tamanho da amostra com um percentual correspondente a tal estimativa (ex.: 10%), embora ainda não corrija potenciais vieses pela perda.¹⁰⁹

11.1.6 Que estratégias podem ser utilizadas na análise quando há dados perdidos?

- Na ocorrência de dados perdidos, a análise mais comum comprehende apenas os ‘casos completos’, com exclusão de participantes com algum dado perdido nas variáveis do estudo. Em casos de grande quantidade de dados perdidos, pode-se perder muito poder estatístico (erro tipo II elevado).¹⁰⁹
- A análise de dados completos é válida quando pode-se argumentar que a probabilidade de o participante ter dados completos depende apenas das covariáveis e não dos desfechos.¹¹¹
- A análise de dados completos é eficiente quando todos os dados perdidos estão no desfecho, ou quando cada participante com dados perdidos nas covariáveis também possui dados perdidos nos desfechos.¹¹¹

 O pacote *base*⁴⁸ fornece a função *na.omit*^a para remover dados perdidos de um objeto em um banco de dados.

^a<https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/na.fail>

 O pacote *stats*⁷¹ fornece a função *complete.cases*^a para identificar os casos completos - isto é, sem dados perdidos - em um banco de dados.

^a<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/complete.cases>

11.1.7 Que estratégias podem ser utilizadas na redação de estudos em que há dados perdidos?

- Informar: o número de participantes com dados perdidos; diferenças nas taxas de dados perdidos entre os braços do estudo; os motivos dos dados perdidos; o fluxo de participantes; quaisquer diferenças entre os participantes com e sem dados perdidos; o padrão de ausência (por exemplo, se é aleatória); os métodos para tratamento de dados perdidos das variáveis em análise; os resultados de quaisquer análises de sensibilidade; as implicações dos dados perdidos na interpretação do resultados.¹¹⁵

11.2 Dados imputados

11.2.1 O que são dados imputados?

- ?

11.2.2 Quando a imputação de dados é indicada?

- A análise com imputação de dados pode ser útil quando pode-se argumentar que os dados foram perdidos ao acaso (MAR); quando o desfecho foi observado e os dados perdidos estão nas covariáveis; e variáveis auxiliares — preditoras do desfecho e não dos dados perdidos — estão disponíveis.¹¹¹
- Na ocorrência de dados perdidos, a imputação de dados (substituição por dados simulados plausíveis preditos pelos dados presentes) pode ser uma alternativa para manter o erro tipo II estipulado no plano de análise.¹⁰⁹

11.2.3 Quais os métodos de imputação de dados?

- Modelos lineares e logísticos podem ser utilizados para imputar dados perdidos em variáveis contínuas e dicotômicas, respectivamente.¹¹⁶
- Os métodos de imputação de dados mais robustos incluem a imputação multivariada por equações encadeadas (*multivariate imputation by chained equations*, MICE)¹¹⁷ e a correspondência média preditiva (*predictive mean matching*, PMM)^{118,119}.



Os pacotes *mice*¹¹⁷ e *miceadds*¹²⁰ fornecem funções *mice*^a e *mi.anova*^b para imputação multivariada por equações encadeadas, respectivamente, para imputação de dados.

^a<https://www.rdocumentation.org/packages/mice/versions/3.16.0/topics/mice>

^b<https://www.rdocumentation.org/packages/miceadds/versions/3.16-18/topics/mi.anova>

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

RAASCUNHO

Capítulo 12

Dados anonimizados e sintéticos

12.1 Dados anonimizados

12.1.1 O que são dados anonimizados?

- ?

12.1.2 Com anonimizar os dados de um banco?

- ?

R

O pacote *ids*¹²¹ fornece a função *random_id*^a para criar identificadores aleatórios por criptografia.

^ahttps://www.rdocumentation.org/packages/ids/versions/1.0.1/topics/random_id

R

O pacote *hash*¹²² fornece a função *hash*^a para criar identificadores por objetos *hash*.

^a<https://www.rdocumentation.org/packages/hash/versions/3.0.1/topics/hash>

R

O pacote *anonymizer*¹²³ fornece a função *anonymize*^a para criar uma versão anônima de variáveis em um banco de dados.

^a<https://www.rdocumentation.org/packages/anonymizer/versions/0.2.0/topics/anonymize>

R

O pacote *digest*¹²⁴ fornece a função *digest*^a para criar identificadores por objetos *hash* criptografados ou não.

^a<https://www.rdocumentation.org/packages/digest/versions/0.6.33/topics/digest>

12.2 Dados sintéticos

12.2.1 O que são dados sintéticos?

- ?

 O pacote *synthpop*¹²⁵ fornece a função *syn*^a para criar bancos de dados sintéticos a partir de um banco de dados real.

^a<https://www.rdocumentation.org/packages/synthpop/versions/1.8-0/topics/syn>

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

PARTE 3: ANÁLISE EXPLORATÓRIA E DESCRIPTIVA

Analizando padrões

RASCUNHO

Capítulo 13

Descrição

13.1 Análise de descrição

13.1.1 O que é análise de descrição de dados?

- A análise descritiva utiliza métodos para calcular, descrever e resumir os dados coletados da(s) amostra(s) de modo que sejam interpretadas adequadamente.⁷⁶
- As análises descritivas geralmente compreendem a apresentação quantitativa (numérica) em tabelas e/ou gráficos.⁷⁶

R O pacote *explore*¹²⁶ fornece a função *explore*^a para análise exploratória de um banco de dados.

^a<https://www.rdocumentation.org/packages/explore/versions/1.0.2/topics/explore>

R O pacote *dataMaid*¹²⁷ fornece a função *makeDataReport*^a para criar um relatório de análise exploratória de um banco de dados.

^a<https://www.rdocumentation.org/packages/dataMaid/versions/1.4.1/topics/makeDataReport>

R O pacote *DataExplorer*¹²⁸ fornece a função *create_report*^a para criar um relatório de análise exploratória de um banco de dados.

^ahttps://www.rdocumentation.org/packages/DataExplorer/versions/0.8.2/topics/create_report

R O pacote *SmartEDA*¹²⁹ fornece a função *ExpReport*^a para criar um relatório de análise exploratória de um banco de dados.

^a<https://www.rdocumentation.org/packages/SmartEDA/versions/0.3.9/topics/ExpReport>

R O pacote *esquisse*¹³⁰ fornece a função *esquisser*^a para executar uma interface interativa para visualização de dados.

^a<https://www.rdocumentation.org/packages/esquisse/versions/1.1.2/topics/esquisser>

13.2 Estimação

13.2.1 O que é estimativa?

- Estimativa é o valor de uma variável de interesse calculado a partir de uma amostra.⁷

13.2.2 O que é estimativa pontual?

- Estimativa pontual é o valor único de uma variável de interesse calculado a partir de uma amostra.⁷

13.2.3 O que é estimativa intervalar?

- Estimativa intervalar é um intervalo de valores de uma variável de interesse calculado a partir de uma amostra.⁷

13.2.4 O que é estimativa de parâmetro?

- Estimativa de parâmetro é o valor de uma variável de interesse calculado a partir de uma amostra que representa o valor da população.⁷

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

Capítulo 14

Análise inicial de dados

14.1 Análise inicial de dados

14.1.1 O que é análise inicial de dados?

- Análise inicial de dados¹³¹ é uma sequência de procedimentos que visam principalmente a transparência e integridade das pré-condições do estudo para conduzir a análise estatística apropriada de modo responsável para responder aos problemas da pesquisa.⁷⁷
- O objetivo da análise inicial de dados é propiciar dados prontos para análise estatística, incluindo informações confiáveis sobre as propriedades dos dados.⁷⁷
- A análise inicial de dados pode ser dividida nas seguintes etapas:⁷⁷
 - Configuração dos metadados
 - Limpeza dos dados
 - Verificação dos dados
 - Relatório inicial dos dados
 - Refinamento e atualização do plano de análise estatística
 - Documentação e relatório da análise inicial de dados
- A análise inicial de dados não deve ser confundida com análise exploratória¹³², nem deve ser utilizada para hipotetizar após os dados serem coletados (conhecido como *Hypothesizing After Results are Known, HARKing*)¹³³.

14.1.2 Como conduzir uma análise inicial de dados?

- Desenvolva um plano de análise inicial de dados consistente com os objetivos da pesquisa. Por exemplo, verifique a distribuição e escala das variáveis, procure por observações não-usuais ou improváveis, avalie possíveis padrões de dados perdidos.⁷⁷
- Não altere diretamente os dados de uma tabela obtida de uma fonte. Use scripts para implementar eventuais alterações, de modo a manter o registro de todas as modificações realizadas no banco de dados.⁷⁷
- Use os metadados do estudo para guiar a análise inicial dos dados e compartilhe com os dados para maior transparência e reproduzibilidade.⁷⁷
- Representação gráfica dos dados pode ajudar a identificar características e padrões no banco de dados, tais como suposições e tendências.⁷⁷
- Verifique a frequência e proporção de dados perdidos em cada variável, e depois examine por padrões de dados perdidos simultaneamente por duas ou mais variáveis.⁷⁷
- Verifique a frequência e proporção de dados perdidos em cada variável, e depois examine por padrões de dados perdidos simultaneamente por duas ou mais variáveis.⁷⁷

- Exclusão de dados *ad hoc* baseada no desfecho pode influenciar os resultados do estudo, portanto os critérios de exclusão de dados antes da análise estatística (descritiva e/ou inferencial) devem ser reportados.¹³⁴

14.1.3 Quais problemas podem ser detectados na análise inicial de dados?

- Ocorrência de dados perdidos, que podem ser excluídos ou imputados para não reduzir o poder do estudo.⁷



O pacote *stats*⁷¹ fornece a função *na.omit*^a para retornar os dados sem os dados perdidos.

^a<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/na.fail>



O pacote *stats*⁷¹ fornece a função *complete.cases*^a para identificar os casos completos - isto é, sem dados perdidos - em um banco de dados.

^a<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/complete.cases>

- Registros duplicados, que devem ser excluídos para não inflar a amostra.¹³⁵



O pacote *base*⁴⁸ fornece a função *duplicated*^a para identificar elementos duplicados de um banco de dados.

^a<https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/duplicated>

- Codificação 0 ou 1 para variáveis dicotômicas para representar a direção esperada da associação entre elas.¹³⁵
- Ordenação cronológica de variáveis com registros temporais (retrospectivos ou prospectivos).¹³⁵
- A distribuição das variáveis para verificação das suposições das análises planejadas.¹³⁵
- Ocorrência de efeitos teto e piso nas variáveis.¹³⁵

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

Capítulo 15

Distribuições e parâmetros

15.1 Distribuições de probabilidade

15.1.1 O que são distribuições de probabilidade?

- Uma distribuição de probabilidade é uma função que descreve os valores possíveis ou o intervalo de valores de uma variável (eixo horizontal) e a frequência com que cada valor é observado (eixo vertical).⁷⁶

15.1.2 Como representar distribuições de probabilidade?

- Tabelas de frequência, polígonos de frequência, gráficos de barras, histogramas e *boxplots* são formas de representar distribuições de probabilidade.¹³⁶
- Tabelas de frequência mostram as categorias de medição e o número de observações em cada uma. É necessário conhecer o intervalo de valores (mínimo e máximo), que é dividido em intervalos arbitrários chamados “intervalos de classe”.¹³⁶
- Se houver muitos intervalos, não haverá redução significativa na quantidade de dados, e pequenas variações serão perceptíveis. Se houver poucos intervalos, a forma da distribuição não poderá ser adequadamente determinada.¹³⁶
- A quantidade de intervalos pode ser determinada pelo método de Sturges, que é dado pela fórmula $k = 1 + 3.322 \times \log_{10}(n)$, onde k é o número de intervalos e n é o número de observações.¹³⁷
- A quantidade de intervalos pode ser determinada pelo método de Scott, que é dado pela fórmula $h = 3.5 \times \text{sd}(x) \times n^{-1/3}$, onde h é a largura do intervalo, $\text{sd}(x)$ é o desvio padrão e n é o número de observações.¹³⁸
- A quantidade de intervalos pode ser determinada pelo método de Freedman-Diaconis, que é dado pela fórmula $h = 2 \times \text{IQR}(x) \times n^{-1/3}$, onde h é a largura do intervalo, $\text{IQR}(x)$ é o intervalo interquartil e n é o número de observações.¹³⁹
- A largura das classes pode ser determinada dividindo o intervalo total de observações pelo número de classes. Recomenda-se larguras iguais, mas larguras desiguais podem ser usadas quando existirem grandes lacunas nos dados ou em contextos específicos. Os intervalos devem ser mutuamente exclusivos e não sobrepostos, evitando intervalos abertos (ex.: $<5, >10$).¹³⁶
- Polígonos de frequência são gráficos de linhas que conectam os pontos médios de cada barra do histograma. Eles são úteis para comparar duas ou mais distribuições de frequência.¹³⁶
- Gráficos de barra verticais ou horizontais representam a distribuição de frequências de uma variável categórica. A altura de cada barra é proporcional à frequência da classe. A largura da barra é igual à largura da classe. A área de cada barra é proporcional à frequência da classe. A área total do gráfico de barras é igual ao número total de observações.¹³⁶
- Histogramas representam a distribuição de frequências de uma variável contínua. A altura de cada barra é proporcional à frequência da classe. A largura da barra é igual à largura da classe. A área de cada barra é proporcional à frequência da classe. A área total do histograma é igual ao número total de observações.¹³⁶

- *Boxplots* representam a distribuição de frequências de uma variável contínua. A linha central divide os dados em duas partes iguais (mediana ou Q2). A caixa inferior representa o primeiro quartil (Q1) e a caixa superior representa o terceiro quartil (Q3). A linha inferior é o mínimo e a linha superior é o máximo. Os valores atípicos são representados por pontos individuais.¹³⁶

R

O pacote *grDevices*¹⁴⁰ fornece a função *nclass*^a para determinar a quantidade de classes de um histograma com os métodos de Sturge¹³⁷, Scott¹³⁸ ou Freedman-Diaconis¹³⁹.

^a<https://www.rdocumentation.org/packages/grDevices/versions/3.6.2/topics/nclass>

R

O pacote *graphics*¹⁴¹ fornece a função *hist*^a para criar histogramas.

^a<https://www.rdocumentation.org/packages/graphics/versions/3.6.2/topics/hist>

15.1.3 Quais características definem uma distribuição?

- Uma distribuição pode ser definida por modelos matemáticos e caracterizada por parâmetros de tendência central, dispersão, simetria e curtose.

15.1.4 Quais são as distribuições mais comuns?

- Distribuições discretas:

- Uniforme: resultados (finitos) que são igualmente prováveis.?
- Binomial: número de sucessos em k tentativas.?
- Poisson: número de eventos em um intervalo de tempo fixo.?
- Bernoulli: .?
- Geométrica: número de testes até o 1º sucesso.?
- Binomial negativa: número de testes até o k -ésimo sucesso.?
- Hipergeométrica: número de indivíduos na amostra tomados sem reposição.?

- Distribuições contínuas:

- Uniforme: resultados que possuem a mesma densidade.?
- Exponencial: tempo entre eventos.?
- Normal: .?
- Normal padrão: .?
- Aproximação binomial: número de sucessos em uma grande quantidade de tentativas.?
- Aproximação Poisson: número de ocorrências em um intervalo de tempo fixo.?
- Qui-quadrado: .?
- t-Student: .?
- Weibull: .?
- Log-normal: .?
- Beta: .?
- Gama: .?
- Logística: .?
- Pareto.?

15.1.5 Quais são as funções de uma distribuição?

- Função de massa de probabilidade (*probability mass function*, pmf).⁷
- Função de distribuição cumulativa (*cumulative distribution function*, cdf).⁷
- Função quantílicas (*quantile function*, qf).⁷
- Função geradora de números aleatórios (*random function*, rf).⁷

R

O pacote *stats*⁷¹ fornece funções de distribuição de probabilidade (p), funções de densidade (d), funções quantílicas (q) e funções geradoras de números aleatórios (r) para as distribuições normal^a, Student t^b, binomial^c, qui-quadrado^d, uniforme^e, dentre outras.

^a<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/Normal>

^b<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/TDist>

^c<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/Binomial>

^d<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/Chisquare>

^e<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/Uniform>

R

O pacote *ggdist*¹⁴² fornece a função *geom_slabinterval*^a para criar gráficos de distribuição de probabilidade (p) e funções de densidade (d) as distribuições.

^ahttps://www.rdocumentation.org/packages/ggdist/versions/3.3.0/topics/geom_slabinterval

R

O pacote *ggfortify*¹⁴³ fornece a função *ggdistribution*^a para criar gráficos de distribuição de probabilidade (p), funções de densidade (d), funções quantílicas (q) e funções geradoras de números aleatórios (r) para as distribuições.

^a<https://www.rdocumentation.org/packages/ggfortify/versions/0.4.16/topics/ggdistribution>

15.1.6 O que é a distribuição normal?

- A distribuição normal (ou gaussiana) é uma distribuição com desvios simétricos positivos e negativos em torno de um valor central.⁸⁸
- Em uma distribuição normal, o intervalo de 1 desvio-padrão ($\pm 1DP$) inclui cerca de 68% dos dados; de 2 desvios-padrão ($\pm 2DP$) cerca de 95% dos dados; e no intervalo de 3 desvios-padrão ($\pm 3DP$) cerca de 99% dos dados.⁸⁸

15.1.7 Que métodos podem ser utilizados para identificar a normalidade da distribuição?

- Histogramas.⁷⁶
- Gráficos Q-Q.⁷⁶
- Testes de hipótese nula:⁷⁶
 - Kolmogorov-Smirnov
 - Shapiro-Wilk
 - Anderson-Darling

15.1.8 O que são distribuições não-normais?

- ?

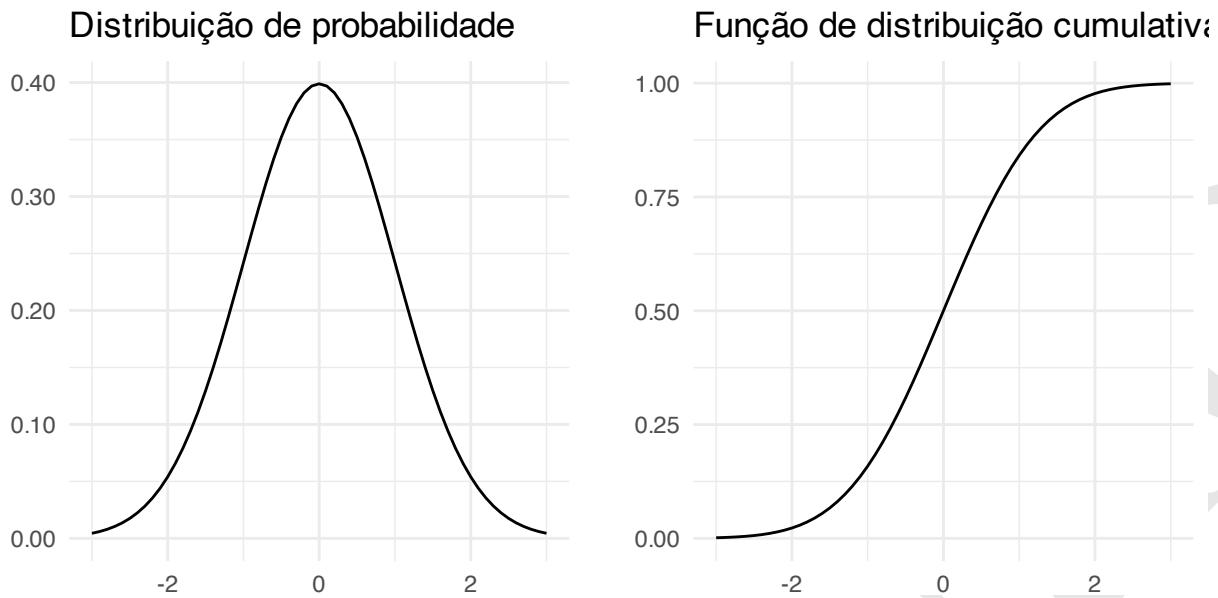


Figura 15.1: Distribuições e funções de probabilidade

15.2 Parâmetros

15.2.1 O que são parâmetros?

- Parâmetros são informações que definem um modelo teórico, como propriedades de uma coleção de indivíduos.⁸⁷
- Parâmetros definem características de uma população inteira, tipicamente não observados por ser inviável ter acesso a todos os indivíduos que constituem tal população.⁷⁶

 O pacote *base*⁴⁸ fornece a função *summary*^a para calcular diversos parâmetros descritivos.

^a<https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/summary>

15.2.2 O que é uma análise paramétrica?

- Testes paramétricos possuem suposições sobre as características e/ou parâmetros da distribuição dos dados na população.⁷⁶
- Testes paramétricos assumem que: a variável é quantitativa numérica (contínua); os dados foram amostrados de uma população com distribuição normal; a variância da(S) amostra(s) é igual à da população; as amostras foram selecionadas de modo aleatório na população; os valores de cada amostra são independentes entre si.^{76,88}
- Testes paramétricos são baseados na suposição de que os dados amostrais provêm de uma população com parâmetros fixos determinando sua distribuição de probabilidade.⁸

15.2.3 O que é uma análise não paramétrica?

- Testes não-paramétricos fazem poucas suposições, ou menos rigorosas, sobre as características e/ou parâmetros da distribuição dos dados na população.^{76,88}
- Testes não-paramétricos são úteis quando as suposições de normalidade não podem ser sustentadas.⁸⁸

15.2.4 Devemos testar as suposições de normalidade?

- Testes preliminares de normalidade não são necessários para a maioria dos testes paramétricos de comparação, pois eles são robustos contra desvios moderados da normalidade. Normalidade da distribuição deve ser

estabelecida para a população.¹⁴⁴

15.2.5 Por que as análises paramétricas são preferidas?

- Em geral, testes paramétricos são mais robustos (isto é, possuem menores erros tipo I e II) que seus testes não-paramétricos correspondentes.^{76,145}
- Testes não-paramétricos apresentam menor poder estatístico (maior erro tipo II) comparados aos testes paramétricos correspondentes.⁸⁸

15.2.6 Que parâmetros podem ser estimados?

- Parâmetros de tendência central.^{88,146}
- Parâmetros de dispersão.^{88,146,147}
- Parâmetros de proporção.^{88,146,148,148}
- Parâmetros de distribuição.¹⁴⁶
- Parâmetros de extremos.⁸⁸

R O pacote *base*⁴⁸ fornece a função *summary*^a para calcular diversos parâmetros descritivos.

^a<https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/summary>

15.3 Tendência central

15.3.1 Que parâmetros de tendência central podem ser estimados?

- *Média*: aritmética, ponderada, geométrica ou harmônica.^{88,146,149}
- *Mediana*.^{88,146,150}
- *Moda*.^{88,146,150}
- A posição relativa das medidas de tendência central (média, mediana e moda) depende da forma da distribuição.¹⁵⁰
- Em uma distribuição normal, as três medidas são idênticas.¹⁵⁰
- A média é sempre puxada para os valores extremos, por isso é deslocada para a cauda em distribuições assimétricas.¹⁵⁰
- A mediana fica entre a média e a moda em distribuições assimétricas.¹⁵⁰
- A moda é o valor mais frequente e, portanto, se localiza no pico da distribuição assimétrica.¹⁵⁰

R O pacote *base*⁴⁸ fornece a função *summary*^a para calcular diversos parâmetros descritivos.

^a<https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/summary>

15.3.2 Como escolher o parâmetro de tendência central?

- A mediana é preferida à média quando existem poucos valores extremos na distribuição, alguns valores são indeterminados, ou há uma distribuição aberta, ou os dados são medidos em uma escala ordinal.¹⁵⁰
- A moda é preferida quando os dados são medidos em uma escala nominal.¹⁵⁰
- A média geométrica é preferida quando os dados são medidos em uma escala logarítmica.¹⁵⁰

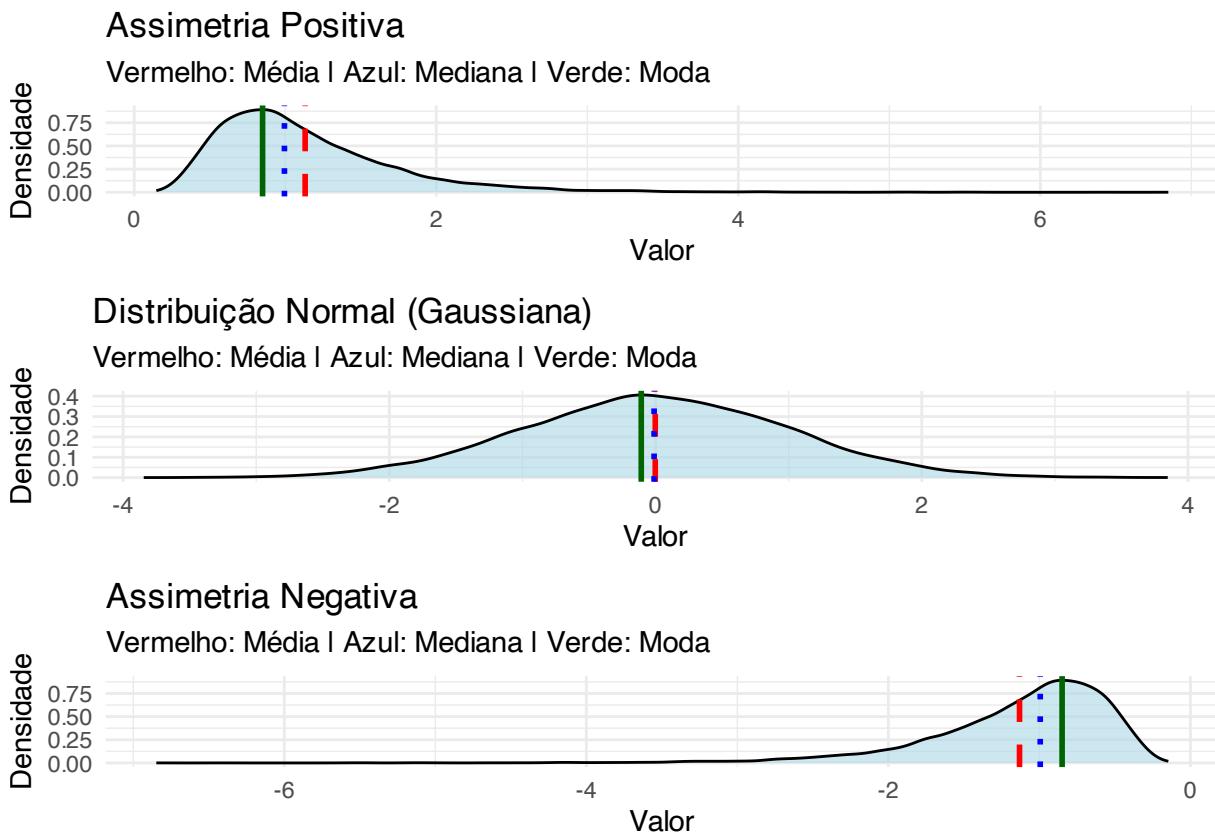


Figura 15.2: Parâmetros de tendência central em distribuições assimétricas e normais.

15.4 Dispersão

15.4.1 Que parâmetros de dispersão podem ser estimados?

- *Variância.*^{88,146}
- *Desvio-padrão:* Informam sobre a dispersão da população e são, portanto, úteis como preditores da variação em novas amostras.^{147,151,152}
- *Erro-padrão:* Refletem a incerteza na média e sua dependência do tamanho da amostra.^{147,151}
- *Amplitude.*^{88,146,152}
- *Intervalo interquartil.*^{88,146,152}
- *Intervalo de confiança:* Captura a média populacional correspondente ao nível de significância α pré-estabelecido.^{88,146,151,153}

R O pacote *base*⁴⁸ fornece a função *summary*^a para calcular diversos parâmetros descritivos.

^a<https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/summary>

R O pacote *stats*⁷¹ fornece a função *confint*^a para calcular o intervalo de confiança em um nível de significância α .

^a<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/confint>

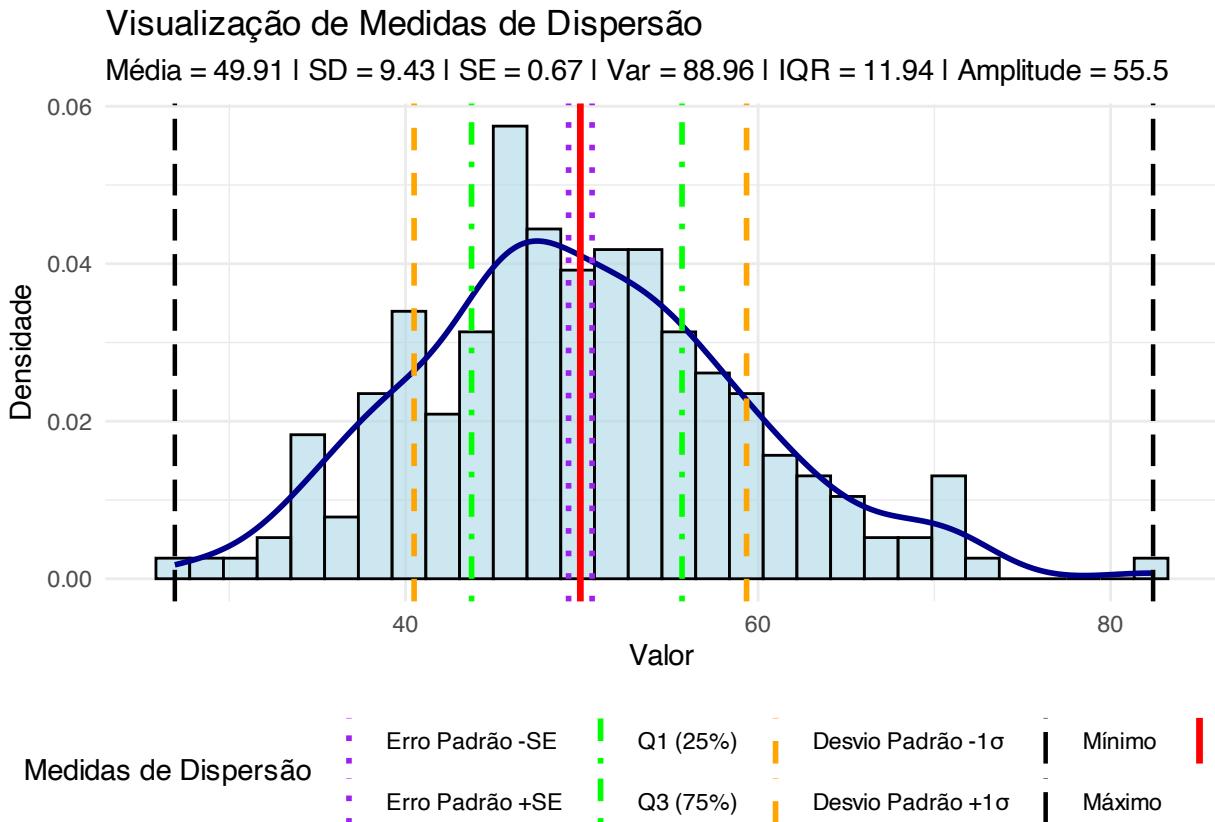


Figura 15.3: Parâmetros de dispersão em distribuições normais.

15.4.2 Como escolher o parâmetro de dispersão?

- Desvio-padrão é apropriado quando a média é utilizada como parâmetro de tendência central em distribuições simétricas.¹⁵²
- Amplitude ou intervalo interquartil são apropriadas para variáveis ordinais ou distribuições assimétricas.¹⁵²

15.4.3 O que é a correção de Bessel para variância?

- Correção de Bessel é um ajuste feito no denominador da fórmula de variância da amostra — ou seja, o número de graus de liberdade — para evitar que a variância amostral seja menor do que a variância populacional.¹⁵⁴
- A correção de Bessel é feita subtraindo-se 1 do número de observações da amostra, ou seja, $n - 1$.¹⁵⁴

15.4.4 Por que a correção de Bessel para variância é importante?

- A correção de Bessel é importante porque a variância amostral tende a ser menor do que a variância populacional, especialmente em amostras pequenas.¹⁵⁴
- A correção de Bessel ajuda a garantir que a variância amostral seja uma estimativa mais precisa da variância populacional, o que é fundamental para a validade dos testes estatísticos e das inferências feitas a partir da amostra.¹⁵⁴

15.5 Proporção

15.5.1 Que parâmetros de proporção podem ser estimados?

- *Frequência absoluta.*^{88,146,148}
- *Frequência relativa.*^{88,146,148}
- *Percentil.*^{88,146,148}

- *Quantil*: é o ponto de corte que define a divisão da amostra em grupos de tamanhos iguais. Portanto, não se referem aos grupos em si, mas aos valores que os dividem.¹⁴⁸
 - Tercil: 2 valores que dividem a amostra em 3 grupos de tamanhos iguais.¹⁴⁸
 - Quartil: 3 valores que dividem a amostra em 4 grupos de tamanhos iguais.¹⁴⁸
 - Quintil: 4 valores que dividem a amostra em 5 grupos de tamanhos iguais.¹⁴⁸
 - Decil: 9 valores que dividem a amostra em 10 grupos de tamanhos iguais.¹⁴⁸



O pacote *base*⁴⁸ fornece a função *summary*^a para calcular diversos parâmetros descritivos.

^a<https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/summary>



O pacote *base*⁴⁸ fornece a função *table*^a para calcular proporções.

^a<https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/table>



O pacote *stats*⁴⁸ fornece a função *quantile*^a para executar análise de percentis.

^a<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/quantile>

15.6 Distribuição

15.6.1 Que parâmetros de distribuição podem ser estimados?

- *Assimetria*.¹⁴⁶
- *Curtose*.¹⁴⁶

15.7 Extremos

15.7.1 O que são extremos?

- Valores extremos podem constituir valores legítimos ou ilegítimos de uma distribuição.¹⁵⁵

15.7.2 Que parâmetros extremos podem ser estimados?

- *Mínimo*.⁸⁸
- *Máximo*.⁸⁸



O pacote *base*⁴⁸ fornece a função *summary*^a para calcular diversos parâmetros descritivos.

^a<https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/summary>

15.8 Valores discrepantes

15.8.1 O que são valores discrepantes (*outliers*)?

- Em termos gerais, um valor discrepante - “fora da curva” ou *outlier* - é uma observação que possui um valor relativamente grande ou pequeno em comparação com a maioria das observações.¹⁵⁶
- Um valor discrepante é uma observação incomum que exerce influência indevida em uma análise.¹⁵⁶
- Valores discrepantes são dados com valores altos de resíduos.¹⁵⁵

15.8.2 Quais são os tipos de valores discrepantes?

- Valores discrepantes podem ser categorizados em três subtipos: *outliers* de erro, *outliers* interessantes e *outliers* aleatórios.¹⁵⁵
- Os valores discrepantes de erro são observações claramente não legítimas, distantes de outros dados devido a imprecisões por erro de mensuração e/ou codificação.¹⁵⁵
- Os valores discrepantes interessantes não são claramente erros, mas podem refletir um processo/mecanismo potencialmente interessante para futuras pesquisas.¹⁵⁵
- Os valores discrepantes aleatórios são observações que resultam por acaso, sem qualquer padrão ou tendência conhecida.¹⁵⁵
- Valores discrepantes podem ser univariados ou multivariados.¹⁵⁵

15.8.3 Por que é importante avaliar valores discrepantes?

- Excluir o valor discrepante implica em reduzir inadequadamente a variância, ao remover um valor que de fato pertence à distribuição considerada.¹⁵⁵
- Manter os dados inalterados (mantendo o valor discrepante) implica em aumentar inadequadamente a variância, pois a observação não pertence à distribuição que fundamenta o experimento.¹⁵⁵
- Em ambos os casos, uma decisão errada pode influenciar o erro do tipo I (α — rejeitar uma hipótese verdadeira) ou o erro do tipo II (β — não rejeitar uma hipótese falsa).¹⁵⁵

15.8.4 Como detectar valores discrepantes?

- Na maioria das vezes, não há como saber de qual distribuição uma observação provém. Por isso, não é possível ter certeza se um valor é legítimo ou não dentro do contexto do experimento.¹⁵⁵
- Recomenda-se seguir um procedimento em duas etapas: detectar possíveis candidatos a *outliers* usando ferramentas quantitativas; e gerenciar os outliers, decidindo manter, remover ou recodificar os valores, com base em informações qualitativas.¹⁵⁵
- A detecção de outliers deve ser aplicada apenas uma vez no conjunto de dados; um erro comum é identificar e tratar os outliers (como remover ou recodificar) e, em seguida, reaplicar o procedimento no conjunto de dados já modificado.¹⁵⁵
- A detecção ou o tratamento dos *outliers* não deve ser realizada após a análise dos resultados, pois isso introduz viés nos resultados.¹⁵⁵

15.8.5 Quais são os métodos para detectar valores discrepantes?

- Valores univariados são comumente considerados *outliers* quando são mais extremos do que a média \pm (desvio padrão \times constante), podendo essa constante ser 3 (99,7% das observações estão dentro de 3 desvios-padrão da média) ou 3,29 (99,9% estão dentro de 3,29 desvios-padrão).¹⁵⁵
- Para detectar *outliers* univariados, recomenda-se o uso da Mediana da Desviação Absoluta (Median Absolute Deviation, MAD), calculado a partir de um intervalo em torno da mediana, multiplicado por uma constante (valor padrão: 1,4826).^{155,157}
- Para detectar *outliers* multivariados, comumente utiliza-se a distância de Mahalanobis, que identifica valores muito distantes do centróide formado pela maioria dos dados (por exemplo, 99%).¹⁵⁵
- Para detectar *outliers* multivariados, recomenda-se o Determinante de Mínima Covariância (Minimum Covariance Determinant, MCD), pois possui o maior ponto de quebra possível e utiliza a mediana, que é o indicador mais robusto em presença de outliers.^{155,158}

15.8.6 Como manejar os valores discrepantes?

- Manter *outliers* pode ser uma boa decisão se a maioria desses valores realmente pertence à distribuição de interesse. Manter *outliers* que pertencem a uma distribuição alternativa pode ser problemático, pois um teste pode se tornar significativo apenas por causa de um ou poucos outliers.¹⁵⁵

- Remover *outliers* pode ser eficaz quando eles distorcem a estimativa dos parâmetros da distribuição. Remover *outliers* que pertencem legitimamente à distribuição pode reduzir artificialmente a estimativa do erro.¹⁵⁵
- Remover *outliers* leva à perda de observações, especialmente em conjuntos de dados com muitas variáveis, quando outliers univariados são excluídos em cada variável.¹⁵⁵
- Recodificar *outliers* evita a perda de uma grande quantidade de dados, mas deve ser baseada em argumentos razoáveis e convincentes.¹⁵⁵
- Erros de observação e de medição são uma justificativa válida para descartar observações discrepantes.¹⁵⁶

15.8.7 Como conduzir análises com valores discrepantes?

- É importante reportar se existem valores discrepantes e como foram tratados.¹⁵⁶
- Valores discrepantes na variável de desfecho podem exigir uma abordagem mais refinada, especialmente quando representam uma variação real na variável que está sendo medida.¹⁵⁶
- Valores discrepantes em uma (co)variável podem surgir devido a um projeto experimental inadequado; nesse caso, abandonar a observação ou transformar a covariável são opções adequadas.¹⁵⁶
- Valores discrepantes podem ser recodificados usando a Winsorização,¹⁵⁹ que transforma os *outliers* em valores de percentis específicos (como o 5º e o 95º).¹⁵⁵

R O pacote *outliers*¹⁶⁰ fornece a função *outlier*^a para identificar os valores mais distantes da média.

