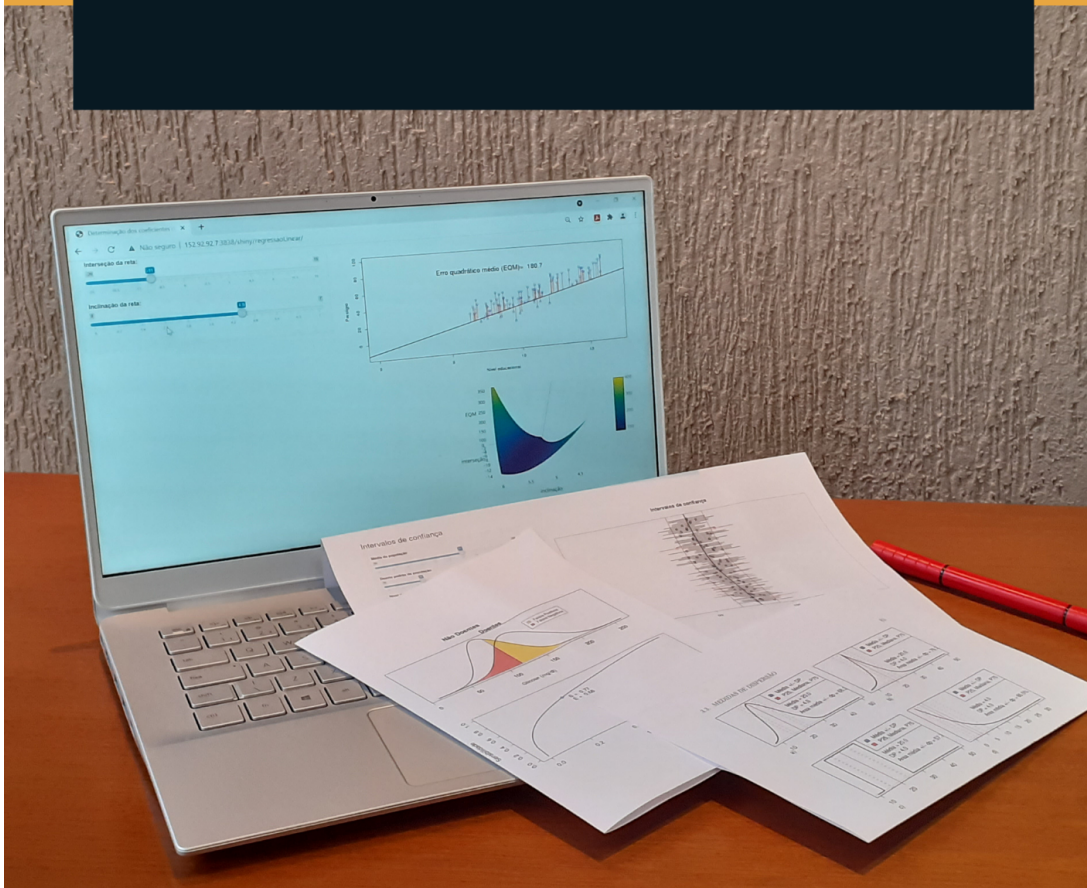


# BIOESTATÍSTICA BÁSICA

SERGIO MIRANDA FREIRE



*e-book*

# Bioestatística Básica

Sergio Miranda Freire

# Bioestatística Básica

Última atualização

“2022-09-28”

Endereço de acesso

<http://www.lampada.uerj.br/bioestatisticabasica>

Capa

Rosimary T. Almeida

**Dados Internacionais de Catalogação na Publicação (CIP)**  
**(Câmara Brasileira do Livro, SP, Brasil)**

Freire, Sergio Miranda  
Bioestatística básica [livro eletrônico] /  
Sergio Miranda Freire. -- Rio de Janeiro :  
Ed. do Autor, 2021.  
PDF  
  
Bibliografia.  
ISBN 978-65-00-35696-0  
  
1. Estatística - Métodos 2. Estatística médica  
I. Título.

21-93574

CDD-570.15195

**Índices para catálogo sistemático:**

1. Bioestatística para ciências da saúde 570.15195

Cibele Maria Dias - Bibliotecária - CRB-8/9427



Exceto onde indicado, esta obra está licenciada com uma Licença Creative Commons Atribuição-Não Comercial 4.0 Internacional.

# Prefácio

Os profissionais de saúde devem conhecer os princípios básicos de Estatística para planejar a realização de estudos, interpretar estatísticas vitais, dados epidemiológicos e resultados de estudos publicados na literatura científica, interagir com estatísticos etc. Diversos textos voltados para o público da área de saúde que abordam os conceitos básicos de Estatística estão disponíveis, alguns com excelente conteúdo e uma abordagem didática. Então cabe a pergunta: por que mais um texto sobre Estatística?

Após alguns anos ensinando Estatística para alunos da pós-graduação em Medicina, percebi a necessidade de tornar as aulas mais eficientes por meio da combinação de teoria e prática, fazendo uso de um pacote estatístico para aplicar os conceitos teóricos em dados reais. Além disso, diversos conceitos básicos em Bioestatística, como intervalo de confiança, teorema do limite central, curva ROC etc., podem ser ilustrados por meio de aplicações onde o aluno pode visualizar o conceito interagindo com o computador.

O presente texto vem sendo desenvolvido nos últimos três anos, por ocasião da reforma curricular do curso de Medicina da Universidade do Estado do Rio de Janeiro. Este texto combina a apresentação dos conceitos básicos de Estatística com o uso do ambiente para análise de dados R e aplicações que permitem ao leitor interagir com as mesmas, alterando parâmetros e verificando a resposta.

A opção pela adoção do R se justifica pelo fato de sua disponibilização como código aberto (*open source*), por sua ampla utilização em nível mundial, pelo constante aperfeiçoamento de seus pacotes e constante surgimento de novos pacotes. Apesar de sua utilização por meio de linhas de comando amedrontar alunos e profissionais da área de saúde com pouca familiaridade com programação de computadores, este texto utiliza principalmente um pacote do R que oferece uma interface gráfica para as funcionalidades e as análises estatísticas mais utilizadas. O texto ilustra o passo a passo de como utilizar o *R Commander*, e eventualmente o *RStudio*, para realizar as operações no R, seguida de uma explicação do comando gerado a partir da interface gráfica. Em algumas situações, nas quais o *R Commander* não dispõe de recursos, o texto mostra como escrever um comando que realizar a função desejada.

Além do R, ao longo do texto são inseridas 24 aplicações, desenvolvidas por meio do pacote *shiny* do R, que ilustram diversos conceitos básicos de Estatística. Espera-se que, com essas aplicações, o aluno possa apreender de maneira mais efetiva alguns dos conceitos abordados.

Para quem deseja aprofundar os conhecimentos do R, o texto disponível neste [endereço](#) mostra como utilizar o *RStudio* e *R Commander* para manipular um arquivo de dados, criar



gráficos, gerenciar uma sessão e obter um conhecimento básico sobre funções e estruturas de controle do R.

Para o aluno que entra em contato pela primeira vez com um conteúdo de Estatística, recomenda-se que os capítulos sejam lidos na ordem apresentada. As seções marcadas em negrito podem ser omitidas numa primeira leitura do texto, sem perda de continuidade.

O capítulo 1 apresenta uma introdução sobre como as variáveis são organizadas em arquivos para a realização de análises estatísticas e as escalas de medidas de variáveis.

Em seguida, uma sequência de três capítulos são relativos à obtenção de estatísticas descritivas e visualização de dados. O capítulo 2 mostra como obter e interpretar tabelas de frequências uni e multivariada no R. O capítulo 3 apresenta as medidas de tendência central e dispersão para variáveis numéricas mais utilizadas na literatura médica. O capítulo 4 apresenta diversos gráficos utilizados para visualizar a distribuição dos dados, tanto para variáveis categóricas quanto para variáveis numéricas.

O capítulo 5 resume os principais desenhos de estudos utilizados em epidemiologia clínica.

O capítulo 6 faz uma introdução à inferência estatística, apresentando os conceitos de teste de hipótese e intervalo de confiança, concluindo por apresentar a importante distinção entre relevância clínica e significância estatística.

O capítulo 7 introduz a noção de probabilidades e alguns conceitos fundamentais, como probabilidade condicional e o teorema de Bayes.

O capítulo 8 apresenta diversas medidas de associação utilizadas em epidemiologia clínica para verificar a associação entre duas variáveis categóricas.

Os capítulos 9, 10 e 11 introduzem, respectivamente, o conceito de variável aleatória, algumas distribuições de probabilidades para variáveis numéricas discretas e o conceito de função densidade de probabilidade, com ênfase na distribuição normal ou gaussiana.

O capítulo 12 apresenta as métricas utilizadas para avaliar a acurácia de testes diagnósticos, tanto para testes cujos resultados são categorias de uma variável, quanto para aqueles baseados em uma variável numérica contínua.

Os capítulos 13, 14 e 15 retomam o tema de inferência estatística, sendo que o capítulo 13 introduz o conceito de estimadores e algumas de suas propriedades e apresenta um teorema importante na análise estatística que é o teorema do limite central.

O capítulo 14 aprofunda o conceito de intervalo de confiança, desta vez mostrando o seu cálculo para a média e a variância de uma distribuição normal. No capítulo 15, são desenvolvidos os conceitos de teste de hipótese, valor de  $p$  e poder estatístico.

Os capítulos seguintes introduzem algumas análises estatísticas frequentemente utilizadas em saúde. O capítulo 16 apresenta o teste  $t$  de Student para dois grupos independentes ou dependentes e os testes não paramétricos alternativos quando as suposições para a realização do teste  $t$  não são satisfeitas. O capítulo 17 continua o tema iniciado no capítulo 8 (medidas de associação), desta vez mostrando os cálculos dos intervalos de confiança e o teste qui ao quadrado, tanto para amostras não pareadas quanto para amostras pareadas.

O capítulo 18 introduz a análise de variância, que trata da comparação de médias de uma variável numérica em mais de duas populações.

O capítulo 19 apresenta o modelo de regressão linear simples que trata do relacionamento linear entre duas variáveis numéricas.

Finalmente o capítulo 20 faz uma breve introdução à análise de sobrevida, com ênfase no método de Kaplan-Meier para estimar as curvas de sobrevida em uma ou mais populações de pacientes.

Sergio Miranda Freire, Rio de Janeiro  
Dezembro de 2021

# Agradecimentos

À Universidade do Estado do Rio de Janeiro (UERJ), que me propiciou as condições para me dedicar ao ensino e pesquisa ao longo de minha carreira.

A Mário João Jr, analista de sistemas do Departamento de Tecnologia da Informação e Educação em Saúde, da Faculdade de Ciências Médicas da UERJ, que zela para que este *e-book* esteja disponível via *web*.

À professora Rosimary T. Almeida e aos alunos do curso de Estatística Aplicada à Física Médica da Universidade Federal do Rio de Janeiro, pelos comentários e sugestões.

Aos alunos dos cursos de Bioestatística Básica da Pós-Graduação em Ciências Médicas, e do curso de Métodos Estatísticos Aplicados à Medicina Laboratorial, do Mestrado Profissional em Saúde, Medicina Laboratorial e Tecnologia Forense, ambos da Universidade do Estado do Rio de Janeiro, pelos comentários e revisões do texto.

A todos aqueles que, com atos e palavras, trabalham para que a humanidade seja regida pela seguinte máxima: a cada um de acordo as suas necessidades, de cada um de acordo com as suas possibilidades.

# Sumário

<b>Bioestatística Básica</b>	<b>ii</b>
<b>Prefácio</b>	<b>iii</b>
<b>Agradecimentos</b>	<b>vi</b>
<b>1 Introdução</b>	<b>1</b>
1.1 O que é Estatística?	1
1.2 Por que estudar Estatística?	1
1.3 Dado, informação e conhecimento	2
1.4 Variáveis em estudos clínicos	5
1.5 Arquivos de dados	9
1.6 Escalas de medidas	14
1.6.1 Escala nominal	14
1.6.2 Escala ordinal	15
1.6.3 Escala intervalar	16
1.6.4 Escala de razão	16
1.7 Transformação de variáveis	17
1.8 Escalas e índices	18
1.9 Identificação de variáveis em estudos clínicos	21
1.10 Exercício	23
<b>2 Tabelas de frequências</b>	<b>24</b>
2.1 Introdução	24
2.2 Carregando conjuntos de dados de pacotes do R	26
2.3 Tabelas de frequências no conjunto de dados <i>stroke</i>	32
2.3.1 Uma única variável categórica	32
2.3.2 Tabelas de frequências para duas variáveis categóricas	34
2.3.3 Tabelas de frequência para mais de duas variáveis categóricas	39
2.3.4 Entrando diretamente com as frequências das células	46
2.4 Exercício	49
<b>3 Medidas de tendência central e dispersão</b>	<b>50</b>
3.1 Introdução	50
3.2 Medidas de tendência central	53

3.2.1	Média . . . . .	53
3.2.2	Mediana . . . . .	55
3.2.3	Moda . . . . .	56
3.2.4	Discussão sobre medidas de tendência central . . . . .	58
3.3	Medidas de dispersão . . . . .	58
3.3.1	Amplitude . . . . .	58
3.3.2	Distância interquartil . . . . .	60
3.3.3	Percentis . . . . .	63
3.3.4	Desvio padrão e variância . . . . .	65
3.3.5	Discussão sobre as medidas de dispersão . . . . .	66
3.4	Apresentação das estatísticas descritivas em publicações . . . . .	68
3.4.1	Exemplos de formas inadequadas de apresentação da média e desvio padrão . . . . .	69
3.4.2	Exemplos de formas adequadas de apresentação da média, mediana, desvio padrão e primeiro e terceiro quartis . . . . .	70
3.5	<b>Escore z ou Escore padrão</b> . . . . .	70
3.6	Obtendo estatísticas descritivas no R . . . . .	72
3.6.1	Carregando conjuntos de dados de pacotes do R . . . . .	72
3.6.2	Obtendo resumos numéricos pelo R Commander . . . . .	75
3.6.3	R Markdown . . . . .	78
3.6.4	Salvando scripts e arquivos do R Markdown . . . . .	81
3.7	<b>Executando scripts no R Commander</b> . . . . .	82
3.8	Exercícios . . . . .	83
4	<b>Visualização de dados</b> . . . . .	86
4.1	Convertendo uma variável numérica para fator . . . . .	91
4.2	Diagrama de barras . . . . .	94
4.3	<b>Usando a linha de comando</b> . . . . .	101
4.3.1	Especificação dos rótulos dos eixos x e y e do título . . . . .	101
4.3.2	Alteração dos tamanhos dos eixos X e Y . . . . .	101
4.3.3	Alteração do título da legenda do diagrama . . . . .	102
4.3.4	Alteração do espaçamento entre as barras . . . . .	103
4.3.5	Tamanhos dos rótulos dos eixos X e Y, dos números no eixo Y e das categorias das barras . . . . .	104
4.3.6	Alteração das categorias da variável do eixo X . . . . .	105
4.3.7	Alteração das cores . . . . .	106
4.3.8	Gráfico de barras horizontais . . . . .	111
4.4	<b>Diagrama de setores, torta ou pizza</b> . . . . .	111
4.5	Diagrama de caixa ( <i>boxplot</i> ou <i>box and whisker plot</i> ) . . . . .	114
4.6	Histograma . . . . .	118
4.6.1	Histograma de frequência x frequência relativa x densidade de frequência relativa . . . . .	121
4.6.2	Histograma por grupos . . . . .	124
4.7	Diagrama de pontos e <i>strip chart</i> . . . . .	125
4.8	Diagrama de dispersão ou espalhamento . . . . .	128

4.8.1	Alterando a espessura e cor da linha de regressão e o tipo dos pontos . . . . .	130
4.9	Salvando gráficos em um arquivo . . . . .	132
4.10	Recursos gráficos de outros plugins . . . . .	135
4.11	Exercícios . . . . .	138
<b>5</b>	<b>Amostragem e delineamentos de pesquisas</b>	<b>142</b>
5.1	Introdução . . . . .	142
5.2	População e amostra . . . . .	143
5.3	Amostragem . . . . .	144
5.3.1	Amostragem probabilística . . . . .	144
5.3.2	Amostragem não probabilística . . . . .	149
5.4	Delineamentos de estudos clínico-epidemiológicos . . . . .	151
5.5	Ensaio controlado randomizado . . . . .	151
5.6	Ensaio controlado não randomizado . . . . .	155
5.7	Série de casos . . . . .	155
5.8	Estudo de coortes . . . . .	156
5.9	Estudo de caso-controle . . . . .	158
5.10	Estudo transversal . . . . .	159
5.11	Revisão sistemática e metanálise . . . . .	160
5.12	Gradação da evidência científica . . . . .	161
5.13	Exercícios . . . . .	163
<b>6</b>	<b>Introdução à Inferência Estatística</b>	<b>165</b>
6.1	Introdução . . . . .	165
6.2	Apresentação de resultados de estudos . . . . .	165
6.3	Teste de hipótese usando randomização . . . . .	168
6.3.1	Contexto do problema . . . . .	168
6.3.2	Hipótese nula e nível de significância . . . . .	172
6.4	Valor de p . . . . .	174
6.5	Intervalo de confiança (IC) . . . . .	175
6.6	Exemplo de teste sem rejeição da hipótese nula . . . . .	178
6.7	Uso inadequado de testes de hipótese . . . . .	182
6.8	Uso de modelos para o cálculo do intervalo de confiança . . . . .	183
6.9	Interpretação do intervalo de confiança . . . . .	185
6.10	Significância estatística e relevância clínica . . . . .	190
6.11	Exercício . . . . .	192
<b>7</b>	<b>Probabilidade</b>	<b>193</b>
7.1	Introdução . . . . .	193
7.2	Conceito de probabilidade . . . . .	193
7.3	Probabilidade da união de eventos . . . . .	196
7.4	Probabilidade condicional . . . . .	198
7.5	Eventos independentes . . . . .	200
7.6	Teorema de Bayes . . . . .	201

7.7	Exercícios . . . . .	204
<b>8</b>	<b>Medidas de associação</b>	<b>205</b>
8.1	Introdução . . . . .	205
8.2	Medidas de associação . . . . .	205
8.2.1	Diferença absoluta de riscos (DAR) . . . . .	208
8.2.2	Número necessário para tratar . . . . .	208
8.2.3	Risco relativo . . . . .	209
8.2.4	Diferença relativa de riscos . . . . .	210
8.2.5	Resumo das medidas de associação apresentadas até o momento . . .	211
8.2.6	Razão de chances ( <i>odds ratio</i> ) . . . . .	212
8.2.7	Razão de chances e risco relativo . . . . .	214
8.3	Medidas de associação no R . . . . .	217
8.4	Exercícios . . . . .	223
<b>9</b>	<b>Variáveis aleatórias</b>	<b>225</b>
9.1	Noção geral de variável aleatória . . . . .	225
9.2	Valor esperado de uma variável aleatória discreta . . . . .	229
9.3	Variância de uma variável aleatória discreta . . . . .	231
9.4	Transformação linear . . . . .	232
9.5	Soma de variáveis aleatórias . . . . .	235
9.6	Independência de variáveis aleatórias . . . . .	238
9.7	Exercício . . . . .	240
<b>10</b>	<b>Distribuições de variáveis aleatórias discretas</b>	<b>241</b>
10.1	Introdução . . . . .	241
10.2	Distribuição binomial . . . . .	241
10.2.1	Probabilidades de uma distribuição binomial . . . . .	242
10.2.2	Valor esperado e variância de uma distribuição binomial . . . . .	253
10.3	Distribuição de Poisson . . . . .	255
10.3.1	Valor esperado e variância de uma distribuição de Poisson . . . . .	260
10.3.2	Aproximação da distribuição binomial pela de Poisson . . . . .	260
10.4	<b>Distribuição geométrica</b> . . . . .	261
10.4.1	<b>Probabilidades de uma distribuição geométrica</b> . . . . .	262
10.4.2	<b>Valor esperado e variância de uma distribuição geométrica</b> .	263
10.5	Exercícios . . . . .	264
<b>11</b>	<b>Funções densidade de probabilidades</b>	<b>265</b>
11.1	Introdução . . . . .	265
11.2	Histograma de variáveis contínuas. Recordação . . . . .	265
11.3	Função densidade de probabilidade . . . . .	268
11.4	<b>Integral da função densidade de probabilidade</b> . . . . .	272
11.5	Propriedades da função densidade de probabilidade . . . . .	274
11.6	Distribuição uniforme . . . . .	275
11.7	Distribuição normal ou gaussiana . . . . .	279

11.7.1	Valores importantes da variável Z padronizada . . . . .	284
11.7.2	<b>Aproximação da distribuição binomial pela normal</b> . . . . .	285
11.7.3	<b>Aproximação da distribuição de Poisson pela normal</b> . . . . .	288
11.8	<b>Distribuição exponencial</b> . . . . .	290
11.9	Transformação de variáveis e variáveis independentes . . . . .	292
11.10	Exercícios . . . . .	294
<b>12</b>	<b>Avaliação de testes diagnósticos</b>	<b>295</b>
12.1	Introdução . . . . .	295
12.2	Teste dicotômico . . . . .	296
12.2.1	Sensibilidade e especificidade . . . . .	298
12.2.2	Valores preditivo positivo e negativo . . . . .	298
12.2.3	Influência dos fatores que afetam os valores preditivos positivo e negativo	300
12.2.4	Aplicações que mostram a influência dos determinantes de VPP e VPN	303
12.2.5	Razão de verossimilhança . . . . .	306
12.2.6	Influência da razão de verossimilhança sobre a probabilidade pós-teste	310
12.3	Variável de teste categórica ordinal . . . . .	311
12.4	Variável de teste contínua . . . . .	313
12.4.1	Curva ROC . . . . .	313
12.4.2	Comparação de testes . . . . .	318
12.4.3	Uso da razão de verossimilhança em testes com variáveis contínuas . .	320
12.5	Análise de testes diagnósticos no R . . . . .	322
12.6	Exercícios . . . . .	338
<b>13</b>	<b>Estimadores</b>	<b>342</b>
13.1	Introdução . . . . .	342
13.2	Estimativas de parâmetros populacionais . . . . .	342
13.2.1	Amostras de uma distribuição de probabilidades . . . . .	343
13.2.2	Propriedades de estimadores . . . . .	346
13.2.3	Estimadores da variância de uma população . . . . .	349
13.3	Teorema do limite central . . . . .	351
13.4	Aproximação pela normal da proporção de eventos . . . . .	357
13.5	Exercícios . . . . .	361
<b>14</b>	<b>Intervalo de confiança</b>	<b>362</b>
14.1	Introdução . . . . .	362
14.2	Intervalo de confiança - IC . . . . .	362
14.3	Interpretação do intervalo de confiança . . . . .	366
14.4	IC para a média quando a variância não é conhecida . . . . .	370
14.5	Intervalo de confiança para a variância . . . . .	373
14.6	<b>Distribuição qui ao quadrado</b> . . . . .	375
14.7	Intervalo de confiança para proporções . . . . .	376
14.8	Resumo para obtenção de intervalos de confiança de um parâmetro . . . . .	378
14.9	Exercícios . . . . .	381



<b>15 Testes de hipóteses</b>	<b>382</b>
15.1 Introdução . . . . .	382
15.2 Exemplo inicial (primeiro cenário) . . . . .	382
15.3 Processo para realizar um teste de hipótese . . . . .	383
15.3.1 Segundo cenário . . . . .	386
15.4 Relação entre o intervalo de confiança e o teste de hipótese . . . . .	387
15.5 <b>Interpretação alternativa para o IC</b> . . . . .	389
15.6 <b>IC para proporções em pequenas amostras</b> . . . . .	392
15.7 Tipos de testes (bilateral ou unilateral) . . . . .	394
15.7.1 Exemplos de testes unilaterais . . . . .	395
15.8 Valor de p (p-value) . . . . .	398
15.9 Erro tipo I (erro $\alpha$ ) e erro tipo II (erro $\beta$ ) . . . . .	405
15.10 Exemplo de um teste hipótese no <i>R Commander</i> . . . . .	409
15.11 Poder de um teste e tamanho amostral . . . . .	410
15.11.1 <b>Cálculo do tamanho amostral</b> . . . . .	415
15.12 <b>Teste de hipótese para pequenas amostras</b> . . . . .	417
15.13 Interpretações incorretas do valor p . . . . .	421
15.14 Exercícios . . . . .	423
<b>16 Comparação de médias entre dois grupos</b>	<b>425</b>
16.1 Introdução . . . . .	425
16.2 Comparação de médias de amostras independentes . . . . .	427
16.2.1 Teste t de Student para amostras independentes . . . . .	430
16.2.2 Teste de igualdade de variâncias . . . . .	435
16.2.3 Normalidade dos dados . . . . .	438
16.2.4 Testes de normalidade . . . . .	441
16.2.5 Teste não paramétrico de Wilcoxon para duas amostras . . . . .	443
16.3 Comparação de médias de amostras dependentes . . . . .	447
16.3.1 Teste t para amostras dependentes (teste t pareado) . . . . .	449
16.3.2 Teste de Wilcoxon para amostras pareadas . . . . .	453
16.4 Teste t pareado x Teste t não pareado . . . . .	456
16.5 Resumo das análises para comparar médias entre 2 grupos . . . . .	459
16.5.1 Amostras Independentes . . . . .	460
16.5.2 Amostras dependentes . . . . .	461
16.6 Exercícios . . . . .	461
<b>17 Comparação de proporções</b>	<b>463</b>
17.1 Introdução . . . . .	463
17.2 Comparação de proporções em duas amostras independentes . . . . .	467
17.2.1 Teste qui ao quadrado . . . . .	467
17.2.2 Intervalos de confiança para a DAR, o RR e a RC . . . . .	469
17.2.3 Usando o epiR para o teste do qui ao quadrado e cálculo das medidas de associação . . . . .	471
17.2.4 <b>Alternativas ao teste qui ao quadrado tradicional</b> . . . . .	471
17.2.5 Teste exato de Fisher-Irwin . . . . .	473

17.3	Comparação de proporções em duas amostras dependentes . . . . .	478
17.3.1	Teste de McNemar . . . . .	480
17.3.2	<b>Intervalos de confiança para a diferença de proporções, risco relativo e razão de chances</b> . . . . .	481
17.3.3	Comparação de proporções entre duas amostras dependentes no R . .	482
17.4	Poder estatístico e tamanho amostral . . . . .	486
17.4.1	Usando o R Commander para calcular o tamanho amostral . . . . .	487
17.5	Tabelas r x c . . . . .	492
17.5.1	Análise de uma tabela r x c no R Commander . . . . .	494
17.6	Exercícios . . . . .	497
<b>18</b>	<b>Análise de variância</b>	<b>500</b>
18.1	Introdução . . . . .	500
18.2	Múltiplas comparações . . . . .	503
18.3	Análise de variância com um fator . . . . .	503
18.3.1	Modelo de efeitos fixos . . . . .	503
18.3.2	Teste de hipótese . . . . .	505
18.3.3	Comparação de médias . . . . .	511
18.3.4	Análise de resíduos . . . . .	514
18.3.5	Teste não paramétrico de Kruskal-Wallis . . . . .	517
18.3.6	Análise de variância com um fator no R . . . . .	517
18.4	Análise de variância com medidas repetidas . . . . .	532
18.4.1	Modelo . . . . .	532
18.4.2	Teste de hipótese . . . . .	534
18.4.3	Diagnósticos para verificar o modelo de medidas repetidas . . . . .	537
18.4.4	Intervalos de confiança . . . . .	538
18.4.5	Teste de Friedman . . . . .	538
18.4.6	Análise de variância com medidas repetidas no R . . . . .	538
18.5	Outros tipos de análise de variância . . . . .	557
18.6	Exercícios . . . . .	557
<b>19</b>	<b>Regressão linear</b>	<b>559</b>
19.1	Introdução . . . . .	559
19.2	Equação da reta . . . . .	561
19.3	Método dos mínimos quadrados . . . . .	561
19.4	Modelo de regressão linear . . . . .	565
19.4.1	Teste de hipótese . . . . .	565
19.4.2	Intervalos de confiança para os coeficientes de regressão . . . . .	569
19.4.3	Coefficiente de determinação . . . . .	572
19.4.4	Validação do modelo de regressão linear . . . . .	573
19.5	Análise de Regressão no <i>R Commander</i> . . . . .	574
19.6	Coefficiente de correlação linear . . . . .	577
19.6.1	Teste de hipótese bilateral e intervalos de confiança para o coeficiente de correlação . . . . .	579
19.6.2	Cálculo do coeficiente de correlação no <i>R Commander</i> . . . . .	580

19.6.3	Coeficiente de correlação de Spearman . . . . .	582
19.7	Exercícios . . . . .	583
<b>20</b>	<b>Análise de sobrevida</b>	<b>585</b>
20.1	Introdução . . . . .	585
20.2	Conjunto de dados utilizado neste capítulo . . . . .	585
20.3	Obtendo a curva de sobrevida no R . . . . .	588
20.4	Estimando a probabilidade de sobrevida . . . . .	591
20.5	Obtendo as probabilidades de sobrevida em instantes específicos . . . . .	595
20.6	Obtendo a curva de sobrevida para diferentes estratos . . . . .	596
20.7	Comparação de funções de sobrevida em diferentes estratos . . . . .	599
20.8	Exercício . . . . .	604
<b>A</b>	<b>Instalação do R, <i>RStudio</i> e <i>R Commander</i></b>	<b>605</b>
A.1	O que é o R? . . . . .	605
A.2	Vantagens do R . . . . .	606
A.3	Instalação do R e do pacote <i>R Commander</i> . . . . .	606
A.4	Instalação do <i>RStudio</i> . . . . .	608
A.5	Console do RStudio . . . . .	610
A.6	Instalação do pacote do <i>R Commander</i> a partir do <i>RStudio</i> . . . . .	613
<b>B</b>	<b>Código da função <code>paired_proportions</code></b>	<b>620</b>
	<b>Referências Bibliográficas</b>	<b>623</b>

# Capítulo 1

## Introdução

### 1.1 O que é Estatística?

Uma definição comum em livros de Estatística considera a Estatística como a ciência que se preocupa com a organização, descrição, análise e interpretação dos dados experimentais (Costa Neto, 1977). Pode-se acrescentar a essa definição o próprio planejamento de experimentos, já que o delineamento experimental influencia a análise e a interpretação dos resultados. Ao longo deste texto, cada um dos itens dessa definição serão esmiuçados.

### 1.2 Por que estudar Estatística?

Há diversos motivos por que os profissionais de saúde devam conhecer os princípios básicos de Estatística, alguns dos quais são citados abaixo:

- Planejamento de estudos. O conhecimento dos princípios básicos de Estatística é importante antes mesmo de um experimento ser realizado. Não é raro estudos serem realizados sem atenção aos princípios básicos de planejamento de experimentos e, ao final, os autores verificarem que o tamanho amostral ou o delineamento do experimento não foram adequados para responder às perguntas formuladas;
- Interpretação de estatísticas vitais e dados epidemiológicos;
- Interpretação de resultados de estudos publicados na literatura ou em outras fontes que visam a avaliar efeitos de tratamentos, qualidade de testes de diagnóstico e etiologia de doenças.

Embora a maior parte das análises estatísticas publicadas na literatura da área de saúde ainda sejam relativamente simples (por exemplo, teste t e teste qui-quadrado), tem havido um crescimento paulatino de análises mais complexas (regressão logística, modelos proporcionais de Cox, análises multivariadas) (Yi et al., 2015). Isso implica em uma maior necessidade de algum conhecimento de Estatística para aqueles que desejem compreender os estudos realizados na área de saúde. Por outro lado, muitos trabalhos têm avaliado a qualidade das análises estatísticas e da descrição dessas análises na área de saúde. Apesar de haver uma

melhoria nesses aspectos, diversos problemas ainda são detectados em um bom número de publicações: tamanho amostral insuficiente, análises equivocadas, descrição deficiente dos métodos utilizados, conclusões não suportadas pelos dados, etc. (Altman, 1998), (Parsons, Nick R et al., 2012), (Fernandes-Taylor et al., 2011). A situação é ainda mais alarmante em estudos que envolvem pesquisas com animais (Kikenny et al., 2009).

## 1.3 Dado, informação e conhecimento

Na prática clínica, obter dados, interpretá-los e relacioná-los são atividades centrais no processo de assistência à saúde. Shortliffe e Barnet (Shortliffe and Cimino, 2014) conceituam dado clínico como toda a observação isolada de um paciente: temperatura axilar, hematócrito, história pregressa de sarampo, cor dos olhos, etc. Um dado pode ser caracterizado pelas seguintes propriedades: a entidade à qual ele se aplica; o que está sendo observado (parâmetro), o seu valor, o instante da observação e o método de observação. Assim, ao medirmos a temperatura axilar do paciente P por meio de um termômetro clínico, às 7:00 horas do dia 10 de agosto de 2014, e obtendo o valor de 38 Graus Celsius, teríamos uma observação (dado), com as seguintes propriedades:

Entidade – paciente P

Parâmetro – temperatura axilar

Valor – 38°C

Instante – 10/08/2014, às 7:00

Método – termômetro clínico

Van Bemmél e Musen (van Bemmél and Musen, 1997) definem dado como a representação de observações ou conceitos de modo adequado para a comunicação, a interpretação e o processamento por seres humanos ou máquinas. Esses mesmos autores definem informação como fatos úteis e com significados extraídos dos dados. Para um médico, o valor da temperatura acima (38°C) em um ser humano adulto indica que esse paciente está febril. O médico realizou uma interpretação desse dado; ele tem agora uma informação.

Assim a representação de um dado define uma forma ou sintaxe para o dado. O conteúdo ou semântica dessa representação define o significado ou interpretação do dado. A informação é o dado mais o seu significado (Bernstam et al., 2010).

O conhecimento é composto de fatos e relacionamentos usados ou necessários para se obter um *insight* ou resolver problemas (van Bemmél and Musen, 1997), ou informação considerada verdadeira (Bernstam et al., 2010). Pelo raciocínio indutivo, com os dados interpretados, coletados de muitos pacientes (ou por processos similares), novas informações são adicionadas ao corpo do conhecimento na medicina. Esse conhecimento é usado para a interpretação de outros dados. Por exemplo, voltando ao paciente em questão, a informação de que ele possui febre juntamente com outros fatos observados sobre o paciente e da associação desses fatos com determinadas doenças podem levar o médico a um diagnóstico para o paciente.

Podemos então considerar a informação como a interpretação de um dado dentro de um determinado contexto e o conhecimento a interpretação e/ou relacionamento de informações em um contexto mais amplo (Figura 1.1).

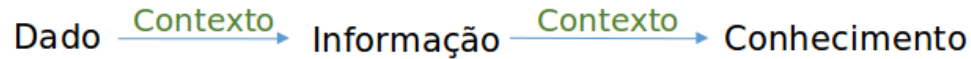


Figura 1.1: Relacionamento entre dado, informação e conhecimento.

Há diversas formas como os dados se apresentam nas diversas atividades humanas. Vamos apresentar abaixo alguns exemplos da área de saúde.

### **Narrativa**

A narrativa é uma das principais formas de comunicação entre os seres humanos. Um exemplo de um trecho de um prontuário de um paciente é mostrado abaixo:

“9:00hs - apresenta-se consciente, comunicativo, icterico, aceitou o desjejum oferecido, tomou banho de aspersão, deambulando, afebril, dispneico, normotenso, taquicárdico, mantendo venóclise por scalp em MSE, com bom fluxo, sem sinais flogísticos, abdômen ascítico, doloroso à palpação, SVD com débito de 200ml de coloração alaranjada, eliminação intestinal ausente há 1 dia. Refere algia generalizada.”

Embora seja extremamente útil para a comunicação entre os seres humanos, é muito difícil representar a narrativa de uma forma que possa ser processada pelos computadores, independentemente da intervenção humana. Essa é uma das áreas mais interessantes da inteligência artificial.

### **Medidas numéricas**

Muitos dados utilizados no dia a dia são numéricos: peso, altura, idade, glicose. Esses dados são facilmente processados pelos computadores e analisados por métodos estatísticos.

### **Dados Categóricos**

Dados categóricos são aqueles cujos valores são selecionados dentro de um conjunto limitado de categorias. Por exemplo, o tipo sanguíneo é um dado categórico com quatro níveis ou categorias (A, B, AB e O). Embora essas categorias possam ser codificadas numericamente, por exemplo, A - 1, B - 2, AB - 3, O - 4, elas não podem ser tratadas numericamente. Assim não há sentido em tirar uma média dos tipos sanguíneos de um grupo de pessoas, por exemplo.

### **Sinais biológicos**

Diversos sinais podem ser obtidos a partir das atividades fisiológicas dos seres vivos: eletrocardiograma, eletroencefalograma, eletromiograma, etc. Esses sinais são utilizados para diversas finalidades, entre elas o estabelecimento de um diagnóstico. A figura 1.2 mostra um exemplo de um eletrocardiograma. Em geral esses sinais são representados por uma série temporal de dados numéricos onde, para cada instante, é registrada a intensidade do sinal.

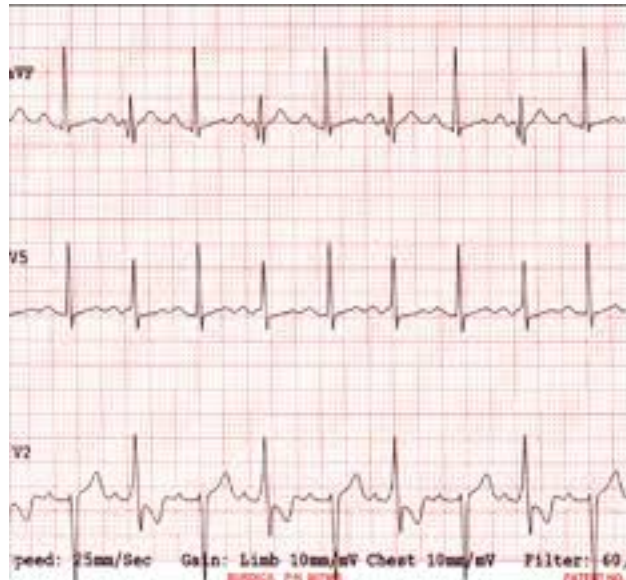


Figura 1.2: Exemplo de um sinal biológico – o eletrocardiograma.

## Imagens

A utilização de imagens para fins de diagnóstico e acompanhamento da evolução de doenças tem avançado muito nas últimas décadas. Atualmente, imagens são geradas por diversos processos: Raios X, Tomografia Computadorizada, Ressonância Magnética Nuclear, Ultrassom, etc. A figura 1.3 mostra um exemplo de uma imagem de Raio X do pulmão. Para imagens monocromáticas, os dados de imagens são registrados por meio da sua localização e intensidade do nível de cinza. Para imagens coloridas, além da localização, são registrados valores relacionados às cores em cada local.



Figura 1.3: Exemplo de uma imagem de Raio X do pulmão.

## Sequência genética

A bioinformática é, possivelmente, uma das áreas que mais tem avançado nos últimos anos. Cada gene é composto de uma sequência de elementos, onde cada elemento da sequência é um dos 4 nucleotídeos: Adenina - A, Guanina - G, Citosina - C e Timina - T. Uma possível sequência genética é mostrada abaixo:

**CTGTGCGGCTCACACCTGGTGGGAAGCTCTCTACCTAGTGTGCGGGGA**

Para o processamento desses dados pelos computadores digitais, todos eles devem ser representados em uma forma que seja compreensível pelos computadores, utilizando o sistema binário.

Em Estatística, os dados são também chamados de variáveis. Para fins de análise estatística, é útil distinguir os tipos de variáveis mais utilizados e as operações que podem ser realizadas com cada tipo.

## 1.4 Variáveis em estudos clínicos

A figura 1.4 é a primeira página de um artigo publicado no *British Medical Journal Open* em 2013 (Rahman et al., 2013). Esse artigo não foi selecionado por alguma razão especial, mas apenas para dar um exemplo de um estudo clínico-epidemiológico. Nesses estudos, em geral, busca-se verificar, entre outras possibilidades, a associação entre variáveis relacionadas ao estado clínico e/ou tratamento aplicado a pacientes e algum desfecho clínico de interesse. Nesse exemplo, é verificada a associação entre duas doenças, osteoartrite e doença cardiovascular.

Não é objetivo deste texto o de discutir o estudo em si, mas sim mostrar os diversos tipos de variáveis que usualmente são coletadas em estudos clínico-epidemiológicos.

A figura 1.5 mostra a tabela 2 do estudo mostrado na figura 1.4, a qual apresenta diversas variáveis que foram medidas no estudo e as respectivas associações com o desfecho, que é a ocorrência ou não de doença cardiovascular. A medida de associação utilizada foi a razão de chances (“*OR - Odds Ratio*”).



# The relationship between osteoarthritis and cardiovascular disease in a population health survey: a cross-sectional study

M Mushfiqur Rahman,<sup>1,2</sup> Jacek A Kopec,<sup>1,2</sup> Jolanda Cibere,<sup>2,3</sup>  
Charlie H Goldsmith,<sup>2,4</sup> Aslam H Anis<sup>1,5</sup>

**To cite:** Rahman MM, Kopec JA, Cibere J, et al. The relationship between osteoarthritis and cardiovascular disease in a population health survey: a cross-sectional study. *BMJ Open* 2013;3:e002624. doi:10.1136/bmjopen-2013-002624

► Prepublication history for this paper are available online. To view these files please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2013-002624>).

Received 23 January 2013  
Revised 14 April 2013  
Accepted 16 April 2013

This final article is available for use under the terms of the Creative Commons Attribution Non-Commercial 2.0 Licence; see <http://bmjopen.bmj.com>

For numbered affiliations see end of article.

**Correspondence to**  
Mr M Mushfiqur Rahman;  
[mrahman@arthritisresearch.ca](mailto:mrahman@arthritisresearch.ca)

## ABSTRACT

**Objectives:** Our objective was to determine the relationship between osteoarthritis (OA) and heart diseases (myocardial infarction (MI), angina, congestive heart failure (CHF)) and stroke using population-based survey data.

**Design:** Cross-sectional study.

**Setting:** Canadian Community Health Survey (CCHS). **Participants:** Adult participants in the CCHS cycles 1.1, 2.1 and 3.1 were included. CCHS provides nationally representative data on health determinants, health status and health system utilisation. We have identified 40 817 self-reported OA subjects and selected 1:1 matched non-OA respondents by age, sex and CCHS cycles.

**Main outcome measures:** Self-reported heart disease was the primary outcome and MI, angina, CHF and stroke were considered as secondary outcomes. Multivariable logistic regression models were used to estimate the ORs after adjusting for sociodemographic status, obesity, physical activity, smoking status, fruit and vegetable consumption, medication use, diabetes, hypertension and chronic obstructive pulmonary disease.

**Results:** The mean age of OA cases was 66 years and 71.6% were women. OA exhibited increased odds of prevalent heart disease, and adjusted overall OR (95% CI) was 1.45 (1.36 to 1.54), 1.35 (1.21 to 1.50) among men and 1.51 (1.39 to 1.64) among women with OA. OA showed increased ORs for angina and CHF in both men and women, and for MI in women. ORs (95% CI) for men and women, respectively, were 1.08 (0.91 to 1.28) and 1.49 (1.28 to 1.75) for MI, 1.76 (1.43 to 2.17) and 1.84 (1.59 to 2.14) for angina, 1.50 (1.13 to 1.97) and 1.81 (1.49 to 2.21) for CHF, and 1.08 (0.83 to 1.40) and 1.13 (0.93 to 1.37) for stroke.

**Conclusions:** Prevalent OA was associated with self-reported heart disease, particularly angina, and CHF in both men and women, after controlling for established risk factors for these conditions. This study provides a rationale for further investigation of the association between OA and heart disease in longitudinal studies for investigating possible biological and behavioural mechanisms.

## ARTICLE SUMMARY

### Article focus

- The purpose of this study was to determine the association between osteoarthritis (OA) and cardiovascular disease (CVD) using data from a large population survey in Canada.
- We analysed the association between OA and CVD, myocardial infarction (MI), angina, congestive heart failure (CHF) and stroke.
- All analyses were carried out for the entire population and separately for men and women.

### Key messages

- OA was significantly associated with any heart disease, angina and CHF in both men and women after controlling for potential confounders.
- We observed that OA was significantly associated with MI among women only and was not associated with stroke.
- The odds of heart disease were 45% higher in persons with OA, compared with age-matched persons without OA and most associations appeared stronger in women than in men.

### Strengths and limitations of this study

- We used a large and representative sample from the Canadian population.
- The results were adjusted for age, body mass index, income, education, physical activity, smoking status, fruit and vegetable consumption, pain medication use, diabetes, hypertension and chronic obstructive pulmonary disease.
- The cross-sectional data prevented us from assessing the temporal exposure–outcome sequence between OA and CVD.
- Self-reported data tend to contain both false-positive and false-negative values, and therefore may introduce bias in the estimates.

## INTRODUCTION

Osteoarthritis (OA) is a highly prevalent chronic disorder and a leading cause of disability among the elderly.<sup>1–3</sup> Although the

Variables	Levels	Overall OR (95% CI)	Men OR (95% CI)	Women OR (95% CI)
Osteoarthritis unadjusted	Yes	1.54 (1.45 to 1.64)	1.47 (1.33 to 1.63)	1.59 (1.47 to 1.72)
Osteoarthritis adjusted	Yes	1.45 (1.36 to 1.54)	1.35 (1.21 to 1.50)	1.51 (1.39 to 1.64)
Age	20–39	Reference	Reference	Reference
	40–49	2.24 (1.50 to 3.33)	1.67 (0.93 to 3.02)	2.62 (1.52 to 4.52)
	50–59	4.28 (2.95 to 6.21)	5.58 (3.27 to 9.52)	3.41 (2.03 to 5.72)
	60–69	7.19 (4.97 to 10.41)	8.47 (4.98 to 14.41)	6.09 (3.64 to 10.19)
	70–79	11.87 (8.19 to 17.20)	13.29 (7.79 to 22.69)	10.28 (6.14 to 17.21)
	≥80	19.33 (13.30 to 28.11)	18.10 (10.51 to 31.18)	18.35 (10.92 to 30.82)
Income	<30000	Reference	Reference	Reference
	30000–50000	0.93 (0.85 to 1.01)	0.90 (0.78 to 1.04)	0.94 (0.85 to 1.04)
	50000–80000	0.82 (0.74 to 0.90)	0.89 (0.76 to 1.04)	0.77 (0.68 to 0.87)
	≥80000	0.69 (0.62 to 0.78)	0.65 (0.55 to 0.78)	0.69 (0.59 to 0.81)
Education	Elementary	Reference	Reference	Reference
	Secondary	0.87 (0.79 to 0.96)	0.97 (0.81 to 1.15)	0.84 (0.74 to 0.94)
	Some postsecondary	0.91 (0.78 to 1.06)	0.98 (0.77 to 1.26)	0.88 (0.73 to 1.07)
	Graduation	0.96 (0.89 to 1.03)	1.05 (0.92 to 1.19)	0.92 (0.83 to 1.01)
Body mass index	<18.4	1.05 (0.82 to 1.35)	0.85 (0.45 to 1.58)	1.06 (0.81 to 1.39)
	18.5–24.9	Reference	Reference	Reference
	25–29.9	0.99 (0.92 to 1.08)	1.09 (0.95 to 1.24)	0.94 (0.84 to 1.04)
	≥30	1.14 (1.03 to 1.26)	1.23 (1.04 to 1.45)	1.09 (0.96 to 1.23)
Physical activity	Active	Reference	Reference	Reference
	Moderate	1.11 (0.99 to 1.24)	1.19 (1.01 to 1.41)	1.07 (0.91 to 1.25)
	Inactive	1.33 (1.21 to 1.47)	1.28 (1.11 to 1.48)	1.37 (1.20 to 1.57)
Smoking	Non-smoker	Reference	Reference	Reference
	Currently	1.16 (1.04 to 1.29)	1.40 (1.16 to 1.69)	1.09 (0.96 to 1.25)
	Former	1.19 (1.11 to 1.29)	1.39 (1.20 to 1.61)	1.16 (1.06 to 1.26)
Fruits and vegetables	0–3 Servings daily	Reference	Reference	Reference
	4–6 Servings daily	1.03 (0.96 to 1.10)	1.10 (0.99 to 1.23)	0.98 (0.89 to 1.07)
	6+ Servings daily	1.15 (1.07 to 1.25)	1.48 (1.31 to 1.68)	1.01 (0.91 to 1.11)
Pain medication use	Yes	1.13 (1.03 to 1.24)	1.22 (1.04 to 1.43)	1.08 (0.96 to 1.21)
Hypertension	Yes	1.98 (1.86 to 2.12)	1.92 (1.72 to 2.14)	2.01 (1.84 to 2.18)
COPD	Yes	2.79 (2.39 to 3.26)	2.98 (2.35 to 3.78)	2.70 (2.19 to 3.31)
Diabetes	Yes	1.90 (1.75 to 2.07)	1.80 (1.57 to 2.06)	1.96 (1.76 to 2.19)

COPD, chronic obstructive pulmonary disease.

Figura 1.5: Tabela 2 do estudo da figura 1.4, com os valores das variáveis analisadas e a medida de associação entre cada variável e o desfecho clínico (doença cardiovascular).

Olhando atentamente a figura 1.5, vamos discutir os tipos e valores das variáveis que foram medidas. Vamos começar pela variável idade (*age*). Pense um pouco sobre como vocês iriam medir a idade de uma pessoa para realizar algum estudo antes de continuar a leitura.

A idade pode ser medida de diversas formas. Pode-se, por exemplo, perguntar à pessoa ou a algum parente a sua idade. Outra possibilidade é a de registrar a data de nascimento e calcular a idade a partir dessa data. Finalmente podem-se criar faixas etárias e, para cada pessoa, registrar a faixa etária a que ela pertence. Essa última opção foi a utilizada para apresentar os valores da idade no estudo; foram utilizadas as seguintes faixas: 20–39, 40–49 e assim por diante.

A idade é uma variável numérica. Uma boa forma de coletá-la é por meio da data de nascimento. Assim nós podemos calculá-la a qualquer momento, com o grau de precisão que desejarmos. Depois, se for desejado, ela pode ser agrupada em qualquer configuração de faixas etárias que quisermos. Caso a idade tenha sido coletada em faixas etárias desde o início, não poderemos depois agrupá-la em outra configuração de faixas etárias e muito menos poderemos saber a idade de cada pessoa, apenas uma aproximação definida pelas faixas etárias criadas.

**Observação importante:** apesar de esse estudo apresentar os valores de idade agrupados por faixas etárias, isso não quer dizer que os autores coletaram os dados sobre idade dessa forma. Eles poderiam ter coletado o valor da idade (número de anos) e, para apresentar os valores de idade no estudo, preferiram agrupá-la em faixas etárias. O mesmo vale para outras variáveis apresentadas na tabela da figura 1.5, como índice de massa corporal, renda, etc.

Outra variável numérica avaliada no estudo é o índice de massa corporal - IMC - (*Body mass index* em inglês). Ela é calculada pela fórmula  $IMC = \text{Peso} / \text{Altura}^2$ . Nesse caso, a melhor opção seria medir o peso e a altura da pessoa e, então, calcular o IMC pela fórmula. Depois, poderemos agrupá-la do jeito que quisermos.

O nível educacional (*Education*) é frequentemente registrado como uma variável categórica, onde os valores da variável são categorias previamente selecionadas. No Brasil, poderíamos usar, por exemplo, analfabeto, fundamental, nível médio, nível superior, etc.

As variáveis *Diabetes*, *Hipertensão*, *DPOC* - doença pulmonar obstrutiva crônica (COPD em inglês) - são binárias no estudo. O indivíduo é ou não diabético, é ou não hipertenso, etc. Essas variáveis poderiam ter sido coletadas de outras formas. Por exemplo, poder-se-ia ter registrado o tipo de diabetes e as medidas de pressão arterial sistólica e diastólica. Como as variáveis são medidas e analisadas em um estudo específico depende de diversos fatores: possibilidades de coleta, tipo de análise a ser realizada, custos, etc.

Verifiquem na figura 1.5 as outras variáveis apresentadas e pensem em outras formas em que elas poderiam ser registradas.

A seção seguinte apresenta como geralmente os dados de um estudo clínico-epidemiológico são organizados em arquivos.

## 1.5 Arquivos de dados

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

Em geral, ao ser realizada uma pesquisa clínica (figura 1.6), tem-se em mente um problema de saúde para o qual o estudo se propõe a investigar algum aspecto. Para isso, é preciso fazer um planejamento do estudo e, em geral, se estabelece um protocolo, detalhando como o estudo será realizado: qual o delineamento ou desenho do estudo, que variáveis serão medidas, o tempo de realização do estudo, as análises que serão realizadas, o número de pacientes que irão compor a amostra do estudo e outros aspectos do trabalho, etc.

Durante a realização do estudo, são definidos um ou mais instrumentos de coleta de dados, que se referem a uma amostra de pessoas que são selecionadas em uma determinada população por meio de um determinado processo de amostragem. Os dados coletados são organizados em estruturas conhecidas como **arquivos de dados**, **conjuntos de dados** ou **bases de dados** e, em seguida, são analisados por meio de pacotes estatísticos. Finalmente os resultados do estudo podem ser publicados em diversos meios: relatórios técnicos, artigos científicos, vídeos, páginas web, etc.

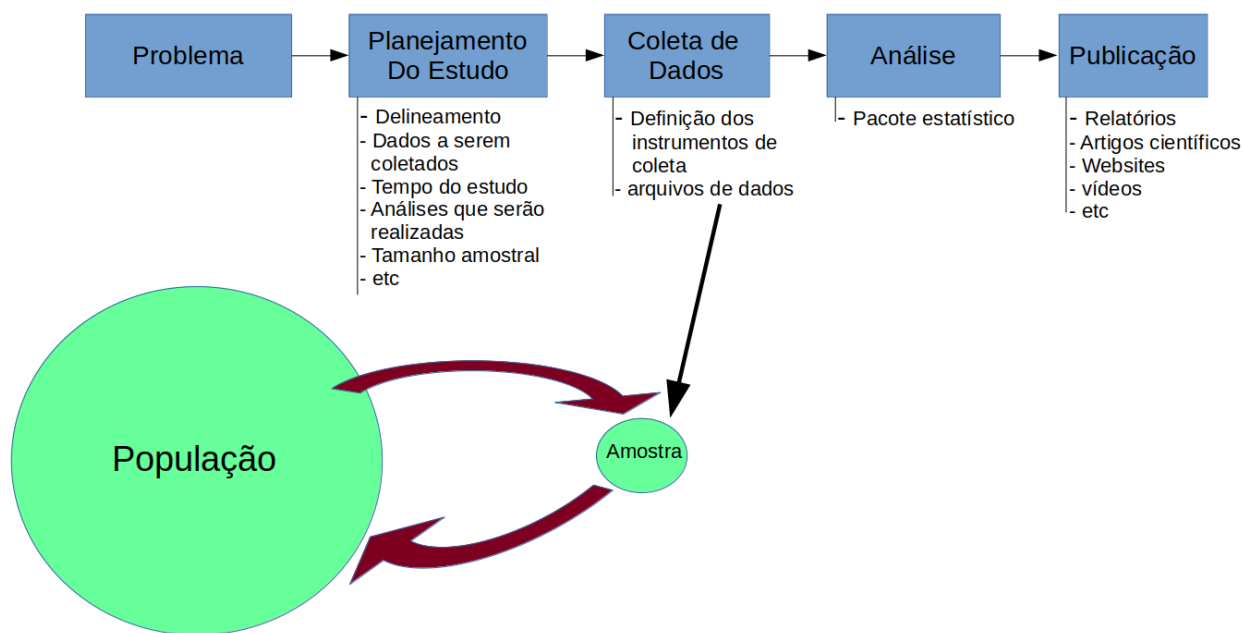


Figura 1.6: Fases de uma pesquisa clínica.

Definidas as variáveis que serão avaliadas no estudo, é preciso criar instrumentos de como essas variáveis serão coletadas para cada entidade participante do estudo. A figura 1.7 mostra uma declaração de nascido vivo. No Brasil, essa declaração é preenchida para todas as pessoas recém-nascidas. A declaração de nascido vivo contém uma série de variáveis que são coletadas por ocasião do nascimento da pessoa: data de nascimento, sexo, peso ao nascer, índice de Apgar, duração da gestação, etc. Esse é um exemplo de um formulário em papel.

## ANEXO A - Modelo da Declaração de Nascido Vivo


 <b>República Federativa do Brasil</b> <b>Ministério da Saúde</b> 1ª VIA - SECRETARIA DE SAÚDE		<b>Declaração de Nascido Vivo</b>		
I	1 Nome do Recém-nascido			
	2 Data e hora do nascimento		3 Sexo	
	2 Data : : : :		M - Masculino F - Feminino I - Ignorado	
II	4 Peso ao nascer	5 Índice de Apgar	6 Detectada alguma anomalia congênita?	
	em gramas	1º minuto 2º minuto	Caso afirmativo, usar o bloco anomalia congênita para descrevê-la. 1 Sim 2 Não 9 Ignorado	
	7 Local da ocorrência	8 Estabelecimento	9 Código CNES	
III	10 Endereço da ocorrência, se fora do estab. ou da resid. da Mãe (rua, praça, avenida, etc)		11 CEP	
	12 Bairro/Distrito	13 Município de ocorrência	14 UF	
	15 Nome da Mãe		16 Cartão SUS	
IV	17 Escolaridade (última série concluída)		18 Ocupação habitual	
	Nível	Série	(Informar anterior, se aposentada/desempregada)	Código CBO 2002
	19 Data nascimento da Mãe	20 Idade (anos)	21 Situação conjugal	22 Raça / Cor da Mãe
V	23 Residência da Mãe		24 CEP	
	25 Logradouro	26 Município	27 UF	
	28 Nome do Pai		29 Idade do Pai	
VI	30 Gestações anteriores		31 Parto	
	30 Gestações anteriores • Nº gestações anteriores • Nº de partos vaginais • Nº de cesáreas • Nº de nascidos vivos • Nº de perdas fetais / abortos		31 Parto 32 Apresentação 33 O Trabalho de parto foi induzido? 34 Tipo de parto 35 Cesárea ocorreu antes do trabalho de parto iniciar? 36 Nascimento assistido por	
	37 Data da última Menstruação (DUM)		38 Número de consultas de pré-natal	39 Mês de gestação em que iniciou o pré-natal
VII	40 Descrever todas as anomalias congênicas observadas			
	41 Data do preenchimento		42 Nome do responsável pelo preenchimento	
	43 Tipo documento		44 Nº do documento	
VIII	45 Cartório		46 Registro	47 Data
	48 Município		49 UF	
	<b>ATENÇÃO: ESTE DOCUMENTO NÃO SUBSTITUI A CERTIDÃO DE NASCIMENTO</b> O Registro de Nascimento é obrigatório por lei. Para registrar esta criança, o pai ou responsável deverá levar este documento ao cartório de registro civil.			

Figura 1.7: Declaração de nascido vivo. Fonte: anexo A do manual de instruções para o preenchimento da declaração de nascido vivo (Ministério da Saúde, 2011).

Os dados também podem ser coletados por meio eletrônico. A figura 1.8 mostra uma tela de um sistema de registro eletrônico de saúde, onde diversos dados relativos à condição clínica do paciente são coletados: queixa principal, doença atual, exames físicos e clínicos, hipóteses diagnósticas, etc.

The screenshot displays the 'Avaliações' (Assessments) window in the GnuHealth system. At the top, there's a header with patient identification: 'Paciente: este1, teste1', 'Gender: Female', 'Age: [blank]', 'Visita: [blank]', 'Prof. da Saúde: Dr. Leonardo', and 'Code: [blank]'. Below this, a tabbed interface shows 'Informações importantes' with sub-tabs for 'Clínico', 'Estado Mental', 'Extra Info', and 'Validation'. The 'Clínico' tab is active, showing sections for 'Queixa Principal' (Main Complaint), 'Doença atual' (Current Disease), 'Clinical and Physical exam', 'Main Condition', 'Other Conditions', 'Hipóteses / DDx' (Hypotheses / Differential Diagnosis), 'Procedimentos' (Procedures), and 'Plano de Tratamento' (Treatment Plan). Each section has a text area for notes and a 'Patologia' (Pathology) or 'Procedimento' (Procedure) dropdown. The bottom of the window shows the 'Estado' (Status) as 'Em progresso' (In progress), 'Início' (Start) as '18-10-2016 11:02:21', and a 'Fim' (End) field. A 'Discharge' button is visible on the right.

Figura 1.8: Interface do [GnuHealth - Hospital Management Information System \(GNU GPL\)](#).

Para fins de análise estatística, os dados coletados são organizados em estruturas denominadas **arquivos de dados** (figura 1.9). Diversos termos são utilizados para se referirem aos componentes de um arquivo de dados. A **unidade de análise**, ou **unidade de observação**, é a menor entidade a ser considerada em um estudo. Na grande maioria dos estudos clínico-epidemiológicos, a unidade de análise é uma pessoa, mas, em outros estudos, a unidade de análise pode ser a escola, uma área geográfica (município, bairro), um animal, etc., dependendo do escopo do trabalho. Para cada unidade de análise, são coletadas uma ou mais variáveis de interesse na investigação.

As variáveis coletadas para cada unidade de análise são armazenadas em um ou mais registros. O conjunto de todos os registros forma o **arquivo de dados, ou conjunto de dados, ou base de dados**.



Unidade de Análise {

registro →

Base de Dados, Conjunto de dados

	age	height	menarche	sex	igf1	tanner	testvol	weight
1	NA	NA	NA	NA	90	NA	NA	NA
2	NA	NA	NA	NA	88	NA	NA	NA
3	NA	NA	NA	NA	164	NA	NA	NA
4	NA	NA	NA	NA	166	NA	NA	NA
5	NA	NA	NA	NA	131	NA	NA	NA
6	0.17	NA	NA	1	101	1	NA	NA
7	0.17	NA	NA	1	97	1	NA	NA
8	0.17	NA	NA	1	106	1	NA	NA
9	0.17	NA	NA	1	111	1	NA	NA
10	0.17	NA	NA	1	79	1	NA	NA
11	0.17	NA	NA	1	43	1	NA	NA
12	0.17	NA	NA	1	64	1	NA	NA
13	0.25	NA	NA	1	90	1	NA	NA
14	0.25	NA	NA	1	141	1	NA	NA
15	0.42	NA	NA	1	42	1	NA	NA
16	0.50	NA	NA	1	43	1	NA	NA
17	0.67	NA	NA	1	132	1	NA	NA
18	0.75	NA	NA	1	43	1	NA	NA
19	0.75	NA	NA	1	36	1	NA	NA
20	1.00	NA	NA	1	86	1	NA	NA
21	1.16	NA	NA	1	44	1	NA	NA
22	1.50	NA	NA	1	68	1	NA	NA
23	1.50	NA	NA	1	89	1	NA	NA
24	1.58	NA	NA	1	101	1	NA	NA
25	1.67	NA	NA	1	115	1	NA	NA
26	1.67	NA	NA	1	53	1	NA	NA
27	1.75	NA	NA	1	94	1	NA	NA
28	1.83	NA	NA	1	95	1	NA	NA
29	1.92	NA	NA	1	76	1	NA	NA
30	2.00	NA	NA	1	79	1	NA	NA

Figura 1.9: Arquivo de dados em estudos transversais. Fonte: conjunto de dados *juul2* do pacote *ISwR* (GPL-2 | GPL-3).

Na figura 1.9, temos um exemplo de um estudo transversal que coletou dados de 1339 pessoas, principalmente pessoas em idade escolar, e as variáveis que foram coletadas são: **idade** (*age*) em anos, **altura** (*height*) em cm, **menarca** (*menarche*: se já ocorreu ou não, codificada como 1, correspondente a não; e 2, correspondente a sim), **sexo** (*sex*: 1 – menino, 2 – menina), **igf1**, que é o fator de crescimento parecido com a insulina, em ug/l, **classificação de tanner** (*tanner*), que se refere aos estágios da puberdade, classificados de I a V, **volume testicular** (*testvol*) em ml e **peso** (*weight*) em kg. A figura mostra o arquivo de dados com as 1339 observações, ou unidades de análises. Cada unidade de análise possui um registro ou linha nesse arquivo, porque as variáveis para cada unidade de análise foram medidas uma única vez, o que caracteriza o estudo como transversal. Cada variável é representada em uma coluna no arquivo e cada linha ou registro contém os valores de cada variável para a entidade

correspondente.

O *NA* que aparece no arquivo significa não disponível (*not available*, em inglês), indicando que o valor da variável não foi coletado para a unidade de observação correspondente, ou não faz sentido em coletar aquela variável para a entidade a que ela se refere. Por exemplo a altura e idade não foram coletadas para os 5 primeiros indivíduos no arquivo da figura. Todos os valores de menarca para os meninos assumem o valor *NA*, já que não há sentido em verificar a ocorrência de menarca em meninos. Da mesma forma, a variável volume testicular assume o valor *NA* para as meninas.

Em estudos longitudinais, as variáveis são medidas em mais de um instante para cada unidade de análise. A figura 1.10 mostra duas formas de organizar os dados em um estudo longitudinal. Os dados nesse exemplo se referem a medidas da frequência cardíaca realizadas antes e após a administração de enalaprilato em 9 pacientes com insuficiência cardíaca congestiva. A frequência cardíaca foi medida nos instantes 0 (antes da administração do medicamento), 30, 60 e 120 min após o uso do medicamento.

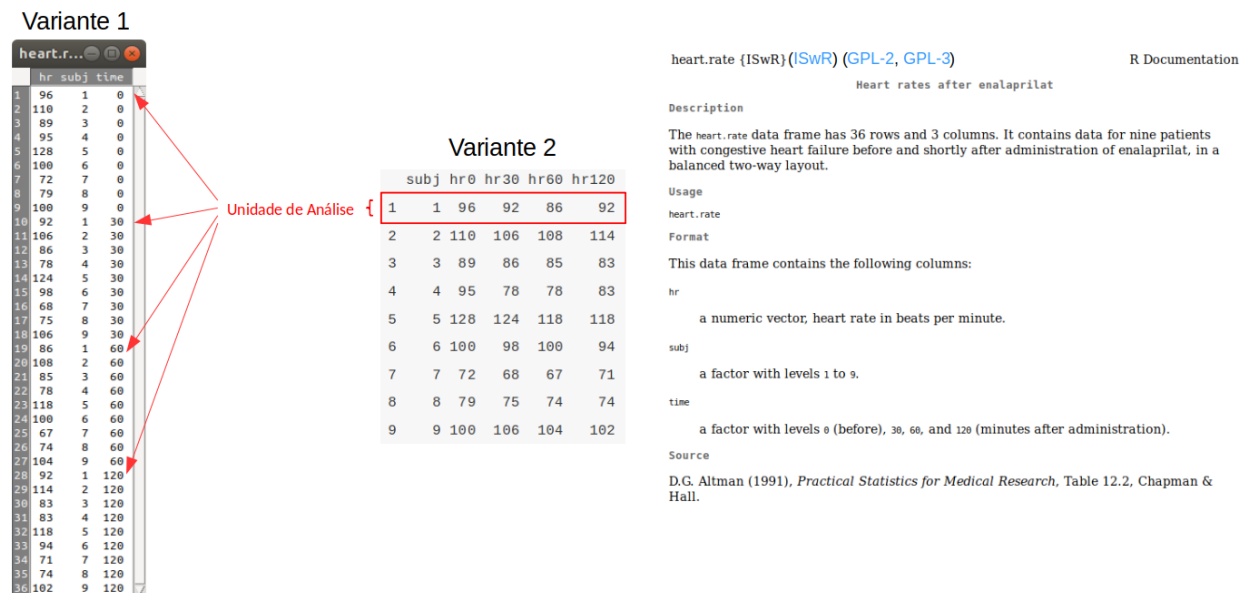


Figura 1.10: Duas formas de organização dos dados em estudos longitudinais. Fonte: conjunto de dados *heart.rate*, do pacote *ISwR* (GPL-2 | GPL-3).

No arquivo indicado pela *variante 1*, há uma única variável (*hr*) para registrar os valores da frequência cardíaca. Uma outra variável (*time*), indica o instante em que a frequência cardíaca foi medida e, finalmente, a variável *subj* identifica cada paciente do estudo. Assim, nesse arquivo, cada paciente possui 4 registros, um para cada instante em que a frequência cardíaca foi medida. Então, o arquivo contém um total de 36 registros, 4 registros por unidade de análise (ou pacientes).

No arquivo indicado pela *variante 2*, há 5 variáveis, onde a variável *subj* identifica cada indivíduo e as variáveis *hr0*, *hr30*, *hr60* e *hr120* correspondem às frequências cardíacas de



cada indivíduo nos instantes 0, 30, 60 e 120 min após a administração do enalaprilato. Nesse arquivo, há um registro para cada unidade de análise, ou paciente.

Os arquivos podem ser armazenados em diversos formatos:

- Arquivos de pacotes estatísticos
  - SPSS, Stata, R, SAS, Prisma, Statistica, etc.
- Planilhas eletrônicas
  - Excel, Calc (Libre-office)
- Bancos de dados
  - Oracle, PostgreSQL, MySQL, SQL-Server, Access, bancos XML, bancos NoSQL, etc.

Cada pacote estatístico geralmente possui um formato específico de armazenamento dos dados. Também podem ser utilizadas planilhas eletrônicas para coletar os dados, que podem ser exportados para serem analisados em algum pacote estatístico. Também é possível obter os dados a partir de bancos de dados que dão suporte a sistemas de prontuários eletrônicos ou a sistemas de gerenciamento da pesquisa clínica, por exemplo.

Neste livro, será utilizado o *R*, que é uma linguagem e um ambiente para a realização de análises estatísticas e construção de gráficos, altamente extensível. O *R* é disponível como software livre sob os termos da Licença Pública Geral GNU da Free Software Foundation.

O apêndice A apresenta o passo a passo para a instalação do R, de um programa que oferece um ambiente integrado de desenvolvimento baseado no R (*R Studio*), e de um pacote que fornece uma interface gráfica para a utilização do R (*Rcmdr*).

## 1.6 Escalas de medidas

O conteúdo desta seção, nos trechos relativos à classificação de variáveis, pode ser visualizado neste [vídeo](#).

Existem diversos níveis de mensuração das variáveis. Esses níveis de mensuração são denominados escalas de medida. A escala na qual uma variável é medida tem implicações na forma como os dados são apresentados e resumidos e nas técnicas estatísticas utilizadas para analisá-los.

As escalas de medida são: escala nominal, escala ordinal, escala intervalar e escala de razão.

### 1.6.1 Escala nominal

Essa é a escala de medição mais simples. Nessa escala, os valores da variável são categorias mutuamente exclusivas e exaustivas. As categorias não possuem uma ordem.

Ex: religião, tipo sanguíneo, nacionalidade.

Variáveis medidas na escala nominal são denominadas **variáveis categóricas nominais**. Em geral essas variáveis são descritas por meio do percentual de cada um dos seus valores possíveis e visualizadas graficamente por meio de diagramas de barras ou diagramas de setores.

### 1.6.2 Escala ordinal

Nessa escala, os valores das variáveis são também categorias como nas variáveis nominais, mas essas categorias podem ser ordenadas de acordo com algum critério.

Ex: escolaridade (quando medida pelas categorias fundamental, médio e superior), estágio de câncer (0, I, II, III, IV).

**A diferença entre duas categorias adjacentes não é a mesma ao longo da escala.**

Assim não se pode afirmar que a variação de gravidade do estágio de câncer I para o II é a mesma do estágio II para o III ou do III para o IV.

Variáveis medidas na escala ordinal são denominadas **variáveis categóricas ordinais**. Em geral essas variáveis são descritas por meio do percentual de cada um dos seus valores possíveis e visualizadas graficamente por meio de diagramas de barras ou diagramas de setores.

Quando o número de categorias é maior que 2, também utiliza-se o termo **variáveis multicategóricas**.

#### Observações:

- Quando há somente duas categorias, a variável é chamada **categórica binária** ou **categórica dicotômica**.
- Frequentemente, as variáveis dicotômicas utilizam as categorias *Sim/Não* ou *0/1*. A designação *0/1* vem da forma como se codificam essas variáveis: em geral, ao indivíduo que apresenta a característica de interesse (*Sim*), atribui-se o valor *1*, e para o que não a apresenta (*Não*), atribui-se o valor *0*. Por exemplo, num estudo onde a variável *fumante* é dicotômica, em geral, os fumantes recebem o valor *1* (correspondente a *Sim*) e os não-fumantes recebem o valor *0* (correspondente a *Não*).
- Muitas variáveis dicotômicas assumem dois valores, mas um não é a negação do outro. Nesses casos, outras codificações podem ser utilizadas, como atribuir arbitrariamente *1* e *2* a cada uma das duas categorias. Por exemplo, para a variável *Sexo*, pode-se atribuir o valor *1* para o sexo masculino e o valor *2* para o sexo feminino.
- Variáveis multicategóricas podem ser transformadas em binárias, combinando-se as categorias. Por exemplo, considere a variável autopercepção de saúde do indivíduo, com cinco categorias ordenadas:

- 1 - excelente
- 2 - muito boa
- 3 - boa
- 4 - razoável
- 5 - ruim

Essa variável pode ser transformada para:

- 1 - ruim ou razoável
- 0 - bom, muito bom ou excelente.

O contrário é impossível: não se pode dividir em mais categorias uma variável que foi originalmente registrada como dicotômica.

- Algumas variáveis apresentam algumas categorias ordenadas e outras que não se encaixam em nenhuma ordenação. Por exemplo: na variável *emprego*, as categorias *integral*, *parcial* e *desempregado* podem ser ordenadas, mas onde posicionar a categoria *aposentado*?

### 1.6.3 Escala intervalar

A escala intervalar é uma escala numérica. A mesma diferença entre dois valores possui o mesmo significado ao longo da escala. Por exemplo, seja a variável idade de uma pessoa, medida em anos. Então, a diferença entre os valores de idade 18 e 20 (2) representa a mesma diferença, em termos de tempo decorrido, que entre as idades 60 e 62.

### 1.6.4 Escala de razão

A escala de razão é uma escala numérica que possui um zero absoluto. Para entender melhor o que isso significa, considere duas escalas muito utilizadas para medir a temperatura: escala Kelvin (K) e escala Celsius (C).

A relação entre essas duas medidas é dada pela fórmula:

$$K = C + 273,16$$

Consideremos as temperaturas

$$K_1 = 20 \text{ K}$$

$$K_2 = 40 \text{ K}$$

A razão entre  $K_2$  e  $K_1$  é 2, o que quer dizer que  $K_2$  representa um estado cuja agitação térmica é o dobro do estado designado por  $K_1$ .

Entretanto, dadas as temperaturas

$$C_1 = 20^\circ\text{C}$$

$$C_2 = 40^\circ\text{C}$$

apesar de a razão entre  $C_2$  e  $C_1$  ser numericamente igual a 2, o estado designado por  $C_2$  não apresenta o dobro de agitação térmica que o estado designado por  $C_1$ . Na escala Kelvin, teríamos:

$$C_1 \rightarrow K_1 = 293,16\text{K}$$

$$C_2 \rightarrow K_2 = 313,16\text{K}$$

$K_2/K_1 = 1,07$ . O estado designado por  $C_2$  apresenta uma agitação térmica 1,07 vezes maior que a agitação térmica do estado designado por  $C_1$ .

Por outro lado, a diferença entre as duas temperaturas em ambas as escalas (20K e 20°C) representam a mesma variação de agitação térmica entre os dois estados. Desse modo, dizemos que a escala Celsius é uma escala intervalar e a escala Kelvin é uma escala de razão.

As variáveis numéricas também podem ser classificadas como **discretas** ou **contínuas**. As variáveis contínuas podem assumir qualquer valor numérico dentro de um intervalo dado.

Ex: peso (kg), altura (m), glicemia de jejum (mg/dl), pH (adimensional).

As variáveis numéricas discretas são aquelas cujos valores pertencem a um conjunto enumerável, sendo esse frequentemente o conjunto dos números inteiros não negativos.

Ex: número de fraturas, número de extrassístoles, número de consultas médicas por ano, etc.

Em epidemiologia, os numeradores e denominadores que compõem os indicadores de saúde são frequentemente variáveis discretas (contagem de óbitos, contagem da população etc).

Algumas variáveis contínuas são *discretizadas*, sendo a principal delas o **tempo**. O tempo flui continuamente, mas a idade, por exemplo, é geralmente apresentada em anos completos, desprezando-se a fração de tempo além dos anos.

**Observação:** Como os instrumentos de medida possuem sempre algum limite de precisão, na prática, as variáveis contínuas não podem assumir um número infinito de valores em um intervalo. Por exemplo, em uma balança com precisão de 1 kg, uma medida de 62 kg na verdade expressa um valor entre 61,5 kg e 62,5 kg. Assim, rigorosamente falando, uma variável contínua pode, muitas vezes, ser tratada como discreta. Porém a ciência se baseia em modelos e, nos modelos estatísticos, tratamos variáveis cuja precisão depende do aparelho de medição como contínuas. Mesmo variáveis numéricas que são inerentemente discretas muitas vezes são tratadas como contínuas quando podem assumir um número grande de valores possíveis.

## 1.7 Transformação de variáveis

Como já vimos, as variáveis categóricas com mais de duas categorias podem ser transformadas em binárias. De maneira semelhante, variáveis numéricas podem ser recodificadas como categóricas ordinais ou mesmo binárias. Por exemplo, embora a *glicemia de jejum* seja uma variável contínua, ela pode ser transformada em uma variável binária: *glicemia alta*. Essa variável é obtida, dividindo-se os valores de glicemia por meio de um *ponto de corte*. A indivíduos com glicemia de jejum  $< 100$  mg/dl poderia ser atribuído o valor 0, e para os com glicemia de jejum  $\geq 100$  mg/dl, o valor 1.

Variáveis com múltiplas categorias também podem ser criadas dessa forma. Por exemplo, um investigador pode preferir utilizar a seguinte escala para classificar a pressão arterial: baixa/normal/alta. Nesse caso, ele precisaria criar dois pontos de corte.

O que não é possível é desagregar uma variável originalmente registrada num formato mais agregado. O gradiente de agregação de variáveis é:

Continua  $\Rightarrow$  Discreta  $\Rightarrow$  Multicategórica  $\Rightarrow$  Binária

A figura 1.11 resume a classificação de variáveis para finalidades estatísticas.

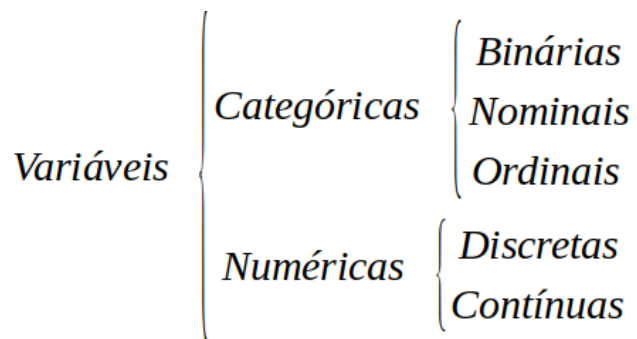


Figura 1.11: Resumo dos tipos de variáveis.

Um grupo especial de variáveis que é muito comum na área de saúde são as escalas ou índices.

## 1.8 Escalas e índices

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

Índices e escalas são dispositivos de redução de dados, onde as várias respostas de um respondente podem ser resumidas num único escore. Não se deve confundir o conceito de escalas utilizado nesta seção com o conceito de escalas de medidas, apresentado na seção 1.6.

Existem inúmeras escalas propostas nas diversas áreas da saúde. Alguns exemplos de escalas são listados abaixo:

- *Geboes Score*
- *Apache II*
- *Charlson Comorbidity Index*
- *Short Form Health Survey (SF-36)*
- *TIMI*
- *Nutritional Risk Screening*

Uma das escalas mais simples é a escala de **Apgar**. Essa escala foi proposta por Virgínia Apgar, uma médica estadunidense e é calculada como mostra a figura 1.12.

	0	1	2
Frequência cardíaca	Ausente	Lenta (< 100bpm)	rápida (> 100bpm)
Respiração	Ausente	Lenta, irregular	Forte, choro
Tonus muscular	Flácido	Flexões nas extremidades	Movimento ativo
Irritabilidade reflexa ao cateter nasal	Sem resposta	Careta	Tosse, espirro
Cor da pele	Cianose central / palidez	Corpo rosado, cianose nas extremidades	Corpo e extremidades rosados

*\*Fonte: Organização mundial da saúde*

-

0

+

-

0

+

-

0

+

-

0

+

-

0

+

Calcular Apgar

0

Figura 1.12: Componentes da escala de Apgar.

Uma criança recém-nascida é avaliada em 5 itens (frequência cardíaca, respiração, tônus muscular, irritabilidade reflexa ao cateter nasal e cor da pele) e recebe uma pontuação de 0 a 2 para cada dimensão de acordo com os critérios mostrados no quadro da figura 1.12. Assim, se a frequência cardíaca estiver ausente, a criança recebe 0 nesse quesito. Se a frequência cardíaca estiver lenta (abaixo de 100 bpm), a criança recebe 1. Se a frequência cardíaca estiver rápida (acima de 100 bpm), a criança recebe 2. Atribuindo a pontuação correspondente para os outros 4 itens e somando-se todos os pontos obtemos o escore de Apgar. Portanto o escore de Apgar é um número inteiro de 0 a 10.

O teste é geralmente realizado no primeiro e quinto minutos após o nascimento (chamados de Apgar de 1 minuto e Apgar de 5 minutos, respectivamente) e é repetido posteriormente se o índice permanecer baixo.

Observando como o valor de Apgar é obtido, podemos considerá-lo como uma variável numérica, medida na escala intervalar, ou os números refletem mais o grau de saúde da criança, sem necessariamente quantificá-la? Por exemplo, se o Apgar fosse uma variável numérica que refletisse o quanto a criança é saudável, então a diferença entre os valores de Apgar 5 e 6, por exemplo, deveria refletir a mesma variação no nível de saúde que a diferença entre os valores de Apgar 9 e 10, ou seja, a diferença de 1 unidade entre dois valores de Apgar deveria significar a mesma variação na quantidade de saúde ao longo de toda a escala de Apgar. Assim, possivelmente, seria mais conveniente considerar a escala de Apgar como uma variável categórica ordinal. Entretanto, há trabalhos publicados na literatura que tratam o Apgar como uma variável numérica discreta. Essa mesma observação possivelmente pode ser aplicada a um bom número das escalas que são propostas na área de saúde.

Há escalas mais complexas do que a escala de Apgar. Uma escala frequentemente utilizada para avaliar a qualidade de vida de pacientes é a escala conhecida como SF36, **Medical**

**Outcomes Short-Form Health Survey.** Essa escala é composta de 36 perguntas (itens) que abordam os domínios:

- capacidade funcional (10 itens)
- aspectos físicos (4 itens)
- dor (2 itens)
- estado geral da saúde (5 itens)
- vitalidade (4 itens)
- aspectos sociais (2 itens)
- aspectos emocionais (3 itens)
- saúde mental (5 itens)
- um item que compara as condições de saúde atual e a de um ano atrás.

O arquivo [SF-36](#) ([CC BY](#)) mostra o questionário e como o score é calculado para cada domínio.

No início, o arquivo mostra como pontuar a resposta a cada questão. Há um total de 11 questões, sendo que algumas questões possuem mais de uma pergunta, totalizando 36 perguntas. Assim, por exemplo, a questão 1 do questionário é mostrada abaixo:

1. Em geral você diria que sua saúde é (circule uma):

Excelente - 1

Muito boa - 2

Boa - 3

Ruim - 4

Muito ruim - 5

Cada resposta da pergunta 1 recebe uma pontuação de acordo com a lista abaixo:

1 - 5,0

2 - 4,4

3 - 3,4

4 - 2,0

5 - 1,0

Em seguida, o arquivo mostra como calcular os scores para cada domínio e as questões que estão relacionadas a cada um dos oito domínios (figura 1.13).

DOMÍNIO	PONTUAÇÃO DA(S) QUESTÃO (ÕES) CORRESPONDENTES	LIMITE INFERIOR	VARIAÇÃO (ESCORE RANGE)
Capacidade funcional	03	10	20
Limitação por aspectos físicos	04	4	4
Dor	07+08	2	10
Estado geral de saúde	01+11	5	20
Vitalidade	09 (somente p/ os itens a + e + g + i )	4	20
Aspectos sociais	06+10	2	8
Limitação por aspectos emocionais	05	3	3
Saúde mental	09 ( somente p/ os itens b + c + d + f + h )	5	25

Figura 1.13: Quadro que mostra como calcular os escores para cada domínio do SF36.

No quadro da figura 1.13, a fórmula para o cálculo do escore para cada domínio é dada por:

$$\text{Domínio} : \frac{(\text{Valor obtido nas questões correspondentes} - \text{limite inferior}).100}{\text{Variação (Score Range)}}$$

Na fórmula acima, os valores de *limite inferior* e *variação* são fixos e especificados no quadro da figura 1.13.

Dois escores (resumo do componente físico e resumo do componente mental) podem também ser derivados do SF-36.

Finalmente o arquivo mostra o questionário com as 11 questões.

Os domínios do SF36 são frequentemente tratados como variáveis numéricas em trabalhos científicos, mas questões semelhantes às levantadas para o Apgar quanto à validade de se tratar os domínios do SF36 como variáveis numéricas podem ser levantadas.

## 1.9 Identificação de variáveis em estudos clínicos

Ao realizar a leitura de um estudo clínico-epidemiológico, é importante identificar as variáveis que foram analisadas e os respectivos tipos. As variáveis podem estar identificadas no próprio texto do trabalho, ou podem aparecer nos resultados gráficos ou em tabelas. Vejam a figura 1.14. A elipse em vermelho destaca as variáveis que aparecem na tabela. O retângulo em amarelo destaca os valores que cada variável à esquerda pode assumir. Assim a variável idade pode assumir os seguintes valores: 20-39, 40-49, 50-59, 60-69, 70-79,  $\geq 80$ . A idade, nesse estudo, é apresentada como uma variável categórica ordinal.



Há uma outra variável não explicitamente identificada na tabela mostrada na figura 1.14. Observe a elipse em verde, onde existe uma coluna que fornece os valores da razão de chances (OR – *Odds Ratio* em inglês) para homens e uma outra para as mulheres. Então temos mais uma variável, gênero, que foi considerada no estudo.

**Table 2** ORs and 95% CIs of heart diseases for osteoarthritis and non-osteoarthritis 1:1 matched samples by age and sex

Variables	Levels	Overall OR (95% CI)	Men OR (95% CI)	Women OR (95% CI)
Osteoarthritis unadjusted	Yes	1.54 (1.45 to 1.64)	1.47 (1.33 to 1.63)	1.59 (1.47 to 1.72)
Osteoarthritis adjusted	Yes	1.45 (1.36 to 1.54)	1.35 (1.21 to 1.50)	1.51 (1.39 to 1.64)
Age	20–39	Reference	Reference	Reference
	40–49	2.24 (1.50 to 3.33)	1.67 (0.93 to 3.02)	2.62 (1.52 to 4.52)
	50–59	4.28 (2.95 to 6.21)	5.58 (3.27 to 9.52)	3.41 (2.03 to 5.72)
	60–69	7.19 (4.97 to 10.41)	8.47 (4.98 to 14.41)	6.09 (3.64 to 10.19)
	70–79	11.87 (8.19 to 17.20)	13.29 (7.79 to 22.69)	10.28 (6.14 to 17.21)
	≥80	19.33 (13.30 to 28.11)	18.10 (10.51 to 31.18)	18.35 (10.92 to 30.82)
Income	<30000	Reference	Reference	Reference
	30000–50000	0.93 (0.85 to 1.01)	0.90 (0.78 to 1.04)	0.94 (0.85 to 1.04)
	50000–80000	0.82 (0.74 to 0.90)	0.89 (0.76 to 1.04)	0.77 (0.68 to 0.87)
	≥80000	0.69 (0.62 to 0.78)	0.65 (0.55 to 0.78)	0.69 (0.59 to 0.81)
Education	Elementary	Reference	Reference	Reference
	Secondary	0.87 (0.79 to 0.96)	0.97 (0.81 to 1.15)	0.84 (0.74 to 0.94)
	Some postsecondary	0.91 (0.78 to 1.06)	0.98 (0.77 to 1.26)	0.88 (0.73 to 1.07)
	Graduation	0.96 (0.89 to 1.03)	1.05 (0.92 to 1.19)	0.92 (0.83 to 1.01)
Body mass index	<18.4	1.05 (0.82 to 1.35)	0.85 (0.45 to 1.58)	1.06 (0.81 to 1.39)
	18.5–24.9	Reference	Reference	Reference
	25–29.9	0.99 (0.92 to 1.08)	1.09 (0.95 to 1.24)	0.94 (0.84 to 1.04)
	≥30	1.14 (1.03 to 1.26)	1.23 (1.04 to 1.45)	1.09 (0.96 to 1.23)
Physical activity	Active	Reference	Reference	Reference
	Moderate	1.11 (0.99 to 1.24)	1.19 (1.01 to 1.41)	1.07 (0.91 to 1.25)
	Inactive	1.33 (1.21 to 1.47)	1.28 (1.11 to 1.48)	1.37 (1.20 to 1.57)
Smoking	Non-smoker	Reference	Reference	Reference
	Currently	1.16 (1.04 to 1.29)	1.40 (1.16 to 1.69)	1.09 (0.96 to 1.25)
	Former	1.19 (1.11 to 1.29)	1.39 (1.20 to 1.61)	1.16 (1.06 to 1.26)
Fruits and vegetables	0–3 Servings daily	Reference	Reference	Reference
	4–6 Servings daily	1.03 (0.96 to 1.10)	1.10 (0.99 to 1.23)	0.98 (0.89 to 1.07)
	6+ Servings daily	1.15 (1.07 to 1.25)	1.48 (1.31 to 1.68)	1.01 (0.91 to 1.11)
Pain medication use	Yes	1.13 (1.03 to 1.24)	1.22 (1.04 to 1.43)	1.08 (0.96 to 1.21)
Hypertension	Yes	1.98 (1.86 to 2.12)	1.92 (1.72 to 2.14)	2.01 (1.84 to 2.18)
COPD	Yes	2.79 (2.39 to 3.26)	2.98 (2.35 to 3.78)	2.70 (2.19 to 3.31)
Diabetes	Yes	1.90 (1.75 to 2.07)	1.80 (1.57 to 2.06)	1.96 (1.76 to 2.19)

COPD, chronic obstructive pulmonary disease.

Figura 1.14: Tabela 2 do estudo de Rahman et al (Rahman et al., 2013) (CC BY-NC), com os valores das variáveis analisadas e a medida de associação entre cada variável e o desfecho clínico (doença cardiovascular).

A tabela 1.1 mostra a classificação das variáveis da figura 1.14.

Tabela 1.1: Classificação das variáveis da tabela mostrada na figura 1.14

Variável	Classificação
osteoartrite	Categórica Binária
Idade	Categórica Ordinal
Renda	Categórica Ordinal
Educação	Categórica Ordinal
Índice de Massa Corporal	Categórica Ordinal
Atividade Física	Categórica Ordinal
Tabagismo	Categórica Ordinal
Frutas e Verduras	Categórica Ordinal
Uso de Medicação	Categórica Binária
Hipertensão	Categórica Binária
Diabetes	Categórica Binária
DPOC	Categórica Binária
Gênero	Categórica Binária

## 1.10 Exercício

- 1) Classifique as variáveis presentes nas tabelas do artigo intitulado “Fatores associados à qualidade de vida sob a perspectiva da terapia medicamentosa em pacientes com asma grave” (Souza et al., 2015) ([CC BY-NC](#)).

# Capítulo 2

## Tabelas de frequências

### 2.1 Introdução

Os conteúdos desta seção e das seções 2.2 e 2.3.1 podem ser visualizados neste [vídeo](#).

Neste capítulo, serão apresentadas diversas formas para apresentar em tabelas a frequência ou porcentagem das categorias de uma variável categórica ou de combinação de categorias de variáveis categóricas em um conjunto de dados. As tabelas assim obtidas são também chamadas de **tabelas de contingência**.

A figura 2.1 mostra a tabela 2 do estudo de Barata e Valet (Barata and Valet, 2018), intitulado “Perfil clínico-epidemiológico de 106 pacientes pediátricos portadores de urolitíase no Rio de Janeiro”. Essa tabela mostra a frequência de ocorrência das categorias das variáveis sexo, cor da pele, idade no início dos sintomas e idade no diagnóstico.

A **frequência de uma categoria** de uma variável categórica em um conjunto de dados é o número de observações daquela categoria da variável no conjunto de dados. Na figura 2.1, vemos que a frequência de mulheres no estudo é 52 e de homens, 54. Em relação à cor da pele, 67 pessoas eram brancas, 30 pardas, 8 negras e 1 amarela.

Se dividirmos a frequência de uma categoria de uma variável pelo número total de observações, obtemos a **proporção da respectiva categoria** da variável no conjunto de dados. Multiplicando essa proporção por 100, obtemos então a **porcentagem da categoria da variável** no conjunto de dados. Na figura 2.1, vemos que a porcentagem de mulheres no estudo é 49,1% e de homens, 50,9%.

**Tabela 2** Características demográficas dos pacientes em seguimento no Hospital Federal dos Servidores do Estado, entre janeiro de 2012 e dezembro de 2014.

	n	%
<b>Gênero</b>		
Feminino	52	49,1
Masculino	54	50,9
<b>Cor da pele</b>		
Branca	67	63,2
Parda	30	28,3
Amarela	1	0,9
Negra	8	7,6
<b>Idade no início dos sintomas<sup>a</sup> (anos)</b>		
<5	17	16,0
≥5 a ≤10	54	50,9
>10 a ≤18	35	33,0
<b>Idade no diagnóstico<sup>b</sup> (anos)</b>		
<5	8	7,5
≥5 a ≤10	52	49,1
>10 a ≤18	46	43,4

Total da amostra: n=106; <sup>a</sup>média 8,9±3,8; <sup>b</sup>média 9,9±3,6.

Figura 2.1: Características demográficas dos pacientes pediátricos portadores de urolitíase no Rio de Janeiro em seguimento no Hospital Federal dos Servidores do Estado. Fonte: (Barata and Valet, 2018) (CC BY).

O conjunto dos pares formados pelas categorias de uma variável em um conjunto de dados, juntamente com as suas respectivas frequências, é chamado de **distribuição de frequências da variável**.

A figura 2.2 mostra a tabela 2 do estudo de Vanin et al. (Vanin et al., 2019), intitulado “Fatores de risco materno-fetais associados à prematuridade tardia”. Essa tabela mostra, entre outras, a distribuição de frequência conjunta das variáveis *sexo* e *maturidade*. A variável *maturidade* possui dois níveis ou categorias: *RNPT* - recém-nascido prematuro tardio e *RNT* - recém-nascido a termo. Para cada categoria da variável *maturidade*, a tabela mostra as frequências de cada categoria da variável *sexo* e entre parênteses a porcentagem da respectiva categoria de *sexo* em relação à frequência total do nível de maturidade correspondente. Por exemplo, entre os *recém-nascidos a termo*, a frequência do sexo feminino é 124, correspondendo a 44,1% do total de crianças *recém-nascidas a termo* (124 de 281).

**Tabela 2** Análise comparativa entre as características fetais e neonatais com o nascimento prematuro tardio e a termo.

	RNPT		RNT		p-valor
	n	(%)	n	(%)	
Pequeno para IG					
Sim	34	(24,1)	22	(7,8)	<0,001
Não	107	(75,9)	260	(92,2)	
Grande para IG					
Sim	5	(3,5)	36	(12,8)	<0,001
Não	136	(96,5)	246	(87,2)	
Sexo					
Masculino	64	(45,4)	157	(55,9)	0,042
Feminino	77	(54,6)	124	(44,1)	

RNPT: recém-nascido prematuro tardio; RNT: recém-nascido a termo; IG: idade gestacional.

Frequência

Porcentagem em relação ao total da coluna

Figura 2.2: Tabela 2 do estudo de Vanin et al., mostrando a frequência conjunta de diversas variáveis e o nível de maturidade das crianças recém-nascidas do estudo. Fonte: (Vanin et al., 2019) (CC BY).

Para mostrar como obtemos distribuições de frequências de variáveis ou combinação de variáveis em um conjunto de dados no ambiente R, vamos utilizar o conjunto de dados *stroke* do pacote *ISwR* (GPL-2 | GPL-3). Esse conjunto de dados contém todos os casos de AVC (acidente vascular cerebral) em Tartu, Estônia, durante 1991-1993, com acompanhamento até 1º de janeiro de 1996.

Na seção seguinte, vamos carregar o conjunto de dados *stroke* no R para podermos analisá-lo.

## 2.2 Carregando conjuntos de dados de pacotes do R

Ao abrir o *R Commander* via *RStudio* ou diretamente a partir da tela de entrada do R, temos acesso ao menu principal e uma janela com duas abas: *Script* e *Markdown*.

Muitos pacotes do R contêm conjuntos de dados que podem ser utilizados para ilustrar os recursos disponíveis no pacote. O conjunto de dados *stroke* pertence ao pacote *ISwR*.

O pacote *ISwR* precisa ser instalado. Os passos para a instalação desse pacote são os mesmos utilizados para a instalação do *R Commander*, seção A.6.

De outra maneira, podemos digitar o comando a seguir na console do *RStudio* e pressionar a tecla *Enter*. De maneira alternativa, com o mesmo comando na área de script do *R Commander* e, **com o cursor na linha do comando**, clicamos no botão *Submit*. Como o nome indica, a função *install.packages* instala o pacote especificado entre aspas.

```
install.packages("ISwR")
```

Antes de abrirmos o conjunto de dados *stroke*, é preciso carregar o pacote *ISwR*. Na sequência deste capítulo, utilizaremos o *R Commander*, carregado a partir do R e não a partir do *RStudio*.

Para carregarmos o pacote *ISwR* a partir do *R Commander*, digitamos `library(ISwR)` na área de *script* do *R Commander* e, **com o cursor na linha do comando**, clicamos no botão *Submeter* (figura 2.3).

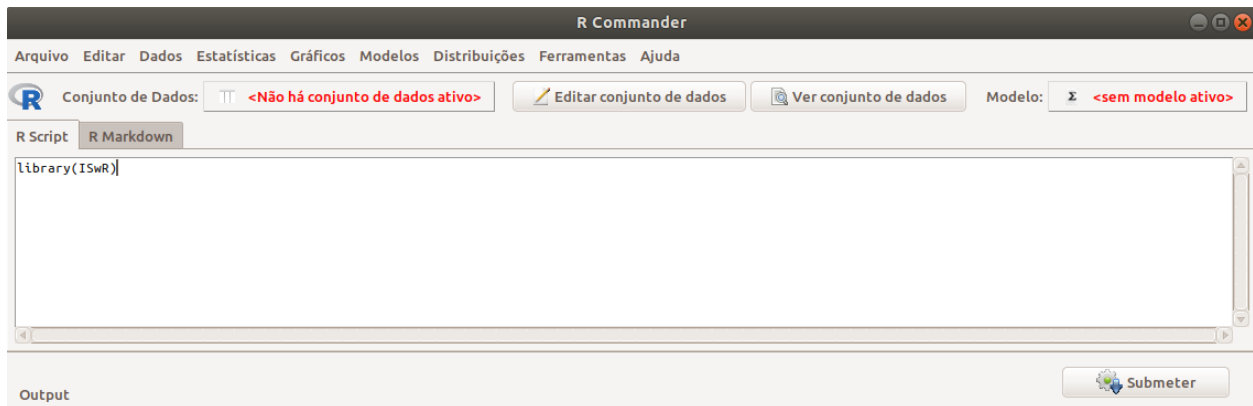


Figura 2.3: Tela do *R commander*, com a digitação da função `library(ISwR)` na área de *Script*.

Ao submetermos a função, ela aparece na área de output do *R Commander* (figura 2.4) e, se houver alguma coisa errada, uma mensagem de erro apareceria na área de mensagens do *R Commander*.

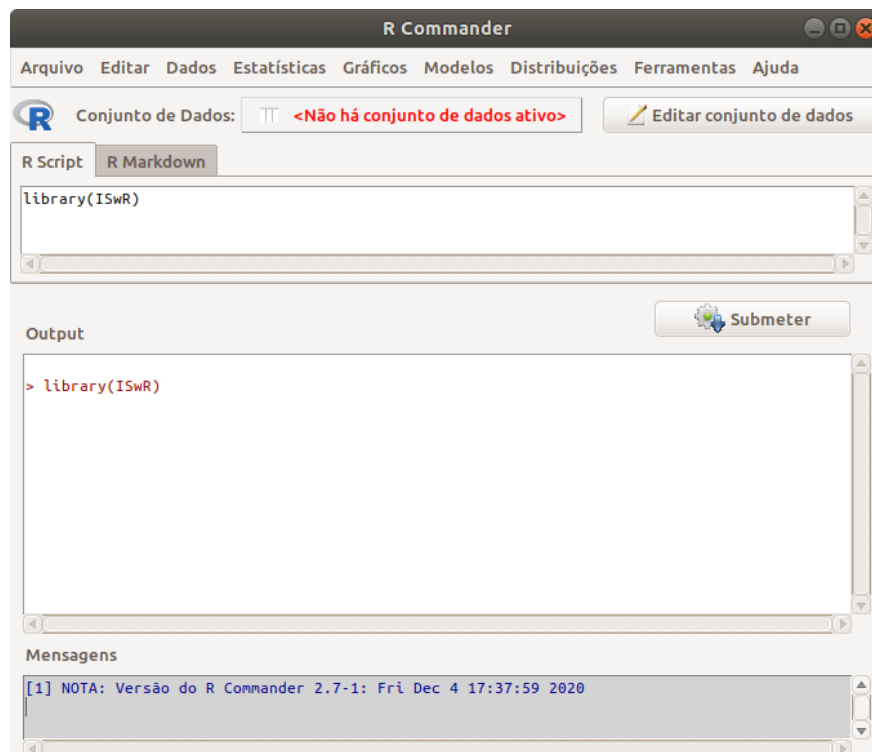


Figura 2.4: Tela do *R Commander* após a execução da função `library` conforme mostrado na figura 2.3.

Alternativamente o pacote *ISwR* poderia ser carregado por meio da opção de menu do *R Commander*:

Ferramentas  $\Rightarrow$  Carregar pacote(s)...

A função `library(ISwR)` carregou a biblioteca *ISwR*, a qual contém uma série de conjuntos de dados que podemos utilizar. Para visualizar e, eventualmente, selecionar um desses conjuntos de dados, selecionamos a opção abaixo no *R Commander* (figura 2.5):

Dados  $\Rightarrow$  Conjunto de dados em pacotes  $\Rightarrow$  Ler dados de pacote 'attachado'

A partir de agora, toda opção a ser selecionada no menu será apresentada como uma sequência de itens a serem selecionados como acima.

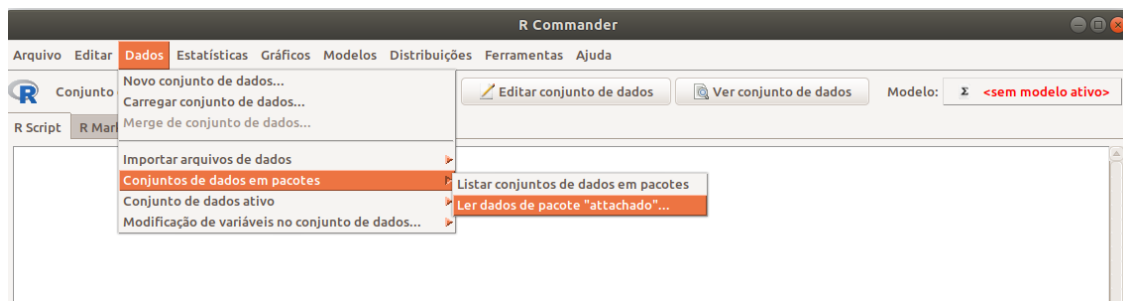


Figura 2.5: Menu do *R Commander* com a opção para carregar arquivos de pacotes do R.

Na tela *Leia dados do pacote* (figura 2.6), observem que alguns pacotes aparecem na área à esquerda da figura: *carData*, *datasets*, *ISwR* e *sandwich*. Para ver a lista dos conjuntos de dados em *ISwR*, demos um duplo clique nesse pacote e uma lista de conjuntos de dados será mostrada à direita (figura 2.6). Rolamos essa lista e clicamos no conjunto *stroke* para selecioná-lo. Para conhecermos a estrutura desse conjunto de dados, clicamos no botão *Ajuda para o conjunto de dados selecionado* (seta verde na figura). Uma descrição desse conjunto de dados será exibida no seu navegador padrão (figura 2.7). Ao clicarmos no botão OK na figura 2.6, após termos selecionado *stroke*, esse conjunto de dados será carregado no *R commander* (figura 2.8).

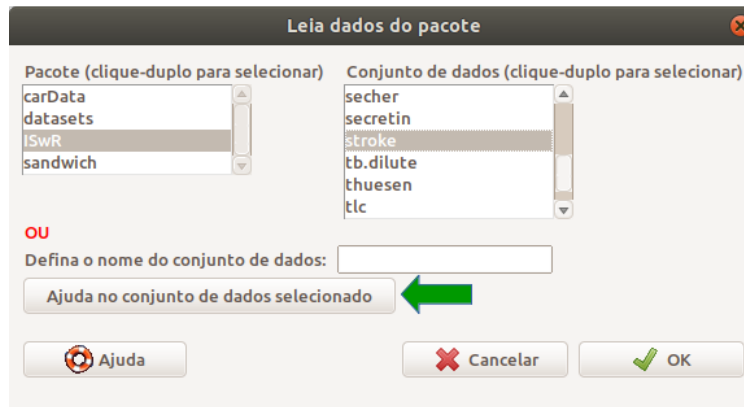


Figura 2.6: Visualizando a lista de conjuntos de dados do pacote *ISwR* e selecionando o conjunto *stroke*.

stroke {ISwR}	Estonian stroke data
Description	All cases of stroke in Tartu, Estonia, during the period 1991-1993, with follow-up until January 1, 1996.
Usage	
stroke	
Format	A data frame with 829 observations on the following 10 variables.
sex	a factor with levels Female and Male.
died	a Date, date of death.
dstr	a Date, date of stroke.
age	a numeric vector, age at stroke.
dgn	a factor, diagnosis, with levels ICH (intracranial haemorrhage), ID (unidentified), INF (infarction, ischaemic), SAH (subarchnoid haemorrhage).
coma	a factor with levels No and Yes, indicating whether patient was in coma after the stroke.
diab	a factor with levels No and Yes, history of diabetes.
minf	a factor with levels No and Yes, history of myocardial infarction.
han	a factor with levels No and Yes, history of hypertension.
obsmonths	a numeric vector, observation times in months (set to 0.1 for patients dying on the same day as the stroke).
dead	a logical vector, whether patient died during the study.
Source	
Original data.	
References	J. Kory, M. Roose, and A.E. Kaasik (1997). Stroke Registry of Tartu, Estonia, from 1991 through 1993. Cerebrovascular Disorders 7:154-162.

[Package *ISwR* version 2.0-8 [index](#)]

Figura 2.7: Texto com a descrição do conjunto de dados *stroke*.



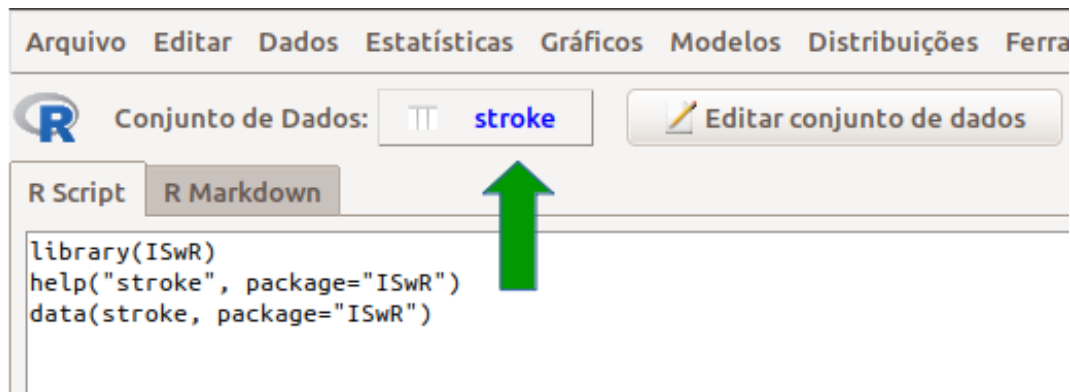


Figura 2.8: Tela do *R commander* após o carregamento do conjunto de dados *stroke*. Observem a função que foi executada – `data(stroke, package="ISwR")` – e o nome do conjunto selecionado (seta verde).

Observem as funções que foram executadas no *R Commander*:

```
library(ISwR)
help("stroke", package="ISwR")
data(stroke, package="ISwR")
```

A função `help` mostra uma ajuda sobre o conjunto de dados *stroke* do pacote *ISwR*.

A função `data(stroke, package="ISwR")` carrega o conjunto de dados *stroke* que passa a ser o conjunto de dados ativo no *R Commander*. Observem o nome dele ao lado do rótulo conjunto de dados (seta verde na figura 2.8). Esse objeto pode ser acessado pelo próprio nome (*stroke* nesse caso).

Na área de mensagens do *R Commander*, aparece a seguinte mensagem abaixo do comando, indicando o número de registros e de variáveis no conjunto de dados *stroke*:

NOTA: Os dados *stroke* tem 829 linhas e 11 colunas.

Para visualizarmos o conteúdo do conjunto de dados *stroke*, clicamos no botão *Ver conjunto de dados* (seta verde na figura 2.9).

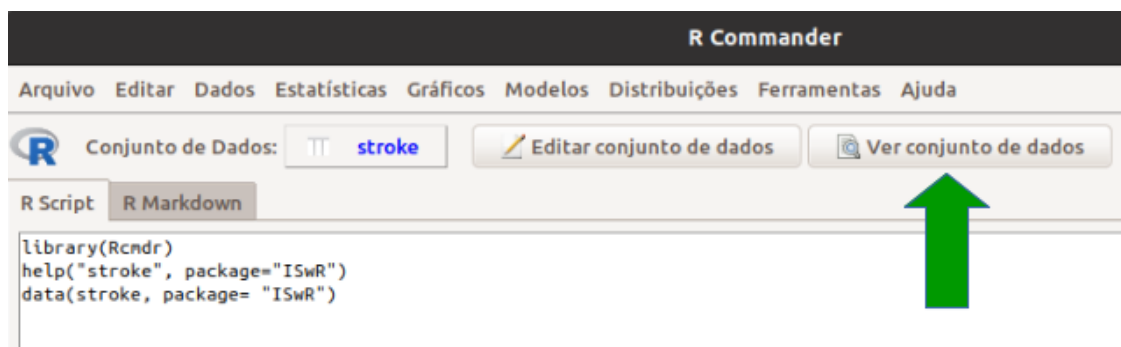


Figura 2.9: Botão do *R Commander* (seta verde) para exibir o conteúdo do conjunto de dados ativo.

A figura 2.10 mostra as observações do conjunto de dados *stroke*.

	sex	died	dstr	age	dgn	coma	diab	minf	han	dead	obsmonths
1	Male	1991-01-07	1991-01-02	76	INF	No	No	Yes	No	TRUE	0.16339869
2	Male	<NA>	1991-01-03	58	INF	No	No	No	No	FALSE	59.60784314
3	Male	1991-06-02	1991-01-08	74	INF	No	No	Yes	Yes	TRUE	4.73856209
4	Female	1991-01-13	1991-01-11	77	ICH	No	Yes	No	Yes	TRUE	0.06535948
5	Female	<NA>	1991-01-13	76	INF	No	Yes	No	Yes	FALSE	59.28104575
6	Male	1991-01-13	1991-01-13	48	ICH	Yes	No	No	Yes	TRUE	0.10000000
7	Female	1993-12-01	1991-01-14	81	INF	No	No	No	Yes	TRUE	34.37908497
8	Male	1991-12-12	1991-01-14	53	INF	No	No	Yes	Yes	TRUE	10.84967320
9	Female	<NA>	1991-01-15	73	ID	No	No	No	Yes	FALSE	59.21568627
10	Female	1993-11-10	1991-01-15	69	INF	No	No	No	Yes	TRUE	33.66013072
11	Female	1991-01-20	1991-01-16	86	ID	No	No	No	No	TRUE	0.13071895
12	Female	<NA>	1991-01-16	79	INF	No	No	No	Yes	FALSE	59.18300654
13	Male	1994-01-26	1991-01-21	69	INF	No	No	No	No	TRUE	35.98039216
14	Female	<NA>	1991-01-22	58	INF	No	No	No	Yes	FALSE	58.98692810
15	Female	<NA>	1991-01-23	71	INF	No	No	No	Yes	FALSE	58.95424837
16	Female	1991-02-04	1991-01-26	84	INF	No	No	No	Yes	TRUE	0.29411765
17	Male	1995-07-27	1991-01-27	63	INF	No	No	No	No	TRUE	53.66013072
18	Female	1991-02-11	1991-01-28	85	ID	No	No	No	No	TRUE	0.45751634
19	Female	1991-01-29	1991-01-28	81	ICH	No	Yes	Yes	Yes	TRUE	0.03267974
20	Female	1991-02-08	1991-01-29	77	INF	Yes	No	No	Yes	TRUE	0.32679739
21	Male	<NA>	1991-01-30	62	INF	No	No	No	Yes	FALSE	58.72549020
22	Female	1991-04-20	1991-01-31	84	INF	No	No	No	Yes	TRUE	2.58169935
23	Female	1991-07-04	1991-02-03	77	INF	No	No	No	Yes	TRUE	4.93464052
24	Male	1995-05-19	1991-02-03	61	ICH	No	No	No	Yes	TRUE	51.17647059
25	Female	<NA>	1991-02-04	79	INF	No	No	No	Yes	FALSE	58.56209150
26	Male	1994-06-08	1991-02-04	71	INF	No	No	Yes	Yes	TRUE	39.86928105
27	Female	1991-10-25	1991-02-06	84	ID	No	No	No	Yes	TRUE	8.52941176
28	Male	1991-02-14	1991-02-07	86	ID	No	No	No	No	TRUE	0.22875817
29	Male	<NA>	1991-02-10	62	INF	No	No	Yes	Yes	FALSE	58.36601307
30	Male	1991-02-23	1991-02-11	68	INF	No	No	Yes	No	TRUE	0.39215686

Figura 2.10: Conteúdo do conjunto de dados *stroke*.

Para mostrar como obter distribuições de frequências de variáveis categóricas e gerar tabulações de dados com combinações de variáveis categóricas, vamos trabalhar com as seguintes variáveis do conjunto de dados *stroke* (figura 2.7):

- *dead*: variável categórica binária, com os valores *TRUE*, se o paciente faleceu, e *FALSE*, se o paciente continuava vivo ao final do estudo;
- *dgn*: diagnóstico do paciente, variável categórica nominal, com as categorias ICH (hemorragia intracranial), ID (não identificado), INF (infarto), SAH (hemorragia subaracnóide);
- *minf*: história de infarto do miocárdio, variável categórica binária, com os valores *No* e *Yes*;
- *diab*: história de diabetes, variável categórica binária, com os valores *No* e *Yes*.

## 2.3 Tabelas de frequências no conjunto de dados *stroke*

### 2.3.1 Uma única variável categórica

Para construirmos uma tabela que mostra a frequência ou porcentagem de cada categoria de uma variável categórica no *R Commander*, usamos a opção:

Estatísticas ⇒ Resumos ⇒ Distribuições de frequência...

Na tela dessa opção (figura 2.11), selecionamos uma ou mais variáveis para as quais desejamos a distribuição de frequências. Serão montadas tabelas para cada variável separadamente. Nesse exemplo, selecionamos a variável *dead*.

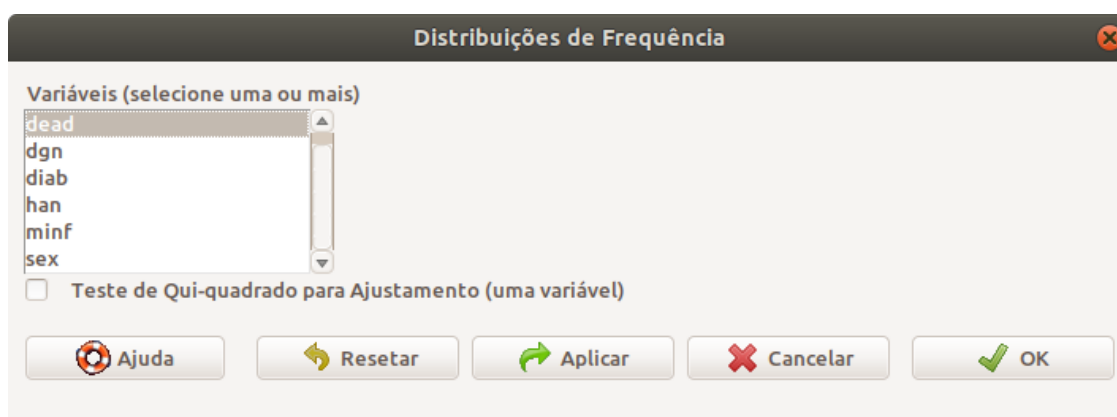


Figura 2.11: Caixa de diálogo para a seleção das variáveis categóricas cujas distribuições marginais serão exibidas.

Ao clicarmos em OK, os comandos abaixo serão executados, com os resultados mostrados logo a seguir.

```
local({
  .Table <- with(stroke, table(dead))
  cat("\ncounts:\n")
  print(.Table)
  cat("\npercentages:\n")
  print(round(100*.Table/sum(.Table), 2))
})
```

```
##
## counts:
## dead
## FALSE  TRUE
##   344   485
##
## percentages:
```

```
## dead
## FALSE TRUE
## 41.5 58.5
```

São mostradas duas tabelas, uma com a frequência de cada categoria da variável *dead* (TRUE, FALSE) no conjunto de dados *stroke*, e outra com as porcentagens de cada categoria.

No primeiro comando da sequência acima, mostrado novamente abaixo, a função *with* possui dois argumentos: o primeiro indica o conjunto de dados que será utilizado (*stroke*), e o segundo argumento indica a função que será executada com variáveis do conjunto de dados especificado no primeiro argumento. Nesse caso, é executada a função *table* para gerar uma tabela de frequência da variável entre parênteses.

```
.Table <- with(stroke, table(dead))
```

O resultado da execução do comando é armazenado no objeto *.Table*.

O comando seguinte *cat* imprime uma linha na tela para informar que a tabela mostrada a seguir é relativa à contagem (*counts:*) ou frequência das categorias da variável. A barra invertida (“\”), seguida da letra “n” antes e depois da expressão “counts:”, indica que uma linha deve ser pulada antes e depois de escrever a expressão “counts:”.

```
cat("\ncounts:\n")
```

O comando seguinte, *print*, mostra então a tabela de frequências da variável *dead*.

```
print(.Table)
```

Vemos que 485 pacientes morreram e 344 continuavam vivos até o final do estudo. Esses números são obtidos, contando-se o número de observações no conjunto de dados *stroke* cujos valores da variável *dead* são *TRUE* ou *FALSE*, respectivamente.

Os dois comandos seguintes imprimem a expressão “percentages:” na tela e, em seguida, a tabela com as porcentagens de cada categoria da variável *dead*.

```
cat("\npercentages:\n")
print(round(100*.Table/sum(.Table), 2))
```

Vamos entender como as porcentagens foram calculadas. A função *sum* aplicada ao objeto *.Table* irá somar as frequências das categorias da variável *dead* ( $344 + 485 = 829$ ). A expressão *.Table/sum(.Table)* então divide a frequência de cada categoria da variável *dead* pela soma das frequências, resultando nas proporções de cada categoria ( $344/829=0,41496$ ;  $485/829=0,58504$ ). Esses dois valores são então multiplicados por 100 para fornecer as porcentagens de cada categoria (41,496; 58,504). A função *round* irá arredondar esses valores com duas casas decimais.

## 2.3.2 Tabelas de frequências para duas variáveis categóricas

Os conteúdos desta seção e da seção 2.3.3 podem ser visualizados neste [vídeo](#).

Para gerarmos tabelas de frequências ou porcentagens para combinação das categorias de duas variáveis categóricas no *R Commander*, usamos a opção:

Estatísticas  $\Rightarrow$  Tabelas de contingência  $\Rightarrow$  Tabela de dupla entrada...

Esse tipo de tabela é chamada também de **tabela de dupla entrada** (por envolver duas variáveis).

Na aba *Dados* da tela dessa opção (figura 2.12), selecionamos as duas variáveis (uma cujas categorias irão aparecer nas linhas e outra cujas categorias irão aparecer nas colunas da tabela), para as quais desejamos a distribuição conjunta de frequências. Nesse exemplo, selecionamos a variável *dgn* (diagnóstico) para as linhas e *dead* para as colunas.



Figura 2.12: Tela para a seleção das variáveis categóricas que comporão a tabela de dupla entrada.

### 2.3.2.1 Obtendo os percentuais de cada célula em relação ao total da linha correspondente

Na aba *Estatísticas* (figura 2.13), vamos marcar a opção *percentual nas linhas* e desmarcar a opção *Teste de independência de Qui-Quadrado*. Nesse caso, a tabela de dupla entrada irá mostrar os percentuais de cada célula da tabela em relação ao total da linha correspondente.

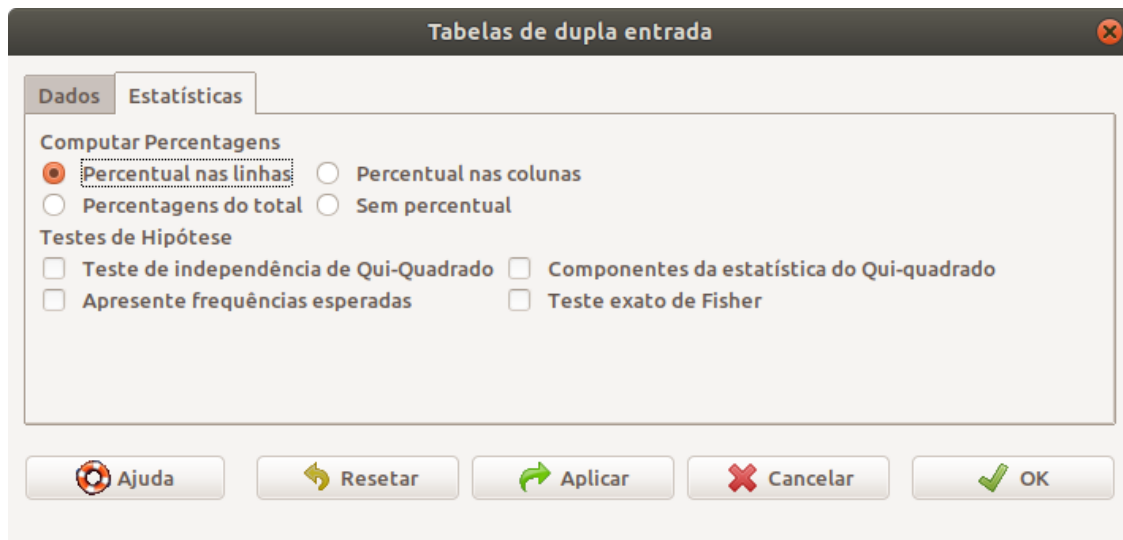


Figura 2.13: Tela para especificar que a tabela de dupla entrada irá mostrar os percentuais de cada célula em relação ao total da linha correspondente.

Ao clicarmos em OK, os comandos abaixo serão executados, com os resultados mostrados logo a seguir.

```
local({
  .Table <- xtabs(~dgn+dead, data=stroke)
  cat("\nFrequency table:\n")
  print(.Table)
  cat("\nRow percentages:\n")
  print(rowPercents(.Table))
})
```

```
##
## Frequency table:
##      dead
## dgn  FALSE TRUE
##  ICH    25   54
##  ID     54  148
##  INF   239  262
##  SAH    26   21
##
## Row percentages:
##      dead
## dgn  FALSE TRUE Total Count
##  ICH  31.6 68.4   100     79
##  ID   26.7 73.3   100    202
##  INF  47.7 52.3   100    501
##  SAH  55.3 44.7   100     47
```

Dessa vez, a tabela foi gerada por meio da função *xtabs*. Essa função utiliza uma fórmula para especificar as variáveis que comporão a tabela. Nessa fórmula, as variáveis são especificadas após o sinal  $\sim$ , sendo a primeira variável aquela cujas categorias aparecerão nas linhas e a segunda variável aquela cujas categorias aparecerão nas colunas. As variáveis são separadas pelo sinal “+”. O argumento *data* especifica o conjunto de dados que contém as variáveis descritas na fórmula.

A função *rowPercents*, do pacote *RcmdrMisc*, gera os percentuais de cada célula em relação ao total da linha correspondente.

Assim a tabela de frequência mostra que, dos 79 pacientes cujo diagnóstico era “ICH”, 54 morreram e 25 não morreram no período de tempo considerado. Esses valores são obtidos, contando-se no conjunto de dados quantos pacientes que possuem o diagnóstico de hemorragia intracranial vieram ou não a óbito, respectivamente. A tabela com as porcentagens nas linhas mostra que, dos 79 pacientes com diagnóstico “ICH”, 31,6% (25) não morreram e 68,4% (54) morreram. Interpretação análoga se aplica às demais linhas da tabela.

### 2.3.2.2 Obtendo os percentuais de cada célula em relação ao total da coluna correspondente

Se, na aba *Estatísticas* para gerar uma tabela de dupla entrada, marcarmos a opção *percentual nas colunas* (figura 2.14), a tabela de dupla entrada irá mostrar os percentuais de cada célula em relação ao total da coluna correspondente.

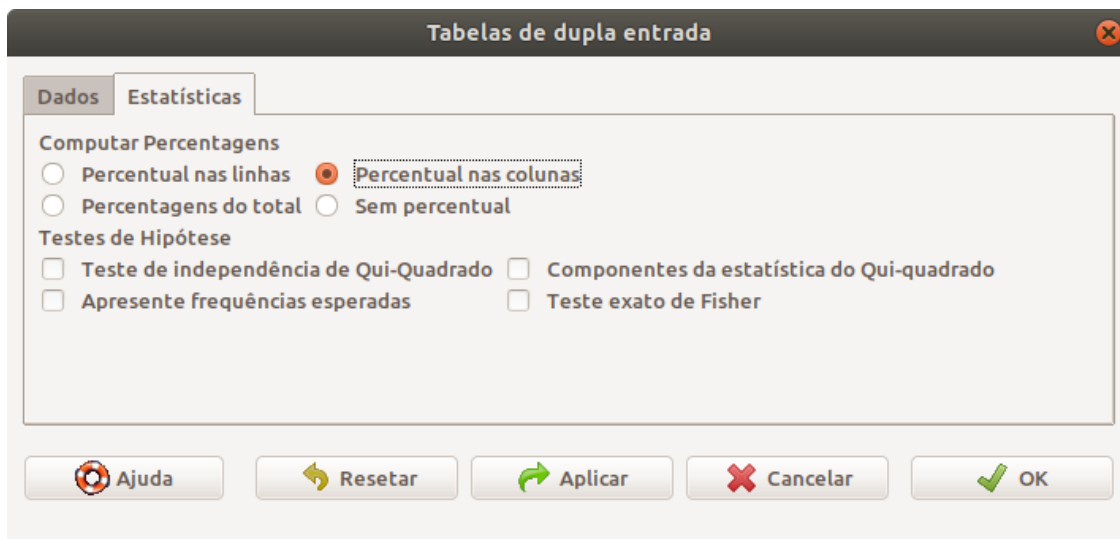


Figura 2.14: Tela para especificar que a tabela de dupla entrada irá mostrar os percentuais de cada célula em relação ao total da coluna correspondente à célula.

Ao clicarmos em OK, os comandos a seguir serão executados, com os resultados mostrados logo após.

```

local({
  .Table <- xtabs(~dgn+dead, data=stroke)
  cat("\nFrequency table:\n")
  print(.Table)
  cat("\nColumn percentages:\n")
  print(colPercents(.Table))
})

```

```

##
## Frequency table:
##      dead
## dgn  FALSE TRUE
##  ICH    25   54
##  ID     54  148
##  INF    239  262
##  SAH     26   21
##
## Column percentages:
##      dead
## dgn  FALSE TRUE
##  ICH    7.3  11.1
##  ID    15.7  30.5
##  INF    69.5  54.0
##  SAH     7.6   4.3
##  Total 100.1  99.9
##  Count 344.0 485.0

```

A função *colPercents*, do pacote *RcmdrMisc*, gera os percentuais de cada célula em relação ao total da coluna correspondente.

Assim a tabela de frequência mostra novamente que, dos 79 pacientes cujo diagnóstico era “ICH”, 54 morreram e 25 não morreram no período de tempo considerado. A tabela com as porcentagens nas colunas mostra que, dos 344 pacientes que não morreram, 7,3% (25) tiveram o diagnóstico “ICH”, e dos 485 que morreram, 11,1% (54) tiveram o diagnóstico “ICH”. Interpretação análoga se aplica às demais linhas da tabela.

### 2.3.2.3 Obtendo os percentuais de cada célula em relação ao total da tabela

Se, na aba *Estatísticas* para gerar uma tabela de dupla entrada, marcarmos a opção *percentagens do total* (figura 2.15), a tabela de dupla entrada irá mostrar os percentuais de cada célula em relação ao total da tabela.



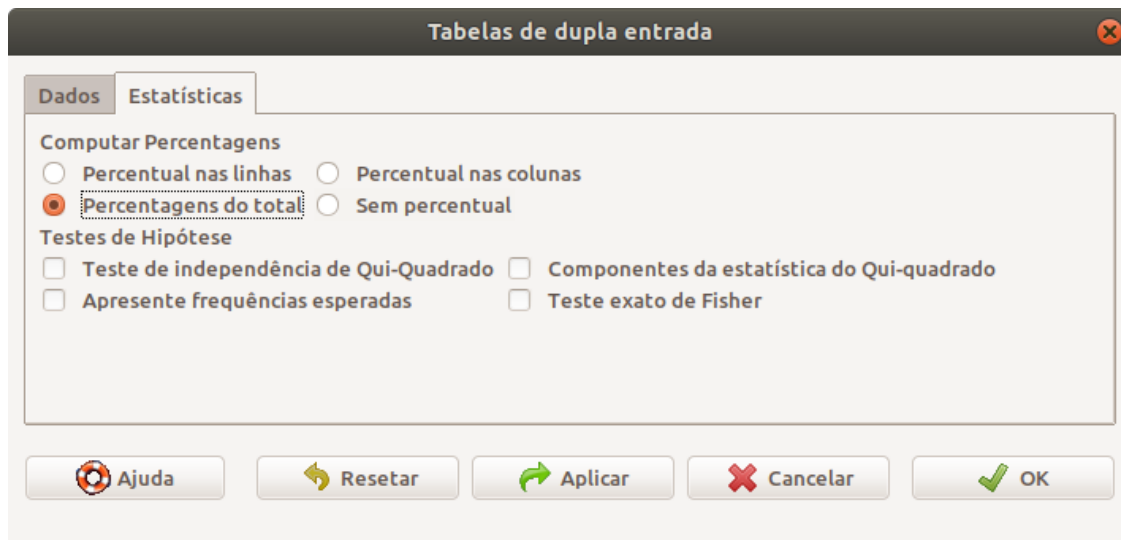


Figura 2.15: Tela para especificar que a tabela de dupla entrada irá mostrar os percentuais de cada célula em relação ao total da tabela.

Ao clicarmos em OK, os comandos abaixo serão executados, com os resultados mostrados logo a seguir.

```
local({
  .Table <- xtabs(~dgn+dead, data=stroke)
  cat("\nFrequency table:\n")
  print(.Table)
  cat("\nTotal percentages:\n")
  print(totPercents(.Table))
})
```

```
##
## Frequency table:
##      dead
## dgn  FALSE TRUE
##  ICH    25   54
##   ID    54  148
##  INF   239  262
##  SAH    26   21
##
## Total percentages:
##      FALSE TRUE Total
##  ICH    3.0  6.5   9.5
##   ID    6.5 17.9  24.4
##  INF   28.8 31.6  60.4
##  SAH    3.1  2.5   5.7
## Total  41.5 58.5 100.0
```

A função *totPercents*, do pacote *RcmdrMisc*, gera os percentuais de cada célula em relação ao total da tabela.

Assim a tabela de frequência mostra que 54 pacientes cujo diagnóstico era “ICH” morreram. A tabela com as porcentagens de cada célula mostra que, do total de 829 pacientes, 6,5% (54) tiveram o diagnóstico “ICH” e morreram. Interpretação análoga se aplica às demais células da tabela.

### 2.3.3 Tabelas de frequência para mais de duas variáveis categóricas

Para gerar tabelas de frequências ou porcentagens para combinação das categorias de três ou mais variáveis categóricas no *R Commander*, usamos a opção:

Estatísticas ⇒ Tabelas de contingência ⇒ Tabela multientrada...

Esse tipo de tabela é chamada também de **tabela de múltiplas entradas** (por envolver mais de duas variáveis).

Na tela dessa opção (figura 2.16), selecionamos uma variável cujas categorias irão aparecer nas linhas da tabela, outra variável cujas categorias irão aparecer nas colunas da tabela e outra(s) variável(is) (chamadas de variáveis de controle no *R Commander*) para as quais desejamos a distribuição conjunta de frequências. Nesse exemplo, selecionamos a variável *dgn* (diagnóstico) para as linhas, *dead* para as colunas e *minf* (história de infarto do miocárdio) para a variável de controle. Também selecionamos a opção de mostrar os percentuais de cada célula na linha correspondente da tabela.

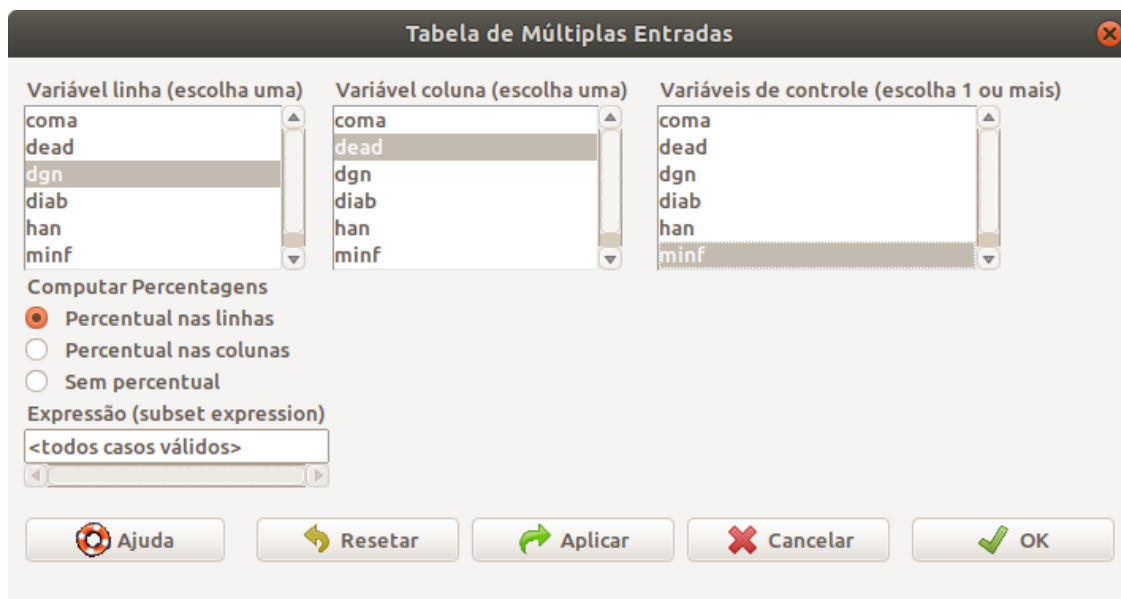


Figura 2.16: Tela para a seleção das variáveis categóricas que comporão a tabela multientrada com três variáveis.

Ao clicarmos em OK, a sequência de comandos abaixo será executada e os resultados mostrados a seguir.

```
local({
  .Table <- xtabs(~dgn+dead+minf, data=stroke)
  cat("\nFrequency table:\n")
  print(.Table)
  cat("\nRow percentages:\n")
  print(rowPercents(.Table))
})
```

```
##
## Frequency table:
## , , minf = No
##
##      dead
## dgn  FALSE TRUE
##  ICH    25   46
##  ID     49  131
##  INF    222  207
##  SAH     24   21
##
## , , minf = Yes
##
##      dead
## dgn  FALSE TRUE
##  ICH     0    8
##  ID      4   12
##  INF     17   54
##  SAH      2    0
##
##
## Row percentages:
## , , minf = No
##
##      dead
## dgn  FALSE TRUE Total Count
##  ICH  35.2 64.8   100     71
##  ID   27.2 72.8   100    180
##  INF  51.7 48.3   100    429
##  SAH  53.3 46.7   100     45
##
## , , minf = Yes
##
##      dead
```

##	dgn	FALSE	TRUE	Total	Count
##	ICH	0.0	100.0	100	8
##	ID	25.0	75.0	100	16
##	INF	23.9	76.1	100	71
##	SAH	100.0	0.0	100	2

Uma tabela de frequências de dupla entrada (com as variáveis *dgn* e *dead*) é construída para cada categoria da variável *minf* e as porcentagens nas linhas são calculadas para cada uma das tabelas separadamente. Na função *xtabs*, a primeira variável após o “~” é aquela cujas categorias aparecerão nas linhas de cada tabela, a segunda variável após o “~” é aquela cujas categorias aparecerão nas colunas de cada tabela. Uma tabela de frequências será construída com as duas primeiras variáveis para cada categoria da terceira variável.

Se desejássemos os percentuais nas colunas, bastaria selecionar a opção correspondente na figura 2.16.

Podemos selecionar mais de três variáveis para montar uma tabela de múltiplas entradas. Na tela mostrada na figura 2.17, selecionamos quatro variáveis: *dgn* para as linhas, *dead* para as colunas, *diab* (história de diabetes) e *minf* (história de infarto do miocárdio) para as variáveis de controle. Dessa vez, foi selecionada a opção de mostrar os percentuais de cada célula na coluna correspondente da tabela. Uma tabela será construída para cada combinação das categoriais das variáveis *diab* e *minf*. Nesse caso, serão mostradas quatro tabelas.

Figura 2.17: Tela para a seleção das variáveis categóricas que compoñão a tabela multientrada com quatro variáveis.

Ao clicarmos em OK, a sequência de comandos abaixo será executada e os resultados mostrados em seguida.

```
local({
  .Table <- xtabs(~dgn+dead+diab+minf, data=stroke)
  cat("\nFrequency table:\n")
  print(.Table)
  cat("\nColumn percentages:\n")
  print(colPercents(.Table))
})
```

```
##
## Frequency table:
## , , diab = No, minf = No
##
##      dead
## dgn  FALSE TRUE
##  ICH    23   43
##  ID     39  114
##  INF    203  171
##  SAH     24   21
##
## , , diab = Yes, minf = No
##
##      dead
## dgn  FALSE TRUE
##  ICH     2    3
##  ID     10   14
##  INF     19   35
##  SAH      0    0
##
## , , diab = No, minf = Yes
##
##      dead
## dgn  FALSE TRUE
##  ICH     0    6
##  ID      4   12
##  INF     13   46
##  SAH      2    0
##
## , , diab = Yes, minf = Yes
##
##      dead
## dgn  FALSE TRUE
##  ICH     0    2
```

```

##      ID      0      0
##      INF      4      8
##      SAH      0      0
##
##
## Column percentages:
## , , diab = No, minf = No
##
##      dead
## dgn      FALSE  TRUE
##      ICH      8.0 12.3
##      ID       13.5 32.7
##      INF      70.2 49.0
##      SAH      8.3  6.0
##      Total 100.0 100.0
##      Count 289.0 349.0
##
## , , diab = Yes, minf = No
##
##      dead
## dgn      FALSE  TRUE
##      ICH      6.5  5.8
##      ID       32.3 26.9
##      INF      61.3 67.3
##      SAH      0.0  0.0
##      Total 100.1 100.0
##      Count  31.0  52.0
##
## , , diab = No, minf = Yes
##
##      dead
## dgn      FALSE  TRUE
##      ICH      0.0  9.4
##      ID       21.1 18.8
##      INF      68.4 71.9
##      SAH      10.5  0.0
##      Total 100.0 100.1
##      Count  19.0  64.0
##
## , , diab = Yes, minf = Yes
##
##      dead
## dgn      FALSE  TRUE
##      ICH      0    20
##      ID       0     0

```

```
##    INF      100   80
##    SAH        0    0
##    Total    100  100
##    Count     4   10
```

### 2.3.3.1 Forma alternativa de apresentar tabelas de frequência para mais de duas variáveis categóricas

Vimos na seção anterior que a função *xtabs* (assim como a função *table*) irá apresentar, para mais de duas variáveis, tantas tabelas quantas forem as combinações possíveis das categorias das variáveis de controle (variáveis após a segunda na fórmula ou na lista de variáveis).

Existe uma outra forma de apresentar os resultados da distribuição de frequências conjuntas, por meio da função *ftable*.

Se quisermos obter a distribuição de frequências conjuntas das variáveis *dgn*, *dead* e *minf*, com *dgn* nas linhas, usamos o comando a seguir.

```
ftable(minf + dead ~ dgn, data=stroke)
```

```
##      minf      No      Yes
##      dead FALSE TRUE FALSE TRUE
## dgn
## ICH      25   46     0    8
## ID       49  131     4   12
## INF     222  207    17   54
## SAH      24   21     2    0
```

As variáveis após o “~” serão mostradas nas linhas e aquelas antes do “~” serão mostradas nas colunas. Nesse exemplo, uma tabela relacionando *dgn* com *dead* será mostrada para a categoria *No* de *minf* ao lado de outra tabela relacionando *dgn* com *dead* para a categoria *Yes* de *minf*.

Se colocarmos *dead* antes de *minf* na fórmula, como no comando a seguir, uma tabela relacionando *dgn* com *minf* será mostrada para a categoria *FALSE* de *dead* ao lado de outra tabela relacionando *dgn* com *minf* para a categoria *TRUE* de *dead*.

```
ftable(dead + minf ~ dgn, data=stroke)
```

```
##      dead FALSE      TRUE
##      minf      No Yes      No Yes
## dgn
## ICH      25   0   46   8
## ID       49   4  131  12
## INF     222  17  207  54
## SAH      24   2   21   0
```

Para obtermos as proporções de cada célula em relação ao total da linha correspondente, podemos utilizar a função *prop.table*, especificando como primeiro argumento o objeto correspondente à tabela que estamos trabalhando (gerado pela função *ftable*) e, como segundo argumento, o valor 1, que indica a primeira dimensão da tabela.

```
prop.table(ftable(dead + minf ~ dgn, data=stroke), 1)
```

##	dead	FALSE		TRUE	
##	minf	No	Yes	No	Yes
##	dgn				
##	ICH	0.31645570	0.00000000	0.58227848	0.10126582
##	ID	0.25000000	0.02040816	0.66836735	0.06122449
##	INF	0.44400000	0.03400000	0.41400000	0.10800000
##	SAH	0.51063830	0.04255319	0.44680851	0.00000000

Para expressar as proporções como porcentagens, basta multiplicar os resultados por 100 como mostrado a seguir.

```
prop.table(ftable(dead + minf ~ dgn, data=stroke), 1) * 100
```

##	dead	FALSE		TRUE	
##	minf	No	Yes	No	Yes
##	dgn				
##	ICH	31.645570	0.000000	58.227848	10.126582
##	ID	25.000000	2.040816	66.836735	6.122449
##	INF	44.400000	3.400000	41.400000	10.800000
##	SAH	51.063830	4.255319	44.680851	0.000000

Para obtermos as proporções de cada célula em relação ao total da coluna correspondente, utilizamos a função *prop.table*, especificando como segundo argumento o valor 2, que indica a segunda dimensão da tabela.

```
prop.table(ftable(dead + minf ~ dgn, data=stroke), 2)
```

##	dead	FALSE		TRUE	
##	minf	No	Yes	No	Yes
##	dgn				
##	ICH	0.07812500	0.00000000	0.11358025	0.10810811
##	ID	0.15312500	0.17391304	0.32345679	0.16216216
##	INF	0.69375000	0.73913043	0.51111111	0.72972973
##	SAH	0.07500000	0.08695652	0.05185185	0.00000000



O comando abaixo mostra a função *fTable* com quatro variáveis, sendo as categorias de *dgn* mostradas nas linhas.

```
fTable(diab + minf + dead ~ dgn, data=stroke)
```

```
##      diab      No      Yes
##      minf      No      Yes      No      Yes
##      dead FALSE TRUE FALSE TRUE FALSE TRUE FALSE TRUE
## dgn
## ICH      23    43      0    6      2    3      0    2
## ID       39   114      4   12     10   14      0    0
## INF      203   171     13   46     19   35      4    8
## SAH       24    21      2    0      0    0      0    0
```

O comando a seguir mostra a função *fTable* com quatro variáveis, sendo as categorias de *dgn* e *diab* mostradas nas linhas. Observem a flexibilidade nos arranjos das tabelas, de acordo com a posição das variáveis (antes e depois do “~”) e da ordem em que elas são colocadas.

```
fTable(minf + dead ~ dgn + diab, data=stroke)
```

```
##      minf      No      Yes
##      dead FALSE TRUE FALSE TRUE
## dgn diab
## ICH No      23    43      0    6
##      Yes      2    3      0    2
## ID  No      39   114      4   12
##      Yes     10   14      0    0
## INF No     203   171     13   46
##      Yes     19   35      4    8
## SAH No     24    21      2    0
##      Yes      0    0      0    0
```

### 2.3.4 Entrando diretamente com as frequências das células

Às vezes, temos disponível uma tabela de frequências já construída, por exemplo, publicada em artigos científicos ou relatórios técnicos e desejamos obter as proporções ou percentuais nas linhas ou colunas e realizar alguma análise estatística.

A opção a seguir nos permite digitar as frequências de cada célula de uma tabela de dupla entrada no *R Commander* e obter os percentuais nas linhas, colunas ou células, além de realizar um teste estatístico de associação entre as duas variáveis:

Estatísticas ⇒ Tabelas de contingência ⇒ Digite e analise tabela dupla entrada...

Na aba *Tabela* dessa opção (figura 2.18), podemos escrever os nomes para as variáveis cujas categorias irão aparecer nas linhas e colunas da tabela, selecionar o número de categorias nas linhas e colunas e digitar os valores das células da tabela de dupla entrada. Nesse exemplo, reproduzimos a tabela de frequências conjuntas das variáveis *dgn* (diagnóstico) para as linhas e *dead* para as colunas (seção 2.3.2). Na aba *Estatísticas* (não mostrada aqui), selecionamos a opção *Percentuais nas linhas* e desmarcamos a opção *Teste de independência de Qui-Quadrado*, porque esse teste será visto com detalhes no capítulo 17.

**Digite tabela de dupla-entrada (two-way)**

**Tabela** | Estatísticas

Name for Row Variable (optional):

Name for Column Variable (optional):

Número de linhas:  4

Número de Colunas:  2

Entrar número:

	1	2
1	25	54
2	54	148
3	239	262
4	26	21

Ajuda Resetar Aplicar Cancelar OK

Figura 2.18: Tela para a digitação das frequências das células que comporão uma tabela de dupla entrada.

Ao clicarmos em OK, a sequência de comandos a seguir será executada e os resultados mostrados logo após.

```
.Table <- matrix(c(25,54,54,148,239,262,26,21), 4, 2, byrow=TRUE)
dimnames(.Table) <- list("dgn"=c("1", "2", "3", "4"), "dead"=c("1", "2"))
.Table # Counts
```

```
##      dead
## dgn   1   2
##   1  25  54
##   2  54 148
##   3 239 262
##   4  26  21
```

```
rowPercents(.Table) # Row Percentages
```

```
##      dead
## dgn    1    2 Total Count
##   1 31.6 68.4   100    79
##   2 26.7 73.3   100   202
##   3 47.7 52.3   100   501
##   4 55.3 44.7   100    47
```

Observamos que os resultados são idênticos aos mostrados na seção 2.3.2, porém os nomes das categorias nas linhas e colunas aparecem como números.

A tabela foi construída por meio da função *matrix*. Podemos tornar os nomes das categorias mais descritivos por meio das funções *colnames* e *rownames* aplicadas ao objeto que representa a tabela gerada (*.Table*), como mostrado a seguir.

```
colnames(.Table) <- c("Não", "Sim")
rownames(.Table) <- c("ICH", "ID", "INF", "SAH")
.Table
```

```
##      dead
## dgn  Não Sim
##   ICH  25  54
##   ID   54 148
##   INF 239 262
##   SAH  26  21
```

Podemos alterar os nomes das dimensões da tabela para valores mais descritivos, usando a função *names*, como mostrado a seguir.

```
names(dimnames(.Table)) <- c("diagnóstico", "morte")
.Table
```

```
##      morte
## diagnóstico Não Sim
##           ICH  25  54
##           ID   54 148
##           INF 239 262
##           SAH  26  21
```

Poderíamos combinar os comandos anteriores (*colnames*, *rownames* e *names*) em um único comando, como mostrado a seguir.

```
dimnames(.Table) <- list("diagnóstico"=c("ICH","ID","INF", "SAH"),
                          "morte"=c("Não", "Sim"))
```

```
.Table
```

```
##           morte
## diagnóstico Não Sim
##           ICH  25  54
##           ID   54 148
##           INF 239 262
##           SAH  26  21
```

Se quisermos trocar as linhas pelas colunas na tabela de frequência, podemos usar a função `t` (de transposta), como mostra o comando a seguir.

```
t(.Table)
```

```
##      diagnóstico
## morte ICH  ID INF SAH
##   Não  25  54 239  26
##   Sim  54 148 262  21
```

## 2.4 Exercício

- 1) Carregue o conjunto de dados *births14* do pacote *openintro* (GPL-3). Para abrir esse conjunto de dados, é necessário que o pacote *openintro* esteja instalado. Responda às questões abaixo.
  - a) Abra a ajuda desse conjunto de dados e verifique o significado das suas variáveis.
  - b) Monte tabelas de frequências e de porcentagens para as variáveis *lowbirthweight*, *whitemom* e *habit*.
  - c) Monte uma tabela de contingência relacionando as variáveis *habit* e *lowbirthweight*. Quais são as porcentagens de baixo peso ao nascer para fumantes e não fumantes?
  - d) Monte uma tabela de contingência relacionando as variáveis *habit* e *lowbirthweight* para cada nível da variável *whitemom* com porcentagens obtidas a partir das linhas. Comente os resultados.

# Capítulo 3

## Medidas de tendência central e dispersão

### 3.1 Introdução

Os conteúdos desta seção e das suas subseções podem ser visualizados neste [vídeo](#).

Abaixo são apresentados os valores das variáveis idade gestacional, crib (*Clinical Risk Index for Babies*), fração inspirada de O<sub>2</sub> máxima (*fiO2\_maximo*), peso ao nascimento e óbito de 10 recém-nascidos tratados em uma UTI neonatal. Esse arquivo de dados, chamado *neonato*, contém dados de mais de 300 recém-nascidos. Há uma série de outras variáveis que foram coletadas para esses recém-nascidos, mas não foram mostradas. Os dados apresentados de uma forma tabular não nos fornece uma boa ideia de como estão distribuídos e suas principais características.

##	ig_parto	crib	fio2_maximo	peso_nascimento	obito
## 1	232	2	NA	1185	No
## 3	235	NA	NA	1500	No
## 4	200	1	NA	1110	No
## 5	218	NA	21	1315	No
## 6	248	NA	21	1390	No
## 7	205	10	100	980	No
## 9	219	NA	30	1060	No
## 10	221	1	60	1095	Yes
## 11	186	16	100	500	Yes
## 12	211	1	60	1255	Yes

O primeiro passo para realizar uma análise de dados envolve a inspeção das variáveis individualmente. Essa exploração dos dados pode ser realizada por meio de tabelas de contingência para variáveis categóricas (capítulo 2), ou por meio de medidas numéricas para variáveis numéricas, além de recursos gráficos para ambos os tipos de variáveis. Os principais recursos gráficos serão apresentados no capítulo 4. Para as variáveis numéricas, as medidas

numéricas mais utilizadas são as de localização (tendência central) e de dispersão.

Um gráfico bastante utilizado para mostrar a distribuição dos valores de uma variável numérica é o histograma que, de uma maneira resumida, mostra a frequência ou percentual dos valores em cada faixa de valores da variável numérica. A figura 3.1 mostra o histograma de frequência da variável idade gestacional do conjunto de dados *neonato* (gráfico superior) e da variável *crib* (gráfico inferior). As retas perpendiculares em ambos os gráficos indicam a média (vermelha) e a mediana (azul), respectivamente. As médias e medianas serão discutidas mais adiante. Especialmente o segundo histograma indica uma assimetria à direita no gráfico. Os valores das médias e medianas não coincidem e são razoavelmente distintos no histograma do *crib*.

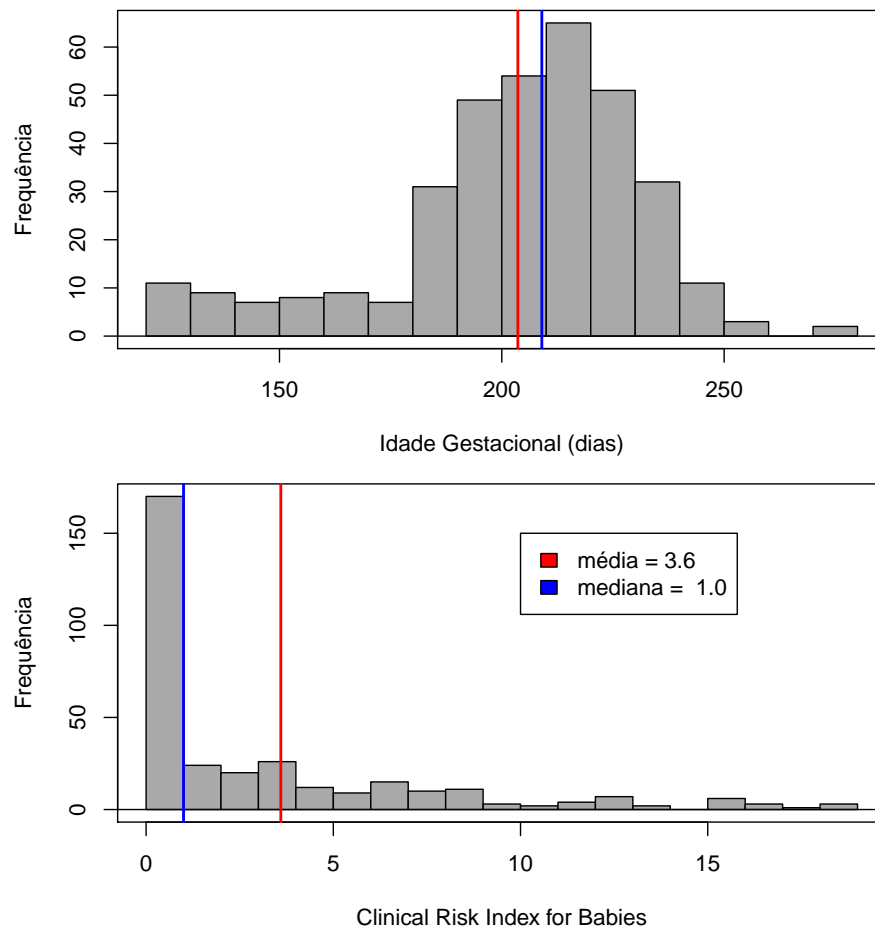


Figura 3.1: Exemplos de duas variáveis com distribuição assimétrica.

A figura 3.2 mostra o histograma da variável fração inspirada de  $O_2$  máxima. Nesse caso, os valores parecem se concentrar em dois locais diferentes.

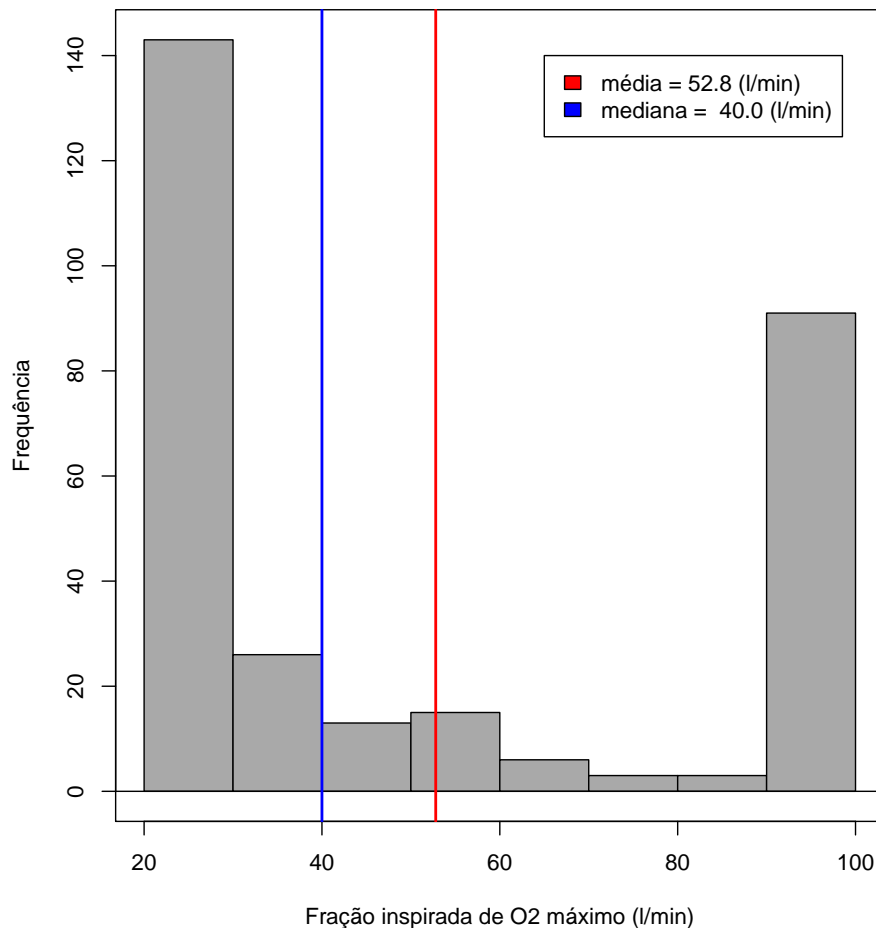


Figura 3.2: Exemplo de uma variável não distribuída em torno de um ponto central.

A figura 3.3 mostra a distribuição do peso ao nascimento para crianças que sobreviveram e as que foram a óbito. Pode-se observar, que a dispersão das duas distribuições são parecidas, mas o peso das crianças que sobrevivem tendem a ser maiores do que o peso das que vão a óbito. Ambas as distribuições são assimétricas, sendo uma para a direita e outra para a esquerda.

As medidas numéricas mais comuns utilizadas para resumir os dados são as medidas de tendência central, dispersão, simetria e curtose. Neste capítulo, serão apresentadas as principais medidas de tendência central e dispersão, que são as mais frequentes em relatórios e artigos científicos.

O primeiro conjunto de medidas numéricas caracteriza a tendência central ou ‘centro de massa’ dos dados ou distribuição. O segundo conjunto procura refletir o espalhamento dos dados em torno do centro; são as medidas de dispersão.

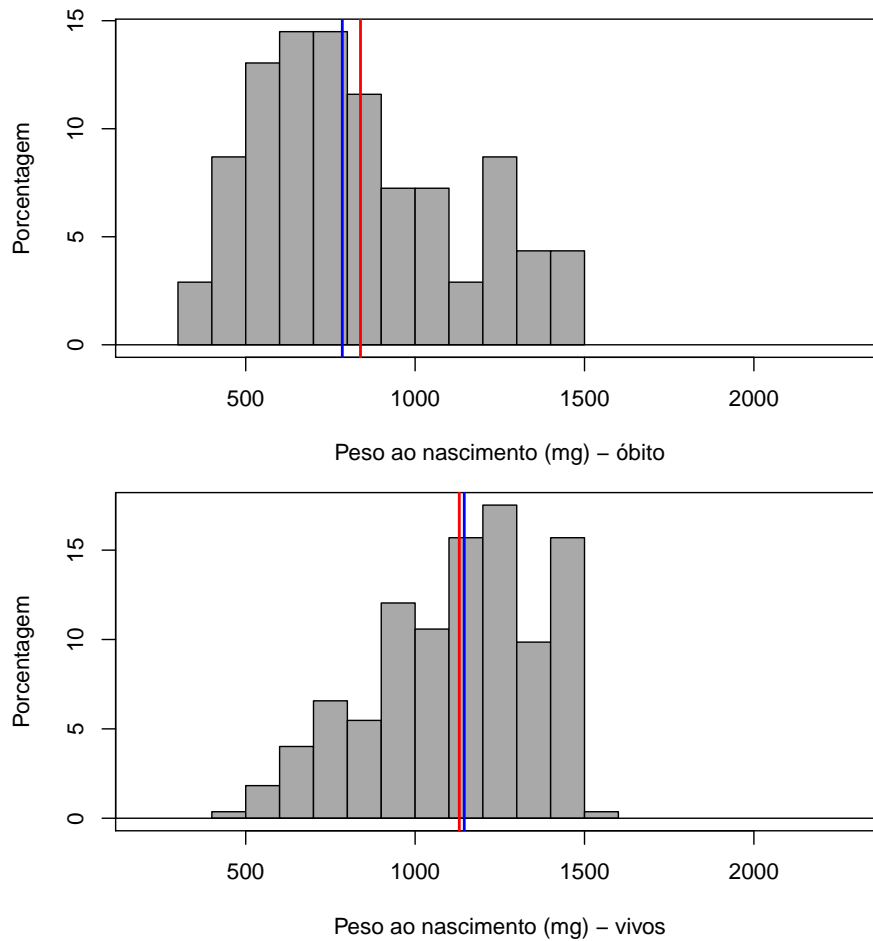


Figura 3.3: Variáveis com distribuições com medidas de tendência central diferentes.

## 3.2 Medidas de tendência central

As medidas de tendência central são aquelas que buscam refletir o ponto de equilíbrio dos dados. Não há uma única medida de tendência central, e a razão para isso será mostrada com exemplos.

Os conteúdos das subseções desta seção podem ser visualizados neste [vídeo](#).

### 3.2.1 Média

A **média**, ou **média aritmética** (*mean* em inglês), é a mais conhecida, sendo de fácil obtenção. Assumindo que a variável  $x$  possua  $n$  valores  $x_i$ ,  $i = 1, 2, \dots, n$ , a média aritmética é calculada por meio da expressão:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$



Vamos considerar a variável  $x$  que contém 15 valores, criada por meio do comando abaixo. O sinal “<-” ou “=” significa que o resultado da operação à direita do sinal é atribuído ao objeto à esquerda do sinal. A função `c` cria um vetor com os elementos dentro dos parênteses, separados por vírgula.

```
x <- c(10, 9, 8, 12, 11, 7, 10, 8.5, 9.5, 6, 14, 13, 11, 12, 9)
```

Para obtermos a média de uma variável  $x$  no R, utilizamos a função `mean`:

```
mean(x)
```

```
## [1] 10
```

Nesse exemplo, a média dá uma boa ideia da tendência central dos dados, mas em outras situações ela pode ser enganosa. Um exemplo é a renda per capita média em um país com uma distribuição da renda bastante desigual, como o Brasil. Nesse caso, a renda média é bastante influenciada por indivíduos que possuem renda bem acima da maioria da população.

No exemplo da variável  $x$  acima, se substituirmos o primeiro valor de  $x$  por um que seja dez vezes a média, o que aconteceria com a média? Utilizando o R, podemos verificar como a média de  $x$  se altera.

O comando abaixo altera o primeiro valor de  $x$  para um valor igual a 10 vezes a média de  $x$ :

```
# substituindo o primeiro elemento de x por 10 vezes a média da variável  
x[1] = 10*mean(x)
```

Vamos interpretar esse comando. Lembrem que a variável  $x$  contém os valores 10, 9, 8, 12, 11, 7, 10, 8.5, 9.5, 6, 14, 13, 11, 12 e 9. Cada valor da variável  $x$  pode ser acessado, bastando indicar a posição do valor (índice do valor) entre colchetes. Assim  $x[1]$  aponta para o primeiro elemento de  $x$ .

Então o comando  $x[1] = 10 * mean(x)$  atribui ao primeiro elemento de  $x$  o valor da média multiplicado (\*) por 10. Na linha de comando do R, tudo que aparece após o sinal # é interpretado como comentário, apenas para esclarecer ao usuário a intenção do comando. Ele é ignorado pelo R.

Para visualizar o novo valor de  $x[1]$ , execute o comando abaixo:

```
x[1]
```

```
## [1] 100
```

Agora, a nova média da variável  $x$  é:

```
mean(x)
```

```
## [1] 16
```

Assim um único valor bastante acima dos demais valores de uma variável deslocou a média para um valor acima de todos os outros valores da variável. Aqui temos um caso de valores

extremos influenciando a média. Uma medida de tendência central que não é influenciada por valores extremos é a mediana.

### 3.2.2 Mediana

A **mediana** (*median* em inglês) é definida como o valor tal que 50% dos valores da variável estão acima da mediana e 50% estão abaixo. A obtenção da mediana é feita ordenando-se os dados e escolhendo-se o valor do meio. Por exemplo, se temos 21 valores, a mediana estará na 11ª posição. No caso de o número  $N$  de dados ser par, computamos a média dos dois valores ‘centrais’ (com 10 valores, a mediana será a média do 5º e 6º valor).

No R, temos a função *median* para fazer esse cálculo. Para a variável  $x$ , definida na seção anterior, o valor da mediana será obtido assim:

```
median(x)
```

```
## [1] 10
```

Observem que a mediana, mesmo com os dados tendo um valor 10 vezes maior que sua média, resulta em uma estimativa mais típica dos valores da variável  $x$ . A mediana é considerada uma medida robusta da tendência central, por não sofrer influências de valores extremos. Vamos verificar como a mediana foi calculada para a variável  $x$ . A função *sort* ordena a variável, conforme mostra a figura 3.4.

```
> sort(x)
[1]  6.0  7.0  8.0  8.5  9.0  9.0  9.5 10.0 11.0 11.0 12.0 12.0 13.0 14.0 100.0
```

Figura 3.4: Ordenação dos valores da variável  $x$  em ordem crescente por meio do comando *sort(x)*.

Como  $x$  possui 15 elementos, a mediana será o valor apontado pelo elemento na posição 8, depois que  $x$  foi ordenado. Esse valor é 10, como indicado pela seta azul na figura 3.5. O fato de o maior valor da variável ser 100, 1000 ou 10000000000 não terá nenhuma influência na determinação da mediana.

```
6.0 7.0 8.0 8.5 9.0 9.0 9.5 10.0 11.0 11.0 12.0 12.0 13.0 14.0 100.0
```



Figura 3.5: Posição da mediana da variável  $x$  indicada pela seta azul.

Assim, para dados que são assimétricos, a mediana é um melhor representante de um valor típico dos dados.

### 3.2.3 Moda

Considere a seguinte variável  $y$  que pode representar, por exemplo, a idade de pessoas que fazem parte de um grupo de pais que ensinam os seus filhos bem novos a nadar:

```
y = c(1, 1, 1, 2, 2, 2, 2, 3, 3, 31, 31, 32, 32, 32, 32, 33, 33, 33)
```

```
y
```

```
## [1] 1 1 1 2 2 2 2 3 3 31 31 32 32 32 32 33 33 33
```

```
median(y)
```

```
## [1] 17
```

```
mean(y)
```

```
## [1] 17
```

A média e a mediana de  $y$  não refletem o valor típico de  $y$ , que contém idades de crianças bem pequenas e dos respectivos pais. Essa variável possui mais de um valor típico: um para as crianças e outro para os pais. Uma medida que pode ser usada nesses casos é a moda.

A **moda** é a medida de maior frequência em um conjunto de dados. Os passos para se obter a moda são:

- 1) encontrar todos os valores distintos da variável;
- 2) obter a frequência de cada valor distinto;
- 3) selecionar o valor (ou valores) com a maior frequência para obter a moda.

Seguindo esses passos para o exemplo acima, temos:

1) valores distintos: 1, 2, 3, 31, 32, 33

2) frequências:

1  $\rightarrow$  3

2  $\rightarrow$  4

3  $\rightarrow$  2

31  $\rightarrow$  2

32  $\rightarrow$  4

33  $\rightarrow$  3

3) os dois valores com a maior frequência são 2 e 32. Assim a variável  $y$  possui duas modas.

Essa variável é bimodal.

Como obter a moda no R? O R não possui uma função específica para obtenção da moda, entretanto podemos usar uma composição de funções e métodos de indexação para obtermos esse valor. Uma possibilidade é usar os seguintes comandos:

```
temp <- table(y)      # a função table gera uma tabela de frequências
                      # para a variável y
temp
```

```
## y
##  1  2  3 31 32 33
##  3  4  2  2  4  3
```

Armazenamos a tabela de frequências de  $y$ , gerada a partir da função `table` em `temp`. Ao visualizarmos o conteúdo de `temp`, vemos a tabela de frequência igual à que calculamos manualmente. Na primeira linha aparece os elementos distintos de  $y$  e na linha abaixo as frequências de cada valor. Agora basta pegar os valores com contagem máxima, usando o comando a seguir:

```
names(temp)[temp == max(temp)]
```

```
## [1] "2"  "32"
```

Os valores obtidos para a moda foram 2 e 32, os mesmos obtidos manualmente. Vamos entender o comando acima. A função `names(temp)` retorna os valores distintos da variável  $y$ , ou a primeira linha da tabela `temp`:

```
names(temp)
```

```
## [1] "1"  "2"  "3"  "31" "32" "33"
```

A função `max(temp)` retorna a frequência máxima da tabela `temp` (4 nesse exemplo):

```
max(temp)
```

```
## [1] 4
```

A expressão `temp == max(temp)` vai retornar todos os itens de `temp` cujas contagens sejam iguais à frequência máxima. Logo `temp[temp == max(temp)]` retorna os elementos da tabela de frequência com frequência máxima:

```
temp[temp == max(temp)]
```

```
## y
##  2 32
##  4  4
```

Finalmente `names(temp)[temp == max(temp)]` retorna os valores com frequência máxima.

Mesmo que, ao aplicarmos o processo acima, identifiquemos um único valor que possua a frequência máxima, pode acontecer que, ao inspecionarmos o histograma da variável,

identificamos que a distribuição dos dados parece se concentrar em mais de um ponto, por exemplo a distribuição da variável *fiO2\_maximo* (figura 3.2). Nesses casos, dizemos que a função possui mais de uma moda, porque existe mais de um local onde os valores parecem se concentrar. No exemplo do *fiO2\_maximo*, diríamos que a variável é *bimodal*.

### 3.2.4 Discussão sobre medidas de tendência central

Que medida de tendência central deve ser usada ou apresentada em relatório ou artigo científico?

Depende dos dados! Como vimos nos exemplos acima, podem ocorrer diversas situações:

- 1) os dados estão distribuídos simetricamente em torno de um ponto central;
- 2) existem valores no conjunto de dados que se desviam marcadamente dos valores típicos (conhecidos em inglês por “outliers”);
- 3) as variáveis podem estar distribuídas de maneira assimétrica em maior ou menor grau;
- 4) os dados podem estar distribuídos em torno de mais de um valor;
- 5) etc.

No primeiro caso acima, a média é uma medida adequada. Já no terceiro caso, e às vezes no segundo, a mediana seria mais apropriada e, finalmente, no quarto caso, a moda seria a medida escolhida.

Assim o primeiro passo é inspecionar os dados e obter não somente as medidas de tendência central, como também conhecer como os dados estão dispersos em torno destas medidas, ou seja, precisamos também trabalhar com as medidas de dispersão, tema da seção seguinte.

## 3.3 Medidas de dispersão

Medidas de tendência central ou localização dos dados não dão a visão completa dos mesmos. Para melhor interpretarmos os dados, também precisamos saber como estes estão ‘espalhados’, isto é, se os dados estão localizados em sua maioria em torno da medida de tendência central ou estão mais dispersos.

### 3.3.1 Amplitude

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

Suponhamos que, em um curso de estatística, tenhamos 12 alunos que obtiveram as seguintes notas finais: 50, 66, 72, 76, 78, 84, 85, 86, 88, 89, 92, 100. Para uma segunda turma, também de 12 alunos, as notas foram: 50, 78, 78, 80, 80, 83, 84, 84, 83, 83, 83, 100. No R, podemos entrar estes dados e produzir uma representação gráfica que irá refletir a distribuição das notas de cada turma.

```
x1 = c(50, 66, 72, 76, 78, 84, 85, 86, 88, 89, 92, 100)
x2 = c(50, 78, 78, 80, 80, 83, 84, 84, 83, 83, 83, 100)
mean(x1); mean(x2)
```

```
## [1] 80.5
```

```
## [1] 80.5
```

Os dois primeiros comandos criaram duas variáveis, cada uma com as notas de estatística de cada turma. O terceiro comando calculou as médias de cada turma que, coincidentemente, são iguais. Observem que, numa única linha de comando, foram executados duas funções de cálculo de médias, separadas por “;”. No R, o sinal “;” representa um separador de comandos.

Uma medida de dispersão simples é a **amplitude** (*range* em inglês), que é a *distância entre o maior e o menor valor do conjunto de dados*. Usualmente, ela é calculada por  $\max(X) - \min(X)$ , onde  $X$  é a variável de interesse. A função *range* no R fornece o menor e o maior valor da variável entre parênteses:

```
range(x1); range(x2)
```

```
## [1] 50 100
```

```
## [1] 50 100
```

A partir da função *range*, a amplitude de cada uma das variáveis poderia ser obtida, subtraindo-se o primeiro valor do segundo valor gerado pela função *range*. A função *diff* calcula essa diferença. Podemos ver que a amplitude é a mesma para as variáveis  $x1$  e  $x2$ .

```
diff(range(x1))
```

```
## [1] 50
```

```
diff(range(x2))
```

```
## [1] 50
```

Vamos ver uma representação gráfica simples para esses dados. Os três comandos abaixo criam dois gráficos para variáveis discretas, o primeiro para a variável  $x1$  e o segundo para a variável  $x2$  (figura 3.6).

```
library(RcmdrMisc)
par(mar = c(4, 4, 1, 1), mfcol=c(1, 2))
discretePlot(x1, scale="frequency", ylab = "Frequência")
discretePlot(x2, scale="frequency", ylab = "Frequência")
```

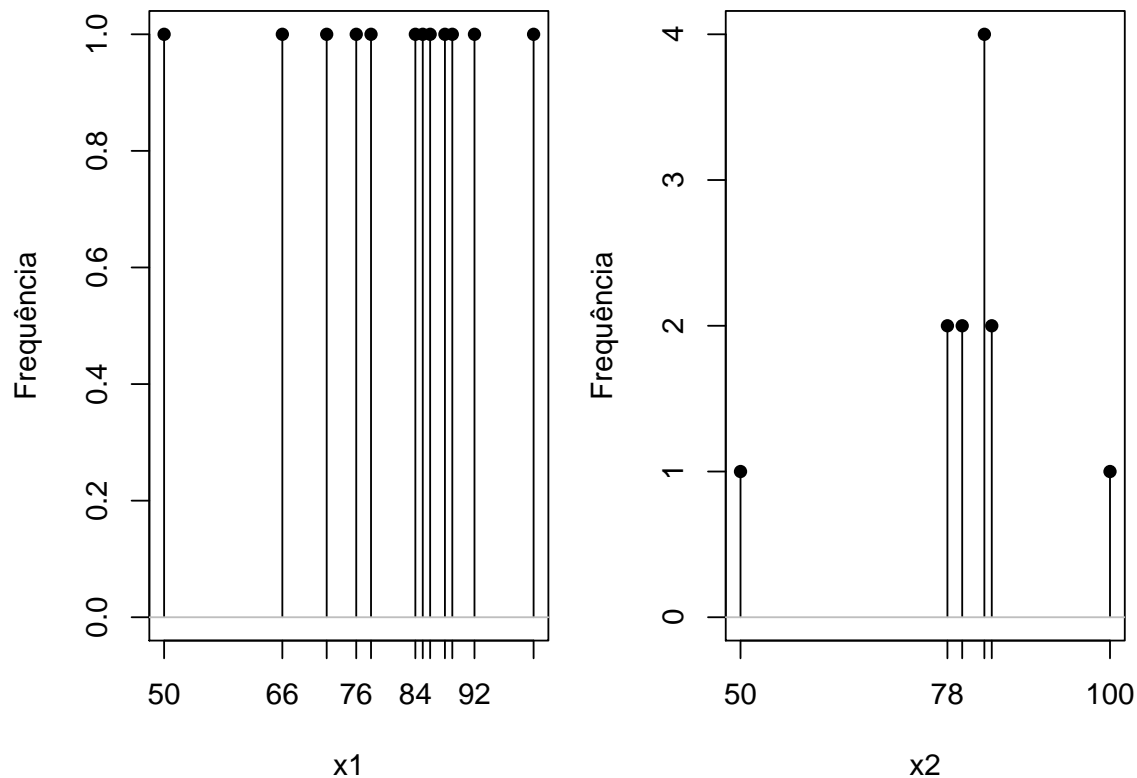


Figura 3.6: Gráfico de barras para as variáveis x1 e x2.

Observem que as distribuições têm diferentes formas, a mesma média e também a mesma amplitude, mas os valores da variável x1 estão mais espalhados do que os da variável x2. Devemos notar que a amplitude só considera os valores extremos, portanto não dá uma boa descrição da distribuição dos dados. Há muitas formas diferentes de construir variáveis com a mesma amplitude, mas diferentes distribuições dos valores. Uma vez que a amplitude considera os valores mais extremos de um conjunto de dados, ela não permite saber como eles estão distribuídos, ou se existem *outliers*. Precisamos de outras medidas de dispersão.

### 3.3.2 Distância interquartil

Os conteúdos desta seção e da seção 3.3.3 podem ser visualizados neste [vídeo](#).

Uma medida de dispersão que não sofre a influência de valores extremos é a amplitude interquartil, ou **distância interquartil** (DIQ ou IQR - *Interquartile range*, em inglês). Antes vamos apresentar o conceito de quantil.

Os **quantis** dividem os valores ordenados de uma variável numérica em q partes essencialmente iguais, ou em q partes com a mesma proporção de valores. Essa divisão dá origem a **q-quantis**.

Alguns quantis têm nomes especiais:

- Os 100-quantis são chamados percentis
- Os 10-quantis são chamados decis
- Os 5-quantis são chamados quintis

- Os 4-quantis são chamados quartis
- Os 3-quantis são chamados tercis

Podemos interpretar os quantis como medidas de posição. Diversas medidas de dispersão podem ser definidas a partir de quantis. Possivelmente, a mais utilizada delas é a distância interquartil.

Para entendermos a distância interquartil, precisamos inicialmente definir os quartis. **Quartis** são quantis que dividem os valores da variável em quatro partes:

- 1) Q1 (primeiro quartil, ou quartil inferior) define o valor para o qual 25% dos valores estão abaixo dele;
- 2) Q2 (segundo quartil) é o valor que tem 50% dos valores abaixo e 50% acima (é a mediana);
- 3) Q3 (terceiro quartil ou quartil superior) define o valor que possui 75% dos dados abaixo dele.