^a<https://www.rdocumentation.org/packages/outliers/versions/0.15/topics/outlier>

R O pacote *outliers*¹⁶⁰ fornece a função *rm.outlier*^a para remover os valores mais distantes da média detectados por testes de hipótese e/ou substitui-los pela média ou mediana.

^a<https://www.rdocumentation.org/packages/outliers/versions/0.15/topics/rm.outlier>

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

Capítulo 16

Análise exploratória de dados

16.1 Análise exploratória de dados

16.1.1 O que é análise exploratória de dados?

- Análise exploratória de dados consiste em um processo iterativo de elaboração e interpretação da síntese de dados, tabelas e gráficos, considerando os aspectos teóricos do estudo.¹³²
- Análise exploratória deve ser separada da análise inferencial de testes de hipóteses; a decisão sobre os modelos a testar deve ser feita *a priori*.¹⁵⁶

16.1.2 Por que conduzir a análise exploratória de dados?

- A condução de análise exploratória de dados pode ajudar a identificar padrões e pode orientar trabalhos futuros, mas os resultados não devem ser interpretados como inferências sobre uma população.¹⁵⁶
- A análise exploratória não deve ser usada para definir as questões e hipóteses científicas do estudo.¹⁵⁶

R

O pacote *explore*¹²⁶ fornece a função *explore*^a para análise exploratória de um banco de dados.

^a<https://www.rdocumentation.org/packages/explore/versions/1.0.2/topics/explore>

R

O pacote *dataMaid*¹²⁷ fornece a função *makeDataReport*^a para criar um relatório de análise exploratória de um banco de dados.

^a<https://www.rdocumentation.org/packages/dataMaid/versions/1.4.1/topics/makeDataReport>

R

O pacote *DataExplorer*¹²⁸ fornece a função *create_report*^a para criar um relatório de análise exploratória de um banco de dados.

^ahttps://www.rdocumentation.org/packages/DataExplorer/versions/0.8.2/topics/create_report

R

O pacote *SmartEDA*¹²⁹ fornece a função *ExpReport*^a para criar um relatório de análise exploratória de um banco de dados.

^a<https://www.rdocumentation.org/packages/SmartEDA/versions/0.3.9/topics/ExpReport>

R O pacote *gtExtras*¹⁶¹ fornece a função *gt_plt_summary*^a para criar uma tabela descritiva síntese com histogramas ou gráficos de barra a partir de um banco de dados.

^ahttps://www.rdocumentation.org/packages/gtExtras/versions/0.5.0/topics/gt_plt_summary

R O pacote *radiant*¹⁶² fornece a função *radiant*^a para executar uma interface interativa para análise exploratória de dados.

^a<https://www.rdocumentation.org/packages/radiant/versions/1.5.0/topics/radiant>

16.1.3 Quais etapas constituem a análise exploratória de dados?

- Cada combinação de problema de pesquisa e delineamento de estudo pode demandar um plano de análise exploratório distinto.¹⁵⁶
- Verifique a existência e/ou influência de valores discrepantes (“fora da curva” ou *outliers*):^{131,132,156}
 - Boxplots
 - Gráficos quantil-quantil (Q-Q)

R O pacote *graphics*¹⁴¹ fornece a função *boxplot*^a para construção de gráficos *boxplot*.

^a<https://www.rdocumentation.org/packages/graphics/versions/3.6.2/topics/boxplot>

- Verifique a homocedasticidade (homogeneidade da variância):¹⁵⁶
 - Boxplots condicionais (por fator de análise)
 - Análise dos resíduos do modelo de regressão
 - Gráfico resíduos vs. valores ajustados
- Verifique a normalidade da distribuição dos dados:^{131,156}
 - Histograma das variáveis (por fator de análise)
 - Histograma dos resíduos da regressão
- Verifique a existência de grande quantidade de valores nulos (=0):¹⁵⁶
 - Histograma das variáveis (por fator de análise)
- Verifique a existência de colinearidade entre variáveis independentes de um modelo de regressão:¹⁵⁶
 - Fator de inflação de variância (*variance inflation factor*, VIF)
 - Coeficiente de correlação de Pearson (*r*)
 - Gráfico de dispersão entre variáveis
- Verifique possíveis relações entre as variáveis dependente(s) e independente(s) de um modelo de regressão:¹⁵⁶
 - Gráfico de dispersão entre variáveis independente e dependente
- Verifique possíveis interações entre as variáveis dependente(s) de um modelo de regressão:¹⁵⁶
 - Gráfico *coplot* de dispersão entre variáveis dependentes

R O pacote *graphics*¹⁴¹ fornece a função *coplot*^a para construção de gráficos *boxplot* condicionais.

^a<https://www.rdocumentation.org/packages/graphics/versions/3.6.2/topics/coplot>

- Verifique por dependência entre variáveis de um modelo de regressão:¹⁵⁶

- Gráfico de série temporal das variáveis
- Gráfico de autocorrelação entre as variáveis

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

RAASCUNHO

RASCUNHO

Capítulo 17

Análise descritiva

17.1 Análise descritiva

17.1.1 O que é análise descritiva?

- Análise descritiva é usada para compreendermos algum aspecto de um conjunto de dados, respondendo a perguntas do tipo “quando?”, “onde?”, “quem?”, “o quê?”, “como?” e “e daí?”.^{76,163}

17.1.2 Como apresentar os resultados descritivos?

- Variáveis categóricas: Reporte valores de frequência absoluta e relativa (n, percentual).¹⁶⁴
- Organização das tabelas: as variáveis são exibidas em linhas e os grupos são exibidos em colunas.¹⁶⁴
- Calcule percentagens para as colunas (isto é, entre grupos) e não entre linhas.¹⁶⁴
- Em caso de dados perdidos, não inclua uma linha com total de dados perdidos, pois distorce as proporções entre colunas e as análises de tabela de contingência. Indique no texto ou em uma coluna separada o total de dados perdidos por variável.¹⁶⁴

17.2 Apresentação de resultados numéricos

17.2.1 O que são casas decimais?

- O número de casas decimais refere-se à quantidade de dígitos que aparecem após a vírgula decimal.^{165,166}

17.2.2 O que são dígitos significativos?

- O termo “dígitos significativos” é preferido a “algarismos significativos” ou “dígitos efetivos” e não se relaciona com significância estatística.^{165,166}
- O número de dígitos significativos é a soma total de dígitos, desconsiderando a vírgula decimal e os zeros à esquerda; os zeros à direita são considerados informativos, salvo exceções.^{165,166}

17.2.3 Como arredondar dados numéricos?

- Apresentar dados com quantidade excessiva de casas decimais pode dificultar a interpretação e induzir erroneamente uma precisão espúria.^{165,166}
- A precisão é determinada pelo grau de arredondamento aplicado, medido em casas decimais ou dígitos significativos.^{165,166}
- O arredondamento também introduz erros, uma vez que aumenta a imprecisão (isto é, incerteza) em torno do valor original.^{165,166}

Tabela 17.1: Quantidade de casas decimais e dígitos significativos.

Valor	Casas Decimais	Dígitos Significativos
0,00789	5	0
0,0456	4	0
45,6	1	2
123,456	3	3
7890,0000	4	4

Tabela 17.2: Valores originais, arredondamentos e erros de arredondamento por casas decimais.

Valor	Casas Decimais	Dígitos Significativos	2 Casas decimais [Margem de erro]	1 Casa decimal [Margem de erro]	Sem casa decimal [Margem de erro]
0,00789	5	0	0,01 [0,005, 0,015]	0,0 [-0,05, 0,05]	0 [-0,5, 0,5]
0,0456	4	0	0,05 [0,045, 0,055]	0,0 [-0,05, 0,05]	0 [-0,5, 0,5]
45,6	1	2	45,60 [45,595, 45,605]	45,6 [45,55, 45,65]	46 [45,5, 46,5]
123,456	3	3	123,46 [123,455, 123,465]	123,5 [123,45, 123,55]	123 [122,5, 123,5]
7890,0000	4	4	7890,00 [7889,995, 7890,005]	7890,0 [7889,95, 7890,05]	7890 [7889,5, 7890,5]

- A regra geral é utilizar 2 ou 3 dígitos significativos para tamanhos de efeito e 1 ou 2 dígitos significativos para medidas de variabilidade.¹⁶⁶
- Regra dos 3 dígitos significativos para proporção de risco: em média, o erro de arredondamento é menor que os 0,5% exigidos, de modo que três dígitos significativos são mais precisos do que o necessário.¹⁶⁵
- Regra dos 4 dígitos significativos para proporção de risco: divida a proporção de risco por quatro e arredonde para dois dígitos significativos e, em seguida, relate a proporção para esse número de casas decimais.¹⁶⁵

17.3 Tabelas

17.3.1 Por que usar tabelas?

- Tabelas complementam o texto (e vice-versa), e podem apresentar os dados de modo mais acessível e informativo.¹⁶⁷

17.3.2 Que informações incluir nas tabelas?

- Título ou legenda, uma síntese descritiva (geralmente por meio de parâmetros descritivos), intervalos de confiança e/ou P-valores conforme necessário para adequada interpretação.^{167,168}

17.3.3 Quais são os erros mais comuns de preenchimento de tabelas?

- Erros tipográficos.¹⁶⁹
- Ausência de rótulos ou unidades nas variáveis.¹⁶⁹
- Relatar estatísticas incorretamente, tais como rotular variáveis contínuas como porcentagens.¹⁶⁹
- Estatísticas descritivas de tendência central (ex.: médias) relatadas sem a estatística de dispersão correspondente (ex.: desvio-padrão).¹⁶⁹
- Desvio-padrão nulo ($\sigma = 0$).¹⁶⁹

- Valores porcentuais que não correspondem ao numerador dividido pelo denominador.¹⁶⁹

R

O pacote *flextable*¹⁷⁰ fornece as funções *flextable*^a, *as_flextable*^b e *save_as_docx*^c para criar e salvar tabelas formatadas em DOCX.

^a<https://search.r-project.org/CRAN/refmans/flextable/html/flextable.html>

^bhttps://search.r-project.org/CRAN/refmans/flextable/html/as_flextable.html

^chttps://search.r-project.org/CRAN/refmans/flextable/html/save_as_docx.html

R

O pacote *rempsyc*¹⁷¹ fornece a função *nice_table*^a para criar tabelas formatadas.

^ahttps://search.r-project.org/CRAN/refmans/rempsyc/html/nice_table.html

R

O pacote *table1*¹⁷² fornece a função *table1*^a para construção de tabelas.

^a<https://search.r-project.org/CRAN/refmans/table1/html/table1.html>

R

O pacote *gtsummary*¹⁷³ fornece a função *tbl_summary*^a para construção da ‘Tabela 1’ com dados descritivos.

^ahttps://search.r-project.org/CRAN/refmans/gtsummary/html/tbl_summary.html

17.4 Tabela 1

17.4.1 O que é a ‘Tabela 1’?

- A ‘Tabela 1’ descreve as características demográficas, sociais e clínicas da amostra, completa ou agrupada por algum fator, geralmente por meio de parâmetros de tendência central e dispersão.^{174,175}

17.4.2 Qual a utilidade da ‘Tabela 1’?

- Descrever (conhecer) as características da amostra e dos grupos sendo comparados, quando aplicável.¹⁷⁵
- Verificar aderência ao protocolo do estudo, incluindo critérios de inclusão/exclusão, tamanho da amostra e perdas amostrais.¹⁷⁵
- Permitir a replicação do estudo.¹⁷⁵
- Meta-analisar os dados junto a estudos similares.¹⁷⁵
- Avaliar a generalização (validade externa) das conclusões do estudo.¹⁷⁵

17.4.3 O que é a falácia da ‘Tabela 1’?

- Falácia da Tabela 1 ocorre pela interpretação errônea dos P-valores na comparação entre grupos, na linha de base, de um ensaio clínico aleatorizado.¹⁷⁶

17.4.4 Como construir a ‘Tabela 1’?

- A Tabela 1 geralmente é utilizada para descrever as características da amostra estudada, possibilitando a análise de ameaças à validade interna e/ou externa ao estudo.^{145,177}

R

O pacote *table1*¹⁷² fornece a função *table1*^a para construção de tabelas.

^a<https://search.r-project.org/CRAN/refmans/table1/html/table1.html>

 O pacote *gtsummary*¹⁷³ fornece a função *tbl_summary*^a para construção da ‘Tabela 1’ com dados descritivos.

^ahttps://search.r-project.org/CRAN/refmans/gtsummary/html/tbl_summary.html

17.5 Tabela 2

17.5.1 Qual a utilidade da ‘Tabela 2’?

- A Tabela 2 mostra associações ajustadas multivariadas com o resultado para variáveis resumidas na Tabela 1.¹⁷⁴

17.5.2 O que é a falácia da ‘Tabela 2’?

- A Tabela 2 pode induzir ao erro de interpretação pelas estimativas de efeitos para covariáveis do modelo também serem utilizados para controlar a confusão da exposição.^{174,178}
- Ao apresentar estimativas de efeito ajustadas para covariáveis juntamente com a estimativa de efeito ajustada para a exposição primária, a Tabela 2 sugere implicitamente que todas estas estimativas podem ser interpretadas de forma semelhante, se não de forma idêntica, como estimativa do efeito total.^{174,178}
- A falácia da Tabela 2 pode ser evitada limitando-se a tabela a estimativas das medidas primárias do efeito de exposição nos diferentes modelos, com as covariáveis secundárias de “ajuste” relatadas em uma nota de rodapé, juntamente com a forma como foram categorizadas ou modeladas.¹⁷⁴

17.5.3 Como construir a ‘Tabela 2’?

- A Tabela 2 pode ser utilizada para apresentar estimativas de múltiplos efeitos ajustados de um mesmo modelo estatístico.¹⁷⁴

 O pacote *table1*¹⁷² fornece a função *table1*^a para construção de tabelas.

^a<https://search.r-project.org/CRAN/refmans/table1/html/table1.html>

 O pacote *gtsummary*¹⁷³ fornece a função *tbl_summary*^a para construção da ‘Tabela 1’ com dados descritivos.

^ahttps://search.r-project.org/CRAN/refmans/gtsummary/html/tbl_summary.html

17.6 Gráficos

17.6.1 O que são gráficos?

- Gráficos são utilizados para apresentar dados (geralmente em grande quantidade) de modo mais intuitivo e fácil de compreender.¹⁷⁹

17.6.2 Que elementos incluir em gráficos?

- Título, eixos horizontal e vertical com respectivas unidades, escalas em intervalos representativos das variáveis, legenda com símbolos, síntese descritiva dos valores e respectiva margem de erro, conforme necessário para adequada interpretação.¹⁷⁹

R

Os pacotes *ggplot2*¹⁸⁰, *plotly*¹⁸¹ e *corrplot*¹⁸² fornecem diversas funções para construção de gráficos tais como *ggplot*^a, *plot_ly*^b e *corrplot*^c respectivamente.

^a<https://www.rdocumentation.org/packages/ggplot2/versions/3.4.3/topics/ggplot>

^bhttps://www.rdocumentation.org/packages/plotly/versions/4.10.2/topics/plot_ly

^c<https://www.rdocumentation.org/packages/corrplot/versions/0.92/topics/corrplot>

17.6.3 Para que servem as barras de erro em gráficos?

- Barras de erro ajudam ao autor a apresentar as informações que descrevem os dados (por exemplo, em uma análise descritiva) ou sobre as inferências ou conclusões tomadas a partir de dados.^{151,153}
- Barras de erro mais longas representam mais imprecisão (maiores erros), enquanto barras mais curtas representam mais precisão na estimativa.¹⁵³
- Barras de erro descritivas geralmente apresentam a amplitude (mínimo-máximo) ou desvio-padrão.¹⁵³
- Barras de erro inferenciais geralmente apresentam o erro-padrão ou intervalo de confiança no nível de significância α pré-estabelecido.^{151,153}
- Barras de erro com desvio-padrão são úteis para descrever a variabilidade dos dados, enquanto as barras de erro com erro padrão da média são úteis para descrever a precisão do parâmetro estimado (média) e sua relação com o tamanho da amostra.¹⁵¹
- Barras de erro com intervalo de confiança são úteis para fornecer uma estimativa da incerteza da estimativa do parâmetro populacional.¹⁵¹
- O comprimento das barras de erro sugere graficamente a imprecisão dos dados do estudo, uma vez que o valor verdadeiro da população pode estar em qualquer nível do intervalo da barra.¹⁵³
- De modo contraintuitivo, um espaço entre as barras não garante significância, nem a sobreposição a descarta—depende do tipo de barra.¹⁵¹
- Para amostras pequenas é preferível apresentar os dados brutos, uma vez que as barras de erro não serão muito informativas.¹⁵¹

17.6.4 Quais são as boas práticas na elaboração de gráficos?

- O tamanho da amostra total e subgrupos, se houver, deve estar descrito na figura ou na sua legenda.¹⁵³
- Para análise inferencial de figuras, as barras de erro representadas por erro-padrão ou intervalo de confiança no nível de significância α pré-estabelecido são preferíveis à amplitude ou desvio-padrão.^{151,153}
- Evite gráficos de barra e mostre a distribuição dos dados sempre que possível.¹⁸³
- Exiba os pontos de dados em boxplots.¹⁸³
- Use *jitter* simétrico em gráficos de pontos para permitir a visualização de todos os dados.¹⁸³
- Prefira palhetas de cor adaptadas para daltônicos.¹⁸³

R

O pacote *ggsci*¹⁸⁴ fornece palhetas de cores tais como *pal_lancet*^a, *pal_nejm*^b e *pal_npg*^c inspiradas em publicações científicas para uso em gráficos.

^ahttps://www.rdocumentation.org/packages/ggsci/versions/3.0.0/topics/pal_lancet

^bhttps://www.rdocumentation.org/packages/ggsci/versions/3.0.0/topics/pal_nejm

^chttps://www.rdocumentation.org/packages/ggsci/versions/3.0.0/topics/pal_npg

R

O pacote *grDevices*¹⁴⁰ fornece a função *dev.new*^a para controlar diversos aspectos do gráfico, tais como tamanho e resolução.

^a<https://www.rdocumentation.org/packages/grDevices/versions/3.6.2/topics/dev>

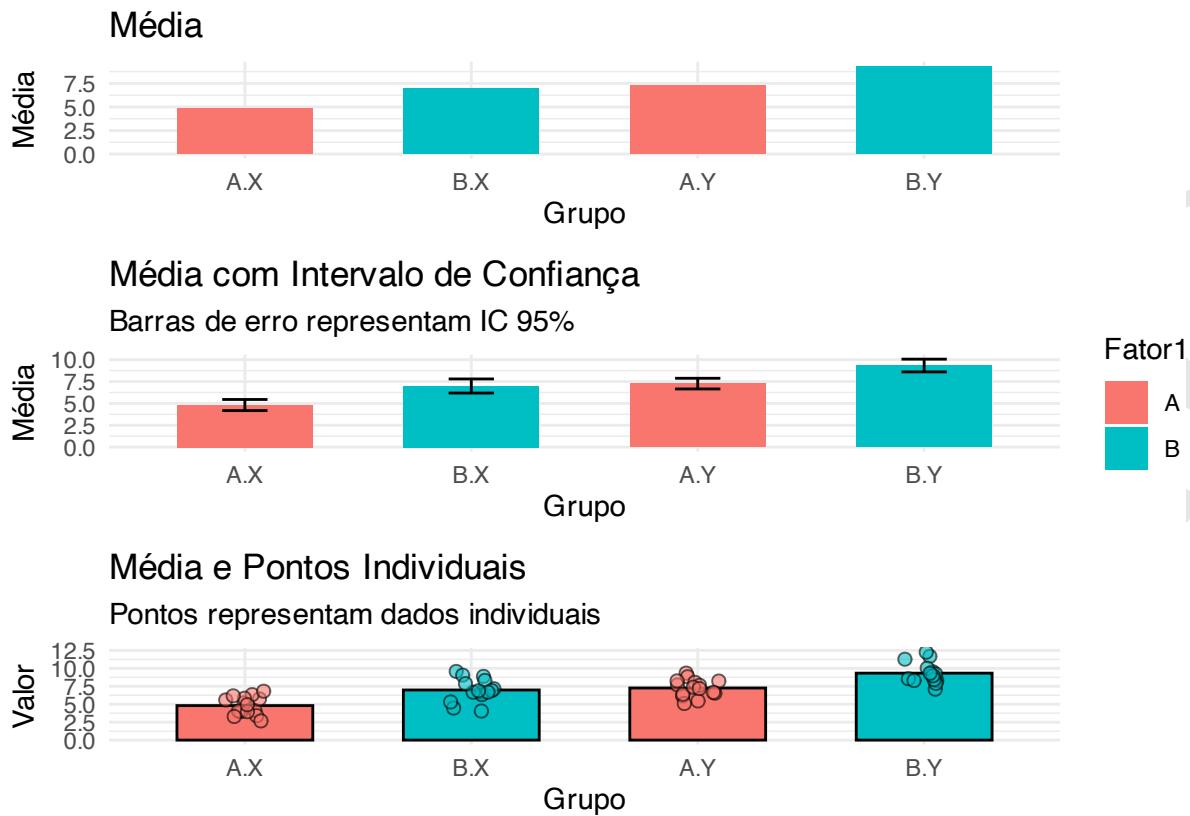


Figura 17.1: Gráficos de barras representando médias, barras de erro e dados individuais.

17.6.5 Como exportar figuras em formato TIFF?

- ?

O pacote *tiff*¹⁸⁵ fornece a função *writeTIFF*^a para exportar gráficos em formato TIFF.

^a<https://www.rdocumentation.org/packages/tiff/versions/0.1-11/topics/writeTIFF>

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

Capítulo 18

Análise robusta

18.1 Raciocínio inferencial robusto

18.1.1 O que é análise robusta?

- Análise robusta é uma abordagem estatística que busca fornecer resultados confiáveis mesmo quando as suposições clássicas dos modelos estatísticos são violadas, como normalidade e homocedasticidade. Ela utiliza métodos que são menos sensíveis a outliers e outras irregularidades nos dados.¹⁸⁶

18.1.2 Por que usar análise robusta?

- Métodos clássicos como ANOVA e regressão por mínimos quadrados assumem normalidade e homocedasticidade — suposições frequentemente violadas na prática. Violações dessas suposições podem comprometer os resultados, reduzindo o poder estatístico, distorcendo os intervalos de confiança e obscurecendo as reais diferenças entre grupos.¹⁸⁶
- Testar previamente as suposições não é suficiente: testes de homocedasticidade têm baixo poder e não garantem segurança analítica.¹⁸⁶
- Métodos estatísticos robustos oferecem uma solução mais segura e eficaz, lidando melhor com dados não ideais.¹⁸⁶

18.1.3 Quando usar análise robusta?

- Em alguns casos, os métodos robustos confirmam os resultados clássicos; em outros, revelam interpretações completamente diferentes. A única forma de saber o impacto real dos métodos robustos é usá-los e comparar com os métodos tradicionais.¹⁸⁶

18.1.4 O que é Winsorização?

- Winsorização é uma técnica que substitui os valores extremos (outliers) por valores menos extremos, preservando a estrutura dos dados. Isso é feito definindo limites superior e inferior e substituindo os valores que ultrapassam esses limites pelos próprios limites.¹⁸⁶

R O pacote *WRS2*¹⁸⁷ fornece as funções *winmean*^a e *winvar*^b para calcular a média e variância Winsorizadas.

^a<https://www.rdocumentation.org/packages/WRS2/versions/1.1-6/topics/trimse>

^b<https://www.rdocumentation.org/packages/WRS2/versions/1.1-6/topics/trimse>

R O pacote *WRS2*¹⁸⁷ fornece a função *yuen*^a para realizar o teste de comparação de Yuen de médias Winsorizadas para amostras independentes ou dependentes.

^a<https://www.rdocumentation.org/packages/WRS2/versions/1.1-6/topics/yuen>

R O pacote *WRS2*¹⁸⁷ fornece a função *wincor*^a para calcular a correlação Winsorizada.

^a<https://www.rdocumentation.org/packages/WRS2/versions/1.1-6/topics/pbcor>

R O pacote *WRS2*¹⁸⁷ fornece as funções *t1way*^a, *t2way*^b e *t3way*^c para realizar testes de comparação de médias Winsorizadas para análise de variância para 1, 2 ou 3 fatores, respectivamente.

^a<https://www.rdocumentation.org/packages/WRS2/versions/1.1-6/topics/t1way>

^b<https://www.rdocumentation.org/packages/WRS2/versions/1.1-6/topics/t2way>

^c<https://www.rdocumentation.org/packages/WRS2/versions/1.1-6/topics/t3way>

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

PARTE 4: INFERÊNCIA E TESTES ESTATÍSTICOS

De amostras para populações

RASCUNHO

Capítulo 19

Análise inferencial

19.1 Raciocínio inferencial

19.1.1 O que é análise inferencial?

- Na análise inferencial são utilizados dados da(s) amostra(s) para fazer uma inferência válida (isto é, estimativa) sobre os parâmetros populacionais desconhecidos.⁷⁶
- No paradigma de Jerzy Neyman e Egon Pearson, um teste de hipótese científica envolve a tomada de decisão sobre hipóteses nulas (H_0) e alternativa (H_1) concorrentes e mutuamente exclusivas.¹⁸⁸

19.1.2 Quais são os tipos de raciocínio inferencial?

- Inferência deductiva: Uma dada hipótese inicial é utilizada para prever o que seria observado caso tal hipótese fosse verdadeira.¹⁸⁹
- Inferência induktiva: Com base nos dados observados, avalia-se qual hipótese é mais defensável (isto é, mais provável).¹⁸⁹

19.1.3 Quais são as questões fundamentais da análise inferencial?

- A direção do efeito¹⁹⁰
- A magnitude do efeito¹⁹⁰
- A importância do efeito¹⁹⁰

19.2 Hipóteses científicas

19.2.1 O que é hipótese científica?

- Hipótese científica é uma ideia que pode ser testada.¹⁸⁸
- Definir claramente os problemas e os objetivos da pesquisa são o ponto de partida de todos os estudos científicos.⁷⁵

19.2.2 Quais são as fontes de ideias para gerar hipóteses científicas?

- Revisão das práticas atuais.¹⁹¹
- Desafio a ideias aceitas.¹⁹¹
- Conflito entre ideias divergentes.¹⁹¹
- Variações regionais, temporais e populacionais.¹⁹¹
- Experiências dos próprios pesquisadores.¹⁹¹

- Imaginação sem fronteiras ou limites convencionais.¹⁹¹

19.3 Hipóteses estatísticas

19.3.1 O que é hipótese nula?

- A hipótese nula (H_0) é uma expressão que representa o estado atual do conhecimento (*status quo*), em geral a não existência de um determinado efeito.¹⁴⁶

19.3.2 O que é hipótese alternativa?

- A hipótese alternativa (H_1) é uma expressão que contém as situações que serão testadas, de modo que um resultado positivo indique alguma ação a ser conduzida.¹⁴⁶

19.3.3 Qual hipótese está sendo testada?

- A hipótese nula (H_0) é a hipótese sob teste em análises inferenciais.⁸⁸
- Pode-se concluir sobre rejeitar ou não rejeitar a hipótese nula (H_0).⁸⁸
- Não se conclui sobre a hipótese alternativa (H_1).¹⁴⁶
- Para testar a hipótese nula, deve-se selecionar o nível de significância crítica (P-valor de corte); a probabilidade de rejeitarmos uma hipótese nula verdadeira (α); e a probabilidade de não rejeitarmos uma hipótese nula falsa (β).¹⁸⁸

19.4 Testes de hipóteses

19.4.1 Quais são os tipos de teste de hipóteses?

- Teste (clássico) de significância da hipótese nula.¹⁹²
- Teste de mínimos efeitos.¹⁹²
- Teste de equivalência.¹⁹²
- Teste de inferioridade.¹⁹²
- Teste de não-inferioridade.[?]
- Teste de superioridade.[?]

19.4.2 O que é uma família de hipóteses?

- ?

19.4.3 O que são testes *ad hoc* e *post hoc*?

- ?

19.4.4 Como ajustar a análise inferencial para hipóteses múltiplas?

- ?

O pacote *stats*⁷¹ fornece a função *p.adjust*^a para ajustar o P-valor utilizando diversos métodos.

^a<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/p.adjust>

19.4.5 O que são testes unicaudais e bicaudais?

- ?

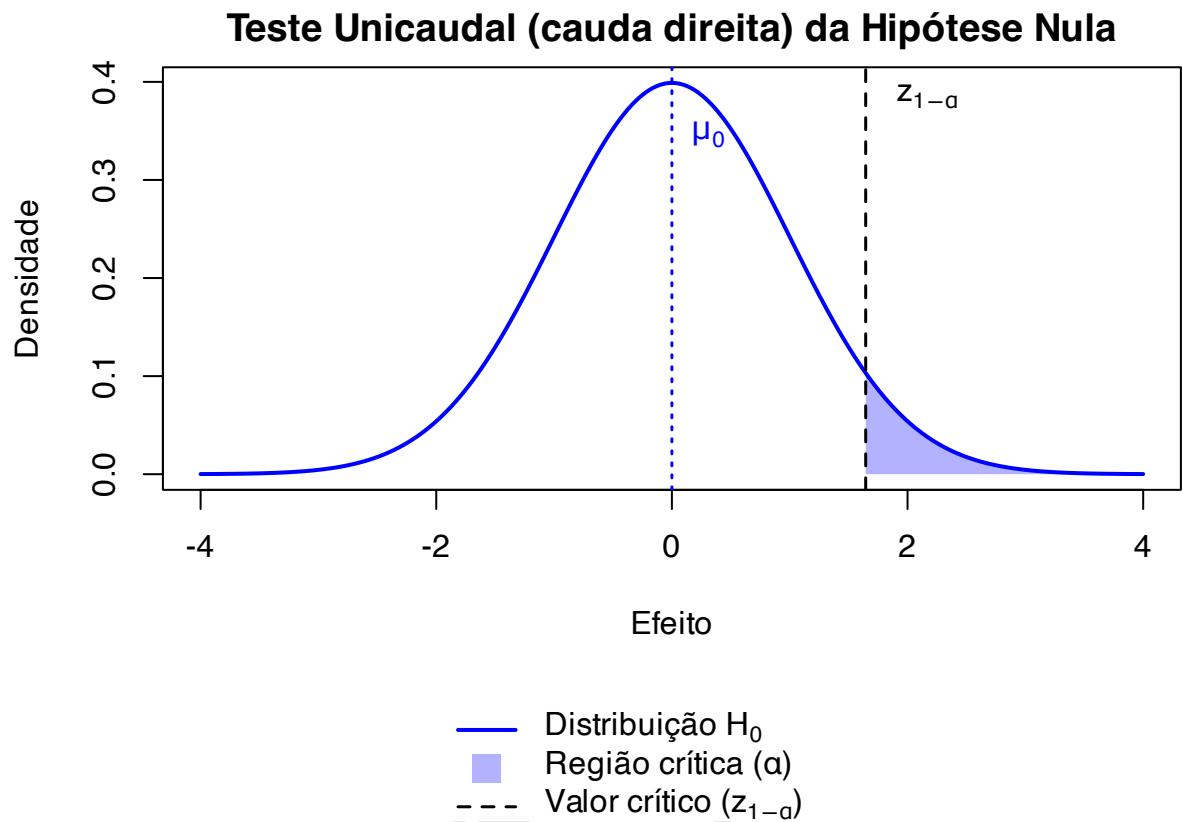


Figura 19.1: Representação gráfica de um teste de hipótese (unicaudal).

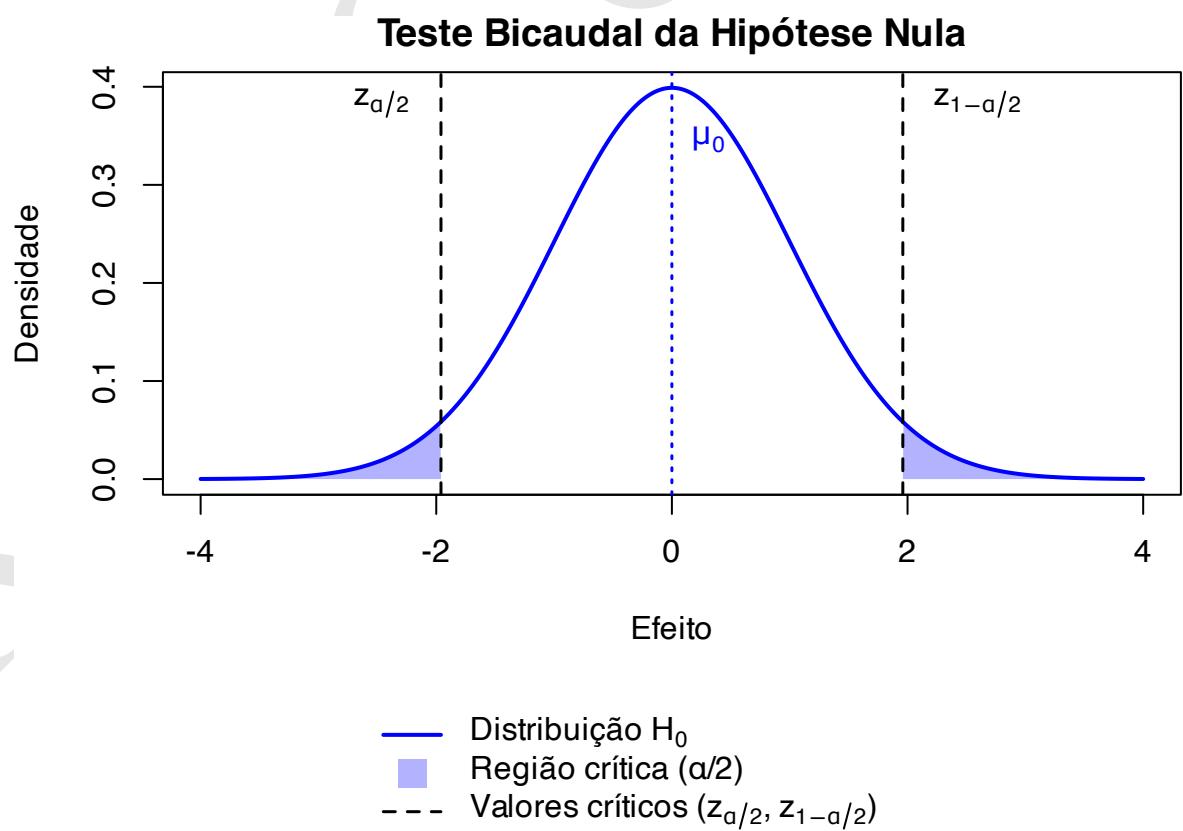


Figura 19.2: Representação gráfica de um teste de hipótese (bicaudal).

19.4.6 O que reportar após um teste de hipótese?

- P-valores, como estimativa da significância estatística.¹⁹³
- Tamanho do efeito, como estimativa de significância substantiva (clínica).¹⁹³

19.5 Poder do teste

19.5.1 O que é poder do teste?

- Poder do teste é a probabilidade de rejeitar corretamente a hipótese nula (H_0) quando esta é falsa.¹⁸⁸
- Poder do teste pode ser calculado como $(1 - \beta)$.¹⁸⁸

19.5.2 O que é análise de poder do teste?

- Poder é a probabilidade de que um dado tamanho de efeito será observado em um experimento futuro sob um conjunto de hipóteses - tamanho de efeito real e erro tipo I - para um dado tamanho de amostra.¹⁹⁴
- O objetivo geral da análise de poder ao projetar um estudo é escolher um tamanho de amostra que controle os 2 tipos de erros de inferência estatística: tipo I (α , resultado falso-positivo) e tipo II (β , resultado falso-negativo).¹⁹⁴
- Numericamente, o poder de um estudo é calculado como $1 - \beta$ e reportado em valor percentual.¹⁹⁴

19.5.3 Quando realizar a análise de poder do teste?

- Na fase de projeto de pesquisa: a análise de poder para determinar o tamanho da amostra objetiva que o tamanho da amostra permita uma probabilidade razoável de detectar um efeito significativo pré-especificado.¹⁹⁴
- Após a coleta de dados: a análise de poder objetiva informar estudos futuros a respeito do tamanho da amostra necessário para a detecção de um efeito significativo pré-especificado.¹⁹⁴

R O pacote *pwr*¹⁹⁵ fornece a função *pwr.2p.test*^a para cálculo do poder do teste de proporção balanceado (2 amostras com mesmo número de participantes).