O intervalo interquartil é a amplitude (distância) entre o primeiro e terceiro quartil

$$\text{IQR} = \text{Q3} - \text{Q1}$$

Há diversos algoritmos para o cálculo dos quartis. Assim, dependendo do algoritmo, valores ligeiramente diferentes poderão ser obtidos, tanto para os quartis quanto para a distância interquartil. Um algoritmo simples funciona como a seguir.

- 1) Determinando o primeiro quartil:
  - a) ordene os valores em ordem crescente;
  - b) seja  $n$  o número de valores. Divida  $n$  por 4 (obter 25% de  $n$ );
  - c) se essa divisão for um número inteiro, então o primeiro quartil é o valor obtido pela média aritmética entre o valor na posição indicada por essa divisão e o valor seguinte;
  - d) se a divisão não for inteira, arredonde o número para cima. Esse número dá a posição do primeiro quartil.

Vamos aplicar esse algoritmo à variável  $x1$  no exemplo da seção anterior. Temos 12 elementos, logo  $12/4 = 3$ . O número na posição 3 é 72, como mostra a seta azul na figura 3.7.

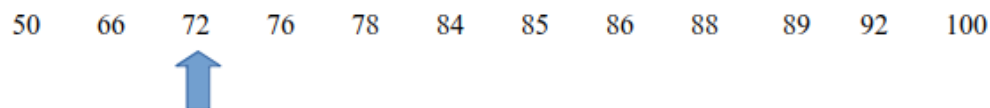


Figura 3.7: Posição indicada pela seta azul para o cálculo do primeiro quartil da variável  $x1$ .

Assim o primeiro quartil de  $x1$  é a média aritmética entre 72 e o valor na posição seguinte (76). Logo:

$$\text{Q1} = (72 + 76) / 2 = 74$$



2) Determinando o terceiro quartil:

- a) Ordene os valores em ordem crescente;
- b) seja  $n$  o número de valores. Divida  $3n$  por 4 (obter 75% de  $n$ );
- c) se essa divisão for um número inteiro, então o terceiro quartil é o valor obtido pela média aritmética entre o valor na posição indicada por essa divisão e o valor seguinte;
- d) se a divisão não for inteira, arredonde o número para cima. Esse número dá a posição do terceiro quartil.

Vamos aplicar esse algoritmo à variável  $x1$  no exemplo da seção anterior. Temos 12 elementos, logo  $(3 \times 12)/4 = 9$ . O número na posição 9 é 88, como mostra a seta na figura 3.8.

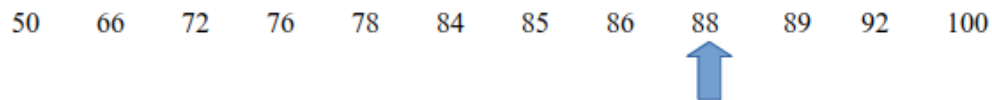


Figura 3.8: Posição indicada pela seta azul para o cálculo do terceiro quartil da variável  $x1$ .

Assim o terceiro quartil de  $x1$  é a média aritmética entre 88 e o valor na posição seguinte (89). Logo:

$$Q3 = (88 + 89) / 2 = 88,5$$

A distância interquartil para a variável  $x1$  é  $Q3 - Q1 = 14,5$  ( $88,5 - 74$ ). Para obtermos os quartis no R (ou qualquer quantil), usamos a função *quantile*. Para obtermos o primeiro quartil de  $x1$ , usamos a função *quantile* conforme mostrado a seguir:

```
quantile(x1, 0.25, type=2)
```

```
## 25%  
## 74
```

Para obtermos o terceiro quartil de  $x1$ , usamos a função *quantile* conforme mostrado a seguir:

```
quantile(x1, 0.75, type=2)
```

```
## 75%  
## 88.5
```

A função *quantile* admite 9 diferentes algoritmos, especificado pelo argumento *type*. Fazendo *type = 2*, será utilizado o algoritmo explicado acima.

Podemos obter o primeiro e o terceiro quartil com uma única chamada da função, da seguinte forma:

```
quantile(x1, probs = c(0.25, 0.75), type=2)
```

```
## 25% 75%  
## 74.0 88.5
```

Para a variável  $x2$ , os valores do primeiro e do terceiro quartil serão:

```
quantile(x2, probs = c(0.25, 0.75), type=2)
```

```
## 25% 75%  
## 79.0 83.5
```

A função *IQR* fornece a distância interquartil. Obtendo a distância interquartil de  $x1$ :

```
IQR(x1, type=2)
```

```
## [1] 14.5
```

Obtendo a distância interquartil de  $x2$ :

```
IQR(x2, type=2)
```

```
## [1] 4.5
```

A variável  $x2$  possui uma dispersão menor do que a variável  $x1$ , quando utilizamos a distância interquartil, refletindo melhor a distribuição das variáveis  $x1$  e  $x2$ .

A distância interquartil é uma medida de variabilidade mais estável ou ‘robusta’, pois ela não é influenciada por valores extremos, como é o caso da amplitude.

### 3.3.3 Percentis

Os **percentis** são quantis utilizados com bastante frequência. O percentil 10, indicado por  $P_{10}$ , designa o valor para o qual 10% dos valores da variável estão abaixo dele e assim por diante.

O primeiro quartil ( $Q1$ ) é, portanto, igual ao  $P_{25}$ ; a mediana é igual ao  $P_{50}$ , e o terceiro quartil ( $Q3$ ) é igual ao  $P_{75}$ . Um algoritmo semelhante ao utilizado para o cálculo do quartil pode ser utilizado para os percentis.

Assim para calcular o percentil  $\alpha\%$ :

- ordene os valores em ordem crescente;
- seja  $n$  o número de valores. Multiplique  $n$  por  $\alpha/100$ ;
- se essa divisão for um número inteiro, então  $P_\alpha$  é o valor obtido pela média aritmética entre o valor na posição indicada por essa divisão e o valor seguinte;
- se a divisão não for inteira, arredonde o número para cima. Esse número dá a posição de  $P_\alpha$ .

Vamos calcular o  $P_{10}$  para a variável  $x1$  no exemplo da seção anterior. Temos 12 elementos, logo  $12 \times 10/100 = 1,2$ . Como essa divisão não é inteira, arredondamos o quociente para cima. Essa é a posição do  $P_{10}$  (seta azul na figura 3.9). Logo  $P_{10} = 66$ .

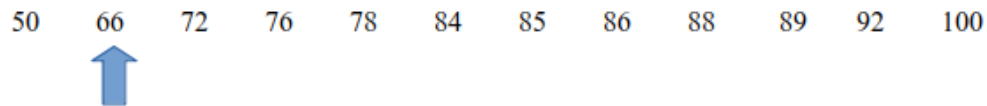


Figura 3.9: Posição indicada pela seta azul para o cálculo do percentil 10 da variável  $x_1$ .

No R, para obtermos qualquer percentil, usamos a função *quantile* com o valor do percentual desejado. Logo, para obtermos o  $P_{10}$  de  $x_1$ , fazemos:

```
quantile(x1, 0.10, type=2)
```

```
## 10%
## 66
```

Um exemplo da utilização de percentis são as curvas de crescimento, utilizadas para comparar o desenvolvimento de uma criança, por exemplo, com o crescimento esperado de uma população de crianças. A figura 3.10 mostra as curvas de crescimento (peso por idade) para meninas do nascimento até 5 anos, padronizadas pela Organização Mundial de Saúde (OMS). As curvas indicam, de baixo para cima, os percentis 3%, 15%, 50%, 85% e 97%, respectivamente. Na idade de 3 anos, por exemplo, a curva do percentil 3% indica que 3% das crianças pesam menos de 11 Kg (indicado pela seta vermelha na figura).

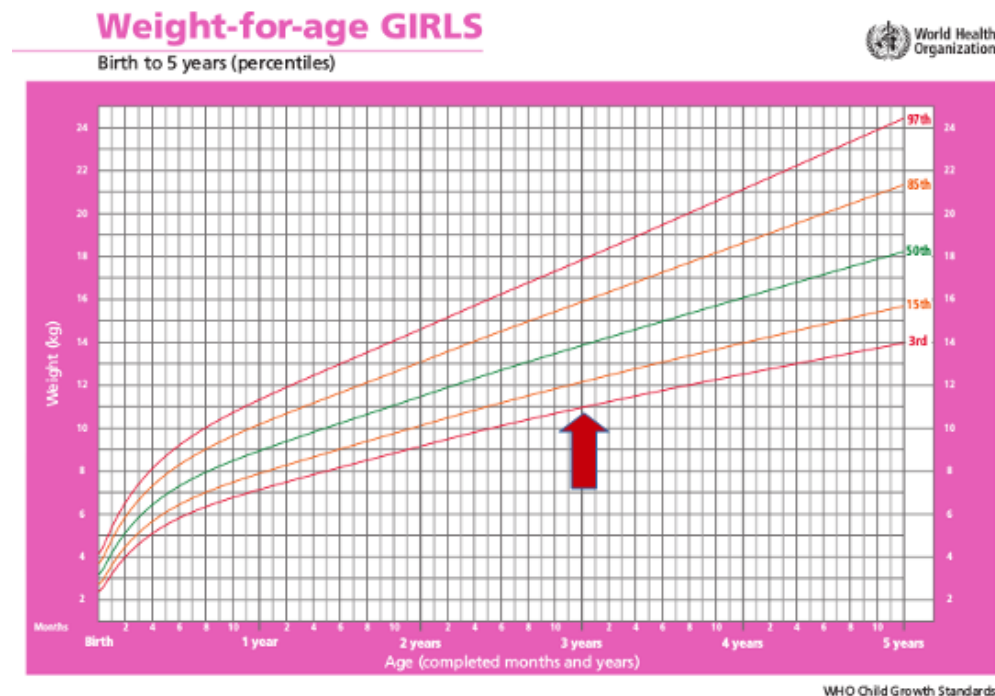


Figura 3.10: Curvas de crescimento padronizadas pela OMS. A seta indica o  $P_3$  do peso para a idade de 3 anos. Fonte: [Organização Mundial de Saúde](#) (CC BY-NC-SA 3.0 IGO).

### 3.3.4 Desvio padrão e variância

Os conteúdos desta seção e da seção 3.3.5 podem ser visualizados neste [vídeo](#).

Um problema com as medidas de dispersão anteriores, como a amplitude e a distância interquartil, é que elas mostram somente as diferenças entre valores em determinadas posições, quando se ordenam os valores da variável. Existe alguma medida que indica a variabilidade de uma variável, mas que leva em conta todos os seus valores?

Sim. Uma primeira ideia é calcular o desvio de cada valor em relação à média ( $x - \bar{x}$ ), somá-los e dividir o resultado pelo número de valores. O problema é que o desvio médio será sempre 0. Desvios negativos serão cancelados por desvios na outra direção, como mostra a expressão abaixo:

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) = \frac{1}{n} \left( \sum_{i=1}^n x_i - \sum_{i=1}^n \bar{x} \right) = \frac{1}{n} \sum_{i=1}^n x_i - \frac{1}{n} \sum_{i=1}^n \bar{x} = \bar{x} - \bar{x} = 0$$

Uma segunda possibilidade é calcular a média das distâncias de cada valor à média. A distância de cada valor à média é sempre positiva. Se o valor está acima da média, a distância é igual ao desvio. Se o valor estiver abaixo da média, a distância é igual ao oposto do desvio.

Uma outra alternativa é elevar cada desvio em relação à média ao quadrado, somá-los e dividir essa soma por  $n-1$ . Nesse caso, todas as parcelas da soma serão positivas e a medida de variabilidade assim obtida é chamada de variância ( $s^2$ ).

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

A razão por que divide-se por  $n-1$  e não por  $n$  será vista no capítulo sobre propriedades de estimadores (capítulo 13).

O desvio padrão ( $s$ ) é calculado, extraindo-se a raiz quadrada da variância  $s^2$ , e é medido na mesma unidade da variável à qual ele se refere.

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

O desvio padrão e a variância consideram todos os valores da variável para o seu cálculo e são bastante utilizados na análise estatística.

No R, temos como obter essas medidas, usando a função `sd` para o desvio padrão ou `var` para a variância.

```
var(x1); var(x2)
```

```
## [1] 176.6364
```

```
## [1] 124.8182
```

```
sd(x1); sd(x2)
```

```
## [1] 13.29046
```

```
## [1] 11.17221
```

Como nas medidas de distância interquartil, os desvios padrões de  $x1$  e  $x2$  também indicam uma maior dispersão para  $x1$ .

### 3.3.5 Discussão sobre as medidas de dispersão

Diversas medidas de dispersão foram apresentadas nas seções anteriores. A amplitude é uma medida bastante simplista de dispersão, já que ela mede a distância entre o maior e o menor valor, mas não fornece nenhuma indicação de como os valores estão distribuídos em torno de uma medida de tendência central, como a média e a mediana.

A variância e o desvio padrão são medidas bem mais adequadas para indicar a variabilidade dos dados e como eles estão dispersos, já que elas consideram como os dados estão agrupados.

Há uma conexão entre a média, o desvio padrão e amplitude. Se  $X$  for uma variável,  $n$  o número de valores,  $m$  a média aritmética e  $s$  o desvio padrão, Harding (Harding, 1996) mostrou que:

$$m - s\sqrt{n-1} \leq X \leq m + s\sqrt{n-1}$$

Logo a amplitude é, no máximo,  $2s\sqrt{n-1}$

A apresentação do 1º, 2º (Mediana) e 3º quartis dá uma boa noção da dispersão dos dados e também indica se os dados estão distribuídos simetricamente em torno da mediana. Essas medidas são as mais frequentemente utilizadas para a construção dos diagramas de caixa (*boxplot* em inglês), que serão apresentados no próximo capítulo.

Para variáveis que seguem uma distribuição normal (em forma de sino), o desvio padrão tem uma interpretação simples: 68,3% dos valores da variável estão situados a uma distância de um desvio padrão em torno da média aritmética (figura 3.11). Essa figura superpõe uma curva normal em um histograma com forma de sino. A distribuição normal será apresentada no capítulo 11.

Entretanto, para outras distribuições, a interpretação pode ser diferente. A figura 3.12 mostra quatro distribuições diferentes, todas com o mesmo valor do desvio padrão, mas que a probabilidade de selecionarmos aleatoriamente um valor da variável e ele estar na região compreendida entre a média  $\pm$  desvio padrão varia de uma distribuição para outra. A distribuição indicada pela letra d é assimétrica em relação à média, ou seja, há uma probabilidade maior de se obter aleatoriamente valores abaixo da média do que acima dela; nesse caso, a mediana e os percentis  $P_{25}$  e  $P_{75}$  são melhores medidas de tendência central e dispersão do que a média e o desvio padrão.

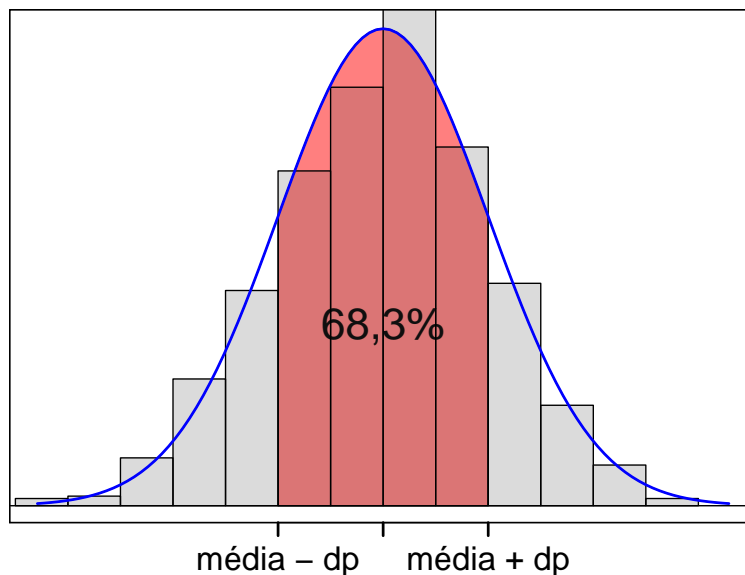


Figura 3.11: Em variáveis que seguem uma distribuição normal, 68,3% dos valores da variável estão situados a uma distância de um desvio padrão em torno da média aritmética.

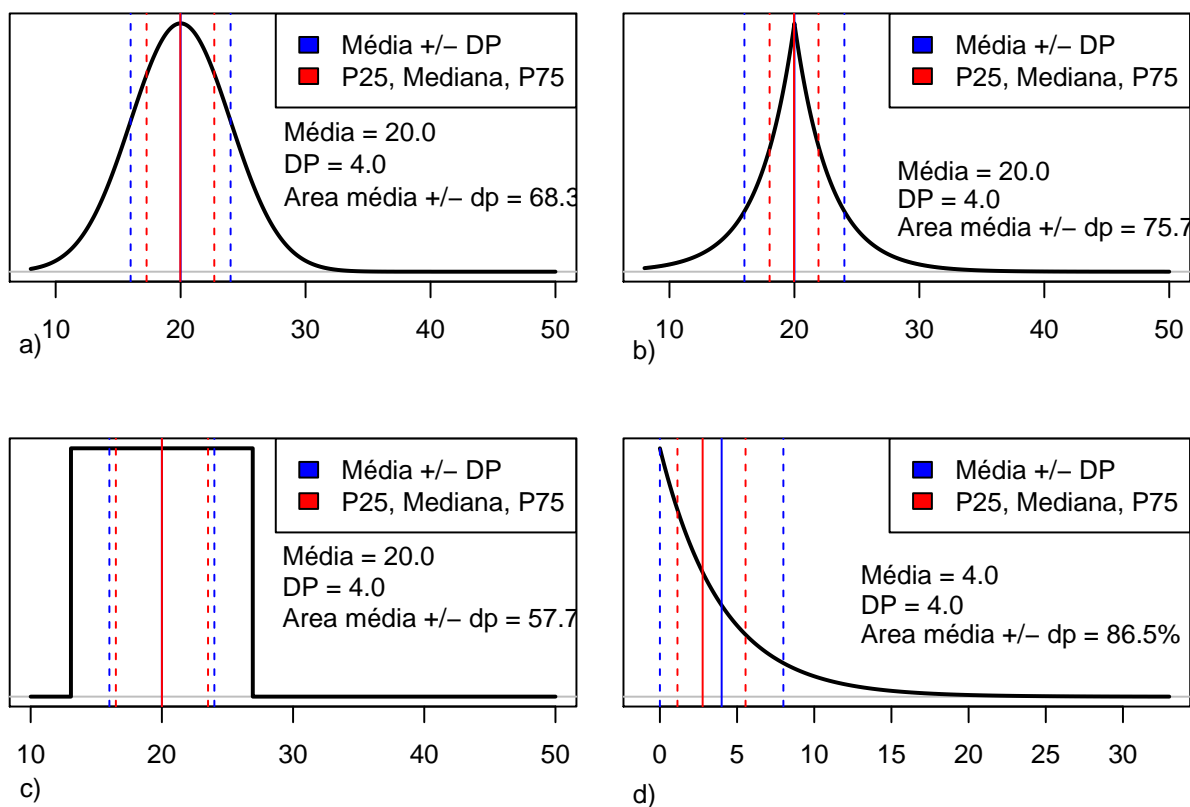


Figura 3.12: Diferentes distribuições de dados que mostram diferentes probabilidades de os valores se situarem na região compreendida entre a média  $\pm$  desvio padrão. Quando a distribuição é assimétrica (d), a mediana e os percentis  $P_{25}$  e  $P_{75}$  dos dados são melhores medidas de tendência central e dispersão.

A figura 3.13 mostra uma distribuição empírica onde os dados se concentram em duas regiões distintas. Nesse exemplo, nem a mediana,  $P_{25}$  e  $P_{75}$  e a média e desvio padrão são boas medidas de tendência central e dispersão. É melhor caracterizar essa distribuição como bimodal.

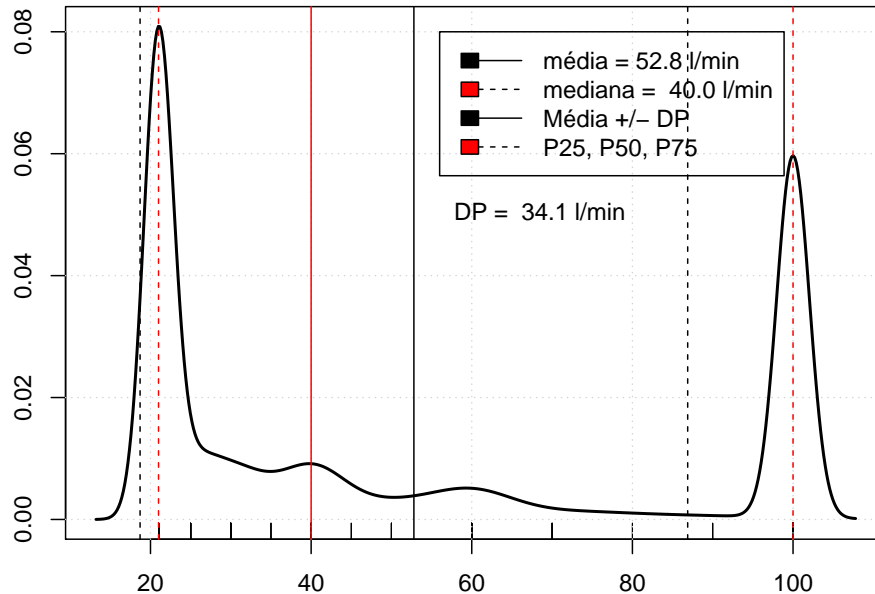


Figura 3.13: Distribuição bimodal.

O R tem uma função, *summary*, que fornece de uma única vez várias estatísticas descritivas discutidas aqui. Vamos utilizá-la para os valores de  $x2$ .

```
summary(x2)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  50.00   79.50   83.00   80.50   83.25   100.00
```

A função *summary* fornece o valor mínimo, o valor máximo, a média, a mediana, o primeiro e o terceiro quartil. Observem que os valores do primeiro e do terceiro quartil são ligeiramente diferentes daqueles que calculamos anteriormente. Isso acontece porque, na chamada da função *summary*, se não especificarmos o tipo de algoritmo, é utilizado por padrão um algoritmo diferente do que utilizamos para calcular os quartis.

## 3.4 Apresentação das estatísticas descritivas em publicações

Os conteúdos desta seção e de suas subseções podem ser visualizados neste [vídeo](#).

Os relatórios e artigos científicos de estudos clínico-epidemiológicos geralmente apresentam as medidas de tendência central e dispersão de variáveis numéricas. A seguir, serão mostrados exemplos de apresentações inadequadas e adequadas dessas medidas.

### 3.4.1 Exemplos de formas inadequadas de apresentação da média e desvio padrão

A frase a seguir foi extraída de um parágrafo de um trabalho científico: **“Foram avaliados 33 pacientes. A idade variou de 32 a 74 anos (média:  $51,4 \pm 10,5$ ).”**

Inicialmente, os autores apresentam a amplitude da idade (32 a 74 anos). O que significa a expressão entre parênteses: (média:  $51,4 \pm 10,5$ )? Se somarmos 51,4 com 10,5, o valor não é igual a 74, o valor máximo de idade. Também se subtrairmos 10,5 de 51,4, o valor não é igual a 32, o mínimo de idade. Possivelmente, os autores querem dizer que o valor 10,5 é o desvio padrão da idade. Não seria mais claro para os leitores se os autores simplesmente escrevessem que o desvio padrão é 10,5?

O vício apresentado acima é bastante frequente. A tabela 3.1 mostra um exemplo de como a média e o desvio padrão são frequentemente apresentados. O problema dessa forma de apresentação é que induz a uma interpretação equivocada do desvio padrão. O desvio padrão é uma medida da dispersão dos dados, não da faixa de variação em torno da média.

Tabela 3.1: Exemplo de uma apresentação inadequada da média e da dispersão de uma variável.

Variável	Média $\pm$ DP
Idade (anos)	$56,4 \pm 5,1$
Glicemia de jejum (mg/dl)	$90,2 \pm 21,4$

A tabela 3.2 mostra outro exemplo que coloca o desvio padrão entre parênteses após a média, que seria uma forma adequada, porém coloca o desvio padrão precedido do sinal  $\pm$ . O que os autores querem dizer com o sinal  $\pm$  antes do desvio padrão?

Tabela 3.2: Outro exemplo de uma apresentação inadequada da média e da dispersão de uma variável.

Variável	Média (DP)
Idade (anos)	$56,4 (\pm 5,1)$
Glicemia de jejum (mg/dl)	$90,2 (\pm 21,4)$

A sentença a seguir também foi extraída de um estudo: **“Os dados foram apresentados como médias  $\pm$  desvios-padrão (DP), mediana e variação, de acordo com a distribuição de normalidade.”** A que se refere a palavra variação? Ao desvio padrão, distância interquartil, ou amplitude? Os autores não esclarecem.

A tabela 3.3 mostra uma tabela com o mesmo problema do texto acima.



Tabela 3.3: Exemplo de uma apresentação inadequada da média e da dispersão de uma variável.

Variável	Mediana (Variação)
Estatura (m2)	1,60 (1,5 - 1,7)
Massa corporal (kg)	52,7 (44,8-60,8)

### 3.4.2 Exemplos de formas adequadas de apresentação da média, mediana, desvio padrão e primeiro e terceiro quartis

A tabela 3.4 mostra um exemplo que segue a boa prática para a apresentação da média e do desvio padrão: média seguida do desvio padrão entre parênteses.

Tabela 3.4: Exemplo de uma apresentação adequada da média e da dispersão de uma variável.

Variável	Média (DP)
Idade (anos)	56,4 (5,1)
Glicemia de jejum (mg/dl)	90,2 (21,4)

A tabela 3.5 mostra uma tabela que mostra a mediana e entre parênteses os valores do primeiro e terceiro quartil, que é uma forma clara e não ambígua de mostrar a dispersão de uma variável.

Tabela 3.5: Exemplo de uma apresentação adequada da mediana e da dispersão de uma variável.

Variável	Mediana (1º quartil - 3º quartil)
Estatura (m2)	1,60 (1,5 - 1,7)
Massa corporal (kg)	52,7 (44,8-60,8)

## 3.5 Escore z ou Escore padrão

Vamos considerar o seguinte exemplo, extraído do livro “Head First Statistics” (Griffiths, 2008). Dois jogadores de um time de basquetebol, com diferentes habilidades, possuem os seguintes desempenhos:

Jogador 1: acerta 70% das cestas com um desvio padrão de 20%.

Jogador 2: acerta 40% das cestas com um desvio padrão de 10%.

Em um certo jogo, o jogador 1 teve 75% de acertos e o jogador 2, 60%. Qual jogador teve o melhor desempenho, levando em conta o seu desempenho histórico?

O jogador 1 acertou um maior percentual de cestas do que o jogador 2, mas isso era esperado em função de seu retrospecto. Entretanto, se levarmos em conta a média e o desvio padrão de cada jogador, verificamos que o jogador 2 teve uma atuação marcante nesse jogo: ele acertou 20% a mais do que o esperado (60% - 40%), 2 desvios padrões (2 x 10%) a mais do que a sua média.

O jogador 1 acertou apenas 5% a mais do que a sua média (75% - 70%), um quarto do desvio padrão (0,25 x 20%) acima de sua média.

Olhando por essa perspectiva, o jogador 2 teve o melhor desempenho, quando comparado com o seu desempenho normal. Essa medida que foi utilizada para a comparação é chamada **escore z** (*z-score* em inglês) ou escore padrão e é obtida pela fórmula:

$$z = \frac{x - \bar{x}}{s}$$

onde:

z = escore z

$\bar{x}$  = média

s = desvio padrão

x = valor a partir do qual iremos calcular o escore z.

Para o jogador 2, o escore z é:

$$z = \frac{60 - 40}{10} = 2$$

O escore z fornece um padrão para compararmos valores de diferentes conjuntos de dados. Ele informa quantos desvios padrões um dado valor está distante em relação à média. Um escore z positivo indica um valor acima da média, enquanto que um escore z negativo indica um valor abaixo da média.

O escore z é muito utilizado na análise estatística. A figura 3.14 mostra curvas de crescimento (peso por idade) para meninas do nascimento até 5 anos, padronizadas pela OMS, semelhantes às da figura 3.10, porém agora as curvas são construídas por meio dos escores z. Então, para uma menina de certa idade, mede-se o seu peso e verificamos qual o seu escore z aproximado, observando entre quais escores z o seu peso se situa na curva de crescimento. Assim é possível verificar como está o desenvolvimento dessa criança em relação ao esperado na população.

Os escores padrão (escores z) não são propriamente medidas de dispersão, mas sim uma forma de indicar como um valor se situa em relação aos valores da variável de onde ele se origina. Nesse sentido, ele pode ser utilizado para detectar possíveis dados desviantes (*outliers*) no conjunto de dados. Por exemplo, às vezes *outliers* são considerados aqueles dados que se afastam mais de 3 desvios padrões da média (escore z < -3 ou escore z > 3).

O escore z também pode ser usado para comparar diferentes valores em variáveis diferentes, mesmo quando essas variáveis possuem diferentes médias e desvios padrões.

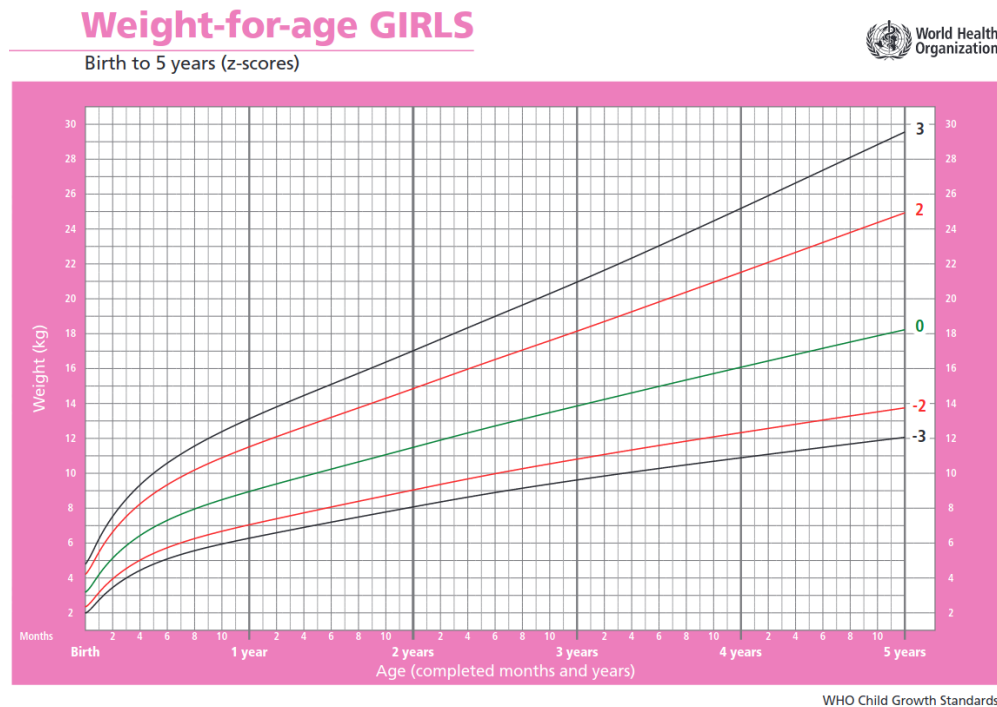


Figura 3.14: Curvas de crescimento padronizadas pela OMS, semelhante à da figura 3.10, porém as curvas são construídas a partir dos escores z. Fonte: [Organização Mundial de Saúde \(CC BY-NC-SA 3.0 IGO\)](#).

## 3.6 Obtendo estatísticas descritivas no R

Nesta seção, iremos utilizar o *R Commander* para obter as medidas de tendência central e dispersão de variáveis e também aprender mais alguns recursos dessa interface gráfica.

Os conteúdos das subseções desta seção podem ser visualizados neste [vídeo](#).

### 3.6.1 Carregando conjuntos de dados de pacotes do R

Nesta seção, vamos carregar um outro conjunto de dados já disponível em um pacote do R. Trata-se do conjunto de dados *juul2* do pacote *ISwR* (GPL-2 | GPL-3).

Nós já instalamos o pacote *ISwR* na seção 2.2.

Antes de abrirmos o conjunto de dados *juul2*, é preciso carregar o pacote *ISwR*. Na área de *script* do R Commander, digitamos `library(ISwR)` e, com o cursor na linha do comando, clicamos no botão *Submeter* (figura 3.15).

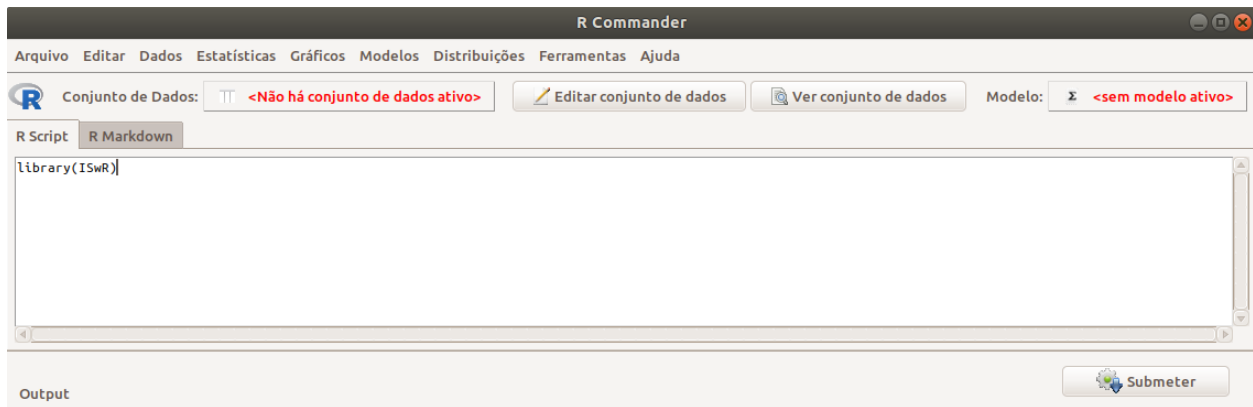


Figura 3.15: Tela do *R commander*, com a digitação da função *library(ISwR)* na área de *Script*.

A função *library(ISwR)* carregou a biblioteca *ISwR*. Para abrir o conjunto de dados *juul2*, vamos seguir um procedimento análogo ao utilizado para abrir o conjunto de dados *stroke* na seção 2.2, que será mostrado novamente a seguir.

Vamos selecionar a opção a seguir no *R Commander*:

Dados  $\Rightarrow$  Conjunto de dados em pacotes  $\Rightarrow$  Ler dados de pacotes 'atachados'

Na tela *Leia dados do pacote*, para ver a lista dos conjuntos de dados em *ISwR*, damos um duplo clique nesse pacote e uma lista de conjuntos de dados será mostrada à direita (figura 3.16). Rolamos essa lista e clicamos no conjunto de dados *juul2* para selecioná-lo (figura 3.16).

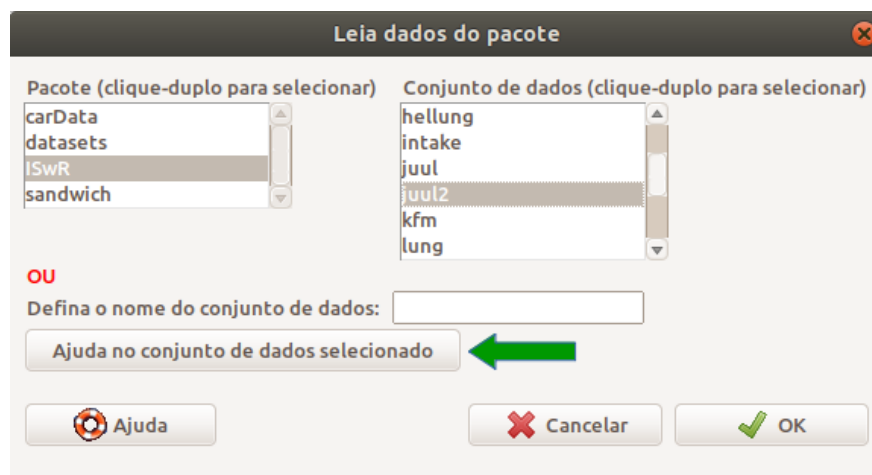


Figura 3.16: Visualizando a lista de conjuntos de dados do pacote *ISwR* e selecionando o conjunto *juul2*.

Para conhecermos a estrutura desse conjunto de dados, clicamos no botão *Ajuda para o conjunto de dados selecionado* (seta verde na figura 3.16). Uma descrição desse conjunto de dados será exibida na janela de seu navegador padrão (figura 3.17). Ao clicarmos no botão OK na figura 3.16, após termos selecionado *juul2*, esse conjunto de dados será carregado no *R Commander* (figura 3.18).

juul2 {ISwR}

R Documentation

Juul's IGF data, extended version

Description

The `juul2` data frame has 1339 rows and 8 columns; extended version of `[juul]`.

Usage

`juul2`

Format

This data frame contains the following columns:

age

a numeric vector (years).

height

a numeric vector (cm).

menarche

a numeric vector. Has menarche occurred (code 1: no, 2: yes)?

sex

a numeric vector (1: boy, 2: girl).

igf1

a numeric vector, insulin-like growth factor (*microgram per liter*).

tanner

a numeric vector, codes 1-5: Stages of puberty ad modum Tanner.

testvol

a numeric vector, testicular volume (ml).

weight

a numeric vector, weight (kg).

Source

Original data.

Examples

`plot(igf1~age, data=juul2)`

---

[Package *ISwR* version 2.0-8 [Index](#)]

Figura 3.17: Texto com a descrição do conjunto de dados *juul2* exibido no navegador padrão.

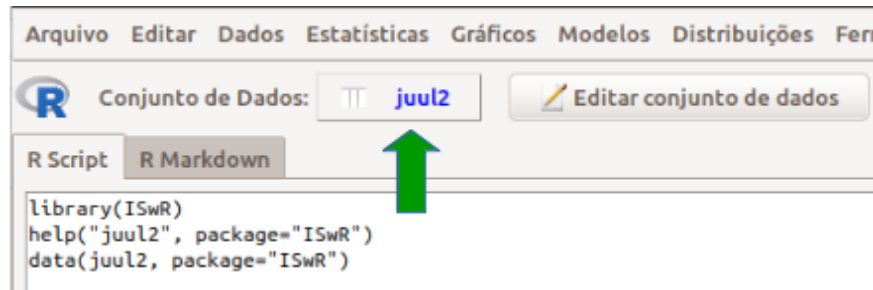


Figura 3.18: Tela do *R Commander* após o carregamento do conjunto de dados *juul2*. Observem a função que foi executada – *data(juul2, package="ISwR")* – e o nome do conjunto selecionado (seta verde).

Na área de mensagens do *R Commander*, aparece a seguinte mensagem abaixo do comando, indicando o número de registros e de variáveis no conjunto de dados *juul2*:

NOTA: Os dados *juul2* tem 1339 linhas e 8 colunas.

### 3.6.2 Obtendo resumos numéricos pelo R Commander

O conjunto de dados *juul2* possui 1339 registros, cada registro com valores de 8 variáveis. Ele contém uma amostra da distribuição da variável *insulin-like growth factor (igf1)*, com os dados coletados em exames físicos, sendo a maior parte dos dados de pessoas em idade escolar, mas também inclui outras faixas etárias. Vamos obter algumas medidas de tendência central e dispersão para as variáveis idade e *igf1*.

No item de menu *Estatística*, clique em *Resumos* e, a seguir, em *Resumos numéricos*:

Estatísticas  $\Rightarrow$  Resumo...  $\Rightarrow$  Resumos numéricos...

Na tela *Resumos Numéricos*, selecionamos as variáveis na aba *Dados*. Para selecionarmos mais de uma variável, mantemos a tecla Ctrl pressionada enquanto clicamos nas variáveis desejadas. Nesse exemplo, vamos selecionar as variáveis *age* e *igf1* (Figura 3.19). Em seguida, clicamos na aba *Estatísticas* (seta verde).

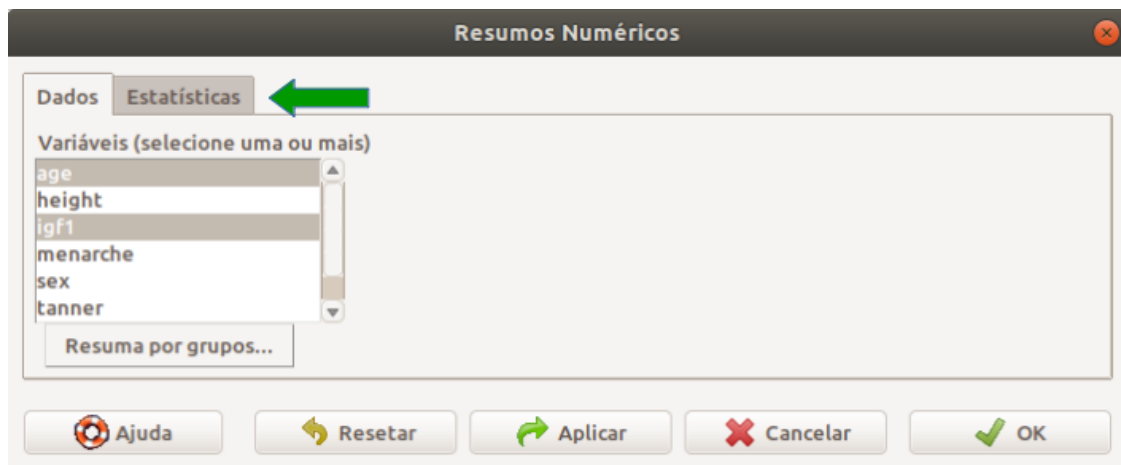


Figura 3.19: Seleção das variáveis para as quais um resumo numérico será mostrado. A seta verde indica a aba onde podem ser selecionadas as medidas que serão apresentadas.

Na aba *Estatísticas* (figura 3.20), observem que as medidas média, desvio padrão, distância interquartil e quantis já estão marcadas. Se desejarmos outros quantis, basta digitá-los na caixa de texto com o rótulo *Quantis*, separados por vírgula, não esquecendo que o separador de decimal no R é o ponto. Ao clicarmos em OK, os resultados serão apresentados ou na console do *RStudio* (figura 3.21) ou na área de resultados (*Output*) do *R Commander*, caso o *R Commander* tenha sido carregado a partir do *RStudio* ou não, respectivamente.

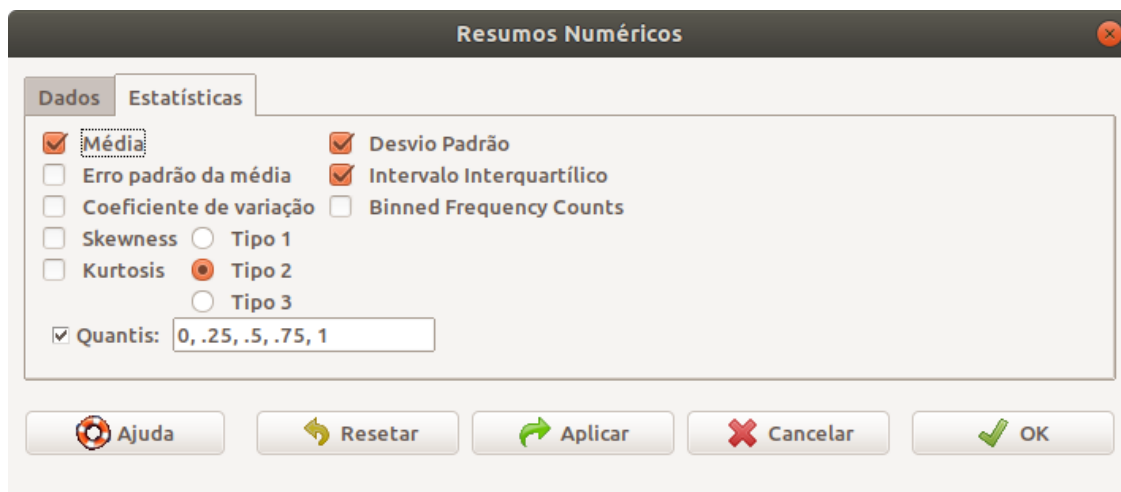


Figura 3.20: Tela para a seleção das medidas que serão apresentadas no resumo numérico.

```
Console Terminal x Jobs x
~/Estatistica/livro/bookdown/estatistica/ ↵

Rcmdr> data(juul2, package="ISwR")
RcmdrMsg: [3] NOTA: Os dados juul2 tem 1339 linhas e 8 colunas.

Rcmdr> library(abind, pos=18)

Rcmdr> library(e1071, pos=19)

Rcmdr> numSummary(juul2[,c("age", "igf1"), drop=FALSE], statistics=c("mean", "sd",
Rcmdr+   "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
      mean      sd    IQR   0%    25%   50%   75% 100%   n  NA
age  15.09535 11.25288  7.8025 0.17  9.0525 12.56 16.855 83 1334  5
igf1 340.16798 171.03560 260.5000 25.00 202.2500 313.50 462.750 915 1018 321
> |
```

Figura 3.21: Resumos numéricos para as variáveis *age* e *igf1*.

Observem que as medidas são calculadas corretamente, mesmo considerando que, em alguns registros, alguns dados não foram preenchidos. Dados ausentes são indicados por *NA* (*not available*, em inglês). Os resultados indicam que 5 registros não possuem valores para idade e 321 registros não possuem valores de *igf1*.

Se formos usar as funções para obter essas medidas individualmente, poderemos ter problemas. Por exemplo, digitando a função `mean(juul2$age)` na área de *script* do *R Commander* e clicando no botão *Submeter*, o resultado será *NA* (figura 3.22). Isso ocorreu devido aos 5 registros sem valores de idade. O comando `mean(juul2$age)` também poderia ser executado na console do *RStudio*.

```
Console Terminal x Jobs x
~/Estatistica/livro/bookdown/estatistica/ ↵

Rcmdr> numSummary(juul2[,c("age", "igf1"), drop=FALSE], statistics=c("mean", "sd",
Rcmdr+   "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
      mean      sd    IQR   0%    25%   50%   75% 100%   n  NA
age  15.09535 11.25288  7.8025 0.17  9.0525 12.56 16.855 83 1334  5
igf1 340.16798 171.03560 260.5000 25.00 202.2500 313.50 462.750 915 1018 321

Rcmdr> mean(juul2$age)
[1] NA
> |
```

Figura 3.22: Resultado da função `mean(juul2$age)` na presença de valores ausentes.

Na função `mean(juul2$age)`, a variável *age* é precedida do nome do conjunto de dados e do símbolo \$. Essa é a sintaxe utilizada para referenciar uma variável que faz parte de um conjunto de dados:

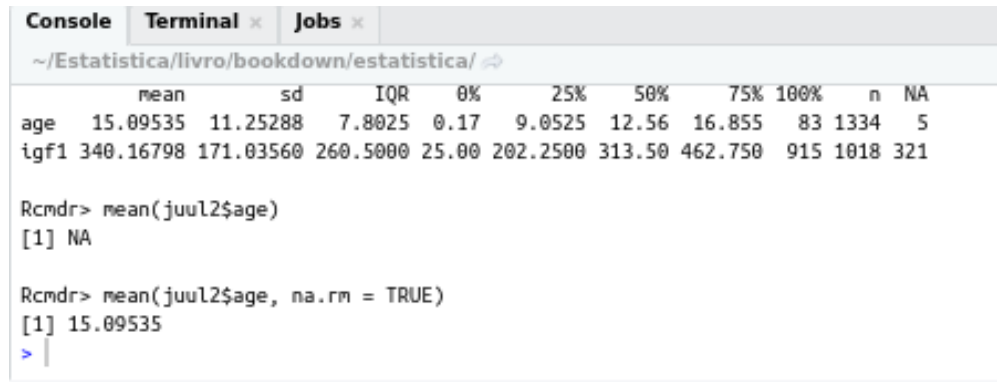
conjuntoDeDados\$nomeDaVariável



Para calcular a média, não incluindo os registros com valores ausentes, usamos o argumento `na.rm=TRUE`, o qual sinaliza para remover do cálculo os registros com valores ausentes. Assim, para obtermos a média de idade, usamos a função abaixo:

```
mean(juul2$age, na.rm=TRUE)
```

Agora a média será obtida corretamente (figura 3.23).



The screenshot shows the R Commander interface with three tabs: Console, Terminal, and Jobs. The Console tab is active, displaying a summary of the 'juul2' dataset and the results of two R commands. The summary table includes columns for mean, sd, IQR, 0%, 25%, 50%, 75%, 100%, n, and NA. The first command, `mean(juul2$age)`, returns `[1] NA`. The second command, `mean(juul2$age, na.rm = TRUE)`, returns `[1] 15.09535`.

	mean	sd	IQR	0%	25%	50%	75%	100%	n	NA
age	15.09535	11.25288	7.8025	0.17	9.0525	12.56	16.855	83	1334	5
igf1	340.16798	171.03560	260.5000	25.00	202.2500	313.50	462.750	915	1018	321

```
Rcmdr> mean(juul2$age)
[1] NA

Rcmdr> mean(juul2$age, na.rm = TRUE)
[1] 15.09535
>
```

Figura 3.23: Obtendo a média de idade corretamente por meio da função `mean(juul2$age, na.rm=TRUE)`.

### 3.6.3 R Markdown

O *R Markdown* é uma linguagem que permite que um relatório possa ser gerado a partir dos comandos que vão sendo executados no R. No *R Commander*, ele pode ser visualizado na aba *R Markdown* (seta verde na figura 3.24).



Figura 3.24: Acessando a aba *R Markdown* no *R Commander*.

Esse relatório pode ser personalizado pelo usuário. Por exemplo, no texto da figura 3.25, alteramos o título e o autor (seta verde na figura), depois selecionamos o comando `help...` (figura 3.26) e o apagamos (figura 3.27) para não exibir a ajuda do conjunto de dados em

uma outra página web quando for gerado o relatório. Ao clicarmos no botão *Gerar relatório*, o relatório será apresentado no navegador padrão de seu computador (figura 3.28).

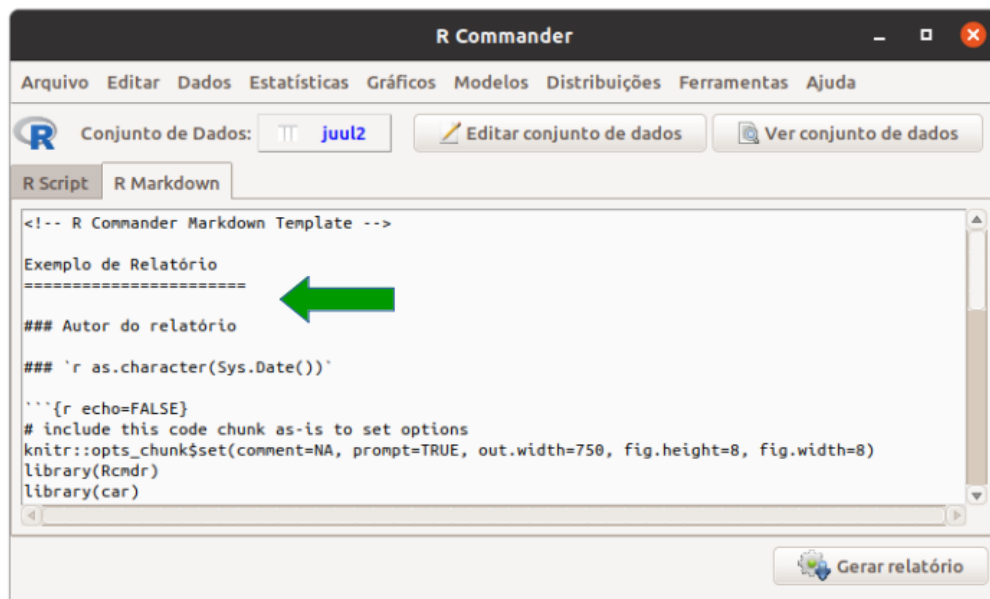


Figura 3.25: Personalizando o título e o autor do relatório no *R Markdown*.



Figura 3.26: Selecionando partes do relatório para edição.



Figura 3.27: Remoção da área selecionada na figura 3.26.

## Exemplo de Relatório

### Autor do relatório

2021-11-04

```
> library(ISwR)
```

```
> data(juul2, package="ISwR")
```

```
> library(abind, pos=18)
```

```
> library(e1071, pos=19)
```

```
> numSummary(juul2[,c("age", "igf1"), drop=FALSE], statistics=c("mean", "sd",  
+ "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
```

	mean	sd	IQR	0%	25%	50%	75%	100%	n	NA
age	15.09535	11.25288	7.8025	0.17	9.0525	12.56	16.855	83	1334	5
igf1	340.16798	171.03560	260.5000	25.00	202.2500	313.50	462.750	915	1018	321

```
> mean(juul2$age)
```

```
[1] NA
```

```
> mean(juul2$age, na.rm=TRUE)
```

```
[1] 15.09535
```

Figura 3.28: Relatório gerado pelo *R Markdown* em html para os comandos utilizados nesta seção.

### 3.6.4 Salvando scripts e arquivos do R Markdown

Todos os comandos utilizados numa seção do *R Commander* podem ser salvos em um arquivo. É possível também editar os comandos, inclusive removê-los, na janela de *Script* antes de salvá-lo. Em outra seção do *R Commander* ou *R Studio*, o arquivo salvo pode ser reaberto e os comandos executados, sem necessidade de executar cada um deles individualmente.

Para salvar o script no *R Commander*, selecionamos a seguinte opção:

Arquivo ⇒ Salvar script como...

Na janela *Salvar como* (figura 3.29), selecionamos a pasta onde o arquivo será gravado e digitamos o nome do mesmo na caixa de texto *Nome do Arquivo*. Vamos manter a extensão do arquivo (.R). Em seguida, clicamos no botão *Salvar* e o arquivo será gravado na pasta desejada.

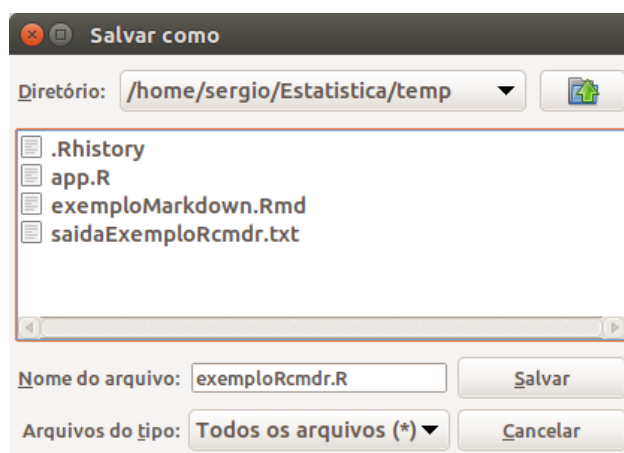


Figura 3.29: Selecionando a pasta e digitando o nome do *script* a ser gravado.

Para salvar o arquivo gerado na aba *R Markdown*, selecionamos a seguinte opção.

Arquivo ⇒ Salvar arquivo R Markdown como...

Na janela *Salvar como* (figura 3.30), selecionamos a pasta onde o arquivo será gravado e digitamos o nome do mesmo na caixa de texto *Nome do Arquivo*. Vamos manter a extensão do arquivo (.Rmd). Em seguida, clicamos no botão *Salvar* e o arquivo será gravado na pasta desejada.

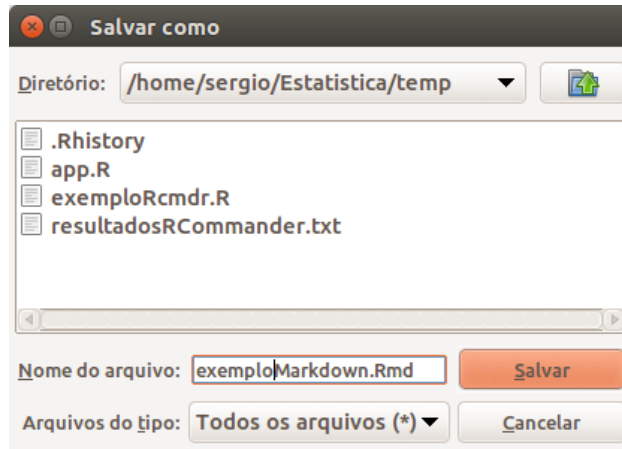


Figura 3.30: Selecionando a pasta e digitando o nome do arquivo do *R Markdown* a ser gravado.

### 3.7 Executando scripts no R Commander

Se, na aba *R Script*, selecionarmos um conjunto de comandos ao mesmo tempo e clicarmos no botão *Submeter*, os comandos serão executados em sequência de maneira automática (figura 3.31) e os resultados serão exibidos ou na console do *R Studio* ou na área de resultados (*Output*) do *R Commander*, caso o *R Commander* tenha sido carregado a partir do *RStudio* ou não, respectivamente.



Figura 3.31: Execução automática e em sequência de um conjunto de comandos selecionados na aba *R Script*.

## 3.8 Exercícios

- 1) No estudo intitulado “Capacidade preditiva de indicadores antropométricos para o rastreamento da dislipidemia em crianças e adolescentes” (Quadros et al., 2015), os resultados de colesterol total, HDL, LDL e triglicerídios são apresentados na figura 3.32 por sexo e faixa etária.
  - a) A apresentação dos valores de média e desvio padrão são adequados? Justifique.
  - b) Qual a sua opinião sobre o por que de os autores apresentarem os resultados de triglicerídios como mediana e percentis 25 e 75 e não como média e desvio padrão?
  - c) Qual a distância interquartil de triglicerídeos para o sexo feminino?

	n	Colesterol total (mg/dL) <sup>a</sup>	HDL-C (mg/dL) <sup>a</sup>	LDL-C (mg/dL) <sup>a</sup>	Triglicérides (mg/dL) <sup>b</sup>
<b>Sexo</b>					
Masculino	506	146,6 (29,3)	47,1 (11,4)	82,4 (24,0)	77 (60, 104)
Feminino	633	151,1 (30,2)	47,1 (10,7)	85,4 (25,4)	82 (65, 111)
p <sup>c</sup>		0,501	0,296	0,336	0,006
<b>Faixa etária (anos)</b>					
6-7	155	155,9 (30,1)	50,2 (10,8)	88,9 (26,4)	80 (62, 106)
8-9	208	148,9 (28,5)	47,8 (10,7)	83,7 (23,7)	75 (60, 107)
10-12	314	149,5 (31,7)	48,1 (11,2)	84,0 (25,5)	77 (63, 105)
13-15	285	157,1 (29,7)	45,6 (10,5)	82,8 (24,9)	85 (64, 114)
16-18	177	145,9 (27,7)	44,4 (11,0)	82,5 (23,1)	84 (65, 108)
p <sup>d</sup>		0,002	0,001	0,022	0,009
Total	1,139	149,1 (29,9)	47,1 (11,0)	84,1 (24,8)	80 (63, 108)

HDL-C, lipoproteína de alta densidade; LDL-C, lipoproteína de baixa densidade.

<sup>a</sup> Média (desvio padrão).

<sup>b</sup> Mediana (percentis 25, 75).

<sup>c</sup> Nível de significância para o colesterol total, HDL-C e LDL-C (teste t de Student) e triglicérides (teste de Mann-Whitney).

<sup>d</sup> Nível de significância para o colesterol total, HDL-C e LDL-C (tendência linear) e triglicérides (teste de Jonckheere-Terpstra).

Figura 3.32: Tabela 2 do estudo de (Quadros et al., 2015) (CC-BY-NC-ND).

- 2) A tabela a seguir (figura 3.33) foi extraída de um estudo que avalia a associação entre o distúrbio ventilatório restritivo (DVR) e o risco cardiovascular e nível de atividade física em adultos assintomáticos.
  - a) Como são apresentadas as medidas de tendência central e dispersão para as variáveis numéricas?
  - b) Essa forma de apresentação é adequada? Justifique a sua resposta.
  - c) Se a resposta à questão (c) é negativa, que modo de apresentação você sugeriria?
  - d) Para a variável CVF, é correto afirmar que 68,3% dos valores estão situados entre 2,92 e 4,92 litros no grupo dos pacientes com padrão espirométrico normal? Justifique a sua resposta.

Característica	Padrão espirométrico	
	Normal (n = 337)	Restritivo (n = 37)
Idade, anos	42 ± 15	47 ± 16*
Gênero (%)		
Feminino	53,3	73,0**
Masculino	43,8	27,0
CVF, l	3,92 ± 1,00	2,75 ± 0,85**
CVF, % do previsto	98 ± 12	74 ± 9**
VEF <sub>1</sub> , l	3,21 ± 0,80	2,13 ± 0,64**
VEF <sub>1</sub> , % do previsto	98 ± 11	71 ± 6**
VEF <sub>1</sub> /CVF, %	82 ± 5	80 ± 5*
Raça (%)		
Branca	59,7	57,1
Negra	7,6	5,7
Parda	30,0	31,4
Oriental	2,1	0,0
Indígena	0,6	5,7*
Peso, kg	75 ± 18	77 ± 20
Estatura, cm	165 ± 97	161 ± 98
IMC, kg/m <sup>2</sup>	27 ± 6	29 ± 7**
Massa gorda, %	28 ± 8	33 ± 10*
Massa magra, kg	53 ± 12	50 ± 11
Pico de VO <sub>2</sub> , ml/min	2,383 ± 863	1,928 ± 814**
Pico de VO <sub>2</sub> , ml/min/kg	32 ± 10	25 ± 10**
Pico de VO <sub>2</sub> , % do previsto	100 ± 20	91 ± 19**
DTC6, m	605 ± 90	519 ± 118**
DTC6, % do previsto	105 ± 13	93 ± 17**
Fatores de risco de DCV (%)		
História familiar	24,6	16,2
Obesidade	26,1	40,5**
Hipertensão	9,5	21,6*
Dislipidemia	21,1	27,0
Diabetes	5,9	16,2*
Tabagismo atual	10,4	21,6*
Inatividade física	16,8	36,4*
Uso de medicações (%)	26,7	43,2*
Ocorrência de quedas (%)	5,3	18,9*
Ensino superior completo (%)		
Sim	40,2	27,0
Não	59,7	72,9**

VO<sub>2</sub>: captação pulmonar de oxigênio; DTC6: distância percorrida no teste de caminhada de seis minutos; e DCV: doença cardiovascular. \*Valores expressos em forma de média ± dp, exceto onde indicado.

Figura 3.33: Tabela 1 do estudo de (Sperandio et al., 2016) (CC-BY-NC).

- 3) Com o conjunto de dados *VA* do pacote *MASS* (GPL-2 | GPL-3) do R, faça as atividades abaixo.
  - a) Verifique a ajuda para o conjunto de dados.
  - b) Carregue o conjunto de dados.
  - c) Visualize os registros do conjunto de dados.
  - d) Verifique as seguintes estatísticas para as variáveis *diag.time* e *stime*: mínimo, máximo, média, mediana, P<sub>25</sub>, P<sub>75</sub> e número de observações.

- 4) Com o conjunto de dados *Melanoma* do pacote *MASS* do R, faça as atividades abaixo.
- a) Verifique a ajuda para o conjunto de dados.
  - b) Carregue o conjunto de dados.
  - c) Visualize os registros do conjunto de dados.
  - d) Verifique as seguintes estatísticas para as variáveis *time* e *thickness*: mínimo, máximo, média, mediana,  $P_{25}$ ,  $P_{75}$  e número de observações.



# Capítulo 4

## Visualização de dados

Uma introdução sobre diversos diagramas que são frequentemente utilizados para explorar os dados de um conjunto de dados pode ser visualizada neste [vídeo](#).

Os dois capítulos anteriores deram início à exploração de dados com as tabelas de frequência e as medidas de tendência central e de dispersão. Neste capítulo, serão abordados alguns recursos gráficos para visualizar a distribuição dos valores das variáveis de um determinado conjunto de dados. Os diagramas que serão abordados são:

diagrama de barras

diagrama de setores (pizza ou torta)

diagrama de caixa (*boxplot*)

histograma

histograma de densidade de frequência

diagrama de pontos

diagrama de *strip chart*

diagrama de dispersão

Os dois primeiros diagramas (barras e tortas) são utilizados para indicar a contagem (frequência) ou proporção de cada categoria de uma variável categórica. Os demais são utilizados para variáveis numéricas.

A figura 4.1 é um dos resultados do estudo de Furuta et al. (Furuta et al., 2003), que consistiu de um estudo clínico randomizado duplo-cego em crianças com adenoide obstrutiva submetidas a tratamento homeopático. Ela mostra um diagrama de barras das frequências dos tratamentos de acordo com a evolução, considerando a avaliação nasofibroscópica.

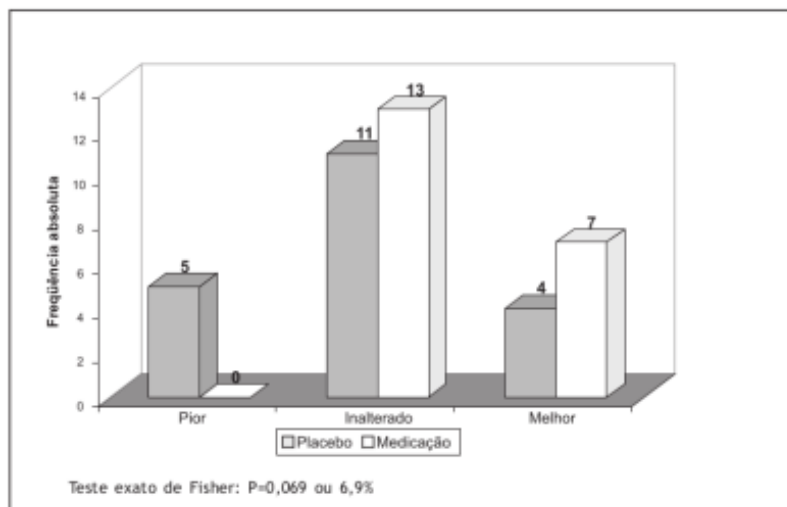


Figura 4.1: Diagrama de barras das frequências dos tratamentos segundo a evolução dos pacientes de acordo com a avaliação nasofibrosópica inicial e final. Fonte: (Furuta et al., 2003) (CC BY-NC).

As figuras 4.2 e 4.3 apresentam os diagramas de pizza mostrando o percentual de cada uma das categorias da evolução de acordo com a avaliação nasofibrosópica inicial e final para os tratamentos homeopático (figura 4.2) e para o placebo (figura 4.3). O percentual de inalterado no tratamento homeopático foi de 65%, enquanto que, no placebo, ele foi de 55%. Nenhum paciente teve uma piora na homeopatia e 25% teve um piora no grupo placebo. Finalmente 35% por cento tiveram uma melhora com o tratamento homeopático versus 20% com o placebo.

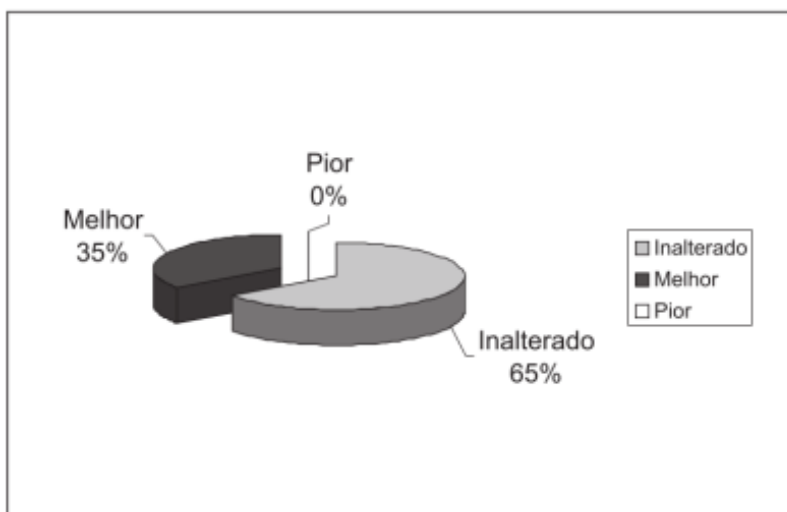


Figura 4.2: Diagrama de pizza da evolução dos pacientes do grupo I (medicamento homeopático), comparando as nasofibroskopias inicial e final do tratamento. Fonte: (Furuta et al., 2003) (CC BY-NC).

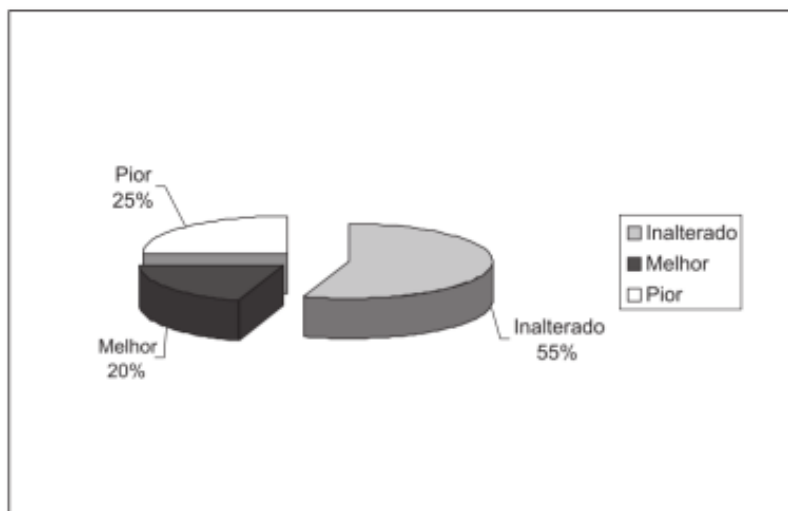


Figura 4.3: Diagrama de pizza da evolução dos pacientes do grupo II (placebo), comparando as nasofibroscopias inicial e final do tratamento. Fonte: (Furuta et al., 2003) (CC BY-NC).

Os dois diagramas de pizza (figuras 4.2 e 4.3) fornecem em conjunto a mesma informação que o diagrama de barras (figura 4.1) de que o tratamento homeopático tem o melhor desempenho do que o placebo em relação à avaliação nasofibroscópica, porém a análise estatística mostrou que essas diferenças não foram estatisticamente significativas ao nível de 5%.

O *boxplot* é utilizado para verificar a distribuição dos valores de uma variável numérica. A figura 4.4 apresenta diagramas de *boxplot* da contagem de diversos marcadores imunológicos em pacientes com a pele exposta à radiação ultravioleta (em vermelho) ou coberta (em azul). O diagrama sugere que as contagens dos marcadores CD45 e CD68 assim como o número de mastócitos ATM são maiores nos indivíduos expostos do que nos indivíduos não expostos à radiação ultra-violeta.

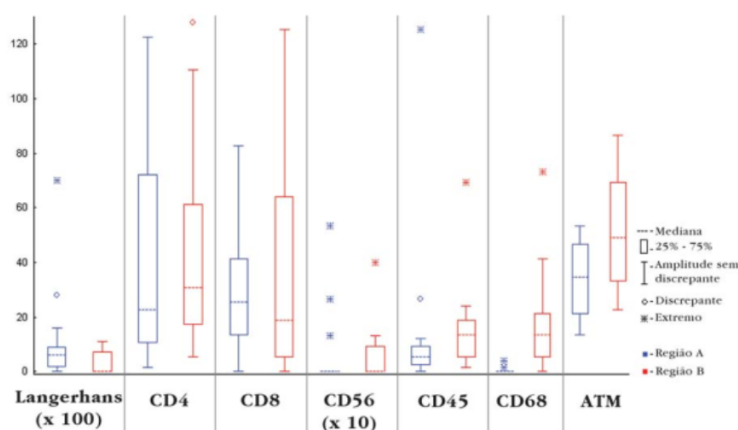


Figura 4.4: *Boxplots* dos marcadores imunológicos em pele coberta e exposta. Fonte: (Bezerra et al., 2011) (CC BY-NC).

A figura 4.5 mostra quatro gráficos de pontos dos níveis séricos das citocinas TNF-fator de necrose tumoral, interleucina-6, interleucina 8 assim como da RANTES (CCL5) para os seguintes grupos de pacientes: controles saudáveis, pacientes com DPOC (Doença Pulmonar Obstrutiva Crônica) e limitação do fluxo aéreo não reversível e pacientes com DPOC e limitação ao fluxo aéreo parcialmente reversível. O diagrama de pontos também é utilizado para verificar a distribuição dos valores de uma variável numérica.

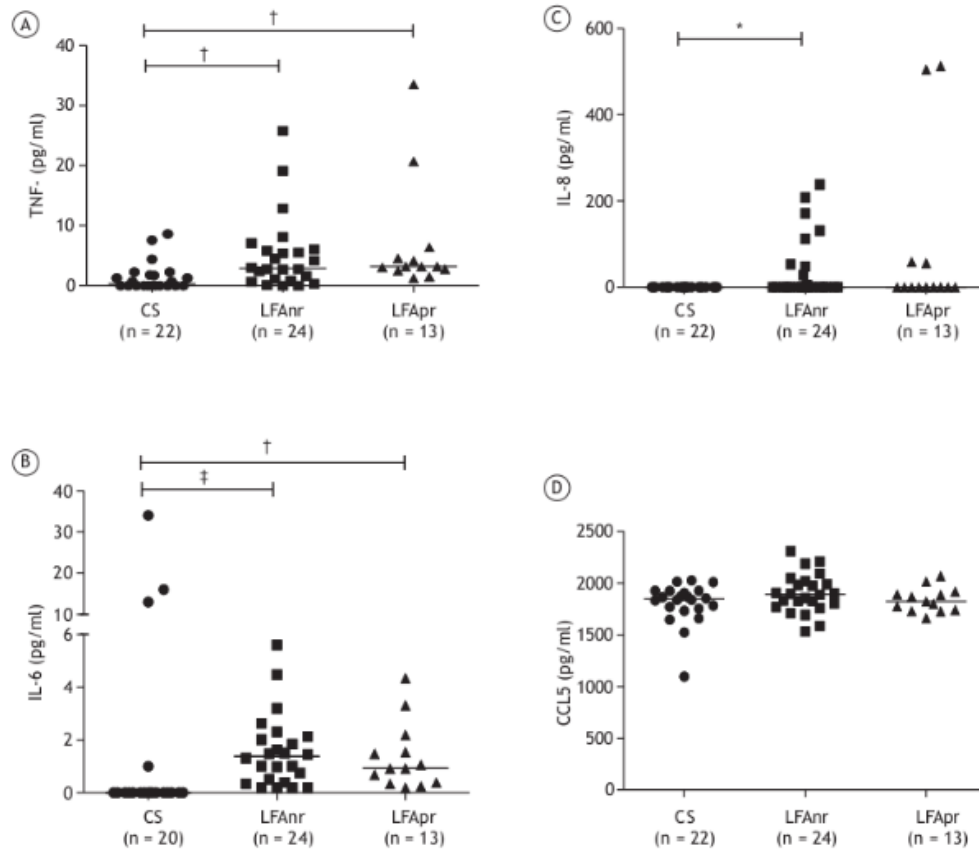


Figura 4.5: Gráfico de pontos dos níveis séricos das citocinas TNF (A), IL-6 (B) e IL-8 (C), assim como da RANTES (CCL5; D), em controles saudáveis (CS), pacientes com DPOC e limitação ao fluxo aéreo não reversível (LFAnr) e pacientes com DPOC e limitação ao fluxo aéreo parcialmente reversível (LFApr). Fonte: (Queiroz et al., 106) ([CC BY-NC](#)).

A figura 4.6 mostra um histograma do escore SYNTAX em pacientes com doença arterial coronariana multiarterial com indicação de intervenção coronária percutânea eletiva ou na vigência de síndrome coronariana aguda sem elevação do segmento ST com stents farmacológicos de primeira ou segunda gerações. O escore SYNTAX permite a estratificação do paciente quanto à complexidade angiográfica das lesões coronarianas.

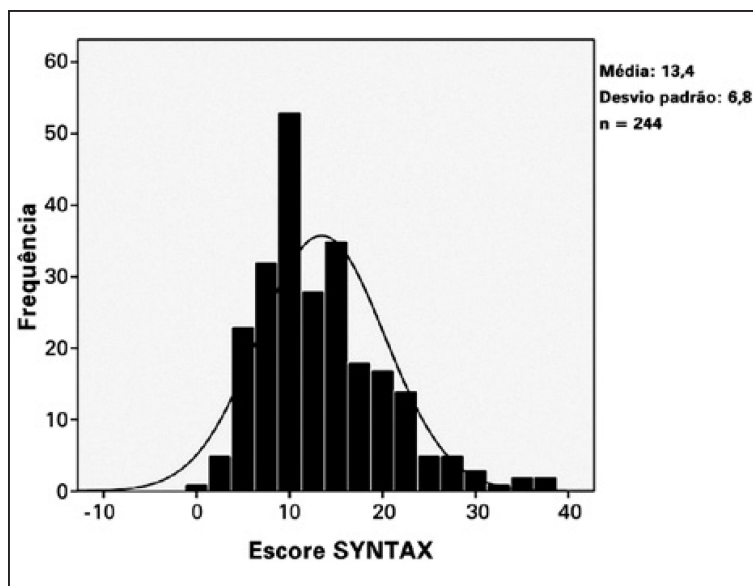


Figura 4.6: Histograma do escore SYNTAX em pacientes com doença arterial coronariana multiarterial. Fonte: (Silva et al., 2014) ([CC BY-NC](#)).

A figura 4.7 mostra um diagrama de dispersão ou espalhamento da relação fração de ejeção do ventrículo esquerdo versus relação neutrófilo-linfócito. Cada ponto corresponde aos valores da relação neutrófilo-linfócito na abscissa e a fração de ejeção do ventrículo esquerdo na ordenada para cada paciente do estudo. Esse gráfico sugere que a fração de ejeção do ventrículo esquerdo tende a diminuir à medida que a relação neutrófilo-linfócito aumenta e os autores apresentaram no gráfico a reta que melhor se ajusta a esses pontos.

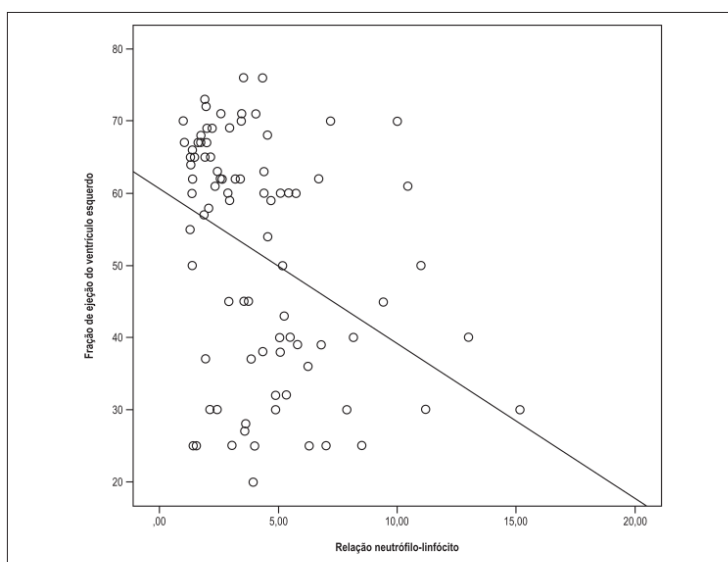


Figura 2 – Análise de correlação da relação neutrófilo-linfócito com a fração de ejeção do ventrículo esquerdo.

Figura 4.7: Diagrama de dispersão da relação neutrófilo-linfócito versus fração de ejeção do ventrículo esquerdo. Fonte: (Durmus et al., 2015) ([CC BY](#)).

Vamos utilizar neste capítulo o *R Commander*, carregado a partir do *RStudio*, mas também poderíamos utilizar o *R Commander*, carregado a partir da tela de entrada do R. Para a construção dos diagramas, vamos utilizar o conjunto de dados *juul2*, após a conversão das variáveis *sex*, *menarche* e *tanner* de numéricas para categóricas (*fator*).

Inicialmente, vamos executar o *RStudio*.

Em seguida, digitamos os comandos abaixo na console do *RStudio* e pressionamos a tecla *Enter* (figura 4.8).

```
plot.new(); library(Rcmdr)
```

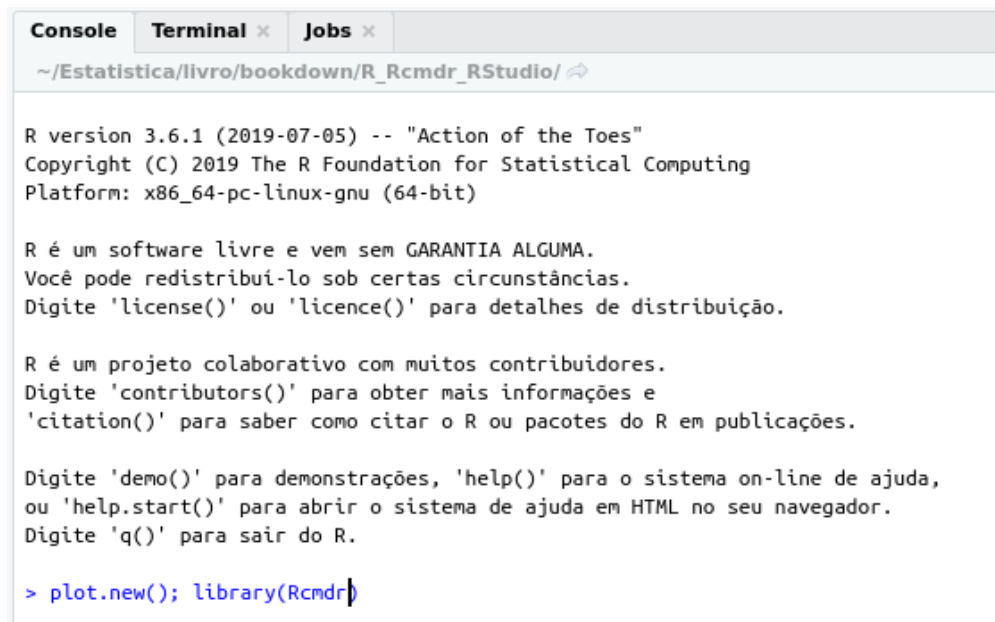


Figura 4.8: Console do *RStudio*, após a digitação dos comandos para carregar o pacote *Rcmdr*.

**Observação:** A função *plot.new()*, é executada antes do carregamento do *R Commander* para garantir que os gráficos gerados pelo *R Commander* sejam mostrados na interface do *RStudio*. Caso ela (ou alguma função que gere um gráfico no *RStudio*) não seja executada antes do carregamento do *R Commander*, os gráficos gerados pelo *R Commander* serão exibidos em uma janela separada.

## 4.1 Convertendo uma variável numérica para fator

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

No *R Commander*, vamos carregar o pacote *ISwR* (GPL-2 | GPL-3) e, em seguida, o conjunto de dados *juul2* (seção 3.6.1).

```
library(ISwR)
data(juul2, package="ISwR")
```

Para analisarmos e visualizarmos as variáveis categóricas corretamente no R, elas têm que serem da classe *factor* (fator), outro nome utilizado em análises estatísticas para variáveis categóricas. Vamos nesta seção converter as variáveis *tanner*, *sex* e *menarche* para fator, porque elas estão codificadas no conjunto de dados como números. A seguir, será mostrado o passo a passo para converter a variável *tanner* para fator, usando o *R Commander*.

Para realizarmos a conversão de uma variável numérica em fator no *R Commander*, selecionamos a opção:

Dados ⇒ Modificação var. conj. dados ⇒ Converter var. numérica para fator

Na tela *Converter Variáveis Numéricas p/ Fator* (figura 4.9), selecionamos a variável que será convertida e escolhemos uma das opções: *manter as categorias expressas como números*, ou *fornecer nomes às categorias*. Vamos dar nomes às categorias neste exemplo. Na caixa de texto, digitamos o nome da variável que será criada. Se nenhum nome for especificado nessa caixa, os nomes serão sobrescritos aos valores numéricos na própria variável que será convertida e não será criada uma nova variável.

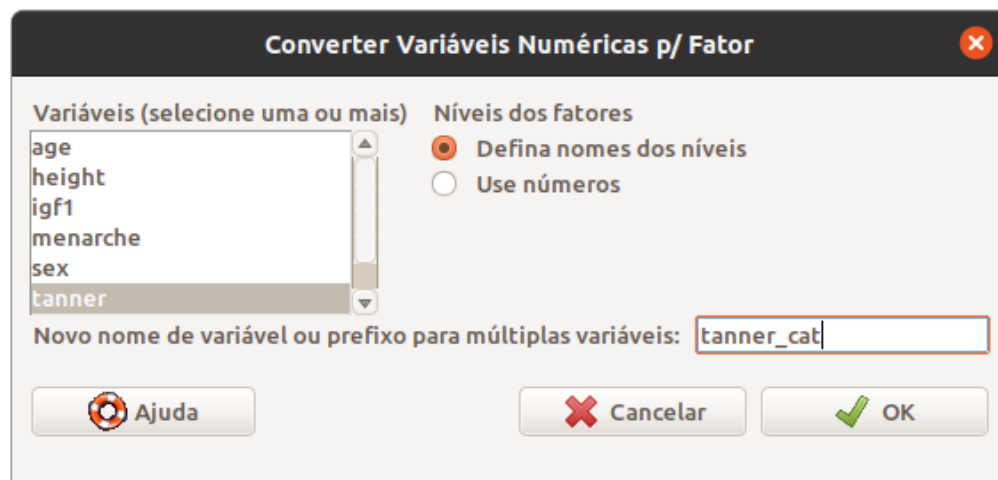


Figura 4.9: Passos para criar as categorias de uma variável: selecionamos a variável na lista da esquerda, escolhemos se as categorias serão dadas como texto e fornecemos o nome da nova variável. Clicamos em OK.

Como selecionamos a opção de fornecer os nomes para as categorias, ao clicarmos em OK na figura 4.9, uma nova tela aparece para darmos os nomes das categorias para cada valor numérico (figura 4.10). Finalmente, ao clicarmos em OK, a nova variável, *tanner\_cat*, será criada com as categorias apropriadas.

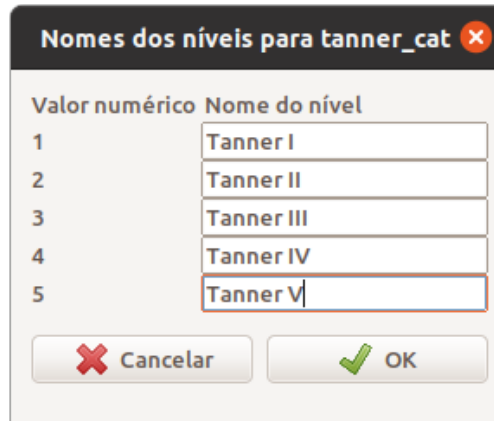


Figura 4.10: Especificação das categorias para a variável *tanner*.

O comando executado é mostrado a seguir:

```
juul2 <- within(juul2, {
  tanner_cat <- factor(tanner,
    labels=c('Tanner I', 'Tanner II', 'Tanner III', 'Tanner IV', 'Tanner V'))
})
```

Vamos entender como essa conversão foi efetuada. A função *factor* converte a variável expressa pelo primeiro argumento da função em fator. O argumento *labels* define os rótulos para cada valor da variável que será convertida. A função *c* cria um vetor com os rótulos que serão atribuídos aos números 1 a 5, na sequência.

A função *within* nesse exemplo possui dois argumentos: o conjunto de dados *juul2* e a expressão que será executada no primeiro argumento. Essa expressão aparece entre `{}`. Assim a função *factor* vai operar sobre variáveis do conjunto de dados *juul2* e, após a conversão, a variável resultante (*tanner\_cat*) será incorporada ao próprio *juul2*. Poderíamos criar um outro conjunto de dados, bastando mudar o nome do objeto antes do sinal de atribuição (`<-`).

Observem os registros do conjunto de dados após a recodificação e a classe da variável *tanner\_cat*, usando a função *class* conforme a seguir:

```
class(juul2$tanner_cat)
```

```
## [1] "factor"
```

Repitam o procedimento acima para a variável *sex*, criando uma nova variável, *sexo\_cat*, com a conversão abaixo:

- 1 - *masculino*
- 2 - *feminino*



O comando para a conversão da variável *sex* para fator é mostrado a seguir:

```
juul2 <- within(juul2, {  
  sexo_cat <- factor(sex, labels=c('masculino','feminino'))  
})
```

Repitam o procedimento acima para a variável *menarche*, criando uma nova variável, *menarca\_cat*, com a conversão abaixo:

- 1 - *não*
- 2 - *sim*

O comando para a conversão da variável *menarche* é mostrado a seguir:

```
juul2 <- within(juul2, {  
  menarca_cat <- factor(menarche, labels=c('não','sim'))  
})
```

## 4.2 Diagrama de barras

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

As variáveis categóricas nominais ou ordinais podem ser visualizadas graficamente por diagramas que fornecem as contagem (frequência) ou a proporção (ou porcentagem) de cada categoria da variável no conjunto de dados. Um dos diagramas mais utilizados com essa finalidade é o diagrama de barras, o qual mostra para cada categoria uma barra com altura proporcional à frequência ou proporção da respectiva categoria. Para criar um diagrama de barras no *R Commander*, selecionamos a opção:

Gráficos  $\Rightarrow$  Gráfico de barras

Na aba *Dados* da caixa de diálogo *Gráfico de Barras*, é possível selecionar a variável categórica desejada. Nesse exemplo, vamos criar um diagrama de barras para a variável categórica *tanner\_cat* (figura 4.11). Na aba *Opções* dessa caixa de diálogo (figura 4.12), é possível especificar a escala do eixo Y (porcentagem ou frequência), selecionar a cor e a posição das legendas, especificar as legendas dos eixos X e Y e o título do gráfico e selecionar se as frequências ou porcentagens serão exibidas nas barras. As demais opções serão discutidas logo adiante. Ao clicarmos em OK, o gráfico será exibido na aba *Plots* do *RStudio* (figura 4.13), mostrando a frequência de cada barra. Se utilizarmos o *R Commander*, carregado a partir do R, o gráfico será exibido em uma outra janela ou na janela do R.

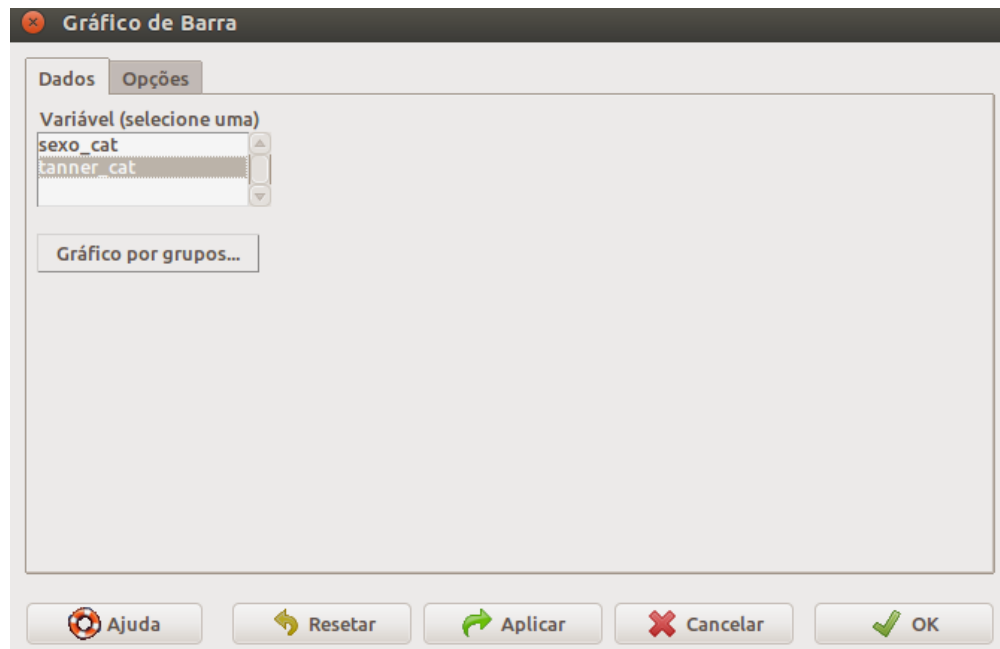


Figura 4.11: Caixa de diálogo para a geração de um diagrama de barras: selecionando a variável.



Figura 4.12: Caixa de diálogo para a geração de um diagrama de barras: especificando o título do gráfico e as legendas dos eixos X e Y.

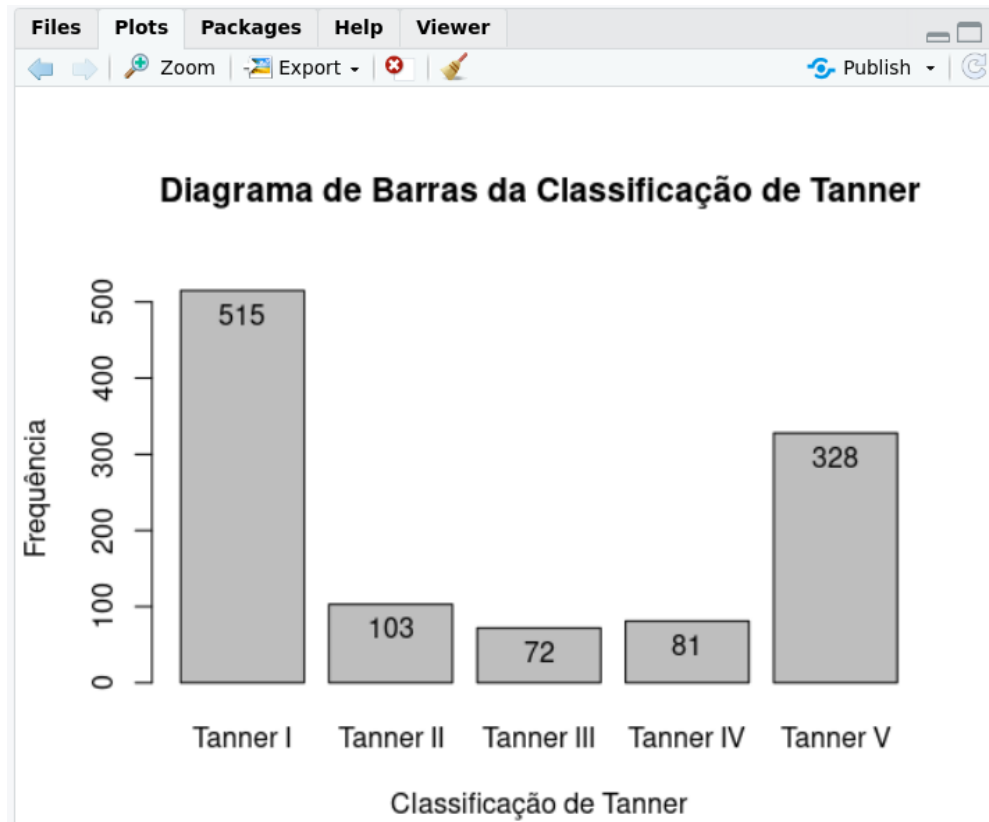


Figura 4.13: Diagrama de barras para a variável *tanner\_cat*. São mostradas as frequências de cada categoria de Tanner.

O gráfico da figura 4.13 mostra as frequências de cada uma das cinco categorias da classificação de Tanner no conjunto de dados *juul2*. A categoria I é a mais frequente, seguida da categoria V. As categorias II, III e IV apresentam frequências próximas umas das outras, mas com frequências bem menores do que as categorias I e V.

Caso desejemos visualizar o diagrama de barras da variável *sexo\_cat* separadamente para cada categoria de Tanner, precisamos selecionar *sexo\_cat* como uma variável de agrupamento. Para isso, clicamos na opção *Gráfico por grupos* na figura 4.11. Seremos, então, apresentados à caixa de diálogo da figura 4.14, onde selecionamos a variável de agrupamento (*sexo\_cat*). Ao clicarmos em OK, voltamos à tela da figura 4.11.

Clicando na aba *Opções*, mostrada novamente na figura 4.15, podemos, em *Estilo de barras agrupadas*, escolher entre duas opções de como o diagrama de barras será construído: barras de cada categoria da variável *sexo\_cat* empilhadas (particionada) para cada categoria da variável *tanner\_cat*, ou lado a lado. Selecionando a segunda opção e clicando em OK, será plotado o gráfico da figura 4.16. Ao selecionarmos a primeira opção, obteremos o gráfico da figura 4.17, onde desmarcamos a opção *Show counts or percentages in bars* na figura 4.15.

Ambos os gráficos mostram que há uma predominância de homens nas categorias I, II e IV (especialmente na I), e mais mulheres nas categorias III e V (principalmente na V). Para cada sexo, as frequências de cada categoria de Tanner seguem o padrão mostrado na figura 4.13.

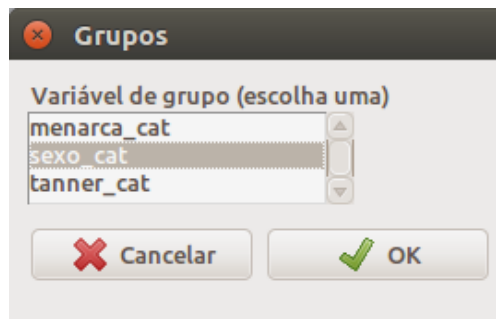


Figura 4.14: Selecionando uma variável de agrupamento para o diagrama de barras das frequências das categorias da variável *sexo\_cat* para cada categoria da classificação de Tanner.



Figura 4.15: Selecionando a forma como as barras serão apresentadas (lado a lado ou empilhadas). Neste exemplo, foi selecionada a opção lado a lado.

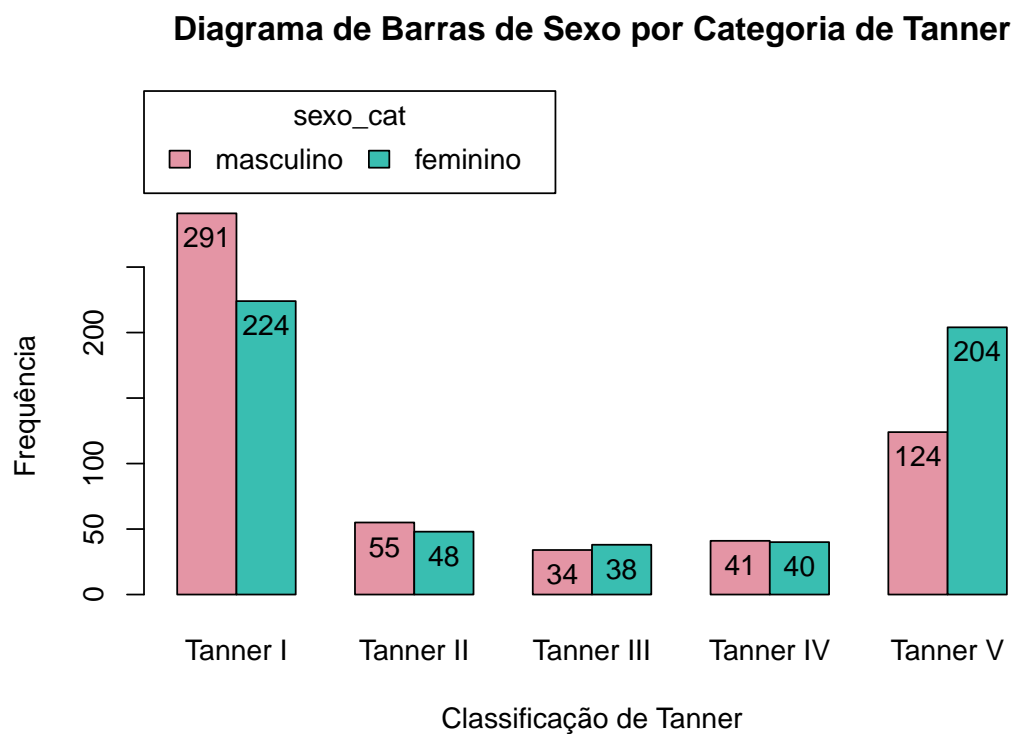


Figura 4.16: Diagrama de barras lado a lado das frequências das categorias da variável *sexo\_cat* para cada categoria da variável *tanner\_cat*.

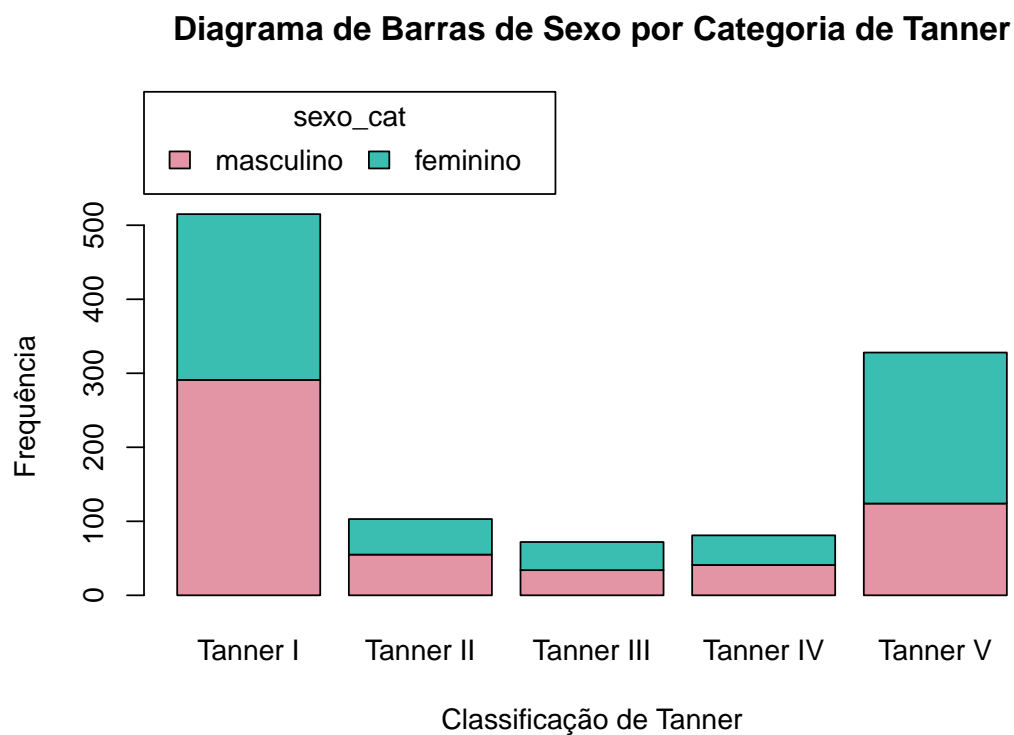


Figura 4.17: Diagrama de barras empilhadas das frequências das categorias da variável *sexo\_cat* para cada categoria da variável *tanner\_cat*.

Vamos supor que queiramos criar um diagrama de barras com percentuais de cada categoria, em vez de frequências. Na caixa de diálogo de opções do comando para a geração de um diagrama de barras, selecionamos as opções como mostrado na figura 4.18. Observem que foram selecionadas as opções *Percentagens* em *Escala do eixo*, e *Total* em *Percentages for Group Bars*.



Figura 4.18: Configuração para gerar um diagrama de barras com percentagens do total para as categorias da variável *sexo\_cat* para cada categoria da variável *tanner\_cat*.

O comando gerado pelo *R Commander* é mostrado abaixo e o gráfico é apresentado na figura 4.19. Para cada categoria de Tanner, as barras mostram o percentual do total de observações para cada sexo naquela categoria. A soma de todos os percentuais é igual a 100%.

```
with(juul2, Barplot(tanner_cat, by=sexo_cat, style="parallel",
  legend.pos="above", xlab="Classificação de Tanner",
  ylab="Porcentagens", conditional=FALSE, label.bars=TRUE,
  main="Diagrama de Barras de Sexo por Categoria de Tanner",
  scale="percent"))
```

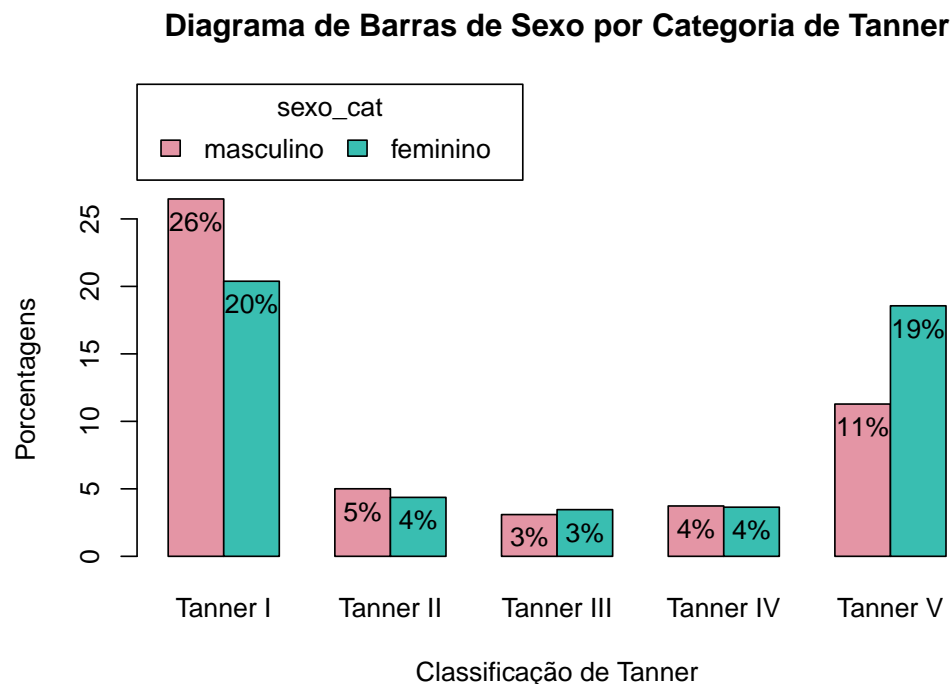


Figura 4.19: Diagrama de barras lado a lado para as porcentagens do total de observações de cada categoria de *sexo\_cat* por categoria da variável *tanner\_cat*.

Caso tivéssemos selecionado a opção *Conditional* em *Percentages for Group Bars* na figura 4.18, obteríamos o diagrama da figura 4.20. Em cada categoria de Tanner, os percentuais de cada sexo somam 100%.

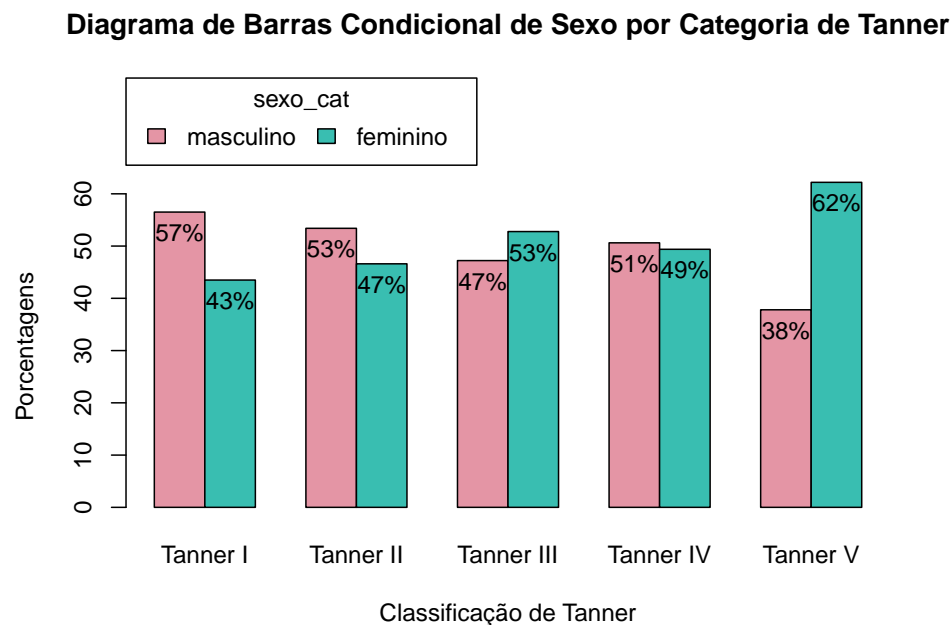


Figura 4.20: Diagrama de barras lado a lado para as porcentagens das categorias da variável *sexo\_cat* em cada categoria da variável *tanner\_cat*.

## 4.3 Usando a linha de comando

O conteúdo das seções 4.3.2, 4.3.3, 4.3.4, 4.3.5 e 4.3.6 podem ser visualizados neste [vídeo](#).

Vamos aproveitar o comando anterior e verificar como podemos alterar diversos outros aspectos do gráfico, alguns dos quais não podem ser configurados via menu do *R Commander*. Alguns argumentos apresentados nas funções a seguir são comuns a diversos tipos de gráficos, outros são específicos do diagrama de barras.

### 4.3.1 Especificação dos rótulos dos eixos x e y e do título

Os argumentos *xlab*, *ylab* e *main* descrevem respectivamente os rótulos dos eixos x, y e título dos gráficos. Eles são comuns a todos os gráficos exibidos pelo *R Commander*.

### 4.3.2 Alteração dos tamanhos dos eixos X e Y

O gráfico da figura 4.19 mostra a altura de uma barra maior do que o tamanho do eixo Y. O argumento *ylim* é usado para alterar os limites do eixo Y. Ele é especificado como um vetor de dois elementos, onde o primeiro elemento representa o limite inferior e o segundo elemento representa o limite superior do eixo.

No comando a seguir, é acrescentado o argumento *ylim = c(0, 30)* à função *Barplot*, e as alturas das barras ficarão menor do que o valor máximo do eixo Y (figura 4.21). Um argumento análogo, *xlim*, seria usado para o eixo X.

```
with(juul2, Barplot(tanner_cat, by=sexo_cat, style="parallel",
  legend.pos="above", xlab="Classificação de Tanner",
  ylab="Porcentagens", conditional=FALSE, label.bars=TRUE,
  main="Diagrama de Barras de Sexo por Categoria de Tanner",
  scale="percent", ylim = c(0, 30)))
```



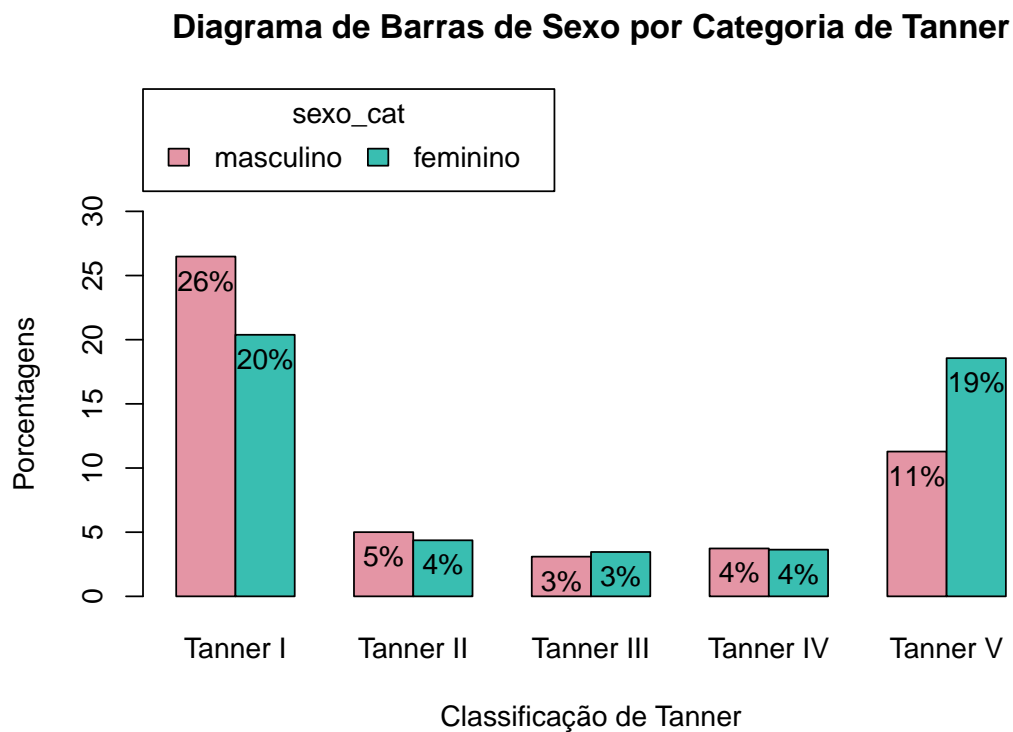


Figura 4.21: Diagrama de barras lado a lado para as porcentagens do total de observações de cada categoria de *sexo\_cat* por categoria da variável *tanner\_cat*, com a altura do eixo Y alterada para uma altura maior do que a altura da maior barra.

### 4.3.3 Alteração do título da legenda do diagrama

O argumento *legend.title* permite a alteração do título da legenda de um gráfico. O comando a seguir substitui o título original da legenda (“*sexo\_cat*”) do gráfico 4.21 por “Sexo”. O resultado é mostrado na figura 4.22.

```
with(juul2, Barplot(tanner_cat, by=sexo_cat, style="parallel",
  legend.pos="above", xlab="Classificação de Tanner",
  ylab="Porcentagens", conditional=FALSE, label.bars=TRUE,
  main="Diagrama de Barras de Sexo por Categoria de Tanner",
  scale="percent", ylim = c(0, 30), legend.title = "Sexo"))
```

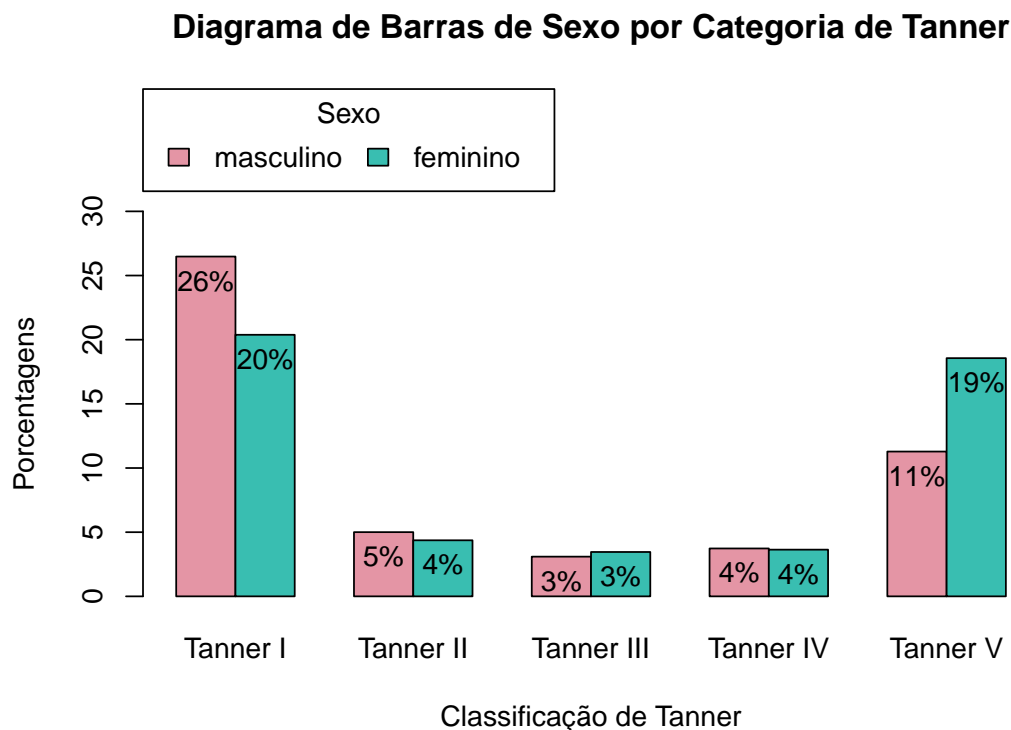


Figura 4.22: Diagrama de barras lado a lado para as porcentagens do total de observações de cada categoria de *sexo\_cat* por categoria da variável *tanner\_cat*, com a alteração do título da legenda.

#### 4.3.4 Alteração do espaçamento entre as barras

O argumento *space* é utilizado para especificar o espaçamento entre as barras. Indiretamente, ele define a largura das barras. O espaçamento é expresso como uma fração da largura das barras. *space* pode ser expresso por dois números, sendo o primeiro o espaçamento entre as barras de um mesmo grupo e o segundo o espaçamento entre os grupos. Se *space* não for especificado, os seguintes valores são assumidos por padrão:

- 1) se o diagrama de barras incluir mais de um fator e as barras forem desenhadas lado a lado, então  $space = c(0,1)$ , ou seja, o espaçamento entre os grupos de barras é igual à largura das mesmas;
- 2) para os demais casos  $space = 0.2$ , significando que o espaçamento entre as barras é igual a 20% da largura das barras.

O comando a seguir altera o espaçamento das barras para 20% da largura da barra dentro de cada grupo e para 1,5 vezes a largura da barra para a separação entre os grupos (figura 4.23).

```
with(juul2, Barplot(tanner_cat, by=sexo_cat, style="parallel",
  legend.pos="above", xlab="Classificação de Tanner",
  ylab="Porcentagens", conditional=FALSE, label.bars=TRUE,
  main="Diagrama de Barras de Sexo por Categoria de Tanner",
  scale="percent", ylim = c(0, 30), legend.title = "Sexo",
  space = c(.2, 1.5)))
```

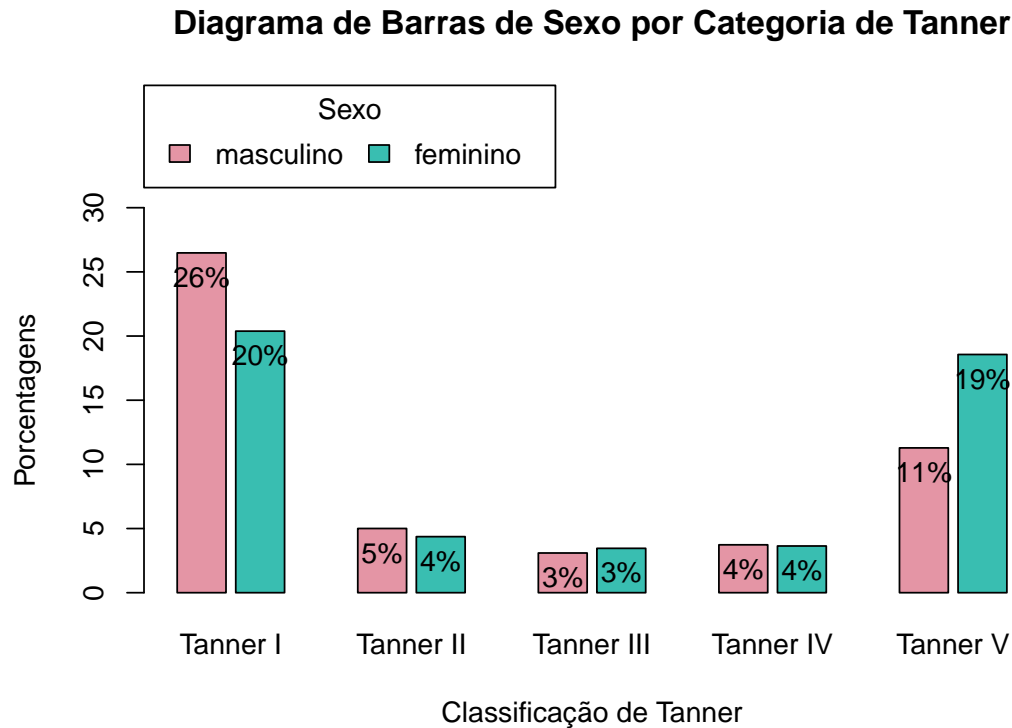


Figura 4.23: Diagrama de barras lado a lado para as porcentagens do total de observações de cada categoria de *sexo\_cat* por categoria da variável *tanner\_cat*, com a alteração do espaçamento entre as barras.

#### 4.3.5 Tamanhos dos rótulos dos eixos X e Y, dos números no eixo Y e das categorias das barras

Para alterar o tamanho dos rótulos que aparecem nos eixos X e Y, o tamanho das categorias da variável do eixo X e o tamanho dos números das escalas dos eixos X e Y, são usados os argumentos *cex.lab*, *cex.names*, e *cex.axis*, respectivamente, sendo o número 1 o padrão.

Ao fazermos os argumentos *cex.lab* = 1.8, *cex.names* = 1.2 e *cex.axis* = 1.5 na função *Barplot* (comando a seguir), o gráfico resultante mostra os rótulos dos eixos X e Y com tamanho 80% maior, as categorias de Tanner 20% maiores, e os números das escalas do eixo Y 50% maiores (figura 4.24).

```
with(juul2, Barplot(tanner_cat, by=sexo_cat, style="parallel",
  legend.pos="above", xlab="Classificação de Tanner",
  ylab="Porcentagens", conditional=FALSE, label.bars=TRUE,
  main="Diagrama de Barras de Sexo por Categoria de Tanner",
  scale="percent", ylim = c(0, 30), legend.title = "Sexo",
  space = c(.2, 1.5),
  cex.lab = 1.8, cex.names = 1.2, cex.axis = 1.5))
```

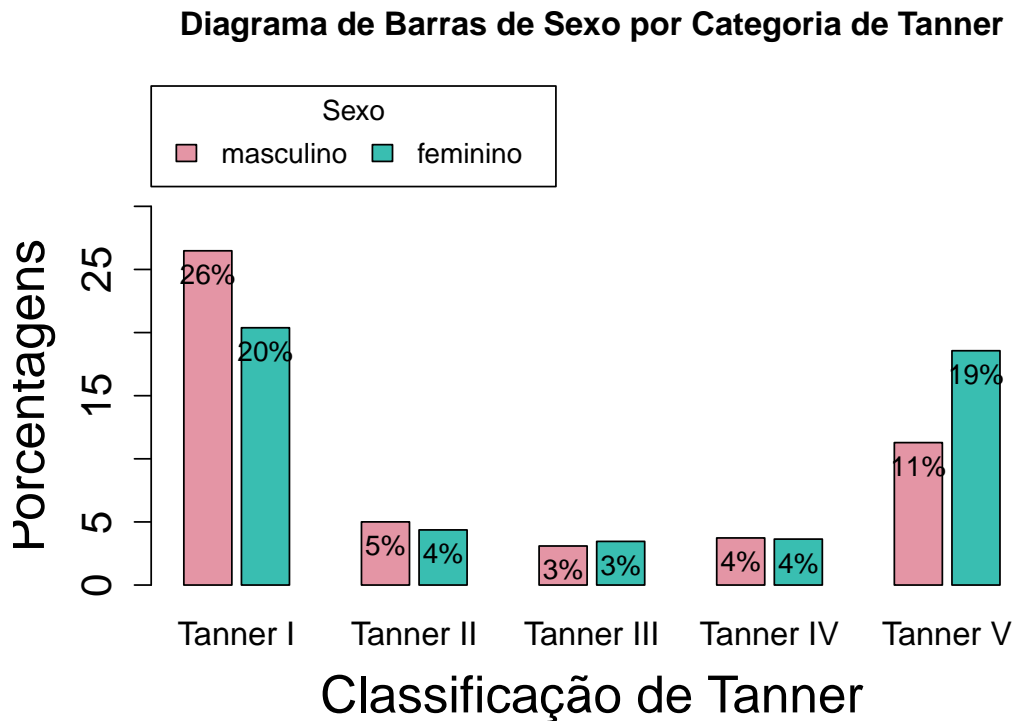


Figura 4.24: Diagrama de barras lado a lado para as porcentagens do total de observações de cada categoria de *sexo\_cat* por categoria da variável *tanner\_cat*, com alterações dos tamanhos dos rótulos dos eixos, das categorias de Tanner e da escala do eixo Y.

#### 4.3.6 Alteração das categorias da variável do eixo X

Os rótulos das categorias da variável do eixo X podem ser alterados por meio do argumento *names.arg*, o qual deve ser especificado como um vetor com um valor do tipo *character* para cada categoria, ou seja, cada categoria deve ser especificada entre aspas.

O comando a seguir altera os nomes das categorias de Tanner de I a V para 'T1', 'T2', 'T3', 'T4', e 'T5', respectivamente. O gráfico resultante é mostrado na figura 4.25.

```
with(juul2, Barplot(tanner_cat, by=sexo_cat, style="parallel",
  legend.pos="above", xlab="Classificação de Tanner",
  ylab="Porcentagens", conditional=FALSE, label.bars=TRUE,
  main="Diagrama de Barras de Sexo por Categoria de Tanner",
  scale="percent", ylim = c(0, 30), legend.title = "Sexo",
  space = c(.2, 1.5),
  cex.lab = 1.8, cex.names = 1.2, cex.axis = 1.5,
  names.arg = c('T1', 'T2', 'T3', 'T4', 'T5')))
```

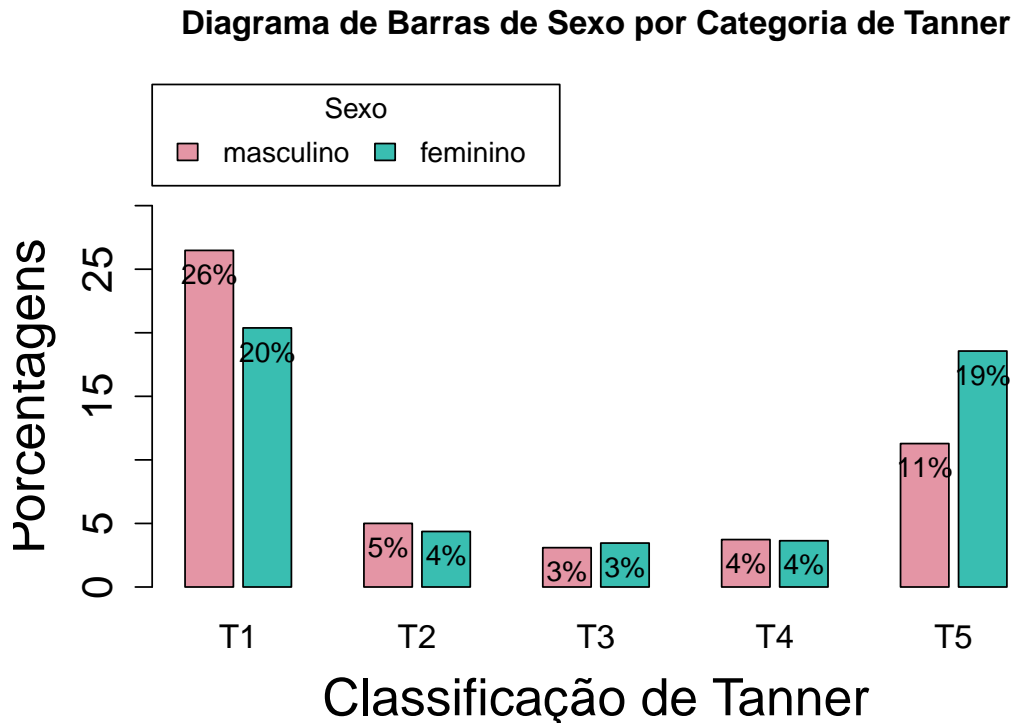


Figura 4.25: Diagrama de barras lado a lado para as porcentagens do total de observações de cada categoria de *sexo\_cat* por categoria da variável *tanner\_cat*, com a alteração dos nomes das categorias de Tanner.

### 4.3.7 Alteração das cores

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

O gráfico exibido na figura 4.13 mostra as barras na cor cinza. Já o gráfico da figura 4.19 mostra as barras nas cores rosa para o sexo masculino e verde para o feminino.

A função *Barplot*, por padrão, mostra as barras em cinza se um fator somente for especificado. Se dois fatores forem especificados, o segundo por meio do argumento *by*, as cores das barras são obtidas por meio da função *rainbow\_hcl* do pacote *colorspace*.

Como fazer para alterar as cores das barras? O argumento *col* é usado para alterar as cores. Vamos ver algumas possibilidades.

No comando a seguir, foi acrescentado o argumento `col` com o valor `c("blue", "orange")`, indicando as duas cores que serão utilizadas agora. Também foi acrescentado o argumento `ylim = c(0, 30)` para alterar os limites do eixo Y. Ao executarmos o comando, selecionando todas as linhas do comando e clicando no botão *Submeter*, obteremos o gráfico da figura 4.26.

```
with(juul2, Barplot(tanner_cat, by=sexo_cat, style="parallel",
  legend.pos="above", xlab="Classificação de Tanner",
  ylab="Porcentagens", conditional=FALSE, label.bars=TRUE,
  main="Diagrama de Barras de Sexo por Categoria de Tanner",
  scale="percent", ylim = c(0, 30), col = c("blue", "orange")))
```

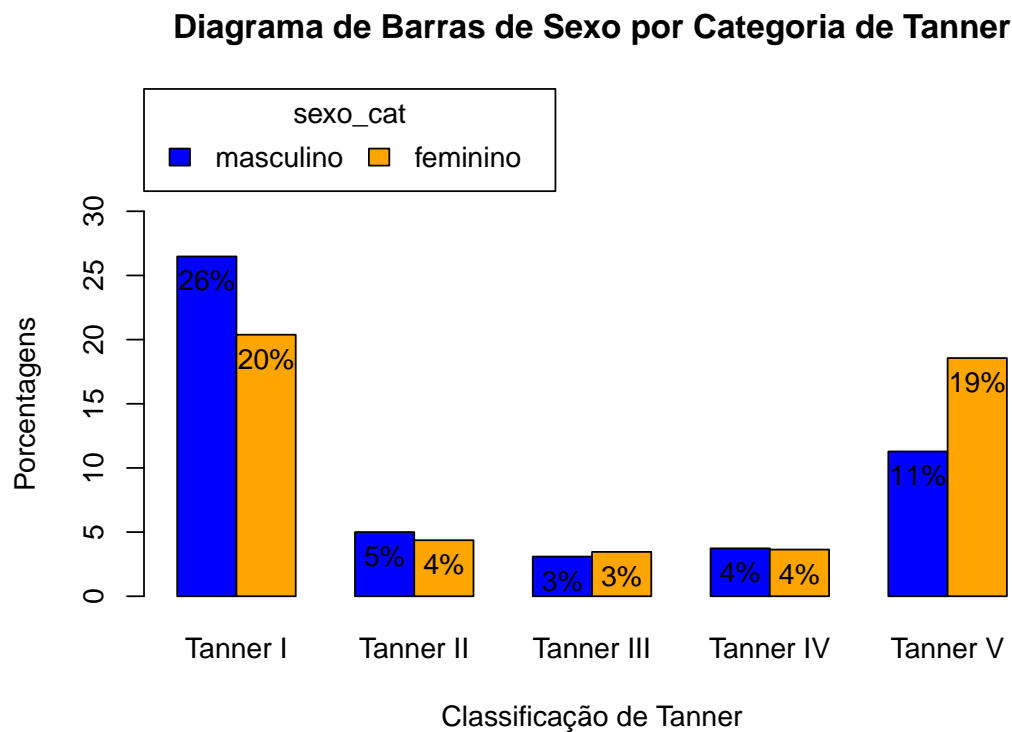


Figura 4.26: Diagrama de barras da figura 4.21, após a alteração das cores das barras.

Caso desejemos alterar as cores das barras em um diagrama de barras simples, sem variável de agrupamento, há diversas opções. Na aba de opções da figura 4.12, podemos marcar a opção *From color palette* no item *Color Selection*. O comando resultante é mostrado abaixo e o gráfico com barras vermelhas para cada categoria de Tanner é mostrado na figura 4.27.

```
with(juul2, Barplot(tanner_cat, xlab="Classificação de Tanner",
  ylab="Frequência", label.bars=TRUE,
  main="Diagrama de Barras da Classificação de Tanner",
  col=palette()[2])
)
```

### Diagrama de Barras da Classificação de Tanner

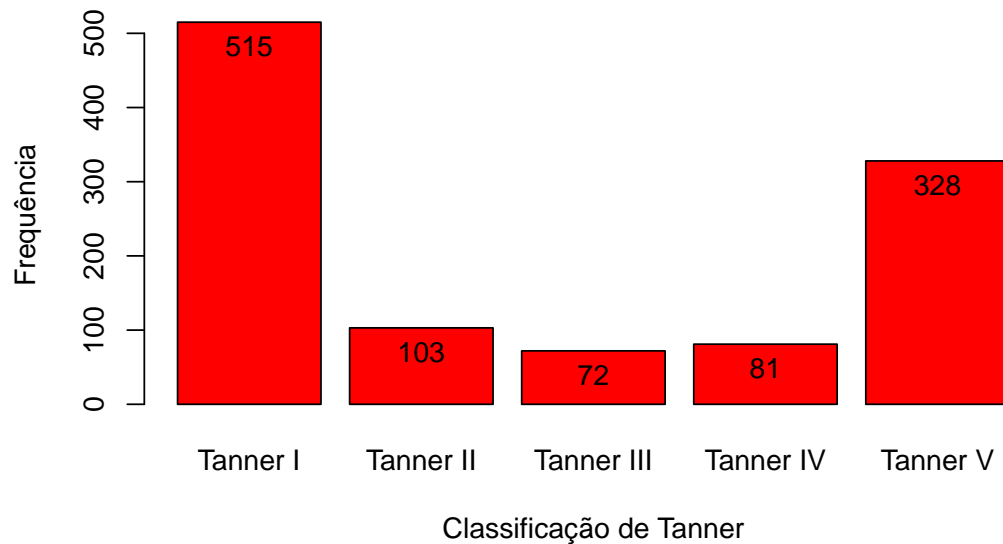


Figura 4.27: Diagrama de barras da figura 4.13, com as barras vermelhas.

A cor vermelha foi selecionada, porque o argumento *col* no comando anterior especificou a segunda cor da paleta de cores do *R Commander* (*palette()[2]*). A função *palette* retorna a paleta de cores do *R Commander* e o índice 2 indica a segunda cor desta paleta.

A figura 4.28 mostra a paleta de cores padrão do *R Commander*. Ela pode ser acessada pela opção do menu:

Gráficos  $\Rightarrow$  Gradiente de cores (color palette)



Figura 4.28: Paleta de cores do *R Commander*. Ao clicarmos na cor indicada pela seta verde, podemos substituí-la por outra.

Se quiséssemos a quarta cor da paleta nas barras, fariamos *col = palette()[4]* na chamada da função *Barplot*, ou simplesmente *col = 4*.

Podemos alterar cada cor dessa paleta, bastando clicar na cor que desejamos alterar. Se clicarmos na cor indicada pela seta verde na figura 4.28, poderemos alterá-la por meio da

caixa de diálogo *Select a color* (figura 4.29). Após selecionarmos a cor e clicarmos no botão OK, a paleta será alterada (figura 4.30).

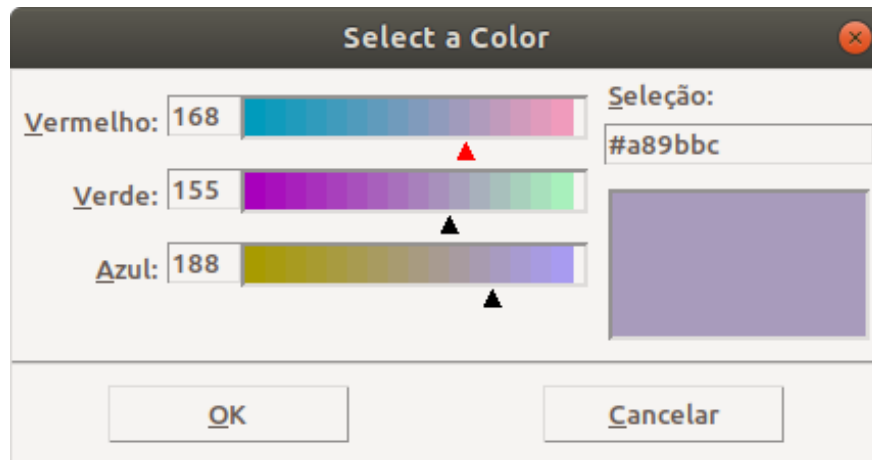


Figura 4.29: Caixa de diálogo para alterar uma cor da paleta.

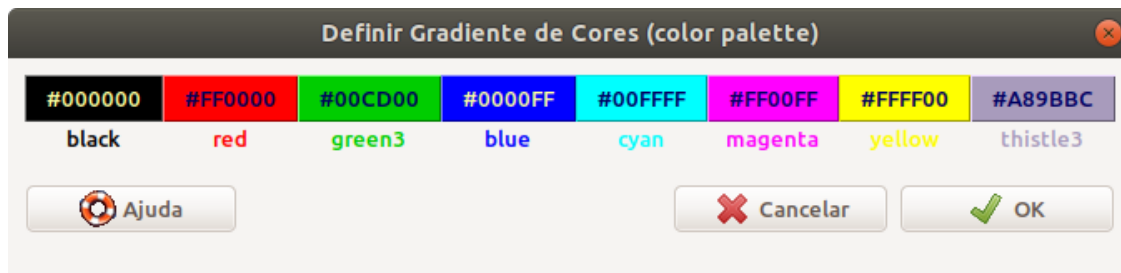


Figura 4.30: Paleta com a última cor alterada.

Se fizermos `col=c(2:6)` para gerar o diagrama de barras das categorias de Tanner, obteremos o comando a seguir, que gera barras com cores diferentes (figura 4.31).

```
with(juul2, Barplot(tanner_cat, xlab="Classificação de Tanner",
                    ylab="Frequência", label.bars=TRUE,
                    main="Diagrama de Barras da Classificação de Tanner",
                    col=c(2:6)))
```



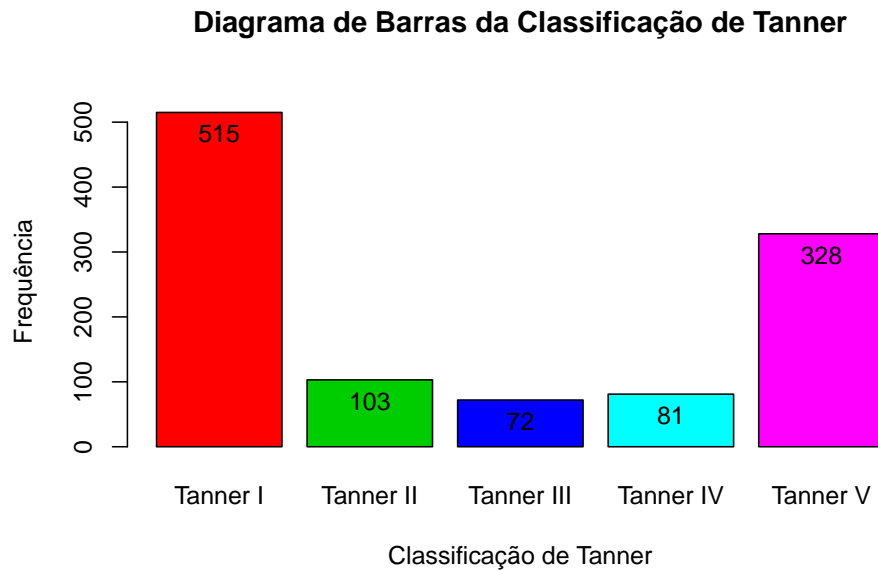


Figura 4.31: Diagrama de barras da figura 4.13, com barras de cores diferentes.

Observem que as cores foram especificadas como  $c(2:6)$ , significando que foram utilizadas as cores de números 2 a 6 da paleta corrente. Cada barra será de uma cor. Experimentem executar o comando. Outra possibilidade seria especificarmos as cores pelos nomes como mostrado a seguir (figura 4.32).

```
with(juul2, Barplot(tanner_cat, xlab="Classificação de Tanner",
  ylab="Frequência", label.bars=TRUE,
  main="Diagrama de Barras da Classificação de Tanner",
  col=c("red", "green", "blue", "cyan", "yellow")))
```

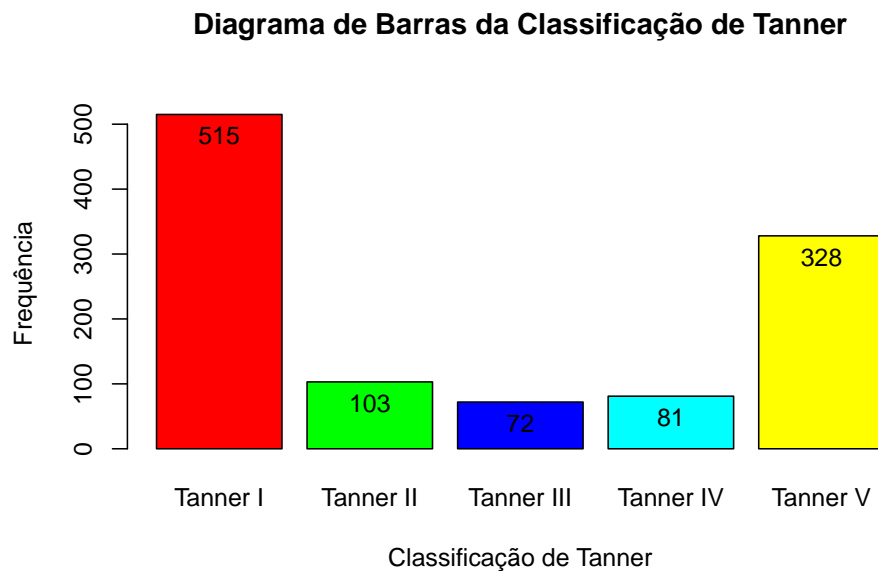


Figura 4.32: Diagrama de barras da figura 4.13, com barras de cores diferentes, especificadas pelo nome.

**Observação:** As interfaces gráficas não oferecem todos os recursos de cada função gráfica. Para conhecermos os argumentos disponíveis para cada função, usamos `help(nome_da_função)`. A internet é uma excelente fonte de ajuda para entender como conseguir os efeitos desejados.

### 4.3.8 Gráfico de barras horizontais

Para plotar as barras na direção horizontal, usa-se o argumento `horiz = TRUE`.

## 4.4 Diagrama de setores, torta ou pizza

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

O diagrama de setores também é utilizado para a visualização de variáveis categóricas. Nesse diagrama, um círculo é dividido em fatias, onde a área de cada fatia é proporcional à frequência de cada categoria da variável no conjunto de dados. Essa informação também é transmitida pelo diagrama de barras.

Para construirmos um diagrama de setores no *R Commander*, selecionamos a opção do menu:

Gráficos  $\Rightarrow$  Gráfico de Pizza

Em seguida, selecionamos a variável categórica para a qual o diagrama será construído, digitamos um título para o gráfico e clicamos em OK (figura 4.33). A figura 4.34 mostra o gráfico resultante.

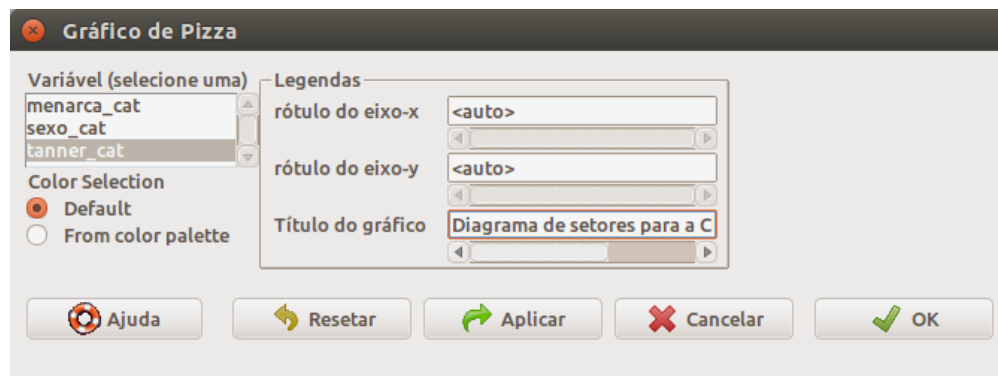


Figura 4.33: Caixa de diálogo para a geração de um diagrama de setores. Selecionamos a variável e digitamos o título do gráfico (opcional).

Observem que o gráfico da figura 4.34 não mostra os percentuais ou as frequências de cada fatia, apesar de dar uma ideia da frequência relativa (ou porcentagens) de cada categoria no conjunto de dados.

**Diagrama de Setores para a Classificação de Tanner**

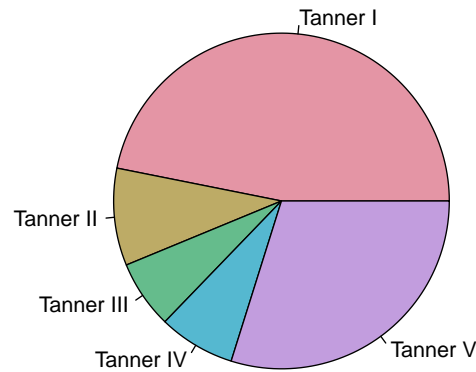


Figura 4.34: Diagrama de setores das categorias de Tanner.

Para obter um gráfico com as frequências (ou percentuais) de cada categoria, podemos utilizar a sequência de comandos a seguir, e o resultado é o gráfico da figura 4.35.

```
par(mar=c(1,1,1,1))
frequencias <- as.vector(with(juul2, table(tanner_cat)))
piepercent<- paste(as.character(
    round(100*frequencias/sum(frequencias), 1)), "%")
with(juul2, pie(table(tanner_cat), labels=piepercent,
    main="Classificação de Tanner",
    col=c(1:length(levels(tanner_cat)))))
legend(.9, .1, legend=c(levels(juul2$tanner_cat)), cex = 0.7,
    fill = c(1:length(levels(juul2$tanner_cat)) ))
```

**Classificação de Tanner**

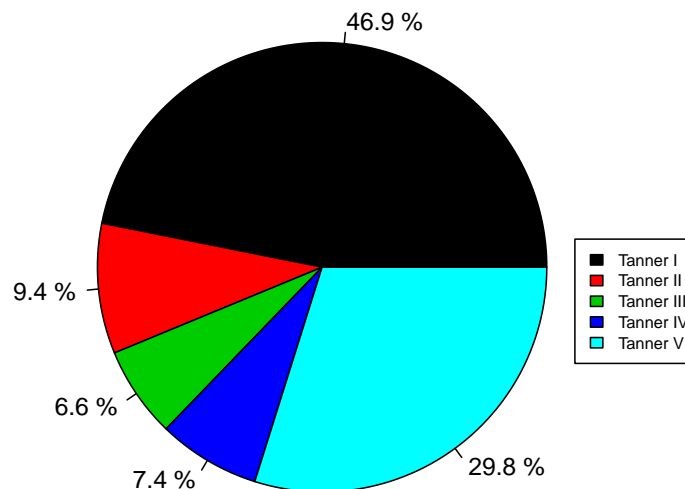


Figura 4.35: Diagrama de setores das categorias de Tanner, com os percentuais de cada categoria.

Vamos entender a sequência de comandos para gerar o gráfico de setores com porcentagens:

- 1) criamos um vetor com as frequências de cada categoria com o comando a seguir:

```
frequencias <- as.vector(with(juul2, table(tanner_cat)))
```

A função *table* cria uma tabela com a frequência de cada categoria de Tanner no conjunto de dados.

- 2) a partir dessas frequências, criamos um outro vetor (chamado *piepercent*) que fornece os percentuais de cada categoria, com o comando a seguir:

```
piepercent <- paste(as.character(round(100*frequencias/sum(frequencias), 1)), "%")
```

A função *round* arredonda o número especificado no primeiro argumento de acordo com o número de decimais especificado pelo segundo argumento. A função *as.character* transforma o número arredondado para *character* e *paste* concatena essa string com o sinal de porcentagem.

- 3) Então usamos o comando gerado pelo *R Commander* para a criação do diagrama e o modificamos conforme a seguir. Observem que alteramos o argumento *labels*, substituindo o seu valor por *piepercent*:

```
with(juul2, pie(table(tanner_cat), labels=piepercent, main="Classificação de Tanner",  
col=c(1:length(levels(tanner_cat)))))
```

- 4) Finalmente adicionamos uma legenda, indicando as cores de cada categoria com o comando a seguir:

```
legend(.9, .1, legend=c(levels(juul2$tanner_cat)), cex = 0.7,  
fill = c(1:length(levels(juul2$tanner_cat)) ))
```

Se for desejado criar um gráfico de pizza com frequências em vez de percentuais, basta alterar o comando do passo 3 acima, substituindo o valor do argumento *labels*, como a seguir. Lembrem-se de que o vetor *frequencias* foi obtido no passo 1. O gráfico é exibido na figura 4.36.

```
par(mar=c(1,1,1,1))  
with(juul2, pie(table(tanner_cat), labels=frequencias,  
main="Classificação de Tanner",  
col=c(1:length(levels(tanner_cat)) )))  
legend(.9, .1, legend=c(levels(juul2$tanner_cat)), cex = 0.7,  
fill = c(1:length(levels(juul2$tanner_cat)) ))
```

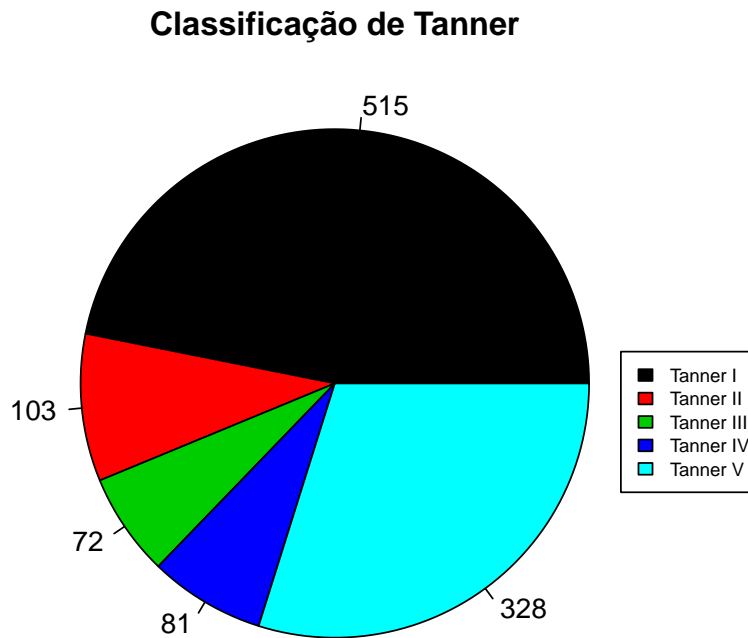


Figura 4.36: Diagrama de setores das categorias de Tanner, com as frequências de cada categoria.

A partir da próxima seção, serão mostrados recursos para a visualização da distribuição dos valores de variáveis numéricas.

## 4.5 Diagrama de caixa (*boxplot* ou *box and whisker plot*)

O conteúdo desta seção pode ser visualizado neste [vídeo](#), seguido deste [vídeo](#).

O diagrama de caixa (em inglês, *box and whisker plot*, ou simplesmente *boxplot*) é um dos mais úteis diagramas para visualizar a distribuição de dados numéricos. Para explicar como o mesmo é construído, vamos criar um diagrama de *boxplot*, selecionando a opção:

Gráficos  $\Rightarrow$  Boxplot

A figura 4.37 mostra a tela de configuração do *boxplot*. Na aba *Dados*, selecionamos a variável. Neste exemplo, selecionamos a variável *igf1* (fator de crescimento parecido com a insulina tipo 1). Na aba *Opções*, digitamos um título para o gráfico e marcamos a opção de não identificar os *outliers* (figura 4.38). O gráfico é mostrado na figura 4.39.



Figura 4.37: Caixa de diálogo para a geração do *boxplot*. Nesse exemplo, estamos selecionando a variável *igf1*.



Figura 4.38: Aba *Opções* da caixa de diálogo para a geração do *boxplot*.

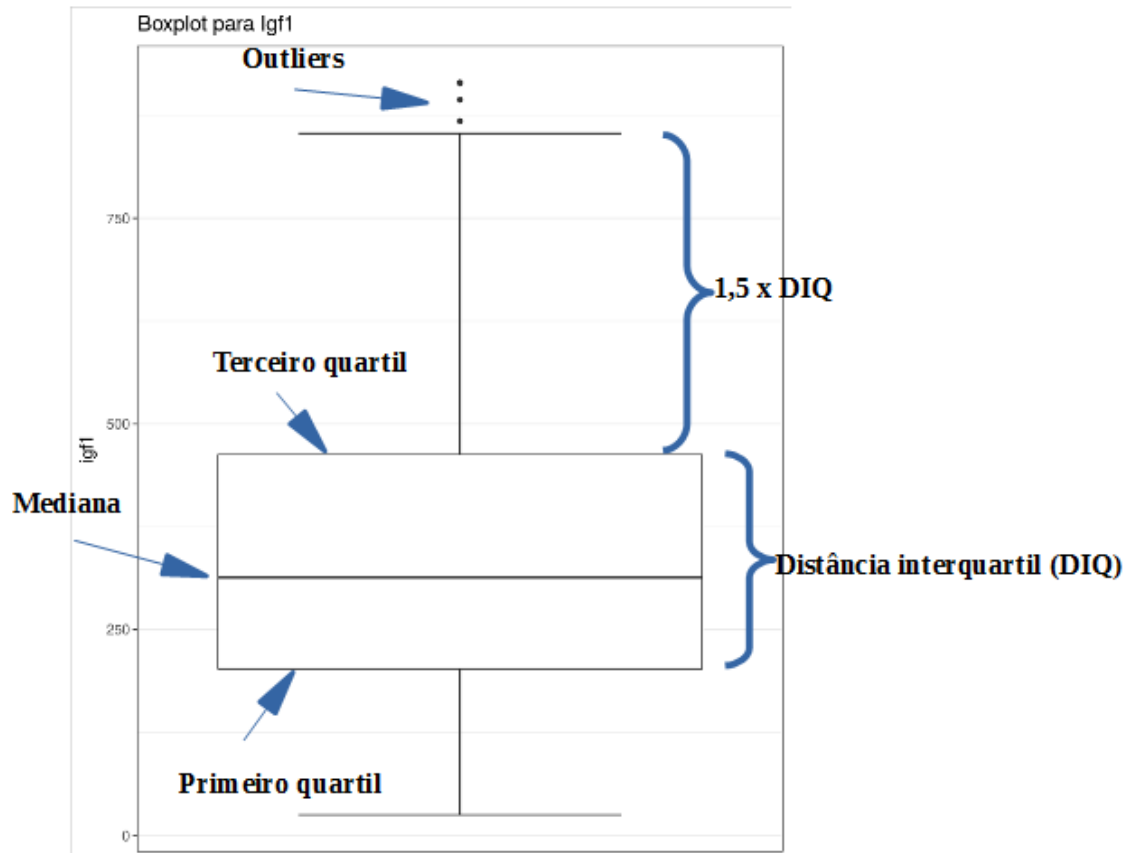


Figura 4.39: *Boxplot* da variável *igf1*.

O *boxplot* consiste de uma caixa cuja linha inferior indica o valor do primeiro quartil da variável, e a linha superior indica o terceiro quartil. Logo a altura da caixa indica a distância interquartil (DIQ). Uma terceira linha horizontal, a mediana, divide a caixa em duas partes. Partindo do meio da linha superior da caixa, uma linha vertical (*whisker* = bigode) liga o terceiro quartil ao valor imediatamente inferior ou igual ao valor do terceiro quartil somado a  $1,5 \times \text{DIQ}$ . Valores acima do *whisker* são considerados *outliers* e indicados por pontos. De maneira semelhante, uma linha vertical parte do meio da linha inferior da caixa e liga o primeiro quartil ao valor imediatamente acima ou igual ao primeiro quartil subtraído de  $1,5 \times \text{DIQ}$ . Pontos inferiores a esse valor também seriam considerados *outliers* e representados por pontos. Nesse exemplo, o menor valor de *igf1* está a uma distância do primeiro quartil menor que  $1,5 \times \text{DIQ}$ . Por isso o *whisker* inferior não possui o mesmo tamanho do superior e não aparece *outliers* na porção inferior do diagrama.

O *boxplot* fornece diversas informações: os valores do primeiro e terceiro quartis, a mediana, a simetria ou assimetria dos dados e a presença de outliers. No diagrama da figura 4.39, verificamos uma certa assimetria dos dados de *igf1* e a presença de *outliers*.

É possível construir os *whiskers* com diferentes tamanhos do que aqui mostrado, assim como podem ser usados a média e desvio padrão para construir os limites da caixa. Porém o método apresentado acima é o mais utilizado e é o padrão na função *Boxplot*.

O *boxplot* da figura 4.39 mostra a distribuição de todos os valores de *igf1*. Podemos construir um *boxplot* de *igf1* para cada categoria da classificação de Tanner, ou por sexo, ou por cada combinação de sexo e classificação de Tanner.

Para mostrarmos o *boxplot* de *igf1* para cada categoria da classificação de Tanner, clicamos no botão *Gráfico por grupos...* na caixa de diálogo do *boxplot* (figura 4.37) e selecionamos a variável *tanner\_cat* para compor os grupos. Na aba *Opções* (figura 4.38), podemos digitar uma legenda para o eixo X. O resultado é mostrado na figura 4.40.

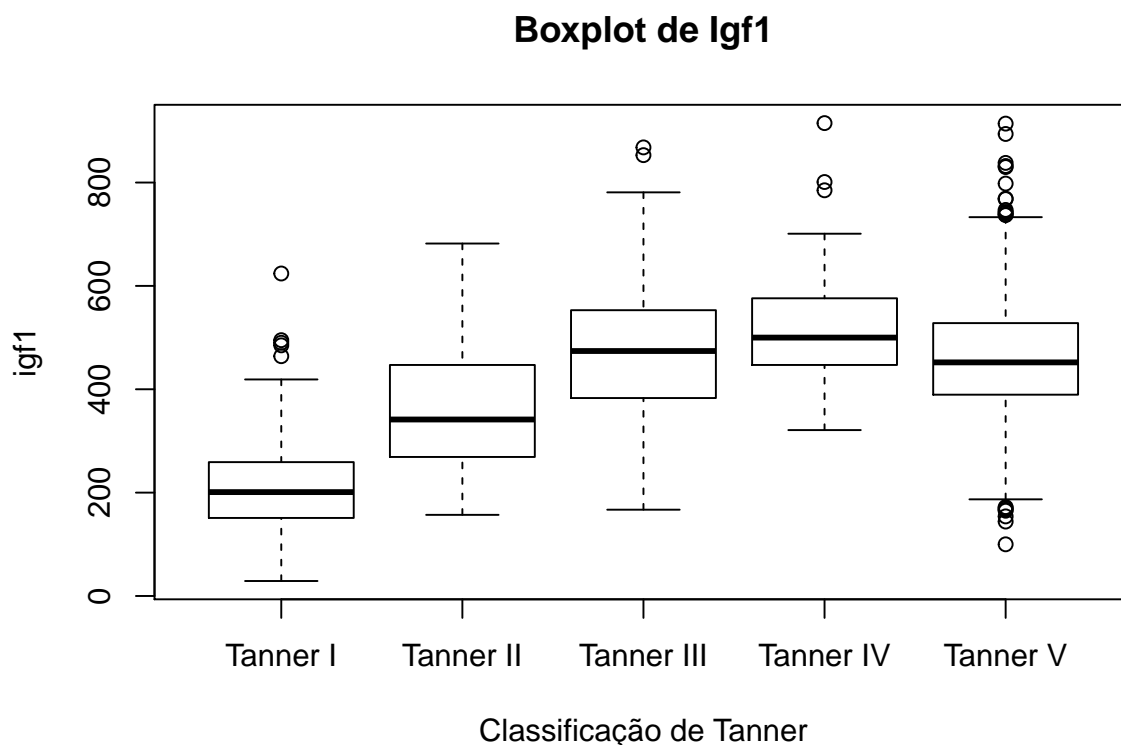


Figura 4.40: *Boxplots* para a variável *igf1* para cada categoria de Tanner.

Esse gráfico nos fornece uma visão melhor de como os valores de *igf1* estão distribuídos. À medida que as pessoas vão crescendo, os valores de *igf1* tendem a aumentar até o nível III da classificação de Tanner, tendendo a se estabilizarem a partir desta categoria. As distribuições dos valores de *igf1* tendem a ser simétricas nas categorias III e V de Tanner e ligeiramente assimétricas nas demais categorias.

**Observação:** Devemos ter cautela, porém, com as afirmações do parágrafo anterior, porque os dados de *igf1* foram coletados a partir de um estudo transversal. Um conjunto de pessoas foram selecionadas e os valores de idade e *igf1* foram coletados para cada uma delas. Um estudo mais apropriado para verificar a dependência de *igf1* com a idade seria um estudo longitudinal, onde um grupo de pessoas fosse acompanhado ao longo do tempo e os valores de *igf1* fossem medidos em diversos instantes para cada indivíduo à medida que ele ou ela fosse envelhecendo.



## 4.6 Histograma

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

Os histogramas, ao lado dos *boxplots*, são os gráficos mais utilizados para visualizarmos dados numéricos. Os histogramas são construídos, agrupando os valores numéricos da variável em faixas de valores e desenhando uma barra com largura igual ao tamanho da faixa e com altura, por exemplo, igual à frequência relativa de valores (percentual de valores) na faixa correspondente. As faixas são contíguas e o conjunto de barras compõem o histograma.

Para construirmos um histograma no *R Commander*, selecionamos a opção:

Gráficos  $\Rightarrow$  Histograma

Em seguida, selecionamos a variável desejada, *igf1* nesse exemplo (figura 4.41). Na aba *Opções* (figura 4.42), vamos selecionar *percentagens* em *Escala do eixo* e digitar a legenda do eixo Y. Ao clicarmos em OK, o gráfico resultante é mostrado na figura 4.43.

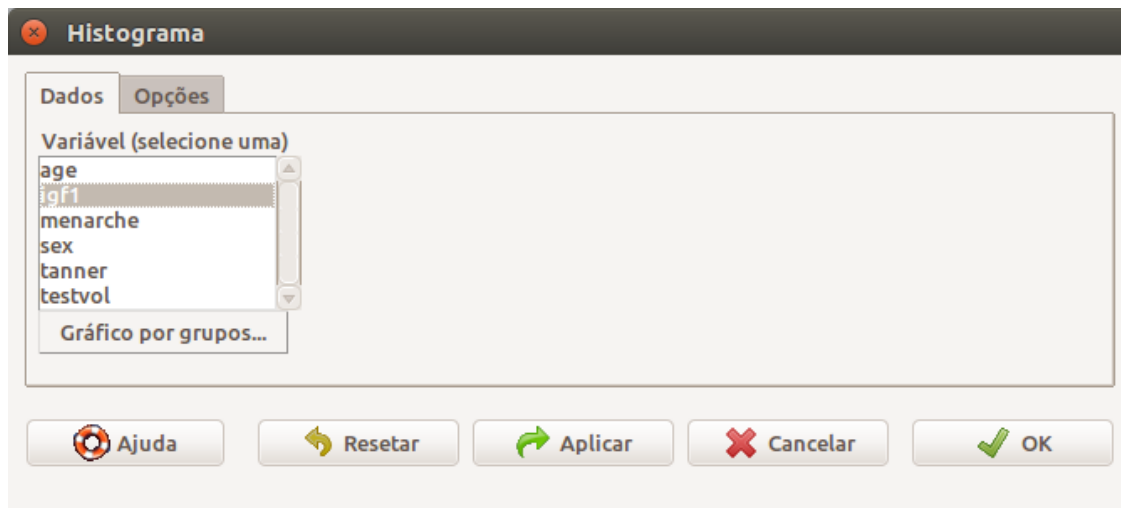


Figura 4.41: Caixa de diálogo para a criação de um histograma. Na aba *Dados*, selecionamos a variável numérica desejada.

**Histograma**

**Dados** **Opções**

**Opções gráficas**

Número de classes: <auto>

Escala do eixo

☐ Contagens de frequência

☒ Percentagens

☐ Densidades

**Legendas**

rótulo do eixo-x: <auto>

rótulo do eixo-y: Frequência relativa (%)

Título do gráfico: <auto>

Ajuda Resetar Aplicar Cancelar OK

Figura 4.42: Caixa de diálogo para a criação de um histograma. Na aba *Opções*, podemos especificar o número de faixas de valores (classes), a escala do eixo e as legendas.

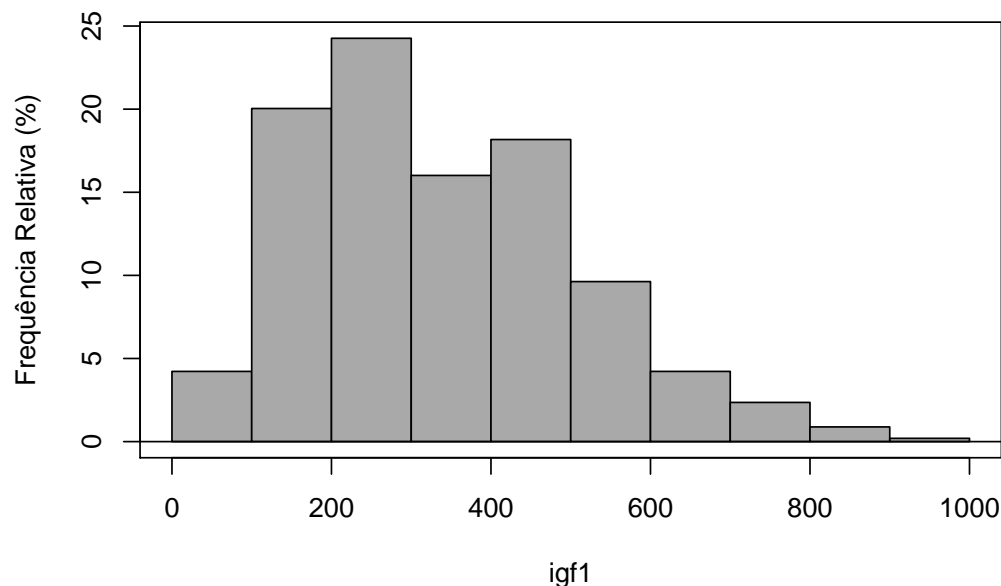


Figura 4.43: Histograma de frequência relativa da variável *igf1*.

Os passos para a construção de um histograma são:

- 1) definir um conjunto exaustivo de faixas de valores para a variável em questão. Cada faixa de valores é usualmente denominada classe. No exemplo acima, os valores de *igf1* foram distribuídos em 10 classes de amplitude 100 cada uma;
- 2) para cada classe, calcular a frequência dos valores nela contidas;
- 3) no caso de um histograma de frequência relativa, dividir a frequência de cada classe pelo número total de valores;
- 4) plotar uma barra para cada classe, com altura igual à frequência relativa (ou a mesma multiplicada por 100 para mostrar os percentuais de cada classe).

No exemplo dado, os limites das classes são:  
0, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000

Os valores da primeira classe são aqueles que pertencem ao intervalo (0, 100]. Desse modo, todos os valores da primeira classe devem satisfazer a seguinte desigualdade:  $0 < x \leq 100$ . As demais classes devem ser interpretadas de maneira análoga. Essa é a forma padrão que a função *hist* do R monta as classes do histograma. A tabela 4.1 mostra, para cada classe, os seus limites, contagem de valores (frequência) e a correspondente frequência relativa em porcentagem.

Tabela 4.1: Definição das classes e frequência relativa de cada uma delas. Cada classe é definida por um intervalo da forma (a, b], onde a é o limite inferior e b o limite superior.

Classe	Limite Inferior (>)	Limite Superior (≤)	Frequência	Frequência Relativa (%)
1	0	100	43	4,22
2	100	200	204	20,04
3	200	300	247	24,26
4	300	400	163	16,01
5	400	500	185	18,17
6	500	600	98	9,63
7	600	700	43	4,22
8	700	800	24	2,36
9	800	900	9	0,88
10	900	1000	2	0,2
			1018	100

No exemplo apresentado, o número de classes utilizado para agrupar os valores de *igf1* foi 10, determinado automaticamente pela função de geração do histograma. Entretanto esse número pode ser escolhido pelo usuário: basta digitar o número de classes desejado na opção *Numero de classes* da figura 4.42. O número de classes não deve ser um número muito baixo, de modo que tenhamos uma visão muito grosseira da distribuição de dados, nem muito alto, de modo que cada classe tenha poucos valores. O número de classes deve fornecer uma boa ideia de como os dados estão distribuídos, geralmente um número entre 10 e 20.

O comando que foi utilizado para a criação do histograma da figura 4.43 é mostrado a seguir:

```
with(juul2, Hist(igf1, scale="percent", breaks="Sturges", col="darkgray",
                ylab="Frequência Relativa (%)))
```

O argumento *breaks* indica como as classes serão definidas. O valor nesse exemplo, *Sturges*, indica o nome de um algoritmo utilizado para calcular o número de classes do histograma. Outros nomes de algoritmos são *Scott* e *Freedman-Diaconis*. Não vamos entrar em detalhes desses algoritmos. Além deles, o usuário pode especificar o nome de uma função qualquer

que calcule o número de classes, ou fixar o número de classes, ou mesmo especificar os limites das classes. Nós veremos essa última opção mais adiante.

#### 4.6.1 Histograma de frequência x frequência relativa x densidade de frequência relativa

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

Na seção anterior, criamos um histograma de frequência relativa para a variável *igf1*. Porém um histograma também pode ser criado, partindo-se da frequência ou da densidade de frequência relativa. Quando a amplitude de cada classe é a mesma, a aparência dos histogramas é a mesma, variando apenas a escala do eixo vertical.

Vamos alterar as classes do histograma para a variável *igf1* de modo que as suas amplitudes sejam diferentes e vamos ver as diferenças entre os três tipos de histogramas. A tabela 4.2 mostra as classes, a **frequência**, a **frequência relativa** e a **densidade de frequência relativa** para cada classe. Esse último termo será explicado mais adiante. Observem que as classes 1 e 10 possuem amplitude igual a 100, a classe 11 possui amplitude igual a 400 e as demais classes possuem amplitude igual a 50.

Tabela 4.2: Definição das classes de um histograma e respectivas frequências, frequências relativas e densidade de frequência relativa para a variável *igf1* do conjunto de dados *juul2*.

Classe	Limite Inferior (>)	Limite Superior (≤)	Frequência	Frequência Relativa (%)	Densidade de Frequência Relativa (x 10 <sup>-3</sup> )
1	0	100	43	4,22	0,42
2	100	150	74	7,27	1,45
3	150	200	130	12,77	2,55
4	200	250	129	12,67	2,53
5	250	300	118	11,59	2,32
6	300	350	69	6,78	1,36
7	350	400	94	9,23	1,85
8	400	450	93	9,14	1,82
9	450	500	92	9,04	1,80
10	500	600	98	9,63	0,96
11	600	1000	78	7,66	0,19
			1018	100	

A figura 4.44 mostra diversos histogramas para a variável *igf1*. Essa figura foi construída com a sequência de comandos a seguir:

```

par(mfrow = c(2,2))
with(juul2, Hist(igf1, scale="percent", breaks="Sturges", col="darkgray",
  ylab="Frequência Relativa (%)))
text(-150, -80, labels="a)", pos = 1, xpd = T, cex = 1.5)
with(juul2, Hist(igf1, scale="frequency", freq = TRUE,
  breaks=c(0,100,150,200,250,300,350,400,450,500, 600, 1000),
  col="darkgray", ylab="Frequência"))
text(-150, -40, labels="b)", pos = 1, xpd = T, cex = 1.5)
with(juul2, Hist(igf1, scale="percent", freq = TRUE,
  breaks=c(0,100,150,200,250,300,350,400,450,500, 600, 1000),
  col="darkgray", ylab="Frequência Relativa(%)))
text(-150, -40, labels="c)", pos = 1, xpd = T, cex = 1.5)
with(juul2, Hist(igf1, scale="density",
  breaks=c(0,100,150,200,250,300,350,400,450,500, 600, 1000),
  col="darkgray", ylab="Densidade de Frequência Relativa"))
text(-150, -0.0010, labels="d)", pos = 1, xpd = T, cex = 1.5)

```

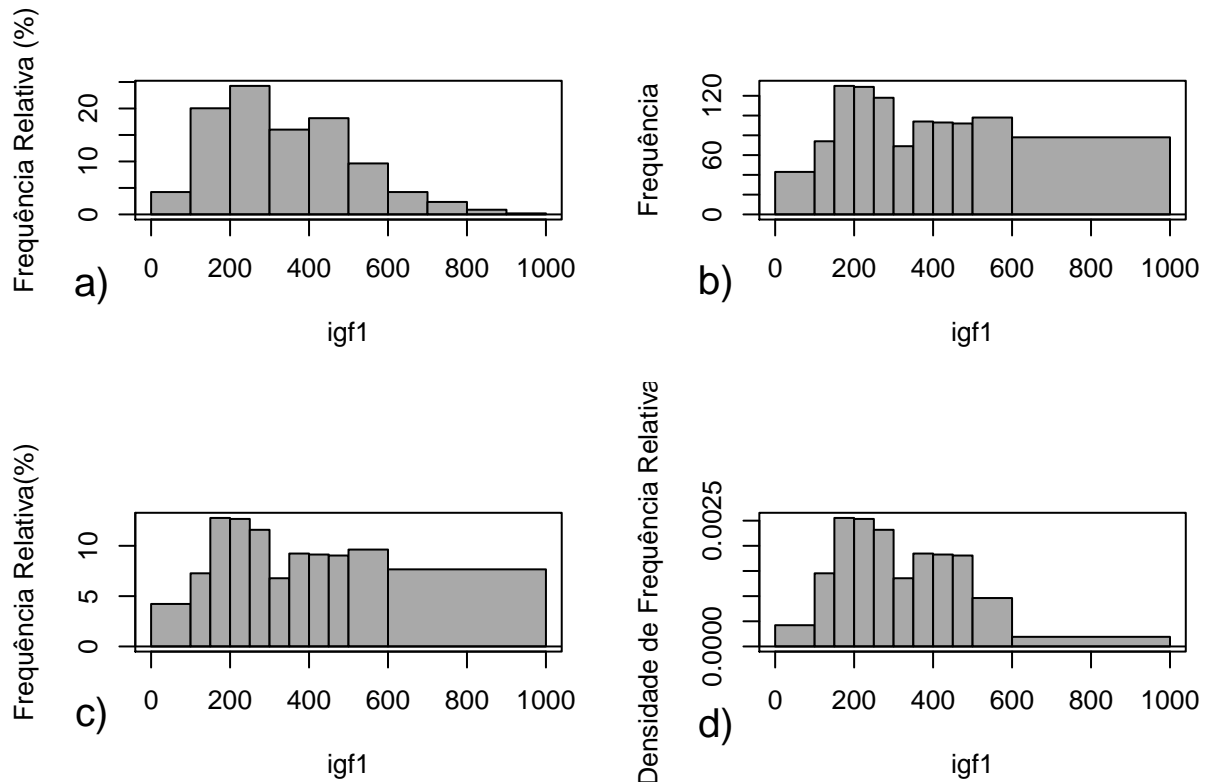


Figura 4.44: a) Histograma de frequência relativa da variável *igf1* para 10 classes com igual amplitude (igual ao da figura 4.43); b) histograma de frequência da variável *igf1* para as classes definidas conforme a tabela 4.2; c) histograma de frequência relativa da variável *igf1* para as classes definidas conforme a tabela 4.2; d) histograma de densidade de frequência relativa da variável *igf1* para as classes definidas conforme a tabela 4.2.

A função `par(mfrow = c(2,2))` indica que os gráficos serão exibidos em duas colunas sendo cada linha preenchida antes de avançar para a próxima.

O histograma da figura 4.44a é idêntico ao da figura 4.43 (10 classes de mesma amplitude). Repetimos a seguir o comando usado para gerá-lo:

```
with(juul2, Hist(igf1, scale="percent", breaks="Sturges", col="darkgray",  
                ylab="Frequência Relativa (%)))
```

O histograma da figura 4.44b é o histograma de frequência de *igf1* para as classes definidas na tabela 4.2 gerado pelo comando:

```
with(juul2, Hist(igf1, scale="frequency", freq = TRUE,  
                breaks=c(0,100,150,200,250,300,350,400,450,500, 600, 1000),  
                col="darkgray", ylab="Frequência"))
```

O histograma da figura 4.44c é o histograma de frequência relativa de *igf1* para as classes definidas na tabela 4.2 gerado pelo comando:

```
with(juul2, Hist(igf1, scale="percent", freq = TRUE,  
                breaks=c(0,100,150,200,250,300,350,400,450,500, 600, 1000),  
                col="darkgray", ylab="Frequência Relativa(%)))
```

No histograma de frequência, a altura de cada classe é igual ao número de valores nela contidos. No histograma de frequência relativa, a altura é a proporção de valores contidos na classe (contagem de valores / número total de valores), eventualmente multiplicada por 100 para ser expressa em porcentagem. Assim tanto o histograma de frequência quanto o de frequência relativa possuem a mesma forma, diferindo apenas na escala do eixo Y.

Quando a amplitude das classes são diferentes, porém, o histograma de frequência (e o de frequência relativa) dão uma visão distorcida da distribuição dos valores da variável. Observem que uma mensagem aparece na área de mensagens para esses dois histogramas com o seguinte teor:

*AVISO: Warning in plot.histogram(r, freq = freq1, col = col, border = border, angle = angle, the AREAS in the plot are wrong – rather use ‘freq = FALSE’*

**Essa mensagem indica que a distorção é causada porque cada classe de um histograma deve ter a altura tal que a área de cada classe (altura x amplitude) deve ser proporcional à frequência (ou frequência relativa) de cada uma delas.** Não é o que acontece nos histogramas das figuras 4.44b e 4.44c. Por exemplo, as áreas das classes 10 e 11 na figura 4.44b deveriam ser proporcionais a 78 (frequência da classe 11) e 98 (frequência da classe 1) respectivamente, ou seja, as áreas deveriam ser iguais, respectivamente, a 78 e 98, multiplicadas por uma constante qualquer. No entanto a área da classe 11 é igual a 400 x 78, e a área da classe 10 é 100 x 98, indicando que o número que multiplica a altura da classe 11 é 4 vezes maior do que o número que multiplica a altura da classe 10, distorcendo a distribuição dos valores da variável. Fato semelhante ocorre no histograma de frequência relativa (figura 4.44c). Para contornar esse problema, quando as classes possuem amplitude diferentes, utiliza-se o **conceito de densidade de frequência relativa**. Nesse caso, para

cada classe, divide-se a sua frequência relativa pela sua amplitude, obtendo-se os valores na última coluna da tabela 4.2.

Com o comando a seguir, obtemos o histograma de densidade de frequência relativa para a variável *igf1*, com as classes definidas na tabela 4.2 (figura 4.44d):

```
with(juul2, Hist(igf1, scale="density",  
                breaks=c(0,100,150,200,250,300,350,400,450,500, 600, 1000),  
                col="darkgray", ylab="Densidade de Frequência Relativa(%)))
```

Observem agora que a altura da classe 11 é relativamente bem menor do que a das demais classes, refletindo o fato de que os 78 valores dessa classe estão distribuídos em uma faixa maior de valores do que as demais classes.

A execução da função *par(mfrow = c(1,1))* a seguir indica que os gráficos voltarão a ser exibidos na forma normal.

```
par(mfrow = c(1,1)) # volta a exibir os gráficos da forma normal
```

## 4.6.2 Histograma por grupos

Da mesma forma que *boxplots*, é possível gerar histogramas de uma variável numérica para cada categoria de uma variável categórica. Por exemplo, para criarmos um histograma de *igf1* para cada categoria de Tanner, clicamos no botão *Gráfico por grupos* na figura 4.41 e selecionamos a variável *tanner\_cat* na caixa de diálogo *Grupos* (figura 4.45). Os histogramas de *igf1* para cada categoria de Tanner são mostrados na figura 4.46.

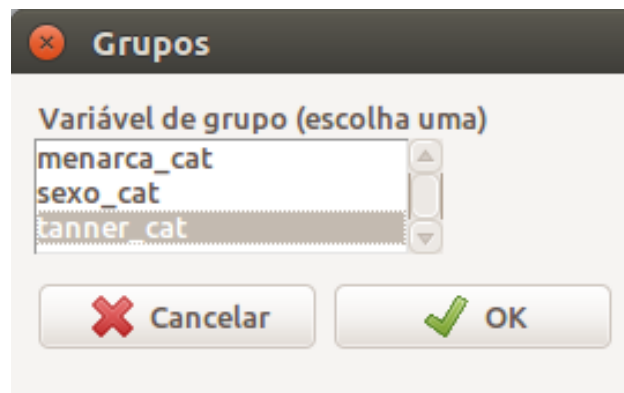


Figura 4.45: Selecionando uma variável de agrupamento para a construção de histogramas.

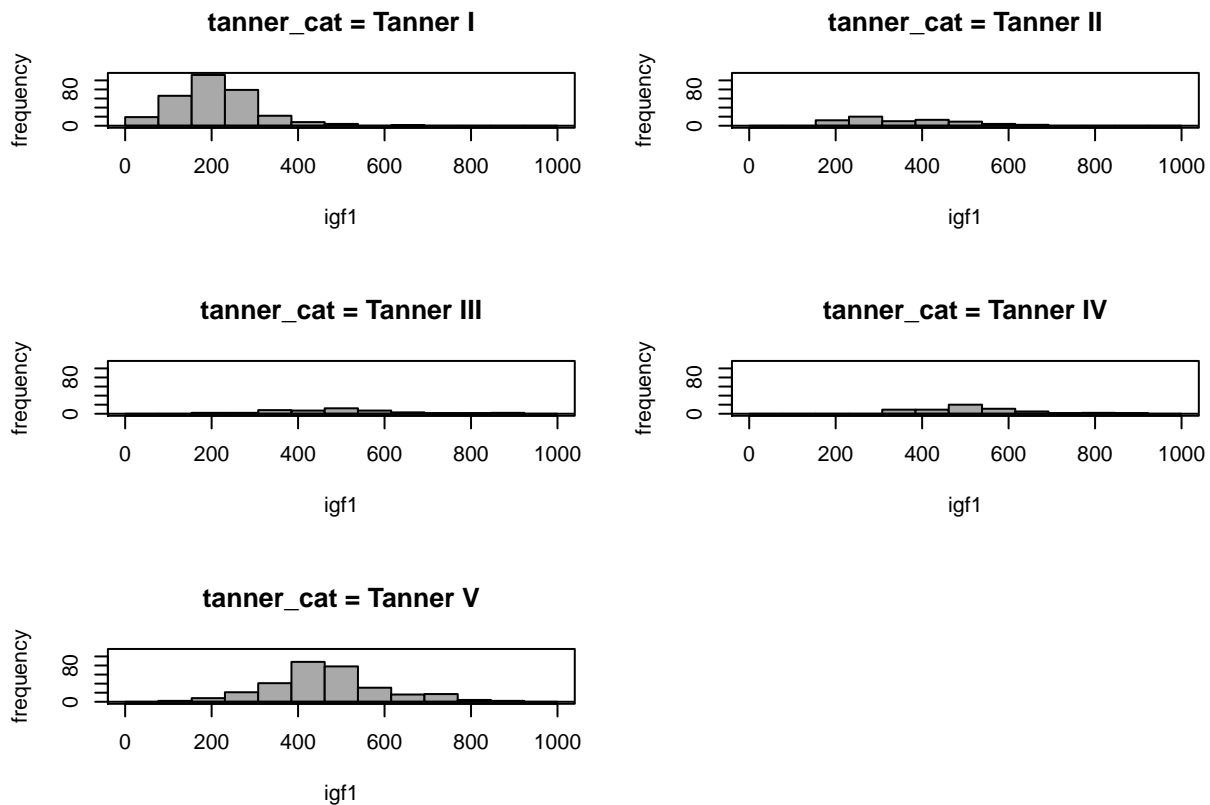


Figura 4.46: Histogramas de *igf1* para cada categoria de Tanner.

## 4.7 Diagrama de pontos e *strip chart*

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

O diagrama de pontos também fornece uma visão da distribuição dos valores de uma variável numérica. Ele pode ser apresentado de formas diferentes. Para criar um diagrama de pontos, acessamos como sempre o menu *Gráficos* e selecionamos a opção:

Gráficos  $\Rightarrow$  Diagrama de Pontos

Na caixa de diálogo do gráfico de pontos (figura 4.47), podemos selecionar a variável desejada e uma variável para criar um gráfico de pontos para cada categoria de outra variável (Gráfico por grupos). Nesse exemplo, selecionamos *igf1* como a variável desejada e *tanner\_cat* como a variável de agrupamento. O diagrama resultante é mostrado na figura 4.48.





Figura 4.47: Caixa de diálogo para a criação de um diagrama de pontos. Na aba *Dados*, selecionamos a variável numérica desejada. Ao clicarmos no botão *Gráfico por grupos*, podemos selecionar uma variável de agrupamento.

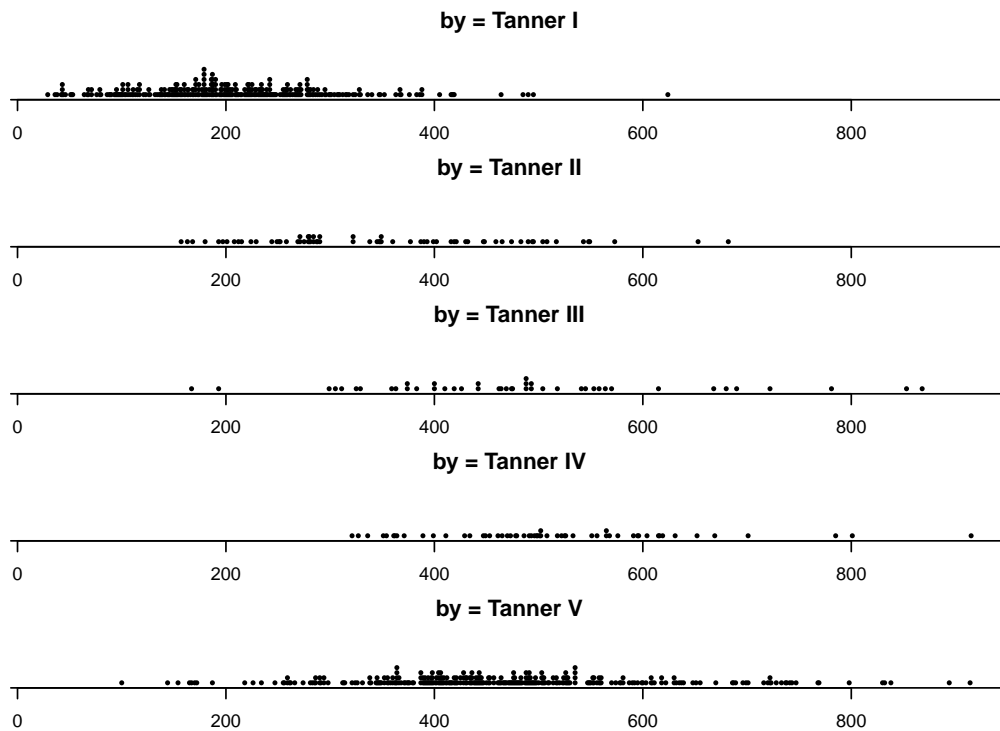


Figura 4.48: Diagrama de pontos de *igf1* para cada categoria de Tanner.

Uma forma alternativa de visualizar as distribuições de pontos é plotar os pontos ao longo de uma linha vertical para cada grupo. Isso pode ser feito por meio do diagrama de *strip chart*, que pode ser configurado com a opção:

Gráficos  $\Rightarrow$  Gráfico Strip Chart

Na caixa de diálogo da figura 4.49, selecionamos a variável numérica (variável resposta) e a(s) variável(is) de agrupamento (fatores), se desejado.



Figura 4.49: Caixa de diálogo para a criação de um diagrama de *strip chart*. Na aba *Dados*, selecionamos a variável numérica e a variável de agrupamento, se desejado.

O gráfico é mostrado na figura 4.50.

### Diagrama de Strip Chart de *igf1* por categorias de Tanner

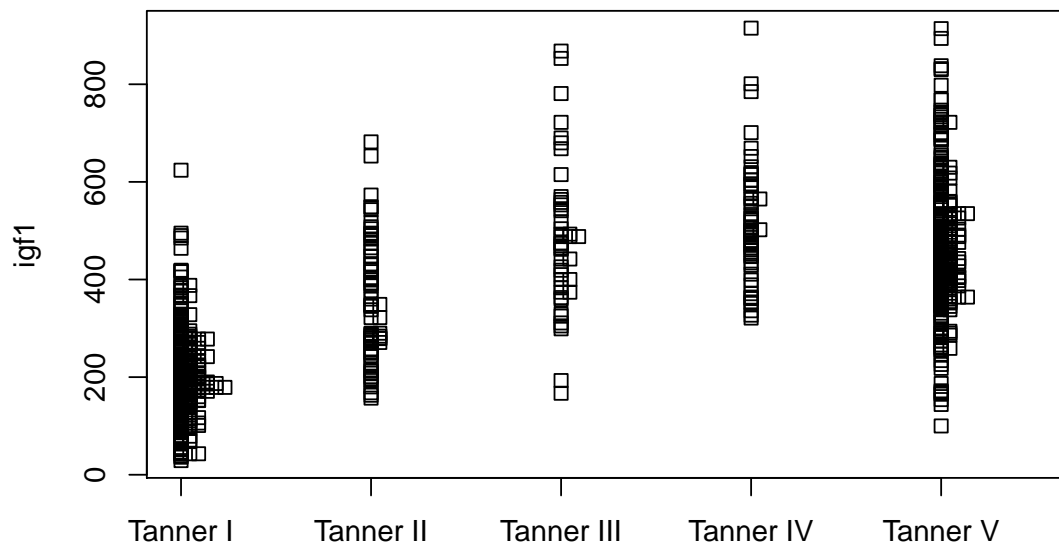


Figura 4.50: Diagrama de *strip chart* de *igf1* para cada categoria de Tanner.

Essencialmente os diagramas 4.48 e 4.50 apresentam a mesma mensagem do que os *boxplots* da figura 4.40.

## 4.8 Diagrama de dispersão ou espalhamento

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

O diagrama de dispersão ou espalhamento é bastante utilizado para mostrar o relacionamento entre duas variáveis numéricas. Ele simplesmente plota os pontos no plano de coordenadas XY, sendo uma variável escolhida para o eixo X e outra para o eixo Y. Vamos criar o diagrama de dispersão das variáveis *igf1* x *idade*. No menu *Gráficos* do *R Commander*, selecionamos a opção:

Gráficos  $\Rightarrow$  Diagrama de Dispersão

Na caixa de diálogo *Gráfico de Dispersão*, selecionamos as variáveis dos eixo X e Y na aba *Dados* (figura 4.51). Também podemos selecionar uma variável de agrupamento, para criar um diagrama de dispersão para cada categoria de uma outra variável. Nesse exemplo, não vamos fazer gráficos por grupos. Na aba *Opções* (figura 4.52), há uma série de opções que podem ser selecionadas. Marquemos a opção *Linha de quadrados mínimos*. Ao clicarmos em OK, o gráfico resultante é mostrado na figura 4.53.

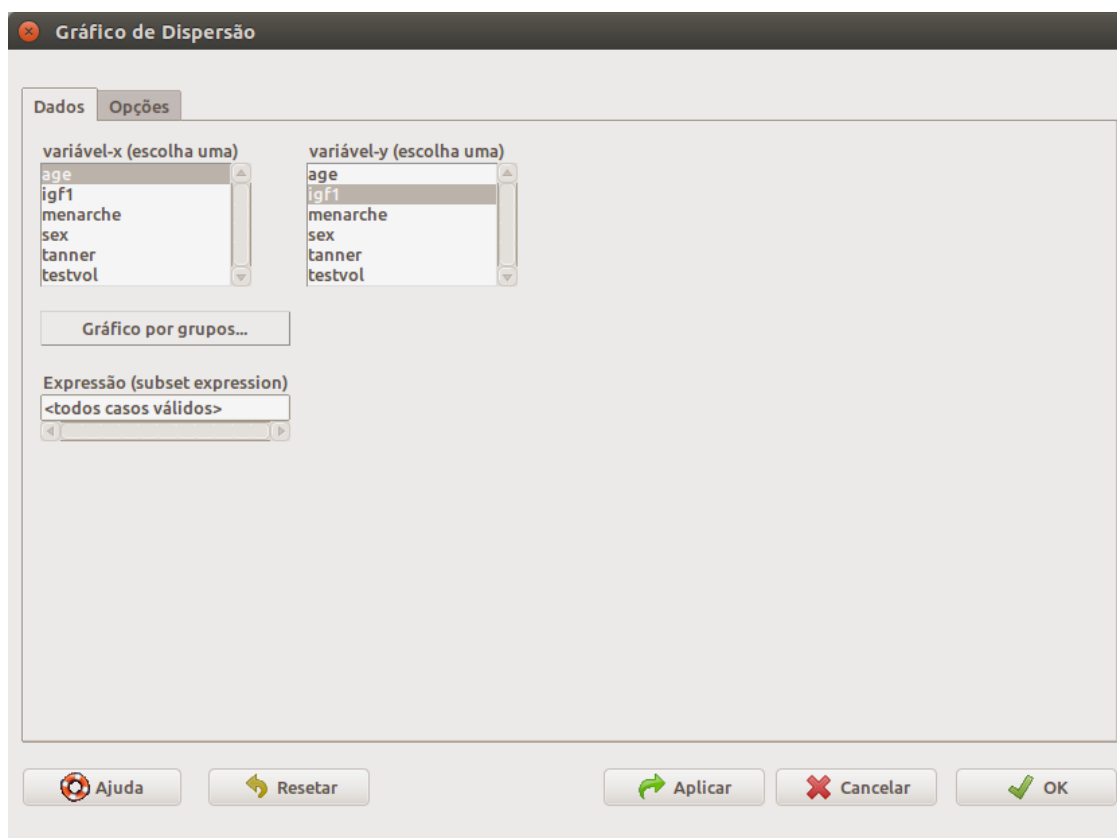


Figura 4.51: Caixa de diálogo para a criação de um diagrama de dispersão. Na aba *Dados*, selecionamos as variáveis numéricas dos eixos X e Y.

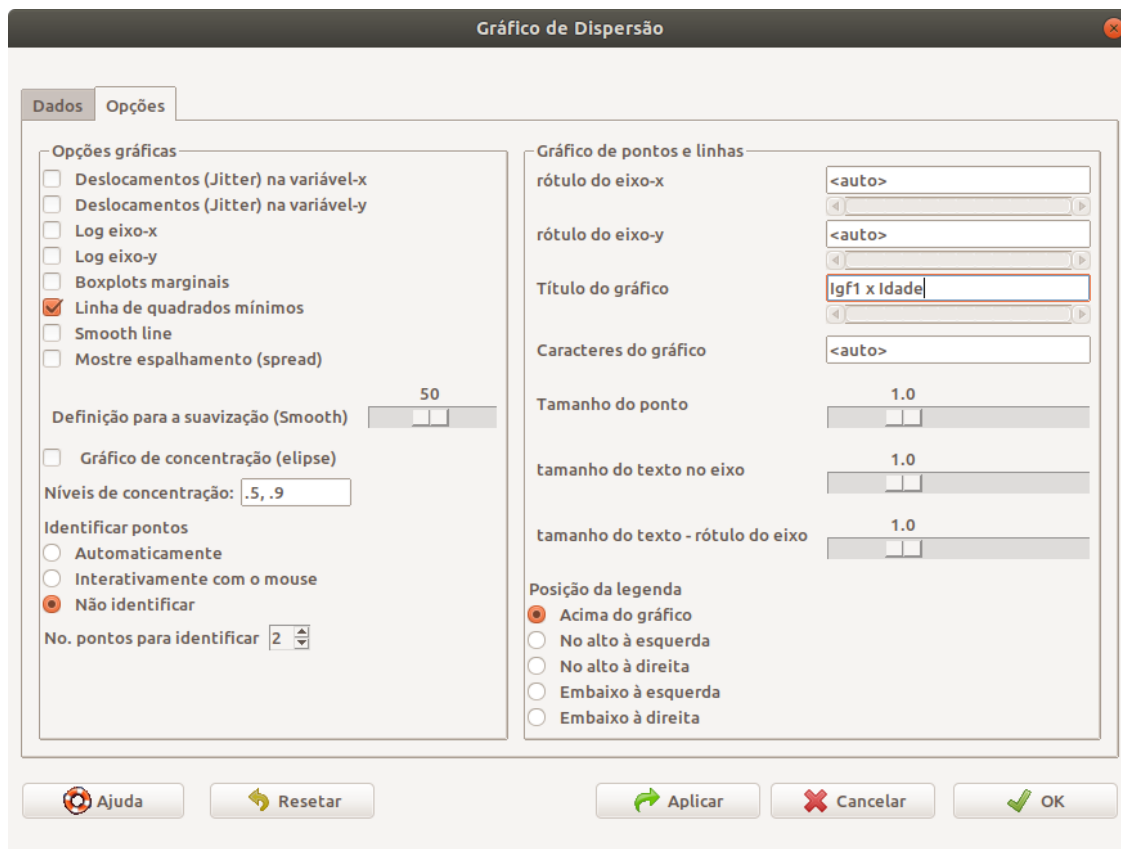


Figura 4.52: Aba *Opções* da caixa de diálogo do gráfico de dispersão.

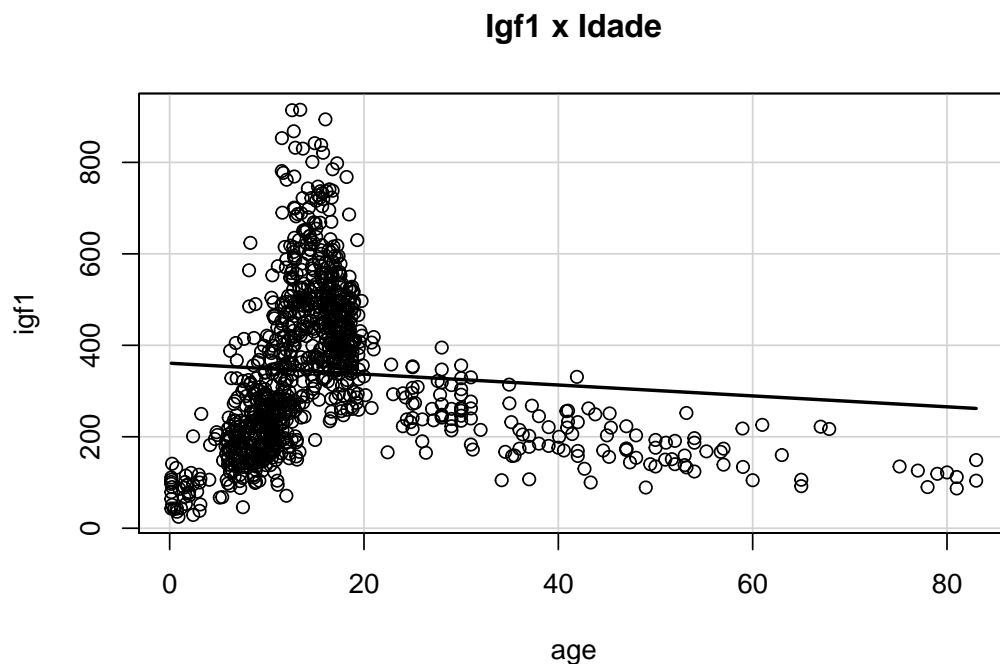


Figura 4.53: Diagrama de dispersão das variáveis *lgf1* x *age*.

Esse diagrama de dispersão sugere que o hormônio *igf1* tende a aumentar com a idade até o final da adolescência e, após esse período, ele tende a decrescer com a idade. Entretanto a mesma observação ao final da seção do diagrama de *boxplot* se aplica aqui.

A reta de regressão é construída de modo a ajustar uma reta aos dados em um diagrama de espalhamento. Como visivelmente a idade não está linearmente associada a *igf1*, a reta de regressão nesse exemplo não reflete como a idade está associada à variável *igf1*.

#### 4.8.1 Alterando a espessura e cor da linha de regressão e o tipo dos pontos

A espessura e a cor da linha de regressão podem ser alteradas por meio do argumento *regLine*. Se quisermos fazer a espessura ser o dobro da original e a cor azul, por exemplo, temos que fazer *regLine=list(lwd = 4, col = "blue")*. O argumento *lwd* (*line width*) especifica a espessura e o padrão é igual a 2.

O tipo dos pontos pode ser alterado por meio do argumento *pch*. Fazendo *pch = 19*, por exemplo, irá plotar pontos cheios. Esta [página](#) mostra uma tabela dos símbolos com os respectivos números. Vide comando a seguir e figura 4.54:

```
scatterplot(igf1~age, regLine=list(lwd = 2, col = "blue"), smooth=FALSE,
            boxplots=FALSE, main="Igf1 x Idade", data=juul2,
            col = "black", pch = 19)
```

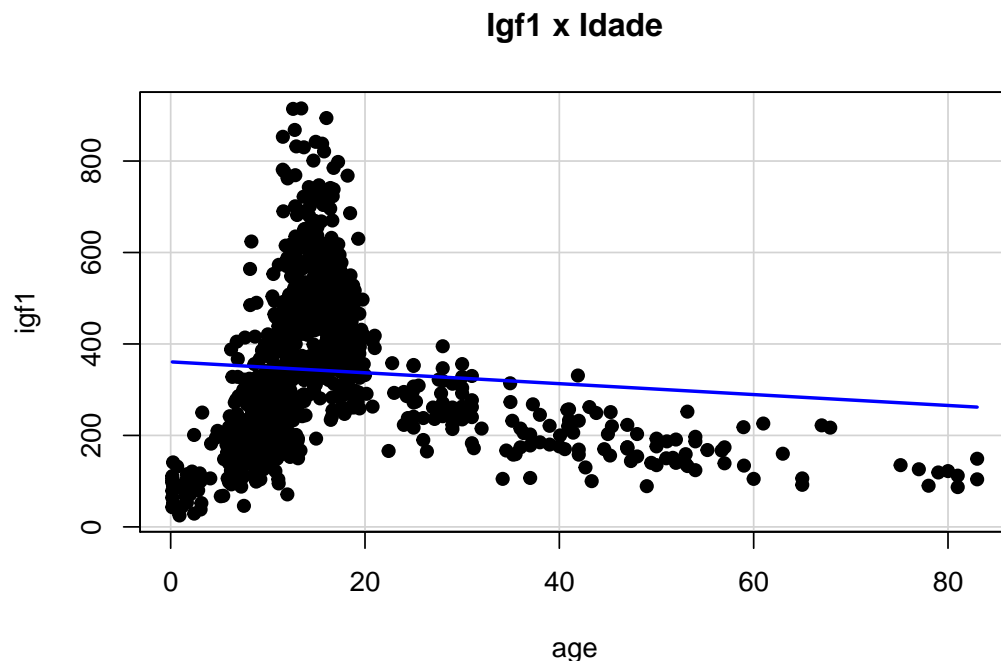


Figura 4.54: Diagrama de dispersão das variáveis *igf1* x *age*, com pontos cheios e reta azul.

Ao selecionarmos as variáveis numéricas que serão plotadas no diagrama de dispersão, podemos distinguir os pontos e as retas de regressão por categorias de uma variável categórica. Ao

clicarmos no botão *Gráfico por grupos...* (figura 4.51), podemos seleccionar uma variável de agrupamento. Seleccionando a variável *tanner\_cat* e configurando as opções do gráfico como mostra a figura 4.55, o resultado é mostrado na figura 4.56.



Figura 4.55: Aba *Opções* da caixa de diálogo do gráfico de dispersão.

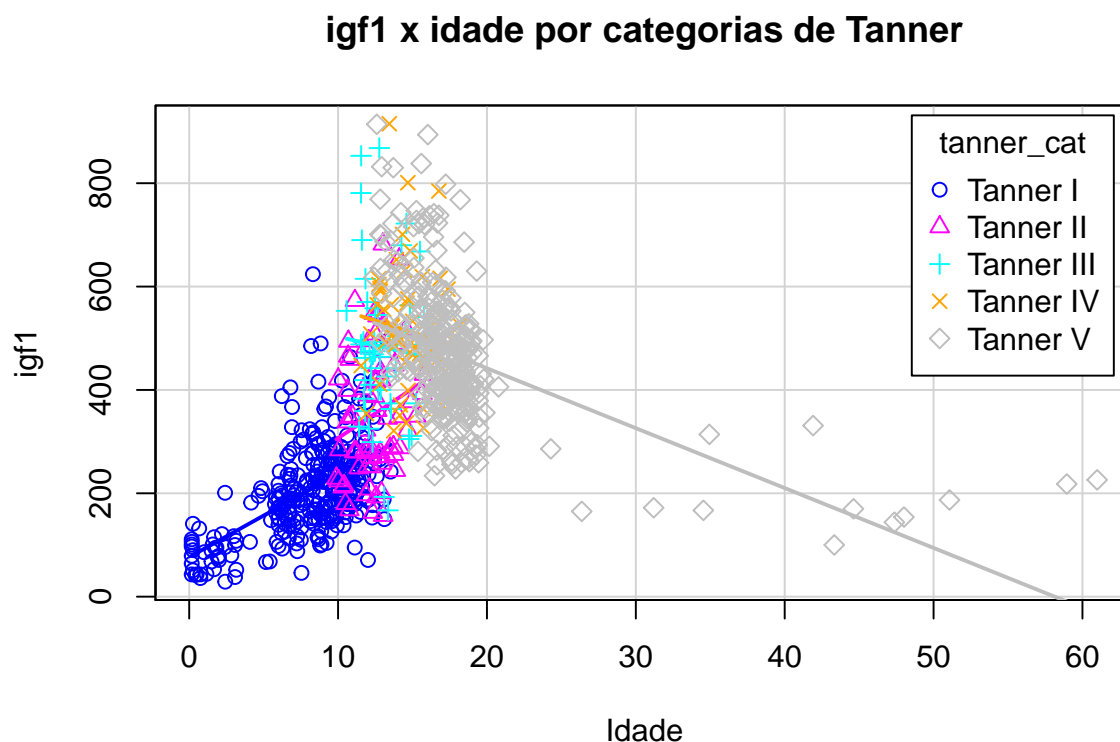


Figura 4.56: Diagrama de dispersão das variáveis *igf1* x *age* por categorias de Tanner.

É possível observar que, dentro de cada categoria de Tanner, com exceção da categoria I, não é tão evidente qual é a relação entre *igf1* e a idade. Na categoria I, aparentemente, *igf1* tende a aumentar com a idade.

## 4.9 Salvando gráficos em um arquivo

No *RStudio*, os gráficos são exibidos na aba *Plots* (figura 4.57). Nessa aba, o usuário pode navegar entre os diversos gráficos que foram gerados na sessão corrente (seta vermelha na figura), bem como salvar o gráfico como imagem, pdf ou copiar para a área de transferência (seta verde). Ao selecionarmos a opção *Salvar como imagem*, surge uma tela que permite ao usuário configurar uma série de opções sobre como o gráfico será gravado.

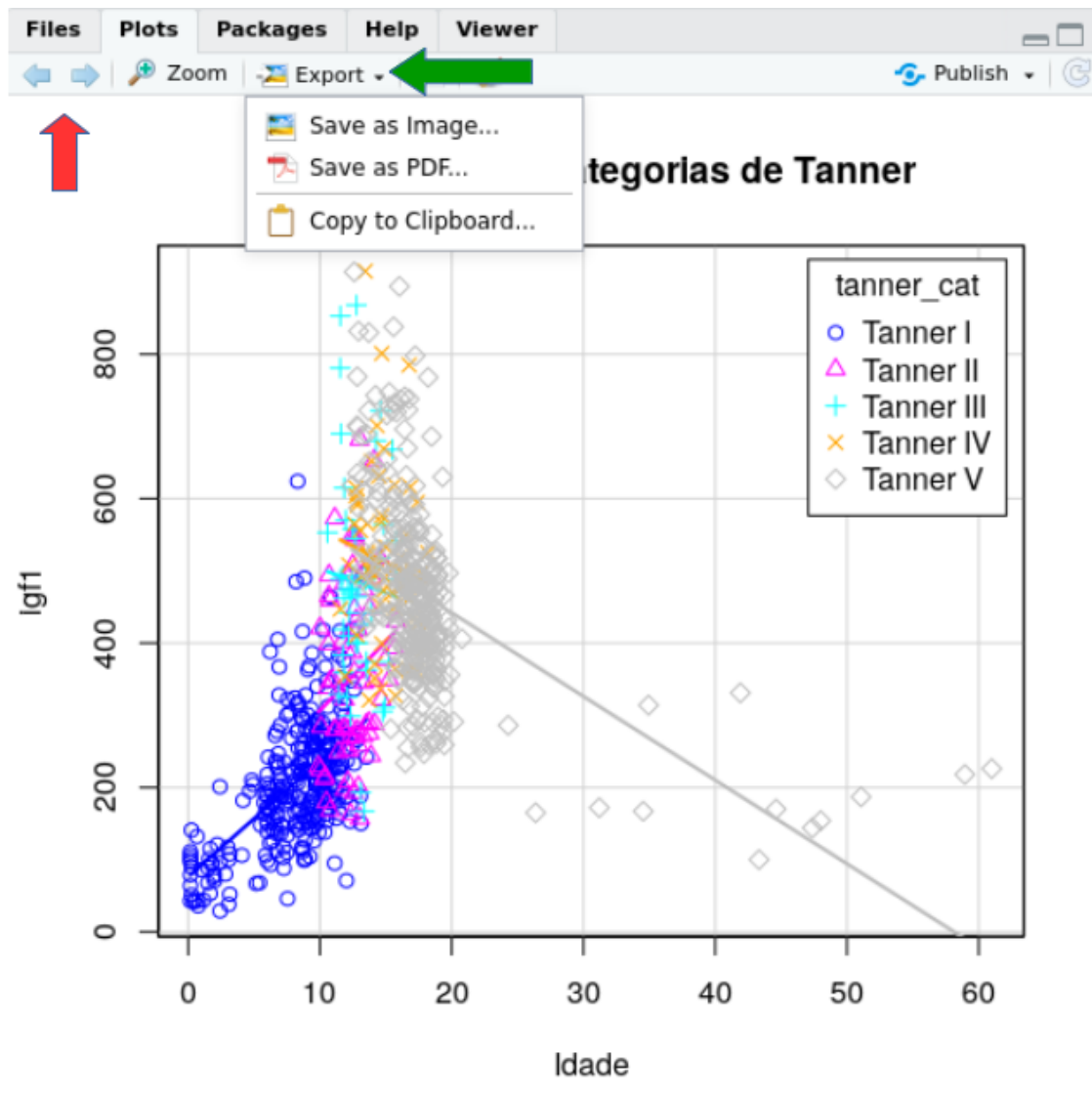


Figura 4.57: Aba *Plots* do *RStudio* com opções para navegar pelos gráficos, ampliar e exportar os gráficos para diferentes formatos.

A caixa de diálogo mostrada na figura 4.58 permite ao usuário selecionar o formato da imagem a ser gravada (png, tiff, jpeg, bmp, svg, eps), a largura e a altura da figura, o nome do arquivo e o diretório (pasta) onde o arquivo será gravado. Ao clicarmos em *Save*, o arquivo será gravado no local especificado.



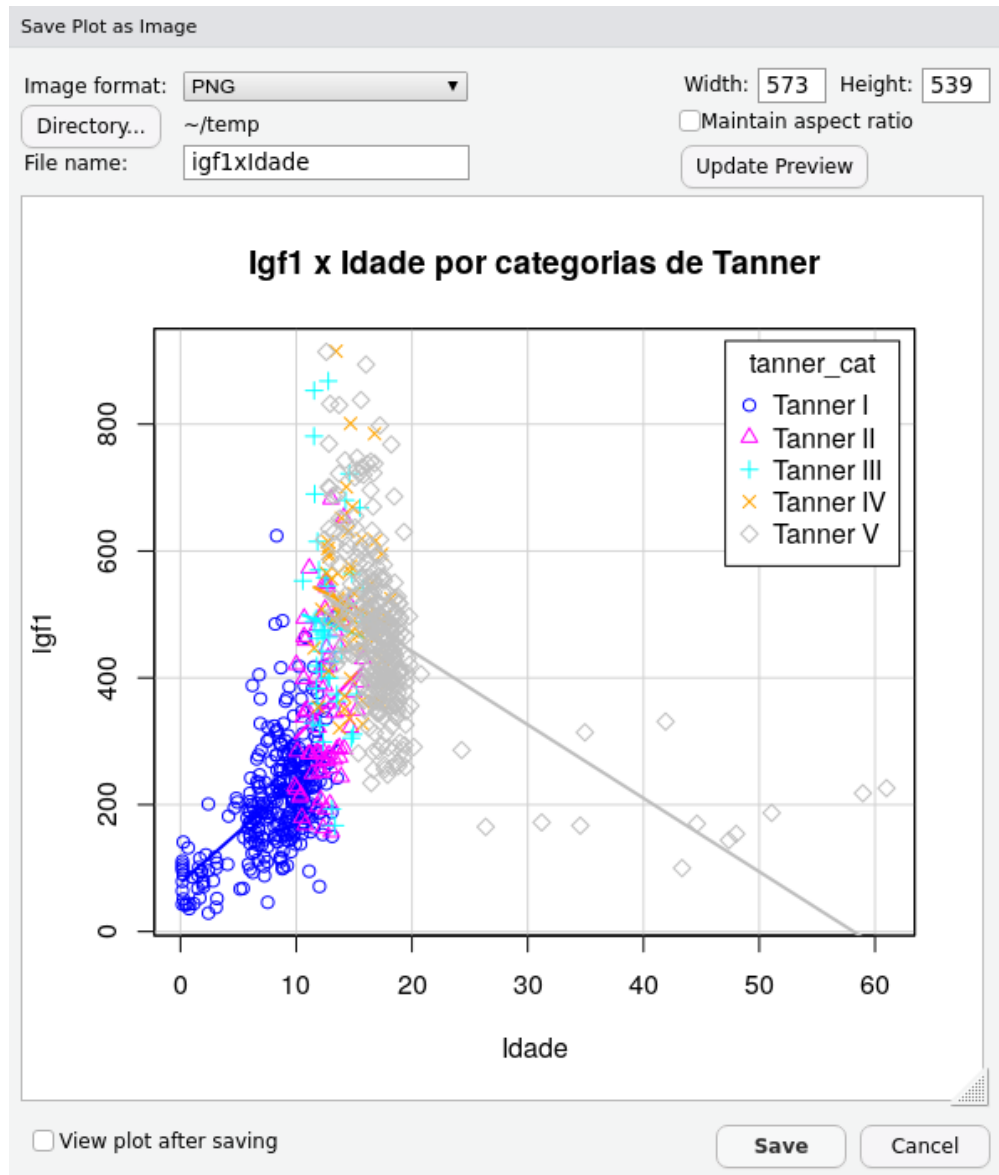


Figura 4.58: Tela para configurar as opções de exportação de um gráfico no *RStudio*.

No *R Commander*, para gravarmos o gráfico em um arquivo, utilizamos a opção:

Gráficos  $\Rightarrow$  Salvar Gráfico em arquivo  $\Rightarrow$  Como Bitmap

A caixa de diálogo mostrada na figura 4.59 permite ao usuário selecionar o formato da imagem a ser gravada (png ou jpeg), a largura e a altura da figura, o tamanho do texto e a resolução. Ao clicarmos em OK, será mostrada uma tela para o usuário especificar o nome do arquivo e o local no disco onde o mesmo será gravado.



Figura 4.59: Tela para configurar as opções de exportação de um gráfico no *R Commander*.

O comando gerado pelo *R Commander* é mostrado a seguir:

```
dev.print(png, filename="/home/sergio/temp/grafico.png", width=6, height=6,
          pointsize=12, units="in", res=300)
```

Para salvarmos o gráfico com uma resolução maior do que 300 dpi, basta alterarmos o valor do argumento *res*.

## 4.10 Recursos gráficos de outros plugins

O *R Commander* oferece uma maneira fácil de construir alguns tipos de gráficos. Porém cada tipo de gráfico possui diversos outros recursos que podem ser utilizados, mas que não são apresentados na interface gráfica. Nesse caso, há outras possibilidades: a) utilizar outros *plugins* que possuem mais recursos na sua interface gráfica; b) digitar o comando que cria o gráfico na janela de script, com as alterações para criar o efeito gráfico desejado; c) utilizar outros pacotes com mais recursos, como o *ggplot2* ([MIT License](#)). Nesta seção, vamos mostrar como carregar outros *plugins* no *R Commander*.

Na sua própria interface gráfica, o *R Commander* pode acrescentar outros *plugins*, com recursos variáveis, inclusive gráficos. Vamos aqui mostrar um desses *plugins*, o *RcmdrPlugin.KMggplot2*. O processo de carregamento desse *plugin* é semelhante para os demais.

Para carregarmos um novo *plugin*, é preciso que ele esteja instalado. Para instalarmos um

*plugin*, o procedimento é o mesmo para instalar qualquer pacote do R. Uma vez instalado o *plugin*, acessamos a opção:

Ferramentas  $\Rightarrow$  Carregar plugins do Rcmdr

Na caixa de diálogo *Carregar Plug-ins*, selecionamos o *plugin* desejado, ou os *plugins*, e clicamos em OK (figura 4.60). Caso um *plugin* desejado não apareça nessa lista, ele precisará ser instalado.

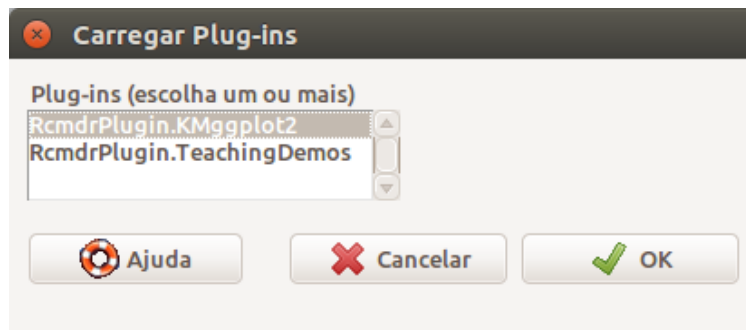


Figura 4.60: Diálogo para selecionar e carregar um novo *plugin* no *R Commander*.

Para carregarmos qualquer *plugin*, será necessário reiniciar o *R Commander* (figura 4.61).

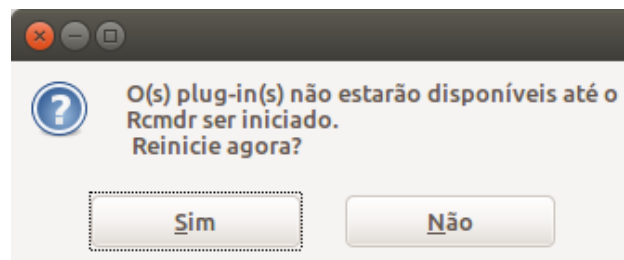


Figura 4.61: É necessário reiniciar o *R Commander* para carregar um novo *plugin*.

Ao reiniciarmos o *R Commander*, se algum conjunto de dados estava sendo usado, ele não está mais ativo. É preciso ativá-lo. Para escolhermos o conjunto de dados que estará ativo no *R Commander*, selecionamos a opção:

Dados  $\Rightarrow$  Conjunto de dados ativo  $\Rightarrow$  Selecionar conjunto de dados ativo...

Neste capítulo, estamos trabalhando com o conjunto de dados *juul2*. Ele aparece na lista de conjuntos de dados disponíveis (figura 4.62). Selecionamos o conjunto *juul2* e clicamos no botão OK. *juul2* será ativado e poderemos acessar o menu do *Kmggplot2* (figura 4.63).

Observem as opções de gráficos desse plugin, experimentem e vejam as diferenças em relação aos recursos padrões do *R Commander*.

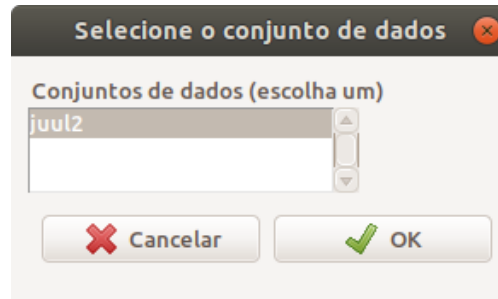


Figura 4.62: Seleção do conjunto de dados a ser ativado.

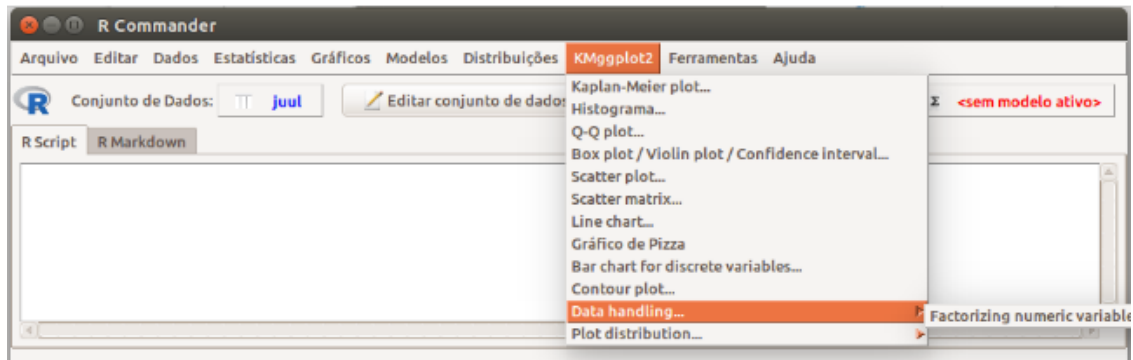


Figura 4.63: *Plugin Kmgplot2* com diversos recursos para a criação de gráficos.

## 4.11 Exercícios

- 1) O diagrama de barras da figura 4.64 mostra a porcentagem de mulheres cujos bebês foram internados em uma UTI neonatal em diversas faixas de escolaridade e estratificada de acordo com o critério se tiveram ou não doença hipertensiva específica da gravidez (DHEG). Responda às questões abaixo.
- Qual a faixa de escolaridade mais frequente?
  - Qual a porcentagem aproximada da faixa de escolaridade de 4-7 anos?
  - Qual a porcentagem aproximada de mulheres na faixa de escolaridade  $>12$  anos e que tiveram DHEG?
  - A partir do diagrama, que relação é sugerida entre a escolaridade e a DHEG? Comente sobre essa relação.

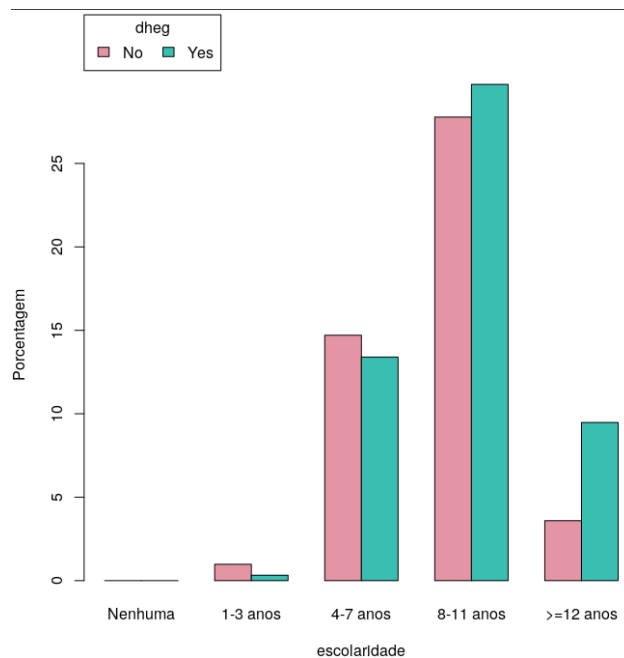


Figura 4.64: Diagrama de barras Escolaridade X DHEG.

- 2) Os dois gráficos da figura 4.65 representam um diagrama de barras para as variáveis peso ao nascimento de recém-nascidos internados em uma UTI neonatal e do número de visitas no pré-natal das respectivas mães.
- Faça uma crítica sobre a adequabilidade dos dois gráficos em mostrar a distribuição dos dados das respectivas variáveis.
  - Que conclusões você pode tirar sobre o uso do diagrama de barras para ilustrar a distribuição dos dados para variáveis numéricas?

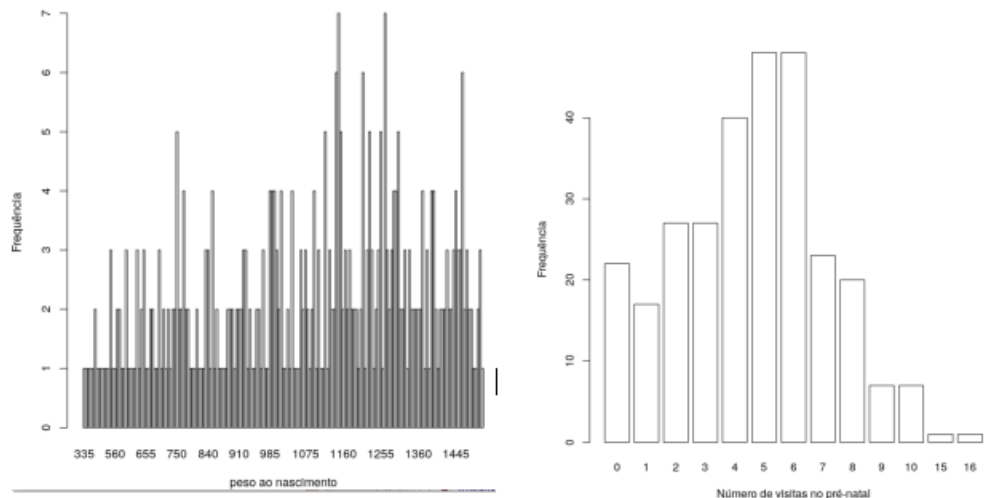


Figura 4.65: Diagrama de barras para variáveis numéricas: a) peso ao nascimento , b) número de visitas no pré-natal.

3) Os três gráficos da figura 4.66 são histogramas da idade gestacional de recém-nascidos que foram internados em uma UTI neonatal. Todos os histogramas foram criados com classes de mesma amplitude.

- O que representa cada barra no histograma superior à esquerda?
- Qual a diferença entre os histogramas?

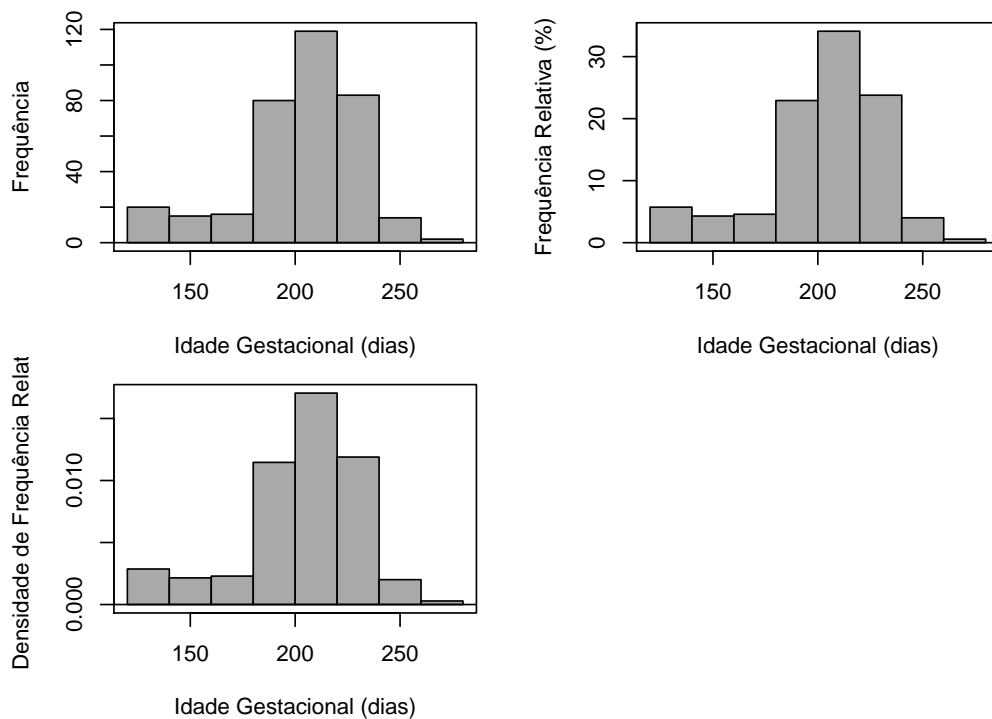


Figura 4.66: Histogramas da idade gestacional de recém-nascidos internados em uma UTI neonatal.

- 4) Os quatro gráficos da figura 4.67 são histogramas da idade gestacional de recém-nascidos que foram internados em uma UTI neonatal.
- Quais são as diferenças entre os histogramas?
  - Que histogramas melhor refletem a distribuição dos dados? Justifique.

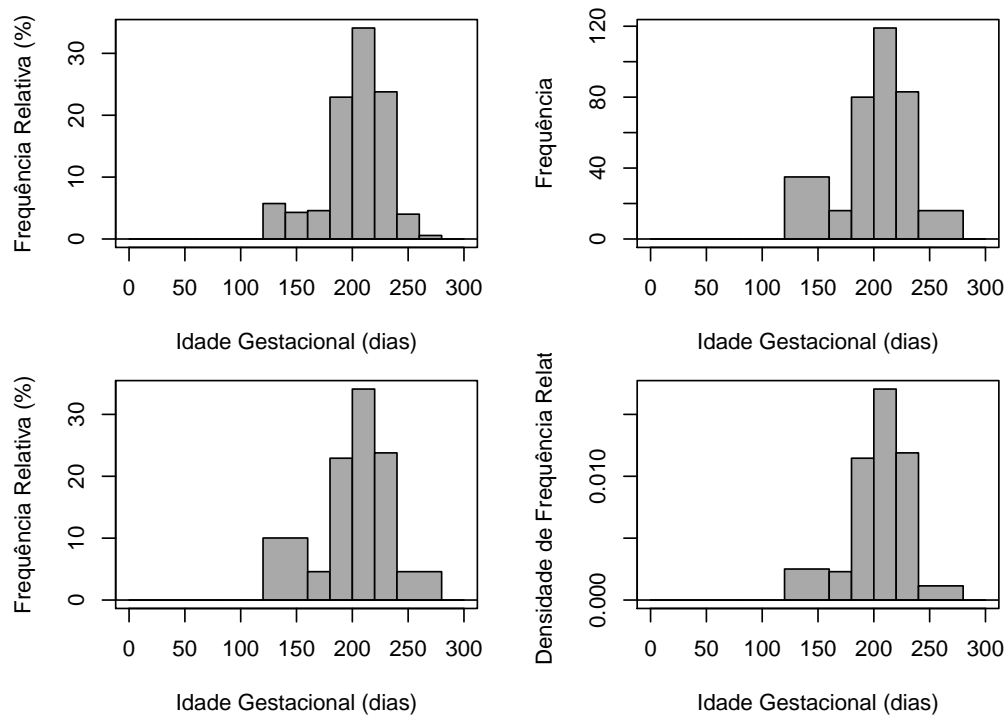


Figura 4.67: Histogramas da idade gestacional de recém-nascidos internados em uma UTI neonatal.

- 5) O gráfico da figura 4.68 mostra os resultados de 5 experimentos para medir a velocidade da luz. Responda às questões abaixo.
- O que significa a linha mais espessa no interior de cada caixa?
  - O que significam as linhas superior e inferior que delimitam as caixas?
  - O que significam as linhas superior e inferior fora das caixas?
  - Por que os experimentos 1 e 3 contêm alguns pontos representados por círculos e os demais não?
  - Por que as linhas tracejadas no experimento 2 não possuem o mesmo comprimento?
  - Qual experimento mostrou maior variabilidade? E a menor? Que critério você utilizou para avaliar a variabilidade dos dados?
  - Que experimento teve a menor mediana? E a maior?
  - Indique os experimentos em que a média e o desvio padrão poderiam ser utilizadas como boas medidas de tendência central e dispersão.

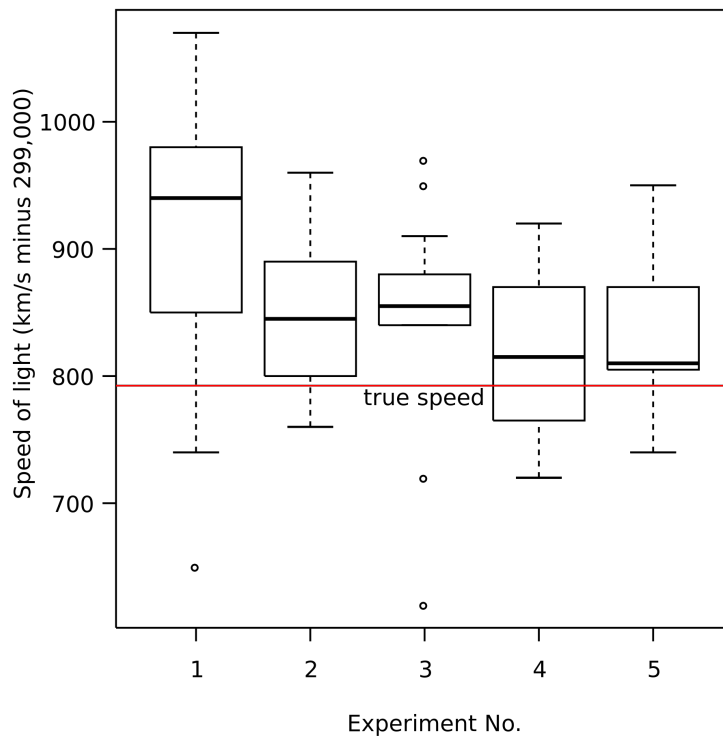


Figura 4.68: Boxplots de experimentos para medir a velocidade da luz. Fonte: [wikipedia](#) (Domínio Público).

- 6) Carregue o conjunto de dados *Pima.te* do pacote *MASS* ([GPL-2](#) | [GPL-3](#)).
  - a) Veja a ajuda do conjunto de dados.
  - b) Use o comando ao final do exercício para gerar uma variável, *glu\_cat*, com as seguintes faixas de glicose: (0-99], (99-120] e >120 mg/dl.
  - c) Faça um diagrama de barras diagrama de barras condicional, lado a lado com as porcentagens da variável *type* para cada categoria da variável *glu\_cat*. Comente o gráfico.
  - d) Faça um *boxplot* da variável *glu* para cada categoria da variável *type*. Comente o diagrama.
  - e) Faça um histograma da variável *glu* para cada categoria da variável *type*. Comente o diagrama.
  - f) Faça um diagrama de *stripchart* da variável *bmi* por categoria da variável *type*. Comente o gráfico.
  - g) Faça um diagrama de dispersão de *glu* por *bmi*. Comente o gráfico.

```
Pima.te$glu_cat <- with(Pima.te, cut(x = glu,
                                   breaks = c(0, 99, 120, Inf),
                                   labels=c('(0,99]', '(99-120]', '>120)'),
                                   ordered_result = TRUE, right = TRUE))
```



# Capítulo 5

## Amostragem e delineamentos de pesquisas

### 5.1 Introdução

Os conteúdos desta seção, da seção 5.2, e da seção 5.3 e suas subseções podem ser visualizados neste [vídeo](#).

O dogma central da inferência estatística é que podemos caracterizar propriedades de uma população de indivíduos a partir de dados colhidos a partir de uma amostra de indivíduos dessa população (figura 5.1).

Se quiséssemos, por exemplo, saber como se distribui a glicemia de jejum numa população de diabéticos, é impraticável medir a glicemia de jejum em todos os diabéticos. Então o que se faz geralmente é extrair uma amostra dessa população, calcular a média e a variância da glicose nesta amostra e, a partir da inferência estatística, extrair informações sobre a média e a variância da população de diabéticos. Se supusermos que a distribuição da glicemia de jejum fosse uma distribuição normal, algumas expressões analíticas nos dão, com certa confiança, intervalos que contêm os valores reais da média e variância da distribuição na população.

Durante a pandemia causada pelo novo agente da família coronavírus (SARS-CoV-2), em que se estudam diversos tratamentos contra o agente causador da doença, esses tratamentos não são aplicados indiscriminadamente na população para se conhecer os seus efeitos. Estudos são conduzidos em amostras de pacientes onde são comparadas as diversas propostas terapêuticas e, a partir dos resultados obtidos, tenta-se inferir o que aconteceria na população. Geralmente não é um único estudo que vai dar a resposta definitiva.

Este capítulo se inicia com conceitos básicos de amostragem em estatística. Em seguida, serão apresentados de maneira simplificada, os principais delineamentos de estudos clínico-epidemiológicos, que são estudos que visam a demonstrar efeitos de tecnologias em saúde sobre desfechos clínicos, relacionamentos entre variáveis clínicas, prognósticos de doenças, ou a influência de possíveis fatores etiológicos sobre doenças ou outros desfechos

clínicos. Para responder a essas questões, diversas arquiteturas de estudos têm sido propostas na literatura científica. Delineamentos de estudos são também chamados de desenhos de estudos ou tipos de estudos.

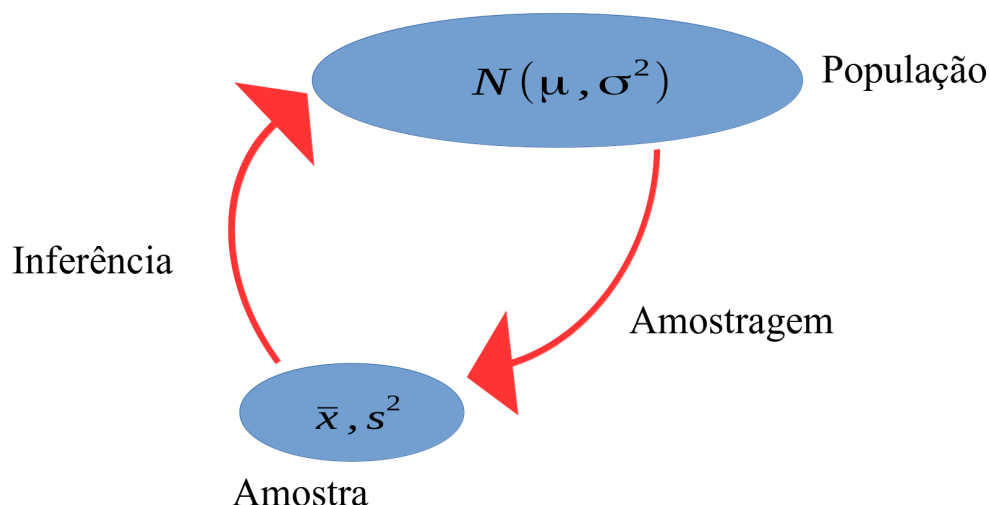


Figura 5.1: Dogma central da estatística: inferência sobre a população a partir de uma amostra extraída dessa população.

## 5.2 População e amostra

Quando se realiza um estudo clínico-epidemiológico, há a intenção de aplicar os resultados do estudo a uma população de indivíduos, definida por características que podem ser sócio-demográficas, físicas, psicológicas ou clínicas. Essa população é chamada de **população alvo** e as características da mesma são estabelecidas no protocolo do estudo. Assim a **população** é um conjunto de indivíduos ou objetos que contém uma ou mais características de interesse. Esses indivíduos podem ser pessoas, escolas, animais, etc., dependendo do tipo de estudo que está sendo realizado. Por exemplo, em um estudo sobre as características de pacientes com urolitíase, a população seria os pacientes com diagnóstico de urolitíase.

Uma vez que geralmente é impossível estudarmos uma população-alvo inteira (todos os pacientes com infarto do miocárdio, ou todos os fumantes, por exemplo), os pesquisadores recorrem a uma amostra da população para realizar, então, um estudo experimental ou observacional. Assim uma **amostra** é um subconjunto da população que também contém as características de interesse. No exemplo citado de pacientes com urolitíase, uma amostra poderia ser os pacientes acompanhados no ambulatório de nefropediatria do Hospital Federal dos Servidores do Estado do Rio de Janeiro (HFSE), no período de janeiro de 2012 a dezembro de 2014 e com o diagnóstico de urolitíase.

A figura 5.2 mostra uma amostra de sangue, a partir da qual exames clínicos são realizados para inferir a concentração de diversos compostos químicos no sangue a partir da concentração na amostra.



Figura 5.2: Exemplo de um processo de amostragem: coleta de uma amostra de sangue.  
Fonte: [Wikipedia](#) (CC BY-SA-3.0).

## 5.3 Amostragem

A **amostragem** é um processo de seleção de um subconjunto da população de interesse que gera a amostra. A amostragem é uma área da estatística que estuda métodos de como determinar o tamanho de uma determinada amostra para se atingir determinado objetivo e técnicas sobre como selecionar amostras da população de modo a realizar inferências sobre a população a partir da análise da amostra. Nesse sentido, duas características desejáveis de uma amostra são: que elas sejam representativas da população de onde são extraídas e que elas sejam geradas de preferência de maneira aleatória.

Quando se extrai uma amostra de uma população, em geral há alguma diferença entre as características da amostra e as da população. Por exemplo, ao se medir a média da glicemia de jejum em uma amostra de pacientes diabéticos, esse valor irá diferir geralmente do valor da média da glicemia de jejum na população. Assim o **erro amostral** é a diferença entre o valor de um parâmetro na população e a sua estimativa a partir de uma amostra.

Há diversos métodos de amostragem que podem ser divididos basicamente em duas categorias: **amostragem probabilística** e **amostragem não probabilística**.

### 5.3.1 Amostragem probabilística

Na amostragem probabilística, a seleção dos itens que comporão a amostra é realizada de maneira aleatória, relacionada à probabilidade de ocorrência dos itens na população. O uso da aleatorização no processo de amostragem nos permite a análise dos resultados usando os métodos de **inferência estatística**. A inferência estatística é baseada nas leis da probabilidade e permite ao pesquisador inferir conclusões acerca de uma dada população com base nos resultados obtidos com a amostragem aleatória.

Entre os métodos de amostragem probabilística, vamos considerar brevemente quatro:

- aleatória simples;
- aleatória estratificada;
- por conglomerados;
- sistemática.

#### 5.3.1.1 Amostragem aleatória simples

É a técnica básica de amostragem. Corresponde a uma amostra de elementos retirados ao acaso da população, isto é, cada indivíduo é escolhido completamente ao acaso e cada membro da população tem a mesma probabilidade de ser incluído na amostra (figura 5.3).

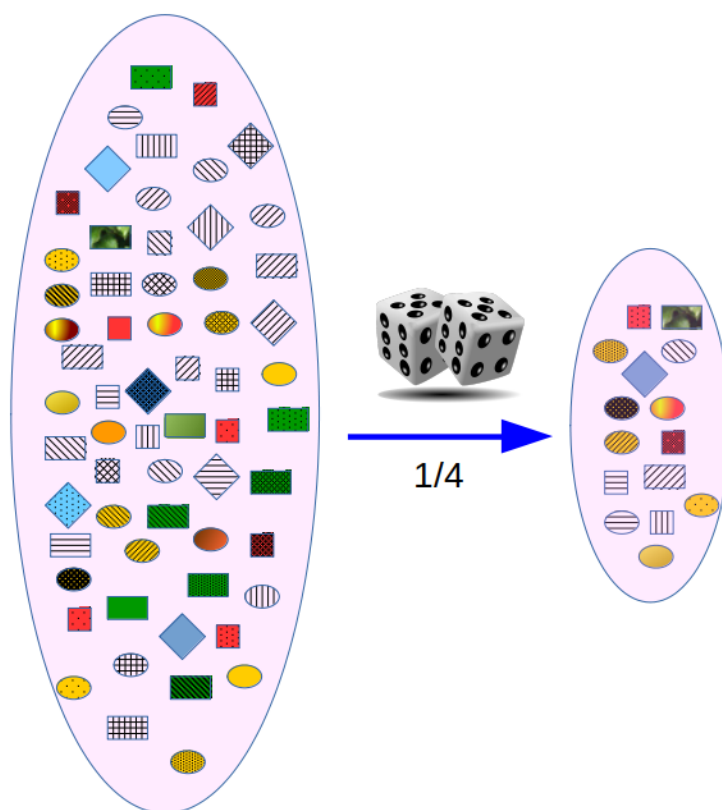


Figura 5.3: Amostragem aleatória simples. Neste exemplo, a amostra contém  $1/4$  dos itens da população selecionados aleatoriamente.

**Exemplo:** o estudo “O perfil socioeconômico e a percepção ambiental dos pescadores da Lagoa do Apodi, Rio Grande do Norte, Brasil” (Pinto Filho et al., 2020) ([CC BY](#)) realiza uma amostragem aleatória simples de indivíduos que fazem uso da lagoa do Apodi para pesca, a partir da informação de que, na colônia de pescadores de Apodi, existem 405 pescadores cadastrados na Confederação Nacional de Pescadores e Aquicultores(CNPA), dos quais 240 atuam diretamente na lagoa. Desses 240 pescadores, 52 foram selecionados aleatoriamente para participarem do estudo.

### 5.3.1.2 Amostragem estratificada

A amostragem estratificada pode ser usada quando a população alvo do estudo é heterogênea e pode ser dividida em estratos ou subgrupos homogêneos, e deseja-se que a amostra extraída da população tenha em cada estrato uma proporção de indivíduos igual ou semelhante à correspondente proporção na população. Por exemplo, a população de uma região pode ser dividida em faixas etárias e uma amostra da população pode ser extraída de tal modo que a proporção de pessoas em cada faixa etária na amostra seja semelhante à proporção de pessoas naquela faixa etária na população alvo.

Na figura 5.4, a amostragem foi construída de modo que a amostra refletisse a proporção de elementos com o mesmo formato na população.

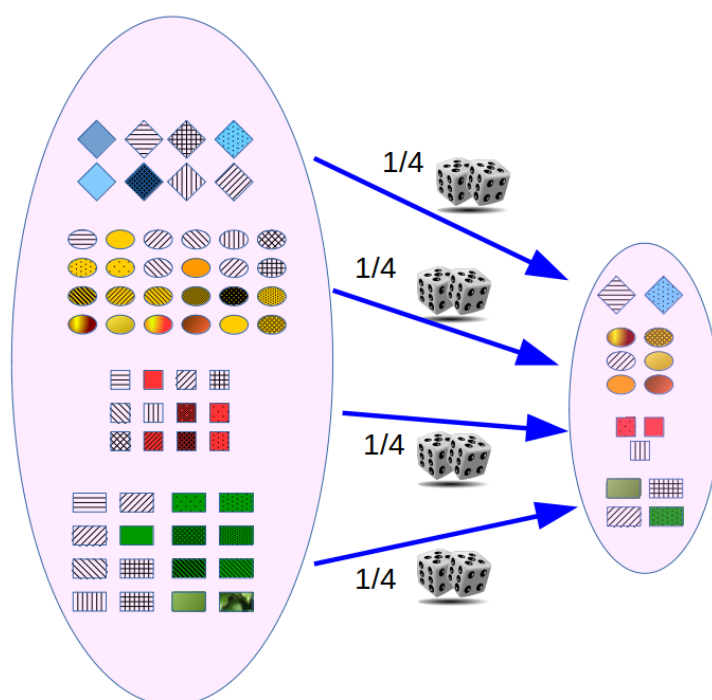


Figura 5.4: Amostragem estratificada.

A amostragem estratificada realizada a partir de uma certa divisão em estratos não garante proporções semelhantes para outras formas de divisão da população em subgrupos. Por exemplo, uma vez realizada uma amostragem estratificada de uma dada população a partir de uma estratificação por faixas etárias, não se pode garantir que a amostra tenha a mesma proporção nos estratos formados por renda que a população alvo.

**Exemplo:** no estudo multicêntrico “Hipertensão Arterial e Diabetes Mellitus entre trabalhadores da saúde: associação com hábitos de vida e estressores ocupacionais” (Novaes Neto et al., 2020) (CC BY), a amostra dos participantes no estudo foi realizada da seguinte forma: a partir de listas nominais de todos os trabalhadores em atividade nos serviços de saúde da atenção básica e da média complexidade, fornecidas pelas Secretarias de Saúde

dos municípios estudados, estratificou-se a amostra em três níveis: área geográfica, nível de assistência (atenção básica e média complexidade) e grupo ocupacional. A composição da amostra foi definida com base na participação percentual de cada grupo por nível de estratificação estabelecido, seguindo-se o sorteio dos trabalhadores para comporem a amostra (o sorteio foi feito com base em listagem de números aleatórios).

### 5.3.1.3 Amostragem por conglomerados

Na amostragem por conglomerados (*clusters* em inglês), a seleção aleatória é realizada em grupos previamente existentes, como, por exemplo, escolas, unidades de saúde, bairros, etc. Nesse tipo de amostragem, a população é dividida em grupos (conglomerados) e uma amostra aleatória simples dos grupos é selecionada. Em seguida, os elementos em cada conglomerado selecionado são amostrados. Se todos os elementos de cada conglomerado for selecionado, então essa amostragem é chamada de amostragem por conglomerados de um estágio. Se uma amostra aleatória é extraída de cada conglomerado selecionado, então essa amostragem é chamada de amostragem por conglomerados de dois estágios.

Na figura 5.5, os objetos a serem amostrados estão distribuídos em caixas com 5 objetos cada. Para selecionar 15 objetos dessa população, foram selecionadas aleatoriamente três caixas e os objetos contidos nas caixas selecionadas formaram a amostra da população. Essa foi uma amostragem por conglomerados de um estágio.

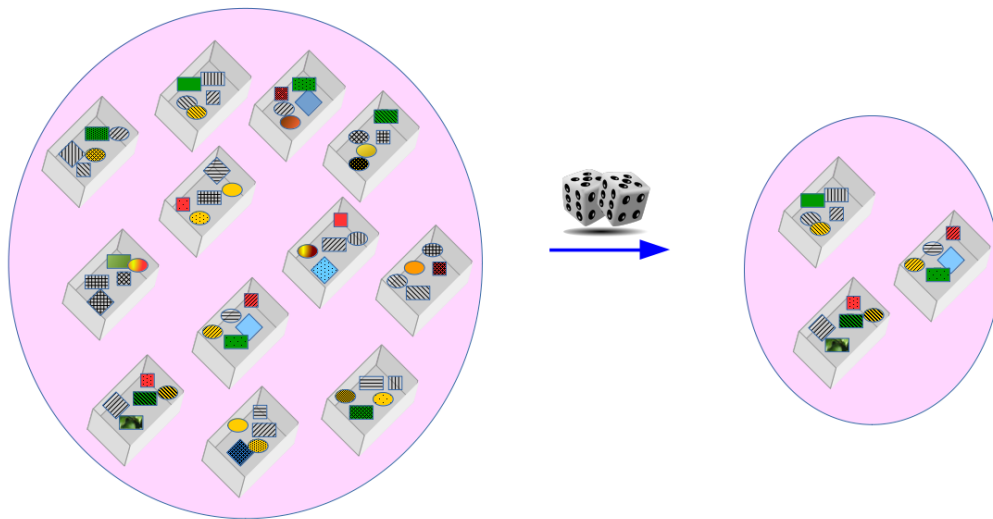


Figura 5.5: Amostragem por conglomerados de 1 estágio. Os grupos são selecionados aleatoriamente e todos os elementos de cada grupo são selecionados.

**Exemplo:** no estudo “Associação da depressão com as características sociodemográficas, qualidade do sono e hábitos de vida em idosos do Nordeste brasileiro: estudo seccional de base populacional” (Lopes et al., 2015) (CC BY), a população-alvo constou de aproximadamente 40 mil idosos, no ano de 2010, residentes na zona urbana da cidade de Campina Grande-PB. O tamanho da amostra de idosos foi estimada em 205 idosos. Esse estudo realizou uma

amostragem por conglomerados, na qual o primeiro estágio consistiu na seleção aleatória de unidades básicas de saúde (UBS) em quatro distritos sanitários urbanos.

No segundo estágio, foi realizada uma amostragem sistemática das residências nas ruas das unidades básicas de saúde selecionadas, conforme detalhado na seção seguinte.

#### 5.3.1.4 Amostragem sistemática

Numa amostragem sistemática, como o nome indica, há um certo procedimento sistemático para escolher os elementos da amostra. Por exemplo, na figura 5.6, para selecionarmos 15 dos 60 objetos da população, criamos uma lista ordenada dos objetos e sorteamos um número de 1 a 4 para indicar o primeiro item a ser selecionado. A partir desse item, selecionamos os demais a partir do item selecionado anteriormente, pulando 4 elementos na lista.

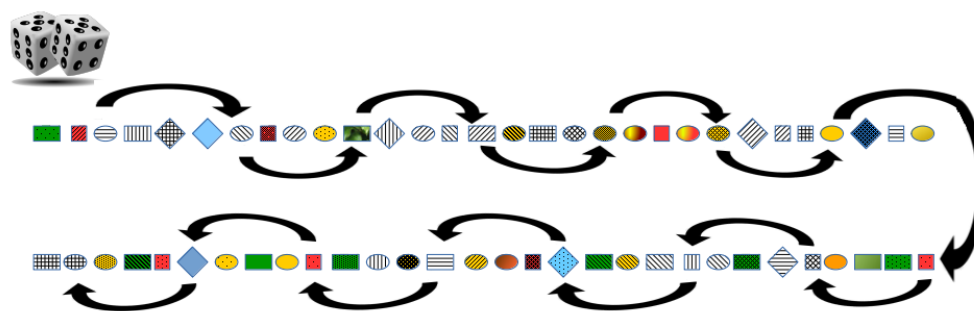


Figura 5.6: Amostragem sistemática.

**Exemplo:** no exemplo mostrado na seção anterior, após a seleção das unidades básicas de saúde (UBS), as ruas das UBS foram percorridas de uma extremidade a outra, nas duas laterais, saltando-se nove casas a partir da esquina escolhida como início. Essa amostragem sistemática dos domicílios foi determinada pela proporção de idosos e casas da cidade a serem visitadas, estratégia semelhante à Pesquisa Nacional por Amostragem de Domicílios (PNAD). Caso não houvesse idoso no domicílio selecionado, devia-se procurar na residência posterior e, se necessário, na anterior. Tendo mais de um idoso no local, realizava-se a coleta de dados com todos.

A amostragem sistemática é uma forma de simplificar o processo de amostragem, mas deve ser realizada com cuidado, pois pode gerar uma amostra não representativa da população. Na figura 5.7, pelo fato de os itens estarem dispostos de uma maneira regular, todos os itens selecionados possuem a mesma forma. Portanto a amostra nesse caso não é representativa da população de objetos.

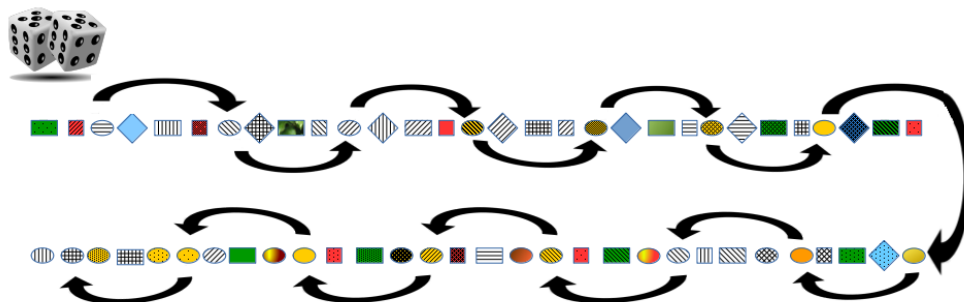


Figura 5.7: Amostragem sistemática que gera uma amostra não representativa da população.

Como outro exemplo, vamos supor que, em um ambulatório, foi realizada uma enquete com pacientes atendidos na unidade para saber o nível de satisfação dessas pessoas com o atendimento prestado. Os pacientes selecionados foram aqueles que compareceram ao ambulatório em 5 terças-feiras seguidas. Uma limitação desse processo é que pode ser que os pacientes de terça-feira sejam atendidos por uma equipe diferente de outros dias ou que os pacientes das terças-feiras possuam um perfil diferente dos pacientes dos demais dias.

### 5.3.2 Amostragem não probabilística

Na amostragem não probabilística, há uma escolha deliberada dos elementos que comporão a amostra, de acordo com critérios e julgamentos estabelecidos pelos pesquisadores.

Há diversas formas de realizar uma amostragem não probabilística, entre elas:

- por conveniência;
- cotas;
- bola de neve;
- julgamento.

Neste texto, vamos tratar somente da amostragem por conveniência.

#### 5.3.2.1 Amostragem por conveniência

Essa técnica de amostragem é muito comum na pesquisa clínico-epidemiológica. Ela consiste em formar uma amostra da população a partir de itens que estejam mais facilmente disponíveis (figura 5.8). Em estudos na área clínica, frequentemente as amostras são obtidas simplesmente identificando um número de pacientes que atendem aos critérios para inclusão em um estudo.



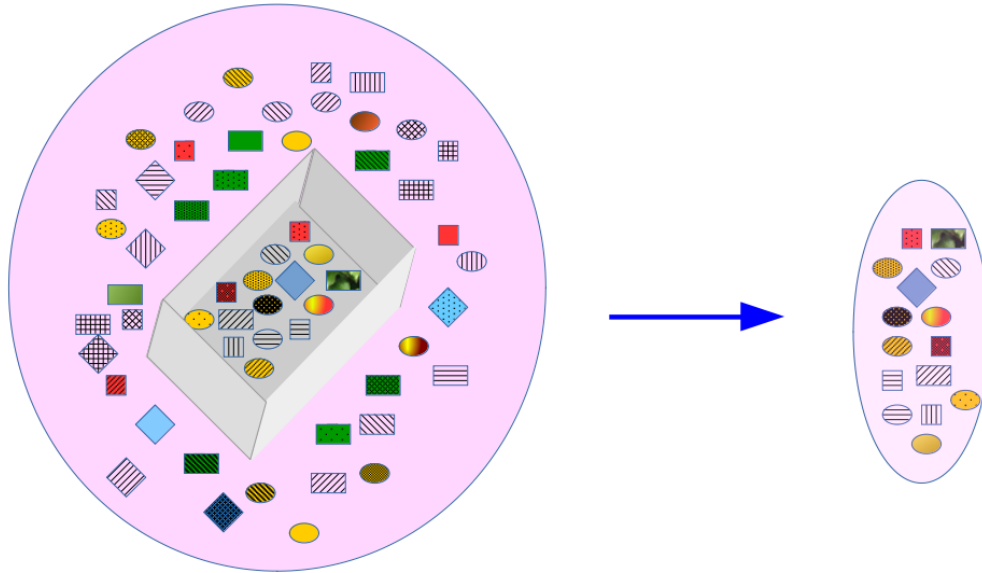


Figura 5.8: Amostragem de conveniência. Nesta figura, a amostra é formada por todos os itens que já estão agrupados dentro de uma caixa.

**Exemplo:** No estudo “Perfil clínico-epidemiológico de 106 pacientes pediátricos portadores de urolitíase no Rio de Janeiro” (Barata and Valette, 2018) (CC BY), os pacientes que constituíram a amostra foram selecionados por meio de consulta ao setor de Estatística e Arquivo do Hospital Federal dos Servidores do Estado do Rio de Janeiro (HFSE), sendo identificados todos os pacientes acompanhados no ambulatório de nefropediatria do hospital, no período de janeiro de 2012 a dezembro de 2014, e selecionados os prontuários com o diagnóstico de urolitíase. Os critérios de inclusão utilizados foram:

1. idade entre 1 mês e 18 anos;
2. confirmação do diagnóstico clínico por pelo menos um exame radiológico, podendo ser a radiografia simples de abdome, a ultrassonografia abdominal ou de aparelho urinário (que pode identificar cálculos  $\geq 5$  mm) e a tomografia computadorizada helicoidal de abdome sem contraste;
3. ser assistido no ambulatório de nefropediatria do HFSE no período citado anteriormente.

Muito cuidado deve ser tomado ao se tentar generalizar os resultados obtidos a partir de um estudo onde a amostra é por conveniência, já que a amostra não é uma amostra aleatória da população e, a rigor, os métodos de inferência estatística somente poderiam ser aplicados a uma amostra aleatória da população de interesse. O problema é que, em geral, é difícil caracterizar na prática essa população e determinar se ela é semelhante à população para a qual os pesquisadores desejam inferir os resultados.

## 5.4 Delineamentos de estudos clínico-epidemiológicos

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

Nas próximas seções, serão apresentados em linhas gerais os seguintes desenhos de estudo: ensaio controlado randomizado, ensaio controlado não randomizado, estudo de coortes, estudo de caso-controle, estudo transversal e série de casos.

Abaixo são definidos quatro conceitos que serão mencionados em diversos pontos do texto:

- **estudo longitudinal:** um estudo é longitudinal quando as variáveis clínicas são mensuradas em mais de um instante do tempo, ou seja, quando as unidades de observação (frequentemente pacientes) são acompanhadas ao longo do tempo e diversas variáveis são coletadas para verificar o estado clínico em cada instante;
- **estudo transversal:** um estudo é transversal quando as variáveis clínicas são mensuradas em um único instante do tempo para cada unidade de observação;
- **estudo prospectivo:** neste texto, por estudo prospectivo entende-se um estudo onde as variáveis serão coletadas a partir do início do estudo em diante;
- **estudo retrospectivo:** neste texto, por estudo retrospectivo entende-se um estudo onde as variáveis analisadas já foram coletadas a priori, utilizando-se de diversas fontes de dados como, por exemplo, prontuários, bancos de dados etc., ou entrevistas são realizadas com as unidades de observação para verificar os valores das variáveis estudadas em instantes anteriores à realização do estudo.

## 5.5 Ensaio controlado randomizado

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

O ensaio controlado randomizado (ECR) é considerado como o padrão ouro para determinar a evidência científica sobre os efeitos de tecnologias em saúde. Um ensaio controlado randomizado bem planejado e conduzido é o tipo de delineamento que apresenta menos possibilidade de ocorrência de vieses. Viés, vício ou tendenciosidade (*bias* em inglês) é um processo em qualquer estágio de inferência que tende a produzir resultados que se desviam sistematicamente dos valores verdadeiros (Fletcher et al., 2014).

Um ECR deve ser precedido de um protocolo que justifique e descreva como o estudo será realizado. Pocock (Pocock, 1983) sugere os seguintes itens como os mais importantes de um protocolo de um ensaio randomizado:

- contexto e objetivos do estudo;
- objetivos específicos;
- critérios de seleção dos pacientes;
- aplicação dos tratamentos;
- métodos de avaliação dos pacientes;
- delineamento do estudo;
- registro e randomização dos pacientes;
- consentimento dos pacientes;
- cálculo do tamanho amostral necessário;

- monitoramento do progresso do ensaio;
- formulários e manipulação dos dados;
- como tratar desvios do protocolo;
- planos para a análise estatística;
- responsabilidades administrativas.

De uma maneira simplificada, um ECR *paralelo* pode ser esquematizado como mostra a figura 5.9. Vamos tomar como exemplo o estudo de Rocha et al. (Rocha et al., 2009). O objetivo deste estudo foi avaliar a segurança na redução do tempo de repouso no leito, de seis para três horas, após cateterismo cardíaco diagnóstico com introdutor arterial 6 F. As duas intervenções (tratamentos) comparadas no estudo eram: *deambulação após três horas após a remoção da bainha* e *deambulação após seis horas após a remoção da bainha*. Vamos chamar a primeira intervenção de A e a segunda de B. Os desfechos clínicos avaliados no estudo foram possíveis complicações vasculares: 1) hematoma no local da punção arterial; 2) sangramento; 3) correção cirúrgica da complicação vascular; 4) reação vasovagal.

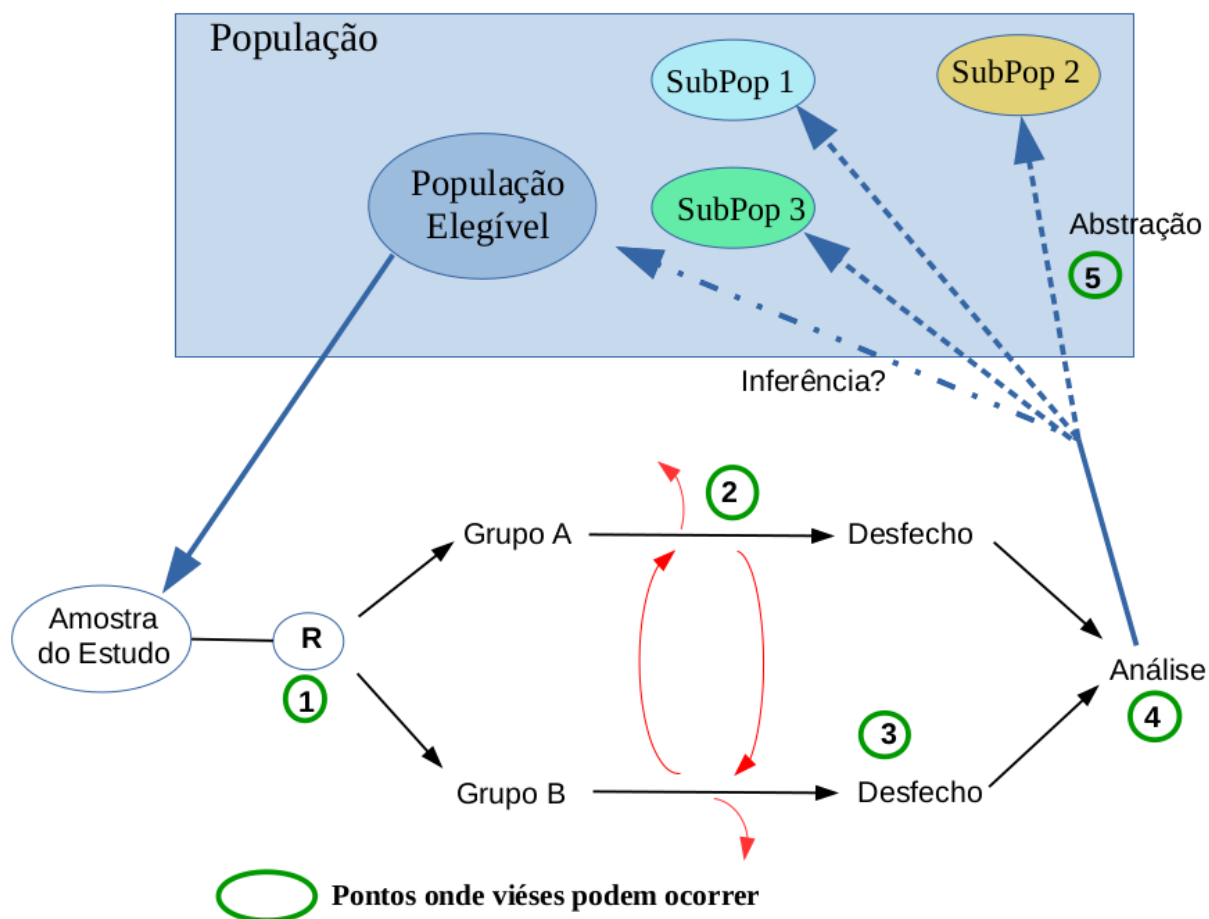


Figura 5.9: Diagrama de um ensaio controlado randomizado paralelo.

Nesse estudo, a população-alvo são *pacientes submetidos a cateterismo cardíaco diagnóstico, utilizando introdutores arteriais 6 F*. Em geral nem todos os membros dessa população alvo

podem ou devem participar do estudo. Para isso, são estabelecidos critérios de inclusão no estudo, os quais definem a população elegível do estudo. No ensaio de Rocha et al., a população elegível são os *pacientes submetidos a cateterismo cardíaco diagnóstico, com idade > 18 anos, de ambos os sexos, excluindo-se pacientes em uso de anticoagulantes, com obesidade mórbida e história de discrasias sanguíneas, doenças da aorta ou com hipertensão arterial grave não-controlada*.

Da população de elegíveis, uma amostra é extraída por algum método de amostragem. Em geral amostras de conveniência são extraídas, por exemplo, pacientes que são tratados na unidade (ou unidades de saúde) onde o ensaio é realizado. No estudo de Rocha et al., a amostra do estudo consistiu de *406 pacientes submetidos a cateterismo cardíaco eletivo, por via femoral, utilizando cateteres e bainha 6 F, no período de agosto de 2007 a novembro de 2008, em um laboratório de hemodinâmica em Santa Maria (RS)*.

A amostra de pacientes é então distribuída de **maneira aleatória (randomização)** entre as intervenções que estão sendo comparadas. Cada intervenção corresponde a um grupo (braço) do ensaio. O estudo de Rocha et al. avaliou duas intervenções: 200 pacientes foram alocados para o grupo de intervenção (deambulação após três horas após a remoção da bainha) e 206 pacientes foram alocados ao grupo controle (deambulação após seis horas após a remoção da bainha).

Após a randomização, os pacientes são submetidos aos respectivos tratamentos e acompanhados pelo tempo de duração do estudo para avaliar a ocorrência e ou realizar mensurações dos desfechos previstos no protocolo do estudo. No exemplo em questão, os pacientes foram acompanhados por meio de um *contato telefônico, realizado pela enfermeira em 24, 48 e 72 horas após a alta, no qual era questionado o aspecto do local da punção (presença de edema, sangramento, hematoma ou outra reação)*.

Após a coleta de dados, os mesmos são analisados por meio de técnicas estatísticas apropriadas ao tipo de estudo e de acordo com as escalas de medidas das variáveis analisadas. Quando as variáveis que definem as intervenções realizadas e o desfecho clínico são categóricas binárias, os resultados podem ser organizados em uma tabela 2x2, como mostra a figura 5.10: *a* representa o número de indivíduos expostos ao tratamento A e que tiveram o desfecho clínico de interesse, *c* representa o número de indivíduos expostos ao tratamento A e que não tiveram o desfecho clínico de interesse, *b* representa o número de indivíduos expostos ao tratamento B e que tiveram o desfecho clínico de interesse e *d* representa o número de indivíduos expostos ao tratamento B e que não tiveram o desfecho clínico de interesse.

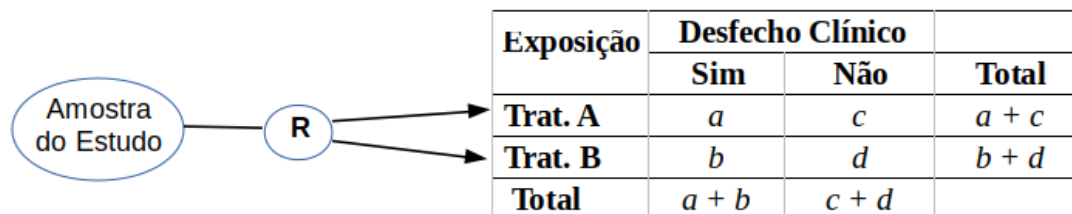


Figura 5.10: Tabela 2x2 de um ensaio controlado randomizado paralelo.

Como exemplo, no estudo de Rocha et al., os resultados são mostrados na tabela 5.1 para o desfecho clínico *ocorrência de reação vasovagal*.

Tabela 5.1: Versão simplificada da tabela 3 do estudo de Rocha et al. (Rocha et al., 2009) (CC BY-NC).

Intervenção	Reação vasovagal		Total
	Sim	Não	
<i>Deambulação após 3 horas</i>	<i>4</i>	<i>196</i>	<i>200</i>
<i>Deambulação após 6 horas</i>	<i>7</i>	<i>199</i>	<i>206</i>
<i>Total</i>	<i>11</i>	<i>395</i>	

Se a amostra de pacientes do estudo for uma amostra elatória da população de pacientes elegíveis para o estudo, os resultados podem ser aplicados por inferência estatística à população elegível para o estudo. Porém geralmente a amostra de pacientes do estudo é uma amostra de conveniência da população elegível, como no exemplo acima. Assim, em geral, a aplicação dos resultados do estudo à população elegível e a outras populações que não a elegível não é uma questão de inferência estatística; é mais uma questão de juízo clínico, que pode ou não corresponder à realidade. Essa é mais uma das razões para se realizar o monitoramento dos efeitos de uma tecnologia após a sua difusão.

Desde o início do estudo até a análise final dos resultados, diversos vieses podem ser introduzidos no estudo: falhas no método de randomização, não mascaramento da alocação, não mascaramento dos pacientes e profissionais em relação aos tratamentos aplicados, não mascaramento da avaliação de desfechos, dados de desfecho incompletos, publicação seletiva de resultados (Higgins et al., 2011), etc.

A randomização em um ensaio controlado visa a eliminar o viés de seleção de pacientes. A randomização tende a equilibrar os fatores que podem interferir no desfecho clínico entre os grupos de estudo, sejam eles conhecidos ou não. Esse equilíbrio tende a ser maior à medida que o tamanho amostral aumenta. Por essa razão, o ensaio controlado randomizado é considerado um **estudo experimental**. Assim é importante que o método de randomização seja livre de manipulação por parte dos integrantes do estudo e que o ato da randomização seja separado das pessoas que recrutam os pacientes (mascaramento da alocação).

Deve-se evitar ou reduzir ao mínimo possível o abandono de pacientes do estudo ou a migração de pacientes de um grupo para outro ao longo do estudo (círculo com 2 na figura 5.9). Um abandono significativo de pacientes pode afetar os resultados, se o abandono estiver associado a efeitos dos tratamentos em estudo. Também é desejável que os pacientes e os profissionais de saúde não conheçam que tratamento estão recebendo ou aplicando, quando isso for possível e especialmente na avaliação de desfechos clínicos, principalmente quando essa avaliação é subjetiva como, por exemplo, percepção da dor ou qualidade de vida. Finalmente técnicas estatísticas apropriadas devem ser empregadas para analisar os dados gerados no estudo e os autores devem publicar todos os resultados especificados no protocolo do estudo e não somente aqueles que são favoráveis ao tratamento em estudo.

## 5.6 Ensaio controlado não randomizado

Os conteúdos desta seção e da seção seguinte (5.7) podem ser visualizados neste [vídeo](#).

Em um ensaio controlado não randomizado, como o nome indica, a alocação dos pacientes aos grupos de estudo não é aleatória. Um exemplo de ensaio controlado não randomizado é o estudo de Andrade et al. (Andrade et al., 2018a) que comparou a intensidade de sangramento de procedimentos odontológicos em pacientes anticoagulados com Varfarina (grupo I) ou Dabigatrana (grupo II).

Diversos estudos mostraram que, em geral, ensaios controlados não randomizados tendem a produzir resultados mais favoráveis à inovação do que ensaios controlados randomizados (Bero and Rennie, 1996).

## 5.7 Série de casos

Numa série de casos, não há um grupo controle. Um conjunto de pacientes é submetido a um tratamento e acompanhado por um período de tempo para se avaliar os efeitos do tratamento (figura 5.11).

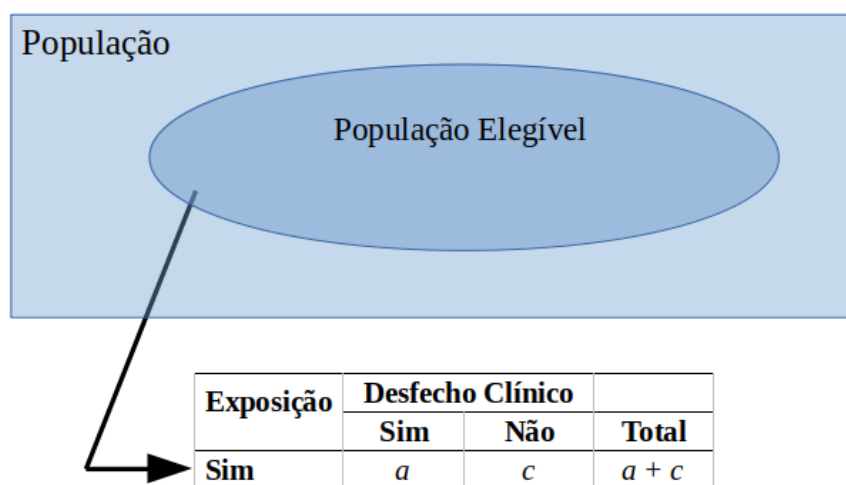


Figura 5.11: Diagrama de um estudo de série de casos.

Pela ausência do grupo controle, as evidências de uma série de casos são fracas, já que não é possível comparar os efeitos do tratamento com um outro tratamento alternativo ou placebo, ou mesmo a ausência de tratamento. Por outro lado, uma série de casos pode sugerir tratamentos para serem avaliados em estudos mais rigorosos.

O estudo de Serpa et al. (Serpa et al., 2014) avaliou parâmetros de resposta à terapia anti-IgE com omalizumabe em pacientes com asma de difícil controle. Foram avaliados 24 pacientes com asma de difícil controle, em uso de omalizumabe há pelo menos 32 semanas e considerados como respondentes à terapia. Avaliaram-se a pontuação do teste de controle de

asma (TCA), a presença de sintomas de asma, a frequência de uso de  $\beta_2$  agonista de curta ação, as doses de corticoide inalatório e oral e o percentual previsto do volume expiratório forçado no 1º minuto (VEF1), antes e com 16 e 32 semanas de tratamento.

## 5.8 Estudo de coortes

Os conteúdos desta seção e das seções 5.9 e 5.10 podem ser visualizados neste [vídeo](#).

Quando se deseja estudar a associação entre possíveis fatores etiológicos (álcool, fumo, radiação, por exemplo) e desfechos clínicos (cirrose, câncer de pulmão, leucemia), não é ético expor deliberadamente pessoas a esses agentes. Nesses casos, deve-se recorrer a estudos observacionais, onde os fatores são estudados comparando-se populações que são expostas ao fator (por opção, tipo de ocupação, acidente, etc.) com populações que não foram expostas.

O estudo de coortes é um estudo longitudinal onde pessoas expostas a dois ou mais níveis de um fator de exposição e inicialmente não apresentando o desfecho (ou desfechos) clínico de interesse são acompanhadas ao longo do tempo para verificar a ocorrência ou não dos desfechos clínicos estudados em algum instante posterior. Um estudo de coortes pode ser retrospectivo, prospectivo ou uma combinação dos dois.

Basicamente, dois métodos de amostragem são utilizados em estudos de coortes. O primeiro é mostrado na figura 5.12 onde uma amostra é extraída da população elegível para o estudo e, então, as unidades de observação são classificadas nos diferentes níveis de exposição ao fator(es) estudado(s) e as ocorrências de desfechos clínicos são então avaliadas. Quando as variáveis de exposição e desfecho são binárias, a análise dos dados levaria a uma tabela 2x2 como mostrada na figura 5.12.

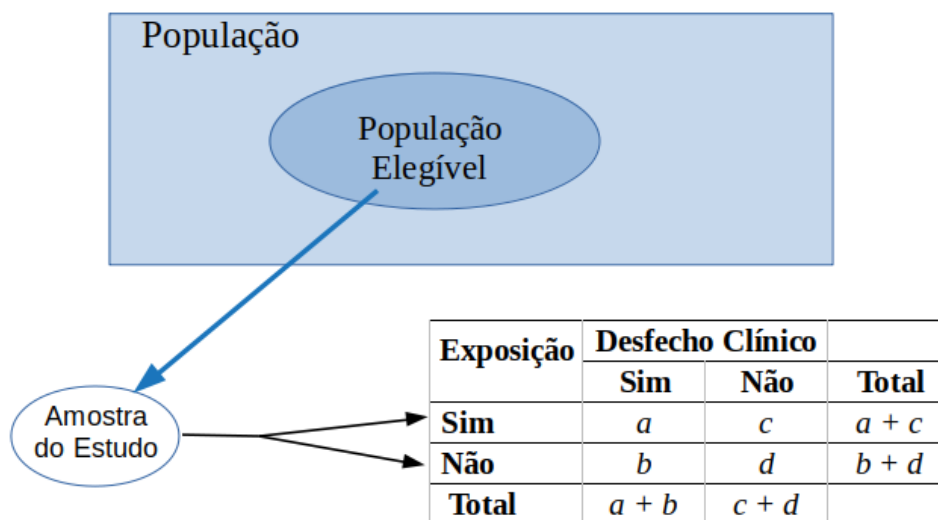


Figura 5.12: Diagrama de um estudo de coortes.

O estudo de Pereira et al. (Pereira et al., 2015) ilustra um estudo de coortes onde uma amostra de conveniência de uma população foi extraída e foram analisadas diversas associações entre

fatores de risco e o desfecho clínico infecção do sítio cirúrgico. A amostra foi composta de 432 pacientes submetidos à cirurgia eletiva para correção de fratura de fêmur. Entre os fatores de risco analisados foram: tempo de internação pré-operatório e ocorrência de acidente vascular cerebral.

Outro possível método de amostragem é mostrado na figura 5.13 onde uma amostra de pessoas expostas e outra amostra de pessoas não expostas ao fator em estudo são extraídas separadamente da população elegível para o estudo e as ocorrências de desfechos clínicos são então avaliadas. Quando as variáveis de exposição e desfecho são binárias, a análise dos dados nos levaria a uma tabela 2x2 semelhante à mostrada na figura 5.12.

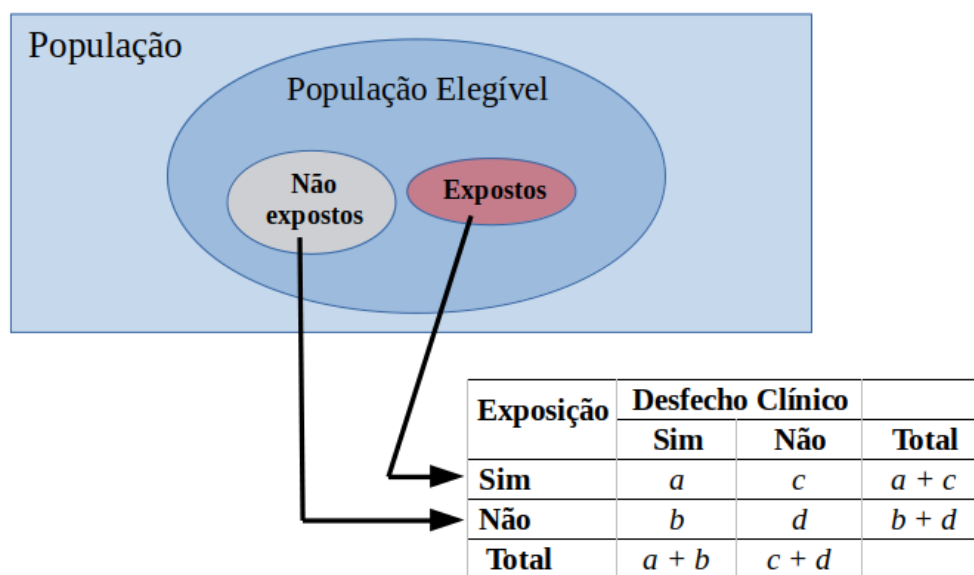


Figura 5.13: Diagrama de um estudo de coortes com grupos expostos e não-expostos.

Um estudo de coortes está sujeito aos mesmos tipos de vieses de um ensaio controlado não randomizado, porque, essencialmente, os grupos de estudo não são alocados aleatoriamente. Quando o desfecho clínico estudado não é frequente e ainda requer um longo tempo desde a exposição para ser detectado, uma amostra grande deve ser coletada, o que aumenta o tempo e custo do estudo e pode levar à perda significativa de pacientes ao longo do tempo em um estudo de coortes prospectivo. Em coortes retrospectivos, o tempo de realização do estudo pode ser reduzido, mas outros problemas podem ocorrer como, por exemplo, limitação dos dados ou baixa qualidade dos dados que podem ser coletados em prontuários ou outra fonte de dados utilizada no estudo.



## 5.9 Estudo de caso-controle

Em um estudo de caso-controle, uma amostra de pacientes que apresentam o desfecho clínico de interesse (casos) é extraída da população elegível. Em seguida, uma outra amostra da população elegível (controle) é extraída para verificar a distribuição do fator de exposição nessa população (figura 5.14). Dependendo do tipo de estudo caso-controle, a amostra de controles pode ou não incluir casos. Observem que o processo de amostragem é diferente daquele utilizado no estudo de coortes.

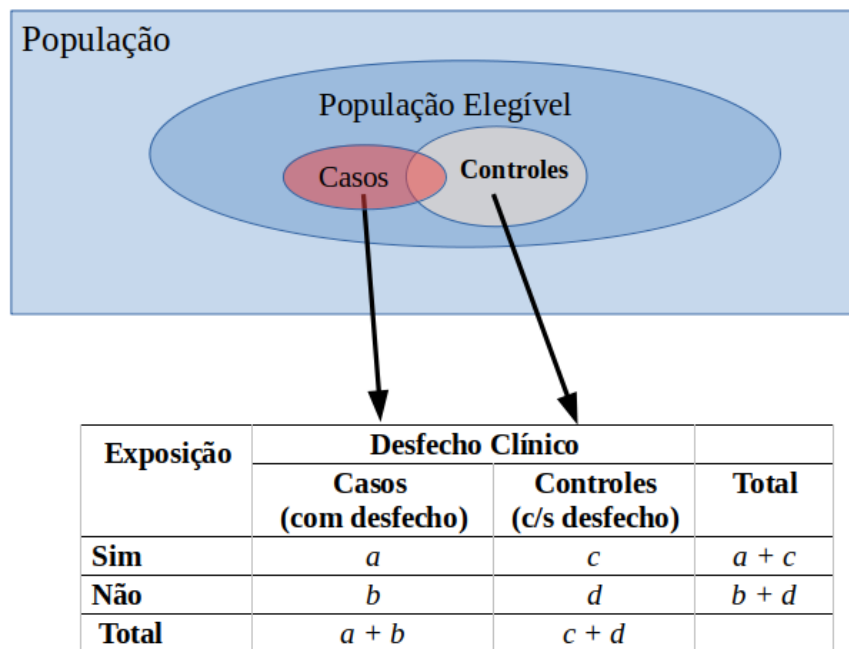


Figura 5.14: Diagrama de um estudo de caso-controle.

Por partir de uma amostra de casos, o estudo de caso-controle pode ser bem mais eficiente do que um estudo de coortes, no sentido de que os tamanhos amostrais tendem a ser menores, assim como o tempo de realização do estudo. Por outro lado, um estudo de caso-controle deve ser conduzido com muito cuidado, porque ele está sujeito a um grande número de vieses.

O estudo de Medeiros et al. (Medeiros et al., 2003) é um exemplo de um estudo de caso-controle que teve como objetivo estudar a relação entre a exposição precoce ao leite de vaca e a ocorrência de diabetes mellitus tipo 1 entre menores de 18 anos atendidos no Hospital Universitário Alcides Carneiro, em Campina Grande. A amostra foi constituída por 128 indivíduos de ambos os sexos, sendo 64 mães de portadores de diabetes mellitus e 64 mães de controles não portadores de diabetes mellitus.

A tabela 5.2 mostra a tabela 2x2 resultante do estudo.

Tabela 5.2: Reprodução da tabela 5 do estudo de Medeiros et al. (Medeiros et al., 2003) (CC BY-NC).

Exposição	Diabetes		Total
	Sim	Não	
<i>Aleitamento materno exclusivo <math>\geq 4</math> meses</i>	10	23	33
<i>Exposição precoce ao leite de vaca <math>&lt; 4</math> meses</i>	54	41	95
<i>Total</i>	64	64	

## 5.10 Estudo transversal

Um estudo transversal é um estudo observacional no qual é realizada uma única medição das variáveis clínicas dos pacientes que compõem a amostra do estudo, geralmente em instantes próximos. A figura 5.15 ilustra um processo típico de amostragem em um estudo transversal e a montagem de uma tabela 2x2 que relaciona duas variáveis categóricas binárias.

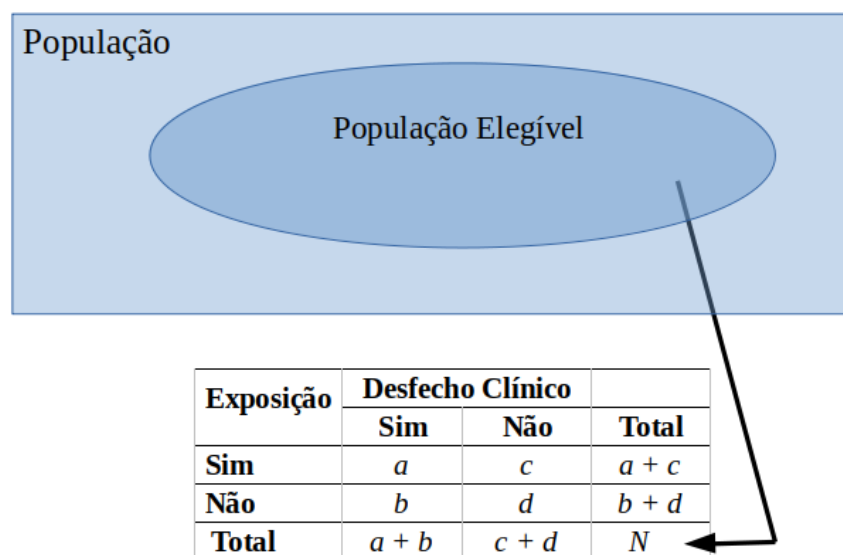


Figura 5.15: Diagrama de um estudo transversal.

Um exemplo de estudo transversal é o estudo de Souza, Noblat e Santos (Souza et al., 2015). Esse estudo avaliou 49 pacientes maiores de 18 anos, portadores de asma grave não controlada ou asma refratária, atendidos em um ambulatório especializado do Sistema Único de Saúde, em uso regular de altas doses de corticoides inalatórios (CIs) e/ou de diversos medicamentos e com comorbidades. Obtiveram-se as medidas de qualidade de vida por meio da aplicação do questionário *Asthma Quality of Life Questionnaire* (AQLQ) num único momento. O escore global e dos domínios do AQLQ foram relacionados com variáveis demográficas (gênero e

idade), escore do *Asthma Control Questionnaire*, terapia medicamentosa (dose inicial de CI, dispositivos inalatórios e politerapia) e comorbidades.

Em um estudo transversal é mais difícil determinar a partir de uma associação de variáveis, que variáveis compõem a causa e que variáveis são consequências, já que, em geral, as medidas efetuadas refletem um mesmo momento na história dos pacientes.

## 5.11 Revisão sistemática e metanálise

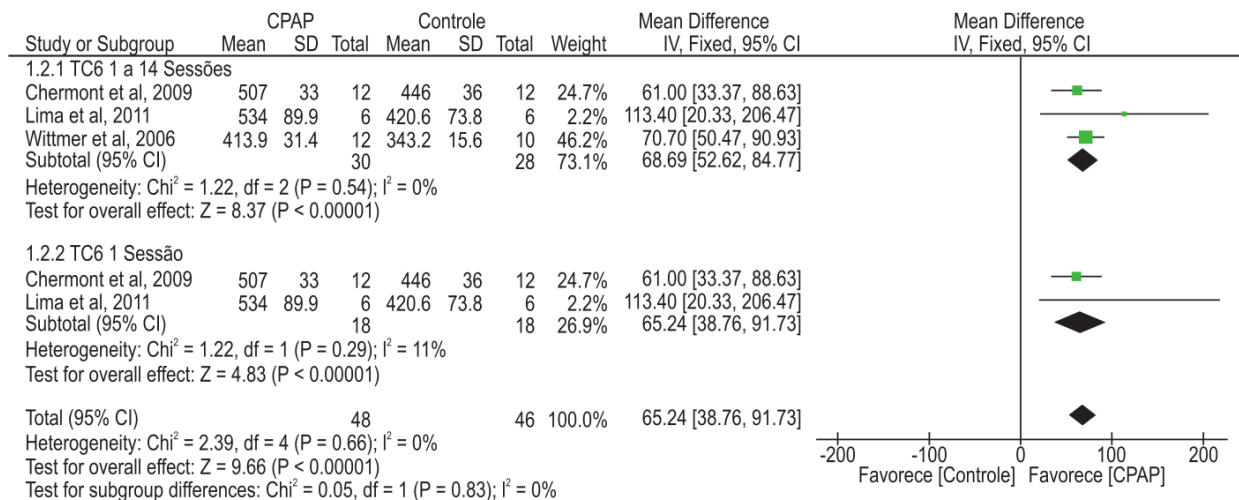
Os conteúdos desta seção e da seção seguinte (5.12) podem ser visualizados neste [vídeo](#).

Em geral um único estudo não fornece a evidência definitiva sobre um determinado tema. Frequentemente vários estudos são realizados sobre um mesmo assunto em locais diferentes e ao longo do tempo. Não necessariamente esses estudos são totalmente idênticos em relação à população estudada, os tratamentos avaliados, níveis dos fatores de exposição ou desfechos.

A revisão narrativa é um trabalho de revisão de um determinado tema na qual os autores se baseiam em estudos anteriores e em suas experiências, sendo a seleção de estudos, apresentação dos argumentos e resultados realizadas de maneira subjetiva. Desse modo, as revisões narrativas tendem a refletir as opiniões, vieses e preconceitos de seus autores.

A revisão sistemática busca uma síntese de estudos semelhantes ou explicar possíveis divergências por meio de um método reprodutível e rigoroso. Caso os resultados dos estudos possam ser integrados por meio de uma análise estatística, a revisão sistemática é chamada de metanálise. A revisão sistemática pode resolver conflitos de estudos individuais e fornecer estimativas mais precisas do que os estudos individuais.

A figura 5.16 mostra o resultado de uma metanálise de ensaios controlados randomizados que estudaram a influência da ventilação não invasiva (VNI) sobre a capacidade funcional de pacientes com insuficiência cardíaca. Quatro estudos foram incluídos na metanálise. Três dos estudos optaram pela CPAP para a aplicação da VNI e um estudo utilizou CPAP e a pressão suporte (PS). A figura mostra os efeitos do CPAP x placebo sobre a distância do teste de caminhada de seis minutos (TC6) para cada estudo individualmente e os efeitos combinados por meio da metanálise.



**FIG. 2. Mudança no TC6 – CPAP versus Controle**

Figura 5.16: Resultados de um estudo de metanálise. Fonte: (Bittencourt et al., 2017) (CC BY).

## 5.12 Gradação da evidência científica

Como visto nas seções anteriores, foram desenvolvidos diversos delineamentos de estudos para fornecer evidências sobre efetividade de tratamentos, qualidade de testes diagnósticos, prognóstico e fatores etiológicos de doenças. Todos os estudos estão sujeitos a vieses em maior ou menor grau. Diversas propostas têm sido apresentadas sobre como graduar a evidência científica em função da qualidade e do delineamento do estudo. A figura 5.17 mostra uma pirâmide de evidência, adaptada do trabalho de Hadorn et al. (Hadorn et al., 1996).

Essa pirâmide distingue 6 níveis de evidência, sendo os níveis mais fortes situados no topo da pirâmide:

- Nível 1:
  - ECRs bem conduzidos de grande porte
  - Metanálise de ECRs envolvendo no total um grande número de pacientes
- Nível 2:
  - ECRs bem conduzidos de pequeno porte
  - Metanálise de ECRs de pequeno porte
- Nível 3:
  - Estudos de coortes bem conduzidos
  - Metanálise de estudos de cortes

- Nível 4:
  - Estudos de caso-controle bem conduzidos
- Nível 5: Estudos pobremente controlados ou não controlados:
  - ECRs com falhas metodológicas graves
  - Estudos observacionais com alto potencial de vieses
  - Série de casos
- Nível 6:
  - Opinião de Especialistas

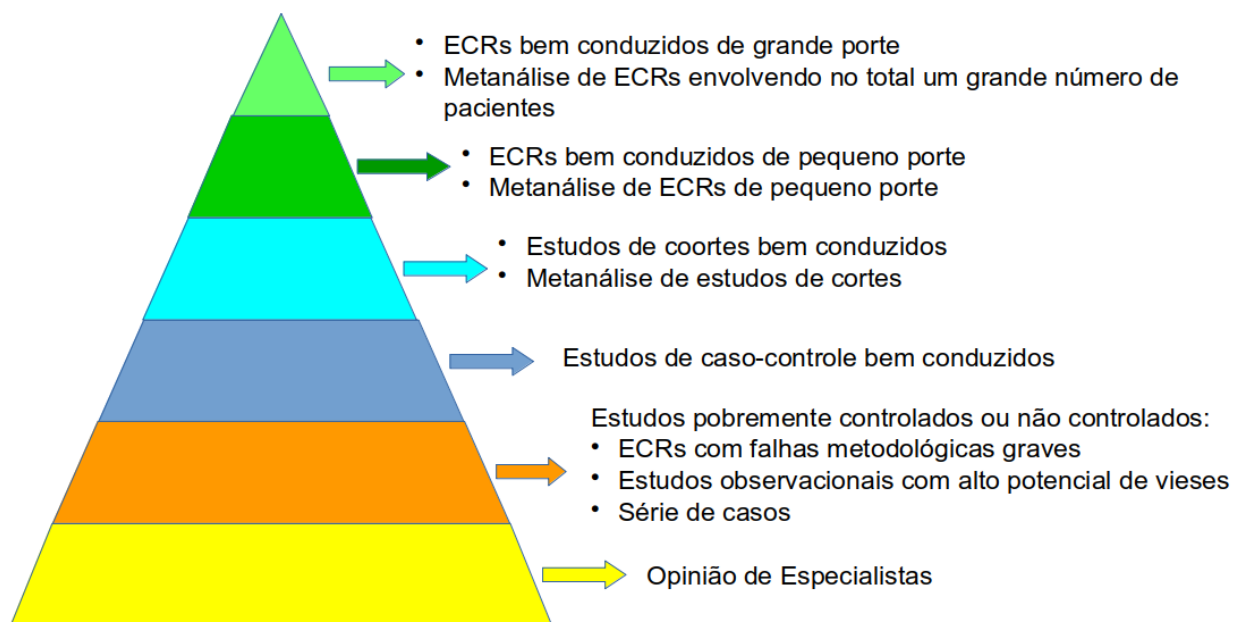


Figura 5.17: Pirâmide de evidência da pesquisa clínica-epidemiológica.

Uma ligeira busca na internet irá mostrar diversas outras propostas, com diferentes nuances. Em geral as propostas concordam em considerar os estudos mais robustos metodologicamente (ECR e Revisões Sistemáticas) com evidências mais fortes, ficando os estudos de coortes e caso-controle no meio do caminho.

O livro de Guyatt et al. (Guyatt et al., 2008) é uma excelente referência sobre como utilizar a literatura médica para apoiar a prática clínica.

## 5.13 Exercícios

1) Dê a classificação do desenho para os estudos abaixo.

- a) 61 pacientes submetidos à remoção de dispositivos eletrônicos cardíacos implantáveis (DCEI) em um hospital quaternário brasileiro foram avaliados para descrever a taxa de sucesso e complicações da remoção percutânea de DCEI. Fonte: (Nubila et al., 2021) ([CC BY](#)).
- b) O presente estudo usa informações do Nord-Trøndelag Health Study (HUNT) na Noruega, onde todos os residentes com 20 anos ou mais foram convidados para participar. Todos os participantes responderam a um questionário abrangente e submetidos a exame clínico. Os dados coletados incluíram informações sobre estilo de vida e fatores relacionados à saúde, como medidas auto-relatadas de atividade física, tabagismo, consumo de álcool e nível de educação, bem como medições clínicas da pressão sanguínea, níveis de glicose no sangue, frequência cardíaca, altura e peso. Informações sobre a variável de desfecho, se um indivíduo desenvolveu diabetes após 11 anos de acompanhamento, foram obtidas a partir da questão “Você tem ou já teve diabetes?”. Esse diagnóstico de diabetes, baseado em um questionário, foi validado por comparação com registros médicos e foram verificados em 96% dos casos. Fonte: (Hjerkind et al., 2017) ([CC BY-NC](#)).
- c) Uma amostra de 33 estudantes foram alocados mediante sorteio aleatório nos grupos intervenção (n=17) e controle (n=16). A intervenção avaliada foi o ensino da técnica de aspiração de vias aéreas inferiores, utilizando a simulação (oficinas individuais e debriefing), e o controle foi o ensino da técnica pelo método tradicional (aula expositiva e treinamento em grupo). Fonte: (Salgado et al., 2018) ([CC BY](#)).
- d) 1.139 crianças e adolescentes de ambos os sexos entre 6 e 18 anos foram avaliados para analisar a capacidade preditiva dos indicadores antropométricos e os seus valores de corte para a triagem da dislipidemia em crianças e adolescentes. O peso corporal, estatura, circunferência da cintura (CC) e prega cutânea subescapular (PCSE) e prega cutânea tricipital (PCT) foram medidos. O índice de massa corporal (IMC) e a relação cintura-estatura (RCE) foram calculados. As crianças e os adolescentes que tinham pelo menos uma das seguintes alterações lipídicas foram definidos como tendo dislipidemia: elevados níveis de colesterol total, HDL-C baixo, LDL-C elevado e concentração elevada de triglicérides. Uma curva ROC (Receiver Operating Characteristics) foi construída e a área sob a curva, a sensibilidade e a especificidade foram calculadas para os parâmetros analisados. Fonte: (Quadros et al., 2015) ([CC-BY-NC-ND](#)).
- e) Estudos epidemiológicos recentes demonstraram que alterações na microbiota e seus metabólitos estão associadas à hipertensão arterial sistêmica. A *Helicobacter pylori* (*H. pylori*) é um dos patógenos bacterianos mais comuns, e a possível associação entre a infecção por *H. pylori* e a hipertensão é controversa. Este estudo teve o objetivo de esclarecer a associação entre eles e proporcionar uma nova base teórica para detectar a patogênese da hipertensão. Foram selecionados

estudos caso-controle e transversais sobre a associação entre *H. pylori* e hipertensão, publicados de 1996 a 2019 indexados nos bancos de dados PubMed, Google Scholar, Chinese Wan Fang Data e Chinese National Knowledge Infrastructure (CNKI). As razões de chance (RC) combinadas e o intervalo de confiança (IC) 95% foram estimados. O  $I^2$  foi realizado para avaliar a heterogeneidade estatística. O viés de publicação foi avaliado utilizando-se os testes de Beggs e de Egger. Fonte: (Huang et al., 2021) (CC BY).

- f) Esse estudo foi conduzido para identificar a possível associação entre valores laboratoriais, comorbidades, tratamento farmacológico, alterações hemodinâmicas, resultado da diálise e alterações estabilométricas com uma maior probabilidade de quedas em pacientes de hemodiálise. Foram analisados os casos de pacientes de uma unidade de hemodiálise que sofreram uma ou mais quedas. Os controles foram pacientes da mesma unidade que não sofreram quedas. Os dados foram obtidos a partir do histórico clínico dos pacientes e, também, de um teste de equilíbrio realizado seis meses antes nesses pacientes. Fonte: (Perez-Gurbindo et al., 2021) (CC BY).
  - g) Simulações estão se tornando amplamente utilizadas na educação médica, mas há poucas evidências de sua eficácia nos cuidados neurocríticos. Como o AVC agudo é uma emergência neurológica que exige atenção imediata, ele é um candidato promissor para o treinamento de simulação. Esse estudo foi realizado para avaliar o impacto de um curso de simulação realista de AVC na autopercepção dos médicos quanto à confiança no manejo do AVC agudo. Para a nossa intervenção, participaram 17 profissionais de saúde em um curso de simulação realista de acidente vascular cerebral. Como controles, os participantes foram escolhidos a partir de uma amostra de conveniência dos participantes dos cursos Suporte Neurológico de Vida de Emergência (18 participantes) e Neurosonologia (20 participantes). Todos os participantes responderam questionários pré e pós-teste, avaliando suas autopercepções de confiança no cuidado do AVC agudo, variando de 10 a 50 pontos. Fonte: (Farias da Guarda et al., 2021) (CC BY).
- 2) Por que a maioria dos estudos clínico-epidemiológicos utilizam amostras de conveniência? Que implicação isso traz para as inferências realizadas a partir dos resultados de cada estudo?
  - 3) Qual a diferença entre revisão sistemática e revisão narrativa? E entre metanálise e revisão sistemática?
  - 4) Como é, em geral, a gradação do nível da evidência dos estudos clínico-epidemiológicos?

# Capítulo 6

## Introdução à Inferência Estatística

### 6.1 Introdução

Neste capítulo, serão apresentados os conceitos básicos de teste de hipótese, valor de  $p$  e intervalo de confiança a partir de dados de um estudo prospectivo que avalia os níveis de ácido fólico em três grupos de pacientes. Esses conceitos serão ilustrados por meio de um teste de randomização, que não faz nenhuma suposição sobre a distribuição dos dados, de modo que nenhum conhecimento prévio de distribuição de probabilidades será necessário para a compreensão dos conceitos. Capítulos posteriores irão aprofundar esses temas. A forma como a randomização foi utilizada para realizar o teste de hipótese e calcular o intervalo de confiança neste capítulo foi inspirada em Manly (Manly, 1997).

### 6.2 Apresentação de resultados de estudos

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

A figura 6.1 mostra os resultados de um estudo de coortes realizado para verificar a influência do diabetes mellitus sobre a perviabilidade da fístula arteriovenosa para hemodiálise (Cruz et al., 2015). Os pacientes foram divididos em dois grupos: 26 pacientes com diabetes mellitus e 66 pacientes sem diabetes mellitus. Diversos desfechos foram avaliados e comparados entre os dois grupos. Dois dos desfechos foram: 1) se houve uma oclusão precoce (variável binária); 2) o tempo (meses) até a oclusão (variável numérica contínua). Para a primeira variável, os autores apresentaram o número de pacientes que tiveram oclusão precoce em cada grupo com o respectivo percentual: 20 (40,82%) no grupo de diabéticos e 50 (41,82%) no grupo dos não diabéticos. Para a segunda variável, os autores apresentaram o tempo médio de oclusão em cada grupo com o respectivo desvio padrão, 9,03 (11,6) no grupo de diabéticos e 15,97 (27,92) no grupo dos não diabéticos. Para cada variável de desfecho, os autores realizaram um teste de hipótese para verificar se as diferenças observadas entre os dois grupos são estatisticamente significativas. Na última coluna, os autores apresentam o valor de  $p$  obtido em cada teste. Esses termos serão explicados mais adiante.



**Tabela 3.** Análise das fístulas arteriovenosas previamente ocluídas.

Variável	Grupo DM (n = 26)	Grupo NDM (n = 66)	Valor de p
Número de oclusões	49	121	
Oclusão precoce*	20 (40,82%)	50 (41,82)	1,0000
Oclusão tardia†	29 (59,18%)	71 (58,68%)	1,0000
Tempo médio até oclusão (meses)	9,03 (± 11,60)	15,97 (± 27,92)	0,0952
Tempo médio até oclusão (meses) para acessos de oclusão tardia	14,62 (± 12,39)	26,15 (± 32,58)	0,0338
Sobrevida FAVs com oclusão tardia	n = 29	n = 71	
Perviedade em 12 meses	15 (51,72%)	38 (53,52%)	1,0000
Perviedade em 24 meses	5 (17,24%)	28 (39,44%)	0,0300

FAV - fístulas arteriovenosas. \*Oclusão do acesso antes que se inicie o uso para hemodiálise (falência primária). †Oclusão de acessos que foram utilizados com sucesso para hemodiálise.

Figura 6.1: Exemplo de apresentação de resultados de testes de hipóteses para diversos tipos de variáveis. Fonte: tabela 3 do estudo de (Cruz et al., 2015) (CC BY).

O estudo de Haijanen et al. (Haijanen et al., 2019) realiza uma comparação de custos do tratamento com antibióticos x apendicetomia para o tratamento da apendicite aguda sem complicações. Parte dos resultados estão mostrados na figura 6.2. Por exemplo, para custos hospitalares em 5 anos de acompanhamento, os autores apresentaram o custo médio para cada grupo de tratamento (2730 x 2056 euros), bem como a diferença de custos entre os dois grupos (674). Ao lado de cada custo, foram mostrados entre parênteses os intervalos de confiança ao nível de 95%. Para a diferença de custos hospitalares em 5 anos de acompanhamento, o intervalo de confiança é dado pelo intervalo [465, 883]. Na última coluna, os autores apresentam o valor de p resultante dos testes de hipóteses para verificar a significância estatística da diferença em cada desfecho analisado. Para todos os desfechos da tabela, o valor de p foi menor que 0,001.

**Table 1.** Mean hospital charges, productivity losses and overall costs in Euros per patient for appendectomy and antibiotic therapy group patients with uncomplicated acute appendicitis at five-year follow-up.

	Appendectomy Group € (95% CI, €)	Antibiotic therapy Group € (95% CI, €)	Difference € (95% CI, €)	p<
<b>One-year follow-up</b>				
Hospital charges	2718 (2636–2799)	1707 (1547–1865)	1010 (835–1186)	0.001
Productivity losses	2962 (2806–3118)	1845 (1712–1976)	1117 (911–1322)	0.001
Overall costs	5680 (5489–5872)	3552 (3334–3769)	2127 (1840–2417)	0.001
<b>Five-year follow-up</b>				
Hospital charges	2730 (2645–2817)	2056 (1861–2251)	674 (465–883)	0.001
Productivity losses	2986 (2822–3149)	2115 (1950–2280)	871 (639–1104)	0.001
Overall costs	5716 (5510–5925)	4171 (3879–4463)	1545 (1193–1899)	0.001

<https://doi.org/10.1371/journal.pone.0220202.t001>

Figura 6.2: Apresentação do valor de p e intervalo de confiança para a diferença de custos entre dois tratamentos para apendicite aguda. Fonte: tabela 1 do estudo de (Haijanen et al., 2019) (CC BY).

Comparando as duas figuras anteriores, é possível verificar que a figura 6.2 apresenta além do valor de p, a diferença dos desfechos com o respectivo intervalo de confiança. Veremos mais adiante que o intervalo de confiança é mais informativo do que o valor de p.

A figura 6.3 mostra os resultados de um estudo de caso-controle sobre a exposição precoce ao leite de vaca e ocorrência de Diabetes Mellitus tipo 1 (Medeiros et al., 2003). 64 diabéticos (casos) e 64 não diabéticos (controles) foram avaliados para verificar se houve exposição precoce ao leite de vaca (antes de 4 meses). 54 pacientes diabéticos foram expostos precocemente ao leite de vaca contra 41 no grupo controle. A razão de chances para esse estudo é igual 3,03 com um intervalo de confiança ao nível de 95% igual a [1,21, 7,72]. O valor de p foi igual 0,01.

**Tabela 5**

Comparação entre o tempo de exposição precoce ao leite de vaca nos grupos estudados.

	Grupo de diabéticos (n = 64)		Grupo controle (n = 64)	
	n	%	n	%
Aleitamento materno exclusivo ≥ 4 meses	10	15,6	23	35,9
Exposição precoce ao leite de vaca ** < 4 meses	54*	84,4	41*	64,1
Total	64	100,0	64	100,0

\* ( $\chi^2 = 5,9$ ; p = 0,01), \*\* Odds ratio = 3,03 (IC95%: 1,21 - 7,72; p = 0,01) entre diabéticos e controles

Figura 6.3: Exemplo de um teste de hipótese e intervalo de confiança para a razão de chances. Fonte: tabela 5 do estudo de (Medeiros et al., 2003) ([CC BY-NC](#)).

O estudo de Kho et al. (Kho et al., 2019) é um ensaio controlado randomizado piloto multicêntrico que comparou a ergometria na cama juntamente com fisioterapia x fisioterapia somente em pacientes ventilados mecanicamente. Alguns resultados são mostrados na figura 6.4 em três instantes diferentes: ao acordar na UTI (*ICU Awakening*), após alta da UTI (*ICU discharge*) e após alta hospitalar (*Hospital discharge*). As variáveis são numéricas (PFIT, Blinded PFIT-s, MRCSS, 30STS, MWT, *Quadriceps strength*), com exceção de MRCSS < 48, que é dicotômica. A legenda da tabela mostra o significado de cada sigla. Para as variáveis numéricas, para cada momento em que foram avaliadas, os resultados são apresentados como média e desvio padrão para cada grupo de estudo (ergometria x rotina), bem como a diferença de médias com o respectivo intervalo de confiança para as variáveis PFIT, Blinded PFIT-s e MRCSS. Já para a variável dicotômica MRCSS < 48, os resultados são mostrados como o número de pacientes que apresentaram o valor de MRCSS < 48 em cada instante e o respectivo percentual em parênteses. Também são apresentados o risco relativo e o respectivo intervalo de confiança em cada instante de avaliação.

### Electronic Supplement 3: Outcome measures

Outcome	ICU Awakening <sup>a</sup>			ICU discharge <sup>b</sup>			Hospital discharge <sup>c</sup>		
Performance-based outcomes	Cycling	Routine		Cycling	Routine		Cycling	Routine	
PFIT-s, mean (SD)	4.4 (2.0)	4.4 (2.2)	0.03 (-1.2, 1.2)	5.7 (2.0)	6.0 (2.5)	-0.3 (-1.6, 1.0)	7.8 (1.9)	8.1 (1.7)	-0.3 (-1.5, 0.8)
Blinded PFIT-s, mean (SD)							7.8 (1.9)	8.1 (1.7)	-0.3 (-1.5, 0.9)
MRCSS, mean (SD)	42.8 (12.1)	43.5 (12.1)	-0.7 (-7.8, 6.4)	46.3 (12.1)	52.4 (5.2)	-6.1 (-11.4, -0.7)	53.7 (5.1)	53.4 (5.6)	0.4 (-8.5, 16.5)
MRCSS <48, n (%)	16/23 (69.6)	14/24 (58.3)	RR = 1.19 (0.77, 1.84)	13/28 (46.4)	3/16 (18.8)	RR = 2.48 (0.83, 7.40)	3/23 (13.0)	4/18 (22.2)	RR = 0.59 (0.15, 2.30)
30STS, median (IQR) # reps	1 (1-2)	1 (1-2)		1.5 (1-4)	1.5 (1-3)		4.5 (2-8)	5 (4-8)	
2MWT, median (IQR) meters				30.5 (24-56)	33.5 (20-49.5)		76 (51-116)	61 (60-90)	
Quadriceps strength, mean (SD), Newtons				102.3 (69.2)	108.9 (50.6)		117.2 (57.1)	147.1 (57.7)	

**Legend:** This table outlines the patients' ICU and hospital outcomes, and their performance-based and patient-reported outcomes recorded at ICU awakening, ICU discharge, and hospital discharge (as applicable). Abbreviations: ICU=Intensive care unit; CI=confidence interval; LOS=length of stay; IQR=interquartile range; PFIT-s=Physical Function ICU Test-scored (maximum=10; higher scores, better function); MRCSS=Medical Research Council sum score (maximum=60; higher scores, better strength); 30STS30=second sit to stand (more repetitions=better function); reps=repetitions; 2MWT=2 minute walk test (further distance=better function); Katz=Katz activities of daily living (maximum=6; higher scores, more independence in function)

<sup>a</sup>Sample size for assessments performed at ICU Awakening (Cycling, routine): PFIT-s (27, 25); MRC Sum Score and MRC total score <48, (23, 24); 30STS (15, 13)

<sup>b</sup>Sample size for assessments performed at ICU Discharge (Cycling, routine): PFIT-s (28, 20); MRC Sum Score and MRC total score <48, (28, 16); 30STS (24, 18), 2MWT (10, 8); quadriceps strength (15, 14); Katz ADLs (28, 21); IPAT (27, 22); PRFS-ICU (25, 19); EQ5D5L Utility (24, 19); EQ5D5L VAS (24, 18).

<sup>c</sup>Sample size for assessments performed at Hospital Discharge (Cycling, routine): PFIT-s (25, 18); PFIT-s, blinded assessors (20, 17); MRC Sum Score and MRC total score <48, (23, 18); 30STS (22, 17), 2MWT (18, 16); quadriceps strength (16, 14); Katz ADLs (26, 18); PRFS-ICU (23, 18); EQ5D5L Utility and EQ5D5L VAS (21, 17).

Figura 6.4: Exemplo de intervalos de confiança para a diferença de médias e o risco relativo para diversas variáveis. Fonte: adaptado do suplemento eletrônico 3 do estudo de (Kho et al., 2019) (CC BY-NC).

Os diversos exemplos acima mostram que o uso de testes de hipótese e o cálculo de intervalos de confiança são comumente utilizados para apresentar os resultados de estudos clínico-epidemiológicos. Nas próximas seções, serão apresentados os conceitos de teste de hipótese, valor de p e intervalo de confiança, sem o uso de qualquer expressão matemática, utilizando um teste de randomização.

## 6.3 Teste de hipótese usando randomização

Os conteúdos das subseções desta seção e da seção 6.4 podem ser visualizados neste [vídeo](#), sendo recomendada a visualização prévia deste [vídeo](#).

### 6.3.1 Contexto do problema

Amess et al. (Amess et al., 1978) realizaram um estudo prospectivo, onde avaliaram os níveis de ácido fólico (microgramas por litro) nas células vermelhas em pacientes com *bypass* cardíaco que receberam três métodos diferentes de ventilação durante a anestesia:

- N2O+O2,24h: 50% de óxido nitroso e 50% de oxigênio, continuamente por 24 horas (8 pacientes);
- N2O+O2,op: 50% de óxido nitroso e 50% de oxigênio, somente durante a operação (9 pacientes);
- O2,24h: sem óxido nitroso, mas com 35%–50% de oxigênio por 24 horas (5 pacientes).

Os dados de cada paciente são mostrados abaixo.

##	folate	ventilation
## 1	243	N2O+O2,24h
## 2	251	N2O+O2,24h
## 3	275	N2O+O2,24h
## 4	291	N2O+O2,24h
## 5	347	N2O+O2,24h
## 6	354	N2O+O2,24h
## 7	380	N2O+O2,24h
## 8	392	N2O+O2,24h
## 9	206	N2O+O2,op
## 10	210	N2O+O2,op
## 11	226	N2O+O2,op
## 12	249	N2O+O2,op
## 13	255	N2O+O2,op
## 14	273	N2O+O2,op
## 15	285	N2O+O2,op
## 16	295	N2O+O2,op
## 17	309	N2O+O2,op
## 18	241	O2,24h
## 19	258	O2,24h
## 20	270	O2,24h
## 21	293	O2,24h
## 22	328	O2,24h

Vamos desconsiderar o grupo O2, 24h e verificar se, estatisticamente, existe alguma diferença entre os níveis de ácido fólico entre os grupos N2O+O2,24h e N2O+O2,op.

Estatisticamente, devemos conceber duas populações de pacientes com *bypass* cardíaco, uma delas sendo submetida ao tratamento N2O+O2,24h e outra ao tratamento N2O+O2,op. Os 8 pacientes submetidos ao tratamento N2O+O2,24h podem ser pensados como constituindo uma amostra aleatória extraída da população de pacientes com *bypass* cardíaco que seriam tratados com N2O+O2,24h. Analogamente, os 9 pacientes submetidos ao tratamento N2O+O2,op podem ser pensados como constituindo uma amostra aleatória extraída da população de pacientes com *bypass* cardíaco que seriam tratados com N2O+O2,op.

Que podemos inferir para as populações mais amplas de pacientes submetidos aos dois tratamentos a partir da análise das duas amostras do estudo?

Vamos abrir a aplicação [Teste de Hipótese e Intervalo de Confiança](#) (figura 6.5).

## Teste de Hipótese e Intervalo de Confiança

**Variável**

**Grupo 1**

**Grupo 2**

**Valores no Grupo 1**

**Valores no Grupo 2**

**Nível de confiança:**

Figura 6.5: Aplicação que permite realizar um teste hipótese para a comparação de médias e calcular o intervalo de confiança por meio da randomização.

O painel à esquerda da aplicação mostra a variável numérica que está sendo avaliada, seguida da descrição dos grupos de tratamento e os valores da variável numérica para os pacientes alocados em cada grupo (amostras).

O campo nível de confiança especifica o nível de confiança que será utilizado para calcular o intervalo de confiança. O seu complemento em relação a 100 fornece o nível de significância do teste de hipótese.

Cada botão na porção inferior do painel irá gerar um gráfico na área principal da aplicação. Ao clicarmos no primeiro botão (*Mostrar um diagrama Stripchart*), o gráfico de *Stripchart* dos valores de ácido fólico nos dois grupos de tratamento é mostrado no painel principal (figura 6.6). É possível observar que os valores de ácido fólico tendem a ser mais elevados no grupo N2O+O2,24h, mas existe uma superposição entre os valores de ácido fólico nos dois grupos.

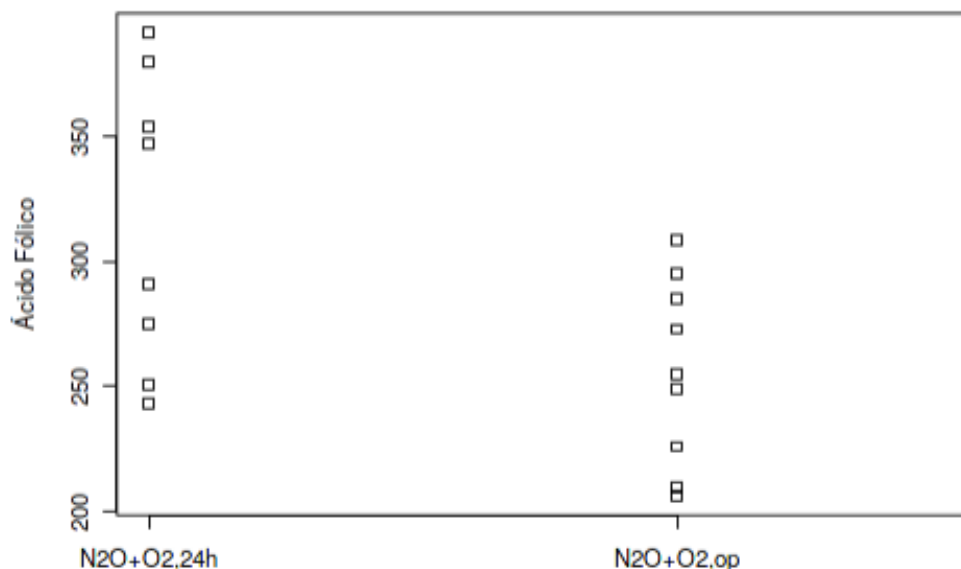


Figura 6.6: *Stripchart* dos valores de ácido fólico em dois grupos submetidos a diferentes tipos de ventilação.

As médias dos valores de ácido fólico nos dois grupos são mostradas abaixo:

```
##                mean n
## N2O+O2,24h 316.6250 8
## N2O+O2,op  256.4444 9
```

Ao realizar um estudo como esse, geralmente os autores desejam responder, dentre outras, às seguintes perguntas:

- 1) Será que podemos afirmar que a diferença observada nos valores de ácido fólico nos dois grupos ( $60,2 \mu\text{g/l}$  – N2O+O2,24h - N2O+O2,op) pode ser generalizada para as populações de pacientes com *bypass* cardíaco submetidos a um ou outro tratamento, ou essa diferença é apenas fruto do acaso?
- 2) Será que, se repetíssemos o mesmo estudo em outra amostra de pacientes, resultados semelhantes seriam observados, ou talvez nenhuma diferença entre os grupos seria

evidente?

- 3) Nesse estudo, a média da diferença dos valores de ácido fólico entre os dois grupos foi igual a  $60,2 \mu\text{g/l}$ . Se repetíssemos o estudo com o mesmo número de pacientes em cada grupo, provavelmente as diferenças observadas seriam diferentes. Que faixa de valores da diferença entre as médias de ácido fólico dos dois grupos conteria o real valor da diferença dos valores de ácido fólico nas populações submetidas aos dois tipos de tratamento?

### 6.3.2 Hipótese nula e nível de significância

Em estatística, um procedimento utilizado para responder às duas primeiras perguntas acima é o de realizar um teste de hipótese. Em tais testes, uma hipótese é formulada, um nível de significância é estabelecido e, em seguida, verifica-se o quão os dados obtidos no estudo são compatíveis com a hipótese formulada.

Vamos ilustrar cada um desses passos com o exemplo da seção anterior. Para testar se as médias de ácido fólico em pacientes submetidos aos dois diferentes métodos de ventilação são diferentes, iremos partir da hipótese de que as médias não são diferentes, ou seja, a diferença de médias é nula, e que os resultados observados nesse estudo foram simplesmente devido à aleatoriedade das amostras de pacientes submetidas aos dois tratamentos. Essa hipótese é chamada de **hipótese nula**. Como iremos verificar o quanto os dados obtidos são compatíveis com essa hipótese nula?

Vamos considerar o seguinte argumento: supondo que a hipótese nula seja verdadeira, ou seja, que não há diferença entre as duas populações de valores de ácido fólico de pacientes submetidos aos tratamentos  $\text{N}_2\text{O}+\text{O}_2, \text{op}$  ou  $\text{N}_2\text{O}+\text{O}_2, 24\text{h}$ , então podemos considerar as duas amostras do estudo acima como proveniente da mesma população. Assim sendo, podemos juntar as duas amostras numa só contendo 17 valores e obter a distribuição das diferenças de valores de ácido fólico para todas as maneiras possíveis de esse conjunto de 17 valores ser dividido aleatoriamente em duas amostras de 8 pacientes (grupo  $\text{N}_2\text{O}+\text{O}_2, 24\text{h}$ ) e 9 pacientes (grupo  $\text{N}_2\text{O}+\text{O}_2, \text{op}$ ).

Na verdade, não é necessário obter todas as divisões possíveis, um número muito grande delas, por exemplo, 20000, é suficiente para os nossos propósitos. A figura 6.7, obtida ao clicarmos no botão *Distribuição sob a hipótese nula* da aplicação da figura 6.5, mostra a distribuição dos valores da diferença de ácido fólico, supondo que a hipótese nula fosse verdadeira.

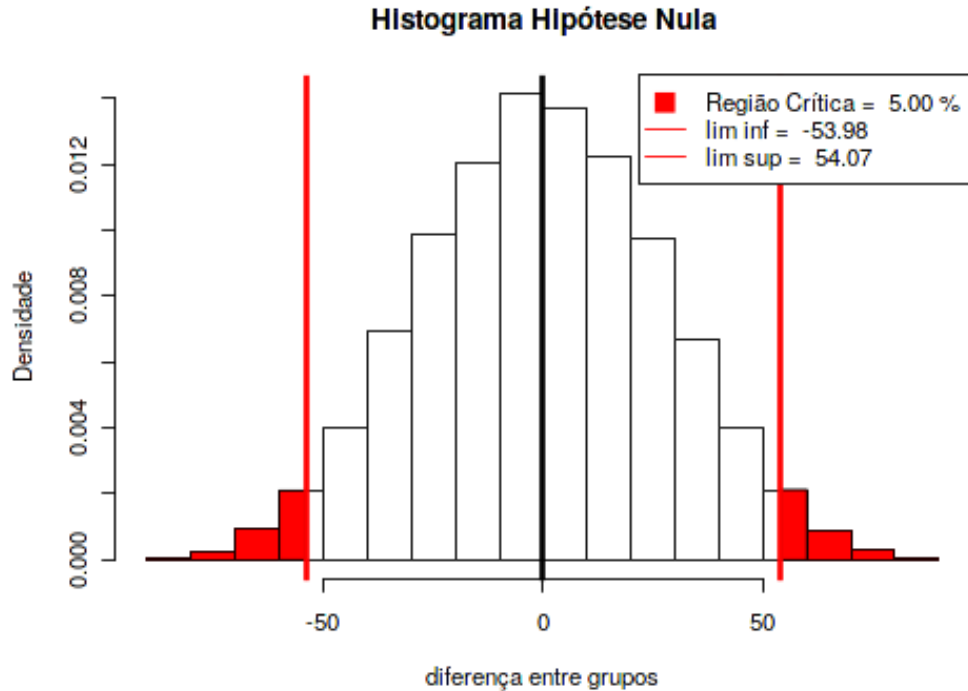


Figura 6.7: Teste para verificar a hipótese nula de igualdade das médias de ácido fólico nos dois tipos de ventilação cujas amostras são mostradas na figura 6.6.

Podemos observar que, se a hipótese nula for verdadeira, na maior parte das vezes, as diferenças de médias de ácido fólico entre as duas amostras se situa na área branca da figura, mas que, eventualmente, diferenças maiores (tanto positivas quanto negativas) podem ocorrer com uma certa probabilidade. A área em vermelho representa duas regiões onde os valores de diferenças de médias de ácido fólico entre as duas amostras são maiores do que  $54,07 \mu\text{g/l}$  ou abaixo de  $-53,98 \mu\text{g/l}$ . A área destas duas regiões corresponde à probabilidade de 5% (2,5% de cada lado), ou seja, se extraíssemos aleatoriamente duas amostras de 9 e 8 elementos respectivamente do conjunto de 17 valores, calculássemos a diferença das médias do ácido fólico nas duas amostras e repetíssemos esse procedimento um número muito grande de vezes, em aproximadamente 5% das vezes a diferença das médias cairia na região em vermelho. Esse valor de 5% é o **nível de significância** do teste de hipótese e a região em vermelho é chamada de **região crítica**. Outros valores poderiam ser usados, como 1%, 10% ou qualquer outro valor. O mais comumente usado é 5%.

Se usássemos o nível de significância igual a 10%, a área em vermelho seria maior e os valores que delimitam as duas áreas em vermelho seriam menor em valor absoluto do que os obtidos para o nível de 5%. O nível de significância é escolhido a priori, antes de coletar os dados e realizar os cálculos.

Para testar a hipótese nula ao nível de 5%, verificamos em qual região da distribuição dos valores de diferenças de médias sob a hipótese nula a diferença observada no estudo se situa. Se o valor estiver na região crítica, rejeitamos a hipótese nula de igualdade das médias de ácido fólico entre os dois tipos de ventilação. O argumento para rejeitar a hipótese nula



nesse caso é que a probabilidade de observar uma diferença de médias na região crítica é 5% se a hipótese nula for verdadeira e consideramos essa probabilidade tão baixa que, se o valor de diferença de médias observado no estudo cair nessa região, preferimos acreditar que a hipótese nula é falsa e a diferença observada no estudo é considerada **estatisticamente significativa**.

Se o valor observado da diferença de médias **não** estiver na região crítica, não rejeitamos a hipótese nula de igualdade das médias de ácido fólico entre os dois tipos de ventilação e consideramos que a diferença observada foi devida ao acaso, sendo a diferença observada considerada **não estatisticamente significativa**.

Como, nesse exemplo, a região crítica possui uma área inferior e uma área superior, ou seja, a hipótese nula será rejeitada se uma diferença de média observada for suficientemente grande em valor absoluto tanto para um lado quanto para o outro, esse teste é chamado de **teste bilateral**.

## 6.4 Valor de p

Ao clicarmos no botão *Testar hipótese nula* da aplicação da figura 6.5, iremos obter a figura 6.8. Nessa figura, a linha vertical azul indica o valor da diferença de médias do ácido fólico entre os dois grupos do estudo, situada à direita da linha vertical vermelha que indica o limite da região crítica superior. Como o valor observado da diferença de médias do ácido fólico está situado na região crítica, **a hipótese nula é rejeitada**.

A área em azul à direita do valor observado no estudo é o valor de p unilateral superior (0,013 na legenda), que indica a probabilidade de se obter, supondo que a hipótese nula é verdadeira, um valor de diferença de médias igual ou superior ao valor observado no estudo. Em um teste bilateral, como esse, o **valor de p** é o dobro do valor de p unilateral superior (ou unilateral inferior, o que for menor). Assim, nesse estudo, o valor de  $p = 0,026$ .

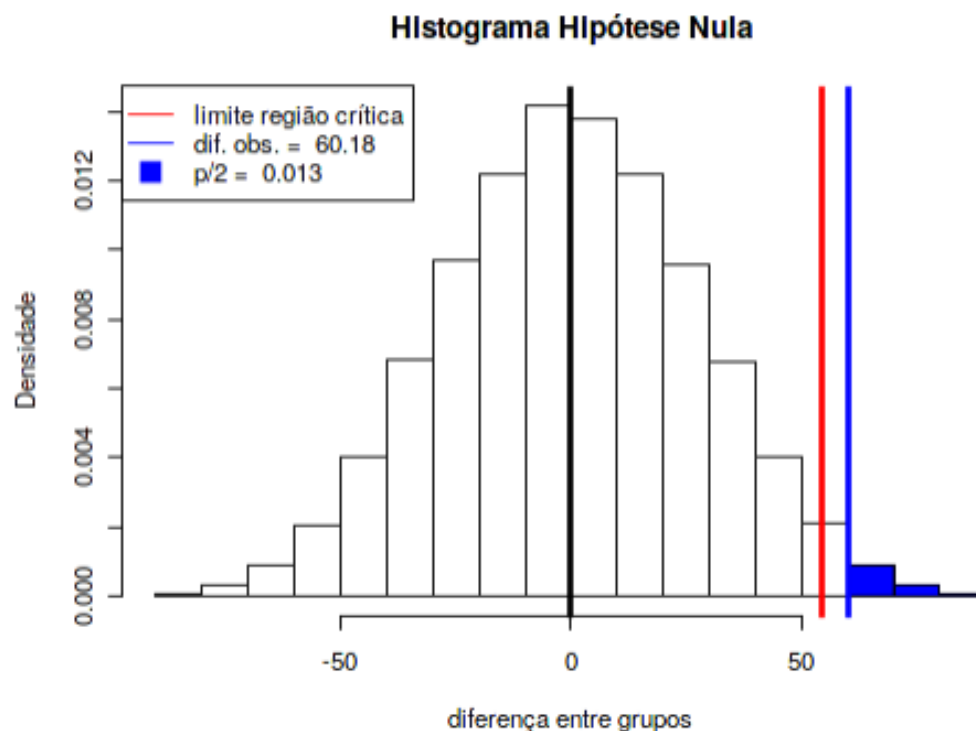


Figura 6.8: Teste para verificar a hipótese nula de igualdade das médias de ácido fólico nos dois tipos de ventilação cujas amostras são mostradas na figura 6.6.

Outra forma de decidir sobre a rejeição ou não da hipótese nula é comparar o valor de  $p$  com o nível de significância. **Se o valor de  $p$  for menor do que o nível de significância, a hipótese nula é rejeitada, caso contrário, a hipótese nula não é rejeitada.** Nesse exemplo, como o valor de  $p = 0,026 < \text{nível de significância} = 0,05$ , a hipótese nula é rejeitada.

## 6.5 Intervalo de confiança (IC)

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

Na seção anterior, rejeitamos a hipótese nula de que a diferença observada entre as médias de ácido fólico para os dois métodos de ventilação é nula. Porém a diferença de médias observada no estudo não necessariamente é o valor real da diferença de médias nas populações de pacientes submetidos aos dois tratamentos, já que ela foi obtida em duas amostras pequenas de pacientes das duas populações. Como poderíamos responder à terceira questão apresentada ao final da seção 6.3.1:

*que faixa de valores da diferença entre as médias de ácido fólico dos dois grupos conteria o real valor da diferença dos valores de ácido fólico nas populações submetidas aos dois tipos de tratamento?*

Vamos ver que não podemos afirmar com certeza que uma faixa de valores contém o valor real da diferença das médias entre os dois grupos. Na seção anterior, se a diferença de médias

observada no estudo se situasse fora da região crítica do teste de hipótese, não rejeitaríamos a hipótese de igualdade de médias entre os dois grupos. Podemos considerar que a hipótese nula é compatível com todos os valores de diferença de médias compreendidos entre os limites inferior e superior da região crítica. Como a diferença observada no estudo está dentro dos limites da região crítica, então a hipótese nula não é compatível com a diferença observada entre as médias das duas amostras do estudo.

Apesar de a hipótese nula frequentemente se referir a diferenças de médias igual a zero, nada impede que se teste uma diferença de médias com qualquer valor. Então poderíamos pensar em verificar um conjunto de hipóteses nulas que não seriam rejeitadas pelo valor da diferença de médias observado nesse estudo, ou dito de outra forma, um conjunto de hipóteses nulas que seriam compatíveis com a diferença de médias observada nesse estudo. Esse conjunto de diferenças de médias correspondentes a hipóteses nulas compatíveis com a diferença de médias observada no estudo é o intervalo de confiança para a real diferença de médias dos valores de ácido fólico entre os dois tipos de ventilação.

Vamos supor que o efeito da ventilação  $N2O+O2,op$  fosse reduzir o valor de ácido fólico de um valor D em relação à ventilação  $N2O+O2,24h$ . Então a diferença entre as distribuições dos valores de ácido fólico dos grupos  $N2O+O2,24h$  e  $N2O+O2,op$  pode ser removida simplesmente subtraindo D do valor de ácido fólico para cada paciente do grupo  $N2O+O2,24h$ , e o teste de randomização poderia ser aplicado como na seção anterior. Para obtermos o limite inferior do intervalo de confiança, identificamos o valor de D tal que a diferença de médias observada no estudo seja igual ao valor crítico superior para a hipótese nula cuja diferença de médias é D. Vamos chamar esse valor de **LI**. Esse valor é obtido por tentativa e erro, testando-se diversos valores de D até encontrar aquele para o qual a distribuição das diferenças de médias sob a hipótese nula tenha como valor crítico superior a diferença de médias observada no estudo ( $60,2 \mu g/l$ ).

Ao clicarmos no botão *Limite Inferior do Intervalo de Confiança* da aplicação da figura 6.5, o valor de LI será calculado para o exemplo do ácido fólico. A linha verde na figura 6.9 mostra o valor de LI ( $10,34 \mu g/l$ ), sendo esse a média da diferença de médias sob a hipótese nula para a qual a área sob o histograma acima do valor observado da diferença de médias no estudo (linha vermelha) é igual a 2,5% (metade do nível de significância). Como esse processo se baseia em amostras aleatórias, podem ser observados valores não exatamente iguais ao valor de LI acima cada vez que a aplicação é executada.

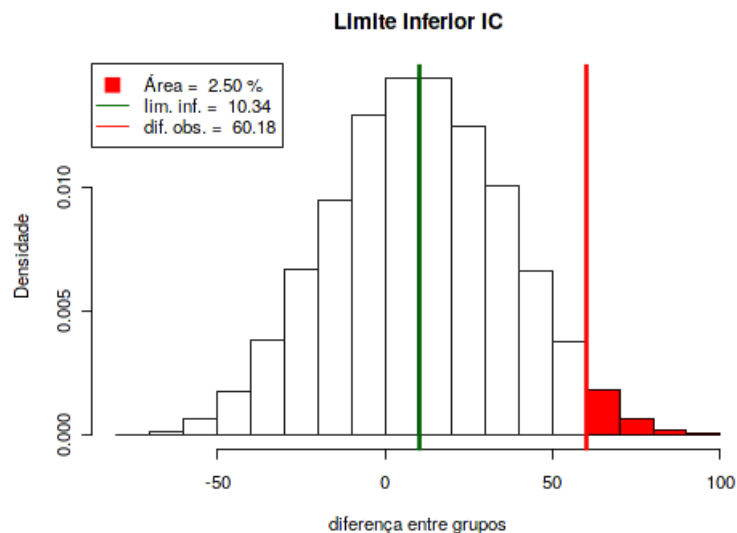


Figura 6.9: Cálculo do limite inferior do intervalo de confiança para a diferença das médias de ácido fólico entre os dois tipos de ventilação cujas amostras são mostradas na figura 6.6.

Processo análogo é feito para calcularmos o limite superior (**LS**) do intervalo de confiança para a diferença de médias do ácido fólico entre os dois métodos de ventilação. Ao clicarmos no botão *Limite Superior do Intervalo de Confiança* da aplicação da figura 6.5, o valor de LS será calculado para o exemplo do ácido fólico. A linha verde na figura 6.10 mostra o valor de LS ( $110,02 \mu\text{g/l}$ ), sendo esse a média da diferença de médias sob a hipótese nula para a qual a área sob o histograma abaixo do valor observado da diferença de médias no estudo (linha vermelha) é igual a 2,5% (metade do nível de significância). Novamente, como esse processo se baseia em amostras aleatórias, podem ser observados valores não exatamente iguais ao valor de LS acima cada vez que a aplicação é executada.

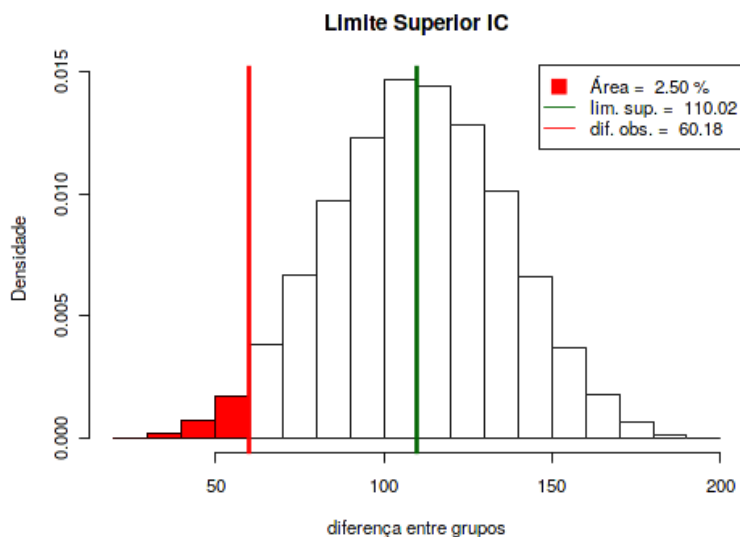


Figura 6.10: Cálculo do limite superior do intervalo de confiança para a diferença das médias de ácido fólico entre os dois tipos de ventilação cujas amostras são mostradas na figura 6.6.

Assim o intervalo de confiança ao nível de 95% para a diferença de médias de ácido fólico entre os métodos de ventilação, N2O+O2,24h e N2O+O2,op, é dado pelo intervalo  $[10,34 - 110,02]$   $\mu\text{g/l}$ . Esse intervalo é bastante amplo, porque as duas amostras desse estudo possuem poucos pacientes. **Observemos que esse intervalo de confiança não inclui a hipótese nula de igualdade de médias entre os dois métodos de ventilação.**

Assim, para esse exemplo, o intervalo de confiança com nível de confiança 95% pode ser interpretado como:

1) O IC  $[10,34 - 110,02]$   $\mu\text{g/l}$  é o conjunto de valores da diferença de médias de ácido fólico entre os dois métodos de ventilação que são compatíveis com a diferença de médias observada no estudo (60,18  $\mu\text{g/l}$ ), com um nível de confiança de 95%, no sentido de que esse IC inclui todas as diferenças de médias entre os dois grupos que correspondem a hipóteses nulas que não seriam rejeitadas com o nível de significância de 5% estabelecido no teste.

Todo intervalo de confiança está associado a um nível de confiança que é o complemento do nível de significância do teste de hipótese. Um intervalo de confiança ao nível de 90% seria mais estreito do que o intervalo de confiança ao nível de 95%.

## 6.6 Exemplo de teste sem rejeição da hipótese nula

Vamos supor que a amostra de pacientes submetidos à ventilação *N2O+O2,24h* contivesse os valores mostrados no campo Grupo 1 do painel lateral da figura 6.11. O nível de confiança também foi alterado para 90%.

## Teste de Hipótese e Intervalo de Confiança

**Variável**

**Grupo 1**

**Grupo 2**

**Valores no Grupo 1**

**Valores no Grupo 2**

**Nível de confiança:**  

90% ▼

Mostrar um diagrama Stripchart

Distribuição sob a hipótese nula

Testar hipótese nula

Limite inferior do Intervalo de Confiança

Limite superior do Intervalo de Confiança

Limpar

Figura 6.11: Alteração na aplicação da figura 6.5 nos valores do grupo 1 (N2O+O2,24h) e do nível de confiança para 90%.

O gráfico de *Stripchart* dos valores de ácido fólico nos dois grupos de tratamento é mostrado na figura 6.12. É possível observar que os valores de ácido fólico tendem a ser mais elevados no grupo N2O+O2,24h, mas não tão elevados quanto no exemplo anterior.

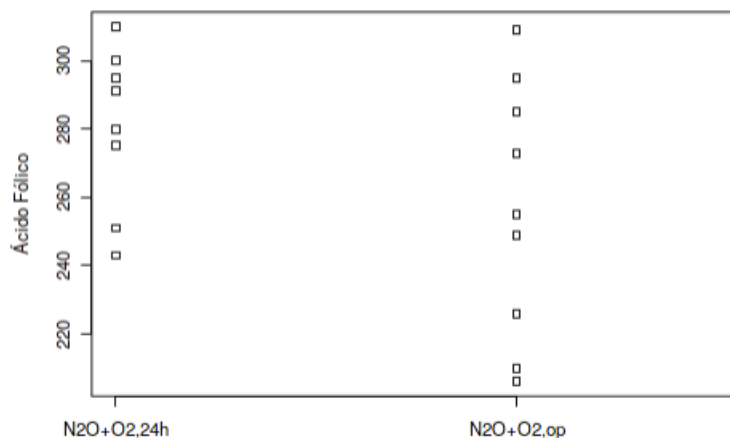


Figura 6.12: *Stripchart* dos valores de ácido fólico mostrados na figura 6.11 para os dois métodos de ventilação.

As médias dos valores de ácido fólico nos dois grupos são mostradas abaixo, sendo a diferença entre elas de  $24,18 \mu\text{g/l}$  :

```
##                mean n
## N2O+O2,24h 280.6250 8
## N2O+O2,op  256.4444 9
```

A figura 6.13, mostra a distribuição dos valores da diferença de médias de ácido fólico entre os dois métodos de ventilação, supondo que a hipótese nula fosse verdadeira.

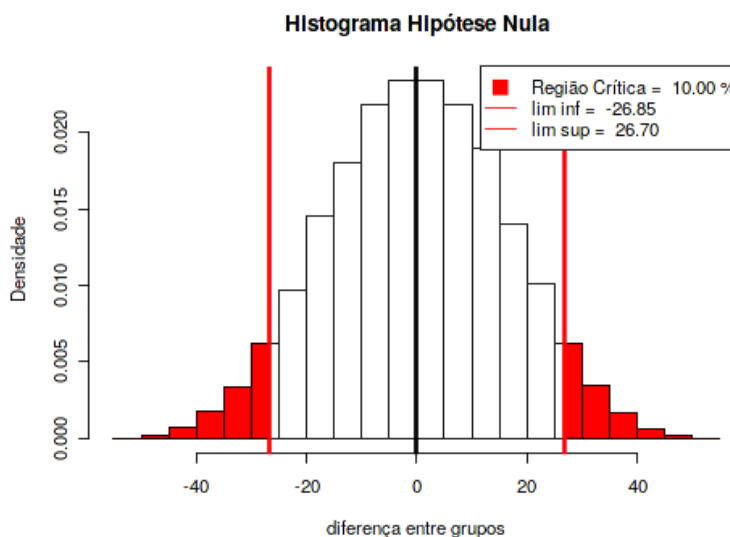


Figura 6.13: Teste para verificar a hipótese nula de igualdade das médias de ácido fólico entre os dois tipos de ventilação cujas amostras são mostradas na figura 6.12.

Podemos observar que a diferença de médias do ácido fólico para os dois grupos observada nesta seção se situa fora da região crítica do teste, o que é confirmado pela figura 6.14, obtida ao

clicarmos no botão *Testar hipótese nula* da aplicação. O valor de  $p = 2 \times 0,068 = 0,136 > 0,10$ . A hipótese nula não é rejeitada.

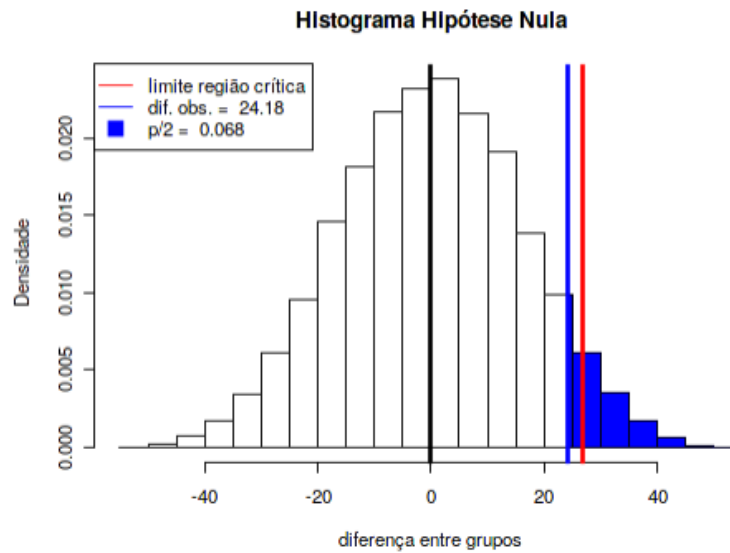


Figura 6.14: Teste para verificar a hipótese nula de igualdade das médias de ácido fólico nos dois tipos de ventilação cujos valores são mostrados na figura 6.11.

Os limites inferior e superior do intervalo de confiança nesse exemplo são mostrados nas figuras 6.15 e 6.16, respectivamente.

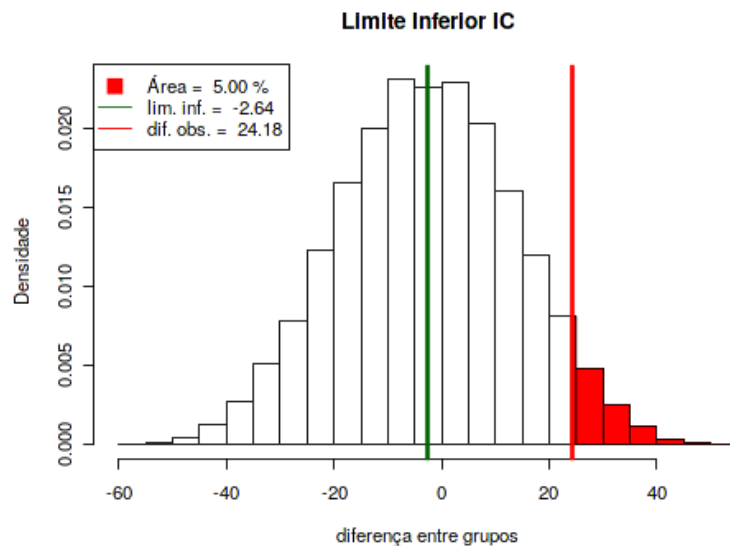


Figura 6.15: Cálculo do limite inferior do intervalo de confiança para a diferença das médias de ácido fólico entre os dois tipos de ventilação cujas amostras são mostradas na figura 6.11.



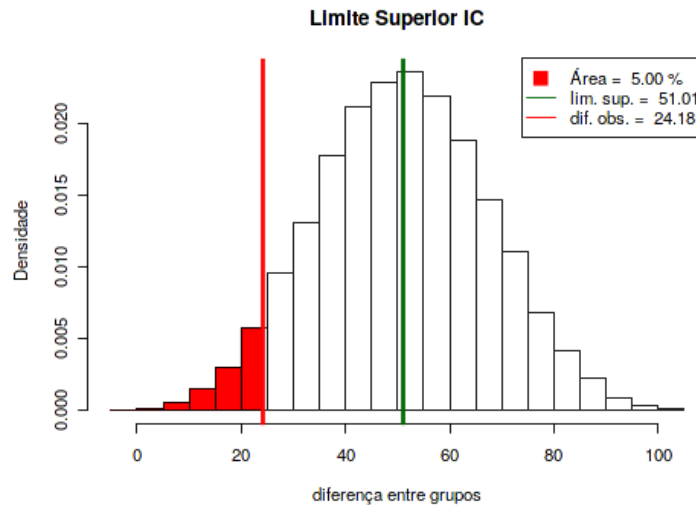


Figura 6.16: Cálculo do limite superior do intervalo de confiança para a diferença das médias de ácido fólico entre os dois tipos de ventilação cujas amostras são mostradas na figura 6.11.

Assim o intervalo de confiança ao nível de 90% para a diferença de médias de ácido fólico entre os métodos de ventilação desse exemplo, N2O+O2,24h e N2O+O2,op, é dado pelo intervalo  $[-2,64; 51,01] \mu\text{g/l}$ . **Observemos que esse intervalo de confiança inclui a hipótese nula de igualdade de médias entre os dois métodos de ventilação.**

Assim quando o intervalo de confiança inclui o valor estabelecido pela hipótese nula, a mesma não é rejeitada. Quando o intervalo de confiança não contém o valor estabelecido pela hipótese nula, a mesma é rejeitada.

Os intervalos de confiança são mais informativos do que o valor de p, já que a sua inspeção permite não somente a tomada de decisão sobre a rejeição ou não da hipótese nula, como também dá uma ideia da precisão da estimativa do parâmetro que está sendo estimado, nesse exemplo a diferença de médias de ácido fólico entre os dois métodos de ventilação.

## 6.7 Uso inadequado de testes de hipótese

O conteúdo desta seção e a apresentação de valores de p e intervalos de confiança na literatura médica podem ser visualizados neste [vídeo](#).

O uso de testes de hipótese é pervasivo na literatura na área de saúde. Às vezes, ele é realizado em situações que não justificam a sua utilização. A figura 6.17 mostra um exemplo. Trata-se de um estudo controlado randomizado, onde dois grupos de pacientes foram alocados aleatoriamente em dois grupos: um grupo de intervenção e um grupo controle. A tabela mostrada na figura realiza testes de hipótese para avaliar se a diferença de médias (ou proporções) entre os dois grupos para cada uma das variáveis clínicas da tabela, logo imediatamente após a randomização, é igual a zero. Na última coluna, a tabela apresenta o valor de p para o teste realizado para a variável correspondente. Tais testes de hipótese não

fazem sentido. Pense um pouco por que.

**TABELA 1**  
**Características demográficas e clínicas**

Características	Todos (n = 406)	Grupo intervenção (n = 200)	Grupo controle (n = 206)	P
Idade, anos	64 ± 9,4	63,5 ± 9,5	64,5 ± 9,3	0,32
Sexo feminino, n (%)	192 (47,3)	101 (50,5)	91 (44,2)	0,20
Peso, kg	75 ± 13,7	75,2 ± 13,8	74,8 ± 13,6	0,76
Altura, cm	1,67 ± 0,08	1,67 ± 0,09	1,67 ± 0,08	0,39
Índice de massa corporal, kg/m <sup>2</sup>	27 ± 4,3	27,1 ± 4,4	27 ± 4,1	0,92
Diabetes, n (%)	69 (17)	35 (17,5)	34 (16,5)	0,78
Insulina, n (%)	7 (1,72)	4 (2)	3 (1,5)	0,27
Hipertensão arterial sistêmica, n (%)	245 (60,3)	114 (57)	131 (63,6)	0,17
Doença periférica vascular, n (%)	35 (8,6)	20 (10)	15 (7,3)	0,32
Medicações em uso, n (%)				
Ácido acetilsalicílico	226 (55,7)	108 (54)	118 (57,3)	0,50
Clopidogrel	49 (12,1)	24 (12)	25 (12,1)	0,96
Ticlopidina	5 (1,2)	2 (1)	3 (1,5)	0,31
PAS, mmHg	127,6 ± 15,2	126,5 ± 15,6	128,7 ± 14,8	0,14
PAD, mmHg	78,8 ± 8,7	77,9 ± 8,6	79,8 ± 8,6	0,06

n = número de pacientes; PAD = pressão arterial diastólica; PAS = pressão arterial sistólica.

Figura 6.17: Situação em que a realização de um teste de hipótese não é adequado. Fonte: tabela 1 do estudo de (Rocha et al., 2009) ([CC BY-NC](#)).

Ora, o que significa o valor de p? Se a hipótese nula for verdadeira, então p está relacionado à probabilidade se observar um valor da estatística utilizada igual ou mais afastado do valor observado na amostra. Nesse exemplo, pela própria natureza de um ensaio controlado randomizado, os dois grupos foram criados aleatoriamente com elementos provenientes da mesma população. Assim, **logo após a randomização**, qualquer diferença observada entre os dois grupos é fruto do acaso e a **hipótese nula é necessariamente verdadeira**. Não há cabimento em testá-la.

## 6.8 Uso de modelos para o cálculo do intervalo de confiança

O conteúdo desta seção e da seção 6.9 podem ser visualizados neste [vídeo](#).

Na seção 6.5, foi utilizado um método de randomização para a obtenção do intervalo de confiança para a diferença de médias do ácido fólico entre os dois métodos de ventilação. Em determinadas situações, existem expressões analíticas que fornecem os limites do intervalo de confiança.

A seção 3.3.5 do capítulo 3 introduziu a distribuição normal para variáveis numéricas. Essa distribuição será vista com mais profundidade no capítulo 11.

Se supusermos que os valores de uma variável numérica para cada uma de duas populações de pacientes tenha uma distribuição normal com mesma variância e com médias  $\mu_1$  e  $\mu_2$  (figura

6.18), então o intervalo com nível de confiança  $(100 - \alpha)\%$  da diferença de médias entre os grupos é dado por:

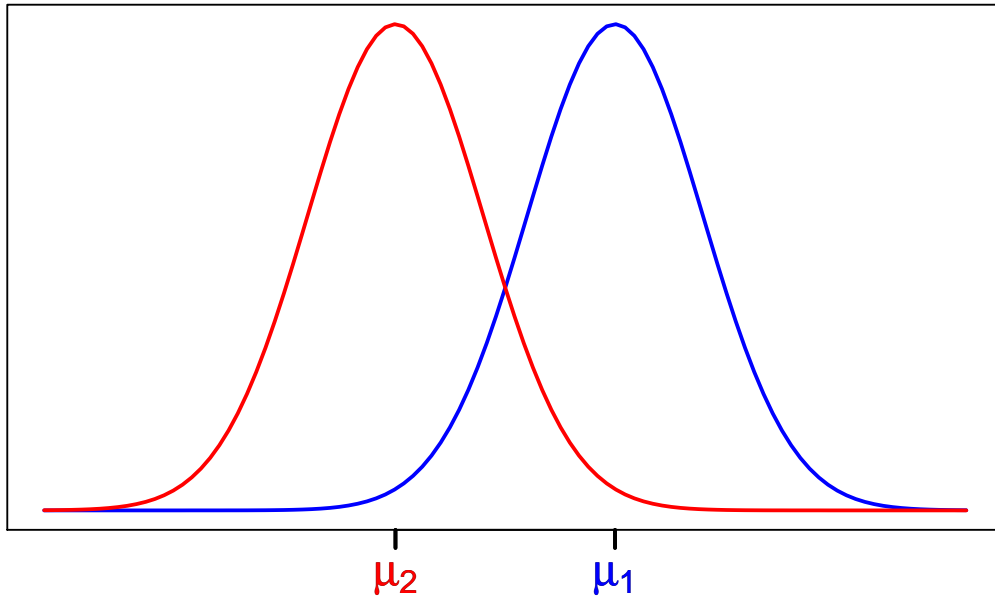


Figura 6.18: Distribuições normais para uma variável numérica em duas populações com mesma variância, mas médias diferentes.

$$(\bar{x}_1 - \bar{x}_2) - t_{gl,1-\alpha/2} s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq (\mu_1 - \mu_2) \leq (\bar{x}_1 - \bar{x}_2) + t_{gl,1-\alpha/2} s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad (6.1)$$

onde  $\bar{x}_1$  e  $\bar{x}_2$  são as médias amostrais dos grupos 1 e 2, respectivamente, e  $s$  é uma estimativa do desvio padrão comum das duas distribuições, obtida a partir da média ponderada das estimativas  $s_1^2$  e  $s_2^2$  das variâncias das duas amostras extraídas de cada uma das duas populações. Os pesos nessa média são respectivamente iguais a  $(n_1 - 1)$  e  $(n_2 - 1)$ , onde  $n_1$  é o tamanho da amostra do grupo 1 e  $n_2$  é o tamanho da amostra do grupo 2, respectivamente:

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

$t_{gl,1-\alpha/2}$  é o quantil  $1 - \alpha/2$  da distribuição *t de Student* com  $(n_1 + n_2 - 2)$  graus de liberdade (gl).