^a<https://www.rdocumentation.org/packages/pwr/versions/1.3-0/topics/pwr.2p.test>

R O pacote *pwr*¹⁹⁵ fornece a função *pwr.2p2n.test*^a para cálculo do poder do teste de proporção não balanceado (2 amostras com diferente número de participantes).

^a<https://www.rdocumentation.org/packages/pwr/versions/1.3-0/topics/pwr.2p.test>

R O pacote *pwr*¹⁹⁵ fornece a função *pwr.anova.test*^a para cálculo do poder do teste de análise de variância balanceado (3 ou mais amostras com mesmo número de participantes).

^a<https://www.rdocumentation.org/packages/pwr/versions/1.3-0/topics/pwr.anova.test>

R O pacote *pwr*¹⁹⁵ fornece a função *pwr.chisq.test*^a para cálculo do poder do teste de qui-quadrado χ^2 .

^a<https://www.rdocumentation.org/packages/pwr/versions/1.3-0/topics/pwr.chisq.test>

R O pacote *pwr*¹⁹⁵ fornece a função *pwr.f2.test*^a para cálculo do poder do teste com modelo linear geral.

^a<https://www.rdocumentation.org/packages/pwr/versions/1.3-0/topics/pwr.f2.test>

R

O pacote *pwr*¹⁹⁵ fornece a função *pwr.norm.test*^a para cálculo do poder do teste de média de uma distribuição normal com variância conhecida.

^a<https://www.rdocumentation.org/packages/pwr/versions/1.3-0/topics/pwr.norm.test>

R

O pacote *pwr*¹⁹⁵ fornece a função *pwr.p.test*^a para cálculo do poder do teste de proporção (1 amostra).

^a<https://www.rdocumentation.org/packages/pwr/versions/1.3-0/topics/pwr.p.test>

R

O pacote *pwr*¹⁹⁵ fornece a função *pwr.r.test*^a para cálculo do poder do teste de correlação (1 amostra).

^a<https://www.rdocumentation.org/packages/pwr/versions/1.3-0/topics/pwr.r.test>

R

O pacote *pwr*¹⁹⁵ fornece a função *pwr.t.test*^a para cálculo do poder do teste *t* de diferença de 1 amostra, 2 amostras dependentes ou 2 amostras independentes (grupos balanceados).

^a<https://www.rdocumentation.org/packages/pwr/versions/1.3-0/topics/pwr.t.test>

R

O pacote *pwr*¹⁹⁵ fornece a função *pwr.t2n.test*^a para cálculo do poder do teste *t* de diferença de 2 amostras independentes (grupos não balanceados).

^a<https://www.rdocumentation.org/packages/pwr/versions/1.3-0/topics/pwr.t2n.test>

R

O pacote *longpower*¹⁹⁶ fornece a função *power.mrm*^a para calcular o poder de testes com análises por modelo de regressão linear misto.

^a<https://www.rdocumentation.org/packages/longpower/versions/1.0.24/topics/power.mrm>

R

O pacote *Superpower*¹⁹⁷ fornece a função *power.ftest*^a para calcular o poder do teste por análise de testes F.

^a<https://www.rdocumentation.org/packages/Superpower/versions/0.2.0/topics/power.ftest>

R

O pacote *Superpower*¹⁹⁷ fornece a função *power_oneway_between*^a para calcular o poder do teste por análise de variância (ANOVA) de 1 fator entre-sujeitos.

^ahttps://www.rdocumentation.org/packages/Superpower/versions/0.2.0/topics/power_oneway_between

R

O pacote *Superpower*¹⁹⁷ fornece a função *power_oneway_within*^a para calcular o poder do teste por análise de variância (ANOVA) de 1 fator intra-sujeitos.

^ahttps://www.rdocumentation.org/packages/Superpower/versions/0.2.0/topics/power_oneway_within

R

O pacote *Superpower*¹⁹⁷ fornece a função *power_oneway_ancova*^a para calcular o poder do teste por análise de covariância (ANCOVA).

^ahttps://www.rdocumentation.org/packages/Superpower/versions/0.2.0/topics/power_oneway_ancova

R O pacote *Superpower*¹⁹⁷ fornece a função *power_twoway_between*^a para calcular o poder do teste por análise de covariância (ANOVA) de 2 fatores entre-sujeitos.

^ahttps://www.rdocumentation.org/packages/Superpower/versions/0.2.0/topics/power_twoway_between

R O pacote *Superpower*¹⁹⁷ fornece a função *power_threeway_between*^a para calcular o poder do teste por análise de covariância (ANOVA) de 3 fatores entre-sujeitos.

^ahttps://www.rdocumentation.org/packages/Superpower/versions/0.2.0/topics/power_threeway_between

R O pacote *InteractionPower*¹⁹⁸ fornece a função *power_interaction*^a para calcular o poder do teste por análise de efeito de interações.

^ahttps://www.rdocumentation.org/packages/InteractionPower/versions/0.2.1/topics/power_interaction

19.5.4 Por que a análise de poder do teste *post hoc* é inadequada?

- A análise do poder é teoricamente incorreta, uma vez que a probabilidade calculada $1 - \beta$ expressa a probabilidade de um evento futuro, o que não é mais relevante quando o evento de interesse já ocorreu.^{164,194}

19.5.5 O que pode ser realizado ao invés da análise de poder?

- Após a coleta e análise de dados, recomenda-se realizar a análise e interpretação dos resultados a partir do tamanho do efeito e do seu intervalo de confiança no nível de significância α pré-estabelecido.¹⁹⁴

19.6 Inferência visual

19.6.1 O que é inferência visual?

- Inferência visual consiste na interpretação de dados apresentados em gráficos.¹⁹⁹
- Para inferência visual, recomenda-se a apresentação dos dados em gráficos com estimativas de tendência central e seu intervalo (preferencialmente intervalo de confiança no nível de significância α pré-estabelecido).¹⁹⁹

19.6.2 Por que usar intervalos de confiança para inferência visual?

- Intervalos de confiança fornecem estimativas pontuais e intervalares na mesma unidade de medida da variável.¹⁹⁹
- Existe uma relação entre o intervalo de confiança e o valor de P obtido pelo teste de significância de hipótese nula, em que ambos consideram o mesmo nível de significância α pré-estabelecido.¹⁹⁹

19.6.3 Como interpretar intervalos de confiança em uma figura?

- Identifique o que as tendências centrais e as barras de erro representam. Qual é a variável dependente? É expressa em unidades originais ou é padronizada? A figura mostra intervalos de confiança, erro-padrão ou desvio-padrão? Qual é o desenho experimental?¹⁹⁹
- Faça uma interpretação substantiva dos valores de tendência central e dos intervalos de confiança.¹⁹⁹
- O intervalo de confiança é uma faixa de valores plausíveis para a tendência central. Valores fora do intervalo são relativamente implausíveis, no nível de significância α pré-estabelecido.¹⁹⁹
- Qualquer valor fora do intervalo de confiança, quando considerado como hipótese nula (H_0), equivale a $P < \alpha$ pré-estabelecido (bicaudal).¹⁹⁹
- Qualquer valor dentro do intervalo, quando considerado como hipótese nula (H_0), equivale a $P > \alpha$ pré-estabelecido (bicaudal).¹⁹⁹

19.7 Interpretação de análise inferencial

19.7.1 Como interpretar uma análise inferencial?

- Testes de hipótese nula (H_0) vs. alternativa (H_1) a partir de um nível de significância (α) pré-especificado.²⁰⁰
- P-valor como evidência estatística sobre (H_0).²⁰⁰
- Estimação de intervalos de confiança de um nível de significância (α) pré-especificado bicaudal ($IC_{1-\alpha/2}$) ou unicaudal ($IC_{1-\alpha}$).²⁰⁰
- Análise Bayesiana.²⁰⁰

19.7.2 O que são resultados ‘positivos’ e ‘negativos’ ou inconclusivos em teste de hipótese?

- Resultados ‘positivos’ compreendem um P-valor dentro da zona crítica estatisticamente significativa (ex.: $P < 0,05$ ou outro ponto de corte) e sugerem que os autores rejeitem a hipótese nula H_0 , confirmando assim sua hipótese científica.²⁰¹
- Resultados ‘negativos’ ou inconclusivos compreendem um P-valor fora da zona crítica estatisticamente significativa (ex.: $P \geq 0,05$ ou outro ponto de corte) e sugerem que os autores não rejeitem a hipótese nula H_0 porque o efeito observado é nulo (logo, *negativo*), ou porque o estudo não possui poder suficiente para detectá-lo, não permitindo portanto afirmar a hipótese científica (logo, *inconclusivo*).²⁰¹

19.7.3 Qual a importância de resultados ‘negativos’?

- Conhecer resultados negativos contribui com uma visão mais ampla do campo de estudo junto aos resultados positivos.²⁰²
- Resultados negativos permitem um melhor planejamento das pesquisas futuras e pode aumentar suas chances de sucesso.²⁰²

19.7.4 Resultados inconclusivos: Ausência de evidência ou evidência de ausência?

- Em estudos (geralmente com amostras grandes), resultados estatisticamente significativos (com P-valores menores do limiar pré-estabelecido, $P < \alpha$) podem não ser clinicamente relevantes.²⁰³
- Em estudos (geralmente com amostras pequenas), resultados estatisticamente não significativos (com P-valores iguais ou maiores do limiar pré-estabelecido, $P \geq \alpha$) não devem ser interpretados como evidência de inexistência do efeito.²⁰³
- Geralmente é razoável aceitar uma nova conclusão apenas quando há dados a seu favor (‘resultados positivos’). Também é razoável questionar se apenas a ausência de dados a seu favor (‘resultados negativos’) justifica suficientemente a rejeição de tal conclusão.²⁰³

19.8 Erros de inferência

19.8.1 O que são erros de inferência estatística?

- Um erro de inferência é a tomada de decisão incorreta, seja a favor ou contra a hipótese nula H_0 .¹⁸⁸

19.8.2 O que são erros Tipo I e Tipo II?

- Erro Tipo I significa a rejeição de uma hipótese nula (H_0) quando esta é verdadeira.¹⁸⁸
- Erro Tipo II significa a não rejeição de uma hipótese nula (H_0) quando esta é falsa.¹⁸⁸

19.8.3 O que são erros Tipo S e Tipo M?

- Erro Tipo S (do inglês *sign*) significa a identificação errônea da direção - positiva ou negativa - do efeito observado.^{204,205}

Tabela 19.1: Tabela de erros tipos I e II de inferência estatística.

		Hipótese nula H_0 é falsa	Hipótese nula H_0 é verdadeira
Hipótese nula H_0 foi rejeitada	Decisão correta		Decisão incorreta (erro tipo I)
Hipótese nula H_0 não foi rejeitada	Decisão incorreta (erro tipo II)		Decisão correta

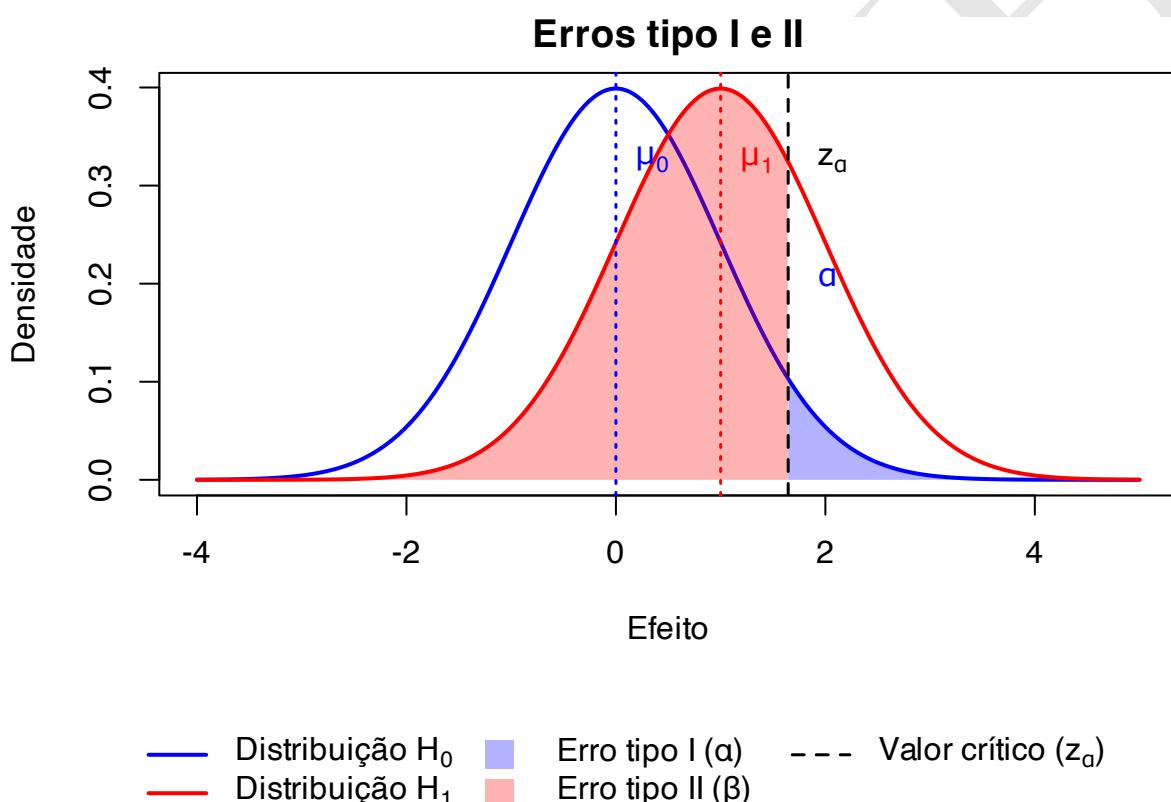


Figura 19.3: Representação gráfica dos erros tipo I e tipo II em um teste de hipótese (bicaudal).

Tabela 19.2: Tabela de erro tipo S de inferência estatística.

		Sinal positivo	Sinal negativo
Sinal positivo	Decisão correta		Decisão incorreta (erro tipo S)
Sinal negativo	Decisão incorreta (erro tipo S)		Decisão correta

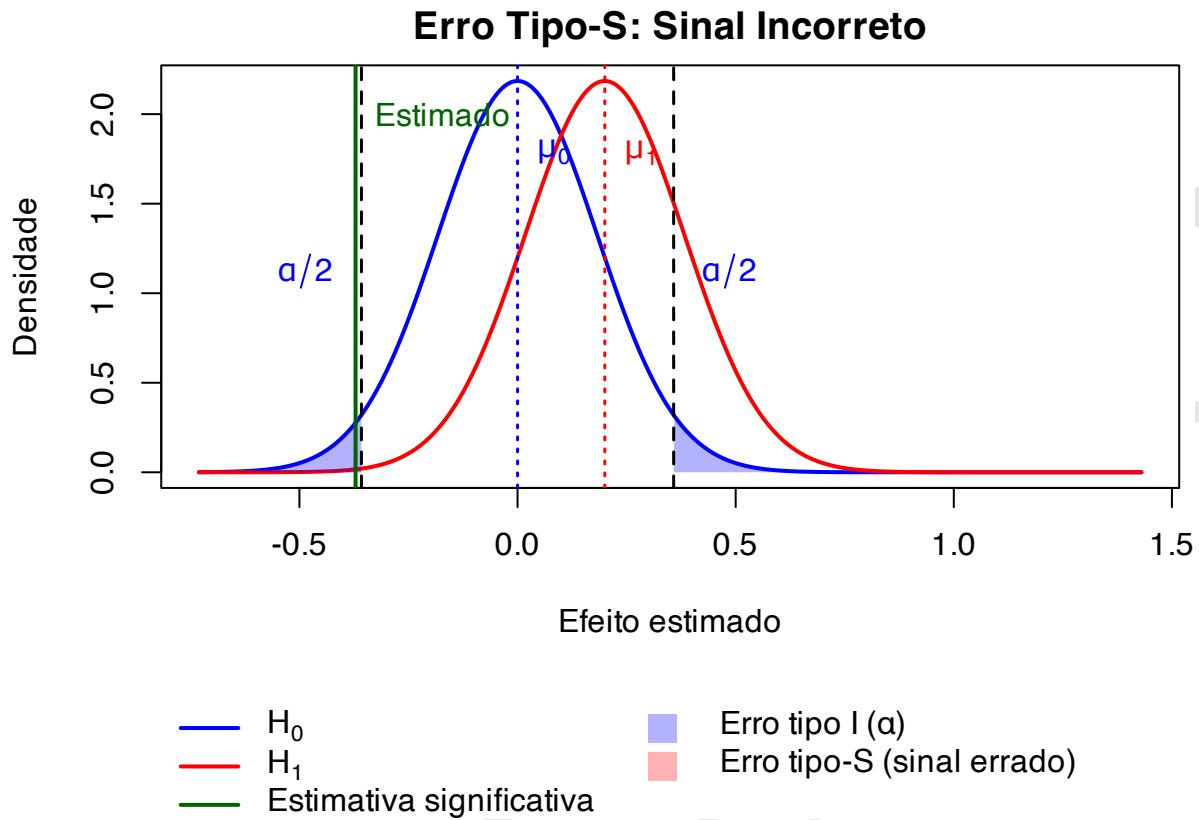


Figura 19.4: Representação gráfica do erro tipo S (sinal) em um teste de hipótese (bicaudal).

Tabela 19.3: Tabela de erro tipo M de inferência estatística.

	Magnitude alta	Magnitude baixa
Magnitude alta	Decisão correta	Decisão incorreta (erro tipo M)
Magnitude baixa	Decisão incorreta (erro tipo M)	Decisão correta

- Erro Tipo M (do inglês *magnitude*) significa a identificação errônea - em geral, exagerada - da magnitude do efeito observado.^{204,205}

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

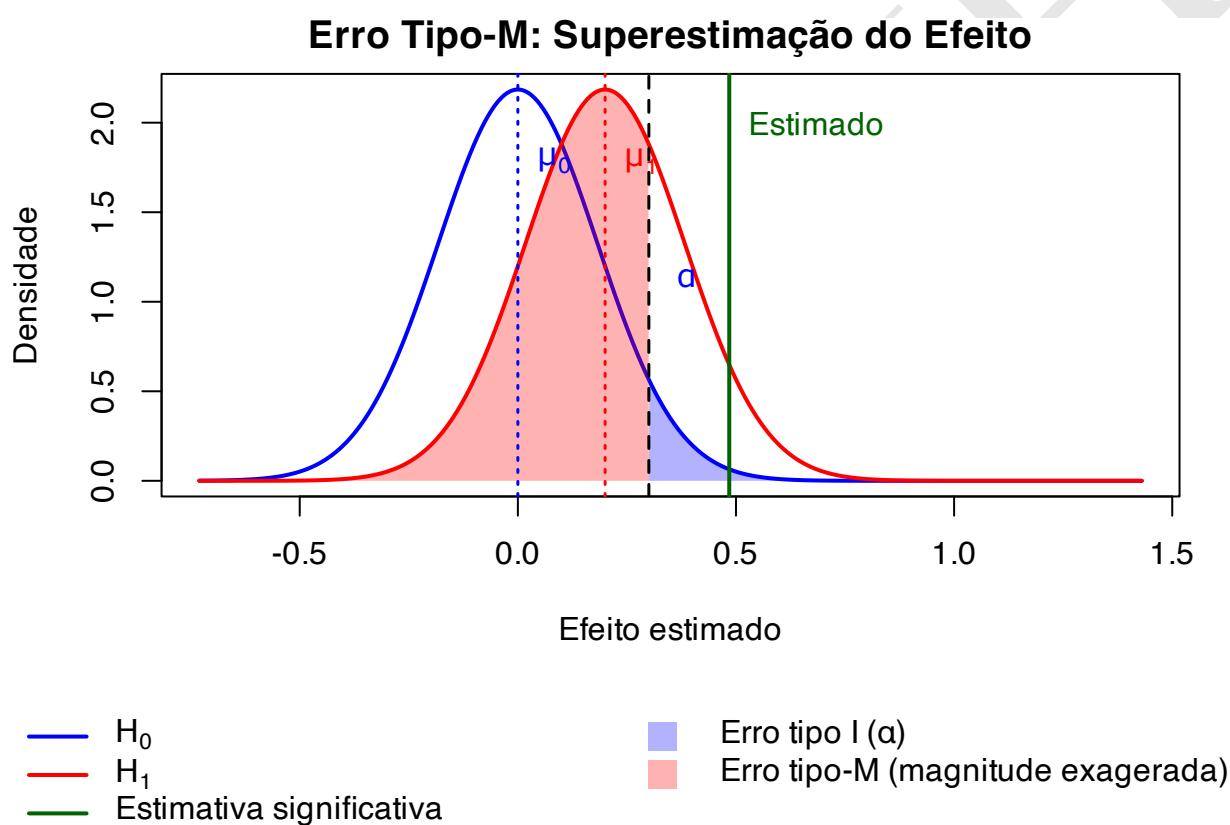


Figura 19.5: Representação gráfica do erro tipo M (magnitude) em um teste de hipótese (bicaudal).

Capítulo 20

Tamanho do efeito e P-valor

20.1 Tamanho do efeito

20.1.1 O que é o tamanho do efeito?

- Tamanho do efeito quantifica a magnitude de um efeito real da análise, expressando uma importância descritiva dos resultados.²⁰⁶

20.1.2 Quais são os tipos de tamanho do efeito?

- Diferenças padronizadas entre grupos:^{193,206}
 - Cohen's d
 - Glass's Δ
 - Razão de chances (RC ou OR)
 - Risco relativo ou razão de risco (RR)
- Medidas de associação:^{193,206}
 - Coeficiente de correlação de Pearson (r), ponto-bisserial (r_s), Spearman (ρ), Kendall (τ), Cramér (V) e ϕ .
 - Coeficiente de determinação (R^2)

20.1.3 Como converter um tamanho de efeito em outro?

- ²⁰⁶

R O pacote *effectsize*²⁰⁷ fornece diversas funções para conversão de diferentes estimativas de tamanhos de efeito.

20.1.4 Como interpretar um tamanho do efeito?

- Tamanhos de efeito podem ser comparadores entre diferentes estudos.¹⁹³

R O pacote *effectsize*²⁰⁷ fornece a função *rules*^a para criar regras de interpretação de tamanhos de efeito.

^a<https://www.rdocumentation.org/packages/effectsize/versions/0.8.3/topics/rules>

R O pacote *effectsize*²⁰⁷ fornece a função *interpret*^a para interpretar os tamanhos de efeito com base em uma lista de regras pré-definidas.

^a<https://www.rdocumentation.org/packages/effectsize/versions/0.8.3/topics/interpret>

R O pacote *pwr*¹⁹⁵ fornece a função *cohen.ES*^a para obter os tamanhos de efeito “pequeno”, “médio” e “grande” para diversos testes de hipóteses.

^a<https://www.rdocumentation.org/packages/pwr/versions/1.3-0/topics/cohen.ES>

20.2 Efeito fixo

20.2.1 O que é efeito fixo?

- ²⁰⁸

20.3 Efeito aleatório

20.3.1 O que é efeito aleatório?

- ²⁰⁸

20.4 Efeito principal

20.4.1 O que é efeito principal?

- ²⁰⁸

20.5 Efeito de modificação

20.5.1 O que é um modificador de efeito?

- ²⁰⁸

20.5.2 O que é efeito de modificação?

- ²⁰⁸

20.6 Efeito de interação

20.6.1 O que é efeito de interação?

- A interação - representada pelo símbolo * - é o termo estatístico empregado para representar a heterogeneidade de um determinado efeito.²⁰⁹
- ²⁰⁸

R O pacote *nlme*²¹⁰ fornece a função *nlme*^a para ajustar um modelo de regressão misto não linear.

^a<https://www.rdocumentation.org/packages/nlme/versions/3.1-163/topics/nlme>

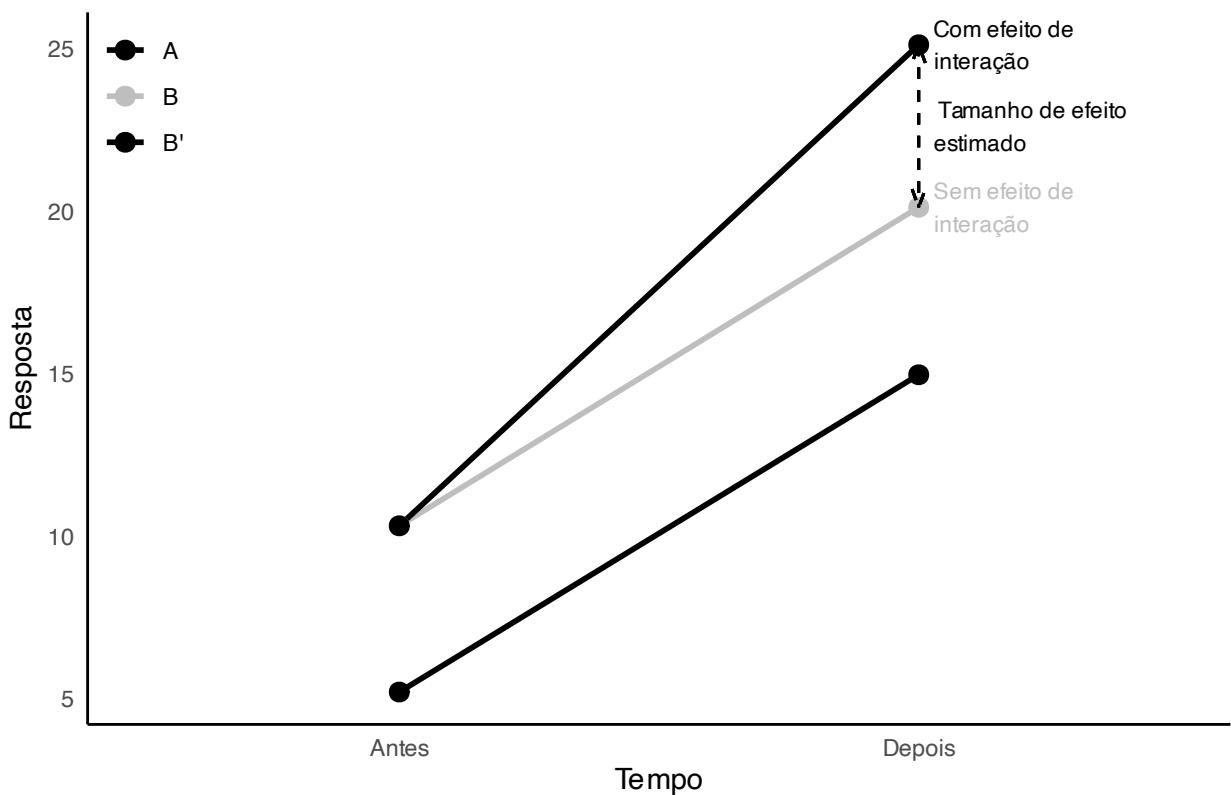


Figura 20.1: Análise de efeito de interação (direta) entre grupos e tempo. Retas paralelas sugerem ausência de efeito de interação.

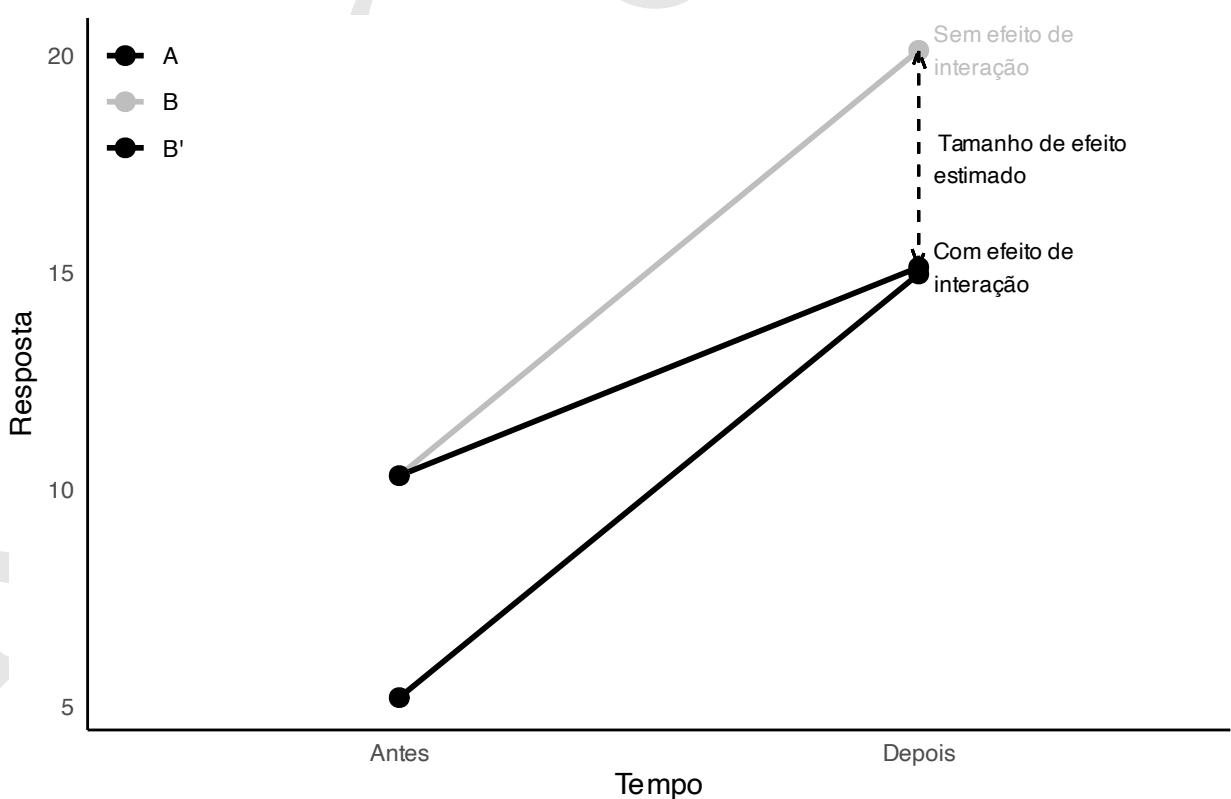


Figura 20.2: Análise de efeito de interação (inversa) entre grupos e tempo. Retas paralelas sugerem ausência de efeito de interação.

R

O pacote *mmrm*²¹¹ fornece a função *mmrm*^a para ajuste de um modelo de regressão misto linear.

^a<https://rdrr.io/cran/mmrm/man/mmrm.html>

R

O pacote *emmeans*²¹² fornece a função *emmeans*^a para calcular as médias marginais dos fatores e suas combinações de um modelo de regressão misto linear.

^a<https://www.rdocumentation.org/packages/emmeans/versions/1.8.7/topics/emmeans>

20.7 Efeito de mediação

20.7.1 O que é um mediador de efeito?

- 213
- 208

20.7.2 O que é efeito de mediação?

- 213
- 208

20.7.3 O que é efeito direto?

- 213
- 208

20.7.4 O que é efeito indireto?

- 213
- 208

20.7.5 O que é efeito total?

- 213
- 208

20.8 Efeitos brutos e padronizados

20.8.1 O que é efeito bruto?

- 214
- 215

20.8.2 O que é efeito padronizado?

- 214
- 215

20.9 P-valor

20.9.1 O que é significância estatística?

- A expressão “significância estatística”²¹⁶ ou “evidência estatística de significância” sugere apenas que um experimento merece ser repetido, uma vez que um baixo P-valor (calculado a partir dos dados, modelos e demais suposições do estudo) sugere ser improvável que os dados coletados sejam coletados no contexto de que a hipótese nula H_0 assumida é verdadeira.²¹⁷

20.9.2 Como justificar o nível de significância estatística de um teste?

- ?

R O pacote *Superpower*¹⁹⁷ fornece a função *optimal_alpha*^a para calcular e justificar o nível de significância α por balanço dos erros tipo I e II.

^ahttps://www.rdocumentation.org/packages/Superpower/versions/0.2.0/topics/optimal_alpha

R O pacote *Superpower*¹⁹⁷ fornece a função *ANOVA_compromise*^a para calcular e justificar o nível de significância α por balanço dos erros tipo I e II em análise de variância (ANOVA).

^ahttps://www.rdocumentation.org/packages/Superpower/versions/0.2.0/topics/ANOVA_compromise

20.9.3 O que é o P-valor?

- P-valor é a probabilidade, assumindo-se um dado modelo estatístico, de que um efeito calculado a partir dos dados seria igual ou mais extremo do que o seu valor observado.²¹⁸
- P-valor é uma variável aleatória que possui distribuição uniforme quando a hipótese nula H_0 é verdadeira.²¹⁹

20.9.4 Como interpretar o P-valor?

- P-valores abaixo de um nível de significância estatística pré-especificado representam que um experimento merece ser repetido, com a rejeição da hipótese nula H_0 justificada apenas quando experimentos adicionais frequentemente reportem igualmente resultados positivos (rejeição da hipótese nula H_0).²⁰⁰
- P-valor resulta da coleta e análise de dados, e assim quantifica a plausibilidade dos dados observados sob a hipótese nula H_0 .²²⁰
- P-valores podem indicar quantitativamente a incompatibilidade entre os dados obtidos e o modelo estatístico especificado a priori (geralmente constituído pela hipótese nula H_0).²¹⁸
- P-valores menores/maiores do que o nível de significância estatístico pré-estabelecido não devem ser utilizados como única fonte de informação para tomada de decisão em ciência.²¹⁸

20.9.5 O que o P-valor não é?

- P-valor não representa a probabilidade de que a hipótese nula H_0 seja verdadeira, nem a probabilidade de que os dados tenham sido produzidos pelo acaso.²¹⁸
- P-valor não mede o tamanho do efeito ou a relevância da sua observação.²¹⁸
- P-valor sozinho não provê informação suficiente sobre a evidência sobre um modelo teórico. A sua interpretação correta requer uma descrição ampla sobre o delineamento, métodos e análises estatísticas aplicados no estudo.²¹⁸
- Evidência estatística de significância não provê informação sobre a magnitude do efeito observado e não necessariamente implica que o efeito é robusto.^{134,219}

20.9.6 Qual a origem do ‘P<0,05’?

- ?

20.9.7 Quais são os complementos ou alternativas ao P-valor?

- Intervalos de confiança, credibilidade ou predição.²¹⁸
- Razão de verossimilhança.²¹⁸
- Métodos Bayesianos, fator Bayes.²¹⁸

20.10 P-hacking

20.10.1 O que é P-hacking?

- ?