A origem da expressão (6.1) e mais detalhes sobre a distribuição *t de Student* serão objetos dos capítulos 14 e 16.

Por ora, vamos aplicar a fórmula (6.1) ao exemplo do ácido fólico. Nesse exemplo, as médias, variâncias e tamanhos amostrais dos grupos N2O+O2,24h (1) e N2O+O2,op (2), são:

$$\begin{aligned}
\bar{x}_1 &= 316,6 \text{ } \mu\text{g/l} \\
\bar{x}_2 &= 256,4 \text{ } \mu\text{g/l} \\
s_1^2 &= 3447,7(\mu\text{g/l})^2 \\
s_2^2 &= 1378,0(\mu\text{g/l})^2 \\
s &= 48,41\mu\text{g/l} \\
n_1 &= 8 \\
n_2 &= 9
\end{aligned}$$

O valor de  $gl = 15$  e  $t_{15,0,975} = 2,13$ .

Substituindo os dados acima na expressão (6.1), iremos obter o seguinte intervalo de confiança.

IC:  $[10,04 - 110,3] \text{ } \mu\text{g/l}$

Esse intervalo é bastante próximo daquele que obtivemos na seção 6.5:  $[10,34 - 110,02] \text{ } \mu\text{g/l}$ .

## 6.9 Interpretação do intervalo de confiança

Uma interpretação do intervalo de confiança para o exemplo do ácido fólico foi fornecida ao final da seção 6.5 e vamos repeti-la aqui:

**1) O IC  $(10,34 - 110,02) \text{ } \mu\text{g/l}$  é o conjunto de valores da diferença de médias de ácido fólico entre os dois métodos de ventilação que são compatíveis com a diferença de médias observada no estudo  $(60,18 \text{ } \mu\text{g/l})$ , com um nível de confiança de 95%, no sentido de que esse IC inclui todas as diferenças de médias entre os dois grupos que correspondem a hipóteses nulas que não seriam rejeitadas com o nível de significância de 5% estabelecido no teste.**

No caso geral, onde um determinado parâmetro (média de uma população, diferença de médias entre duas populações, risco relativo, etc.) está sendo estudado, o intervalo de confiança para esse parâmetro com um nível de confiança igual a  $(100 - \alpha)\%$  é o conjunto de valores do parâmetro estudado que são compatíveis com a estimativa do parâmetro obtida no estudo, no sentido de que esse IC inclui todos os valores do parâmetro que correspondem a hipóteses nulas que não seriam rejeitadas com o nível de significância estabelecido no teste.

Para uma variável numérica que possui uma distribuição normal com média  $\mu$  e desvio padrão  $\sigma$  (figura 6.19) numa população, o intervalo de confiança para a média da população é dado pela expressão:

$$\left[ \bar{x} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right] \quad (6.2)$$

onde  $\bar{x}$  é a média da amostra de tamanho  $n$  extraída aleatoriamente da população,  $\sigma$  é a variância da população, e  $z_{1-\alpha/2}$  é o quantil  $1 - \alpha/2$  da distribuição normal padrão.

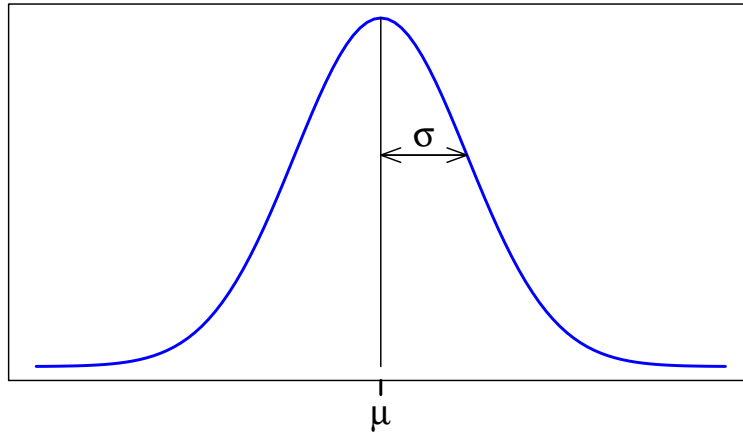


Figura 6.19: Variável numérica com uma distribuição normal em uma população.

A expressão (6.2) é utilizada na aplicação [Intervalos de confiança](#) (figura 6.20), que nos fornece uma outra interpretação para o intervalo de confiança. Essa aplicação calcula e exibe intervalos de confiança para a média de uma distribuição normal, a partir de um certo número de amostras extraídas dessa distribuição. Os parâmetros da distribuição normal, bem como o nível de confiança, o tamanho de cada amostra e o número de amostras são especificados pelo usuário. O painel principal é atualizado sempre que o usuário pressiona o botão *Reamostrar* (mais intervalos de confiança são exibidos) ou *Limpar* (limpa a tela).

#### Intervalos de confiança

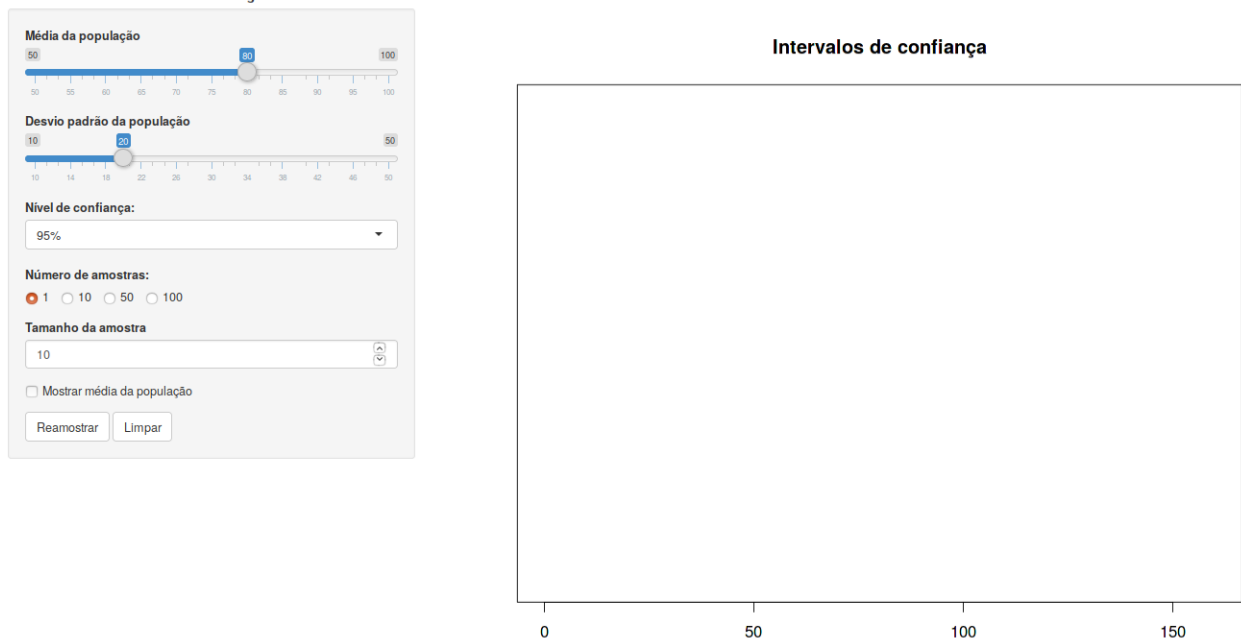


Figura 6.20: Aplicação que calcula e exibe intervalos de confiança para a média de uma distribuição normal calculados a partir de um certo número de amostras extraídas dessa distribuição.

A figura 6.21 exibe intervalos de confiança para 50 amostras de tamanho 10 de uma distribuição normal  $N(80, 400)$ . Para cada amostra, foi calculado o intervalo de confiança ao nível de 95% conforme a expressão (6.2), com  $z_{1-\alpha/2} = 1,96$  e  $\sigma = 20$ . Os 50 intervalos de confiança são exibidos no painel principal da figura.

#### Intervalos de confiança

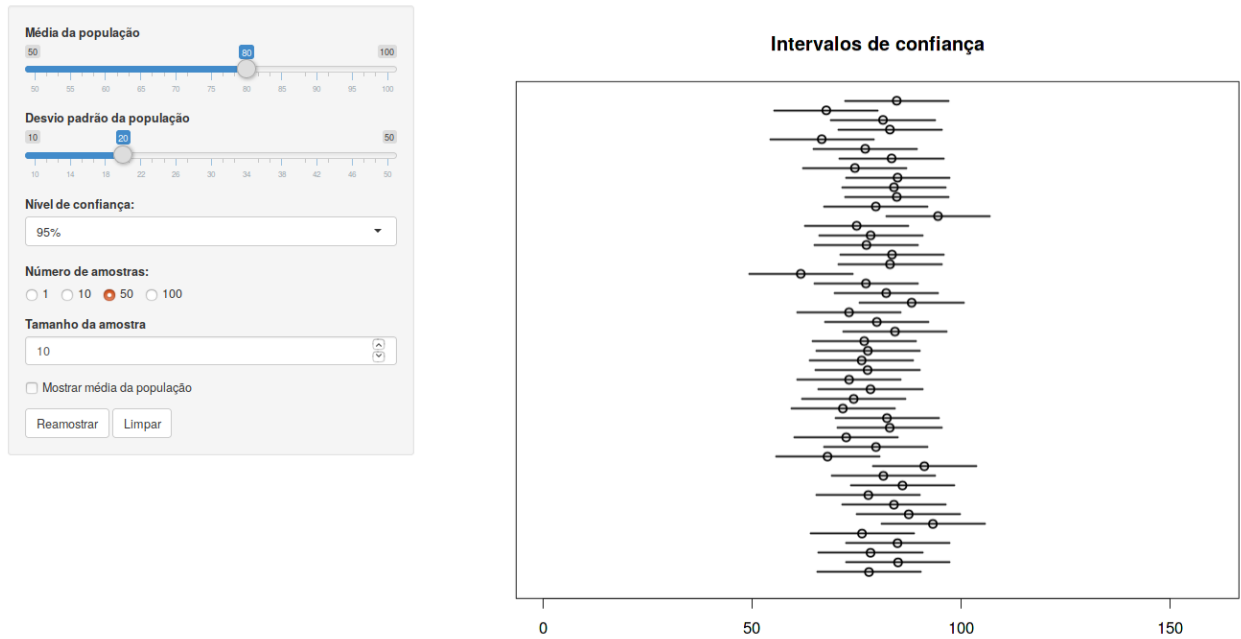


Figura 6.21: Intervalos com 95% de confiança para a média de uma distribuição normal  $N(80, 400)$ , calculados a partir de 50 amostras de tamanho 10.

Ao selecionarmos a opção *Mostrar média da população* na aplicação, uma linha preta, indicando a média real da população, é exibida, e duas linhas verticais em vermelho indicam distâncias iguais a  $z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$  acima e abaixo da média da distribuição (figura 6.22). Para cada intervalo de confiança, o centro com uma marcação representa a média da respectiva amostra. Observem que a maioria dos intervalos de confiança contém a média da distribuição, mas alguns deles (em vermelho) não contêm a média.

## Intervalos de confiança

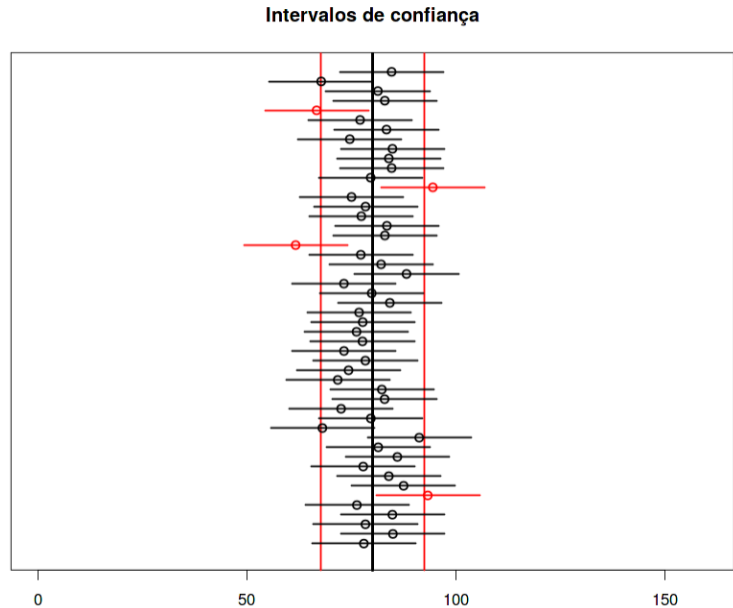
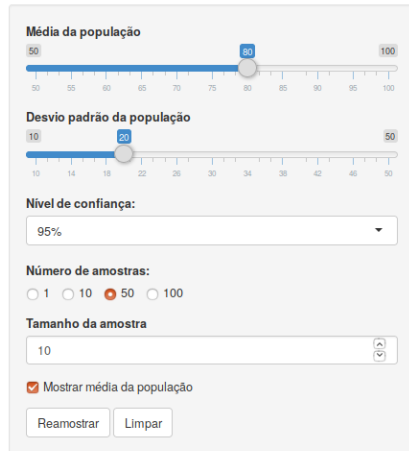


Figura 6.22: Figura 6.21 com retas que mostram a média real da população (linha vertical preta) e indicam distâncias iguais a  $z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$  acima e abaixo da média da distribuição (linhas verticais em vermelho).

É de se esperar que nem sempre o intervalo de confiança contenha o valor real do parâmetro que ele estima. No exemplo da figura 6.22, o nível de confiança é de 95%. Isso significa que, se extraíssemos um número infinito de amostras aleatórias da população e calculássemos os respectivos intervalos de confiança, em 95% das vezes o intervalo de confiança irá incluir a média real da população e em 5% das vezes, o intervalo de confiança não irá incluir a média. Isso equivale a dizer que, a cada 100 intervalos de confiança calculados, em média 5 (5%) não contêm a média da distribuição. Na figura 6.22, 4 intervalos em 50 não contêm a média real da distribuição.

A figura 6.23, mostra o uso da aplicação com a mesma distribuição normal da figura 6.20, com o mesmo número de amostras, mas com três tamanhos amostrais diferentes (1, 10 e 50). Observem que a precisão dos intervalos de confiança aumenta à medida que o tamanho das amostras aumenta de 1 para 10 e de 10 para 50. Isso é de se esperar, porque o erro padrão  $\frac{\sigma}{\sqrt{n}}$ , utilizado no cálculo do intervalo de confiança, diminui com  $n$ .

O leitor deve experimentar com diferentes níveis de confiança, número de amostras, tamanhos amostrais e parâmetros da distribuição normal.

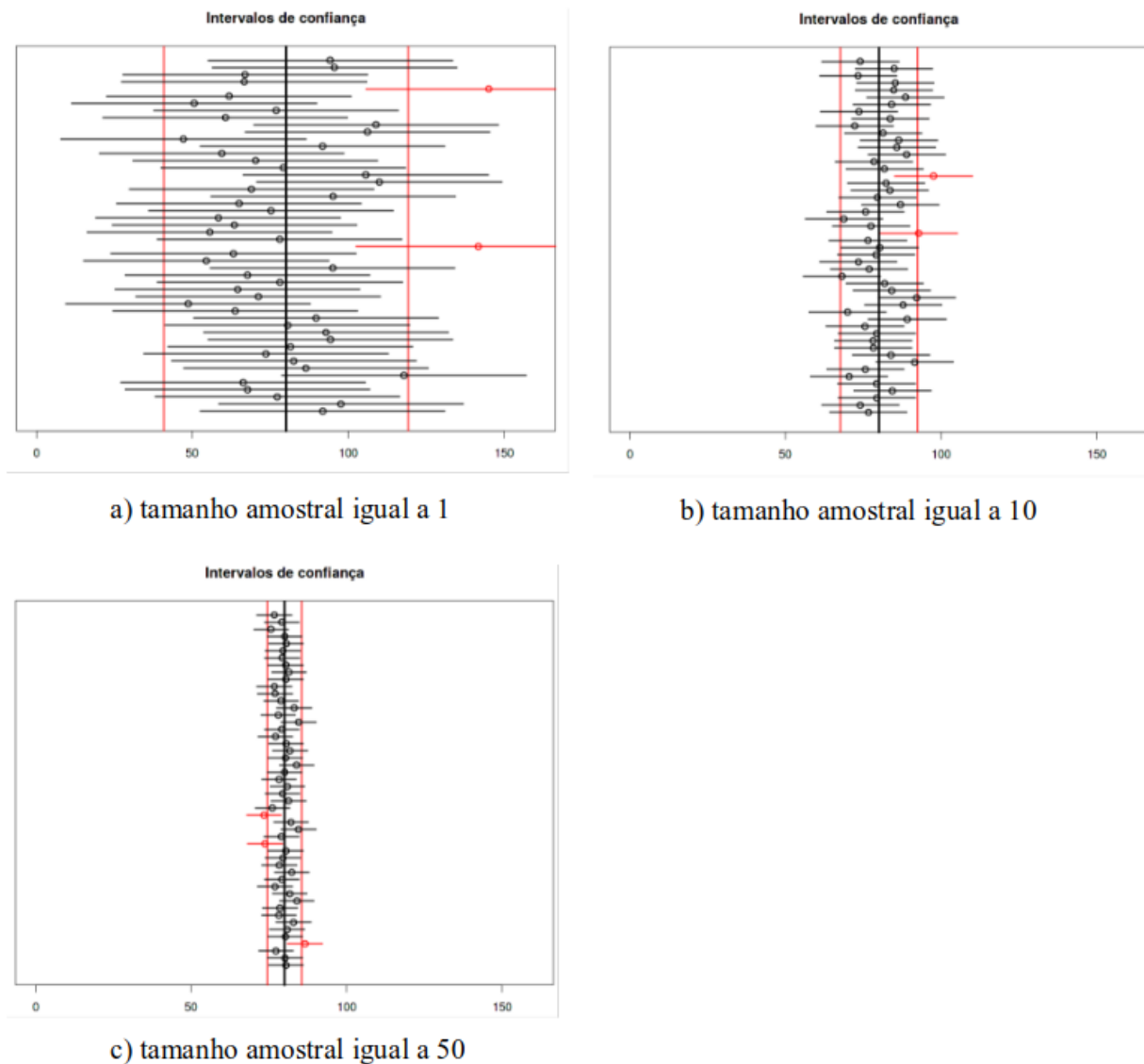


Figura 6.23: Intervalos de confiança para diferentes tamanhos de amostra (a-1, b-10, c-50).

Para qualquer estudo específico, não é possível garantir que o intervalo de confiança calculado contenha o parâmetro real da população.

Resumindo esta seção, podemos interpretar o intervalo de confiança das duas formas seguintes:

- 1) o intervalo de confiança para um determinado parâmetro com um nível de confiança igual a  $(100 - \alpha)\%$  é o conjunto de valores do parâmetro estudado que são compatíveis com a estimativa do parâmetro obtida no estudo, no sentido de que esse IC inclui todos os valores do parâmetro que correspondem a hipóteses nulas que não seriam rejeitadas com o nível de significância estabelecido no teste.
- 2) dado um nível de confiança estabelecido a priori  $(100 - \alpha)\%$ , temos uma confiança de  $(100 - \alpha)\%$  que o IC contenha o real valor do parâmetro estudado.



Essa confiança deve ser interpretada no sentido de que, se repetíssemos o estudo um número infinito de vezes e, em cada vez, calculássemos o IC, em  $(100 - \alpha)\%$  das vezes, o IC conteria o real valor do parâmetro estudado.

Para o exemplo específico do ácido fólico, temos as seguintes interpretações para o intervalo de confiança obtido na seção 6.5:

1) O IC (10,34 – 110,02) ug/l é o conjunto de valores da diferença de médias de ácido fólico entre os dois métodos de ventilação que são compatíveis com a diferença de médias observada no estudo (60,18 ug/l), com um nível de confiança de 95%.

2) Com uma confiança de 95%, o intervalo de valores entre 10,34 e 110,02 ug/l inclui o real valor da diferença entre as médias de ácido fólico para os dois métodos de ventilação.

## 6.10 Significância estatística e relevância clínica

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

Na expressão (6.1) para o cálculo do intervalo de confiança para a diferença de médias entre dois grupos, quando a variável aleatória em cada grupo segue uma distribuição normal com mesma variância, vemos que a largura do intervalo de confiança vai diminuir se aumentarmos os valores de  $n_1$  e  $n_2$ , ou seja, se aumentarmos o número de elementos da amostra de pacientes nos dois grupos. Nesse caso, dizemos que a precisão do intervalo de confiança aumenta à medida que o tamanho da amostra aumenta.

Também na expressão (6.2) para o cálculo do intervalo de confiança para a média de uma população, supondo que a variável aleatória segue uma distribuição normal, vemos que a precisão do intervalo de confiança vai aumentar se aumentarmos o tamanho amostral.

Esse é um comportamento geral.

Isso nos leva à conclusão de que podemos fazer com que qualquer diferença entre o valor observado de um parâmetro em uma amostra e o valor do parâmetro sob a hipótese nula seja estatisticamente significativa desde que tenhamos amostras com tamanho suficientemente grande. Por outro lado, mesmo que uma hipótese nula seja rejeitada em um estudo, isso não quer dizer que o efeito observado na amostra seja clinicamente relevante.

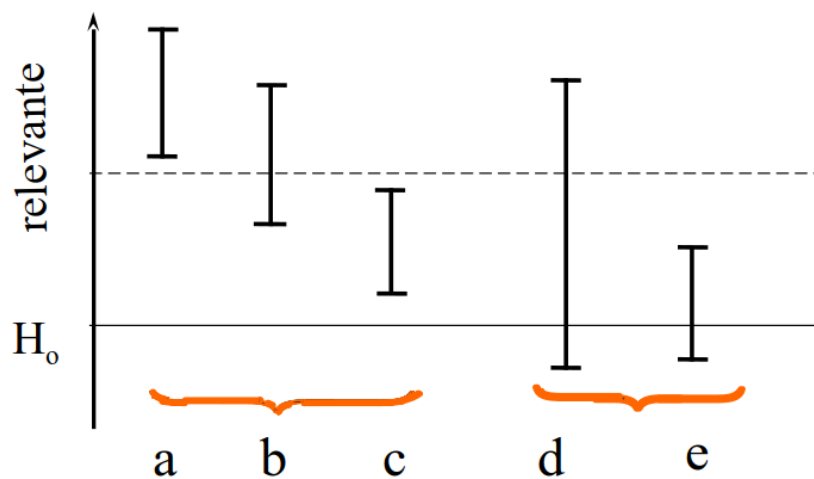
Vamos considerar um exemplo fictício para entendermos essa afirmação. Vamos supor que, clinicamente, uma redução da pressão arterial sistólica mínima de 15 mmHg seja um valor considerado clinicamente relevante e reduções abaixo desse valor não sejam interessantes do ponto de vista clínico. Então vamos supor que dois medicamentos, um experimental e outro utilizado como controle, tenham sido utilizados em duas amostras de pacientes hipertensos e, depois de um tempo, verificou-se os valores de pressão arterial sistólica em ambos os grupos, e o intervalo de confiança foi construído para a diferença de médias de pressão arterial sistólica entre os dois grupos.

A figura 6.24 mostra 5 situações possíveis de acontecer, dependendo dos efeitos dos medicamentos e do tamanho amostral utilizado no estudo.

A linha horizontal contínua representa a hipótese nula (o medicamento experimental não altera os valores da média da pressão arterial sistólica em relação ao medicamento controle). A linha horizontal tracejada representa uma diferença média de pressão de 15 mmHg (o valor mínimo considerado clinicamente relevante).

Cada linha vertical representa o intervalo de confiança para a diferença de médias de pressão arterial sistólica entre os dois grupos em um estudo hipotético. Vamos discutir cada um dos possíveis resultados.

## Relevância Clínica x Significância Estatística



### Relevância clínica:

- a – definitivamente relevante
- b – possivelmente relevante
- c – não relevante
- d – inconclusivo
- e – não relevante

### Significância estatística:

- a – estatisticamente significativo
- b – estatisticamente significativo
- c – estatisticamente significativo
- d – estatisticamente não significativo
- e – estatisticamente não significativo

Figura 6.24: Diferentes situações que mostram que não há relação entre a relevância clínica e a significância estatística.

Em 6.24a, o intervalo de confiança está todo acima do valor mínimo considerado clinicamente relevante. Nesse caso, consideramos que o resultado do estudo é clinicamente relevante e, como a hipótese nula é rejeitada, o estudo é também estatisticamente significativo. Dizemos que o medicamento experimental é mais efetivo do que o medicamento controle.

Em 6.24b, o intervalo de confiança não inclui o valor 0 (hipótese nula). Portanto o estudo é estatisticamente significativo, porém o intervalo de confiança contém valores abaixo e acima do mínimo considerado clinicamente relevante. Nesse caso, consideramos que possivelmente o efeito do medicamento experimental é clinicamente relevante, mas seria necessária uma

amostra maior para sabermos se o intervalo de confiança estaria todo acima ou todo abaixo do mínimo considerado clinicamente relevante.

Em 6.24c, o intervalo de confiança não inclui o valor 0 (hipótese nula). Portanto o estudo é estatisticamente significativo, porém o intervalo de confiança contém somente valores abaixo do mínimo considerado clinicamente relevante. Apesar de o estudo mostrar um efeito do medicamento experimental na redução da pressão arterial, esse efeito não é considerado clinicamente relevante.

Em 6.24d, o intervalo de confiança inclui o valor 0 (hipótese nula). Portanto o resultado do estudo não é estatisticamente significativo, ou seja, a hipótese nula não é rejeitada, porém o intervalo de confiança contém valores acima e abaixo do mínimo considerado clinicamente relevante. Nesse caso, consideramos que possivelmente o efeito do medicamento experimental é clinicamente relevante, mas seria necessária uma amostra maior para sabermos se o intervalo de confiança estaria todo acima ou todo abaixo do mínimo considerado clinicamente relevante.

Em 6.24e, o intervalo de confiança contém o valor da hipótese nula e está todo abaixo do valor mínimo considerado clinicamente relevante. Nesse caso, consideramos que o resultado do estudo não é clinicamente relevante e, como a hipótese nula não é rejeitada, o estudo também não é estatisticamente significativo.

Concluindo, significância estatística não implica relevância clínica e uma possível relevância clínica não implica significância estatística. É importante observarmos o intervalo de confiança para o efeito que estamos estudando para extrairmos conclusões sobre um estudo e não somente verificarmos a significância estatística do estudo.

Na avaliação da relevância clínica, também devemos levar em conta o contexto. Por exemplo, no caso c, mesmo que a redução de pressão causada pelo medicamento experimental não seja relevante clinicamente, pode ser que esse medicamento provoque menos efeitos adversos do que o medicamento controle, ou custe menos. Nesse caso, o medicamento experimental pode ser mais eficiente, levando em conta um contexto mais amplo do que somente o efeito sobre a pressão arterial sistólica.

## 6.11 Exercício

- 1) Nas tabelas apresentadas nas figuras 6.1, 6.2, 6.3 e 6.4, interprete os intervalos de confiança apresentados e indique, para cada valor de  $p$  apresentado, se a respectiva hipótese nula foi ou não rejeitada.

# Capítulo 7

## Probabilidade

### 7.1 Introdução

Os conteúdos desta seção e das seções 7.2 e 7.3 podem ser visualizados neste [vídeo](#).

Em física, alguns modelos da mecânica clássica permitem prever como os fenômenos irão se comportar dentro de determinadas condições. Assim é possível prever com boa precisão a trajetória de satélites lançados no espaço, por exemplo. Esses modelos são chamados determinísticos.

Em saúde, em geral é difícil prever com precisão como as pessoas irão evoluir sob determinados tratamentos, ou quando expostas a determinadas condições. Os fatores que podem interferir em determinados eventos são frequentemente tão numerosos e, em muitos casos, até desconhecidos, que impedem a criação de modelos determinísticos. Entre a diversidade de fatores, podem-se citar a variação genética na população, diferentes condições socioeconômicas, variáveis comportamentais, etc. Por isso, mesmo sendo universalmente aceito que o fumo é um fator etiológico para o câncer do pulmão, nem todas as pessoas que fumam desenvolvem o câncer do pulmão. Assim sendo, em geral, há alguma incerteza associada aos fenômenos biológicos.

Há diversas formas de se lidar com a incerteza. Um dos enfoques mais utilizados é a teoria da probabilidade, que é um método sistemático de descrever a aleatoriedade e a incerteza. A teoria apresenta um conjunto de regras para a manipulação e o cálculo de probabilidades. Ela tem sido aplicada em muitas áreas do conhecimento: física, epidemiologia, economia, computação, etc.

### 7.2 Conceito de probabilidade

Usualmente nos referimos a uma situação em que os desfechos, ou resultados, possuem algum componentes aleatório como um experimento. Um enfoque convencional para a teoria da probabilidade começa com o conceito de **espaço amostral**, que corresponde a um conjunto de todos os possíveis resultados de um experimento. Subconjuntos do espaço amostral são

denominados **eventos**. Por exemplo, no caso bastante simples em que 3 moedas são lançadas, o espaço amostral dos resultados possíveis será:

$$S = \{hhh; hht; hth; htt; thh; tht; tth; ttt\}$$

onde h - cara, t - coroa. Um subconjunto qualquer do espaço amostral é chamado de evento. Por exemplo, podemos definir um evento correspondente a “a segunda moeda é cara” e, nesse caso, teremos o evento

$$E = \{hhh; hht; thh; tht\}$$

Quando lançamos um dado, temos para o espaço amostral:

$$S = \{1, 2, 3, 4, 5, 6\}$$

O evento de que o número observado seja menor que 4 é dado pelo subconjunto:

$$E = \{1, 2, 3\}$$

Ao estudarmos uma fonte radioativa, se medirmos o número de partículas emitidas em um minuto, teremos que o espaço amostral será definido por um conjunto infinito de possíveis resultados:

$$S = \{0, 1, 2, \dots\}$$

Um evento de interesse seria ter em um minuto menos que cinco partículas emitidas:

$$E = \{0, 1, 2, 3, 4\}$$

Cada elemento do espaço amostral  $S$  corresponde a um único resultado. A construção do espaço amostral é feita de modo a nos auxiliar a pensar de forma mais coerente sobre os eventos. Em muitas situações, não há necessidade de definir explicitamente o espaço amostral; em geral é suficiente manipular os eventos por meio de um conjunto de regras sem identificar explicitamente os eventos como um subconjunto do espaço amostral. Por exemplo, ao jogarmos três moedas e observarmos o desfecho  $\{h, h, h\}$  temos então que os seguintes eventos ocorreram:  $\{\text{não coroa}\}$ ,  $\{\text{pelo menos uma cara}\}$ ,  $\{\text{mais caras que coroas}\}$ . Entretanto o evento  $\{\text{número par de caras}\}$  não ocorreu.

A incerteza sobre a ocorrência de um evento é modelada pela probabilidade associada ao mesmo. A probabilidade de ocorrência de um evento  $E$  é representada usualmente por  $P[E]$  ou  $P(E)$ .

Possivelmente, a forma mais aceita de se definir probabilidade é a partir da adoção de um conjunto de **axiomas**. De uma maneira simplificada, a **definição axiomática** de probabilidade afirma que a probabilidade de um evento é um número,  $P[E]$ , associado ao evento, que segue um conjunto básico de axiomas ou postulados.

### Axiomas da probabilidade:

(P1):  $0 \leq P[E] \leq 1$ , isto é, a probabilidade é um número não negativo entre 0 e 1;

(P2): Para o subconjunto vazio  $\emptyset$ , ou seja, um evento impossível,  $P[\emptyset] = 0$ ;

(P3): Para o espaço amostral  $S$ , que corresponde ao evento certo,  $P[S] = 1$ ;

(P4): Se um evento  $E$  for dividido em eventos disjuntos, ou mutuamente exclusivos,  $E_1, E_2, \dots$ , ou seja, se um dos eventos  $E_i$ ,  $i = 1, 2, \dots$ , ocorrer, os demais não ocorrerão, então a probabilidade do evento  $E$  será a soma das probabilidades de cada um dos subeventos  $E_i$ , matematicamente expressa por  $P[E] = \sum P[E_i]$

Para a regra P4, podemos entender o evento  $E$  como a união dos eventos  $E_1, E_2, \dots$  e reescrever o evento de interesse como  $E = E_1 \cup E_2 \cup E_3 \cup \dots$ , onde o símbolo  $\cup$  corresponde à união.

Em geral, se temos um espaço amostral com  $N$  resultados **mutuamente exclusivos**, então, pelas regras P3 e P4:

$$P[E_1] + P[E_2] + \dots + P[E_N] = P[S] = 1$$

Se os eventos são equiprováveis, então  $P[E_i] = 1/N$ ,  $i = 1, 2, \dots, N$ . Nesse caso, o cálculo das probabilidades se reduz à contagem. Se um evento  $A$  consiste de  $k$  resultados, a partir do espaço amostral teremos  $P[A] = k/N$ . A regra P4 é conhecida como **regra da adição para eventos mutuamente exclusivos**.

Dois eventos são complementares quando são mutuamente exclusivos e a sua união é o espaço amostral. Se  $D$  é um evento, o seu complemento será representado por  $\bar{D}$ , ou  $D^c$ , ou  $D^-$ , e temos a seguinte relação:

$$P[D] = 1 - P[\bar{D}]$$

A definição axiomática não estabelece como as probabilidades de eventos podem ser estimadas. Dependendo da situação, a probabilidade de um evento pode ser estimada por considerações de simetria, considerando todos os resultados de um experimento como equiprováveis e calculando a probabilidade de um evento como o número de resultados favoráveis ao evento dividido pelo número total de resultados possíveis.

Em um lançamento de um dado, por exemplo, se considerarmos que o dado não é viciado, podemos supor que a probabilidade de qualquer uma das faces estar voltada para cima ao cair é  $1/6$ . Assim a probabilidade de ocorrer o evento número ímpar em um lançamento do dado, seria dada por:

$$P(impar) = \frac{3}{6} = 0,5 = 50\%$$

Em outros cenários, a probabilidade pode ser estimada pela proporção da ocorrência de um evento em um certo número de experimentos, ou por algum outro meio. Assim, por exemplo, se 120 cirurgias de um determinado cirurgião foram bem sucedidas em um total de 150 cirurgias realizadas pelo mesmo, podemos estimar a probabilidade de sucesso desse cirurgião como:

$$P(sucesso) = \frac{120}{150} = 0,8 = 80\%$$

Outras regras de probabilidades permitem o cálculo de probabilidades de eventos combinados por meio da operação de união e interseção de conjuntos ou a probabilidade de um evento, sabendo-se que um outro evento ocorreu. As seções seguintes apresentam essas situações.

## 7.3 Probabilidade da união de eventos

Vamos considerar o lançamento de um dado não viciado e os seguintes eventos:

- evento A: valor par -  $\{2, 4, 6\}$
- evento B: número primo -  $\{2, 3, 5\}$
- evento C: valor = 1 -  $\{1\}$
- evento D: valor  $> 4$  -  $\{5, 6\}$

As probabilidades associadas aos eventos A, B, C e D podem ser calculadas como:

$$P(A) = \frac{3}{6} = \frac{1}{2}$$

$$P(B) = \frac{3}{6} = \frac{1}{2}$$

$$P(C) = \frac{1}{6}$$

$$P(D) = \frac{2}{6} = \frac{1}{3}$$

Os eventos C e D não possuem nenhum elemento em comum, logo eles são mutuamente exclusivos, e a probabilidade da união dos eventos C e D,  $\{1, 5, 6\}$ , é dada pela soma das probabilidades dos eventos C e D, pela propriedade P4 acima:

$$P(C \cup D) = P(C) + P(D) = \frac{1}{6} + \frac{2}{6} = \frac{1}{2}$$

Os eventos A e B não são mutuamente exclusivos. A interseção de A e B, representada pela notação  $A \cap B$ , é o conjunto formado pelos elementos comuns a A e B. Nesse exemplo:

$$A \cap B = \{2\}$$

Para dois eventos que não são mutuamente exclusivos, a probabilidade da união pode ser calculada pela seguinte expressão:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (7.1)$$

Assim, no exemplo acima, temos:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{1}{2} + \frac{1}{2} - \frac{1}{6} = \frac{5}{6}$$

A fórmula para a probabilidade da união de dois eventos pode ser facilmente entendida a partir do diagrama de Venn (figura 7.1).

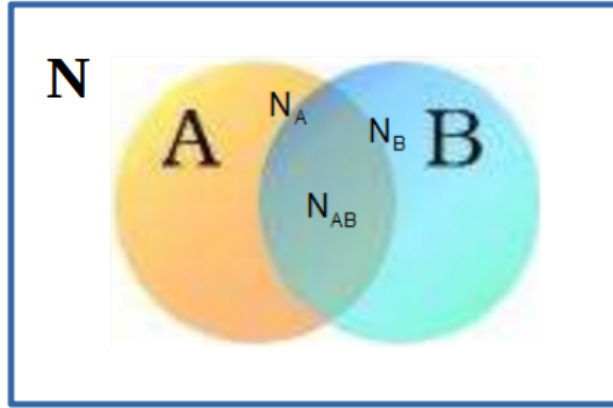


Figura 7.1: Dois eventos A e B.  $N_A$  significa o número de maneiras (eventos elementares) que o evento A pode ocorrer,  $N_B$  significa o número de maneiras que o evento B pode ocorrer,  $N_{AB}$  significa o número de maneiras que o evento A e B ocorrem simultaneamente. N é o número total de eventos elementares distintos no espaço amostral.

A partir dessa figura, temos:

$$P(A) = \frac{N_A}{N}$$

$$P(B) = \frac{N_B}{N}$$

$$P(A \cap B) = \frac{N_{AB}}{N}$$

$$P(A \cup B) = \frac{\text{Número de eventos elementares em } A \cup B}{N} = \frac{N_A + N_B - N_{AB}}{N}$$

$$P(A \cup B) = \frac{N_A}{N} + \frac{N_B}{N} - \frac{N_{AB}}{N}$$

Logo:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

A razão para a subtração de  $N_{AB}$  é que  $N_A$  inclui os eventos elementares que estão em A mas não estão em B e os eventos elementares que também estão em B, e  $N_B$  inclui os eventos elementares que estão em B mas não estão em A e os eventos elementares que também estão em A. Assim, ao somarmos  $N_A + N_B$ , estamos somando  $N_{AB}$  duas vezes. Logo é preciso subtrair  $N_{AB}$  uma vez da soma  $N_A + N_B$ .

Para A e B serem eventos disjuntos ou mutuamente exclusivos, a intersecção desses eventos deve ser o conjunto vazio, isto é, dois eventos disjuntos nunca podem ocorrer juntos.

Para três eventos A, B e C, a probabilidade da união dos três eventos é dada por:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$$



Essa fórmula pode também facilmente ser entendida por meio de um diagrama de Venn com três eventos A, B e C. A extensão dessa fórmula para n eventos  $E_1, E_2, E_3, \dots, E_n$  pode ser obtida por indução matemática.

**Exemplo 1:** Sendo  $P[C] = 0,48$  a probabilidade de que um médico se encontre em seu consultório e  $P[D] = 0,27$  a probabilidade de que ele se encontre no hospital, pergunta-se: qual a probabilidade de que ele não se encontre em nenhum desses dois lugares  $P[A]$ ?

Assumindo que esse médico não tenha consultório no hospital, os eventos são mutuamente exclusivos, já que o médico não pode estar no hospital e no consultório ao mesmo tempo. Assim:

$$P(C \cup D) = P(C) + P(D) = 0,75$$

e, portanto, temos como calcular a probabilidade do evento desejado:

$$P(A) = P[(C \cup D)^-] = 1 - P(C \cup D) = 1 - 0,75 = 0,25$$

## 7.4 Probabilidade condicional

Os conteúdos desta seção e da seção 7.5 podem ser visualizados neste [vídeo](#).

Nem todos os problemas de cálculo de probabilidades são resolvidos como nos exemplos anteriores, dividindo o evento em partes cujas probabilidades conhecemos e, então, somando as probabilidades. Em geral a obtenção das probabilidades depende do que é conhecido e do que foi aprendido ou assumido sobre a situação que estamos trabalhando. Por exemplo, poderíamos ter representado o exemplo anterior sobre o lançamento de dado e a obtenção de um resultado ímpar como:

$$P[\text{resultado} = \text{ímpar} \mid \text{dado não é viciado}]$$

para indicar que a obtenção da probabilidade é condicionada a algumas informações (ou hipóteses). A barra vertical é lida como *supondo que*, e nos referimos à *probabilidade de* ocorrer um resultado ímpar em um lançamento de um dado, *supondo que* o dado não é viciado. Se a informação condicionante não varia durante a análise, então usualmente não temos que nos preocupar com o componente *supondo que* da probabilidade condicional. Entretanto, se a informação condicionante varia, então essa notação é fundamental na análise.

Vamos considerar um exemplo simples:

**Exemplo 2:** Seja um lote de 100 peças, com 20 peças defeituosas e 80 peças boas. Suponhamos que escolhamos duas peças aleatoriamente desse lote sem reposição, ou seja, retiramos uma peça aleatoriamente e, a seguir, retiramos aleatoriamente outra peça das restantes no lote.

Sejam dois eventos A e B, definidos como:

$$A = \{1^{\text{a}} \text{ peça é defeituosa} \},$$

$$B = \{2^{\text{a}} \text{ peça é defeituosa} \}$$

$$\text{Temos que: } P(A) = \frac{20}{100} = \frac{1}{5}$$

Vamos calcular a probabilidade de B ocorrer dado que A ocorreu, ou seja, a primeira peça retirada é defeituosa. Denotamos por  $P(B|A)$  a probabilidade condicional de o evento B ocorrer quando A tiver ocorrido:

$$P(B|A) = 19/99$$

A probabilidade das duas peças serem defeituosas pode ser calculada como:

$$P(A \cap B) = \frac{\binom{20}{2}}{\binom{100}{2}} = \frac{1}{5} \cdot \frac{19}{99} = P(A)P(B|A)$$

Assim podemos usar a seguinte definição de probabilidade condicional:

Se A e B são eventos, então:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad (7.2)$$

O diagrama de Venn (figura 7.2) nos ajuda a compreender a razão da definição (7.2).

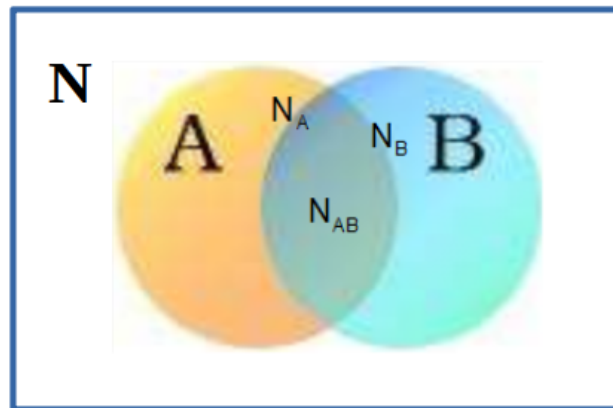


Figura 7.2: Diagrama de Venn para visualizar o cálculo da probabilidade condicional.

Seja N o número de eventos elementares no espaço amostral. Então:

$$P(B|A) = \frac{N_{AB}}{N_A} \text{ e } P(A) = \frac{N_A}{N}$$

e, nesse caso, temos que, em  $N_A$  vezes em que A ocorreu,  $N_{AB}$  vezes A e B ocorreram, logo:

$$P(A \cap B) = \frac{N_{AB}}{N} = \frac{N_{AB}}{N_A} \cdot \frac{N_A}{N} = P(B|A) \cdot P(A)$$

$$\text{Logo : } P(B|A) = \frac{P(A \cap B)}{P(A)}$$

A expressão

$$P(A \cap B) = P(B|A).P(A)$$

é conhecida como **regra da multiplicação**.

**Exemplo 3:** Vamos considerar os resultados apresentados pelo estudo de Hjerkind, Stenehjem e Nilsen (Hjerkind et al., 2017), que avaliou a associação entre a adiposidade e a atividade física com o diabetes mellitus por meio de um estudo de coortes. A tabela 7.1 é uma versão simplificada dos resultados desse estudo, que mostra a associação entre dois níveis de atividade física e diabetes mellitus entre os homens.

Tabela 7.1: Versão simplificada da tabela 3 do estudo de Hjerkind, Stenehjem e Nilsen (Hjerkind et al., 2017). Essa tabela mostra somente dois níveis de atividade física e os dados são referentes a pessoas do sexo masculino.

Atividade Física	Diabetes Mellitus		Total
	Sim	Não	
<i>Inativo</i>	73	1875	<b>1948</b>
<i>Exercita 4+ vezes/semana</i>	45	1836	<b>1881</b>
<i>Total</i>	<b>118</b>	<b>3711</b>	

Supondo que o estudo tenha uma boa validade interna e que não houvesse outros fatores que influenciassem o desfecho, qual a probabilidade de um homem com o perfil da população considerada nesse estudo desenvolver diabetes mellitus se ele for inativo?

Nesse caso, estamos interessados na  $P(\text{diabetes} | \text{Inativo})$ , ou seja, a probabilidade de um desfecho de diabetes dado que a pessoa é inativa. Analogamente, poderíamos estar interessados na probabilidade  $P(\text{diabetes} | \text{Exercita } 4 \text{ ou mais vezes por semana})$ , ou seja, a probabilidade de um desfecho de diabetes dado que a pessoa exercita 4 ou mais vezes na semana.

A partir dos dados tabulados, a  $P(\text{diabetes} | \text{Inativo})$  é estimada dividindo-se o número de pessoas inativas que tiveram diabetes pelo total de pessoas inativas acompanhadas.

$$P(\text{diabetes}|\text{inativo}) = \frac{73}{1948} = 0,037$$

$P(\text{diabetes}|\text{Exercita } \geq 4)$  é estimada dividindo-se o número de pessoas que exercitam 4 ou mais vezes por semana e que tiveram diabetes pelo total de pessoas que exercitam 4 ou mais vezes por semana.

$$P(\text{diabetes}|\text{Exercita } \geq 4) = \frac{45}{1881} = 0,024$$

## 7.5 Eventos independentes

Voltando ao exemplo do lote de peças defeituosas (exemplo 2), vamos supor agora que a retirada das peças é com reposição, ou seja, a segunda peça é retirada após a primeira peça retirada ser reposta ao lote. Considerando os eventos A e B, definidos como:

$A = \{1^{\text{a}} \text{ peça é defeituosa} \},$   
 $B = \{2^{\text{a}} \text{ peça é defeituosa} \}$

Com reposição, temos que:

$$P(A) = P(B) = \frac{20}{100} = \frac{1}{5}$$

$P(B|A) = P(B|\bar{A}) = P(B)$ , ou seja, a probabilidade de ocorrência da segunda peça ser defeituosa independe do fato de a primeira ser defeituosa ou não

$$P(A \cap B) = P(A).P(B|A) = P(A).P(B)$$

Dizemos que os eventos A e B são condicionalmente independentes se:

$P[B|A] = P[B]$  e, nesse caso:

$$P(A \cap B) = P(A)P(B) \quad (7.3)$$

A independência de eventos é uma hipótese frequentemente usada na modelagem estatística. Ela permite reduzir considerações sobre sequências complexas de eventos a uma análise de cada evento isoladamente.

## 7.6 Teorema de Bayes

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

O teorema de Bayes consiste de uma manipulação da regra da multiplicação, reescrita de duas maneiras equivalentes:

$$P[A \cap B] = P[A]P[B|A] = P[B]P[A|B]$$

Logo temos:

$$P[B|A] = \frac{P[B]P[A|B]}{P[A]} \quad (7.4)$$

Esse é o teorema de Bayes, o qual permite o cálculo de  $P[B|A]$  se conhecermos  $P[A]$ ,  $P[B]$  e  $P[A|B]$ .

**Exemplo 4:** Vamos considerar o estudo de Malacarne et al. (Malacarne et al., 2019), que avaliou o desempenho de testes para o diagnóstico de tuberculose pulmonar em populações indígenas no Brasil. Os resultados para o teste rápido molecular (TRM) em comparação à cultura de escarro (teste padrão) para todas as amostras de escarro combinadas são mostrados na tabela 7.2.

Tabela 7.2: Avaliação do teste rápido molecular (TRM) para detectar tuberculose (TB).  
Fonte: (Malacarne et al., 2019) ([CC BY-NC](#)).

Teste Rápido Molecular (TRM)	Cultura do escarro		Totais
	Com TB (D)	Sem TB ( $\bar{D}$ )	
<i>Teste positivo (<math>T^+</math>)</i>	54	7	61
<i>Teste negativo (<math>T^-</math>)</i>	4	401	405
	58	408	466

A partir dessa tabela, podemos obter um conjunto de informações importantes referentes à avaliação do uso do teste rápido molecular (TRM):

- a) **Verdadeiros positivos:** são os pacientes onde tanto o TRM quanto a cultura de escarro indicaram tuberculose pulmonar (54 pacientes);
- b) **Verdadeiros negativos:** são os pacientes onde tanto o TRM quanto a cultura de escarro não indicaram tuberculose pulmonar (401 pacientes);
- c) **Falsos positivos:** são os pacientes que o TRM indicou tuberculose pulmonar, mas a cultura de escarro deu negativo (7 pacientes);
- d) **Falsos negativos:** são os pacientes que o TRM deu negativo, mas a cultura de escarro indicou tuberculose pulmonar (4 pacientes);
- e) **Sensibilidade:** é a indicação da capacidade de o exame (TRM) de identificar pacientes que de fato tenham a condição e corresponde a uma probabilidade condicional

$$P(T^+|D) = 54/58 = 0,931$$

- f) **Especificidade:** corresponde à capacidade de o exame de rejeitar corretamente pacientes não portadores da condição, sendo também uma probabilidade condicional

$$P(T^-|\bar{D}) = 401/408 = 0,983$$

- g) Finalmente uma informação fundamental é sabermos que, uma vez tendo-se um resultado positivo, qual é a probabilidade de que o paciente tenha a condição, isto é, qual é a probabilidade de um paciente ter tuberculose pulmonar, dado que o exame foi positivo,  $P(D|T^+)$ . Devido à forma como o estudo foi realizado, essa probabilidade não pode ser calculada somente a partir da tabela 7.2 sem o conhecimento da prevalência da doença na população de interesse,  $P(D)$ , também chamada de probabilidade pré-teste.

Suponhamos agora que a prevalência da tuberculose numa população é  $P(D) = 0,1 = 10\%$ . Como poderíamos obter  $P(D|T^+)$ ?

Para efetuarmos esse cálculo, temos que utilizar o teorema de Bayes.

Temos como informações disponíveis: a probabilidade pré-teste de ocorrência da doença  $P(D)$ , a probabilidade de termos um resultado positivo dado que o indivíduo tenha a doença  $P(T^+|D)$ , e também a probabilidade de que o teste dê um resultado positivo se o indivíduo não estiver com a doença  $P(T^+|D^-)$ .

A partir da regra da multiplicação, temos:

$$P(D \cap T^+) = P(T^+) \cdot P(D|T^+) = P(D) \cdot P(T^+|D)$$

onde  $D$  é o evento ter a doença e  $T^+$  é ter um resultado positivo no exame. Com base nessa expressão, podemos escrever:

$$P(D|T^+) = \frac{P(D) \cdot P(T^+|D)}{P(T^+)} \quad (7.5)$$

Observamos que está faltando um dado importante, que é a probabilidade de termos um resultado positivo,  $P(T^+)$ . Uma forma de obtermos essa probabilidade é escrevermos  $P[T^+]$  como:

$$P(T^+) = P(T^+ \cap D) + P(T^+ \cap D^-) = P(D) \cdot P(T^+|D) + P(D^-) \cdot P(T^+|D^-) \quad (7.6)$$

Essa expressão pode ser melhor compreendida a partir da figura 7.3. Nesse diagrama, podemos ver que o evento  $T^+$  pode ser expresso pela união das interseções de  $T^+$  com o evento  $D$  e com o seu complemento ( $\bar{D}$ ), respectivamente.

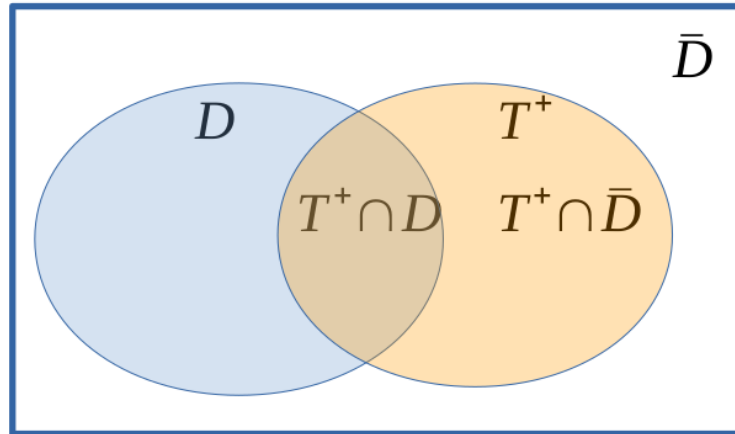


Figura 7.3: Diagrama de Venn que ilustra como obter a probabilidade de um evento  $T^+$  a partir da interseção de  $T^+$  com dois outros eventos complementares.

Temos:

$$P(D) = 0,1$$

$$P(\bar{D}) = 1 - P(D) = 1 - 0,1 = 0,9$$

$$P(T^+|D) = 0,931$$

$$P(T^-|\bar{D}) = 0,983$$

$$P(T^+|\bar{D}) = 1 - P(T^-|\bar{D}) = 1 - 0,983 = 0,017$$

Substituindo os valores acima na expressão (7.6), temos:

$$P[T^+] = 0,1 \cdot 0,931 + (1 - 0,1) \cdot 0,017 = 0,1085$$

$$\text{Finalmente: } P[D|T^+] = \frac{0,1 \cdot 0,931}{0,1085} = 0,858 = 85,8\%$$

Essas e outras métricas para avaliação de testes diagnósticos serão tema do capítulo 12.

## 7.7 Exercícios

- 1) Em uma determinada população de mulheres, 4% tiveram câncer de mama, 20% são fumantes e 3 por cento são fumantes e tiveram câncer de mama. Uma mulher é selecionada aleatoriamente da população. Qual é a probabilidade de ela ter câncer de mama ou fumar ou ambos?
- 2) Em um hospital, os centros cirúrgicos I, II e III são responsáveis por 37%, 42% e 21% do total de cirurgias. Se 0,6% das infecções hospitalares são oriundas do centro cirúrgico I, e as percentagens para os centros cirúrgicos II e III são respectivamente 0,4% e 1,2%, qual é a probabilidade de que um paciente com infecção hospitalar tenha feito cirurgia no centro cirúrgico III?

Seja o evento A a infecção hospitalar e sejam os eventos B1, B2 e B3 correspondendo a pacientes operados na sala I, II e III, respectivamente. Temos:

$$P(B1) = 0,37, P(B2) = 0,42, P(B3) = 0,21$$

$$P(A|B1) = 0,006$$

$$P(A|B2) = 0,004$$

$$P(A|B3) = 0,012$$

- 3) Suponha que 0,5% (0,005) da população apresentam uma doença D. Um teste para detectar essa doença existe, mas não é perfeito. Para pessoas com a doença D, o teste erra o diagnóstico 2% das vezes (falsos negativos). Para pessoas que não possuem a doença, ele indica a doença em 3% das vezes (falsos positivos).
  - a) Determine a probabilidade de que uma pessoa escolhida aleatoriamente da população terá um teste positivo.
  - b) Se o teste for positivo, qual é a probabilidade que a pessoa tenha D.

# Capítulo 8

## Medidas de associação

### 8.1 Introdução

Neste capítulo, serão apresentadas algumas das principais medidas de associação para variáveis categóricas utilizadas em estudos clínico-epidemiológicos: diferença de riscos, número necessário para tratar, risco relativo, diferença relativa de riscos e razão de chances. Seguindo outros capítulos, será mostrado como obter essas medidas a partir de conjuntos de dados no R.

### 8.2 Medidas de associação

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

As medidas de associação entre duas variáveis serão apresentadas para o caso de duas variáveis categóricas dicotômicas. Uma maneira bastante útil de obter essas medidas é apresentar os resultados na forma de uma tabela 2x2 (tabela 8.1). Vamos supor que uma variável de exposição possua duas categorias (*Nível 1* e *Nível 2*) e a variável de desfecho clínico possua as categorias *Sim* e *Não*, que indicam se o desfecho avaliado ocorreu ou não. Por exemplo, no estudo de Brindle et al. (Brindle et al., 2017), *Adjunctive clindamycin for cellulitis: a clinical trial comparing flucloxacillin with or without clindamycin for the treatment of limb cellulitis*, a variável de exposição seria, por exemplo, tratamento para celulite dos membros, com duas categorias: flucloxacilina com clindamicina e flucloxacilina sem clindamicina. Uma variável de desfecho do estudo é a melhoria no quinto dia, com as categorias: *Sim* e *Não*.

Interpretando a tabela 8.1 como o resultado de um estudo de coortes, transversal ou ensaio clínico randomizado (ECR),  $a$  representa o número de indivíduos expostos ao nível 1 e que tiveram o desfecho clínico de interesse,  $b$  representa o número de indivíduos expostos ao nível 1 e que não tiveram o desfecho clínico de interesse,  $c$  representa o número de indivíduos expostos ao nível 2 e que tiveram o desfecho clínico de interesse e  $d$  representa o número de indivíduos expostos ao nível 2 e que não tiveram o desfecho clínico de interesse.

O risco absoluto de ocorrência de um evento quando um indivíduo está exposto a um



Tabela 8.1: Tabela 2x2 que verifica a associação entre duas variáveis dicotômicas.

Exposição	Desfecho Clínico		Total
	Sim	Não	
<i>Nível 1</i>	<i>a</i>	<i>b</i>	<i>a+b</i>
<i>Nível 2</i>	<i>c</i>	<i>d</i>	<i>c+d</i>
<i>Total</i>	<i>a+c</i>	<i>b+d</i>	<i>a+b+c+d</i>

determinado nível de exposição é expresso pela razão entre o número de indivíduos expostos ao nível de exposição nos quais o evento ocorreu e o número total de indivíduos expostos ao correspondente nível de exposição.

Assim, na tabela 8.1, o risco de um indivíduo apresentar o desfecho clínico de interesse quando o indivíduo está exposto ao nível 1 do fator de exposição é dado por:

$$R_{N1} = \frac{a}{a+b} \quad (8.1)$$

onde  $R_{N1}$  significa o risco devido à exposição ao nível 1 do fator de exposição. Conforme visto no capítulo anterior, esse risco é uma estimativa da probabilidade de ocorrência do desfecho clínico condicionado à exposição do indivíduo ao nível 1 do fator de exposição.

Analogamente, o risco de um indivíduo apresentar o desfecho clínico de interesse quando o indivíduo está exposto ao nível 2 do fator de exposição é dado por:

$$R_{N2} = \frac{c}{c+d} \quad (8.2)$$

Onde  $R_{N2}$  significa o risco devido à exposição ao nível 2 do fator de exposição. Analogamente, esse risco é uma estimativa da probabilidade de ocorrência do desfecho clínico condicionado à exposição do indivíduo ao nível 2 do fator de exposição.

Vamos considerar os resultados apresentados pelo estudo de Hjerkind, Stenehjem e Nilsen (Hjerkind et al., 2017), que avaliaram a associação entre a adiposidade e a atividade física com o diabetes mellitus.

A tabela 8.2 é uma versão simplificada da tabela 3 do referido estudo, considerando somente a relação entre a atividade física e a ocorrência de diabetes mellitus entre os homens. Levando em conta somente dois níveis de atividade física (inatividade e exercita 4 ou mais vezes/semana), os riscos seriam expressos como:

$$R_{inativo} = \frac{73}{1948} = 0,037 = 3,7\% \quad (8.3)$$

$$R_{exercita4+vezes/semana} = \frac{45}{1881} = 0,024 = 2,4\% \quad (8.4)$$

Tabela 8.2: Versão simplificada da tabela 3 do estudo de Hjerkind, Stenehjem e Nilsen (Hjerkind et al., 2017) (CC BY-NC). Essa tabela mostra somente dois níveis de atividade física e os dados se referem a pessoas do sexo masculino.

Atividade Física	Diabetes Mellitus		Total
	Sim	Não	
<i>Exercita 4+ vezes/semana</i>	45	1836	<b>1881</b>
<i>Inativo</i>	73	1875	<b>1948</b>
<i>Total</i>	<b>118</b>	<b>3711</b>	

De cada 100 homens inativos, 3,7 em média irão desenvolver diabetes mellitus. De cada 100 homens que se exercitam 4 ou mais vezes por semana, em média, 2,4 irão desenvolver diabetes mellitus. Esse resultado deve ser visto com cautela, já que não leva em conta outros fatores que poderiam afetar o desfecho clínico em estudo.

A tabela 8.3 é uma versão simplificada da tabela 3 do estudo de Hjerkind, Stenehjem e Nilsen, considerando somente a relação entre a adiposidade, expressa pelo índice de massa corporal (IMC) e a ocorrência de diabetes mellitus entre os homens.

Tabela 8.3: Versão simplificada da tabela 2 do estudo de Hjerkind, Stenehjem e Nilsen (Hjerkind et al., 2017) (CC BY-NC). Essa tabela mostra somente dois níveis do índice de massa corporal (IMC) e os dados se referem a pessoas do sexo masculino.

IMC(kg/m <sup>2</sup> )	Diabetes Mellitus		Total
	Sim	Não	
<i><math>\geq 30</math></i>	156	1228	<b>1384</b>
<i>14,5 – 24,9</i>	95	10893	<b>10988</b>
<i>Total</i>	<b>251</b>	<b>12121</b>	

Levando em conta somente dois níveis do IMC ( $14,5 < \text{IMC} < 25 \text{ kg/m}^2$  e  $\text{IMC} \geq 30 \text{ kg/m}^2$ ), os riscos seriam expressos como:

$$R_{14,5 < \text{IMC} < 25} = \frac{95}{10988} = 0,0086 = 0,86\% \quad (8.5)$$

$$R_{\text{IMC} \geq 30} = \frac{156}{1384} = 0,11 = 11\% \quad (8.6)$$

Uma vez estimados os riscos associados a cada nível de uma variável de exposição, vamos definir e exemplificar algumas das principais medidas de associação utilizadas em epidemiologia clínica.

### 8.2.1 Diferença absoluta de riscos (DAR)

Os conteúdos desta seção e da seção 8.2.2 podem ser visualizados neste [vídeo](#).

A **Diferença Absoluta de Riscos** para uma variável de exposição é expressa pela diferença entre os riscos absolutos associados aos respectivos níveis da variável de exposição considerada, tomando um dos níveis como referência. Assim sendo, na tabela 8.1, a DAR para a variável de exposição, tomando o nível 2 como referência, é dada por:

$$DAR = R_{N1} - R_{N2} \quad (8.7)$$

No exemplo da tabela 8.2, para a atividade física, tomando a inatividade como referência, a DAR é dada por:

$$DAR = R_{exercita4+} - R_{inativo} = 0,024 - 0,037 = -0,013 = -1,3\% \quad (8.8)$$

Nesse exemplo, o fato de se exercitar com frequência maior ou igual a 4 vezes por semana reduz o risco de diabetes mellitus em 1,3% em comparação com a inatividade, desconsiderando a influência de outros fatores sobre o diabetes. Nesse caso, a atividade física seria considerada um fator de proteção em relação ao diabetes mellitus. Assim seria preciso 100 homens passarem a se exercitar 4 ou mais vezes por semana para, em média, termos uma redução de 1,3 homens que desenvolverão a doença.

No exemplo da tabela 8.3, tomando o IMC como fator de exposição e o nível do IMC entre 14,5 e 25 kg/m<sup>2</sup> como referência, a DAR, é dada por:

$$DAR = R_{IMC \geq 30} - R_{14,5 < IMC < 25} = 0,11 - 0,0086 = 0,1014 = 10,14\% \quad (8.9)$$

Nesse exemplo, um IMC  $\geq 30$  kg/m<sup>2</sup> aumenta em 10,14% o risco de diabetes mellitus em comparação com o IMC na faixa entre 14,5 e 25 kg/m<sup>2</sup>, desconsiderando a influência de outros fatores sobre o diabetes. Nesse caso, adiposidade é considerada um fator de risco em relação ao diabetes mellitus.

### 8.2.2 Número necessário para tratar

O **Número Necessário para Tratar** (NNT) é o inverso do módulo (desconsiderando o sinal) da diferença absoluta de riscos.

No exemplo da tabela 8.2, o seu valor seria:

$$NNT = \frac{1}{0,013} = 76,9 \quad (8.10)$$

Vimos no item anterior que seria preciso 100 homens passarem a se exercitar 4 ou mais vezes por semana para, em média, se ter uma redução de 1,3 homens que desenvolverão a doença.

O NNT pode ser obtido a partir do número de homens que deveriam se exercitar 4 ou mais vezes por semana para evitar 1 caso de diabetes mellitus. Usando uma regra de três:

$$\begin{array}{rcl} 100 & \text{---} & 1,3 \\ \text{NNT} & \text{---} & 1 \end{array}$$

$$NNT = \frac{1 \cdot 100}{1,3} = 76,9 \quad (8.11)$$

Assim, nesse exemplo, o NNT pode ser interpretado como o número de homens que precisariam passar a realizar a atividade de prevenção (exercitar 4 ou mais vezes por semana) para que tenhamos um caso a menos da doença.

No exemplo da tabela 8.3, como um  $IMC \geq 30 \text{ kg/m}^2$  aumenta o risco de ocorrência de diabetes mellitus, essa medida seria mais convenientemente chamada de **Número Necessário para Causar Dano** (NNH - *Number Needed to Harm*, em inglês). O seu valor seria:

$$NNH = \frac{1}{0,1014} = 9,9 \quad (8.12)$$

Ou seja, seria preciso 9,9 homens terem um  $IMC \geq 30 \text{ kg/m}^2$ , para ocorrer um caso a mais de diabetes mellitus do que se esses homens tivessem o IMC na faixa entre 14,5 e 25  $\text{kg/m}^2$ .

Assim a interpretação do NNT depende se o fator estudado é um fator de proteção ou de risco.

### 8.2.3 Risco relativo

Os conteúdos desta seção e das seções 8.2.4 e 8.2.5 podem ser visualizados neste [vídeo](#).

O **Risco Relativo** (RR) de ocorrência de um evento devido a uma fator de exposição é expresso pela razão entre os riscos absolutos associados aos respectivos níveis da variável de exposição considerada, tomando um dos níveis como referência. Assim sendo, na tabela 8.1, o RR para a variável de exposição, tomando o nível 2 como referência, é dado por:

$$RR = \frac{R_{N1}}{R_{N2}} \quad (8.13)$$

No exemplo da tabela 8.2, o RR de ocorrência de diabetes para a atividade física entre os homens, tomando a inatividade como referência, é dada por:

$$RR = \frac{R_{exercita4+}}{R_{inativo}} = \frac{\frac{45}{1881}}{\frac{73}{1948}} = 0,638 \quad (8.14)$$

Nesse exemplo, o risco de desenvolver diabetes mellitus ao se exercitar com frequência maior ou igual a 4 vezes por semana é igual a 0,64 vezes o risco de desenvolver diabetes mellitus para

homens inativos, desconsiderando a influência de outros fatores sobre o diabetes. **Quando uma variável de exposição exerce um efeito protetor, o RR varia entre 0 e 1.**

No exemplo da tabela 8.3, tomando o nível de IMC 14,5 – 24,9 kg/m<sup>2</sup> como referência, o RR é dada por:

$$RR = \frac{R_{IMC \geq 30}}{R_{14,5 < IMC < 25}} = \frac{\frac{156}{1384}}{\frac{95}{10988}} = 13,0 \quad (8.15)$$

Nesse exemplo, o risco de desenvolver diabetes mellitus com um IMC  $\geq 30$  kg/m<sup>2</sup> é 13,0 vezes maior do que o risco de desenvolver diabetes mellitus em comparação com um IMC na faixa entre 14,5 e 25 kg/m<sup>2</sup>, desconsiderando a influência de outros fatores sobre o diabetes. Nesse caso, a adiposidade é considerada um fator de risco em relação ao diabetes mellitus. **Para uma variável de exposição cujo efeito aumenta o risco, o RR é maior que 1.**

Quando o RR é igual a 1, isso significa que a variável de exposição considerada não tem influência sobre o desfecho clínico.

#### 8.2.4 Diferença relativa de riscos

A **Diferença Relativa de Riscos (DRR)** é uma estimativa do percentual do risco basal (nível de referência) que é removido (ou aumentado) como resultado do fator em estudo; ela é calculada como o oposto da diferença absoluta de riscos entre os grupos (níveis do fator) estudados, dividida pelo risco absoluto nos pacientes no grupo de referência.

Para a tabela 1, a DRR é calculada por:

$$DRR = \frac{R_{N2} - R_{N1}}{R_{N2}} = \frac{-DAR}{R_{N2}} = 1 - RR \quad (8.16)$$

Para os dados da tabela 8.2, a DRR é dada por:

$$DRR = \frac{R_{inativo} - R_{exercita4+}}{R_{inativo}} = \frac{0,037 - 0,024}{0,037} = 0,362 = 36,2\% \quad (8.17)$$

Nesse caso, podemos interpretar a diferença relativa de riscos como a **redução relativa do risco** de desenvolver diabetes mellitus devido à atividade física. Houve uma redução relativa de 36,2% no risco de desenvolver diabetes mellitus no grupo de homens que se exercitam com frequência maior ou igual a 4 vezes por semana em relação ao grupo de inativos, desconsiderando a influência de outros fatores sobre o diabetes.

Para os dados da tabela 8.3, a DRR é dada por:

$$DRR = \frac{R_{14,5 < IMC < 25} - R_{IMC \geq 30}}{R_{14,5 < IMC < 25}} = \frac{0,0086 - 0,11}{0,0086} = -12,04 = -1204\% \quad (8.18)$$

Nesse caso, podemos inverter o sinal e interpretar a diferença relativa de riscos como o **aumento relativo do risco** de desenvolver diabetes mellitus devido ao índice de massa corporal mais elevado. Houve um aumento relativo de 1204% no risco de desenvolver diabetes mellitus com um IMC  $\geq 30$  kg/m<sup>2</sup> em relação ao grupo com IMC na faixa entre 14,5 e 25 kg/m<sup>2</sup>, desconsiderando a influência de outros fatores sobre o diabetes.

### 8.2.5 Resumo das medidas de associação apresentadas até o momento

A tabela 8.4 apresenta os valores das medidas de associação apresentadas até o momento para diferentes configurações dos riscos absolutos para o nível de referência (Risco basal) e para o nível de interesse.

Tabela 8.4: Exemplos de valores das medidas de associação para diferentes configurações dos riscos absolutos.

Situação	Risco Basal (Referência)	Risco nível de interesse	Risco Relativo	Diferença Relativa de Riscos	Diferença Absoluta de Riscos	NNT
1	0,02	0,01	0,5	50%	-0,01	100
2	0,4	0,2	0,5	50%	-0,2	5
3	0,04	0,02	0,5	50%	-0,02	50
4	0,04	0,03	0,75	25%	-0,01	100
5	0,4	0,3	0,75	25%	-0,1	10
6	0,01	0,005	0,5	50%	-0,005	200

As seguintes observações podem ser feitas em relação a essa tabela:

- 1) nas duas primeiras linhas, o risco relativo é igual, o mesmo acontecendo com a diferença relativa de riscos. Porém a diferença absoluta de riscos e, conseqüentemente o NNT, são bastante diferentes;
- 2) nas linhas 1 e 4, as diferenças absolutas de riscos são iguais, mas os valores do risco relativo e diferença relativa de riscos são diferentes;
- 3) para diferentes valores da diferença absoluta de riscos (linhas 1, 2 e 3 ou linhas 4 e 5), os valores do risco relativo e diferença relativa de riscos podem ser iguais (linhas 1, 2 e 3), mas também podem ser diferentes (linhas 1 e 4);
- 4) os valores do risco relativo e diferença relativa de riscos podem ser iguais para diferentes configurações de riscos (linhas 1, 2, 3 e 6 e também linhas 4 e 5), mas também podem ser diferentes (linhas 3 e 4 e também linhas 5 e 6);
- 5) quanto menor a diferença absoluta de riscos, maior o valor do NNT e mais pessoas precisam ser submetidas ao grupo de tratamento para observarmos um caso que se beneficiaria do tratamento.

A diferença relativa de riscos pode ser obtida a partir do risco relativo, e o número necessário para tratar pode ser obtido a partir da diferença absoluta de risco, mas a relação entre a

diferença absoluta de riscos (número necessário para tratar) e o risco relativo (diferença relativa de riscos) depende dos riscos absolutos de cada nível do fator em estudo.

Na apresentação dos dados de um estudo, é boa prática apresentar a tabela sempre que possível para permitir que o usuário derive as diversas medidas de associação e extraia as suas próprias conclusões. Isso evita a situação em que autores chamem atenção para os seus resultados, apresentando somente as medidas com valores “atraentes”.

A diferença absoluta de riscos e o número necessário para tratar são medidas úteis na gestão em saúde, particularmente na decisão sobre alocação de recursos. Quanto menor o NNT de um dado tratamento, por exemplo, maior o impacto da aplicação desse tratamento na população e mais eficiente a alocação de recursos para esse tratamento em relação a um outro tratamento com NNT maior. Obviamente, deve ser levado em conta que nenhuma medida isoladamente deve ser o fator determinante para a alocação de recursos.

O risco relativo é frequentemente utilizado em epidemiologia para expressar a força de associação entre, por exemplo, um fator de exposição e uma dada doença.

## 8.2.6 Razão de chances (*odds ratio*)

Os conteúdos desta seção e da seção seguinte (8.2.7) podem ser visualizados neste [vídeo](#).

### 8.2.6.1 Chance (odds) de ocorrência de um evento

Seja  $p$  a probabilidade (risco) de ocorrência de um evento, um número entre 0 e 1. A chance de ocorrência  $C$  desse evento é dada por:

$$C = \frac{p}{1 - p} \quad (8.19)$$

A tabela 8.5 lista alguns valores para a probabilidade e o correspondente valor para a chance. Observem que a chance pode variar de 0 até  $+\infty$ . Para valores pequenos da probabilidade, digamos abaixo de 10%, o valor da chance é próximo ao valor da probabilidade. À medida que a probabilidade aumenta, os valores da chance se afastam cada vez mais do correspondente valor da probabilidade. Assim não é correto interpretar a chance como probabilidade.

A chance é frequentemente usada em bolsa de apostas. Quando se fala que a chance de um cavalo ganhar a corrida é de 8 para 1 (chance = 8), isso significa que o cavalo ganharia 8 corridas para cada uma que perdesse. Isso corresponderia a uma probabilidade de ganhar de 8/9 (89%). Esse valor poderia ser obtido, expressando a probabilidade em função da chance na expressão (8.19):

$$(1 - p)C = p \Rightarrow p = \frac{C}{1 + C} \quad (8.20)$$

Tabela 8.5: Relação entre Probabilidade (Risco) e Chance.

Risco	Chance
0 (0%)	0
0,05 (5%)	0,053
0,1 (10%)	0,11
0,2 (20%)	0,25
0,3 (30%)	0,43
0,4 (40%)	0,67
0,5 (50%)	1
0,6 (60%)	1,5
0,7 (70%)	2,3
0,8 (80%)	4
0,9 (90%)	9
0,95 (95%)	19
0,99 (99%)	99

Na tabela 8.2, a chance de ocorrência de diabetes para homens inativos é dada por:

$$C_{Inativo} = \frac{R_{Inativo}}{1 - R_{Inativo}} = \frac{0,037}{1 - 0,037} = 0,0384 \quad (8.21)$$

e para pessoas que exercitam 4 ou mais vezes por semana:

$$C_{exercita4+} = \frac{R_{exercita4+}}{1 - R_{exercita4+}} = \frac{0,024}{1 - 0,024} = 0,0246 \quad (8.22)$$

Na tabela 8.3, a chance de ocorrência de diabetes para homens com IMC na faixa 14,5-25 kg/m<sup>2</sup> é dada por:

$$C_{14,5 < IMC < 25} = \frac{R_{14,5 < IMC < 25}}{1 - R_{14,5 < IMC < 25}} = \frac{0,0086}{1 - 0,0086} = 0,0087 \quad (8.23)$$

e para pessoas com  $IMC \geq 30 \text{ kg/m}^2$ :

$$C_{IMC \geq 30} = \frac{R_{IMC \geq 30}}{1 - R_{IMC \geq 30}} = \frac{0,11}{1 - 0,11} = 0,124 \quad (8.24)$$

#### 8.2.6.2 Razão de chances (*odds ratio*)

A **Razão de Chances** (RC) de ocorrência de um evento para uma variável de exposição é expressa pela razão entre as chances associadas aos respectivos níveis da variável de exposição considerada, tomando um dos níveis como referência. Assim sendo, na tabela 8.1, a RC de



ocorrência de diabetes para a variável de exposição, tomando o nível 2 como referência, é dada por:

$$RC = \frac{C_{N1}}{C_{N2}} \quad (8.25)$$

No exemplo da tabela 8.2, a RC de ocorrências de diabetes para a atividade física entre os homens, tomando a inatividade como referência, é dada por:

$$RC = \frac{C_{exercita4+}}{C_{inativo}} = \frac{45 \cdot 1875}{73 \cdot 1836} = 0,630 \quad (8.26)$$

Nesse exemplo, a chance de desenvolver diabetes mellitus ao se exercitar com frequência maior ou igual a 4 vezes por semana é igual a 0,63 vezes maior do que a chance de desenvolver diabetes mellitus para homens inativos, desconsiderando a influência de outros fatores sobre o diabetes, ou seja, o exercício reduz a chance de ocorrência de diabetes mellitus. **Quando uma categoria de exposição exerce um efeito protetor, a RC varia entre 0 e 1. Observem que, nesse caso, a RC é menor do que o RR.**

No exemplo da tabela 8.3, tomando o nível de IMC 14,5 – 24,9 kg/m<sup>2</sup> como referência, a RC é dada por:

$$RC = \frac{C_{IMC \geq 30}}{C_{14,5 < IMC < 25}} = \frac{156 \cdot 10893}{95 \cdot 1228} = 14,6 \quad (8.27)$$

Nesse exemplo, a chance de desenvolver diabetes mellitus com um IMC  $\geq 30$  kg/m<sup>2</sup> é 14,6 vezes maior do que a chance de desenvolver diabetes mellitus em comparação com um IMC na faixa 14,5–24,9 kg/m<sup>2</sup>, desconsiderando a influência de outros fatores sobre o diabetes. Nesse caso, a adiposidade é considerada um fator de risco para o diabetes mellitus. **Para uma categoria de exposição cujo efeito aumenta a chance, a RC é maior que 1. Observem que, nesse caso, a RC é maior do que o RR.**

### 8.2.7 Razão de chances e risco relativo

Foi visto na seção anterior que a razão de chances superestima o risco relativo quando ele é maior que 1 e subestima o risco relativo quando o mesmo é menor que 1. Duas perguntas podem ser feitas em relação à razão de chances:

- 1) Por que usar a razão de chances como medida de associação?
- 2) Quando a razão de chances pode ser utilizada como uma aproximação para o risco relativo?

Em relação à primeira pergunta, a razão de chances é uma das medidas de associação mais utilizadas em estudos de metanálise e estudos de caso-controle. Além disso, ela é obtida diretamente a partir dos modelos de regressão logística, que é um dos modelos estatísticos mais utilizados na literatura médica.

Para responder à segunda pergunta, vamos utilizar a análise apresentada por Davies, Crombie, and Tavakoli (Davies et al., 1998).

Sejam  $P_1$  e  $P_2$  as proporções de indivíduos que experimentam um evento em dois grupos 1 e 2, respectivamente. Seja o grupo 1 o grupo de referência e vamos chamar de  $P_1$  o risco basal ou risco de referência. Então o risco relativo (RR) é dado por:

$$RR = \frac{P_2}{P_1} \quad (8.28)$$

A partir da definição de razão de chances (RC), ela pode ser escrita como:

$$RC = \frac{1 - P_1}{1 - P_2} RR \quad (8.29)$$

Com alguma manipulação algébrica, é possível representar a discrepância entre a razão de chances e o risco relativo como uma proporção do risco relativo.

Para estudos onde a razão de chances é  $> 1$ , essa discrepância é expressa por:

$$RC - RR = P_1 (RC - 1) RR \quad (8.30)$$

Para estudos onde a razão de chances é  $< 1$ , essa discrepância é expressa por:

$$RC - RR = P_1 (1 - RC) RR \quad (8.31)$$

O multiplicador do risco relativo nas expressões acima,  $P_1 (RC - 1)$  ou  $P_1 (1 - RC)$ , fornece o quanto a razão de chances superestima ou subestima o risco relativo, expresso como uma proporção do risco relativo. Essa proporção depende do risco basal (inicial ou referência) e do valor da razão de chances.

A aplicação [Risco Relativo x Razão de Chances](#) mostra o quanto a razão de chances subestima ou superestima o risco relativo, para um determinado nível do risco basal (nível de referência). Nessa aplicação, cuja tela inicial é mostrada na figura 8.1, as linhas vermelhas mostram a discrepância da razão de chances em relação ao risco relativo em função do risco basal, expressa em porcentagem. Para cada valor da razão de chance, uma reta diferente é obtida. Algumas linhas de referência com os respectivos valores da razão de chance são mostradas. O gráfico superior mostra as retas para valores da razão de chance acima de 1. O gráfico da parte inferior mostra as retas para valores da razão de chance abaixo de 1. O ponto em azul no gráfico mostra a discrepância para o valor do risco basal selecionado no painel à esquerda. Experimente a aplicação com diferentes valores da razão de chances e risco basal.

## Risco Relativo x Razão de Chances

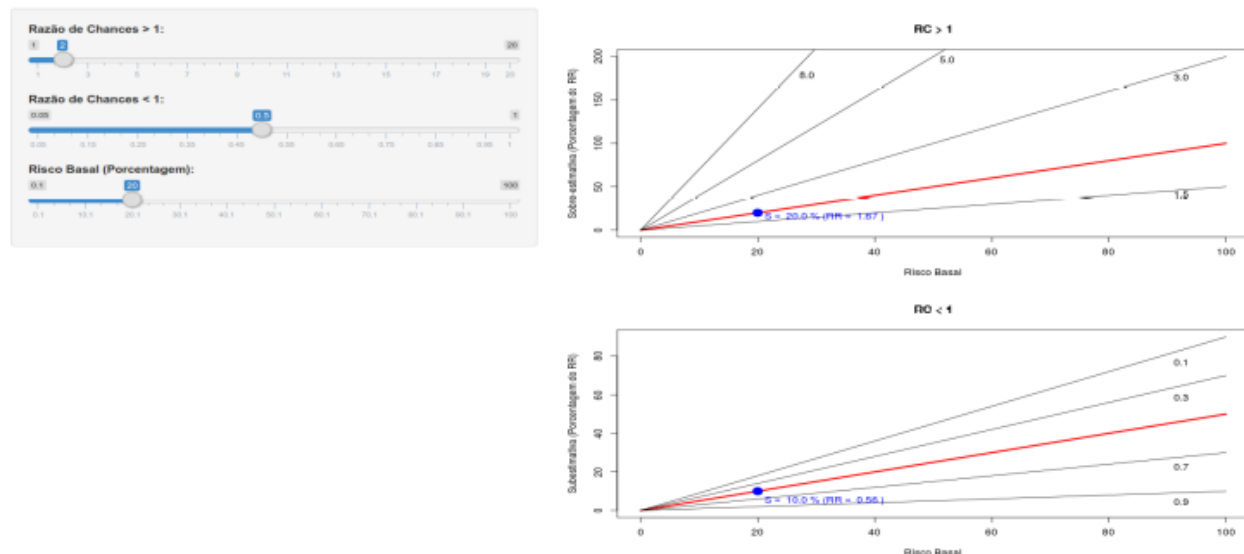


Figura 8.1: Aplicação que mostra o quanto a RC superestima ou subestima o RR em função do risco basal. No gráfico superior, o ponto em azul indica que, para o risco basal de 20%, a RC (2 nesse exemplo) é 20% maior do que o risco relativo. No gráfico inferior, o ponto em azul indica que, para o risco basal de 20%, a RC (0,5 nesse exemplo) é 10% inferior ao risco relativo.

Os gráficos mostram que, para razões de chances menores que 1, mesmo com riscos basais tão grandes quanto 50% e grandes reduções no risco (razão de chances em torno de 0,1), a razão de chances é somente 50% menor do que o risco relativo. De fato, para razão de chances menores que 1, a discrepância entre a razão de chances e o risco relativo nunca será maior do que o risco basal.

Para razões de chances maiores que 1, embora grandes desvios entre a razão de chances e o risco relativo são possíveis, a razão de chances exagera o risco relativo em menos de 50% para uma vasta gama de riscos basais e razão de chances. Para riscos basais de 10% ou menos, mesmo razões de chances de até 6 podem razoavelmente ser interpretadas como riscos relativos (discrepância menor que 50%). Como regra conservadora, se o risco basal multiplicado pela razão de chances for inferior a 100%, a razão de chances será menor do que o dobro do risco relativo.

Resumindo, vale a pena reproduzir aqui a conclusão da análise de Davies, Crombie, and Tavakoli (Davies et al., 1998):

“A razão de chances pode ter uma interpretação não intuitiva, mas, em quase todos os casos realistas, interpretando-as como se fossem riscos relativos, é improvável que qualquer avaliação **qualitativa** dos resultados do estudo seja alterada. A razão de chances sempre irá exagerar o valor quando interpretada como um risco relativo, e o grau de exagero irá aumentar à medida que aumenta o risco basal e o efeito do fator em estudo. No entanto não há nenhum ponto em que o grau de exagero é susceptível de conduzir a diferentes julgamentos qualitativos sobre o estudo. Discrepâncias significativas entre a razão de chances e o risco relativo são

vistas apenas quando os efeitos do fator em estudo são grandes e o risco basal é elevado. Se um grande aumento ou um grande decréscimo do risco é indicado, nossos julgamentos serão provavelmente os mesmos: os efeitos são importantes.”

A razão de chances é utilizada frequentemente nas seguintes situações:

- Como aproximação do risco relativo, particularmente em estudos de caso-controle, onde não se conhece o risco de ocorrência do desfecho no grupo de referência ou não seja possível estimar o risco relativo;
- em modelos de regressão logística, nos quais os coeficientes de cada variável do modelo são relacionados com a razão entre as chances de ocorrência do desfecho para os respectivos níveis da variável;
- em estudos de metanálise, como expressão da medida de associação comum dos diversos estudos analisados.

## 8.3 Medidas de associação no R

O conteúdo desta seção pode ser visualizado neste [vídeo](#), seguido deste [vídeo](#) e deste [vídeo](#).

Nesta seção, será mostrado como obter no R as medidas de associação apresentadas nas seções anteriores. Para isso, será utilizado o conjunto de dados *stroke* do pacote *ISwR* ([GPL-2](#) | [GPL-3](#)). Para ler um conjunto de dados de um pacote do R, siga os passos especificados no capítulo 2, seção 2.2. O conjunto de dados *stroke* contém todos os casos de AVC (acidente vascular cerebral) (829 observações) em Tartu, Estonia, durante 1991-1993, com acompanhamento até 1º de janeiro de 1996.

Vamos obter as medidas de associação entre as variáveis:

- *dead*: variável categórica binária, com os valores *TRUE*, se o paciente faleceu, e *FALSE*, se o paciente continuava vivo ao final do estudo, e
- *diab*: história de diabetes, variável categórica binária, com os valores *No* e *Yes*.

Vamos tomar a ausência de história de diabetes como nível de referência e vamos comparar os riscos de morte entre aqueles com ou sem histórico de diabetes.

Recordando, para abrir o conjunto de dados via *R Commander*, executamos a função

```
library(ISwR)
```

na área de script do *R Commander* e, em seguida, selecionamos o conjunto *stroke* via:

Dados  $\Rightarrow$  Conjunto de dados em pacotes  $\Rightarrow$  Ler conjunto de dados de pacotes 'atachados'

Em seguida, selecionamos o conjunto de dados *stroke* no pacote *ISwR* (figura 8.2).

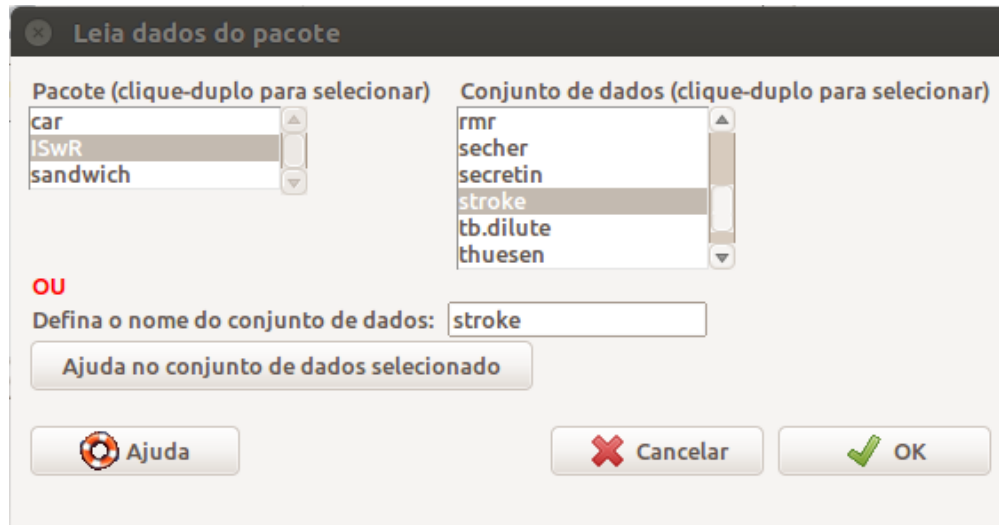


Figura 8.2: Tela para a leitura do conjunto de dados *stroke* do pacote *ISwR* via *R Commander*.

As tabelas que mostram a contagem para cada combinação das categorias das variáveis categóricas são também chamadas tabelas de contingência. Quando há duas variáveis categóricas binárias, a tabela de contingência também é chamada tabela de dupla entrada ou tabela 2x2.

Vamos analisar uma tabela 2x2 no *R Commander*. Selecionamos a opção:

Estatísticas  $\Rightarrow$  Tabelas de contingência  $\Rightarrow$  Tabela de dupla entrada...

Na tela de configuração do comando para analisar uma tabela 2x2, é preciso selecionar a variável cujas categorias aparecerão nas linhas da tabela e a outra variável cujas categorias comporão as colunas da tabela (figura 8.3).



Figura 8.3: Selecionando as variáveis da tabela 2x2.

Na aba *Estatísticas*, vamos marcar as opções *Percentual das linhas*, *Teste do Qui-Quadrado* e *Teste exato de Fisher* (figura 8.4). Vamos clicar no botão OK.

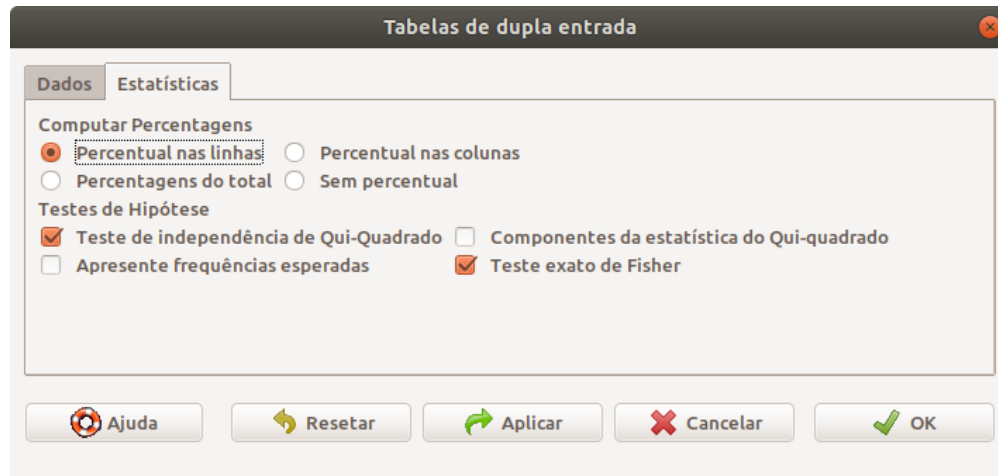


Figura 8.4: Opções para a análise de uma tabela de contingência 2x2 por meio do *R Commander*.

Os resultados são mostrados na figura 8.5. As frequências da tabela 2x2 são mostradas na primeira tabela. Observem que as colunas e linhas são ordenadas, por padrão, em ordem alfabética (linhas vermelhas).

Os riscos de morte com ou sem história de diabetes são mostrados abaixo de *Row percentages* (círculo verde). Em seguida, são mostrados os testes estatísticos do qui-quadrado e teste exato de Fisher e o intervalo de confiança para a razão de chances. Finalmente é mostrada a estimativa da razão de chances para essa tabela (linha alaranjada).

```
Frequency table:
  dead
diab FALSE TRUE
No    308  414
Yes    35   62

Row percentages:
  dead
diab FALSE TRUE Total Count
No    42.7  57.3    100    722
Yes   36.1  63.9    100    97

Pearson's Chi-squared test

data: .Table
X-squared = 1.5196, df = 1, p-value = 0.2177

Fisher's Exact Test for Count Data

data: .Table
p-value = 0.2297
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.8328888 2.1108614
sample estimates:
odds ratio
 1.317437
```

Figura 8.5: Resultado da análise da tabela de contingência 2x2 por meio do *R Commander*.

Observem, porém, que a tabela 2x2 foi montada de tal forma que a primeira linha corresponde àqueles sem história de diabetes e a segunda linha àqueles com história de diabetes. A primeira coluna corresponde aos sobreviventes e a segunda coluna aos mortos. Para calcularmos as medidas de associação entre a ocorrência de morte e histórico ou não de diabetes, precisamos montar a tabela de modo que a primeira linha corresponda aos casos com histórico de diabetes e a primeira coluna aos casos de morte. Para isso, basta alterar a ordem dos níveis das variáveis *diab* e *dead*, utilizando os dois comandos abaixo:

```
stroke$diab <- ordered(stroke$diab, levels=c("Yes", "No"))
stroke$dead <- ordered(stroke$dead, levels=c("TRUE", "FALSE"))
```

O primeiro comando ordena os níveis da variável *diab* do conjunto de dados *stroke*, sendo a ordem especificada pelo argumento *levels*. Recordando, o *\$* é utilizado para separar a variável do conjunto de dados. O comando seguinte ordena os níveis da variável *dead*.

Ao executarmos esses dois comandos e repetirmos a análise da tabela de contingência no *R Commander*, obtemos os resultados da figura 8.6.

```
Frequency table:
  dead
diab TRUE FALSE
Yes   62    35
No   414   308

Row percentages:
  dead
diab TRUE FALSE Total Count
Yes  63.9  36.1    100     97
No   57.3  42.7    100    722

Pearson's Chi-squared test

data: .Table
X-squared = 1.5196, df = 1, p-value = 0.2177

Fisher's Exact Test for Count Data

data: .Table
p-value = 0.2297
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.8328888 2.1108614
sample estimates:
odds ratio
 1.317437
```

Figura 8.6: Análise da tabela 2 x 2 (*diab* x *dead*) após a reordenação dos níveis dos fatores.

Observem agora que as linhas e colunas estão ordenadas da forma desejada: o desfecho de interesse (morte) na primeira coluna e a exposição de interesse (história de diabetes) na primeira linha.

A partir dos valores do risco de morte para histórico ou não de diabetes, podem ser obtidas as demais medidas de associação vistas neste capítulo. Porém vamos verificar uma forma de obtê-las diretamente do R. É imprescindível que os fatores estejam com os níveis ordenados de modo adequado. Caso contrário, as medidas de associação obtidas podem não ser calculadas corretamente.

Para isso, vamos instalar o pacote *epiR*, conforme mostrado na seção A.6. Uma vez instalado

o pacote, vamos executar o seguinte script no *R Commander*:

```
library(epiR)
tab <- table(stroke$diab, stroke$dead)
epi.2by2(tab, method = 'cohort.count', conf.level = 0.95)
```

A primeira função carrega o pacote *epiR* para ser utilizado. A segunda função cria uma tabela 2x2 a partir das variáveis *diab* e *dead* do conjunto de dados *stroke*. A tabela criada é armazenada no objeto *tab*. Finalmente a função *epi.2by2* irá gerar as medidas de associação para a tabela criada e calcular o intervalo de confiança para cada uma das medidas de associação, utilizando o método de coortes, supondo que esse tenha sido o método utilizado para a coleta de dados do conjunto *stroke*. As medidas de associação são mostradas na figura 8.7.

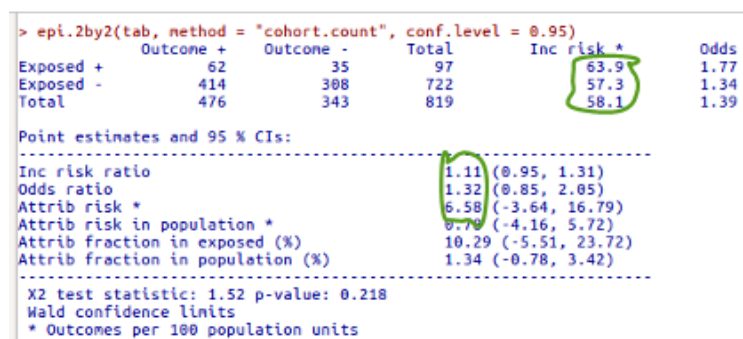


Figura 8.7: Medidas de associação obtidas para a tabela da figura 8.6 por meio do pacote *epiR*. Os riscos absolutos e as medidas de associação são mostrados dentro das linhas verdes.

O *epiR* sempre considera a primeira linha como *Exposed +* (nesse caso com histórico de diabetes) e a segunda linha como *Exposed -* (sem histórico de diabetes). Analogamente, a primeira coluna seria *Outcome +* (Morte) e a segunda coluna *Outcome -* (Vivo). Por isso é fundamental que os níveis dos fatores estejam ordenados adequadamente.

No *epiR*, o risco relativo é expresso como *Inc risk ratio*, abaixo de *Point estimates and 95% CIs*. A diferença absoluta de riscos é expressa como *Attrib risk*. São mostrados as estimativas dessas medidas e os respectivos intervalos de confiança entre parênteses.

Apesar de o *epiR* não fornecer as medidas NNT (no caso seria o NNH) e DRR, é possível obtê-las a partir dos resultados mostrados na figura 8.7, usando uma calculadora, ou então, gravando o resultado do comando *epi.2by2* em um objeto e utilizando esse objeto para calcular o NNH e o RRR. Vamos ver como fazer dessa forma.

Em primeiro lugar, executamos a função *epi.2by2* conforme abaixo:

```
result = epi.2by2(tab, method = 'cohort.count', conf.level = 0.95)
```

As saídas da função *epi.2by2* são enviadas para o objeto *result*, e não para a tela, como anteriormente. Uma consulta à ajuda do *epi.2by2* (função *help(epi.2by2)*) mostra que a saída *massoc* da função *epi.2by2* lista as estimativas das medidas de associação e os respectivos



intervalos de confiança. Assim, se mandarmos listar `result$massoc`, são listadas na tela as medidas de associação geradas pela função `epi.2by2` (figura 8.8).

```
> result$massoc
$RR.strata.wald
      est      lower      upper
1 1.114697 0.9477777 1.311013
$RR.strata.score
      est      lower      upper
1 1.114697 0.9522707 1.29081
$OR.strata.wald
      est      lower      upper
1 1.317874 0.8488883 2.045962
$OR.strata.score
      est      lower      upper
1 1.317874 0.8503403 2.042468
$OR.strata.cfield
      est      lower      upper
1 1.317874 0.8509723 2.061451
$OR.strata.nle
      est      lower      upper
1 1.317437 0.8328888 2.110861
$ARisk.strata.wald
      est      lower      upper
1 6.576806 -3.638403 16.79201
$ARisk.strata.score
      est      lower      upper
1 6.576806 -3.902575 16.20153
```

Figura 8.8: Medidas de associação obtidas pela função `epi.2by2`.

Os comandos a seguir calculam e mostram os valores de NNH e DRR a partir dos resultados mostrados na figura 8.8.

```
NNH=(1/result$massoc$ARisk.strata.score$est)*100
NNH
```

```
## [1] 15.20495
```

```
DRR = 1 - result$massoc$RR.strata.score$est
DRR
```

```
## [1] -0.1146969
```

Para calcular o NNH, precisamos da medida DAR, cuja estimativa é mostrada na linha verde da figura 8.8. Para acessar esse valor, utilizamos o caminho `result$massoc$ARisk.strata.score$est` que é formado por `result$massoc`, seguido do acesso à variável `ARisk`, indicada pela linha verde na figura 8.8, seguido da sua estimativa (`$est`)

Como a DAR está expressa em porcentagem, temos que multiplicar a sua inversa por 100 para obtermos o NNH.

Analogamente, para calcularmos a DRR, utilizamos o caminho mostrado pela região em vermelho na figura 8.8.

## 8.4 Exercícios

- 1) Qual é a relação entre o risco relativo e a razão de chances?
- 2) Considere a tabela 8.6 abaixo. Nela estão representados resultados de diversos estudos controlados com dois grupos, com os resultados em cada grupo expressos na linha correspondente a cada estudo. A partir dos resultados apresentados, calcule e preenche as lacunas para cada medida de associação expressa na tabela.

Tabela 8.6: Exemplos de valores das medidas de associação para diferentes configurações dos riscos absolutos.

Risco Basal (Controle)	Risco Grupo de Estudo	Risco Relativo	Razão de Chances	Diferença Absoluta de Riscos	NNT
0,005	0,01				
0,3	0,6				
0,01	0,005				
0,2	0,1				

- 3) Três estudos diferentes que avaliaram a associação entre duas variáveis dicotômicas obtiveram os seguintes valores para o risco relativo: 0,5; 1 e 3, respectivamente. Que valores de razão de chances são compatíveis com os valores de RR na ordem?
  - a) 1; 0,7 e 4
  - b) 0,4; 1 e 4
  - c) 0,7; 1 e 2
  - d) 0,4; 1 e 2
  - e) 0,7; 1 e 4
  - f) 0,7; 4 e 1
- 4) Carregue o conjunto de dados *stroke* do pacote *ISwR*.
  - a) Veja a ajuda do conjunto de dados.
  - b) Faça uma tabela 2x2, relacionando as variáveis *han* e *coma* e obtenha as medidas de associação: diferença de riscos, risco relativo e razão de chances. Interprete os resultados.

- c) Faça uma tabela 2x2, relacionando as variáveis *minf* e *dead* e obtenha as medidas de associação: diferença de riscos, risco relativo e razão de chances. Interprete os resultados.
- d) Gere o relatório.

# Capítulo 9

## Variáveis aleatórias

### 9.1 Noção geral de variável aleatória

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

Iremos iniciar este capítulo apresentando uma conceituação formal para variável aleatória, tomando como exemplo uma máquina de caça-níquel simples. Esse exemplo foi adaptado do capítulo 5 do livro “Head First Statistics” (Griffiths, 2008). Considerem a figura 9.1. A cada jogada, três figuras aparecem aleatoriamente na tela da máquina.



Figura 9.1: Máquina de caça-níquel. As três figuras à esquerda são os resultados que interessam e que podem dar algum ganho ao jogador. Os demais resultados resultam em perda para o jogador.

Suponhamos que a probabilidade de cada figura aparecer seja dada pela tabela 9.1.

Consideremos que o resultado de cada janela da máquina é independente dos resultados das demais janelas. O custo de cada jogada é R\$ 5,00. Para cada configuração de resultados mostrada na figura 9.2, o jogador ganha algum dinheiro. O jogador não ganha nada se ocorrer

Tabela 9.1: Probabilidades associadas a cada figura que pode aparecer em uma janela da máquina de caça-níquel.

Figura	Probabilidade
Sete	0,1
Sino	0,2
Cereja	0,2
Outra figura	0,5

qualquer outra configuração que não seja uma das mostradas na figura 9.2. Observem que o jogador ganha R\$ 75,00 se ocorrerem dois setes e uma cereja em qualquer ordem, ou seja, dois setes seguidos de uma cereja, ou se ocorrer uma cereja entre dois setes, ou se ocorrer uma cereja seguida de dois setes.

Podemos conceber esse problema como um experimento, cujo espaço amostral  $S$  são as seguintes configurações:

$$S = \{3 \text{ setes}, 3 \text{ cerejas}, 3 \text{ sinos}, 2 \text{ setes/cereja em qualquer ordem, outras configurações}\}$$

Qualquer ordem		→ R\$ 100,00
		→ R\$ 75,00
		→ R\$ 50,00
		→ R\$ 25,00

Figura 9.2: Ganhos do jogador com cada uma das quatro configurações acima da máquina caça-níquel. O jogador não ganha nada com qualquer outro resultado. O jogador ganha R\$ 75,00 com qualquer uma das configurações de uma cereja com dois setes.

O jogador está interessado em saber quais são as probabilidades de ocorrer cada uma das configurações mostradas na figura 9.2. Supondo que a probabilidade de cada figura aparecer é independente das demais, e utilizando os dados da tabela 9.1, podemos calcular facilmente as probabilidades de cada configuração:

$$P[3 \text{ setes}] = P[\text{sete}] \times P[\text{sete}] \times P[\text{sete}] = 0,1 \times 0,1 \times 0,1 = 0,001$$

$$P[3 \text{ sinos}] = P[\text{sino}] \times P[\text{sino}] \times P[\text{sino}] = 0,2 \times 0,2 \times 0,2 = 0,008$$

$$P[3 \text{ cerejas}] = P[\text{cereja}] \times P[\text{cereja}] \times P[\text{cereja}] = 0,2 \times 0,2 \times 0,2 = 0,008$$

$$\begin{aligned}
P[\text{cereja, 2 setes}] &= P[\text{cereja}] \times P[\text{sete}] \times P[\text{sete}] + P[\text{sete}] \times P[\text{cereja}] \times P[\text{sete}] \\
&\quad + P[\text{sete}] \times P[\text{sete}] \times P[\text{cereja}] \\
&= 0,2 \times 0,1 \times 0,1 + 0,1 \times 0,2 \times 0,1 + 0,1 \times 0,1 \times 0,2 \\
&= 0,002 + 0,002 + 0,002 \\
&= 0,006
\end{aligned}$$

$$\begin{aligned}
P[\text{outra configuração}] &= 1 - (P[3 \text{ setes}] + P[3 \text{ sinos}] + P[3 \text{ cerejas}] + P[\text{cereja, 2 setes}]) \\
&= 1 - 0,001 - 0,008 - 0,008 - 0,006 \\
&= 0,977
\end{aligned}$$

A tabela 9.2 mostra as probabilidades de ocorrência de cada configuração do espaço amostral. Por isso tabelas desse tipo são chamadas **Distribuições de Probabilidades**, porque mostram como as probabilidades se distribuem para o conjunto de eventos disjuntos e exaustivos do espaço amostral.

Tabela 9.2: Distribuição de probabilidades das configurações possíveis da máquina de caça-níquel.

Configuração	Probabilidade
3 Setes	0,001
3 Sinos	0,008
3 Cerejas	0,008
2 Setes/1 Cereja	0,006
Outras	0,977

A partir da distribuição de probabilidades da tabela 9.2, podemos calcular então as probabilidades dos possíveis lucros do jogador em cada jogada. Considerando que o jogador deve pagar R\$ 5,00 para cada jogada, os possíveis lucros do jogador são:

**3 Setes:** R\$ 100,00 – R\$ 5,00 = R\$ 95,00

**3 Sinos:** R\$ 25,00 – R\$ 5,00 = R\$ 20,00

**3 Cerejas:** R\$ 50,00 – R\$ 5,00 = R\$ 45,00

**Setes/Cereja:** R\$ 75,00 – R\$ 5,00 = R\$ 70,00

**Outra configuração:** R\$ 0,00 – R\$ 5,00 = – R\$ 5,00

Assim, a partir do espaço amostral de configurações possíveis da máquina de caça-níquel, podemos fazer um mapeamento para valores que são os lucros do jogador em cada configuração (figura 9.3). Vamos chamar esse mapeamento de variável X.

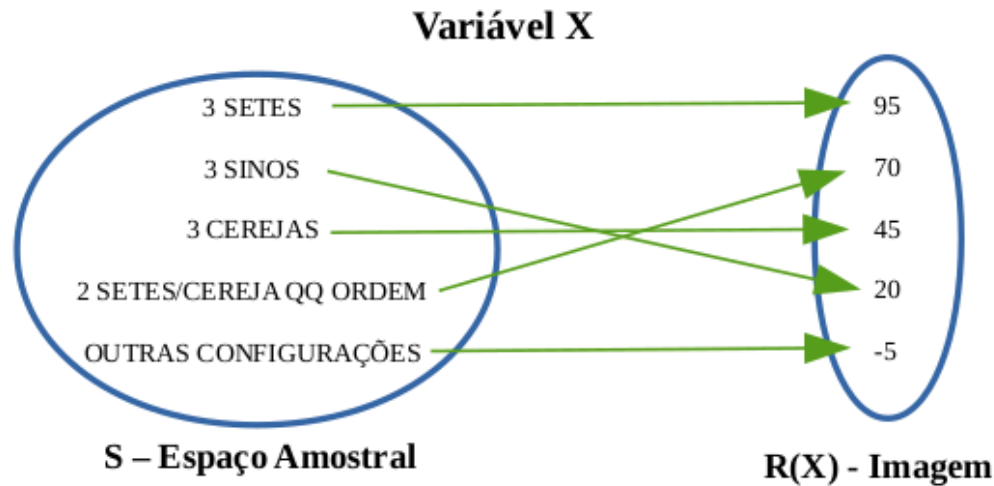


Figura 9.3: Mapeamento do espaço amostral da máquina de caça-níquel para a variável X.

Podemos associar a cada valor da imagem desse mapeamento a sua probabilidade de ocorrência (tabela 9.3). Essa tabela é a distribuição de probabilidades da variável X. Pelo fato dessa variável possuir uma distribuição de probabilidades, ela é chamada de **Variável Aleatória**.

Tabela 9.3: Distribuição de probabilidades da variável X (lucro do jogador).

<b>Lucro</b>	<b>Probabilidade</b>
95	0,001
20	0,008
45	0,008
70	0,006
-5	0,977

Todo esse raciocínio motiva a seguinte

**Definição:** Sejam um experimento e S um espaço amostral associado ao experimento. Uma função X, que associa a cada elemento  $s \in S$  um número real  $X(s)$ , é denominada variável aleatória.

Uma variável aleatória corresponde a um resultado numérico de um experimento e é usualmente representada por letras maiúsculas, como X, Y, Z, T, W. Os valores observados de uma variável aleatória X usualmente são representados por letras minúsculas,  $x_i$ ,  $i = 1, 2, \dots, N$ , onde N representa o número de experimentos.

*Portanto uma variável aleatória é uma variável que está associada a uma distribuição de probabilidades.*

Seja X uma variável aleatória. Se o conjunto de valores possíveis de X for finito ou infinito enumerável, chamamos X de **variável aleatória discreta**, isto é, os valores possíveis de X podem ser dispostos em ordem crescente e serem associados cada um a um número inteiro positivo em ordem crescente.

## 9.2 Valor esperado de uma variável aleatória discreta

Os conteúdos desta seção e da seção 9.3 podem ser visualizados neste [vídeo](#).

Um dos interesses do jogador é saber o quanto ele esperaria lucrar em cada jogada na máquina de caça-níquel. Vamos raciocinar da seguinte forma: se o jogador jogasse 1.000 vezes com a distribuição de probabilidades do lucro expressa na tabela 3, então ele esperaria 1 vez lucrar R\$ 95,00 (0,001 x 1000), R\$ 70,00 em 6 vezes, R\$ 45,00 em 8 vezes, R\$ 20,00 em 8 vezes e perderia R\$ 5,00 em 977 vezes. Então o lucro esperado do jogador (L) em 1.000 jogadas seria:

$$\begin{aligned} L &= 95 \times 1 + 70 \times 6 + 45 \times 8 + 20 \times 8 - 5 \times 977 \\ L &= - \text{R\$ } 3.850,00 \end{aligned}$$

ou seja, o jogador perderia R\$ 3.850,00 em 1.000 jogadas. Assim, em média, ele perderia R\$ 3,85 por jogada.

Usando a seguinte notação

$$\begin{aligned} x_1 &= 95, p_1 = P[X = 95] \\ x_2 &= 70, p_2 = P[X = 70] \\ x_3 &= 45, p_3 = P[X = 45] \\ x_4 &= 20, p_4 = P[X = 20] \\ x_5 &= -5, p_5 = P[X = -5] \end{aligned}$$

vamos denominar  $L_{\text{jogo}}$  como o valor esperado da variável  $X$  e usar a notação  $E[X]$ .

Logo:

$$L_{\text{jogo}} = E[X] = 95 \frac{1}{1000} + 70 \frac{6}{1000} + 45 \frac{8}{1000} + 20 \frac{8}{1000} - 5 \frac{977}{1000} = -3,85 \quad (9.1)$$

$$E[X] = x_1 \cdot p_1 + x_2 \cdot p_2 + x_3 \cdot p_3 + x_4 \cdot p_4 + x_5 \cdot p_5$$

$$E[X] = \sum_{i=1}^5 x_i p_i$$

O lucro esperado em cada jogada ( $L_{\text{jogo}}$ ) pode ser calculado como a média ponderada dos ganhos possíveis, onde o peso é a probabilidade de ocorrência de cada ganho. Assim podemos adotar a seguinte definição de valor esperado para uma variável aleatória discreta qualquer:

**Valor esperado:** Seja  $X$  uma variável aleatória discreta e que tenha associada uma distribuição de probabilidade  $p(x)$ , isto é,  $X$  pode assumir valores  $x_1, x_2, \dots, x_N$  com probabilidades  $p_1, p_2, \dots, p_N$ . O valor esperado de  $X$ ,  $E[X]$ , é definido por:

$$E[X] = \sum_{i=1}^N x_i p_i \quad (9.2)$$



$E[X]$  representa o valor esperado, ou esperança matemática, da variável aleatória  $X$ . Ao contrário da probabilidade, o valor esperado não se restringe ao intervalo  $[0,1]$ . Quando estamos tratando de variáveis aleatórias condicionadas, assim como no caso de probabilidade condicional, escrevemos  $E(X|F)$ , ou seja, o valor esperado de  $X$  dado que o evento  $F$  ocorreu.

Em geral nos referimos ao valor esperado como sendo a média da variável aleatória e representamos  $E[X]$  simplesmente por  $\mu$ .

Vamos simular a máquina caça-níqueis no R? Inicialmente, criamos dois vetores contendo os valores da variável  $x$  e as probabilidades de ocorrência de cada valor, respectivamente:

```
x = c(95, 70, 45, 20, -5)
px = c(.001, .006, .008, .008, .977)
```

Em seguida vamos simular 1000 jogadas por meio da função `sample`:

```
lucro = sample(x, size = 1000, replace = TRUE, prob = px)
```

Esse comando vai tomar 1000 amostras de  $x$  com repetição, onde cada valor de  $x$  tem a probabilidade definida por  $px$ . Em seguida, a função `table` irá indicar a frequência de cada valor de  $x$  na amostra. Um exemplo seria:

```
table(lucro)
```

```
## lucro
##  -5  20  45  70
## 976   6  11   7
```

O lucro médio dessas 1000 jogadas é dado pela função `mean` aplicada à tabela de lucros:

```
mean(lucro)
```

```
## [1] -3.775
```

Observem que o número de ocorrências dos diversos valores possíveis para o lucro em cada jogada são próximos, mas não necessariamente iguais ao número de ocorrências esperados e o lucro médio é próximo ao valor esperado calculado em (9.1) (R\$ -3,85). Se simularmos novamente, os resultados serão ligeiramente diferentes:

```
lucro = sample(x, size = 1000, replace = TRUE, prob = px)
table(lucro)
```

```
## lucro
##  -5  20  45  70  95
## 975   8   6   9   2
```

```
mean(lucro)
```

```
## [1] -3.625
```

Dá para notar que a probabilidade de ganhar no jogo é muito pequena, não é? Também é possível verificar que as frequências observadas são próximas das esperadas, assim como o lucro médio. Se aumentarmos o número de amostras, as frequências dos valores vão se aproximar ainda mais das frequências esperadas. Vamos simular mais uma vez, agora com 1.000.000 de jogadas.

```
lucro = sample(x, size = 1000000, replace = TRUE, prob = px)
table(lucro)
```

```
## lucro
##      -5      20      45      70      95
## 977193  7921  7989  5918  979
```

```
mean(lucro)
```

```
## [1] -3.860775
```

### 9.3 Variância de uma variável aleatória discreta

Nem todas as jogadas de nossa máquina caça-níquel irão resultar no mesmo valor de ganho. Na grande maioria das vezes, iremos perder, mas eventualmente podemos ganhar alguma coisa. Assim poderemos estar interessados em uma medida de variabilidade dos nossos ganhos. Vimos no capítulo 3 que uma das medidas mais utilizadas para avaliar a dispersão dos dados é a variância, ou a sua raiz quadrada, que é o desvio padrão.

Também vimos que a variância de uma amostra de  $n$  elementos é calculada da seguinte forma:

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

Quando se conhece o valor esperado da variável  $X$  ( $\mu$ ), o cálculo da variância é realizado substituindo-se  $(n-1)$  na fórmula acima por  $n$ .

Se houver  $k$  valores diferentes na amostra e o valor  $x_j$  ( $j=1, \dots, k$ ) ocorrer  $n_j$  vezes, multiplicamos  $(x_j - \mu)^2$  por  $n_j$ , somamos os produtos  $n_j (x_j - \mu)^2$ , para todos os valores  $x_j$ , e dividimos a soma por  $n$ . Fazendo  $p_j = P[X=x_j]$  e chamando a variância de  $X$  de  $\text{var}[X]$ , temos:

$$\text{var}[X] = \frac{1}{n} \sum_j n_j (x_j - \mu)^2 = \sum_{j=1}^k \frac{n_j (x_j - \mu)^2}{n} = \sum_{j=1}^k p_j (x_j - \mu)^2 \quad (9.3)$$

Isso sugere a seguinte

**Definição:** Dada uma variável aleatória discreta  $X$ , com  $k$  valores distintos, a variância de  $X$ , denotada por  $\sigma^2$ , é expressa por:

$$\text{var}[X] = \sigma^2 = \sum_{j=1}^k p_j (x_j - \mu)^2 = E[(X - \mu)^2] \quad (9.4)$$

O desvio padrão, denotado por  $\sigma$ , é a raiz quadrada da variância:

$$\sigma[X] = \sqrt{\text{var}[X]} \quad (9.5)$$

Vamos calcular a variância do ganho por jogada na nossa máquina de caça-níquel.

$$\begin{aligned} \text{var}[X] &= (95 - (-3,85))^2 \times 0,001 + (70 - (-3,85))^2 \times 0,006 + (45 - (-3,85))^2 \times 0,008 \\ &\quad + (20 - (-3,85))^2 \times 0,008 + (-5 - (-3,85))^2 \times 0,977 \\ &= 98,85^2 \times 0,001 + 73,85^2 \times 0,006 + 48,85^2 \times 0,008 + 23,85^2 \times 0,008 \\ &\quad + (-1,15)^2 \times 0,977 \\ &= 67,428 \end{aligned}$$

$$\text{Portanto } \sigma[X] = \sqrt{67,428} = 8,21$$

Vamos calcular a variância e o desvio padrão da amostra de 1.000.000 que obtivemos na seção anterior do R:

```
var(lucro)
```

```
## [1] 66.70411
```

```
sd(lucro)
```

```
## [1] 8.167258
```

## 9.4 Transformação linear

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

Suponhamos que tenha sido realizada uma alteração na nossa máquina de caça-níquel. Agora, os valores dos prêmios foram multiplicados por 5 e cada jogada passou a custar R\$ 10,00, sendo as probabilidades mantidas como eram antes. Assim a probabilidade de o jogador ganhar R\$ 500,00 é 0,001; de ganhar R\$ 375,00 é 0,006; e assim por diante. Qual seria o novo ganho esperado e a sua variância?

Vamos chamar de  $Y$  o lucro líquido do jogador nessa nova configuração da máquina caça-níquel. A tabela 9.4 mostra a distribuição de probabilidades de  $Y$ . Recordando como montar esta tabela, se numa jogada o jogador obtivesse 3 setes, então ele ganharia R\$ 500,00 e gastaria R\$ 10,00. Assim o ganho líquido dele seria R\$ 490,00. De maneira análoga, obtemos as demais linhas da tabela 9.4.

Tabela 9.4: Distribuição de probabilidades da variável  $Y$  (lucro do jogador) após a alteração da máquina caça-níquel.

Lucro	Probabilidade
490	0,001
115	0,008
240	0,008
365	0,006
-10	0,977

O valor esperado e a variância de  $Y$  poderiam ser calculados por meio das fórmulas apresentadas na seção anterior. Porém vamos calculá-los de outra maneira para ilustrar o conceito de transformação linear de uma variável aleatória.

Na introdução deste capítulo, foi visto que a variável  $X$ , o lucro em uma jogada com a máquina caça-níquel, podia ser expressa por:

$$X = G - 5 \quad (9.6)$$

onde  $G$  é o ganho na jogada e 5 era o que o jogador pagava em cada jogada.

Na nova configuração da máquina caça-níquel, o lucro (variável  $Y$ ) passa a ser:

$$Y = 5G - 10 \quad (9.7)$$

onde  $G$  era o ganho na configuração anterior.

Substituindo  $G = X + 5$  da expressão (9.6) em (9.7), obtemos:

$$Y = 5(X + 5) - 10$$

Logo:  $Y = 5X + 15$

Uma expressão desse tipo é chamada de transformação linear. Assim temos a seguinte definição:

**Transformação Linear:** dada uma variável aleatória  $X$ , e dois números reais  $a$  e  $b$ , então a variável  $Y = aX + b$  é uma transformação linear da variável  $X$ .

Se conhecermos o valor esperado de  $X$ , o valor esperado de  $Y$  pode ser calculado da seguinte forma, supondo que  $X$  tenha  $N$  elementos distintos:

$$\begin{aligned} E[Y] &= E[aX + b] \\ &= \sum_{i=1}^N (ax_i + b)p_i \\ &= \sum_{i=1}^N ax_i p_i + \sum_{i=1}^N bp_i \\ &= a \sum_{i=1}^N x_i p_i + b \sum_{i=1}^N p_i \end{aligned}$$

Logo:

$$\mathbf{E[Y] = aE[X] + b} \quad (9.8)$$

Então, para o exemplo acima, como  $Y = 5X + 15$ , o valor esperado de  $Y$  será:

$$\begin{aligned} E[Y] &= 5E[X] + 15 = 5 \times (-3,85) + 15 \\ E[Y] &= -4,25 \end{aligned}$$

Se conhecermos a variância de  $X$ , a variância de  $Y$  pode ser calculada da seguinte forma:

$$\begin{aligned} var[Y] &= var[aX + b] = E[(aX + b - E[aX + b])^2] \\ &= E[(aX + b - aE[X] - b)]^2 \\ &= E[(aX - a\mu)^2] \\ &= E[a^2(X - \mu)^2] \\ &= E[a^2(X - \mu)^2] \\ &= a^2 E[(X - \mu)^2] \end{aligned}$$

Logo:

$$\mathbf{var[Y] = a^2 var[X]} \quad (9.9)$$

Desse modo, numa transformação linear, a variância é multiplicada pelo quadrado da constante  $a$ . A constante  $b$  não influencia na dispersão da nova variável. Isso é razoável, já que se apenas somarmos uma constante a todos os valores de uma variável aleatória, então todos os seus valores terão o mesmo deslocamento, inclusive a sua média, e a dispersão em torno da nova média não irá ser alterada.

Para o exemplo acima, como  $Y = 5X + 15$ , a variância de  $Y$  será:

$$\begin{aligned}\text{var}[Y] &= 5^2 \text{ var}[X] = 25 \times 67,428 \\ \text{var}[Y] &= 1685,70 \rightarrow \sigma[Y] = 41,06\end{aligned}$$

Vamos simular esta transformação no R? Lembramos novamente que os valores de variância e desvio padrão que você obtiver irão diferenciar dos obtidos a seguir, devido ao fato de que as suas amostras serão diferentes das obtidas neste exemplo.

```
x = c(95, 70, 45, 20, -5)
px = c(.001, .006, .008, .008, .977)
y = 5*x+15
y
```

```
## [1] 490 365 240 115 -10
```

```
lucroy <- sample(y, size = 1000000, replace = TRUE, prob = px)
var(lucroy)
```

```
## [1] 1678.774
```

```
sd(lucroy)
```

```
## [1] 40.97285
```

## 9.5 Soma de variáveis aleatórias

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

Vamos supor agora que temos duas máquinas caça-níquel com as probabilidades de lucro definidas pela tabela 3. Vamos supor que o resultado de uma jogada em uma das máquinas independe do resultado de uma jogada na outra máquina. Um casal, Ana e Carlos, joga em cada uma das máquinas. Qual é a distribuição de probabilidades do lucro dos dois juntos? Quanto eles esperam lucrar juntos? Qual é a variância do lucro deles?

Vamos considerar  $X$  como o lucro obtido por Ana e  $Y$  o lucro obtido por Carlos em uma jogada. Então o lucro dos dois é a soma de  $X$  e  $Y$ . Vamos chamar essa soma de  $Z$ . Então:

$$Z = X + Y$$

Qual é a distribuição de probabilidades de  $Z$ ?

Em primeiro lugar, vamos verificar quais são os valores que  $Z$  pode assumir. Tanto  $X$  como  $Y$  podem assumir os valores 95, 70, 45, 20 e -5. Como  $X$  e  $Y$  são independentes, então  $Z$  pode assumir a soma de qualquer combinação dos valores de  $X$  e  $Y$ . Logo teríamos:

$X = 95, Y = 95 \rightarrow Z = 190$   
 $X = 95, Y = 70 \rightarrow Z = 165$   
 $X = 95, Y = 45 \rightarrow Z = 140$   
.....  
 $X = -5, Y = 20 \rightarrow Z = 15$   
 $X = -5, Y = -5 \rightarrow Z = -10$

Como os pares (X, Y) (95, 70) e (70, 95) geram o mesmo valor de Z, assim como os pares (95, 45) e (45, 95), (95, 20) e (20, 95) etc., teríamos um total de 9 diferentes valores para Z (tabela 9.5).

Para calcularmos as probabilidades de ocorrência de cada valor de Z, vamos considerar inicialmente a probabilidade de  $Z = 190$ , ou seja,  $X = 95$  e  $Y = 95$ :

$$P[Z=190] = P[X=95, Y=95]$$

Como as variáveis aleatórias X e Y são independentes, então

$$P[Z=190] = P[X=95] \cdot P[Y=95] = 0,001 \cdot 0,001 = 0,000001$$

Analogamente, teríamos:

$$P[Z=165] = P[X=95, Y=70] + P[X=70, Y=95] = 2 \cdot 0,001 \cdot 0,006 = 0,000012$$

e assim por diante. Calculando as probabilidades para cada valor de Z, obteremos a distribuição de probabilidades de Z, mostrada na tabela 9.5.

A partir da distribuição de Z, poderíamos calcular os valores de  $E[Z]$  e  $\text{var}[Z]$ . Porém vamos obtê-los de uma forma diferente, supondo que X e Y possam assumir, respectivamente,  $N_1$  e  $N_2$  elementos distintos (na tabela 9.5,  $N_1 = N_2 = 5$ ). Temos:

$$\begin{aligned}
 E[Z] &= E[X + Y] = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} (x_i + y_j) P[X = x_i, Y = y_j] \\
 &= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} x_i P[X = x_i, Y = y_j] + \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} y_j P[X = x_i, Y = y_j] \\
 &= \sum_{i=1}^{N_1} x_i \sum_{j=1}^{N_2} P[X = x_i, Y = y_j] + \sum_{j=1}^{N_2} y_j \sum_{i=1}^{N_1} P[X = x_i, Y = y_j] \\
 &= \sum_{i=1}^{N_1} x_i P[X = x_i] + \sum_{j=1}^{N_2} y_j P[Y = y_j]
 \end{aligned}$$

Logo:

$$\mathbf{E[Z]} = \mathbf{E[X + Y]} = \mathbf{E[X]} + \mathbf{E[Y]} \quad (9.10)$$

Tabela 9.5: Distribuição de probabilidades da variável  $Z = X + Y$ , sendo  $X$  e  $Y$  independentes e com distribuição dada pela tabela 9.3.

<b>X</b>	<b>Y</b>	<b>Z</b>	<b>P[Z]</b>
95	95	190	0,000001
95	70	165	0,000012
70	95		
95	45	140	0,000052
45	95		
70	70		
95	20	115	0,000112
20	95		
70	45		
45	70		
95	-5	90	0,002114
-5	95		
45	45		
70	20		
20	70		
70	-5	65	0,011852
-5	70		
45	20		
20	45		
45	-5	40	0,015696
-5	45		
20	20		
20	-5	15	0,015632
-5	20		
-5	-5	-10	0,954529



Assim o valor esperado de uma variável aleatória que é a soma de duas outras variáveis aleatórias é a soma dos valores esperados de cada uma das variáveis que compõem as parcelas da soma. Observem que, na demonstração acima, nenhuma consideração foi realizada sobre a independência das variáveis  $X$  e  $Y$ . **Essa propriedade não depende da hipótese de independência das variáveis.**

Para obtermos uma expressão para a variância de  $Z$ , vamos tomar  $E[X] = \mu_x$ ,  $E[Y] = \mu_y$ , e aplicar a fórmula para o cálculo da variância:

$$\begin{aligned} var[Z] &= var[X + Y] = E[(X + Y - E[X + Y])^2] \\ &= E[(X - E[X] + Y - E[Y])^2] \\ &= E[(X - \mu_x + Y - \mu_y)^2] \\ &= E[(X - \mu_x)^2 + (Y - \mu_y)^2 + 2(X - \mu_x)(Y - \mu_y)] \\ &= E[(X - \mu_x)^2] + E[(Y - \mu_y)^2] + 2E[(X - \mu_x)(Y - \mu_y)] \end{aligned}$$

$$var[Z] = var[X] + var[Y] + 2E[(X - \mu_x)(Y - \mu_y)] \quad (9.11)$$

Assim a variância da soma de duas variáveis aleatórias é igual à soma das variâncias de cada uma das variáveis e um terceiro termo. Quando as variáveis são independentes, então o terceiro termo será igual a zero, como mostraremos a seguir.

## 9.6 Independência de variáveis aleatórias

Dadas duas variáveis aleatórias  $X$  e  $Y$ , independentes, temos como uma propriedade básica do valor esperado que  $E[XY] = E[X]E[Y]$ . Para demonstrarmos essa propriedade, recorreremos à propriedade do valor esperado de uma função de uma variável aleatória, no caso a função  $f = XY$ :

$$\begin{aligned} E[XY] &= \sum_{i=1}^{N1} \sum_{j=1}^{N2} (x_i y_j) P[X = x_i, Y = y_j] \\ &= \sum_{i=1}^{N1} \sum_{j=1}^{N2} x_i y_j P[X = x_i] P[Y = y_j] \\ &= \sum_{i=1}^{N1} x_i P[X = x_i] \sum_{j=1}^{N2} y_j P[Y = y_j] \\ &= \sum_{i=1}^{N1} x_i P[X = x_i] \sum_{j=1}^{N2} y_j P[Y = y_j] \end{aligned}$$

$$E[XY] = E[X]E[Y]$$

Para variáveis aleatórias independentes, o valor esperado do produto de duas variáveis aleatórias é igual ao produto dos valores esperados de cada uma das variáveis. Se as variáveis **não** forem independentes, esse resultado não é válido.

O terceiro termo da expressão (9.11) pode ser desenvolvido e, assumindo X e Y independentes, resulta em:

$$\begin{aligned}
 E[(X - \mu_x)(Y - \mu_y)] &= E[XY] - \mu_y E[X] - \mu_x E[Y] + E[\mu_x \mu_y] \\
 &= E[X]E[Y] - \mu_x \mu_y - \mu_x \mu_y + \mu_x \mu_y \\
 &= \mu_x \mu_y - \mu_x \mu_y - \mu_x \mu_y + \mu_x \mu_y \\
 &= 0
 \end{aligned}$$

Logo, **para variáveis independentes**, temos:

$$\mathbf{var}[X + Y] = \mathbf{var}[X] + \mathbf{var}[Y]$$

Para o exemplo da seção anterior,  $\mathbf{var}[Z=X+Y] = 67,428 + 67,428 = 134,856$

Caso a variável aleatória fosse  $Z = X - Y$ , X e Y independentes, podemos mostrar facilmente que:

$$\begin{aligned}
 \mathbf{E}[Z] &= \mathbf{E}[X - Y] = \mathbf{E}[X] - \mathbf{E}[Y] \\
 \mathbf{var}[Z] &= \mathbf{var}[X - Y] = \mathbf{var}[X] + \mathbf{var}[Y]
 \end{aligned}$$

Assim o valor esperado da diferença de duas variáveis aleatórias é a diferença dos valores esperados das duas variáveis e a variância da diferença, assim como a variância da soma, é a soma das variâncias, desde que as variáveis sejam independentes.

Generalizando um pouco mais, se X e Y são duas variáveis aleatórias, e

$$\mathbf{Z} = \mathbf{aX} + \mathbf{bY}$$

então:

$$\mathbf{E}[Z] = \mathbf{aE}[X] + \mathbf{bE}[Y] \quad (9.12)$$

Além disso, se X e Y são variáveis aleatórias independentes, então:

$$\mathbf{var}[X + Y] = \mathbf{a^2 var}[X] + \mathbf{b^2 var}[Y] \quad (9.13)$$

Os resultados acima podem ser estendidos para a soma de mais de duas variáveis, ou seja, o valor esperado da soma de duas ou mais variáveis será sempre a soma dos valores esperados das variáveis. Assumindo independência entre as variáveis aleatórias, a variância da soma de duas ou mais variáveis será sempre a soma das variâncias dessas variáveis.

Resumindo, se  $X$  e  $Y$  são variáveis aleatórias, então:

$$E[kX] = kE[X]$$

$$E[aX \pm bY] = aE[X] \pm bE[Y]; \text{ } a \text{ e } b \text{ constantes}$$

$$E[XY] = E[X]E[Y], \text{ se } X \text{ e } Y \text{ são independentes}$$

$$E[X - E[X]] = 0$$

$$\text{var}[aX \pm bY] = a^2\text{var}[X] + b^2\text{var}[Y], \text{ se } X \text{ e } Y \text{ são independentes}$$

$$\text{var}[cX] = c^2\text{var}[X]$$

Finalmente sejam  $X_1, X_2, X_3, \dots, X_n$   $n$  variáveis independentes, onde  $E[X_i] = \mu$  e  $\text{var}[X_i] = \sigma^2$ , ou seja,  $n$  variáveis aleatórias com a mesma média e a mesma variância. Nesse caso, se  $S_n$  é a soma dessas variáveis aleatórias:

$$S_n = X_1 + X_2 + X_3 + \dots + X_n$$

e  $\bar{X}$  é a média das variáveis  $X_1, X_2, X_3, \dots, X_n$ :

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n},$$

teremos os seguintes resultados importantes:

$$E[S_n] = n\mu$$

$$\text{var}[S_n] = n\sigma^2$$

$$E[\bar{X}] = \mu$$

$$\text{var}[\bar{X}] = \frac{\sigma^2}{n}$$

## 9.7 Exercício

- 1) Um jogo consiste no lançamento de um dado não viciado. Se ocorrer 6, a pessoa ganha R\$ 100,00. Se ocorrer qualquer outro resultado, ela perde R\$ 25,00.
  - a. Se a pessoa lançar o dado uma vez, qual é a distribuição de probabilidades do valor recebido pela pessoa? Qual é o valor esperado do seu ganho?
  - b. Se duas pessoas jogam cada uma o dado, qual a distribuição de probabilidades do ganho das duas pessoas juntas? Qual é o valor esperado do ganho das duas pessoas juntas?
  - c. Se o jogo se altera e os valores são triplicados para cada ocorrência do dado, qual é a distribuição de probabilidades do valor recebido pela pessoa? Qual é o valor esperado do seu ganho?

# Capítulo 10

## Distribuições de variáveis aleatórias discretas

### 10.1 Introdução

Um dos conceitos fundamentais em estatística é o de distribuição de probabilidades.

Como visto no capítulo anterior, uma distribuição de probabilidades, também conhecida como função de probabilidade, associa um valor de probabilidade a cada valor possível de uma variável aleatória discreta. Portanto uma variável aleatória discreta  $X$ , que possa assumir  $k$  valores diferentes, tem associada a cada valor  $X = x_i$ ,  $i = 1, 2, \dots, k$ , uma probabilidade definida por  $P(X = x_i) = p_i$ . Existem diversas funções de probabilidades teóricas que são usadas para modelar problemas envolvendo variáveis discretas.

Quando assumimos uma distribuição teórica para representar o resultado de um experimento, estamos descrevendo um *modelo estatístico* para o problema. Este capítulo apresenta três distribuições de probabilidades para variáveis discretas bastante utilizadas: binomial, Poisson e geométrica.

### 10.2 Distribuição binomial

Os conteúdos desta seção e de suas subseções podem ser visualizados neste [vídeo](#).

A distribuição binomial é uma das distribuições de probabilidades mais utilizadas para modelar fenômenos aleatórios discretos. A distribuição binomial descreve as probabilidades do número de sucessos em um certo número de experimentos ( $n$ ) se as seguintes condições são satisfeitas:

1. O número de experimentos  $n$  é fixo;
2. Cada experimento é independente;
3. O resultado de cada experimento é um de dois possíveis desfechos (sucesso ou fracasso, 0 ou 1, etc). Experimentos desse tipo são conhecidos como experimentos de Bernoulli;
4. A probabilidade de sucesso  $p$  é a mesma em cada experimento.

Se as condições acima são satisfeitas, então  $X$  possui uma distribuição binomial com parâmetros  $n$  e  $p$ , e podemos abreviadamente representá-la por  $X \sim B(n, p)$ .

Recordando, a tabela 8.3 do capítulo 8 compara o risco de desenvolvimento de diabetes mellitus em pessoas com IMC (Índice de Massa Corporal) na faixa de 14,5 - 24,9 kg/m<sup>2</sup> com pessoas com IMC > 30 kg/m<sup>2</sup>. A distribuição binomial é utilizada para modelar o número de pessoas que desenvolvem diabetes em cada uma dessas duas categorias de IMC. Tomando o grupo de pessoas com IMC > 30 kg/m<sup>2</sup>, por exemplo, vimos que a proporção de pessoas que desenvolveram diabetes mellitus no estudo foi de:

$$R_{IMC>30} = \frac{156}{1384} = 11\%$$

Assim poderíamos utilizar uma distribuição binomial para modelar o número de pessoas com IMC > 30 kg/m<sup>2</sup> que desenvolverão diabetes mellitus. Aplicando as condições 1 a 4 acima, esse modelo seria válido, se:

1. fixamos o número de experimentos, ou seja, o número de pessoas com IMC > 30 kg/m<sup>2</sup> que iremos acompanhar. No caso da tabela 3, faremos  $n = 1384$ ;
2. cada experimento é independente, ou seja, o fato de uma pessoa do grupo desenvolver diabetes mellitus independe do desfecho nas outras pessoas;
3. o desfecho de cada experimento é uma variável dicotômica; nesse caso, cada pessoa desenvolve ou não diabetes mellitus;
4. a probabilidade do desfecho  $p$  é a mesma em cada experimento. Aqui consideraremos que a probabilidade de ocorrência de diabetes mellitus é a mesma para todas as pessoas no grupo estudado. Essa probabilidade foi estimada como 0,11.

Portanto o número de pessoas do grupo com IMC > 30 kg/m<sup>2</sup> que terão diabetes mellitus ( $X$ ) é uma variável aleatória discreta que segue uma distribuição binomial, com parâmetros  $n = 1384$  e  $p = 0,11$ . Então:

$$X \sim B(1384; 0,11)$$

A variável aleatória  $X$  pode assumir os valores 0, 1, 2, 3, ..., 1384, ou seja, para as 1384 pessoas com IMC > 30 kg/m<sup>2</sup>, pode acontecer que nenhuma pessoa fique doente, ou 1, 2, 3, ou até todas as pessoas venham a desenvolver diabetes mellitus. Nesse estudo em particular, 156 pessoas ficaram doentes, mas qualquer um dos valores entre 0 e 1384 poderia ser observado.

### 10.2.1 Probabilidades de uma distribuição binomial

Para uma variável aleatória que segue a distribuição binomial, as probabilidades de ocorrência de cada valor  $k$  entre 0 e  $n$  (número de experimentos) são dadas pela seguinte expressão:

$$P(X = k) = \binom{n}{k} p^k q^{n-k} = \frac{n!}{k!(n-k)!} p^k q^{n-k} \quad (10.1)$$

Na expressão acima,  $q = 1 - p$ . O símbolo ! significa fatorial.

Assim  $n! = n.(n - 1).(n - 2) \dots 3.2.1$ .

Para o exemplo da seção anterior, teríamos:

$$P(X = k) = \binom{1384}{k} 0,11^k 0,89^{1384-k} = \frac{1384!}{k!(1384 - k)!} 0,11^k 0,89^{1384-k}$$

Com a fórmula acima, poderíamos calcular a probabilidade de nenhuma pessoa desenvolver o diabetes mellitus ( $k = 0$ ), 1 pessoa ( $k = 1$ ) desenvolver a doença e assim por diante. Apenas como ilustração, a probabilidade de nenhuma pessoa desenvolver a doença será:

$$P(X = 0) = \frac{1384!}{0!(1384 - 0)!} 0,11^0 0,89^{1384-0} = 9,03 \cdot 10^{-71},$$

uma probabilidade extremamente baixa.

Vamos utilizar o R para estudarmos um pouco mais a distribuição binomial. Vamos utilizar uma distribuição binomial com menos experimentos do que o exemplo anterior. Vamos considerar a seguinte situação hipotética: 20 pessoas foram submetidas a uma cirurgia cuja probabilidade de sucesso é de 40%, ou seja,  $p = 0,4$ . Quais seriam as probabilidades de 1, 2, 3, ou todas as 20 cirurgias serem bem sucedidas?

Considerando cada cirurgia como um experimento de Bernoulli, o número de sucessos em 20 cirurgias ( $X$ ) será uma variável aleatória com distribuição binomial com  $n = 20$  e  $p = 0,4$ . Logo:

$$P(X = k) = \frac{20!}{k!(20 - k)!} 0,4^k 0,6^{20-k}$$

Substituindo  $k$  na fórmula acima por 0, 1, 2, ..., 20, obteremos as probabilidades de observarmos 0, 1, 2, ..., 20 cirurgias, supondo que a distribuição binomial seja válida para esses 20 experimentos. Em vez de realizarmos esses cálculos manualmente, vamos utilizar o *R Commander*. A figura 10.1 mostra como acessar o menu que contém diversos recursos para trabalhar com a distribuição binomial. Vamos começar selecionando o item *Probabilidades da binomial*.

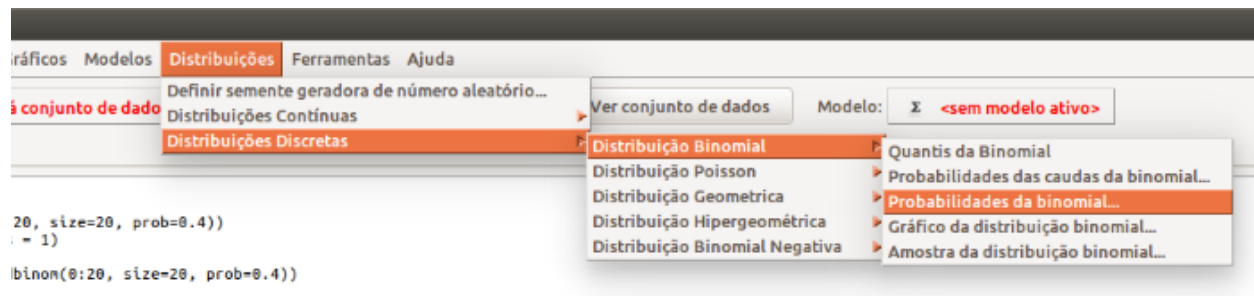


Figura 10.1: Acessando o menu do *R Commander* que nos permite trabalhar com a distribuição binomial.

Como o nome indica, ao selecionarmos a opção *Probabilidades da binomial*, somos levados a uma caixa de diálogo na qual especificamos os parâmetros da distribuição binomial que desejamos (figura 10.2). Nesse exemplo,  $n = 20$  (número de experimentos) e  $p = 0,4$ . Não esqueçamos que o separador de decimais no R é o ponto, não a vírgula. Ao clicarmos em OK, as probabilidades serão mostradas na console do *R Studio* (figura 10.3) ou na área de resultados do *R Commander*.

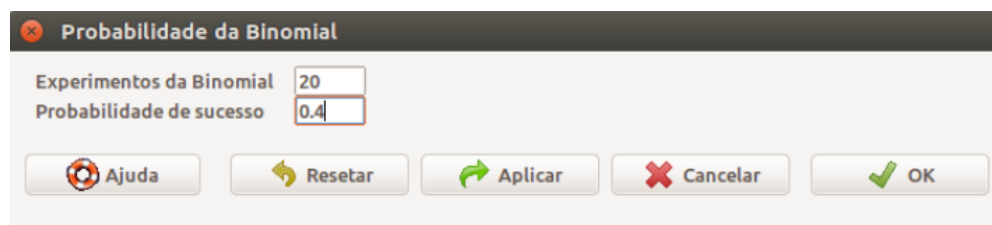


Figura 10.2: Diálogo do *R Commander*, onde especificamos os parâmetros da binomial. Depois, clicamos em OK para obtermos as probabilidades.

```
Console Terminal x Jobs x
~/Estatistica/livro/bookdown/estatistica/

Rcmdr> local({
Rcmdr+   .Table <- data.frame(Probability=dbinom(0:20, size=20, prob=0.4))
Rcmdr+   rownames(.Table) <- 0:20
Rcmdr+   print(.Table)
Rcmdr+ })
      Probability
0 0.00003656158440
1 0.00048748779201
2 0.00308742268272
3 0.01234969073088
4 0.03499079040416
5 0.07464701952887
6 0.12441169921478
7 0.16588226561971
8 0.17970578775469
9 0.15973847800417
10 0.11714155053639
11 0.07099487911296
12 0.03549743955648
13 0.01456305212574
14 0.00485435070858
15 0.00129449352229
16 0.00026968615048
17 0.00004230370988
18 0.00000470041221
19 0.00000032985349
20 0.00000001099512
> |
```

Figura 10.3: Probabilidades de 0, 1, 2, ..., 20 cirurgias serem bem sucedidas, supondo que o número de cirurgias bem sucedidas segue a distribuição  $B(20, 0.4)$ .

Para obtermos o gráfico da distribuição binomial no *R Commander*, selecionamos a opção:

Distribuições  $\Rightarrow$  Distribuições discretas  $\Rightarrow$  Dist. Binomial  $\Rightarrow$  Gráfico Dist. Binomial...

Na caixa de diálogo da figura 10.4, novamente especificamos os parâmetros da distribuição binomial e selecionamos uma das duas opções de gráficos. Selecionando a primeira opção, *Gráfico da função de massa*, obtemos o gráfico mostrado na figura 10.5.



Figura 10.4: Diálogo do *R Commander* para obtermos os gráficos de uma distribuição binomial.



**Binomial Distribution: Binomial trials=20, Probability of success=0.**

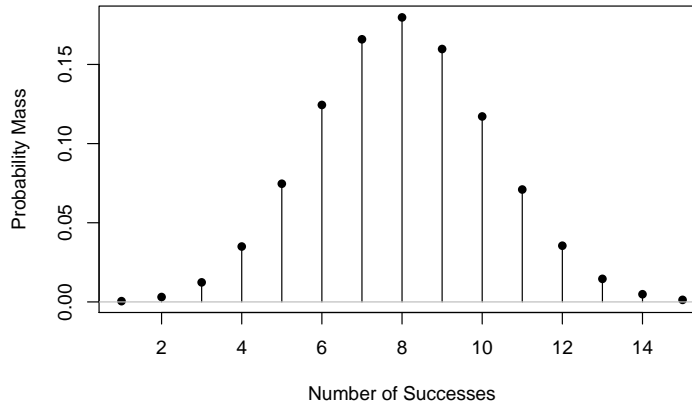


Figura 10.5: Gráfico da distribuição binomial B(20, 0,4).

O gráfico da distribuição de probabilidades mostra, para cada valor possível da variável aleatória, uma linha vertical indicando a probabilidade de ocorrer aquele valor. Assim observem que as probabilidades mostradas correspondem às aquelas observadas na listagem da figura 10.3. As probabilidades para os valores acima de 15 e abaixo de 1 não foram mostradas, porque elas são muito pequenas, quando comparadas com os demais valores.

Ao selecionarmos a segunda opção na figura 10.4, *Gráfico da função cumulativa*, obtemos o gráfico da figura 10.6.

**Binomial Distribution: Binomial trials=20, Probability of success=0.**

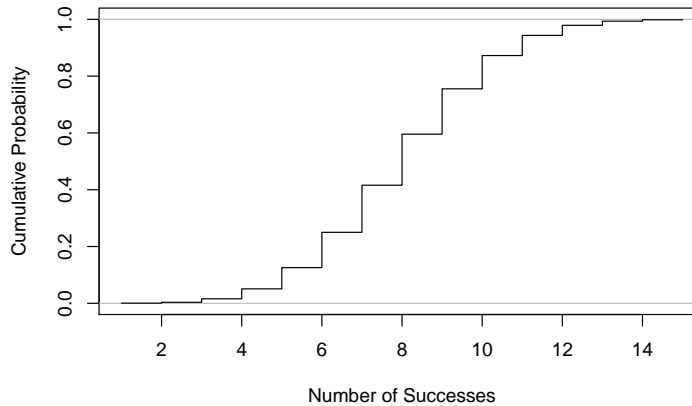


Figura 10.6: Gráfico da distribuição binomial acumulada B(20, 0,4).

O gráfico da distribuição cumulativa, ou acumulada, fornece para cada valor k da variável aleatória discreta, X, a probabilidade de observarmos um número de ocorrências menor ou igual a k. Assim:

$$ProbabilidadeAcumulada = F(X = k) = \sum_{0}^k P(X = k), 0 \leq k \leq n$$

Para ilustrar:

$$F(0) = P(0) = 0,00003656$$

$$F(1) = P(0) + P(1) = 0,00003656 + 0,00048749 = 0,00052405$$

.....

$$F(20) = P(0) + P(1) + P(2) + \dots + P(20) = 1$$

O gráfico da distribuição acumulada possui um formato de escada, pois é sempre crescente. O valor máximo de  $F$  é 1 e o valor mínimo é 0.

A figura 10.7 mostra o gráfico da distribuição binomial  $B(10, 0,5)$ . Observem que o gráfico é simétrico em torno do valor 5. Essa simetria é observada para todas as distribuições binomiais, onde a probabilidade de sucesso é igual a 0,5.

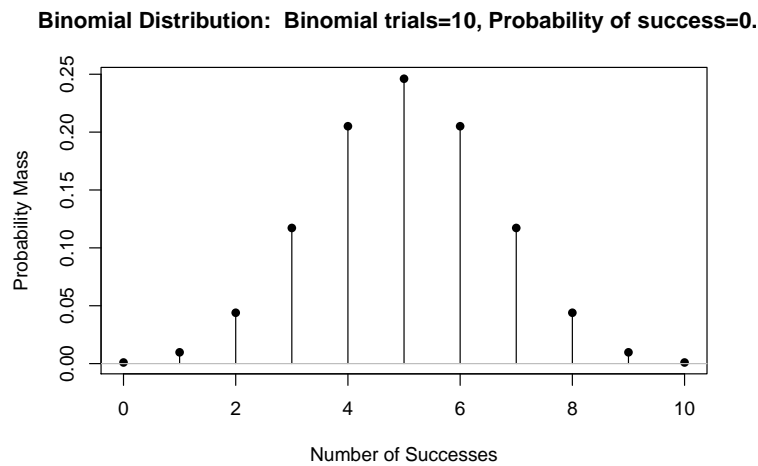


Figura 10.7: Gráfico da distribuição binomial  $B(10, 0,5)$ .

A figura 10.8 mostra o gráfico da distribuição binomial  $B(10, 0,1)$ . Observem agora que o gráfico é bastante assimétrico, já que a probabilidade de sucesso é apenas 0,1.

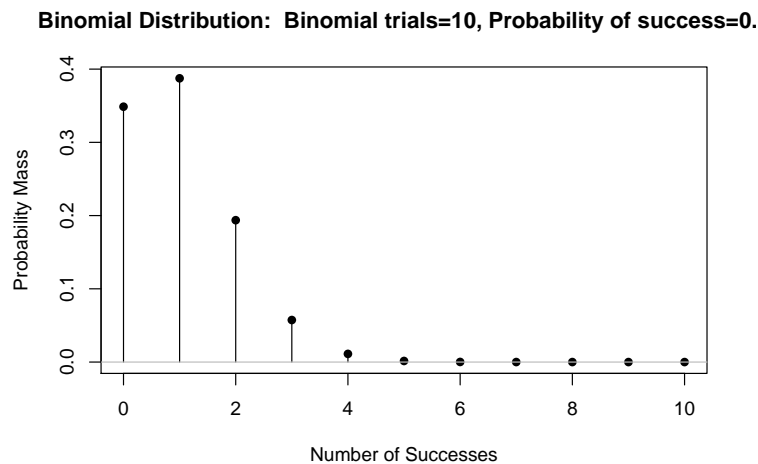


Figura 10.8: Gráfico da distribuição binomial  $B(10, 0,1)$ .

Caso queiramos obter as probabilidades de ocorrência de até  $k$  sucessos, ou seja, a probabilidade acumulada para  $X = k$ , utilizamos a segunda opção no menu da figura 10.1: *Probabilidades das caudas da binomial*. A caixa de diálogo da figura 10.9 nos permite especificar um ou mais valores de  $k$ , os parâmetros da binomial e se vamos utilizar a cauda inferior ou superior da distribuição.



Figura 10.9: Diálogo do *R Commander* para obtermos as probabilidades das caudas de uma distribuição binomial. Ao selecionarmos a opção *cauda inferior*, vamos calcular a probabilidade acumulada para o valor especificado (ou valores especificados) em *Valores da Variável*.

Para ilustrar, especificamos os parâmetros da distribuição do nosso exemplo  $B(20, 0.4)$ , vamos fazer  $k = 2$  (ocorrência de 2 cirurgias em 20) em *Valores da Variável*, e vamos selecionar a opção *Cauda inferior*. Isso significa que desejamos a probabilidade de ocorrer até 2 cirurgias, ou seja, queremos  $F(2) = P(0) + P(1) + P(2)$ . O comando gerado pelo R é mostrado abaixo, seguido do resultado de sua execução:

```
pbinom(c(2), size=20, prob=0.4, lower.tail=TRUE)
```

```
## [1] 0.003611472
```

Ao selecionarmos a opção *Cauda superior*, vamos calcular a probabilidade de observarmos um número de sucessos **acima** do valor especificado (ou valores especificados) em *Valores da Variável*. Observem que a probabilidade fornecida é para o número de ocorrências acima, não igual ou acima, do valor especificado. Isso equivale ao valor de  $1-F(k)$ . Como exemplo, na figura 10.10, selecionamos a *cauda superior* e especificamos vários valores para  $k$  em *Valores da Variável*, separados por vírgula. Isso significa que serão calculadas as probabilidades da cauda superior para cada um dos valores (17, 18, 19, 20)



Figura 10.10: Diálogo do *R Commander* para obtermos as probabilidades das caudas de uma distribuição binomial. Ao selecionarmos a opção *Cauda superior*, vamos calcular a probabilidade de observarmos um número de sucessos **acima** do valor especificado (ou valores especificados) em *Valores da Variável*.

O comando gerado pelo R é mostrado abaixo, seguido das quatro probabilidades, uma para cada um dos quatro valores especificados. Observem que a probabilidade da cauda superior para 20 é zero, porque não pode ocorrer mais de 20 cirurgias bem sucedidas em 20. O valor para  $k = 19$  é igual à probabilidade de observarmos 20 cirurgias (veja o valor de  $P(20)$  na figura 10.3). A probabilidade da cauda superior para  $k = 18$  é igual a  $P(19) + P(20) = 1 - F(18)$ , e assim por diante. A letra *e* mostrada nesses resultados significa a base 10, não o número irracional *e*.

```
pbinom(c(17,18,19,20), size=20, prob=0.4, lower.tail=FALSE)
```

```
## [1] 5.041261e-06 3.408486e-07 1.099512e-08 0.000000e+00
```

Assim como, ao especificarmos um certo número de ocorrências do desfecho de interesse, podemos obter a probabilidade das caudas da distribuição binomial, podemos, inversamente, ao especificar uma probabilidade, obter o número limite de ocorrências nas caudas da distribuição binomial, ou quantis da binomial. Vamos ver alguns exemplos.

A caixa de diálogo para obtermos os quantis da distribuição binomial (figura 10.11) aparece ao selecionarmos a primeira opção no menu da figura 10.1. Ao selecionarmos a opção *Cauda inferior*, vamos obter o **menor** número de ocorrências do desfecho, cuja probabilidade acumulada seja maior ou igual à probabilidade especificada no campo *Probabilidades*. Nesse exemplo, colocamos em *Probabilidades* o valor de probabilidade igual ao resultado obtido ao obtermos a probabilidade acumulada do número de ocorrências igual a 2, que é o resultado gerado a partir da configuração da figura 10.9. A saída do R é gerada pelo comando a seguir.

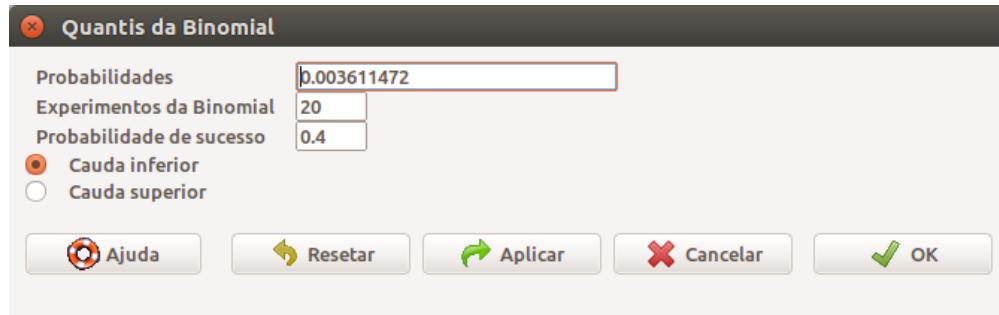


Figura 10.11: Diálogo do *R Commander* para obtermos os quantis da distribuição binomial. Ao selecionarmos a opção *Cauda inferior*, vamos obter o menor número de ocorrências do desfecho cuja **probabilidade acumulada** seja **maior ou igual** à probabilidade especificada no campo *Probabilidades*.

```
qbinom(c(0.003611472), size=20, prob=0.4, lower.tail=TRUE)
```

```
## [1] 2
```

De maneira análoga, ao especificarmos a *Cauda superior*, vamos obter o número mínimo de ocorrências do desfecho onde **1 - probabilidade acumulada** seja menor ou igual à probabilidade especificada no campo *Probabilidades*. Por exemplo, especificando o valor 5.041261e-06 (figura 10.12), obtemos o valor 17, que corresponde ao primeiro valor especificado no campo *Valores da variável* na figura 10.10.



Figura 10.12: Diálogo do *R Commander* para obtermos os quantis da distribuição binomial. Ao selecionarmos a opção cauda superior, vamos obter o número mínimo de ocorrências do desfecho onde **1 - probabilidade acumulada** seja menor ou igual à probabilidade especificada no campo *Probabilidades*.

```
qbinom(c(5.041261e-06), size=20, prob=0.4, lower.tail=FALSE)
```

```
## [1] 17
```

Vamos ver um outro exemplo. Suponhamos que queiramos obter o quantil da  $B(20, 0,4)$  para o qual a probabilidade de valores acima dele seja igual a 0,6 (Figura 10.13). O resultado é 7. Vejamos por que. As probabilidades da cauda superior para  $X = 6, 7$  e  $8$  são respectivamente 0,750, 0,584 e 0,404. O valor  $X = 7$  corresponde ao menor número de ocorrências  $k$  para o qual  $P(X > k) \leq 0,6$ .



Figura 10.13: Diálogo do *R Commander* para obtermos o quantil da  $B(20, 0,4)$  para o qual 1 - probabilidade acumulada é igual a 0,6.

```
qbinom(c(0.6), size=20, prob=0.4, lower.tail=FALSE)
```

```
## [1] 7
```

Finalmente a última opção no menu da figura 10.1 pode ser usada para simular experimentos com a distribuição binomial. A partir da caixa de diálogo *Sample from Binomial Distribution* (figura 10.14), vamos simular um número de experimentos com uma distribuição binomial  $B(20, 0,4)$ . Nessa figura, além dos parâmetros da binomial, também especificamos os seguintes parâmetros:

```
nome do conjunto de dados: BinomialSamples
tamanho da amostra: 15
número de observações: 10
```

Além disso, marcamos a opção *Médias amostrais*. Ao clicarmos em OK, serão geradas 15 linhas, cada linha contendo 10 possíveis números de ocorrências do desfecho em 20 experimentos, onde a probabilidade de ocorrência do evento em cada experimento é 0,4. Cada linha conterá também a média dos valores dos 10 números de ocorrências da linha. Os resultados serão armazenados no conjunto de dados *BinomialSamples*.

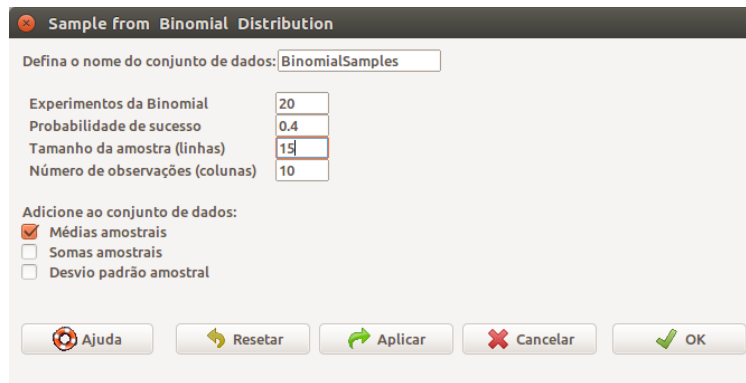


Figura 10.14: Diálogo do *R Commander* para gerar amostras de uma distribuição binomial.

Para observarmos as amostras geradas, clicamos no botão *Ver conjunto de dados* (figura 10.15), certificando-nos que o conjunto de dados *BinomialSamples* seja o conjunto ativo. Os dados gerados são mostrados na figura 10.16. Cada vez que gerarmos amostras da distribuição binomial, obteremos resultados diferentes.

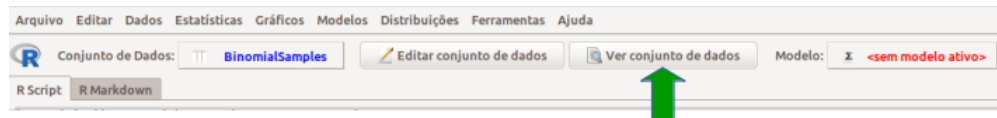


Figura 10.15: Tela principal do *R Commander*. Visualizando o conjunto de dados *BinomialSamples*, obtido a partir da opção *Sample from Binomial Distribution* do menu da figura 10.1.

Observem que os números de ocorrência do desfecho mais frequentes estão na faixa entre 7 e 12, o que era de se esperar, já que as maiores probabilidades de ocorrência se situam nesta faixa (figura 10.3). Observem também que as médias do número de ocorrências em cada linha se situam entre 7,0 e 8,9. Isso também é de se esperar, conforme iremos ver na próxima seção.

	obs1	obs2	obs3	obs4	obs5	obs6	obs7	obs8	obs9	obs10	mean
sample1	7	11	4	9	5	8	10	9	7	7	7.7
sample2	9	6	8	8	10	7	7	8	6	10	7.9
sample3	7	7	6	9	10	4	6	9	9	6	7.3
sample4	10	10	6	6	6	7	6	7	9	12	7.9
sample5	7	9	9	6	6	9	11	6	3	4	7.0
sample6	5	8	10	8	6	10	12	8	8	12	8.7
sample7	6	9	8	5	9	8	7	7	10	5	7.4
sample8	7	8	9	10	8	6	5	3	11	11	7.8
sample9	8	7	9	6	7	7	5	10	10	4	7.3
sample10	9	8	9	9	8	11	6	11	8	10	8.9
sample11	8	9	4	11	6	10	10	6	8	8	8.0
sample12	7	8	9	11	8	6	8	5	7	10	7.9
sample13	7	8	6	8	6	6	6	7	6	9	6.9
sample14	8	10	4	8	8	8	11	6	6	5	7.4
sample15	8	8	7	7	8	8	7	8	6	7	7.4

Figura 10.16: Amostras da distribuição binomial geradas a partir das opções estabelecidas na caixa de diálogo da figura 10.14.

**Outro exemplo de uso da distribuição binomial:** evidências anteriores indicaram que um dado componente eletrônico de um dispositivo biomédico tem a probabilidade 0,98 de funcionar satisfatoriamente. Ao comprar 5 desses componentes, deseja-se saber qual é a probabilidade de encontrarmos dois ou mais com defeito.

Entendendo como sucesso o evento componente sem defeito e considerando que temos 5 experimentos, então, para 2 ou mais equipamentos estarem com defeito, significa que, no máximo, 1 está funcionando corretamente, ou seja, desejamos  $P(X \leq 1)$ :

$$P(X \leq 1) = P(X = 0) + P(X = 1) = F(1)$$

Portanto podemos usar a fórmula (10.1) para calcularmos as probabilidades  $P(0)$  e  $P(1)$  e somarmos os dois valores, ou, usando o *R Commander*, na opção *Probabilidades das caudas da binomial*, preencheremos os campos conforme a figura 10.17 e obtemos o valor  $P(X \leq 1) = 7,872 \cdot 10^{-7}$ .



Figura 10.17: Diálogo do *R Commander* para obtermos a probabilidade acumulada da  $B(5, 0,98)$  para o número de ocorrências igual a 1.

## 10.2.2 Valor esperado e variância de uma distribuição binomial

### 10.2.2.1 Valor esperado

Para uma variável aleatória que segue uma distribuição binomial  $B(n,p)$ , temos que o **número esperado de sucessos** é dado por **np**. Abaixo segue a demonstração desse resultado. O leitor pode saltar essa demonstração diretamente para a seção 10.2.2.2 sem perda de continuidade.

Usando a definição de valor esperado, podemos obter esse valor por aplicação da fórmula:

$$E[X] = \sum_0^n k.P(X = k) = \sum_0^n k.\binom{n}{k}p^kq^{n-k}$$

Porém existe uma maneira mais fácil de calcular o  $E[X]$ . Para facilitar a solução desse problema e de outros, vamos introduzir aqui o uso de variáveis indicadoras. Para obtermos  $E[X]$ , vamos definir  $n$  novas variáveis  $X_i$ ,  $i=1, 2, 3, \dots, n$ :

$$X_i = \begin{cases} 1, & \text{se o } i\text{ésimo experimento for um sucesso} \\ 0, & \text{caso contrário} \end{cases}$$



ou seja, para cada experimento, teremos uma indicação se ele foi um sucesso (1) ou um fracasso (0). Por exemplo, se  $n = 5$ , poderíamos ter como resultado *SSFFS*, ou seja,  $X = 3$  sucessos, e os valores das variáveis indicadoras seriam  $X_1 = 1$ ,  $X_2 = 1$ ,  $X_3 = 0$ ,  $X_4 = 0$  e  $X_5 = 1$ . O número total de sucessos,  $X$ , é igual à soma das variáveis indicadoras, isto é:

$$X = X_1 + X_2 + X_3 + X_4 + X_5 = 3 \quad (10.2)$$

Como o valor esperado de uma soma é igual à soma dos valores esperados, teríamos para  $n$  experimentos:

$$E[X] = E[X_1] + E[X_2] + \dots + E[X_n]$$

O valor esperado para cada  $X_i$  para esse experimento é obtido assim:

$$E[X_i] = 0 \cdot P(X_i = 0) + 1 \cdot P(X_i = 1) \quad (10.3)$$

$$= P(X_i = 1) \quad (10.4)$$

$$= p \quad (10.5)$$

Teremos, então, para o valor esperado da binomial

$$E[X] = p + p + \dots + p \quad (10.6)$$

$$E[X] = np \quad (10.7)$$

### 10.2.2.2 Variância

A variância de uma distribuição binomial  $X \sim B(n, p)$  é dada por:

$$var(X) = np(1 - p) = npq \quad (10.8)$$

**Demonstração:** Para obtermos a variância da  $B(n, p)$  vamos usar uma propriedade da variância:

$$var(X) = E[X^2] - \mu^2 = E[X^2] - (E[X])^2$$

O  $E[X^2]$  corresponde ao valor médio quadrático de uma variável aleatória. Voltando ao nosso problema de estimar a variância da  $B(n, p)$ , usamos o resultado da seção anterior com a variável indicadora (equação (10.2)) para escrever

$$\text{var}(X) = \text{var}(X_1) + \text{var}(X_2) + \dots + \text{var}(X_n)$$

Para obtermos  $\text{var}(X_i)$ , usamos a propriedade que acabamos de apresentar. Como  $X_i$  assume apenas os valores 0 ou 1,  $X_i^2 = X_i$  e, usando o resultado (10.5), temos que:

$$E[X_i^2] = E[X_i] = p$$

Logo:

$$\text{var}(X_i) = E[X_i^2] - (E[X_i])^2 = p - p^2$$

$$\text{var}(X_i) = p(1 - p)$$

Finalmente, como a  $\text{var}(X)$  para a binomial é simplesmente a soma de  $n$  variáveis indicadoras independentes, teremos:

$$\text{var}(X) = np(1 - p) = npq$$

Para a distribuição  $B(20, 0,4)$ , teremos:

$$E[X] = 0,4 \times 20 = 8$$

$$\text{var}(X) = 20 \times 0,4 \times (1-0,4) = 4,8$$

Esses valores são compatíveis com as amostras da  $B(20, 0,4)$ , mostradas na figura 10.16.

## 10.3 Distribuição de Poisson

Os conteúdos desta seção e da subseção 10.3.1 podem ser visualizados neste [vídeo](#).

Outra distribuição de probabilidades de uma variável aleatória discreta bastante utilizada é a distribuição de Poisson, descrita por:

$$P(X = k) = \text{Pois}(\lambda) = \frac{e^{-\lambda} \lambda^k}{k!} \quad (10.9)$$

onde  $\lambda$  é o número médio de eventos em um dado intervalo de tempo (ou outra unidade, por exemplo de distância, ou área). O modelo de Poisson é bastante utilizado para estudos de fila, ocorrência de doenças raras, falhas de equipamentos, etc. Por exemplo, poderíamos ter que o número médio de cirurgias por dia em um certo hospital é de 10, logo  $\lambda=10$ . Com base no modelo de Poisson, teríamos como obter as probabilidades de termos  $k=0,1,2,3,\dots$  cirurgias em um dia. Assim, por exemplo, a probabilidade de ocorrer 8 cirurgias em um dia é dada por:

$$P(X = 8) = \text{Pois}(\lambda) = \frac{e^{-10} 10^8}{8!} = 0,1126$$

Podemos usar o *R Commander* para obter as probabilidades de uma distribuição de Poisson, traçar gráficos, obter os quantis da distribuição, bem como as probabilidades das caudas de uma distribuição de Poisson, de maneira análoga aos procedimentos mostrados na seção sobre a distribuição binomial. Assim, por exemplo, a figura 10.18 mostra como acessar as diversas opções disponíveis para a distribuição de Poisson no *R Commander*. Uma diferença em relação à distribuição binomial é que, para a distribuição de Poisson, somente devemos especificar um parâmetro, a média, que é igual a  $\lambda$ . Na figura 10.19, especificamos o valor de  $\lambda$  igual a 10 para obtermos as probabilidades da distribuição de Poisson do exemplo acima.

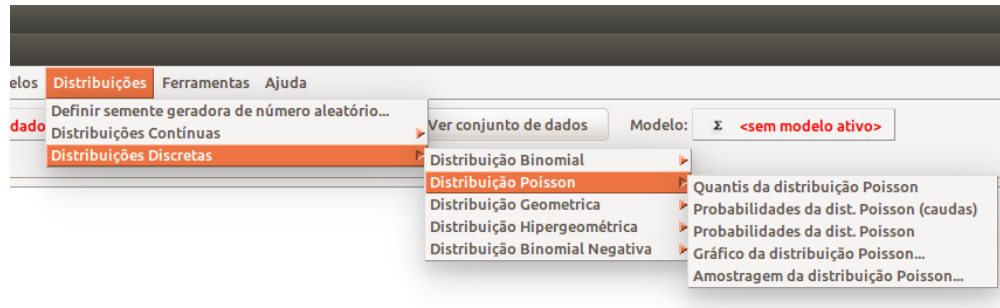


Figura 10.18: Acessando o menu do *R Commander* que nos permite trabalhar com a distribuição de Poisson.

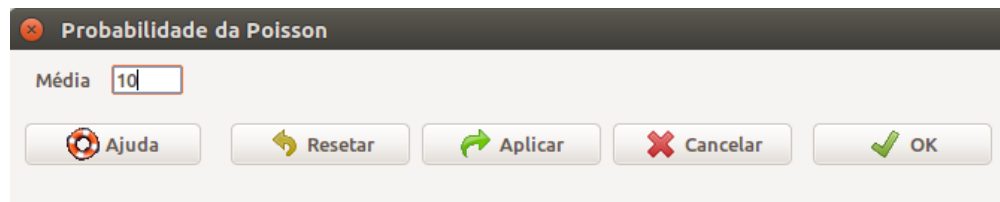


Figura 10.19: Diálogo do *R Commander*, onde especificamos o parâmetro da distribuição de Poisson. A seguir, clicamos em OK para obtermos as probabilidades.

A figura 10.20 mostra as probabilidades para a distribuição de Poisson para  $\lambda = 10$  e para o número de ocorrências variando de 2 a 22. Observem que  $P(X=8) = 0,112599$ , coincidindo com o valor calculado pela fórmula (10.9).

```

> local({
+   .Table <- data.frame(Probability=dpois(2:22, lambda=10))
+   rownames(.Table) <- 2:22
+   print(.Table)
+ })
  Probability
2  0.0022699965
3  0.0075666550
4  0.0189166374
5  0.0378332748
6  0.0630554580
7  0.0900792257
8  0.1125990321
9  0.1251100357
10 0.1251100357
11 0.1137363961
12 0.0947803301
13 0.0729079462
14 0.0520771044
15 0.0347180696
16 0.0216987935
17 0.0127639962
18 0.0070911090
19 0.0037321626
20 0.0018660813
21 0.0008886101
22 0.0004039137

```

Figura 10.20: Probabilidades de ocorrência de 2, 3, ..., 22 cirurgias em um dia, supondo que o número de cirurgias segue a distribuição  $\text{Pois}(10)$ .

A figura 10.21 mostra o gráfico da distribuição de probabilidades para uma variável aleatória que segue a distribuição de Poisson com  $\lambda = 10$ . Observem que os valores da variável aleatória  $X$  podem variar de 0 até  $\infty$ . Porém, para  $\lambda = 10$ , as probabilidades para  $X > 22$  são desprezíveis.

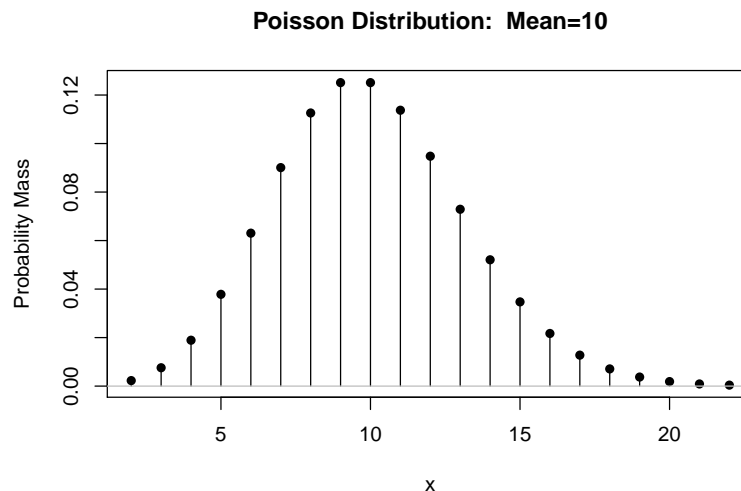


Figura 10.21: Gráfico da distribuição de Poisson para  $\lambda = 10$ .

A figura 10.22 mostra os gráficos das funções distribuição de probabilidades para variáveis aleatórias que seguem a distribuição de Poisson para  $\lambda = 0,5$ ; 1,0; 1,5; e 3, respectivamente.

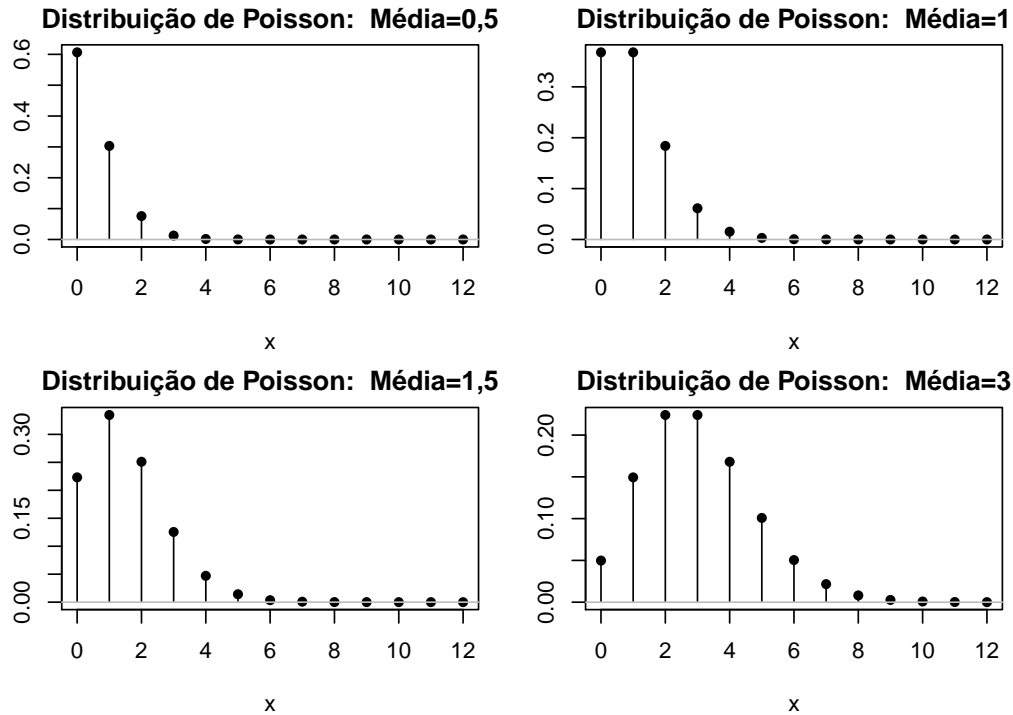


Figura 10.22: Gráficos da distribuição de probabilidades de Poisson para  $\lambda = 0,5$ ; 1,0; 1,5; e 3, respectivamente.