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

Capítulo 21

Seleção de testes

21.1 Multiverso de análises estatísticas

21.1.1 Por que escolher o teste é um problema?

- Analisar a mesma hipótese com o mesmo banco de dados pode resultar em diferenças substanciais nas estimativas estatísticas e nas conclusões.²²¹
- As decisões para especificação das análises estatísticas podem ser tão minuciosas que muitas vezes nem sequer são registadas como decisões e, assim, podem impactar na reproduzibilidade do estudo.²²¹

21.2 Escolha de testes para análise inferencial

21.2.1 Como selecionar os testes para a análise estatística inferencial?

- ²²²
- ²²³
- ²²⁴
- ²²⁵
- ²²⁶
- ²²⁷
- ²²⁸
- ²²⁹

RAASCUNHO

Capítulo 22

Testes estatísticos

22.1 Testes de Qui-quadrado (χ^2)

```
# carrega os pacotes
library("dplyr")
library("gtsummary")

# tabela 2x2
tbl_cross <-
  # banco de dados
  trial %>%
  # cria a tabela de contingência
  gtsummary::tbl_cross(
    row = trt,
    col = response,
    statistic = "{n}",
    digits = 0,
    percent = "cell",
    margin = c("row", "column"),
    missing = "no",
    missing_text = "Dados perdidos",
    margin_text = "Total"
  ) %>%
  # calcula o p-valor do teste
  gtsummary::add_p(
    test = "chisq.test",
    pvalue_fun = function(x) style_pvalue(x, digits = 3)
  ) %>%
  gtsummary::modify_header(
    p.value = "***P-valor***"
  ) %>%
  # calcula o tamanho do efeito
  gtsummary::modify_table_styling(
    rows = NULL,
    footnote = as.character(rstatix::cramer_v(trt, response))
  ) %>%
  # formata o título em negrito
  gtsummary::bold_labels() %>%
  # cria título da tabela
  gtsummary::modify_caption(
    "Teste Qui-quadrado (com correção de Yates)"
  )
```

```
# exibe a tabela
require(huxtable)
tbl_cross %>%
  gtsummary::as_hux_table()
```

Tabela 22.1: Teste Qui-quadrado (com correção de Yates)

		Tumor Response		P-valor
		0	1	
Chemotherapy Treatment				P-valor
Drug A		67	28	95
Drug B		65	33	98
Total		132	61	193

```
# carrega os pacotes
library("dplyr")
library("gtsummary")

# tabela 2x2
tbl_cross <-
  # banco de dados
  trial %>%
  # cria a tabela de contingência
  gtsummary::tbl_cross(
    row = trt,
    col = response,
    statistic = "{n}",
    digits = 0,
    percent = "cell",
    margin = c("row", "column"),
    missing = "no",
    missing_text = "Dados perdidos",
    margin_text = "Total"
  ) %>%
  # calcula o p-valor do teste
  gtsummary::add_p(
    test = "chisq.test.no.correct",
    pvalue_fun = function(x) style_pvalue(x, digits = 3)
  ) %>%
  gtsummary::modify_header(
    p.value = "***P-valor***"
  ) %>%
  # calcula o tamanho do efeito
  gtsummary::modify_table_styling(
    rows = NULL,
    footnote = as.character(rstatix::cramer_v(trt, response))
  ) %>%
  # formata o título em negrito
  gtsummary::bold_labels() %>%
  # cria título da tabela
  gtsummary::modify_caption(
    "Teste Qui-quadrado (sem correção de Yates)"
  )
```

```
# exibe a tabela
require(huxtable)
tbl_cross %>%
  gtsummary::as_hux_table()
```

Tabela 22.2: Teste Qui-quadrado (sem correção de Yates)

		Tumor Response		P-valor
		0	1	
Chemotherapy Treatment				P-valor
		0	1	
Drug A		67	28	95
Drug B		65	33	98
Total		132	61	193

22.2 Teste exato de Fisher

```
# carrega os pacotes
library("dplyr")
library("gtsummary")

# tabela 2x2
tbl_cross <-
  # banco de dados
  trial %>%
  # cria a tabela de contingência
  gtsummary::tbl_cross(
    row = trt,
    col = response,
    statistic = "{n}",
    digits = 0,
    percent = "cell",
    margin = c("row", "column"),
    missing = "no",
    missing_text = "Dados perdidos",
    margin_text = "Total"
  ) %>%
  # calcula o p-valor do teste
  gtsummary::add_p(
    test = "fisher.test",
    pvalue_fun = function(x) style_pvalue(x, digits = 3)
  ) %>%
  gtsummary::modify_header(
    p.value = "***P-valor***"
  ) %>%
  # calcula o tamanho do efeito
  gtsummary::modify_table_styling(
    rows = NULL,
    footnote = as.character(rstatix::cramer_v(trt, response))
  ) %>%
  # formata o título em negrito
  gtsummary::bold_labels() %>%
```

```
# cria título da tabela
gtsummary::modify_caption(
  "Teste exato de Fisher"
)

# exibe a tabela
require(huxtable)
tbl_cross %>%
  gtsummary::as_hux_table()
```

Tabela 22.3: Teste exato de Fisher

		Tumor Response		P-valor
		0	1	
		Chemotherapy Treatment		
	Drug A	67	28	95
	Drug B	65	33	98
	Total	132	61	193

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

Capítulo 23

Comparação

23.1 Análise inferencial de comparação

23.1.1 O que é análise de comparação de dados?

- ?



O pacote *cocor*²³⁰ fornece as funções *cocor.indep.groups*^a, *cocor.dep.groups.overlap*^b e *cocor.dep.groups.nonoverlap*^c para comparar 2 coeficientes de correlação entre grupos independentes, grupos sobrepostos ou independentes, respectivamente.²³⁰

^a<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/cor.test>

^b<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/cor.test>

^c<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/cor.test>

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

RASCUNHO

Capítulo 24

Associação

24.1 Análise inferencial de associação

24.1.1 O que é análise de associação?

- ?

24.2 Associação bivariada

24.2.1 O que são análises de associação bivariada?

- ?

24.2.2 Quais testes podem ser usados para análises de associação bivariada?

- Teste Qui-quadrado (χ^2).^{231,232}
 - O teste qui-quadrado (χ^2) avalia uma hipótese global se a relação entre duas variáveis e/ou fatores é independente ou associada.²³²
 - O teste qui-quadrado é utilizado para comparar a distribuição de uma variável categórica em uma amostra ou grupo com a distribuição em outro. Se a distribuição da variável categórica não for muito diferente nos diferentes grupos, pode-se concluir que a distribuição da variável categórica não está relacionada com a variável dos grupos. Pode-se também concluir que a variável categórica e os grupos são independentes.²³²
 - Tipo: não paramétrico.^{231,232}
 - Suposições:^{231,232}
 - * As variáveis são ordinais ou categóricas nominais, de modo que as células representem frequência.
 - * Os níveis dos fatores (variáveis categóricas) são mutuamente exclusivos.
 - * Tamanho de amostra grande e adequado porque é baseado em uma abordagem de aproximação.
 - * Menos de 20% das células com frequências esperadas < 5
 - * Nenhuma célula com frequência esperada < 1 .
 - Hipóteses:²³²
 - * Nula (H_0): independente (sem associação)
 - * Alternativa (H_1): não independente (associação)
 - Tamanho do efeito:²³²
 - * Phi (ϕ), para tabelas de contingência 2x2

- * Razão de chances (*RC* ou *OR*), para tabelas de contingência 2x2
- * Cramer V (*V*), para tabelas de contingência NxM

 O pacote *gtsummary*¹⁷³ fornece a função *tbl_cross*^a para criar uma tabela NxM.

^ahttps://www.rdocumentation.org/packages/gtsummary/versions/1.6.3/topics/tbl_cross

- Teste Exato de Fisher (χ^2).^{231,232}
 - O teste exato de Fisher avalia a hipótese nula de independência aplicando a distribuição hipergeométrica dos números nas células da tabela.²³²
 - Hipóteses:^{231,232}
 - * Nula (H_0): independente (sem associação)
 - * Alternativa (H_1): não independente (associação)
 - Tamanho do efeito:^{231,232}
 - * Phi (ϕ), para tabelas de contingência 2x2
 - * Razão de chances (*RC* ou *OR*), para tabelas de contingência 2x2
 - * Cramer V (*V*), para tabelas de contingência NxM

 O pacote *gtsummary*¹⁷³ fornece a função *tbl_cross*^a para criar uma tabela NxM.

^ahttps://www.rdocumentation.org/packages/gtsummary/versions/1.6.3/topics/tbl_cross

24.3 Associação multivariada

24.3.1 O que são análises de associação multivariada?

- ?

24.3.2 Quais testes podem ser usados para análises de associação multivariada?

- ?

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

Capítulo 25

Correlação

25.1 Análise inferencial de correlação

25.1.1 O que é covariância?

- ?

25.1.2 O que é correlação?

- ?

25.1.3 Qual é a interpretação das medidas de correlação?

- Os valores de correlação estão no intervalo $[-1; 1]$.^{91,233,234}
- Valores de correlação positivos representam uma relação direta entre as variáveis, tal que valores maiores de uma variável estão associados a valores maiores de outra variável.^{233,234}
- Valores de correlação negativos representam uma relação indireta (ou inversa) entre as variáveis, tal que valores maiores (menores) de uma variável estão associados a valores maiores (menores) de outra variável.^{233,234}
- Valores de correlação próximos de 0 representam a inexistência de relação entre as variáveis.^{233,234}

25.1.4 Quais precauções devem ser tomadas na interpretação de medidas de correlação?

- Tamanhos de efeito grande (ou qualquer outro) não representam necessariamente uma relação causa-efeito entre as variáveis.²³³
- Tamanhos de efeito grande (ou qualquer outro) não representam necessariamente uma relação de concordância ou confiabilidade entre as variáveis.²³³
- Uma escala de medição com representação agregada do constructo na coleta de dados pode subestimar o tamanho do efeito da correlação r em de cerca de 13% e do coeficiente de determinação R^2 de cerca de 30%.⁶⁸ Neste caso, a correlação desatenuada $r_{x'y'}$ pode ser calculada pela equação (25.1), utilizando a correlação observada r_{xy} e os fatores de correção $r_{xx'}$ e $r_{yy'}$ para o número de intervalos nas variáveis X e Y, respectivamente:⁶⁸

$$r_{x'y'} = \frac{r_{xy}}{r_{xx'} r_{yy'}} \quad (25.1)$$



O pacote *psychmeta*²³⁵ fornece a função *correct_r_coarseness*^a para calcular o coeficiente de correlação desatenuado ($r_{x'y'}$).

^ahttps://www.rdocumentation.org/packages/psychmeta/versions/2.7.0/topics/correct_r_coarseness

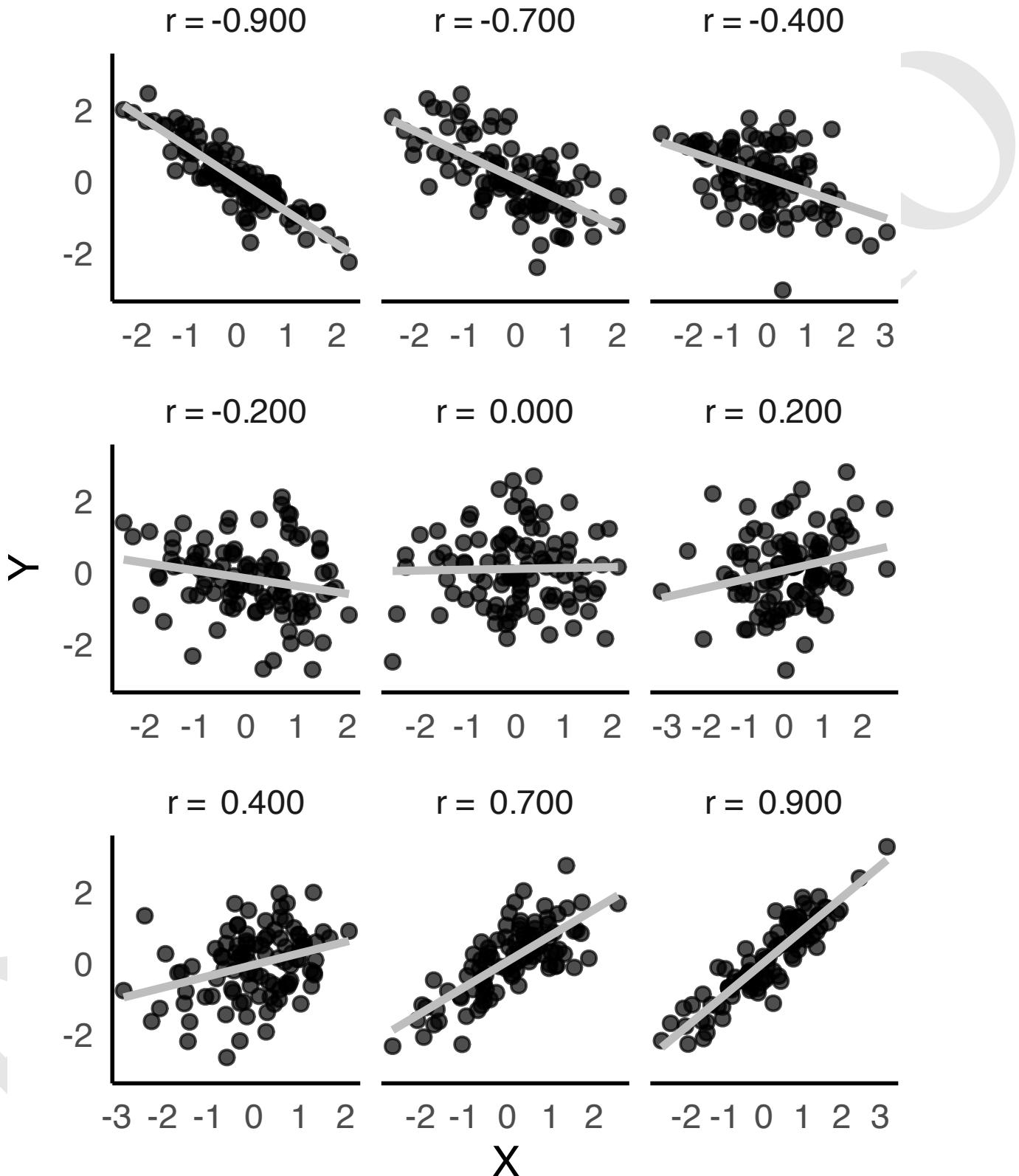


Figura 25.1: Exemplo de diferentes forças e direção de correlação entre duas variáveis X e Y.

Tabela 25.1: Quarteto de Anscombe.

ID	x1	x2	x3	x4	y1	y2	y3	y4
1	10	10	10	8	8.04	9.14	7.46	6.58
2	8	8	8	8	6.95	8.14	6.77	5.76
3	13	13	13	8	7.58	8.74	12.74	7.71
4	9	9	9	8	8.81	8.77	7.11	8.84
5	11	11	11	8	8.33	9.26	7.81	8.47
6	14	14	14	8	9.96	8.10	8.84	7.04
7	6	6	6	8	7.24	6.13	6.08	5.25
8	4	4	4	19	4.26	3.10	5.39	12.50
9	12	12	12	8	10.84	9.13	8.15	5.56
10	7	7	7	8	4.82	7.26	6.42	7.91
11	5	5	5	8	5.68	4.74	5.73	6.89

Tabela 25.2: Análise descritiva do Quarteto de Anscombe demonstrando os conjuntos de dados bivariados com parâmetros quase idênticos.

	X1Y1	X2Y2	X3Y3	X4Y4
Observações	11.00	11.00	11.00	11.00
Média x	9.00	9.00	9.00	9.00
Média y	7.50	7.50	7.50	7.50
Variância x	11.00	11.00	11.00	11.00
Variância y	4.13	4.13	4.12	4.12
Correlação	0.82	0.82	0.82	0.82
Coeficiente angular	0.50	0.50	0.50	0.50
Coeficiente linear	3.00	3.00	3.00	3.00
Coeficiente de determinação	0.67	0.67	0.67	0.67

R

O pacote *psychmeta*²³⁵ fornece a função *correct_r*^a para calcular o coeficiente de correlação em escala restrita e/ou com erro de mensuração ($r_{x'y'}$).

^ahttps://www.rdocumentation.org/packages/psychmeta/versions/2.7.0/topics/correct_r

- Os coeficientes de correlação possuem suposições que, se violadas, podem levar a interpretações equivocadas. Nesses cenários, visualizar os dados e as relações entre as variáveis pode contribuir com a interpretação e utilidade dos coeficientes de correlação.²³⁶
- O *quarteto de Anscombe* é um conjunto de quatro bancos de dados bivariados que possuem a mesma média, variância, correlação e regressão linear (até a 2a casa decimal), mas que são visualmente diferentes e, assim, demonstram a importância da análise gráfica da correlação.²³⁶

R

O pacote *anscombiser*²³⁷ fornece a função *anscombise*^a para gerar bancos de dados que compartilham os mesmos valores de parâmetros do Quarteto de Anscombe.

^a<https://www.rdocumentation.org/packages/anscombiser/versions/1.1.0/topics/anscombise>

25.2 Coeficientes de correlação

25.2.1 Quais coeficientes podem ser usados em análises de correlação?

- Coeficiente de correlação de Pearson (r).^{233,234}
 - O coeficiente de correlação de Pearson (r) avalia a força e direção da relação linear entre duas variáveis quantitativas.^{233,234}

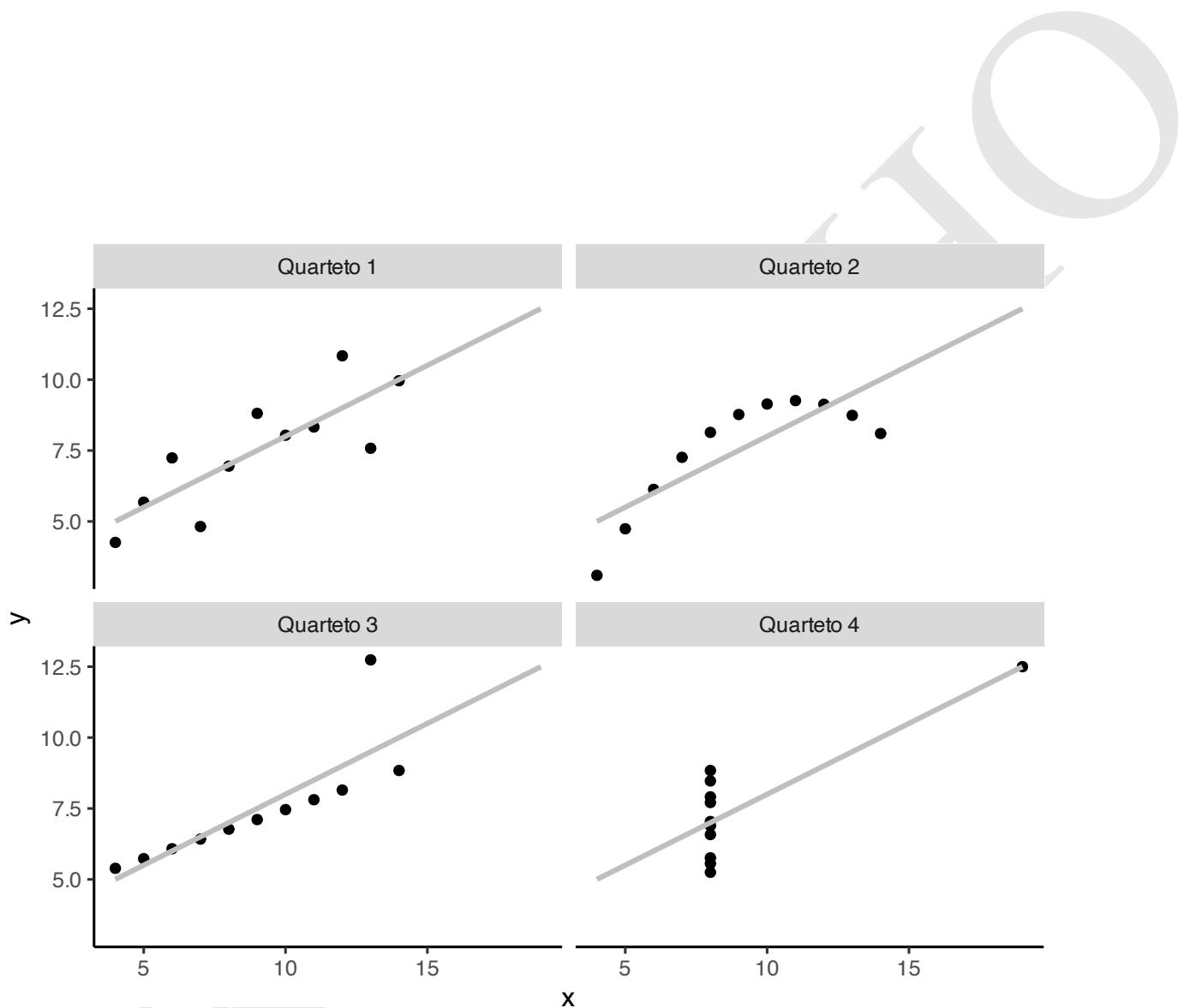


Figura 25.2: Gráfico de dispersão do Quarteto de Anscombe para representação gráfica de conjuntos de dados bivariados com parâmetros quase idênticos e relações muito distintas.

- Tipo: paramétrico.^{233,234}
- Hipóteses:²³⁴
 - * Nula (H_0): $r = 0$
 - * Alternativa (H_1): $r \neq 0$
- Tamanho do efeito:^{233,234}
 - * Coeficiente de correlação de Pearson (r)

R

O pacote *stats*⁷¹ fornece a função *cor.test*^a para calcular o coeficiente de correlação de Pearson (r).

^a<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/cor.test>

R

O pacote *correlation*²³⁸ do projeto *easystats*²³⁹ fornece a função *correlation*^a para calcular o coeficiente de correlação de Pearson (r).

^a<https://cloud.r-project.org/web/packages/correlation/index.html>

- Coeficiente de correlação ponto-bisserial (r_s).²³³
 - O coeficiente de correlação ponto-bisserial (r_s) avalia a força e direção da relação linear entre uma variável quantitativa e outra dicotômica.²³³
 - Tipo: paramétrico.²³³
 - Hipóteses:²³³
 - * Nula (H_0): $r_s = 0$
 - * Alternativa (H_1): $r_s \neq 0$
 - Tamanho do efeito:²³³
 - * Coeficiente de correlação ponto-bisserial (r_s)

R

O pacote *stats*⁷¹ fornece a função *cor.test*^a para calcular o coeficiente de correlação ponto-bisserial (r_s).

^a<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/cor.test>

R

O pacote *correlation*²³⁸ do projeto *easystats*²³⁹ fornece a função *correlation*^a para calcular o coeficiente de correlação ponto-bisserial (r_s).

^a<https://cloud.r-project.org/web/packages/correlation/index.html>

- Coeficiente de correlação de Spearman (ρ).^{233,234}
 - O coeficiente de correlação de Spearman (ρ) avalia a força e direção da relação monotônica entre duas variáveis quantitativas.^{233,234}
 - O coeficiente de correlação de Spearman (ρ) pode ser também definida como a correlação de Pearson (r) entre as classificações (*ranks*) das duas variáveis quantitativas.^{233,234}
 - Tipo: não-paramétrico.^{233,234}
 - Hipóteses:^{233,234}
 - * Nula (H_0): $\rho = 0$
 - * Alternativa (H_1): $\rho \neq 0$
 - Tamanho do efeito:^{233,234}

- * Coeficiente de correlação de Spearman (ρ)

R O pacote *stats*⁷¹ fornece a função *cor.test*^a para calcular o coeficiente de correlação de Spearman (ρ).

^a<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/cor.test>

R O pacote *correlation*²³⁸ do projeto *easystats*²³⁹ fornece a função *correlation*^a para calcular o coeficiente de correlação de Spearman (ρ).

^a<https://cloud.r-project.org/web/packages/correlation/index.html>

- Coeficiente de Kendall (τ).^{233,234}

- O coeficiente Kendall τ avalia a força e direção da relação monotônica entre duas variáveis quantitativas ou qualitativas.^{233,234}
- O coeficiente Kendall τ é definido como a proporção de todos os pares concordantes menos a proporção de todos os pares discordantes.^{233,234}
- Tipo: não-paramétrico.^{233,234}
- Hipóteses:^{233,234}
 - * Nula (H_0): $\tau = 0$
 - * Alternativa (H_1): $\tau \neq 0$
- Tamanho do efeito:^{233,234}
 - * Kendall τ

R O pacote *stats*⁷¹ fornece a função *cor.test*^a para calcular o coeficiente Kendall τ .

^a<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/cor.test>

R O pacote *correlation*²³⁸ do projeto *easystats*²³⁹ fornece a função *correlation*^a para calcular o coeficiente coeficiente Kendall τ .

^a<https://cloud.r-project.org/web/packages/correlation/index.html>

- Coeficiente de Cramér (V).?
 - O coeficiente Cramér (V) avalia a força e direção da relação entre duas variáveis qualitativas.?
 - Tipo: não-paramétrico.?
 - Hipóteses:?
 - * Nula (H_0): $V = 0$
 - * Alternativa (H_1): $V \neq 0$
 - Tamanho do efeito:?
 - * Coeficiente Cramer (V)
- Coeficiente de Sheperd ϕ .?
 - O coeficiente Phi (ϕ) avalia a força e direção da relação entre duas variáveis dicotômicas.?
 - Tipo: não-paramétrico.?
 - Hipóteses:?

- * Nula (H_0): $\phi = 0$
- * Alternativa (H_1): $\phi \neq 0$
- Tamanho do efeito:²³⁸
- * Coeficiente Phi (ϕ)

R

O pacote *correlation*²³⁸ do projeto *easystats*²³⁹ fornece a função *correlation*^a para calcular o coeficiente coeficiente Sheperd ϕ .

^a<https://cloud.r-project.org/web/packages/correlation/index.html>

R

O pacote *corrplot*¹⁸² fornece a função *cor.mtest*^a para calcular os P-valores e intervalos de confiança da matriz de correlação.

^a<https://www.rdocumentation.org/packages/corrplot/versions/0.92/topics/cor.mtest>

R

O pacote *corrplot*¹⁸² fornece a função *corrplot*^a para visualização da matriz de correlação.

^a<https://www.rdocumentation.org/packages/corrplot/versions/0.92/topics/corrplot>

25.3 Colinearidade

25.3.1 O que é colinearidade?

- Colinearidade representa a correlação entre duas variáveis.²⁴⁰
- Colinearidade exata indica uma relação linear perfeita entre duas variáveis.²⁴⁰

25.3.2 Como identificar colinearidade na matriz de correlação?

- A colinearidade pode ser identificada na matriz de correlação por meio da análise dos coeficientes de correlação entre as variáveis.²⁴⁰
- Valores de correlação próximos de 1 ou -1 indicam colinearidade entre as variáveis.²⁴⁰

R

O pacote *GGally*²⁴¹ fornece a função *ggally_cor*^a para estimar a correlação bivariada e exibir o coeficiente de correlação e o P-valor na matriz de correlação.²⁴¹

^ahttps://www.rdocumentation.org/packages/GGally/versions/2.2.1/topics/ggally_cor

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

RAASCUNHO

Capítulo 26

Regressão

26.1 Análise de regressão

26.1.1 O que é regressão?

- Regressão refere-se a uma equação matemática que permite que uma ou mais variável(is) de desfecho (dependentes) seja(m) prevista(s) a partir de uma ou mais variável(is) independente(s). A regressão implica em uma direção de efeito, mas não garante causalidade.²⁰¹
- Para estimar os efeitos imparciais de um fator de exposição primária sobre uma variável de desfecho, frequentemente constroem-se modelos estatísticos de regressão.¹⁷⁸

R

O pacote *modelsummary*²⁴² fornece as funções *modelsummary*^a e *modelplot*^b para gerar tabelas e gráficos de coeficientes de regressão.

^a<https://www.rdocumentation.org/packages/modelsummary/versions/1.4.1/topics/modelsummary>

^b<https://www.rdocumentation.org/packages/modelsummary/versions/1.4.1/topics/modelplot>

R

O pacote *gtsummary*¹⁷³ fornece a função *tbl_regression*^a para construção da ‘Tabela 2’ com dados do modelo de regressão.

^ahttps://www.rdocumentation.org/packages/gtsummary/versions/1.6.3/topics/tbl_regression

26.1.2 Quais são os algoritmos de regressão?

- Linear.?
- Não-linear.?
- Polinomial.?
- Ridge.?
- Lasso.?

26.1.3 O que são análises de regressão simples?

- A análise de regressão simples consiste em modelos estatísticos com 1 variável dependente (desfecho) e 1 variável independente (preditor).²⁴³
- A equação de regressão simples é expressa como (26.1), onde Y é a variável dependente, X é a variável independente, β_0 é o intercepto (constante), β_1 é o coeficiente de regressão da variável independente e ϵ representa o erro aleatório do modelo.²⁴³

$$Y = \beta_0 + \beta_1 X + \epsilon \quad (26.1)$$

26.1.4 O que são análises de regressão multivariável?

- A análise multivariável (ou múltiplo) consiste em modelos estatísticos com 1 variável dependente (desfecho) e duas ou mais variáveis independentes.²⁴³
- A equação de regressão multivariável é expressa como (26.2), onde Y é a variável dependente, X_1, X_2, \dots, X_n são as variáveis independentes, β_0 é o intercepto (constante), $\beta_1, \beta_2, \dots, \beta_n$ são os coeficientes de regressão das variáveis independentes e ϵ representa o erro aleatório do modelo.²⁴³

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \epsilon \quad (26.2)$$

26.1.5 O que são análises de regressão multivariada?

- A análise multivariada consiste em modelos estatísticos com 2 ou mais variáveis dependentes (desfechos) e duas ou mais variáveis independentes.²⁴³
- A equação de regressão multivariada é expressa como (26.3), onde Y_1, Y_2, \dots, Y_m são as variáveis dependentes, X_1, X_2, \dots, X_n são as variáveis independentes, β_{0j} é o intercepto (constante) da variável dependente Y_j , β_{ij} são os coeficientes de regressão das variáveis independentes para a variável dependente Y_j e ϵ_j representa o erro aleatório do modelo para a variável dependente Y_j .²⁴³

$$Y_1 = \beta_{01} + \beta_{11} X_1 + \beta_{12} X_2 + \dots + \beta_{1n} X_n + \epsilon_1 \quad (26.3)$$

$$Y_2 = \beta_{02} + \beta_{21} X_1 + \beta_{22} X_2 + \dots + \beta_{2n} X_n + \epsilon_2 \quad (26.4)$$

$$\vdots \quad (26.5)$$

$$Y_m = \beta_{0m} + \beta_{m1} X_1 + \beta_{m2} X_2 + \dots + \beta_{mn} X_n + \epsilon_m \quad (26.6)$$

26.1.6 O que são análises de regressão linear?

- ?

26.1.7 O que são análises de regressão não-linear?

- ?

26.1.8 O que são análises de regressão polinomial?

- ?

26.1.9 O que são análises de regressão ridge?

- ?

26.1.10 O que são análises de regressão logística?

- ?

26.2 Preparação de variáveis para regressão

26.2.1 Como preparar as variáveis categóricas para análise de regressão?

- Variáveis fictícias (*dummy*) compreendem variáveis criadas para introduzir, nos modelos de regressão, informações contidas em outras variáveis que não podem ser medidas em escala numérica.²⁴⁴

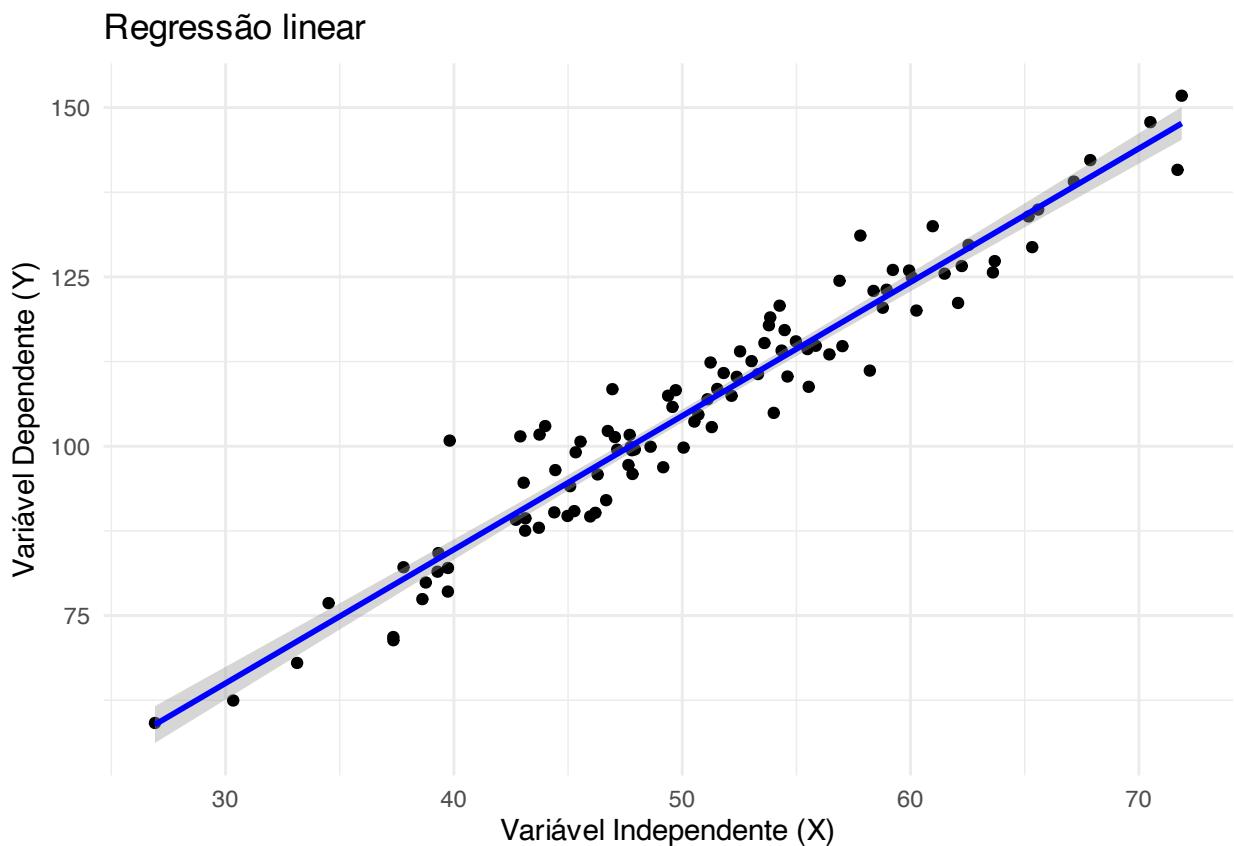


Figura 26.1: Regressão linear.

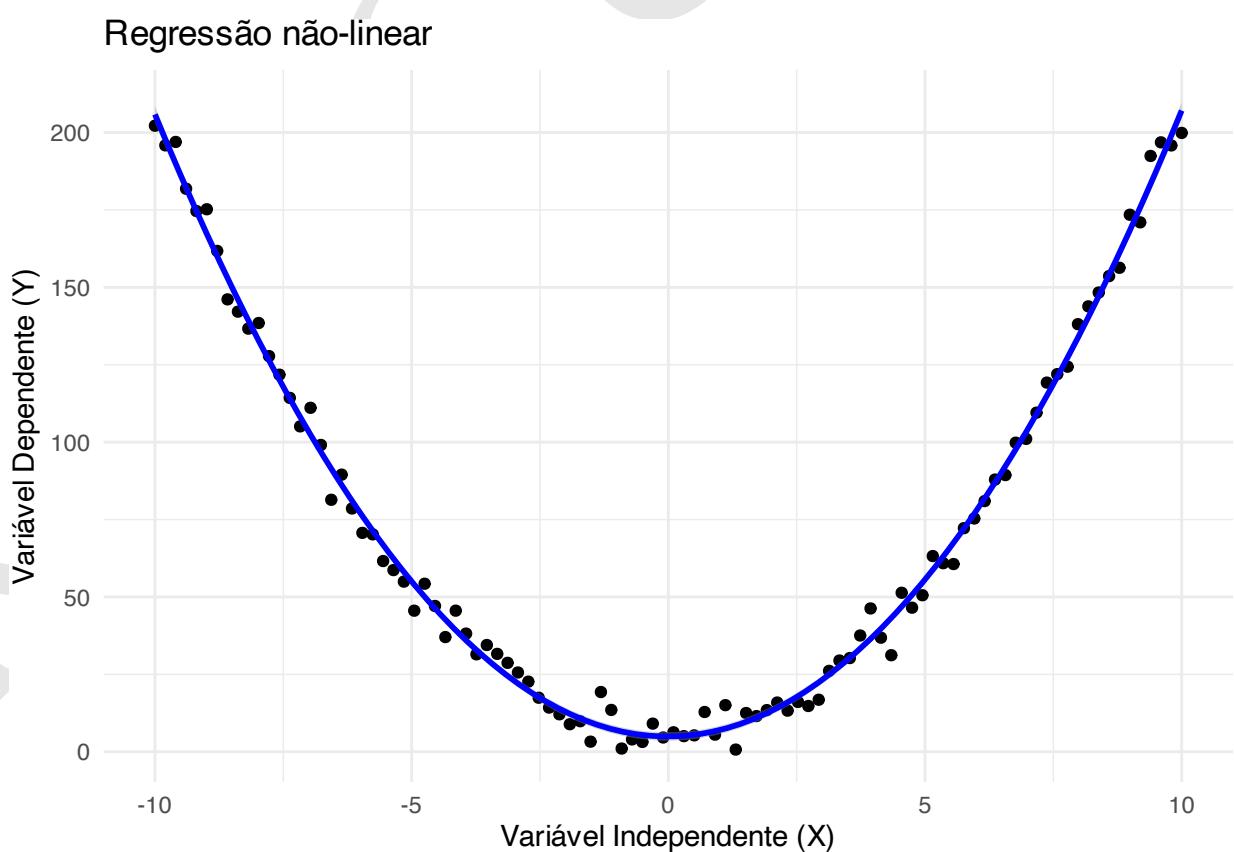


Figura 26.2: Regressão não-linear.

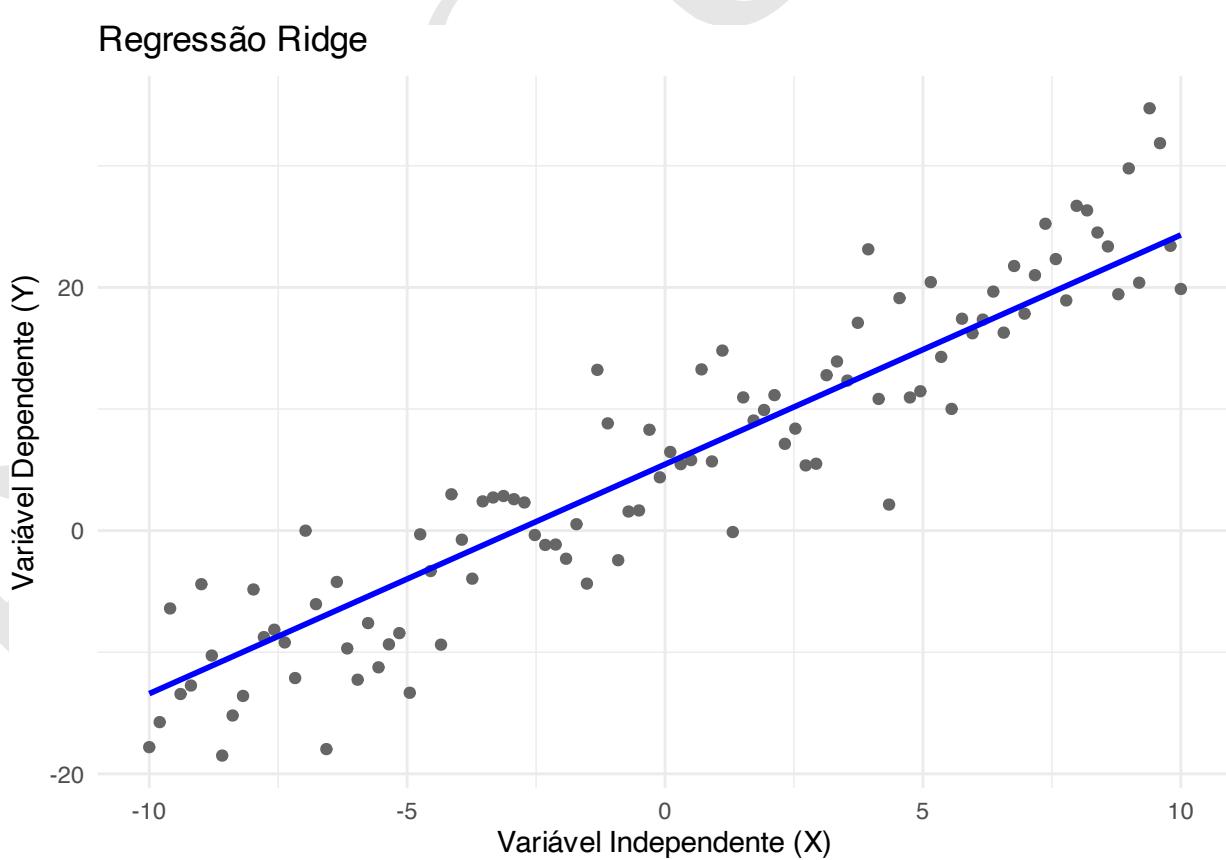
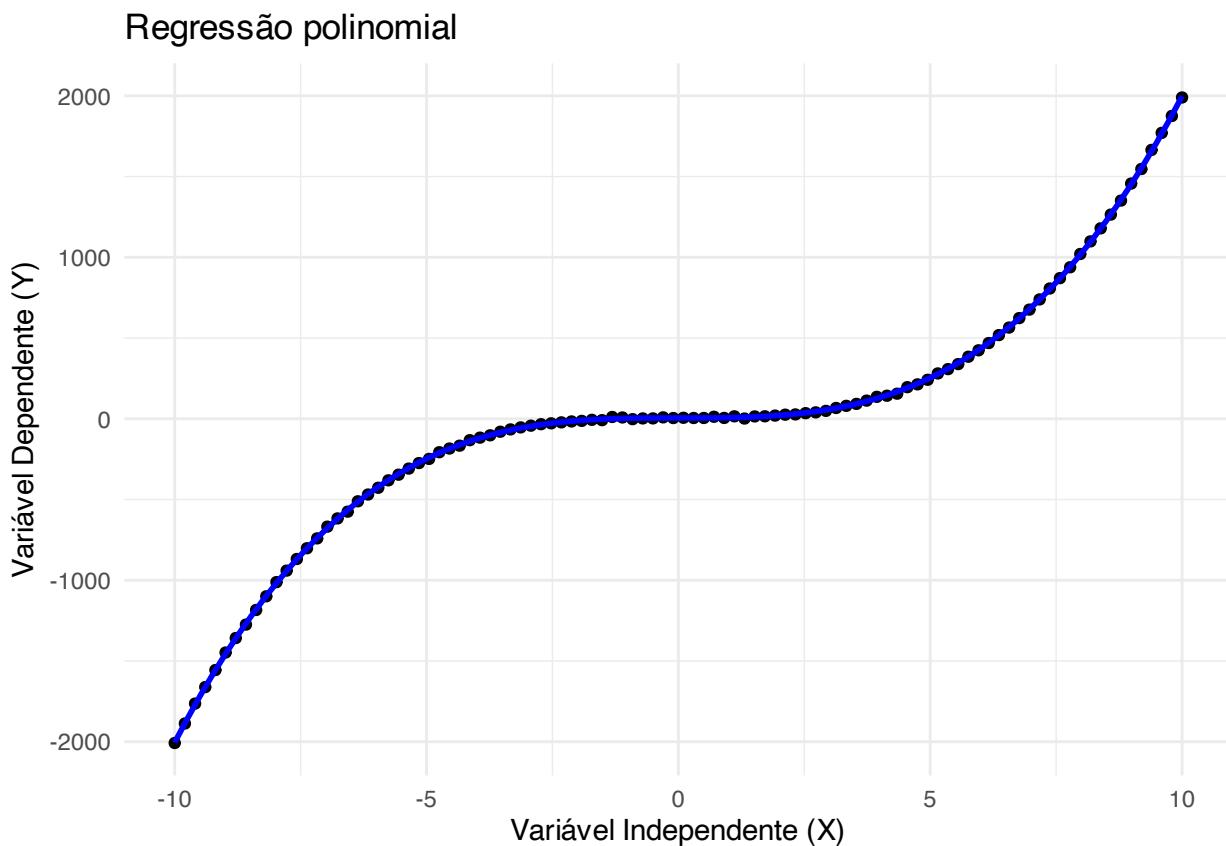


Figura 26.4: Regressão ridge.

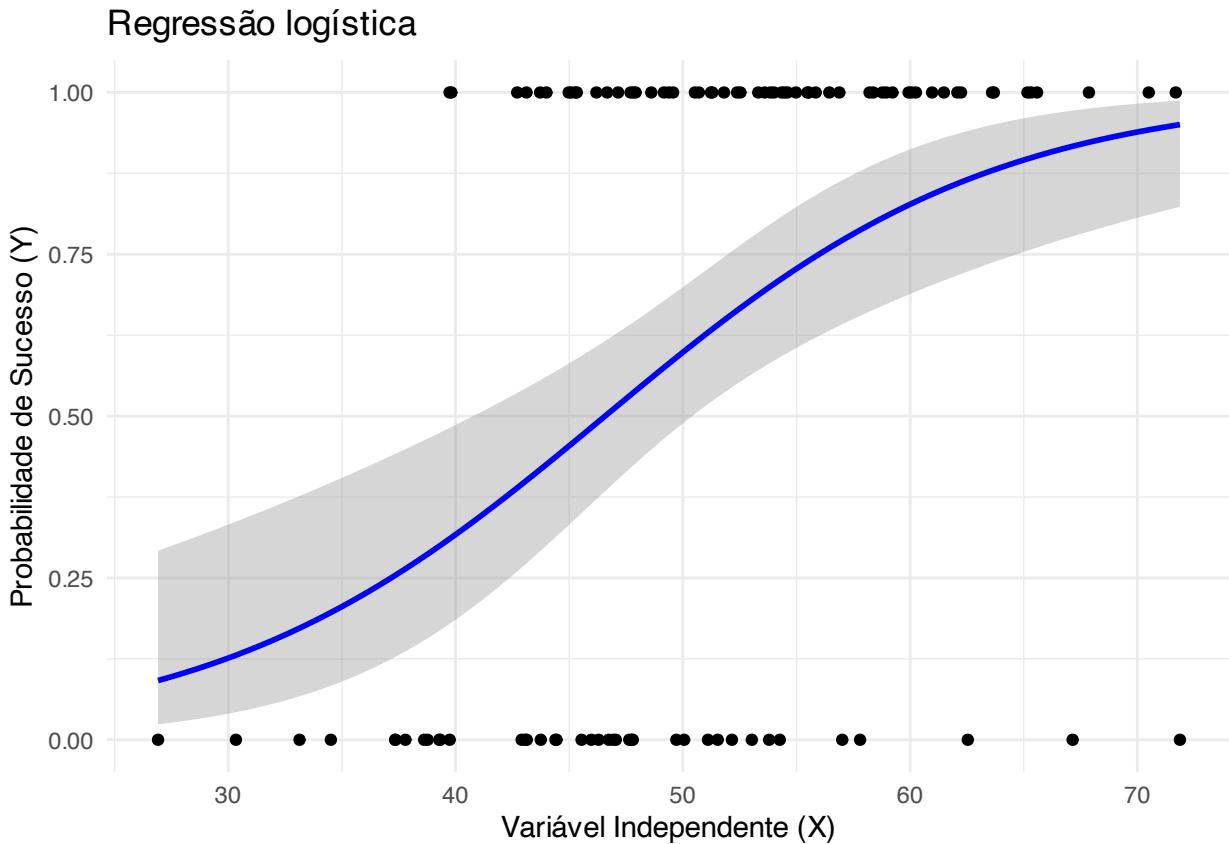


Figura 26.5: Regressão logística.

- Variáveis categóricas nominais, com 2 ou mais níveis, devem ser subdivididas em variáveis fictícias dicotômicas para ser usada em modelos de regressão.²⁴⁵
- Cada nível da variável categórica nominal será convertido em uma nova variável fictícias dicotômica, tal que a nova variável dicotômica assume valor 1 para a presença do nível correspondente e 0 em qualquer outro caso.²⁴⁵

R O pacote *fastDummies*²⁴⁶ fornece a função *dummy_cols*^a para preparar as variáveis categóricas fictícias para análise de regressão.

^ahttps://www.rdocumentation.org/packages/fastDummies/versions/1.7.3/topics/dummy_columns

26.2.2 Por que é comum escolher a categoria mais frequente como referência em modelos epidemiológicos?

- Maior estabilidade estatística: a categoria mais frequente costuma gerar estimativas mais estáveis, com menor erro padrão nos coeficientes das demais categorias?
- A escolha da referência não altera o ajuste nem o valor predito pelo modelo — apenas muda o ponto de comparação?

26.3 Multicolinearidade

26.3.1 O que é multicolinearidade?

- Multicolinearidade representa a intercorrelação entre as variáveis independentes (explanatórias) de um modelo.²⁴⁰

26.3.2 Como diagnosticar multicolinearidade de forma quantitativa?

- Verifique a existência de multicolinearidade entre as variáveis candidatas.²⁴⁷
- O Coeficiente de determinação (R^2) é uma medida de quão bem as variáveis independentes explicam a variabilidade da variável dependente. Valores próximos a 1 indicam que as variáveis independentes estão fortemente correlacionadas entre si, o que pode indicar multicolinearidade.²⁴⁸
- O Fator de Inflação da Variância (*variance inflation factor*, VIF) é uma medida que quantifica o quanto a variância de um coeficiente de regressão é inflacionada devido à multicolinearidade. Valores de VIF maiores que 10 são frequentemente considerados indicativos de multicolinearidade significativa.²⁴⁹
- O recíproco da VIF é chamado de Tolerância, que mede a proporção da variância de uma variável independente que não é explicada pelas outras variáveis independentes. Valores baixos de Tolerância (geralmente abaixo de 0.1) indicam multicolinearidade.²⁴⁹
- O número de condições (*Condition Number*) é uma medida que avalia a estabilidade numérica de um modelo de regressão. Valores altos (entre 10 de 30) indicam multicolinearidade, e valores maiores que 30 indicam forte multicolinearidade.²⁴⁹

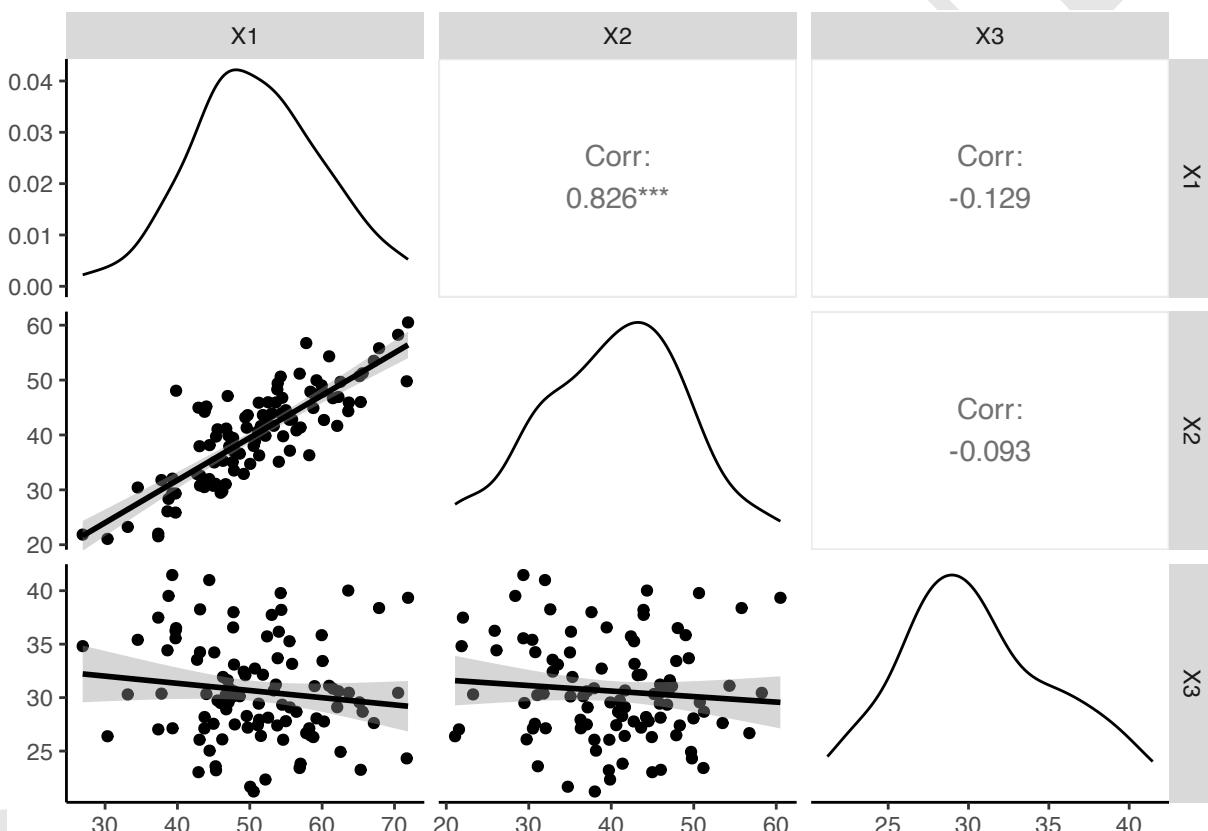


Figura 26.6: Multicolinearidade entre variáveis candidatas em modelos de regressão multivariável.



O pacote *GGally*²⁴¹ fornece a função *ggpairs*^a para criar uma matriz gráfica de correlações bivariadas.

^a<https://www.rdocumentation.org/packages/GGally/versions/2.2.1/topics/ggpairs>



O pacote *car*²⁴⁸ fornece a função *vif*^a para calcular o fator de inflação da variância (VIF).

^a<https://www.rdocumentation.org/packages/car/versions/3.1-3/topics/vif>

26.3.3 O que fazer em caso de multicolinearidade elevada?

- Verifique a transformação (codificação) de variáveis numéricas em categóricas.²⁴⁰
- Aumente o tamanho da amostra, se possível, para reduzir a multicolinearidade.²⁴⁰
- Combine níveis de variáveis categóricas com baixa frequência de ocorrência.²⁴⁰
- Combine variáveis numéricas altamente correlacionadas em uma única variável composta, como a média ou soma das variáveis.²⁴⁰
- Considere a exclusão de variáveis altamente correlacionadas do modelo, especialmente se elas não forem essenciais para a análise.²⁴⁰
- Use técnicas de seleção de variáveis, como seleção passo a passo, para identificar e remover variáveis redundantes.²⁴⁰
- Use técnicas de regularização, como regressão ridge ou lasso, que podem lidar com multicolinearidade ao penalizar coeficientes de regressão.²⁴⁰

26.4 Redução de dimensionalidade

26.4.1 Correlação bivariada pode ser usada para seleção de variáveis em modelos de regressão multivariável?

- Seleção bivariada de variáveis - isto é, aplicação de testes de correlação em pares de variáveis candidatas e variável de desfecho afim de selecionar quais serão incluídas no modelo multivariável - é um dos erros mais comuns na literatura.^{220,247,249}
- A seleção bivariada de variáveis torna o modelo mais suscetível a otimismo no ajuste se as variáveis de confundimento não são adequadamente controladas.^{247,249}

26.4.2 Variáveis sem significância estatística devem ser excluídas do modelo final?

- Eliminar uma variável de um modelo significa anular o seu coeficiente de regressão ($\beta = 0$), mesmo que o valor estimado pelos dados seja outro. Desta forma, os resultados se afastam de uma solução de máxima verossimilhança (que tem fundamento teórico) e o modelo resultante é intencionalmente subótimo.²²⁰
- Os coeficientes de regressão geralmente dependem do conjunto de variáveis do modelo e, portanto, podem mudar de valor (“mudança na estimativa” positiva ou negativa) se uma (ou mais) variável(is) for(em) eliminada(s) do modelo.²²⁰

26.4.3 Por que métodos de regressão gradual não são recomendados para seleção de variáveis em modelos de regressão multivariável?

- Métodos diferentes de regressão gradual podem produzir diferentes seleções de variáveis de um mesmo banco de dados.²⁴⁵
- Nenhum método de regressão gradual garante a seleção ótima de variáveis de um banco de dados.²⁴⁵
- As regras de término da regressão baseadas em P-valor tendem a ser arbitrárias.²⁴⁵

26.4.4 O que pode ser feito para reduzir o número de variáveis candidatas em modelos de regressão multivariável?

- Em caso de uma proporção baixa entre o número de participantes e de variáveis, use o conhecimento prévio da literatura para selecionar um pequeno conjunto de variáveis candidatas.²⁴⁷
- Colapse categorias com contagem nula (células com valor igual a 0) de variáveis candidatas.²⁴⁷
- Use simulações de dados para identificar qual(is) variável(is) está(ão) causando problemas de convergência do ajuste do modelo.²⁴⁷
- A eliminação retroativa tem sido recomendada como a abordagem de regressão gradual mais confiável entre aquelas que podem ser facilmente alcançadas com programas de computador.²²⁰

26.4.5 Quando devemos forçar uma variável no modelo?

- Sempre que houver base teórica ou evidência prévia forte (por exemplo, idade em estudos de câncer), ou se for a variável de exposição principal.²⁵⁰

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

PARTE 5: MODELAGEM ESTATÍSTICA

Ferramentas preditivas e causais

RASCUNHO

Capítulo 27

Modelos

27.1 Modelos estatísticos

27.1.1 O que é modelagem estatística?

- Modelagem é o processo de usar dados para selecionar um modelo matemático explícito que represente o processo gerador dos dados.²⁵⁰

27.1.2 Por que a escolha do modelo é complexa?

- Há inúmeras combinações possíveis de variáveis, formas funcionais (lineares, quadráticas, transformações), interações e formas do desfecho, o que torna o espaço de possibilidades muito amplo.²⁵⁰

R

O pacote *equatiomatic*²⁵¹ fornece a função *extract_eq*^a para extrair a equação dos modelos em formato LaTeX para visualização.

^ahttps://www.rdocumentation.org/packages/equatiomatic/versions/0.3.1/topics/extract_eq

27.2 Suposições dos modelos

27.2.1 Quais suposições são feitas para modelagem?

- ?

27.2.2 Como avaliar as suposições de um modelo?

- ?

R

O pacote *performance*²⁵² fornece a função *check_model*^a para analisar a colinearidade entre variáveis, a normalidade da distribuição das variáveis e a heteroscedasticidade.

^ahttps://www.rdocumentation.org/packages/performance/versions/0.10.4/topics/check_model

27.3 Avaliação de modelos

27.3.1 O que é qualidade de ajuste de um modelo?

- ?

27.3.2 Como avaliar a qualidade de ajuste de um modelo?

- Usando diagnóstico de regressão (ex.: análise de resíduos, gráficos de valores observados vs. preditos) e comparação com análises estratificadas.²⁵⁰

 O pacote *performance*²⁵² fornece a função *model_performance*^a para calcular as métricas de ajuste da regressão adequadas ao modelo pré-especificado.

^ahttps://www.rdocumentation.org/packages/Performance/versions/0.10.4/topics/model_performance

 O pacote *performance*²⁵² fornece a função *compare_performance*^a para comparar o desempenho e a qualidade do ajuste de diversos modelos de regressão pré-especificados.

^ahttps://www.rdocumentation.org/packages/Performance/versions/0.10.4/topics/compare_performance

27.4 Modelos estocásticos

27.4.1 O que são modelos estocásticos?

- ?

27.4.2 O que são cadeias de Markov?

- ?

27.4.3 Como construir uma cadeia de Markov?

- ?

 O pacote *markovchain*²⁵³ fornece a função *markovchainFit*^a ajusta uma cadeia com base em dados observados.

^a<https://www.rdocumentation.org/packages/markovchain/versions/0.9.5/topics/createSequenceMatrix>

27.5 Comparação de modelos

27.5.1 Como comparar modelos de aprendizagem de máquina?

- ?

 O pacote *correctR*²⁵⁴ fornece funções para comparar o desempenho e a qualidade do ajuste de diversos modelos de aprendizagem de máquina em amostras correlacionadas.

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

Capítulo 28

Redes

28.1 Análise de redes

28.1.1 O que é análise de rede?

- ?



O pacote *cooccur*²⁵⁵ fornece a função *cooccur*^a para criar calcular a coocorrência de objetos em um banco de dados.

^a<https://www.rdocumentation.org/packages/cooccur/versions/1.3/topics/cooccur>

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

RAASCUNHO

Capítulo 29

Aprendizado de máquina

29.1 Aprendizado de máquina

29.1.1 O que é aprendizado de máquina?

- .?

29.1.2 Quais são os principais algoritmos de aprendizado de máquina?

- Regressão linear: .?
- Árvores de decisão: .?
- Máquinas de vetores de suporte: .?
- Regressão logística: .?
- K-médias: .?
- K-vizinhos mais próximos: .?
- Redes neurais: .?
- Florestas aleatórias: .?
- Análise de componentes principais: .?
- Naive Bayes: .?

29.2 Aprendizado supervisionado

- .?

29.3 Aprendizado não-supervisionado

- .?

29.4 Aprendizado por reforço

- .?

29.5 Aprendizado profunda

- .?

R O pacote *h2o*²⁵⁴ fornece funções construir modelos de aprendizado de máquina.

R O pacote *correctR*²⁵⁴ fornece as funções *kfold_ttest*^a, *repkfold_ttest*^b e *resampled_ttest*^c para calcular estatística para comparação de modelos de aprendizado de máquina em amostras dependentes.

^a<https://cloud.r-project.org/web/packages/correctR/correctR.pdf>

^b<https://cloud.r-project.org/web/packages/correctR/correctR.pdf>

^c<https://cloud.r-project.org/web/packages/correctR/correctR.pdf>

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

Capítulo 30

Árvore de decisão

30.1 Árvore de decisão

30.1.1 O que é árvore de decisão?

- ?

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

RAASCUNHO

Capítulo 31

Análise preditiva

31.1 Predição via modelagem

31.1.1 O que são predições?

- ?

R O pacote *ggeffects*²⁵⁶ fornece a função *predict_response*^a para calcular valores preditos marginais ou ajustados das variáveis de desfecho.

^ahttps://www.rdocumentation.org/packages/ggeffects/versions/1.6.0/topics/predict_response

R O pacote *ggeffects*²⁵⁶ fornece a função *test_response*^a para testar valores preditos marginais ou ajustados das variáveis de desfecho.

^ahttps://www.rdocumentation.org/packages/ggeffects/versions/1.6.0/topics/test_response

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

RAASCUNHO

Capítulo 32

Análise causal

32.1 Causalidade

32.1.1 O que é análise causal?

- Análise causal é usada para explicar a relação entre causa e efeito em um conjunto de dados, respondendo a perguntas do tipo “por quê?”.¹⁶³
- Análise causal implica em contrafactual, no sentido de que a análise causal é baseada na comparação entre o que realmente aconteceu e o que teria acontecido se uma ou mais variáveis tivessem sido diferentes.¹⁶³

R O pacote *dagitty*²⁵⁷ fornece a função *dagitty*^a para criar um objeto grafo a partir de uma descrição textual.

^a<https://cran.r-project.org/web/packages/dagitty/index.html>

R O pacote *ggdag*²⁵⁸ fornece a função *ggdag*^a para criar figuras de grafos.

^a<https://www.rdocumentation.org/packages/ggdag/versions/0.2.10/topics/ggdag>

R O pacote *performance*²⁵² fornece a função *check_dag*^a para criar, verificar e visualizar os modelos em grafos.

^ahttps://easystats.github.io/performance/reference/check_dag.html

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

RAASCUNHO

PARTE 6: DELINEAMENTO DE PESQUISA

Delineamento antes da análise

RASCUNHO

Capítulo 33

Delineamento de estudos

33.1 Critérios de delineamento

33.1.1 Quais critérios são utilizados para classificar os delineamentos de estudos?

- ?

33.2 Alocação

33.2.1 O que é alocação?

- ?

33.3 Cegamento

- ?

33.3.1 O que é cegamento?

33.4 Pareamento

33.4.1 O que é pareamento?

- Pareamento significa que para cada participante de um grupo (por exemplo, com alguma condição clínica), existe um (ou mais) participantes (por exemplo, grupo controle) que possui características iguais ou similares relativas a algumas variáveis de interesse.²⁵⁹
- As variáveis escolhidas para pareamento devem ter relação com as variáveis de desfecho, mas não são de interesse elas mesmas.²⁵⁹
- O ajuste por pareamento deve ser incluído nas análises estatísticas mesmo que as variáveis de pareamento não sejam consideradas prognósticas ou confundidores na amostra estudada.²⁵⁹
- A ausência de evidência estatística de diferença entre grupos não é considerada pareamento.²⁵⁹

33.5 Aleatorização

33.5.1 O que é aleatorização?

- ?

33.6 Taxonomia de estudos

33.6.1 Como podem ser classificados os estudos científicos?

- Estudos científicos podem ser classificados em *básicos, observacionais, experimentais, acurácia diagnóstica, propriedades psicométricas, avaliação econômica e revisões de literatura*.²⁶⁰⁻²⁶⁹
- *Estudos básicos*^{261,266}
 - Genética
 - Celular
 - Experimentos com animais
 - Desenvolvimento de métodos
- *Estudos de simulação computacional*^{267,269}
- *Estudos de propriedades psicométricas*^{262,264}
 - Validade
 - Concordância
 - Confiabilidade
- *Estudos de desempenho diagnóstico*^{265,268}
 - Transversal
 - Caso-Controle
 - Comparativo
 - Totalmente pareado
 - Parcialmente pareado com subgrupo aleatório
 - Parcialmente pareado com subgrupo não aleatório
 - Não pareado aleatório
 - Não pareado não aleatório
- *Estudos observacionais*^{261,266}
 - Descritivo
 - * Estudo de caso
 - * Série de casos
 - * Transversal
 - Analítico
 - * Transversal
 - * Caso-Controle
 - Caso-Controle aninhado
 - Caso-Coorte
 - Coorte prospectiva ou retrospectiva
- *Estudos quase-experimentais*²⁶³
 - Quase-aleatorizado controlado
 - Estimação de variável instrumental
 - Descontinuidade de regressão
 - Série temporal interrompida controlada

- Série temporal interrompida
- Diferença
- *Estudos experimentais*^{261,266}
 - Fases I a IV
 - * Aleatorizado controlado
 - * Não-aleatorizado controlado
 - * Autocontrolado
 - * Cruzado
 - * Fatorial
 - Campo
 - Comunitário
- *Estudos de avaliação econômica*²⁶¹
 - Análise de custo
 - Análise de minimização de custo
 - Análise de custo-utilidade
 - Análise de custo-efetividade
 - Análise de custo-benefício
- *Estudos de revisão*²⁶⁰
 - Estado-da-arte
 - Narrativa
 - Crítica
 - Mapeamento
 - Escopo
 - Busca e revisão sistemática
 - Sistematizada
 - Sistemática
 - * Meta-análise
 - * Bibliométrica.^{270,271}
 - Sistemática qualitativa
 - Mista
 - Visão geral
 - Rápida
 - Guarda-chuva

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

RAIS DE CUSTO

Capítulo 34

Tamanho da amostra

34.1 Tamanho da amostra

34.1.1 O que é tamanho da amostra?

- Tamanho da amostra n é a quantidade de participantes (ou unidades de análise) necessárias para conduzir um estudo a fim de testar uma hipótese.²⁷²
- ¹³

34.1.2 Por que determinar o tamanho da amostra é importante?

- É virtualmente impossível, devido a limitações de recursos - tempo, acesso, custo, dentre outros - coletar dados da população completa.⁸
- Uma amostra muito pequena para o estudo pode resultar em ajuste exagerado, imprecisão e baixo poder do teste.⁷⁵

34.1.3 Quais fatores devem ser considerados para determinar o tamanho da amostra?

- Tamanho da população (N): O tamanho da amostra depende parcialmente do tamanho da população de origem. Geralmente assume-se que a população tem tamanho desconhecido ou infinito. Em alguns estudos serão amostradas populações de tamanho finito (inferior a 100.000 indivíduos), geralmente em pesquisas descritivas, em que esse tamanho deve ser incorporado nos cálculos.²⁷²
- Delineamento do estudo.²⁷²
- Quantidade e características (dependente vs. independente) dos grupos de participantes do estudo.²⁷²
- Erros tipo I (α) e tipo II (β).²⁷²
- Tipo de variável a ser observada (contínua, intervalo, ordinal, nominal, dicotômica).²⁷²
- Tamanho de efeito mínimo a ser observado.²⁷²
- Variabilidade da(s) variável(eis) coletada(s).²⁷²
- Lateralidade do teste de hipótese (uni- ou bicaudais).²⁷²
- Perdas de dados durante a coleta e/ou acompanhamento dos participantes do estudo.²⁷²

R O pacote *pwr*¹⁹⁵ fornece a função *plot.power.htest*^a para apresentar graficamente a relação entre o tamanho da amostra e o poder de testes de hipóteses.

^a<https://www.rdocumentation.org/packages/pwr/versions/1.3-0/topics/plot.power.htest>

34.1.4 Quais aspectos éticos estão envolvidos no tamanho da amostra?

- Determinar a priori o tamanho da amostra pode diminuir o risco de realizar testes ou intervenções desnecessários, de desperdício de recursos (tempo e dinheiro) associados e, por outro lado, de coletar dados insuficientes para testar as hipóteses do estudo.²⁷²
- O tratamento ético dos participantes do estudo, portanto, não exige que se considere se o poder do estudo é inferior à meta convencional de 80% ou 90%.²⁷³
- Estudos com poder <80% não são necessariamente antiéticos.²⁷³
- Grandes estudos podem ser desejáveis por outras razões que não as éticas.²⁷³

34.2 Cálculo do tamanho da amostra

34.2.1 Como calcular o tamanho da amostra?

- O tamanho amostral pode ser calculado por meio de fórmulas matemáticas que tendem a assegurar margens de erros tipos I (α) e II (β) para a estimação dos parâmetros populacionais (tamanho de efeito) a partir dos dados amostrais.²⁷²
- O tamanho da amostra deve ser calculado para cada um dos objetivos primários e/ou secundários, sendo escolhido o maior tamanho de amostra calculado para o estudo.²⁷²
- Geralmente é recomendado ser cético em relação às regras práticas para o tamanho da amostra, tais como a proporção entre o número de variáveis (ou eventos) e de participantes.⁷⁵

34.2.2 Como especificar o tamanho do efeito esperado?

- Estudo-piloto — realizados nas mesmas condições do estudo, mas envolvendo um tamanho de amostra limitado — pode ser útil na estimativa do tamanho da amostra a partir do tamanho do efeito estimado.²⁷²
- Utilizar os limites dos intervalos de confiança de estudos-piloto de ensaios clínicos como estimativa do tamanho do efeito pode aumentar o poder estatístico da análise se comparado ao uso das estimativas pontuais obtidas no mesmo piloto.²⁷⁴
- Embora os testes de hipótese considerem efeito nulo para a hipótese nula — ex.: diferença de média ($H_0 : \mu_1 - \mu_2 = 0$), correlação ($H_0 : r = 0$), associação ($H_0 : \beta = 0$ ou $H_0 : OR = 1$) —, em geral é improvável que os efeitos populacionais sejam de fato nulos (isto é, exatamente 0).²⁷⁵

R O pacote *pwr*¹⁹⁵ fornece a função *pwr.2p.test*^a para cálculo do tamanho da amostra para testes de proporção balanceados (2 amostras com mesmo número de participantes).

^a<https://www.rdocumentation.org/packages/pwr/versions/1.3-0/topics/pwr.2p.test>

R O pacote *pwr*¹⁹⁵ fornece a função *pwr.2p2n.test*^a para cálculo do tamanho da amostra para testes de proporção não balanceados (2 amostras com diferente número de participantes).

^a<https://www.rdocumentation.org/packages/pwr/versions/1.3-0/topics/pwr.2p2n.test>

R O pacote *pwr*¹⁹⁵ fornece a função *pwr.anova.test*^a para cálculo do tamanho da amostra para testes de análise de variância balanceados (3 ou mais amostras com mesmo número de participantes).

^a<https://www.rdocumentation.org/packages/pwr/versions/1.3-0/topics/pwr.anova.test>

R

O pacote *pwr*¹⁹⁵ fornece a função *pwr.chisq.test*^a para cálculo do tamanho da amostra para testes de qui-quadrado χ^2 .

^a<https://www.rdocumentation.org/packages/pwr/versions/1.3-0/topics/pwr.chisq.test>

R

O pacote *pwr*¹⁹⁵ fornece a função *pwr.f2.test*^a para cálculo do tamanho da amostra para testes com modelo linear geral.

^a<https://www.rdocumentation.org/packages/pwr/versions/1.3-0/topics/pwr.f2.test>

R

O pacote *pwr*¹⁹⁵ fornece a função *pwr.norm.test*^a para cálculo do tamanho da amostra para a média de uma distribuição normal com variância conhecida.

^a<https://www.rdocumentation.org/packages/pwr/versions/1.3-0/topics/pwr.norm.test>

R

O pacote *pwr*¹⁹⁵ fornece a função *pwr.p.test*^a para cálculo do tamanho da amostra para testes de proporção (1 amostra).

^a<https://www.rdocumentation.org/packages/pwr/versions/1.3-0/topics/pwr.p.test>

R

O pacote *pwr*¹⁹⁵ fornece a função *pwr.r.test*^a para cálculo do tamanho da amostra para testes de correlação (1 amostra).

^a<https://www.rdocumentation.org/packages/pwr/versions/1.3-0/topics/pwr.r.test>

R

O pacote *pwr*¹⁹⁵ fornece a função *pwr.t.test*^a para cálculo do tamanho da amostra para testes *t* de diferença de 1 amostra, 2 amostras dependentes ou 2 amostras independentes (grupos balanceados).

^a<https://www.rdocumentation.org/packages/pwr/versions/1.3-0/topics/pwr.t.test>

R

O pacote *pwr*¹⁹⁵ fornece a função *pwr.t2n.test*^a para cálculo do tamanho da amostra para testes *t* de diferença de 2 amostras independentes (grupos não balanceados).

^a<https://www.rdocumentation.org/packages/pwr/versions/1.3-0/topics/pwr.t2n.test>

R

O pacote *longpower*¹⁹⁶ fornece a função *power.mmmr*^a para calcular o tamanho da amostra para estudos com análises por modelo de regressão linear misto.

^a<https://www.rdocumentation.org/packages/longpower/versions/1.0.24/topics/power.mmmr>

34.3 Perdas de amostra

34.3.1 O que é perda de amostra?

- Perda de amostra(s) — isto é, participante(s) ou unidade(s) de análise — pode ocorrer durante a coleta e/ou acompanhamento dos participantes do estudo.²⁷²
- Perda amostral pode ocorrer por: abandono ou desistência do participante, perda de contato com o participante, perda de informação, ocorrência de eventos adversos, morte do participante, entre outros.²⁷²

34.3.2 Por que a perda de amostra é um problema?

- A perda de amostra pode levar a uma redução do poder do estudo, aumentando a probabilidade de erro tipo II (β).²⁷¹
- A perda de amostra pode levar a um viés de seleção, pois os participantes que permanecem no estudo podem ser diferentes daqueles que foram perdidos.²⁷²

34.3.3 Como evitar perda de amostra?

- A perda de amostra pode ser evitada por meio de um planejamento cuidadoso do estudo, incluindo a definição de critérios de inclusão e exclusão claros e apropriados, bem como a definição de estratégias para minimizar a perda de amostra.²⁷³
- A perda de amostra pode ser compensada pelo aumento do tamanho da amostra, desde que o aumento seja suficiente para manter o poder do estudo.²⁷²

34.4 Ajustes no tamanho da amostra

34.4.1 Por que ajustar o tamanho da amostra?

- O tamanho da amostra pode ser ajustado durante o estudo para compensar a perda de amostra, desde que o aumento seja suficiente para manter o poder do estudo.²⁷²

34.4.2 Como ajustar para perda amostral?

- Aumentar o tamanho da amostra estimada n pela porcentagem d de perdas esperada ou prevista, para obter o tamanho da amostra efetiva n' com base na equação (34.1).²⁷²

$$n' = \frac{n}{1 - d} \quad (34.1)$$

34.5 Justificativa do tamanho da amostra

34.5.1 Como justificar o tamanho da amostra de um estudo?

- Em estudos que envolvem condições raras, pode ser difícil recrutar o número necessário de participantes devido à limitada disponibilidade de casos da população. Mesmo assim, é aconselhável determinar o tamanho da amostra.²⁷²
- Quando um estudo deste tipo não é possível, as considerações referentes ao tamanho da amostra são justificadas de acordo com o número máximo de pacientes que podem ser recrutados no decorrer do estudo.²⁷²

34.6 SPARKing

34.6.1 O que é SPARKing?

- SPARKing é um acrônimo para *Sample size Planning After the Results are Known*.²⁷⁶
- SPARKing é uma má prática que envolve o ajuste do tamanho da amostra após a coleta dos dados, com o objetivo de obter resultados estatisticamente significativos.²⁷⁶
- SPARKing é considerado antiético, pois pode levar a resultados enviesados e não confiáveis, além de violar os princípios da pesquisa científica.²⁷⁶

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

Capítulo 35

Estudos observacionais

35.1 Características

35.1.1 Quais são as características de estudos observacionais?

- ?