O modelo de Poisson pode ser entendido como a ocorrência aleatória de eventos ao longo do tempo ou do espaço uni, bi ou tridimensional. A figura 10.23 mostra a situação onde  $x$  denota um evento ocorrendo aleatoriamente ao longo do tempo.



Figura 10.23: Eventos ocorrendo aleatoriamente ao longo do tempo.

Assumindo que ocorram  $\mu$  eventos em média em um intervalo de tempo unitário, então, em um intervalo de tempo  $t$ , teremos  $\mu t$  eventos em média. Supondo que a distribuição de Poisson se aplique a esse caso, então, para calcular as probabilidades de ocorrência de um certo número de eventos no intervalo de tempo  $t$ , utiliza-se a fórmula (10.9) com o parâmetro  $\lambda$  substituído por  $\mu t$ . No exemplo da cirurgia, sabemos que, em média, ocorrem 10 cirurgias por dia. Se desejássemos a distribuição de probabilidades para o número de cirurgias por semana, então utilizaríamos  $\lambda = 10 \times 7 = 70$ . Caso desejássemos a distribuição de probabilidades para o número de cirurgias em 20 dias, então utilizaríamos  $\lambda = 10 \times 20 = 200$ .

**Exemplo:** O número de partículas que atingem um dado dispositivo de medida em 40 períodos consecutivos de um minuto é dado na tabela 10.1 (colunas 1 e 2). Assumindo que as partículas atingem o dispositivo aleatoriamente com uma taxa média constante, podemos assumir como um modelo adequado a distribuição de Poisson para descrever as frequências observadas. Para verificarmos esse modelo, podemos, a partir dos dados, calcular a taxa

média por unidade de tempo (minuto), obter as probabilidades de Poisson e, a seguir, calcular a frequência esperada do número de partículas em cada minuto.

Tabela 10.1: Número de partículas que atingem um dado dispositivo por minuto.

No. Partículas (k)	No. períodos com k partículas (Frequência Observada)	Probabilidades de Poisson	Frequência Esperada
0	13	0,301	12,04
1	13	0,361	14,44
2	8	0,217	8,68
3	5	0,087	3,48
4	1	0,026	1,04
5+	0	0,008	0,32
Total	40		

Com base nos dados da tabela 10.1, o número total de partículas observadas em 40 minutos é  $13 + (2 \times 8) + (3 \times 5) + (4 \times 1) = 48$

Para o modelo de Poisson, estimamos que a taxa média por unidade de tempo (minuto) é dada por  $\lambda = 48/40 = 1,2$ . Logo a distribuição de Poisson é dada por:

$$Pois(\lambda = 1,2) = \frac{e^{-1,2} 1,2^k}{k!}, \quad k = 1, 2, \dots$$

Usando o R, podemos calcular as probabilidades para  $k = 0, 1, 2, \dots$  e, a seguir, multiplicando-as por 40, obtemos as frequências esperadas para comparar com as observadas (quarta coluna da tabela 10.1). Nesse exemplo, vemos que as frequências observadas são próximas das frequências esperadas de acordo com o modelo de Poisson.

### 10.3.1 Valor esperado e variância de uma distribuição de Poisson

O valor esperado e a variância de uma variável aleatória que possui uma distribuição de Poisson são iguais a  $\lambda$ . A demonstração segue abaixo:

O valor esperado é dado por:

$$\begin{aligned} E[X] &= \sum_{k=0}^{\infty} kP(X = k) = \sum_{k=0}^{\infty} k \frac{e^{-\lambda} \lambda^k}{k!} \\ &= e^{-\lambda} \left( \frac{\lambda}{1!} + \frac{2\lambda^2}{2!} + \frac{3\lambda^3}{3!} + \dots \right) \\ &= e^{-\lambda} \lambda \left( 1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \dots \right) \\ &= e^{-\lambda} \lambda e^{\lambda} \\ &= \lambda \end{aligned}$$

Para determinarmos a variância, vamos primeiramente obter  $E[X^2]$ .

$$\begin{aligned} E[X^2] &= \sum_{k=0}^{\infty} k^2 P(X = k) = \sum_{k=0}^{\infty} k^2 \frac{e^{-\lambda} \lambda^k}{k!} \\ &= \sum_{k=1}^{\infty} k \frac{e^{-\lambda} \lambda^k}{(k-1)!} \end{aligned}$$

Fazendo  $k = s + 1$ :

$$\begin{aligned} E[X^2] &= \sum_{s=0}^{\infty} (s+1) \frac{e^{-\lambda} \lambda^{s+1}}{s!} \\ &= \lambda \sum_{s=0}^{\infty} s \frac{e^{-\lambda} \lambda^s}{s!} + \lambda \sum_{s=0}^{\infty} 1 \frac{e^{-\lambda} \lambda^s}{s!} \\ &= \lambda^2 + \lambda \end{aligned}$$

e a variância dessa variável aleatória é dada por:

$$\begin{aligned} \text{var}(X) &= E[X^2] - (E[X])^2 \\ &= \lambda^2 + \lambda - \lambda^2 \\ &= \lambda \end{aligned}$$

### 10.3.2 Aproximação da distribuição binomial pela de Poisson

A distribuição binomial pode ser aproximada pela de Poisson quando os parâmetros  $n$  tende a infinito e  $p$  tende a zero. Existe uma demonstração formal para essa propriedade (Höel, 1971). Podemos ver que quando  $p \rightarrow 0$ ,  $q = 1 - p \rightarrow 1$ . Logo o valor esperado e a variância da distribuição binomial serão iguais, como pode ser visto abaixo:

$$E[X] = np \text{ e } \text{var}(X) = npq = np$$

Para  $n = 100$  e  $p = 0,05$ , a figura 10.24, mostra as probabilidades da binomial  $B(100; 0,05)$  e de  $\text{Pois}(5)$  para valores de  $k=0,1,2, \dots, 14$ . Observem que os valores de probabilidades são bastante próximos, especialmente para valores de  $k < 11$ . Assim é razoável adotar

a política de que, para  $n \geq 100$  e  $p \leq 0,05$ , a distribuição binomial pode ser aproximada pela distribuição de Poisson, com parâmetro igual ao valor esperado da distribuição binomial (Höel, 1971).

a	Probability	b	x Probability
0	0.0067379470	0	0.0059
1	0.0336897350	1	0.0312
2	0.0842243375	2	0.0812
3	0.1403738958	3	0.1396
4	0.1754673698	4	0.1781
5	0.1754673698	5	0.1800
6	0.1462228081	6	0.1500
7	0.1044448630	7	0.1060
8	0.0652780393	8	0.0649
9	0.0362655774	9	0.0349
10	0.0181327887	10	0.0167
11	0.0082421767	11	0.0072
12	0.0034342403	12	0.0028
13	0.0013208616	13	0.0010
14	0.0004717363	14	0.0003

Figura 10.24: a) probabilidades para a distribuição  $B(100; 0,05)$ ; b) probabilidades para a distribuição  $Pois(5)$ , mostradas somente para valores abaixo de 15.

## 10.4 Distribuição geométrica

Consideremos a seguinte situação. A probabilidade de sucesso em uma cirurgia é 0,20 e é constante de cirurgia para cirurgia, ou seja, não há melhoria dos resultados à medida que mais cirurgias são realizadas. Qual será o número médio de cirurgias que terão que ser realizadas até ocorrer o primeiro sucesso.

Essa situação é parecida com a distribuição binomial, porém com a seguinte diferença: nós não estamos interessados no número de sucessos em um certo número de experimentos (no caso cirurgias realizadas), mas sim em quantos experimentos precisam ser realizados até obtermos o primeiro sucesso. Se chamarmos de  $X$  a variável número de cirurgias até conseguirmos o primeiro sucesso, a distribuição de probabilidades de  $X$  é chamada de **distribuição geométrica**.

A distribuição geométrica se aplica quando se deseja saber o número de experimentos até a obtenção de um desfecho de interesse e quando as seguintes condições são satisfeitas:

1. Cada experimento é independente;
2. Cada experimento possui dois possíveis desfechos (sucesso ou fracasso);
3. A probabilidade de sucesso  $p$  é a mesma em cada experimento.

Representamos uma distribuição geométrica, que depende de um parâmetro  $p$  (probabilidade de sucesso em um experimento) por **Geo(p)**.



### 10.4.1 Probabilidades de uma distribuição geométrica

Dada uma variável aleatória  $X$ , com distribuição  $\text{Geo}(p)$ , como calculamos as probabilidades de serem realizados 1, 2, ...,  $k$  experimentos até a obtenção do primeiro sucesso? A figura 10.25 nos auxilia no estabelecimento de uma fórmula geral para o cálculo dessas probabilidades.

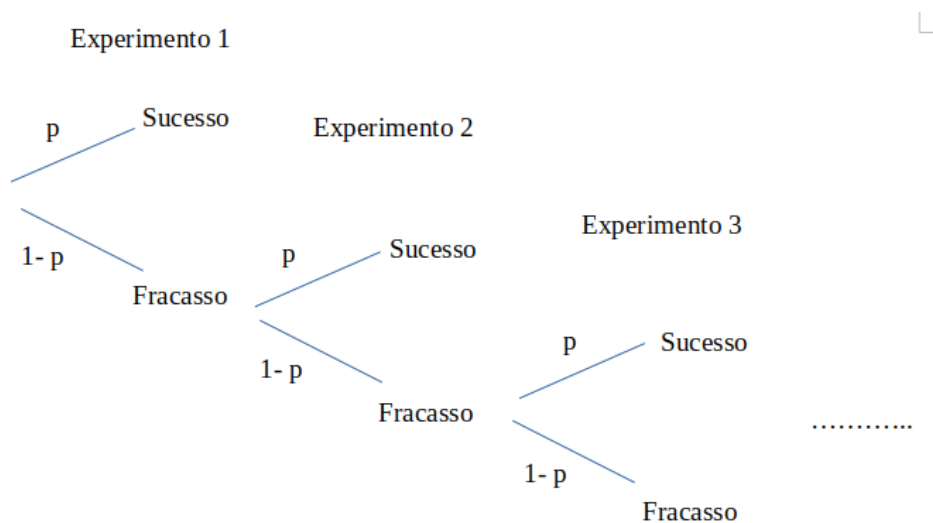


Figura 10.25: Árvore para a obtenção das probabilidades de uma distribuição geométrica com parâmetro  $p$ .

Vamos considerar cada valor da variável aleatória  $X$ . Quando  $X = 1$ , ou seja, sucesso no primeiro experimento, a probabilidade de sucesso no primeiro experimento é  $p$ . Logo:

$$P(X = 1) = p$$

Para termos o primeiro sucesso no segundo experimento, é necessário que o primeiro seja um fracasso e o segundo um sucesso. Então:

$$P(X = 2) = (1 - p)p$$

Para termos um sucesso no terceiro experimento, é necessário que os dois primeiros experimentos sejam um fracasso e o terceiro um sucesso. Então:

$$P(X = 3) = (1 - p)(1 - p)p = (1 - p)^2 p$$

Continuando com esse raciocínio, para o primeiro sucesso ocorrer no  $k$ -ésimo experimento, é necessário que os experimentos de 1 até  $k - 1$  sejam fracassos. Logo:

$$P(X = k) = (1 - p)^{k-1} p \quad (10.10)$$

De modo análogo às distribuições binomial e de Poisson, podemos usar o *R Commander* para obtermos as probabilidades, gráficos, quantis, amostras e probabilidades da cauda de uma distribuição geométrica. A figura 10.26 mostra o gráfico da distribuição de probabilidades da distribuição geométrica com  $p = 0,2$ .

**ATENÇÃO:** observem que o gráfico começa a partir de 0 e que a variável aleatória é o número de fracassos até obter o primeiro sucesso. Assim as probabilidades mostradas no gráfico para  $X = k$  correspondem às probabilidades obtidas da fórmula (10.10) para  $X=k+1$ . Tenha isto sempre em mente quando usar o *R Commander* para trabalhar com a distribuição geométrica.

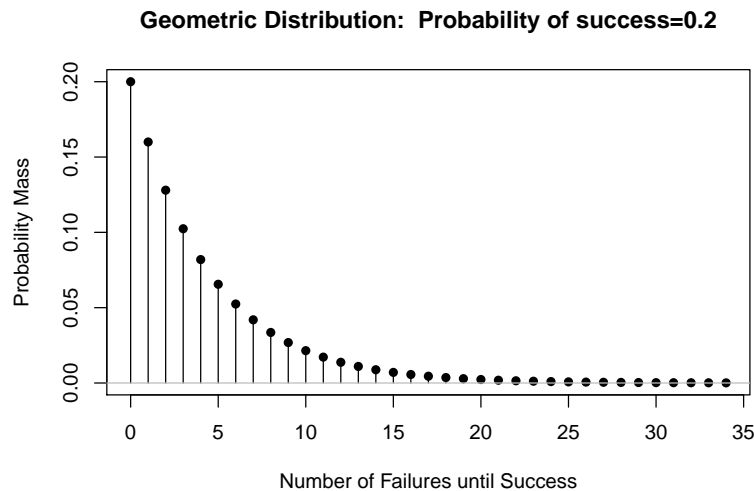


Figura 10.26: Distribuição de probabilidades para a distribuição Geo(0,2).

### 10.4.2 Valor esperado e variância de uma distribuição geométrica

O valor esperado de uma variável aleatória  $X$  com distribuição geométrica com parâmetro  $p$  é obtido a partir da aplicação da expressão de definição do valor esperado.

Pode-se demonstrar (vide (Meyer, 1969)) que o valor esperado é dado por:

$$E[X] = 1/p \quad (10.11)$$

Também não é difícil demonstrar que a variância de uma variável aleatória  $X$  com distribuição geométrica com parâmetro  $p$  é dada por:

$$var(X) = (1 - p)/p^2 \quad (10.12)$$

Voltando ao exemplo inicial, como a probabilidade de sucesso em uma cirurgia é de 0,20, então o valor esperado do número de cirurgias que precisam ser realizadas até se obter o primeiro sucesso é de  $1/0,2 = 5$ .

## 10.5 Exercícios

- 1) Um cirurgião plástico quer comparar o número de enxertos de pele bem sucedidos em sua casuística de pacientes queimados com o número de outros pacientes queimados. Uma pesquisa na literatura indica que aproximadamente 30% dos enxertos infectam, porém que 80% sobrevivem.
  - a. Qual a probabilidade de ocorrer 1 infecção em 8 pacientes?
  - b. Qual a probabilidade de sobrevida de 7 dentre 8 enxertos?
- 2) Estime a probabilidade de um paciente cirúrgico apresentar 5 ou mais hospitalizações nos 10 anos de acompanhamento descritos em um estudo onde, nesse período, 390 pacientes cirúrgicos tiveram um total de 1487 hospitalizações.
- 3) O número de reações gastrointestinais relatadas de um certo medicamento anti-inflamatório foi de 538 por 9.160.000 prescrições. Isso corresponde a uma taxa de 0,58 reações gastrointestinais por 1000 prescrições. Usando o modelo de Poisson, encontre as probabilidades abaixo.
  - a. Exatamente uma reação gastrointestinal em 1000 prescrições;
  - b. Mais de 10 reações gastrointestinais em 1000 prescrições;
  - c. Nenhuma reação gastrointestinal em 1000 prescrições.

# Capítulo 11

## Funções densidade de probabilidades

### 11.1 Introdução

Até agora vimos os conceitos fundamentais da teoria de probabilidades e distribuição de probabilidades, assumindo que temos conhecimento da probabilidade de eventos ou que a variável aleatória é discreta. Vimos no capítulo 1 que, devido às limitações da precisão dos instrumentos de medida, a rigor, todas as variáveis numéricas poderiam ser consideradas discretas na prática. Entretanto é bastante útil realizar uma simplificação aqui e considerar uma situação teórica em que os instrumentos de medida possuem precisão ilimitada. Nesse caso, vamos considerar que variáveis numéricas como tempo, peso, altura, etc, podem assumir qualquer valor real dentro de um dado intervalo. Vamos agora verificar que, para obter probabilidades e distribuições de probabilidades para variáveis aleatórias contínuas, será preciso introduzir os conceitos de densidade de probabilidade e integral de uma função.

### 11.2 Histograma de variáveis contínuas. Recordação

Os conteúdos desta seção e das seções 11.3, 11.4 e 11.5 podem ser visualizados neste [vídeo](#).

Vamos retornar a dois histogramas criados no capítulo 4 para visualizar a distribuição dos valores da variável *igf1* (fator de crescimento semelhante à insulina tipo 1) do conjunto de dados *juul2*. Os histogramas são baseados na tabela 11.1, cópia da tabela 4.2.

O histograma de frequência relativa para a variável *igf1* é mostrado na figura 11.1. Podemos considerar a frequência relativa de cada classe como a probabilidade de obtermos um valor dentro desta classe se escolhermos aleatoriamente um valor de *igf1* dentre os valores possíveis de *igf1* no conjunto de dados.

Tabela 11.1: Definição das classes de um histograma e respectivas frequências, frequências relativas e densidade de frequência relativa para a variável *igf1* do conjunto de dados *juul2*.

Classe	Limite Inferior (>)	Limite Superior (<=)	Frequência	Frequência Relativa (%)	Densidade de Frequência Relativa ( $\times 10^{-3}$ )
1	0	100	43	4,22	0,42
2	100	150	74	7,27	1,45
3	150	200	130	12,77	2,55
4	200	250	129	12,67	2,53
5	250	300	118	11,59	2,32
6	300	350	69	6,78	1,36
7	350	400	94	9,23	1,85
8	400	450	93	9,14	1,82
9	450	500	92	9,04	1,80
10	500	600	98	9,63	0,96
11	600	1000	78	7,66	0,19
			1018	100	

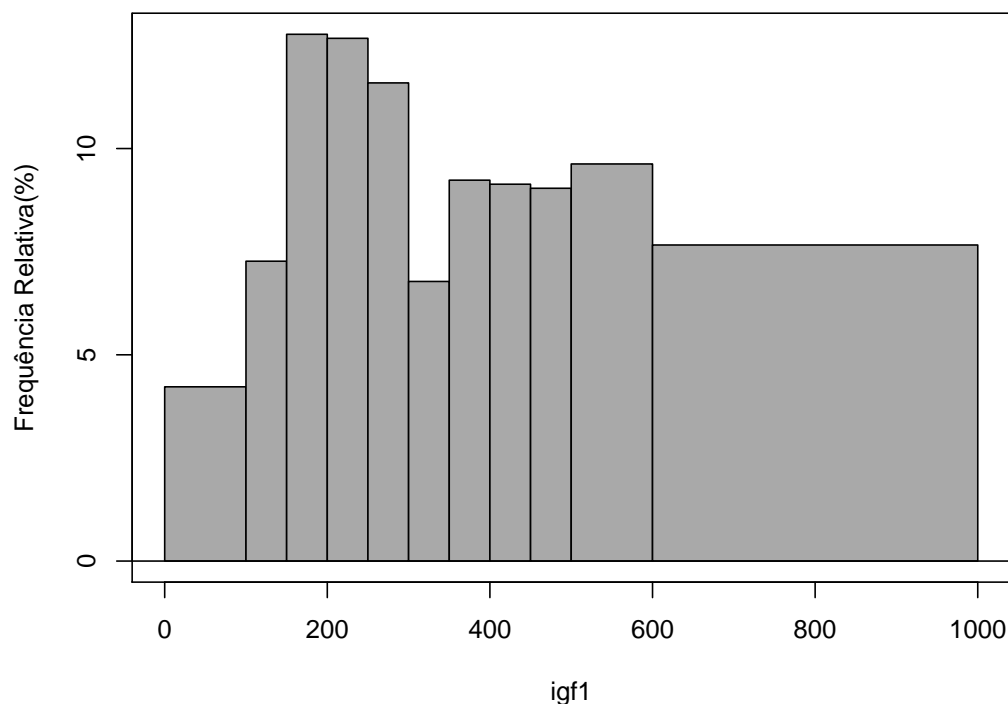


Figura 11.1: Histograma de frequência relativa da variável *igf1* para as classes definidas conforme a tabela 11.1.

Relembrando a discussão apresentada no capítulo 4, observem que a classe 10, que vai de 500 a 600 possui 98 elementos (9,6% do total de valores de *igf1*) e a classe 11, que vai de

600 a 1000 possui 78 elementos (7,7% do total de valores de *igf1*). Apesar de o histograma mostrar que a altura da classe 10 é pouco maior do que a da classe 11, é importante ter em mente que os 98 elementos da classe 10 estão distribuídos em um intervalo de amplitude 100, enquanto que os 78 elementos da classe 11 estão distribuídos em um intervalo de amplitude 400 (1000 - 600), ou seja, a densidade da classe 10 é bem maior do que a da classe 11 e este fato não é mostrado pelo histograma de frequência relativa. Isso pode ser contornado, construindo-se o histograma de tal modo que a área de cada classe (amplitude da classe x altura) seja igual à sua frequência relativa. Para isso, dividimos a frequência relativa de cada classe por sua amplitude, obtendo então a densidade de frequência relativa da classe (coluna 6 da tabela 11.1). O histograma assim construído é denominado **histograma de densidade de frequência relativa** (figura 11.2).

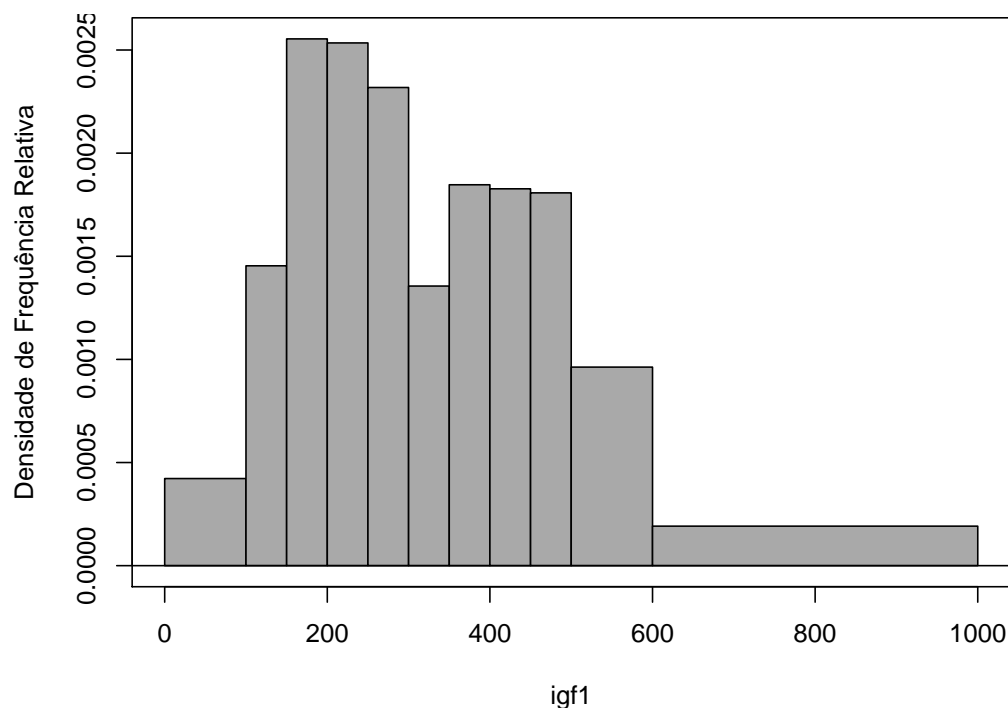


Figura 11.2: Histograma de densidade de frequência relativa da variável *igf1* para as classes definidas conforme a tabela 11.1.

Observem agora que a altura da classe 11 é relativamente bem menor do que a das demais classes, refletindo o fato de que os 78 valores dessa classe estão distribuídos em uma faixa maior de valores do que as demais classes, mas a área dessa classe é igual à sua frequência relativa. Também observem que a área de cada classe sob o histograma de densidade de frequência fornece a frequência relativa daquela classe, ou a probabilidade de obtermos um valor dentro dessa classe se escolhermos aleatoriamente um valor de *igf1* dentre os valores possíveis de *igf1* no conjunto de dados.

Vamos aplicar esse raciocínio para chegarmos às funções densidade de probabilidade e funções de distribuição de probabilidades para variáveis contínuas.

## 11.3 Função densidade de probabilidade

Vamos inicialmente realizar um experimento que você pode reproduzir, acessando a aplicação [Densidade de Frequência e Distribuição Acumulada](#). A tela inicial dessa aplicação é mostrada na figura 11.3. O experimento consiste em considerar uma variável aleatória que é o peso de uma pessoa escolhida aleatoriamente de uma população de pessoas com média 70 kg e desvio padrão de 10 kg. Inicialmente vamos escolher aleatoriamente 20.000.000 pessoas dessa população, de modo que teremos 20.000.000 valores da nossa variável aleatória. Histogramas de frequência relativa, densidade de frequência e distribuição cumulativa da densidade de frequência são então plotados para esses 20.000.000 valores, bem como uma tabela de frequência com as seis classes centrais dos histogramas é construída.

Densidade de Frequência e Distribuição Acumulada

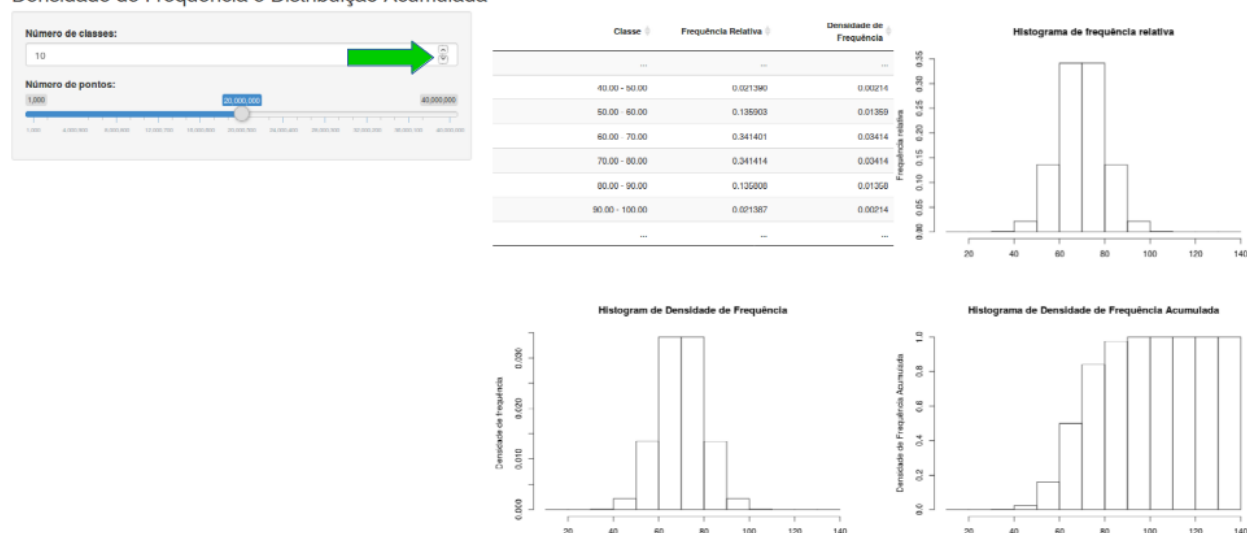
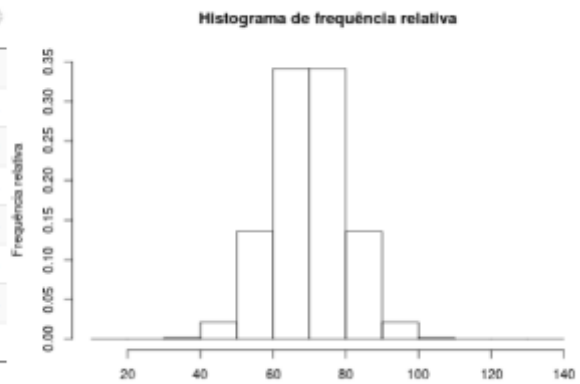


Figura 11.3: Tabelas de frequência, histogramas de frequência relativa, densidade de frequência e distribuição cumulativa de densidade de frequência para uma variável aleatória com média 70 e desvio padrão 10 para um certo número de classes no histograma (10 inicialmente) e um certo número de valores da variável aleatória escolhida aleatoriamente (inicialmente 20.000.000). O número de classes do histograma pode ser alterado ao clicarmos na seta para cima ou para baixo ou digitando o valor desejado na caixa de texto correspondente.

Vamos começar com 10 classes. Vamos variar esse número de classes de 10 para 100, e depois 1000, e ver o que acontece com o histograma da frequência relativa (figura 11.4). Para cada histograma, a amplitude (largura) de cada classe é a mesma.

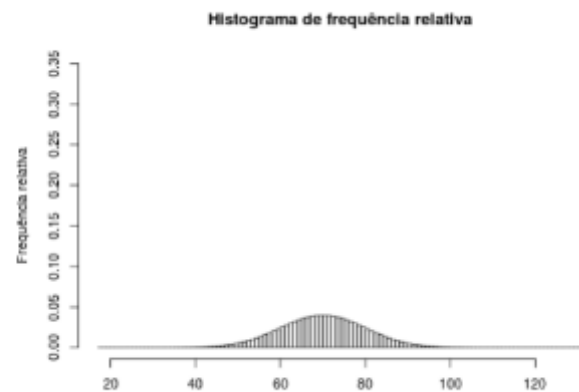
a)  $n = 10$  classes

Classe	Frequência Relativa	Densidade de Frequência
...	...	...
40.00 - 50.00	0.021390	0.00214
50.00 - 60.00	0.135903	0.01359
60.00 - 70.00	0.341401	0.03414
70.00 - 80.00	0.341414	0.03414
80.00 - 90.00	0.135908	0.01359
90.00 - 100.00	0.021387	0.00214
...	...	...



b)  $n = 100$  classes

Classe	Frequência Relativa	Densidade de Frequência	Densidade de Frequência Acumulada
...	...	...	...
66.00 - 67.00	0.037491	0.03749	0.3821
67.00 - 68.00	0.038710	0.03871	0.4208
68.00 - 69.00	0.039439	0.03944	0.4602
69.00 - 70.00	0.039805	0.03981	0.5000
70.00 - 71.00	0.039792	0.03979	0.5398
71.00 - 72.00	0.039441	0.03944	0.5793
...	...	...	...



c)  $n = 1000$  classes

Classe	Frequência Relativa	Densidade de Frequência	Densidade de Frequência Acumulada
...	...	...	...
69.70 - 69.80	0.003985	0.03985	0.4920
69.80 - 69.90	0.003988	0.03997	0.4960
69.90 - 70.00	0.004001	0.04001	0.5000
70.00 - 70.10	0.004006	0.04006	0.5040
70.10 - 70.20	0.003994	0.03995	0.5080
70.20 - 70.30	0.003989	0.03989	0.5120
...	...	...	...

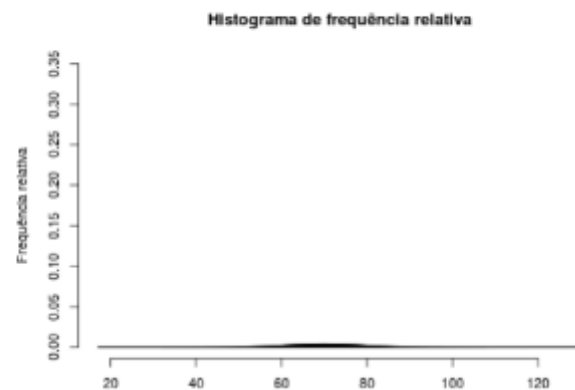


Figura 11.4: Tabela de frequência e histograma de frequência relativa da aplicação da figura 11.3, para diversos valores para o número de classes (a-10, b-100 e c-1000).

Observem que, à medida que o número de classes do histograma aumenta, o histograma se aproxima cada vez mais de uma linha contínua, a amplitude de cada classe diminui, assim como a frequência relativa de cada classe. O valor máximo da frequência relativa cai de  $\sim 0,34$  para 10 classes para menos de 0,004 para 1000 classes.

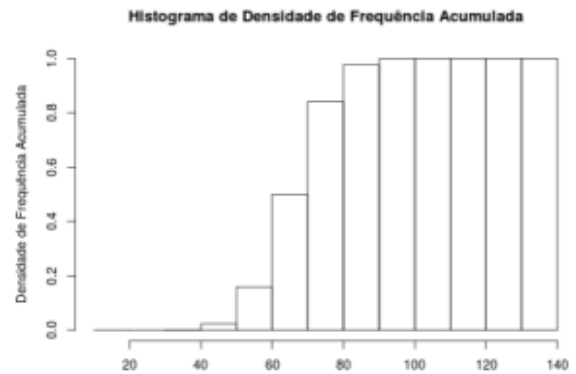
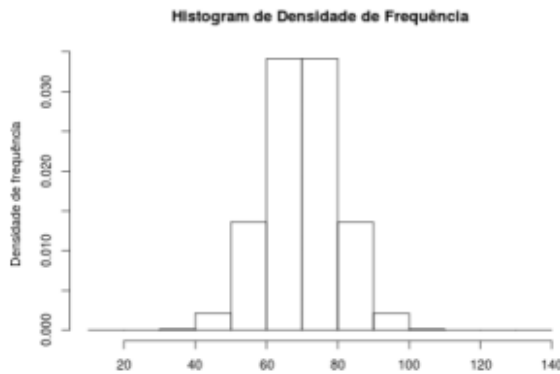


Assim, se fizermos um exercício mental e aumentarmos o número de valores aleatórios extraídos da população indefinidamente e criarmos um histograma de frequência relativa aumentando indefinidamente o número de classes, veremos que a frequência relativa de cada classe irá aproximar-se cada vez mais de zero! Ora, ao aumentarmos indefinidamente o número de classes, a amplitude de cada intervalo irá convergir para zero e cada classe irá ser constituída de um único ponto. **Como a frequência relativa de uma classe é uma estimativa da probabilidade de obtermos um valor dentro da classe, então a probabilidade de obtermos um valor específico dentre os valores possíveis da variável é zero!**

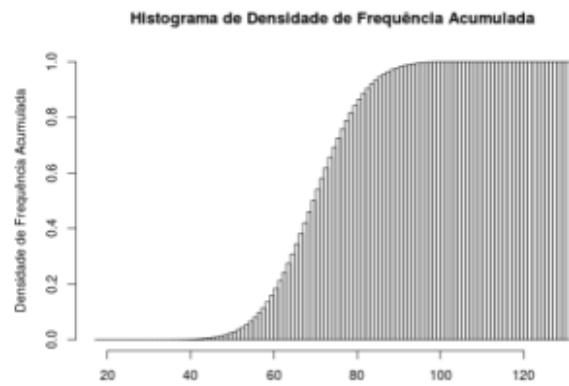
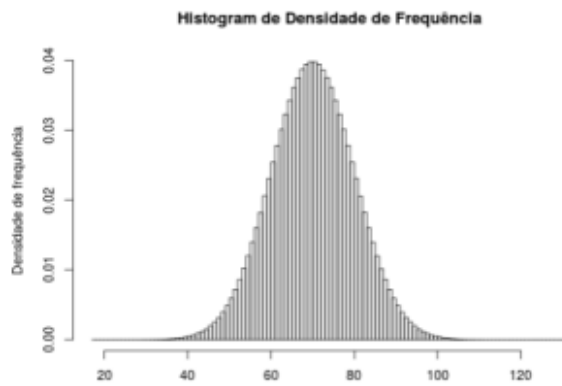
Ao alterarmos o número de classes no histograma, a aplicação também mostra os respectivos histogramas de densidade de frequência e da distribuição cumulativa de densidade de frequência relativa (figura 11.5).

Observem que, analogamente aos histogramas de frequência relativa, os histogramas de densidade e distribuição cumulativa da densidade de frequência se aproximam de uma curva contínua à medida que o número de classes aumenta. Porém, ao contrário dos histogramas de frequência relativa, as alturas das classes no histograma de densidade de frequência não se aproximam de zero, mantendo-se relativamente constantes, à medida que o número de classes aumenta. Lembramos que a densidade de frequência de cada classe é obtida dividindo-se a frequência relativa da classe por sua amplitude. Assim, apesar de a frequência relativa de cada classe diminuir à medida que o número de classes aumenta, a amplitude da classe também diminui proporcionalmente, de modo que a densidade de frequência (divisão da frequência relativa pela amplitude da classe) tende a um valor constante.

a)



b)



c)

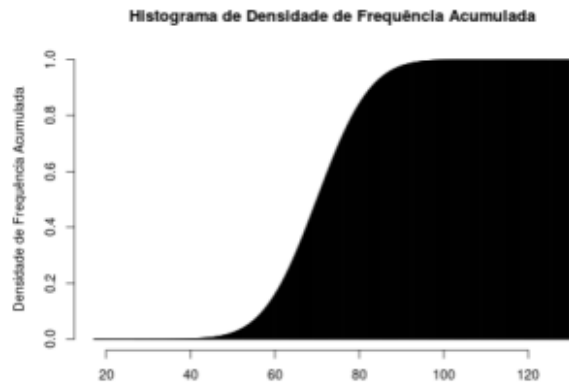
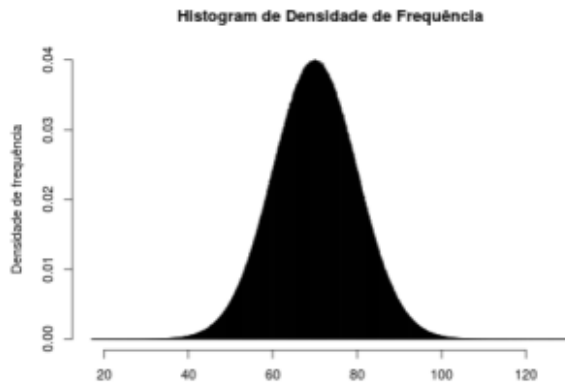


Figura 11.5: Histograma da densidade de frequência relativa e da distribuição cumulativa da densidade de frequência relativa da aplicação da figura 11.3, para diversos valores para o número de classes (a-10, b-100, c-1000).

Da mesma forma que interpretamos a frequência relativa como estimativas de probabilidades, podemos interpretar as densidades de frequências relativas como estimativas das densidades de probabilidades. Portanto não há sentido em termos uma distribuição de probabilidades para variáveis contínuas, já que a probabilidade de ocorrência de cada valor da variável é

zero. Mas podemos falar numa **função densidade de probabilidade**.

Analogamente ao caso de uma variável discreta, a  $P(X \leq x_0)$  para uma variável contínua é igual à probabilidade cumulativa para  $X = x_0$ , correspondendo à probabilidade de se obter um valor menor ou igual a  $x_0$ . À função  $F(x) = P(X \leq x_0)$  para todos os valores de  $x$ , chamamos de **função de distribuição (ou de probabilidade cumulativa)**. Vamos aprofundar um pouco mais a relação entre a função densidade de probabilidade e a função de distribuição, introduzindo o conceito de integral de uma função.

## 11.4 Integral da função densidade de probabilidade

Como obter a função de distribuição a partir da função densidade de probabilidade? Sabemos que o histograma da distribuição cumulativa da densidade de frequência relativa é obtido a partir do histograma da densidade de frequência relativa, somando-se, para cada ponto de uma dada classe, a área dessa classe e as áreas das classes que a precedem. Vamos rodar a aplicação [Integracao](#) (figura 11.6).

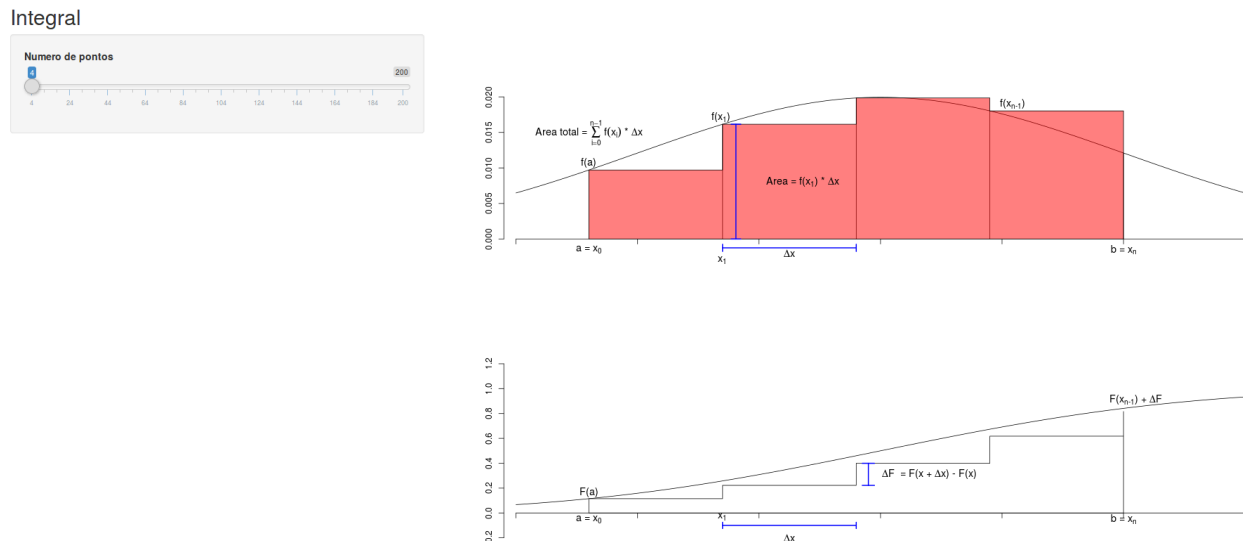


Figura 11.6: Obtenção aproximada da função de distribuição a partir da função densidade de probabilidade, construindo-se histogramas sucessivos com um número crescente de classes. A aplicação é iniciada com um histograma com quatro classes.

Nessa aplicação, vamos considerar que temos uma função densidade de probabilidade  $f(x)$  cuja curva está mostrada no gráfico superior e desejamos saber a probabilidade de obtermos um valor da variável aleatória  $X$  entre  $a$  e  $b$  ( $P(a \leq X \leq b)$ ). A partir da função de distribuição (gráfico inferior), essa probabilidade seria dada por  $F(X = b) - F(X = a)$ . Porém vamos supor que não conhecemos a função de distribuição e temos somente a função densidade de probabilidade. Poderíamos aproximar a área da função densidade de probabilidade entre  $a$  e  $b$  por quatro retângulos, cada um com base igual a  $\Delta x = (b - a)/4$ . Então o valor da função de distribuição em  $b$  seria igual à soma das áreas dos quatro retângulos mais o valor

da função de distribuição em  $a$ . O primeiro retângulo teria área igual a  $f(a) \cdot \Delta x$ , o segundo retângulo área igual a  $f(x_1) \cdot \Delta x$ , e assim por diante, de modo que a soma das áreas dos quatro retângulos poderia ser expressa por:

$$\text{Área} \sim F(X = b) - F(X = a) = \sum_{i=0}^3 f(x_i) \Delta x$$

Podemos observar que essa aproximação é um tanto imprecisa. Podemos melhorá-la, aumentando o número de classes (retângulos). Ao aumentarmos o número de classes para 8, obtemos a figura 11.7.

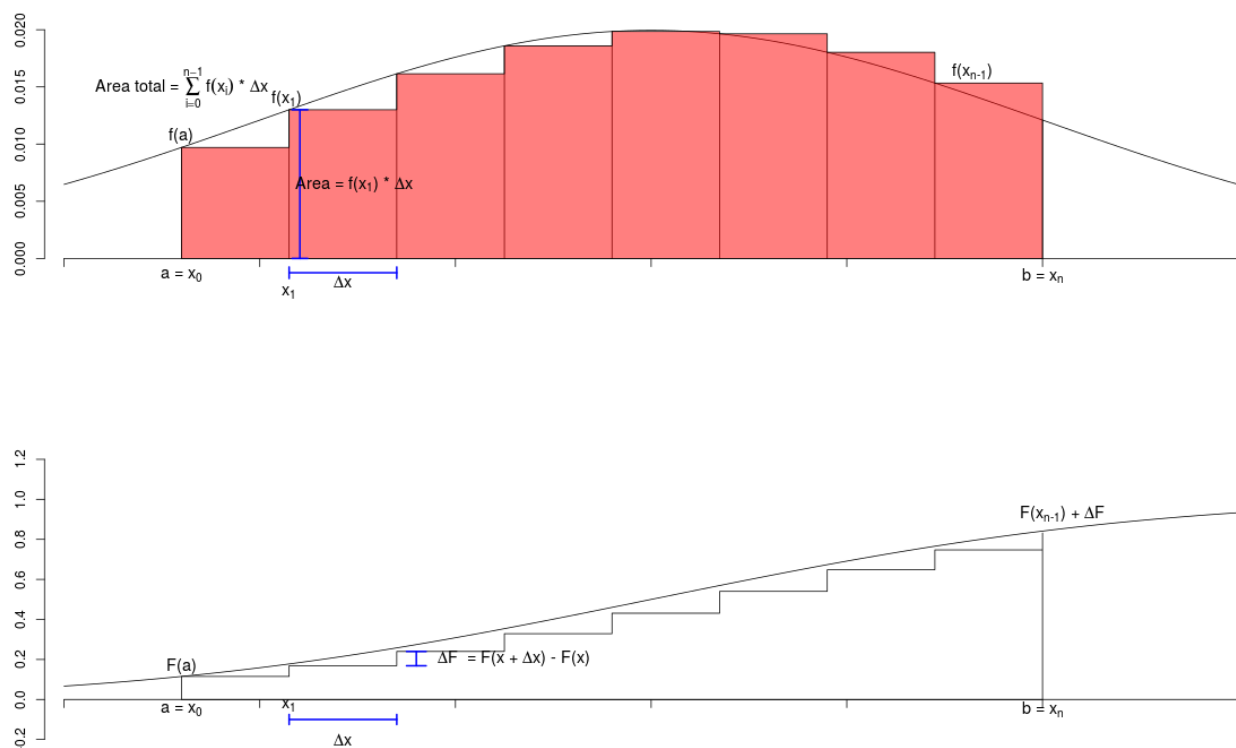


Figura 11.7: Obtenção aproximada da função de distribuição a partir da função densidade de probabilidade, utilizando um histograma com oito classes.

A aproximação é um pouco melhor, e a área dos 8 retângulos seria dada por:

$$\text{Área} \sim F(X = b) - F(X = a) = \sum_{i=0}^7 f(x_i) \Delta x$$

À medida que o número de classes aumenta, a aproximação fica cada vez melhor. Com 200 classes, obtemos a figura 11.8. A aproximação nesse caso já é bastante boa. O limite da soma das áreas dos retângulos obtida quando aumentamos indefinidamente o número de retângulos, sendo a amplitude de cada um deles cada vez menor, é a área sob a curva densidade de probabilidade compreendida entre os pontos  $a$  e  $b$ . Esse limite é a integral da função densidade de probabilidade entre os pontos  $a$  e  $b$  e podemos escrever matematicamente:

$$\text{Área sob } f(x) \text{ entre } a \text{ e } b = F(X = b) - F(X = a) = \int_a^b f(x)dx = \lim_{n \rightarrow \infty, \Delta x \rightarrow 0} \sum_0^n f(x_i)\Delta x \quad (11.1)$$

onde o símbolo  $\int$  é o sinal de integração. O conceito de integração é um conceito fundamental do cálculo e possui inúmeras aplicações, além do cálculo de probabilidades. Por exemplo, dada a função de velocidade de um corpo, a área sob o gráfico da velocidade, ou seja, a integral da velocidade, fornece a distância percorrida. Em um outro exemplo, dada a densidade linear de uma barra, a integral da densidade linear fornece a massa da barra.

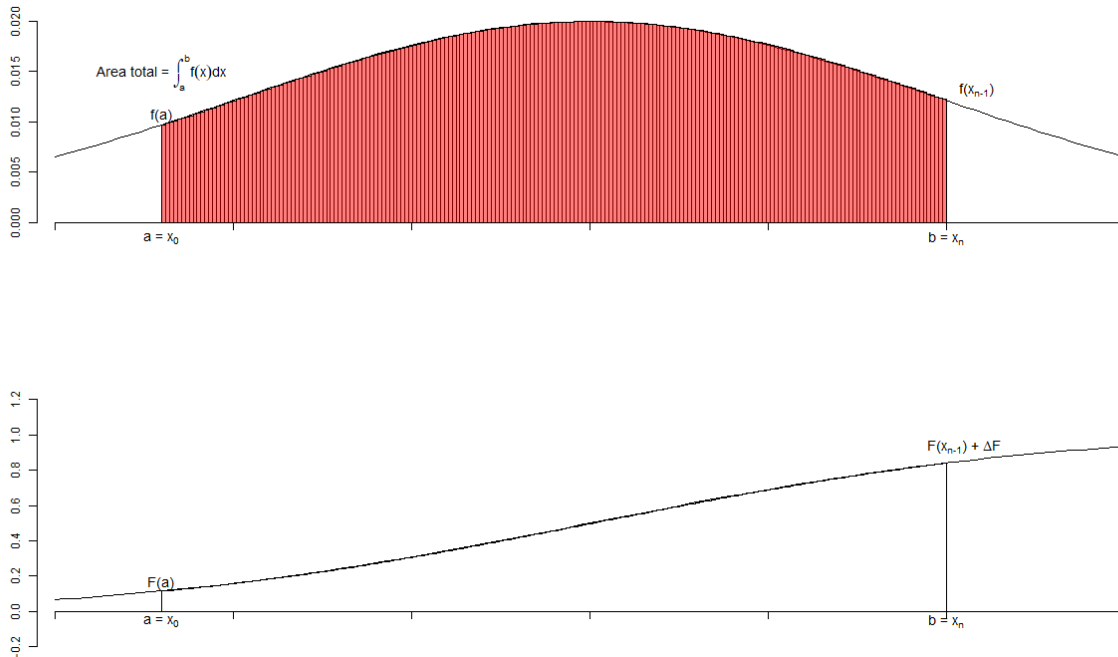


Figura 11.8: Obtenção aproximada da função de distribuição a partir da função densidade de probabilidade, utilizando um histograma com 200 classes.

## 11.5 Propriedades da função densidade de probabilidade

As propriedades descritas no capítulo 7 sobre probabilidades, bem como o teorema de Bayes se aplicam diretamente para funções densidade de probabilidades, lembrando apenas que agora estamos trabalhando com funções de densidade e não mais com valores de probabilidades específicas. Em particular, sendo  $f(x)$  uma função densidade de probabilidade, teremos:

(P1):  $f(x)$  é uma função sempre positiva;

(P2): Para o evento impossível,  $f(x) = 0$ ;

(P3): Para o espaço amostral  $S$  que corresponde ao evento certo,  $\int_{-\infty}^{\infty} f(x)dx = 1$ , ou seja, a área total sob o gráfico da função densidade de probabilidade é sempre igual a 1;

(P4): Se um evento  $X$  for dividido em eventos disjuntos, ou mutuamente exclusivos,  $X_1, X_2, \dots, X_n$  (ou seja, se  $X_i$ ,  $i = 1, 2, \dots, n$ , ocorre, os demais não ocorrerão), então a probabilidade do evento  $X$  será a soma das probabilidades de cada um dos eventos  $X_i$ . Matematicamente, isso será expresso por  $f(x) = \sum_1^n f(X_i)$ . Levem em conta que cada evento  $X_i$  aqui pode se referir a um intervalo do conjunto dos números reais;

(P5): Se  $X$  e  $Y$  são dois eventos, então  $f(x,y) = f(x)f(y|x)$ , onde  $f(y|x)$  é a função densidade de probabilidade condicional

Para o teorema de Bayes, como  $f(x,y) = f(x).f(y|x) = f(y).f(x|y)$ , obtemos:

$$f(x|y) = \frac{f(x).f(y|x)}{f(y)} \quad (11.2)$$

Para funções densidade de probabilidades, estamos trabalhando com funções contínuas, necessitando em muitos casos de resolver integrais para calcular probabilidades.

O valor esperado de uma variável aleatória  $X$  cuja função densidade de probabilidade é dada por  $f(x)$  é definido pela expressão:

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx \quad (11.3)$$

e a variância é definida por :

$$var(X) = \int_{-\infty}^{\infty} (x - E[X])^2 f(x)dx = E[X^2] - (E[X])^2 \quad (11.4)$$

A seguir, serão apresentadas algumas das principais funções densidade de probabilidade, a saber: distribuição uniforme, distribuição exponencial e distribuição normal.

## 11.6 Distribuição uniforme

Uma variável aleatória contínua  $X$  possui uma **distribuição uniforme** quando a sua função densidade de probabilidade possui um valor constante em um intervalo  $[a, b]$  e zero fora desse intervalo. Logo, se  $f(x)$  é a função densidade da variável aleatória  $X$ , então:

$$f(X) = \begin{cases} \frac{1}{b-a}, & \text{se } a \leq X \leq b \\ 0, & \text{caso contrário} \end{cases} \quad (11.5)$$

O valor constante é igual ao inverso da amplitude do intervalo  $[a, b]$ , porque a área sob o retângulo como base igual a  $(b - a)$  tem que ser igual a 1. Lembremos que a área sob o gráfico da função densidade de probabilidade é igual a 1.

Por meio do *R Commander*, podemos visualizar o gráfico de um conjunto de funções de distribuição contínuas, de modo análogo ao caso de distribuições discretas. A figura 11.9 mostra como acessar a caixa de diálogo para obter o gráfico da distribuição uniforme.

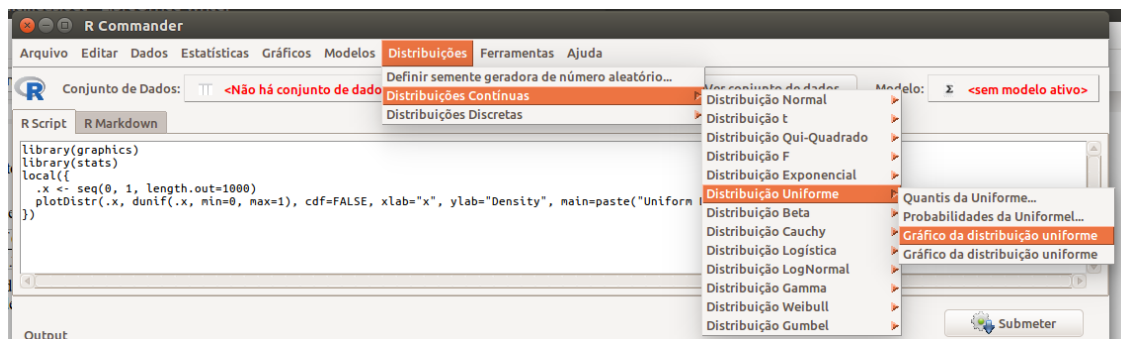


Figura 11.9: Interface do *R Commander* com os menus de acesso para configurar o gráfico de uma distribuição uniforme.

A caixa de diálogo para estabelecer os parâmetros da distribuição uniforme é mostrada na figura 11.10. Os valores mínimo e máximo correspondem aos limites do intervalo  $[a, b]$ . Nesse exemplo, os valores para  $a$  e  $b$  são respectivamente 0 e 1. O gráfico da função de densidade de probabilidade é mostrado na figura 11.11a e o da função de distribuição de probabilidade cumulativa na figura 11.11b.

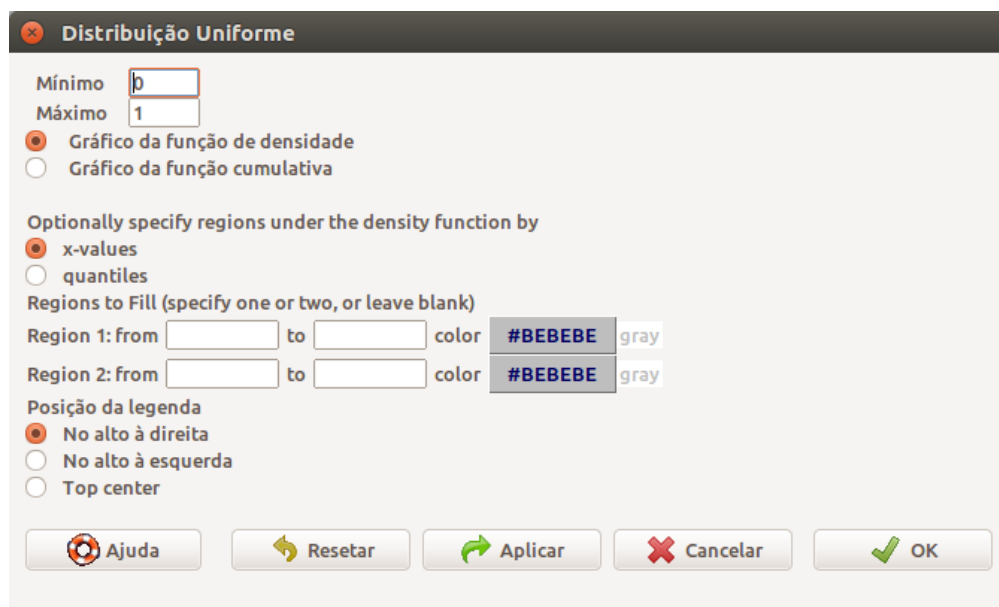


Figura 11.10: Caixa de diálogo do *R Commander* para configurar os parâmetros e gerar o gráfico de uma distribuição uniforme.

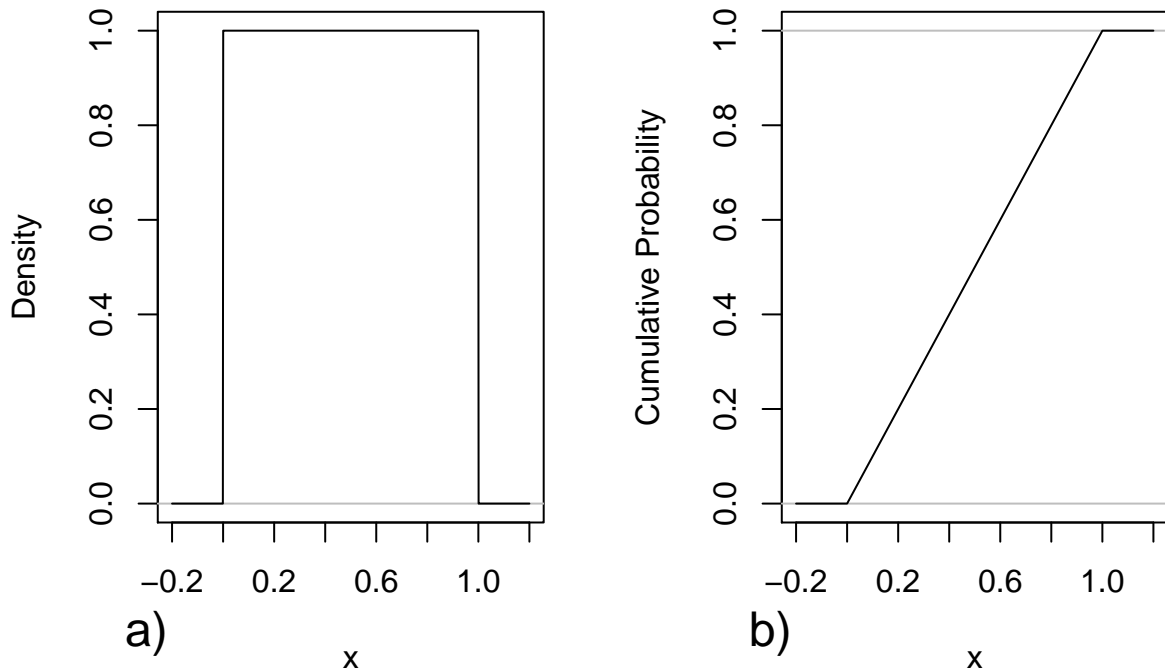


Figura 11.11: Gráfico da função de densidade de probabilidade (a) e o da função de distribuição de probabilidade cumulativa (b) para a distribuição uniforme com mínimo = 0 e máximo = 1.

**Exemplo 1:** dada uma distribuição uniforme com valores mínimo igual a 50 e máximo igual a 70, calcule a probabilidade de se obter um valor entre 50 e 55.

A função densidade de probabilidade dessa distribuição uniforme é dada por:

$$f(X) = \begin{cases} \frac{1}{20}, & \text{se } 50 \leq X \leq 70 \\ 0, & \text{caso contrário} \end{cases}$$

Então a probabilidade de se obter um valor entre 50 e 55 é igual à área sob o gráfico da distribuição uniforme para os valores de  $X$  situados entre 50 e 55. Essa área é representada pela área hachurada da figura 11.12. Essa figura é obtida a partir da caixa de diálogo para obter o gráfico da distribuição uniforme (figura 11.10), estabelecendo os valores mínimo e máximo iguais a 50 e 70, respectivamente, e especificando os valores da *Region1 from 50 to 55*.

Logo:

$$P(50 \leq X \leq 55) = \frac{1}{20}(55 - 50) = 0,25$$



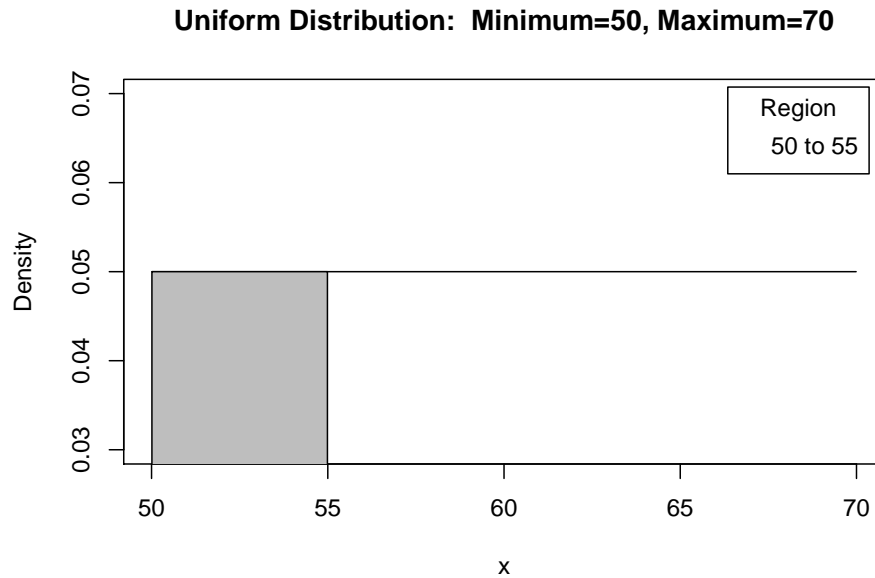


Figura 11.12: Gráfico da função de densidade de probabilidade para a distribuição uniforme com mínimo = 50 e máximo = 70. A área hachurada representa a probabilidade de se obter aleatoriamente um valor entre 50 e 55.

A probabilidade do exemplo 1 poderia também ser obtida a partir do *R Commander* por meio do item de menu *Probabilidades da Uniforme* (vide figura 11.9). Ao selecionarmos essa opção, a caixa de diálogo da figura 11.13 será exibida.



Figura 11.13: Caixa de diálogo para obtermos a probabilidade de selecionarmos aleatoriamente um valor entre 50 e 55 para a distribuição uniforme com mínimo = 50 e máximo = 70.

Fixando os valores mínimo e máximo, digitando 55 em *Valores da Variável*, selecionando a cauda inferior e pressionando o botão OK, obtemos a probabilidade desejada.

O valor esperado de uma distribuição uniforme com valores mínimo =  $a$  e máximo =  $b$  é:

$$E[X] = \frac{a + b}{2} \quad (11.6)$$

ou seja, o ponto médio entre  $a$  e  $b$ .

A variância de uma distribuição uniforme com valores mínimo =  $a$  e máximo =  $b$  é:

$$\text{var}(X) = \frac{(b - a)^2}{12} \quad (11.7)$$

obtida por meio da resolução da integral  $\int_a^b \left(x - \frac{a+b}{2}\right)^2 \frac{1}{b-a} dx$

## 11.7 Distribuição normal ou gaussiana

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

A função densidade de probabilidade **normal** é a mais importante das distribuições contínuas, uma vez que possui vasta aplicação em modelagem e inferência estatística. Essa função tem uma forma de sino e sua equação é:

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2} \quad (11.8)$$

Ela é abreviadamente representada como  $N(\mu, \sigma^2)$ , onde os parâmetros  $\mu$  e  $\sigma$  são respectivamente o valor esperado, ou a média da distribuição, e o desvio padrão. A figura 11.14 mostra o gráfico da função densidade de probabilidade  $N(0,1)$ , que é conhecida como distribuição normal padronizada ou distribuição normal padrão.

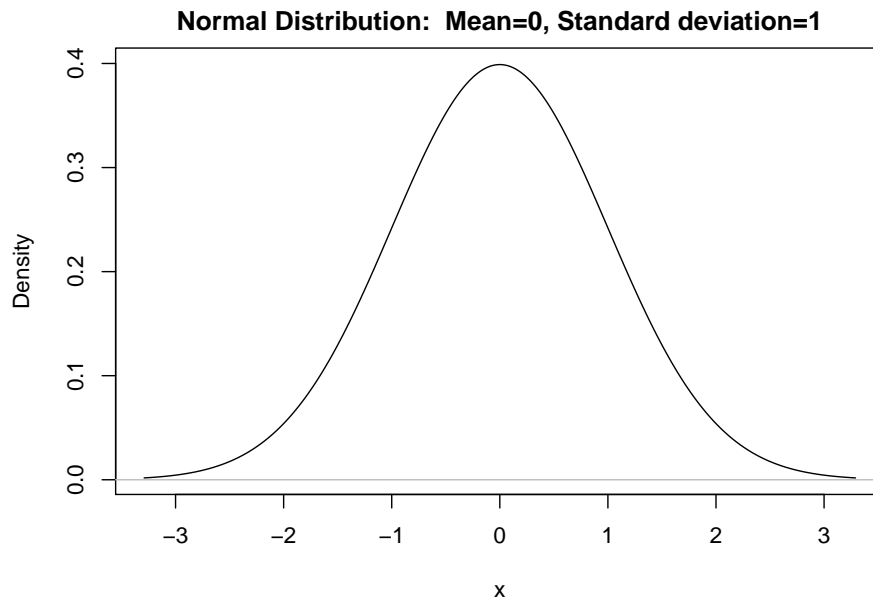


Figura 11.14: Gráfico da função densidade de probabilidade da distribuição normal padrão.

A figura 11.15 mostra os gráficos das funções densidade de probabilidade de duas distribuições normais,  $N(170, 16)$  e  $N(190, 16)$ , que diferem somente no valor da média. A figura 11.16 mostra os gráficos das funções densidade de probabilidade de duas distribuições normais,  $N(170, 4)$  e  $N(170, 16)$ , que diferem somente no valor do desvio padrão. Como é de se esperar, a média indica a localização da função densidade e o desvio padrão indica o quanto o gráfico se concentra em torno da média.

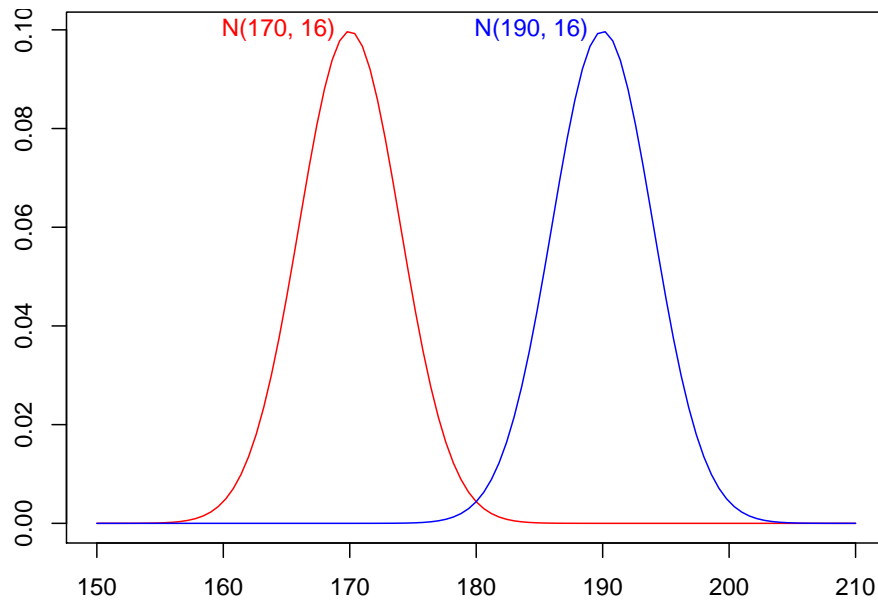


Figura 11.15: Gráfico da função densidade de probabilidade de duas distribuições normais que diferenciam somente no valor da média.

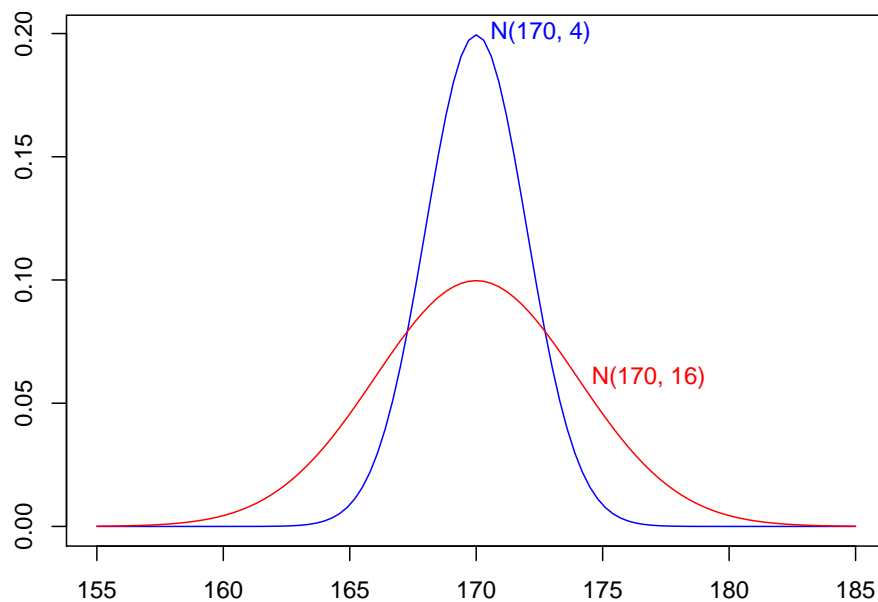


Figura 11.16: Gráfico da função densidade de probabilidade de duas distribuições normais que diferenciam somente no valor do desvio padrão.

Uma variável aleatória  $X$  com distribuição normal  $N(\mu, \sigma^2)$  pode ser convertida à normal padronizada  $Z$ , com distribuição  $N(0,1)$ , usando a expressão:

$$Z = \frac{X - \mu}{\sigma} \quad (11.9)$$

A transformação para a variável  $Z$  padronizada produz um número que expressa quantos desvios padrões o valor da variável original  $X$  está distante da média, ou seja, obtemos uma medida padronizada da distância do valor de  $X$  à média. Por exemplo,  $Z = 1$ , indica que o dado está a 1 desvio padrão da média,  $Z = 1,5$  indica que o valor de  $X$  dista exatamente 1,5 desvios padrões da média e assim por diante.

A função normal não possui uma integral analítica, ou seja, se quisermos obter a probabilidade de  $X \leq x_0$  para uma  $N(\mu, \sigma^2)$ , teríamos que resolver a integral:

$$f(x_0) = P(X \leq x_0) = \int_{-\infty}^{x_0} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

por métodos numéricos. Felizmente, podemos recorrer ao uso de tabelas onde a integral da  $N(0,1)$  para diferentes valores encontra-se tabulada ou, então, recorrer a um programa de computador como o R. O uso de tabelas requer alguns cuidados, fundamentalmente porque não há uma padronização de como os valores são tabelados. O R fornece, por meio da função *pnorm*, a área sob a curva da função densidade de probabilidade de uma variável aleatória com distribuição normal qualquer de  $-\infty$  até o valor  $x_0$  de interesse especificando o parâmetro *lower.tail = TRUE*, ou de  $x_0$  até  $+\infty$  especificando o parâmetro *lower.tail = FALSE*.

**Exemplo 2:** Uma população de homens tem altura com distribuição normal,  $\mu = 170$  cm e  $\sigma = 4$  cm, isto é,  $N(170, 16)$ . Qual a probabilidade de que um indivíduo selecionado aleatoriamente dessa população tenha a altura entre 165 cm a 174 cm?

Estamos interessados em calcular  $P(165 \leq X \leq 174)$ . Para calcularmos essa probabilidade, teríamos que calcular a área sob a função densidade de probabilidade da  $N(170, 16)$ , situada entre 165 e 174. Para obtermos o gráfico da distribuição normal no *R Commander*, selecionamos a opção:

Distribuições  $\Rightarrow$  Distribuições Contínuas  $\Rightarrow$  Dist. Normal  $\Rightarrow$  Gráfico dist. normal

A figura 11.17 mostra como configurar os parâmetros da distribuição normal e selecionar a região que será preenchida (entre 165 e 174) e a figura 11.18 mostra o gráfico da função densidade de probabilidade e a área entre 165 e 174 cm.

**Distribuição Normal**

Média:

Desvio padrão:

☒ Gráfico da função de densidade  
☐ Gráfico da função cumulativa

Optionally specify regions under the density function by

☒ x-values  
☐ quantiles

Regions to Fill (specify one or two, or leave blank)

Region 1: from  to  color

Region 2: from  to  color

Posição da legenda

☒ No alto à direita  
☐ No alto à esquerda  
☐ Top center

Figura 11.17: Caixa de diálogo para gerar um gráfico de uma distribuição normal. Especificando os limites de uma região, fará com que a área sob o gráfico entre esses limites seja preenchida.

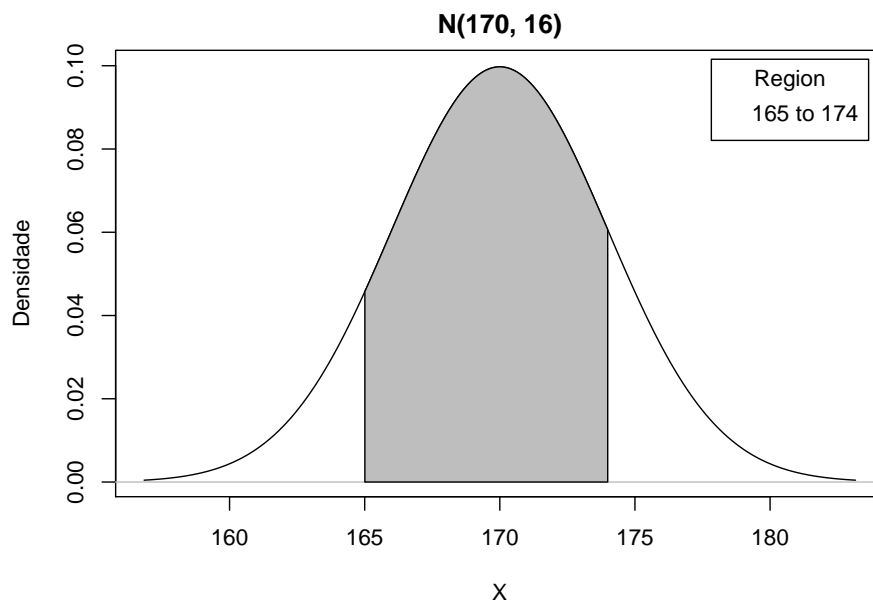


Figura 11.18: A área preenchida representa a probabilidade de se selecionar aleatoriamente um homem da população com altura entre 165 e 174 cm.

A área preenchida na figura 11.18, ou seja, a probabilidade de se selecionar aleatoriamente um indivíduo com altura entre 165 e 174, pode ser calculada por meio da área preenchida do gráfico à direita na figura 11.19 menos a área preenchida do gráfico à esquerda

$$P(165 \leq X \leq 174) = P(X \leq 174) - P(X \leq 165)$$

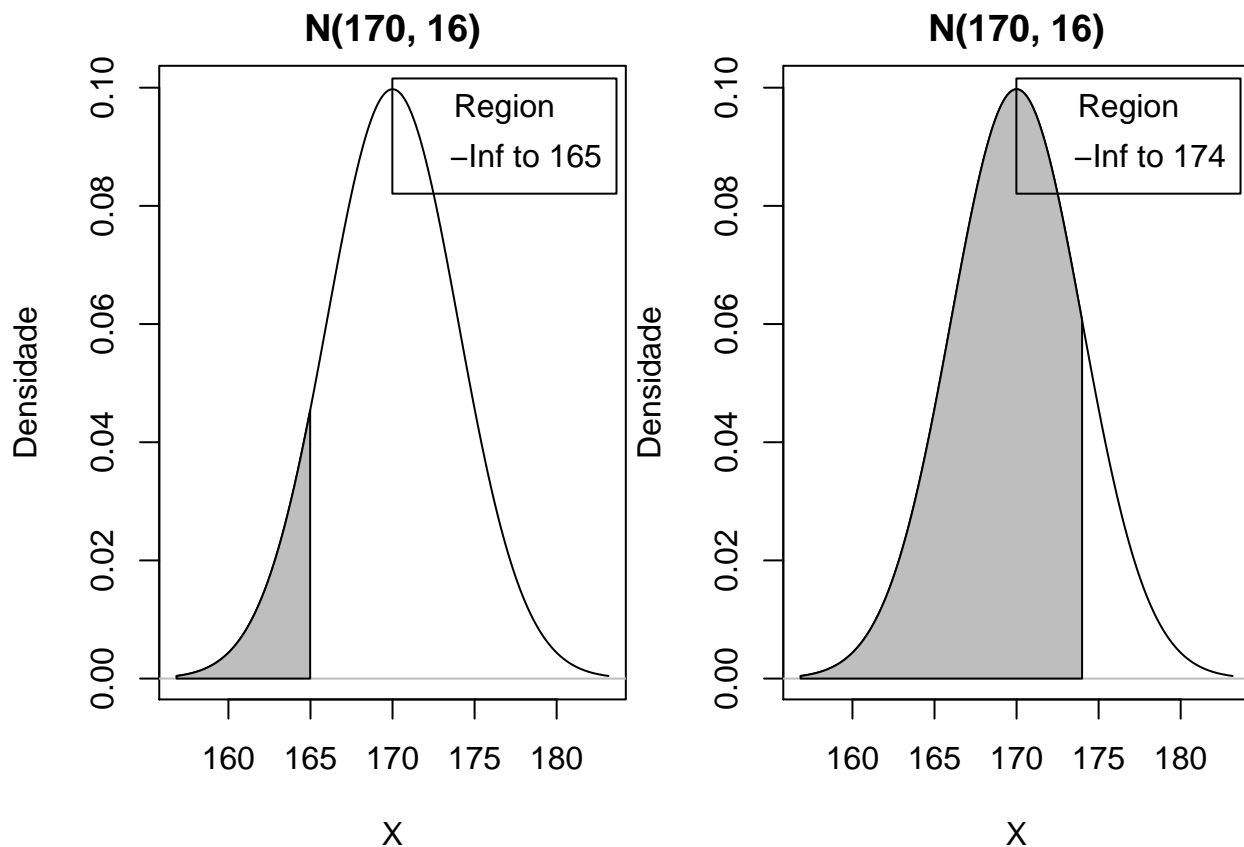


Figura 11.19: A área preenchida na figura 11.18 pode ser calculada por meio da área preenchida do gráfico à direita menos a área preenchida do gráfico à esquerda

As duas probabilidades  $P(X \leq 174)$  e  $P(X \leq 165)$  podem ser calculadas no *R Commander* por meio da opção:

Distribuições  $\Rightarrow$  Distribuições Contínuas  $\Rightarrow$  Dist. Normal  $\Rightarrow$  Probabilidades da Normal...

A figura 11.20 mostra como calcular a probabilidade  $P(X \leq 174)$ .



Figura 11.20: Caixa de diálogo para calcularmos a probabilidade  $P(X \leq 174)$ .

A função para calcular a probabilidade  $P(X \leq 174)$  é dada a seguir:

```
pnorm(c(174), mean=170, sd=4, lower.tail=TRUE)
```

```
## [1] 0.8413447
```

Analogamente, obtemos  $P(X \leq 165)$ :

```
pnorm(c(165), mean=170, sd=4, lower.tail=TRUE)
```

```
## [1] 0.1056498
```

O comando abaixo calcula  $P(165 \leq X \leq 174) = P(X \leq 174) - P(X \leq 165)$ :

```
pnorm(174, mean=170, sd=4, lower.tail=TRUE) -  
  pnorm(165, mean=170, sd=4, lower.tail=TRUE)
```

```
## [1] 0.735695
```

### 11.7.1 Valores importantes da variável Z padronizada

(Z1): Regra 68-95

A probabilidade de extrairmos um indivíduo da população e o valor observado da variável estar dentro do intervalo de um desvio padrão abaixo ou acima da média (isto é, entre -1 e 1 para a distribuição normal padrão) é 68% (0,6826). Similarmente, a probabilidade de extrairmos um indivíduo da população e o valor observado da variável estar dentro do intervalo de dois desvios padrões abaixo ou acima da média é 95% (0,9544).

(Z2): Quartis

Para uma curva normal padronizada (Z), o primeiro quartil é -0,67 e o terceiro quartil é 0,67. Ou seja, para uma variável aleatória com distribuição normal, é de 50% a probabilidade de extrairmos um indivíduo da população e o valor observado da variável estar dentro de 2/3 de desvios padrões em torno da média.

(Z3): Valores extremos (*outliers*)

Em um diagrama de *boxplot* no R, por padrão, *outliers* são valores que estão a mais de 1,5 vezes o intervalo ou faixa inter-quartilica (IQR), antes do Q1 (primeiro quartil) ou após o Q3 (terceiro quartil). Como conhecemos Q1 e Q3 para a normal, o  $IQR = 0,67 - (-0,67) = 1,34$

Usando o valor acima, temos que valores extremos são aqueles menores do que -2,68  $(-0,67 - 1,5 \times 1,34)$  ou maiores do que 2,68  $(0,67 + 1,5 \times 1,34)$  desvios padrões. A probabilidade de se observar valores extremos é, portanto:

$$P(z < -2,68) + P(z > 2,68) = 0,0037 + 0,0037 = 0,0074,$$

isto é, menor que 1%.

No R, para obtermos  $P(X < -x_0 \text{ ou } X > x_0)$ , com  $x_0 = 2,68$ , usamos a seguinte função:

```
pnorm(-2.68) + pnorm(2.68, lower.tail = FALSE)
```

```
## [1] 0.007362216
```

A função distribuição normal, que também é conhecida como gaussiana, tem sua importância em primeiro lugar, porque diversas medições físicas são aproximadas muito bem por essa distribuição. Por exemplo, se o erro de medição de uma quantidade desconhecida é o resultado da soma de diversos pequenos erros que podem ser positivos ou negativos e que ocorrem aleatoriamente, então a distribuição normal pode ser utilizada para modelar esse erro.

Outro aspecto é que diversas variáveis aleatórias que não são normalmente distribuídas podem ser transformadas, por exemplo tomando o logaritmo ou a raiz quadrada da variável de interesse, e o resultado da transformação pode ser aproximadamente normalmente distribuído.

### 11.7.2 Aproximação da distribuição binomial pela normal

Em muitas situações, podemos aproximar a distribuição binomial, que é uma distribuição discreta pela distribuição normal. Essa aproximação é adequada quando, no modelo  $B(n, p)$ ,  $n$  é grande e  $p$  não está muito próximo de 0 ou 1. Por exemplo, para  $n > 20$  e  $0,3 < p < 0,7$ , temos uma boa aproximação. Outros autores consideram que a aproximação é boa para  $np > 5$  e  $n(1-p) > 5$ . A vantagem de usarmos essa aproximação é que, para problemas com  $n$  grande, o cálculo das probabilidades da binomial é trabalhoso, caso sejam feitos manualmente. Atualmente, com o uso de computadores, essa dificuldade foi minimizada mas, mesmo assim, em diversas situações, a aproximação é útil.

Para usarmos a aproximação normal para a binomial, fazemos  $\mu = np$  e  $\sigma^2 = np(1-p)$ , que são justamente a média e a variância da binomial. A figura 11.21 ilustra a aplicação [Binomial x Normal](#), que mostra um gráfico da distribuição normal sobreposto ao gráfico da binomial para diferentes combinações de  $n$  e  $p$ . Observem que, para  $n = 12$  e  $p = 0,1$ , a aproximação não é boa. Já para  $n = 10$  e  $p = 0,5$  (figura 11.22), a aproximação parece bastante razoável. Vamos utilizar essa aproximação para calcularmos probabilidades da binomial, usando a distribuição normal.



## Binomial x Normal

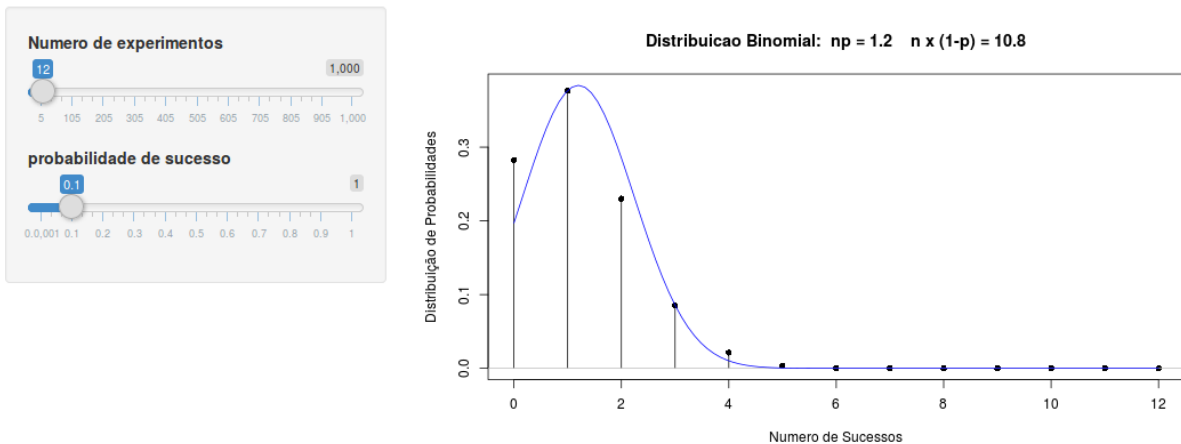


Figura 11.21: Aplicação para verificar a relação entre a distribuição normal e a distribuição binomial.

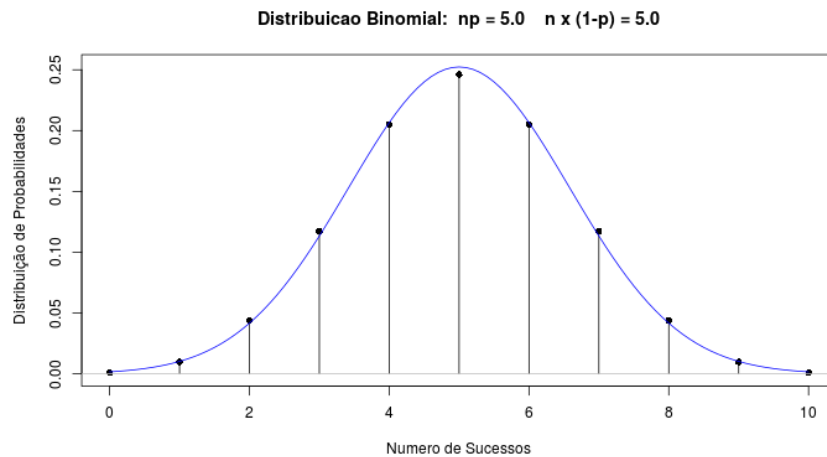


Figura 11.22: Gráficos das distribuições  $N(5, 2,5)$  e  $B(10, 0,5)$ . A aproximação usando a normal parece bastante razoável.

Vamos calcular a probabilidade de obtermos um número de sucessos entre 5 e 8 ( $P(5 \leq X \leq 8)$ ) para a distribuição binomial  $B(12; 0,5)$ . Usando o *R Commander*, conforme visto no capítulo anterior, podemos calcular essa probabilidade, usando o comando a seguir:

```
pbinom(8, size=12, prob=0.5, lower.tail=TRUE) -  
  pbinom(4, size=12, prob=0.5, lower.tail=TRUE)
```

```
## [1] 0.7331543
```

Usando uma aproximação normal para essa distribuição binomial,  $N(6, 3)$ , poderíamos pensar em calcular a área sob a função densidade da normal, compreendida entre 5 e 8 (figura 11.23).

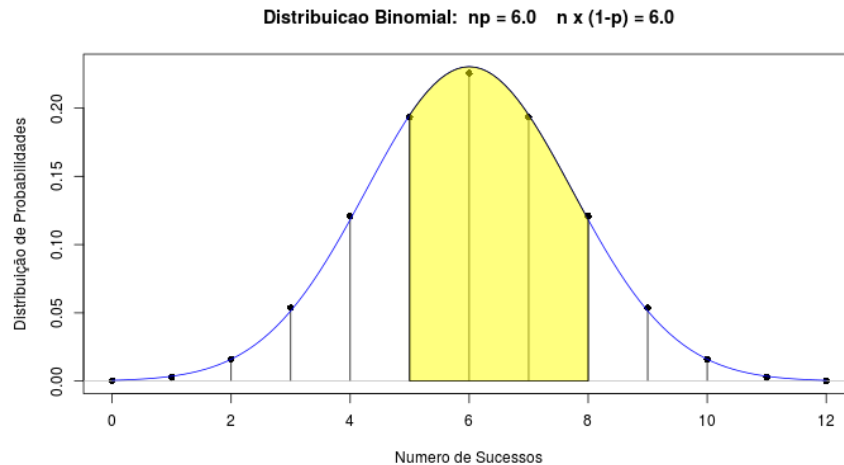


Figura 11.23: Gráficos das distribuições  $N(6, 3)$  e  $B(12, 0,5)$ . A área em amarelo representa a área sob o gráfico da distribuição normal entre 5 e 8.

A área em amarelo na figura 11.23 pode ser calculada por meio da expressão abaixo:

```
pnorm(8, mean=6, sd=1.732, lower.tail=TRUE) -  
pnorm(5, mean=6, sd=1.732, lower.tail=TRUE)
```

```
## [1] 0.5940547
```

O resultado  $P(5 \leq X \leq 8) = 0,594$  é bastante diferente do valor real (0,733), obtido anteriormente. Devemos ter em mente, porém, que a função densidade de probabilidade da distribuição normal é uma função contínua e, para obtermos as probabilidades da binomial, devemos realizar uma correção de continuidade. Para esse exemplo, essa correção implica em calcularmos a área entre 4,5 e 8,5 (figura 11.24).

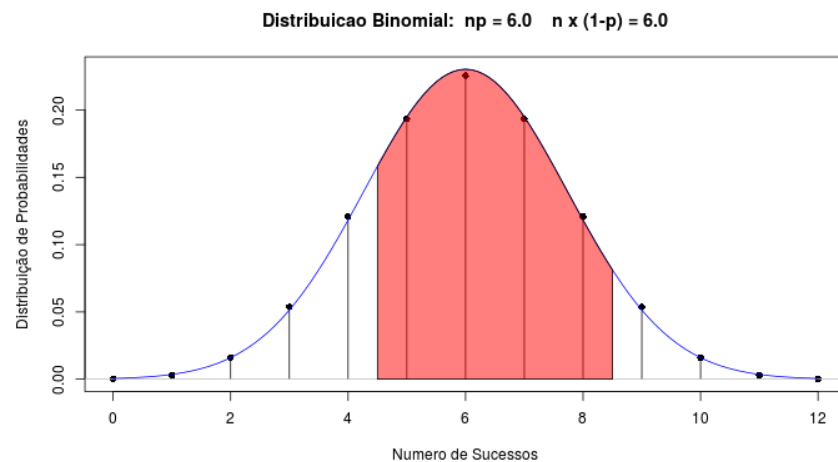


Figura 11.24: Gráficos das distribuições  $N(6, 3)$  e  $B(12, 0,5)$ . A área em vermelho representa a área sob o gráfico da distribuição normal entre 4,5 e 8,5.

Para calcular  $P(4,5 \leq X \leq 8,5)$ , usamos o comando a seguir:

```
pnorm(8.5, mean=6, sd=1.732, lower.tail=TRUE) -  
  pnorm(4.5, mean=6, sd=1.732, lower.tail=TRUE)
```

```
## [1] 0.7323175
```

O resultado,  $P(4,5 \leq X \leq 8,5) = 0,732$ , é bastante próximo do valor real.

Assim, ao aproximarmos a distribuição binomial pela normal, devemos aplicar correções de continuidade. Que tipo de correção você aplicaria para calcular  $P(X \leq 5)$  para a distribuição  $B(12; 0,5)$ ? E para calcular  $P(X \geq 8)$ ? E para calcular  $P(X < 5)$ ? E para calcular  $P(X > 8)$ ? E para calcular  $P(X = 8)$ ?

### 11.7.3 Aproximação da distribuição de Poisson pela normal

A distribuição de Poisson também pode ser aproximada pela distribuição normal para valores suficientemente grandes do parâmetro  $\lambda$ . A figura 11.25 ilustra a aplicação [Poisson x Normal](#), que mostra um gráfico da distribuição normal sobreposto ao gráfico da distribuição de Poisson para diferentes valores de  $\lambda$ . A média e a variância da distribuição normal são iguais a  $\lambda$ . Observem que, para  $\lambda = 2$ , a aproximação não é boa (figura 11.26). Já para  $\lambda = 15$  (figura 11.27), a aproximação melhora bastante. Podemos adotar este critério: a partir de  $\lambda = 15$ , a distribuição normal com média e variância iguais a  $\lambda$  é uma boa aproximação para a distribuição de Poisson com parâmetro  $\lambda$ . Ao calcularmos probabilidades usando essa aproximação, também devemos adotar a correção de continuidade, analogamente ao caso da distribuição binomial.

#### Poisson x Normal

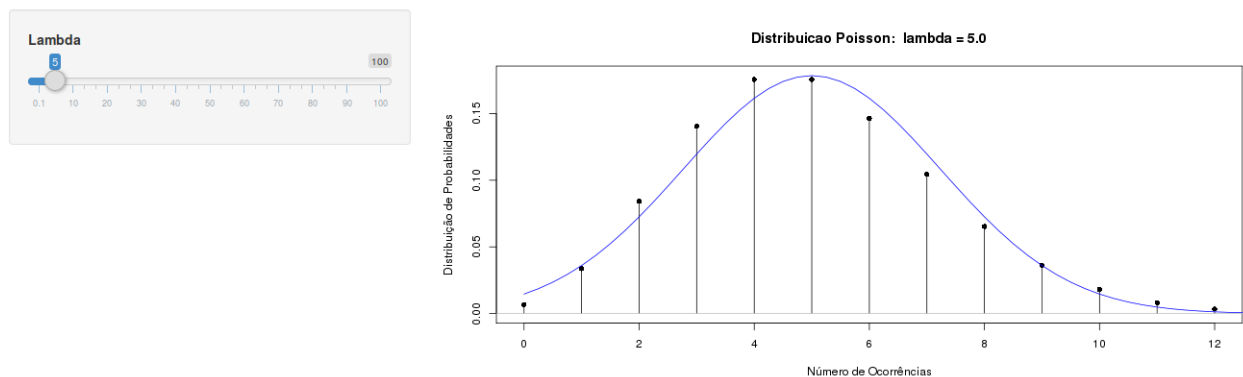


Figura 11.25: Aplicação para verificar a relação entre a distribuição normal e a distribuição de Poisson.

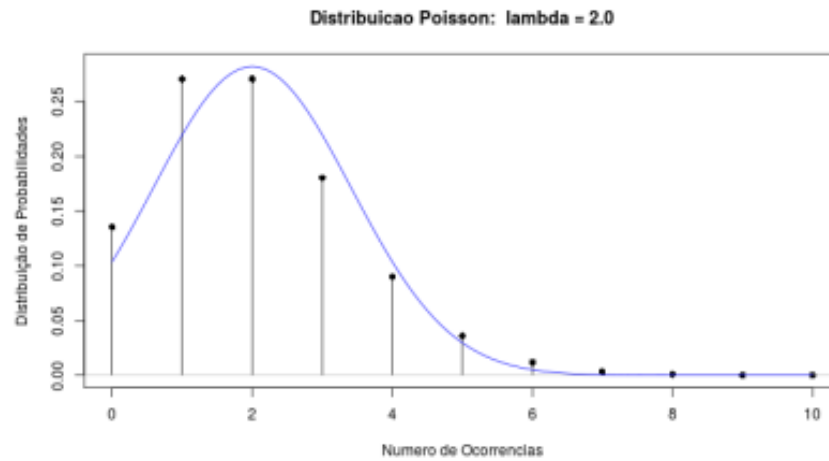


Figura 11.26: Gráficos das distribuições  $\text{Pois}(2)$  e  $N(2, 4)$ . A aproximação da normal não é boa.

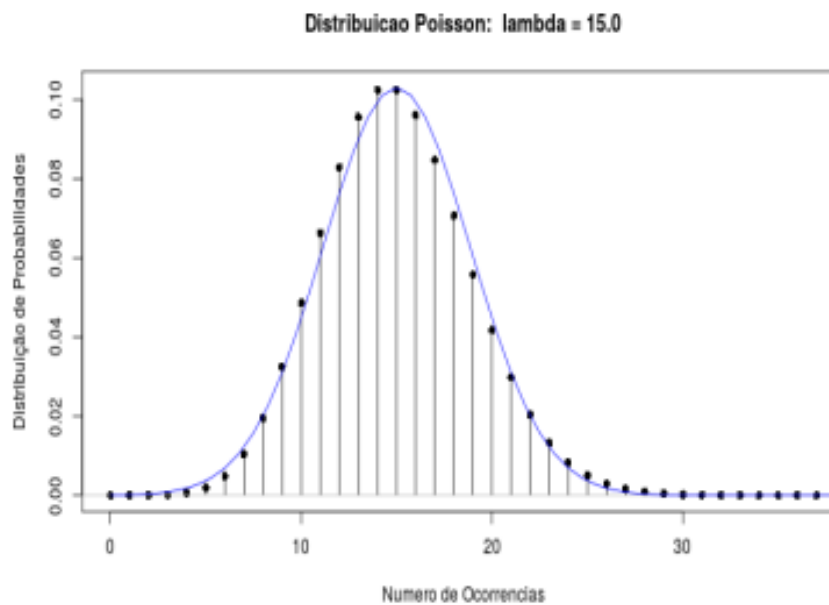


Figura 11.27: Gráficos das distribuições  $\text{Pois}(15)$  e  $N(15, 15)$ . A aproximação da normal parece bastante razoável.

## 11.8 Distribuição exponencial

Uma variável aleatória contínua  $X$  possui uma **distribuição exponencial** com parâmetro  $\lambda$  quando a sua função densidade de probabilidade for dada por:

$$f(X) = \begin{cases} \lambda e^{-\lambda X}, & \text{se } X > 0 \\ 0, & \text{caso contrário} \end{cases} \quad (11.10)$$

Essa é uma distribuição importante, utilizada para estudos que envolvem o intervalo de tempo entre eventos. Está intimamente associada à variável aleatória de Poisson, que estuda o número de eventos que ocorrem em um dado intervalo de tempo. O parâmetro  $\lambda$  corresponde à taxa média de eventos na unidade de tempo (ou distância, ou outra unidade de interesse).

A figura 11.28 mostra gráficos da função exponencial para três valores diferentes do parâmetro  $\lambda$ . Cada um dos gráficos pode ser obtido a partir do *R Commander*, acessando a caixa de diálogo para obter o gráfico de uma distribuição exponencial a partir do menu mostrado na figura 11.9.

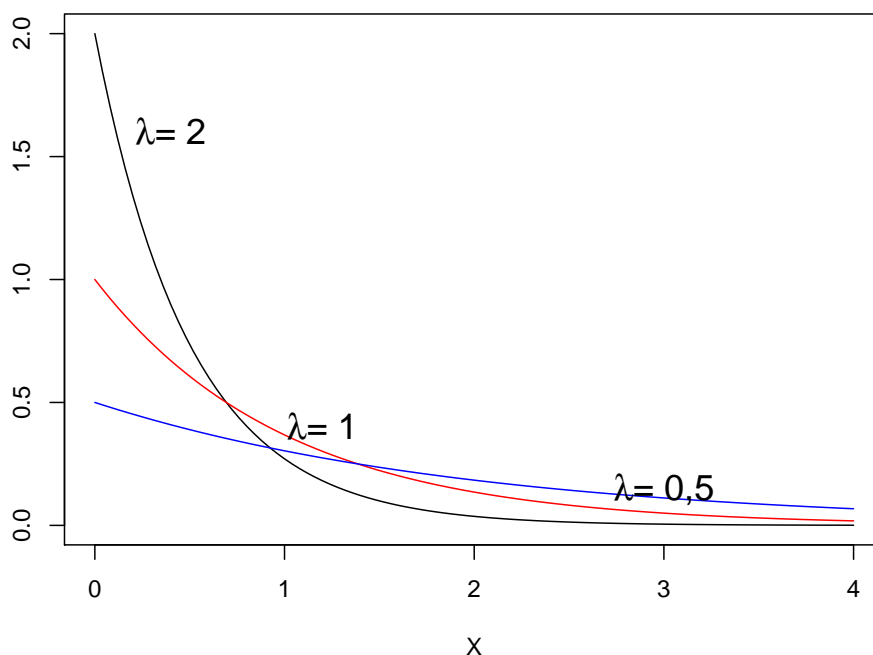


Figura 11.28: Gráficos da função de densidade de uma distribuição exponencial para  $\lambda = 0,5$ ; 1 e 2, respectivamente.

O valor esperado e a variância da distribuição exponencial são dados por:

$$E[X] = \int_0^{\infty} x \lambda e^{-\lambda x} dx = \frac{1}{\lambda} \quad (11.11)$$

$$var(X) = \int_0^{\infty} \left(x - \frac{1}{\lambda}\right)^2 \lambda e^{-\lambda x} dx = \frac{1}{\lambda^2} \quad (11.12)$$

Logo a variância de uma variável aleatória com distribuição exponencial é igual ao quadrado da média.

A função de distribuição cumulativa da distribuição exponencial é facilmente obtida. Seja  $T$  uma variável aleatória com distribuição exponencial e parâmetro  $\lambda$ . Se  $t > 0$ , então temos:

$$F(t) = P(T \leq t) = \int_0^t \lambda e^{-\lambda x} dx = 1 - e^{-\lambda t}$$

Uma propriedade importante da função densidade exponencial é que ela não possui “memória”, ou seja:

$$P(T > r + s | T > r) = P(T > s)$$

ou seja, a probabilidade condicional do evento “encontrarmos um intervalo de tempo maior que a soma de dois intervalos  $r+s$ , dado que o intervalo de tempo é maior do que a primeira parcela da soma” é independente dessa parcela. Essa propriedade pode ser demonstrada, notando que

$$\begin{aligned} P(T > s) &= 1 - P(T \leq s) = 1 - F(s) = e^{-\lambda s} \\ P(T > r + s | T > r) &= \frac{P(T > r+s \text{ e } T > s)}{P(T > r)} = \frac{P(T > r+s)}{P(T > r)} \\ &= \frac{1 - F(r+s)}{1 - F(r)} \\ &= \frac{e^{-\lambda(r+s)}}{e^{-\lambda r}} = e^{-\lambda s} \end{aligned}$$

Existe uma relação importante entre a distribuição de Poisson e a função densidade exponencial. Para a distribuição de Poisson, consideramos o número de eventos em um dado intervalo  $t$ . Se os eventos ocorrem a uma taxa de  $\lambda$  eventos por unidade de tempo, então a média da distribuição de Poisson será  $\mu = \lambda t$ . Lembrando da expressão da distribuição de Poisson, podemos estimar qual é a probabilidade de termos zero eventos em um dado intervalo  $t$ , isto é,

$$P(X = 0) = Pois(\lambda t) = \frac{e^{-\lambda t} (\lambda t)^0}{0!} = e^{-\lambda t}$$

Se agora considerarmos a variável aleatória  $T$ , correspondente ao tempo entre um instante genérico e o instante em que o evento ocorre, esse valor será maior que um dado  $t$ , desde que não ocorra nenhum evento até esse ponto, isto é, ele será igual a  $P(X = 0)$  acima:

$$P(T > t) = e^{-\lambda t}$$

Logo teremos

$$P(T \leq t) = 1 - e^{-\lambda t}$$

que nada mais é do que a função distribuição cumulativa da exponencial. Concluindo, **para um processo de Poisson, o intervalo entre eventos segue uma distribuição exponencial.**

**Exemplo 3:** Se o tempo de vida médio de um componente eletrônico é igual a 100 horas, temos que a taxa de falhas por hora será  $\lambda = 1/100$ . Supondo que o tempo de vida desse componente possua uma distribuição exponencial, para obtermos a probabilidade de que o componente falhe antes de 50 horas, temos que usar a função distribuição de probabilidade (que é a cumulativa)

$$F(x) = 1 - e^{-\lambda x}, \text{ com } x = 50.$$

Logo obtemos para a probabilidade desejada

$$P(T \leq 50) = F(50) = 1 - e^{-50/100} = 0,393$$

## 11.9 Transformação de variáveis e variáveis independentes

Sejam  $X$  e  $Y$  duas variáveis aleatórias contínuas, tal que  $Y = aX + b$ , onde  $a$  e  $b$  são números reais quaisquer, ou seja,  $Y$  é uma transformação linear de  $X$ . Analogamente ao caso das variáveis aleatórias discretas, para variáveis contínuas, ao realizarmos uma transformação linear de variáveis, as seguintes fórmulas para o valor esperado e variância se aplicam:

$$E[Y] = aE[X] + b \tag{11.13}$$

$$var(Y) = a^2 var(X) \tag{11.14}$$

Sejam  $X$  e  $Y$  duas variáveis aleatórias independentes e  $Z$  uma variável aleatória, tal que:

$$Z = X + Y$$

Então:

$$E[Z] = E[X] + E[Y] \tag{11.15}$$

$$var(Z) = var(X) + var(Y) \tag{11.16}$$

Esses resultados se estendem para uma soma de  $n$  variáveis independentes.

Em particular, se  $X$  e  $Y$  são normalmente distribuídas, com distribuição  $N(\mu, \sigma^2)$ , então  $Z$  também será normalmente distribuída com distribuição  $N(2\mu, 2\sigma^2)$ .

**Exemplo 4:** Dadas  $n$  variáveis aleatórias independentes  $X_1, X_2, \dots, X_n$  obtidas de uma distribuição normal  $N(\mu, \sigma^2)$ , obter o valor esperado e a variância das novas variáveis:

$$S_n = X_1 + X_2 + \dots + X_n$$

$$T_n = \frac{S_n}{n}$$

Solução: Como  $S_n$  é uma soma de  $n$  variáveis independentes com a mesma distribuição normal, então ela terá uma distribuição normal cuja média é  $n\mu$  e cuja variância é  $n\sigma^2$ . Logo:

$$S_n \sim N(n\mu, n\sigma^2).$$

Como  $T_n$  é uma transformação linear de  $S_n$ , podemos aplicar as fórmulas (11.13) e (11.14), estendendo-as para  $n$  variáveis. Logo:

$$\text{média de } T_n = \frac{1}{n}E[S_n] = \frac{n\mu}{n} = \mu$$

$$\text{variância de } T_n = \frac{1}{n^2}\text{var}(S_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Portanto  $T_n$  possui a mesma média de cada uma das parcelas da soma, porém a sua variância é a variância de uma das parcelas dividida por  $n$ .  $T_n$  possui uma dispersão menor do que cada uma das parcelas, para  $n > 1$ . Iremos discutir esse resultado com mais detalhes no capítulo 13.

Como  $S_n$  é normalmente distribuída, então:

$$T_n \sim N(\mu, \sigma^2/n).$$

Concluindo este tópico sobre distribuições de probabilidades, ressaltamos que há ainda um número grande de outras distribuições que não foram abordadas aqui. Algumas, como a  $t$  de Student e qui ao quadrado serão abordadas mais adiante quando forem apresentados os conceitos de intervalo de confiança e testes de hipóteses.



## 11.10 Exercícios

- 1) Numa distribuição de probabilidades para uma variável aleatória discreta, é possível saber a probabilidade para cada valor da variável aleatória. E para uma variável contínua, qual é a probabilidade de se obter um valor particular da variável aleatória?
- 2) Em relação à resposta anterior, como conciliar a sua resposta com o fato de que se você extrair uma amostra de uma população, você vai obter um conjunto de valores para a variável aleatória que você está interessado?
- 3) Como você traça o gráfico da função densidade de probabilidade
$$f(x) = \begin{cases} 0.1 - 0.005x, & \text{onde } 0 < x < 20 \end{cases}$$
Qual é a área compreendida entre o gráfico dessa função e o eixo X?
- 4) Suponha que a temperatura em um certo local seja normalmente distribuída com média  $50^\circ$  e variância 4. Indique, usando a curva normal como você calcularia a probabilidade de que a temperatura T em um dado dia esteja entre  $48^\circ$  e  $53^\circ$  C? E a probabilidade de que a temperatura T esteja acima de  $53^\circ$  C ou abaixo de  $44^\circ$  C?
- 5) Vamos supor que a pressão arterial sistólica dos membros de uma certa sociedade acadêmica siga uma distribuição normal com média igual a 125 mmHg e o desvio padrão igual a 10 mmHg.
  - a. Qual a variância dessa população?
  - b. Qual a probabilidade de selecionarmos aleatoriamente um membro dessa população e obtermos o valor de 125 mmHg exatamente para a pressão sistólica?
  - c. É possível extrairmos uma amostra de 10 pessoas e obtermos um valor da média de pressão sistólica acima de 140 mmHg? Explique a resposta.

# Capítulo 12

## Avaliação de testes diagnósticos

### 12.1 Introdução

Os conteúdos desta seção, do início da seção 12.2 e da seção 12.2.1 podem ser visualizados neste [vídeo](#).

Na prática clínica, o uso de um teste diagnóstico necessita de uma avaliação de quanto esse teste é capaz de distinguir quem tem uma doença de quem não tem. Esse teste é comparado com o padrão-ouro, ou seja, o método padrão de diagnóstico utilizado até então.

Na avaliação da qualidade dos estudos de testes diagnósticos, algumas perguntas devem ser respondidas para se conhecer a validade dos resultados:

- a) Foi feita uma comparação independente e cega com o teste padrão (padrão-ouro)?
- b) Os pacientes representam a população do local onde o teste será usado?
- c) O método do teste é adequadamente descrito e permite ser reproduzido?

Nesse sentido, a avaliação de um teste diagnóstico baseia-se na sua relação com algum meio de saber se a doença está ou não realmente presente – um indicador mais fiel da verdade é referido como **padrão-ouro**.

Critérios e métricas para a avaliação de testes diagnósticos são apresentados em diversas publicações, por exemplo (Fletcher et al., 2014) e (Guyatt et al., 2008). As métricas mais utilizadas para avaliar quantitativamente a qualidade de testes diagnósticos são a sensibilidade, especificidade, valores preditivos positivo e negativo, ou alternativas como a razão de verossimilhança e chance pós-teste. Os textos apresentam normogramas para obter a chance pós-teste a partir da razão de verossimilhança e a chance pré-teste (Fletcher et al., 2014), (Guyatt et al., 2008), e gráficos que mostram a influência da probabilidade pré-teste, da sensibilidade e da especificidade sobre os valores preditivos positivo e negativo, bem como a relação entre a sensibilidade e a especificidade para diferentes pontos de corte de um teste cujo resultado é uma variável numérica (curva ROC) (Owens and Sox, 2014).

Neste capítulo, serão apresentadas as métricas mais utilizadas para a avaliação de testes diagnósticos, considerando as seguintes situações:

- 1) o resultado do teste é expresso por uma variável dicotômica;
- 2) o resultado do teste é expresso por uma variável categórica com mais de duas categorias;
- 3) o resultado do teste é expresso por uma variável numérica contínua.

## 12.2 Teste dicotômico

Para avaliar um teste diagnóstico cujo resultado é dicotômico, dois esquemas de amostragem são frequentemente utilizados.

No primeiro esquema, selecionam-se uma amostra de pessoas com a doença e uma outra amostra de pessoas sem a doença, com diagnóstico estabelecido de acordo com o padrão-ouro, às quais são submetidas ao teste diagnóstico a ser avaliado (figura 12.1). Nessa figura,  $a$  pessoas doentes foram consideradas positivas de acordo com o teste,  $c$  pessoas doentes foram consideradas negativas pelo teste,  $b$  pessoas não doentes foram consideradas positivas pelo teste e  $d$  pessoas não doentes foram consideradas negativas pelo teste.

Apesar de a amostragem ser típica de um estudo de caso-controle, as medições são realizadas como em um estudo transversal, ou seja, em um curto lapso de tempo entre uma e outra medição, de modo que as medições possam ser consideradas como contemporâneas.

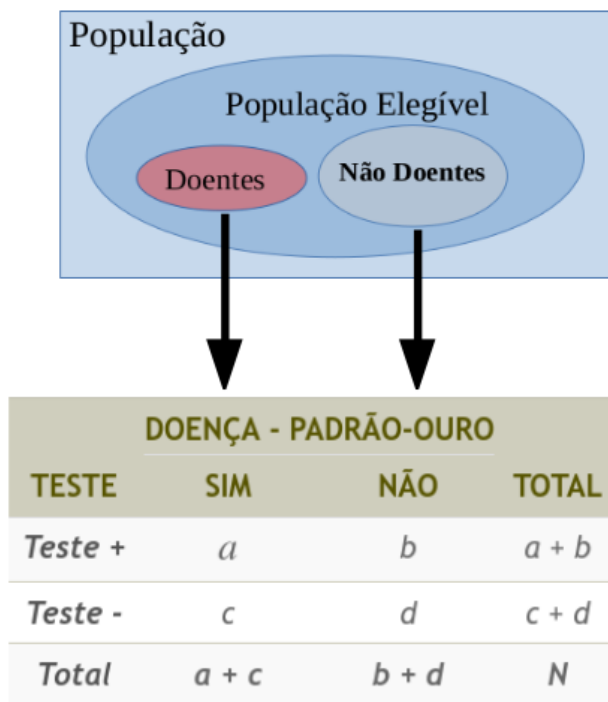


Figura 12.1: Avaliação de um teste diagnóstico com resultado dicotômico, utilizando uma amostragem típica de um estudo de caso-controle.

No segundo esquema, seleciona-se uma amostra de pessoas a partir de uma população elegível, às quais são submetidas tanto ao teste diagnóstico em avaliação quanto ao padrão-ouro (figura 12.2). As frequências das células da tabela 2x2 são calculadas de modo análogo ao da tabela mostrada na figura 12.1.

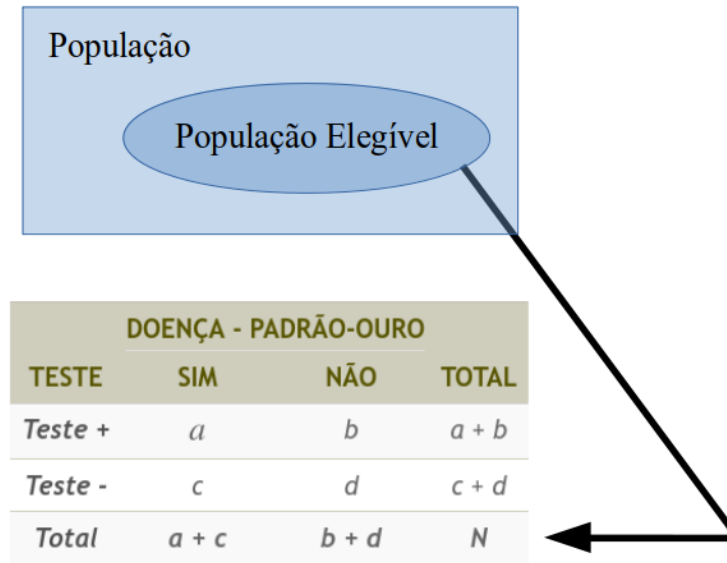


Figura 12.2: Avaliação de um teste diagnóstico com resultado dicotômico, utilizando uma amostragem típica de um estudo transversal.

Independentemente do esquema de amostragem, os resultados podem ser apresentados conforme a tabela 12.1, onde quatro situações podem ocorrer:

- 1) o resultado do teste é positivo e a doença está realmente presente: verdadeiro positivo;
- 2) o resultado do teste é positivo, mas a doença não está presente: falso positivo;
- 3) o resultado do teste é negativo, mas a doença está presente: falso negativo;
- 4) o resultado do teste é negativo e a doença realmente não está presente: verdadeiro negativo.

Tabela 12.1: Apresentação dos resultados de uma avaliação de um teste diagnóstico onde o resultado do teste é uma variável dicotômica.

		Doença	
		Presente ( $D$ )	Ausente ( $\bar{D}$ )
Teste	Positivo ( $T^+$ )	Verdadeiro Positivo ( $a$ )	Falso Positivo ( $b$ )
	Negativo ( $T^-$ )	Falso Negativo ( $c$ )	Verdadeiro Negativo ( $d$ )

Duas medidas intrínsecas ao teste frequentemente usadas são: a **sensibilidade** e a **especificidade**.

### 12.2.1 Sensibilidade e especificidade

A sensibilidade é a fração dos pacientes doentes que tiveram resultado positivo no teste:

$$\textbf{Sensibilidade} = S = P(T^+|D) = \frac{a}{a+c} \quad (12.1)$$

A especificidade é a fração dos pacientes não doentes que tiveram resultado negativo no teste:

$$\textbf{Especificidade} = E = P(T^-|\bar{D}) = \frac{d}{b+d} \quad (12.2)$$

Ao realizar um teste diagnóstico, o médico está interessado em saber qual é a probabilidade de o paciente ter a doença se o teste for positivo ou a probabilidade de o paciente não ter a doença se o teste for negativo. Essas duas probabilidades são conhecidas como **valor preditivo positivo** (VPP) e **valor preditivo negativo** (VPN), respectivamente. Elas não podem ser obtidas diretamente a partir da tabela 12.1 quando os grupos de doentes e não doentes são gerados por meio uma amostragem como em um estudo de caso-controle, porque as duas amostras utilizadas para gerar a tabela (doentes e não doentes) permitem estimar a sensibilidade ( $P[T^+|D]$ ) e a especificidade ( $P[T^-|\bar{D}]$ ), mas não o VPP ( $P[D|T^+]$ ) ou o VPN ( $P[\bar{D}|T^-]$ ).

Os valores preditivos positivo e negativo também não podem ser obtidos diretamente a partir da tabela 12.1 quando os grupos de doentes e não doentes são gerados por meio de uma amostragem típica de estudos transversais, se a prevalência da doença na população à qual o paciente pertence é diferente da prevalência da doença na população alvo do estudo a partir do qual as métricas de sensibilidade e especificidade foram obtidas.

### 12.2.2 Valores preditivo positivo e negativo

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

O valor preditivo positivo (VPP) é a probabilidade de o paciente ter a doença se o teste for positivo:

$$\textbf{Valor Preditivo Positivo} = VPP = P(D|T^+) \quad (12.3)$$

A expressão  $P(D)$  representa a propabilidade de o paciente ter a doença antes de realizar o teste, ou seja, a prevalência da doença (ou a probabilidade pré-teste) em um determinado contexto.

Como visto no capítulo 7, seção 7.6, o valor preditivo positivo pode ser calculado por meio do teorema de Bayes, se supusermos que os valores de sensibilidade e especificidade são independentes da probabilidade pré-teste:

$$\begin{aligned} P(D|T^+) &= \frac{P(T^+|D)P(D)}{P(T^+)} \\ &= \frac{P(T^+|D)P(D)}{P(T^+ \cap D) + P(T^+ \cap \bar{D})} \\ &= \frac{P(T^+|D)P(D)}{P(T^+|D)P(D) + P(T^+|\bar{D})P(\bar{D})} \end{aligned}$$

Portanto:

$$VPP = \frac{S \cdot P_{pre-teste}}{S \cdot P_{pre-teste} + (1 - E) \cdot (1 - P_{pre-teste})} \quad (12.4)$$

O valor preditivo negativo (VPN) é a probabilidade de o paciente não ter a doença se o teste for negativo:

$$\textbf{Valor Preditivo Negativo} = VPN = P(\bar{D}|T^-) \quad (12.5)$$

Seguindo um raciocínio semelhante ao utilizado para obter a expressão do VPP, chegamos à expressão para o VPN:

$$\begin{aligned} VPN = P(\bar{D}|T^-) &= \frac{P(T^-|\bar{D})P(\bar{D})}{P(T^-|D)P(D) + P(T^-|\bar{D})P(\bar{D})} \\ VPN &= \frac{E \cdot (1 - P_{pre-teste})}{E(1 - P_{pre-teste}) + (1 - S) \cdot P_{pre-teste}} \end{aligned} \quad (12.6)$$

**Exemplo:** Vamos considerar o estudo de Malacarne et al. (Malacarne et al., 2019), que avaliou o desempenho de testes para o diagnóstico de tuberculose pulmonar em populações indígenas no Brasil. Os resultados para o teste rápido molecular (TRM) em comparação à cultura de escarro (teste padrão) para todas as amostras de escarro combinadas são mostrados na tabela 12.2.

A partir da tabela 12.2, podemos calcular a sensibilidade e a especificidade do teste:

$$S = P(T^+|D) = 54/58 = 0,931$$

Tabela 12.2: Avaliação do teste rápido molecular (TRM) para detectar tuberculose (TB).  
Fonte: (Malacarne et al., 2019) ([CC BY-NC](#)).

Teste Rápido Molecular (TRM)	Cultura do escarro		Totais
	Com TB (D)	Sem TB ( $\bar{D}$ )	
<b>Teste positivo (<math>T^+</math>)</b>	<del>54</del>	<del>7</del>	<b>61</b>
<b>Teste negativo (<math>T^-</math>)</b>	<del>4</del>	<del>401</del>	<b>405</b>
	<b>58</b>	<b>408</b>	<b>466</b>

$$E = P(T^-|\bar{D}) = 401/408 = 0,983$$

Supondo que a prevalência seja  $P(D) = 0,1 = 10\%$  e substituindo os valores de S, E e  $P(D)$  na fórmula do valor preditivo positivo, temos:

$$VPP = \frac{0,931 \cdot 0,10}{0,931 \cdot 0,10 + (1 - 0,983)(1 - 0,10)} = 0,858 = 85,8\%$$

O fato de o teste diagnóstico dar positivo elevou a probabilidade de o paciente estar doente de 10% para 85,8%. Assim o VPP é influenciado pelos valores da sensibilidade, especificidade e da probabilidade pré-teste.

Substituindo os valores de S, E e  $P(D)$  na fórmula do valor preditivo negativo, temos:

$$VPN = \frac{0,983 \cdot (1 - 0,10)}{0,983 \cdot (1 - 0,10) + (1 - 0,931) \cdot 0,10} = 0,992 = 99,2\%$$

O fato de o teste diagnóstico ter dado negativo elevou a probabilidade de o paciente não estar doente de 90% para 99,2%. Nesse caso, como a probabilidade pré-teste de o indivíduo não ter a doença já era elevada, o fato de o teste dar negativo não alterou muito a probabilidade de o indivíduo estar doente.

### 12.2.3 Influência dos fatores que afetam os valores preditivos positivo e negativo

Os conteúdos desta seção e da seção 12.2.4 podem ser visualizados neste [vídeo](#).

Para entender melhor a influência dos fatores que afetam os valores preditivos positivo e negativo, vamos reconstruir a tabela 12.1, mas agora expressando as suas células em termos de probabilidades (figura 12.3). Assim, na tabela 12.1, substituímos a célula 1 pela sensibilidade e a célula 3 pelo seu complemento. Analogamente, substituímos a célula 4 pela especificidade e a célula 2 pelo seu complemento.

	<b>D</b>		<b>D<sup>-</sup></b>	
<b>T<sup>+</sup></b>	<b>1</b>	S	<b>2</b>	(1-E)
<b>T<sup>-</sup></b>	<b>3</b>	1-S	<b>4</b>	E
	1		1	

Figura 12.3: Preparação da tabela 12.1 para o cálculo do VPP e do VPN.

A partir da figura 12.3, o teorema de Bayes ajusta as células 1 e 3 de modo que a soma das duas probabilidades seja igual à probabilidade pré-teste ( $P[D]$ ) e as células 2 e 4, de modo que as somas das duas probabilidades seja  $1 - P_{\text{pre-teste}}$  (figura 12.4). Esse ajuste faz que com as células 1 a 4 reflitam a distribuição na população que tivesse a prevalência dada por  $P_{\text{pre-teste}}$  e os valores de sensibilidade e especificidade estimados a partir da tabela 12.1.

Na figura 12.4, pode-se estimar o VPP a partir das células 1 e 2, dividindo-se o valor da célula 1 pela soma das células 1 e 2, como mostrado na quarta coluna da tabela. Analogamente, pode-se estimar o VPN a partir das células 3 e 4, dividindo-se o valor da célula 3 pela soma das células 3 e 4.

	<b>D</b>		<b>D<sup>-</sup></b>	
<b>T<sup>+</sup></b>	<b>1</b>	$P_{\text{pre-teste}} \cdot S$	<b>2</b>	$(1-E) \cdot (1-P_{\text{pre-teste}})$
<b>T<sup>-</sup></b>	<b>3</b>	$P_{\text{pre-teste}} \cdot (1-S)$	<b>4</b>	$E \cdot (1-P_{\text{pre-teste}})$
	1	$P_{\text{pre-teste}}$	$1-P_{\text{pre-teste}}$	

$$VPP = \frac{P_{\text{pre-teste}} \cdot S}{P_{\text{pre-teste}} \cdot S + (1-E) \cdot (1-P_{\text{pre-teste}})}$$

$$VPN = \frac{E \cdot (1-P_{\text{pre-teste}})}{P_{\text{pre-teste}} \cdot (1-S) + E \cdot (1-P_{\text{pre-teste}})}$$

Figura 12.4: Utilização do teorema de Bayes para calcular o VPP e o VPN.

A tabela 12.3 mostra os dados do exemplo, substituídos na tabela mostrada na figura 12.4.



Tabela 12.3: Cálculo do VPP e VPN do teste rápido molecular (TRM) para detectar tuberculose.

	<b>D</b>	<b><math>\bar{D}</math></b>	
<b><math>T^+</math></b>	$0,10 \cdot 0,931$	$(1 - 0,983) \cdot (1 - 0,10)$	$VPP = \frac{0,931 \cdot 0,10}{0,931 \cdot 0,10 + (1 - 0,983) \cdot (1 - 0,10)} = 0,858$
<b><math>T^-</math></b>	$0,10 \cdot (1 - 0,931)$	$0,983(1 - 0,10)$	$VPN = \frac{0,983 \cdot (1 - 0,10)}{0,983 \cdot (1 - 0,10) + (1 - 0,931) \cdot 0,10} = 0,992$
<b>1</b>	<b>0,10</b>	<b>1 - 0,10</b>	

Os mesmos valores de VPP e VPN seriam obtidos se multiplicássemos os valores das células 1, 2, 3 e 4 pela mesma quantidade, por exemplo 100, já que essa constante será cancelada nas expressões para o VPP e VPN (tabela 12.4). Apesar de ser apenas uma mudança cosmética, é mais fácil trabalharmos com quantidades inteiras do que fracionárias.

Tabela 12.4: Cálculo do VPP e VPN do teste rápido molecular (TRM) para detectar tuberculose.

	<b>D</b>	<b><math>\bar{D}</math></b>	
<b><math>T^+</math></b>	$10 \cdot 0,931$	$(1 - 0,983) \cdot 90$	$VPP = \frac{0,931 \cdot 10}{0,931 \cdot 10 + (1 - 0,983) \cdot 90} = 0,858$
<b><math>T^-</math></b>	$10 \cdot (1 - 0,931)$	$0,983 \cdot 90$	$VPN = \frac{0,983 \cdot 90}{10 \cdot (1 - 0,931) + 0,983 \cdot 90} = 0,992$
<b>100</b>	<b>10</b>	<b>90</b>	

Quanto maior o valor da sensibilidade (menor a probabilidade de falsos negativos) e maior o valor da especificidade (menor a probabilidade de falsos positivos), mais acurado o teste.

Quanto maior o valor da sensibilidade (mais próximo de 1), mais o valor da célula 3 na figura 12.4 se aproxima de zero e o VPN vai se aproximando de 1.

Assim um teste muito sensível é útil para descartar o diagnóstico se ele der negativo.

Por outro lado, quanto maior o valor de especificidade (mais próximo de 1), mais o valor da célula 2 na figura 12.4 se aproxima de zero e o VPP vai se aproximando de 1.

Assim um teste muito específico é útil para confirmar o diagnóstico se ele der positivo.

Em um processo de triagem para uma certa doença, onde se procura detectar pessoas que possam ter a doença, é interessante a utilização de testes sensíveis e que possam ser utilizados em escala mais ampla de modo a termos poucos falsos negativos. Os indivíduos que forem considerados positivos no teste serão então submetidos a um teste mais específico para confirmar o diagnóstico.

Apesar de a montagem da tabela mostrada na figura 12.4 ser equivalente à aplicação direta das fórmulas (12.4) e (12.6) para os valores VPP e VPN, respectivamente, essa tabela pode ser programada numa planilha eletrônica, por exemplo, de modo a obter rapidamente o VPP e o VPN para quaisquer combinações de sensibilidade, especificidade e probabilidade pré-teste.

## 12.2.4 Aplicações que mostram a influência dos determinantes de VPP e VPN

As fórmulas (12.4) e (12.6) também podem ser programadas em aplicações que permitem visualizar a influência desses parâmetros sobre o VPP e o VPN, como será mostrado nas subseções seguintes.

### 12.2.4.1 Influência da probabilidade pré-teste sobre os valores preditivos positivo e negativo

A figura 12.5 mostra a tela inicial da aplicação [Probabilidade Pós-Teste x Probabilidade Pré-teste, Sensibilidade e Especificidade](#). Essa aplicação permite ao usuário visualizar como as probabilidades pós-teste (VPP e VPN) variam com a prevalência (probabilidade pré-teste) para diferentes valores de sensibilidade e especificidade.

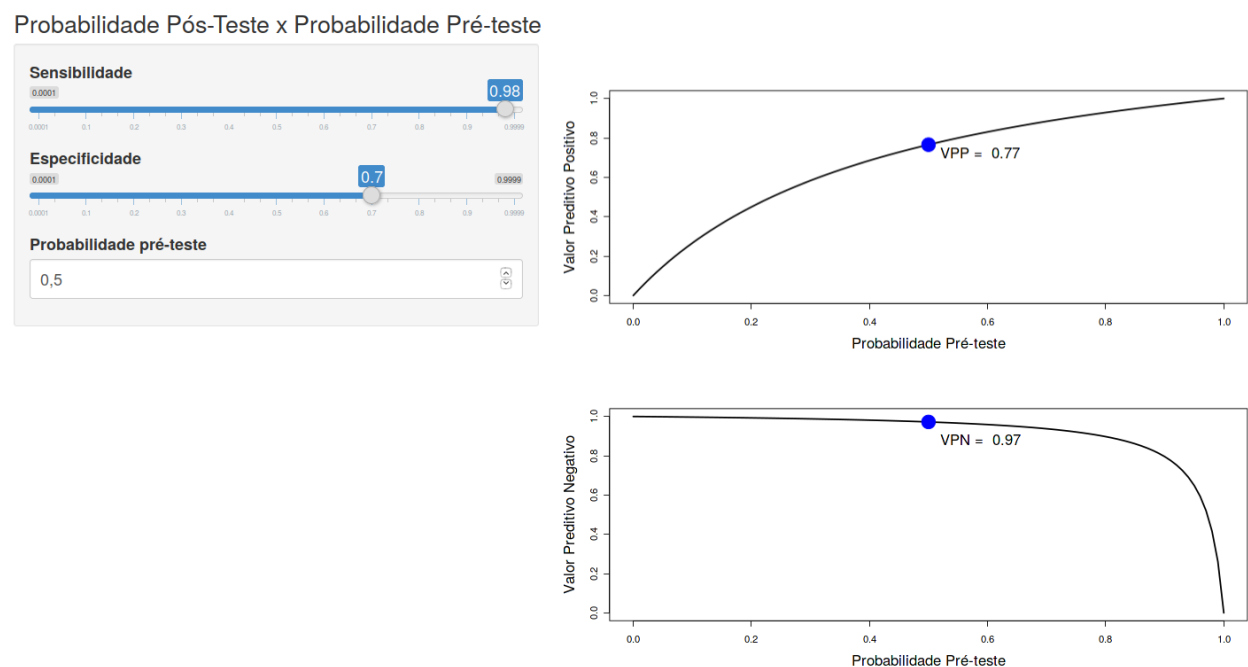


Figura 12.5: Aplicação que permite visualizar a dependência das probabilidades pós-teste em relação à prevalência para valores de sensibilidade e especificidade selecionados pelo usuário.

Para cada valor de sensibilidade e especificidade selecionados no painel à esquerda, a curva no gráfico superior mostra como o valor preditivo positivo, ou a probabilidade pós-teste se relaciona com a probabilidade pré-teste.

A curva no gráfico inferior mostra como o valor preditivo negativo se relaciona com a probabilidade pré-teste.

Os pontos em azul nos gráficos mostram os valores preditivos positivos e negativos, respectivamente, para o valor de probabilidade pré-teste selecionado no painel à esquerda.

Assim os valores de sensibilidade e especificidade determinam a curva nos gráficos à direita e o valor de probabilidade pré-teste determina o ponto azul.

É possível observar que as probabilidades pós-teste são fortemente influenciadas pela probabilidade pré-teste da doença. Mesmo se a sensibilidade e a especificidade forem altas, o VPP pode ser baixo se aplicado em uma população com baixa prevalência da doença.

#### 12.2.4.2 Influência da sensibilidade sobre os valores das probabilidades pós-teste

A figura 12.6 mostra a tela inicial da aplicação [Influência da Sensibilidade sobre a Probabilidade Pós-Teste](#). Essa aplicação permite ao usuário visualizar como a sensibilidade afeta a relação das probabilidades pós-teste (VPP e VPN) e a probabilidade pré-teste para diferentes valores de especificidade.

Influência da Sensibilidade sobre a Probabilidade Pós-Teste

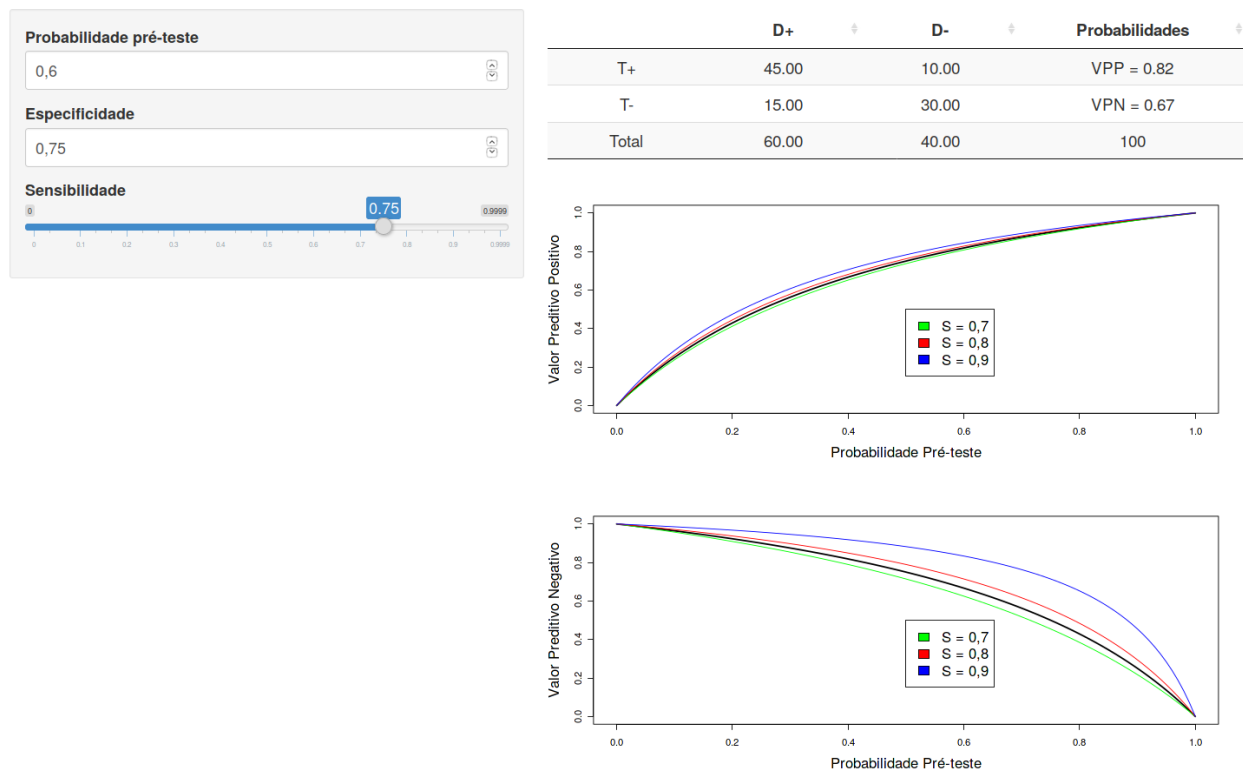


Figura 12.6: Aplicação que mostra a influência da sensibilidade sobre as curvas da probabilidade pós-teste em função da prevalência.

Três curvas de referência foram plotadas em cada gráfico, cada uma com um valor diferente de sensibilidade (curvas nas cores vermelho, verde e azul). Para os valores de sensibilidade e especificidade selecionados, as curvas em preto mostram a relação das probabilidades pós-teste e a probabilidade pré-teste.

É possível observar que, para um valor fixo de especificidade, variando a sensibilidade, a curva VPN x probabilidade pré-teste sofre um maior deslocamento do que a curva VPP x probabilidade pré-teste.

A tabela na figura mostra os valores de VPP e VPN para os valores de sensibilidade, especificidade e probabilidade pré-teste selecionados. Quando a sensibilidade se aproxima de 1, a parcela do denominador da fórmula do VPN (indicada pelo círculo 3 na tabela da figura 12.4) se aproxima de 0 e o VPN se aproxima de 1 (100%).

**Assim um teste com maior sensibilidade tem maior resolução quando o seu resultado é negativo, aumentando o VPN, mas também, por tender a ter menos falsos negativos, ele poderia ser aplicado na triagem para detectar pessoas com teste positivo e que possam ter a doença confirmada por um teste mais específico.**

#### 12.2.4.3 Influência da especificidade sobre os valores das probabilidades pós-teste

A figura 12.7 mostra a tela inicial da aplicação [Influência da Especificidade sobre a Probabilidade Pós-Teste](#). Essa aplicação permite ao usuário visualizar como a especificidade afeta a relação das probabilidades pós-teste (VPP e VPN) e a probabilidade pré-teste para diferentes valores de sensibilidade.

Influência da Especificidade sobre a Probabilidade Pós-Teste

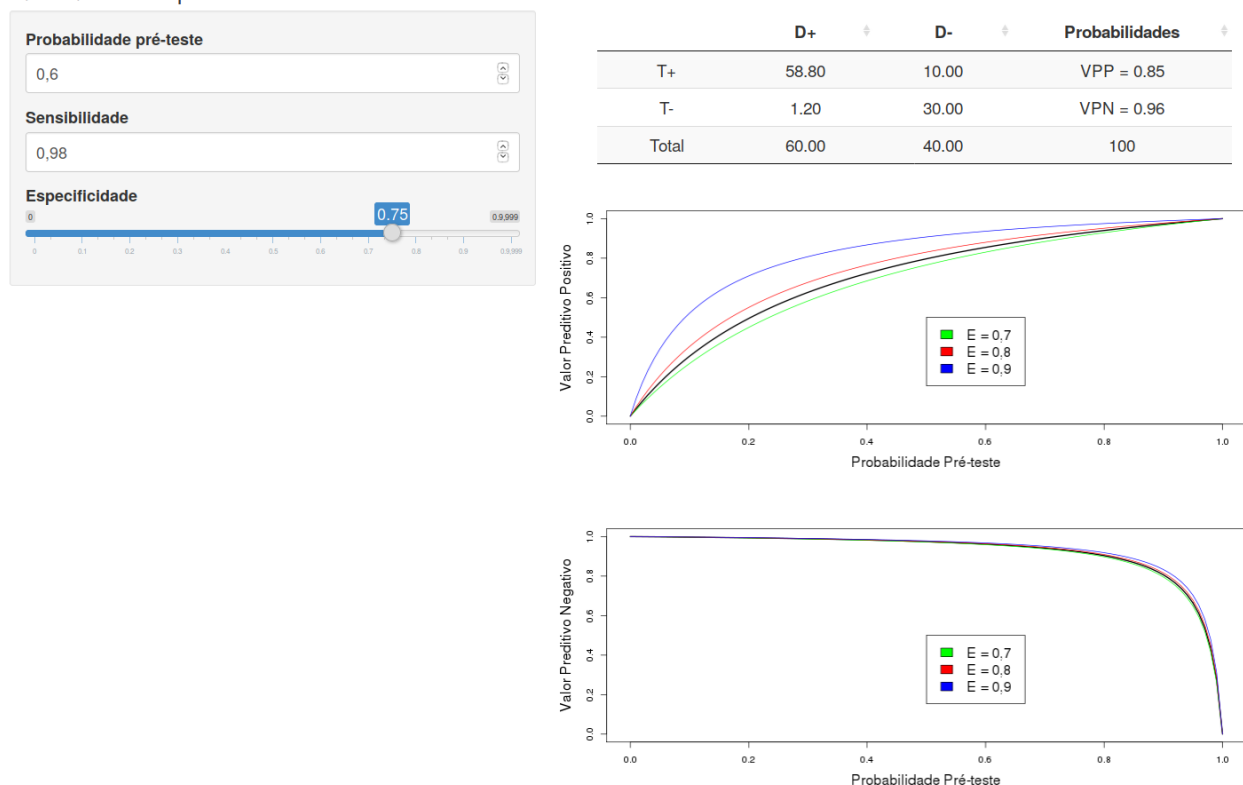


Figura 12.7: Aplicação que mostra a influência da especificidade sobre as curvas da probabilidade pós-teste em função da prevalência.

Três curvas de referência foram plotadas em cada gráfico, cada uma com um valor diferente de especificidade (curvas nas cores vermelho, verde e azul). Para os valores de sensibilidade e especificidade selecionados, as curvas em preto mostram a relação das probabilidades pós-teste e a probabilidade pré-teste.

É possível observar que, para um valor fixo de sensibilidade, variando a especificidade, a curva VPP x probabilidade pré-teste sofre um maior deslocamento do que a curva VPN x probabilidade pré-teste.

A tabela na figura mostra os valores de VPP e VPN para os valores de sensibilidade, especificidade e probabilidade pré-teste selecionados. Quando a especificidade se aproxima de 1, a parcela do denominador da fórmula do VPP (indicada pelo círculo 2 na tabela da figura 12.4) se aproxima de 0 e o VPP se aproxima de 1 (100%).

**Assim um teste com maior especificidade tem maior resolução quando o seu resultado é positivo, aumentando o VPP, e deve ser utilizado para confirmar o diagnóstico de uma doença.**

Uma outra maneira de caracterizar a eficácia de um teste diagnóstico é por meio da medida conhecida como razão de verossimilhança, explicada a seguir.

### 12.2.5 Razão de verossimilhança

Os conteúdos desta seção e das seções 12.2.6 e 12.3 podem ser visualizados neste [vídeo](#).

#### Teste positivo

Supondo que o resultado do teste seja positivo e, a partir do teorema de Bayes, temos:

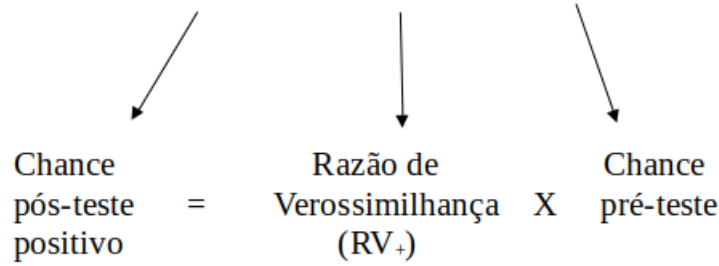
$$P(D|T^+) = \frac{P(T^+|D)P(D)}{P(T^+)}$$

Essa expressão pode ser manipulada da seguinte forma:

$$\begin{aligned} \frac{P(D|T^+)}{1 - P(D|T^+)} &= \frac{P(T^+|D)P(D)}{P(T^+)(1 - P(D|T^+))} = \frac{P(T^+|D)P(D)}{P(T^+)P(\bar{D}|T^+)} = \\ &= \frac{P(T^+|D)P(D)}{P(T^+|\bar{D})P(\bar{D})} \end{aligned}$$

Logo:

$$\frac{P(D|T^+)}{1 - P(D|T^+)} = \frac{P(T^+|D)}{P(T^+|\bar{D})} \frac{P(D)}{1 - P(D)} \quad (12.7)$$



A chance pós-teste positivo é a chance de o indivíduo ter a doença em caso de um resultado positivo do teste. Ela é o produto da razão de verossimilhança para o resultado positivo do teste ( $RV_+$ ) pela chance pré-teste. Podemos escrever a razão de verossimilhança para o resultado positivo do teste como:

$$RV_+ = \frac{P(T^+|D)}{P(T^+|\bar{D})} = \frac{S}{1 - E} \quad (12.8)$$

Uma vez calculada a chance pós-teste, o valor preditivo positivo será dado por:

$$\frac{P(D|T^+)}{1 - P(D|T^+)} = \text{Chance pós-teste} = \frac{VPP}{1 - VPP} \Rightarrow VPP = \frac{\text{Chance pós-teste}}{1 + \text{Chance pós-teste}} \quad (12.9)$$

### Teste negativo

Supondo que o resultado do teste seja negativo e seguindo um processo semelhante ao descrito para o teste positivo, seguimos os seguintes passos:

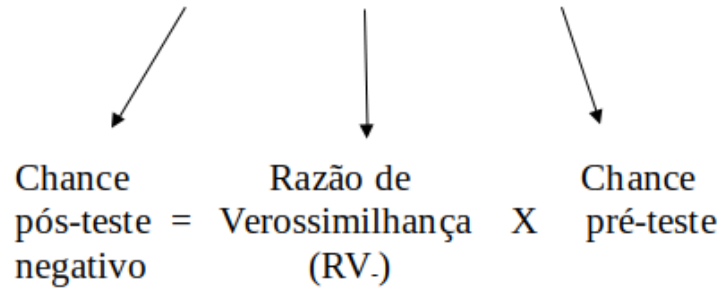
$$P(D|T^-) = \frac{P(T^-|D)P(D)}{P(T^-)}$$

que pode ser manipulada da seguinte forma:

$$\begin{aligned} \frac{P(D|T^-)}{1 - P(D|T^-)} &= \frac{P(T^-|D)P(D)}{P(T^-)(1 - P(D|T^-))} = \frac{P(T^-|D)P(D)}{P(T^-)P(\bar{D}|T^-)} = \\ &= \frac{P(T^-|D)P(D)}{P(T^-|\bar{D})P(\bar{D})} \end{aligned}$$

Logo:

$$\frac{P(D|T^-)}{1 - P(D|T^-)} = \frac{P(T^-|D)}{P(T^-|\bar{D})} \frac{P_{\text{pré-teste}}}{1 - P_{\text{pré-teste}}} \quad (12.10)$$



A chance pós-teste negativo é a chance de o indivíduo ter a doença em caso de resultado negativo do teste. Ela é o produto da razão de verossimilhança para o resultado negativo do teste (RV<sub>-</sub>) pela chance pré-teste. Podemos escrever a razão de verossimilhança para o resultado negativo do teste como:

$$RV_- = \frac{P(T^-|D)}{P(T^-|\bar{D})} = \frac{1 - S}{E} \quad (12.11)$$

Para o resultado negativo do teste, a chance pós-teste é dada por:

$$\frac{P(D|T^-)}{1 - P(D|T^-)} = \frac{1 - VP_N}{VP_N} \Rightarrow 1 - VP_N = \frac{\text{Chance pós-teste}}{1 + \text{Chance pós-teste}} \quad (12.12)$$

Em ambos os casos, chamando de T o resultado do teste (positivo ou negativo), podemos escrever:

$$\frac{P(D|T)}{1 - P(D|T)} = \frac{P(T|D)}{P(T|\bar{D})} \frac{P_{\text{pré-teste}}}{1 - P_{\text{pré-teste}}} \quad (12.13)$$



A chance pós-teste é a chance de o indivíduo ter a doença após o resultado do teste. Ela é o produto da razão de verossimilhança pela chance pré-teste.

$$\frac{P(D|T)}{1 - P(D|T)} = \text{Chance pós-teste} \Rightarrow P(D|T) = \frac{\text{Chance pós-teste}}{1 + \text{Chance pós-teste}}$$

A figura 12.8 mostra os cálculos da razão de verossimilhança, tanto para o resultado positivo do teste quanto para o resultado negativo. Para o resultado positivo, calculamos, em primeiro lugar, as proporções de resultados positivos em cada coluna (setas vermelhas). Em seguida, dividimos as duas proporções para obtermos a razão de verossimilhança para o resultado positivo do teste (seta verde). Analogamente, se procede para o cálculo da razão de verossimilhança para o resultado negativo do teste.

Teste	Doença				Razão de Verossimilhança
	Presente (D)		Ausente (D')		
	N	Proporção	N	Proporção	
Positivo (T <sup>+</sup> )	<div>a</div>	<div><math>P(T^+ D) = a / (a+c)</math></div>	<div>b</div>	<div><math>P(T^+ D') = b / (b+d)</math></div>	<div><math>P(T^+ D) / P(T^+ D')</math></div>
Negativo (T <sup>-</sup> )	<div>c</div>	<div><math>P(T^- D) = c / (a+c)</math></div>	<div>d</div>	<div><math>P(T^- D') = d / (b+d)</math></div>	<div><math>P(T^- D) / P(T^- D')</math></div>
	<div>a + c</div>		<div>b+d</div>		

$$RV_+ = \frac{\frac{a}{a+c}}{\frac{b}{b+d}}, \quad RV_- = \frac{\frac{c}{a+c}}{\frac{d}{b+d}}$$

Figura 12.8: Cálculo da razão de verossimilhança. As setas vermelhas e verdes indicam como calcular as razões de verossimilhança para cada resultado do teste.

A tabela 12.5 mostra os cálculos da razão de verossimilhança para os dados do teste rápido molecular (TRM) para detectar a tuberculose. Para um resultado positivo do teste, a chance pós-teste é 54,3 vezes maior do que a chance pré-teste, enquanto que, para um resultado negativo, a chance pós-teste é apenas 0,07 vezes a chance pré-teste.

Este [vídeo](#) mostra como calcular os valores de sensibilidade, especificidade e razão de verossimilhança para um teste dicotômico no R.



Tabela 12.5: Cálculo da RV do teste rápido molecular (TRM) para detectar tuberculose

	D	$\bar{D}$	Razão de Verossimilhança
$T^+$	54	7	$RV_+ = \frac{\frac{54}{58}}{\frac{7}{408}} = 54,3$
$T^-$	4	401	$RV_- = \frac{\frac{4}{58}}{\frac{401}{408}} = 0,07$
	58	408	

### 12.2.6 Influência da razão de verossimilhança sobre a probabilidade pós-teste

A figura 12.9 mostra a tela inicial da aplicação [Probabilidade Pós-Teste x Probabilidade Pré-teste e Razão de Verossimilhança](#). Essa aplicação permite ao usuário visualizar como a razão de verossimilhança afeta a relação da probabilidade de doença pós-teste e a probabilidade pré-teste. O usuário pode variar os valores da razão de verossimilhança e visualizar como se altera o gráfico da probabilidade de doença pós-teste x probabilidade pré-teste. Para um dado valor de probabilidade pré-teste e razão de verossimilhança, o valor da probabilidade de doença pós-teste é apresentado no gráfico. A razão de verossimilhança pode assumir valores entre 0 e infinito, sendo o valor 1 aquele que não afeta a chance pós-teste.

Probabilidade Pós-Teste x Probabilidade Pré-teste

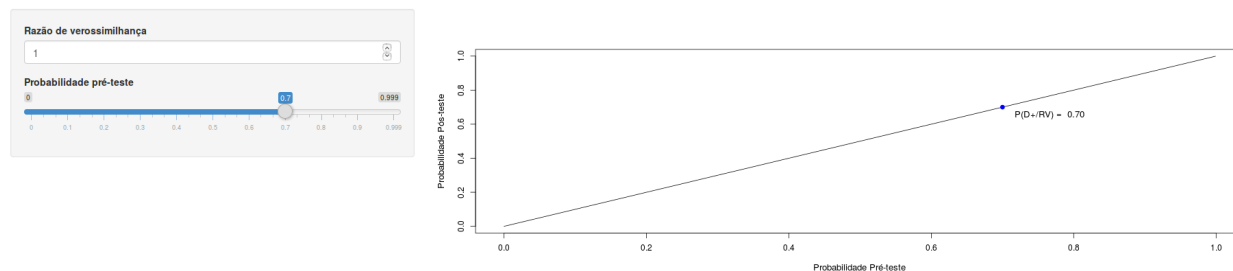


Figura 12.9: Aplicação que permite ao usuário visualizar a influência da razão de verossimilhança sobre a curva probabilidade de doença pós-teste x probabilidade pré-teste.

Quanto mais aumentamos o valor da razão de verossimilhança para o resultado de um teste, mais a curva da probabilidade de doença pós-teste x probabilidade pré-teste se desloca para cima e para a esquerda (figura 12.10). Quanto mais diminuirmos o valor da razão de verossimilhança para o resultado de um teste, mais a curva da probabilidade de doença pós-teste x probabilidade pré-teste se desloca para baixo e para a direita (figura 12.11). Assim um bom teste é aquele que possui alto valor de razão de verossimilhança (bem acima de 1) para o resultado positivo e baixo valor da razão de verossimilhança para o resultado negativo (bem abaixo de 1).

Probabilidade Pós-Teste x Probabilidade Pré-teste

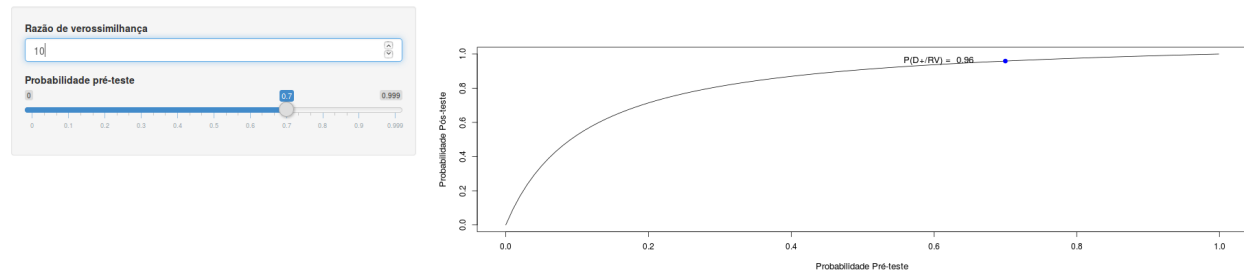


Figura 12.10: Quanto maior a razão de verossimilhança do resultado de um teste, mais a curva da probabilidade de doença pós-teste x prevalência se desloca para cima e para a esquerda.

Probabilidade Pós-Teste x Probabilidade Pré-teste

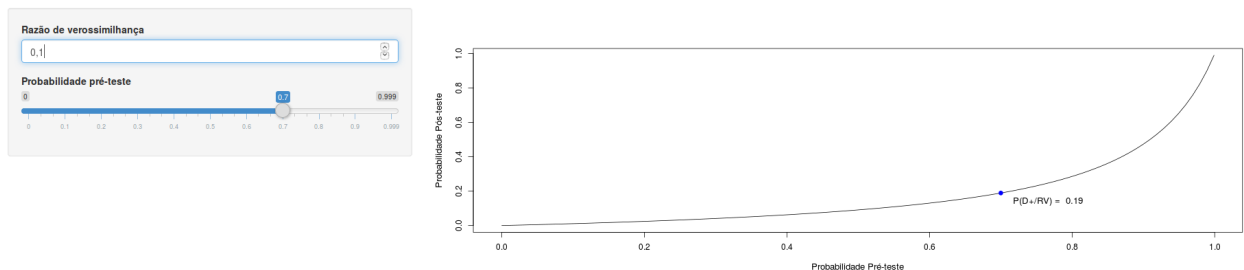


Figura 12.11: Quanto menor a razão de verossimilhança do resultado de um teste, mais a curva da probabilidade de doença pós-teste x prevalência se desloca para baixo e para a direita.

## 12.3 Variável de teste categórica ordinal

A tabela 12.6 mostra a distribuição de valores da glicemia em 2 horas em um teste de tolerância à glicose de mulheres com herança dos indígenas Pima, diabéticas e não diabéticas. Essa tabela foi construída a partir do conjunto de dados *PimaIndiansDiabetes2*, disponível no pacote *mlbench* (Leisch and Dimitriadou, 2010) ([GPL-2](#)).

Nessa tabela, o resultado do teste diagnóstico é expresso por uma variável categórica ordinal, com 4 categorias e não como uma variável binária. Se esse estudo fosse analisado em termos de sensibilidade e especificidade, teríamos que dicotomizar o resultado do teste, escolhendo uma categoria de corte. Uma possibilidade seria considerar como positivo somente os testes com valores de glicemia de 2 horas acima de 160 mg/dl e como negativo os demais resultados. Uma alternativa seria considerar como positivo os testes com valores de glicemia de 2 horas acima de 120 mg/dl e como negativo os demais resultados. Em quaisquer dos casos, a escolha irá considerar como do mesmo grupo (positivo ou negativo) resultados em diferentes faixas de glicemia de 2 horas.

Tabela 12.6: Distribuição de valores da glicemia em 2 horas em um teste de tolerância à glicose de mulheres com herança dos indígenas Pima, diabéticas e não diabéticas.

Faixas de valores da glicemia em 2 horas (mg/dl)	padrão-ouro	
	Diabetes Presente	Diabetes Ausente
<i>(160 - 200]</i>	<i>83</i>	<i>18</i>
<i>(120 - 160]</i>	<i>112</i>	<i>136</i>
<i>(80 - 120]</i>	<i>69</i>	<i>303</i>
<i>(40 - 80]</i>	<i>2</i>	<i>40</i>
	<b>266</b>	<b>497</b>

O uso da razão de verossimilhança evita esse problema. Generalizando o resultado mostrado na seção 12.2.5, a razão de verossimilhança pode ser calculada para cada resultado do teste separadamente, pela fórmula abaixo, onde T é o correspondente resultado do teste:

$$RV = \frac{P(T|D)}{P(T|\bar{D})} \quad (12.14)$$

A chance de o paciente ter a doença para um determinado resultado do teste é dada pela expressão

$$\text{Chance pós-teste} = \frac{\text{Razão de Verossimilhança (RV)}}{\text{Chance pré-teste}} \times$$

A figura 12.12 ilustra o cálculo da razão de verossimilhança para cada uma das faixas de glicemia em 2 horas em um teste de tolerância à glicose da tabela 12.6. Observem como a razão de verossimilhança preserva o caráter discriminatório de cada categoria do resultado. Usando a aplicação da figura 12.9, podemos obter a probabilidade de o paciente ter a doença para cada valor de RV e probabilidade pré-teste.

Faixas de valores da glicemia em 2 horas (mg/dl)	Diabetes				Razão de Verossimilhança
	Presente		Ausente		
	N	Proporção	N	Proporção	
[160-200)	83	83/266 = 0,312	18	18/497 = 0,036	8,62
[120-160)	112	112/266 = 0,421	136	136/497 = 0,274	1,54
[80-120)	69	69/266 = 0,259	303	303/497 = 0,610	0,43
[40-80)	2	2/266 = 0,008	40	40/497 = 0,080	0,09
	266		497		

Figura 12.12: Uso da razão de verossimilhança para avaliar a glicemia em 2 horas em um teste de tolerância à glicose para o diagnóstico de diabetes.

## 12.4 Variável de teste contínua

Os conteúdos das seções 12.4.1 e 12.4.2 podem ser visualizados neste [vídeo](#).

Para variáveis contínuas, como glicemia, pressão arterial sistólica, etc., as medidas de sensibilidade, especificidade, razão de verossimilhança, valores preditivos positivos e negativos também podem ser utilizadas, porém os resultados possíveis terão que ser agrupados em categorias. Para se trabalhar com a especificidade e a sensibilidade, um ponto de corte terá que ser estabelecido, sendo resultados de um lado do ponto de corte considerado como negativo e do outro lado, positivo. Para se trabalhar com a razão de verossimilhança, pode-se dividir a variável de teste em faixas, sendo a razão de verossimilhança calculada para cada faixa do resultado.

Inicialmente será apresentado o conceito da curva ROC.

### 12.4.1 Curva ROC

A figura 12.13 mostra histogramas da concentração de glicose no plasma em 2 horas em um teste de tolerância à glicose de mulheres com herança dos indígenas Pima, diabéticas e não diabéticas. Esses histogramas foram construídos a partir do conjunto de dados *PimaIndiansDiabetes2*, disponível no pacote *mlbench* (GPL-2).

É possível observar que o histograma das diabéticas está deslocado para a direita em relação ao histograma das não diabéticas, mas há uma superposição entre os dois histogramas, que é melhor visualizada quando os dois histogramas são apresentados no mesmo gráfico (figura 12.14). Em ambas as figuras, uma função densidade de probabilidade foi ajustada a cada histograma.

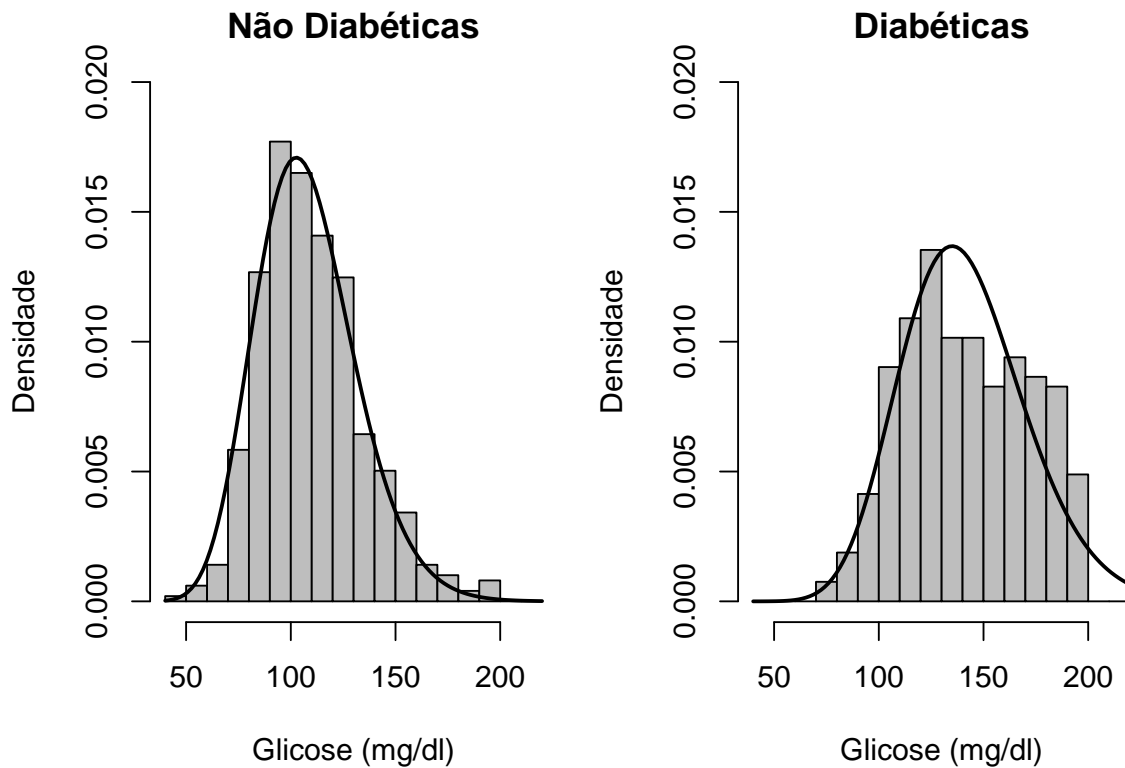


Figura 12.13: Histogramas da glicose de pacientes diabéticas e não diabéticas. Conjunto de dados: *PimaIndiansDiabetes2* do pacote *mlbench* (GPL-2).

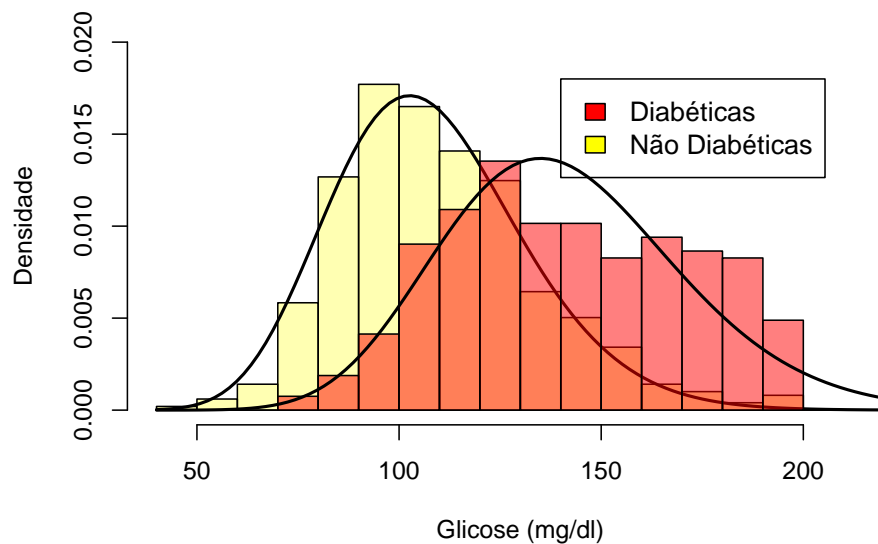


Figura 12.14: Histogramas da glicose de pacientes diabéticas e não diabéticas sobrepostos. Conjunto de dados: *PimaIndiansDiabetes2* (Leisch and Dimitriadou, 2010) (GPL-2).

Vamos considerar, hipoteticamente, duas populações de pessoas: diabéticas (DIAB) e não diabéticas (NAO DIAB) e que as funções densidade de probabilidade para a variável glicose

em 2 horas em um teste de tolerância à glicose para as duas populações sejam como mostradas na figura 12.15.

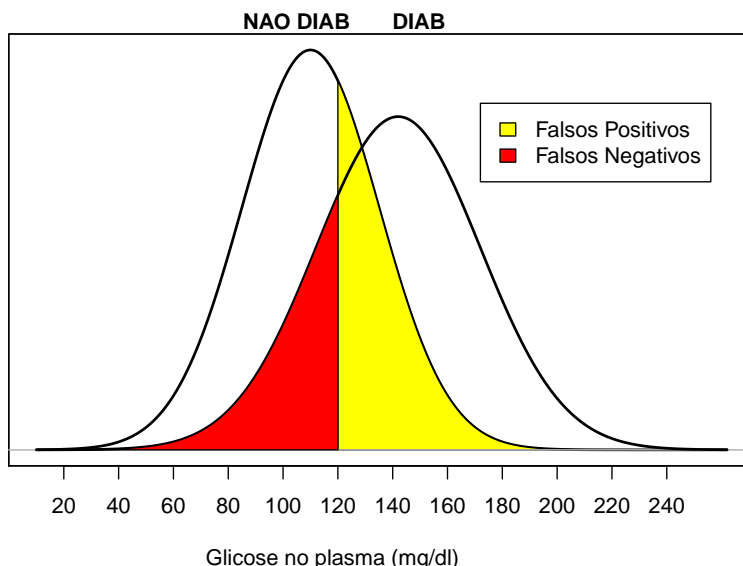


Figura 12.15: Funções densidade de probabilidade da variável glicose em 2 horas em um teste de tolerância à glicose em duas populações: diabéticos (DIAB) e não diabéticos (NAO DIAB). A área em amarelo indica a proporção de falsos positivos e a área em vermelho indica a proporção de falsos negativos, quando o valor 120 mg/dl é escolhido como ponto de corte.

Suponhamos que tenhamos escolhido o ponto de corte do teste igual a 120 mg/dl e consideramos como positivo os resultados acima e negativo os resultados abaixo de 120 mg/dl. Esse ponto divide o gráfico da função densidade das pessoas não diabéticas (NAO DIAB) em duas regiões. A área em amarelo indica a proporção de pessoas não diabéticas que possuem valores de glicose acima de 120 mg/dl. Essas pessoas serão os falsos positivos do teste e a área em amarelo representa a probabilidade de o teste dar positivo em pessoas não diabéticas ( $P(T^+|\bar{D})$ ). A área sob a curva das pessoas não diabéticas abaixo do ponto de corte representa os verdadeiros negativos ou a especificidade. Analogamente o ponto de corte divide o gráfico da função densidade das pessoas diabéticas (DIAB) em duas regiões. A área em vermelho indica a proporção de pessoas diabéticas que possuem valores de glicose abaixo de 120 mg/dl. Essas pessoas serão os falsos negativos do teste e a área em vermelho representa a probabilidade de o teste dar negativo em pessoas diabéticas ( $P(T^-|D)$ ). A área sob a curva das pessoas diabéticas acima do ponto de corte representa os verdadeiros positivos ou a sensibilidade.

A aplicação [Curva ROC](#) (figura 12.16) mostra o efeito sobre a sensibilidade e especificidade quando variamos o ponto de corte de uma variável de teste contínua.

## Curva ROC

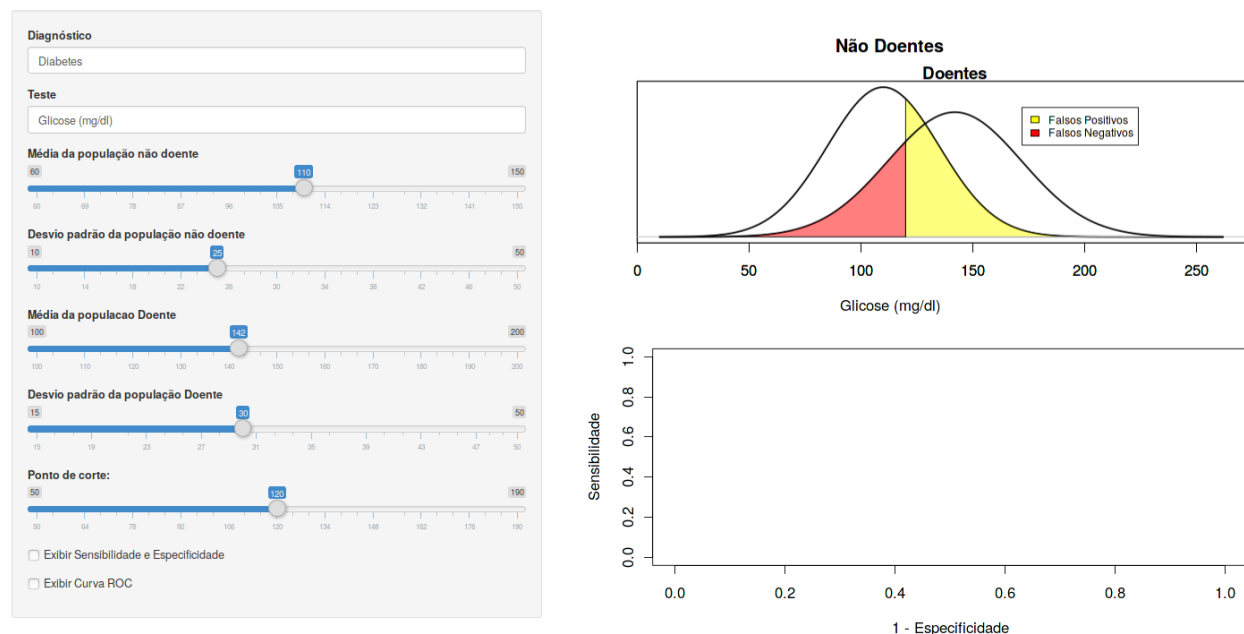


Figura 12.16: Aplicação que permite ao usuário visualizar a construção da curva ROC e a influência do ponto de corte sobre os valores de sensibilidade e especificidade.

Ao marcarmos a caixa de seleção *Exibir Sensibilidade e Especificidade*, o ponto (azul) correspondente aos valores de sensibilidade e especificidade definidos pelo ponto de corte selecionado será mostrado no gráfico da parte inferior (figura 12.17). O eixo X nesse gráfico corresponde aos valores de  $1 - \text{Especificidade}$  e o eixo Y aos valores de *Sensibilidade*. Para cada ponto de corte, teremos valores correspondentes de especificidade e sensibilidade.

## Curva ROC

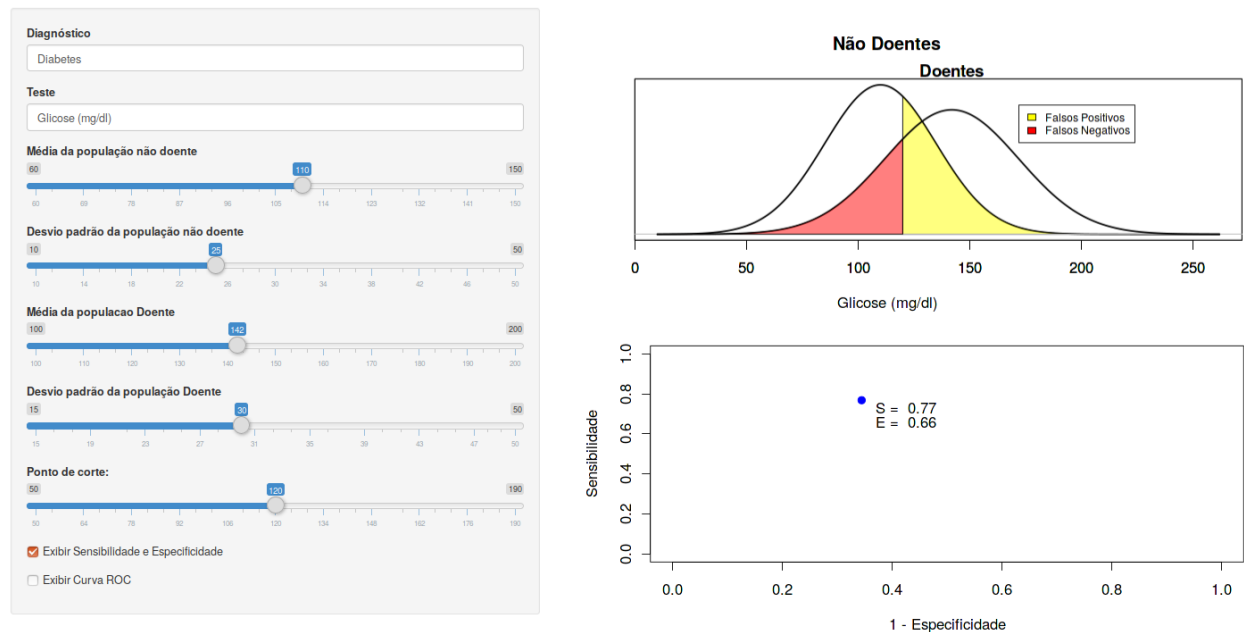


Figura 12.17: Valores da sensibilidade e especificidade para o ponto de corte escolhido na aplicação da figura 12.16.

Ao aumentarmos o ponto de corte, a sensibilidade diminui e a especificidade aumenta. O inverso ocorre se o ponto de corte for diminuído. Se unirmos os pontos cujas coordenadas são  $(1 - \text{especificidade}, \text{sensibilidade})$  ao variarmos o ponto de corte, obteremos uma curva denominada curva ROC (figura 12.18). Essa curva pode ser visualizada ao marcarmos a caixa de seleção *Exibir Curva ROC* na aplicação. O termo ROC significa *Receiving Operating Characteristics* e originou na área de telecomunicação.

O usuário pode variar os parâmetros das duas funções densidades de probabilidade (doentes e não doentes) e o ponto de corte e observar os efeitos sobre a curva ROC.



## Curva ROC

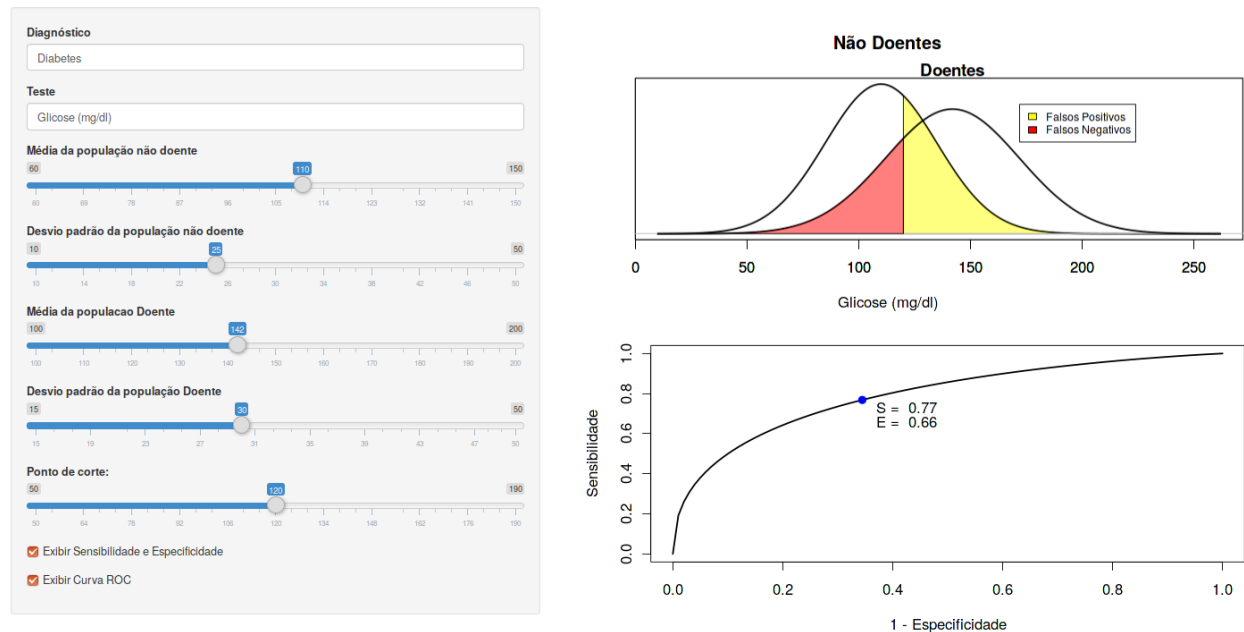


Figura 12.18: Curva ROC gerada a partir da variação do ponto de corte utilizado para classificar o teste como positivo ou negativo.

### 12.4.2 Comparação de testes

Quando houver dois testes diagnósticos cujas variáveis de teste sejam contínuas, como proceder para compará-los, já que os valores de especificidade e sensibilidade dependem do ponto de corte escolhido? Um critério bastante utilizado é comparar as áreas sob cada uma das curvas ROC. A área sob a curva ROC (*AUC* - *Area Under Curve*, em inglês) é a área compreendida entre a curva ROC e o eixo X. Essa área pode variar de 0 a 1. Quanto mais próxima a curva ROC do canto superior esquerdo do gráfico, mais próxima de 1 é a área sob a curva e melhor é o desempenho do teste.

A aplicação [Curvas ROC para dois testes](#) (figura 12.19) mostra as curvas ROC de dois testes para o diagnóstico da diabetes (o primeiro usa a glicemia em 2 horas no teste de tolerância à glicose e o segundo o índice de massa corporal). A curva em preto é a curva ROC da glicemia e a curva em azul corresponde ao IMC. Para os parâmetros selecionados na figura 12.19, as curvas ROC não se interceptam e a curva ROC da glicemia possui a maior área. Assim, para este caso, a glicemia é melhor para discriminar os diabéticos pois, para cada valor de especificidade, a glicemia possui maior sensibilidade do que o IMC.

Deve ser observado que ambas as medidas, glicemia e IMC, devido às características de suas curvas ROC, não devem ser usadas isoladamente para estabelecer um diagnóstico de diabetes.

O usuário pode selecionar os parâmetros das funções de probabilidade dos doentes e não doentes para cada teste e verificar as alterações nas respectivas curvas ROC.

### Curvas ROC para dois testes

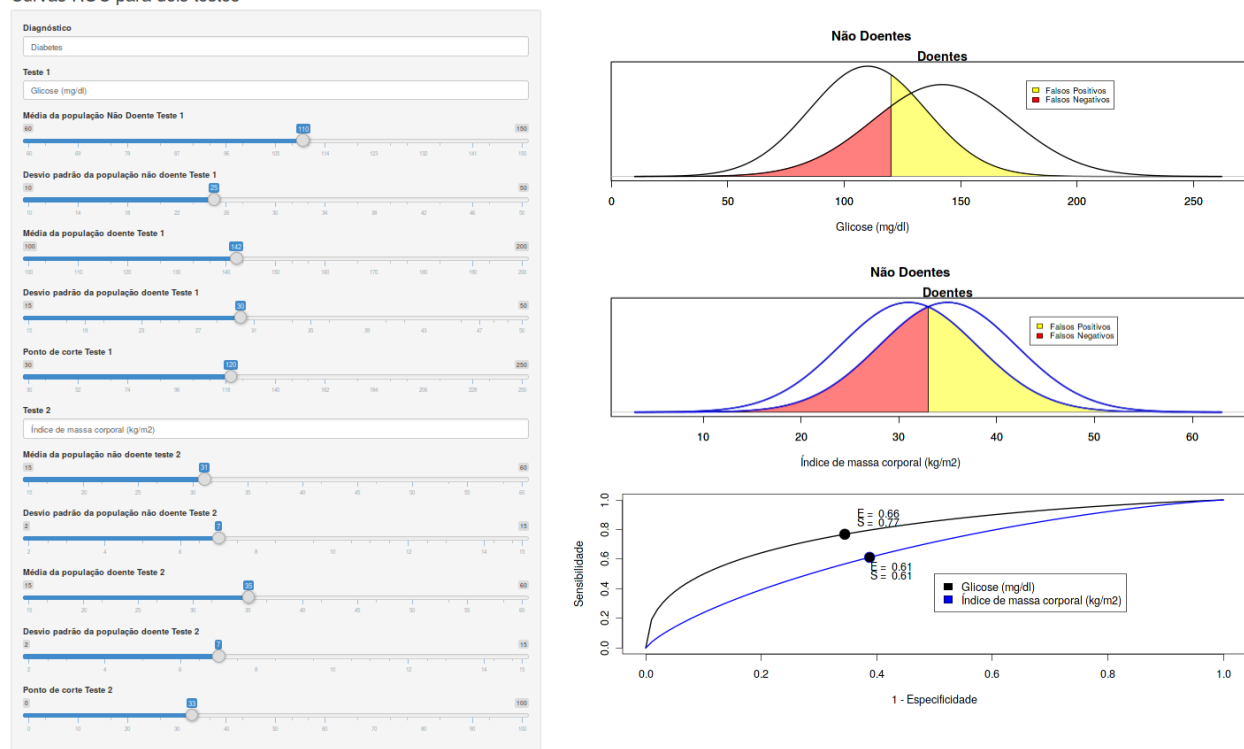


Figura 12.19: Aplicação que permite ao usuário comparar as curvas ROC de dois testes diagnósticos diferentes. A curva ROC preta corresponde ao teste 1 e a azul ao teste 2.

Nem sempre, porém, as curvas ROC se diferenciam como no parágrafo anterior. A figura 12.20 mostra um exemplo de dois testes cujas curvas ROC se interceptam. Nesse caso, a área sob a curva ROC pode não ser o melhor critério para selecionar o melhor teste, pois, à esquerda do ponto de interseção das duas curvas, o teste 2 é melhor, mas o teste 1 é melhor à direita da interseção das duas curvas. Outros fatores terão que ser levados em conta na escolha do melhor teste, tais como: custos, riscos de se tratar pessoas que não possuem a doença (falsos positivos), riscos de não se tratar pessoas que são doentes (falsos negativos), etc.

Outro ponto a ressaltar na comparação de dois testes diagnósticos por meio das respectivas curvas ROC é que as curvas obtidas em um determinado estudo são dependentes da amostra de pacientes do estudo e, portanto, podem variar de amostra para amostra. Assim a comparação de testes diagnósticos cujos resultados são expressos por uma variável numérica deve levar em conta também os intervalos de confiança das diferenças das áreas sob a curva ROC e/ou testes estatísticos que comparam as respectivas curvas ROC, para avaliar tanto a relevância clínica quanto a significância estatística das diferenças observadas.

Curvas ROC para dois testes

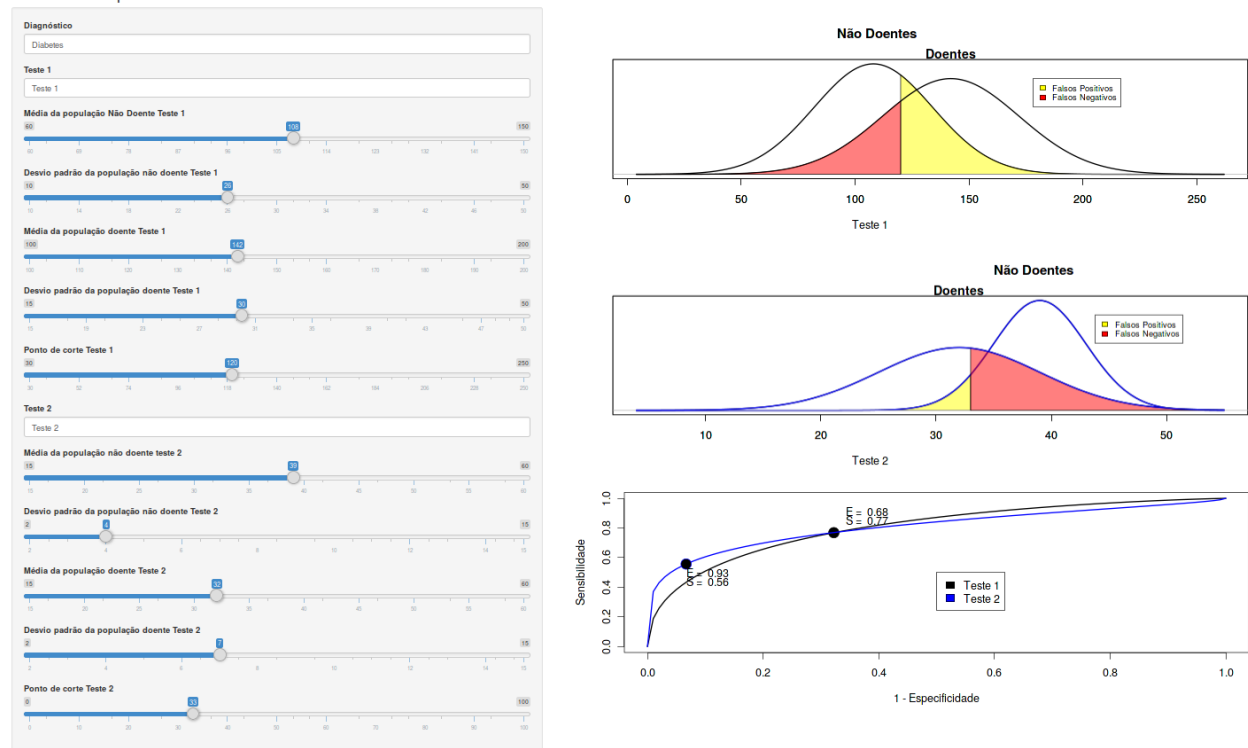


Figura 12.20: Exemplo de dois testes cujas curvas ROC se interceptam. Nesse caso, a área sob a curva ROC pode não ser o melhor critério para selecionar o melhor teste.

### 12.4.3 Uso da razão de verossimilhança em testes com variáveis contínuas

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

De modo análogo ao caso de uma variável de teste categórica com mais de duas categorias de resultados possíveis, pode-se também usar a razão de verossimilhança em um teste cuja variável é contínua. Para isso, divide-se os limites dos valores possíveis da variável de teste em faixas e calcula-se a razão de verossimilhança para cada faixa. Isso evita a necessidade de se estabelecer um ponto de corte, fazendo com que cada faixa de resultado tenha sua própria influência sobre a probabilidade de doença, quando se estima as chances de que uma determinada doença esteja presente. A tabela 12.7 mostra o cálculo da razão de verossimilhança para a glicemia em 2 horas em um teste de tolerância à glicose para discriminar diabéticos de não diabéticos para diversas faixas de valores, a partir dos mesmos dados da seção 12.4.1.

Tabela 12.7: Distribuição de valores da glicemia em 2 horas em um teste de tolerância à glicose em pacientes diabéticos e não diabéticos, com os cálculos das razões de verossimilhança.

Glicose (mg/dl)	Número de pessoas diabéticas (%)	Número de pessoas não diabéticas (%)	Razão de Verossimilhança
(40 - 60]	0 (0)	4 (0,8)	0
(60 - 80]	2 (0,75)	36 (7,2)	0,10
(80 - 100]	16 (6,0)	151 (30,3)	0,20
(100 - 120]	53 (19,9)	152 (30,6)	0,65
(120 - 140]	63 (23,7)	94 (18,9)	1,25
(140 - 160]	49 (18,4)	42 (8,5)	2,18
(160 - 180]	48 (18,0)	12 (2,4)	7,47
(180 - 200]	35 (13,2)	6 (1,2)	10,9
<b>Total</b>	<b>266</b>	<b>497</b>	

A aplicação [Razão de Verossimilhança para Variáveis Contínuas](#) (figura 12.21) mostra os valores da razão de verossimilhança para cada intervalo em que foram distribuídos os valores da variável de um teste diagnóstico. O usuário pode alterar os valores dos parâmetros das funções densidade de probabilidade entre os doentes e não doentes e o número de intervalos em que a variável de teste será dividida. Para cada intervalo, a razão de verossimilhança é calculada dividindo-se a área sob a curva dos doentes (probabilidade de se obter valores no intervalo entre os doentes) pela área sob a curva dos não doentes (probabilidade de se obter valores no intervalo entre os não doentes) no respectivo intervalo.

Razão de Verossimilhança para Variáveis Contínuas

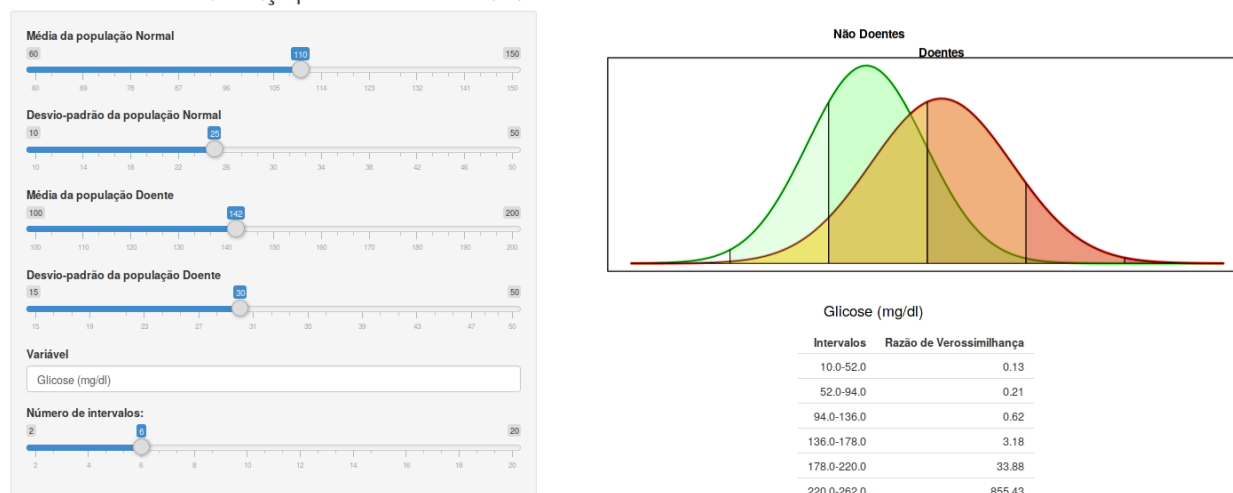


Figura 12.21: Construção da tabela de valores da razão de verossimilhança para cada intervalo em que foi dividida a variável de teste.

## 12.5 Análise de testes diagnósticos no R

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

Diversos pacotes do R podem ser utilizados para analisar dados relativos à avaliação de testes diagnósticos. Vamos utilizar nesta seção os pacotes *epiR* (também utilizado no capítulo 8) e o *RcmdrPlugin.ROC* ([GPL-2](#) | [GPL-3](#)). Vamos continuar a utilizar o conjunto de dados *PimaIndiansDiabetes2*, do pacote *mlbench* (Leisch and Dimitriadou, 2010) ([GPL-2](#)).

Inicialmente, vamos ler o conjunto de dados. Para isso, precisamos instalar o pacote *mlbench*:

```
install.packages("mlbench")
```

Em seguida, carregamos o pacote *mlbench* e o conjunto de dados *PimaIndiansDiabetes2*:

```
library(mlbench)
data(PimaIndiansDiabetes2, package="mlbench")
```

Para visualizar a descrição desse conjunto de dados, utilize o comando:

```
help("PimaIndiansDiabetes2")
```

Vamos trabalhar com três variáveis do conjunto de dados *PimaIndiansDiabetes2*:

- *glucose*: concentração de glicose no plasma em 2 horas em um teste de tolerância à glicose (mg/dl);
- *diabetes*: variável dicotômica, *pos* - diabética, *neg* - não diabética;
- *mass*: índice de massa corporal ( $\text{kg}/\text{m}^2$ ).

Vamos verificar as métricas para *glucose* como variável de teste diagnóstico de *diabetes mellitus gestacional*. Para calcular os valores de sensibilidade, especificidade e razão de verossimilhança para o resultado positivo/negativo, vamos inicialmente utilizar o pacote *epiR*. Como a variável *glucose* é contínua, temos que estabelecer um ponto de corte. Vamos supor que escolhamos o valor 120 mg/dl como ponto de corte. Então temos que montar uma tabela 2 x 2, que relaciona o teste positivo (*glucose* > 120 mg/dl) com o verdadeiro status da doença.

Inicialmente, vamos criar uma variável dicotômica, *glucose\_bin*, que assumirá os valores:

- *Positivo*, caso *glucose* > 120 mg/dl;
- *Negativo*, caso *glucose* ≤ 120 mg/dl.

Para criar essa variável, vamos recodificar a variável *glucose*, utilizando a seguinte opção no *R Commander*:

Dados ⇒ Modificação var. conj. dados ⇒ Recodificar variáveis

A tela para recodificação no *R Commander* deve ser preenchida como mostra a figura 12.22. Nessa tela, o nome da variável recodificada é *glucose\_bin*, a opção *Faça de cada nova variável um fator* foi marcada para que a nova variável seja da classe *factor*. As instruções para

recodificação são:

- 0:120 = “Negativo” -> indica que os valores de *glucose* menores ou iguais a 120 mg/dl serão recodificados para *Negativo* na variável *glucose\_bin*;
- 121:hi = “Positivo” -> indica que os valores de *glucose* maiores ou iguais a 121 mg/dl serão recodificados para *Positivo* na variável *glucose\_bin*.



Figura 12.22: Configuração da tela do *R Commander* para transformar a variável *glucose* em uma variável dicotômica.

O comando executado é mostrado a seguir:

```
PimaIndiansDiabetes2 <- within(PimaIndiansDiabetes2, {  
  glucose_bin <- Recode(glucose, '0:120 = "Negativo"; 121:hi = "Positivo"',  
                        as.factor=TRUE)  
})
```

Agora, temos que montar uma tabela 2 x 2, que relaciona o resultado do teste (*glucose\_bin*) com o status da doença (*diabetes*) e obter as métricas de avaliação do teste. A seguinte sequência de comandos obtém as medidas desejadas:

```
library(epiR)  
PimaIndiansDiabetes2$glucose_bin = ordered(PimaIndiansDiabetes2$glucose_bin,  
                                             levels=c("Positivo", "Negativo"))  
PimaIndiansDiabetes2$diabetes = ordered(PimaIndiansDiabetes2$diabetes,  
                                          levels=c("pos", "neg"))  
tab = table(PimaIndiansDiabetes2$glucose_bin, PimaIndiansDiabetes2$diabetes)  
epi.tests(tab, conf.level = 0.95)
```

```
##           Outcome +      Outcome -      Total
## Test +           195           154          349
## Test -            71           343          414
## Total            266           497          763
##
## Point estimates and 95 % CIs:
## -----
## Apparent prevalence           0.46 (0.42, 0.49)
## True prevalence               0.35 (0.31, 0.38)
## Sensitivity                   0.73 (0.68, 0.79)
## Specificity                   0.69 (0.65, 0.73)
## Positive predictive value     0.56 (0.50, 0.61)
## Negative predictive value     0.83 (0.79, 0.86)
## Positive likelihood ratio     2.37 (2.04, 2.75)
## Negative likelihood ratio     0.39 (0.31, 0.48)
## -----
```

A primeira função acima carrega a biblioteca *epiR*. Os dois comandos seguintes ordenam os níveis das variáveis *glucose\_bin* e *diabetes* para que as células da tabela 2x2 do teste possam ser apresentadas corretamente pelo *epiR*. Esse procedimento é semelhante ao utilizado no capítulo 8.

Em seguida, a função *table* cria a tabela 2x2 correspondente ao ponto de corte selecionado e a função *epi.tests* apresenta a tabela e os valores das medidas procuradas com os respectivos intervalos de confiança.

Os valores preditivos positivo e negativo apresentados são calculados, supondo-se que a prevalência da doença seja a mesma apresentada no teste ( $266/763 = 34,9\%$ ). Para outras prevalências, esses valores teriam que ser calculados por meio das expressões (12.4) e (12.6) da seção 12.2.2.

Em seguida, vamos construir a curva ROC que relaciona a *diabetes* com a variável de teste *glucose*. Para isso, vamos utilizar o plugin *RcmdrPlugin.ROC*. Caso o plugin não esteja instalado, use a seguinte instrução para instalá-lo:

```
install.packages("RcmdrPlugin.ROC")
```

Após a instalação, vamos carregar o plugin *RcmdrPlugin.ROC*, a partir da opção do menu:

Ferramentas ⇒ Carregar plugins do Rcmdr

Na tela mostrada na figura 12.23, selecionamos o plugin *RcmdrPlugin.ROC* e clicamos em *OK*. Em seguida, aparece uma janela solicitando a reinicialização do *R Commander*. Clicamos em *Sim*.

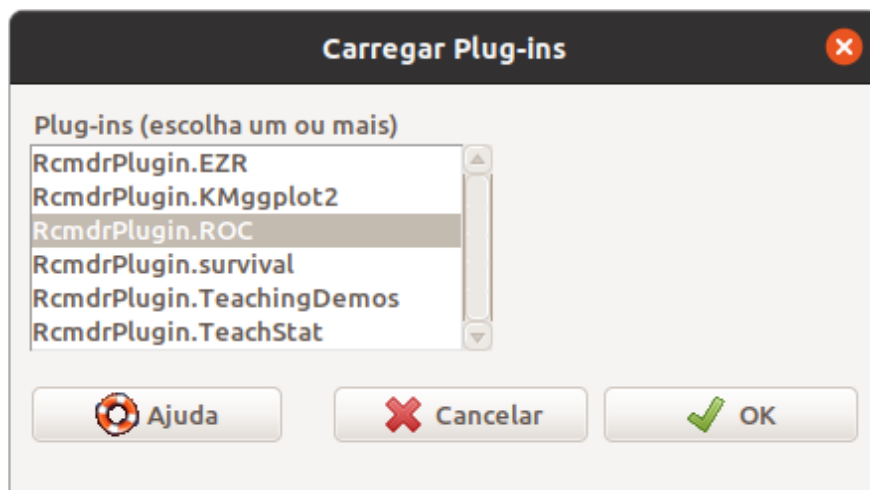


Figura 12.23: Seleção do *plugin RcmdrPlugin.ROC* para carregamento no *R Commander*.

Após a reinicialização do *R Commander*, ativamos novamente o conjunto de dados *PimaIndiansDiabetes2*, clicando no botão indicado pela seta verde na figura 12.24 e selecionando *PimaIndiansDiabetes2* na tela seguinte (figura 12.25).

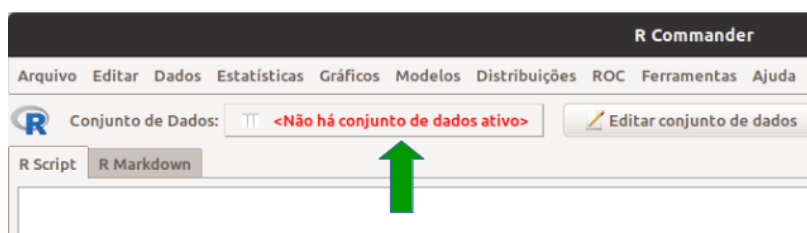


Figura 12.24: Botão do *R Commander* para selecionar um conjunto de dados ativo.

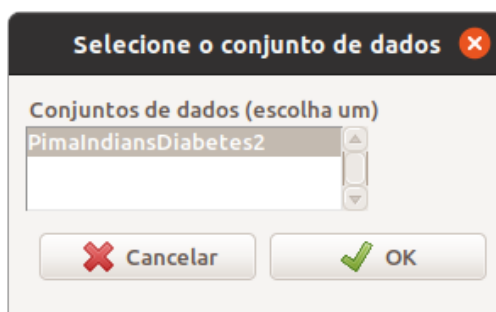


Figura 12.25: Seleção de *PimaIndiansDiabetes2* como o conjunto de dados ativo no *R Commander*.



Para construir a curva ROC, vamos acessar a seguinte opção no menu do *R Commander*:

ROC  $\Rightarrow$  pROC  $\Rightarrow$  plot ROC curve for data...

A figura 12.26 mostra a primeira aba da caixa de diálogo para configurar os parâmetros da curva ROC que será exibida. Nessa aba, selecionamos a variável de teste (*glucose*) e a variável de desfecho (*diabetes*). Na aba seguinte (figura 12.27), não vamos mexer nas opções previamente selecionadas. Na terceira aba (AUC), as opções padrões também não serão alteradas (figura 12.28).

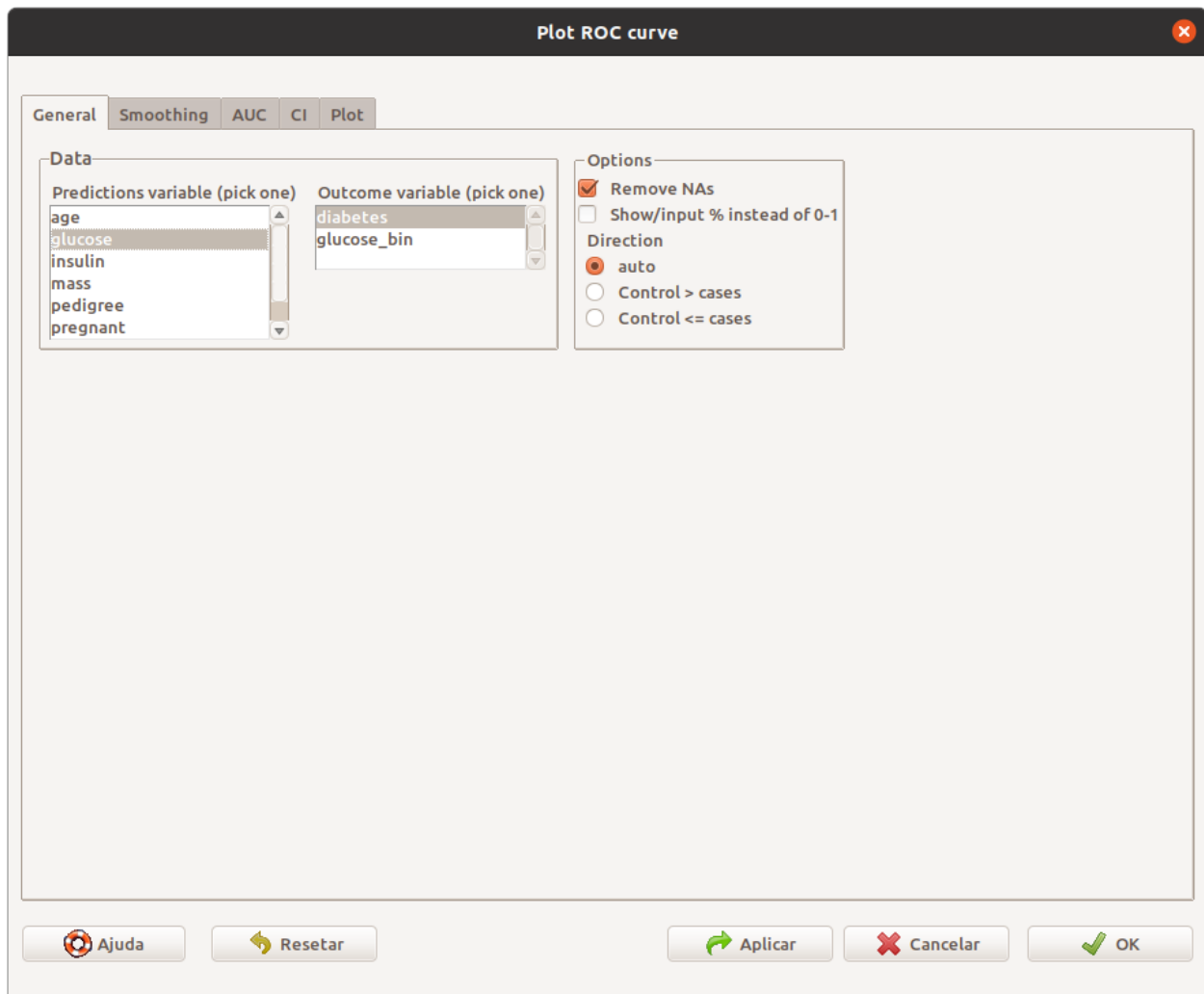


Figura 12.26: Diálogo para configurar a curva ROC: seleção das variáveis.

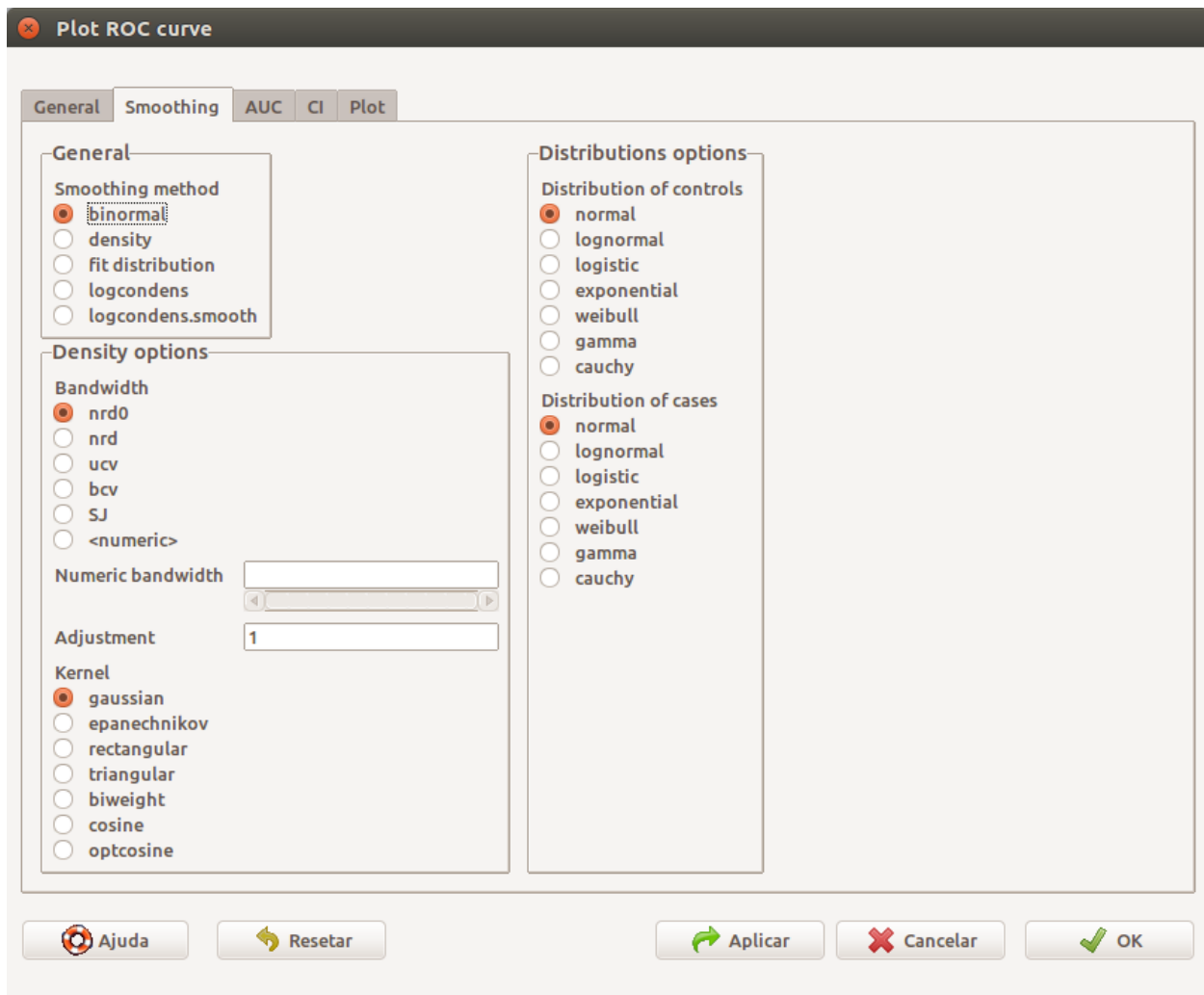


Figura 12.27: Diálogo para configurar a curva ROC: seleção de métodos para alisamento da curva. Não iremos mexer nestas configurações.

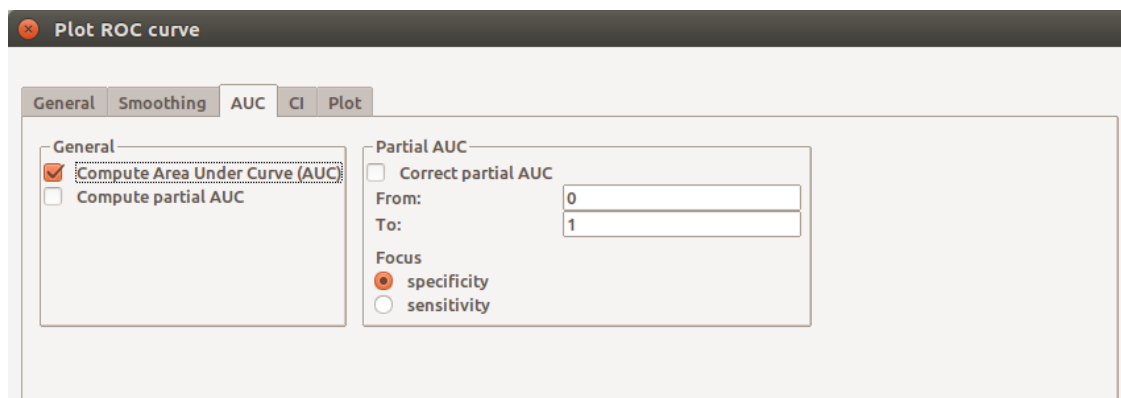


Figura 12.28: Diálogo para configurar a curva ROC: cálculo da área sob a curva ROC. Não iremos mexer nestas configurações.

Na figura 12.29, configuramos que intervalos de confiança serão mostrados e como serão calculados. Vamos marcar a opção *se* (sensibilidade) e deixar as demais opções padrões.

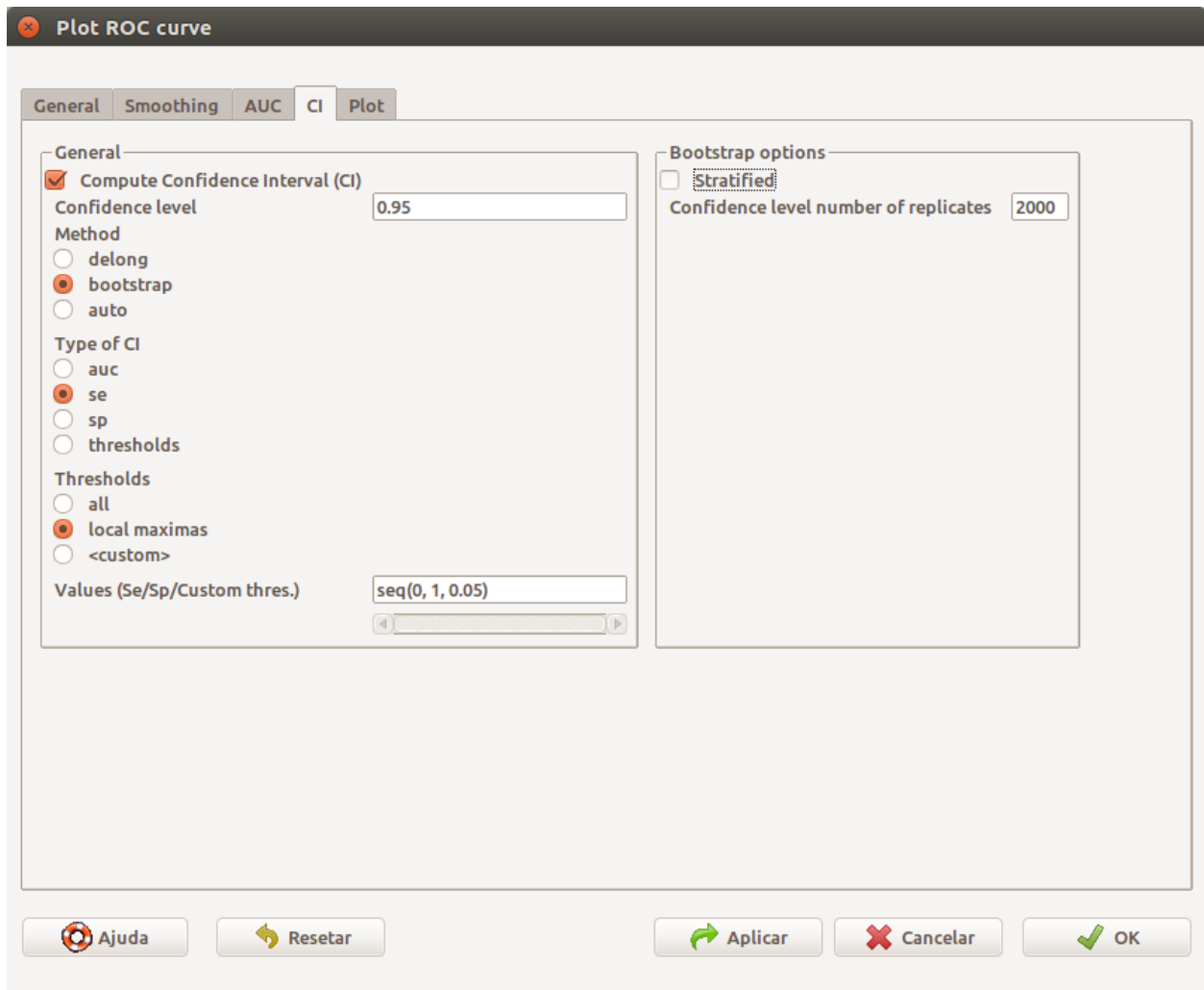


Figura 12.29: Diálogo para configurar a curva ROC: exibição dos intervalos de confiança.

Na figura 12.30, configuramos as opções de plotagem. Vamos marcar as opções *Display confidence interval*, *AUC* e *best:  $\max(\text{Sum}(Sp+Se))$* , e *bars* em *CI plot type*. Também vamos especificar os rótulos que irão ser exibidos nos eixos x e y. A opção *AUC* vai exibir no gráfico o valor da área sob a curva e o respectivo intervalo de confiança. A opção *best:  $\max(\text{Sum}(Sp+Se))$*  vai mostrar na curva ROC o ponto de corte onde é máxima a soma dos valores de especificidade e sensibilidade.

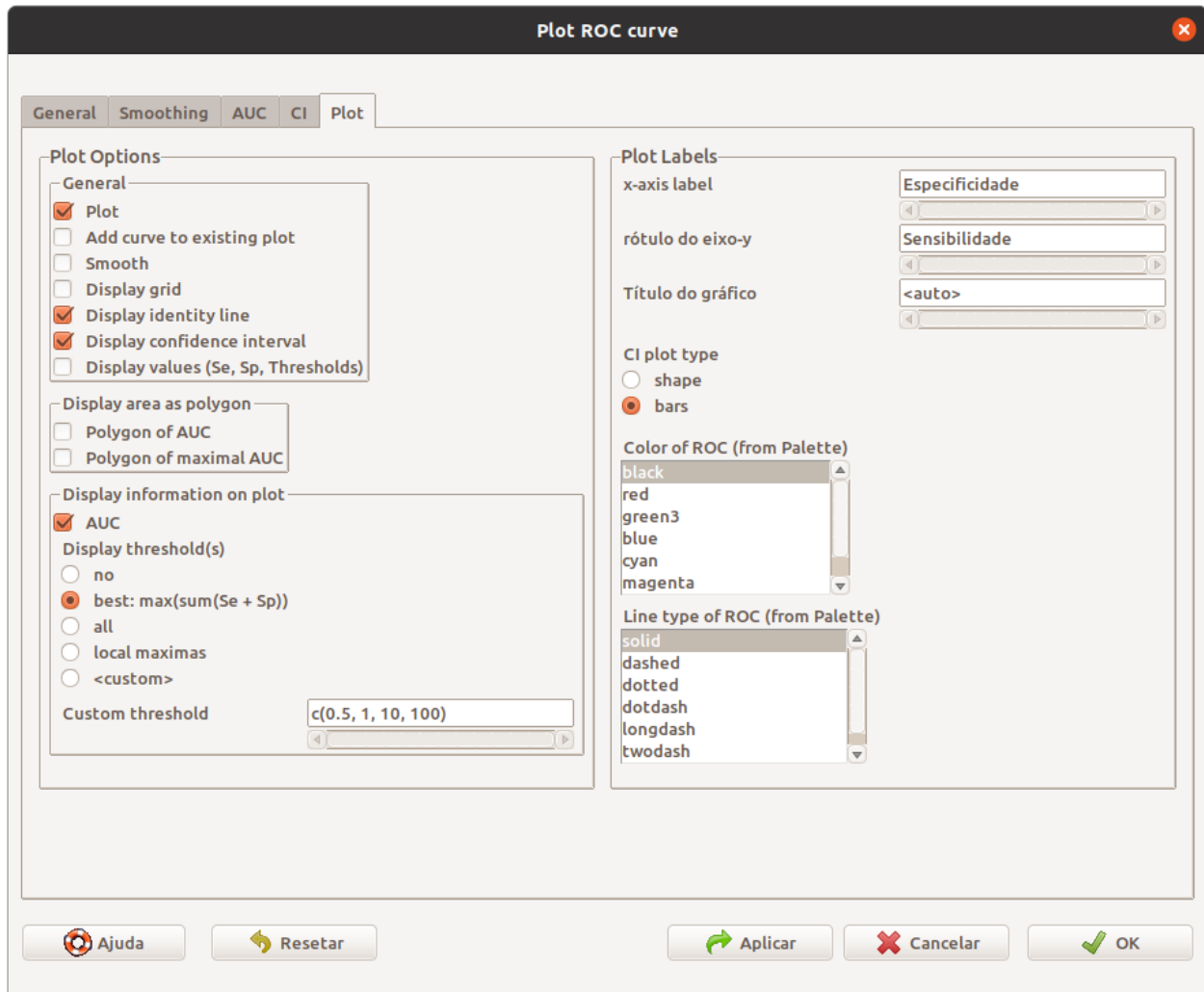


Figura 12.30: Diálogo para configurar a curva ROC: configuração do que será plotado.

Ao pressionarmos o botão OK, o gráfico será exibido (figura 12.31). Observem as barras que delimitam os intervalos de confiança para os valores da sensibilidade. Prestem atenção também que, ao contrário do exibido nas curvas ROC anteriores, o eixo x mostra a especificidade (e não  $1 - \text{especificidade}$ ) e que o valor 1 está no início do eixo X e o valor 0 no final.

A área sob a curva ROC é igual a 0,79, com o intervalo de confiança ao nível de 95% variando de 0,76 a 0,83. A soma da especificidade e sensibilidade é máxima para o valor de glicose igual a 123,50. Nesse ponto de corte, a sensibilidade é igual a 0,73 e a especificidade é igual a 0,71.

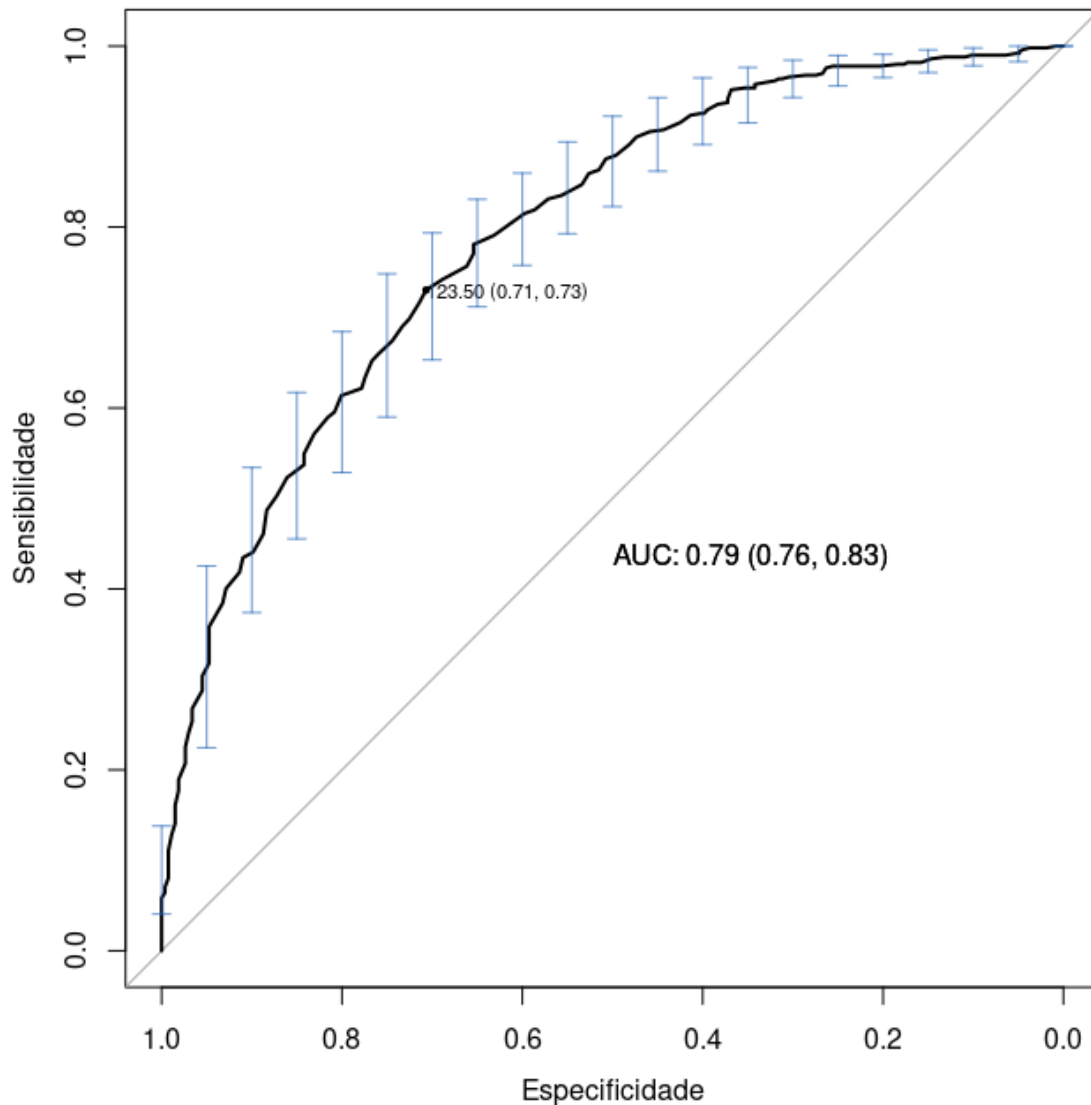


Figura 12.31: Gráfico da curva ROC de diabetes x glicose a partir das configurações estabelecidas nas figuras 12.26 a 12.30.

Ao gerar a curva ROC para o conjunto de dados *PimaIndiansDiabetes2*, o *R Commander* utilizou uma série de comandos. Um deles cria um objeto que será utilizado em outros

comandos. Esse objeto (*roc.obj*) contém uma série de dados para construir a curva ROC. Depois de exibir o gráfico, o *R Commander* remove este objeto, de modo que ele não estará mais acessível. O comando para a criação do objeto é mostrado a seguir:

```
roc.obj <- pROC::roc(diabetes ~ glucose, data=PimaIndiansDiabetes2,
                    na.rm=TRUE, percent=FALSE,
                    direction='auto', partial.auc=FALSE, ci=TRUE,
                    partial.auc.focus='specificity', conf.level=0.95,
                    partial.auc.correct=FALSE, auc=TRUE, plot=FALSE,
                    of='auc', ci.method='bootstrap', boot.n=2000,
                    boot.stratified=FALSE)
```

Vamos executar esse comando novamente e utilizar o objeto *roc.obj* para calcular a sensibilidade para diferentes valores de especificidade. Após a execução do comando, o objeto *roc.obj* estará acessível e o comando abaixo irá calcular os valores de sensibilidade para valores de especificidade entre 0 e 1, com intervalos de 0,05 (figura 12.32):

```
ci(roc.obj, of='se', specificities = seq(0,1,0.05), conf.level = 0.95)
```

Observem que, além dos valores de sensibilidade, os intervalos de confiança também são mostrados.

```
> ci(roc.obj, of='se', specificities = seq(0,1,0.05), conf.level = 0.95)
95% CI (2000 stratified bootstrap replicates):
```

sp	se.low	se.median	se.high
0.00	1.00000	1.00000	1.0000
0.05	0.98240	0.99400	1.0000
0.10	0.97830	0.98990	0.9980
0.15	0.96990	0.98410	0.9960
0.20	0.96580	0.97990	0.9899
0.25	0.95670	0.97590	0.9879
0.30	0.94370	0.96580	0.9823
0.35	0.91540	0.95370	0.9756
0.40	0.89050	0.92780	0.9630
0.45	0.85910	0.90670	0.9393
0.50	0.81890	0.87810	0.9189
0.55	0.79220	0.84080	0.8900
0.60	0.75690	0.81240	0.8566
0.65	0.70880	0.77620	0.8300
0.70	0.65110	0.73130	0.7944
0.75	0.58940	0.67100	0.7441
0.80	0.53220	0.60800	0.6799
0.85	0.45340	0.53440	0.6138
0.90	0.37290	0.44470	0.5338
0.95	0.22130	0.32430	0.4254
1.00	0.04024	0.06439	0.1449

Figura 12.32: Sensibilidades e respectivos intervalos de confiança para diferentes valores de especificidade para glicose x diabetes no conjunto de dados *PimaIndiansDiabetes2*.

O comando abaixo irá calcular os valores de especificidade para valores de sensibilidade entre 0 e 1, com intervalos de 0,05 (figura 12.33):

```
ci(roc.obj, of='sp', sensitivities = seq(0,1,0.05), conf.level = 0.95)
```

```
> ci(roc.obj, of='sp', sensitivities = seq(0,1,0.05), conf.level = 0.95)
95% CI (2000 stratified bootstrap replicates):
  se sp.low sp.median sp.high
0.00 1.0000  1.00000 1.00000
0.05 0.9925  1.00000 1.00000
0.10 0.9808  0.99250 1.00000
0.15 0.9689  0.98500 0.99810
0.20 0.9549  0.97740 0.99490
0.25 0.9419  0.96910 0.98870
0.30 0.9248  0.95490 0.98110
0.35 0.9127  0.94660 0.97370
0.40 0.8801  0.92710 0.96090
0.45 0.8505  0.89700 0.93980
0.50 0.8197  0.87340 0.91730
0.55 0.7862  0.84210 0.89230
0.60 0.7404  0.80780 0.86430
0.65 0.7004  0.76500 0.82520
0.70 0.6583  0.72650 0.78870
0.75 0.6054  0.67780 0.74910
0.80 0.5368  0.61860 0.69550
0.85 0.4649  0.53580 0.61690
0.90 0.3838  0.46420 0.54020
0.95 0.2744  0.35980 0.43000
1.00 0.0000  0.01128 0.06391
```

Figura 12.33: Especificidades e respectivos intervalos de confiança para diferentes valores de sensibilidade para glicose x diabetes no conjunto de dados PimaIndiansDiabetes2.

Se, na figura 12.30, tivéssemos selecionado a opção *shape* em *CI plot type* (figura 12.34), a curva ROC seria exibida como na figura 12.35.

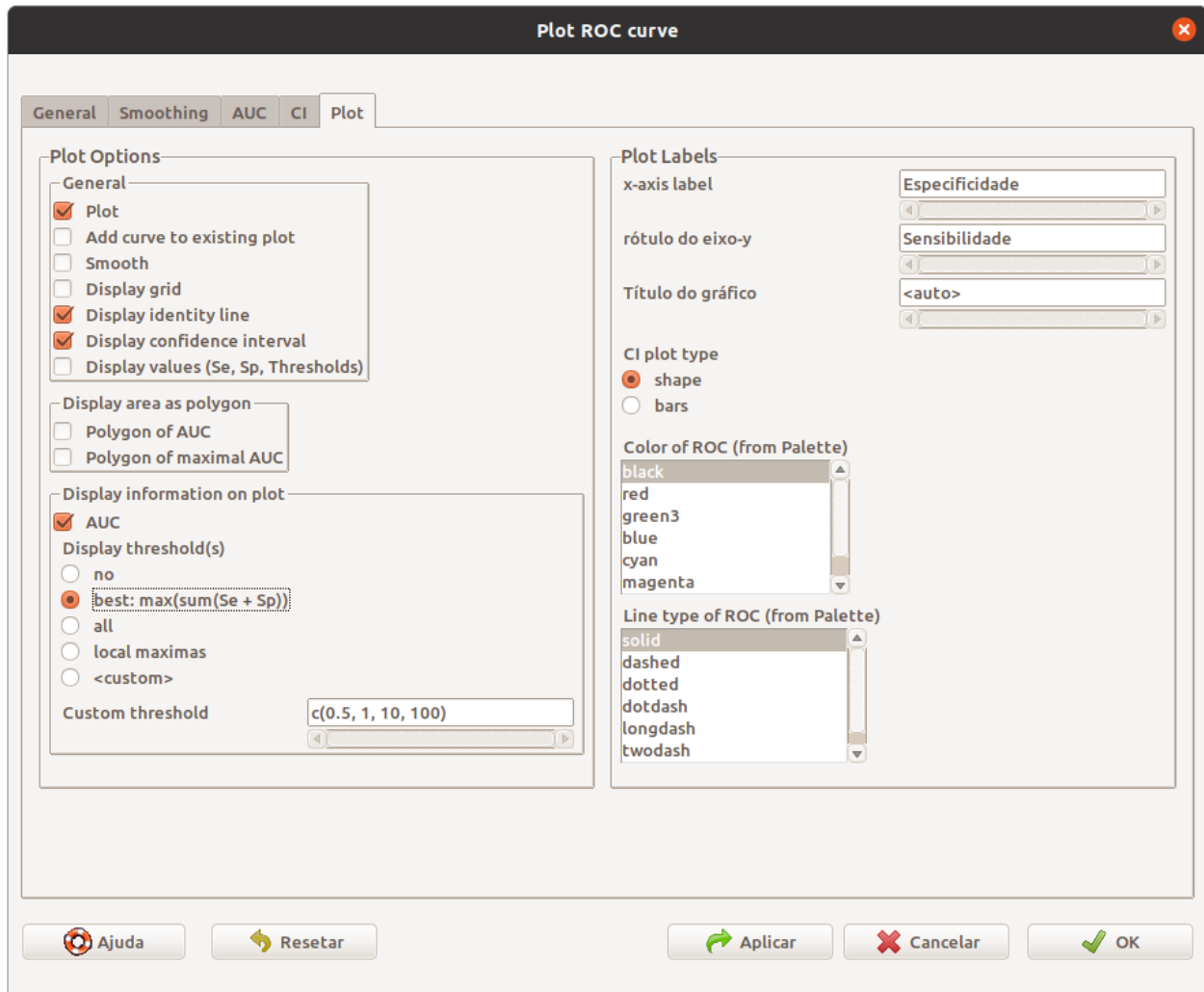


Figura 12.34: Diálogo para configurar a curva ROC: configuração do que será plotado. Agora vamos selecionar a opção *shape* em *CI plot type*.



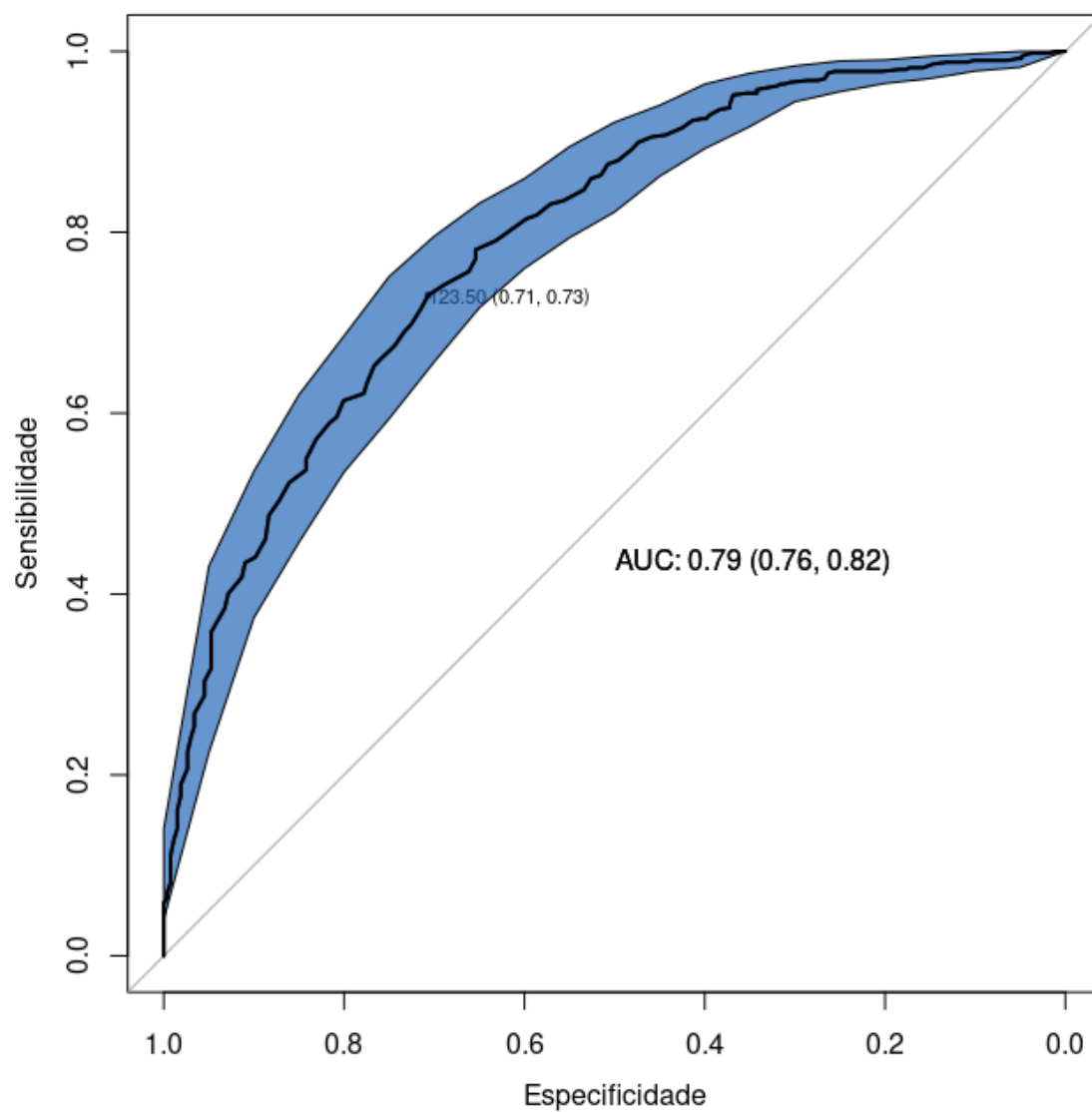


Figura 12.35: Gráfico da curva ROC de diabetes x glicose, dessa vez com os intervalos de confiança desenhados como uma figura sólida.

Para comparar as curvas ROC das variáveis glicemia de 2 horas no teste de tolerância à glicose (*glucose*) e o índice de massa corporal (IMC - *mass*), vamos acessar a seguinte opção no menu do *R Commander*:

ROC  $\Rightarrow$  pROC  $\Rightarrow$  Unpaired ROC curves comparison...

Na aba *General* (figura 12.36), selecionamos as variáveis *glucose* como variável de predição 1, *mass* como variável de predição 2 e *diabetes* como variável de desfecho para as duas curvas, mantendo as demais opções como sugeridas pelo programa.

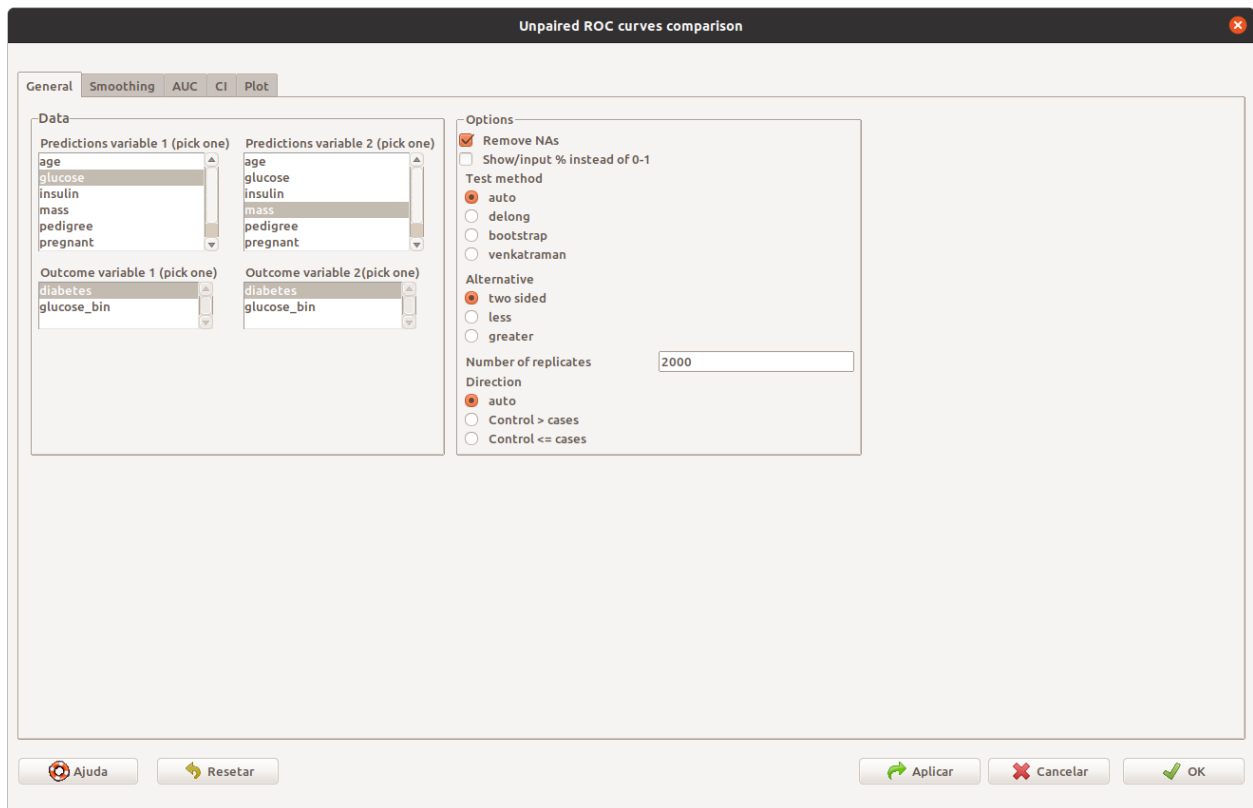


Figura 12.36: Configurações gerais para a construção de duas curvas ROC, relacionando duas variáveis numéricas com uma doença.

Vamos manter as configurações originais nas abas *Smoothing*, *AUC* e *CI*. Na última aba, *Plot*, além das opções que já estão selecionadas, vamos marcar as opções *Display confidence interval*, *Display values (Se, Sp e Thresholds)*, *AUC*, *Test p value* e *best: max(Se + Sp)* (figura 12.37). Ao clicarmos em OK, as duas curvas são mostradas na figura 12.38.

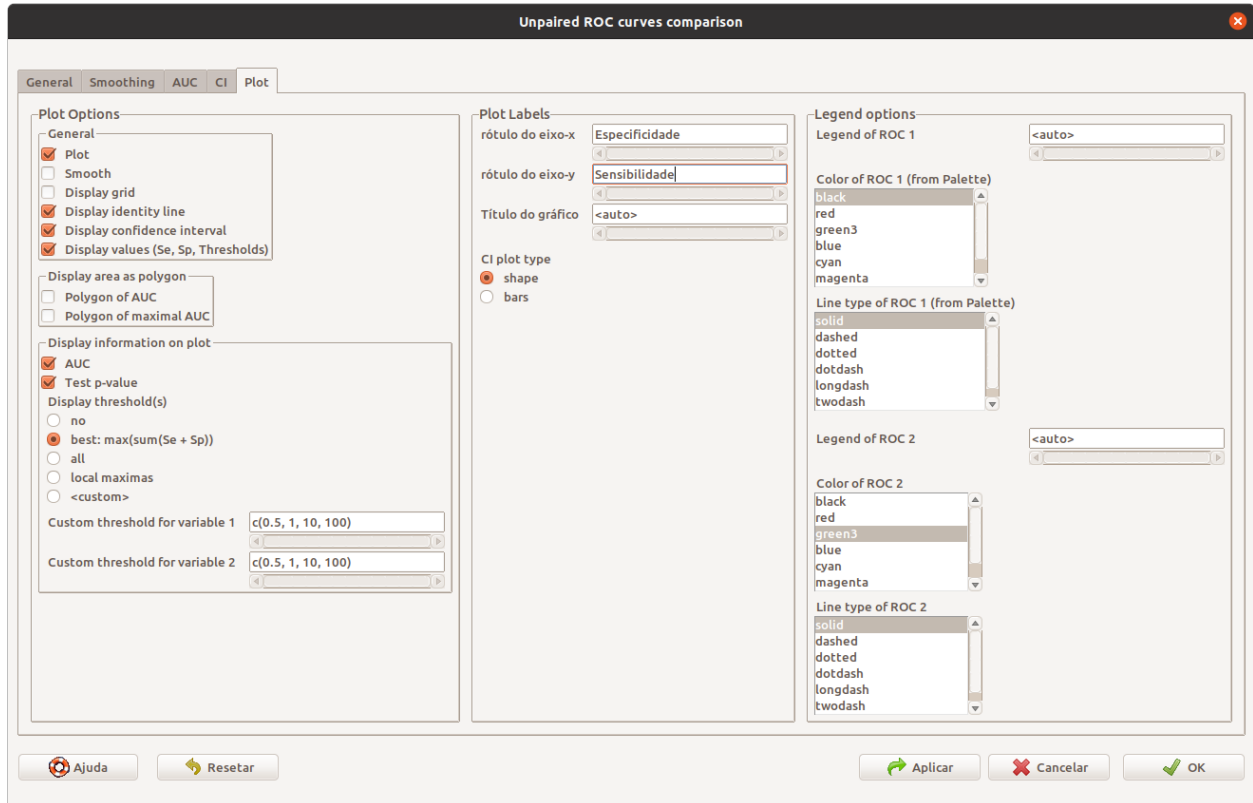


Figura 12.37: Configurações de plotagem para a construção de duas curvas ROC, relacionando duas variáveis numéricas com uma doença.

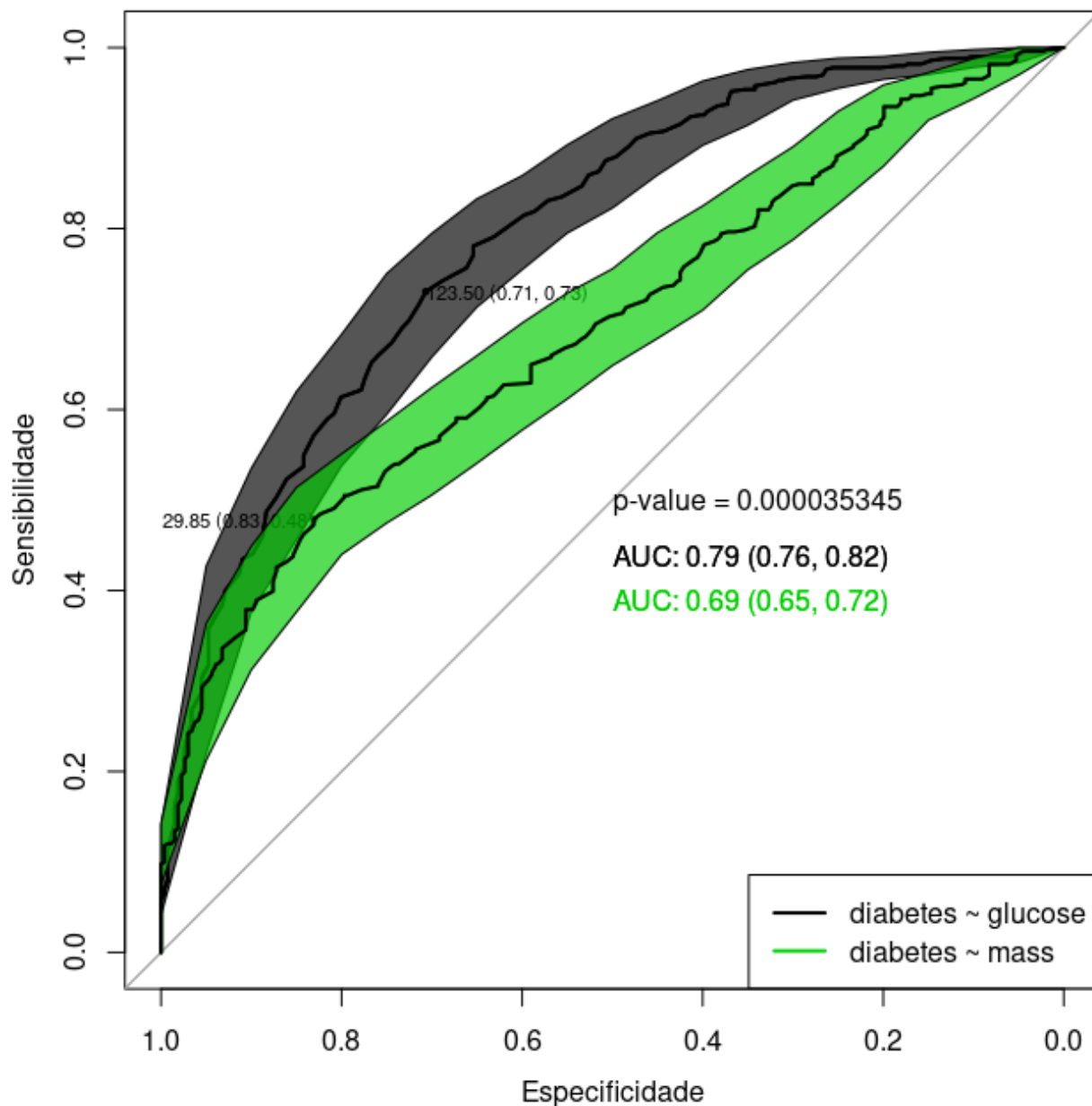


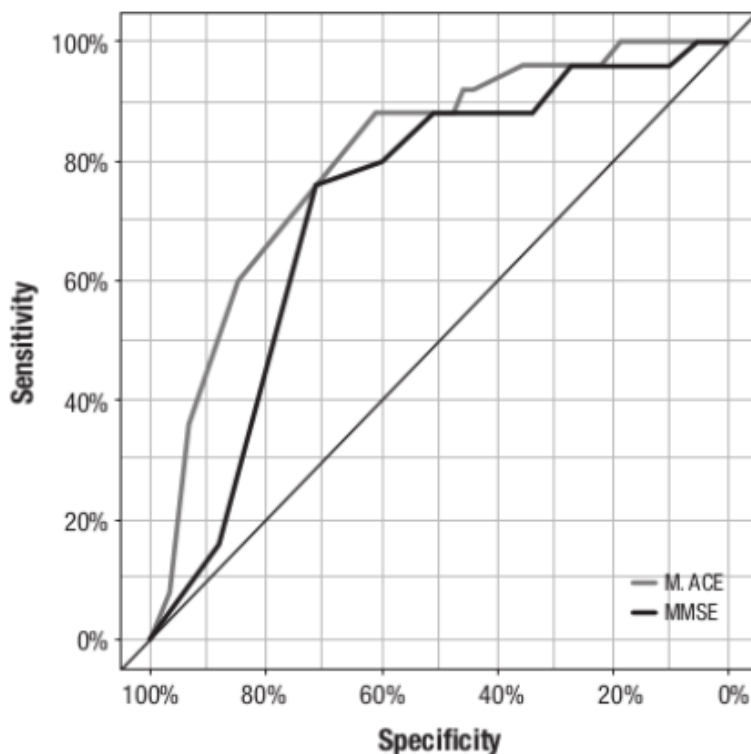
Figura 12.38: Curvas ROC de diabetes x glicose e de diabetes x IMC.

São exibidos os valores da áreas sob a curva e a sensibilidade e a especificidade no ponto onde é maior a soma da especificidade e sensibilidade para cada uma das curvas. Visualmente, vemos que a curva ROC da glicose está mais acima do que a curva ROC do IMC, o que é confirmado pelos valores de AUC para cada curva, os respectivos intervalos de confiança e o valor de  $p$  do teste estatístico que compara as áreas das duas curvas. Nesse conjunto de dados, a glicose é melhor preditora de diabetes mellitus gestacional do que o IMC.

Este [endereço](#) contém diversos exemplos de curvas ROC que podem ser construídas.

## 12.6 Exercícios

- 1) O gráfico abaixo (figura 12.39) foi extraído do artigo: “The Mini-Addenbrooke’s Cognitive Examination (M-ACE) as a brief cognitive screening instrument in Mild Cognitive Impairment and mild Alzheimer’s disease” (Miranda et al., 2018).
  - a) Descreva em linhas gerais como os autores construíram as curvas ROC apresentadas na figura.
  - b) Em cada curva, qual o valor aproximado da sensibilidade correspondente à especificidade igual a 60%? E os respectivos valores da razão de verossimilhança para um teste positivo?
  - c) Julgando somente pela figura, que teste você considera mais acurado para distinguir a doença de Alzheimer do comprometimento cognitivo leve ou sem comprometimento cognitivo? Justifique.
  - d) Que outras informações você considera necessárias para comparar estatisticamente os dois instrumentos?



**Figure 1.** M-ACE and MMSE ROC Curves for differentiating the AD group from the other groups (MCI + Control).

Figura 12.39: Figura 1 do artigo de (Miranda et al., 2018) ([CC BY](#)).

- 2) Quais as vantagens de se usar a razão de verossimilhança em relação à sensibilidade e especificidade?

3) A figura 12.40, mostra a distribuição de valores planimétricos da substância negra entre portadores de Parkinson (PD) e portadores de outras doenças (NPD), respectivamente. Suponha que a linha vermelha seja o ponto de corte ( $\sim 22$ ) utilizado como critério diagnóstico da doença. Abaixo do ponto de corte, a pessoa seria identificada como negativa e acima, positiva. Responda às questões abaixo.

- Mostre no gráfico a área correspondente aos falsos positivos.
- Mostre no gráfico a área correspondente aos falsos negativos.
- Se aumentarmos o ponto de corte para 28, o que acontece com a sensibilidade? E com a especificidade? Qual a implicação disso em termos clínicos?

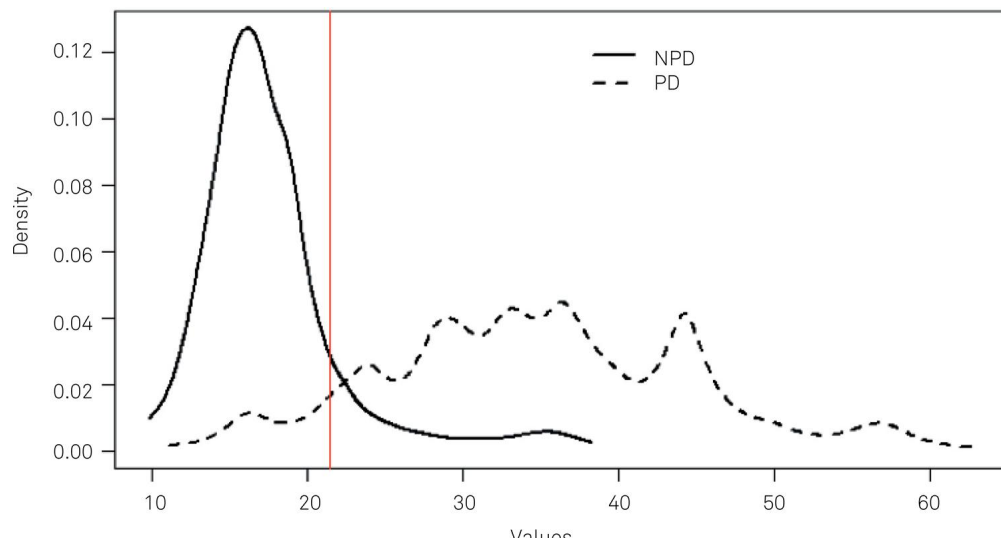


Figura 12.40: Distribuição de valores planimétricos da substância negra entre portadores de Parkinson (PD) e portadores de outras doenças (NPD), respectivamente. Fonte: figura 3 do artigo de (Grippe et al., 2018) ([CC BY](#)).

- Considerando novamente o artigo da questão anterior, a figura 12.41 mostra a tabela 4 desse artigo. Responda às questões abaixo.
  - Use o R para calcular as medidas de sensibilidade, especificidade e as razões de verossimilhança para os resultados positivo e negativo do TCS para o diagnóstico da doença de Parkinson, usando um ponto de corte de 20 mm<sup>2</sup> (números que não estão entre parênteses), e os respectivos intervalos de confiança.
  - Quais são as interpretações dos intervalos de confiança para a especificidade e para a razão de verossimilhança para resultado negativo?
  - O que você tem a comentar sobre a precisão dos intervalos de confiança calculados no item a?
  - Quais os valores preditivos positivo e negativo do teste, supondo uma prevalência da doença de 1%?

**Table 4.** Results of TCS and the definitive clinical diagnoses.

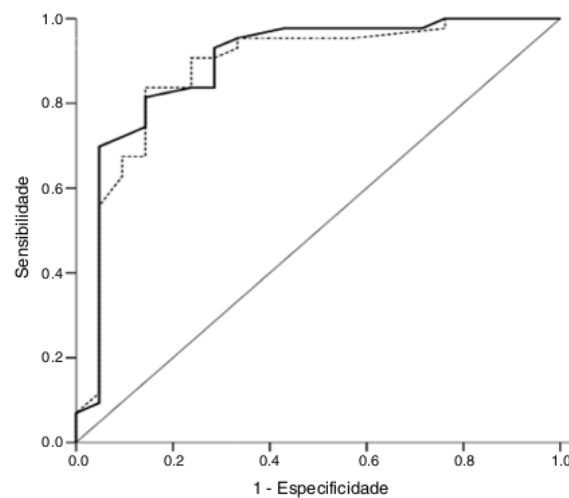
Variable	SN+	SN-
Parkinson's disease	37 (32)	2 (6)
Non-PD	3 (1)	23 (24)

The numbers in parentheses represent data calculated using the alternative cut-off value of 25 mm<sup>2</sup>.

Figura 12.41: Tabela 4 do artigo de (Grippe et al., 2018) (CC BY).

5) Considere o artigo “Diagnóstico rápido da síndrome do desconforto respiratório por aspirado bucal em recém-nascidos prematuros” (Ribeiro et al., 2019). A figura 12.42 abaixo mostra a figura 3 desse artigo. Responda às questões abaixo.

- Comparem na figura as curvas ROC dos testes das microbolhas estáveis no aspirado bucal e no aspirado gástrico para o diagnóstico do desconforto respiratório. É possível apontar um dos aspirados como mais acurado? Considerem também os valores e o intervalo de confiança para as áreas sob a curva na legenda da figura.
- Indiquem na figura os pontos correspondentes às sensibilidades e especificidades indicadas no parágrafo abaixo da figura 3 no artigo.
- Comente sobre a precisão dos intervalos de confiança para a sensibilidade e especificidade apresentados no parágrafo abaixo da figura.



**Figura 3** Curva da característica de operação do receptor da contagem das microbolhas nas amostras de fluido oral (linha contínua) e amostras de fluido gástrico (linha pontilhada) para diagnóstico da síndrome do desconforto respiratório. Área abaixo da curva: TME-AB=0,89 (IC de 95%=0,81-0,97;  $p < 0,001$ ); TME-AG=0,88 (IC de 95%=0,80-0,96; 0,001). TME-AB, teste das microbolhas estáveis no aspirado bucal; TME-AG, teste das microbolhas estáveis no aspirado gástrico.

Figura 12.42: Figura 3 do artigo de (Ribeiro et al., 2019) (CC-BY-NC-ND).

- 6) Carregue o conjunto de dados *elastase* do pacote *GsymPoint* ([GPL-2](#) | [GPL-3](#)) do R.
  - a) Instale o pacote *GsymPoint*.
  - b) Verifique a ajuda do conjunto de dados *elastase*.
  - c) Converta a variável *status* para fator.
  - d) Construa a curva ROC para avaliar a utilidade clínica da determinação da elastase leucocitária no diagnóstico da doença arterial coronariana (DAC).
  - e) Discuta os resultados.



# Capítulo 13

## Estimadores

### 13.1 Introdução

Os conteúdos desta seção, das seções 13.2, 13.2.1 e 13.2.2 podem ser visualizados neste [vídeo](#).

Em capítulos anteriores, foram introduzidos diversos conceitos que serão a partir de agora utilizados para tratar do problema da inferência estatística. O dogma central da inferência estatística é que podemos caracterizar propriedades de uma população de indivíduos a partir de dados colhidos a partir de uma amostra de indivíduos dessa população. A partir de uma amostra, estimativas de parâmetros de uma população, por exemplo a média e a variância, podem ser calculadas. Existem diferentes estimadores para cada parâmetro e algumas propriedades de bons estimadores serão apresentadas. Finalmente será apresentado um teorema fundamental em estatística, denominado teorema do limite central, o qual destaca a importância da distribuição normal para a inferência estatística.

### 13.2 Estimativas de parâmetros populacionais

Três conceitos importantes em inferência estatística são: **parâmetro**, **estatística** e **estimador**. Nos capítulos 10 e 11, foram apresentadas diversas distribuições de probabilidades, tanto para variáveis aleatórias discretas quanto contínuas. A distribuição binomial, por exemplo, é caracterizada pelos parâmetros  $n$  (número de experimentos) e  $p$  (probabilidade de ocorrência do evento de interesse na população). Já a distribuição normal é totalmente determinada pelos parâmetros  $\mu$  (média) e  $\sigma$  (desvio padrão). Conhecendo-se os valores dos parâmetros de uma distribuição, é possível calcular probabilidades de ocorrência de valores (ou faixa de valores) a partir da distribuição de probabilidades.

Os parâmetros de uma distribuição de probabilidades são medidas que descrevem a população. Em geral não conhecemos os parâmetros de uma distribuição de probabilidades já que, para isso, precisaríamos medir a variável descrita pela distribuição de probabilidades em todos os indivíduos da população. Um dos problemas centrais da inferência estatística é justamente o de estimar os parâmetros de uma dada distribuição de probabilidades na população a partir

de amostras extraídas da população referida. Assim, a partir de uma amostra de pacientes de uma população, poderíamos estimar a probabilidade de um dado evento ocorrer ou a média de uma certa variável. Expressões de cálculo que geram números a partir dos elementos de uma amostra são chamadas de estatísticas. As estimativas de um parâmetro obtidas a partir de estatísticas, em geral, não coincidem com o parâmetro correspondente na população. Métodos estatísticos foram desenvolvidos para caracterizar a precisão dessas estimativas. O restante do capítulo irá apresentar diversos estimadores utilizados para estimar alguns parâmetros da população e as propriedades desses estimadores.

Resumindo, podemos assim definir os conceitos de parâmetro e estatística.

**Parâmetro:** é um número ou medida usada para descrever a população, como, por exemplo, a percentagem ou proporção de indivíduos com colesterol acima de 250 mg/dl na população do Rio de Janeiro, ou a média da estatura de mulheres adultas brasileiras.

**Estatística:** é um número que pode ser calculado a partir dos dados de uma amostra, como, por exemplo, a média da amostra. Uma regra para calcular uma estatística que representa uma estimativa de um determinado parâmetro é chamada de **estimador**.

Um exemplo de estimador seria o uso da média aritmética de uma amostra de valores de uma população para estimar a média dessa população.

### 13.2.1 Amostras de uma distribuição de probabilidades

Vamos considerar uma variável aleatória contínua  $X$ , com uma função densidade de probabilidade normal com média 20 e variância 16 (desvio padrão = 4). Vamos a seguir extrair 15 amostras de tamanho 10 (10 observações por amostra) dessa distribuição. A figura 13.1 mostra como acessar a caixa de diálogo no *R Commander* (figura 13.2) para obter essas amostras.

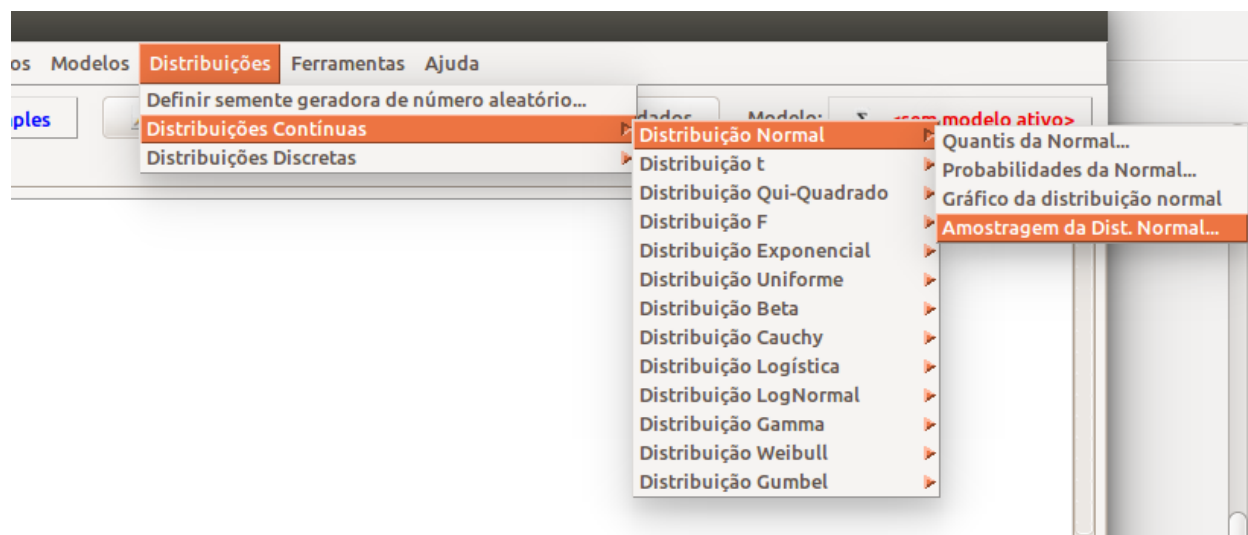


Figura 13.1: Menu para obter amostras de uma distribuição normal no *R Commander*.

Na figura 13.2, demos um nome para o conjunto de dados contendo as 15 amostras que serão geradas, configuramos os parâmetros da distribuição normal (média e desvio padrão) o número de amostras (impropriamente traduzido como tamanho da amostra) e o número de observações em cada amostra (que deveria ser chamado de tamanho da amostra). Também solicitamos que sejam geradas a média e o desvio padrão de cada amostra.

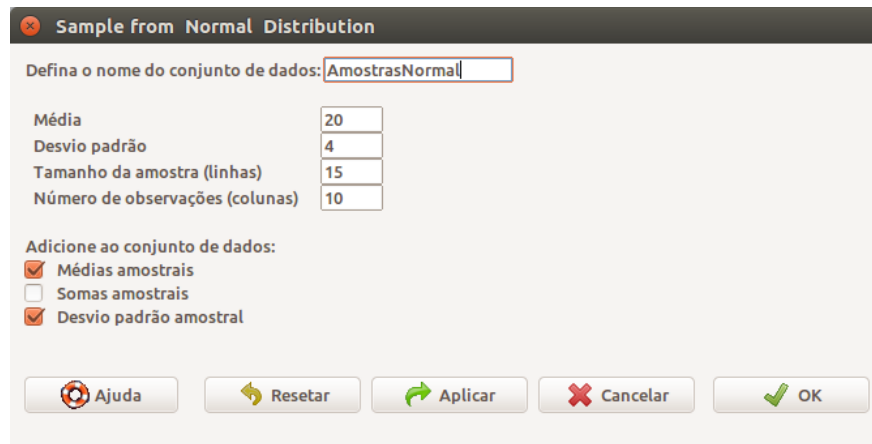


Figura 13.2: Caixa de diálogo para obter amostras de uma distribuição normal. Nesse exemplo, serão obtidas 15 amostras de tamanho 10 (10 observações em cada amostra) de uma distribuição normal com média 20 e variância 16. As amostras serão armazenadas no conjunto de dados *AmostrasNormal*, juntamente com a média e o desvio padrão de cada amostra.

Ao pressionarmos o botão OK na caixa de diálogo da figura 13.2, um conjunto de dados com as amostras geradas, chamado *AmostrasNormal*, é criado (figura 13.3) e basta clicar no botão *Ver conjunto de dados* para visualizá-lo. A figura 13.4 mostra as amostras geradas. Como essas amostras são aleatórias, o leitor obterá valores diferentes dos apresentados na figura 13.4.

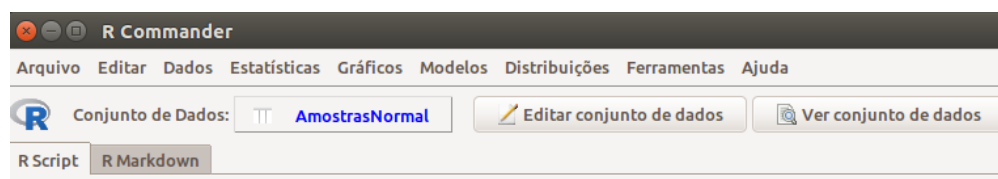


Figura 13.3: Após pressionarmos OK na caixa de diálogo da figura 13.2, um conjunto de dados (*AmostrasNormal*) foi gerado. Para visualizá-lo, basta clicarmos no botão *Ver conjunto de dados*.

Podemos observar na figura 13.4 que a média amostral (e também o desvio padrão) varia de amostra para amostra e, em geral, não coincide com o valor da média da população.

	obs1	obs2	obs3	obs4	obs5	obs6	obs7	obs8	obs9	obs10	sd	mean
sample1	18.18125	26.19573	20.696860	17.670774	17.22056	27.84614	18.75158	27.38635	20.143268	13.01106	4.908376	20.71036
sample2	16.32560	16.57861	21.760398	20.052237	21.35538	17.12521	17.54730	23.44352	24.915382	20.44377	2.992980	19.95474
sample3	28.51721	26.08999	21.207988	28.570355	18.50937	16.62246	24.59718	23.07250	17.531991	21.64515	4.320954	22.63642
sample4	10.81846	21.35367	17.890257	24.796291	17.65164	21.06621	23.44509	24.12428	22.747004	23.47145	4.259527	20.73644
sample5	25.43547	17.51263	24.857576	25.967563	14.14200	15.31555	17.08619	26.44109	20.502373	19.73900	4.673787	20.69994
sample6	22.92231	29.98169	22.812792	26.761180	19.59473	20.78284	18.41269	20.05693	22.299517	22.90898	3.47124	22.65337
sample7	25.02665	21.86488	9.590753	17.127468	12.22573	18.18095	17.91916	21.64872	21.122450	14.25070	4.786331	17.89575
sample8	15.70986	21.48653	23.359438	22.234656	13.20457	18.42736	23.22930	16.71883	15.253777	26.98952	4.434380	19.66138
sample9	15.58773	18.85445	19.773626	21.359721	21.17986	24.58394	24.25409	16.50082	19.722430	23.97753	3.126361	20.57942
sample10	18.65597	19.69001	18.075183	17.569331	25.88544	19.97539	15.04892	23.78814	20.859925	17.55687	3.177842	19.71052
sample11	26.90853	20.14469	18.873196	9.748434	11.39412	12.85149	18.81698	11.20215	21.352296	23.87233	5.899334	17.51642
sample12	10.99195	22.55683	23.171864	14.244268	25.28254	21.58282	17.74235	25.29660	23.032123	26.30354	5.087531	21.02049
sample13	17.26916	31.15118	12.929115	21.869660	17.89211	22.01592	25.38979	24.12667	17.934743	18.42160	5.148274	20.89999
sample14	13.01017	15.99886	17.713598	17.007531	17.24854	17.39457	20.83409	21.23699	13.789049	20.26100	2.773375	17.44944
sample15	19.29366	24.45829	18.022634	15.614837	22.25804	28.41910	18.83871	23.22326	9.130848	23.88590	5.41071	20.31453

Figura 13.4: 15 amostras de tamanho 10 da distribuição normal  $N(20, 16)$ . As duas últimas colunas mostram o desvio padrão e a média de cada amostra (em vermelho).

Vamos supor agora que não conhecemos os parâmetros da distribuição da variável aleatória  $X$  na população (média e variância) e somente temos as amostras para estimar os valores desses parâmetros. Vamos inicialmente considerar a média da população. **Como poderemos estimá-la a partir de uma amostra de tamanho  $n$ ?**

A média da população pode ser estimada de diversas maneiras. Vamos chamar de  $\hat{\mu}$  o estimador da média da população  $\mu$ . Abaixo serão apresentados dois possíveis estimadores da média da população.

- 1) um estimador natural é a média aritmética da amostra,  $\bar{X}$ , chamada de média amostral:

$$\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^n X_i}{n} \quad (13.1)$$

- 2) pode-se também estimar a média da população, tomando-se o primeiro elemento da amostra (ou qualquer outro elemento):

$$\hat{\mu} = X_1 \quad (13.2)$$

Já que podemos ter diversos estimadores para a média da população, que critérios devemos utilizar para escolher um deles. A seguir, serão apresentadas algumas propriedades de estimadores que, em muitos casos, permitem caracterizar quando um estimador é melhor que outro.

## 13.2.2 Propriedades de estimadores

Vamos chamar de  $\theta$  um parâmetro de uma distribuição de probabilidades e  $\hat{\theta}$  um estimador para esse parâmetro.  $\hat{\theta}$  é uma função dos valores de uma amostra. Diversos critérios para avaliar estimadores estão disponíveis na literatura científica: **não tendenciosidade, consistência, mínima variância, suficiência**, etc. Vamos abordar a seguir três desses critérios.

### 13.2.2.1 Estimadores não tendenciosos

Um estimador é considerado não tendencioso se o seu valor esperado é igual ao parâmetro estimado, ou seja,  $E[\hat{\theta}] = \theta$

No caso de estimadores da média de uma distribuição de probabilidades, podemos verificar facilmente que a média aritmética amostral e o primeiro elemento da amostra são estimadores não tendenciosos.

Valor esperado da **média amostral**:

$$E[\hat{\mu}] = E[\bar{X}] = E\left[\frac{\sum_{i=1}^n X_i}{n}\right] = \frac{\sum_{i=1}^n E[X_i]}{n} = \frac{\sum_{i=1}^n \mu}{n} = \frac{n\mu}{n}$$

$$E[\bar{X}] = \mu \tag{13.3}$$

Valor esperado do **primeiro elemento da amostra**:

$$E[\hat{\mu}] = E[X_1] = \mu$$

A aplicação [Propriedades de estimadores da média de uma distribuição normal](#) permite a visualização de propriedades dos estimadores aqui discutidos. A figura 13.5 mostra a tela de entrada dessa aplicação.

### Propriedades de estimadores da média de uma distribuição normal

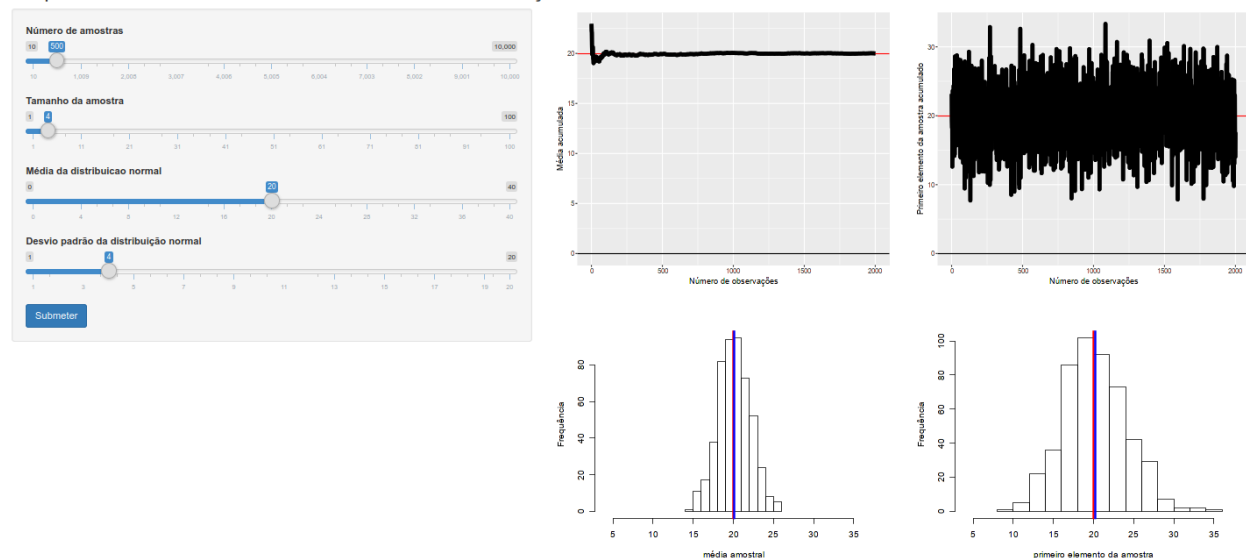


Figura 13.5: Aplicação que permite a visualização de como dois estimadores da média se comportam à medida que o tamanho da amostra aumenta. As amostras foram extraídas de uma distribuição normal.

Os dois estimadores para a média de uma distribuição normal apresentados acima (média amostral e primeiro elemento da amostra) são calculados para um certo número de amostras, sendo cada amostra de um tamanho especificado. A aplicação permite variar os parâmetros da distribuição normal, o número de amostras ( $n\_amostras$ ) e o tamanho de cada amostra ( $tamanho\_amostra$ ). Após selecionarmos os valores dos parâmetros, clicamos no botão *Submeter*.

No gráfico superior à esquerda, o estimador da média da população baseado na média aritmética é calculado para amostras aleatórias de tamanho 1, 2 e assim sucessivamente até uma amostra com tamanho igual a  $n\_amostras \times tamanho\_amostra$ ; nesse caso,  $500 \times 4 = 2000$ .

No gráfico superior à direita, o estimador da média da população baseado somente no primeiro elemento da amostra é obtido para amostras aleatórias de tamanho 1, 2 e assim sucessivamente até uma amostra com tamanho igual a  $n\_amostras \times tamanho\_amostra$ .

Os dois gráficos na parte inferior mostram os histogramas das médias aritméticas, à esquerda, e dos primeiros elementos da amostra para as  $n\_amostras$  de tamanho  $tamanho\_amostra$  extraídas da distribuição normal com os parâmetros da distribuição normal escolhidos.

Os histogramas na parte inferior mostram que cada estimador possui uma certa variabilidade nos valores calculados para a média da distribuição.

Entretanto o ponto médio dos dois histogramas, indicado pelas linhas verticais azuis, coincidem ou estão bastante próximos da média da distribuição (linha vertical vermelha), ilustrando a não tendenciosidade dos estimadores, ou seja, a média dos valores amostrais dos estimadores converge para a média da distribuição de probabilidades de onde as amostras foram extraídas,

à medida que o número de amostras aumenta. Isso é sempre verdadeiro tanto para a média amostral quanto para o primeiro elemento da amostra.

### 13.2.2.2 Variância de estimadores

É desejável que os estimadores de um determinado parâmetro da população possuam um valor de variância que seja o mínimo possível, isso porque uma variância baixa significa uma precisão maior da estimativa do que uma variância alta.

Podemos observar na figura 13.5 que o histograma das médias aritméticas possui uma variabilidade menor do que o histograma dos primeiros valores das amostras. Além disso, ao aumentarmos o tamanho de cada amostra (aumentando o valor da variável *tamanho da amostra* na aplicação e clicando em submeter) a variabilidade da média aritmética da amostra vai diminuindo, enquanto que a variabilidade do primeiro valor da amostra não diminui à medida que o tamanho da amostra aumenta.

Pode-se mostrar que a média aritmética da amostra é o estimador de menor variância entre todos os estimadores lineares da média de uma população.

### 13.2.2.3 Estimadores consistentes

Seja um estimador de  $\theta$  baseado em uma amostra de tamanho  $n$ .  $\hat{\theta}$  será um estimador consistente de  $\theta$  se:

$$\lim_{n \rightarrow \infty} P(|\hat{\theta} - \theta| \geq \epsilon) = 0$$

ou seja, à medida que o tamanho da amostra aumenta, o estimador se aproxima cada vez mais do parâmetro da população, no sentido probabilístico, significando que a variância do estimador tende a zero, assim como o valor do estimador tende para o valor do parâmetro estimado, quando o tamanho da amostra aumenta indefinidamente.

Essa propriedade é mostrada nos gráficos da figura 13.5.

No gráfico acima e à esquerda da figura 13.5, vemos que a média aritmética converge para a média da distribuição normal (indicada pela linha horizontal vermelha) à medida que o tamanho da amostra aumenta. Na seção anterior, vimos que variância do estimador baseado na média aritmética diminui à medida que o tamanho da amostra aumenta. Essas duas características fazem com que a média aritmética da amostra seja um estimador **consistente** para a média da população.

No gráfico acima e à direita, o estimador baseado no primeiro valor da amostra oscila em torno da média da distribuição e não converge para a média da distribuição. Além disso, a sua variância não diminui à medida que o tamanho da amostra aumenta. Essas duas características fazem com que o primeiro elemento da amostra seja um estimador **não consistente** para a média da população.

Resumindo, para a média de uma população, a média amostral (média aritmética da amostra) é um estimador não tendencioso, consistente e de mínima variância. O primeiro valor da

amostra (ou o segundo, terceiro, etc) é um estimador não tendencioso, mas não é consistente e a sua variância não se altera à medida que o tamanho da amostra aumenta.

Portanto a média amostral é um melhor estimador para a média da população do que qualquer estimador que seleciona apenas um elemento da amostra como estimador da média.

### 13.2.3 Estimadores da variância de uma população

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

No capítulo 3, ao calcularmos a variância de um conjunto de dados, utilizamos a seguinte fórmula:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (13.4)$$

Vamos considerar agora que temos uma amostra de tamanho  $n$  de uma variável aleatória  $X$  que segue uma certa distribuição de probabilidades na população e desejamos estimar a variância dessa variável aleatória na população a partir da amostra. Intuitivamente, poderíamos considerar o seguinte estimador para a variância:

$$S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (13.5)$$

obtido a partir de (13.4) substituindo-se  $(n-1)$  por  $n$  no denominador.

Sendo  $\sigma^2$  o valor real da variância de  $X$ , pode-se mostrar que o estimador (13.4) é um estimador não tendencioso dessa variância, enquanto que (13.5) é um estimador tendencioso da mesma variância. Mais precisamente:

$$E[S^2] = \sigma^2 \quad e \quad E[S_n^2] = \frac{n-1}{n} \sigma^2 \quad (13.6)$$

Assim, para amostras pequenas, os valores esperados dos dois estimadores podem apresentar valores bastante diferentes. Por exemplo, para uma amostra de tamanho 4:  $E[S_n^2] = 0,75E[S^2]$ .

A figura 13.6 mostra a tela de entrada da aplicação [Estimadores da variância de uma distribuição](#). Ela mostra as propriedades dos estimadores (13.4) e (13.5) para a variância de uma distribuição normal com média e variância, número de amostras e o tamanho de cada amostra selecionados pelo usuário. Após selecionarmos os valores dos parâmetros, clicamos no botão *Submeter*.



### Estimadores da variância de uma distribuição

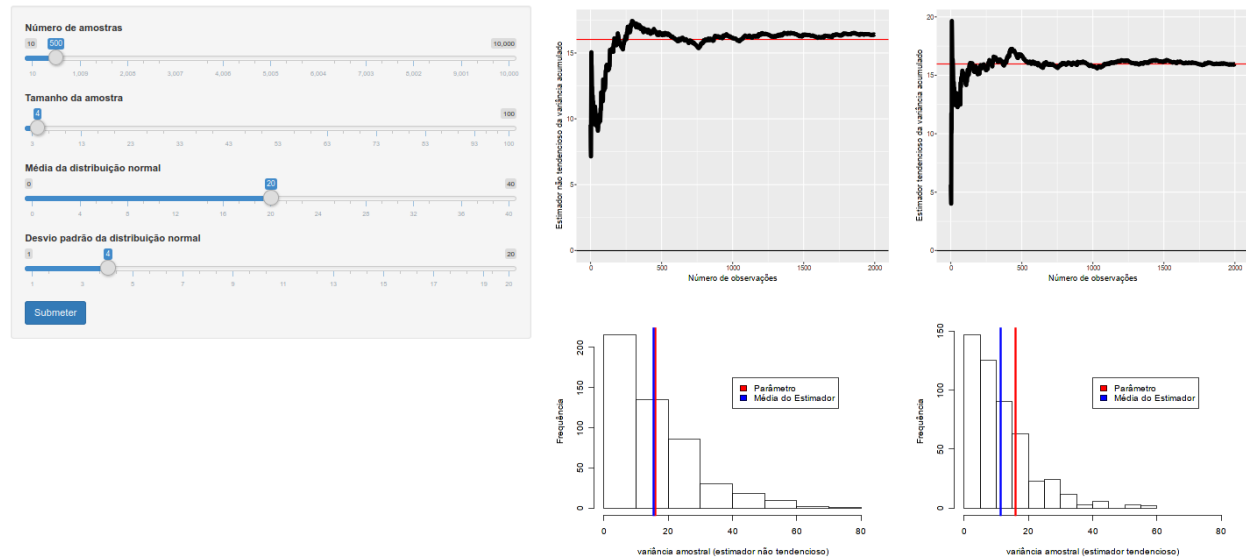


Figura 13.6: Aplicação que permite a visualização de como dois estimadores da variância se comportam à medida que o tamanho da amostra aumenta.

Os gráficos da parte superior da figura 13.6 mostram os valores da variância estimada pelo estimador não tendencioso (à esquerda) e pelo estimador tendencioso (à direita) para amostras aleatórias de tamanho 1, 2 e assim sucessivamente até uma amostra com tamanho igual a  $n\_amostras \times tamanho\_amostra$ ; nesse caso,  $500 \times 4 = 2000$ .

Podemos observar nos dois gráficos da parte superior da figura 13.6 que o valor da variância amostral converge para a variância da distribuição normal (indicada pela linha horizontal vermelha) à medida que o tamanho amostral aumenta, para ambos os estimadores da variância.

Os gráficos da parte inferior mostram os histogramas dos valores dos estimadores (13.4) e (13.5) para a variância amostral para o número de amostras e tamanho de cada amostra escolhidos pelo usuário. Cada histograma foi construído a partir dos 500 valores de variância calculados para cada amostra de tamanho 4 extraídas aleatoriamente da distribuição  $N(20, 16)$ . A linha vertical azul indica a média das variâncias amostrais e a linha vertical vermelha indica o valor da variância da distribuição.

Observamos que a média das variâncias amostrais das 500 amostras é próxima ao valor exato da variância da distribuição para o estimador não tendencioso (histograma à esquerda), mas é bastante diferente do valor exato para o estimador tendencioso (histograma à direita), daí o fato dele ser dito estimador tendencioso da variância.

Ao aumentarmos o tamanho de cada amostra, iremos verificar que a média da variância amostral para o estimador tendencioso vai convergir para o valor exato da variância e que a variabilidade dos valores das variâncias amostrais também irá reduzir.

Assim, para amostras pequenas, os dois estimadores tendem a apresentar valores bastante diferentes, mas, à medida que o tamanho da amostra aumenta, o estimador tendencioso aproxima-se cada vez mais da variância da distribuição.

A variabilidade de ambos os estimadores diminui à medida que o tamanho da amostra aumenta. Assim os dois estimadores são consistentes.

## 13.3 Teorema do limite central

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

Resumindo o que foi visto acima, temos que distinguir duas distribuições de probabilidades. A primeira é a *distribuição da variável aleatória de interesse na população*, que é um modelo teórico para o problema de nosso interesse. A segunda é a *distribuição amostral* que é a distribuição de probabilidades associada a uma estatística obtida a partir de amostras da população.

A média e a variância são duas estatísticas que podemos obter da amostra e, como vimos nas seções anteriores, elas variam de amostra para amostra sendo variáveis aleatórias que possuem uma distribuição associada.

Vamos considerar a média amostral. Vimos na seção 13.2.2.1 que a média amostral é um estimador não tendencioso da média de uma população. Também foi visto na seção 13.2.2.2 que a variância da média amostral diminui à medida que o tamanho da amostra aumenta. Vamos explorar esse tema com mais profundidade.

A aplicação [Teorema do Limite Central](#) (figura 13.7) mostra histogramas da média amostral para diferentes distribuições (normal, uniforme e gama), diferentes parâmetros dessas distribuições e permite variar o tamanho de cada amostra, o número de amostras extraídas da população e o número de classes do histograma. Ao selecionarmos uma distribuição, escolhemos os valores dos parâmetros da distribuição selecionada.

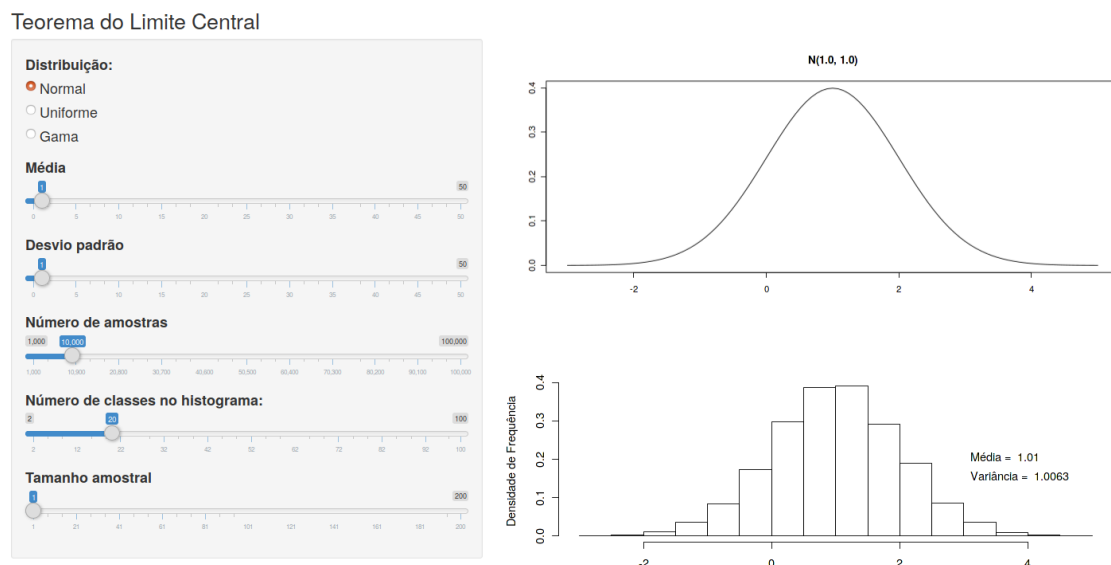


Figura 13.7: Aplicação que mostra o histograma da média amostral para diferentes distribuições de probabilidade e tamanhos amostrais.

Vamos inicialmente considerar uma distribuição normal (média 10 e desvio padrão 1) e variar o tamanho de cada amostra extraída da população e verificar o comportamento da média amostral. Na figura 13.7, fixamos o número de amostras em 10000 e o histograma com 10 classes.

A figura 13.8 mostra três histogramas obtidos a partir de 10000 amostras de tamanhos iguais a 1, 4 e 16, respectivamente. Podemos verificar que as médias das distribuições amostrais são aproximadamente iguais à média da distribuição normal e que as variâncias das médias amostrais são aproximadamente iguais à variância da população (1) dividida pelo tamanho amostral ( $n$ ). Assim a variância da média amostral para  $n = 1$  é aproximadamente 1 ( $1/1$ ), para  $n = 4$ , 0,25 ( $1/4$ ) e para  $n = 16$ , 0,0625 ( $1/16$ ).

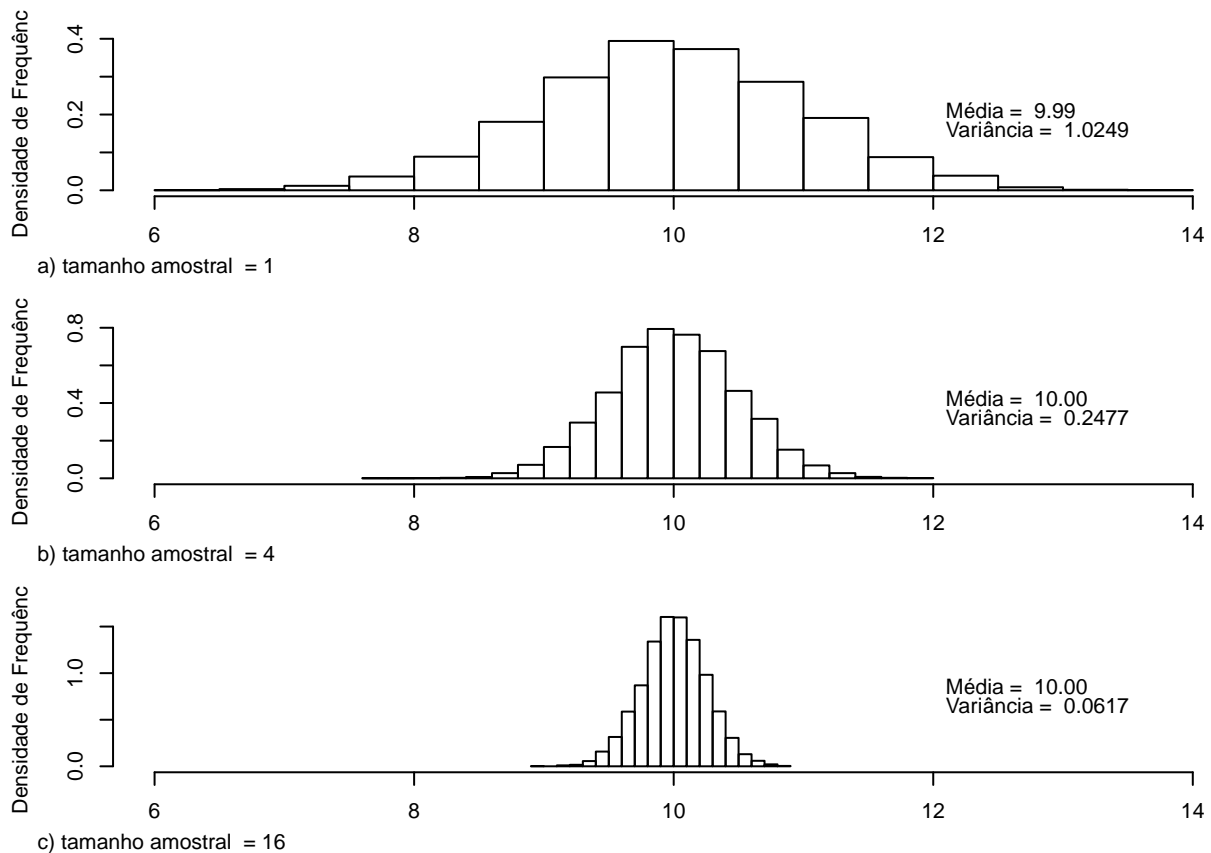


Figura 13.8: Histogramas da média amostral para amostras de tamanho 1, 4 e 16, respectivamente, de uma distribuição normal  $N(10, 1)$ . Observem que as médias das distribuições amostrais são aproximadamente iguais à média da distribuição normal e que as variâncias das médias amostrais são aproximadamente iguais a 1 ( $1/1$ ), 0,25 ( $1/4$ ) e 0,0625 ( $1/16$ ), respectivamente.

Ao selecionarmos uma distribuição uniforme na aplicação da figura 13.7, podemos selecionar os valores mínimo e máximo da distribuição uniforme e repetir o procedimento utilizado para obtermos os histogramas da figura 13.8. A figura 13.9 mostra a distribuição uniforme com valores mínimo e máximo iguais, respectivamente a 5 e 15.

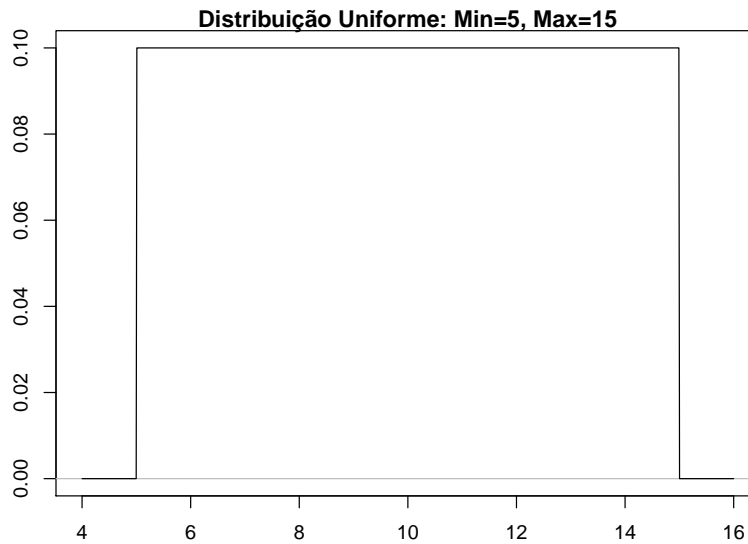


Figura 13.9: Distribuição Uniforme com min = 5,0 e max = 15,0.

A média e variância da distribuição uniforme da figura 13.9 são, respectivamente:

$$\mu = \frac{max - min}{2} = 10$$

$$\sigma^2 = \frac{(max - min)^2}{12} = 8,33$$

A figura 13.10 mostra três histogramas obtidos a partir de 10000 amostras de tamanhos iguais a 1, 4 e 16, respectivamente, da distribuição uniforme da figura 13.9. Novamente, podemos verificar que as médias das distribuições amostrais são aproximadamente iguais à média da distribuição uniforme e que as variâncias das médias amostrais são aproximadamente iguais à variância da população (1) dividida pelo tamanho amostral (n). Assim a variância da média amostral para n = 1 é aproximadamente 8,33 (8,33/1), 2,08 para n = 4 (8,33/4), e 0,52 para n = 16 (8,33/16).

Além disso, pode-se observar que, à medida que o tamanho amostral aumenta, a forma dos histogramas se aproxima cada vez mais da forma de uma distribuição normal.

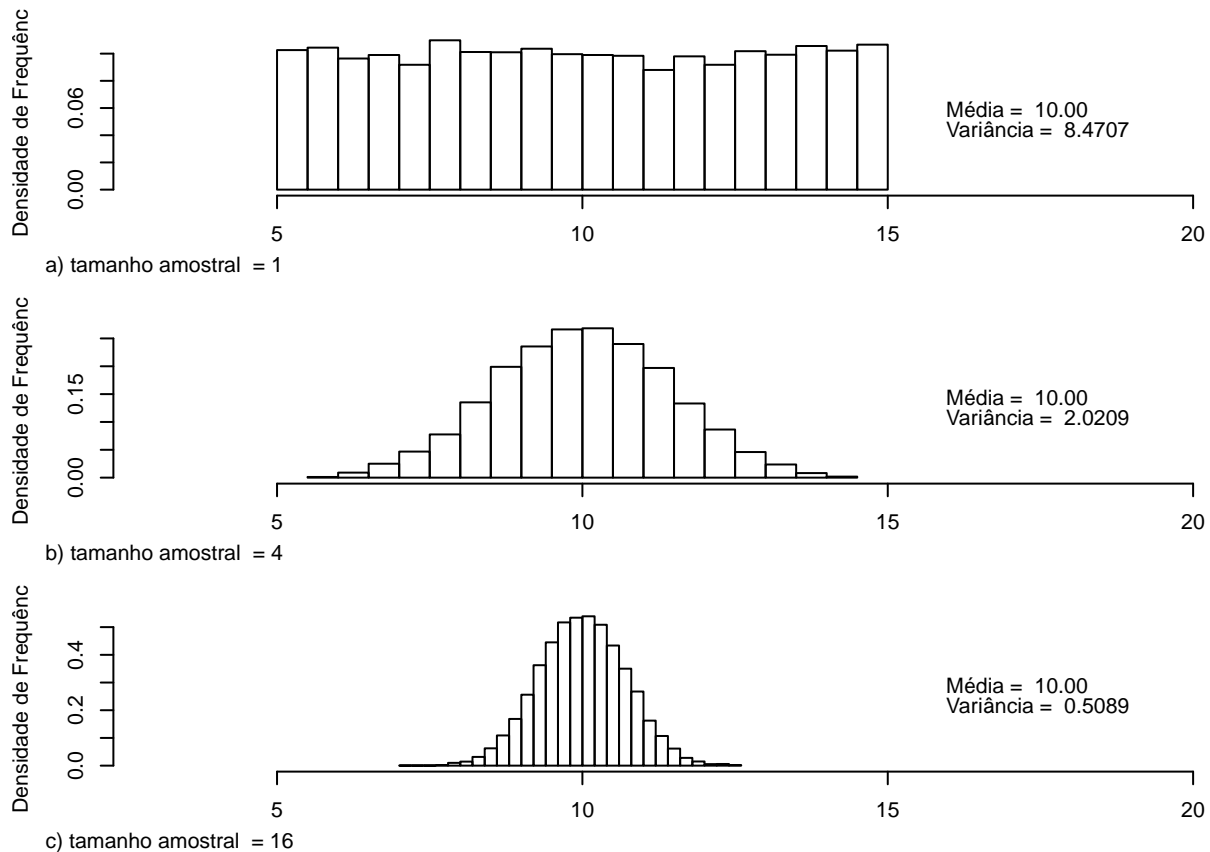


Figura 13.10: Histogramas da média amostral para amostras de tamanho 1, 4 e 16, respectivamente, de uma distribuição uniforme  $U(5, 15)$ . Observem que as médias das distribuições amostrais são aproximadamente iguais à média da distribuição uniforme e que as variâncias das médias amostrais são aproximadamente iguais a 8,33 ( $8,33/1$ ), 2,08 ( $8,33/4$ ) e 0,52 ( $8,33/16$ ), respectivamente.

Vamos selecionar agora a distribuição gama na aplicação da figura 13.7. A distribuição gama é uma distribuição de probabilidades para variáveis contínuas, com dois parâmetros ( $\alpha$  e  $r$ ). Não iremos entrar em detalhes dessa distribuição neste texto, mas é importante salientar que a distribuição gama é bastante utilizada em estatística, da qual diversas outras distribuições são casos especiais, como a distribuição exponencial e a distribuição qui-quadrado. Vamos repetir o procedimento utilizado para obtermos os histogramas da figura 13.8. A figura 13.11 mostra a distribuição gama para os parâmetros  $\alpha = 1$  e  $r = 2$ .

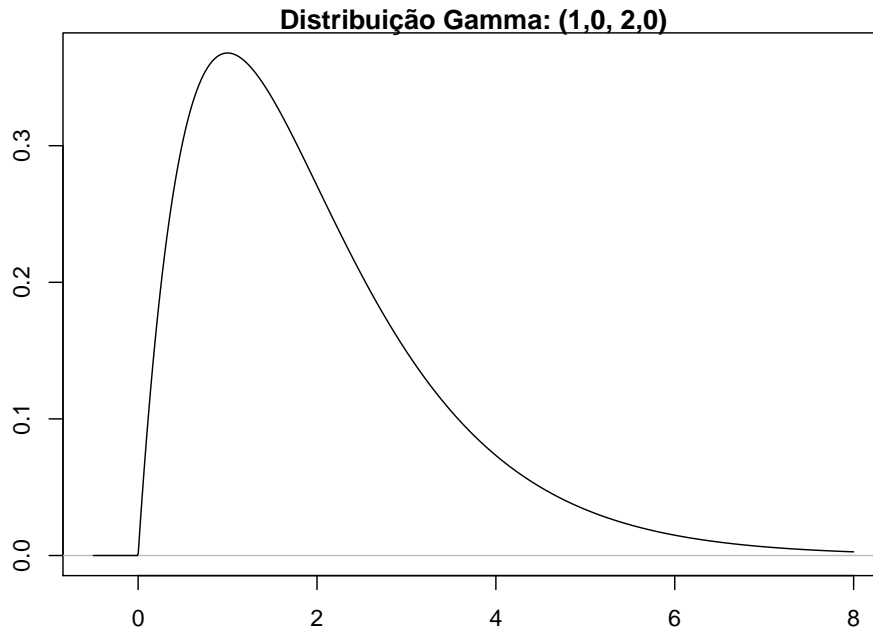


Figura 13.11: Distribuição gama com parâmetros  $\alpha = 1$  e  $r = 2$ .

A média e variância de uma distribuição gama são respectivamente:

$$\mu = \frac{r}{\alpha} = 2$$

$$\sigma^2 = \frac{r}{\alpha^2} = 2$$

A figura 13.12 mostra três histogramas obtidos a partir de 10000 amostras de tamanhos iguais a 1, 4 e 16, respectivamente, da distribuição gama da figura 13.11. Novamente, podemos verificar que as médias das distribuições amostrais são aproximadamente iguais à média da distribuição gama e que as variâncias das médias amostrais são aproximadamente iguais à variância da população (1) dividida pelo tamanho amostral (n). Assim a variância da média amostral para  $n = 1$  é aproximadamente 2 ( $2/1$ ), 0,5 para  $n = 4$  ( $2/4$ ), e 0,125 para  $n = 16$  ( $2/16$ ).

Novamente, pode-se observar que, à medida que o tamanho amostral aumenta, a forma dos histogramas se aproxima cada vez mais de uma distribuição normal.

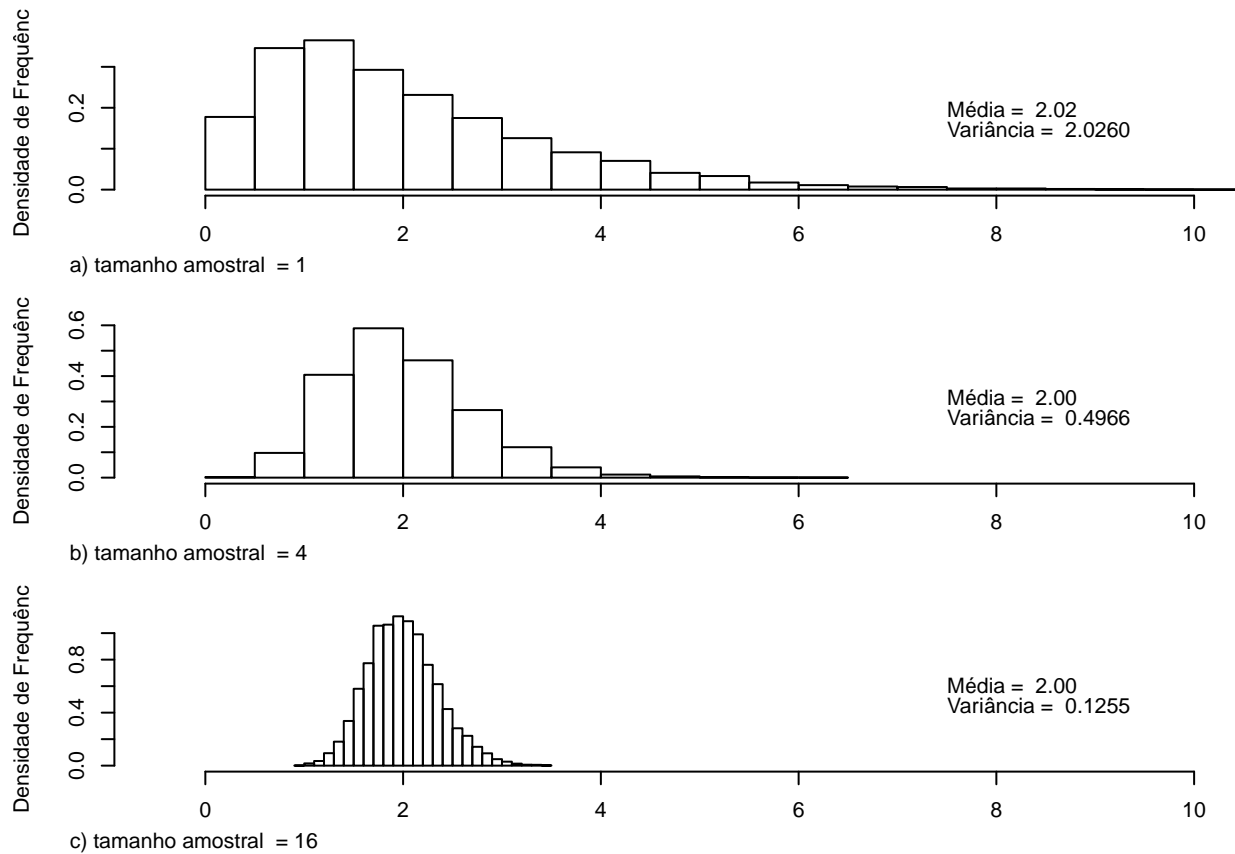


Figura 13.12: Histogramas da média amostral para amostras de tamanho 1, 4 e 16, respectivamente, de uma distribuição gama (1, 2). Observem que as médias das distribuições amostrais são aproximadamente iguais à média da distribuição gama e que as variâncias das médias amostrais são aproximadamente iguais a 2 (2/1), 0,5 (2/4) e 0,125 (2/16), respectivamente.

Esses três exemplos nos leva a intuir que a variância da média amostral de uma variável aleatória  $X$  é igual à variância dessa variável na população dividido pelo tamanho da amostra. De fato, isso pode ser demonstrado facilmente a partir da definição de variância de uma variável aleatória. Assim, sendo  $\sigma^2$  a variância da variável aleatória  $X$ , então a variância da média amostral é calculada por:

$$var(\bar{X}) = var\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} var\left(\sum_{i=1}^n X_i\right)$$

Considerando que cada elemento da amostra possui a mesma distribuição da variável aleatória  $X$  e são independentes uns dos outros, temos que:

$$var(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^n var(X_i) = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 \quad (13.7)$$

$$= \frac{1}{n^2} n \sigma^2 \quad (13.8)$$

$$= \frac{\sigma^2}{n} \quad (13.9)$$

Além disso, os três exemplos também nos induzem a pensar que, à medida que o tamanho amostral aumenta, as distribuições da média amostral tendem para uma distribuição normal com a mesma média da população e variância igual à da população dividida pelo tamanho da amostra. Esse fato é mostrado pelo **Teorema do Limite Central** que é um dos principais teoremas da Estatística. Esse teorema afirma que:

Sejam  $X_1, X_2, \dots, X_n$  **n** variáveis aleatórias independentes, todas com a mesma distribuição. Sejam  $\mu = E[X_i]$  e  $\sigma^2 = var(X_i) < \infty$  a média e a variância comuns. Seja  $S = \sum_{i=1}^n X_i$  a soma das variáveis aleatórias  $X_1, X_2, \dots, X_n$ . Então  $E[S] = n\mu$  e  $var(S) = n\sigma^2$ . Além disso, a distribuição de  $S$  tende para a distribuição normal  $N(n\mu, n\sigma^2)$  à medida que **n** aumenta.

Aplicado à média amostral  $\bar{X} = \frac{S}{n}$ , o teorema do limite central fornece os resultados:

$$E[\bar{X}] = \mu$$

$$var(\bar{X}) = \frac{\sigma^2}{n}$$

O desvio padrão de uma estatística de interesse, no caso a média amostral, é denominada de **erro padrão**. Assim, para a média amostral, o erro padrão é igual  $\frac{\sigma}{\sqrt{n}}$ .

## 13.4 Aproximação pela normal da proporção de eventos

Para valores grandes de **n**, a distribuição de uma variável  $X$  que representa o número de sucessos em **n** experimentos de Bernoulli e da proporção  $\hat{P}$  desses eventos é aproximadamente normal. Esse resultado vem do Teorema do Limite Central. A média e a variância para a distribuição normal aproximada de  $X$  são  $np$  e  $np(1-p)$  respectivamente, sendo  $p$  a probabilidade de ocorrência do evento em um experimento de Bernoulli. Para a proporção amostral, temos que:

$$\hat{P} = \frac{X}{n}$$

Logo:

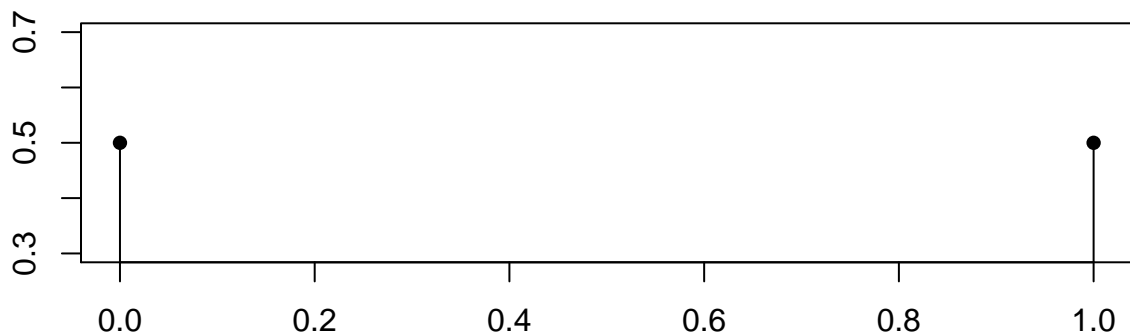


$$E[\hat{P}] = \frac{E[X]}{n} = \frac{np}{n} = p$$

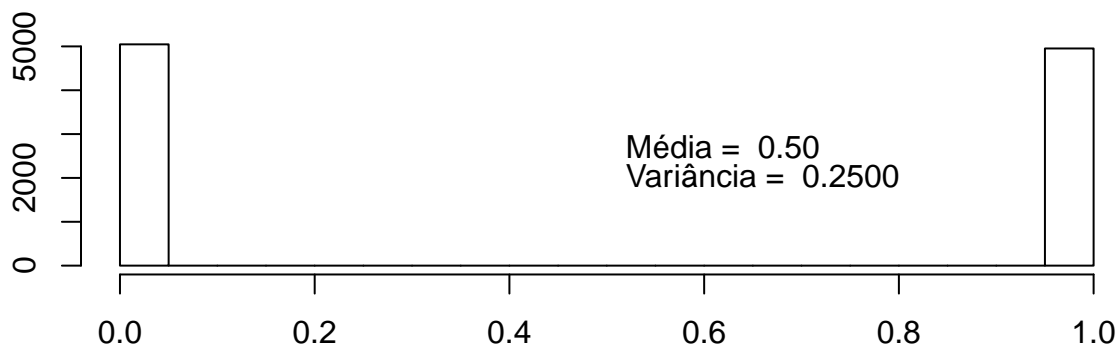
$$\text{var}(\hat{P}) = \frac{\text{var}(X)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

Assim a média e a variância da proporção amostral são então dadas por  $p$  e  $p(1-p)/n$ . A aproximação pela distribuição normal será melhor nos casos em que a média  $np > 10$  e  $np(1-p) > 10$ .

A figura 13.13 mostra uma distribuição binomial (13.13a) e um histograma da proporção de eventos de Bernoulli (13.13b), obtido a partir de 10000 experimentos de Bernoulli. Em cada experimento, a proporção é zero, se o evento não ocorreu, ou 1, se o evento ocorreu. Observem que os valores da média das proporções e da variância são muito próximos dos valores teóricos.



a) Distribuição Binomial (0.5, 1)



b) Histograma da proporção de eventos em 10000 experimentos de Bernoulli

Figura 13.13: a) gráfico da distribuição binomial para  $p = 0,5$  e  $n = 1$ ; b) histograma da proporção de eventos em 1 experimento de Bernoulli ( $p = 0,5$ ). O histograma foi construído a partir de 10000 repetições do experimento de Bernoulli. A média da proporção de eventos foi de 0,5 e a variância 0,25, sendo os valores teóricos iguais a 0,5 e 0,25, respectivamente

A figura 13.14 mostra uma distribuição binomial com  $p = 0,5$  e  $n = 16$  (13.14a) e um histograma da proporção de eventos (13.14b), obtido a partir de 10000 amostras dessa distribuição binomial. Além dos valores da média e variância das proporções serem próximos aos valores teóricos, o histograma já começa a adquirir um formato de uma distribuição normal.

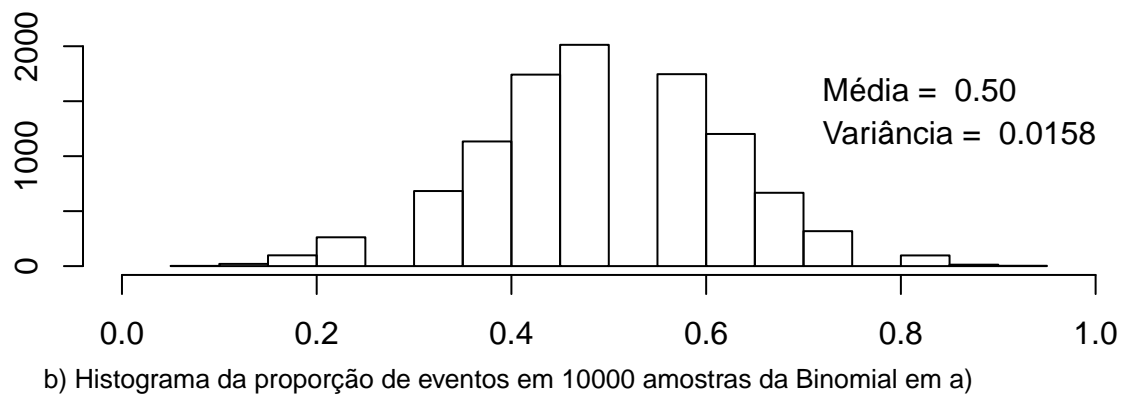
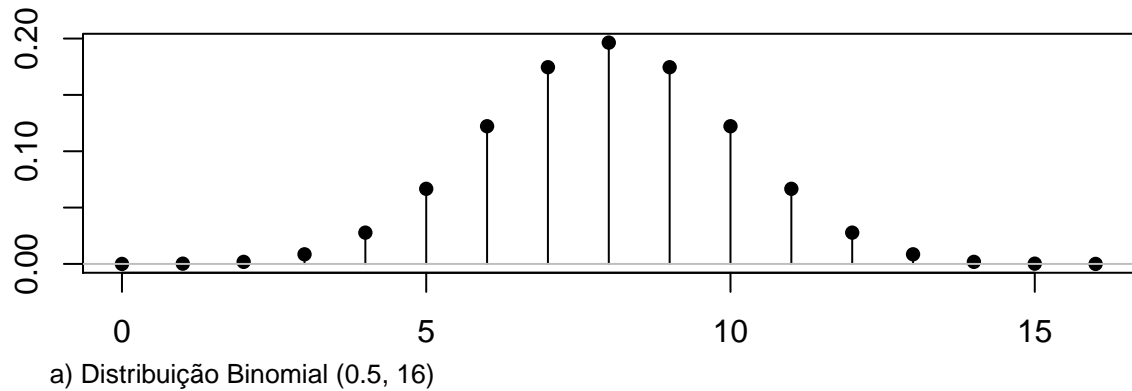
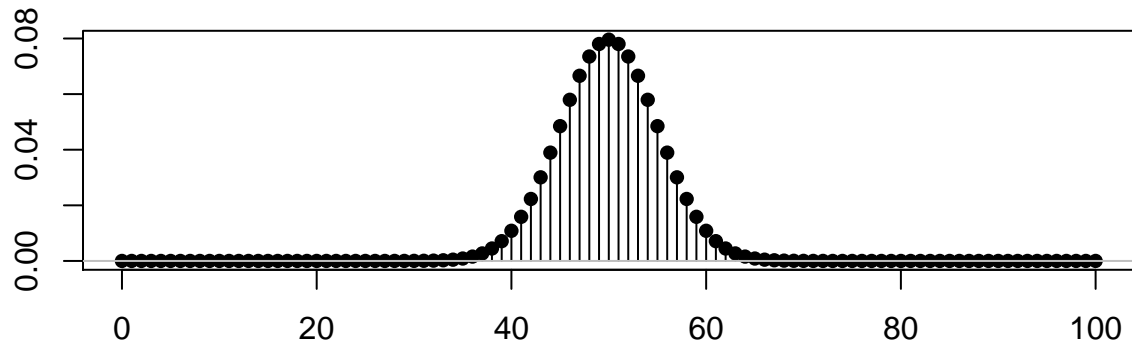
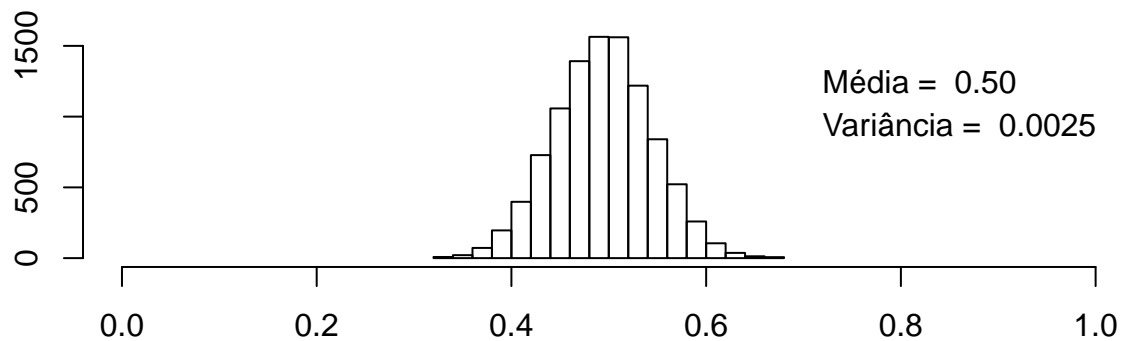


Figura 13.14: a) gráfico da distribuição Binomial para  $p = 0,5$  e  $n = 16$ ; b) histograma da proporção de eventos em 10000 amostras extraídas da distribuição binomial de a). A média da proporção de eventos foi de 0,50 e a variância 0,0158, sendo os valores teóricos iguais a 0,5 e 0,015625, respectivamente.

A figura 13.15 mostra uma distribuição binomial com  $p = 0.5$  e  $n = 100$  (13.15a) e um histograma da proporção de eventos (13.15b), obtido a partir de 10000 amostras dessa distribuição binomial. Além dos valores da média e variância das proporções serem iguais aos valores teóricos, o histograma apresenta um formato de uma distribuição normal.



a) Distribuição Binomial (0.5, 100)



b) Histograma da proporção eventos em 10000 amostras da Binomial em a)

Figura 13.15: a) gráfico da distribuição binomial para  $p = 0,5$  e  $n = 100$ ; b) histograma da proporção de eventos em 10000 amostras extraídas da distribuição binomial de a). A média da proporção de eventos foi de 0,50 e a variância 0,0025, sendo os valores teóricos iguais a 0,5 e 0,0025, respectivamente.

A partir do teorema do limite central, podemos utilizar a distribuição normal como uma aproximação para a distribuição de muitas estatísticas para grandes amostras e, assim, calcular a precisão dessas estatísticas. Assim podemos ter uma estimação dos valores e precisão dos parâmetros de uma distribuição na população a partir de amostras extraídas dessa população, quando não conhecemos esses parâmetros a priori. O próximo capítulo irá mostrar como realizar essas inferências.

## 13.5 Exercícios

- 1) Qual a diferença entre parâmetros de uma distribuição e estatísticas?
- 2) Dê exemplo de dois possíveis estimadores para a média de uma população?
- 3) Há sentido em dizer que um estimador da média possui uma variância? Justifique.
- 4) Se  $X \sim N(\mu, \sigma^2)$ , qual a distribuição da média amostral de  $X$  com  $n$  elementos? Explique porque, intuitivamente, essa distribuição faz sentido.
- 5) Qual o nome que se dá ao desvio padrão da média amostral? Qual a diferença em relação ao desvio padrão da variável?
- 6) Em que casos a média amostral segue uma distribuição normal?
- 7) Qual a importância do teorema do limite central?
- 8) Suponhamos que a variável aleatória  $X$  tenha uma distribuição normal  $N(\mu, \sigma^2)$ , mas não sabemos a variância. Como obter uma estimativa desse parâmetro?
- 9) Dê exemplo de um estimador do parâmetro  $p$  em uma distribuição binomial a partir de uma amostra e da variância desse estimador?

# Capítulo 14

## Intervalo de confiança

### 14.1 Introdução

Os conteúdos desta seção e das seções 14.2 e 14.3 podem ser visualizados neste [vídeo](#).

De acordo com a interpretação frequentista, um intervalo de confiança para um determinado parâmetro da população mostra um intervalo de valores do verdadeiro parâmetro da população compatíveis com os dados da amostra, com um certo nível de confiança, seja o parâmetro a média de uma determinada distribuição, o risco relativo, a razão de chances, a variância, etc. O cálculo do intervalo de confiança é um dos principais resultados de uma análise estatística.

O capítulo 6 mostrou um exemplo de cálculo de um intervalo de confiança para a diferença de médias entre dois grupos, utilizando o método de randomização. Este capítulo aprofundará o conceito de intervalo de confiança e mostrará como calculá-lo em algumas situações onde a distribuição dos dados é conhecida. Inicialmente, iremos mostrar o cálculo e a interpretação de um intervalo de confiança para a média de uma população que segue uma distribuição normal com variância conhecida.

Intervalos de confiança para a média de uma distribuição normal quando a variância da população não é conhecida nos leva à distribuição t de Student. Finalmente serão apresentados intervalos de confiança para a variância de uma distribuição normal e a probabilidade de um evento de Bernoulli.

### 14.2 Intervalo de confiança - IC

Neste capítulo, vamos adotar a seguinte convenção:  $z_p$  significa o valor da variável aleatória  $Z$  que segue a distribuição normal padrão,  $N(0,1)$ , para o qual  $P(z \leq z_p) = p$ , ou seja, a área sob a curva normal padrão à esquerda de  $z_p$  é igual a  $p$ . Assim, para  $p = 0,01$  (1%),  $z_{0,01} = -2,33$ . Como a curva normal padrão é simétrica em torno de 0, a área sob a curva acima de  $z_{1-p}$  é também igual a  $p$  (figura 14.1).

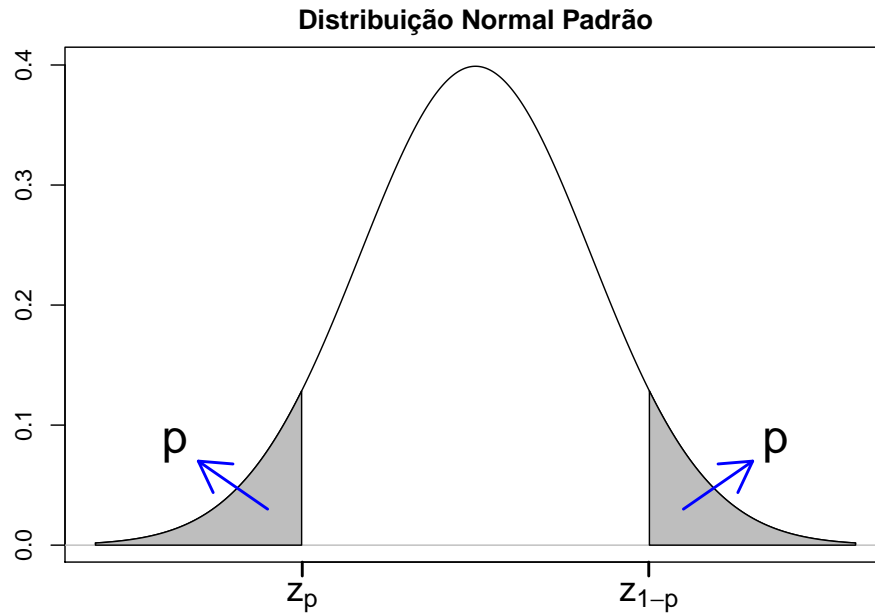


Figura 14.1: Curva normal padrão:  $p$  é a área sob a curva normal à esquerda de  $z_p$  ou à direita de  $z_{1-p}$ .

Para introduzir o conceito de intervalo de confiança de um parâmetro populacional, vamos partir do exemplo a seguir.

**Exemplo 1:** Vamos supor que a glicemia de jejum de uma população de pessoas não diabéticas siga a distribuição normal, com desvio padrão igual a 16, mas não conhecemos o valor da média dessa população. Também vamos supor que extraímos uma amostra aleatória de tamanho 36 da população de pessoas não diabéticas e obtivemos a média amostral da glicemia de jejum igual a 92 mg/dl.

Então:  $\bar{x} = 92$  mg/dl,  $n = 36$  e  $\sigma = 16$  mg/dl.

Como podemos estimar a média da população com um certo nível de confiança estabelecido a priori?

Para uma população com uma distribuição normal padrão, dado um valor  $\alpha$  ( $0 \leq \alpha \leq 1$ ), podemos obter o intervalo  $(z_{\alpha/2}, z_{1-\alpha/2})$  que conterá com probabilidade  $(1 - \alpha)$  o valor de um elemento extraído aleatoriamente dessa população (figura 14.2). Para a distribuição normal padrão,  $z_{\alpha/2} = -z_{1-\alpha/2}$ .

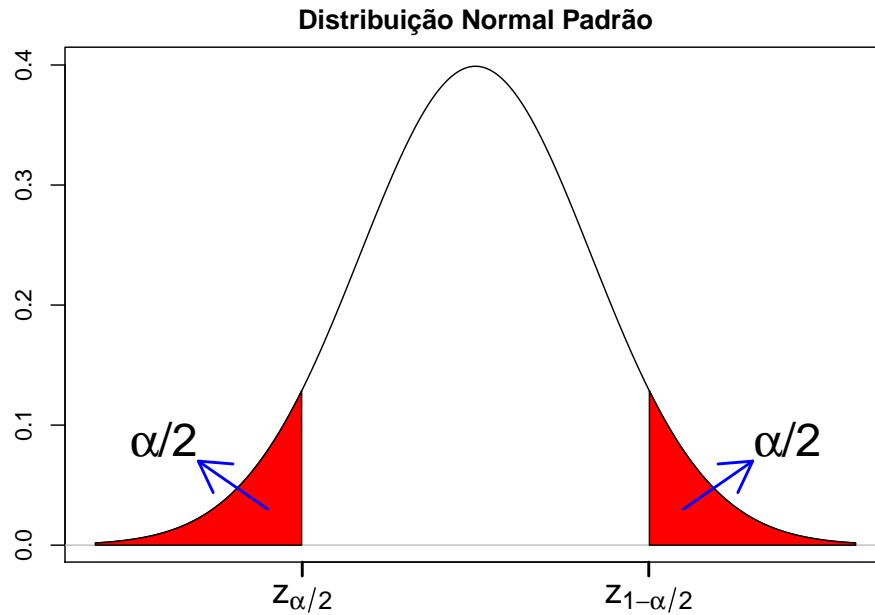


Figura 14.2: Distribuição normal padrão. A área sob a curva compreendida no intervalo  $(z_{\alpha/2}, z_{1-\alpha/2})$  é igual a  $(1 - \alpha)$  e representa a probabilidade de se extrair um elemento da população e obter um valor no intervalo  $(z_{\alpha/2}, z_{1-\alpha/2})$ ,  $z_{\alpha/2} = -z_{1-\alpha/2}$ . Semelhante à figura 14.1, com  $p$  substituído por  $\alpha/2$ .

Usando a função `qnorm` no R, podemos obter os valores de  $z_{1-\alpha/2}$  para qualquer valor de  $\alpha$  entre 0 e 1. Por exemplo, para  $\alpha = 0,05$  (5%), o valor de  $z_{0,975}$  é igual a 1,96:

```
qnorm(0.975, mean = 0, sd = 1, lower.tail=TRUE)
```

```
## [1] 1.959964
```

Outros valores de  $\alpha$  comumente usados são 0,1 (10%) e 0,01 (1%). Os valores correspondentes de  $z_{0,95}$  ( $\alpha = 10\%$ ) e  $z_{0,995}$  ( $\alpha = 1\%$ ) são:

```
qnorm(c(0.95, .995), mean = 0, sd = 1)
```

```
## [1] 1.644854 2.575829
```

A tabela 14.1 resume os três valores de  $\alpha$  e  $z_{1-\alpha/2}$  obtidos acima.

Tabela 14.1: Valores de  $\alpha$  e os correspondentes valores de  $z_{1-\alpha/2}$ .

$\alpha$	$z_{1-\alpha/2}$
1%	2,58
5%	1,96
10%	1,64

Podemos mapear áreas sob o gráfico de uma distribuição normal genérica com média  $\mu$  e desvio padrão  $\sigma$  para áreas sob o gráfico da distribuição normal padrão por meio da expressão:

$$Z = \frac{X - \mu}{\sigma} \quad (14.1)$$

Assim o intervalo  $z_{\alpha/2} \leq Z \leq z_{1-\alpha/2}$  irá ser mapeado para o intervalo

$$z_{\alpha/2} \leq \frac{X - \mu}{\sigma} \leq z_{1-\alpha/2} \quad (14.2)$$

$$\mu + z_{\alpha/2}\sigma \leq X \leq \mu + z_{1-\alpha/2}\sigma \Rightarrow \mu - z_{1-\alpha/2}\sigma \leq X \leq \mu + z_{1-\alpha/2}\sigma \quad (14.3)$$

Portanto a probabilidade de extrairmos um elemento da população com distribuição  $N(\mu, \sigma^2)$  e o valor desse elemento pertencer ao intervalo  $[\mu - z_{1-\alpha/2}\sigma, \mu + z_{1-\alpha/2}\sigma]$  é  $(1 - \alpha)\%$ .

Voltando ao exemplo 1, conforme vimos no capítulo anterior, uma distribuição populacional possui parâmetros os quais podem ser estimados a partir de estatísticas obtidas a partir de amostras. Também vimos que toda estatística amostral tem associada uma *distribuição amostral*. Para uma população que possui uma distribuição normal  $N(\mu, \sigma^2)$  com média  $\mu$  e desvio padrão  $\sigma$ , a distribuição da média amostral para amostras de tamanho  $n$  será  $N(\mu, \sigma^2/n)$ .

Substituindo  $\sigma$  por  $\frac{\sigma}{\sqrt{n}}$  na expressão (14.3), obtemos o intervalo:

$$\mu - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \quad (14.4)$$

**A área sob o gráfico da distribuição da média amostral compreendida no intervalo (14.4) é a probabilidade  $(1 - \alpha)\%$  de extrairmos uma amostra de tamanho  $n$  da população e obtermos uma média amostral dentro desse intervalo.**

Então:

$$P\left(\mu - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = (1 - \alpha)\% \quad (14.5)$$



Como no exemplo 1, e na maioria das situações, não conhecemos a média da população, podemos reescrever a expressão (14.5) como abaixo:

$$P\left(\bar{X} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = (1 - \alpha)\% \quad (14.6)$$

A expressão (14.6) deve ser interpretada da seguinte forma: ela indica que a probabilidade de o intervalo aleatório (14.7) a seguir conter a média  $\mu$  da população é  $(100 - \alpha)\%$ :

$$\left[\bar{X} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right] \quad (14.7)$$

Esse intervalo é o intervalo de confiança ao nível de  $(100 - \alpha)\%$  para a média da população.

### 14.3 Interpretação do intervalo de confiança

O intervalo de confiança  $(100 - \alpha)\%$  para um parâmetro consiste em um intervalo aleatório que possui a propriedade de conter o valor real desse parâmetro com uma probabilidade de  $(100 - \alpha)\%$ . O termo aleatório nessa interpretação indica que, antes de realizarmos a amostragem e calcularmos o intervalo de confiança de acordo com o procedimento apropriado, haverá uma probabilidade de  $(100 - \alpha)\%$  de o intervalo vir a conter o real valor do parâmetro de interesse.

A expressão (14.7) é utilizada na aplicação [Intervalos de confiança](#) (figura 14.3), que nos ajuda a interpretar o intervalo de confiança. A explicação a seguir sobre essa aplicação reproduz o texto do capítulo 6, seção 6.9.

A aplicação [Intervalos de confiança](#) calcula e exibe intervalos de confiança para a média de uma distribuição normal, a partir de um certo número de amostras extraídas dessa distribuição. Os parâmetros da distribuição normal, bem como o nível de confiança, o tamanho de cada amostra e o número de amostras são especificados pelo usuário. O painel principal é atualizado sempre que o usuário pressiona o botão *Reamostrar* (mais intervalos de confiança são exibidos) ou *Limpar* (limpa a tela).

### Intervalos de confiança

**Média da população**  
 50 100

**Desvio padrão da população**  
 10 50

**Nível de confiança:**  
 95%

**Número de amostras:**  
☒ 1 ☐ 10 ☐ 50 ☐ 100

**Tamanho da amostra**

☐ Mostrar média da população

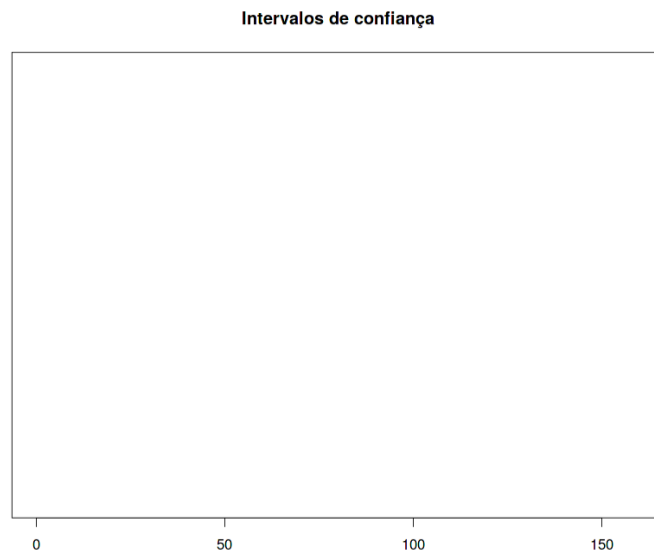


Figura 14.3: Aplicação que calcula e exibe intervalos de confiança para a média de uma distribuição normal calculados a partir de um certo número de amostras extraídas dessa distribuição.

A figura 14.4 exibe intervalos de confiança para 50 amostras de tamanho 10 de uma distribuição normal  $N(80, 400)$ . Para cada amostra, foi calculado o intervalo de confiança ao nível de 95% conforme a expressão (14.7), com  $z_{1-\alpha/2} = 1,96$  e  $\sigma = 20$ . Os 50 intervalos de confiança são exibidos no painel principal da figura.

### Intervalos de confiança

**Média da população**  
 50 100

**Desvio padrão da população**  
 10 50

**Nível de confiança:**  
 95%

**Número de amostras:**  
☐ 1 ☐ 10 ☒ 50 ☐ 100

**Tamanho da amostra**

☐ Mostrar média da população

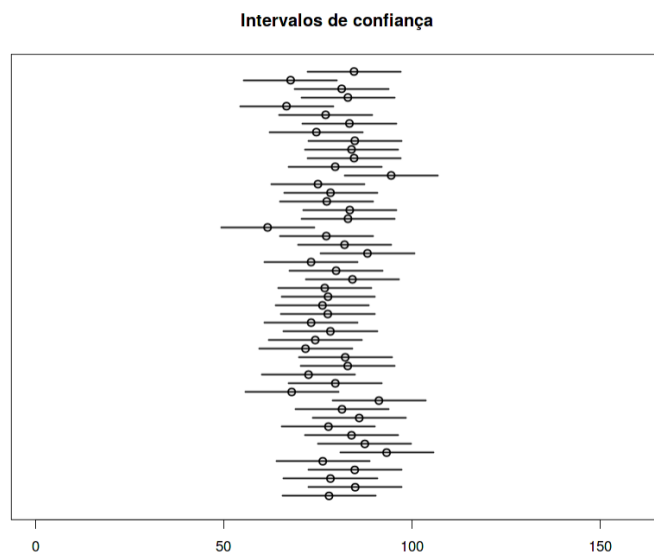


Figura 14.4: Intervalos com 95% de confiança para a média de uma distribuição normal  $N(80, 400)$ , calculados a partir de 50 amostras de tamanho 10.

Ao selecionarmos a opção *Mostrar média da população* na aplicação, uma linha preta, indicando a média real da população, é exibida, e duas linhas verticais em vermelho indicam distâncias iguais a  $z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$  acima e abaixo da média da distribuição (figura 14.5). Para cada intervalo de confiança, o centro com uma marcação representa a média da respectiva amostra. Observe que a maioria dos intervalos de confiança contém a média da distribuição, mas alguns deles (em vermelho) não contêm a média.

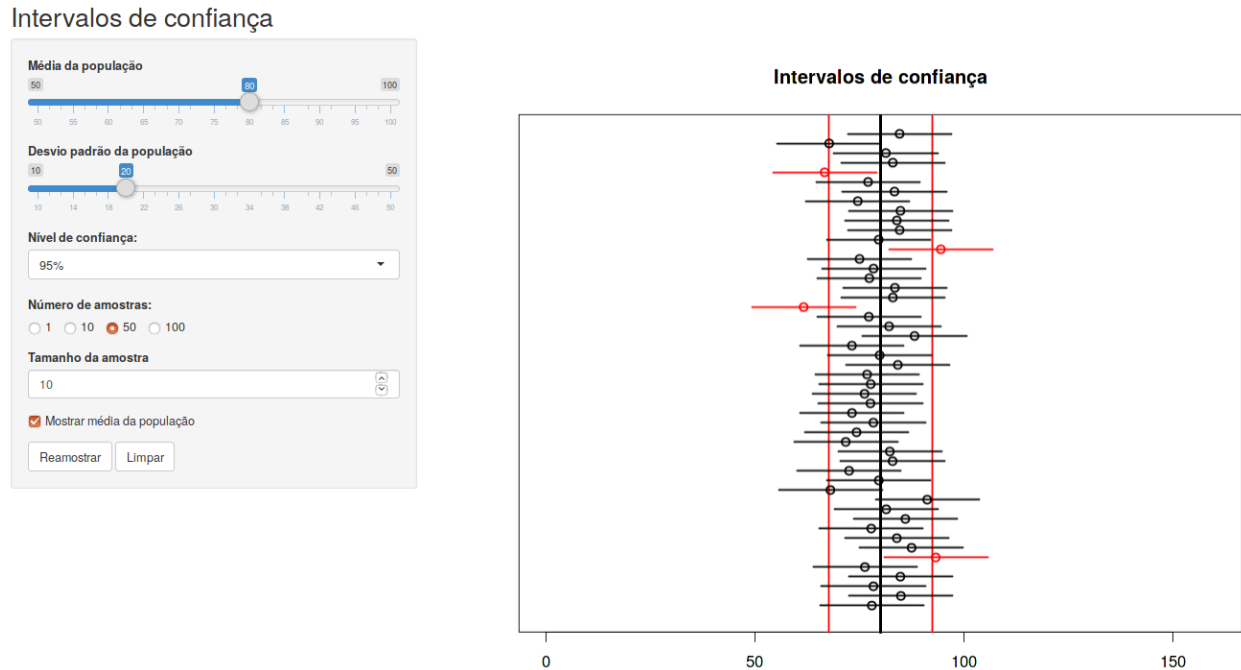


Figura 14.5: Figura 14.4 com linhas que mostram a média real da população (linha vertical preta) e indicam distâncias iguais a  $z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$  acima e abaixo da média da distribuição (linhas verticais vermelhas).

É de se esperar que nem sempre o intervalo de confiança contenha o valor real do parâmetro que ele estima. No exemplo da figura 14.5, o nível de confiança é de 95%. Isso significa que, se extraíssemos um número infinito de amostras aleatórias da população e calculássemos os respectivos intervalos de confiança, em 95% das vezes o intervalo de confiança irá incluir a média real da população e em 5% das vezes, o intervalo de confiança não irá incluir a média. Isso equivale a dizer que, a cada 100 intervalos de confiança calculados, em média 5 (5%) não conterão a média da distribuição. Na figura 14.5, 4 intervalos em 50 não contêm a média real da distribuição.

A figura 14.6, mostra o uso da aplicação com a mesma distribuição normal da figura 14.3, com o mesmo número de amostras, mas com três tamanhos amostrais diferentes (1, 10 e 50). Observe que a precisão dos intervalos de confiança aumenta à medida que o tamanho das amostras aumenta de 1 para 10 e de 10 para 50. Isso é de se esperar, porque o erro padrão  $\frac{\sigma}{\sqrt{n}}$ , utilizado no cálculo do intervalo de confiança, diminui com  $n$ .

O leitor deve experimentar com diferentes níveis de confiança, número de amostras, tamanhos

amostrais e parâmetros da distribuição normal.

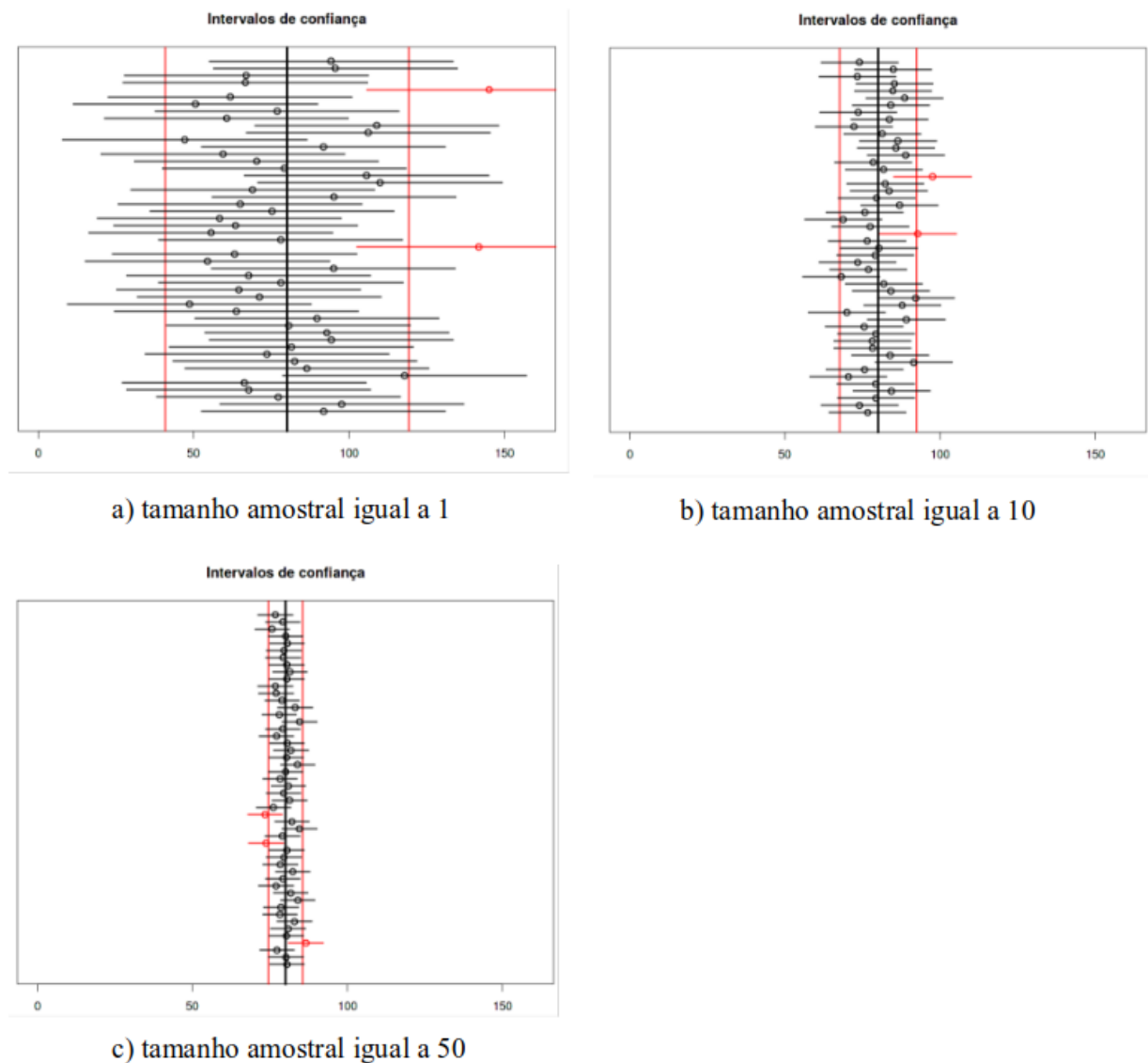


Figura 14.6: Intervalos de confiança para diferentes tamanhos de amostra (a-1, b-10, c-50).

Para qualquer estudo específico, não é possível garantir que o intervalo de confiança calculado contenha o parâmetro real da população.

Resumindo esta seção, podemos interpretar o intervalo de confiança da seguinte forma:

**Dado um nível de confiança estabelecido a priori  $(100 - \alpha)\%$ , temos uma confiança de  $(100 - \alpha)\%$  que o IC contenha o real valor do parâmetro estudado.**

Essa confiança deve ser interpretada no sentido de que, se repetíssemos o estudo um número infinito de vezes e, em cada vez, calculássemos o IC, em  $(100 - \alpha)\%$  das vezes o IC conteria o real valor do parâmetro estudado.

Voltando ao exemplo 1, substituindo os valores

$$\bar{x} = 92 \text{ mg/dl}, n = 36 \text{ e } \sigma = 16 \text{ mg/dl}$$

na expressão (14.7), temos que o intervalo de confiança ao nível de 95% para a média da glicemia de jejum de nossa população de pessoas não diabéticas é dado por:

$$92 - 1,96 \cdot 16/6 \leq \mu \leq 92 + 1,96 \cdot 16/6$$

$$86,8 \leq \mu \leq 97,2 \quad (\text{mg/dl}) \quad (14.8)$$

Se desejássemos o intervalo de confiança ao nível de 99%, bastaria substituir na expressão (14.7) os valores de  $\bar{x}$ ,  $n$ ,  $\sigma$  e  $z_{0,995}$ .

## 14.4 IC para a média quando a variância não é conhecida

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

Recordando mais uma vez, para uma população que possui uma distribuição normal  $N(\mu, \sigma^2)$  com média  $\mu$  e desvio padrão  $\sigma$ , a distribuição da média amostral para amostras de tamanho  $n$  será  $N(\mu, \sigma^2/n)$ . Isso significa que a distribuição da variável aleatória

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

possui a distribuição  $N(0, 1)$ . Isso justifica a utilização de  $z_{1-\alpha/2}$  na expressão (14.7) para o cálculo do intervalo de confiança.

**O que acontece quando não conhecemos a variância da população, que é o caso mais comum?** Nesse caso, podemos estimar  $\sigma^2$  por  $S^2$ , onde:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2,$$

porém a distribuição da variável aleatória

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

não segue a distribuição normal padrão  $N(0,1)$ . Na realidade, a variável aleatória  $T$  segue uma distribuição conhecida como  $t$  de Student, devido ao fato de essa distribuição ter sido publicada por William Gosset, que a publicou sob o pseudônimo de Student.

Matematicamente, a distribuição t de Student é dada pela fórmula:

$$T : f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi} \Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad (14.9)$$

onde  $\Gamma$  é a função Gama e  $\nu$  é o número de graus de liberdade.

A função Gama é dada pela seguinte integral:

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t} dt \quad (14.10)$$

O valor esperado de T é igual a zero ( $E[T] = 0$ ) e seu desvio padrão é maior 1.

A variável aleatória T segue a distribuição t de Student com n-1 graus de liberdade.

A figura 14.7 mostra o gráfico da distribuição normal padrão superposto ao gráfico da distribuição t de Student com  $\nu = 1$ . É possível observar que o gráfico da distribuição t de Student com 1 grau de liberdade é menos concentrado em torno da média do que o gráfico da distribuição normal padrão e se espalha mais para as laterais.

A figura 14.8 mostra o gráfico da distribuição normal padrão superposto aos gráficos da distribuição t de Student com  $\nu = 1, 5, 15$  e  $30$ , respectivamente. Observem que o gráfico da distribuição t de Student se aproxima cada vez mais do gráfico da distribuição normal padrão à medida que o número de graus de liberdade aumenta.

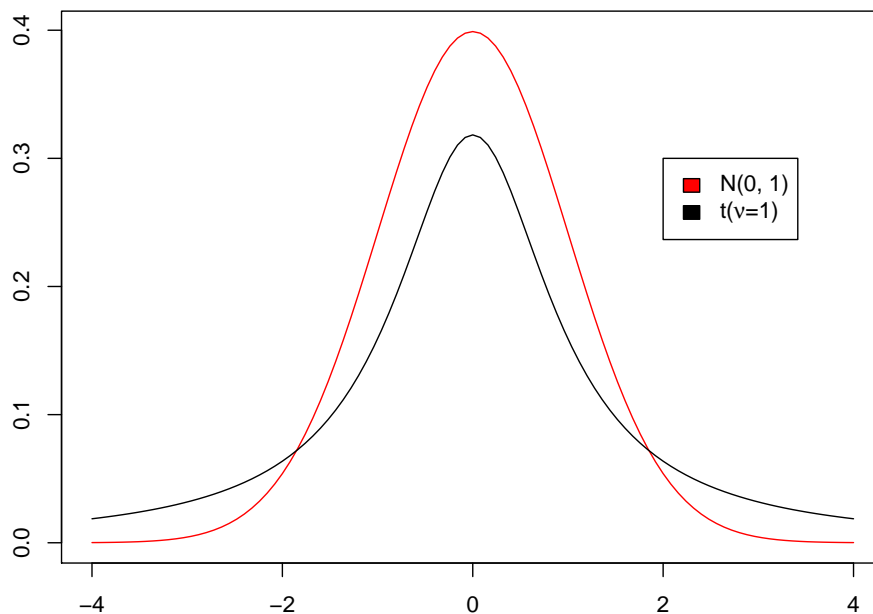


Figura 14.7: Gráficos da densidade de probabilidade para a distribuição normal padrão (vermelho) e da distribuição t de Student com 1 grau de liberdade.

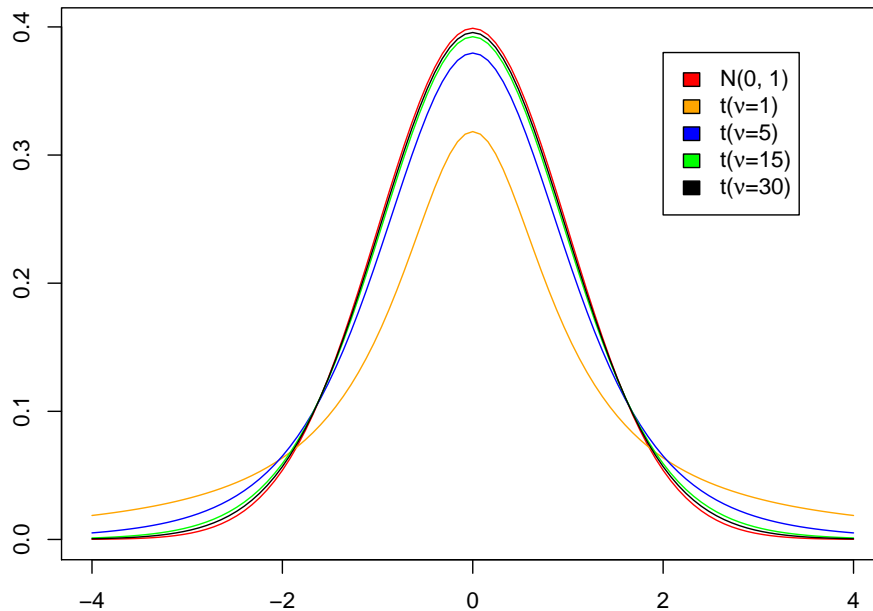


Figura 14.8: Gráficos da densidade de probabilidade para a distribuição normal padrão (vermelho) e da distribuição t de Student com graus de liberdade iguais a 1, 5, 15 e 30, respectivamente.

Para interpretarmos o conceito de graus de liberdade, vamos imaginar que obtivemos 1 amostra com  $n$  elementos  $x_1, x_2, \dots, x_n$ , extraídos de maneira independente de uma mesma população. Temos, portanto, que os valores da amostra são *iid*, ou seja, *independente e identicamente distribuídos*. Para essa amostra, ao obtermos sua média, somamos os  $n$  valores e, nesse caso, dizemos que a média amostral tem  $n$  graus de liberdade, porque foi calculada usando-se os  $n$  valores independentes da amostra.

Para a variância da amostra, temos que calcular:

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

Observemos que, no somatório, temos agora não  $n$  valores independentes, pois a média amostral é função dos  $n$  valores da amostra. Como  $\sum_{i=1}^n (x_i - \bar{x}) = 0$ , um dos termos  $(x_i - \bar{x})$  é o oposto da soma dos demais. Nesse caso, perdemos um grau de liberdade e a variância amostral tem, portanto,  $n - 1$  graus de liberdade.

É muito raro conhecermos o desvio padrão da população, portanto a construção do intervalo de confiança quase sempre envolve estimativas pontuais de  $\mu$  e  $\sigma$ . Quando  $\sigma$  não é conhecido, usamos sua estimativa  $S$  para a construção do intervalo de confiança da média, ou seja, usamos  $S_{\bar{X}} = \frac{S}{\sqrt{n}}$  como estimativa do desvio padrão da média amostral. Quando o desvio padrão é estimado, devemos usar a distribuição t ao invés da distribuição normal, embora, para amostras suficientemente grandes,  $z$  possa também ser utilizado, uma vez que a distribuição t se aproxima da distribuição normal à medida que  $n$  aumenta. Os valores de t são maiores do que os obtidos com a variável normal padronizada ( $z$ ), portanto o intervalo de confiança

será mais largo do que quando  $\sigma$  é conhecido. A fórmula para o intervalo de confiança para a média quando estimamos  $\sigma$  por  $S$  é dada pela expressão (14.7), substituindo-se  $z_{1-\alpha/2}$  por  $t_{n-1,1-\alpha/2}$  e  $\sigma$  por  $S$ :

$$\left[ \bar{X} - t_{n-1,1-\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + t_{n-1,1-\alpha/2} \frac{S}{\sqrt{n}} \right] \quad (14.11)$$

o valor de  $t_{n-1,1-\alpha/2}$  pode ser obtido no R pela expressão  $qt(1 - \alpha/2, n - 1)$ .

No exemplo 1, vamos supor que o desvio padrão 16 tenha sido estimado a partir da amostra.

Então:  $\bar{x} = 92$  mg/dl,  $n = 36$  e  $s = 16$ . Para  $\alpha = 5\%$  e  $n-1 = 35$ ,  $t_{35,0,975}$  é dado por:

```
qt(.975, 35)
```

```
## [1] 2.030108
```

O intervalo de confiança será dado então por:

$$[92 - 2, 03.16/6, 92 + 2, 03.16/6] = [86, 6 - 97, 4] \text{ (mg/dl)}$$

Esse intervalo é ligeiramente maior, mas próximo, do que o intervalo dado por (14.8).

## 14.5 Intervalo de confiança para a variância

Para a variância, como vimos no capítulo anterior, um bom estimador não tendencioso e consistente é dado por:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (14.12)$$

Se soubermos a distribuição de probabilidades desse estimador, então poderíamos calcular o intervalo de confiança para a variância de modo análogo ao realizado no caso da média da distribuição.

Multiplicando e dividindo o segundo membro de (14.12) por  $\sigma^2$ , podemos reescrever o estimador  $S^2$  como abaixo, :

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{\sigma^2}{n-1} \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2}$$



Quando  $X_i$  for uma variável aleatória com distribuição normal, o somatório na expressão acima terá uma distribuição conhecida como *qui ao quadrado* com  $(n-1)$  graus de liberdade, sendo representada por  $\chi_{n-1}^2$ :

$$S^2 \sim \frac{\sigma^2}{n-1} \chi_{n-1}^2$$

$\chi_{n,p}^2$  significa o valor da variável aleatória  $\chi_n^2$  que segue a distribuição qui ao quadrado para o qual  $P(\chi_{n,p}^2 \leq \chi_n^2) = p$ , ou seja, a área sob a curva qui ao quadrado à esquerda de  $\chi_{n,p}^2$  é igual a  $p$ .

Para obtermos então o intervalo com nível de confiança igual a  $(1 - \alpha)\%$ , a partir da distribuição acima, achamos os valores  $\chi_{n-1,\alpha/2}^2$  e  $\chi_{n-1,1-\alpha/2}^2$  para os quais o intervalo abaixo contém, com probabilidade  $(1 - \alpha)\%$ , a estimativa da variância em uma amostra de tamanho  $n$ :

$$\frac{\sigma^2}{n-1} \chi_{n-1,\alpha/2}^2 \leq S^2 \leq \frac{\sigma^2}{n-1} \chi_{n-1,1-\alpha/2}^2$$

Isolando a variância  $\sigma^2$  na expressão acima, obtemos então o intervalo de confiança ao nível de  $(1 - \alpha)\%$  para a variância da população de uma variável aleatória que segue a distribuição normal:

$$S^2 \frac{n-1}{\chi_{n-1,1-\alpha/2}^2} \leq \sigma^2 \leq S^2 \frac{n-1}{\chi_{n-1,\alpha/2}^2} \quad (14.13)$$

**Exemplo 2:** Uma amostra de 6 cobaias é analisada para verificar a dosagem de um certo composto, obtendo-se a média amostral igual a 14,1 mg/dl e a variância igual a 2,1 (mg/dl)<sup>2</sup>. Obter intervalos de confiança, ao nível de 95%, para a média e a variância populacional, assumindo que os dados seguem uma distribuição normal.