35.2 Diretrizes para redação

35.2.1 Quais são as diretrizes para redação de estudos observacionais?

- Visite a rede *Enhancing the QUAlity and Transparency Of health Research* EQUATOR Network¹ para encontrar diretrizes específicas para cada tipo de estudo observacional.
 - *The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies:*²⁷⁷ <https://www.equator-network.org/reporting-guidelines/strobe/>

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

¹<https://www.equator-network.org/>

RAISGUNHO

Capítulo 36

Propriedades psicométricas

36.1 Características

36.1.1 O que são propriedades psicométricas?

• ?

R

O pacote *lavaan*²⁷⁸ fornece a função *cfa*^a para implementar modelos de análise fatorial confirmatória.

^a<https://www.rdocumentation.org/packages/lavaan/versions/0.6-16/topics/cfa>

R

O pacote *lavaan*²⁷⁸ fornece a função *modificationIndices*^a para calcular os índices de modificação.

^a<https://www.rdocumentation.org/packages/lavaan/versions/0.6-16/topics/modificationIndices>

R

O pacote *semTools*²⁷⁹ fornece a função *reliability*^a para analisar a confiabilidade de um instrumento.

^a<https://www.rdocumentation.org/packages/semTools/versions/0.5-6/topics/reliability-deprecated>

R

O pacote *psych*²⁸⁰ fornece a função *icc*^a para calcular a confiabilidade utilizando coeficientes de correlação intraclass.

^a<https://www.rdocumentation.org/packages/psych/versions/2.3.6/topics/ICC>

36.2 Análise fatorial exploratória

36.2.1 O que é análise fatorial exploratória?

• ?

Tabela 36.1: Tabela de confusão sobre propriedades psicométricas de instrumentos.

	Concordância alta	Concordância baixa
Validade alta	Adequado	Inadequado
Validade baixa	Inadequado	Inadequado

36.3 Análise factorial confirmatória

36.3.1 O que é análise factorial confirmatória?

- ?

36.4 Validade de conteúdo

36.4.1 O que é validade interna?

- 281

36.4.2 O que é validade externa?

- 281

36.4.3 Que fatores afetam a validade?

- A amostragem não probabilística pode dificultar a generalização dos achados da amostra para a população, diminuindo assim a validade externa do estudo.¹³
- Quando as características da amostra obtida por seleção não probabilística forem similares às da população, a validade externa pode ser maior.¹³

36.4.4 Como avaliar a validade de um estudo?

- As características da amostra apresentadas na Tabela 1 são úteis para interpretação da validade interna e externa dos achados do estudo.¹⁷⁴

36.5 Validade de face

36.5.1 O que é validade de face?

- [RF]

36.6 Validade do construto

36.6.1 O que é construto?

- [RF]

36.7 Validade factorial

36.7.1 O que é validade factorial?

- [RF]

36.8 Validade convergente

36.8.1 O que é validade convergente?

- [RF]

36.9 Validade discriminante

36.9.1 O que é validade discriminante?

- [RF]

Tabela 36.2: Tabela de confusão 2x2 para análise de concordância de testes e variáveis dicotômicas.

	Teste positivo	Teste negativo	Total
Teste positivo	a	b	$g = a + b$
Teste negativo	c	d	$h = c + d$
Total	$e = a + c$	$f = b + d$	$N = a + b + c + d$

36.10 Validação de critério

36.10.1 O que é validade de critério?

- .[RF]

36.11 Validação concorrente

36.11.1 O que é concorrente?

- .[RF]

36.11.2 O que é validade concorrente?

- .[RF]

36.11.3 O que é validade preditiva?

- .[RF]

36.12 Responsividade

36.12.1 O que é responsividade?

- ?

36.13 Concordância

36.13.1 O que é concordância?

- ?

36.13.2 Quais métodos são adequados para análise de concordância de variáveis dicotômicas?

- Coeficiente de Cohen κ : mede a concordância corrigida pelo acaso.^{282,283}
- Coeficiente de correlação tetracórica r_{tet} .^{284,285}

 O pacote *psych*²⁸⁰ fornece a função *tetrachoric*^a para calcular o coeficiente de correlação tetracórica (r_{tet}).

^a<https://www.rdocumentation.org/packages/psych/versions/2.3.6/topics/tetrachoric>

36.13.3 Quais métodos não são adequados para análise de concordância de variáveis dicotômicas?

- Concordância absoluta C_A - quantidade de casos em que examinadores concordam - não é recomendada porque não corrige a estimativa para possíveis concordâncias ao acaso.²⁸⁵

Tabela 36.3: Tabela de confusão 3x3 para análise de concordância de testes e variáveis dicotômicas.

	Grave	Moderado	Leve	Total
Grave	a	b	c	$j = a + b + c$
Moderado	d	e	f	$k = d + e + f$
Leve	g	h	i	$l = g + h + i$
Total	$j = a + d + g$	$k = b + e + h$	$l = c + f + i$	$N = a + b + c + d + e + f + g + h + i$

- Concordância percentual $C\%$ - proporção de casos em que examinadores concordam pela quantidade total de casos - não é recomendada porque não corrige a estimativa para possíveis concordâncias ao acaso.²⁸⁵
- Qui-quadrado χ^2 a partir da tabela de contigência não é recomendado porque tal teste analisa associação.²⁸⁵
- A família de coeficientes de Cohen κ não é adequada para analisar concordância quando as variáveis são aparentemente (e não originalmente) dicotômicas.²⁸⁵

36.13.4 Quais métodos são adequados para análise de concordância de variáveis categóricas?

- Coeficiente de Cohen κ : mede a concordância corrigida pelo acaso.^{282,283}
- Coeficiente de Cohen ponderado κ_w : mede a concordância corrigida pelo acaso.^{282,283}
- Coeficiente de correlação policórica r_{pol} .²⁸⁵

R O pacote *psych*²⁸⁰ fornece a função *tetrachoric*^a para calcular o coeficiente de correlação policórica (r_{pol}).

^a<https://www.rdocumentation.org/packages/psych/versions/2.3.6/topics/tetrachoric>

36.13.5 Quais métodos são adequados para análise de concordância de variáveis categóricas e contínuas?

- Coeficiente de correlação bisserial r_s .²⁸⁵

R O pacote *psych*²⁸⁰ fornece a função *tetrachoric*^a para calcular o coeficiente de correlação bisserial (r_s).

^a<https://www.rdocumentation.org/packages/psych/versions/2.3.6/topics/tetrachoric>

36.13.6 Quais métodos são adequados para análise de concordância de variáveis ordinais?

- Coeficiente de Cohen ponderado κ_w : mede a concordância corrigida pelo acaso.^{282,283}

36.13.7 Quais métodos são adequados para análise de concordância de variáveis contínuas?

- Gráfico de dispersão com a reta de regressão.⁷³
- Gráfico de limites de concordância (média dos testes vs. diferença entre testes) com a reta de regressão do viés e respectivo no nível de significância α pré-estabelecido.⁷³

R

O pacote *BlandAltmanLeh*²⁸⁶ fornece as funções *bland.altman.stats*^a e *bland.altman.plot*^b para calcular e apresentar, respectivamente, o gráfico com os limites de concordância entre dois métodos.

^a<https://www.rdocumentation.org/packages/BlandAltmanLeh/versions/0.3.1/topics/bland.altman.stats>

^b<https://www.rdocumentation.org/packages/BlandAltmanLeh/versions/0.3.1/topics/bland.altman.plot>

36.13.8 Quais métodos não são adequados para análise de concordância de variáveis contínuas?

- Comparação de médias: dois métodos apresentarem médias similares - isto é, 'sem diferença estatística' após um teste inferencial de hipótese nula $H_0 : \mu_1 = \mu_2$ - não informa sobre a concordância deles. Métodos com maior erro de medida tendem a ter menos chance de rejeição da hipótese nula.⁷³
- Correlação bivariada: o coeficiente de correlação dependente tanto da variação entre indivíduos (isto é, entre os valores verdadeiros) quanto da variação intraindividual (isto é, erro de medida). Se a variância dos erros de medida de ambos os métodos não for pequena comparadas à variância dos valores verdadeiros, o tamanho do efeito da correlação será pequeno mesmo que os métodos possuam boa concordância.⁷³
- Regressão linear: o teste da hipótese nula da inclinação da reta de regressão ($H_0 : \beta = 0$) é equivalente a testar a correlação bivariada ($H_0 : \rho = 0$).⁷³

36.13.9 Quais métodos são adequados para modelagem de concordância?

- Modelo log-linear.²⁸⁵

36.14 Confiabilidade

36.14.1 O que é confiabilidade?

- ?

36.14.2 Quais métodos são adequados para análise de confiabilidade?

- ?

36.15 Diretrizes para redação

36.15.1 Quais são as diretrizes para redação de estudos de propriedades psicométricas?

- Visite a rede *Enhancing the QUAlity and Transparency Of health Research* EQUATOR Network¹ para encontrar diretrizes específicas para cada tipo de estudo de propriedades psicométricas.
 - *COSMIN reporting guideline for studies on measurement properties of patient-reported outcome measures*:²⁸⁷ <https://www.equator-network.org/reporting-guidelines/cosmin-reporting-guideline-for-studies-on-measurement-properties-of-patient-reported-outcome-measures/>
 - *Recommendations for reporting the results of studies of instrument and scale development and testing*:²⁸⁸ <https://www.equator-network.org/reporting-guidelines/recommendations-for-reporting-the-results-of-studies-of-instrument-and-scale-development-and-testing/>
 - *Guidelines for reporting reliability and agreement studies (GRRAS) were proposed*:²⁸⁹ <https://www.equator-network.org/reporting-guidelines/guidelines-for-reporting-reliability-and-agreement-studies-grras-were-proposed/>

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

¹<https://www.equator-network.org/>

RAASCUNHO

Capítulo 37

Desempenho diagnóstico

37.1 Características

37.1.1 Quais são as características de estudos de desempenho diagnóstico?

- ?

37.2 Tabelas 2x2

37.2.1 O que é uma tabela de confusão 2x2?

- Tabela de confusão é uma matriz de 2 linhas por 2 colunas que permite analisar o desempenho de classificação de uma variável dicotômica (padrão-ouro ou referência) versus outra variável dicotômica (novo teste).²⁹⁰

37.2.2 Como analisar o desempenho diagnóstico em tabelas 2x2?

- Verdadeiro-positivo (*VP*): caso com a condição presente e corretamente identificado como tal.²⁹¹
- Falso-negativo (*FN*): caso com a condição presente e erroneamente identificado como ausente.²⁹¹
- Verdadeiro-negativo (*VN*): controle sem a condição presente e corretamente identificados como tal.²⁹¹
- Falso-positivo (*FP*): controle sem a condição presente e erroneamente identificado como presente.²⁹¹
- Tabelas de confusão também podem ser visualizadas em formato de árvores de frequência.²⁹⁰

R O pacote *riskyR*²⁹² fornece a função *plot_prism*^a para construir árvores de frequência a partir de diferentes cenários.

^ahttps://www.rdocumentation.org/packages/riskyR/versions/0.4.0/topics/plot_prism

37.2.3 Quais probabilidades caracterizam o desempenho diagnóstico de um teste em tabelas 2x2?

- Sensibilidade (*SEN*), equação (37.1): Proporção de verdadeiro-positivos dentre aqueles com a condição.²⁹¹

Tabela 37.1: Tabela de confusão 2x2 para análise de desempenho diagnóstico de testes e variáveis dicotômicas.

	Condição presente	Condição ausente	Total
Teste positivo	VP	FP	$VP + FP$
Teste negativo	FN	VN	$FN + VN$
Total	$VP + FN$	$FP + VN$	$N = VP + VN + FP + FN$

Scenario

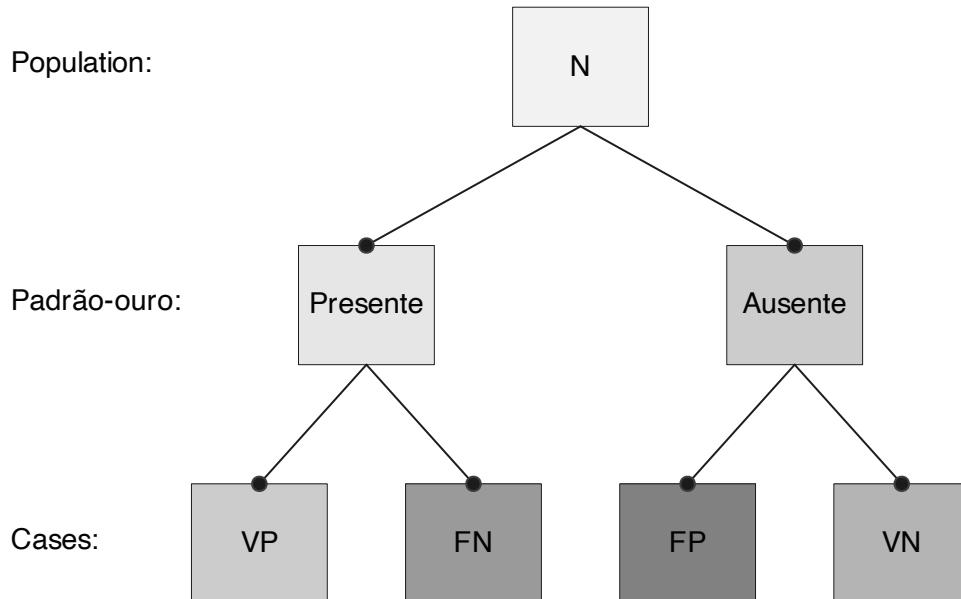


Figura 37.1: Árvore de frequência do desempenho diagnóstico de uma tabela de confusão 2x2 representando um método novo (dicotômico) comparado ao método padrão-ouro ou referência (dicotômico).

$$SEN = \frac{VP}{VP + FN} \quad (37.1)$$

- Especificidade (*ESP*), equação (37.2): Proporção de verdadeiro-negativos dentre aqueles sem a condição.²⁹¹

$$ESP = \frac{VN}{VN + FP} \quad (37.2)$$

- Acurácia (*ACU*), equação (37.3): Proporção de casos e controle corretamente identificados.²⁹¹

$$ACU = \frac{VP + VN}{VP + VN + FP + FN} \quad (37.3)$$

- Valor preditivo positivo (*VPP*), equação (37.4): Proporção de casos corretamente identificados como verdadeiro-positivos.²⁹¹

$$VPP = \frac{VP}{VP + FP} \quad (37.4)$$

- Valor preditivo negativo (*VPN*), equação (37.5): Proporção de controles corretamente identificados como verdadeiro-negativos.²⁹¹

$$VPN = \frac{VN}{VN + FN} \quad (37.5)$$

O pacote *riskyR*²⁹² fornece a função *comp_prob*^a para estimar 13 probabilidades relacionadas ao desempenho diagnóstico em tabelas 2x2.

^ahttps://www.rdocumentation.org/packages/riskyR/versions/0.4.0/topics/comp_prob

Tabela 37.2: Probabilidades calculados a partir da tabela de confusão 2x2 para análise de desempenho diagnóstico de testes e variáveis dicotômicas.

	Condição presente	Condição ausente	Total	Probabilidades
Teste positivo	VP	FP	$VP + FP$	$VPP = \frac{VP}{VP+FP}$
Teste negativo	FN	VN	$FN + VN$	$VPN = \frac{VN}{VN+FN}$
Total	$VP + FN$	$FP + VN$	$N = VP + VN + FP + FN$	
Probabilidades	$SEN = \frac{VP}{VP+FN}$	$ESP = \frac{VN}{VN+FP}$		$ACU = \frac{VP+VN}{VP+VN+FP+FN}$

R

O pacote *caret*²⁹³ fornece a função *confusionMatrix*^a para estimar 11 probabilidades relacionadas ao desempenho diagnóstico em tabelas 2x2.

^a<https://www.rdocumentation.org/packages/caret/versions/3.45/topics/confusionMatrix>

37.3 Gráficos *crosshair*

37.3.1 O que um gráfico *crosshair*?

• ²⁹⁴

R

O pacote *mada*²⁹⁵ fornece a função *crosshair*^a para criar um gráfico *crosshair*²⁹⁴ a partir de dados de verdadeiro-positivo, falso-positivo, verdadeiro-negativo e verdadeiro-positivo de tabelas de confusão 2x2.

^a<https://www.rdocumentation.org/packages/mada/versions/0.5.11/topics/crosshair>

37.4 Curvas ROC

37.4.1 O que é a área sob a curva (AUROC)?

- A área sob a curva ROC (AUC ou AUROC) quantifica o poder de discriminação ou desempenho diagnóstico na classificação de uma variável dicotômica.²⁹⁶

R

O pacote *proc*²⁹⁷ fornece a função *plot.roc*^a para criar uma curva ROC.

^a<https://www.rdocumentation.org/packages/pROC/versions/1.18.4/topics/plot.roc>

37.4.2 Como interpretar a área sob a curva (ROC)?

- A área sob a curva AUC varia no intervalo $[0.5; 1]$, com valores mais elevados indicando melhor discriminação ou desempenho do modelo de classificação.²⁹⁶
- As interpretações qualitativas (isto é: pobre/fraca/baixa, moderada/razoável/aceitável, boa ou muito boa/alta/excelente) dos valores de área sob a curva são arbitrários e não devem ser considerados isoladamente.²⁹⁶
- Modelos de classificação com valores altos de área sob a curva podem ser enganosos se os valores preditos por esses modelos não estiverem adequadamente calibrados.²⁹⁶

37.4.3 Como analisar o desempenho diagnóstico em desfechos com distribuição trimodal na população?

- Limiares duplos podem ser utilizados para análise de desempenho diagnóstico de testes com distribuição trimodal.²⁹⁸

37.5 Interpretação da validade de um teste

37.5.1 Que itens devem ser verificados na interpretação de um estudo de validade?

- O novo teste foi comparado junto ao método padrão-ouro.²⁹¹
- As probabilidades pontuais estimadas que caracterizam o desempenho diagnóstico do novo teste são altas e adequadas para sua aplicação clínica.²⁹¹
- Os intervalos de confiança estimados para as probabilidades do novo teste são estreitos e adequados para sua aplicação clínica.²⁹¹
- O novo teste possui adequada confiabilidade intra/inter examinadores.²⁹¹
- O estudo de validação incluiu um espectro adequado da amostra.²⁹¹
- Todos os participantes realizaram ambos o novo teste e o padrão-ouro no estudo de validação.²⁹¹
- Os examinadores do novo teste estavam cegados para o resultado do teste padrão-ouro.²⁹¹

37.6 Diretrizes para redação

37.6.1 Quais são as diretrizes para redação de estudos diagnósticos?

- Visite a rede *Enhancing the QUAlity and Transparency Of health Research* EQUATOR Network¹ para encontrar diretrizes específicas para cada tipo de estudo de desempenho diagnóstico.
 - *STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies*:²⁹⁹ <https://www.equator-network.org/reporting-guidelines/stard/>

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

¹<https://www.equator-network.org/>

Capítulo 38

Ensaios quase-experimentais

38.1 Características

38.1.1 Quais são as características de ensaios quase-experimentais?

- ?

38.2 Diretrizes para redação

38.2.1 Quais são as diretrizes para redação de ensaios quase-experimentais?

- Visite a rede *Enhancing the QUAlity and Transparency Of health Research* EQUATOR Network¹ para encontrar diretrizes específicas para cada tipo de estudo de ensaio quase-experimental.
- *Guidelines for reporting non-randomised studies*:³⁰⁰ <https://www.equator-network.org/reporting-guidelines/guidelines-for-reporting-non-randomised-studies/>

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

¹<https://www.equator-network.org/>

RAASCUNHO

Capítulo 39

Ensaios experimentais

39.1 Ensaio clínico aleatorizado

39.1.1 Quais são as características de ensaios clínicos aleatorizados?

- A característica essencial de um ensaio clínico aleatorizado é a comparação entre grupos.³⁰¹
- Quanto à unidade de alocação:³⁰²
 - Individual
 - Agrupado
- Quanto ao número de braços:³⁰²
 - Único*
 - Múltiplos
- Quanto ao número de centros:³⁰²
 - Único
 - Múltiplos
- Quanto ao cegamento:³⁰²
 - Aberto*
 - Simples-cego
 - Duplo-cego
 - Tripo-cego
 - Quádruplo-cego
- Quanto à alocação:³⁰²
 - Sem sorteio
 - Estratificada (centro apenas)
 - Estratificada
 - Minimizada
 - Estratificada e minimizada

39.1.2 Quais são as estratégias para metodológicas para reduzir vieses metodológicos?

- Grupo controle.[?]
- Grupo placebo.[?]

- Controle sham.[?]
- Cegamento.[?]

39.2 Modelos de análise de comparação

39.2.1 Que modelos podem ser utilizados para comparações?

- As abordagens compreendem a comparação da variável de desfecho medida entre os momentos antes e depois ou da sua mudança (pré - pós) entre os momentos.³⁰³
- Se a média da variável é igual entre grupos no início do acompanhamento, ambas abordagens estimam o mesmo efeito. Caso contrário, o efeito será influenciado pela correlação entre as medidas antes e depois. A análise da mudança não controla para desbalanços no início do estudo.³⁰³
- A abordagem mais recomendada é a análise de covariância (ANCOVA) - equação (39.1) - pois ajusta os valores pós-intervenção (Y_{ij}) aos valores pré-intervenção (X_{ij}) para cada participante (i) de cada grupo $\{Z_{ij}\}$, e portanto não é afetada pelas diferenças entre grupos no início do estudo.^{10,303}

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \beta_2 Z_j + \epsilon_{ij} \quad (39.1)$$

- A ANCOVA modelando seja a mudança (pré - pós: $\Delta = X_{ij} - Y_{ij}$) quando o desfecho no pós-tratamento parece ser o método mais efetivo considerando-se o viés de estimação dos parâmetros, a precisão das estimativas, a cobertura nominal (isto é, intervalo de confiança) e o poder do teste.³⁰⁴
- Quando a ANCOVA - equação (39.2) - é utilizada com a mudança (pré - pós) com variável de desfecho (Y_{ij}), o coeficiente de regressão β_1 é diminuído em 1 unidade.^{10,305}

$$(X_{ij} - Y_{ij}) = \beta_0 + \beta_1 Z_j + \epsilon_{ij} \quad (39.2)$$

- Análise de variância (ANOVA) e modelos lineares mistos (MLM) são outras opções de métodos, embora apresentem maior variância, menor poder, e cobertura nominal comparados à ANCOVA.³⁰⁴
- ³⁰⁶
- ³⁰⁷

39.3 Comparação na linha de base

39.3.1 O que são dados na linha de base?

- Dados sociodemográficos, clínicos e funcionais são coletados na linha de base sobre cada participante no momento da aleatorização.³⁰⁸
- Os dados de linha de base são usados para caracterizar os pacientes incluídos no estudo e para mostrar que os grupos de tratamento estão bem equilibrados.³⁰⁸
- Dados da linha de base podem ser utilizados para a aleatorização de modo a equilibrar ou estratificar os grupos considerando alguns fatores-chave.³⁰⁸
- Dados da linha de base podem ser utilizados como ajuste de covariável para análise do resultado por grupo de tratamento.³⁰⁸

39.3.2 O que é comparação entre grupos na linha de base em ensaios clínicos aleatorizados?

- A comparação se refere ao teste de hipótese nula de não haver diferença ('balanço' ou 'equilíbrio') entre grupos de tratamento nas (co)variáveis na linha de base, geralmente apresentadas apenas de modo descritivo na 'Tabela 1'.³⁰⁹
- A interpretação isolada do P-valor da comparação entre grupos na linha de base não permite identificar as razões para eventuais diferenças.³⁰⁹

39.3.3 Para quê comparar grupos na linha de base em ensaios clínicos aleatorizados?

- Os P-valores estão relacionados à aleatorização dos participantes em grupos.³¹⁰
- Em ensaios clínicos aleatorizados, a comparação de (co)variáveis na linha de base é usada para avaliar se aleatorização foi ‘bem sucedida’.³¹⁰

39.3.4 Quais são as razões para diferenças entre grupos de tratamento nas (co)variáveis na linha de base?

- Acaso.^{175,309}
- Viés.^{175,309}
- Tamanho da amostra.^{175,309}
- Má conduta científica.¹⁷⁵

39.3.5 Quais cenários permitem a comparação entre grupos na linha de base em ensaios clínicos aleatorizados?

- Em ensaios clínicos aleatorizados agregados, os P-valores possuem interpretação diferente de estudos aleatorizados individualmente.³¹⁰
- Em ensaios clínicos com agrupamento, nos quais o recrutamento ocorreu após a aleatorização, os P-valores já não estão inteiramente relacionados ao processo de aleatorização, mas sim ao método de recrutamento, o que pode resultar na comparação de amostras não aleatórias.³¹⁰

39.3.6 Por que não se deve comparar grupos na linha de base em ensaios clínicos aleatorizados?

- A interpretação errônea dos P-valores na comparação entre grupos, na linha de base, de um ensaio clínico aleatorizado constitui a ‘falácia da Tabela 1’.¹⁷⁶
- Quando a aleatorização é bem-sucedida, a hipótese nula de diferença entre grupos na linha de base é verdadeira.³¹¹
- Testes de significância estatística na linha de base avaliam a probabilidade de que as diferenças observadas possam ter ocorrido por acaso; no entanto, já sabemos - pelo delineamento do experimento - que quaisquer diferenças são causadas pelo acaso.³¹²

39.3.7 Quais estratégias podem ser adotadas para substituir a comparação entre grupos na linha de base em ensaios clínicos aleatorizados?

- Na fase de projeto: identifique as variáveis prognósticas do desfecho de acordo com a literatura.³¹¹
- Na fase de análise: inclua as variáveis prognósticas nos modelos para ajuste.³¹¹

39.4 Comparação intragrupos

39.4.1 Por que não se deve comparar intragrupos (pré - pós) em ensaios clínicos aleatorizados?

- Testar por mudanças a partir da linha de base separadamente em cada grupo aleatorizado não permite concluir sobre diferenças entre grupos; não se pode fazer inferências a partir da comparação de P-valores.³⁰¹

39.5 Comparação entre grupos

39.5.1 O que é comparação entre grupos em ensaios clínicos aleatorizados?

- A comparação se refere ao teste de hipótese nula de não haver diferença (‘alteração’ ou ‘mudança’) pós-tratamento entre grupos de tratamento.³⁰¹

39.5.2 O que pode ser comparado entre grupos?

- Valores pós-tratamento; mudança entre linha de base e pós-tratamento; mudança percentual da linha de base.³¹³

39.5.3 Qual é a comparação entre grupos mais adequada em ensaios clínicos aleatorizados?

- Análise de covariância (ANCOVA) que analisa o pós-tratamento com a linha de base como covariável é a opção que possui maior poder estatístico.³¹³
- Mudança entre linha de base e pós-tratamento tem poder adequado apenas quando a correlação entre linha de base e pós-tratamento é alta.³¹³
- Mudança percentual da linha de base é a opção menos eficiente em termos de poder estatístico.³¹³

39.6 Comparação de subgrupos

39.6.1 O que é comparação de subgrupos em ensaios clínicos aleatorizados?

- Análises de subgrupos podem ser realizadas para avaliar se as diferenças no resultado do tratamento (ou a falta delas) dependem de algumas características na linha de base dos pacientes.³⁰⁸

39.6.2 Como realizar a comparação de subgrupos em ensaios clínicos aleatorizados?

- Testes estatísticos de interação (que avaliam se um efeito de tratamento difere entre subgrupos) devem ser usados, e não apenas a inspeção dos P-valores do subgrupo. Somente se o teste de interação estatística apoiar um efeito de subgrupo as conclusões poderão ser influenciadas.^{308,314}

39.6.3 Como interpretar a comparação de subgrupos em ensaios clínicos aleatorizados?

- Análises de subgrupos devem ser consideradas de natureza exploratória e raramente elas afetam as conclusões obtidas a partir do estudo.^{308,314}
- A credibilidade das análises de subgrupos é melhor se restrita ao desfecho primário e a alguns subgrupos predefinidos e baseadas em hipóteses biologicamente plausíveis.³⁰⁸
- Deve-se verificar se o estudo possui poder estatístico suficiente para detectar tamanhos de efeitos realistas em subgrupos e interpretar com cautela uma diferença de tratamento em um subgrupo quando a comparação global do tratamento não é significativa.³⁰⁸

39.7 Efeito de interação

39.7.1 Por que analisar o efeito de interação?

- Em ensaios clínicos aleatorizados, o principal problema de pesquisa é se há uma diferença pré - pós maior em um grupo do que em outro(s).³⁰¹
- A comparação de subgrupos por meio de testes de significância de hipótese nula separados é enganosa por não testar (comparar) diretamente os tamanhos dos efeitos dos tratamentos.³¹⁵
- 208

39.7.2 Quando usar o termo de interação?

- Análise de efeito de interação pode ser usada para testar se o efeito de um tratamento varia entre dois ou mais subgrupos de indivíduos, ou seja, se um efeito é modificado pelo(s) outros(s) efeito(s).²⁰⁹
- A interação entre duas (ou mais) variáveis pode ser utilizada para comparar efeitos do tratamento em subgrupos de ensaios clínicos.³¹⁶
- O poder estatístico para detectar efeitos de interação é limitado.³¹⁶

39.8 Ajuste de covariáveis

39.8.1 Quais variáveis devem ser utilizadas no ajuste de covariáveis?

- A escolha das características de linha de base pelas quais uma análise é ajustada deve ser determinada pelo conhecimento prévio de uma possível influência no resultado, em vez da evidência de desequilíbrio entre os grupos de tratamento no estudo.³¹¹

39.8.2 Quais os benefícios do ajuste de covariáveis?

- Ajustar por covariáveis ajuda a estimar os efeitos do tratamento para o indivíduo, assim como aumenta a eficiência dos testes para hipótese nula e a validade externa do estudo.³¹⁷
- Incluir a variável de desfecho medida na linha de base como covariável - independentemente de a análise ser realizada com a medida pós-tratamento da mesma variável ou a diferença para a linha de base - pode aumentar o poder estatístico do estudo.³¹⁸
- Incluir outras variáveis medidas na linha de base, com potencial para serem desbalanceadas entre grupos após a aleatorização, diminui a chance de afetar as estimativas de efeito dos tratamentos.³¹⁸

39.8.3 Quais os riscos do ajuste de covariáveis?

- Incluir covariáveis que não são prognósticas do desfecho pode reduzir o poder estatístico do estudo.³¹⁸
- Incluir covariáveis com dados perdidos pode reduzir o tamanho amostral e consequentemente o poder estatístico do estudo (análise de casos completos) ou levar a desvios do plano de análise por exclusão de covariáveis prognósticas.³¹⁸

39.9 Imputação de dados perdidos

39.9.1 Como lidar com os dados perdidos em desfechos?

- Em dados longitudinais com um pequeno número de ‘ondas’ (medidas repetidas) e poucas variáveis, para análise com modelos de regressão univariados, a imputação paramétrica via especificação condicional completa - também conhecido como imputação multivariada por equações encadeadas (*multivariate imputation by chained equations*, MICE) - é eficiente do ponto de vista computacional e possui acurácia e precisão para estimação de parâmetros.^{310,319}
- Para dados perdidos em desfechos dicotômicos, o desempenho dos métodos de imputação multivariada por equações encadeadas (*multivariate imputation by chained equations*, MICE)³¹⁷ e por correspondência média preditiva (*predictive mean matching*, PMM)^{318,319} é similar.³¹⁶

39.9.2 Como lidar com os dados perdidos em covariáveis?

- Imputação de dados perdidos de uma covariável pela média dos dados do respectivo grupo permite estimativas não enviesadas dos efeitos do tratamento, preserva o erro tipo I e aumenta o poder estatístico comparado à análise de dados completos.³¹⁸

R Os pacotes *mice*³¹⁷ e *miceadds*³¹⁹ fornecem funções *mice*^a e *mi.anova*^b para imputação multivariada por equações encadeadas, respectivamente, para imputação de dados.

^a<https://www.rdocumentation.org/packages/mice/versions/3.16.0/topics/mice>

^b<https://www.rdocumentation.org/packages/miceadds/versions/3.16-18/topics/mi.anova>

39.10 Diretrizes para redação

39.10.1 Quais são as diretrizes para redação de ensaios experimentais?

- Visite a rede *Enhancing the QUAlity and Transparency Of health Research* EQUATOR Network¹ para encontrar diretrizes específicas para cada tipo de ensaio experimental.
- CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials.*³²⁰ <https://www.equator-network.org/reporting-guidelines/consort/>

R

O pacote *consort*³²¹ fornece a função *consort_plot*^a para elaboração do fluxograma de ensaios experimentais no formato padrão.

^a%60r%20paste0(%22https://search.r-project.org/CRAN/refmans/%22,%20%22consort%22,%20%22/html/%22,%20%22consort_plot%22,%20%22.html%22)%60

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

¹<https://www.equator-network.org/>

Capítulo 40

Meta-análise

40.1 Características

40.1.1 O que é meta-análise?

- ?

R

O pacote *metagear*³²² fornece a função *plot_PRISMA*^a para gerar o fluxograma de uma revisão sistemática de acordo com o *Preferred Reporting Items for Systematic Reviews and Meta-Analyses*³²³.

^ahttps://www.rdocumentation.org/packages/metagear/versions/0.7/topics/plot_PRISMA

R

O pacote *PRISMA2020*³²⁴ fornece a função *PRISMA_flowdiagram*^a para elaboração do fluxograma de revisões sistemáticas no formato padrão.

^ahttps://www.rdocumentation.org/packages/PRISMA2020/versions/1.1.1/topics/PRISMA_flowdiagram

40.2 Interpretação de efeitos em meta-análise

40.2.1 Como avaliar a variação do tamanho do efeito?

- O intervalo de predição contém informação sobre a variação do tamanho do efeito.³²⁵
- Se o intervalo de predição não contém a hipótese nula (H_0), podemos concluir que (a) o tratamento funciona igualmente bem em todas as populações, ou que ele funciona melhor em algumas populações do que em outras.³²⁵
- Se o intervalo de predição contém a hipótese nula (H_0), podemos concluir que o tratamento pode ser benéfico em algumas populações, mas prejudicial em outras, de modo que a estimativa pontual (geralmente a média) torna-se amplamente irrelevante. Nesse caso, é recomendado investigar em que populações o tratamento seria benéfico e em quais causaria danos.³²⁵

40.2.2 Como avaliar a heterogeneidade entre os estudos?

- A heterogeneidade - variação não-aleatória - no efeito do tratamento entre os estudos incluídos em uma meta-análise pode ser avaliada pelo I^2 .^{325,326}
- I^2 representa qual proporção da variância observada reflete a variância nos efeitos verdadeiros em vez do erro de amostragem.³²⁵
- I^2 mede a proporção da variância total que pode ser atribuída à heterogeneidade entre os estudos incluídos.³²⁶

- I^2 não depende da quantidade de estudos incluídos na meta-análise. Entretanto, I^2 aumenta com a quantidade de participantes incluídos nos estudos meta-analisados.³²⁶
- A heterogeneidade entre estudos é explicada de modo mais confiável utilizando dados de pacientes individuais, uma vez que a direção verdadeira da modificação de efeito não pode ser observada a partir de dados agregados no estudo.³²⁷

R O pacote *psychmeta*²³⁵ fornece a função *ma_d*^a para meta-analisar valores *d*.

^ahttps://www.rdocumentation.org/packages/psychmeta/versions/2.7.0/topics/ma_d

R O pacote *psychmeta*²³⁵ fornece a função *ma_r*^a para meta-analisar correlações.

^ahttps://www.rdocumentation.org/packages/psychmeta/versions/2.7.0/topics/ma_r

R O pacote *psychmeta*²³⁵ fornece a função *plot_forest*^a para criar figuras tipo *forest plot*.

^ahttps://www.rdocumentation.org/packages/psychmeta/versions/2.7.0/topics/plot_forest

R O pacote *psychmeta*²³⁵ fornece a função *plot_funnel*^a para criar figuras tipo *funnel plot*.

^ahttps://www.rdocumentation.org/packages/psychmeta/versions/2.7.0/topics/plot_funnel

40.3 Diretrizes para redação

40.3.1 Quais são as diretrizes para redação de meta-análises?

- Visite a rede *Enhancing the QUAlity and Transparency Of health Research* EQUATOR Network¹ para encontrar diretrizes específicas para cada tipo de meta-análises.
 - *The PRISMA 2020 statement: An updated guideline for reporting systematic reviews*:³²⁸ <https://www.equator-network.org/reporting-guidelines/prisma/>

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

¹<https://www.equator-network.org/>

Capítulo 41

Simulação computacional

41.1 Características

41.1.1 Quais são as características de estudos de simulação computacional?

- ?

41.2 Método de Monte Carlo

41.2.1 O que é o método de Monte Carlo?

- ?

R

O pacote *base*⁴⁸ fornece a função *set.seed*^a para especificar uma semente para reproduzibilidade de computações que envolvem números aleatórios.

^a<https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/Random>

R

O pacote *simstudy*³²⁹ fornece as funções *defData*^a e *genData*^b para criar variáveis e simular um banco de dados de acordo com o delineamento pré-especificado, respectivamente.

^a<https://www.rdocumentation.org/packages/simstudy/versions/0.7.0/topics/defData>

^b<https://www.rdocumentation.org/packages/simstudy/versions/0.7.0/topics/genData>

R

O pacote *faux*³³⁰ fornece a função *sim_design*^a para simular um banco de dados de acordo com o delineamento pré-especificado.