Como temos uma amostra pequena,  $n = 6$ , para determinar o IC (intervalo de confiança), temos que usar a distribuição  $t$  de Student, com  $\alpha = (1 - 0,95) = 0,05$  e número de graus de liberdade igual a 5 (6-1).

$$\left[ \bar{x} - t_{n-1,\alpha/2} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1,\alpha/2} \frac{s}{\sqrt{n}} \right] \quad (14.14)$$

O valor de  $t_{n-1,\alpha/2}$ , para  $n = 6$  e  $\alpha = 0,05$  pode ser obtido no R com a função `qt(0.975, df=5)`, sendo igual a 2,57.

```
qt(0.975, df=5)
```

```
## [1] 2.570582
```

Logo, entrando com os valores na expressão (14.11), teremos que o intervalo de confiança para a média será dado por

$$14,1 - 2,57\sqrt{\frac{2,1}{6}} \leq \mu \leq 14,1 + 2,57\sqrt{\frac{2,1}{6}}$$

$$12,58 \leq \mu \leq 15,62 \text{ (mg/dl)}$$

Para a variância, usamos a expressão (14.13), mas precisamos obter os valores  $\chi_{5;0,025}^2$  e  $\chi_{5;0,975}^2$ . No R, as expressões `qchisq(0.025, df=5)` e `qchisq(0.975, df=5)` nos darão os valores desejados de  $\chi_{5;0,025}^2$  e  $\chi_{5;0,975}^2$ .

```
qchisq(0.025,df=5); qchisq(0.975, df=5)
```

```
## [1] 0.8312116
```

```
## [1] 12.8325
```

Entrando com esses e os demais valores necessários na equação (14.13), teremos o intervalo de confiança para a variância:

$$2,1 \frac{5}{12,8325} \leq \sigma^2 \leq 2,1 \frac{5}{0,8312}$$

$$0,82 \leq \sigma^2 \leq 12,63 \text{ (mg/dl)}^2$$

## 14.6 Distribuição qui ao quadrado

Seja  $X$  uma variável aleatória com distribuição  $N(\mu, \sigma)$  e  $Z = \frac{X-\mu}{\sigma}$  a correspondente variável normal padronizada. Então a variável aleatória  $Z^2$  é uma variável aleatória não negativa e sua distribuição é denominada distribuição qui ao quadrado ( $\chi_1^2$ ) com um grau de liberdade.

Agora consideremos duas variáveis normais padronizadas  $Z_1$  e  $Z_2$  independentes. Nesse caso, a soma  $Z_1^2 + Z_2^2$  segue uma distribuição  $\chi_2^2$  com dois graus de liberdade. Analogamente, a soma de  $n$  variáveis aleatórias normais padronizadas independentes terá uma distribuição  $\chi_n^2$  com  $n$  graus de liberdade.

$$\chi_n^2 = Z_1^2 + Z_2^2 + \cdots + Z_n^2$$

O valor esperado e a variância de  $\chi_n^2$  são, respectivamente

$$E[\chi_n^2] = n$$

$$var[\chi_n^2] = 2n$$

Usa-se como representação geral da variável qui ao quadrado a expressão  $\chi_\nu^2$ , onde  $\nu$  corresponde ao número de graus de liberdade. Para  $n > 30$ , temos que a distribuição qui ao quadrado se aproxima bem de uma distribuição normal com média  $n$  e variância  $2n$ .

Apesar de não usarmos nesse texto a função matemática da distribuição qui ao quadrado, ela é dada pela expressão:

$$f(\chi_\nu^2) = \frac{1}{2^{\frac{\nu}{2}} \Gamma(\frac{\nu}{2})} (\chi_\nu^2)^{\frac{\nu}{2}-1} e^{-\frac{\chi_\nu^2}{2}}, \chi_\nu^2 > 0$$

onde  $\Gamma(.)$  é a função gama que aparece na distribuição t de Student (equação (14.9)).  $\chi_\nu^2$  é a variável da função  $f(.)$ , ou seja, ela não é uma variável elevada ao quadrado, mas sim **a própria variável na expressão**.

A figura 14.9 mostra gráficos da distribuição qui ao quadrado para diferentes graus de liberdade ( $\nu = 2, 10, 20$  e  $30$ ).

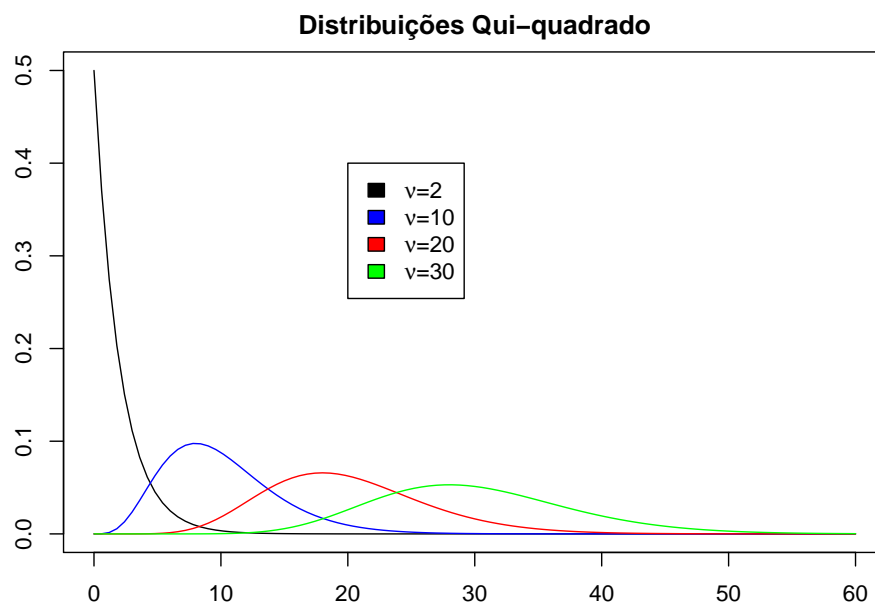


Figura 14.9: Gráficos de distribuições qui ao quadrado com graus de liberdade iguais a 2, 10, 20 e 30, respectivamente.

## 14.7 Intervalo de confiança para proporções

Vimos no capítulo anterior que a probabilidade  $p$  de sucesso um experimento de Bernoulli pode ser estimada pela proporção de sucessos ( $\hat{P}$ ) em  $n$  experimentos de Bernoulli. Assim, sendo  $X$  o número de sucessos nos  $n$  experimentos, temos:

$$\hat{P} = \frac{X}{n} \quad (14.15)$$

Para  $n$  suficientemente grande, a distribuição de  $\hat{P}$  pode ser aproximada por uma distribuição normal com média e variância dadas por:

$$E[\hat{P}] = \frac{E[X]}{n} = \frac{np}{n} = p \quad (14.16)$$

$$\text{var}[\hat{P}] = \frac{\text{var}[X]}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n} \quad (14.17)$$

Se realizarmos  $n$  experimentos de Bernoulli e observarmos a proporção de sucessos, podemos então estimar o intervalo de confiança para a proporção real de sucessos  $p$ , aplicando a fórmula para o intervalo de confiança da média de uma distribuição normal (equação (14.7)), substituindo  $\mu$  por  $p$ ,  $\bar{X}$  por  $\hat{P}$ , e  $\sigma$  por  $\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$ . Assim o intervalo de confiança ao nível  $(100 - \alpha)\%$  para a proporção de sucessos em  $n$  experimentos de Bernoulli para  $n$  suficientemente grande é dado por:

$$\hat{P} - z_{1-\alpha/2} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \leq p \leq \hat{P} + z_{1-\alpha/2} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \quad (14.18)$$

**Exemplo 3:** vamos considerar que desejamos estimar a proporção de aparelhos de raios-X que estejam com defeito e produzam um excesso de radiação. Tomando-se uma amostra de 40 aparelhos, identificou-se que 12 estavam com defeito. O nosso problema aqui é determinar um intervalo de confiança para a proporção populacional  $p$  de aparelhos que possam estar com defeito. Na terminologia utilizada neste texto, o sucesso, nesse caso, é considerado um aparelho com defeito.

A partir da amostra, obtemos  $\hat{p} = 12/40 = 0,30$ . Para um intervalo de confiança de 95%, obtemos  $z_{1-\alpha/2} = 1,96$ . Com esses valores, obtemos o IC (intervalo de confiança):

$$0,3 - 1,96 \sqrt{\frac{0,3(1-0,3)}{40}} \leq p \leq 0,3 + 1,96 \sqrt{\frac{0,3(1-0,3)}{40}}$$

$$0,16 \leq p \leq 0,44$$

Nesse caso, estamos assumindo uma aproximação para a normal. Um resultado mais acurado pode ser obtido, introduzindo-se a correção de continuidade, somando  $0,5/n$  para o limite superior e subtraindo esse valor do limite inferior, obtendo os valores 0,1475 e 0,4525 para os limites inferior e superior, respectivamente. Assim, com 95% de confiança, entre 14,75% e 45,25% dos equipamentos de raios x apresentam o defeito.

## 14.8 Resumo para obtenção de intervalos de confiança de um parâmetro

Resumindo o que vimos até aqui, podemos generalizar os passos para obter um intervalo de confiança de um determinado parâmetro. Basicamente são quatro passos, que serão ilustrados com os casos discutidos nas seções precedentes.

### 1) Encontrar um estimador para o parâmetro

Exemplos:

- a) Para a média de uma distribuição de uma variável aleatória  $X$ :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (14.19)$$

- b) Para a variância de uma distribuição:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (14.20)$$

- c) para a probabilidade de sucesso em uma distribuição binomial:

$$\hat{P} = \frac{\text{contagem de ocorrências do evento}}{\text{número de experimentos}} \quad (14.21)$$

### 2) Encontrar a distribuição do estimador do parâmetro de interesse

- a) Para a média de uma distribuição normal com variância conhecida  $\sigma$ :

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

onde  $Z$  segue a distribuição normal padrão.

- b) Para a média de uma distribuição normal com variância desconhecida:

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

onde T segue a distribuição t de Student com n-1 graus de liberdade

c) Para a variância de uma distribuição normal com média desconhecida:

$$S^2 = \frac{\sigma^2}{n-1} \chi_{n-1}^2$$

onde  $\chi_{n-1}^2$  segue a distribuição qui ao quadrado com n-1 graus de liberdade.

d) para a probabilidade de sucesso em uma distribuição binomial, com n suficientemente grande:

$$Z = \frac{\hat{P} - p}{\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}}$$

onde Z segue a distribuição normal padrão

**3) Especificar o nível de confiança  $(1 - \alpha)$  e determinar o intervalo que contém, com probabilidade  $(1 - \alpha)$ , a estimativa do parâmetro**

a) Para a média de uma distribuição normal com variância conhecida  $\sigma$ :

$$\mu - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \quad (14.22)$$

b) Para a média de uma distribuição normal com variância desconhecida:

$$\mu - t_{n-1, 1-\alpha/2} \frac{S}{\sqrt{n}} \leq \bar{X} \leq \mu + t_{n-1, 1-\alpha/2} \frac{S}{\sqrt{n}} \quad (14.23)$$

c) Para a variância de uma distribuição normal com média desconhecida:

$$\frac{\sigma^2}{n-1} \chi_{n-1, \alpha/2}^2 \leq S^2 \leq \frac{\sigma^2}{n-1} \chi_{n-1, 1-\alpha/2}^2 \quad (14.24)$$

d) para a probabilidade de sucesso em uma distribuição binomial, com n suficientemente grande:

$$p - z_{1-\alpha/2} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \leq \hat{P} \leq p + z_{1-\alpha/2} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \quad (14.25)$$

#### 4) Nas expressões em 3, inverter as relações para obter o intervalo de confiança

a) Para a média de uma distribuição normal com variância conhecida  $\sigma$ :

$$\bar{X} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \quad (14.26)$$

b) Para a média de uma distribuição normal com variância desconhecida:

$$\bar{X} - t_{n-1, 1-\alpha/2} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1, 1-\alpha/2} \frac{S}{\sqrt{n}} \quad (14.27)$$

c) Para a variância de uma distribuição normal com média desconhecida:

$$S^2 \frac{n-1}{\chi_{n-1, 1-\alpha/2}^2} \leq \sigma^2 \leq S^2 \frac{n-1}{\chi_{n-1, \alpha/2}^2} \quad (14.28)$$

d) para a probabilidade de sucesso em uma distribuição binomial, com n suficientemente grande:

$$\hat{P} - z_{1-\alpha/2} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \leq p \leq \hat{P} + z_{1-\alpha/2} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \quad (14.29)$$

Portanto, **para qualquer parâmetro que estimamos usando uma estatística, não somente os exemplificados acima**, podemos determinar o intervalo de confiança se conhecermos a distribuição amostral da estatística de interesse.

## 14.9 Exercícios

- 1) As idades de uma amostra aleatória de 50 membros de uma certa sociedade acadêmica são obtidas e encontra-se que  $\bar{x} = 53,8$  anos e  $s = 9,89$  anos. A idade dessa população segue uma distribuição normal, mas não conhecemos nem a média nem a variância.
  - a) Calcule o intervalo com 90% de confiança para a média de idade de todos os membros da sociedade.
  - b) Se a média acima tivesse sido obtida de uma amostra aleatória de 100 indivíduos da população, qual seria o intervalo com 90% de confiança para a média da população?
  - c) Calcule o intervalo com 99% de confiança para a variância da idade de todos os membros da sociedade.
- 2) Quais são os passos para construir o intervalo de confiança para a média amostral de uma população que segue a distribuição normal e você não sabe qual é a variância populacional?



# Capítulo 15

## Testes de hipóteses

### 15.1 Introdução

Este capítulo irá aprofundar diversos conceitos ligados aos testes de hipóteses, introduzidos no capítulo 6, enfatizando a importância do tamanho das amostras, o poder estatístico do teste e a interpretação correta dos valores de p.

### 15.2 Exemplo inicial (primeiro cenário)

Os conteúdos desta seção e da seção 15.3 podem ser visualizados neste [vídeo](#).

Vamos iniciar com um exemplo. Supondo que a distribuição dos valores da glicemia de jejum em uma população de pessoas não diabéticas seja normal, um pesquisador deseja testar a hipótese de que a média de glicemia de jejum nessa população seja igual a 85 mg/dl. Como proceder?

Vamos considerar esse assunto, utilizando o que aprendemos de estatística até agora.

Sabemos que, se uma variável aleatória  $X$  possui uma distribuição normal com média  $\mu$ , mas não sabemos a variância dessa distribuição, então a estatística:

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

terá uma distribuição t de Student com  $n-1$  graus de liberdade e

$$P\left[\mu - t_{n-1, 1-\alpha/2} \frac{S}{\sqrt{n}} \leq \bar{X} \leq \mu + t_{n-1, 1-\alpha/2} \frac{S}{\sqrt{n}}\right] = 1 - \alpha$$

ou seja, a probabilidade de extrairmos uma amostra de tamanho  $n$  dessa população e a média da amostra cair no intervalo a seguir é  $1 - \alpha$ :

$$\left[ \mu - t_{n-1, 1-\alpha/2} \frac{S}{\sqrt{n}}, \mu + t_{n-1, 1-\alpha/2} \frac{S}{\sqrt{n}} \right] \quad (15.1)$$

Então uma forma de testarmos a hipótese de que a média da população é de 85 mg/dl é adotar o seguinte enfoque: escolhemos um nível de confiança  $(1 - \alpha)$ , selecionamos uma amostra de um determinado tamanho da população, e verificamos se a média da amostra está fora do intervalo definido pela expressão (15.1). Se a média da amostra estiver fora do intervalo, a hipótese é rejeitada. Caso contrário, ela não é rejeitada.

Vamos aplicar esse raciocínio ao exemplo inicial. Vamos supor que o pesquisador selecionou o nível de confiança igual a 95%, ou seja,  $\alpha = 5\%$ , e obteve na amostra de 36 pessoas uma média de 92 mg/dl e desvio padrão,  $s$ , igual a 16 mg/dl. Então, com 95% de probabilidade, o valor da média de uma amostra estaria no intervalo:

$$\left[ 85 - t_{35, 0,975} \frac{16}{\sqrt{36}}, 85 + t_{35, 0,975} \frac{16}{\sqrt{36}} \right] = \left[ 85 - 2,03 \frac{16}{\sqrt{36}}, 85 + 2,03 \frac{16}{\sqrt{36}} \right] \quad (15.2)$$

$$= [79,6 - 90,4] \text{ mg/dl} \quad (15.3)$$

O valor de  $t_{35, 0,975}$  acima pode ser obtido no R com o comando:

```
qt(.975, df=35)
```

```
## [1] 2.030108
```

O valor da média da amostra, 92 mg/dl, está fora do intervalo dado por (15.3). Portanto o pesquisador irá rejeitar a hipótese de que a média da glicemia de jejum nessa população seja de 85 mg/dl. A razão para essa rejeição é que, caso 85 mg/dl fosse a média de glicemia de jejum na população, a probabilidade de selecionar aleatoriamente uma amostra de 36 pacientes dessa população e a média cair fora do intervalo  $[79,6 - 90,4]$  é de apenas 5%. Nesse caso, o pesquisador prefere acreditar que a média da população é diferente de 85 mg/dl.

Vamos a seguir abordar esse mesmo problema sob o ponto de vista de um teste de hipótese.

## 15.3 Processo para realizar um teste de hipótese

A forma exata como um teste de hipótese é conduzido depende de cada problema específico. Usando o exemplo anterior, vamos sistematizar os passos gerais para realizar um teste de hipótese, aproveitando para reforçar alguns conceitos e notações comumente utilizados e introduzidos no capítulo 6.

### 1) Passo 1: expressar o tema da pesquisa em termos de uma hipótese estatística

O primeiro passo é o de estabelecer uma hipótese que será avaliada. Essa hipótese é chamada de **hipótese nula**, representada por  $H_0$ . No exemplo anterior, a hipótese nula é de que a

distribuição de probabilidades da glicemia de jejum na população que estamos estudando é normal com média igual a 85 mg/dl. Nada especificamos sobre o valor da variância. Ao estabelecermos uma hipótese a ser testada, a mesma será confrontada com uma **hipótese alternativa**, representada por  $H_1$ . No exemplo acima, a hipótese alternativa é que a distribuição de probabilidades da glicemia de jejum na população que estamos estudando é normal com média diferente de 85 mg/dl.

Assim temos:

$H_0$ : glicemia de jejum  $\sim$  distribuição normal  $N(\mu, \sigma^2)$ , com  $\mu = 85 \text{ mg/dl}$

$H_1$ : glicemia de jejum  $\sim$  distribuição normal  $N(\mu, \sigma^2)$ , com  $\mu \neq 85 \text{ mg/dl}$

## 2) Passo 2: decidir sobre um teste estatístico apropriado para testar a hipótese nula (escolha de uma estatística)

A partir da hipótese nula, podemos pensar em um teste estatístico apropriado para testá-la. No exemplo considerado, uma possibilidade é o de selecionarmos uma amostra aleatória da população com um certo número de elementos ( $n$ ), calcularmos a média da amostra e compararmos com a média estabelecida pela hipótese nula.

Sabemos que a estatística

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \quad (15.4)$$

terá uma distribuição t de Student com  $n-1$  graus de liberdade, se a hipótese nula for verdadeira.

## 3) Passo 3: selecionar o nível de significância ( $\alpha$ ) para o teste estatístico

O **nível de significância**, também chamado de **alfa**, é um valor de probabilidade que o pesquisador considera suficientemente baixo que definirá uma **região crítica** da distribuição da estatística definida no passo 2, tal que, se o valor da estatística calculada a partir da amostra selecionada cair na região crítica, a hipótese nula será rejeitada. Caso contrário, a hipótese nula não será rejeitada. Escolhemos valores baixos para  $\alpha$ , porque desejamos que seja baixa a probabilidade de rejeitarmos a hipótese nula quando ela for verdadeira. Valores tradicionais de  $\alpha$  são 0,1 (10%), 0,05 (5%) e 0,01 (1%), sendo o nível de 5% o mais frequentemente utilizado. Vamos tentar entender isso melhor.

Considerando que a estatística  $T$ , definida no passo 2, segue a distribuição t de Student, a figura 15.1 mostra o gráfico de  $T$  e dois pontos nessa distribuição  $-t_{n-1, 1-\alpha/2}$  e  $t_{n-1, 1-\alpha/2}$  que delimitam uma região (em vermelho) cuja área é igual a  $\alpha$  (a probabilidade de se obter um valor de  $t$  abaixo de  $-t_{n-1, 1-\alpha/2}$  ou acima de  $t_{n-1, 1-\alpha/2}$ ). Essa região é denominada de região crítica e os pontos  $-t_{n-1, 1-\alpha/2}$  e  $t_{n-1, 1-\alpha/2}$  são chamados de **pontos críticos**.

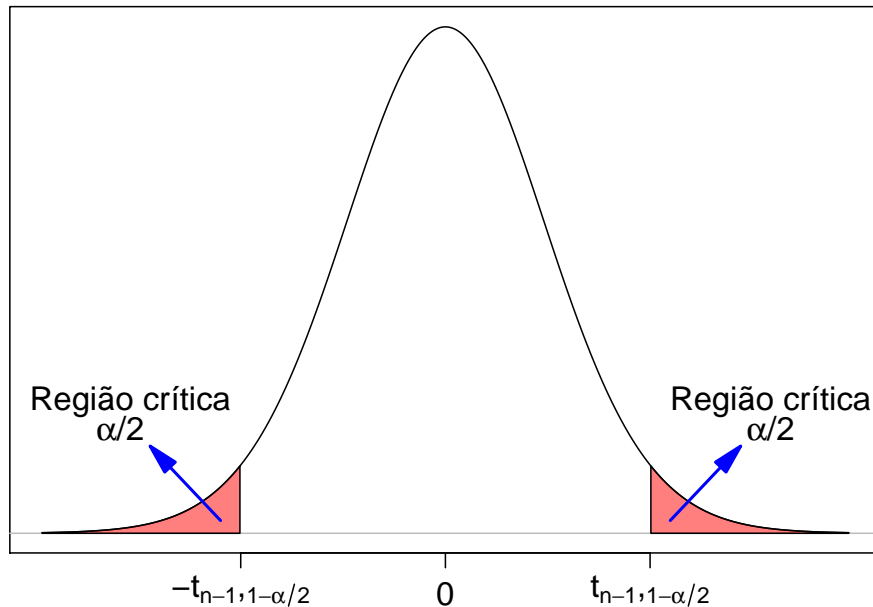


Figura 15.1: Definição da região crítica para o teste de hipótese definido no passo 2.

No exemplo acima,  $\alpha$  foi definido como 5% e  $n = 36$ . Assim o valor crítico  $t_{35, 1-0,05/2} = t_{35;0,975}$  é calculado no R da seguinte forma:

```
qt(.975, 35)
```

```
## [1] 2.030108
```

Então  $t_{35;0,975} = 2,03$ .

#### 4) Passo 4: selecionar a amostra e realizar os cálculos

Ao definirmos a estatística a ser utilizada, o valor de  $\alpha$  e, conseqüente, a região crítica do teste, procedemos à seleção da amostra do estudo, a partir da qual o valor da estatística será calculado e comparado com os valores críticos. Caso o valor da estatística caia dentro da região crítica do teste, a hipótese nula será rejeitada. Nesse caso, dizemos que o resultado do teste é **estatisticamente significativo**. Caso contrário, ela não será rejeitada e o resultado do teste não é estatisticamente significativo.

No exemplo considerado, a amostra da população gerou os seguintes resultados:

$\bar{x} = 92$  mg/dl

$n = 36$

$s = 16$  mg/dl

Substituindo esses valores na expressão (15.4), obtemos:

$$t = \frac{92 - 85}{\frac{16}{\sqrt{36}}} = 2,62$$

### 5) Passo 5: tomar a decisão

A partir do cálculo do valor de  $t$  na amostra, vemos que  $t > t_{35;0,975} = 2,03$ . O valor de  $t$  caiu na região crítica e a hipótese nula é então rejeitada (figura 15.2). O resultado é estatisticamente significativo.

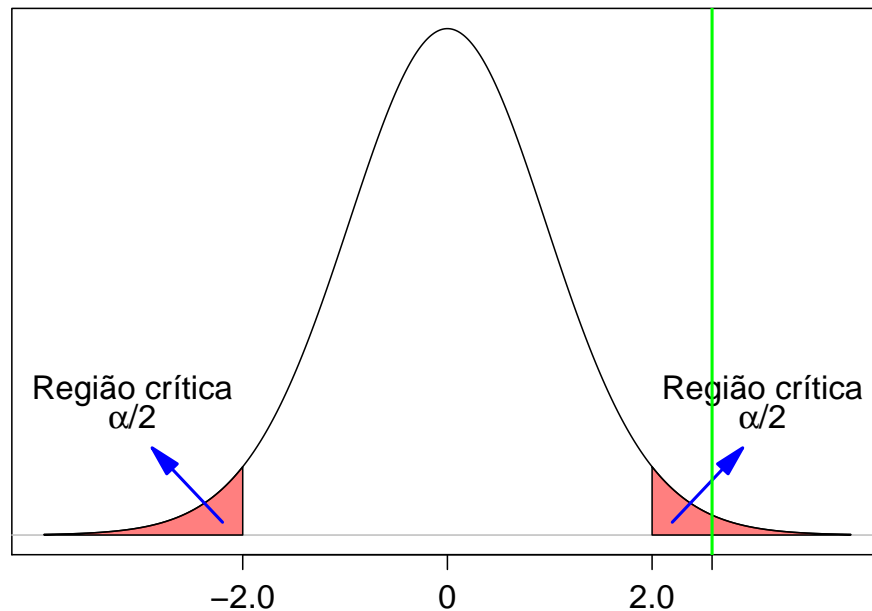


Figura 15.2: O valor da estatística no teste de hipótese é 2,62 (reta vertical verde) e está localizado na região crítica do teste. Portanto a hipótese nula é rejeitada.

### 15.3.1 Segundo cenário

Vamos supor que, ao extrairmos uma amostra aleatória da população no passo 4 da seção anterior, obtivéssemos os seguintes resultados:

$$\bar{x} = 89 \text{ mg/dl}$$

$$n = 36$$

$$s = 16 \text{ mg/dl}$$

Substituindo esses valores na expressão (15.4), obtemos:

$$t = \frac{89 - 85}{\frac{16}{\sqrt{36}}} = 1,50$$

A partir do cálculo do valor de  $t$  nessa amostra, vemos que  $-t_{35;0,975} \leq t \leq t_{35;0,975}$ . O valor de  $t$  não caiu na região crítica e a hipótese nula **não** é rejeitada (figura 15.3). O resultado **não é estatisticamente significativo**.

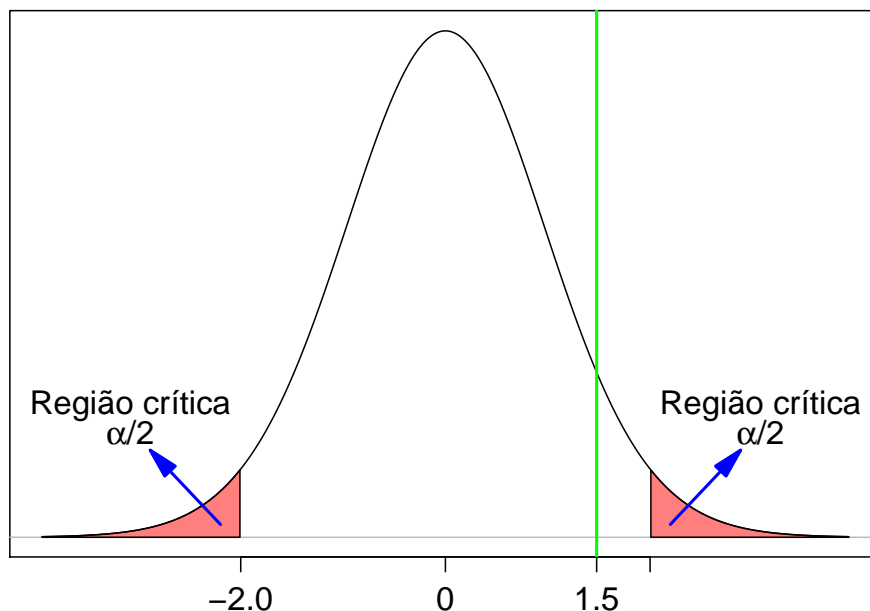


Figura 15.3: O valor da estatística no teste de hipótese é 1,50 (reta vertical verde) e está localizado fora da região crítica do teste. Portanto a hipótese nula não é rejeitada.

## 15.4 Relação entre o intervalo de confiança e o teste de hipótese

Os conteúdos desta seção e da seção 15.5 podem ser visualizados neste [vídeo](#).

Também podemos, em muitas situações, usar o intervalo de confiança para decidir sobre a significância ou não de um resultado de um teste estatístico.

Retomando o exemplo da introdução, vamos novamente calcular o intervalo de confiança com nível de confiança de 95% ( $1 - \alpha = 0,95$ ). Para uma variável aleatória  $X$  que segue uma distribuição normal com média  $\mu$  e variância desconhecida, o intervalo de confiança, obtido a partir de uma amostra de tamanho  $n$ , é dado pela expressão (14.11) do capítulo 14:

$$\left[ \bar{X} - t_{n-1, 1-\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + t_{n-1, 1-\alpha/2} \frac{S}{\sqrt{n}} \right]$$

Substituindo os valores produzidos pela amostra:

$$\bar{x} = 92 \text{ mg/dl}$$

$$n = 36$$

$$s = 16 \text{ mg/dl}$$

$$t_{35;0,975} = 2,03$$

na expressão acima, resulta no seguinte intervalo de confiança:

$$\text{IC}(95\%): 86,6 \leq \mu \leq 97,4$$

A linha horizontal em azul na figura 15.4 mostra o intervalo de confiança calculado acima. Para esse teste de hipótese, a hipótese nula foi rejeitada e o intervalo de confiança não contém o valor estabelecido pela hipótese nula (85 mg/dl).

**Assim, se o intervalo de confiança não contiver o valor do parâmetro sob a hipótese nula que está sendo testada, então a hipótese nula será rejeitada.**

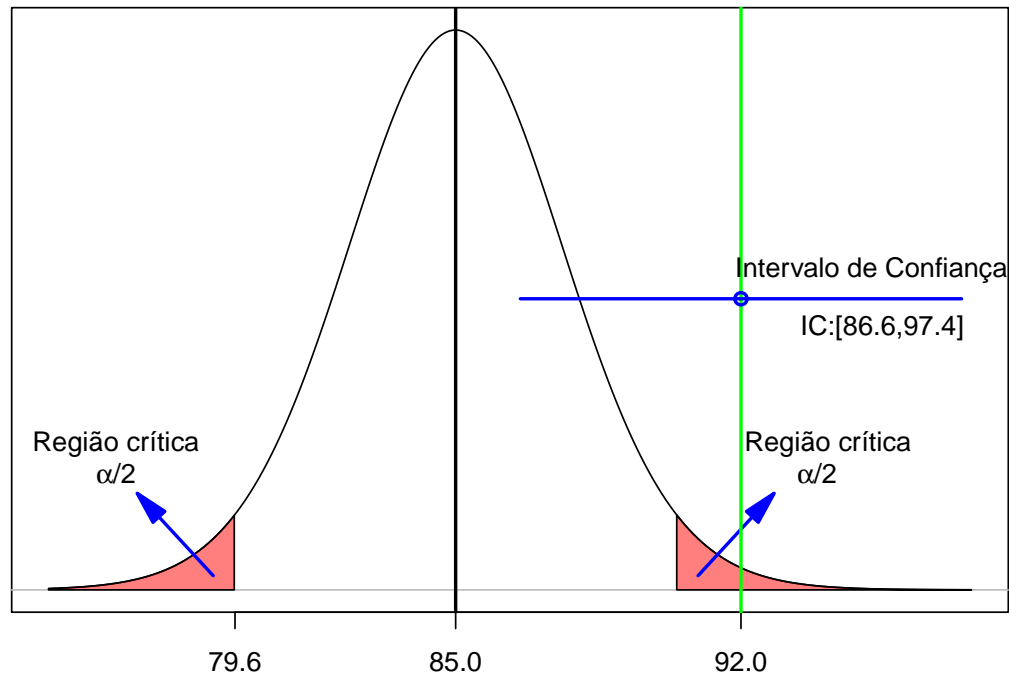


Figura 15.4: Relação entre um teste de hipótese e o intervalo de confiança quando a hipótese nula é rejeitada. A linha verde indica o valor da média amostral.

No outro cenário mostrado ao final da seção anterior, onde a amostra gerou os seguintes valores:

$$\bar{x} = 89 \text{ mg/dl}$$

$$n = 36$$

$$s = 16 \text{ mg/dl}$$

o intervalo de confiança será:

$$\text{IC}(95\%): 83,6 \leq \mu \leq 94,4$$

A linha horizontal em azul na figura 15.5 mostra o intervalo de confiança calculado acima. Para esse teste de hipótese, a hipótese nula não foi rejeitada e o intervalo de confiança contém o valor estabelecido pela hipótese nula (85 mg/dl).

**Assim, se o intervalo de confiança incluir o valor do parâmetro sob a hipótese nula que está sendo testada, então ela não será rejeitada.**

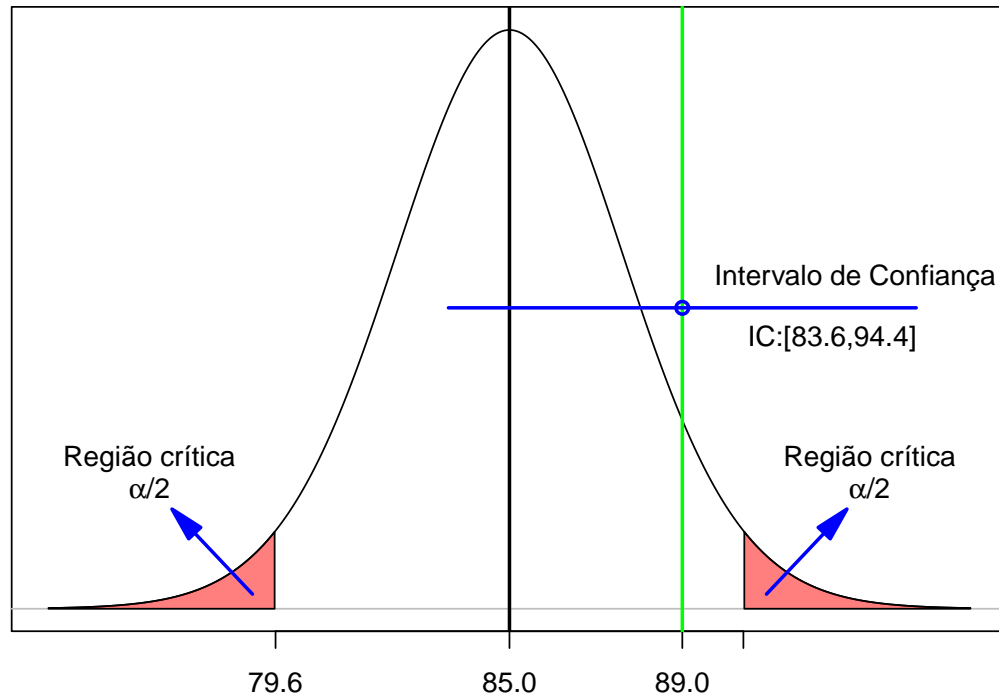


Figura 15.5: Relação entre um teste de hipótese e o intervalo de confiança quando a hipótese nula não é rejeitada. A linha verde indica o valor da média amostral.

Novamente vamos enfatizar que o intervalo de confiança fornece mais informações do que a simples rejeição ou não de uma hipótese nula. Ele também nos fornece a precisão da estimativa do parâmetro considerado, nos permitindo também avaliar a relevância clínica do achado (seção 6.10).

## 15.5 Interpretação alternativa para o IC

Os limites de um intervalo de confiança podem ser obtidos de maneira análoga à utilizada para obter o intervalo de confiança para a diferença de médias dos valores de ácido fólico para dois métodos diferentes de ventilação, conforme mostrado no capítulo 6, seção 6.5.

Vamos supor que estamos interessados em calcular o intervalo de confiança para a média de uma variável que segue uma distribuição normal com variância conhecida,  $\sigma$ , e que calculamos a média amostral ( $\bar{x}$ ) de uma amostra de tamanho  $n$ , extraída aleatoriamente da população com essa distribuição.

A figura 15.6 mostra uma maneira diferente de interpretar os limites superior e inferior do intervalo de confiança.

O limite inferior do intervalo de confiança,  $\mu_i$ , pode ser obtido traçando o gráfico da distribuição normal da média amostral com variância igual a  $\sigma/\sqrt{n}$ , de tal modo que a área sob essa distribuição acima de  $\bar{x}$  barra seja igual a  $\alpha/2$ , que é igual à probabilidade de obtermos aleatoriamente uma amostra de tamanho  $n$  da população com média igual a  $\mu_i$  e a média



amostral ser maior do que  $\bar{x}$ . Para qualquer valor de média da distribuição normal abaixo de  $\mu_i$ , a probabilidade de se obter uma média de uma amostra com tamanho  $n$  maior que a média amostral encontrada,  $\bar{x}$ , será menor que  $\alpha/2$ . Assim todas as hipóteses nulas cujos valores de médias fossem menores do que  $\mu_i$  seriam rejeitadas no teste de hipótese realizado a partir dessa amostra.

O limite superior do intervalo de confiança,  $\mu_s$ , pode ser obtido traçando o gráfico da distribuição normal da média amostral com variância igual a  $\sigma/\sqrt{n}$ , de tal modo que a área sob essa distribuição abaixo de  $\bar{x}$  seja igual a  $\alpha/2$ , que é igual à probabilidade de obtermos aleatoriamente uma amostra de tamanho  $n$  da população com média igual a  $\mu_s$  e a média amostral ser menor do que  $\bar{x}$ . Para qualquer valor de média da distribuição normal acima de  $\mu_s$ , a probabilidade de se obter uma média de uma amostra com tamanho  $n$  menor que a média amostral encontrada,  $\bar{x}$ , será menor que  $\alpha/2$ . Assim todas as hipóteses nulas cujos valores de médias fossem maiores do que  $\mu_s$  seriam rejeitadas no teste de hipótese realizado a partir dessa amostra.

Logo o intervalo de confiança é dado por todos os valores de média,  $\mu$ , tal que é  $\mu_i \leq \mu \leq \mu_s$ . A distância de  $\bar{x}$  até  $\mu_i$  ou  $\mu_s$  é igual  $z_{1-\alpha/2} \cdot \sigma/\sqrt{n}$ . Esse intervalo de confiança é o mesmo que obtivemos na seção 14.2, capítulo 14, e pode ser interpretado como o conjunto de valores de médias da distribuição normal correspondentes a hipóteses nulas que não seriam rejeitadas por um teste de hipótese realizado a partir dos dados obtidos na amostra do estudo.

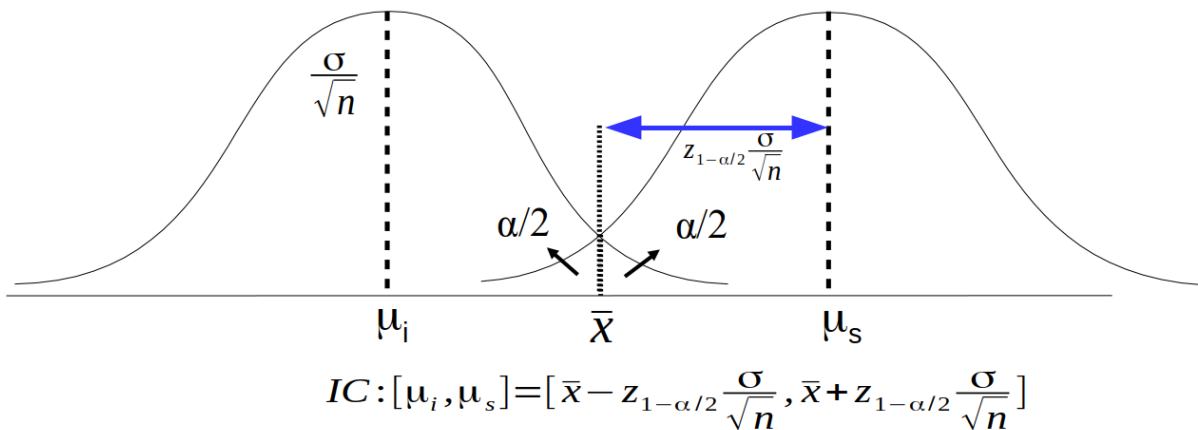


Figura 15.6: Interpretação alternativa dos limites do intervalo de confiança. Sendo  $\bar{x}$  a média amostral,  $\mu_i$  e  $\mu_s$  correspondem aos limites inferior e superior do intervalo de confiança para a média da população.

A aplicação [Limites do intervalo de confiança distribuição normal](#) (figura 15.7) permite ao usuário encontrar os limites do intervalo de confiança para a média de uma população com uma distribuição normal, com variância conhecida. Ao especificar o desvio padrão da população, o tamanho amostral, o nível de confiança e a média amostral, o usuário pode variar os valores do limite superior e limite inferior. O valor de  $\alpha/2$  corresponde à área em vermelho para cada um dos limites. Quando o valor de  $\alpha/2$  para o correspondente

limite for igual a  $(1 - \text{nível de confiança}/100)/2$ , a área vermelha ficará verde e o valor do correspondente limite do intervalo de confiança será mostrado em azul na tela, como visto na figura 15.8. Experimente.

#### Limites do intervalo de confiança

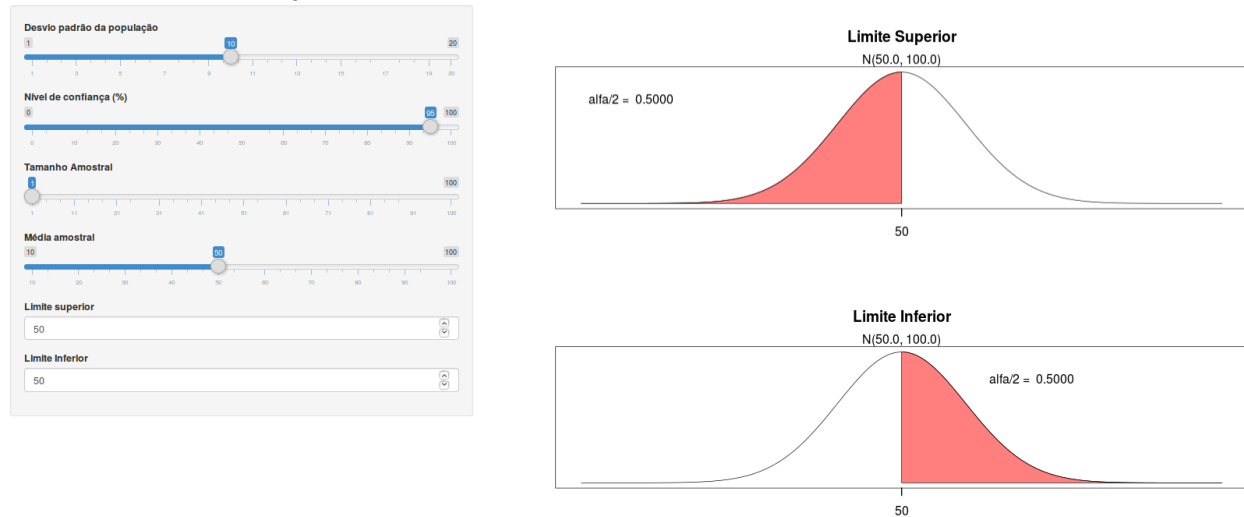


Figura 15.7: Aplicação que permite encontrar os limites do intervalo de confiança, manipulando os valores do limite superior e inferior à esquerda da tela, uma vez selecionados os valores do nível de confiança, desvio padrão da população, média amostral e tamanho da amostra.

#### Limites do intervalo de confiança

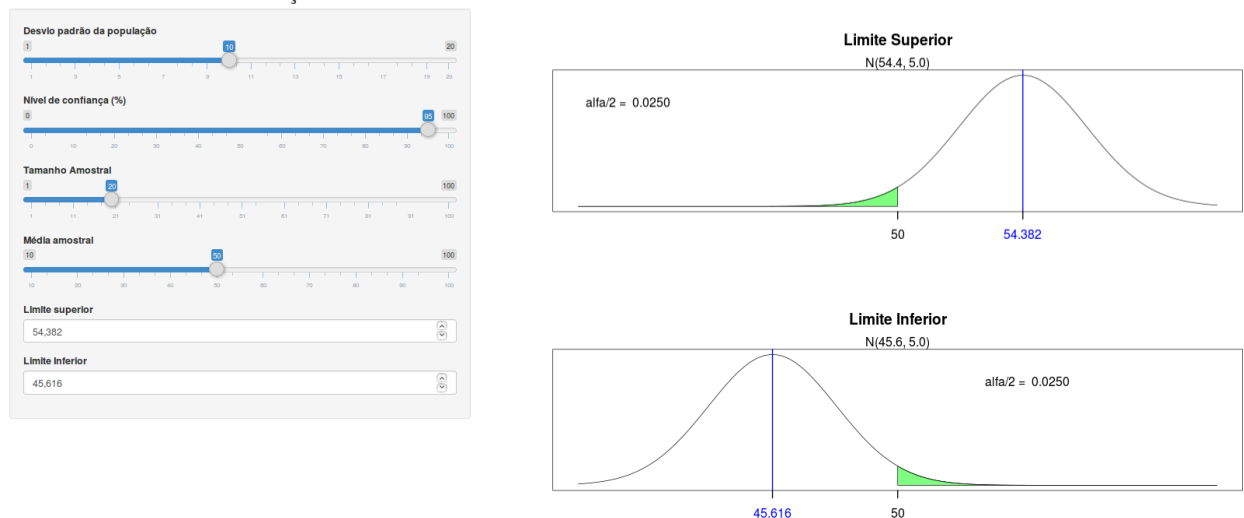


Figura 15.8: Usando a aplicação da figura 15.7, com o nível de confiança igual 95%, desvio padrão = 10, tamanho amostral = 20 e média amostral = 50, os limites do intervalo de confiança são limite inferior = 45,616 e limite superior = 54,382.

Essa interpretação alternativa para o intervalo de confiança fornece um método para o cálculo dos limites do intervalo para situações nas quais esses limites não são expressos por expressões

analíticas como mostrado na seção seguinte.

## 15.6 IC para proporções em pequenas amostras

Para a proporção de sucessos em um experimento de Bernoulli, usamos na seção 14.7 uma aproximação pela normal. Conforme já vimos, essa é uma boa aproximação para um número grande de experimentos. Para amostras pequenas, ou mesmo para amostras grandes, podemos obter o intervalo de confiança exato por meio do cálculo de probabilidades de um modelo binomial.

Por exemplo, consideremos 8 experimentos de Bernoulli, sendo que o evento de interesse tenha ocorrido 6 vezes ( $\hat{p} = 0,75$ ). Caso utilizássemos a aproximação para a normal para calcularmos o intervalo de confiança da proporção de sucessos (nível de 95%), obteríamos:

$$0,75 - 1,96\sqrt{\frac{0,75(1-0,75)}{8}} \leq p \leq 0,75 + 1,96\sqrt{\frac{0,75(1-0,75)}{8}}$$

$$0,4500 \leq p \leq 1,200$$

ou, usando a correção de continuidade:  $0,3875 \leq p \leq 1,2625$ .

Obviamente, a probabilidade  $p$  não pode ser maior do que 1. O intervalo exato com nível de confiança  $(1 - \alpha)$  pode ser obtido da seguinte forma:

- 1) o limite inferior do IC é obtido aplicando a distribuição binomial para 8 experimentos e experimentando com valores da probabilidade de sucesso até que a probabilidade de se observar um número de eventos maior ou igual a 6 seja  $\alpha/2$ ;
- 2) o limite superior do IC é obtido aplicando a distribuição binomial para 8 experimentos e experimentando com valores da probabilidade de ocorrência de sucesso até que a probabilidade de se observar um número de eventos menor ou igual a 6 seja  $\alpha/2$ .

A aplicação [Limites do intervalo de confiança distribuição binomial](#) (figura 15.9) permite ao usuário encontrar os limites do intervalo de confiança para a probabilidade de uma distribuição binomial. Ao especificar o número de experimentos de Bernoulli, o número de ocorrências do evento de interesse, e o nível de confiança, o usuário pode variar os valores do limite superior e limite inferior da probabilidade de ocorrência do evento. O valor de  $\alpha/2$  é igual à soma das probabilidades dos segmentos em vermelho. Quando o valor de  $\alpha/2$  para o correspondente limite for igual a  $(100 - \text{nível de confiança})/2$  %, os segmentos para valores acima ou abaixo do número de ocorrência do evento ficarão verdes e os valores dos correspondentes limites do intervalo de confiança serão as probabilidades selecionadas na tela à esquerda (figura 15.10). Experimente!

### Limites do intervalo de confiança

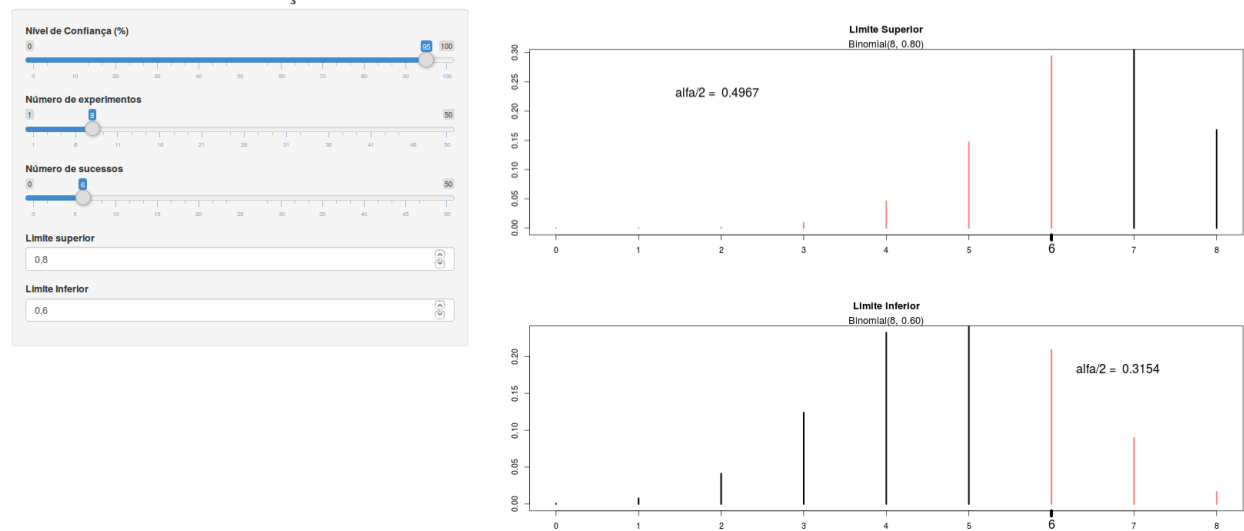


Figura 15.9: Aplicação que permite encontrar os limites do intervalo de confiança para a probabilidade de uma distribuição binomial, manipulando os valores do limite superior e inferior à esquerda da tela, uma vez selecionados os valores do nível de confiança, número de experimentos de Bernoulli e número de sucessos.

### Limites do intervalo de confiança

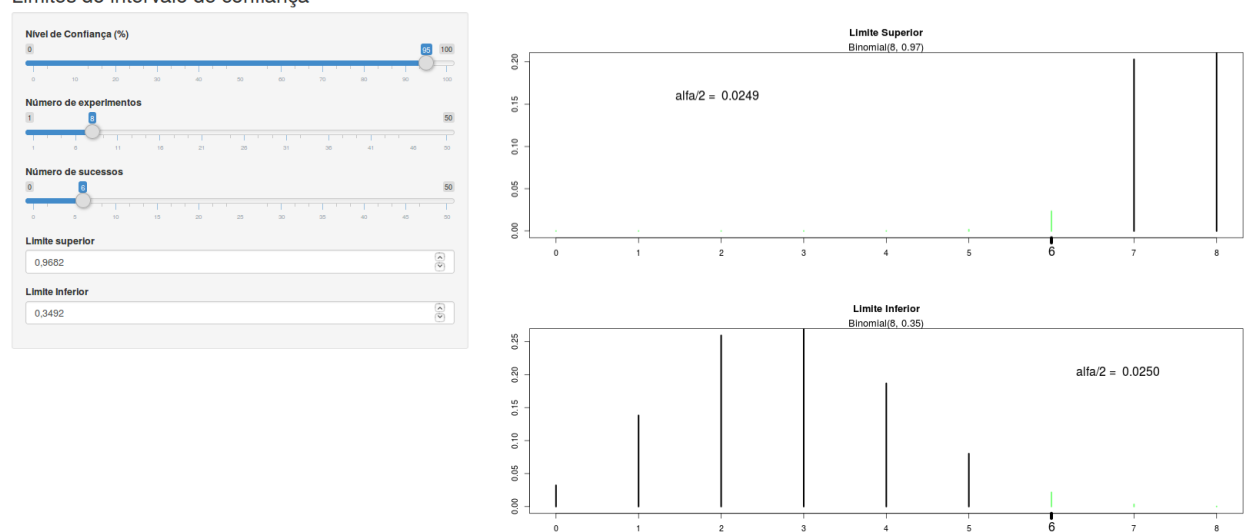


Figura 15.10: Usando a aplicação da figura 15.9, com o nível de confiança igual 95%, número de experimentos = 8 e número de sucessos = 6. Os limites do intervalo de confiança são mostrados no painel à esquerda.

Para o exemplo considerado, o intervalo de confiança exato é dado por:

$$0,3492 \leq p \leq 0,9682$$

O limite inferior é ligeiramente inferior ao obtido utilizando a aproximação da normal com

a correção de continuidade, mas o limite superior é bastante inferior ao calculado pela aproximação, e obviamente menor que 1.

## 15.7 Tipos de testes (bilateral ou unilateral)

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

Existem duas formas de realizarmos um teste estatístico. Se nosso interesse for em avaliar se há uma diferença em relação ao valor definido para a  $H_0$ , tanto positiva como negativa, temos o **teste bilateral (two-sided test)**. No exemplo das seções anteriores, esse foi o tipo de teste que realizamos, porque, para rejeitá-la, é preciso um desvio suficiente grande em relação à hipótese nula em qualquer sentido.

Às vezes, porém, pode-se supor que uma diferença real possa ocorrer somente em um sentido, de tal forma que, se ocorrer uma diferença no outro sentido, isso é devido ao acaso. Nesse caso, a hipótese alternativa se restringe a um efeito em um único sentido. Por exemplo, vamos supor que um medicamento esteja sendo comparado com o placebo para o tratamento de alguma condição de saúde, e um desfecho que está sendo avaliado é algum efeito adverso que esse medicamento possa causar. Os investigadores acreditam que, se houver diferença na proporção de efeitos adversos devido ao medicamento em relação ao placebo, o medicamento deverá ter uma maior proporção de efeitos adversos do que o placebo, e qualquer diferença observada no sentido contrário é devida ao acaso. Sendo  $RR$  o risco relativo para o efeito adverso do medicamento em relação ao placebo, a hipótese nula será  $RR \leq 1$ , e a hipótese alternativa será  $RR > 1$ , ou seja, a hipótese nula será rejeitada somente se o  $RR$  foi maior do que valor crítico, determinado a partir do nível de significância ( $\alpha$ ) e da distribuição da estatística utilizada. Nesse caso, o teste é chamado de **teste unilateral (one-sided test)**.

A figura 15.11 compara a região crítica de um teste bilateral com a região crítica de um teste unilateral com o mesmo nível de significância, supondo que a região crítica do teste unilateral corresponda a desvios positivos. Nesse teste unilateral, o nível de significância corresponde a uma única cauda da distribuição (área vermelha + área rosa na cauda superior), enquanto que, em um teste bilateral, o valor de  $\alpha$  é dividido entre as duas caudas da distribuição. Podemos verificar que, para valores entre  $t_{1-\alpha}$  e  $t_{1-\alpha/2}$ , a hipótese nula seria rejeitada no teste unilateral, mas não seria rejeitada em um teste bilateral. Daí a importância de se especificar o tipo de teste de antemão, para que o resultado do estudo não influencie a escolha do tipo de teste.

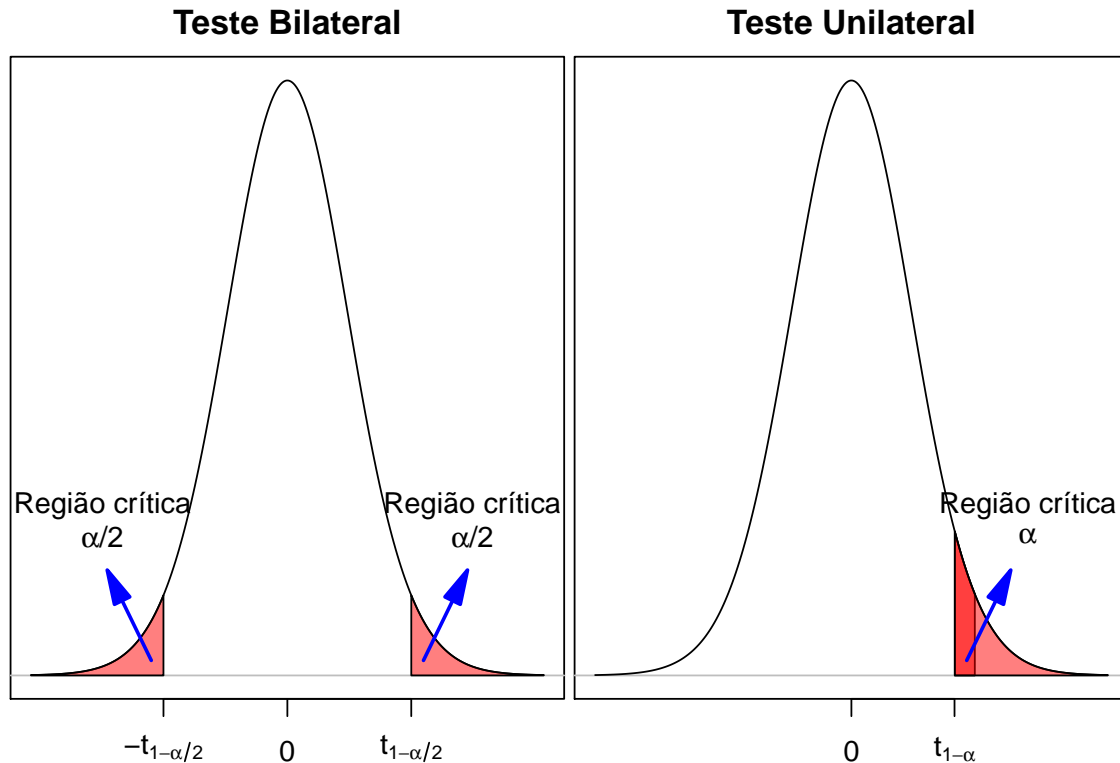


Figura 15.11: Comparação das regiões críticas de um teste unilateral (à direita) e bilateral (à esquerda). Para um teste unilateral superior, a região crítica é a união das áreas de cor vermelha e rosa na cauda superior do gráfico.

### 15.7.1 Exemplos de testes unilaterais

Vamos fazer uma pequena alteração no cenário que utilizamos na seção 15.3. Vamos supor que a distribuição dos valores da glicemia de jejum em uma população de pessoas não diabéticas seja gaussiana ou normal, mas que queiramos testar a hipótese de que **a média de glicemia de jejum nessa população não seja maior do que 85 mg/dl**, a partir de uma amostra de tamanho 36, extraída dessa população. Um teste de hipótese formalizado dessa forma é chamado de teste unilateral, porque nós iremos rejeitar a hipótese nula se o valor da média de glicose obtida na amostra for suficientemente maior do que a média estabelecida pela hipótese nula. A diferença desse cenário para o anterior é que, em um teste bilateral, a hipótese nula era de que o parâmetro avaliado, nesse caso a média da glicose, era igual a um dado valor (85 mg/dl).

Os passos para a realização de um teste de hipótese unilateral são os mesmos de um teste bilateral. Nesse exemplo, no primeiro passo, nós expressamos a hipótese nula, dizendo que a variável glicemia de jejum, denominada por  $X$ , segue uma distribuição normal, onde a média é menor ou igual a 85 mg/dl:

$H_0$ : glicemia de jejum  $\sim$  distribuição normal  $N(\mu, \sigma^2)$ , com  $\mu \leq 85 \text{ mg/dl}$

$H_1$ : glicemia de jejum  $\sim$  distribuição normal  $N(\mu, \sigma^2)$ , com  $\mu > 85 \text{ mg/dl}$

No segundo passo, vamos utilizar a mesma estatística utilizada no teste bilateral. Vamos selecionar uma amostra aleatória da população com um certo número de elementos ( $n$ ), vamos calcular a média da amostra e vamos obter a estatística  $T$  a partir da amostra:

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \quad (15.5)$$

Vamos, em seguida, escolher o nível de significância igual a 5%. Com isso, e considerando que a estatística  $T$ , definida acima, segue a distribuição  $t$  de Student com  $n-1$  graus de liberdade, a figura 15.12 mostra o gráfico da distribuição de  $t$ , supondo que a hipótese nula seja verdadeira, e um ponto correspondente ao quantil  $t_{n-1,1-\alpha}$  da distribuição  $t$  que delimita uma região (em vermelho) cuja área é igual a  $\alpha$  (a probabilidade de se obter um valor de  $t$  acima  $t_{n-1,1-\alpha}$ ). Essa é a região crítica desse teste unilateral, correspondente à região de rejeição da hipótese nula.

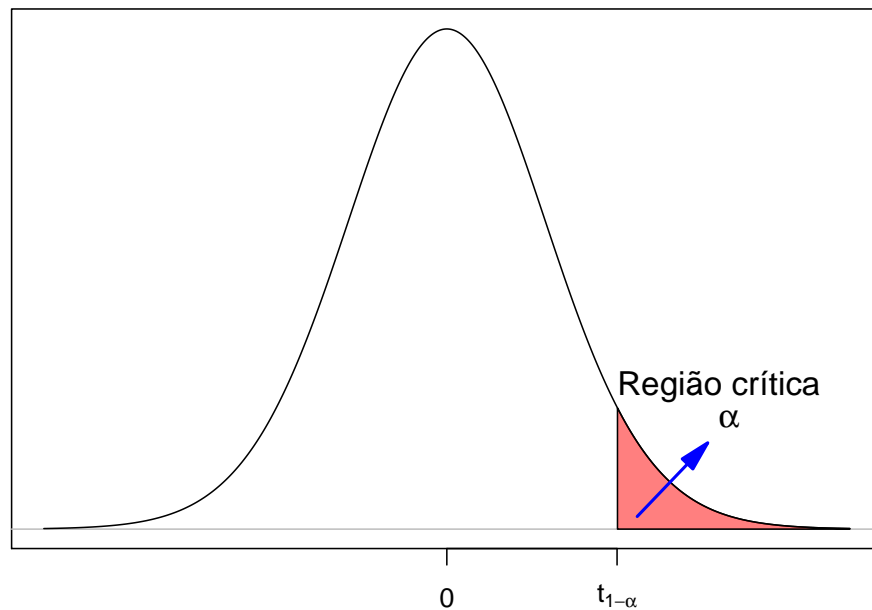


Figura 15.12: Região crítica na cauda superior em um teste unilateral.

Para esse exemplo, o valor crítico  $t_{n-1,1-\alpha} = t_{35;0,95}$  é calculado no R da seguinte forma ( $n = 36$ ):

```
qt(.95, 35)
```

```
## [1] 1.689572
```

Então  $t_{35;0,95} = 1,69$ .

Ao definirmos a estatística a ser utilizada, o valor de  $\alpha$  e, conseqüente, a região crítica do teste, procedemos à seleção da amostra do estudo, a partir da qual o valor da estatística será calculado e comparado com o valor crítico.

Vamos supor que a amostra da população gerou os seguintes resultados:

$$\bar{x} = 92 \text{ mg/dl}$$

$$n = 36$$

$$s = 16 \text{ mg/dl}$$

Substituindo os valores na expressão (15.4), obtemos:

$$t = \frac{92 - 85}{\frac{16}{\sqrt{36}}} = 2,62$$

Como esse valor de  $t$  cai na região crítica (figura 15.13), nós rejeitamos a hipótese nula e dizemos que o resultado é estatisticamente significativo.

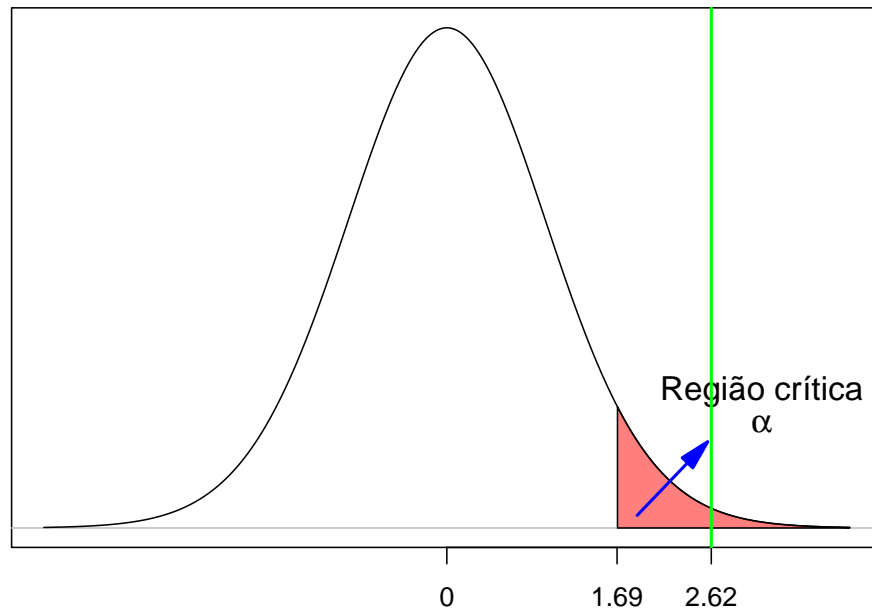


Figura 15.13: Exemplo de teste unilateral onde ocorre a rejeição da hipótese nula. A linha verde indica o valor da estatística  $t$  calculada a partir da amostra.

Vamos supor uma outra situação onde a amostra coletada gerou os seguintes resultados:

$$\bar{x} = 89 \text{ mg/dl},$$

$$n = 36,$$

$$s = 16 \text{ mg/dl},$$

Substituindo os valores na expressão (15.4), obtemos:

$$t = \frac{89 - 85}{\frac{16}{\sqrt{36}}} = 1,50.$$



Como esse valor de  $t$  é menor que o valor crítico (figura 15.14), nós não rejeitamos a hipótese nula e dizemos que o resultado não é estatisticamente significativo.

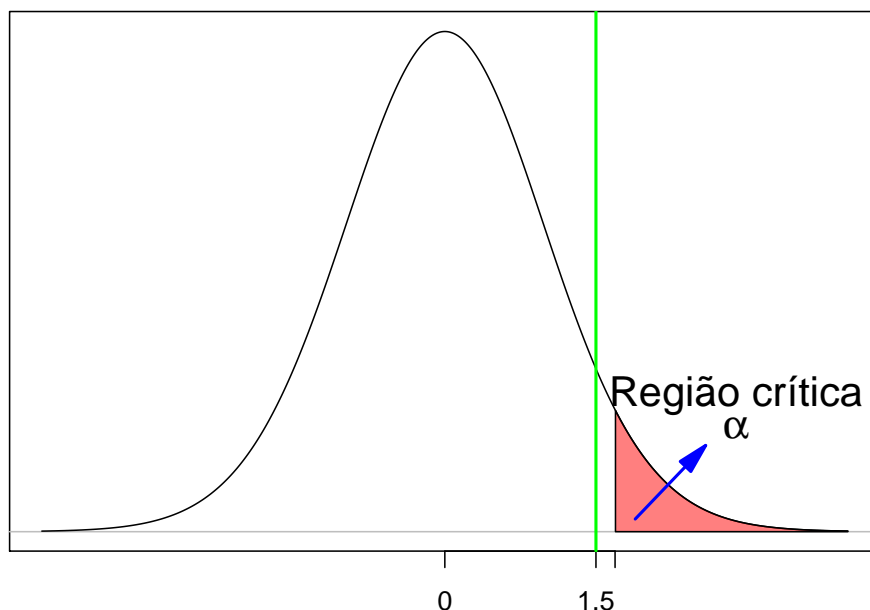


Figura 15.14: Exemplo de teste unilateral onde não ocorre a rejeição da hipótese nula. A linha verde indica o valor da estatística  $t$  calculada a partir da amostra.

Em geral, a menos que haja uma razão plausível para se utilizar um teste unilateral, recomenda-se a utilização de testes bilaterais. Mesmo quando há uma grande expectativa de que uma diferença de efeitos ocorra somente em um sentido, nós não podemos estar certos disso, e devemos considerar todas as possibilidades.

## 15.8 Valor de $p$ (p-value)

O conteúdo desta seção e da seção 15.13 podem ser visualizados neste [vídeo](#).

As probabilidades, sob a hipótese nula, de se obter um valor igual ou maior que o calculado para a estatística do teste ou de se obter um valor igual ou menor do que o calculado para a estatística do teste a partir da amostra são a base para obter o que se denomina valor de  $p$  (*p value* em inglês). Os programas estatísticos usualmente fornecem diretamente o valor de  $p$  quando realizamos testes de hipótese, naturalmente representado pela letra  $p$ .

Em um teste de hipótese para a média de uma variável aleatória  $X$ , que possui uma distribuição normal com média  $\mu$ , mas não sabemos a variância dessa distribuição, calcula-se a estatística abaixo a partir de uma amostra aleatória de tamanho  $n$  da população:

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

A figura 15.15 mostra a distribuição da estatística  $T$ , supondo que  $H_0$  seja verdadeira. O

valor da estatística  $t$  calculada a partir da amostra divide a região sob a distribuição da estatística  $T$  em duas partes.

A área em azul representa a probabilidade de se obter um valor da estatística de teste maior ou igual a  $t$  sob  $H_0$  ( $p_{\text{superior}}$ ). Para um teste unilateral, onde a região crítica é a cauda superior da distribuição da estatística de teste, o valor de  $p$  é igual a  $p_{\text{superior}}$  e representa, portanto, a probabilidade de se obter um valor da estatística de teste maior ou igual ao valor da estatística observado na amostra sob  $H_0$ .

A área em amarelo representa a probabilidade de se obter um valor da estatística de teste menor ou igual a  $t$  sob  $H_0$  ( $p_{\text{inferior}}$ ). Para um teste unilateral, onde a região crítica é a cauda inferior da distribuição da estatística de teste, o valor de  $p$  é igual a  $p_{\text{inferior}}$  e representa, portanto, a probabilidade de se obter um valor da estatística de teste menor ou igual ao valor da estatística observado na amostra sob  $H_0$ .

Para testes bilaterais, um dos critérios para estabelecer o valor de  $p$  é considerá-lo como o dobro da menor das probabilidades  $p_{\text{inferior}}$  e  $p_{\text{superior}}$ . Assim, para testes bilaterais:

$$\text{valor de } p = 2 \cdot \min(p_{\text{superior}}, p_{\text{inferior}})$$

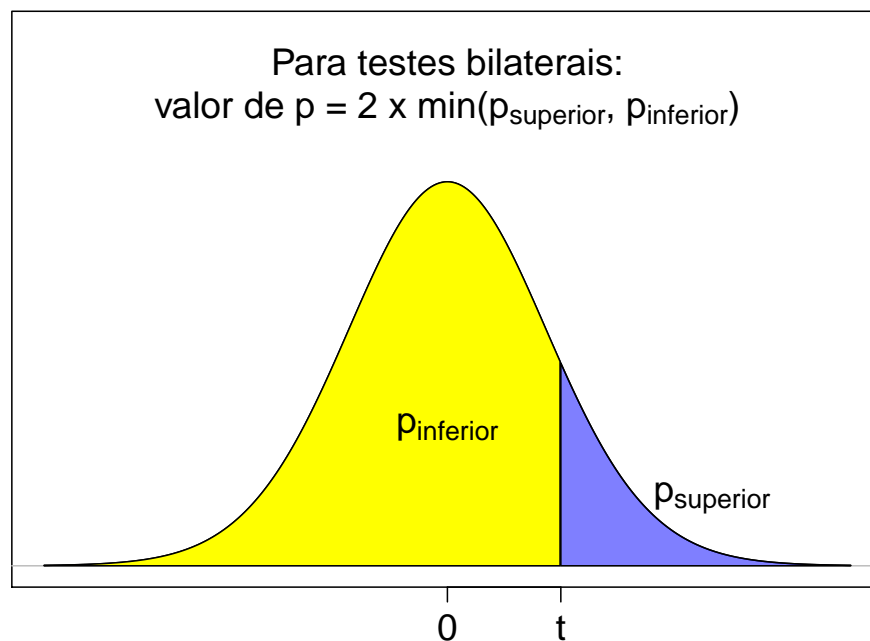


Figura 15.15: Definição do valor de  $p$ . A estatística calculada a partir da amostra,  $t$ , é mostrada no eixo  $X$ , juntamente com as áreas acima e abaixo dessa estatística.

Vamos calcular o valor de  $p$  para o teste de hipótese para a média da glicemia de jejum em uma população de não diabéticos, que possui uma distribuição normal com média  $\mu$ , mas não sabemos a variância dessa distribuição, em quatro situações diferentes, já apresentadas nas seções anteriores.

As duas primeiras situações consideram um **teste de hipótese unilateral**, com nível de significância igual a 5%, com as seguintes hipóteses nulas e alternativas:

$H_0$ : glicemia de jejum  $\sim$  distribuição normal  $N(\mu, \sigma^2)$ , com  $\mu \leq 85 \text{ mg/dl}$

$H_1$ : glicemia de jejum  $\sim$  distribuição normal  $N(\mu, \sigma^2)$ , com  $\mu > 85 \text{ mg/dl}$

Vamos supor que a amostra extraída aleatoriamente da população gerou os seguintes resultados:

$$\bar{x} = 92 \text{ mg/dl}$$

$$n = 36$$

$$s = 16 \text{ mg/dl}$$

$$t_{\text{critico}} = t_{35;0,95} = 1,69$$

Substituindo os valores acima na expressão (15.4), obtemos:

$$t = \frac{92 - 85}{\frac{16}{\sqrt{36}}} = 2,62$$

O valor de  $p$  é a área em azul na figura 15.16, área sob a distribuição além do valor de  $t$  calculado a partir da amostra (2,62), que é a probabilidade de a estatística  $T$  assumir um valor maior ou igual a 2,62. O valor de  $t$  é maior do que o  $t_{\text{critico}}$ . Logo a hipótese nula é rejeitada.

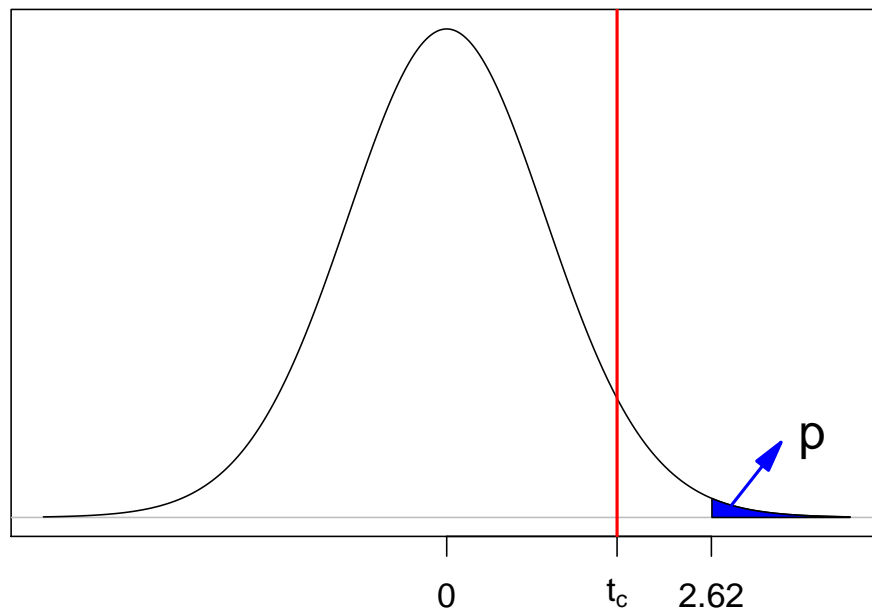


Figura 15.16: Valor de  $p$  para um teste unilateral onde a região crítica é a cauda superior da distribuição. A linha vermelha indica o valor crítico. Nesse exemplo, a hipótese nula é rejeitada.

O valor de  $p$  pode ser obtido no R por meio da expressão:

```
pt(2.62, df = 35, lower.tail = FALSE)
```

```
## [1] 0.006458247
```

Verificamos que o valor de  $p$  (0,006) é menor do que o nível de significância (5%) do teste.

Vamos agora manter o mesmo teste unilateral acima, mas vamos supor que a amostra extraída aleatoriamente da população gerou os seguintes resultados:

$$\bar{x} = 89 \text{ mg/dl}$$

$$n = 36$$

$$s = 16 \text{ mg/dl}$$

$$t_{\text{critico}} = t_{35;0,95} = 1,69$$

Substituindo os valores acima na expressão (15.4), obtemos:

$$t = \frac{89 - 85}{\frac{16}{\sqrt{36}}} = 1,5$$

O valor de  $p$  é a área em azul na figura 15.17, área sob a distribuição além do valor de  $t$  calculado a partir da amostra (1,5), que é a probabilidade de a estatística  $T$  assumir um valor maior ou igual a 1,5. O valor de  $t$  é menor do que o  $t_{\text{critico}}$ . Logo a hipótese nula não é rejeitada.

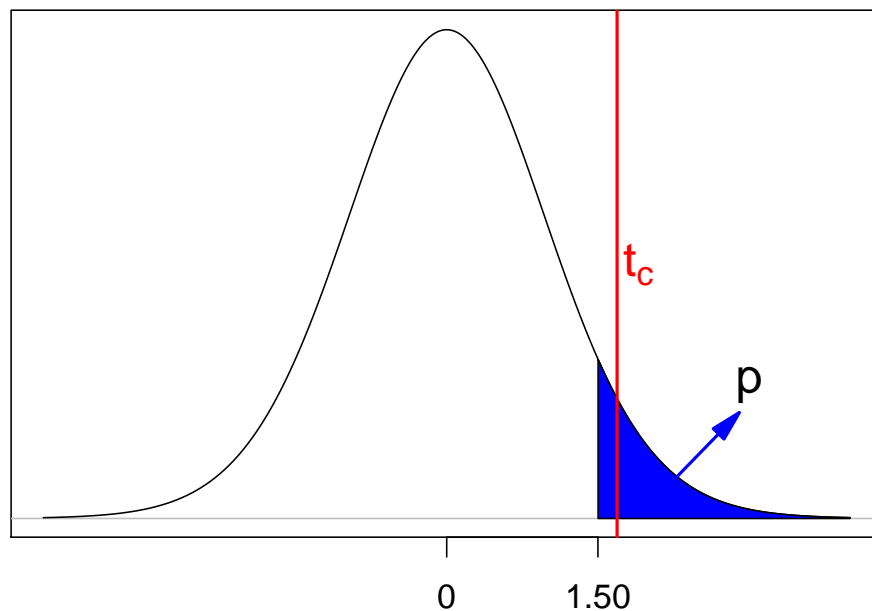


Figura 15.17: Valor de  $p$  para um teste unilateral onde a região crítica é a cauda superior da distribuição. A linha vermelha indica o valor crítico. Nesse exemplo a hipótese nula não é rejeitada.

O valor de p pode ser obtido no R por meio da expressão:

```
pt(1.5, df = 35, lower.tail = FALSE)
```

```
## [1] 0.07129092
```

Verificamos que o valor de p (0,07) é maior do que o nível de significância (5%) do teste.

Vamos considerar agora um **teste de hipótese bilateral**, com nível de significância igual a 5%, com as seguintes hipóteses nulas e alternativas:

$H_0$ : glicemia de jejum  $\sim$  distribuição normal  $N(\mu, \sigma^2)$ , com  $\mu = 85 \text{ mg/dl}$

$H_1$ : glicemia de jejum  $\sim$  distribuição normal  $N(\mu, \sigma^2)$ , com  $\mu \neq 85 \text{ mg/dl}$

Vamos supor que a amostra extraída aleatoriamente da população gerou os seguintes resultados:

$$\bar{x} = 92 \text{ mg/dl}$$

$$n = 36$$

$$s = 16 \text{ mg/dl}$$

$$t_{\text{critico}} = t_{35;0,975} = 2,03$$

Substituindo os valores acima na expressão (15.4), obtemos:

$$t = \frac{92 - 85}{\frac{16}{\sqrt{36}}} = 2,62$$

Esse valor de t é maior do que o  $t_{\text{critico}}$ . Logo a hipótese nula é rejeitada.

O valor de  $p_{\text{superior}}$  é a área em azul na figura 15.18, área sob a distribuição além do valor de t calculado a partir da amostra (2,62), que é a probabilidade de a estatística T assumir um valor maior ou igual a 2,62. O valor de  $p_{\text{inferior}}$  é a área em amarelo na figura 15.18, área sob a distribuição aquém do valor de t calculado a partir da amostra (2,62), que é a probabilidade de a estatística T assumir um valor menor ou igual a 2,62.

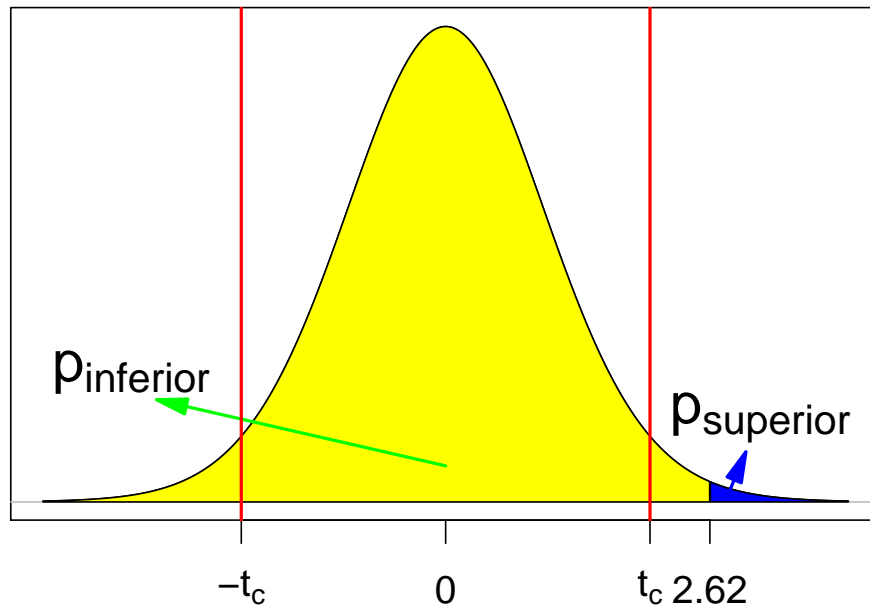


Figura 15.18: Valor de p para um teste bilateral. As linhas vermelhas indicam os valores críticos. Nesse exemplo, a hipótese nula é rejeitada.

O valor de  $p_{\text{superior}}$  pode ser obtido no R por meio da expressão a seguir, dando um valor igual a 0,0065:

```
pt(2.62, df = 35, lower.tail = FALSE)
```

```
## [1] 0.006458247
```

O valor de  $p_{\text{inferior}}$  pode ser obtido no R por meio da expressão a seguir, dando um valor igual a 0,9935 (complemento de  $p_{\text{superior}}$ ):

```
pt(2.62, df = 35, lower.tail = TRUE)
```

```
## [1] 0.9935418
```

Como o valor de  $p_{\text{superior}}$  é menor do que o valor de  $p_{\text{inferior}}$ , o valor de p para esse teste é o dobro do valor de  $p_{\text{superior}}$ , sendo igual a 0,013. Verificamos que o valor de p (1,3%) é menor do que o nível de significância do teste (5%).

Finalmente vamos agora manter o mesmo teste bilateral acima, mas vamos supor que a amostra extraída aleatoriamente da população gerou os seguintes resultados:

$$\bar{x} = 89 \text{ mg/dl}$$

$$n = 36$$

$$s = 16 \text{ mg/dl}$$

$$t_{\text{critico}} = t_{35;0,975} = 2,03$$

Substituindo os valores acima na expressão (15.4), obtemos:

$$t = \frac{92 - 85}{\frac{16}{\sqrt{36}}} = 1,5$$

Esse valor de  $t$  está fora da região crítica. Logo a hipótese nula não é rejeitada.

O valor de  $p_{\text{superior}}$  é a área em azul na figura 15.19, área sob a distribuição além do valor de  $t$  calculado a partir da amostra (1,5), que é a probabilidade de a estatística  $t$  assumir um valor maior ou igual a 1,5. O valor de  $p_{\text{inferior}}$  é a área em amarelo na figura 15.19, área sob a distribuição aquém do valor de  $t$  calculado a partir da amostra, que é a probabilidade de a estatística  $t$  assumir um valor menor ou igual a 1,5.

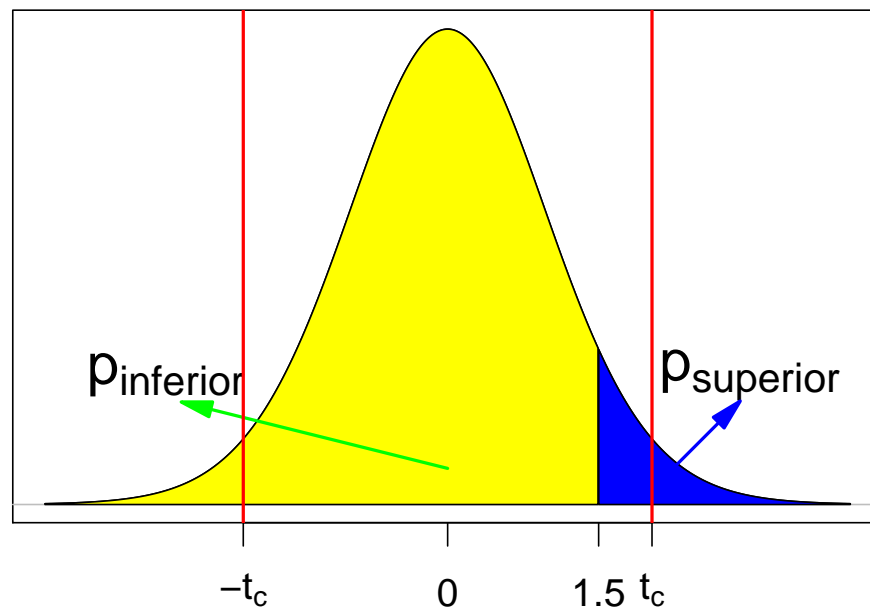


Figura 15.19: Valor de  $p$  para um teste bilateral. As linhas vermelhas indicam os valores críticos. Nesse exemplo, a hipótese nula não é rejeitada.

O valor de  $p_{\text{superior}}$  pode ser obtido no R por meio da expressão a seguir, dando um valor igual a 0,071:

```
pt(1.5, df = 35, lower.tail = FALSE)
```

```
## [1] 0.07129092
```

O valor de  $p_{\text{inferior}}$  pode ser obtido no R por meio da expressão a seguir, dando um valor igual a 0,929 (complemento de  $p_{\text{superior}}$ ):

```
pt(1.5, df = 35, lower.tail = TRUE)
```

```
## [1] 0.9287091
```

O valor de  $p$  para esse teste é o dobro do valor de  $p_{\text{superior}}$ , sendo igual a 0,142. Verificamos que o valor de  $p$  (14,2%) é maior do que o nível de significância do teste (5%).

Ao calcularmos o valor de  $p$  para um teste de hipótese, se ele for menor do que  $\alpha$ , a hipótese nula é rejeitada; se  $p$  for maior ou igual a  $\alpha$ , a hipótese nula não é rejeitada. É sempre mais conveniente apresentar o valor de  $p$  em um teste estatístico do que simplesmente dizer se ele é maior ou menor do que  $\alpha$ . Se o valor de  $p$  for pequeno, significa que os dados amostrais obtidos são muito improváveis de terem ocorrido se a hipótese nula fosse verdade, ou seja, essa hipótese ou é falsa ou temos uma amostra muito improvável.

Para testes bilaterais, há mais de uma proposta para calcular o valor de  $p$ . Uma alternativa à apresentada nesta seção é mostrada na seção 15.12.

## 15.9 Erro tipo I (erro $\alpha$ ) e erro tipo II (erro $\beta$ )

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

Em todo processo decisório onde temos de escolher entre dois cursos de ação, sempre podemos cometer erros. O mesmo acontece em um teste de hipótese, quando utilizamos o processo descrito nas seções anteriores. Basicamente dois tipos de erros podem ocorrer, não simultaneamente:

**1) Erro tipo I (erro  $\alpha$ ):** esse erro ocorre quando **rejeitamos** a hipótese nula quando **de fato ela é verdadeira**. A probabilidade de ocorrer esse erro é  $\alpha$ . Isso pode acontecer quando extraímos uma amostra da população e a estatística calculada a partir dessa amostra **cai** na região crítica. Ao fixarmos  $\alpha$ , fixamos a probabilidade desse erro.

No primeiro cenário da seção 15.3, a hipótese nula foi rejeitada. Nesse caso, podemos ter cometido o erro tipo I.

**2) Erro tipo II (erro  $\beta$ ):** esse erro ocorre quando **não rejeitamos** a hipótese nula quando **de fato ela é falsa**. A probabilidade de ocorrer esse erro é  $\beta$ . Isso acontece quando extraímos uma amostra da população e a estatística calculada a partir dessa amostra **não cai** na região crítica.

No segundo cenário, apresentado na seção 15.3.1, a hipótese nula não foi rejeitada. O valor da estatística  $t$  caiu fora da região crítica (linha vertical verde na figura 15.20) e a hipótese nula não é rejeitada. Se a distribuição verdadeira da glicemia de jejum nessa população tivesse a média 90 mg/dl (e não 85 mg/dl), estaríamos cometendo o erro tipo II, porque não rejeitamos a hipótese nula. O gráfico para a estatística  $t$  para a hipótese  $H_1$  ( $\mu = 90$  mg/dl) seria a curva sob  $H_1$  na figura 15.20. A probabilidade do erro tipo II para a hipótese  $H_1$  é dada pela área do gráfico sob  $H_1$  situada fora da região crítica do teste (área amarela na figura 15.20).



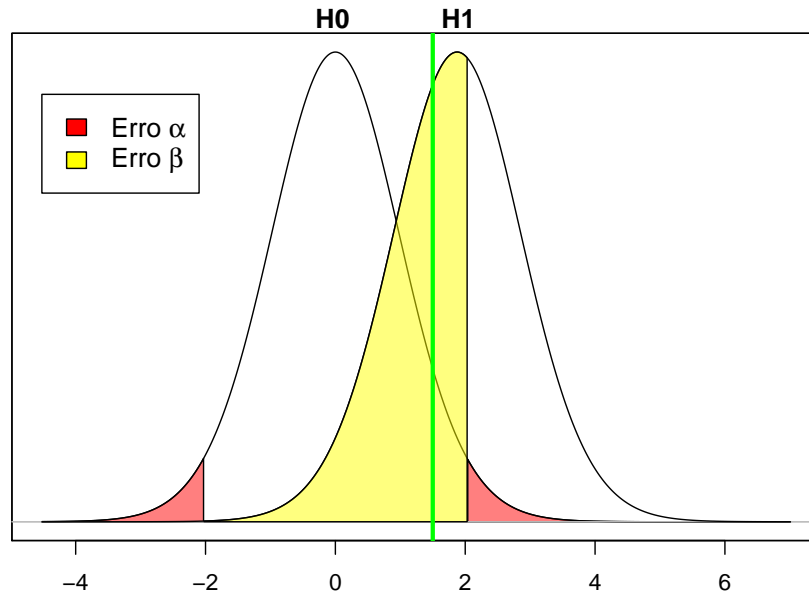


Figura 15.20: O valor da estatística no teste de hipótese é 1,50 (reta vertical verde) e está localizada fora da região crítica do teste. Portanto a hipótese nula não é rejeitada. Se a hipótese  $H_1$  fosse a verdadeira, então estaríamos cometendo o erro  $\beta$ , cuja probabilidade seria a área amarela sob o gráfico definido por  $H_1$ .

Usando o R, a probabilidade do erro tipo II (se a média 90 mg/dl for a verdadeira) é dada pela expressão a seguir e o valor é de 55,4%.

```
pt(2.03, 35, ncp = 1.875, lower.tail=TRUE) -  
  pt(-2.03, 35, ncp = 1.875, lower.tail=TRUE)
```

```
## [1] 0.5541548
```

O valor da função `pt(2.03, 35, ncp = 1.875, lower.tail=TRUE)` fornece a probabilidade de obtermos um valor de  $t$  abaixo de 2,03 (limite superior da região crítica) para a distribuição  $t$  de Student definida por  $H_1$  (por isso o parâmetro `ncp = 1.875`). Vamos chamar esse valor de  $p_1$ . Se utilizássemos `ncp = 0`, estaríamos calculando a probabilidade de  $t$  abaixo de 2,06 para a distribuição  $t$  de Student definida por  $H_0$ .

O valor de `ncp` define o deslocamento da distribuição  $t$  de Student em relação à distribuição sob a hipótese nula. esse deslocamento é calculado, substituindo-se na estatística  $t$  dada pela expressão (15.4) o valor de  $\bar{x}$  pela média da distribuição  $H_1$  (90 mg/dl) para a qual desejamos calcular a probabilidade do erro tipo II e o valor de  $\mu$  pela média da distribuição sob a hipótese nula. Logo:

$$ncp = \frac{90 - 85}{\frac{16}{\sqrt{36}}} = 1,875$$

O valor da função `pt(-2.03, 35, ncp = 1.875, lower.tail=TRUE)` fornece a probabilidade de obtermos um valor de  $t$  abaixo de -2,03 (limite inferior da região crítica) para a distribuição  $t$  de Student definida por  $H_1$ . Esse valor (muito pequeno) tem que ser subtraído do valor  $p_1$ ,

para dar a área da distribuição de  $H_1$  fora da região crítica do teste.

Assim sempre podemos estar cometendo um dos dois erros em um teste de hipótese. Se rejeitarmos a hipótese nula, podemos estar cometendo o erro tipo I. Se não a rejeitarmos, podemos cometer o erro tipo II.

Enquanto a probabilidade do erro tipo I ( $\alpha$ ) é definida a priori pelos investigadores, há infinitos valores para a probabilidade do erro tipo II, dependendo de qual hipótese é a verdadeira (no exemplo acima, ela depende de qual é a verdadeira média da glicemia de jejum na população). Para cada valor diferente para a média real, teríamos um valor diferente para a probabilidade do erro tipo II ( $\beta$ ).

A figura 15.21 mostra a variação da probabilidade do erro  $\beta$  em função da hipótese alternativa. A probabilidade do erro  $\beta$  aumenta quando a média sob a hipótese alternativa se aproxima da média sob a hipótese nula (sendo igual a  $1 - \alpha$  quando a diferença entre elas é infinitesimal, ponto em azul no gráfico). À medida que a média sob a hipótese alternativa se afastar da média sob a hipótese nula, a probabilidade do erro tipo II irá diminuir.

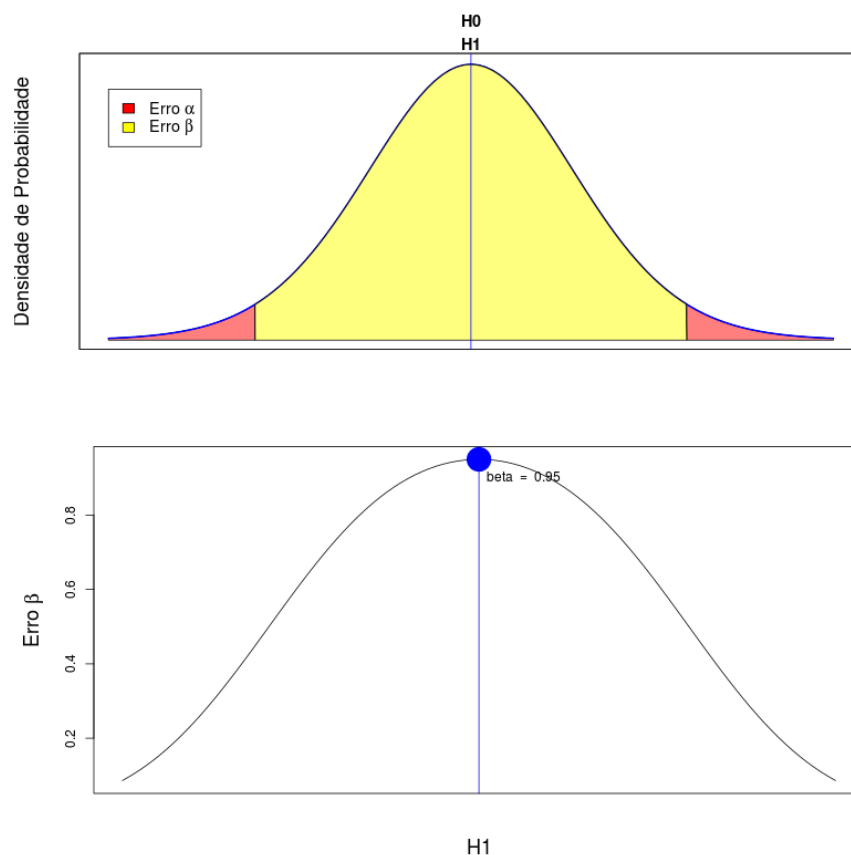


Figura 15.21: O gráfico na parte inferior mostra a relação entre a probabilidade do erro  $\beta$  em função de diferentes hipóteses alternativas ( $\alpha$  foi fixado em 5% neste gráfico). O gráfico na parte superior indica a situação quando a hipótese alternativa difere infinitesimalmente da hipótese nula. Nessa situação, a probabilidade do erro  $\beta$  corresponde ao ponto azul no gráfico na parte inferior.

A aplicação [Erros tipo I e tipo II](#) (figura 15.22) permite ao leitor experimentar com diferentes hipóteses alternativas e valores de  $\alpha$  e verificar a variação da probabilidade do erro tipo II, tomando como hipótese nula uma distribuição normal, com média 85, desvio padrão igual a 16 e tamanho amostral igual a 36 (erro padrão da média amostral é igual a 2,67).

Erros tipo I e tipo II

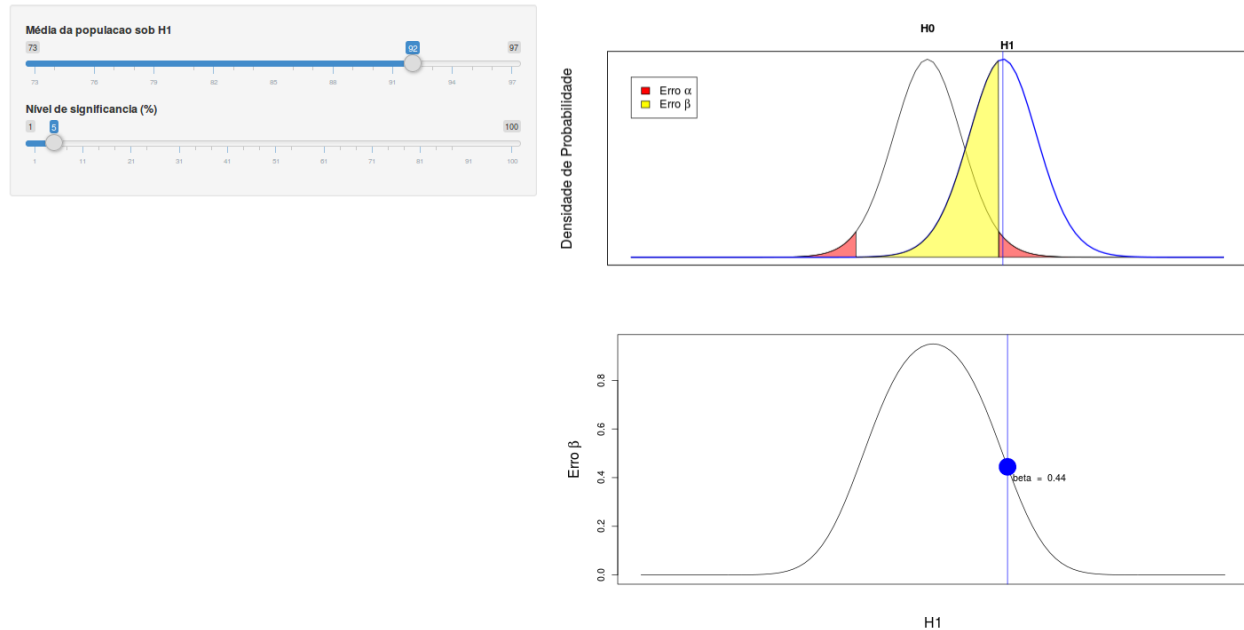


Figura 15.22: Aplicação que mostra a variação da probabilidade do erro tipo II (gráfico da parte inferior), tomando como hipótese nula uma distribuição normal com média 85 e desvio padrão igual a 16, e tamanho amostral igual a 36. O usuário pode selecionar o valor do erro tipo I e diferentes hipóteses alternativas ( $H_1$ , gráfico superior) e verificar a variação da probabilidade do erro tipo II.

A tabela 15.1 resume as quatro situações possíveis que podem ocorrer ao realizar um teste de hipótese.

Tabela 15.1: Situações possíveis em um teste de hipótese.

		Realidade	
		$H_0$ é Verdadeira	$H_0$ é Falsa
Teste	Não Rejeitar $H_0$	Decisão Correta	Erro Tipo II ( $\beta$ )
	Rejeitar $H_0$	Erro Tipo I ( $\alpha$ )	Decisão Correta

## 15.10 Exemplo de um teste hipótese no *R Commander*

Vamos fazer um teste t, utilizando o *R Commander*. Para isso vamos abrir o conjunto de dados *juul2* do pacote *ISwR* (GPL-2 | GPL-3), conforme mostrado no capítulo 3. Vamos testar se a média da variável aleatória *igf1* (*insulin-like growth factor*) é  $360 \mu\text{g/l}$  na população estudada, considerando o nível de significância igual a 10%. Para isso, selecionamos a opção:

Estatísticas  $\Rightarrow$  Médias  $\Rightarrow$  Teste t para uma amostra

Na caixa de diálogo para realização do teste, selecionamos a variável *igf1*, o valor da média para a hipótese nula, o nível de confiança e clicamos em Ok (Figura 15.23).



Figura 15.23: Parâmetros para a realização do teste t para uma amostra: variável, média sob a hipótese nula, nível de confiança ( $1 - \alpha$ ), e tipo de teste (nesse exemplo, a hipótese alternativa é que a média é diferente de  $360 \mu\text{g/l}$ ).

O resultado do teste é mostrado na figura 15.24 abaixo:

```
> with(juul2, (t.test(igf1, alternative='two.sided', mu=360.0, conf.level=.90)))

One Sample t-test

data:  igf1
t = -3.6996, df = 1017, p-value = 0.0002275
alternative hypothesis: true mean is not equal to 360
90 percent confidence interval:
 331.3426 348.9934
sample estimates:
mean of x
 340.168
```

Figura 15.24: Resultado do teste t especificado na figura 15.23.

O resultado mostra o valor da estatística  $t$  calculada para a amostra, o número de graus de liberdade ( $df$ ), o valor de  $p$ , a média da amostra e o intervalo de confiança com nível de confiança de 90%. O resultado é estatisticamente significativo e a hipótese nula é rejeitada.

## 15.11 Poder de um teste e tamanho amostral

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

Nos exemplos apresentados até agora, utilizamos amostras de tamanhos arbitrários. Nesta seção, vamos mostrar por que a escolha do tamanho da amostra, ou amostras, de um estudo é tão importante. Ao realizar uma análise estatística, seja ela um teste de hipótese, seja o cálculo do intervalo de confiança de um parâmetro populacional, estamos interessados em obter erros tipo I ou tipo II os menores possíveis, bem como intervalos de confiança com boa precisão.

Em um teste de hipótese, o erro tipo II ( $\beta$ ) representa a probabilidade de não rejeitarmos a hipótese nula quando ela é falsa. O complemento de  $\beta$  ( $1 - \beta$ ) significa então a probabilidade de rejeitarmos a hipótese nula quando ela é falsa.  $1 - \beta$  é chamado de **poder estatístico do teste**.

Na seção 15.9, vimos que o erro  $\beta$ , e consequentemente o poder estatístico, depende do valor do parâmetro sob a hipótese alternativa. A aplicação [Poder Estatístico e Tamanho Amostral](#), cuja tela inicial é apresentada na figura 15.25, mostra o poder estatístico em função do desvio padrão, do tamanho amostral, do erro alfa e da distância entre a hipótese nula e uma hipótese alternativa possível. A hipótese nula corresponde a uma variável  $X \sim N(100, \sigma^2)$ , ou seja, uma distribuição normal com média 100 e desvio padrão  $\sigma$ , escolhido pelo usuário. A hipótese  $H_1$  possui o mesmo desvio padrão da hipótese nula e média definida pelo usuário. O gráfico inferior mostra o poder estatístico para diferentes valores da média sob  $H_1$  e o ponto em azul mostra o valor do poder do teste para a média sob a hipótese  $H_1$  selecionada pelo usuário no painel à esquerda.

## Poder Estatístico e Tamanho Amostral

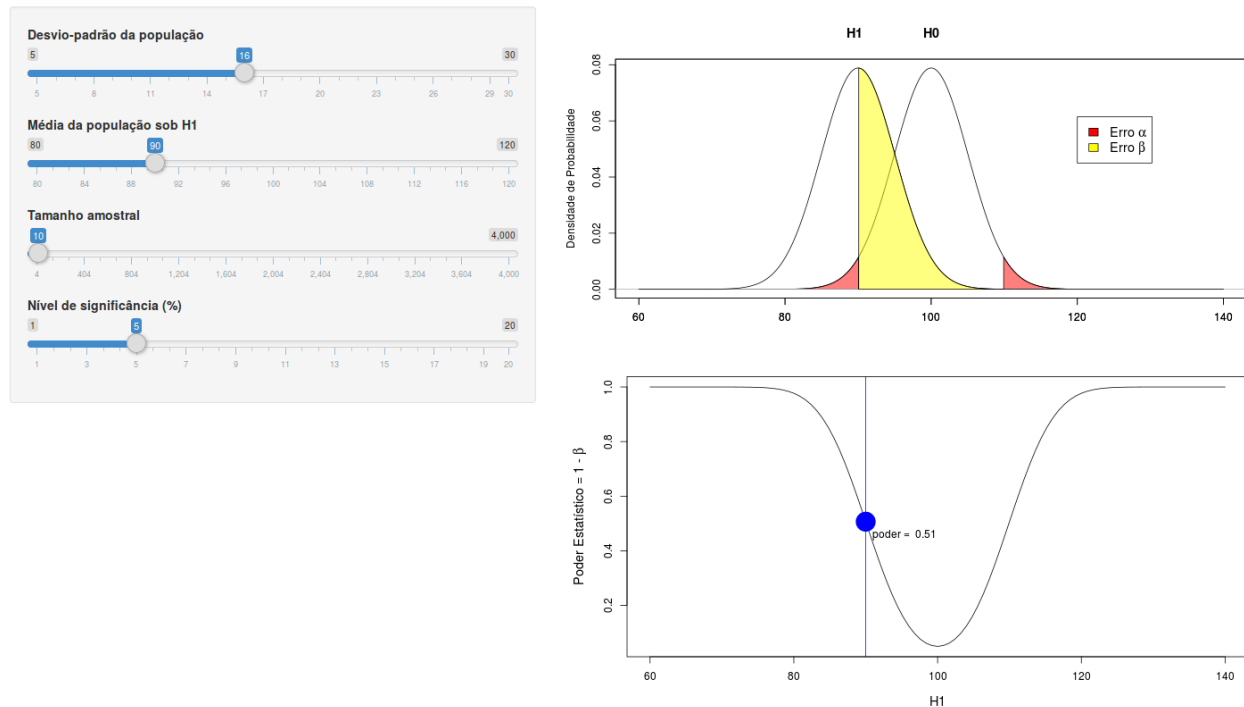


Figura 15.25: Aplicação que mostra o poder estatístico em função do desvio padrão, do tamanho amostral, do erro alfa e da distância entre a hipótese nula e uma hipótese alternativa possível. A hipótese nula corresponde a uma variável  $X \sim N(100, \sigma^2)$ , ou seja, uma distribuição normal com média 100 e desvio padrão  $\sigma$ , escolhido pelo usuário.

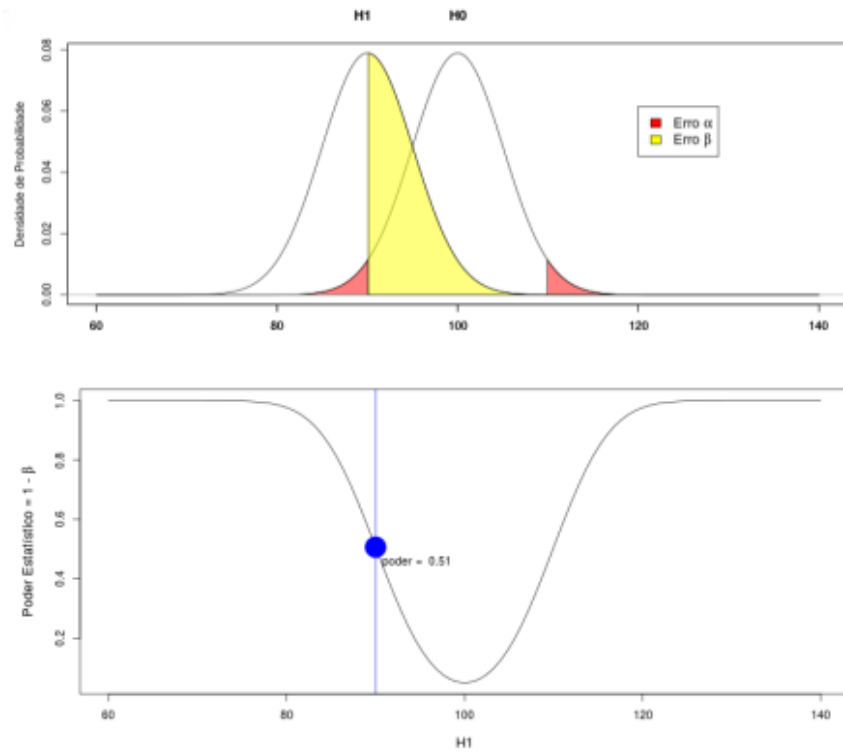
A figura 15.26 ilustra o efeito do desvio padrão sobre o poder estatístico. Ao reduzirmos o desvio padrão à metade, o poder estatístico subiu de 51% para 98%.

A figura 15.27 ilustra o efeito da diferença entre as médias sob a hipótese alternativa e nula, respectivamente, sobre o poder estatístico. Ao dobrarmos o valor da diferença, o poder estatístico também subiu de 51% para 98%.

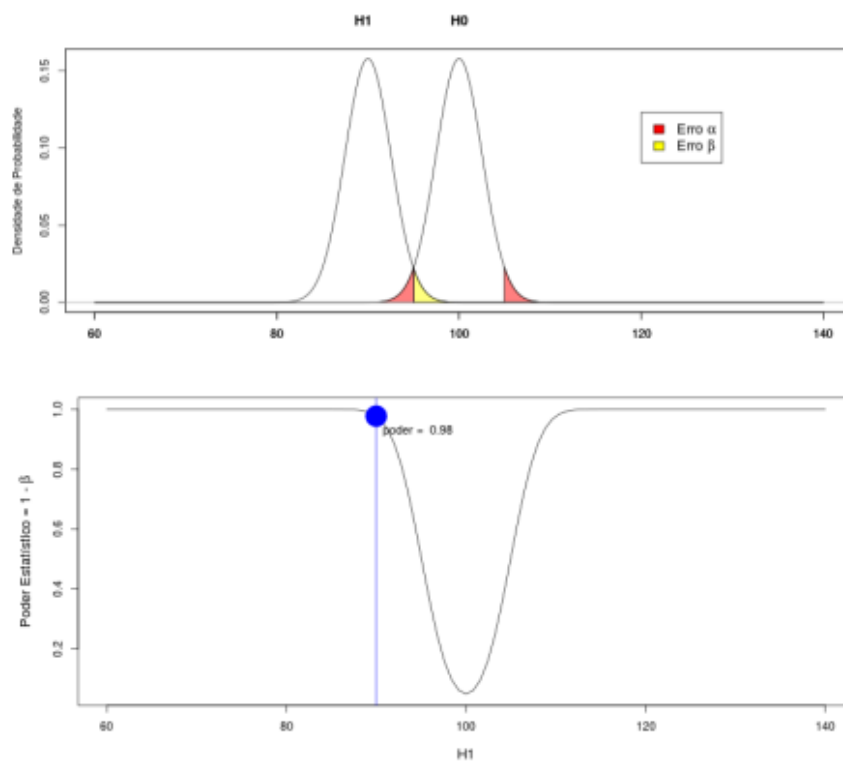
A figura 15.28 ilustra o efeito do tamanho da amostra sobre o poder estatístico. Ao dobrarmos o tamanho da amostra, o poder estatístico subiu de 51% para 80%.

A figura 15.29 mostra o poder estatístico para diferentes valores da média sob a hipótese  $H_1$  e para diferentes tamanhos amostrais. São mostradas três curvas para valores do tamanho amostral iguais a 4, 8 e 16, respectivamente. Observem que, para cada valor da média sob  $H_1$ , o poder estatístico aumenta com o tamanho amostral. Em um estudo experimental, não é possível controlar o desvio padrão da população, mas os investigadores podem ajustar o tamanho amostral, de modo que, se uma dada hipótese alternativa  $H_1$  for verdadeira, então o estudo pode ter um poder estatístico preestabelecido para rejeitar a hipótese nula.

Um exemplo de cálculo de tamanho amostral é apresentado a seguir.

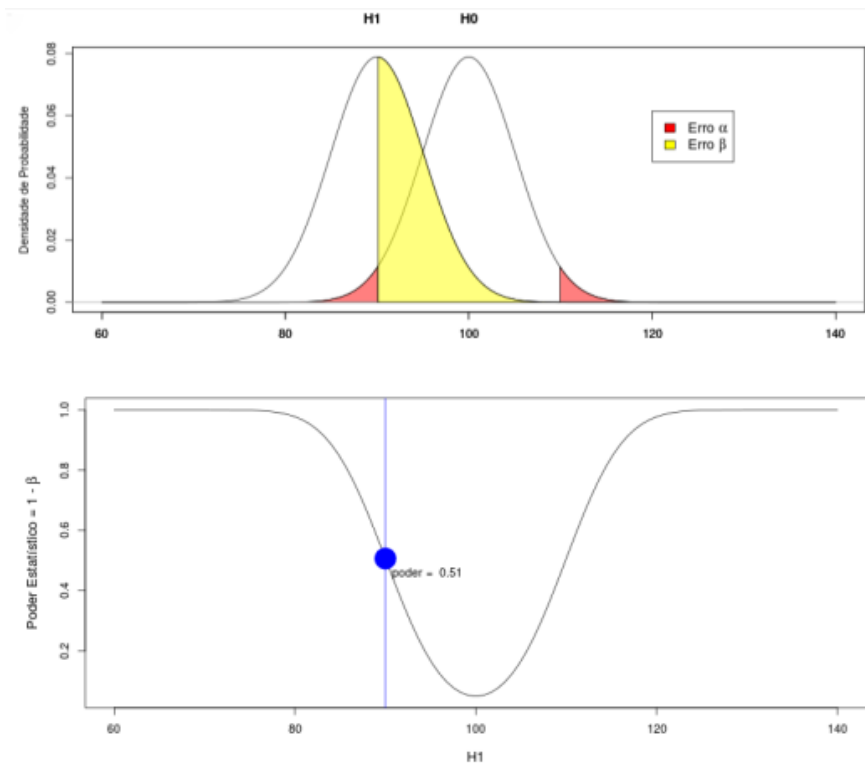


a)  $H_1: N(90, 256), n = 10$

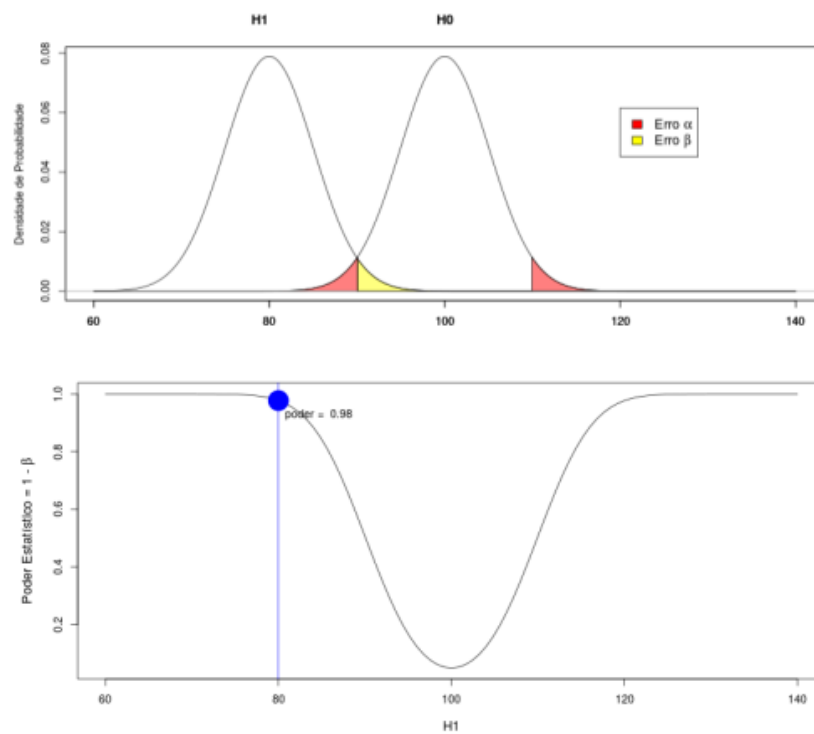


b)  $H_1: N(90, 64), n = 10$

Figura 15.26: Aumento do poder estatístico com a redução do desvio padrão. A figura b corresponde a um desvio padrão igual à metade do desvio padrão da figura a.



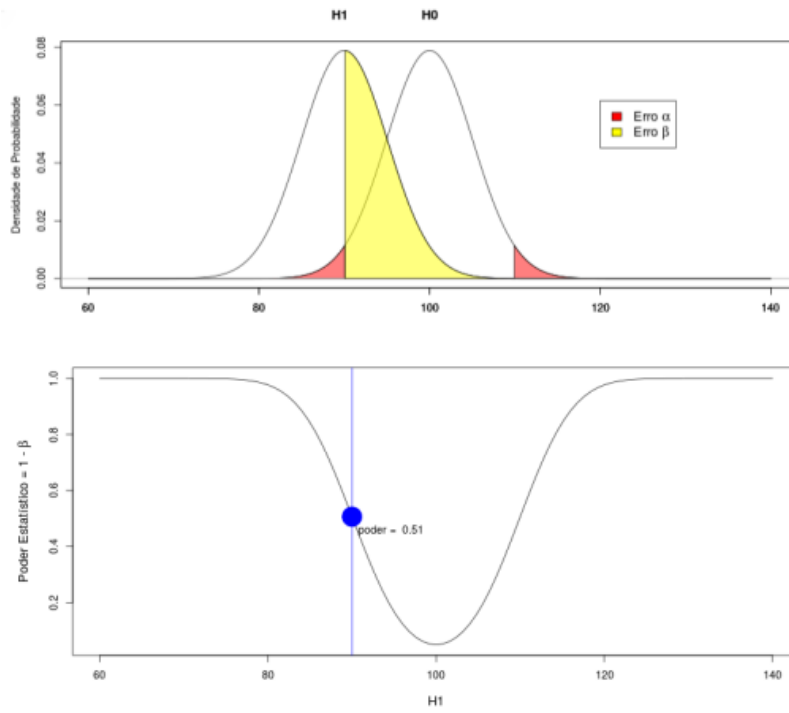
a)  $H_1: N(90, 256), n = 10$



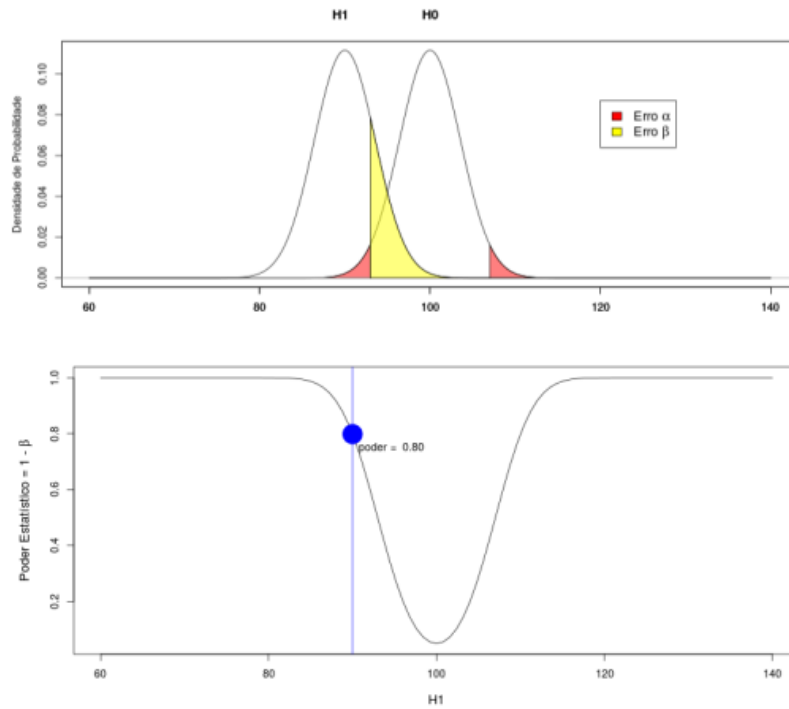
b)  $H_1: N(80, 256), n = 10$

Figura 15.27: Aumento do poder estatístico com o aumento da distância entre a média de  $H_1$  e a média de  $H_0$ . A figura b corresponde a uma distância igual ao dobro da distância na figura a.





a)  $H_1: N(90, 256), n = 10$



b)  $H_1: N(90, 256), n = 20$

Figura 15.28: Aumento do poder estatístico com o aumento do tamanho amostral A figura b corresponde a um tamanho amostral igual ao dobro da amostra na figura a.

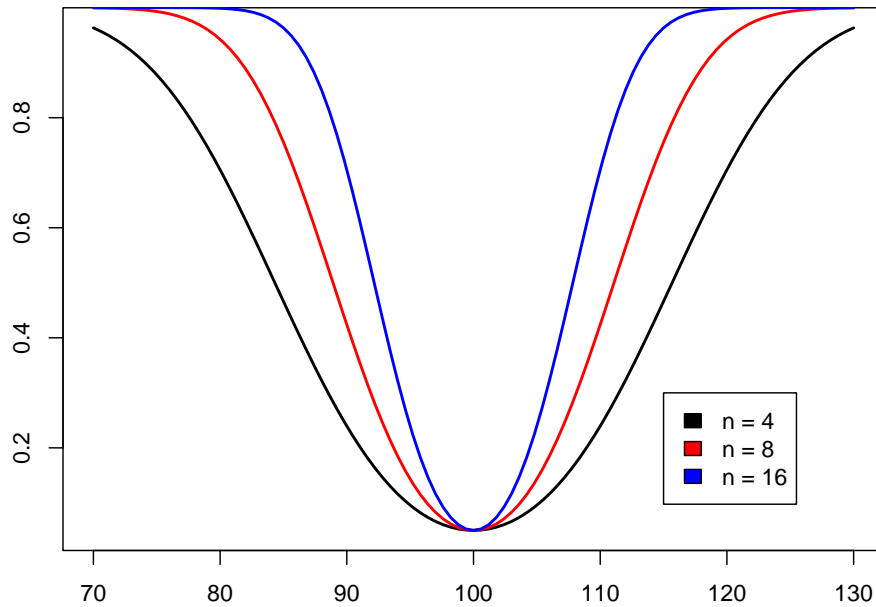


Figura 15.29: Função potência (poder estatístico) para diferentes valores do tamanho da amostra. O poder estatístico aumenta com o tamanho amostral.

### 15.11.1 Cálculo do tamanho amostral

Vamos ilustrar o cálculo do tamanho amostral para a situação onde a hipótese nula é que a amostra venha de uma população normal com média  $\mu_0$  e desvio padrão conhecido  $\sigma$ .

hipótese nula  $\sim N(\mu_0, \sigma^2)$

Vamos supor que, se a média real da população de onde extraímos a amostra seja  $\mu_1$ , então desejamos que a amostra seja tal que tenhamos um poder estatístico  $1 - \beta$  de rejeitar a hipótese nula. Vamos fixar o nível de significância para um teste bilateral igual a  $\alpha$ . A figura 15.30 mostra a distribuição normal para a hipótese nula e para uma hipótese alternativa com média  $\mu_1$ , as áreas correspondentes aos erros alfa e beta, e o valor crítico  $x_c$ .

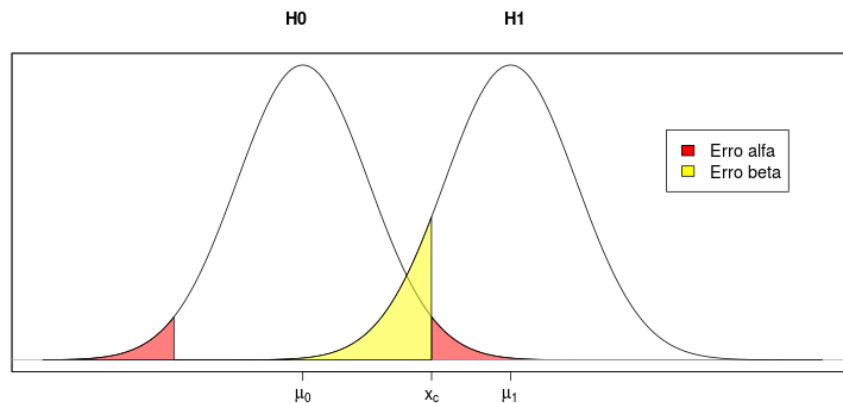


Figura 15.30: Figura auxiliar para o cálculo do tamanho amostral, uma vez fixado o poder estatístico e o erro tipo I.

A equação abaixo expressa o valor do quantil  $1 - \alpha/2$  da distribuição normal padrão ( $z_{1-\alpha/2}$ ), correspondente à área vermelha à direita na figura 15.30, em função de  $x_c$ ,  $\mu_0$ ,  $\sigma$  e do tamanho amostral ( $n$ ):

$$z_{1-\alpha/2} = \frac{x_c - \mu_0}{\frac{\sigma}{\sqrt{n}}} \quad (15.6)$$

A equação abaixo expressa o valor do quantil  $\beta$  da distribuição normal padrão ( $z_\beta$ ), correspondente à área amarela à esquerda na figura 15.30, em função de  $x_c$ ,  $\mu_1$ ,  $\sigma$  e do tamanho amostral ( $n$ ):

$$z_\beta = \frac{x_c - \mu_1}{\frac{\sigma}{\sqrt{n}}} \quad (15.7)$$

Isolando  $x_c$  nas expressões (15.6) e (15.7), obtemos:

$$x_c = z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} + \mu_0 \quad (15.8)$$

$$x_c = z_\beta \frac{\sigma}{\sqrt{n}} + \mu_1 \quad (15.9)$$

Igualando (15.8) e (15.9), temos

$$z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} + \mu_0 = z_\beta \frac{\sigma}{\sqrt{n}} + \mu_1 \Rightarrow (z_{1-\alpha/2} - z_\beta) \frac{\sigma}{\sqrt{n}} = \mu_1 - \mu_0$$

Logo:

$$n = \left[ (z_{1-\alpha/2} - z_\beta) \frac{\sigma}{\mu_1 - \mu_0} \right]^2 \quad (15.10)$$

Confirmando o que foi mostrado na seção anterior, o tamanho amostral aumenta proporcionalmente ao quadrado do desvio padrão, ao inverso do quadrado da diferença entre os valores das médias para as duas hipóteses avaliadas e aumenta à medida que o erro beta e/ou alfa diminuam.

**Exemplo:** vamos considerar o problema do início do capítulo onde  $\mu_0 = 85$  mg/dl,  $\sigma = 16$  mg/dl. Supondo que  $\alpha = 5\%$ , vamos estimar o valor de  $n$  tal que tenhamos um erro  $\beta = 20\%$ , se a média real da população for  $\mu_1 = 92$  mg/dl.

Portanto:  $z_{1-\alpha/2} = 1,96$  e  $z_\beta = -0.84$

Substituindo os valores em (15.10), obtemos:

$$n = \left[ (1,96 - (-0,84)) \frac{16}{92 - 85} \right]^2 \Rightarrow n = 40,96 \Rightarrow n = 41$$

## 15.12 Teste de hipótese para pequenas amostras

Nesta seção, vamos realizar um teste de hipótese para a proporção de ocorrência de um evento que segue a distribuição binomial, porém sem poder recorrer a uma aproximação pela distribuição normal, pelo fato de que o tamanho amostral é pequeno. Nesse caso, temos que recorrer à distribuição exata do parâmetro de interesse.

Vamos considerar então uma população hipotética, de pessoas que sofrem de uma certa doença e que existe um tratamento padrão para essa doença com efetividade de 40%, ou seja 40% das pessoas que sofrem da doença e se submetem ao tratamento são curadas. Dessa forma, consideramos que a probabilidade de extrairmos uma pessoa ao acaso da população de doentes, submetê-la ao tratamento padrão e ela ficar curada é 40%.

Suponhamos agora que temos um tratamento experimental que acreditamos ser mais efetivo que o convencional. Vamos testar esse tratamento em uma amostra aleatória de 20 pessoas da população de doentes. Vamos supor também que as características da população não mudaram, que os pacientes não conseguem distinguir o tratamento novo do convencional e que os critérios para determinar a cura da doença são objetivos, de modo que a avaliação do estado do paciente após o tratamento não seja influenciada pelo tratamento recebido. Suponhamos que, dos 20 pacientes submetidos ao tratamento, 13 foram curados. Que conclusão devemos tirar?

Podemos realizar um teste de hipótese nesse cenário. Vamos seguir os passos típicos de um teste de hipótese.

**PASSO 1:** expressar o tema da pesquisa em termos de hipóteses estatísticas

Nesse exemplo, vamos considerar como hipótese nula a de que o tratamento novo não é melhor do que o convencional. A hipótese alternativa é que o tratamento novo é mais efetivo do que o convencional. Em termos estatísticos, podemos dizer que, para a hipótese nula, a proporção de pessoas curadas com o tratamento novo ( $p$ ) é menor ou igual a 0,4. A hipótese alternativa é que essa proporção é acima de 0,4. Assim temos:

Hipótese Nula ( $H_0$ ):  $p \leq 0,4$

Hipótese Alternativa ( $H_1$ ):  $p > 0,4$

Nesse caso, o teste de hipótese será unilateral.

**PASSO 2:** decidir sobre um teste estatístico apropriado para testar a hipótese nula

Para testar a hipótese nula, vamos utilizar o número de curas observadas no estudo (13). De acordo com a hipótese nula, a variável número de curas em uma amostra de 20 pessoas segue uma distribuição binomial,  $X \sim \text{Binomial}(0,4; 20)$ . Assim a estatística será o número de curas que observamos quando aplicamos o tratamento novo a uma amostra de 20 pessoas.

Vamos chamar de  $X$  a variável número de curas em uma amostra de 20 pessoas extraídas ao acaso da população de doentes e  $p$  a probabilidade de uma pessoa ficar curada após o tratamento padrão. A probabilidade de observarmos  $r$  curas na amostra de  $n = 20$  é dada pela tabela 15.2. O gráfico dessa distribuição é mostrado na figura 15.31.

Tabela 15.2: Probabilidades associadas à observação de  $r$  curas ( $r = 0, 1, \dots, 20$ ) em uma amostra de 20 pessoas, supondo a probabilidade de uma cura igual a 0,4.

<b>X</b>	<b>P(X)</b>
0	0.00003656
1	0.00048749
2	0.00308742
3	0.01234969
4	0.03499079
5	0.07464702
6	0.12441170
7	0.16588227
8	0.17970579
9	0.15973848
10	0.11714155
11	0.07099488
12	0.03549744
13	0.01456305
14	0.00485435
15	0.00129449
16	0.00026969
17	0.00004230
18	0.00000470
19	0.00000033
20	0.00000001

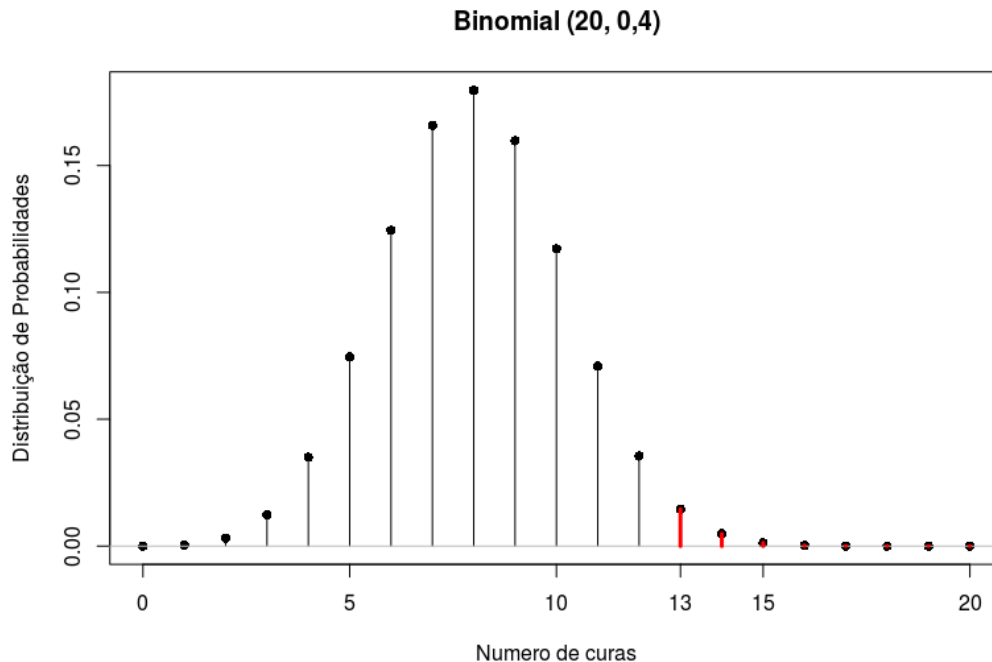


Figura 15.31: Distribuição de probabilidades para o número de curas em uma amostra de 20 pessoas, com  $p = 0,4$ . Em vermelho são as probabilidade de observarmos 13 ou mais curas em uma amostra aleatória de 20 pessoas.

**PASSO 3:** selecionar o nível de significância  $\alpha$

Nesse exemplo, vamos selecionar  $\alpha = 5\%$ .

**PASSOS 4 e 5:** Realizar os cálculos e tomar a decisão

Nesse caso, o valor de  $p$  será a probabilidade de observarmos um número de curas igual ou maior que 13 se a hipótese nula é verdadeira. Esse valor será a soma das probabilidades da tabela II para os valores de  $X$  no intervalo de 13 a 20, ou seja:

$$p = P(X=13)+P(X=14)+P(X=15)+P(X=16)+P(X=17)+P(X=18)+P(X=19)+P(X=20)$$

$$p = 0,0146+0,0049+0,0013+0,00027+0,000042+0,0000047+0,00000033+0,000000011$$

$$p = 0,021 = 2,1\%$$

$$p < \alpha$$

**Assim nós rejeitamos a hipótese nula e concluimos que o tratamento novo é mais efetivo que o convencional.**

**Como calcularíamos o valor de  $p$  se o teste acima fosse bilateral?** Se antes de realizarmos o estudo, tivéssemos a suspeita de que o tratamento novo poderia também ser inferior ao convencional, nós realizaríamos um teste bilateral. Nesse caso, a hipótese nula seria que a efetividade do tratamento novo é igual à do tratamento convencional e a hipótese

alternativa seria de que a efetividade do tratamento novo fosse diferente do tratamento convencional, ou seja:

Hipótese Nula ( $H_0$ ):  $p = 0,4$

Hipótese Alternativa ( $H_1$ ):  $p \neq 0,4$

### Como seria o teste de hipótese nesse caso?

O passo 1 seria alterado, sendo a hipótese nula e a hipótese alternativa como mostradas acima. O passo 2 e o passo 3 continuariam inalterados.

No passo 4, o cálculo do valor de  $p$  seria alterado. Ainda não existe um consenso sobre a forma de calculá-lo. Uma possível proposta é considerar como resultados extremos todos aqueles cuja probabilidade seja igual ou inferior à probabilidade do valor observado (**definição 1**). No exemplo, o valor de  $p$  para o teste bilateral seria então a soma de todas as probabilidades iguais ou inferiores a 0,0146, que é a probabilidade de observarmos 13 curas, supondo que a hipótese nula seja verdadeira.

Desse modo, o valor de  $p$  seria igual à soma:

$$p = P(X=0)+P(X=1)+P(X=2)+P(X=3)+P(X=13)+P(X=14)+P(X=15)+P(X=16)+ \\ P(X=17)+P(X=18)+P(X=19)+P(X=20)$$

$$p = 0,000037 + 0,000487 + 0,0031 + 0,012 + 0,0146 + 0,0049 + 0,0013 + 0,00027 + 0,000042 \\ + 0,0000047 + 0,00000033 + 0,000000011 = 0,0367$$

ou seja  $p = 3,67\%$

Uma outra definição (**definição 2**), é a que utilizamos na seção 15.8, que parte das seguintes definições:

**Valor  $p$  unilateral superior:** é a probabilidade de se observar um valor igual ou acima do valor obtido no estudo, considerando a hipótese nula verdadeira.

**Valor  $p$  unilateral inferior:** é a probabilidade de se observar um valor igual ou abaixo do valor obtido no estudo, considerando a hipótese nula verdadeira.

O valor de  $p$  para um teste bilateral é definido como o dobro do menor dos dois valores  $p$  unilaterais. Por essa definição, o valor de  $p$  seria  $2 \times 2,1\% = 4,2\%$

Existem outras definições do valor de  $p$  para testes bilaterais que não serão aqui abordadas.

Nesse exemplo, a conclusão continua inalterada, uma vez que  $p < \alpha$  para as duas definições de valor de  $p$  mostradas acima.

## 15.13 Interpretações incorretas do valor p

É importante chamar a atenção para algumas interpretações incorretas do valor p e mostrar por que elas estão equivocadas.

**Interpretação errada 1: o valor p é a probabilidade de a hipótese nula ser verdadeira.**

O valor p é justamente calculado com a suposição de que a hipótese nula seja verdadeira e a probabilidade expressa por p indica a compatibilidade entre a estatística observada na análise e a distribuição de probabilidades expressa pela hipótese nula.

**Interpretação nem sempre correta 2:** o valor p bilateral é a probabilidade de se observar um valor tão ou mais extremo que o obtido na amostra se a hipótese nula for verdadeira.

Essa interpretação é correta para o valor p em testes unilaterais. Em testes bilaterais, nem sempre ela produz valores corretos de acordo com as definições de valores p bilaterais apresentadas aqui, apesar de muitos livros textos fazerem essa interpretação. Vamos tomar um exemplo para mostrar por que essa interpretação não gera os mesmos valores de p que as duas definições de valor p bilateral dadas na seção anterior. Esse exemplo é extraído de Rothman e Greenland ((Rothman et al., 2011), página 225). Suponhamos que uma enquête com 1000 pessoas tenha sido realizada para verificar a prevalência de HIV, e os investigadores estão testando a hipótese de que a prevalência de HIV naquela população seja de 0,005 (0,5%), utilizando um teste bilateral e a distribuição binomial. Nessa amostra de 1000 pessoas, 1 pessoa foi identificada com o vírus. Vamos calcular o valor p bilateral, segundo a interpretação acima.

Precisamos em primeiro lugar identificar o que seria um valor tão ou mais extremo que o obtido na amostra. A proporção de ocorrência do evento (pessoa com HIV) sob a hipótese nula é 0,005. Na amostra estudada, ocorreu 1 evento em 1000, logo a proporção de ocorrência do evento na amostra foi de  $1/1000 = 0,001$ . A diferença entre essa proporção e a proporção sob a hipótese nula é:  $0,005 - 0,001 = 0,004$ . Assim valores tão ou mais extremos na amostra seriam aqueles cujas proporções fossem abaixo de ou iguais a 0,001 ou acima de ou iguais a  $0,005 + 0,004 = 0,009$ . Uma proporção igual a 0,009 corresponde a 9 pessoas com HIV em 1000. Assim, para obtermos o valor de p bilateral com essa interpretação, somaríamos as probabilidades de ocorrência de 9 ou mais pessoas com HIV, ou 1 ou menos pessoas com HIV, supondo que a prevalência fosse 0,005. Então o valor p bilateral seria:

$$p = P[X \leq 1] + P[X \geq 9]$$

Usando o R, o valor de  $P[X \leq 1]$  é dado por:

```
pbinom(1, 1000, .005, lower.tail=TRUE)
```

```
## [1] 0.040091
```

O valor de  $P[X \geq 9]$  é dado por:



```
pbinom(8, 1000, .005, lower.tail=FALSE)
```

```
## [1] 0.06760297
```

Assim  $p = 0,04 + 0,067 = 0,11$

Ao final da seção anterior, utilizamos duas definições do valor p bilateral frequentemente utilizadas:

**definição 1:** soma das probabilidades de todos os valores cuja probabilidade seja igual ou inferior à probabilidade do valor observado.

A probabilidade do valor observado (1) é igual a:

```
dbinom(1, 1000, .005)
```

```
## [1] 0.03343703
```

Observamos que as probabilidades de ocorrerem 9 ou 10 pessoas com HIV sob a hipótese nula são:

```
dbinom(9:10, 1000, .005)
```

```
## [1] 0.03613774 0.01799623
```

Assim os valores cujas probabilidades sejam menores do que 0,033 são 0, 1, e 10 em diante. Portanto, por essa definição, o valor de p seria igual a  $P[X \leq 1] + P[X \geq 10]$

```
pbinom(1, 1000, .005, lower.tail=TRUE) + pbinom(9, 1000, .005,
                                                lower.tail=FALSE)
```

```
## [1] 0.07155624
```

**definição 2:** o dobro do menor valor entre o p unilateral superior e o p unilateral inferior.

Nesse caso, o valor de p seria o dobro de  $P[X \leq 1]$

$p = 2 \times 0,04 = 0,08$

Vemos, portanto, que os valores de p para as duas definições dadas não coincidem com o valor de p obtido usando uma interpretação comumente utilizada. Devemos levar em conta, porém, que em distribuições de probabilidades simétricas, a interpretação de p bilateral apresentada aqui geraria o mesmo valor de p do que os obtidos pelas definições 1 e 2 acima. Mesmo em distribuições assimétricas, como o exemplo da seção anterior, a interpretação de p bilateral apresentada aqui geraria o mesmo valor de p do que a definição 1.

**Interpretação errada 3:** O valor de p é a probabilidade de ocorrência da estatística calculada a partir da amostra sob a hipótese nula.

Essa interpretação é totalmente incorreta. O valor de  $p$  unilateral inclui não somente a probabilidade da estatística calculada a partir da amostra sob a hipótese nula, como também as probabilidades sob a hipótese nula de todas as possíveis configurações de dados em que a estatística de teste seja mais extrema que a observada, para cima ou para baixo, dependendo do tipo de teste unilateral. No caso do valor de  $p$  bilateral, vide a discussão para a interpretação errada 2 acima.

**Interpretação errada 4:** Associar o valor  $p$  à significância clínica de um resultado. Vide seção 6.10.

## 15.14 Exercícios

- 1) Um ensaio controlado de um novo tratamento com o placebo levou ao resultado que o tratamento reduz o desfecho adverso estudado em relação ao placebo ( $p < 0,05$ , unilateral). Qual das afirmações abaixo você prefere? Comente.
  - a) Foi provado que o tratamento é melhor que o placebo.
  - b) Se o tratamento não é efetivo, há menos que 5% de probabilidade de se observar em uma amostra uma redução no valor do desfecho clínico menor ou igual à observada no estudo.
  - c) O efeito observado do tratamento é tão grande que há menos que 5% de probabilidade que o tratamento não é melhor que o placebo.
- 2) O que significa o valor  $p$  em um teste de hipótese unilateral? E no bilateral?
- 3) Indique se cada afirmação abaixo é verdadeira ou falsa e justifique a resposta.
  - a) O valor de  $p$  é a probabilidade de se obter um valor improvável.
  - b) O valor de  $p$  está relacionado à qualidade do estudo.
  - c) O valor de  $p$  abaixo do nível de significância indica que o estudo é importante clinicamente.
  - d) Em geral, quanto maior o tamanho amostral, maior é o poder estatístico de um teste.
  - e) O erro tipo II nunca pode ser menor que o erro tipo I.
  - f) O erro tipo II é igual ao poder estatístico de um teste.
  - g) O valor de  $p$  significa a probabilidade de se obter o valor da estatística avaliada na amostra, supondo que a hipótese nula é verdadeira.
  - h) O erro tipo II está relacionado a uma hipótese alternativa, portanto seu valor depende com que hipótese alternativa você está trabalhando.
  - i) O erro tipo I independe da hipótese alternativa.
  - j) O intervalo de confiança é mais informativo do que o valor de  $p$ .
- 4) Suponhamos que você tenha “chutado” todos os itens da questão anterior e você está interessado em estimar as probabilidades de acertar um certo número de questões. Que distribuição de probabilidades você usaria? Qual a probabilidade de acertar 5 questões? E de acertar menos de 2?

- 5) Quais são os passos para se realizar um teste de hipótese?
- 6) As idades de uma amostra aleatória de 50 membros de uma certa população são obtidas e encontra-se que  $\bar{x} = 53,8$  anos e  $s = 9,89$  anos. A idade dessa população segue uma distribuição normal, mas não conhecemos nem a média nem a variância.
  - a) Explique como você utilizaria os dados do enunciado para realizar um teste de hipótese bilateral de que a média de idade da população de membros da população é 52 anos.
  - b) Em face da decisão tomada, que tipo de erro você pode ter cometido?
  - c) Como você obteria o valor de p do teste e o que ele significa?
- 7) O que é nível de significância? Quais são os valores frequentemente usados?
- 8) Qual é o outro nome para o erro tipo II? O que é poder estatístico? Como encontrar o erro tipo II?
- 9) Que fatores influenciam o poder de um teste estatístico?
- 10) Por que intervalos de confiança são mais informativos do que testes de hipótese?
- 11) Um estudo controlado randomizado comparou o levamisol com o placebo para o tratamento de aftas. Os autores apresentaram a tabela exibida na figura 15.32, a qual realiza um teste estatístico (Mann-Whitney) para comparar os dois grupos de pacientes antes do tratamento. Por que isso não faz sentido?

Quadro 1 - Características clínicas dos pacientes antes do tratamento			
Aspecto observado	Levamisol	Placebo	
Média do número de episódios mensais de aftas	2,2	2	Teste de Mann-Whitney p= 0,9068
Média do número de lesões	2,9	2,9	Teste de Mann-Whitney p= 0,9068
Média do tempo de duração das lesões (dias)	12,5	16,4	Teste de Mann-Whitney p= 0,8606
Paciente com dor moderada ou intensa	12	9	Teste Exato de Fisher p= 0,7852

Figura 15.32: Exemplo de uma situação onde um teste de hipótese não faz sentido. Fonte: quadro 1 do estudo intitulado “Levamisol não previne lesões de estomatite aftosa recorrente: um ensaio controlado randomizado, duplo-cego e controlado por placebo” (Weckx et al., 2009) (CC BY).

# Capítulo 16

## Comparação de médias entre dois grupos

### 16.1 Introdução

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

Este capítulo discute técnicas para estudar a relação entre uma variável binária e uma variável quantitativa. Vamos considerar duas situações distintas.

A primeira situação diz respeito à estimativa da diferença das médias dos valores de uma variável numérica em dois grupos distintos de pacientes. O estudo de Haijanen et al. (Haijanen et al., 2019) foi um ensaio controlado randomizado multicêntrico que fez uma comparação de custos de antibióticos x apendectomia para o tratamento da apendicite aguda sem complicações. Parte dos resultados estão mostrados na figura 16.1. Por exemplo, para custos hospitalares em 5 anos de acompanhamento, os autores apresentaram o custo médio para cada grupo de tratamento (2730 x 2056 euros), bem como a diferença de custos entre os dois grupos (674). Ao lado de cada custo, foram mostrados entre parênteses o intervalo de confiança ao nível de 95%. Na última coluna, os autores apresentaram os valores de p resultante dos testes de hipóteses para verificar a significância estatística da diferença entre os custos de cada tratamento.

A análise estatística para a diferença entre os custos médios nos dois grupos de tratamento foi baseada no **teste t de Student**. Diferenças entre os grupos em relação ao tempo de internação e licença médica (não apresentadas na tabela) foram testadas por meio do **teste de Mann-Whitney**.

Os dois grupos de pacientes nesse estudo são chamados **independentes**, porque as unidades de observação (os pacientes) foram alocados aos grupos de maneira independente. Em um arquivo de dados, usualmente uma variável binária é usada para designar o grupo a que cada unidade de análise pertence e outra variável designa a variável numérica que está sendo medida em cada unidade de análise (vide figura 1.9).

**Table 1. Mean hospital charges, productivity losses and overall costs in Euros per patient for appendectomy and antibiotic therapy group patients with uncomplicated acute appendicitis at five-year follow-up.**

	Appendectomy Group € (95% CI, €)	Antibiotic therapy Group € (95% CI, €)	Difference € (95% CI, €)	p<
<b>One-year follow-up</b>				
Hospital charges	2718 (2636–2799)	1707 (1547–1865)	1010 (835–1186)	0.001
Productivity losses	2962 (2806–3118)	1845 (1712–1976)	1117 (911–1322)	0.001
Overall costs	5680 (5489–5872)	3552 (3334–3769)	2127 (1840–2417)	0.001
<b>Five-year follow-up</b>				
Hospital charges	2730 (2645–2817)	2056 (1861–2251)	674 (465–883)	0.001
Productivity losses	2986 (2822–3149)	2115 (1950–2280)	871 (639–1104)	0.001
Overall costs	5716 (5510–5925)	4171 (3879–4463)	1545 (1193–1899)	0.001

<https://doi.org/10.1371/journal.pone.0220202.t001>

Figura 16.1: Comparação de diversos custos entre dois tratamentos para apendicite aguda. Fonte: tabela 1 do estudo de (Haijanen et al., 2019) (CC BY).

Outra situação é quando as duas amostras ou os dois grupos são dependentes. Isso pode ocorrer nos cenários apresentados a seguir.

O primeiro cenário diz respeito à estimativa da diferença de efeitos sobre uma variável numérica em um grupo de pacientes quando dois tratamentos distintos são aplicados em sequência aos pacientes (a ordem de aplicação pode ser aleatória) e, então, uma variável numérica é medida após cada tratamento e os valores da variável após cada tratamento são comparados. Cada conjunto de medidas da variável após cada tipo de tratamento forma um grupo.

Outro cenário é quando uma variável numérica é medida em cada par de indivíduos, sendo que cada par é formado por indivíduos semelhantes de acordo com algum critério estabelecido. Os primeiros elementos de cada par formam um grupo e os segundos elementos de cada par formam o outro grupo.

Um terceiro cenário é quando uma variável numérica é medida em dois instantes diferentes, ou em posições diferentes, em um mesmo grupo de indivíduos e os valores dessa variável nos dois instantes (posições) são comparados. Cada instante, ou cada posição, representa um grupo de medidas.

Por exemplo, o estudo de Andrade et al. (Andrade et al., 2018b) avaliou o modelo de avaliação da homeostase do índice de resistência à insulina (HOMA-IR) em pacientes com hepatite C crônica tratados com medicação antiviral de ação direta na resposta virológica sustentada (RVS). Os dados foram coletados no início do tratamento (t-base) e na décima segunda semana após o término do tratamento (t-RVS12). O HOMA-IR foi calculado como insulinemia de jejum ( $\mu U/mL$ ) x glicemia de jejum (mmol/L)/22,5. A tabela 4 desse estudo (figura 16.2) mostra as médias das diferenças dos valores das variáveis glicemia de jejum, insulinemia de jejum e HOMA-IR entre o início do tratamento e a décima segunda semana após o término do tratamento para pacientes não diabéticos e com valores de HOMA-IR > 2,5. Foi realizado o **teste t pareado** para cada uma dessas médias e os valores de p dos testes são mostrados na última coluna da tabela.

**TABLE 4.** Values of delta glucose, delta insulin and delta HOMA-IR in patients with values of HOMA-IR >2.5 and non-diabetics.

	Mean delta (n=75)	P-value
Delta glucose	2.6 ± 12.49	0.07
Delta insulin	2.9 ± 9.88	0.01
Delta HOMA-IR	0.76 ± 2.81	0.02

Delta: t-base–t-SVR; HOMA-IR: homeostasis model assessment of insulin resistance; t-base: baseline; t-SVR: sustained virological response.

Figura 16.2: Comparação dos valores das variáveis glicemia de jejum, insulinemia de jejum e HOMA-IR entre o início do tratamento e a décima segunda semana após o término do tratamento para pacientes não diabéticos e com valores de HOMA-IR > 2,5. Fonte: tabela 4 do estudo de (Andrade et al., 2018b) ([CC BY-NC](#)).

Nos três cenários acima, dizemos que os grupos são **dependentes**, ou **pareados**, porque os valores da variável numérica tendem a estar correlacionados em cada indivíduo ou em cada par de indivíduos. Por exemplo, no estudo de Andrade et al., indivíduos que possuem valores de glicemia de jejum mais baixos antes do tratamento tendem a ter valores mais baixos de glicemia de jejum após o tratamento do que indivíduos que possuem valores mais altos de glicemia de jejum antes do tratamento.

Em amostras dependentes, há duas variantes para a organização do arquivo de dados. Numa variante, uma variável identifica cada indivíduo, uma segunda variável identifica os grupos (instante ou posição da medida, um dos elementos de cada par de indivíduos, ou tratamento aplicado) e uma variável numérica identifica o desfecho. Na segunda variante, duas variáveis numéricas identificam as duas medidas da variável (uma para um instante - posição da medida, um dos elementos de cada par de indivíduos ou um dos tratamentos aplicados - e outra para o outro instante - posição, elemento do par ou tratamento) (vide figura 1.10). Este capítulo discute as condições para a realização de cada uma das técnicas de análise indicadas acima, começando pela situação onde os dois grupos, ou amostras, são independentes.

## 16.2 Comparação de médias de amostras independentes

O conteúdo desta seção e da seção 16.2.1 podem ser visualizados neste [vídeo](#).

Vamos utilizar o conjunto de dados *energy* da biblioteca *ISwR* ([GPL-2](#) | [GPL-3](#)). Esse conjunto de dados contém o consumo de energia de 22 pessoas, sendo 13 magras e 9 obesas. As duas variáveis são: *expend*, que representa o consumo de energia, e *stature*, que indica se a pessoa é magra ou obesa. Para ler esse conjunto de dados, podemos utilizar o *R Commander*, seguindo o procedimento mostrado no capítulo 3, seção 3.6.1, ou por meio dos comandos:

```
library(ISwR)
data(energy)
```

## energy

```
##      expend stature
## 1      9.21   obese
## 2      7.53    lean
## 3      7.48    lean
## 4      8.08    lean
## 5      8.09    lean
## 6     10.15    lean
## 7      8.40    lean
## 8     10.88    lean
## 9      6.13    lean
## 10     7.90    lean
## 11     11.51   obese
## 12     12.79   obese
## 13      7.05    lean
## 14     11.85   obese
## 15      9.97   obese
## 16      7.48    lean
## 17      8.79   obese
## 18      9.69   obese
## 19      9.68   obese
## 20      7.58    lean
## 21      9.19   obese
## 22      8.11    lean
```

A primeira função carrega a biblioteca *ISwR* e a segunda função carrega o conjunto de dados *energy*. A última função mostra os dados das 22 pessoas.

Em relação ao conjunto de dados *energy*, podemos realizar as seguintes perguntas:

- 1) Existe alguma relação entre o consumo de energia e o fato de a pessoa ser magra ou obesa? Colocados em termos estatísticos, existe diferença estatisticamente significativa entre os níveis de consumo de energia entre pessoas magras e obesas?
- 2) Como podemos quantificar o valor e a precisão dessa diferença?

A figura 16.3 mostra os *boxplots* da variável *expend* para as mulheres magras e obesas respectivamente. Os diagramas sugerem que os consumos de energia, em geral, são maiores nas mulheres obesas do que nas mulheres magras. Vamos analisar esses dados estatisticamente.

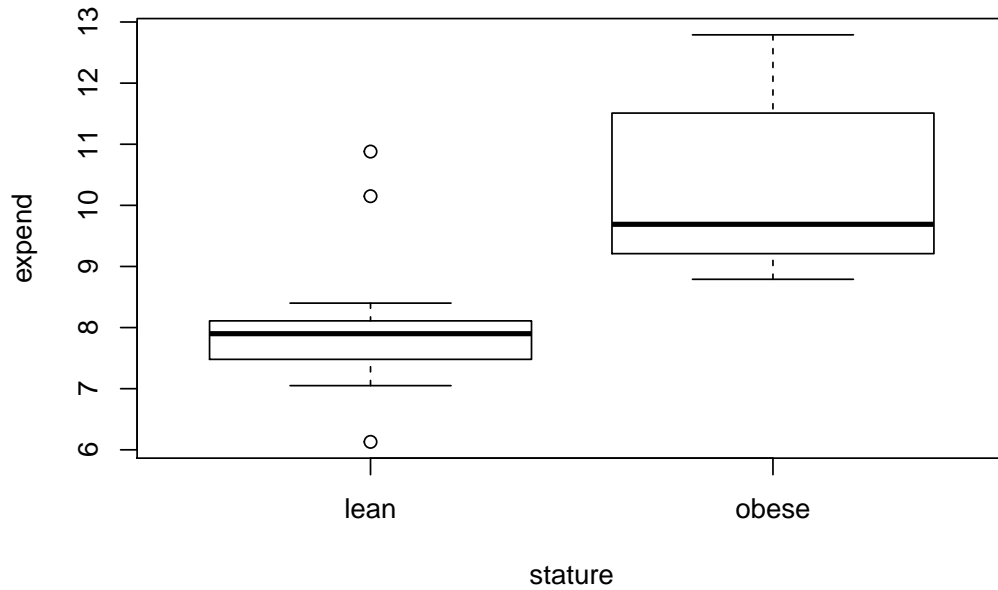


Figura 16.3: *Boxplots* da variável *expend* do conjunto de dados *energy* para as mulheres magras e obesas respectivamente.

De um modo geral, consideremos o seguinte problema: dadas duas populações, 1 e 2, que se distinguem por uma característica (por exemplo, magros e obesos), uma amostra de tamanho  $n_1$  é extraída aleatoriamente da população 1 e uma amostra aleatória de tamanho  $n_2$  é extraída da população 2. Sejam  $X_1$  a variável de interesse (por exemplo, consumo de energia), medida em cada unidade da amostra 1 e  $X_2$  a mesma variável medida nas unidades da amostra 2.

Vamos supor que:

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

e que as variâncias  $\sigma_1^2$  e  $\sigma_2^2$  sejam conhecidas.

Como  $X_1$  e  $X_2$  são variáveis aleatórias independentes, vimos na seção 9.6 que a variável

$$X = X_1 - X_2$$

terá uma distribuição  $N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$ .

Consequentemente, a partir dos resultados da mesma seção 9.6, a diferença de médias amostrais

$$\bar{X} = \bar{X}_1 - \bar{X}_2$$

terá uma distribuição  $N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$ .

Vamos considerar diferentes situações.



Quando as variâncias  $\sigma_1^2$  e  $\sigma_2^2$  são conhecidas, a estatística

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

possui uma distribuição normal padrão (vide capítulo 14, seção 14.2). A estatística

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (16.1)$$

pode ser utilizada para realizar um teste de hipótese bilateral para  $H_0 : \mu_1 - \mu_2 = 0$  ou testes unilaterais para  $H_0 : \mu_1 - \mu_2 \geq 0$  ou  $H_0 : \mu_1 - \mu_2 \leq 0$ .

O intervalo de confiança para  $\mu_1 - \mu_2$ , sendo  $(1 - \alpha)$  o nível de confiança, é dado por:

$$(\bar{X}_1 - \bar{X}_2) - z_{1-\alpha/2} \sigma \leq (\mu_1 - \mu_2) \leq (\bar{X}_1 - \bar{X}_2) + z_{1-\alpha/2} \sigma \quad (16.2)$$

onde

$$\sigma = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Como, em geral, as variâncias não são conhecidas, então elas devem ser estimadas por meio das variâncias amostrais. Sob determinadas condições, uma análise frequentemente utilizada quando não se conhece as variâncias nas duas populações se baseia na distribuição *t de Student*.

### 16.2.1 Teste t de Student para amostras independentes

Quando as variáveis possuem distribuições normais com a mesma variância,  $X_1 \sim N(\mu_1, \sigma^2)$  e  $X_2 \sim N(\mu_2, \sigma^2)$ , mas a variância não é conhecida, um estimador da variância comum pode ser obtido a partir da média ponderada dos estimadores das variâncias nas amostras 1 ( $S_1^2$ ) e 2 ( $S_2^2$ ), com pesos respectivamente iguais a  $n_1 - 1$  e  $n_2 - 1$ :

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

onde:

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_{1i} - \bar{X}_1)^2$$

e

$$S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (X_{2i} - \bar{X}_2)^2$$

A estatística

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (16.3)$$

possui uma distribuição t de Student com  $n_1 + n_2 - 2$  graus de liberdade (gl). A estatística:

$$T = \frac{(\bar{X}_1 - \bar{X}_2)}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

pode ser utilizada para realizar um teste de hipótese bilateral para  $H_0 : \mu_1 - \mu_2 = 0$  ou testes unilaterais para  $H_0 : \mu_1 - \mu_2 \geq 0$  ou  $H_0 : \mu_1 - \mu_2 \leq 0$ .

O intervalo com nível de confiança  $(1 - \alpha)$  para a diferença de médias entre os dois grupos é dado por:

$$(\bar{X}_1 - \bar{X}_2) - t_{gl, 1-\alpha/2} S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq (\mu_1 - \mu_2) \leq (\bar{X}_1 - \bar{X}_2) + t_{gl, 1-\alpha/2} S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad (16.4)$$

Vamos utilizar o *R Commander* para realizar um teste de hipótese bilateral de igualdade de médias e calcular o intervalo de confiança ao nível de 90% para o conjunto de dados *energy*. Tendo selecionado o conjunto de dados *energy*, utilizamos a seguinte opção do menu do *R Commander* para realizar um teste t para amostras independentes:

Estatísticas  $\Rightarrow$  Médias  $\Rightarrow$  Teste t para amostras independentes...

Após a seleção do teste, é preciso definir a variável que define os grupos e a variável resposta (figura 16.4).



Figura 16.4: Seleção das variáveis de resposta e da variável que define os grupos. O conjunto de dados *energy* somente tem uma variável como fator e uma variável quantitativa como resposta.

Ao clicarmos na guia *Opções* na caixa de diálogo da figura 16.4, podemos selecionar o tipo de teste (bilateral/unilateral), o nível de confiança e se as variâncias são iguais ou não (figura 16.5). Vamos especificar o nível de confiança igual a 90% (0.9) e marcar a opção que as variâncias são iguais.

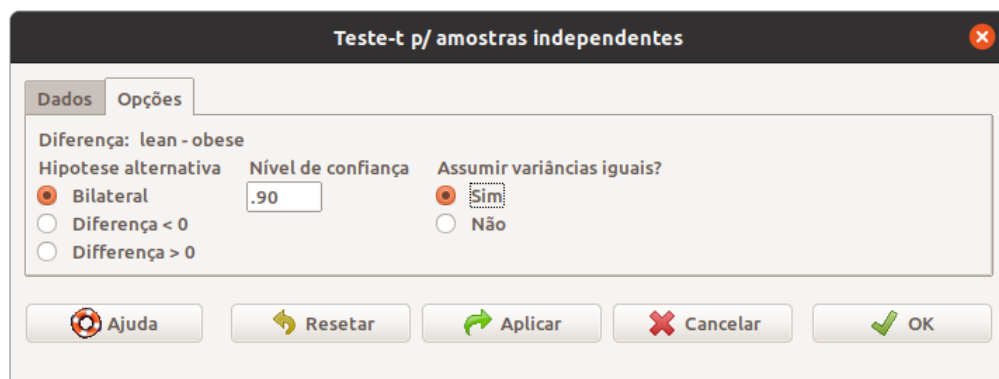


Figura 16.5: Definindo o tipo de teste, o nível de confiança e especificando que as variâncias são iguais.

Ao clicarmos em OK na figura 16.5, o teste t é realizado e os resultados são mostrados a seguir.

```
t.test(expend~stature, alternative='two.sided', conf.level=.90,
       var.equal=TRUE, data=energy)
```

```
##
## Two Sample t-test
##
## data: expend by stature
## t = -3.9456, df = 20, p-value = 0.000799
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
## -3.207130 -1.256118
## sample estimates:
## mean in group lean mean in group obese
## 8.066154 10.297778
```

Observem a sintaxe do comando que é executado para a realização do teste.

A saída mostra que o valor de  $p$  é 0,0008. Nesse caso, a hipótese nula deve ser rejeitada. O intervalo de confiança ao nível de 90% para a diferença de médias do consumo de energia entre as populações de mulheres magras e obesas varia de -3,2 a -1,26 MJ. Observem que o intervalo de confiança não inclui o zero (hipótese nula).

Se as variáveis  $X_1$  e  $X_2$  possuem distribuição normal, mas com variâncias desconhecidas e diferentes, um procedimento confiável é conhecido como *teste t para duas amostras de Welch*. Por essa aproximação, a variável aleatória

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad (16.5)$$

segue uma distribuição *t de Student* com graus de liberdade dado pela seguinte expressão:

$$gl = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2-1}} \quad (16.6)$$

A estatística

$$T = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

pode ser utilizada para realizar um teste de hipótese bilateral para  $H_0 : \mu_1 - \mu_2 = 0$  ou testes unilaterais para  $H_0 : \mu_1 - \mu_2 \geq 0$  ou  $H_0 : \mu_1 - \mu_2 \leq 0$ .

O intervalo com nível de confiança  $(1 - \alpha)$  para a diferença de médias entre as duas amostras é dado por:

$$(\bar{X}_1 - \bar{X}_2) - t_{gl,1-\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \leq (\mu_1 - \mu_2) \leq (\bar{X}_1 - \bar{X}_2) + t_{gl,1-\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \quad (16.7)$$

Para realizarmos um teste de hipótese bilateral e calcularmos o intervalo de confiança ao nível de 90% para o conjunto de dados *energy*, supondo que as variâncias sejam diferentes, seguimos os mesmos passos das figuras 16.4 e 16.5, porém, não assumimos que as variâncias são iguais na aba *Opções* (figura 16.6).

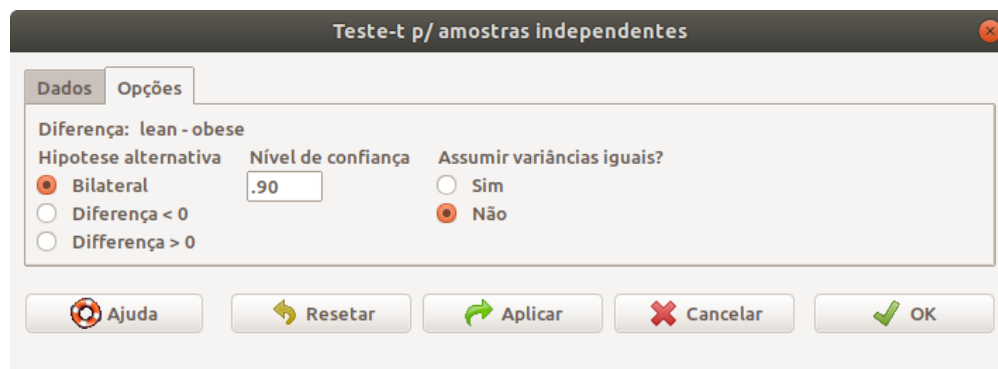


Figura 16.6: Definindo o tipo de teste, o nível de confiança e especificando que as variâncias são diferentes

Ao clicarmos em OK na figura 16.6, o teste t para duas amostras de Welch é realizado e a saída é mostrada a seguir.

```
t.test(expend~stature, alternative='two.sided', conf.level=.9,
      var.equal=FALSE, data=energy)

##
##  Welch Two Sample t-test
##
## data:  expend by stature
## t = -3.8555, df = 15.919, p-value = 0.001411
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##  -3.242484 -1.220763
## sample estimates:
##  mean in group lean mean in group obese
##           8.066154           10.297778
```

Os resultados nesse exemplo são bastante semelhantes aos obtidos com a suposição de que as variâncias são iguais.

Quando as duas amostras possuem o mesmo número de elementos ( $n_1 = n_2 = n$ ), o número de graus de liberdade, calculado pela expressão (16.6), é igual  $2n - 2$ , as estatísticas (16.3) e (16.5) são iguais, assim como os intervalos de confiança (16.7) e (16.4). Isso significa que quando as amostras possuem o mesmo tamanho, o teste t de Student é idêntico ao teste t de Welch.

Uma condição necessária para se realizar um teste t de Student ou o teste t aproximado de Welch é que as variáveis  $X_1$  e  $X_2$  sejam normalmente distribuídas. O teste t é robusto para desvios consideráveis da hipótese de normalidade dos dados, especialmente se os tamanhos das amostras são iguais ou aproximados e especialmente quando os testes são bilaterais.

Mesmo quando as variáveis possuem grandes desvios em relação à distribuição normal, como a distribuição da média amostral tende a uma distribuição normal à medida que o tamanho da amostra aumenta (Teorema do Limite Central), se as amostras são suficientemente grandes (digamos  $n_1, n_2 \geq 30$ ), podemos usar a estatística (16.1) para realizarmos um teste de hipótese bilateral para  $H_0 : \mu_1 - \mu_2 = 0$  ou testes unilaterais para  $H_0 : \mu_1 - \mu_2 \geq 0$  ou  $H_0 : \mu_1 - \mu_2 \leq 0$ , e a expressão (16.2) para o cálculo do intervalo de confiança para a diferença de médias, com  $\sigma_1^2$  e  $\sigma_2^2$  substituídos por suas estimativas amostrais  $S_1^2$  e  $S_2^2$ .

Para amostras pequenas, digamos  $n_1$  ou  $n_2 < 30$ , é necessário verificar a normalidade das variáveis  $X_1$  e  $X_2$  e a igualdade de suas variâncias.

### 16.2.2 Teste de igualdade de variâncias

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

Na seção 16.2.1 foram realizados dois testes t para a comparação de médias entre dois grupos independentes cujas variáveis seguem uma distribuição normal: um supondo que as variâncias dos grupos fossem iguais e outro na suposição de que as variâncias fossem diferentes. Há vários testes estatísticos para verificar a suposição de que as variâncias de duas populações sejam iguais. Para realizar tais testes, como sempre, temos que definir a hipótese nula  $H_0$  e a hipótese alternativa  $H_1$ , bem como qual a estatística a ser utilizada no teste. Para a variância, sendo  $S_1^2$  e  $S_2^2$  as estimativas amostrais das variâncias  $\sigma_1^2$  e  $\sigma_2^2$ , respectivamente, **de duas variáveis com distribuição normal**, teríamos:

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

**Estatística de teste:** um dos testes para verificação de igualdade de variâncias é o **teste F para duas variâncias**. Nesse caso, utiliza-se uma das duas estatísticas a seguir:

$$F_1 = \frac{S_1^2}{S_2^2} \quad (16.8)$$

ou

$$F_2 = \frac{S_2^2}{S_1^2} \quad (16.9)$$

A estatística (16.8) segue uma distribuição chamada F de Fisher, com  $(n_1 - 1)$  e  $(n_2 - 1)$  graus de liberdade, que é a razão entre duas distribuições qui-quadrado, a primeira com  $(n_1 - 1)$  graus de liberdade e a segunda com  $(n_2 - 1)$  graus de liberdade:

$$F_1 \sim \frac{\chi_{\nu_1}^2/\nu_1}{\chi_{\nu_2}^2/\nu_2}$$

A estatística (16.9) segue uma distribuição F de Fisher, com  $(n_2 - 1)$  e  $(n_1 - 1)$  graus de liberdade, que é a razão entre duas distribuições qui-quadrado, a primeira com  $(n_2 - 1)$  graus de liberdade e a segunda com  $(n_1 - 1)$  graus de liberdade:

$$F_2 \sim \frac{\chi_{\nu_2}^2/\nu_2}{\chi_{\nu_1}^2/\nu_1}$$

A estatística utilizada no teste F é o maior valor entre  $F_1$  e  $F_2$ . Sob a hipótese nula de igualdade de variâncias, espera-se que o maior valor entre  $F_1$  e  $F_2$  esteja próximo de 1. Se essa razão for acima do valor crítico, então a hipótese de igualdade de variâncias é rejeitada.

No exemplo do conjunto de dados *energy*, os graus de liberdade  $\nu_1$  e  $\nu_2$  são dados por 12 e 8, respectivamente. Para realizarmos esse teste no *R Commander*, selecionamos a seguinte opção:

Estatísticas  $\Rightarrow$  Variâncias  $\Rightarrow$  Teste F para 2 variâncias

Na caixa de diálogo do teste F para 2 variâncias (figura 16.7), selecionamos a variável que define os grupos e a variável resposta. Em *Opções*, selecionamos o tipo de teste (bilateral/unilateral) e o nível de confiança.



Figura 16.7: Seleção das variáveis de resposta e da variável que define os grupos. O conjunto de dados *energy* somente tem uma variável como fator e uma variável quantitativa como resposta.

Ao clicarmos no botão OK, o teste é realizado de acordo com a função abaixo.

```
var.test(expend ~ stature, alternative='two.sided', conf.level=.90,
         data=energy)
```

```
##
## F test to compare two variances
##
## data:  expend by stature
## F = 0.78445, num df = 12, denom df = 8, p-value = 0.6797
## alternative hypothesis: true ratio of variances is not equal to 1
## 90 percent confidence interval:
##  0.2388735 2.2345455
## sample estimates:
## ratio of variances
##           0.784446
```

Com base nos resultados, vemos que o valor para a maior das estatísticas (16.8) e (16.9) está dentro do intervalo de não rejeição da hipótese de que as variâncias sejam iguais. Portanto não há evidência suficiente para rejeitarmos a hipótese nula de que as variâncias sejam iguais. Porém o intervalo de confiança para a razão entre as variâncias  $[0, 24 - 2, 2]$  é bastante amplo, de modo que esse teste possui pouco poder estatístico, ou seja, pouca capacidade de rejeitar a hipótese de igualdade de variâncias, se elas forem diferentes.

Há diversos outros testes para verificar a igualdade de 2 variâncias. No *R Commander*, dois outros podem ser executados (teste de Bartlett e teste de Levene) via mesma opção do menu que leva ao *teste F para 2 variâncias*. Em geral esses testes não possuem grande poder estatístico, especialmente quando as distribuições das variáveis  $X_1$  e  $X_2$  não são normais. **Então, se houver dúvidas de que as variâncias são iguais, o mais indicado é realizar o teste t para duas amostras de Welch, na suposição de que as variâncias**



são diferentes.

### 16.2.3 Normalidade dos dados

O conteúdo desta seção e da seção 16.2.4 podem ser visualizados neste [vídeo](#).

O teste t para diferença de duas médias para grupos independentes supõe que os dados sejam normalmente distribuídos. Um instrumento visual útil para checar a normalidade de dados é o gráfico de probabilidade normal (*normal probability plot* ou *qqplot*). No *R Commander*, esse gráfico é obtido por meio da opção *Gráfico de comparação de quantis...*

O gráfico de probabilidade normal é construído a partir da ordenação dos valores da variável em ordem crescente e a plotagem em um gráfico do *i*-ésimo valor contra o quantil esperado desse valor em uma distribuição normal. Ao plotar todos os pontos assim obtidos, obteríamos uma linha reta se os dados seguissem uma distribuição normal. Diferentes fontes usam diferentes aproximações para o cálculo do quantil esperado do *i*-ésimo valor.

A fórmula usada pelo R é dada por:

$$z_i = \Phi^{-1} \left( \frac{i - a}{n + 1 - 2a} \right) \quad (16.10)$$

para  $i = 1, 2, \dots, n$ , onde:

$$a = \begin{cases} 3/8, & n \leq 10 \\ 0,5, & n > 10 \end{cases}$$

e  $\Phi^{-1}$  é a função quantil da distribuição normal.

Vamos mostrar como seriam obtidos os pares de pontos para construir o *qqplot* para o grupo de obesas do conjunto de dados *energy*. Primeiramente iremos selecionar todos os valores do consumo energético para o grupo de mulheres obesas, usando o comando abaixo.

```
obesas_exp <- energy$expend[energy$stature == 'obese']
```

Nesse comando, criamos uma variável (*obesas\_exp*) que irá conter os valores do consumo energético das mulheres obesas. A expressão entre colchetes (*stature == 'obese'*) testa cada valor de *stature* e será verdadeira somente para as observações de mulheres obesas. Então a expressão *expend[stature == 'obese']* retorna todos os valores de *expend* das mulheres obesas.

Vamos ordenar a variável *obesas\_exp* em ordem crescente:

```
sort(obesas_exp)
```

```
## [1] 8.79 9.19 9.21 9.68 9.69 9.97 11.51 11.85 12.79
```

Temos nove valores na variável *obesas\_exp*. Substituindo  $a = 3/8 = 0,375$  e  $n = 9$  na expressão (16.10), temos:

$$z_i = \Phi^{-1} \left( \frac{i - 0,375}{9,25} \right)$$

A tabela 16.1 mostra os valores de  $z_i$  correspondentes a cada valor de *expnd* para o grupo de mulheres obesas.

Tabela 16.1: Correspondência entre os valores de *expnd* e  $z_i$  para a construção do *qqplot* do grupo de mulheres obesas.

<i>obesas_exp</i>	<i>i</i>	$(i-0,375)/9,25$	$z_i$
8,79	1	0,0675	-1,49
9,19	2	0,176	-0,93
9,21	3	0,284	-0,57
9,68	4	0,392	-0,27
9,69	5	0,500	0,00
9,97	6	0,608	0,27
11,51	7	0,716	0,57
11,85	8	0,824	0,93
12,79	9	0,932	1,49

Para gerar o gráfico de comparação de quantis no *R Commander*, usamos a seguinte opção:

Gráficos ⇒ Gráfico de comparação de quantis...

Na figura 16.8, selecionamos a variável desejada (*expnd*). Em seguida, clicamos em *Gráfico por grupos* e selecionamos a variável *stature* (figura 16.9). Clicamos em OK e, a seguir, clicamos na aba *Opções* para verificar as opções disponíveis (figura 16.10). Vamos selecionar a distribuição normal e a opção *não identificar* em *Identificar pontos*.