^ahttps://www.rdocumentation.org/packages/faux/versions/1.2.1/topics/sim_design

R

O pacote *InteractionPower*¹⁹⁸ fornece a função *generate_interaction*^a para simular bancos de dados com efeitos de interação.

^ahttps://www.rdocumentation.org/packages/InteractionPower/versions/0.2.1/topics/generate_interaction

41.3 Diretrizes para redação

41.3.1 Quais são as diretrizes para redação de estudos de simulação computacional?

- Visite a rede *Enhancing the QUAlity and Transparency Of health Research* EQUATOR Network¹ para encontrar diretrizes específicas para cada tipo de estudo de simulação computacional.

– *Reporting Guidelines for Health Care Simulation Research: Extensions to the CONSORT and STROBE Statements*.³³¹ <https://www.equator-network.org/reporting-guidelines/reporting-guidelines-for-health-care-simulation-research-extensions-to-the-consort-and-strobe-statements/>

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

¹<https://www.equator-network.org/>

PARTE 7: APLICAÇÕES E COMUNICAÇÃO

Da análise ao impacto

RASCUNHO

Capítulo 42

Plano de análise

42.1 Plano de análise estatística

42.1.1 O que é plano de análise estatística?

- ?

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

RAASCUNHO

Capítulo 43

Redação de resultados

43.1 Resultados da análise estatística

43.1.1 Como redigir os resultados da análise estatística?

- ?

 O pacote *report*³³² fornece a função *report*^a para redigir a descrição de diversas análises estatísticas.

^a<https://www.rdocumentation.org/packages/report/versions/0.5.8/topics/report>

43.2 Diretrizes e Listas

43.2.1 Quais diretrizes estão disponíveis para redação estatística?

- *Review of guidance papers on regression modeling in statistical series of medical journals.*³³³
- *Principles and recommendations for incorporating estimands into clinical study protocol templates.*³³⁴
- *How to write statistical analysis section in medical research.*²²³
- *Recommendations for Statistical Reporting in Cardiovascular Medicine: A Special Report From the American Heart Association.*³³⁵
- *Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework.*³³⁶
- *Guidelines for reporting of figures and tables for clinical research in urology.*³³⁷
- *Who is in this study, anyway? Guidelines for a useful Table 1.*¹⁷⁷
- *Guidelines for Reporting of Statistics for Clinical Research in Urology.*³³⁸
- *Reveal, Don't Conceal: Transforming Data Visualization to Improve Transparency.*¹⁸³
- *Guidelines for the Content of Statistical Analysis Plans in Clinical Trials.*³³⁹
- *Basic statistical reporting for articles published in Biomedical Journals: The 'Statistical Analyses and Methods in the Published Literature' or the SAMPL Guidelines.*³⁴⁰
- *Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm.*³⁴¹
- *STRengthening analytical thinking for observational studies: the STRATOS initiative.*³⁴²
- *Research methods and reporting.*³⁴³
- *How to ensure your paper is rejected by the statistical reviewer.*³⁴⁴

43.2.2 Quais listas de verificação estão disponíveis para redação estatística?

- *A Checklist for statistical Assessment of Medical Papers (the CHAMP statement): explanation and elaboration.*³⁴⁵
- *Checklist for clinical applicability of subgroup analysis.*³⁴⁶
- *Evidence-based statistical analysis and methods in biomedical research (SAMBR) checklists according to design features.*²²²

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

PARTE 8: RECURSOS

Material complementar

RASCUNHO

Capítulo 44

Shiny Apps

Aplicativos por delineamento de estudo

Ensaios clínicos

- *RCTapp*¹

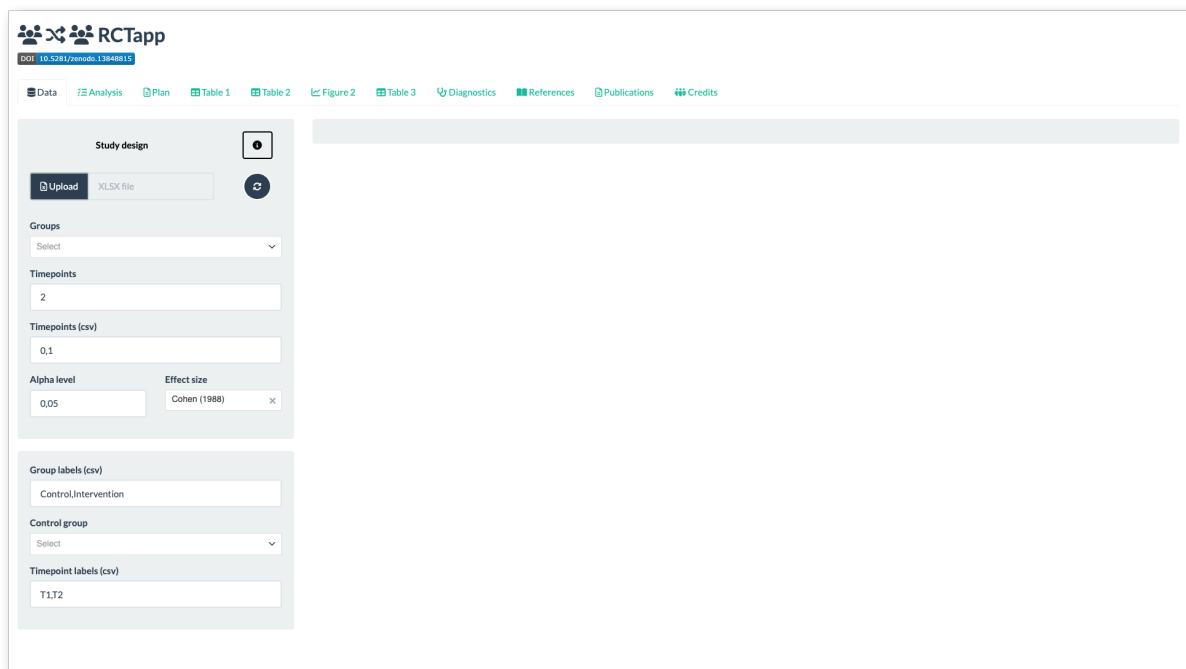


Figura 44.1: RCTapp: Shiny app para análise de ensaios clínicos aleatorizados.

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

¹<https://scicrusher.shinyapps.io/RCTapp/>

RAASCUNHO

Capítulo 45

Fontes externas

45.1 Fontes de informação externas

45.1.1 American Heart Association

- *Statistical Reporting Recommendations - AHA/ASA journals*¹

45.1.2 American Physiological Society

- *Statistics*²
- *Exploration in Statistics*³
- *General Statistics*⁴
- *Reporting Statistics*⁵

45.1.3 American Statistical Association

- *Statistical Inference in the 21st Century: A World Beyond $p < 0.05$* - The American Statistical Association⁶

45.1.4 British Medicine Journal

- *Statistics - Latest from The BMJ*⁷
- *Statistics notes - Latest from The BMJ*⁸
- *Statistics and research methods - Latest from The BMJ*⁹
- *Statistics at Square One*¹⁰
- *Research methods & reporting*¹¹

45.1.5 Enhancing the Quality And Transparency Of health Research Network

- *Enhancing the Quality and Transparency of health research* EQUATOR Network¹²

¹<https://www.ahajournals.org/statistical-recommendations>

²<https://journals.physiology.org/topic/advances-collections/statistics?seriesKey=&tagCode=&>

³<https://journals.physiology.org/topic/advances-collections/explorations-in-statistics?seriesKey=&tagCode=&>

⁴<https://journals.physiology.org/topic/advances-collections/general-statistics?seriesKey=&tagCode=&>

⁵<https://journals.physiology.org/topic/advances-collections/reporting-statistics?seriesKey=&tagCode=&>

⁶<https://www.tandfonline.com/toc/utas20/73/sup1?nav=tocList>

⁷<https://www.bmj.com/specialties/statistics>

⁸<https://www.bmj.com/specialties/statistics-notes>

⁹<https://www.bmj.com/specialties/statistics-and-research-methods>

¹⁰<https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one>

¹¹<https://www.bmj.com/research/research-methods-and-reporting>

¹²<https://www.equator-network.org>

45.1.6 Journal of the American Medical Association

- *JAMA Guide to Statistics and Methods - JAMA*¹³

45.1.7 Nature Publishing Group

- *Statistics for Biologists - Nature Publishing Group*¹⁴

45.1.8 Oxford Reference

- *A Dictionary of Statistics*¹⁵

45.1.9 Royal Statistical Society

- *Best Practices for Data Visualisation - Royal Statistical Society*¹⁶

45.1.10 Statistics in Medicine

- *Tutorials in Biostatistics Papers*¹⁷

45.1.11 BMC Trials

- *Design and analysis of n-of-1 trials*¹⁸

45.1.12 The Journal of Applied Statistics in the Pharmaceutical Industry

- *Tutorial Papers*¹⁹

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

¹³<https://jamanetwork.com/collections/44042/jama-guide-to-statistics-and-methods>

¹⁴<https://www.nature.com/collections/qghhqm>

¹⁵<https://www.oxfordreference.com/display/10.1093/acref/9780199679188.001.0001/acref-9780199679188>

¹⁶<https://royal-statistical-society.github.io/datavisguide>

¹⁷<https://onlinelibrary.wiley.com/page/journal/10970258/homepage/tutorials.htm>

¹⁸<https://www.biomedcentral.com/collections/DANT>

¹⁹https://onlinelibrary.wiley.com/page/journal/15391612/homepage/tutorial_papers.htm

Capítulo 46

Diretrizes e Listas

46.1 Diretrizes

46.1.1 Quais são as diretrizes para relatórios estatísticos em pesquisas?

- *Review of guidance papers on regression modeling in statistical series of medical journals.*³³³
- *Principles and recommendations for incorporating estimands into clinical study protocol templates.*³³⁴
- *How to write statistical analysis section in medical research.*²²³
- *A Guideline for Reporting Mediation Analyses of Randomized Trials and Observational Studies: The AGReMA Statement.*³⁴⁷
- *Recommendations for Statistical Reporting in Cardiovascular Medicine: A Special Report From the American Heart Association.*³³⁵
- *Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework.*³³⁶
- *Guidelines for reporting of figures and tables for clinical research in urology.*³³⁷
- *Who is in this study, anyway? Guidelines for a useful Table 1.*¹⁷⁷
- *Guidelines for Reporting of Statistics for Clinical Research in Urology.*³³⁸
- *Reveal, Don't Conceal: Transforming Data Visualization to Improve Transparency.*¹⁸³
- *Guidelines for the Content of Statistical Analysis Plans in Clinical Trials.*³³⁹
- *Basic statistical reporting for articles published in Biomedical Journals: The 'Statistical Analyses and Methods in the Published Literature' or the SAMPL Guidelines.*³⁴⁰
- *Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm.*³⁴¹
- *STRengthening analytical thinking for observational studies: the STRATOS initiative.*³⁴²
- *Research methods and reporting.*³⁴³
- *How to ensure your paper is rejected by the statistical reviewer.*³⁴⁴

46.2 Listas de verificação

46.2.1 Quais são as listas de verificação para relatórios estatísticos em pesquisas?

- *A CHecklist for statistical Assessment of Medical Papers (the CHAMP statement): explanation and elaboration.*³⁴⁵
- *Checklist for clinical applicability of subgroup analysis.*³⁴⁶

- *Evidence-based statistical analysis and methods in biomedical research (SAMBR) checklists according to design features.*²²²

Citar como: Ferreira, Arthur de Sá. Ciência com R: Perguntas e respostas para pesquisadores e analistas de dados. Rio de Janeiro: 1a edição,

RAASCUNHO

Referências

1. Grami A. Discrete probability. In: Elsevier; 2023:285-305. doi:10.1016/b978-0-12-820656-0.00016-2
2. Viti A, Terzi A, Bertolaccini L. A practical overview on probability distributions. *Journal of Thoracic Disease*. 2015;7(3). <https://jtd.amegroups.org/article/view/4086>.
3. Benford F. The law of anomalous numbers. *Proceedings of the American Philosophical Society*. 1938;78(4):551-572. <http://www.jstor.org/stable/984802>. Accessed November 24, 2024.
4. Tversky A, Kahneman D. Belief in the law of small numbers. *Psychological Bulletin*. 1971;76(2):105-110. doi:10.1037/h0031322
5. Bishop DVM, Thompson J, Parker AJ. Can we shift belief in the ‘Law of Small Numbers’? *Royal Society Open Science*. 2022;9(3). doi:10.1098/rsos.211028
6. Guy RK. The strong law of small numbers. *The American Mathematical Monthly*. 1988;95(8):697. doi:10.2307/2322249
7. Guy RK. The Second Strong Law of Small Numbers. *Mathematics Magazine*. 1990;63(1):3-20. doi:10.1080/0025570x.1990.11977475
8. Kwak SG, Kim JH. Central limit theorem: the cornerstone of modern statistics. *Korean Journal of Anesthesiology*. 2017;70(2):144. doi:10.4097/kjae.2017.70.2.144
9. Galton F. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*. 1886;15:246. doi:10.2307/2841583
10. Barnett AG. Regression to the mean: what it is and how to deal with it. *International Journal of Epidemiology*. 2004;34(1):215-220. doi:10.1093/ije/dyh299
11. Senn S. Francis Galton and Regression to the Mean. *Significance*. 2011;8(3):124-126. doi:10.1111/j.1740-9713.2011.00509.x
12. Recchia D. *Regtomean: Regression Toward the Mean.*; 2022. <https://CRAN.R-project.org/package=regtomean>.
13. Banerjee A, Chaudhury S. Statistics without tears: Populations and samples. *Industrial Psychiatry Journal*. 2010;19(1):60. doi:10.4103/0972-6748.77642
14. Bland JM, Altman DG. Statistics Notes: Bootstrap resampling methods. *BMJ*. 2015;350(jun02 13):h2622-h2622. doi:10.1136/bmj.h2622
15. Altman DG, Bland JM. Statistics Notes: Units of analysis. *BMJ*. 1997;314(7098):1874-1874. doi:10.1136/bmj.314.7098.1874

16. Matthews JN, Altman DG, Campbell MJ, Royston P. Analysis of serial measurements in medical research. *BMJ*. 1990;300(6719):230-235. doi:10.1136/bmj.300.6719.230
17. Amatuzzi MLL, Barreto M do CC, Litvoc J, Leme LEG. Linguagem metodológica: Parte 1. *Acta Ortopédica Brasileira*. 2006;14(1):53-56. doi:10.1590/s1413-78522006000100012
18. Amatuzzi MLL, Barreto M do CC, Litvoc J, Leme LEG. Linguagem metodológica: Parte 2. *Acta Ortopédica Brasileira*. 2006;14(2):108-112. doi:10.1590/s1413-78522006000200012
19. Munafò MR, Nosek BA, Bishop DVM, et al. A manifesto for reproducible science. *Nature Human Behaviour*. 2017;1(1). doi:10.1038/s41562-016-0021
20. Resnik DB, Shamoo AE. Reproducibility and Research Integrity. *Accountability in Research*. 2016;24(2):116-123. doi:10.1080/08989621.2016.1257387
21. Hofner B, Schmid M, Edler L. Reproducible research in statistics: A review and guidelines for the *Biometrical Journal*. *Biometrical Journal*. 2015;58(2):416-427. doi:10.1002/bimj.201500156
22. Mair P. Thou shalt be reproducible! A technology perspective. *Frontiers in Psychology*. 2016;7. doi:10.3389/fpsyg.2016.01079
23. Ihaka R, Gentleman R. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*. 1996;5(3):299. doi:10.2307/1390807
24. Introduction to r and RStudio. *Practical Machine Learning in R*. April 2020:25-52. doi:10.1002/9781119591542.ch2
25. R Core Team. The comprehensive r archive network. 2021. <https://cran.r-project.org>.
26. Racine JS. RStudio: A Platform-Independent IDE for R and Sweave. *Journal of Applied Econometrics*. 2011;27(1):167-172. doi:10.1002/jae.1278
27. Aden-Buie G, Schloerke B, Allaire J, Rossell Hayes A. *Learnr: Interactive Tutorials for r*; 2023. <https://CRAN.R-project.org/package=learnr>.
28. Love J, Selker R, Marsman M, et al. **JASP**: Graphical Statistical Software for Common Statistical Designs. *Journal of Statistical Software*. 2019;88(2). doi:10.18637/jss.v088.i02
29. SAHİN M, AYBEK E. Jamovi: An easy to use statistical software for the social scientists. *International Journal of Assessment Tools in Education*. 2020;6(4):670-692. doi:10.21449/ijate.661803
30. Selker R, Love J, Dropmann D. *Jmv: The {Jamovi} Analyses*; 2023. <https://CRAN.R-project.org/package=jmv>.
31. Love J. *Jmvconnect: Connect to the {Jamovi} Statistical Spreadsheet*; 2022. <https://CRAN.R-project.org/package=jmvconnect>.
32. Hinsen K. A data and code model for reproducible research and executable papers. *Procedia Computer Science*. 2011;4:579-588. doi:10.1016/j.procs.2011.04.061
33. All r CRAN packages [full list]. 2025. <https://r-packages.io/packages>. Accessed February 11, 2025.
34. R Core Team. R: A language and environment for statistical computing. 2023. <https://www.R-project.org/>.
35. Schwab, Simon, Held, Leonhard. Statistical programming: Small mistakes, big impacts. *Wiley-Blackwell Publishing, Inc*. 2021. doi:10.5167/UZH-205154

36. Eglen SJ, Marwick B, Halchenko YO, et al. Toward standard practices for sharing computer code and programs in neuroscience. *Nature Neuroscience*. 2017;20(6):770-773. doi:10.1038/nn.4550
37. Xie Y. *formatR: Format r Code Automatically.*; 2022. <https://CRAN.R-project.org/package=formatR>.
38. Müller K, Walthert L. *Styler: Non-Invasive Pretty Printing of r Code.*; 2023. <https://CRAN.R-project.org/package=styler>.
39. Hester J, Angly F, Hyde R, et al. *Lintr: A {Linter} for r Code.*; 2023. <https://CRAN.R-project.org/package=lintr>.
40. Trisovic A, Lau MK, Pasquier T, Crosas M. A large-scale study on research code quality and execution. *Scientific Data*. 2022;9(1). doi:10.1038/s41597-022-01143-6
41. Allaire J, Xie Y, Dervieux C, et al. *Rmarkdown: Dynamic Documents for r.*; 2023. <https://CRAN.R-project.org/package=rmarkdown>.
42. Gohel D, Ross N. *Officedown: Enhanced {r Markdown} Format for {Word} and {PowerPoint}.*; 2023. <https://CRAN.R-project.org/package=officedown>.
43. Xie Y. *Bookdown: Authoring books and technical documents with r markdown.* 2023. <https://github.com/rstudio/bookdown>.
44. Holmes DT, Mobini M, McCudden CR. Reproducible manuscript preparation with RMarkdown application to JMSACL and other Elsevier Journals. *Journal of Mass Spectrometry and Advances in the Clinical Lab*. 2021;22:8-16. doi:10.1016/j.jmsacl.2021.09.002
45. Ioannidis JPA. How to Make More Published Research True. *PLoS Medicine*. 2014;11(10):e1001747. doi:10.1371/journal.pmed.1001747
46. Krieger N, Perzynski A, Dalton J. *Projects: A Project Infrastructure for Researchers.*; 2021. <https://CRAN.R-project.org/package=projects>.
47. Schultze A, Tazare J. The role of programming code sharing in improving the transparency of medical research. *BMJ*. October 2023:p2402. doi:10.1136/bmj.p2402
48. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2023. <https://www.R-project.org/>.
49. Zhao Y, Xiao N, Anderson K, Zhang Y. Electronic common technical document submission with analysis using R. *Clinical Trials*. 2022;20(1):89-92. doi:10.1177/17407745221123244
50. Francisco Rodríguez-Sánchez, Connor P. Jackson, Shaurita D. Hutchins. Grateful: Facilitate citation of r packages. 2023. <https://github.com/Pakillo/grateful>.
51. Meng XL. Statistical paradises and paradoxes in big data (i): Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics*. 2018;12(2). doi:10.1214/18-aoas1161sf
52. Abelson RP. A variance explanation paradox: When a little is a lot. *Psychological Bulletin*. 1985;97(1):129-133. doi:10.1037/0033-2909.97.1.129
53. Berkson J. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*. 1946;2(3):47. doi:10.2307/3002000
54. Ellsberg D. Risk, ambiguity, and the savage axioms. *The Quarterly Journal of Economics*. 1961;75(4):643. doi:10.2307/1884324

55. Freedman DA, Freedman DA. A Note on Screening Regression Equations. *The American Statistician*. 1983;37(2):152-155. doi:10.1080/00031305.1983.10482729
56. Freedman LS, Pee D. Return to a note on screening regression equations. *The American Statistician*. 1989;43(4):279. doi:10.2307/2685389
57. Hand DJ. On Comparing Two Treatments. *The American Statistician*. 1992;46(3):190-192. doi:10.1080/00031305.1992.10475881
58. LINDLEY DV. A STATISTICAL PARADOX. *Biometrika*. 1957;44(1-2):187-192. doi:10.1093/biomet/44.1-2.187
59. Lord FM. A paradox in the interpretation of group comparisons. *Psychological Bulletin*. 1967;68(5):304-305. doi:10.1037/h0025105
60. Lord FM. Statistical adjustments when comparing preexisting groups. *Psychological Bulletin*. 1969;72(5):336-337. doi:10.1037/h0028108
61. Simpson EH. The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1951;13(2):238-241. doi:10.1111/j.2517-6161.1951.tb00088.x
62. Blyth CR. On Simpson's Paradox and the Sure-Thing Principle. *Journal of the American Statistical Association*. 1972;67(338):364-366. doi:10.1080/01621459.1972.10482387
63. Pearl J. Comment: Understanding Simpson's Paradox. *The American Statistician*. 2014;68(1):8-13. doi:10.1080/00031305.2014.876829
64. Stein C. INADMISSIBILITY OF THE USUAL ESTIMATOR FOR THE MEAN OF a MULTIVARIATE NORMAL DISTRIBUTION. In: University of California Press; 1956:197-206. doi:10.1525/9780520313880-018
65. De S, Sen A. The generalised Gamow-Stern problem. *The Mathematical Gazette*. 1996;80(488):345-348. doi:10.2307/3619568
66. Feld SL. Why Your Friends Have More Friends Than You Do. *American Journal of Sociology*. 1991;96(6):1464-1477. doi:10.1086/229693
67. Polin BA, Benisaac E. A longitudinal analysis of the hot hand and gambler's fallacy biases. *Judgment and Decision Making*. 2023;18. doi:10.1017/jdm.2023.23
68. Aguinis H, Pierce CA, Culpepper SA. Scale Coarseness as a Methodological Artifact. *Organizational Research Methods*. 2008;12(4):623-652. doi:10.1177/1094428108318065
69. Bryer J, Speerschneider K. *Likert: Analysis and Visualization Likert Items.*; 2016. <https://CRAN.R-project.org/package=likert>.
70. Ferris TLJ. A new definition of measurement. *Measurement*. 2004;36(1):101-109. doi:10.1016/j.measurement.2004.03.001
71. R Core Team. *R: A Language and Environment for Statistical Computing*.; 2023. <https://www.R-project.org/>.
72. Healy MJR, Goldstein H. Regression to the mean. *Annals of Human Biology*. 1978;5(3):277-280. doi:10.1080/03014467800002891
73. Altman DG, Bland JM. Measurement in medicine: The analysis of method comparison studies. *The Statistician*. 1983;32(3):307. doi:10.2307/2987937

74. Olson K. What Are Data? *Qualitative Health Research.* 2021;31(9):1567-1569. doi:10.1177/10497323211015960
75. Smeden M van. A very short list of common pitfalls in research design, data analysis, and reporting. *PRIMER.* 2022;6. doi:10.22454/PRIMER.2022.511416
76. Vetter TR. Fundamentals of Research Data and Variables. *Anesthesia & Analgesia.* 2017;125(4):1375-1380. doi:10.1213/ane.0000000000002370
77. Baillie M, Cessie S le, Schmidt CO, Lusa L, Huebner M. Ten simple rules for initial data analysis. *PLOS Computational Biology.* 2022;18(2):e1009819. doi:10.1371/journal.pcbi.1009819
78. Buttliere B. Adopting standard variable labels solves many of the problems with sharing and reusing data. *Methodological Innovations.* 2021;14(2):205979912110266. doi:10.1177/20597991211026616
79. Pebesma E, Mailund T, Hiebert J. Measurement units in {r}. 2016;8. doi:10.32614/RJ-2016-061
80. Firke S. *Janitor: Simple Tools for Examining and Cleaning Dirty Data.*; 2023. <https://CRAN.R-project.org/package=janitor>.
81. Harrell Jr FE. *Hmisc: Harrell Miscellaneous.*; 2023. <https://CRAN.R-project.org/package=Hmisc>.
82. Tierney N, Cook D. Expanding Tidy Data Principles to Facilitate Missing Data Exploration, Visualization and Assessment of Imputations. *Journal of Statistical Software.* 2023;105(7). doi:10.18637/jss.v105.i07
83. Hammill D. *DataEditR: An Interactive Editor for Viewing, Entering, Filtering & Editing Data.*; 2022. <https://CRAN.R-project.org/package=DataEditR>.
84. Broman KW, Woo KH. Data Organization in Spreadsheets. *The American Statistician.* 2018;72(1):2-10. doi:10.1080/00031305.2017.1375989
85. Juluru K, Eng J. Use of Spreadsheets for Research Data Collection and Preparation: *Academic Radiology.* 2015;22(12):1592-1599. doi:10.1016/j.acra.2015.08.024
86. Dowle M, Srinivasan A. *Data.table: Extension of 'Data.frame'.*; 2023. <https://CRAN.R-project.org/package=data.table>.
87. Altman DG, Bland JM. Statistics notes Variables and parameters. *BMJ.* 1999;318(7199):1667-1667. doi:10.1136/bmj.318.7199.1667
88. Ali Z, Bhaskar Sb. Basic statistical tools in research and data analysis. *Indian Journal of Anaesthesia.* 2016;60(9):662. doi:10.4103/0019-5049.190623
89. Dettori JR, Norvell DC. The Anatomy of Data. *Global Spine Journal.* 2018;8(3):311-313. doi:10.1177/2192568217746998
90. Kaliyadan F, Kulkarni V. Types of variables, descriptive statistics, and sample size. *Indian Dermatology Online Journal.* 2019;10(1):82. doi:10.4103/idoj.idoj_468_18
91. Barkan H. Statistics in clinical research: Important considerations. *Annals of Cardiac Anaesthesia.* 2015;18(1):74. doi:10.4103/0971-9784.148325
92. Bland JM, Altman DG. Statistics Notes: Transforming data. *BMJ.* 1996;312(7033):770-770. doi:10.1136/bmj.312.7033.770
93. Fedorov V, Mannino F, Zhang R. Consequences of dichotomization. *Pharmaceutical Statistics.* 2009;8(1):50-61. doi:10.1002/pst.331

94. Osborne J. Improving your data transformations: Applying the box-cox transformation. *University of Massachusetts Amherst*. 2010. doi:10.7275/QBPC-GK17
95. Box GEP, Cox DR. An Analysis of Transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1964;26(2):211-243. doi:10.1111/j.2517-6161.1964.tb00553.x
96. Venables WN, Ripley BD. Modern applied statistics with s. 2002. <https://www.stats.ox.ac.uk/pub/MASS4/>.
97. MacCallum RC, Zhang S, Preacher KJ, Rucker DD. On the practice of dichotomization of quantitative variables. *Psychological Methods*. 2002;7(1):19-40. doi:10.1037/1082-989x.7.1.19
98. Altman DG, Royston P. The cost of dichotomising continuous variables. *BMJ*. 2006;332(7549):1080.1. doi:10.1136/bmj.332.7549.1080
99. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine*. 2005;25(1):127-141. doi:10.1002/sim.2331
100. Collins GS, Ogundimu EO, Cook JA, Manach YL, Altman DG. Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. *Statistics in Medicine*. 2016;35(23):4124-4135. doi:10.1002/sim.6986
101. Nelson SLP, Ramakrishnan V, Nietert PJ, Kamen DL, Ramos PS, Wolf BJ. An evaluation of common methods for dichotomization of continuous variables to discriminate disease status. *Communications in Statistics – Theory and Methods*. 2017;46(21):10823-10834. doi:10.1080/03610926.2016.1248783
102. Bennette C, Vickers A. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. *BMC Medical Research Methodology*. 2012;12(1). doi:10.1186/1471-2288-12-21
103. Barnier J, Briatte F, Larmarange J. *Questionr: Functions to Make Surveys Processing Easier*; 2023. <https://CRAN.R-project.org/package=questionr>.
104. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3(1):32-35. doi:10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3
105. Strobl C, Boulesteix AL, Augustin T. Unbiased split selection for classification trees based on the Gini Index. *Computational Statistics & Data Analysis*. 2007;52(1):483-501. doi:10.1016/j.csda.2006.12.030
106. Pearson K. X. *On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*. 1900;50(302):157-175. doi:10.1080/14786440009463897
107. Greiner M, Pfeiffer D, Smith RD. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Preventive Veterinary Medicine*. 2000;45(1-2):23-41. doi:10.1016/s0167-5877(00)00115-x
108. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychological Bulletin*. 1971;76(5):378-382. doi:10.1037/h0031619
109. Altman DG, Bland JM. Missing data. *BMJ*. 2007;334(7590):424-424. doi:10.1136/bmj.38977.682025.2c
110. Heymans MW, Twisk JWR. Handling missing data in clinical research. *Journal of Clinical Epidemiology*. September 2022. doi:10.1016/j.jclinepi.2022.08.016
111. Carpenter JR, Smuk M. Missing data: A statistical framework for practice. *Biometrical Journal*. 2021;63(5):915-947. doi:10.1002/bimj.202000196

112. Yanagida T. *Misty: Miscellaneous Functions.*; 2023. <https://CRAN.R-project.org/package=misty>.
113. Little RJA. A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association*. 1988;83(404):1198-1202. doi:10.1080/01621459.1988.10478722
114. Tierney N, Cook D. Expanding tidy data principles to facilitate missing data exploration, visualization and assessment of imputations. 2023;105. doi:10.18637/jss.v105.i07
115. Akl EA, Shawwa K, Kahale LA, et al. Reporting missing participant data in randomised trials: systematic survey of the methodological literature and a proposed guide. *BMJ Open*. 2015;5(12):e008431. doi:10.1136/bmjopen-2015-008431
116. Austin PC, Buuren S van. Logistic regression vs. predictive mean matching for imputing binary covariates. *Statistical Methods in Medical Research*. September 2023. doi:10.1177/09622802231198795
117. Buuren S van, Groothuis-Oudshoorn K. {Mice}: Multivariate imputation by chained equations in r. 2011;45:1-67. doi:10.18637/jss.v045.i03
118. Rubin DB. Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*. 1986;4(1):87. doi:10.2307/1391390
119. Little RJA. Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics*. 1988;6(3):287-296. doi:10.1080/07350015.1988.10509663
120. Robitzsch A, Grund S. *Miceadds: Some Additional Multiple Imputation Functions, Especially for {Mice}.*; 2023. <https://CRAN.R-project.org/package=miceadds>.
121. FitzJohn R. *Ids: Generate Random Identifiers.*; 2017. <https://CRAN.R-project.org/package=ids>.
122. Brown C. *Hash: Full Featured Implementation of Hash Tables/Associative Arrays/Dictionaries.*; 2023. <https://CRAN.R-project.org/package=hash>.
123. Hendricks P. Anonymizer: Anonymize data containing personally identifiable information. 2023. <https://github.com/paulhendricks/anonymizer>.
124. Lucas DE with contributions by A, Tuszyński J, Bengtsson H, et al. *Digest: Create Compact Hash Digests of r Objects.*; 2023. <https://CRAN.R-project.org/package=digest>.
125. Nowok B, Raab GM, Dibben C. {Synthpop}: Bespoke creation of synthetic data in {r}. 2016;74. doi:10.18637/jss.v074.i11
126. Krasser R. *Explore: Simplifies Exploratory Data Analysis.*; 2023. <https://CRAN.R-project.org/package=explore>.
127. Petersen AH, Ekstrøm CT. {dataMaid}: Your assistant for documenting supervised data quality screening in {r}. 2019;90. doi:10.18637/jss.v090.i06
128. Cui B. *DataExplorer: Automate Data Exploration and Treatment.*; 2020. <https://CRAN.R-project.org/package=DataExplorer>.
129. Dayanand Ubrangala, R K, Prasad Kondapalli R, Putatunda S. *SmartEDA: Summarize and Explore the Data.*; 2022. <https://CRAN.R-project.org/package=SmartEDA>.
130. Meyer F, Perrier V. *Esquisse: Explore and Visualize Your Data Interactively.*; 2022. <https://CRAN.R-project.org/package=esquisse>.

131. Chatfield C. Exploratory data analysis. *European Journal of Operational Research*. 1986;23(1):5-13. doi:10.1016/0377-2217(86)90209-2
132. Ferketich S, Verran J. Technical Notes. *Western Journal of Nursing Research*. 1986;8(4):464-466. doi:10.1177/019394598600800409
133. Kerr NL. HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*. 1998;2(3):196-217. doi:10.1207/s15327957pspr0203_4
134. Landis SC, Amara SG, Asadullah K, et al. A call for transparent reporting to optimize the predictive value of preclinical research. *Nature*. 2012;490(7419):187-191. doi:10.1038/nature11556
135. Huebner M, Vach W, Cessie S le. A systematic approach to initial data analysis is good research practice. *The Journal of Thoracic and Cardiovascular Surgery*. 2016;151(1):25-27. doi:10.1016/j.jtcvs.2015.09.085
136. S M. Frequency distribution. *Journal of Pharmacology and Pharmacotherapeutics*. 2011;2(1):54-56. doi:10.4103/0976-500x.77120
137. Sturges HA. The Choice of a Class Interval. *Journal of the American Statistical Association*. 1926;21(153):65-66. doi:10.1080/01621459.1926.10502161
138. SCOTT DW. On optimal and data-based histograms. *Biometrika*. 1979;66(3):605-610. doi:10.1093/biomet/66.3.605
139. Freedman D, Diaconis P. On the histogram as a density estimator:L 2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*. 1981;57(4):453-476. doi:10.1007/bf01025868
140. R Core Team. R: A language and environment for statistical computing. 2024. <https://www.R-project.org/>.
141. R Core Team. R: A language and environment for statistical computing. 2023. <https://www.R-project.org/>.
142. Kay M. {Ggdist}: Visualizations of distributions and uncertainty in the grammar of graphics. 2024;30. doi:10.1109/TVCG.2023.3327195
143. Tang Y, Horikoshi M, Li W. *Ggfortify: Unified Interface to Visualize Statistical Result of Popular r Packages*. Vol 8.; 2016. doi:10.32614/RJ-2016-060
144. Rochon J, Gondan M, Kieser M. To test or not to test: Preliminary assessment of normality when comparing two independent samples. *BMC Medical Research Methodology*. 2012;12(1). doi:10.1186/1471-2288-12-81
145. Greenhalgh T. How to read a paper: Statistics for the non-statistician. I: Different types of data need different statistical tests. *BMJ*. 1997;315(7104):364-366. doi:10.1136/bmj.315.7104.364
146. Kanji G. *100 Statistical Tests*; 2006. doi:10.4135/9781849208499
147. Curran-Everett D. Explorations in statistics: standard deviations and standard errors. *Advances in Physiology Education*. 2008;32(3):203-208. doi:10.1152/advan.90123.2008
148. Altman DG, Bland JM. Statistics Notes: Quartiles, quintiles, centiles, and other quantiles. *BMJ*. 1994;309(6960):996-996. doi:10.1136/bmj.309.6960.996
149. S. M. Measures of central tendency: The mean. *Journal of Pharmacology and Pharmacotherapeutics*. 2011;2(2):140-142. doi:10.4103/0976-500x.81920
150. S. M. Measures of central tendency: Median and mode. *Journal of Pharmacology and Pharmacotherapeutics*. 2011;2(3):214-215. doi:10.4103/0976-500x.83300

151. Krzywinski M, Altman N. Error bars. *Nature Methods*. 2013;10(10):921-922. doi:10.1038/nmeth.2659
152. Manikandan S. Measures of dispersion. *Journal of Pharmacology and Pharmacotherapeutics*. 2011;2(4):315-316. doi:10.4103/0976-500x.85931
153. Cumming G, Fidler F, Vaux DL. Error bars in experimental biology. *The Journal of Cell Biology*. 2007;177(1):7-11. doi:10.1083/jcb.200611141
154. Sahai H, Misra S. Definitions of sample variance: Some teaching problems to be overcome. *The Statistician*. 1992;41(1):55. doi:10.2307/2348636
155. Leys C, Delacre M, Mora YL, Lakens D, Ley C. How to Classify, Detect, and Manage Univariate and Multivariate Outliers, With Emphasis on Pre-Registration. *International Review of Social Psychology*. 2019;32(1). doi:10.5334/irsp.289
156. Zuur AF, Ieno EN, Elphick CS. A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*. 2009;1(1):3-14. doi:10.1111/j.2041-210x.2009.00001.x
157. Leys C, Ley C, Klein O, Bernard P, Licata L. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*. 2013;49(4):764-766. doi:10.1016/j.jesp.2013.03.013
158. Leys C, Klein O, Dominicy Y, Ley C. Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. *Journal of Experimental Social Psychology*. 2018;74:150-156. doi:10.1016/j.jesp.2017.09.011
159. Tukey JW, McLaughlin DH. Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/winsorization 1. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*. 1963;25(3):331-352. <http://www.jstor.org/stable/25049278>. Accessed April 11, 2025.
160. Komsta L. *Outliers: Tests for Outliers*; 2022. <https://CRAN.R-project.org/package=outliers>.
161. Mock T. *gtExtras: Extending {Gt} for Beautiful HTML Tables*; 2023. <https://CRAN.R-project.org/package=gtExtras>.
162. Nijs V. *Radiant: Business Analytics Using r and Shiny*; 2023. <https://CRAN.R-project.org/package=radiant>.
163. Gerring J. Mere Description. *British Journal of Political Science*. 2012;42(4):721-746. doi:10.1017/s0007123412000130
164. Cummings P, Rivara FP. Reporting Statistical Information in Medical Journal Articles. *Archives of Pediatrics & Adolescent Medicine*. 2003;157(4):321. doi:10.1001/archpedi.157.4.321
165. Cole TJ. Setting number of decimal places for reporting risk ratios: rule of four. *BMJ*. 2015;350(apr27 3):h1845-h1845. doi:10.1136/bmj.h1845
166. Cole TJ. Too many digits: the presentation of numerical data. *Archives of Disease in Childhood*. 2015;100(7):608-609. doi:10.1136/archdischild-2014-307149
167. Inskip H, Ntani G, Westbury L, et al. Getting started with tables. *Archives of Public Health*. 2017;75(1). doi:10.1186/s13690-017-0180-1
168. Kwak SG, Kang H, Kim JH, et al. The principles of presenting statistical results: Table. *Korean Journal of Anesthesiology*. 2021;74(2):115-119. doi:10.4097/kja.20582

169. Barnett A. Automated detection of over- and under-dispersion in baseline tables in randomised controlled trials. *F1000Research*. 2023;11:783. doi:10.12688/f1000research.123002.2
170. Gohel D, Skintzos P. *Flextable: Functions for Tabular Reporting.*; 2023. <https://CRAN.R-project.org/package=flextable>.
171. Thériault R. {Rempscy}: Convenience functions for psychology. 2023;8:5466. doi:10.21105/joss.05466
172. Rich B. *Table1: Tables of Descriptive Statistics in HTML.*; 2023. <https://CRAN.R-project.org/package=table1>.
173. Sjoberg DD, Whiting K, Curry M, Lavery JA, Larmarange J. Reproducible summary tables with the gtsummary package. 2021;13:570-580. doi:10.32614/RJ-2021-053
174. Westreich D, Greenland S. The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients. *American Journal of Epidemiology*. 2013;177(4):292-298. doi:10.1093/aje/kws412
175. Chen H, Lu Y, Slye N. Testing for baseline differences in clinical trials. *International Journal of Clinical Trials*. 2020;7(2):150. doi:10.18203/2349-3259.ijct20201720
176. Pijls BG. The Table I Fallacy: P Values in Baseline Tables of Randomized Controlled Trials. *Journal of Bone and Joint Surgery*. 2022;104(16):e71. doi:10.2106/jbjs.21.01166
177. Hayes-Larson E, Kezios KL, Mooney SJ, Lovasi G. Who is in this study, anyway? Guidelines for a useful Table 1. *Journal of Clinical Epidemiology*. 2019;114:125-132. doi:10.1016/j.jclinepi.2019.06.011
178. Bandoli G, Palmsten K, Chambers CD, Jelliffe-Pawlowski LL, Baer RJ, Thompson CA. Revisiting the Table 2 fallacy: A motivating example examining preeclampsia and preterm birth. *Paediatric and Perinatal Epidemiology*. 2018;32(4):390-397. doi:10.1111/ppe.12474
179. Park JH, Lee DK, Kang H, et al. The principles of presenting statistical results using figures. *Korean Journal of Anesthesiology*. 2022;75(2):139-150. doi:10.4097/kja.21508
180. Wickham H. *ggplot2: Elegant graphics for data analysis*. 2016. <https://ggplot2.tidyverse.org>.
181. Sievert C. Interactive web-based data visualization with r, plotly, and shiny. 2020. <https://plotly-r.com>.
182. Wei T, Simko V. R package {corrplot}: Visualization of a correlation matrix. 2021. <https://github.com/taiyun/corrplot>.
183. Weissgerber TL, Winham SJ, Heinzen EP, et al. Reveal, Don't Conceal. *Circulation*. 2019;140(18):1506-1518. doi:10.1161/circulationaha.118.037777
184. Xiao N. *Ggsci: Scientific Journal and Sci-Fi Themed Color Palettes for {Ggplot2}*; 2023. <https://CRAN.R-project.org/package=ggsci>.
185. Urbanek S, Johnson K. *Tiff: Read and Write TIFF Images.*; 2022. <https://CRAN.R-project.org/package=tiff>.
186. Mair P, Wilcox R. Robust statistical methods in r using the WRS2 package. 2020;52. doi:10.3758/s13428-019-01246-w
187. Mair P, Wilcox R, Indrajeet P. *A Collection of Robust Statistical Methods.*; 2025. <https://CRAN.R-project.org/package=WRS2>.
188. Curran-Everett D. Explorations in statistics: hypothesis tests and P values. *Advances in Physiology Education*. 2009;33(2):81-86. doi:10.1152/advan.90218.2008

189. Goodman SN. Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy. *Annals of Internal Medicine*. 1999;130(12):995. doi:10.7326/0003-4819-130-12-199906150-00008
190. McCaskey K, Rainey C. Substantive importance and the veil of statistical significance. *Statistics, Politics and Policy*. 2015;6(1-2). doi:10.1515/spp-2015-0001
191. Vandenbroucke JP, Pearce N. From ideas to studies: how to get ideas and sharpen them into research questions. *Clinical Epidemiology*. 2018;Volume 10:253-264. doi:10.2147/clep.s142940
192. Lakens D, Scheel AM, Isager PM. Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*. 2018;1(2):259-269. doi:10.1177/2515245918770963
193. Sullivan GM, Feinn R. Using Effect Size—or Why the *P* Value Is Not Enough. *Journal of Graduate Medical Education*. 2012;4(3):279-282. doi:10.4300/jgme-d-12-00156.1
194. Heckman MG, Davis JM, Crowson CS. Post Hoc Power Calculations: An Inappropriate Method for Interpreting the Findings of a Research Study. *The Journal of Rheumatology*. 2022;49(8):867-870. doi:10.3899/jrheum.211115
195. Champely S. *Pwr: Basic Functions for Power Analysis*; 2020. <https://CRAN.R-project.org/package=pwr>.
196. Iddi S, Donohue MC. Power and sample size for longitudinal models in r-the longpower package and shiny app. 2022;14:264-281.
197. Lakens D, Caldwell A. Simulation-based power analysis for factorial analysis of variance designs. 2021;4:251524592095150. doi:10.1177/2515245920951503
198. Baranger DAA, Finsaas MC, Goldstein BL, Vize CE, Lynam DR, Olino TM. Tutorial: Power analyses for interaction effects in cross-sectional regressions. 2022. doi:10.31234/osf.io/5ptd7
199. Cumming G, Finch S. Inference by Eye: Confidence Intervals and How to Read Pictures of Data. *American Psychologist*. 2005;60(2):170-180. doi:10.1037/0003-066x.60.2.170
200. Goodman SN. Aligning statistical and scientific reasoning. *Science*. 2016;352(6290):1180-1181. doi:10.1126/science.aaf5406
201. Greenhalgh T. How to read a paper: Statistics for the non-statistician. II: Significant relations and their pitfalls. *BMJ*. 1997;315(7105):422-425. doi:10.1136/bmj.315.7105.422
202. Weintraub PG. The Importance of Publishing Negative Results. *Journal of Insect Science*. 2016;16(1):109. doi:10.1093/jisesa/iew092
203. Altman DG, Bland JM. Statistics notes: Absence of evidence is not evidence of absence. *BMJ*. 1995;311(7003):485-485. doi:10.1136/bmj.311.7003.485
204. Gelman A, Carlin J. Beyond Power Calculations. *Perspectives on Psychological Science*. 2014;9(6):641-651. doi:10.1177/1745691614551642
205. Lu J, Qiu Y, Deng A. A note on Type S/M errors in hypothesis testing. *British Journal of Mathematical and Statistical Psychology*. 2018;72(1):1-17. doi:10.1111/bmsp.12132
206. Kim HY. Statistical notes for clinical researchers: effect size. *Restorative Dentistry & Endodontics*. 2015;40(4):328. doi:10.5395/rde.2015.40.4.328
207. Ben-Shachar MS, Lüdecke D, Makowski D. *Effectsize: Estimation of effect size indices and standardized parameters*. 2020;5:2815. doi:10.21105/joss.02815

208. Bours MJL. Using mediators to understand effect modification and interaction. *Journal of Clinical Epidemiology*. September 2023. doi:10.1016/j.jclinepi.2023.09.005
209. Altman DG, Matthews JNS. Statistics Notes: Interaction 1: heterogeneity of effects. *BMJ*. 1996;313(7055):486-486. doi:10.1136/bmj.313.7055.486
210. Pinheiro J, Bates D, R Core Team. *Nlme: Linear and Nonlinear Mixed Effects Models.*; 2023. <https://CRAN.R-project.org/package=nlme>.
211. Sabanes Bove D, Dedic J, Kelkhoff D, et al. *Mmrm: Mixed Models for Repeated Measures.*; 2022. <https://CRAN.R-project.org/package=mmrm>.
212. Lenth RV. *Emmeans: Estimated Marginal Means, Aka Least-Squares Means.*; 2023. <https://CRAN.R-project.org/package=emmeans>.
213. Baron RM, Kenny DA. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*. 1986;51(6):1173-1182. doi:10.1037/0022-3514.51.6.1173
214. GREENLAND S, SCHLESSELMAN JJ, CRIQUI MH. THE FALLACY OF EMPLOYING STANDARDIZED REGRESSION COEFFICIENTS AND CORRELATIONS AS MEASURES OF EFFECT. *American Journal of Epidemiology*. 1986;123(2):203-208. doi:10.1093/oxfordjournals.aje.a114229
215. Greenland S, Maclure M, Schlesselman JJ, Poole C, Morgenstern H. Standardized Regression Coefficients. *Epidemiology*. 1991;2(5):387-392. doi:10.1097/00001648-199109000-00015
216. LATTER OH. THE EGG OF CUCULUS CANORUS: AN ENQUIRY INTO THE DIMENSIONS OF THE CUCKOO'S EGO AND THE RELATION OF THE VARIATIONS TO THE SIZE OF THE EGGS OF THE FOSTER-PARENT, WITH NOTES ON COLORATION, &c. *Biometrika*. 1902;1(2):164-176. doi:10.1093/biomet/1.2.164
217. Aylmer Fisher R. The arrangement of field experiments. *Ministry of Agriculture and Fisheries*. 1926. doi:10.23637/ROTHAMSTED.8V61Q
218. Wasserstein RL, Lazar NA. The ASA Statement on *p*-Values: Context, Process, and Purpose. *The American Statistician*. 2016;70(2):129-133. doi:10.1080/00031305.2016.1154108
219. Altman N, Krzywinski M. P values and the search for significance. *Nature Methods*. 2017;14(1):3-4. doi:10.1038/nmeth.4120
220. Heinze G, Dunkler D. Five myths about variable selection. *Transplant International*. 2016;30(1):6-10. doi:10.1111/tri.12895
221. Breznau N, Rinke EM, Wuttke A, et al. Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences*. 2022;(44):e2203150119. doi:10.1073/pnas.2203150119
222. Dwivedi AK, Shukla R. Evidence-based statistical analysis and methods in biomedical research (SAMBR) checklists according to design features. *CANCER REPORTS*. 2019;3(4). doi:10.1002/cnr2.1211
223. Dwivedi AK. How to Write Statistical Analysis Section in Medical Research. *Journal of Investigative Medicine*. 2022;70(8):1759-1770. doi:10.1136/jim-2022-002479
224. Kim N, Fischer AH, Dyring-Andersen B, Rosner B, Okoye GA. Research Techniques Made Simple: Choosing Appropriate Statistical Methods for Clinical Research. *Journal of Investigative Dermatology*. 2017;137(10):e173-e178. doi:10.1016/j.jid.2017.08.007

225. Marusteri M, Bacarea V. Comparing groups for statistical differences: How to choose the right statistical test? *Biochimia Medica*. 2010;15-32. doi:10.11613/bm.2010.004
226. Mishra P, Pandey C, Singh U, Keshri A, Sabaretnam M. Selection of appropriate statistical methods for data analysis. *Annals of Cardiac Anaesthesia*. 2019;22(3):297. doi:10.4103/aca.aca_248_18
227. Ray A, Najmi A, Sadasivam B. How to choose and interpret a statistical test? An update for budding researchers. *Journal of Family Medicine and Primary Care*. 2021;10(8):2763. doi:10.4103/jfmpc.jfmpc_433_21
228. Nayak B, Hazra A. How to choose the right statistical test? *Indian Journal of Ophthalmology*. 2011;59(2):85. doi:10.4103/0301-4738.77005
229. Shankar S, Singh R. Demystifying statistics: How to choose a statistical test? *Indian Journal of Rheumatology*. 2014;9(2):77-81. doi:10.1016/j.injr.2014.04.002
230. Diedenhofen B, Musch J. Cocor: A comprehensive solution for the statistical comparison of correlations. 2015;10:e0121945. doi:10.1371/journal.pone.0121945
231. McHugh ML. The chi-square test of independence. *Biochimia Medica*. 2013;143-149. doi:10.11613/bm.2013.018
232. Kim HY. Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test. *Restorative Dentistry & Endodontics*. 2017;42(2):152. doi:10.5395/rde.2017.42.2.152
233. Khamis H. Measures of Association: How to Choose? *Journal of Diagnostic Medical Sonography*. 2008;24(3):155-162. doi:10.1177/8756479308317006
234. Allison JS, Santana L, (Jaco) Visagie IJH. A primer on simple measures of association taught at undergraduate level. *Teaching Statistics*. 2022;44(3):96-103. doi:10.1111/test.12307
235. Dahlke JA, Wiernik BM. {Psychmeta}: An r package for psychometric meta-analysis. 2019;43. doi:10.1177/0146621618795933
236. Anscombe FJ. Graphs in Statistical Analysis. *The American Statistician*. 1973;27(1):17-21. doi:10.1080/00031305.1973.10478966
237. Northrop PJ. *Anscombiniser: Create Datasets with Identical Summary Statistics.*; 2022. <https://CRAN.R-project.org/package=anscombiniser>.
238. Makowski D, Wiernik BM, Patil I, Lüdecke D, Ben-Shachar MS. {{Correlation}}: Methods for Correlation Analysis.; 2022. <https://CRAN.R-project.org/package=correlation>.
239. Lüdecke D, Ben-Shachar MS, Patil I, et al. *Easystats: Framework for Easy Statistical Modeling, Visualization, and Reporting.*; 2022. <https://easystats.github.io/easystats/>.
240. Kim JH. Multicollinearity and misleading statistical results. *Korean Journal of Anesthesiology*. 2019;72(6):558-569. doi:10.4097/kja.19087
241. Schloerke B, Cook D, Larmarange J, et al. GGally: Extension to 'ggplot2'. 2024. doi:10.32614/CRAN.package.GGally
242. Arel-Bundock V. {Modelsummary}: Data and model summaries in {r}. 2022;103. doi:10.18637/jss.v103.i01
243. Hidalgo B, Goodman M. Multivariate or Multivariable Regression? *American Journal of Public Health*. 2013;103(1):39-40. doi:10.2105/ajph.2012.300897

244. Suits DB. Use of Dummy Variables in Regression Equations. *Journal of the American Statistical Association*. 1957;52(280):548-551. doi:10.1080/01621459.1957.10501412
245. Healy MJ. Statistics from the inside. 16. Multiple regression (2). *Archives of Disease in Childhood*. 1995;73(3):270-274. doi:10.1136/adc.73.3.270
246. Kaplan J. *fastDummies: Fast Creation of Dummy (Binary) Columns and Rows from Categorical Variables*.; 2023. <https://CRAN.R-project.org/package=fastDummies>.
247. Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *Journal of Clinical Epidemiology*. 1996;49(8):907-916. doi:10.1016/0895-4356(96)00025-x
248. Fox J, Weisberg S. An {r} companion to applied regression. 2019. <https://www.john-fox.ca/Companion/>.
249. DALES LG, URY HK. An Improper Use of Statistical Significance Testing in Studying Covariates. *International Journal of Epidemiology*. 1978;7(4):373-376. doi:10.1093/ije/7.4.373
250. Greenland S. Modeling and variable selection in epidemiologic analysis. *American Journal of Public Health*. 1989;79(3):340-349. doi:10.2105/ajph.79.3.340
251. Anderson D, Heiss A, Sumners J. *Equatiomatic: Transform Models into {LaTeX} Equations*.; 2024. <https://CRAN.R-project.org/package=equatiomatic>.
252. Lüdecke D, Ben-Shachar MS, Patil I, Waggoner P, Makowski D. {Performance}: An {r} package for assessment, comparison and testing of statistical models. 2021;6:3139. doi:10.21105/joss.03139
253. Spedicato GA. Discrete time markov chains with r. 2017;9. <https://journal.r-project.org/archive/2017/RJ-2017-036/index.html>.
254. Henderson T. correctR: Corrected test statistics for comparing machine learning models on correlated samples. 2025. <https://CRAN.R-project.org/package=correctR>.
255. Griffith DM, Veech JA, Marsh CJ. {Cooccur}: Probabilistic species co-occurrence analysis in {r}. 2016;69. doi:10.18637/jss.v069.c02
256. Lüdecke D. Ggeffects: Tidy data frames of marginal effects from regression models. 2018;3:772. doi:10.21105/joss.00772
257. Textor J, Zander B van der, Gilthorpe MS, Liskiewicz M, Ellison GT. Robust causal inference using directed acyclic graphs: The r package {dagitty}. 2016;45. doi:10.1093/ije/dyw341
258. Barrett M. *Ggdag: Analyze and Create Elegant Directed Acyclic Graphs*.; 2024. <https://CRAN.R-project.org/package=ggdag>.
259. Bland JM, Altman DG. Statistics notes: Matching. *BMJ*. 1994;309(6962):1128-1128. doi:10.1136/bmj.309.6962.1128
260. Grant MJ, Booth A. A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal*. 2009;26(2):91-108. doi:10.1111/j.1471-1842.2009.00848.x
261. Sut N. Study designs in medicine. *Balkan Medical Journal*. 2015;31(4):273-277. doi:10.5152/balkanmedj.2014.1408
262. Souza AC de, Alexandre NMC, Guirardello E de B, Souza AC de, Alexandre NMC, Guirardello E de B. Propriedades psicométricas na avaliação de instrumentos: avaliação da confiabilidade e da validade. *Epidemiologia e Serviços de Saúde*. 2017;26(3):649-659. doi:10.5123/s1679-49742017000300022

263. Reeves BC, Wells GA, Waddington H. Quasi-experimental study designs series—paper 5: a checklist for classifying studies evaluating the effects on health interventions—a taxonomy without labels. *Journal of Clinical Epidemiology*. 2017;89:30-42. doi:10.1016/j.jclinepi.2017.02.016
264. Echevarría-Guanilo ME, Gonçalves N, Romanoski PJ. PSYCHOMETRIC PROPERTIES OF MEASUREMENT INSTRUMENTS: CONCEPTUAL BASIS AND EVALUATION METHODS – PART II. *Texto & Contexto – Enfermagem*. 2019;28. doi:10.1590/1980-265x-tce-2017-0311
265. Chassé M, Fergusson DA. Diagnostic Accuracy Studies. *Seminars in Nuclear Medicine*. 2019;49(2):87-93. doi:10.1053/j.semnuclmed.2018.11.005
266. Chidambaram AG, Josephson M. Clinical research study designs: The essentials. *PEDIATRIC INVESTIGATION*. 2019;3(4):245-252. doi:10.1002/ped4.12166
267. Erdemir A, Mulugeta L, Ku JP, et al. Credible practice of modeling and simulation in healthcare: ten rules from a multidisciplinary perspective. *Journal of Translational Medicine*. 2020;18(1). doi:10.1186/s12967-020-02540-4
268. Yang B, Olsen M, Vali Y, et al. Study designs for comparative diagnostic test accuracy: A methodological review and classification scheme. *Journal of Clinical Epidemiology*. 2021;138:128-138. doi:10.1016/j.jclinepi.2021.04.013
269. Chipman H, Bingham D. Let's practice what we preach: Planning and interpreting simulation studies with design and analysis of experiments. *Canadian Journal of Statistics*. 2022;50(4):1228-1249. doi:10.1002/cjs.11719
270. Donthu N, Kumar S, Mukherjee D, Pandey N, Lim WM. How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*. 2021;133:285-296. doi:10.1016/j.jbusres.2021.04.070
271. Lim WM, Kumar S. Guidelines for interpreting the results of bibliometric analysis: A sensemaking approach. *Global Business and Organizational Excellence*. August 2023. doi:10.1002/joe.22229
272. Rodríguez del Águila M, González-Ramírez A. Sample size calculation. *Allergologia et Immunopathologia*. 2014;42(5):485-492. doi:10.1016/j.aller.2013.03.008
273. Bacchetti P. Ethics and Sample Size. *American Journal of Epidemiology*. 2005;161(2):105-110. doi:10.1093/aje/kwi014
274. Ying X, Robinson KA, Ehrhardt S. Re-evaluating the role of pilot trials in informing effect and sample size estimates for full-scale trials: a meta-epidemiological study. *BMJ Evidence-Based Medicine*. 2023;28(6):383-391. doi:10.1136/bmjebm-2023-112358
275. Andrade C. Sample Size and its Importance in Research. *Indian Journal of Psychological Medicine*. 2020;42(1):102-103. doi:10.4103/ijpsym.ijpsym_504_19
276. Sasaki K, Yamada Y. SPARKing: Sample-size planning after the results are known. *Frontiers in Human Neuroscience*. 2023;17. doi:10.3389/fnhum.2023.912338
277. Elm E von, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies. *Annals of Internal Medicine*. 2007;147(8):573. doi:10.7326/0003-4819-147-8-200710160-00010
278. Rosseel Y. {Lavaan}: An {r} package for structural equation modeling. 2012;48. doi:10.18637/jss.v048.i02
279. Contributors semTools. *semTools: Useful Tools for Structural Equation Modeling*; 2016. <https://CRAN.R-project.org/package=semTools>.

280. William Revelle. *Psych: Procedures for Psychological, Psychometric, and Personality Research.*; 2023. <https://CRAN.R-project.org/package=psych>.
281. Findley MG, Kikuta K, Denly M. External Validity. *Annual Review of Political Science*. 2021;24(1):365-393. doi:10.1146/annurev-polisci-041719-102556
282. Scott WA. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*. 1955;19(3):321. doi:10.1086/266577
283. Cohen J. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*. 1960;20(1):37-46. doi:10.1177/001316446002000104
284. I. Mathematical contributions to the theory of evolution. —VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London Series A, Containing Papers of a Mathematical or Physical Character*. 1901;195(262-273):1-47. doi:10.1098/rsta.1900.0022
285. Banerjee M, Capozzoli M, McSweeney L, Sinha D. Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics*. 1999;27(1):3-23. doi:10.2307/3315487
286. Lehnert B. *BlandAltmanLeh: Plots (Slightly Extended) Bland-Altman Plots.*; 2015. <https://CRAN.R-project.org/package=BlandAltmanLeh>.
287. Gagnier JJ, Lai J, Mokkink LB, Terwee CB. COSMIN reporting guideline for studies on measurement properties of patient-reported outcome measures. *Quality of Life Research*. 2021;30(8):2197-2218. doi:10.1007/s11136-021-02822-4
288. Streiner DL, Kottner J. Recommendations for reporting the results of studies of instrument and scale development and testing. *Journal of Advanced Nursing*. 2014;70(9):1970-1979. doi:10.1111/jan.12402
289. Kottner J, Audigé L, Brorson S, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *Journal of Clinical Epidemiology*. 2011;64(1):96-106. doi:10.1016/j.jclinepi.2010.03.002
290. Steckelberg A, Balgenorth A, Berger J, Mühlhauser I. Explaining computation of predictive values: 2×2 table versus frequency tree. A randomized controlled trial [ISRCTN74278823]. *BMC Medical Education*. 2004;4(1). doi:10.1186/1472-6920-4-13
291. Greenhalgh T. How to read a paper: Papers that report diagnostic or screening tests. *BMJ*. 1997;315(7107):540-543. doi:10.1136/bmj.315.7107.540
292. Neth H, Gaisbauer F, Gradwohl N, Gaissmaier W. *RiskyR: Rendering Risk Literacy More Transparent.*; 2022. <https://CRAN.R-project.org/package=riskyR>.
293. Kuhn, Max. Building predictive models in r using the caret package. *Journal of Statistical Software*. 2008;28(5):1-26. doi:10.18637/jss.v028.i05
294. Phillips B, Stewart LA, Sutton AJ. 'Cross hairs' plots for diagnostic meta-analysis. *Research Synthesis Methods*. 2010;1(3-4):308-315. doi:10.1002/jrsm.26
295. Sousa-Pinto PD with contributions from B. *Mada: Meta-Analysis of Diagnostic Accuracy.*; 2022. <https://CRAN.R-project.org/package=mada>.
296. Hond AAH de, Steyerberg EW, Calster B van. Interpreting area under the receiver operating characteristic curve. *The Lancet Digital Health*. 2022;4(12):e853-e855. doi:10.1016/s2589-7500(22)00188-1
297. Robin X, Turck N, Hainard A, et al. pROC: An open-source package for r and s+ to analyze and compare ROC curves. 2011;12:77.

298. Ferreira ADS, Meziat-Filho N, Ferreira APA. Double threshold receiver operating characteristic plot for three-modal continuous predictors. *Computational Statistics*. 2021;36(3):2231-2245. doi:10.1007/s00180-021-01080-9
299. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ*. October 2015:h5527. doi:10.1136/bmj.h5527
300. Reeves BC, Gaus W. Guidelines for Reporting Non-Randomised Studies. *Complementary Medicine Research*. 2004;11(1):46-52. doi:10.1159/000080576
301. Bland JM, Altman DG. Comparisons within randomised groups can be very misleading. *BMJ*. 2011;342(may06 2):d561-d561. doi:10.1136/bmj.d561
302. Bruce CL, Juszczak E, Ogollah R, Partlett C, Montgomery A. A systematic review of randomisation method use in RCTs and association of trial design characteristics with method selection. *BMC Medical Research Methodology*. 2022;22(1). doi:10.1186/s12874-022-01786-4
303. Vickers AJ, Altman DG. Statistics Notes: Analysing controlled trials with baseline and follow up measurements. *BMJ*. 2001;323(7321):1123-1124. doi:10.1136/bmj.323.7321.1123
304. O Connell NS, Dai L, Jiang Y, et al. Methods for analysis of pre-post data in clinical research: A comparison of five common methods. *Journal of Biometrics & Biostatistics*. 2017;08(01). doi:10.4172/2155-6180.1000334
305. Laird N. Further Comparative Analyses of Pretest-Posttest Research Designs. *The American Statistician*. 1983;37(4a):329-330. doi:10.1080/00031305.1983.10483133
306. Cnaan A, Laird NM, Slasor P. Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in Medicine*. 1997;16(20):2349-2380. doi:10.1002/(sici)1097-0258(19971030)16:20<2349::aid-sim667>3.0.co;2-e
307. Mallinckrodt CH, Lane PW, Schnell D, Peng Y, Mancuso JP. Recommendations for the Primary Analysis of Continuous Endpoints in Longitudinal Clinical Trials. *Drug Information Journal*. 2008;42(4):303-319. doi:10.1177/009286150804200402
308. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *The Lancet*. 2000;355(9209):1064-1069. doi:10.1016/s0140-6736(00)02039-0
309. Stang A, Baethge C. Imbalance *p* values for baseline covariates in randomized controlled trials: a last resort for the use of *p* values? A pro and contra debate. *Clinical Epidemiology*. 2018;Volume 10:531-535. doi:10.2147/cep.s161508
310. Bolzern JE, Mitchell A, Torgerson DJ. Baseline testing in cluster randomised controlled trials: should this be done? *BMC Medical Research Methodology*. 2019;19(1). doi:10.1186/s12874-019-0750-8
311. Roberts C, Torgerson DJ. Understanding controlled trials: Baseline imbalance in randomised controlled trials. *BMJ*. 1999;319(7203):185-185. doi:10.1136/bmj.319.7203.185
312. Gruijters SLK. Baseline comparisons and covariate fishing: Bad statistical habits we should have broken yesterday. July 2020. <http://dx.doi.org/10.31234/osf.io/qftwg>.
313. Vickers AJ. The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study. *BMC Medical Research Methodology*. 2001;1(1). doi:10.1186/1471-2288-1-6

314. Brookes ST, Whitely E, Egger M, Smith GD, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; *Journal of Clinical Epidemiology*. 2004;57(3):229-236. doi:10.1016/j.jclinepi.2003.08.009
315. Matthews JNS, Altman DG. Statistics Notes: Interaction 2: compare effect sizes not P values. *BMJ*. 1996;313(7060):808-808. doi:10.1136/bmj.313.7060.808
316. Altman DG. Statistics notes: Interaction revisited: The difference between two estimates. *BMJ*. 2003;326(7382):219-219. doi:10.1136/bmj.326.7382.219
317. Hauck WW, Anderson S, Marcus SM. Should We Adjust for Covariates in Nonlinear Regression Analyses of Randomized Trials? *Controlled Clinical Trials*. 1998;19(3):249-256. doi:10.1016/s0197-2456(97)00147-5
318. Kahan BC, Jairath V, Doré CJ, Morris TP. The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. *Trials*. 2014;15(1). doi:10.1186/1745-6215-15-139
319. Cao Y, Allore H, Vander Wyk B, Gutman R. Review and evaluation of imputation methods for multivariate longitudinal data with mixed-type incomplete variables. *Statistics in Medicine*. October 2022. doi:10.1002/sim.9592
320. Schulz KF. CONSORT 2010 Statement: Updated Guidelines for Reporting Parallel Group Randomized Trials. *Annals of Internal Medicine*. 2010;152(11):726. doi:10.7326/0003-4819-152-11-201006010-00232
321. Dayim A. *Consort: Create Consort Diagram*; 2023. <https://CRAN.R-project.org/package=consort>.
322. Lajeunesse MJ. Facilitating systematic reviews, data extraction, and meta-analysis with the metagear package for r. 2016;7:323-330.
323. Moher D, Shamseer L, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*. 2015;4(1). doi:10.1186/2046-4053-4-1
324. Haddaway NR, Page MJ, Pritchard CC, McGuinness LA. PRISMA2020: An r package and shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. 2022;18:e1230. doi:10.1002/cl2.1230
325. Borenstein M. In a meta-analysis, the I-squared statistic does not tell us how much the effect size varies. *Journal of Clinical Epidemiology*. October 2022. doi:10.1016/j.jclinepi.2022.10.003
326. Rücker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I² in assessing heterogeneity may mislead. *BMC Medical Research Methodology*. 2008;8(1). doi:10.1186/1471-2288-8-79
327. Grooth HJ de, Parienti JJ. Heterogeneity between studies can be explained more reliably with individual patient data. *Intensive Care Medicine*. July 2023. doi:10.1007/s00134-023-07163-z
328. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *PLOS Medicine*. 2021;18(3):e1003583. doi:10.1371/journal.pmed.1003583
329. Goldfeld K, Wujciak-Jens J. Simstudy: Illuminating research methods through data generation. 2020;5:2763. doi:10.21105/joss.02763
330. DeBruine L. *Faux: Simulation for Factorial Designs*; 2023. doi:10.5281/zenodo.2669586
331. Cheng A, Kessler D, Mackinnon R, et al. Reporting Guidelines for Health Care Simulation Research. *Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare*. 2016;11(4):238-248. doi:10.1097/sih.00000000000000150

332. Makowski D, Lüdecke D, Patil I, Thériault R, Ben-Shachar MS, Wiernik BM. *Automated Results Reporting as a Practical Tool to Improve Reproducibility and Methodological Best Practices Adoption.*; 2023. <https://easystats.github.io/report/>.
333. Wallisch C, Bach P, Hafermann L, et al. Review of guidance papers on regression modeling in statistical series of medical journals. Mathes T, ed. *PLOS ONE*. 2022;17(1):e0262918. doi:10.1371/journal.pone.0262918
334. Lynggaard H, Bell J, Lösch C, et al. Principles and recommendations for incorporating estimands into clinical study protocol templates. *Trials*. 2022;23(1). doi:10.1186/s13063-022-06515-2
335. Althouse AD, Below JE, Claggett BL, et al. Recommendations for Statistical Reporting in Cardiovascular Medicine: A Special Report From the American Heart Association. *Circulation*. 2021;144(4). doi:10.1161/circulationaha.121.055393
336. Lee KJ, Tilling KM, Cornish RP, et al. Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework. *Journal of Clinical Epidemiology*. 2021;134:79-88. doi:10.1016/j.jclinepi.2021.01.008
337. Vickers AJ, Assel MJ, Sjoberg DD, et al. Guidelines for Reporting of Figures and Tables for Clinical Research in Urology. *Urology*. 2020;142:1-13. doi:10.1016/j.urology.2020.05.002
338. Assel M, Sjoberg D, Elders A, et al. Guidelines for Reporting of Statistics for Clinical Research in Urology. *Journal of Urology*. 2019;201(3):595-604. doi:10.1097/ju.0000000000000001
339. Gamble C, Krishan A, Stocken D, et al. Guidelines for the Content of Statistical Analysis Plans in Clinical Trials. *JAMA*. 2017;318(23):2337. doi:10.1001/jama.2017.18556
340. Lang TA, Altman DG. Basic statistical reporting for articles published in Biomedical Journals: The “Statistical Analyses and Methods in the Published Literature” or the SAMPL Guidelines. *International Journal of Nursing Studies*. 2015;52(1):5-9. doi:10.1016/j.ijnurstu.2014.09.006
341. Weissgerber TL, Milic NM, Winham SJ, Garovic VD. Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm. *PLOS Biology*. 2015;13(4):e1002128. doi:10.1371/journal.pbio.1002128
342. Sauerbrei W, Abrahamowicz M, Altman DG, Cessie S, Carpenter J. STREngthening Analytical Thinking for Observational Studies: the STRATOS initiative. *Statistics in Medicine*. 2014;33(30):5413-5432. doi:10.1002/sim.6265
343. Groves T. Research methods and reporting. *BMJ*. 2008;337(oct22 1):a2201-a2201. doi:10.1136/bmj.a2201
344. Stratton IM, Neil A. How to ensure your paper is rejected by the statistical reviewer. *Diabetic Medicine*. 2005;22(4):371-373. doi:10.1111/j.1464-5491.2004.01443.x
345. Mansournia MA, Collins GS, Nielsen RO, et al. A Checklist for statistical Assessment of Medical Papers (the CHAMP statement): explanation and elaboration. *British Journal of Sports Medicine*. 2021;55(18):1009-1017. doi:10.1136/bjsports-2020-103652
346. Gil-Sierra MD, Fénix-Caballero S, Abdel kader-Martin L, et al. Checklist for clinical applicability of subgroup analysis. *Journal of Clinical Pharmacy and Therapeutics*. 2019;45(3):530-538. doi:10.1111/jcpt.13102
347. Lee H, Cashin AG, Lamb SE, et al. A Guideline for Reporting Mediation Analyses of Randomized Trials and Observational Studies. *JAMA*. 2021;326(11):1045. doi:10.1001/jama.2021.14075

Ciência com R

Você está pronto para desbloquear o poder da análise estatística de dados e elevar sua pesquisa a novos patamares? Não procure mais. Em “Ciência com R”, o Dr. Arthur de Sá Ferreira, um pesquisador experiente, oferece um guia indispensável que capacitará pesquisadores, analistas de dados e estudantes a tomarem decisões informadas e baseadas em evidências em seus empreendimentos científicos.

ORIENTAÇÃO ESPECIALIZADA: Beneficie-se da ampla experiência do Dr. Arthur de Sá Ferreira enquanto ele responde às perguntas mais fundamentais: *O que é isso? Por que usá-lo? Quando usar? Quando não usar? Como fazer?* Cada capítulo se aprofunda em questões específicas, oferecendo explicações claras e concisas e exemplos práticos.

FORMATO DE PERGUNTAS E RESPOSTAS: Mantenha uma conversa direta e objetiva com o autor. Descubra respostas para as perguntas comumente feitas por estudantes, pesquisadores e profissionais em todas as fases de sua jornada acadêmica e científica.

APRENDIZADO PROGRESSIVO: Navegue por uma progressão de conceitos e aplicações. Capítulos são estruturados didaticamente para maior clareza educacional, com referências cruzadas para garantir uma compreensão coesa dos tópicos inter-relacionados, reduzindo a fragmentação do conteúdo.

INSIGHTS ATUALIZADOS: Fique à frente da curva com as referências e insights mais recentes. Dr. [Seu nome] lança luz sobre preconceitos, paradoxos, mitos e práticas ilícitas na área, oferecendo uma clareza inestimável até mesmo para os pesquisadores mais experientes.

Quer você seja um estudante de pós-graduação em busca de métodos para analisar seus projetos de pesquisa, um pesquisador que precisa de informações e referências para o desenvolvimento de projetos ou um analista de dados experiente que deseja se manter atualizado, este livro é seu melhor companheiro. Além disso, pessoas de diversas áreas encontrarão neste livro uma porta de entrada para compreender a importância de fazer e responder perguntas no mundo da ciência.

Tome decisões informadas, evite armadilhas e destaque-se em sua pesquisa científica com “Ciência com R”. Os insights profundos do Dr. Arthur de Sá Ferreira permitirão que você transforme seus dados em descobertas significativas, colocando você no caminho da excelência em pesquisa.