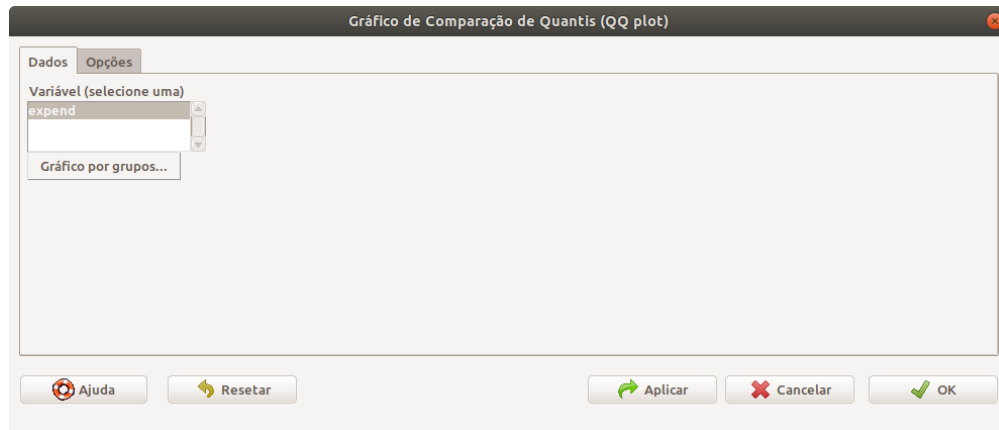


Figura 16.8: Diálogo do gráfico de comparação de quantis para selecionar a variável cujo gráfico será construído.



Figura 16.9: Diálogo para a seleção da variável de agrupamento para o gráfico de comparação de quantis.

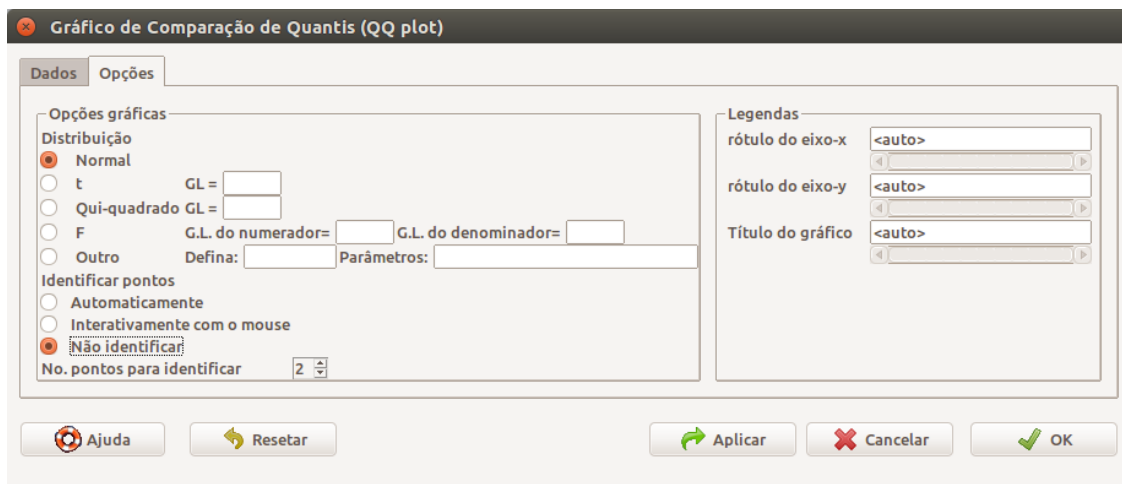


Figura 16.10: Caixa de diálogo de opções do gráfico de comparação de quantis.

Ao pressionarmos o botão OK, o comando a seguir é executado e o gráfico é mostrado na

figura 16.11. Observando os dois gráficos, verificamos que há três pontos no grupo das mulheres magras que se desviam do que seria esperado em uma distribuição normal.

```
with(energy, qqPlot(expend, dist="norm", id=list(method="y", n=0,
labels=rownames(energy)), groups=stature))
```

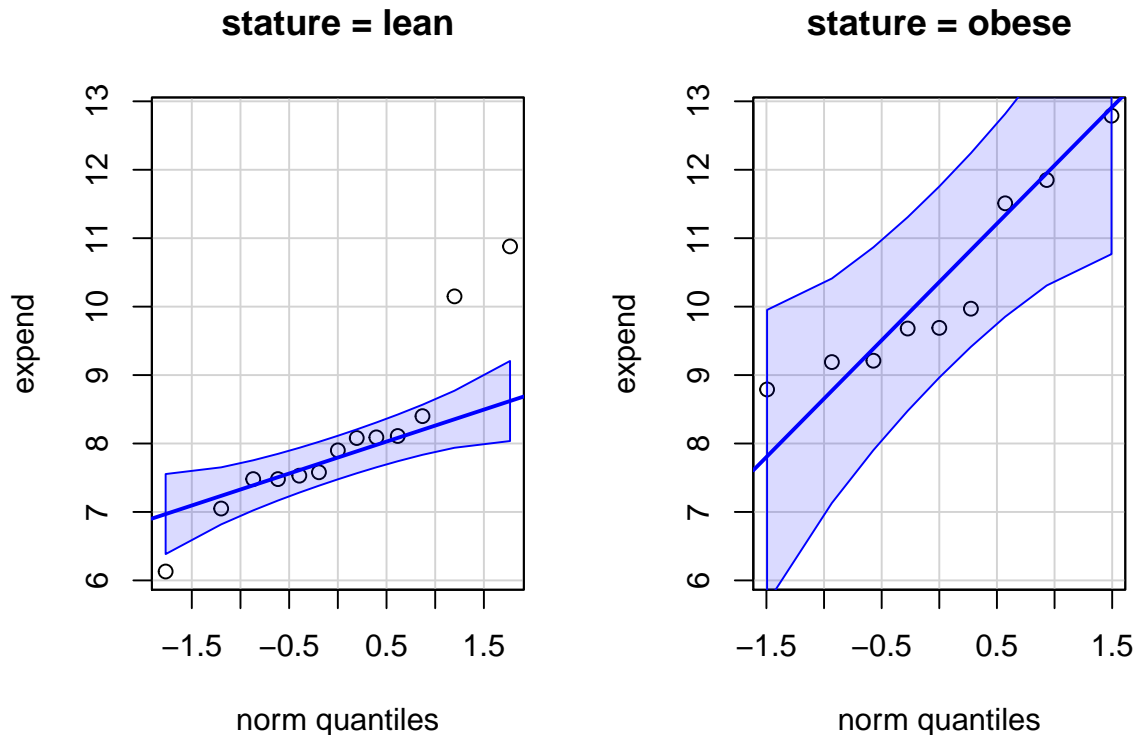


Figura 16.11: Gráfico de comparação de quantis da variável *expend* para os dois grupos de mulheres (obesas e magras).

## 16.2.4 Testes de normalidade

Além do gráfico de comparação de quantis, há vários testes estatísticos para verificar a hipótese de normalidade de dados.

Dois testes frequentemente utilizados são [Kolmogorov-Smirnov](#) e [Shapiro-Wilk](#). Detalhes desses testes podem ser vistos nos links indicados. O teste de Shapiro-Wilk tende a ser mais poderoso para um mesmo nível de significância do que o teste de Kolmogorov-Smirnov, mas ambos podem detectar desvios insignificantes da normalidade quando a amostra for grande.

Para realizar um teste de normalidade no *R Commander*, usamos a seguinte opção:

Estatísticas  $\Rightarrow$  Resumos  $\Rightarrow$  Test of normality...

Na figura 16.12, selecionamos a variável desejada (*expend*). Em seguida, clicamos no botão *Test by groups* e selecionamos a variável *stature* como variável de agrupamento, porque

precisamos testar a normalidade da variável *expend* em cada um dos grupos. Vamos realizar o teste de *Shapiro-Wilk*.

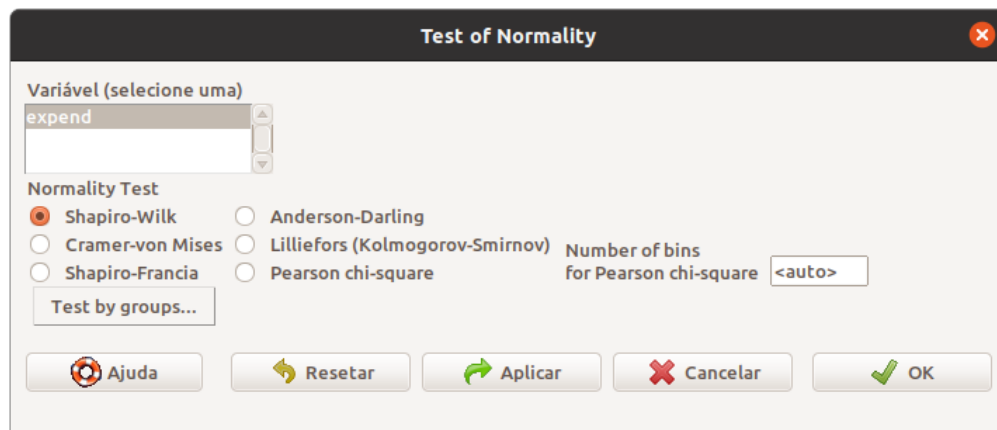


Figura 16.12: Caixa de diálogo para selecionar a variável cuja normalidade será testada, o teste a ser realizado e, eventualmente a variável de agrupamento.

Ao pressionarmos o botão OK, o comando a seguir é executado.

```
normalityTest(expend ~ stature, test="shapiro.test", data=energy)
```

```
##
## -----
## stature = lean
##
## Shapiro-Wilk normality test
##
## data:  expend
## W = 0.86733, p-value = 0.04818
##
## -----
## stature = obese
##
## Shapiro-Wilk normality test
##
## data:  expend
## W = 0.87603, p-value = 0.1426
##
## -----
##
## p-values adjusted by the Holm method:
##      unadjusted adjusted
## lean  0.048184   0.096367
## obese 0.142574   0.142574
```

Os resultados mostram o valor de  $p$  resultante da aplicação do teste de normalidade de *Shapiro-Wilk* para a variável *expend* em cada um dos grupos da variável *stature* ( $p = 0,048$  para as magras e  $p = 0,14$  para as obesas). Como são realizados dois testes, há uma correção dos valores de  $p$  para múltiplos testes, usando o *método de Holm*, resultando em valores de  $p = 0,096$  para as magras e  $p = 0,14$  para as obesas. Continuando com o nível de significância de 10% que estamos usando neste capítulo, mesmo com a correção de *Holm*, rejeitaríamos a hipótese de normalidade dos valores da variável *energy* para as mulheres magras ( $p = 0,096 < 0,1$ ). Esses resultados estão de acordo com os obtidos no gráfico de comparação de quantis, indicando um ligeiro desvio da normalidade para as mulheres magras.

### 16.2.5 Teste não paramétrico de Wilcoxon para duas amostras

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

Para a comparação de médias entre duas amostras independentes, quando as distribuições das variáveis  $X_1$  e  $X_2$  se desviam marcadamente da normal e as amostras são pequenas ( $n_1$  ou  $n_2 < 30$ ), então pode-se utilizar o teste não paramétrico denominado **teste de Mann-Whitney, que é equivalente ao teste de Wilcoxon para duas amostras**.

O teste de Wilcoxon para duas amostras também é conhecido como teste de Wilcoxon-Mann-Whitney ou teste de Wilcoxon da soma dos postos (*Wilcoxon rank-sum test*). Esse teste relaxa a exigência de normalidade dos dados, porém requer que as seguintes condições sejam verdadeiras:

- 1) as duas amostras foram aleatoriamente e independentemente extraídas das suas respectivas populações;
- 2) a escala da medição é pelo menos ordinal;
- 3) se as distribuições das populações diferem, elas diferem somente em relação à sua localização.

Vamos retornar ao conjunto de dados *energy*, apresentado na seção 16.2. Nessa seção, foi realizado um teste  $t$  para amostras independentes. Vamos utilizar esse conjunto de dados para ilustrar o teste de Mann-Whitney.

Os valores da variável dependente (*expend*) são ordenados em ordem crescente (ou decrescente), independentemente do grupo ao qual pertencem. A tabela 16.2 mostra os valores ordenados em ordem crescente nas colunas 1 (grupo “magras”) e 2 (grupo “obesas”). Após a ordenação, cada valor é substituído pela sua ordem (posto) na sequência de valores. Por exemplo, o menor valor recebe o posto 1, o segundo menor valor recebe o posto 2 e assim por diante. Quando houver mais de um valor iguais entre si (empates), cada um deles recebe a média dos postos que receberiam se fossem diferentes. Por exemplo, há dois valores iguais a 7,48 nas posições 3 e 4. Portanto eles recebem o posto igual a 3,5. Os postos para a variável *expend* em cada grupo são mostrados nas colunas 3 e 4 para os grupos “magras” e “obesas”, respectivamente.

Tabela 16.2: Cálculo da soma dos postos da variável *expend* para o grupo “magras” e “obesas”. A última linha mostra a soma dos postos nas respectivas colunas.

Consumo de energia Grupo 'magras'	Consumo de energia Grupo 'obesas'	Posto Grupo 'magras'	Posto Grupo 'obesas'
6.13		1	
7.05		2	
7.48		3,5	
7.48		3,5	
7.53		5	
7.58		6	
7.9		7	
8.08		8	
8.09		9	
8.11		10	
8.4		11	
	8.79		12
	9.19		13
	9.21		14
	9.68		15
	9.69		16
	9.97		17
10.15		18	
10.88		19	
	11.51		20
	11.85		21
	12.79		22
		<b>103</b>	<b>150</b>

Sejam  $n_1$  e  $n_2$  o número de valores nos grupos 1 e 2,  $R_1$  a soma dos postos no grupo 1,  $R_2$  a soma dos postos no grupo 2. As estatísticas a seguir são utilizadas pelo teste de Mann-Whitney:

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 \quad (16.11)$$

$$U' = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2 \quad (16.12)$$

U pode ser obtido a partir de U' e vice-versa. Na hipótese nula de igualdade de localização das duas distribuições, por exemplo, se uma das duas estatísticas (U ou U') for maior ou igual ao valor crítico, rejeita-se a hipótese nula. No teste de Wilcoxon para duas amostras, definem-se estatísticas alternativas e equivalentes, que são dadas por:

$$W = R_1 - \frac{n_1(n_1 + 1)}{2} \quad (16.13)$$

$$W' = R_2 - \frac{n_2(n_2 + 1)}{2} \quad (16.14)$$

Vamos realizar o teste de Wilcoxon para duas amostras no *R Commander* para comparar a variável *expend* nos grupos de mulheres “magras” e “obesas”. Vamos realizar um teste de hipótese bilateral. Tendo selecionado o conjunto de dados *energy*, utilizamos a seguinte opção do menu do *R Commander*:

Estatísticas  $\Rightarrow$  Testes Não-Paramétricos  $\Rightarrow$  Teste de Wilcoxon (2 amostras)

Após a seleção do teste, é preciso definir a variável que define os grupos e a variável resposta (figura 16.13).



Figura 16.13: Seleção das variáveis de resposta e da variável que define os grupos.

Ao clicarmos na guia *Opções* na caixa de diálogo da figura 16.13, podemos selecionar se o teste é bilateral ou unilateral e o tipo de teste (figura 16.14). Vamos selecionar a opção *Exato* para tipo de teste.



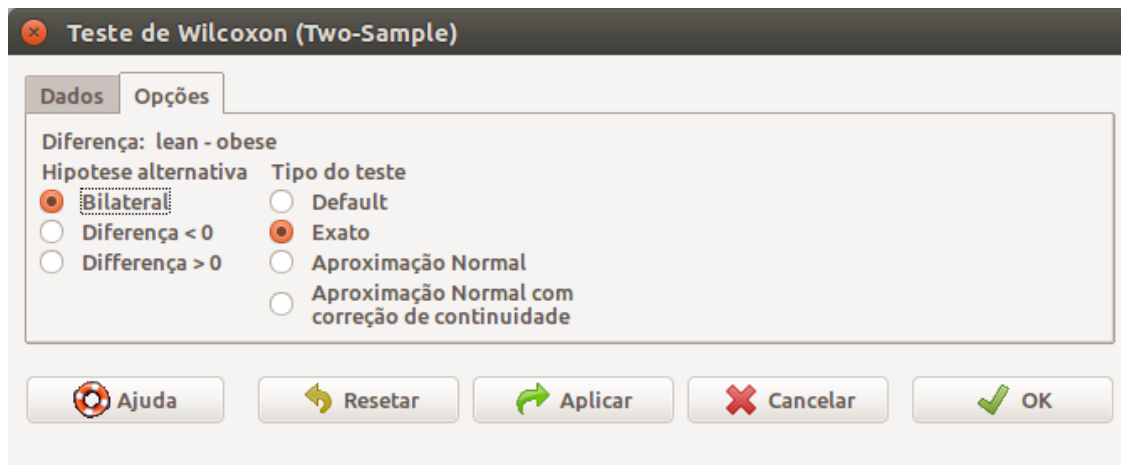


Figura 16.14: Definindo as opções para o teste de Wilcoxon para duas amostras. Observem que não é possível especificar o nível de confiança nessa caixa de diálogo.

Ao clicarmos em OK na figura 16.14, o teste de Wilcoxon para duas amostras é realizado conforme a seguir.

```
with(energy, tapply(expend, stature, median, na.rm=TRUE))
```

```
## lean obese
## 7.90 9.69
```

```
wilcox.test(expend ~ stature, alternative='two.sided', exact=TRUE,
            correct=FALSE, data=energy)
```

```
##
## Wilcoxon rank sum test
##
## data: expend by stature
## W = 12, p-value = 0.001896
## alternative hypothesis: true location shift is not equal to 0
```

Os resultados mostram as medianas dos grupos das mulheres “magras” e “obesas”, e o resultado do teste estatístico, com valor de  $p = 0,001896$ , rejeitando a hipótese nula de igualdade de localização (medianas) das distribuições dos dois grupos de mulheres.

Para calcular o intervalo de confiança para a diferença de localização entre as duas populações, é necessário utilizar a linha de comando. A função `wilcox.test` com a especificação do nível de confiança (`conf.level = 0.90`) e determinando que o intervalo de confiança seja calculado (`conf.int = TRUE`) é mostrada a seguir seguida dos resultados.

```
wilcox.test(expend ~ stature, alternative='two.sided', exact=TRUE,
            correct=FALSE, data=energy, conf.int=TRUE, conf.level=0.90)
```

```
##
## Wilcoxon rank sum test
##
## data: expend by stature
## W = 12, p-value = 0.001896
## alternative hypothesis: true location shift is not equal to 0
## 90 percent confidence interval:
## -3.419949 -1.310093
## sample estimates:
## difference in location
## -1.909972
```

Há diversos detalhes sobre o teste de Wilcoxon para duas amostras realizado pelo R que são apresentados na página de ajuda da função *wilcox.test* (acessada por meio do comando *?wilcox.test*). **Devemos chamar a atenção que a diferença de localização mostrada nos resultados acima não estima a diferença de medianas dos grupos, mas sim a mediana da diferença entre um item de  $X_1$  e um item de  $X_2$ .**

No exemplo acima, podemos dizer, com confiança de 90%, que a mediana do consumo de energia de uma mulher magra é pelo menos 1,31 menor do que a mediana de uma mulher obesa.

## 16.3 Comparação de médias de amostras dependentes

O conteúdo desta seção e da seção 16.3.1 podem ser visualizados neste [vídeo](#).

Para amostras dependentes (pareadas), vamos usar como exemplo o conjunto de dados *intake* da biblioteca *ISwR*.

```
data(intake, package="ISwR")
intake
```

```
##      pre post
## 1  5260 3910
## 2  5470 4220
## 3  5640 3885
## 4  6180 5160
## 5  6390 5645
## 6  6515 4680
## 7  6805 5265
## 8  7515 5975
## 9  7515 6790
## 10 8230 6900
## 11 8770 7335
```

O conjunto de dados *intake* contém dados de consumo de energia (kJ) em 11 pacientes antes e depois da menstruação.

Como os dados são pareados, a melhor forma de analisá-los estatisticamente para verificar se existe uma diferença de médias entre as medidas antes ( $X_1$ ) e depois ( $X_2$ ) é criar uma variável aleatória consistindo da diferença entre as medidas antes e depois:

$$D = X_1 - X_2$$

Vamos considerar diferentes situações.

**Se a variável D possui uma distribuição  $N(\mu_D, \sigma_D^2)$  com variância conhecida**, a estatística

$$Z = \frac{\bar{D} - \mu_D}{\sqrt{\frac{\sigma_D^2}{n}}} \quad (16.15)$$

possui uma distribuição normal padrão. A estatística

$$Z = \frac{\bar{D}}{\sqrt{\frac{\sigma_D^2}{n}}} \quad (16.16)$$

onde  $n$  é o tamanho da amostra, pode ser utilizada para realizar um teste de hipótese bilateral para  $H_0 : \mu_D = 0$  ou testes unilaterais para  $H_0 : \mu_D \geq 0$  ou  $H_0 : \mu_D \leq 0$ .

O intervalo de confiança para  $\mu_D$  (diferença de médias entre as duas medidas), sendo  $(1 - \alpha)$  o nível de confiança, é dado por:

$$\bar{D} - z_{1-\alpha/2} \sqrt{\frac{\sigma_D^2}{n}} \leq \mu_D \leq \bar{D} + z_{1-\alpha/2} \sqrt{\frac{\sigma_D^2}{n}} \quad (16.17)$$

Como, em geral, a variância de D não é conhecida, ela deve ser estimada por meio da variância amostral. Sob a condição de normalidade, uma análise frequentemente utilizada quando não se conhece a variância da diferença das observações se baseia na distribuição *t de Student*.

### 16.3.1 Teste t para amostras dependentes (teste t pareado)

Se a variável D possui uma distribuição  $N(\mu_D, \sigma_D^2)$  com variância desconhecida, a variância da diferença entre as medidas pode ser estimada pela expressão:

$$S_D^2 = \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2 \quad (16.18)$$

Nesse caso, a estatística

$$T = \frac{\bar{D} - \mu_D}{\sqrt{\frac{S_D^2}{n}}} \quad (16.19)$$

possui uma distribuição t de Student com n-1 graus de liberdade. A estatística

$$t = \frac{\bar{D}}{\sqrt{\frac{S_D^2}{n}}} \quad (16.20)$$

pode ser utilizada para realizar um teste de hipótese bilateral para  $H_0 : \mu_D = 0$  ou testes unilaterais para  $H_0 : \mu_D \geq 0$  ou  $H_0 : \mu_D \leq 0$ .

O intervalo de confiança para  $\mu_D$  (diferença de médias entre as duas medidas), sendo  $(1 - \alpha)$  o nível de confiança, é dado por:

$$\bar{D} - t_{n-1, 1-\alpha/2} \sqrt{\frac{S_D^2}{n}} \leq \mu_D \leq \bar{D} + t_{n-1, 1-\alpha/2} \sqrt{\frac{S_D^2}{n}} \quad (16.21)$$

Vamos utilizar o *R Commander* para realizar um teste de hipótese bilateral e calcular o intervalo de confiança ao nível de 90% para o conjunto de dados *intake*. Tendo selecionado o conjunto de dados *intake*, utilizamos a seguinte opção do menu do *R Commander* para realizar um teste t para amostras dependentes:

Estatísticas  $\Rightarrow$  Médias  $\Rightarrow$  Teste t (dados pareados)

Após a seleção do teste, é preciso definir as variáveis que definem as duas medidas efetuadas em cada unidade de observação (figura 16.15).

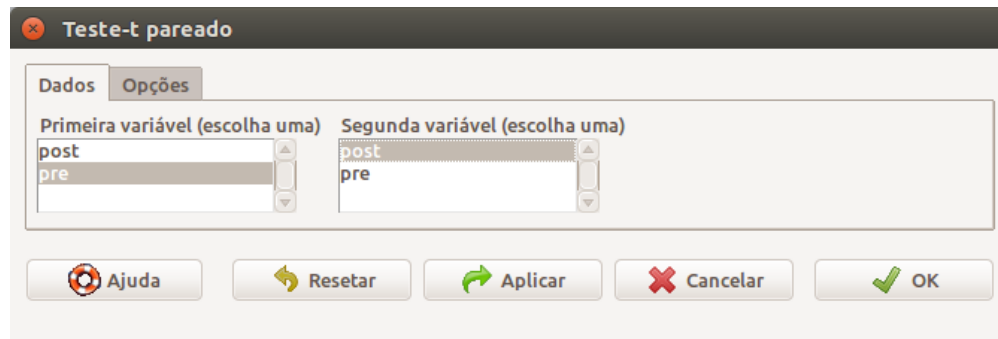


Figura 16.15: Seleção das variáveis que definem as duas medidas efetuadas em cada unidade de observação.

Ao clicarmos na guia *Opções* na caixa de diálogo da figura 16.15, podemos selecionar o tipo de teste (bilateral/unilateral) e o nível de confiança. Vamos especificar o nível de confiança igual a 90% (0.9) (figura 16.16). O teste verificará a diferença (*post* – *pre*).



Figura 16.16: Definindo o tipo de teste e o nível de confiança.

Ao clicarmos em OK na figura 16.16, o teste t pareado é realizado por meio da função *t.test* abaixo, com os resultados mostrados a seguir.

```
with(intake, (t.test(pre, post, alternative='two.sided',  
                    conf.level=.90, paired=TRUE)))
```

```
##
## Paired t-test
##
## data:  pre and post
## t = 11.941, df = 10, p-value = 0.0000003059
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##  1120.036 1520.873
## sample estimates:
## mean of the differences
##                1320.455
```

Os resultados mostram que o intervalo com confiança de 90% da diferença *pre* – *post* é dado por [1120,0; 1520,9] kJ. O valor de  $p$  é  $3,1 \cdot 10^{-7}$  (bastante significativo), rejeitando a hipótese nula de que o consumo de energia é o mesmo antes e depois da menstruação.

Uma condição necessária para se realizar o teste  $t$  pareado é que a diferença das variáveis,  $D = X_1 - X_2$ , seja normalmente distribuída. O teste  $t$  é robusto para desvios consideráveis da hipótese de normalidade dos dados e especialmente quando os testes são bilaterais.

Mesmo quando diferença das variáveis possui grandes desvios em relação à distribuição normal, se a amostra for suficientemente grande (digamos  $n \geq 30$ ), podemos usar a estatística (16.16) para realizarmos um teste de hipótese bilateral para  $H_0 : \mu_D = 0$  ou testes unilaterais para  $H_0 : \mu_D \geq 0$  ou  $H_0 : \mu_D \leq 0$ , e a expressão (16.17) para o cálculo do intervalo de confiança para a diferença de médias, com  $\sigma_D^2$  substituída por sua estimativa amostral,  $S_D^2$ .

Para amostras pequenas, digamos  $n < 30$ , é necessário verificar a normalidade da variável  $D$ .

Vamos verificar por meio do diagrama de comparação de quantis a normalidade da diferença (*pre* – *post*). Como não temos essa variável no conjunto de dados, precisamos criá-la. Vamos mostrar como fazer isso utilizando o *R Commander*. Para criar uma nova variável a partir de outras variáveis existentes no conjunto de dados ativo, acessamos a seguinte opção no menu:

Dados  $\Rightarrow$  Modificação variáveis no conj. de dados...  $\Rightarrow$  Computar nova variável...

Na caixa de diálogo da figura 16.17, damos o nome para a variável que vai ser criada e especificamos a fórmula de cálculo dessa nova variável (*pre* – *post*).

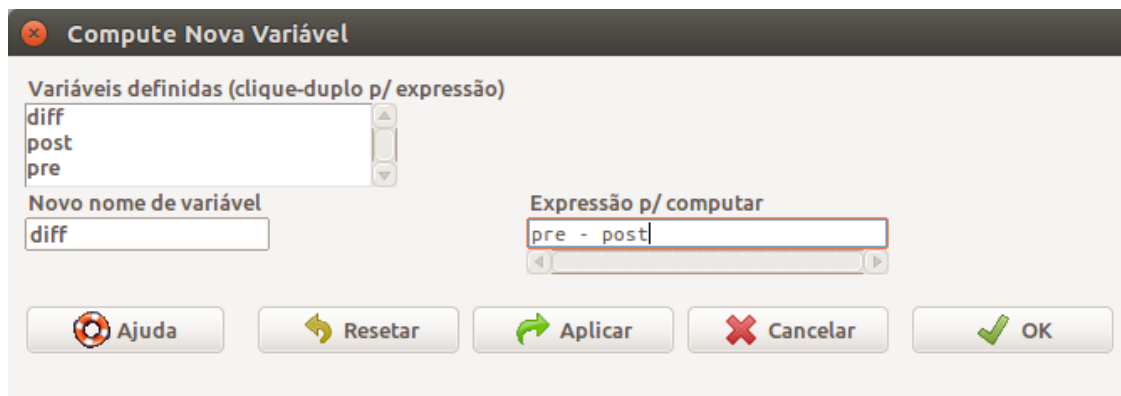


Figura 16.17: Nessa caixa de diálogo, damos o nome para a variável que vai ser criada e especificamos a fórmula de cálculo dessa nova variável.

Ao clicarmos em OK, a variável será criada por meio do comando a seguir e é adicionada ao conjunto de dados, como mostra a nova listagem de *intake*.

```
intake$diff <- with(intake, pre - post)
intake
```

```
##      pre post diff
## 1  5260 3910 1350
## 2  5470 4220 1250
## 3  5640 3885 1755
## 4  6180 5160 1020
## 5  6390 5645  745
## 6  6515 4680 1835
## 7  6805 5265 1540
## 8  7515 5975 1540
## 9  7515 6790  725
## 10 8230 6900 1330
## 11 8770 7335 1435
```

Agora podemos gerar o gráfico de comparação de quantis para a variável *diff*, como na seção anterior, ou usando diretamente a função *qqPlot* como a seguir. O gráfico é mostrado na figura 16.18, indicando que a variável *diff* não possui grandes desvios em relação a uma distribuição normal.

```
with(intake, qqPlot(diff, dist="norm", id=list(method="y", n=0,
labels=rownames(intake))))
```

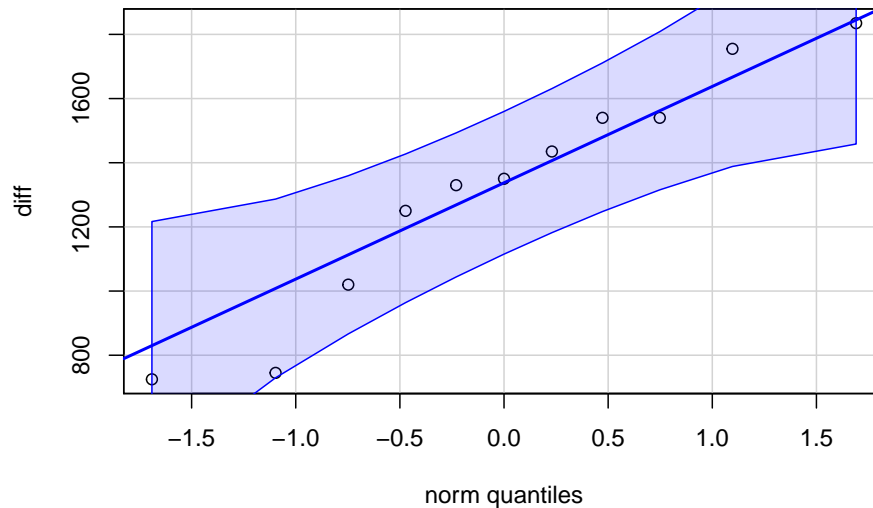


Figura 16.18: Gráfico de comparação de quantis da normal para a variável *diff* (*pre* – *post*).

Quando a diferença de valores da variável resposta não possui uma distribuição normal e o tamanho da amostra é pequeno ( $n < 30$ ), então pode-se utilizar o teste de Wilcoxon para amostras pareadas, apresentado na seção seguinte.

### 16.3.2 Teste de Wilcoxon para amostras pareadas

O conteúdo desta seção e da seção 16.5 podem ser visualizados neste [vídeo](#).

Para duas amostras dependentes, o teste não paramétrico mais utilizado é denominado teste de Wilcoxon para amostras pareadas (*Wilcoxon paired-sample test*, *Wilcoxon matched pairs test*, *signed-rank test*). Esse teste consiste em colocar em ordem crescente ou decrescente as diferenças entre os valores da variável em cada par. O posto correspondente a cada diferença recebe o sinal positivo se a diferença for positiva e o sinal negativo se a diferença for negativa. Em seguida são somados os postos positivos (ou negativos). A soma é então comparada com a distribuição da soma dos postos sob a hipótese nula.

Vamos retornar ao conjunto de dados *intake*, seção 16.3.1. A tabela 16.3 ordena os valores da diferença entre as variáveis *pre* e *post* na primeira coluna e soma os postos positivos na segunda coluna. Nesse exemplo, todos os postos são positivos, porque todas as diferenças são positivas.



Tabela 16.3: Cálculo da soma dos postos positivos para a diferença entre as variáveis *pre* e *post*. A última linha mostra a soma dos postos positivos.

pre-post	Postos pre-post
725	1
745	2
1020	3
1250	4
1330	5
1350	6
1435	7
1540	8,5
1540	8,5
1755	10
1835	11
	<b>66</b>

Vamos realizar o teste de Wilcoxon para amostras pareadas no *R Commander* para comparar as variáveis *pre* e *post* do conjunto de dados *intake*. Vamos realizar um teste de hipótese bilateral e calcular o intervalo de confiança ao nível de 90%. Tendo selecionado o conjunto de dados *intake*, utilizamos a seguinte opção do menu do *R Commander* para realizar de Wilcoxon para amostras pareadas:

Estatísticas ⇒ Testes Não-Paramétricos ⇒ Teste de Wilcoxon (amostras pareadas)

Após a seleção do teste, é preciso definir as variáveis que correspondem à medida do consumo de energia antes e depois da menstruação (figura 16.19).

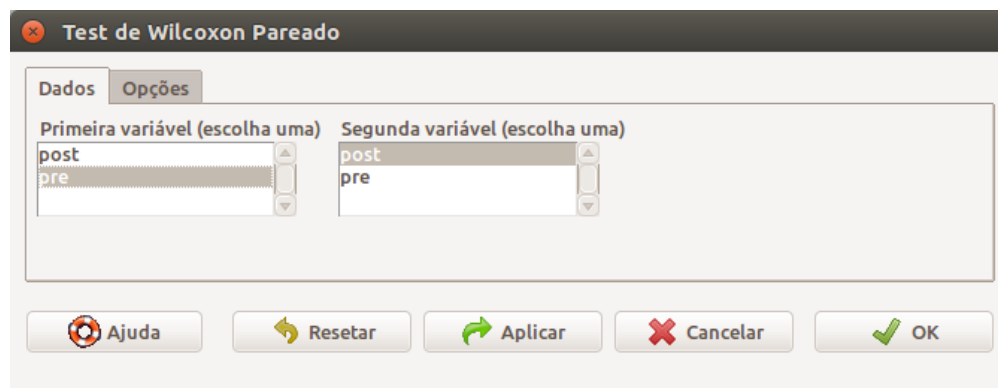


Figura 16.19: Seleção das variáveis que correspondem à medida do consumo de energia antes e depois da menstruação.

Ao clicarmos na guia *Opções* na caixa de diálogo da figura 16.19, podemos selecionar se o teste será bilateral ou unilateral e o tipo de teste (figura 16.20). Vamos selecionar a opção *Exato*.

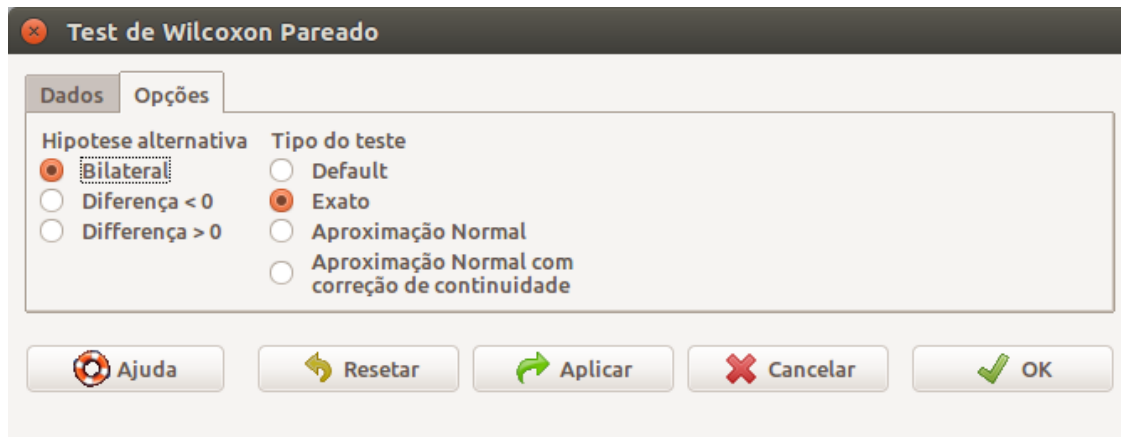


Figura 16.20: Definindo as opções do teste de Wilcoxon para amostras pareadas. Observem que não é possível especificar o nível de confiança nessa caixa de diálogo.

Ao clicarmos em OK na figura 16.20, o teste de Wilcoxon para amostras pareadas é realizado como a seguir. Observem que o parâmetro *paired* = *TRUE*.

```
with(intake, median(pre - post, na.rm=TRUE)) # median difference
```

```
## [1] 1350
```

```
with(intake, wilcox.test(pre, post, alternative='two.sided',
                        exact=TRUE, paired=TRUE))
```

```
##
## Wilcoxon signed rank test with continuity correction
##
## data: pre and post
## V = 66, p-value = 0.00384
## alternative hypothesis: true location shift is not equal to 0
```

Os resultados mostram a mediana da diferença entre os valores de *pre* e *post* e o resultado do teste estatístico, com valor de  $p = 0,00384$ . Considerando o nível de significância igual a 10%, a hipótese nula de igualdade de localização das distribuições das variáveis *pre* e *post* é rejeitada.

Para calcularmos o intervalo de confiança para a diferença de localização entre as medidas das duas populações, é necessário utilizar a linha de comando. A função *wilcox.test* com as especificações do nível de confiança (*conf.level* = 0.90) e determinando que o intervalo de confiança seja calculado (*conf.int* = *TRUE*) é mostrada a seguir, e os resultados são apresentados logo após.

```
with(intake, wilcox.test(pre, post, alternative='two.sided', exact=TRUE,
                        paired=TRUE, conf.int=TRUE, conf.level=0.90))
```

```
##
## Wilcoxon signed rank test with continuity correction
##
## data: pre and post
## V = 66, p-value = 0.00384
## alternative hypothesis: true location shift is not equal to 0
## 90 percent confidence interval:
## 1132.5 1540.0
## sample estimates:
## (pseudo)median
## 1341.332
```

A pseudo mediana mostrada acima é o estimador de Hodges-Lehmann (Wikipedia, 2019). Ele é um estimador não paramétrico e robusto de um parâmetro de localização da população. Para populações que são simétricas em relação à mediana, como a distribuição gaussiana ou t de Student, o estimador de Hodges-Lehmann é um estimador consistente e não enviesado da mediana da população. Para populações não simétricas, o estimador de Hodges-Lehmann é um estimador da *pseudo mediana*, que é proximamente relacionada à mediana da população.

## 16.4 Teste t pareado x Teste t não pareado

Vamos considerar o conjunto de dados *sleep* da biblioteca *datasets* (GPL-3). As funções abaixo carregam o pacote *datasets*, o conjunto de dados *sleep* e mostram as observações de 2 pacientes (2 linhas por paciente) do conjunto de dados *sleep*.

```
library(datasets)
data(sleep, package="datasets")
sleep[c(1,2,11,12),]
```

```
##      extra group ID
## 1      0.7      1  1
## 2     -1.6      1  2
## 11     1.9      2  1
## 12     0.8      2  2
```

O conjunto de dados *sleep* mostra o efeito de dois medicamentos soporíficos em 10 pacientes. A variável *extra* indica o aumento de horas de sono após o uso dos medicamentos. A variável *group* possui dois valores: 1 indica um dos medicamentos e 2 indica o outro medicamento. A variável *ID* indica o paciente. Assim a primeira linha mostra o número de horas extras de sono do paciente 1 após o uso do medicamento 1. A linha 11 mostra o número de horas extras de sono do paciente 1 após o uso do medicamento 2 e assim por diante.

Observem que são realizadas duas medidas em cada paciente, cada medida com um medicamento. Essas medidas tendem a ser correlacionadas. Vamos ver o que acontece se tratássemos esses dados como se fossem duas amostras de pacientes distintos e realizássemos um teste t para amostras independentes (não pareado). Utilizando a variável *group* para separar os dois grupos de pacientes e seguindo passos análogos aos da seção 16.2, executaríamos a função *t.test* conforme a seguir. Dessa vez, vamos definir o nível de confiança igual a 95%.

```
t.test(extra~group, alternative='two.sided', conf.level=.95,
       var.equal=FALSE, data=sleep)

##
##  Welch Two Sample t-test
##
## data:  extra by group
## t = -1.8608, df = 17.776, p-value = 0.07939
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.3654832  0.2054832
## sample estimates:
## mean in group 1 mean in group 2
##           0.75           2.33
```

O intervalo de confiança para a diferença de médias entre os dois medicamentos inclui o valor 0, o valor de p é 0,079 e, portanto, a hipótese nula de igualdade de médias não é rejeitada ao nível de significância de 5%. Porém, como os pacientes são os mesmos nos dois grupos, o teste t para amostras independentes não é o mais indicado para essa situação.

A interface gráfica do *R Commander* não permite realizar um teste pareado com os dados de *sleep*, porque os dados não estão organizados conforme os dados da seção anterior (conjunto de dados *intake*), onde cada medida é representada por uma variável diferente. Precisamos usar a linha de comando para realizar o teste t pareado. A seguinte função realiza o teste t pareado para os dados de *sleep*.

```
with(sleep, t.test(extra[group == 1], extra[group == 2], paired = TRUE))

##
##  Paired t-test
##
## data:  extra[group == 1] and extra[group == 2]
## t = -4.0621, df = 9, p-value = 0.002833
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.4598858 -0.7001142
## sample estimates:
## mean of the differences
```

##

-1.58

Agora o valor de  $p$  é 0,0028 (estatisticamente significativo) e o intervalo de confiança não inclui o zero. Para essa situação, o teste  $t$  pareado é mais poderoso do que o teste  $t$  não pareado, porque o desvio padrão da diferença de variáveis é calculado levando-se em conta somente a diferença dos valores em cada indivíduo. Assim o desvio padrão é bem menor do que o calculado no teste  $t$  para amostras independentes, que leva em conta a diferença entre indivíduos.

Vamos entender o comando. A função *with* possui dois parâmetros: um conjunto de dados (nesse exemplo *sleep*), e uma outra função (*t.test*), significando que a função *t.test* será aplicada sobre o conjunto de dados *sleep*. Nesse exemplo, três parâmetros são especificados para a função *t.test*. O primeiro parâmetro, *extra[group == 1]*, seleciona as medidas (*extra*) do medicamento 1 (*group == 1*). A expressão entre colchetes funciona como um seletor dos valores da variável que precede os colchetes. O segundo parâmetro, *extra[group == 2]*, seleciona as medidas (*extra*) do medicamento 2 (*group == 2*). Finalmente o terceiro parâmetro, *paired = TRUE*, informa que o teste  $t$  é pareado.

Vamos obter o diagrama de comparação de quantis para a diferença da variável *extra* para os dois medicamentos. Os seguintes comandos geram o diagrama (figura 16.21):

```
extra1 <- subset(sleep$extra, sleep$group == 1)
extra2 <- subset(sleep$extra, sleep$group == 2)
diff <- extra1 - extra2
qqPlot(diff, dist="norm", id=list(method="y", n=0))
```

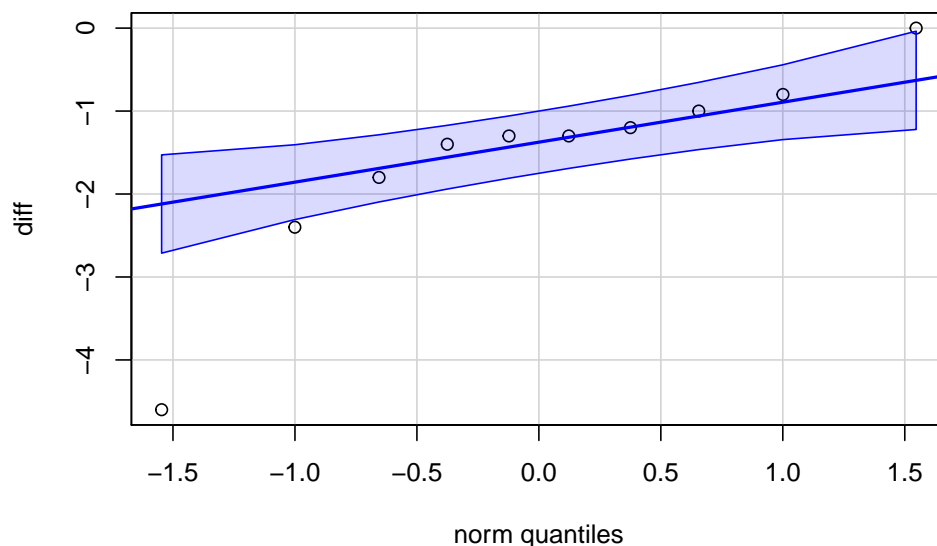


Figura 16.21: Gráfico de comparação de quantis da normal para a diferença dos valores da variável *extra* do conjunto de dados *sleep* para cada um dos medicamentos.

O primeiro comando cria a variável *extra1* que contém os valores de *extra* para o medicamento 1. A função *subset* tem dois parâmetros: o primeiro informa a variável de onde o subconjunto será extraído (*sleep\$extra*); o segundo parâmetro informa o critério de seleção (*sleep\$group == 1*).

O segundo comando cria a variável *extra2* que contém os valores de *extra* para o medicamento 2. O terceiro comando faz a subtração das duas variáveis *extra1* e *extra2*, e a última função cria o diagrama. O diagrama indica que a variável *diff* não possui grandes desvios de uma distribuição normal, especialmente se considerarmos que a amostra é pequena.

Para o conjunto de dados *sleep*, não é possível fazer o teste de Wilcoxon para amostras pareadas, usando diretamente o menu do *R Commander*, já que os valores da variável *extra* não estão separados em duas variáveis, uma para cada medicamento. A função a seguir mostra como realizar o teste de Wilcoxon para amostras pareadas com os dados de *sleep*, especificando o nível de confiança de 0,95 e solicitando o cálculo do intervalo de confiança da *pseudo mediana*. Observem que os resultados levam às mesmas conclusões do teste t pareado.

```
with(sleep, wilcox.test(extra[group == 1], extra[group == 2],
                        alternative='two.sided', exact=TRUE, paired=TRUE,
                        conf.int = TRUE, conf.level = 0.95))
```

```
##
## Wilcoxon signed rank test with continuity correction
##
## data: extra[group == 1] and extra[group == 2]
## V = 0, p-value = 0.009091
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
## -2.949921 -1.050018
## sample estimates:
## (pseudo)median
## -1.400031
```

## 16.5 Resumo das análises para comparar médias entre 2 grupos

Diante do exposto nas seções anteriores, podemos fazer um resumo dos principais testes utilizados para comparar medidas de localização entre dois grupos. Vamos separar em duas seções: amostras independentes e amostras dependentes.

### 16.5.1 Amostras Independentes

A tabela da figura 16.22 resume aproximadamente as análises que se aplicam a cada situação, quando se comparam medidas de localização entre dois grupos independentes.

Suposições	Estatística	Intervalo de Confiança
<ul style="list-style-type: none"> <li>Populações normalmente distribuídas</li> <li>Variâncias conhecidas</li> </ul>	<b>Z Normal Padrão</b> $Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}}$	$(\bar{X}_1 - \bar{X}_2) \pm z_{1-\alpha/2} \sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}$
<ul style="list-style-type: none"> <li>Populações normalmente distribuídas</li> <li>Variâncias desconhecidas, mas supostas iguais baseado nos dados da amostra ou supostamente diferentes, mas com <math>n_1 \approx n_2</math></li> </ul>	<b>t de Student</b> $T = \frac{(\bar{X}_1 - \bar{X}_2)}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ $gl = n_1 + n_2 - 2$ $S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$	$(\bar{X}_1 - \bar{X}_2) \pm t_{gl, 1-\alpha/2} S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
<ul style="list-style-type: none"> <li>Populações não normalmente distribuídas</li> <li>Variâncias desconhecidas</li> <li><math>n_1, n_2 \geq 30</math></li> </ul>	<b>Z Normal Padrão</b> $Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}}$	$(\bar{X}_1 - \bar{X}_2) \pm z_{1-\alpha/2} \sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}$
<ul style="list-style-type: none"> <li>Populações normalmente distribuídas</li> <li>Variâncias desconhecidas, mas supostamente diferentes com base nos dados da amostra</li> <li><math>n_1 \neq n_2</math></li> </ul>	<b>t de Student</b> $T = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}}$ $gl = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$	$(\bar{X}_1 - \bar{X}_2) \pm t_{gl, 1-\alpha/2} \sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}$
<ul style="list-style-type: none"> <li>Populações divergindo bastante da Normal</li> <li><math>n_1, n_2 &lt; 30</math></li> </ul>	<b>Mann-Whitney-Wilcoxon</b>	

Figura 16.22: Análises aplicáveis para a comparação de médias entre dois grupos independentes.

## 16.5.2 Amostras dependentes

A tabela da figura 16.23 resume aproximadamente as análises que se aplicam a cada situação, quando se comparam medidas de localização entre dois grupos dependentes.

Suposições	Estatística	Intervalo de Confiança
<ul style="list-style-type: none"> <li>População das diferenças entre os grupos normalmente distribuída</li> <li><math>n &lt; 30</math></li> </ul>	<b>t de Student</b> $T = \frac{\bar{D} - \mu_D}{E_D}$ $E_D = \frac{S_D}{\sqrt{n}}$	$\bar{D} \pm t_{n-1, \alpha/2} E_d$
<ul style="list-style-type: none"> <li>População das diferenças entre os grupos não normalmente distribuída</li> <li><math>n \geq 30</math></li> </ul>	<b>Z Normal Padrão</b> $Z = \frac{\bar{D} - \mu_D}{E_d}$ $E_D = \frac{S_D}{\sqrt{n}}$	$\bar{D} \pm z_{1-\alpha/2} E_D$
<ul style="list-style-type: none"> <li>População das diferenças divergindo bastante da Normal</li> <li><math>n &lt; 30</math></li> </ul>	<b>Teste de Wilcoxon para amostras pareadas</b>	

Figura 16.23: Análises aplicáveis para a comparação de médias entre dois grupos dependentes.

## 16.6 Exercícios

- Com o conjunto de dados *birthwt*, do pacote [MASS](#) (GPL-2 | GPL-3), faça as atividades abaixo.
  - Verifique a ajuda para o conjunto de dados;
  - Compare as médias da variável *bwt* (peso ao nascer) entre os grupos das mães que fumavam ou não durante a gestação.
  - Obtenha o intervalo de confiança ao nível de 95% para a diferença das médias de peso ao nascer entre os dois grupos.
  - Verifique as suposições para a realização do teste t para amostras independentes.
  - Repita os itens “b” a “d” para a comparação das médias de peso ao nascer entre os grupos das mães com ou sem histórico de hipertensão.



- 2) O conjunto de dados *WeightLoss* do pacote *carData* ([GPL-2](#) | [GPL-3](#)) contém dados artificiais sobre perda de peso e auto-estima ao longo de três meses, para três grupos de indivíduos: Controle, Dieta e Dieta + Exercício.
- Verifique a ajuda para o conjunto de dados.
  - Crie um subconjunto de dados de *WeightLoss*, chamado *wlDietEx*, com dados somente dos indivíduos que fizeram dieta + exercício, por meio do comando ao final deste exercício.
  - Com o conjunto de dados *wlDietEx*, compare as médias de perdas de peso no 2º e 3º mês. Utilize o nível de significância igual a 10%.
  - Compare as médias de perdas de peso no 1º e 2º mês. Utilize o nível de significância igual a 10%.
  - Verifique as suposições para a realização do teste t pareado nos itens “c” e “d”.

```
wlDietEx <- subset(WeightLoss, subset=group == "DietEx",  
                  select=c(wl1,wl2,wl3))
```

# Capítulo 17

## Comparação de proporções

### 17.1 Introdução

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

Os métodos apresentados sobre comparação de médias de uma variável numérica entre duas amostras são aplicáveis quando se compara a média de uma variável contínua em duas amostras independentes ou dependentes.

Em muitos estudos, estamos interessados em relações entre variáveis categóricas nominais ou ordinais.

No capítulo 8 (Medidas de associação), foram apresentadas diferentes medidas de associação entre variáveis dicotômicas a partir de uma tabela 2x2 (tabela 17.1).

Tabela 17.1: Tabela 2x2 que verifica a associação entre duas variáveis dicotômicas.

Exposição	Desfecho Clínico		Total
	Sim	Não	
<i>Nível 1</i>	$n_{11}$	$n_{12}$	$n_{1+} = n_{11} + n_{12}$
<i>Nível 2</i>	$n_{21}$	$n_{22}$	$n_{2+} = n_{21} + n_{22}$
<i>Total</i>	$n_{+1} = n_{11} + n_{21}$	$n_{+2} = n_{12} + n_{22}$	$n = n_{1+} + n_{2+}$

O risco de um indivíduo apresentar o desfecho clínico de interesse quando está exposto ao nível 1 de um fator é dado por:

$$R_{N1} = \frac{n_{11}}{n_{1+}} \quad (17.1)$$

A expressão (17.1) é uma estimativa da probabilidade de um indivíduo apresentar o desfecho clínico de interesse ao estar exposto ao nível 1 do fator.

Analogamente, o risco de um indivíduo apresentar o desfecho clínico de interesse quando está exposto ao nível 2 de um fator é dado por:

$$R_{N2} = \frac{n_{21}}{n_{2+}} \quad (17.2)$$

A expressão (17.2) é uma estimativa da probabilidade de um indivíduo apresentar o desfecho clínico de interesse ao estar exposto ao nível 2 do fator.

Neste capítulo, vamos chamar  $R_{N1}$  e  $R_{N2}$  de  $p_1$  e  $p_2$ , respectivamente.

$$p_1 = R_{N1}, \quad p_2 = R_{N2}$$

Ao realizarmos um estudo e observarmos uma tabela 2x2 como a tabela 17.1, estamos interessados em responder às seguintes questões:

- 1) obter medidas de associação (diferença absoluta de riscos - DAR, risco relativo - RR, razão de chances - RC, etc.) para o estudo e testar a hipótese de que não haja associação entre as duas variáveis;
- 2) obter os intervalos de confiança das medidas de associação.

Recordando:

$$DAR = p_1 - p_2$$

$$RR = \frac{p_1}{p_2}$$

$$RC = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}$$

Neste capítulo, vamos responder a essas perguntas, considerando as três medidas de associação listadas acima a partir de duas situações distintas:

- 1) quando as amostras de indivíduos expostos aos diferentes níveis do fator em estudo são independentes;
- 2) quando as amostras de indivíduos expostos aos diferentes níveis do fator em estudo são dependentes (ou pareadas).

Testar a hipótese de que não havia associação entre a variável de exposição e o desfecho é equivalente a testar que o fator de exposição não tem influência sobre o desfecho, ou seja, que as variáveis fator de exposição e desfecho clínico são independentes, ou de maneira equivalente:

$$DAR = 0, \text{ ou } RR = RC = 1.$$

As três igualdades acima implicam que as proporções  $p_1$  e  $p_2$  são iguais.

Também vamos realizar um teste de independência das variáveis quando as variáveis de exposição ou de desfecho possuem mais de duas categorias.

O estudo de Hjerkind, Stenehjem e Nilsen (Hjerkind et al., 2017) é um exemplo de um estudo com amostras independentes. Trata-se de um estudo prospectivo que verifica a associação entre as variáveis categóricas adiposidade e atividade física e diabetes mellitus. A figura 17.1 mostra a tabela 3 desse estudo. Essa tabela mostra os resultados, expressos como risco relativo, do acompanhamento de 38.413 pessoas estratificadas em diversos grupos, ou níveis de atividade física, e por sexo.

Os estratos ou grupos são independentes, porque os grupos foram formados naturalmente a partir do nível de atividade física de cada pessoa que participou do estudo. Não houve, por exemplo, nenhum pareamento entre indivíduos de um estrato com outro por sexo, idade ou outra variável.

**Table 3** RR of diabetes according to physical activity measures, by gender

	Men				Women			
	C	Pers.	RR* (95% CI)	p Value	C	Pers.	RR* (95% CI)	p Value
Frequency per week								
No activity	73	1948	1.00 (reference)		97	1972	1.00 (reference)	
<1 per week	162	5518	0.96 (0.73 to 1.25)		129	5701	0.83 (0.64 to 1.07)	
1 per week	110	4610	0.77 (0.57 to 1.03)		124	5778	0.70 (0.53 to 0.92)	
2–3 per week	98	4062	0.84 (0.62 to 1.13)		94	4663	0.67 (0.50 to 0.88)	
≥4 per week	45	1881	0.55 (0.38 to 0.80)		63	2280	0.71 (0.52 to 0.96)	
P <sub>trend</sub>				0.001				0.005
Minutes per session								
No activity	73	1948	1.00 (reference)		97	1972	1.00 (reference)	
1–15	37	934	0.89 (0.61 to 1.29)		46	1384	0.73 (0.52 to 1.02)	
16–30	119	3509	0.94 (0.71 to 1.25)		121	5350	0.65 (0.50 to 0.86)	
31–60	120	5208	0.75 (0.56 to 1.00)		128	6349	0.75 (0.57 to 0.97)	
≥60	66	3081	0.70 (0.51 to 0.98)		28	1396	0.74 (0.49 to 1.12)	
P <sub>trend</sub>			0.011				0.056	
Intensity								
No activity	73	1948	1.00 (reference)		97	1972	1.00 (reference)	
Low	206	5149	0.91 (0.70 to 1.18)		279	9061	0.83 (0.66 to 1.04)	
Medium/high	133	7272	0.71 (0.53 to 0.95)		57	5285	0.62 (0.44 to 0.87)	
P <sub>trend</sub>				0.011				0.005

\*Adjusted for BMI (continuous); age (continuous); education (≤9 years, 10–12 years, >12 years; unknown); alcohol frequency in the past 2 weeks (no, 1–4, ≥5, abstainer; unknown); smoking (never, former, current, unknown); BP medication use (yes, no, unknown); prevalent CVD (yes, no, unknown).

BMI, body mass index; BP, blood pressure; C, cases; CVD, cardiovascular disease; NA, not applicable; Pers., persons; RR, risk ratio.

Figura 17.1: Exemplo de um estudo com amostras independentes. Tabela 3 do estudo de (Hjerkind et al., 2017) ([CC BY-NC](#)).

Para testar hipóteses e calcular intervalos de confiança para parâmetros populacionais, quando as observações nas amostras são independentes e as variáveis de exposição e desfecho são variáveis categóricas, a análise mais simples envolve a utilização do teste qui ao quadrado para amostras independentes quando as amostras são suficientemente grandes. Quando as amostras são pequenas, utiliza-se o chamado teste exato de Fisher-Erwin.

O estudo de Severo et al. (Severo et al., 2018) é um exemplo de um estudo com amostras dependentes. A figura 17.2 mostra parte da tabela 1 desse estudo, que mostra a associação entre fatores de risco e quedas em pacientes adultos hospitalizados. Por exemplo, entre os casos de queda 81% tinham limitação para caminhar contra 67% no grupo controle (não tiveram queda).

Apesar de a apresentação da tabela ser semelhante ao exemplo anterior, esse estudo é um estudo de caso-controle, onde cada caso foi pareado com um controle em relação ao sexo, à unidade e data da internação, ou seja, para cada caso de queda durante a hospitalização, foi identificado um controle (que não teve queda) que fosse do mesmo sexo e com a mesma data e unidade de internação. Dizemos que, neste caso, as amostras são dependentes e veremos que a análise é diferente da análise para o caso de amostras dependentes.

Tabela 1 - Distribuição dos fatores de risco intrínsecos e extrínsecos para as quedas (n=358). Porto Alegre, RS, Brasil, 2013-2014

Fatores de risco	Caso		Controle		Total	
	(n=179)	%	(n=179)	%	(n=358)	%
Fatores intrínsecos:						
Limitação para caminhar	145	81,0	120	67,0	265	74,0
Queda prévia	80	44,6	54	30,1	134	37,4
Desorientação/confusão	73	40,7	31	17,3	104	29,0
Micção frequente	57	31,8	31	17,3	88	24,5
Urgência urinária/intestinal	54	30,2	30	16,8	84	23,4
Período pós-operatório	41	22,9	58	32,4	99	27,6
Sonolência	37	20,7	24	13,4	61	17,0
Agitação	24	13,4	5	2,7	29	8,1

Figura 17.2: Exemplo de um estudo de caso-controle com amostras dependentes (estudo pareado). Parte da tabela 1 do estudo de (Severo et al., 2018) ([CC BY](#)).

Em um estudo com amostras dependentes, as observações nas 2 amostras são relacionadas, seja porque dois tratamentos distintos são aplicados em sequência a um conjunto de pacientes (a ordem de aplicação pode ser aleatória) e, então, uma variável de desfecho categórica é medida após cada tratamento, seja porque um controle é pareado com um caso em um estudo de caso-controle, sendo que cada par é formado por indivíduos semelhantes de acordo com um critério estabelecido, como no estudo de Severo et al. (Severo et al., 2018).

Uma análise simples, frequentemente utilizada nesses casos, é o teste de McNemar.

Na próxima seção, vamos ver como são realizados os testes qui ao quadrado e o teste exato de Fisher-Erwin para tabelas 2x2 com amostras independentes e como são calculados o intervalo de confiança para as medidas de associação DAR, RR e RC para amostras suficientemente grandes.

## 17.2 Comparação de proporções em duas amostras independentes

Nesta seção, serão abordados dois testes estatísticos para verificar a associação entre duas variáveis categóricas: o teste qui ao quadrado e o teste exato de Fisher. Também serão apresentados os intervalos de confiança aproximados, quando as amostras não são pequenas, para as medidas de associação: diferença de risco, risco relativo e razão de chances. Nesta seção, consideraremos que as duas amostras, correspondentes aos níveis 1 e 2 do fator de exposição, foram obtidas de maneira independente.

### 17.2.1 Teste qui ao quadrado

Os conteúdos desta seção e das seções 17.2.2 a 17.2.5 podem ser visualizados neste [vídeo](#).

O teste qui ao quadrado é utilizado para verificar se a associação entre duas variáveis dicotômicas é estatisticamente significativa, quando as amostras são suficientemente grandes.

A partir da tabela 17.1, podemos derivar a tabela 17.2, dividindo-se cada célula da tabela 17.1 por  $n$ , o tamanho total das duas amostras, obtendo-se então as proporções **observadas** em cada célula:

$$p_{11} = \frac{n_{11}}{n}, p_{12} = \frac{n_{12}}{n}, p_{21} = \frac{n_{21}}{n}, p_{22} = \frac{n_{22}}{n}$$

$$p_{1+} = \frac{n_{1+}}{n} = p_{11} + p_{12}, p_{+1} = \frac{n_{+1}}{n} = p_{11} + p_{21}$$

$$p_{2+} = \frac{n_{2+}}{n} = p_{21} + p_{22}, p_{+2} = \frac{n_{+2}}{n} = p_{12} + p_{22}$$

Tabela 17.2: Tabela 2x2 que verifica a associação entre duas variáveis dicotômicas, obtida da tabela 17.1, dividindo-se cada célula por  $n$ .

Exposição	Desfecho Clínico		Total
	Sim	Não	
<i>Nível 1</i>	$p_{11}$	$p_{12}$	$p_{1+} = p_{11} + p_{12}$
<i>Nível 2</i>	$p_{21}$	$p_{22}$	$p_{2+} = p_{21} + p_{22}$
<i>Total</i>	$p_{+1} = p_{11} + p_{21}$	$p_{+2} = p_{12} + p_{22}$	<b>1</b>

Sob a hipótese de independência das variáveis Exposição e Desfecho Clínico, as proporções **esperadas** de indivíduos em cada célula da tabela seriam:

$$p_{11} = P[(\text{Exposição} = \text{Nível 1}) \cap (\text{Desfecho Clínico} = \text{Sim})] = p_{1+} p_{+1}$$

$$p_{21} = P[(\text{Exposição} = \text{Nível 2}) \cap (\text{Desfecho Clínico} = \text{Sim})] = p_{2+} p_{+1}$$

$$p_{12} = P[(\text{Exposição} = \text{Nível 1}) \cap (\text{Desfecho Clínico} = \text{Não})] = p_{1+} p_{+2}$$

$$p_{22} = P[(\text{Exposição} = \text{Nível 2}) \cap (\text{Desfecho Clínico} = \text{Não})] = p_{2+} p_{+2}$$

A frequência esperada na célula correspondente ao *nível 1* da exposição e o valor *Sim* para o desfecho clínico seria dada por:

$$E_{11} = n p_{1+} p_{+1} = n \frac{n_{1+}}{n} \frac{n_{+1}}{n} = \frac{n_{1+} n_{+1}}{n}$$

De modo análogo, são obtidas as frequências esperadas para as demais células (tabela 17.3).

Tabela 17.3: Valores esperados em cada célula da tabela 17.1, sob a hipótese de independência das variáveis.

Exposição	Desfecho Clínico		Total
	Sim	Não	
<b>Nível 1</b>	$E_{11} = (n_{1+} n_{+1})/n$	$E_{12} = (n_{1+} n_{+2})/n$	<b><math>n_{1+}</math></b>
<b>Nível 2</b>	$E_{21} = (n_{2+} n_{+1})/n$	$E_{22} = (n_{2+} n_{+2})/n$	<b><math>n_{2+}</math></b>
<b>Total</b>	<b><math>n_{+1}</math></b>	<b><math>n_{+2}</math></b>	<b><math>n</math></b>

Sob a hipótese nula (independência dos eventos), espera-se que as frequências observadas ( $O_{ij}$ ) em cada célula não sejam muito diferentes das frequências esperadas. A estatística abaixo:

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (17.3)$$

segue aproximadamente uma distribuição qui ao quadrado, com 1 grau de liberdade, quando a hipótese nula é verdadeira. Assim valores suficientemente altos de  $\chi^2$  levam à rejeição da hipótese nula. Esse teste foi proposto pelo estatístico Karl Pearson (Pearson, 1900).

Vamos voltar à tabela 8.2 do capítulo 8, derivada do estudo de Hjerkind, Stenehjem e Nilsen (Hjerkind et al., 2017) e mostrada novamente na tabela 17.4.

Vamos realizar o teste qui ao quadrado para esse exemplo. A tabela 17.5 mostra as frequências esperadas para as células, caso a atividade física não tenha influência sobre a ocorrência de diabetes mellitus.

Então:

$$\chi^2 = \frac{(45 - 58)^2}{58} + \frac{(73 - 60)^2}{60} + \frac{(1836 - 1823)^2}{1823} + \frac{(1875 - 1888)^2}{1888} = 5,91$$

Usando o R, podemos obter a probabilidade de se obter um valor tão ou mais elevado do que 5,91 na distribuição qui ao quadrado com 1 grau de liberdade:

Tabela 17.4: Versão simplificada da tabela 3 do estudo de Hjerkind, Stenehjem e Nilsen (Hjerkind et al., 2017) (CC BY-NC). Essa tabela mostra somente dois níveis de atividade física.

Atividade Física	Diabetes Mellitus		Total
	Sim	Não	
<i>Exercita 4+ vezes/semana</i>	45	1836	1881
<i>Inativo</i>	73	1875	1948
<i>Total</i>	118	3711	3829

Tabela 17.5: Valores esperados nas células da tabela 17.4 sob a hipótese de independência das variáveis.

Atividade Física	Diabetes Mellitus		Total
	Sim	Não	
<i>Inativo</i>	58	1823	1881
<i>Exercita 4+ vezes/semana</i>	60	1888	1948
<i>Total</i>	118	3711	3829

```
pchisq(5.91, 1, lower.tail = FALSE)
```

```
## [1] 0.01505517
```

O valor de p é igual a 0,015. Se o nível de significância do teste fosse de 5%, nós rejeitaríamos a hipótese nula de igualdade de riscos.

## 17.2.2 Intervalos de confiança para a DAR, o RR e a RC

### Diferença de riscos

Para amostras suficientemente grandes, o erro padrão (ep) e o intervalo de confiança (IC) para a diferença de riscos (DAR) podem ser aproximados por (Fleiss, 1981):

$$DAR = p_1 - p_2, \quad ep(DAR) = \sqrt{\frac{p_1(1-p_1)}{n_{1+}} + \frac{p_2(1-p_2)}{n_{2+}}}$$

$$IC(DAR) : (p_1 - p_2) \pm z_{\alpha/2} \sqrt{\frac{p_1(1-p_1)}{n_{1+}} + \frac{p_2(1-p_2)}{n_{2+}}}$$

Para os dados da tabela 17.4 e para um IC com 95% de confiança, temos:

$$p_1 = \frac{45}{1881} = 0,0239, \quad p_2 = \frac{73}{1948} = 0,0375$$

$$IC(DAR) : (0,0239 - 0,0375) \pm 1,96 \sqrt{\frac{0,0239 \cdot 0,9761}{1881} + \frac{0,0375 \cdot 0,9625}{1948}}$$



$$IC(DAR) = [-0,0246; -0,0026] = [-2,46\%; -0,26\%]$$

### Risco relativo

Para amostras suficientemente grandes, o erro padrão (ep) e o intervalo de confiança (IC) para o logaritmo neperiano do risco relativo (RR) podem ser aproximados por (Rothman et al., 2011):

$$RR = \frac{p_1}{p_2}, \quad ep[\ln(RR)] = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{1+}} + \frac{1}{n_{21}} + \frac{1}{n_{2+}}}$$

$$IC[\ln(RR)] : \ln(RR) \pm z_{\alpha/2} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{1+}} + \frac{1}{n_{21}} + \frac{1}{n_{2+}}}$$

Para os dados da tabela 17.4 e para um IC com 95% de confiança, temos:

$$RR = \frac{0,0239}{0,0375} = 0,64$$

$$IC[\ln(RR)] : \ln(0,64) \pm 1,96 \sqrt{\frac{1}{45} + \frac{1}{1881} + \frac{1}{73} + \frac{1}{1948}} = -0,45 \pm 0,366$$

$$IC[\ln(RR)] = [-0,81; -0,08]$$

Para obtermos o intervalo de confiança para o risco relativo, precisamos elevar a base “e” aos limites do intervalo de confiança de  $\ln(RR)$ :

$$IC(RR) = [e^{-0,81} - e^{-0,08}] = [0,44 - 0,92]$$

### Razão de chances

Para amostras suficientemente grandes, o erro padrão (ep) e o intervalo de confiança (IC) para o logaritmo neperiano da razão de chances (RC) podem ser aproximados por (Fleiss, 1981):

$$RC = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}, \quad ep[\ln(RC)] = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

$$IC[\ln(RC)] : \ln(RC) \pm z_{\alpha/2} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

Para os dados da tabela 17.4 e para um IC com 95% de confiança, temos:

$$RC = \frac{\frac{0,0239}{1-0,0239}}{\frac{0,0375}{1-0,0375}} = \frac{0,03896}{0,02449} = 0,63$$

$$IC[\ln(RC)] : \ln(0,63) \pm 1,96 \sqrt{\frac{1}{45} + \frac{1}{1836} + \frac{1}{73} + \frac{1}{1875}} = -0,462 \pm 0,377$$

$$IC[\ln(RC)] = [-0,84; -0,085]$$

Para obtermos o intervalo de confiança para a razão de chances, precisamos elevar a base “ $e$ ” aos limites do intervalo de confiança de  $\ln(RC)$ :

$$IC(RC) = [e^{-0,84} - e^{-0,085}] = [0,43 - 0,92]$$

### 17.2.3 Usando o epiR para o teste do qui ao quadrado e cálculo das medidas de associação

Recordando o capítulo 8, seção 8.3, podemos realizar as análises das duas seções anteriores por meio do *epiR*, utilizando os seguintes comandos:

```
library(epiR)
table <- matrix(c(45, 1836, 73, 1875), 2, 2, byrow=TRUE)
epi.2by2(table, method = "cohort.count", conf.level = 0.95)
```

##	Outcome +	Outcome -	Total	Inc risk *	Odds
## Exposed +	45	1836	1881	2.39	0.0245
## Exposed -	73	1875	1948	3.75	0.0389
## Total	118	3711	3829	3.08	0.0318

```
##
## Point estimates and 95 % CIs:
## -----
## Inc risk ratio          0.64 (0.44, 0.92)
## Odds ratio             0.63 (0.43, 0.92)
## Attrib risk *          -1.36 (-2.45, -0.27)
## Attrib risk in population * -0.67 (-1.67, 0.34)
## Attrib fraction in exposed (%) -56.64 (-125.88, -8.63)
## Attrib fraction in population (%) -21.60 (-39.84, -5.74)
## -----
## X2 test statistic: 5.883 p-value: 0.015
## Wald confidence limits
## * Outcomes per 100 population units
```

Nos resultados mostrados, o risco relativo é identificado por *Inc risk ratio*, e a diferença de riscos é denominada *Attrib risk*. Observem que os intervalos de confiança são próximos aos calculados pelas fórmulas apresentadas.

A seção 8.3 do capítulo sobre medidas de associação mostra como analisar uma tabela 2x2 para amostras independentes a partir de um conjunto de dados, tanto no *R Commander* quanto usando o pacote *epiR*.

### 17.2.4 Alternativas ao teste qui ao quadrado tradicional

A estatística apresentada pela expressão (17.3) é a mais usada para realizar um teste de hipótese em tabelas 2 x 2. Porém outras alternativas foram propostas. Pode-se mostrar que a estatística (17.3) pode também ser expressa como:

$$\chi^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{+1}n_{+2}n_{1+}n_{2+}} \quad (17.4)$$

onde  $n$  é a soma das frequências na tabela 17.1.

Egon Pearson (Pearson, 1947) propôs a substituição de  $n$  em (17.4) por  $(n - 1)$ :

$$\chi_{EP}^2 = \frac{(n - 1)(n_{11}n_{22} - n_{12}n_{21})^2}{n_{+1}n_{+2}n_{1+}n_{2+}} = \frac{n - 1}{n} \chi^2 \quad (17.5)$$

À medida que  $n$  aumenta, a razão  $(n-1)/n$  se aproxima de 1 e a expressão (17.5) se aproxima da expressão (17.4).

Yates (Yates, 1934) propôs uma outra alternativa, que ele chamou de **correção de continuidade**, que está disponível em diversos softwares estatísticos:

$$\chi_Y^2 = \frac{n(|n_{11}n_{22} - n_{12}n_{21}| - \frac{n}{2})^2}{n_{+1}n_{+2}n_{1+}n_{2+}} \quad (17.6)$$

Quão grande devem ser as amostras para que as análises apresentadas nas duas seções anteriores sejam válidas?

Essa é uma questão que tem suscitado longos debates. Campbell (Campbell, 2007) realizou uma comparação de diversas propostas e chegou à conclusão que a melhor política para a análise de tabelas 2x2 para estudos de coortes, ensaios randomizados ou estudos transversais é a seguinte:

- (1) quando todos os valores esperados nas células são maiores ou iguais a 1, utilize a proposta de Egon Pearson (expressão (17.5)) que é o qui ao quadrado tradicional com  $n$  substituído por  $n-1$ ;
- (2) para os demais casos, use o teste exato de Fisher–Irwin, com o teste bilateral realizado de acordo com a regra de Irwin (como mostrado na seção 17.2.5).

O *R Commander* não oferece uma opção para realizar o teste qui ao quadrado de acordo com a proposta de Egon Pearson. Entretanto pode-se obter o resultado do teste qui ao quadrado do *R Commander* e modificá-lo para implementar o teste de Egon Pearson (Pearson, 1947).

Vamos repetir o teste qui ao quadrado da seção 8.3 do capítulo 8 (figuras 8.2 a 8.4). Os comandos a seguir foram executados para montar a tabela 2 x 2 e realizar o teste qui ao quadrado para as variáveis *diab* (diabético) e *dead* (óbito) do conjunto de dados *stroke* do pacote *ISwR* (GPL-2 | GPL-3).

```
data(stroke, package="ISwR")
local({
  .Table <- xtabs(~diab+dead, data=stroke)
  cat("\nFrequency table:\n")
  print(.Table)
  .Test <- chisq.test(.Table, correct=FALSE)
  print(.Test)
})
```

```
##
## Frequency table:
##      dead
## diab FALSE TRUE
## No    308  414
## Yes   35   62
##
## Pearson's Chi-squared test
##
## data:  .Table
## X-squared = 1.5196, df = 1, p-value = 0.2177
```

A tabela 2 x 2 gerada (*.Table*) nesse teste pode ser utilizada para, então, substituir  $n$  por  $n-1$  na expressão (17.4) e realizar o teste qui ao quadrado segundo Egon Pearson. A sequência de comandos abaixo realiza este procedimento:

```
.Table <- xtabs(~diab+dead, data=stroke)
n <- sum(.Table)
r <- as.numeric(chisq.test(.Table, correct = FALSE)$statistic)
pchisq(r*(n-1)/n, 1, lower.tail = FALSE)

## [1] 0.2179651
```

O primeiro comando recria a tabela 2 x 2, que é armazenada no objeto *.Table*. O comando seguinte obtém a frequência total da tabela (soma de todas as células). O terceiro comando recupera a estatística do teste qui ao quadrado tradicional, expressão (17.4). O último comando substitui  $n$  por  $n-1$  no cálculo da expressão (17.4) para a tabela 2 x 2 e calcula o valor de  $p$ . O resultado é 0,218, bastante próximo ao do teste qui ao quadrado tradicional, já que o valor de  $n$  (819) é muito próximo de  $n - 1$ .

### 17.2.5 Teste exato de Fisher-Irwin

Para tabelas 2x2 com valores esperados muito pequenos nas células, uma alternativa bastante utilizada é o chamado teste exato de Fisher (ou teste exato de Fisher-Irwin). Esse método consiste em calcular as probabilidades associadas a todas as tabelas possíveis que possuem os mesmos totais nas linhas e colunas que os valores observados na tabela, com a suposição de

que a hipótese nula é verdadeira, ou seja, que as variáveis representadas nas linhas e colunas são independentes.

Considerando a tabela 17.6, supondo que a variável de exposição não seja relacionada ao desfecho e que os totais nas linhas e colunas ( $n_{+1}$ ,  $n_{+2}$ ,  $n_{1+}$ ,  $n_{2+}$ ) sejam fixos, então a frequência da célula correspondente ao *nível 1* da variável exposição e ao nível *Sim* da variável desfecho é uma variável aleatória que segue uma distribuição conhecida por **distribuição hipergeométrica**.

Tabela 17.6: Tabela 2x2, fixando os totais nas linhas e colunas.

Exposição	Desfecho Clínico		Total
	Sim	Não	
<b>Nível 1</b>	$n_{11}$	$n_{12}$	$n_{1+}$
<b>Nível 2</b>	$n_{21}$	$n_{22}$	$n_{2+}$
<b>Total</b>	$n_{+1}$	$n_{+2}$	$n = n_{1+} + n_{2+}$

Pode-se mostrar nesse caso que a probabilidade de se observar a frequência  $n_{11}$  é dada por:

$$P(n_{11}) = \frac{\binom{n_{1+}}{n_{11}} \binom{n_{2+}}{n_{+1}-n_{11}}}{\binom{n}{n_{+1}}} = \frac{n_{1+}!n_{+1}!n_{2+}!n_{+2}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!} \quad (17.7)$$

Como exemplo, suponhamos que um estudo gerou a tabela 17.7.

Tabela 17.7: Tabela 2x2 com valores esperados muito pequenos nas células.

Exposição	Desfecho Clínico		Total
	Sim	Não	
<b>Nível 1</b>	<b>1</b>	<b>6</b>	<b>7</b>
<b>Nível 2</b>	<b>5</b>	<b>3</b>	<b>8</b>
<b>Total</b>	<b>6</b>	<b>9</b>	<b>15</b>

Supondo que os totais nas linhas e colunas sejam fixos (7, 8, 6, 9), então podem ser montadas 7 tabelas possíveis com esses totais (figura 17.3). A tabela 17.7 corresponde à segunda tabela (em negrito) na figura. Para cada tabela na figura 17.3, é mostrada a probabilidade de serem observados os seus valores, supondo que as variáveis sejam independentes. A partir das probabilidades de cada tabela, nós podemos calcular a probabilidade de obter a tabela observada, ou uma tabela menos provável, quando a hipótese nula é verdadeira.

Por exemplo, a probabilidade de se observar a tabela 17.7 é dada por:

$$P(n_{11} = 1) = \frac{7!6!8!9!}{15!1!6!5!3!} = 0,07832$$

<table><tr><td>0</td><td>7</td></tr><tr><td>6</td><td>2</td></tr></table> <p>P = 0,00559</p>	0	7	6	2	<table><tr><td>1</td><td>6</td></tr><tr><td>5</td><td>3</td></tr></table> <p>P = 0,07832</p>	1	6	5	3	<table><tr><td>2</td><td>5</td></tr><tr><td>4</td><td>4</td></tr></table> <p>P = 0,29371</p>	2	5	4	4
0	7													
6	2													
1	6													
5	3													
2	5													
4	4													
<table><tr><td>3</td><td>4</td></tr><tr><td>3</td><td>5</td></tr></table> <p>P = 0,39161</p>	3	4	3	5	<table><tr><td>4</td><td>3</td></tr><tr><td>2</td><td>6</td></tr></table> <p>P = 0,19580</p>	4	3	2	6	<table><tr><td>5</td><td>2</td></tr><tr><td>1</td><td>7</td></tr></table> <p>P = 0,03357</p>	5	2	1	7
3	4													
3	5													
4	3													
2	6													
5	2													
1	7													
<table><tr><td>6</td><td>1</td></tr><tr><td>0</td><td>8</td></tr></table> <p>P = 0,0014</p>	6	1	0	8										
6	1													
0	8													

Figura 17.3: Todas as 7 tabelas possíveis de serem formadas, mantendo-se fixos os valores totais das linhas e colunas na tabela 17.7.

O valor de  $p$  do teste de Fisher-Irwin é obtido adicionando-se as probabilidades da tabela observada e de todas as demais que possuem probabilidades menores ou iguais à da tabela observada no estudo. A tabela observada possui probabilidade igual a 0,07832. As outras tabelas com probabilidades menores ou iguais a essa são as tabelas 1, 6 e 7 na figura 17.3. Logo:

$$\text{valor de } p = 0,07832 + 0,00559 + 0,03357 + 0,0014 = 0,1189$$

#### 17.2.5.1 Teste exato de Fisher-Irwin no R

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

Pode-se realizar o teste exato de Fisher no *R Commander* para um certo conjunto de dados, seguindo o procedimento para realizar o teste qui ao quadrado mostrado na seção 8.3 do capítulo sobre medidas de associação. Nesta seção, vamos mostrar como realizar o teste exato de Fisher partindo diretamente da tabela 2x2, tomando como exemplo a tabela 17.7.

No *R Commander*, selecionamos a opção:

Estatísticas  $\Rightarrow$  Tabelas de Contingência  $\Rightarrow$  Digite e analise tabela dupla entrada

Na tela da figura 17.4, digitamos as células da tabela e selecionamos o teste exato de Fisher na aba *Estatísticas* (figura 17.5).

**Digite tabela de dupla-entrada (two-way)**

**Tabela** | Estatísticas

Name for Row Variable (optional):

Name for Column Variable (optional):

Número de linhas:  2

Número de Colunas:  2

Entrar número:

	1	2
1	1	6
2	5	3

Ajuda | Resetar | Aplicar | Cancelar | OK

Figura 17.4: Entrada dos valores na tabela 2x2.

**Digite tabela de dupla-entrada (two-way)**

**Tabela** | **Estatísticas**

Computar Percentagens

- ☒ Percentual nas linhas
- ☐ Percentual nas colunas
- ☐ Percentagens do total
- ☐ Sem percentual

Teste de Hipóteses

- ☒ Teste de independência de Qui-Quadrado
- ☐ Componentes da estatística do Qui-quadrado
- ☐ Apresente frequências esperadas
- ☒ Teste exato de Fisher

Ajuda | Resetar | Aplicar | Cancelar | OK

Figura 17.5: Seleção do teste exato de Fisher.

Os resultados são mostrados a seguir. Observem que primeiramente é realizado o teste qui ao quadrado de Pearson. Uma mensagem (não mostrada aqui) alerta que a aproximação pode estar incorreta. Ao final é mostrado o teste de Fisher, juntamente com a medida da razão de chances.

```
.Table <- matrix(c(1,6,5,3), 2, 2, byrow=TRUE)
dimnames(.Table) <- list("rows"=c("1", "2"), "columns"=c("1", "2"))
```

```

.Table # Counts

##      columns
## rows 1 2
##      1 1 6
##      2 5 3

.Test <- chisq.test(.Table, correct=FALSE)
.Test

##
##  Pearson's Chi-squared test
##
## data:  .Table
## X-squared = 3.6161, df = 1, p-value = 0.05722

remove(.Test)
fisher.test(.Table)

##
##  Fisher's Exact Test for Count Data
##
## data:  .Table
## p-value = 0.1189
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.001827082 1.768053629
## sample estimates:
## odds ratio
##  0.1189474

remove(.Table)

```

Podemos ver que os valores de p, obtidos pelo teste de Fisher e pelo teste qui ao quadrado, são bastante diferentes, indicando que a aproximação do teste qui ao quadrado não é adequada neste exemplo.



## 17.3 Comparação de proporções em duas amostras dependentes

Os conteúdos desta seção e de suas subseções podem ser visualizados neste [vídeo](#).

Nesta seção, consideraremos que as duas amostras, correspondentes aos níveis 1 e 2 do fator de exposição, ou correspondentes aos diferentes níveis do desfecho, não foram obtidas de maneira independente. Essa situação pode acontecer, por exemplo, nos seguintes casos:

- 1) Ensaio controlado randomizado *cross-over*: nesses estudos, o mesmo paciente experimenta os tratamentos investigados em sequência, de tal modo que cada paciente é o próprio controle;
- 2) Estudo de coortes onde, para cada indivíduo exposto ao fator em estudo, um indivíduo não exposto ao fator é selecionado com base em uma ou mais características cujos valores concordam com os valores do indivíduo exposto, utilizando um critério definido a priori, por exemplo, a idade do indivíduo não exposto não deve ser diferente daquela do indivíduo exposto por mais de dois anos e que ambos sejam do mesmo sexo. Dizemos que cada indivíduo não exposto foi pareado com um indivíduo exposto;
- 3) estudo de caso-controle onde, para cada caso, um controle é selecionado com base em uma ou mais características cujos valores concordam com os valores do caso. Dizemos que cada controle foi pareado com um caso.

Em todos esses casos, dizemos que as amostras são pareadas.

Supondo que estamos com um estudo de coortes, com o fator em estudo e o desfecho dicotômicos, onde cada indivíduo exposto ao fator em estudo foi pareado com outro indivíduo não exposto ao fator, num total de  $n$  pares, a tabela 2 x 2 usualmente é montada como mostra a tabela 17.8.

Tabela 17.8: Montagem de uma tabela 2x2 para comparação de proporções em estudos de coortes com amostras pareadas.

		Fator ausente		
		<i>Desfecho positivo</i>	<i>Desfecho negativo</i>	
<i>Fator presente</i>	<i>Desfecho positivo</i>	<i>a</i>	<i>b</i>	<i>a+b</i>
	<i>Desfecho negativo</i>	<i>c</i>	<i>d</i>	<i>c+d</i>
		<i>a+c</i>	<i>b+d</i>	<i>n = a+b+c+d</i>

A primeira célula da tabela,  $a$ , representa o número de pares onde os dois elementos do par tiveram o desfecho (ou o nível 1 do desfecho).

A segunda célula da tabela,  $b$ , representa o número de pares onde o elemento do par exposto ao fator (ou o nível 1 do fator) teve o desfecho (ou o nível 1 do desfecho) e o outro elemento do par não teve o desfecho (ou teve o nível 2 do desfecho).

A terceira célula da tabela,  $c$ , representa o número de pares onde o elemento do par exposto ao fator (ou o nível 1 do fator) não teve o desfecho (ou teve o nível 2 do desfecho) e o outro elemento do par teve o desfecho (ou o nível 1 do desfecho).

A quarta célula da tabela,  $d$ , representa o número de pares onde os dois elementos do par não tiveram o desfecho (ou tiveram o nível 2 do desfecho).

O total da tabela é o total de pares do estudo ( $n$ ), ou  $2n$  indivíduos.

Verifiquem a diferença entre essa tabela e a tabela para amostras independentes (tabela 17.1).

A partir da tabela 17.8, o risco do desfecho positivo entre os que possuem o fator é:

$$p_1 = \frac{a + b}{n}$$

O risco do desfecho positivo entre os que não possuem o fator é:

$$p_2 = \frac{a + c}{n}$$

Logo a **diferença entre as duas proporções** é dada por:

$$DAR = p_1 - p_2 = \frac{b - c}{n} \quad (17.8)$$

O **risco relativo** é dado por:

$$RR = \frac{p_1}{p_2} = \frac{a + b}{a + c} \quad (17.9)$$

A **razão de chances** é estimada por [Rothman et al. (2011), página 286]:

$$RC = \frac{b}{c} \quad (17.10)$$

Para um estudo de caso-controle pareado, a tabela seria construída de maneira semelhante, apenas trocando as posições das variáveis fator de exposição e desfecho (tabela 17.9). A mesma fórmula (17.10) se aplica para o cálculo da razão de chances.

Tabela 17.9: Montagem de uma tabela 2x2 para comparação de proporções em um estudo de caso-controle com amostras pareadas.

		Desfecho negativo		
		<i>Fator presente</i>	<i>Fator ausente</i>	
<i>Desfecho positivo</i>	<i>Fator presente</i>	<i>a</i>	<i>b</i>	<i>a+b</i>
	<i>Fator ausente</i>	<i>c</i>	<i>d</i>	<i>c+d</i>
		<i>a+c</i>	<i>b+d</i>	<i>n = a+b+c+d</i>

### 17.3.1 Teste de McNemar

Há diversos testes propostos para testar a hipótese nula de independência das variáveis em amostras pareadas. Fagerland et al. (Fagerland et al., 2014) discutem algumas delas. Vamos apresentar duas delas, possivelmente os dois testes mais frequentemente usados.

#### a) Teste de McNemar assintótico:

Para amostras suficientemente grandes, o erro padrão de  $p_1 - p_2$  é dado por:

$$EP(p_1 - p_2) = \frac{\sqrt{b+c}}{n}$$

e a estatística

$$\chi^2 = \left( \frac{p_1 - p_2}{EP[p_1 - p_2]} \right)^2 = \frac{(b-c)^2}{b+c} \quad (17.11)$$

segue uma distribuição qui ao quadrado com 1 grau de liberdade.

#### b) Teste de McNemar assintótico com correção de continuidade:

Este teste é semelhante ao anterior, porém com uma correção de continuidade:

$$\chi^2 = \left( \frac{|p_1 - p_2| - 1/n}{EP[p_1 - p_2]} \right)^2 = \frac{(|b-c| - 1)^2}{b+c} \quad (17.12)$$

A estatística  $\chi^2$  segue uma distribuição qui ao quadrado com 1 grau de liberdade.

Para se testar uma hipótese com um nível  $\alpha$  de significância, compara-se o valor obtido de (17.11) ou (17.12) com o valor correspondente a  $\chi^2_{1,\alpha}$ .

### 17.3.2 Intervalos de confiança para a diferença de proporções, risco relativo e razão de chances

Para a estimativa dos intervalos de confiança para as medidas de associação expressas em (17.8), (17.9) e (17.10) também existem diversas propostas na literatura. Neste texto, vamos utilizar algumas recomendadas por Fagerland et al. (Fagerland et al., 2014).

#### 17.3.2.1 Intervalo de confiança para a diferença de proporções

Um dos métodos recomendados por Fagerland et al. para o cálculo do intervalo de confiança para a diferença de proporções é chamado de *Wald with Bonnett–Price Laplace adjustment*. Para o cálculo dos limites do intervalo de confiança, calcula-se inicialmente as duas quantidades:

$$\tilde{p}_{12} = \frac{b+1}{n+2}, \tilde{p}_{21} = \frac{c+1}{n+2}$$

Então os limites do intervalo de confiança para  $p_1 - p_2$  serão:

$$P_i = (\tilde{p}_{12} - \tilde{p}_{21}) - z_{\alpha/2} \sqrt{\frac{\tilde{p}_{12} + \tilde{p}_{21} - (\tilde{p}_{12} - \tilde{p}_{21})^2}{n+2}}$$

$$P_s = (\tilde{p}_{12} - \tilde{p}_{21}) + z_{\alpha/2} \sqrt{\frac{\tilde{p}_{12} + \tilde{p}_{21} - (\tilde{p}_{12} - \tilde{p}_{21})^2}{n+2}}$$

Limites fora do intervalo  $[-1, 1]$  são truncados.

#### 17.3.2.2 Intervalo de confiança para o risco relativo

Um dos métodos recomendados por Fagerland et al. para o cálculo do intervalo de confiança para o risco relativo é chamado de *Bonnett–Price hybrid Wilson score*.

Seja  $nd = a + b + c$  e definamos:

$$A = \sqrt{\frac{b+c+2}{(a+b+1)(a+c+1)}}$$

$$B = \sqrt{\frac{1 - \frac{a+b+1}{nd+2}}{(a+b+1)}}$$

$$C = \sqrt{\frac{1 - \frac{a+c+1}{nd+2}}{(a+c+1)}}$$

$$z = \frac{A}{B+C} z_{\alpha/2}$$

O intervalo do escore de Wilson para  $p_1$  é dado por:

$$[l_1, u_1] = \frac{2(a+b) + z^2 \pm \sqrt{z^2 + 4(a+b) \left(1 - \frac{a+b}{nd}\right)}}{2(nd + z^2)}$$

O intervalo do escore de Wilson para  $p_2$  é dado por:

$$[l_2, u_2] = \frac{2(a+c) + z^2 \pm \sqrt{z^2 + 4(a+c) \left(1 - \frac{a+c}{nd}\right)}}{2(nd + z^2)}$$

Finalmente o intervalo do risco relativo pelo método de *Bonett–Price hybrid Wilson score* é dado por:

$$[RR_i, RR_s] = \left[ \frac{l_1}{u_2}, \frac{u_1}{l_2} \right]$$

### 17.3.2.3 Intervalo de confiança para a razão de chances

Um dos métodos recomendados por Fagerland et al. para o cálculo do intervalo de confiança para a razão de chances é chamado de *Transformed Wilson score*. Por esse método, calcula-se inicialmente os valores de L e S por meio da expressão abaixo:

$$[L, S] = \frac{2b + z_{\alpha/2}^2 \pm z_{\alpha/2} \sqrt{z_{\alpha/2}^2 + 4b \left(1 - \frac{b}{b+c}\right)}}{2 \left(b + c + z_{\alpha/2}^2\right)}$$

O intervalo de confiança para a razão de chances é dado então por:

$$[RC_i, RC_s] = \left[ \frac{L}{1-L}, \frac{S}{1-S} \right]$$

## 17.3.3 Comparação de proporções entre duas amostras dependentes no R

O pacote *stats* no R possui uma função chamada *mcnemar.test* que realiza as duas versões do teste de McNemar apresentadas na seção 17.3.1. O parâmetro *correct = TRUE* realiza o teste com a correção de continuidade, enquanto se fizermos *correct = FALSE*, o teste será realizado sem a correção de continuidade. Se nada for especificado, a função assumirá que *correct = TRUE*. Essa função não calcula os intervalos de confiança para as medidas de associação. Vide a ajuda para o teste de McNemar para verificar como utilizá-lo.

Uma função que tanto realiza o teste de McNemar quanto calcula os intervalos de confiança apresentados nas seções anteriores é a função *paired\_proportions*, disponibilizada [neste arquivo](#), cujo código fonte é mostrado no apêndice B.

Vamos supor que o arquivo *paired\_proportions.R* tenha sido baixado na pasta *temp* do disco C do Windows. Para carregar esse arquivo e disponibilizar a função *paired\_proportions* para ser utilizada numa sessão do R, podemos executar o comando a seguir:

```
source("C:\\temp\\paired_proportions.R")
```

Vamos ilustrar o uso da função *paired\_proportions*, utilizando o conjunto de dados *backpain* do pacote *HSAUR2* (GPL-2). Para instalar o pacote *HSAUR2*, use a função:

```
install.packages("HSAUR2")
```

Ao solicitarmos a ajuda para o conjunto de dados *backpain*, obtemos a figura 17.6.

```
backpain {HSAUR2} R Documentation
```

**Driving and Back Pain Data**

**Description**

A case-control study to investigate whether driving a car is a risk factor for low back pain resulting from acute herniated lumbar intervertebral discs (AHLID).

**Usage**

```
data("backpain")
```

**Format**

A data frame with 434 observations on the following 4 variables.

**ID**

a factor which identifies matched pairs.

**status**

a factor with levels *case* and *control*.

**driver**

a factor with levels *no* and *yes*.

**suburban**

a factor with levels *no* and *yes* indicating a suburban resident.

**Details**

These data arise from a study reported in Kelsey and Hardy (1975) which was designed to investigate whether driving a car is a risk factor for low back pain resulting from acute herniated lumbar intervertebral discs (AHLID). A case-control study was used with cases selected from people who had recently had X-rays taken of the lower back and had been diagnosed as having AHLID. The controls were taken from patients admitted to the same hospital as a case with a condition unrelated to the spine. Further matching was made on age and sex and a total of 217 matched pairs were recruited, consisting of 89 female pairs and 128 male pairs.

**Source**

Jennifer L. Kelsey and Robert J. Hardy (1975), Driving of Motor Vehicles as a Risk Factor for Acute Herniated Lumbar Intervertebral Disc. *American Journal of Epidemiology*, **102**(1), 63-73.

Figura 17.6: Descrição das variáveis do conjunto de dados *backpain* do pacote *HSAUR2*.

O conjunto de dados *backpain* contém dados de um estudo de caso-controle para investigar se dirigir um carro é um fator de risco para dor lombar resultante de hérnia de disco intervertebral lombar aguda.

Os casos foram selecionados de pessoas que fizeram radiografias da parte inferior das costas e foram diagnosticadas como portadoras de hérnia de disco intervertebral lombar aguda.

Os controles foram retirados de pacientes internados no mesmo hospital que um caso com uma condição não relacionada à coluna vertebral e também pareados em idade e sexo, num total de 217 pares, consistindo de 89 pares femininos e 128 pares masculinos.

As variáveis no conjunto de dados são:

*status* – indica se o indivíduo é um caso ou um controle;

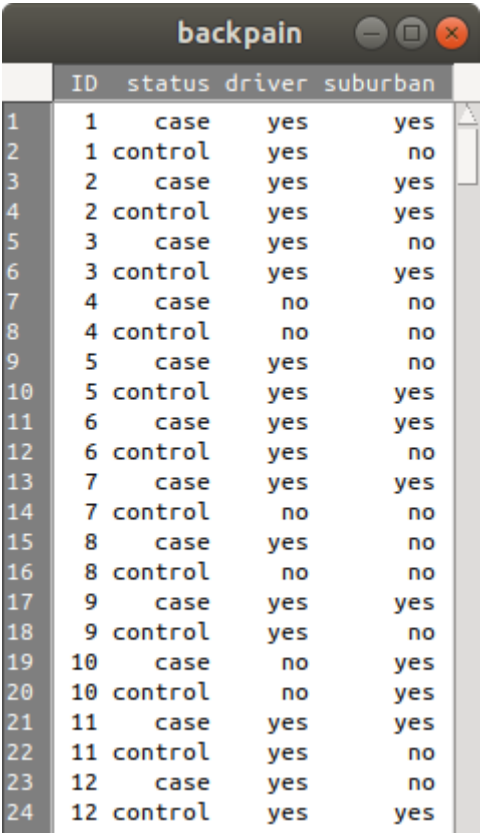
*driver* - indica se o indivíduo dirigia ou não;

*suburban* - indica se o indivíduo morava no subúrbio ou não;

*ID* - indica o par de indivíduos. Cada um dos 217 pares de indivíduos recebe um valor de *ID* diferente.

A presença da variável *ID* indica que se trata de um estudo com amostras pareadas.

A figura 17.7 mostra parte dos dados. Observem que cada valor de *ID* está associado a dois indivíduos, um caso e um controle.



	ID	status	driver	suburban
1	1	case	yes	yes
2	1	control	yes	no
3	2	case	yes	yes
4	2	control	yes	yes
5	3	case	yes	no
6	3	control	yes	yes
7	4	case	no	no
8	4	control	no	no
9	5	case	yes	no
10	5	control	yes	yes
11	6	case	yes	yes
12	6	control	yes	no
13	7	case	yes	yes
14	7	control	no	no
15	8	case	yes	no
16	8	control	no	no
17	9	case	yes	yes
18	9	control	yes	no
19	10	case	no	yes
20	10	control	no	yes
21	11	case	yes	yes
22	11	control	yes	no
23	12	case	yes	no
24	12	control	yes	yes

Figura 17.7: Parte do conjunto de dados *backpain* do pacote *HSAUR2*.

Em seguida, execute a sequência de comandos a seguir para realizar o teste de McNemar e calcular o intervalo de confiança para a razão de chances.

```
library(HSAUR2)
data("backpain", package = "HSAUR2")
source('paired_proportions.R')
paired_proportions(data=backpain, id='ID', row='driver', col='status',
                    row_ref='no', row_trt='yes', col_ref='control',
                    col_out='case', case_control=TRUE, alpha=0.05)

##          control
## case  yes no
##  yes 144 41
##   no   19 13
##
## McNemar's Chi-squared test with continuity correction
##
## data:  mat
## McNemar's chi-squared = 7.35, df = 1, p-value = 0.006706
##
## odds ratio Lower 95% CI Upper 95% CI
##    2.157895    1.260716    3.693543
```

Para utilizar a função *paired\_proportions*, o usuário deve carregar a função no R, por meio da função *source*, especificando entre aspas simples ou duplas o caminho no sistema de arquivos e o nome do arquivo que contém a função. No comando acima, o caminho foi especificado de acordo com a sintaxe do sistema operacional *Linux*.

Na chamada da função, o usuário deve especificar os seguintes parâmetros:

*data* - o *data.frame* a ser utilizado;

*id* - identificação dos pares - deve ser um variável da classe *factor*;

*row* - variável de exposição;

*col* - variável de desfecho;

*row\_ref* - nível de referência da variável de exposição;

*row\_trt* - outro nível de variável de exposição;

*col\_ref* - nível de referência da variável de desfecho;

*col\_out* - outro nível da variável de desfecho;

*case\_control* - indica se o estudo é de caso-controle (*TRUE*) ou não (*FALSE*). O padrão é *FALSE*;

*alpha* - nível de significância (o padrão é 0,05).

A primeira saída da função monta a tabela 2x2 na forma semelhante à apresentada na tabela 17.9. Em seguida, os resultados do teste de McNemar são apresentados, seguido do intervalo de confiança para a razão de chances. Como os dados de *backpain* são relativos a um estudo de caso-controle e o pareamento foi baseado na variável de desfecho (*status*), os intervalos de confiança para a diferença de proporções e risco relativo não são apresentados.

Nesse exemplo, o valor de alfa foi de 5% e o teste de McNemar foi realizado com correção de



continuidade. Caso se deseje outro nível de confiança, especifique o valor de alfa correspondente na variável *alpha*. Para não utilizar a correção de continuidade, especifique o parâmetro *correct = FALSE*. O exemplo abaixo utiliza o nível de confiança igual a 90% (*alpha = 0.1*) e não utiliza a correção de continuidade:

```
paired_proportions(backpain, 'ID', 'driver', 'status', 'no', 'yes', 'control',
                    'case', case_control=TRUE, alpha = 0.1, correct = FALSE)

##          control
## case  yes no
##  yes 144 41
##   no   19 13
##
## McNemar's Chi-squared test
##
## data:  mat
## McNemar's chi-squared = 8.0667, df = 1, p-value = 0.004509
##
## odds ratio Lower 90% CI Upper 90% CI
##    2.157895    1.372335    3.393129
```

## 17.4 Poder estatístico e tamanho amostral

Ao planejar um estudo, é importante estimar previamente o tamanho amostral para que o teste de hipótese utilizado tenha um certo poder estatístico ou o intervalo de confiança tenha uma certa precisão.

Pode-se estimar o tamanho amostral para estudos em que uma única proporção esteja envolvida, quando duas proporções são comparadas em duas amostras independentes ou quando as amostras são pareadas, ou em casos mais genéricos. Não vamos considerar aqui todas as possibilidades. Vamos ilustrar o uso do R para o cálculo do tamanho amostral e poder estatístico para o caso de um estudo experimental onde duas amostras independentes são utilizadas.

Suponhamos que um estudo esteja sendo planejado para comparar um tratamento padrão com um tratamento experimental para um determinado tipo de câncer e o desfecho principal seja se houve ou não a remissão da doença após um determinado intervalo de tempo. Um certo número de pacientes serão submetidos ao tratamento padrão e o mesmo número será submetido ao tratamento experimental. Para estimar o tamanho amostral desse estudo, ou seja, o número  $n$  de pacientes em cada grupo, vamos partir das seguintes suposições:

1. Seja  $p_1$  a proporção de pacientes que esperamos ter remissão da doença com o tratamento padrão;
2. Seja  $p_2$  a proporção de pacientes que esperamos ter remissão da doença com o tratamento experimental;

3. Seja  $\alpha$  o nível de significância do teste e  $1 - \beta$  o poder estatístico do teste para detectar a diferença  $p_2 - p_1$ , caso ela realmente exista.

A partir dos dados acima, calculamos a média das duas proporções:

$$\bar{p} = \frac{p_1 + p_2}{2}$$

Fleiss (Fleiss, 1981, página 41) apresenta a seguinte fórmula para o cálculo do tamanho amostral, quando a correção de continuidade no teste estatístico não é utilizada:

$$n = \left( \frac{(c_{1-\alpha/2} \sqrt{2\bar{p}(1-\bar{p})} - c_{1-\beta} \sqrt{p_1(1-p_1) + p_2(1-p_2)})}{p_2 - p_1} \right)^2 \quad (17.13)$$

onde  $c_{1-\alpha/2}$  corresponde à probabilidade  $P(Z > c_{1-\alpha/2}) = \alpha/2$  (quantil  $1-\alpha/2$ ) na distribuição normal e  $c_{1-\beta}$  à probabilidade  $(P(Z > c_{1-\beta}) = 1 - \beta$  (quantil  $\beta$ )).

Vamos supor que  $p_1 = 0,60$ ;  $p_2 = 0,70$ ,  $\alpha = 5\%$  e  $1 - \beta = 80\%$ .

Então:

$$0,025 = P(Z > c_{1-\alpha/2}) = P(Z > c_{0,975}) \Rightarrow c_{0,975} = 1,96 \text{ e}$$

$$0,80 = P(Z > c_{1-\beta}) = P(Z > c_{0,80}) \Rightarrow c_{0,80} = 0,84$$

$$n = \frac{(1,96 \sqrt{2 \cdot 0,65 \cdot 0,35} - 0,84 \sqrt{0,6 \cdot 0,4 + 0,7 \cdot 0,3})^2}{(0,1)^2} = 356$$

### 17.4.1 Usando o R Commander para calcular o tamanho amostral

No *R Commander*, vamos carregar o plugin *RcmdrPlugin.EZR*. Para isso, selecionamos no menu a opção:

Ferramentas  $\Rightarrow$  Carregar plug-in(s) do Rcmdr...

Na tela com a lista de plugins disponíveis, selecionamos o *RcmdrPlugin.EZR* e pressionamos o botão OK (figura 17.8). Será preciso reiniciar o *R Commander*.



Figura 17.8: Selecionando o *RcmdrPlugin.EZR*.

Após a reinicialização do *R Commander*, selecionamos a opção:

Stat. Analysis  $\Rightarrow$  Calc. sample size  $\Rightarrow$  Calc. sample size for comp. between two proportions

A figura 17.9 mostra a tela para configurar os parâmetros para o cálculo do tamanho amostral. Nessa figura, configuramos os parâmetros de acordo com o exemplo ao final da seção anterior. Como as duas amostras são iguais, colocamos o parâmetro *Sample size ratio* igual a 1.

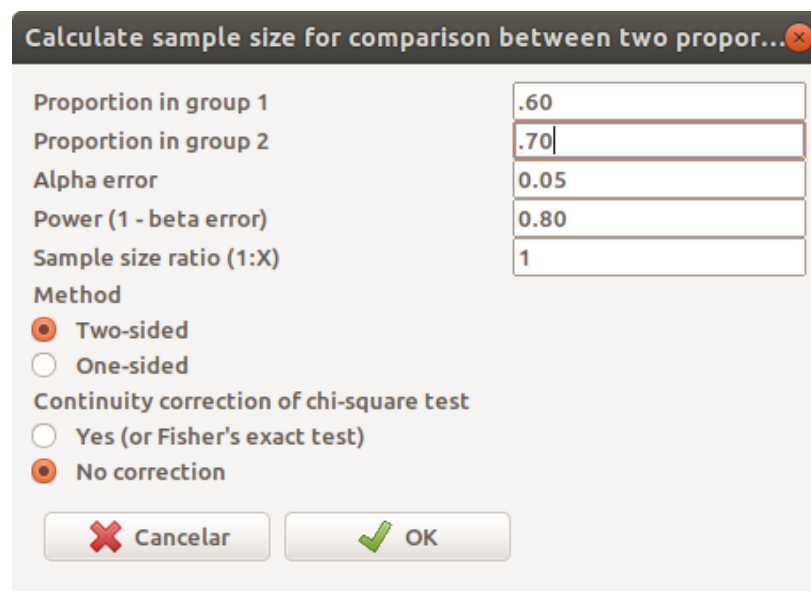


Figura 17.9: Configurando os parâmetros para o cálculo do tamanho amostral para o exemplo usado neste texto.

Ao clicarmos em OK na figura 17.9, o resultado é mostrado no *R Commander*, sendo o mesmo resultado obtido ao aplicarmos a fórmula (17.13). Além do resultado numérico, o *R Commander* exibe um gráfico que mostra o poder estatístico para diferentes valores de  $n$  (figura 17.10). A linha pontilhada horizontal corresponde ao poder estatístico igual a 80%.

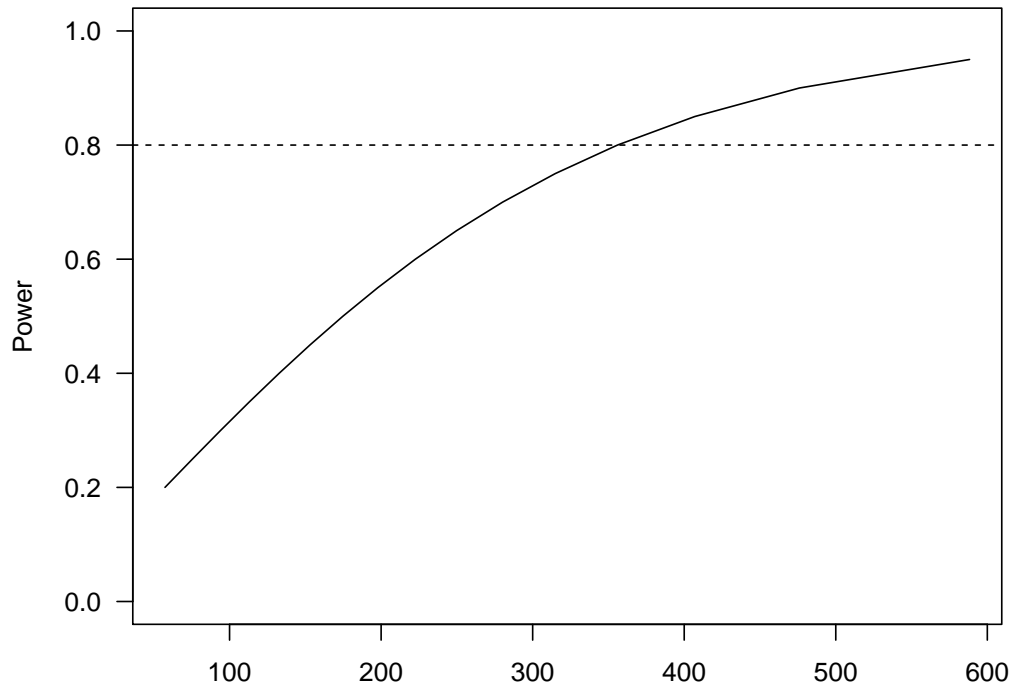


Figura 17.10: Tamanho amostral estimado para o exemplo considerado nesta seção.

```
##                               Assumptions
## P1                           0.6
## P2                           0.7
## Alpha                        0.05
##                               two-sided
## Power                        0.8
## N2/N1                        1
##
## Required sample size    Estimated
## N1                      356
## N2                      356
```

Continuando o exemplo anterior, vamos supor que o grupo experimental foi planejado para ter 1,5 vezes o número de pacientes do grupo do tratamento padrão. Então, modificando a figura 17.9 e fazendo *Sample size ratio* igual 1,5 (figura 17.11), obtemos o resultado mostrado na figura 17.12.

Calculate sample size for comparison between two propor...

Proportion in group 1

0.6

Proportion in group 2

0.7

Alpha error

0.05

Power (1 - beta error)

0.80

Sample size ratio (1:X)

1.5

Method

☒ Two-sided
 ☐ One-sided

Continuity correction of chi-square test

☐ Yes (or Fisher's exact test)
 ☒ No correction

✖ Cancelar

✔ OK

Figura 17.11: Alterando a razão entre o número de pacientes em cada grupo de tratamento para 1,5.

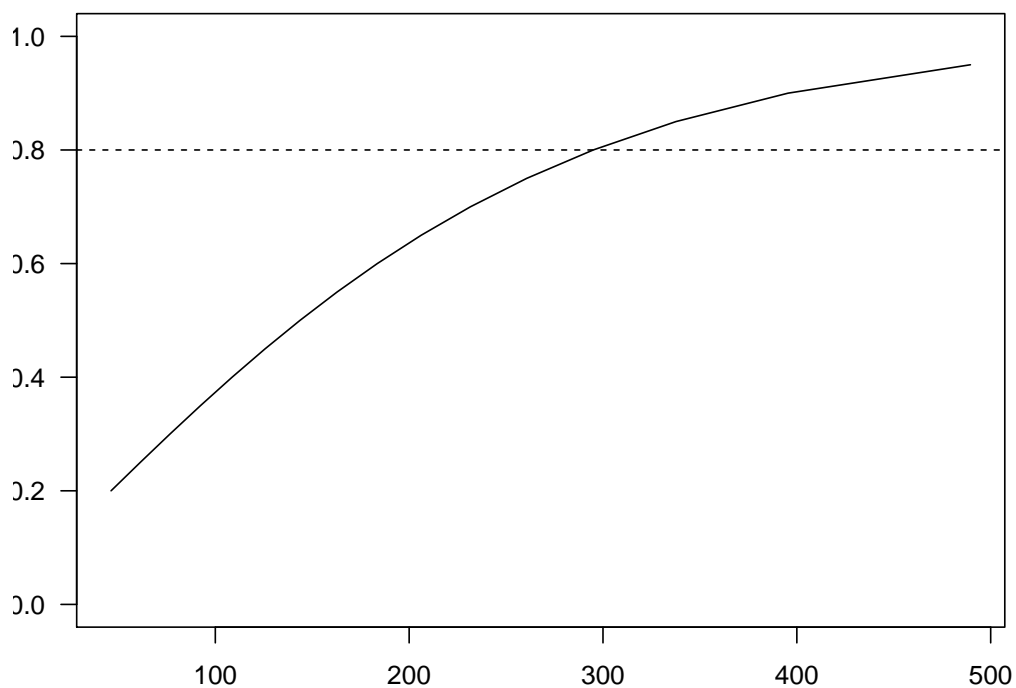


Figura 17.12: Tamanho amostral estimado para a configuração mostrada na figura 17.11.

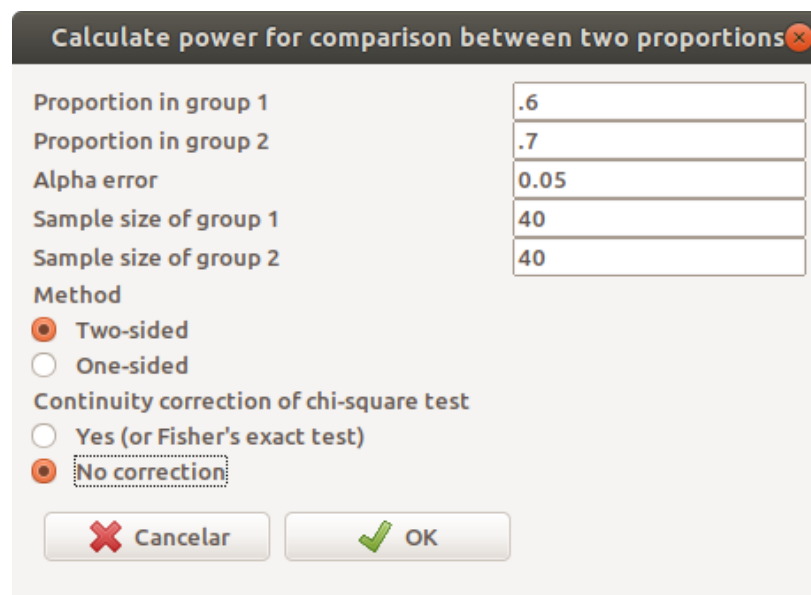
```
##               Assumptions
## P1              0.6
## P2              0.7
## Alpha           0.05
##               two-sided
```

```
## Power          0.8
## N2/N1          1.5
##
## Required sample size  Estimated
## N1              295
## N2              442.5
```

Alternativamente, conhecendo-se o tamanho amostral, pode-se estimar o poder estatístico de um teste que compara duas proporções em duas amostras independentes por meio da opção:

Stat. Analysis  $\Rightarrow$  Calc. sample size  $\Rightarrow$  Calc. power for comp. between two prop.

Na tela de configuração dos parâmetros (figura 17.13), vamos utilizar as mesmas proporções, erro  $\alpha$  e erro  $\beta$  dos exemplos anteriores e vamos supor que os dois grupos tenham 40 pacientes.



**Calculate power for comparison between two proportions**

Proportion in group 1: .6

Proportion in group 2: .7

Alpha error: 0.05

Sample size of group 1: 40

Sample size of group 2: 40

Method:

- ☒ Two-sided
- ☐ One-sided

Continuity correction of chi-square test:

- ☐ Yes (or Fisher's exact test)
- ☒ No correction

Buttons: Cancelar, OK

Figura 17.13: Configuração dos parâmetros para o cálculo do poder estatístico a partir do tamanho amostral.

Ao pressionarmos o botão OK, obtemos a curva poder estatístico x tamanho amostral (figura 17.14) e o poder estatístico igual a 15,2% para a configuração da figura 17.13, um valor bastante baixo.

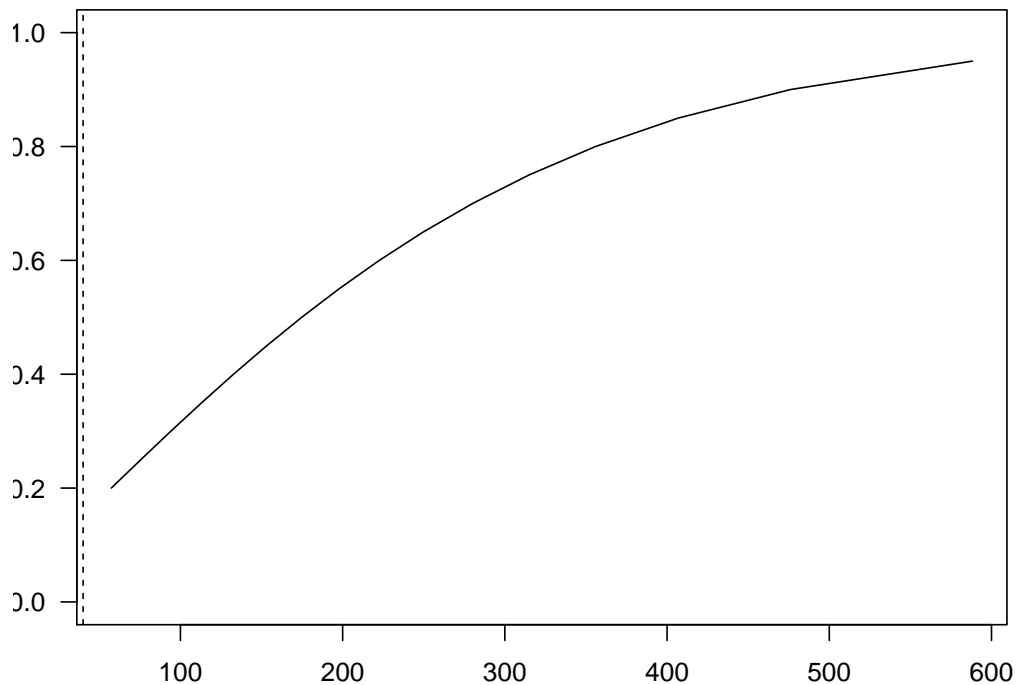


Figura 17.14: Poder estatístico de um teste de comparação de duas proporções com as configurações mostradas na figura 17.13.

```
##           Assumptions
## P1           0.6
## P2           0.7
## Alpha        0.05
##           two-sided
## Sample size
## N1           40
## N2           40
##
##           Estimated
## Power        0.152
```

## 17.5 Tabelas $r \times c$

Os conteúdos desta seção e da subseção 17.5.1 podem ser visualizados neste [vídeo](#).

Nesta seção, iremos considerar a situação onde as variáveis de exposição e desfecho tenham  $r$  e  $c$  categorias ( $r$  ou  $c > 2$ ), respectivamente, em vez de somente duas categorias, como nas seções anteriores. Vamos considerar somente o caso onde as amostras são independentes. No caso mais geral, uma tabela  $r \times c$  ( $r$  linhas e  $c$  colunas) terá a estrutura conforme a tabela 17.10.

Tabela 17.10: Tabela r x c que verifica a associação entre duas variáveis categóricas com r e c níveis respectivamente.

Exposição	Desfecho Clínico			Total
	Nível 1		Nível c	
<b>Nível 1</b>	$n_{11}$	...	$n_{1c}$	$n_{1+} = n_{11} + \dots + n_{1c}$
...	...	...	...	...
<b>Nível r</b>	$n_{r1}$	...	$n_{rc}$	$n_{r+} = n_{r1} + \dots + n_{rc}$
<b>Total</b>	$n_{+1} = n_{11} + \dots + n_{r1}$	...	$n_{+c} = n_{1c} + \dots + n_{rc}$	$n = n_{1+} + \dots + n_{r+}$

As proporções em cada célula da tabela 17.11 são calculadas a partir da tabela 17.10, dividindo-se a frequência em cada célula pela soma de todas as frequências (n) da tabela.

Tabela 17.11: Proporções em cada célula da tabela 17.10, obtidas por meio da divisão da frequência de cada célula por n.

Exposição	Desfecho Clínico			Total
	Nível 1		Nível c	
<b>Nível 1</b>	$p_{11}$	...	$p_{1c}$	$p_{1+} = p_{11} + \dots + p_{1c}$
...	...	...	...	...
<b>Nível r</b>	$p_{r1}$	...	$p_{rc}$	$p_{r+} = p_{r1} + \dots + p_{rc}$
<b>Total</b>	$p_{+1} = p_{11} + \dots + p_{r1}$	...	$p_{+c} = p_{1c} + \dots + p_{rc}$	$p = p_{1+} + \dots + p_{r+}$

Sob a hipótese de independência das variáveis *Exposição* e *Desfecho*, a proporção esperada em cada célula seria dada pela expressão:

$$p_{ij} = p_{i+}p_{+j}, \quad i = 1, \dots, r; \quad j = 1, \dots, c$$

e a frequência esperada em cada célula é calculada de acordo com a tabela 17.12.

Tabela 17.12: Valores esperados em cada célula da tabela 17.10, sob a hipótese de independência.

Exposição	Desfecho Clínico			Total
	Nível 1		Nível c	
<b>Nível 1</b>	$E_{11} = (n_{1+} n_{+1})/n$	...	$E_{1c} = (n_{1+} n_{+c})/n$	$n_{1+}$
...	...	...	...	...
<b>Nível r</b>	$E_{r1} = (n_{r+} n_{+1})/n$	...	$E_{rc} = (n_{r+} n_{+c})/n$	$n_{r+}$
<b>Total</b>	$n_{+1}$	...	$n_{+c}$	$n$

Sob a hipótese nula (independência dos eventos), espera-se que os valores observados não sejam muito diferentes dos valores esperados. A estatística abaixo, generalização da expressão



(17.3), segue aproximadamente uma distribuição qui ao quadrado, com  $(r-1)(c-1)$  graus de liberdade, quando a hipótese nula é verdadeira. Assim valores suficientemente altos de  $\chi^2$  levam à rejeição da hipótese nula.

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (17.14)$$

A aproximação (17.14) é válida para valores suficientemente grandes nas células da tabela. Entretanto o que pode ser considerado “suficientemente grande” é uma questão aberta. Uma recomendação frequentemente citada é que o teste qui ao quadrado não deve ser usado caso uma **frequência esperada seja menor que 2 ou se mais de 20% das frequências esperadas forem menores que 5** (Dawson and Trapp, 2001). Nesses casos, deve-se usar o teste exato de Fisher, ou usar uma estratégia de reduzir o número de categorias em cada variável.

Como temos reforçado nesse texto, o teste de hipótese é apenas um dos elementos da análise estatística. Mais importante em uma tabela  $r \times c$  é verificar os valores das medidas de associação e os respectivos intervalos de confiança. Há que se considerar que mais variáveis podem afetar as associações observadas em uma tabela  $r \times c$  (que contém somente duas variáveis). Análises mais sofisticadas envolvendo variáveis categóricas utilizam modelagem estatística para identificar padrões nos dados observados. Esses modelos estão fora do escopo deste texto (Agresti, 1996).

### 17.5.1 Análise de uma tabela $r \times c$ no R Commander

Vamos utilizar o conjunto de dados *bacteria* do pacote *MASS* (Venables and Ripley, 2002) ([GPL-2](#) | [GPL-3](#)). Esse conjunto de dados contém dados sobre testes sobre a presença da bactéria *H. influenzae* em crianças com otite média no norte da Austrália. As crianças foram randomizadas em três tratamentos (variável *trt*): placebo, medicamento (*drug*) e medicamento + encorajamento (*drug+*) para tomar o medicamento. A variável *y* indica a presença ou não de bactéria.

A presença de bactérias é investigada antes de iniciar o tratamento, na segunda, na quarta, na sexta e na décima-primeira semanas após o início do tratamento.

Vamos analisar a relação entre as variáveis tratamento e presença ou não de bactéria na 6ª semana.

Após carregarmos o conjunto de dados *bacteria*, reordenamos os níveis das variáveis *y* e *trt*:

```
bacteria$y <- with(bacteria, factor(y, levels=c('y','n')))  
bacteria$trt <- with(bacteria, factor(trt,  
                                     levels=c('drug+', 'drug', 'placebo')))
```

Em seguida, selecionamos a opção do menu do *R Commander*:

Estatísticas  $\Rightarrow$  Tabelas Contingência  $\Rightarrow$  Tabela de dupla entrada

Na caixa de diálogo da figura 17.15, selecionamos a variável cujas categorias aparecerão nas linhas da tabela (em geral a variável de exposição) e a variável cujas categorias aparecerão nas colunas da tabela (em geral a variável de desfecho). No campo *expressão (subset expression)*, digitamos a expressão lógica “*week == 6*” para incluir na análise somente os registros da 6ª semana.

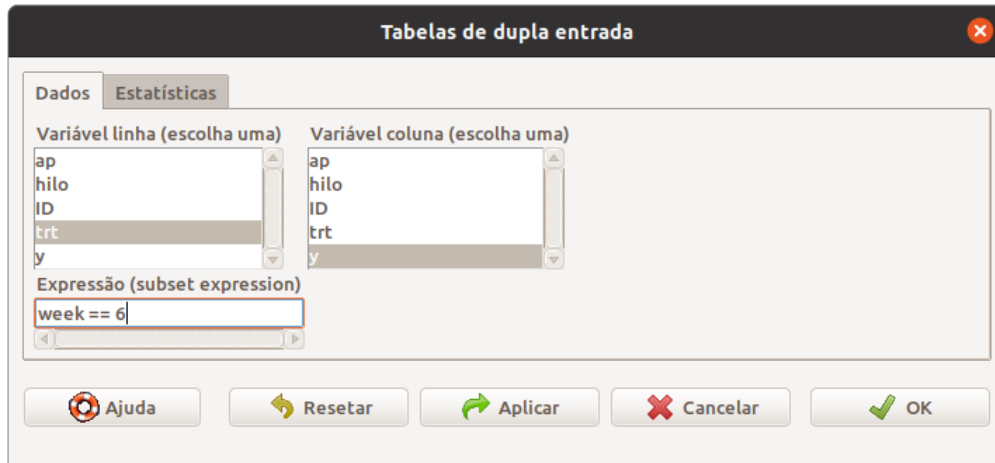


Figura 17.15: Selecionando as duas variáveis para o teste qui ao quadrado para uma tabela  $r \times c$  e os registros que serão incluídos na análise.

Em seguida, na aba *Estatísticas*, o usuário seleciona o modo como a tabela será apresentada e os testes estatísticos (qui ao quadrado e/ou teste exato de Fisher, figura 17.16).

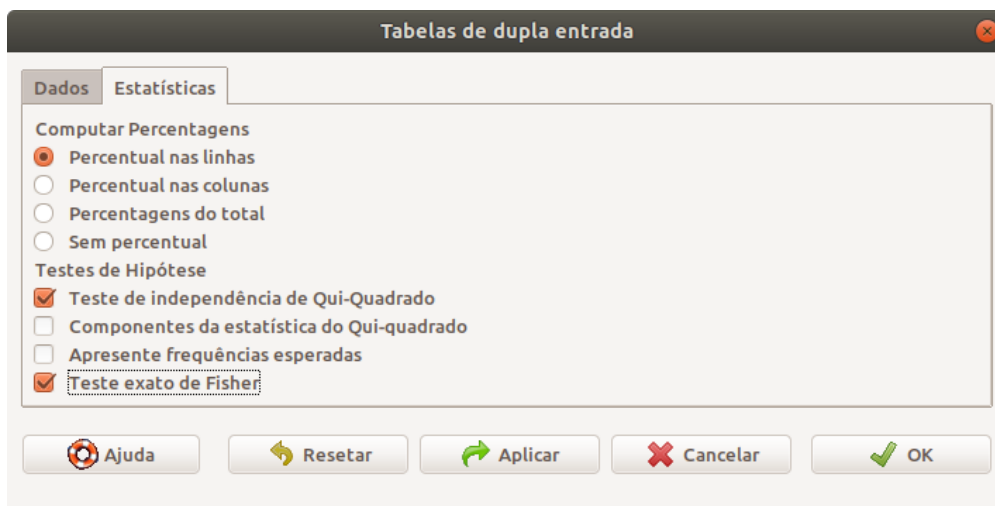


Figura 17.16: Selecionando os testes que serão realizados, bem como que percentuais serão mostrados a partir do teste qui ao quadrado para uma tabela  $r \times c$ .

Ao clicarmos no botão OK, os resultados são apresentados.

```
##
## Frequency table:
##           y
## trt        y  n
##  drug+      7  5
##   drug      6  5
##  placebo 16  1
##
## Row percentages:
##           y
## trt        y    n Total Count
##  drug+    58.3 41.7   100     12
##   drug    54.5 45.5   100     11
##  placebo  94.1  5.9   100     17
##
## Pearson's Chi-squared test
##
## data:  .Table
## X-squared = 6.9712, df = 2, p-value = 0.03064
##
##
## Fisher's Exact Test for Count Data
##
## data:  .Table
## p-value = 0.03146
## alternative hypothesis: two.sided
```

Foram identificadas bactérias em 94 % das crianças do grupo placebo, em apenas 54% das crianças no grupo medicamento e em 58% do grupo medicamento + estímulo para ingerir o medicamento.

O valor de p obtido por meio do teste qui ao quadrado é semelhante ao obtido pelo teste exato de Fisher. Usando o nível de significância de 5%, a hipótese nula de independência entre tratamento e presença de bactéria é rejeitada.

A análise deverá prosseguir para verificar os efeitos dos tratamentos e respectivos intervalos de confiança, mas também deverá considerar os outros instantes de tempo onde a presença de bactérias foram identificadas.

## 17.6 Exercícios

- 1) Com o conjunto de dados *birthwt* do pacote *MASS* ([GPL-2](#) | [GPL-3](#)), faça as atividades abaixo.
  - a) Obtenha as medidas de associação DAR, RR e RC e os respectivos intervalos de confiança ao nível de 95% para a relação entre as variáveis *smoke* (mãe fumante ou não durante a gravidez) e *low* (indicador de baixo peso ao nascer). Discuta os resultados em termos de força e precisão das medidas.
  - b) Repita o item “a” para as variáveis *ht* (histórico de hipertensão da mãe) e *low*.
- 2) Com o mesmo conjunto de dados da questão 1, verifique a associação entre as variáveis *race* (raça da mãe) e *low*.
- 3) O conjunto de dados *retinopathy* do pacote *survival* ([LGPL-2](#) | [LGPL-2.1](#) | [LGPL-3](#)) contém dados sobre um ensaio de coagulação a laser como tratamento para atrasar a retinopatia diabética. Cada paciente teve um dos olhos tratados com um dos dois tipos de laser: xenônio ou argônio. A variável *trt* indica qual dos olhos de cada paciente foi tratado (0 = olho controle, 1 = olho tratado), a variável *status* indica se houve perda de visão no olho correspondente até o final do acompanhamento no estudo (0 = censurado, 1 = perda de visão) e a variável *id* identifica cada paciente (dois registros por paciente, um para cada olho). Use a função *paired\_proportions* (seção 17.3.3) para verificar a associação entre o fato de um olho ser ou não tratado com ocorrência da perda de visão. A variável nas linhas é *trt*, a variável nas colunas é *status*, os níveis das variáveis *trt* e *status* são numéricos e o estudo não é um estudo de caso-controle (é um ensaio controlado randomizado). É preciso converter a variável *id* para fator. Obtenha as medidas de associação DAR, RR e RC e os respectivos intervalos de confiança ao nível de 90%. Discuta os resultados.
- 4) Considere o artigo “Estudo clínico, duplo-cego, randomizado, em crianças com amigdalites recorrentes submetidas a tratamento homeopático”, cujos resultados são apresentados na figura 17.17. Vamos analisar a associação entre os tipos de tratamento e a ocorrência de amigdalite, tomando o placebo como referência. Responda às questões abaixo.
  - a) Você considera a apresentação dos resultados satisfatória? Justifique.
  - b) Obtenha as medidas de associação DAR, RR e RC para esse estudo e os respectivos intervalos de confiança?
  - c) Qual é a interpretação para o intervalo de confiança da DAR?
  - d) Qual é o valor de p para essa associação. O que você pode dizer a respeito da significância estatística desse resultado?
  - e) O que você tem a comentar sobre a precisão da estimativa da diferença absoluta de riscos?
  - f) Obtenha o NNT para esse estudo.

Tabela 1. Evolução dos pacientes, com amigdalite recorrente; Teste exato de Fisher:  $p=0,015$  ou 1,5%\*

Grupo	Não houve amigdalite	Houve amigdalite	Total
Placebo	5	10	15
Medicação	14	4	18
Total	19	14	33

Figura 17.17: Tabela 1 do estudo de (Furuta et al., 2017) (CC BY).

- 5) Suponha que diversos ensaios controlados randomizados comparam o efeito de dois tratamentos A e B sobre um desfecho clínico dicotômico. Considere também que os estudos não apresentam nenhuma evidência de tendenciosidades e os tratamentos comparados não são os mesmos de estudo para estudo. Os resultados dos estudos são mostrados na figura 17.18 a seguir.

**Estudo 1:**

Tratamento	Desfecho		Total
	Sim	Não	
A	20.000	80.000	100.000
B	20.500	80.000	100.500
	40.500	160.000	200.500

**Resultados:**

RR = 0,8  
 IC95% = [0,96 - 0,999]  
 $p = 0,027$

**Estudo 3:**

Tratamento	Desfecho		Total
	Sim	Não	
A	60	240	300
B	120	180	300
	180	420	600

**Resultados:**

RR = 0,50  
 IC95% = [0,38 - 0,65]  
 $p = 0,0000001$

**Estudo 2:**

Tratamento	Desfecho		Total
	Sim	Não	
A	20.000	80.000	100.000
B	20.200	80.000	100.200
	40.200	160.000	200.200

**Resultados:**

RR = 0,99  
 IC95% = [0,97 - 1,01]  
 $p = 0,373$

**Estudo 4:**

Tratamento	Desfecho		Total
	Sim	Não	
A	5	20	25
B	10	15	25
	15	35	50

**Resultados:**

RR = 0,50  
 IC95% = [0,2 - 1,25]  
 $p = 0,373$

Figura 17.18: Diferentes resultados gerados a partir de tabelas 2x2.

Discuta cada resultado separadamente. A seguir, a partir de uma comparação dos resultados, comente a possível incidência de erro tipo I ou tipo II em cada estudo, as limitações de testes de hipótese e o valor de  $p$  e o efeito do tamanho amostral sobre os resultados.

# Capítulo 18

## Análise de variância

### 18.1 Introdução

Os conteúdos desta seção e da seção seguinte (seção 18.2) podem ser visualizados neste [vídeo](#).

Amess et al. (Amess et al., 1978) realizaram um estudo prospectivo, onde avaliaram os níveis de ácido fólico (microgramas por litro) nas células vermelhas em pacientes com *bypass* cardíaco que receberam três métodos diferentes de ventilação durante a anestesia.

Os dados desse estudo estão disponíveis no conjunto de dados *red.cell.folate* do pacote *ISwR* ([GPL-2](#) | [GPL-3](#)) (figura 18.1).

Red cell folate data	
Description	
The <code>folate</code> data frame has 22 rows and 2 columns. It contains data on red cell folate levels in patients receiving three different methods of ventilation during anesthesia.	
Usage	
<code>red.cell.folate</code>	
Format	
This data frame contains the following columns:	
folate	a numeric vector, folate concentration ( <i>microgram per liter</i> ).
ventilation	a factor with levels <code>N2O+O2,24h</code> : 50% nitrous oxide and 50% oxygen, continuously for 24 hours; <code>N2O+O2,op</code> : 50% nitrous oxide and 50% oxygen, only during operation; <code>O2,24h</code> : no nitrous oxide but 35%–50% oxygen for 24 hours.
Source	
D.G. Altman (1991), <i>Practical Statistics for Medical Research</i> , Table 9.10, Chapman & Hall.	

Figura 18.1: Descrição das variáveis que compõem o conjunto de dados *red.cell.folate*.

Os três métodos de ventilação são:

- 1) N<sub>2</sub>O-O<sub>2</sub>, 24h: 50% óxido nitroso e 50% oxigênio continuamente por 24 horas;
- 2) N<sub>2</sub>O-O<sub>2</sub>, op: 50% óxido nitroso e 50% oxigênio somente durante a operação;
- 3) O<sub>2</sub>, op: 35-50% oxigênio continuamente por 24 horas.

O conteúdo do conjunto de dados *red.cell.folate* é mostrado a seguir:

```
library(RcmdrMisc)
library(ISwR)
data(red.cell.folate, package="ISwR")
red.cell.folate
```

```
##      folate ventilation
## 1      243   N20+O2,24h
## 2      251   N20+O2,24h
## 3      275   N20+O2,24h
## 4      291   N20+O2,24h
## 5      347   N20+O2,24h
## 6      354   N20+O2,24h
## 7      380   N20+O2,24h
## 8      392   N20+O2,24h
## 9      206    N20+O2,op
## 10     210    N20+O2,op
## 11     226    N20+O2,op
## 12     249    N20+O2,op
## 13     255    N20+O2,op
## 14     273    N20+O2,op
## 15     285    N20+O2,op
## 16     295    N20+O2,op
## 17     309    N20+O2,op
## 18     241       O2,24h
## 19     258       O2,24h
## 20     270       O2,24h
## 21     293       O2,24h
## 22     328       O2,24h
```

As medidas de tendência central e dispersão da variável *folate* para cada método de ventilação são apresentadas a seguir e a figura 18.2 mostra o correspondente diagrama de pontos.

```
numSummary(red.cell.folate[, "folate", drop=FALSE],
            groups=red.cell.folate$ventilation,
            statistics=c("mean", "sd", "IQR", "quantiles"),
            quantiles=c(0, .25, .5, .75, 1))
```



##		mean	sd	IQR	0%	25%	50%	75%	100%	folate:n
##	N2O+O2,24h	316.6250	58.71709	91.5	243	269	319	360.5	392	8
##	N2O+O2,op	256.4444	37.12180	59.0	206	226	255	285.0	309	9
##	O2,24h	278.0000	33.75648	35.0	241	258	270	293.0	328	5

```
with(red.cell.folate, Dotplot(folate, by=ventilation, bin=FALSE))
```

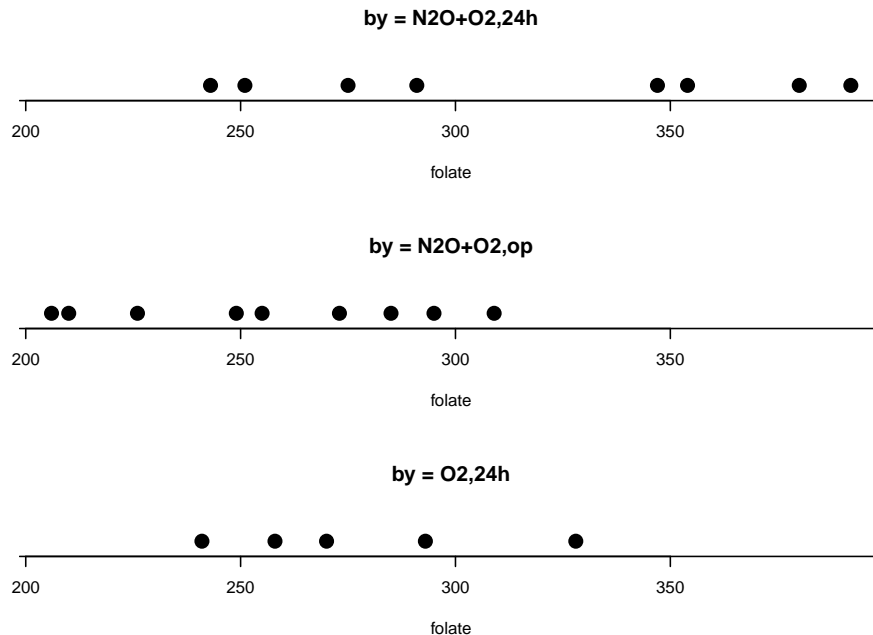


Figura 18.2: Diagrama de pontos da variável ácido fólico para cada método de ventilação.

O capítulo 16 explorou técnicas estatísticas para verificar se uma variável numérica possui a mesma distribuição em duas populações independentes ou não, assim como a determinação do intervalo de confiança para a diferença de médias entre as duas populações.

Generalizando o problema, poderíamos estar interessados em saber se a variável *folate* estaria distribuída da mesma forma em cada método de ventilação ou quais seriam as diferenças de médias de ácido fólico entre os diversos métodos de ventilação. Neste capítulo, será introduzida a técnica estatística conhecida como **Análise de Variância** (ANOVA), utilizada para a comparação das médias de uma variável aleatória numérica em mais de duas populações.

A variável numérica cujas médias estão sendo comparadas em três ou mais populações é chamada de variável **resposta** ou **dependente**. A variável categórica que distingue as diversas populações em estudo é chamada de **variável de tratamento**, ou **fator**, ou **variável explanatória**, ou **variável independente**. As perguntas que devem ser respondidas são:

- 1) Os diferentes níveis do fator em estudo resultam em diferenças de médias na variável resposta?
- 2) Se sim, como quantificar essas diferenças?

Esse é o tipo mais simples de análise de variância e é conhecida como análise de variância com um fator, onde uma única fonte de variação é avaliada em mais de duas amostras independentes, sendo uma extensão do teste t para duas amostras independentes.

A análise de variância com um fator é utilizada para testar a hipótese nula de que as distribuições da variável resposta em cada nível do fator em estudo é a mesma e também para obter intervalos de confiança para contrastes entre as médias da variável resposta nos diversos níveis do fator. No exemplo considerado no início deste capítulo, a ANOVA pode ser utilizada para verificar se as médias do ácido fólico são as mesmas em cada método de ventilação e, eventualmente, estabelecer intervalos de confiança para diferenças das médias de ácido fólico entre os diversos métodos.

## 18.2 Múltiplas comparações

Uma primeira ideia para comparar as médias de ácido fólico entre os três métodos de ventilação no conjunto de dados *red.cell.folate* seria realizar um teste t com um nível  $\alpha$  de significância para cada par de métodos de ventilação e considerar estatisticamente significativas aquelas comparações cujo valor de p fosse menor que  $\alpha$ . Como temos 3 métodos de ventilação, esse raciocínio é equivalente a realizar 3 testes estatísticos, comparando dois a dois todos os pares de métodos. Em cada teste, a probabilidade de não rejeitar a hipótese nula quando ela é verdadeira é  $1 - \alpha$ .

Supondo que os 3 testes sejam independentes e que a hipótese nula seja verdadeira em todos eles, a probabilidade de todos os testes não serem estatisticamente significativos é de  $(1 - \alpha)^3$ . Logo a probabilidade de pelo menos um deles ser estatisticamente significativo é de  $1 - (1 - \alpha)^3$ . Se tomarmos  $\alpha = 5\%$ , por exemplo, o erro tipo I para o conjunto de testes não será 5%, mas  $1 - (1 - 0,05)^3 = 0,14$  (14%).

Esse é o problema quando múltiplos testes são realizados, especialmente quando são realizados de maneira exploratória, sem terem sido previamente planejados. Nesse exemplo, o mais adequado é realizar um único teste para verificar se existe alguma diferença entre os métodos de ventilação tomados em conjunto. Caso o teste indique que as médias sejam diferentes, então uma análise mais detalhada seria realizada para determinar onde estão estas diferenças e calcular os intervalos de confiança de tal forma que, em conjunto, tenham o nível de confiança desejado. Esses são os objetivos da *Análise de Variância*.

## 18.3 Análise de variância com um fator

### 18.3.1 Modelo de efeitos fixos

Os conteúdos desta seção e da seção 18.3.2 podem ser visualizados neste [vídeo](#).

Na análise de variância com um fator, amostras são extraídas das populações correspondentes a cada nível do fator em estudo. A rigor, essas amostras deveriam ser aleatórias para justificar a aplicação das técnicas estatísticas descritas a seguir. Entretanto, em muitos estudos, essas

amostras são obtidas por conveniência, na suposição de que de que não haja vieses na seleção dos pacientes.

Vamos supor que temos  $k$  níveis do fator e que  $n_i$  representa o tamanho da amostra correspondente ao nível  $i$ ,  $i$  variando de 1 a  $k$ . O modelo para a análise de variância com um fator pode ser escrito como:

$$X_{ij} = \mu_i + \epsilon_{ij} = \mu + \alpha_i + \epsilon_{ij} \quad (18.1)$$

onde:

$i$  - indica cada nível do fator em estudo ( $i = 1, 2, \dots, k$ ).

$j$  - indica cada uma das observações da variável resposta em cada amostra ( $j = 1, 2, \dots, n_i$ ).

$X_{ij}$  - valor da variável resposta correspondente à observação  $j$  da amostra  $i$ .

$\mu_i$  - média da variável  $X$  na população correspondente ao nível  $i$  do fator de estudo. Representa o efeito do nível  $i$  do fator em estudo.

$\epsilon_{ij}$  - erro associado à observação  $j$  da amostra  $i$ , diferença entre o valor de  $X_{ij}$  e a média da população  $i$ .

$\mu$  - média geral da variável  $X$  em todas as populações no estudo.

$\alpha_i$  - diferença entre a média da variável  $X$  na população correspondente ao nível  $i$  do fator de estudo e a média geral,  $\alpha_i = \mu_i - \mu$ .

Ao olharmos para esse modelo, podemos ver que uma observação típica do conjunto total de dados em estudo é composta de: 1) média geral de todos os tratamentos ( $\mu$ ); 2) um efeito diferencial de cada tratamento ( $\alpha_i$ ); e 3) um termo de erro ( $\epsilon_{ij}$ ), representando o desvio da observação em relação à média de seu grupo.

Esse modelo é chamado de modelo de efeitos fixos, porque estamos interessados somente nesses níveis do fator em estudo. No caso da relação entre os métodos de ventilação e os níveis de ácido fólico, esse modelo supõe que os autores não estão interessados em nenhum outro método de ventilação além dos três utilizados. Em outros tipos de estudos, os níveis do fator estudado poderiam ser um subconjunto de todos os níveis possíveis. Nesses casos, o modelo seria chamado de modelo de efeitos aleatórios.

As suposições do modelo de efeitos fixos são as seguintes:

- as  $k$  amostras são amostras independentes das respectivas populações;
- cada uma das populações de onde as amostras foram extraídas é normalmente distribuída com média  $\mu_i$  e variância  $\sigma^2$ ;
- as populações possuem a mesma variância, isto é,  $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 = \sigma^2$ ;
- as médias  $\mu_i$  são constantes desconhecidas e  $\sum_{i=1}^k \alpha_i = 0$ , uma vez que a soma dos desvios de  $\mu_i$  em relação à média geral é 0;

- os  $\epsilon_{ij}$  são normalmente e independentemente distribuídos com mesma variância, ou seja,  $\epsilon_{ij} \sim N(0, \sigma^2)$ .

### 18.3.2 Teste de hipótese

Na análise de variância de modelos de efeitos fixos com um fator, a hipótese nula para um teste bilateral é que as médias das populações correspondentes a cada nível do fator são iguais e a hipótese alternativa é que pelo menos uma das médias é diferente das demais:

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

$H_1$ : nem todas as médias  $\mu_i$  são iguais

Se as médias das populações são iguais, então o efeito diferencial de cada nível do fator é igual a 0. Então a hipótese nula e a alternativa poderiam ser escritas como:

$$H_0: \alpha_i = 0, \quad i = 1, 2, \dots, k$$

$H_1$ : nem todos os  $\alpha_i = 0$

Se  $H_0$  for verdadeira e as suposições de igualdade de variâncias e distribuição normal das populações forem satisfeitas, as funções densidade de probabilidade das populações se parecerão com a figura 18.3. Nesse caso, as médias e as variâncias das populações são todas iguais e as respectivas distribuições de probabilidades serão superpostas.

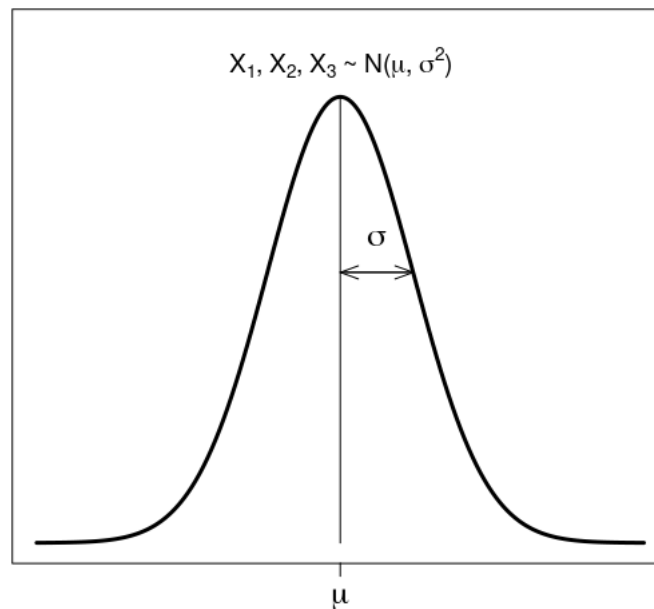


Figura 18.3: Situação em que as médias das populações correspondentes a cada nível do fator em estudo são iguais.

Quando  $H_0$  for falsa, ela pode ser falsa, porque uma das populações possui a média diferente das demais, que são iguais, ou todas as populações possuem médias diferentes; há diversas

outras possibilidades. A figura 18.4 mostra a situação em que as suposições são satisfeitas, mas  $H_0$  é falsa, porque todas as médias são diferentes, considerando que um fator com três níveis esteja sendo estudado.

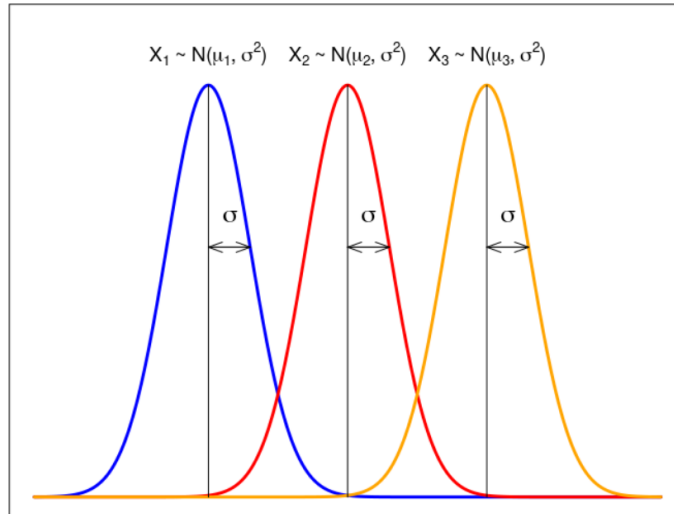


Figura 18.4: Situação em que as médias das populações correspondentes a cada nível do fator em estudo são diferentes para um fator com três níveis.

Para verificar a hipótese  $H_0$ , vamos considerar a situação em que uma variável aleatória  $X$  foi medida em  $k$  populações diferentes, definida por  $k$  níveis de uma variável categórica (fator). Para cada nível do fator, uma amostra aleatória com  $n_i$  elementos foi extraída. O número total de valores medidos da variável  $X$ ,  $n_T$  é igual a:

$$n_T = \sum_{i=1}^k n_i \quad (18.2)$$

A média aritmética de todos os valores de  $X$  (média geral) é dada por:

$$\bar{X} = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} X_{ij}}{n_T} \quad (18.3)$$

A média aritmética dos valores de  $X$  para cada grupo é dada por:

$$\bar{X}_i = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i} \quad (18.4)$$

onde  $i = 1, 2, \dots, k$ .

## Partição da soma dos resíduos em relação à média geral

A expressão a seguir mostra que o desvio de cada valor da variável aleatória em relação à média geral é igual ao desvio em relação à média do respectivo grupo somado ao desvio da média do respectivo grupo em relação à média geral:

$$\underbrace{X_{ij} - \bar{X}}_{\text{desvio em relação à média geral}} = \underbrace{(X_{ij} - \bar{X}_i)}_{\text{desvio em relação à média do grupo}} + \underbrace{(\bar{X}_i - \bar{X})}_{\text{desvio da média do grupo em relação à média geral}}$$

Elevando cada desvio ao quadrado e somando os quadrados de todos os desvios, pode-se mostrar que o resultado é a expressão abaixo:

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{X}_i - \bar{X})^2$$

Essa expressão pode ser escrita como:

$$SQTot = SQE + SQEG$$

onde:

$$SQTot = \text{Soma total dos quadrados} = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2$$

$$SQE = \text{Soma dos quadrados dos desvios dentro de cada grupo} = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$$

$$\begin{aligned} SQEG &= \text{Soma dos quadrados dos desvios dos grupos em relação à média} \\ &= \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{X}_i - \bar{X})^2 \\ &= \sum_{i=1}^k n_i (\bar{X}_i - \bar{X})^2 \end{aligned}$$

## Partição dos graus de liberdade

Correspondendo à partição da soma total dos quadrados dos desvios (SQTot), há uma partição dos respectivos graus de liberdade.

A SQTot possui  $n_T - 1$  graus de liberdade associado a ela. Há  $n_T$  desvios  $X_{ij} - \bar{X}$ , mas um grau de liberdade é perdido, porque esses desvios não são independentes, já que a soma de todos eles é 0.

A SQE possui  $n_T - k$  graus de liberdade associado a ela. Em cada amostra, há  $n_i$  desvios, mas um grau de liberdade é perdido, porque esses desvios não são independentes, já que a soma dos desvios em relação média de cada amostra é nula. Então o número de graus de liberdade é igual a  $\sum_{i=1}^k (n_i - 1) = n_T - k$ .

A SQEG possui  $k - 1$  graus de liberdade associado a ela. Há  $k$  desvios das médias de cada amostra em relação à média geral, mas um grau de liberdade é perdido, porque esses desvios não são independentes, já que a soma dos desvios das médias de cada grupo em relação à média geral é igual a 0.

O **erro quadrático médio residual** (EQMR) é obtido dividindo-se a soma dos quadrados dos resíduos (SQE) por  $n_T - k$ :

$$EQMR = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2}{n_T - k} \quad (18.5)$$

O **erro quadrático médio entre grupos** (EQMEG) é obtido dividindo-se a soma dos quadrados entre grupos por  $k - 1$ :

$$EQMEG = \frac{\sum_{i=1}^k n_i (\bar{X}_i - \bar{X})^2}{k - 1} \quad (18.6)$$

Pode-se mostrar que o valor esperado do erro quadrático médio residual é igual à variância das populações:

$$E[EQMR] = \sigma^2 \quad (18.7)$$

e que o valor esperado do erro quadrático médio entre grupos é igual à expressão abaixo:

$$E[EQMEG] = \sigma^2 + \frac{\sum_{i=1}^k n_i (\mu_i - \mu)^2}{k - 1} \quad (18.8)$$

## Teste F

Quando a hipótese nula é verdadeira, tanto o erro quadrático médio residual quanto o erro quadrático médio entre grupos são estimadores não tendenciosos da variância das populações. Quando a hipótese nula não é verdadeira, o valor esperado do erro quadrático médio entre grupos é maior do que a variância das populações, aumentando à medida que as diferenças entre as médias das populações aumenta.

Assim a divisão do valor do erro quadrático médio entre grupos (EQMEG) pelo erro quadrático médio residual (EQMR) dá uma indicação de quanto a hipótese nula é compatível com os dados. O valor dessa divisão é representado por  $F^*$ :

$$F^* = \frac{EQMEG}{EQMR} \quad (18.9)$$

Pode-se mostrar que a razão  $\frac{EQMEG}{EQMR}$ , se a hipótese nula é verdadeira, segue a distribuição de Fisher,  $F(k - 1, n_T - k)$ . A distribuição  $F(\nu_1, \nu_2)$  possui dois parâmetros onde  $\nu_1$  é o número de graus de liberdade do numerador e  $\nu_2$  é o número de graus de liberdade do denominador. Dado um nível de significância  $\alpha$ , quando o valor de F, obtido da expressão (18.9) for maior que o quantil  $1 - \alpha$  da distribuição  $F(k - 1, n_T - k)$ , então a hipótese nula é rejeitada.

Esse método é conhecido por análise de variância, porque é realizada a comparação das variâncias cujas estimativas são obtidas pelas expressões (18.5) e (18.6) para verificar se as médias das populações correspondentes a cada nível do fator em estudo são diferentes ou não.

A aplicação [Análise de Variância](#) (figura 18.5) ilustra a ideia básica da análise de variância. O painel à esquerda permite ao usuário configurar a média e o tamanho das amostras de cada nível do fator em estudo (3 níveis na aplicação), o desvio padrão comum a cada distribuição dos valores da variável aleatória e o nível de significância do teste estatístico. O painel principal mostra, no primeiro gráfico, a média em cada grupo (indicados pelas cores preta, vermelha e verde) e as distâncias entre cada valor da variável aleatória e a correspondente média de cada grupo. O gráfico superior à direita mostra para cada valor da variável aleatória para um dado grupo a distância entre a média do grupo e a média geral. O gráfico na parte inferior da figura mostra a distância de cada valor da variável aleatória e a média geral. Finalmente a tabela na área inferior à direita mostra os cálculos para a análise de variância.

#### Análise de Variância

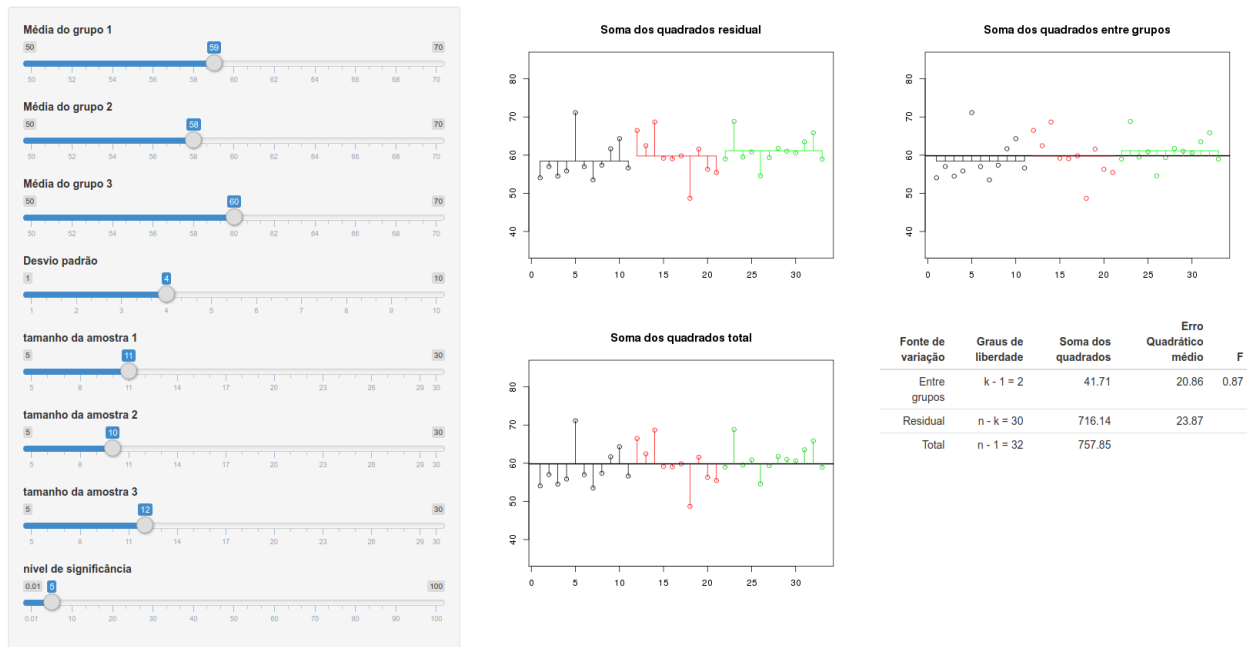


Figura 18.5: Aplicação que ilustra o princípio da ANOVA.



Quando as médias amostrais dos grupos são próximas (figura 18.6: médias amostrais iguais a 59, 58 e 60 em cada grupo, respectivamente) e com o desvio padrão igual a 4, por exemplo, as estimativas da variância por meio da soma dos quadrados dos resíduos e pela soma dos quadrados entre grupos (com tamanhos de amostras iguais a 11, 12 e 11, respectivamente) não são muito diferentes. O valor de F não é estatisticamente significativo.

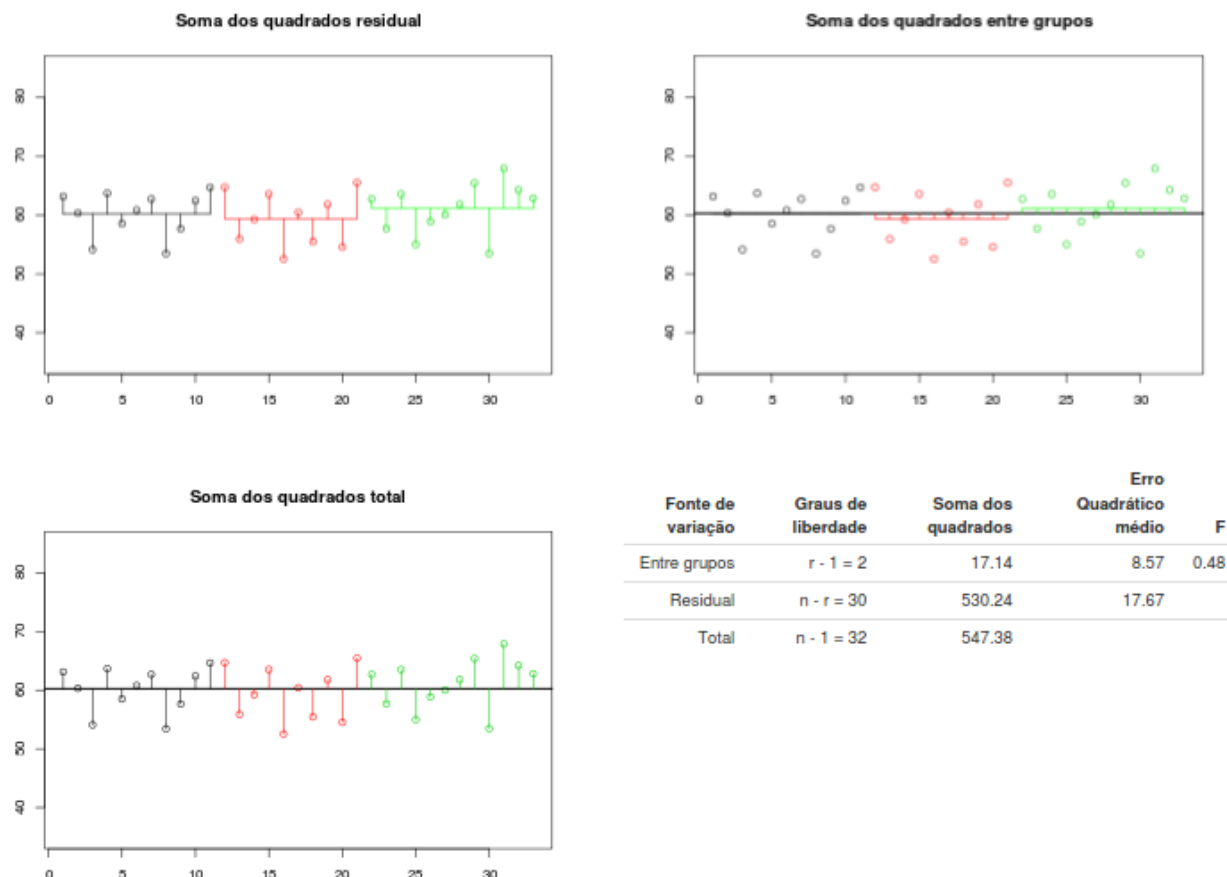


Figura 18.6: Variâncias obtidas na aplicação da figura 18.5 quando as médias são próximas.

Quando as médias amostrais dos grupos não são próximas (figura 18.7: médias amostrais iguais a 69, 50 e 60 em cada grupo, respectivamente) e com o mesmo desvio padrão (4), as estimativas da variância por meio da soma dos quadrados dos resíduos e pela soma dos quadrados entre grupos (com tamanhos de amostras iguais a 11, 10 e 12, respectivamente) são bem diferentes. O valor de F é estatisticamente significativo, indicado por um asterisco após a letra F no gráfico.

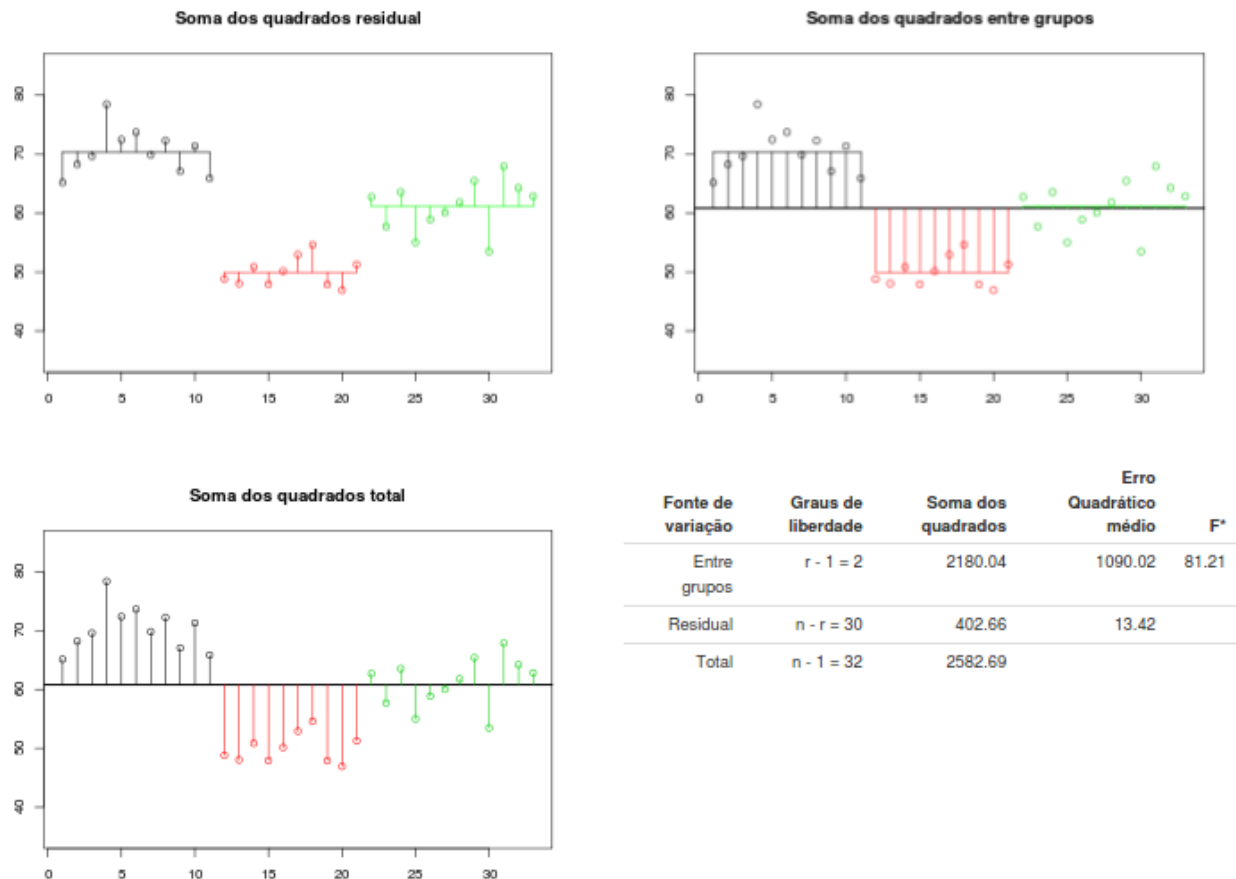


Figura 18.7: Variâncias obtidas na aplicação da figura 18.5 quando as médias são significativamente diferentes.

### 18.3.3 Comparação de médias

Os conteúdos desta seção e das subseções 18.3.3.1 e 18.3.3.2 podem ser visualizados neste [vídeo](#).

Quando uma análise de variância rejeita a hipótese nula de igualdade de médias, ou mesmo na ausência de um teste de hipótese, em geral há interesse em obter intervalos de confiança para a diferença de médias entre os diversos grupos que compõem o estudo.

Na seção 18.2, vimos que realizar comparações de médias duas a duas por meio do teste  $t$  acaba gerando um nível de significância bem acima do nível de significância de uma única comparação. O mesmo ocorre para múltiplos intervalos de confiança, cada um deles calculado com um nível de confiança  $1 - \alpha$ : a cobertura de todos os intervalos de confiança construídos é menor do que  $1 - \alpha$ .

Há diversos procedimentos propostos na literatura estatística para lidar com esse problema, conhecido como **múltiplas comparações**. Alguns métodos frequentemente citados são:

*Bonferroni, Scheffé, Tukey e Newman-Keuls.* Nesta seção, apresentaremos os métodos de Bonferroni, Tukey e Scheffé.

Vamos supor que desejamos obter o intervalo de confiança para a diferença das médias dos grupos  $i$  e  $j$ ,  $\mu_i - \mu_j$ , com amostras de tamanho  $n_i$  e  $n_j$  respectivamente. A estimativa para a diferença das médias é dada por:

$$\bar{D}_{ij} = \bar{X}_i - \bar{X}_j,$$

e a variância para a diferença das médias, supondo que as distribuições das variáveis  $X_i$  e  $X_j$  sejam normais, ou que as amostras sejam suficientemente grandes, será dada por:

$$var(\bar{D}_{ij}) = var(\bar{X}_i) + var(\bar{X}_j) = \sigma^2 \left( \frac{1}{n_i} + \frac{1}{n_j} \right) \quad (18.10)$$

Logo o erro padrão para as diferenças das duas médias é dado por:

$$EP(\bar{D}_{ij}) = \sigma \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \quad (18.11)$$

e é estimado por:

$$EP(\bar{D}_{ij}) = \sqrt{EQMR \left( \frac{1}{n_i} + \frac{1}{n_j} \right)} \quad (18.12)$$

O intervalo de confiança ao nível de  $(1 - \alpha)\%$  para a diferença das médias  $\mu_i - \mu_j$  é então dado por:

$$\bar{D}_{ij} - t_{n_i+n_j-2, 1-\alpha/2} EP(\bar{D}_{ij}) \leq (\mu_i - \mu_j) \leq \bar{D}_{ij} + t_{n_i+n_j-2, 1-\alpha/2} EP(\bar{D}_{ij}) \quad (18.13)$$

Os diversos procedimentos para lidar com múltiplas comparações de médias substituem o elemento  $t_{n_i+n_j-2, 1-\alpha/2}$  na expressão acima por outro valor de modo a que o conjunto de comparações realizadas tenha pelo menos um grau de confiança  $1 - \alpha \%$ . Os intervalos de confiança serão mais largos do que os obtidos com uma só comparação.

Esses métodos são particularmente importantes quando o interesse reside em realizar comparações que foram pensadas após uma inspeção dos dados e que não foram planejadas a priori, antes da realização do estudo. Em investigações exploratórias, muitas questões novas podem ser sugeridas quando os dados são analisados.

### 18.3.3.1 Método de Bonferroni

O método de Bonferroni é o mais simples de todos: no termo  $t_{n_i+n_j-2, 1-\alpha/2}$ , divide-se o valor de  $\alpha/2$  pelo número de comparações realizadas. Assim, se o número de grupos for 3 e se deseja o intervalo de confiança para a diferença de todas as médias entre si, então substitui-se  $t_{n_i+n_j-2, 1-\alpha/2}$  na expressão (18.13) por  $t_{n_i+n_j-2, 1-\alpha/6}$ .

À medida que o número de comparações aumenta, os intervalos de confiança ficam excessivamente conservadores no sentido de que o nível de confiança é na verdade maior do que  $(1 - \alpha)\%$ .

### 18.3.3.2 Método de Tukey

O método de Tukey pode ser utilizado quando se deseja comparar as médias de todos os pares de níveis do fator em estudo, ou seja, quando se deseja estimar  $\bar{D}_{ij} = \mu_i - \mu_j$  para todos os pares  $i, j$ .

Quando os tamanhos das amostras são iguais, o nível de confiança para toda a família de comparações é exatamente  $(1 - \alpha)\%$ . Quando as amostras não são iguais, o nível de confiança da família de intervalos é maior que  $(1 - \alpha)\%$ .

O método de Tukey é baseado na distribuição da estatística conhecida como *amplitude studentizada* (*studentized range*), que é definida pela expressão:

$$q(k, \nu) = \frac{\max(X_i) - \min(X_i)}{s} \quad (18.14)$$

onde  $X_i$ ,  $i = 1, \dots, k$  corresponde a  $k$  observações independentes de uma distribuição normal com média  $\mu$  e variância  $\sigma^2$ . Então o numerador da expressão acima corresponde à amplitude das  $k$  observações e  $s^2$  corresponde à estimativa da variância com  $\nu$  graus de liberdade.

Para a ANOVA com um fator, os intervalos de confiança para as diferenças entre pares de médias, de acordo com o método de Tukey, são obtidos substituindo-se  $t_{n_i+n_j-2, 1-\alpha/2}$  na expressão (18.13) por:

$$T = \frac{q(1 - \alpha, k, n_T - k)}{\sqrt{2}} \quad (18.15)$$

Os valores de  $T$  são então obtidos a partir do quantil  $1 - \alpha$  da distribuição de  $q$  com parâmetros iguais a  $k$  (número de níveis do fator – grupos) e  $n_T - k$ , respectivamente ( $n_T$  é o número total de observações).

### 18.3.3.3 Método de Scheffé

O método de Scheffé se aplica quando se deseja estimativas de intervalos de confiança para o conjunto de todos os contrastes possíveis entre as médias dos níveis do fator em estudo.

Um contraste é uma expressão da forma:

$$C = \sum c_k \mu_k, \text{ onde } \sum c_k = 0.$$

Um exemplo de um contraste para três níveis de um fator seria:  $\frac{\mu_1 + \mu_2}{2} - \mu_3$

Uma estimativa de um contraste é dada por:

$$\bar{C} = \sum c_k \bar{X}_k \quad (18.16)$$

Uma estimativa da variância de um contraste é dada por:

$$var(\bar{C}) = EQMR \sum \frac{c_k^2}{n_k} \quad (18.17)$$

Pode-se mostrar que a probabilidade é  $(1 - \alpha)\%$  de que todos os intervalos do tipo:

$$\bar{C} \pm S \, var(\bar{C})$$

estejam corretos simultaneamente, onde  $\bar{C}$  e  $var(\bar{C})$  são calculados por (18.16) e (18.17), respectivamente, e S é dado por:

$$S^2 = (k - 1)F(1 - \alpha, k - 1; n_T - 1)$$

O método de Scheffé é mais flexível do que os demais, porque permite obter intervalos de confiança com o nível de confiança  $(1 - \alpha)\%$ , quando considerados simultaneamente, para um conjunto muito mais amplo de comparações. Por essa razão, as amplitudes dos intervalos de confiança, em geral, são maiores do que intervalos de confiança obtidos por outros métodos. Quando se deseja realizar poucas comparações, possivelmente os métodos de Bonferroni, Tukey e outros podem fornecer intervalos mais precisos.

### 18.3.4 Análise de resíduos

Os conteúdos desta seção, de suas subseções e da seção 18.3.5 podem ser visualizados neste [vídeo](#).

A análise de variância apresentada nas seções anteriores parte de três suposições básicas:

- 1) as variâncias da variável resposta são as mesmas em todos os grupos;
- 2) os erros são distribuídos normalmente;
- 3) os erros são independentes.

Os desvios, ou resíduos, das observações em relação à média em cada grupo ou nível do fator em estudo

$$r_{ij} = X_{ij} - \bar{X}_i, \quad i = 1, 2, \dots, n_i \quad (18.18)$$

podem ser usados para checar se as suposições acima são satisfeitas. Essa análise é conhecida como **análise de resíduos**.

#### 18.3.4.1 Comparação das variâncias

A igualdade das variâncias da variável resposta em cada nível do fator em estudo pode ser investigada qualitativamente, por exemplo, por um diagrama de pontos que mostram os resíduos para cada nível do fator. A figura 18.8 mostra dois diagramas de resíduos para cada um dos quatro níveis de um fator. No diagrama a, é possível observar que a variabilidade dos dados é diferente nos diversos níveis do fator, enquanto que no diagrama b não é possível identificar claramente uma divergência nos valores das variâncias em cada grupo.

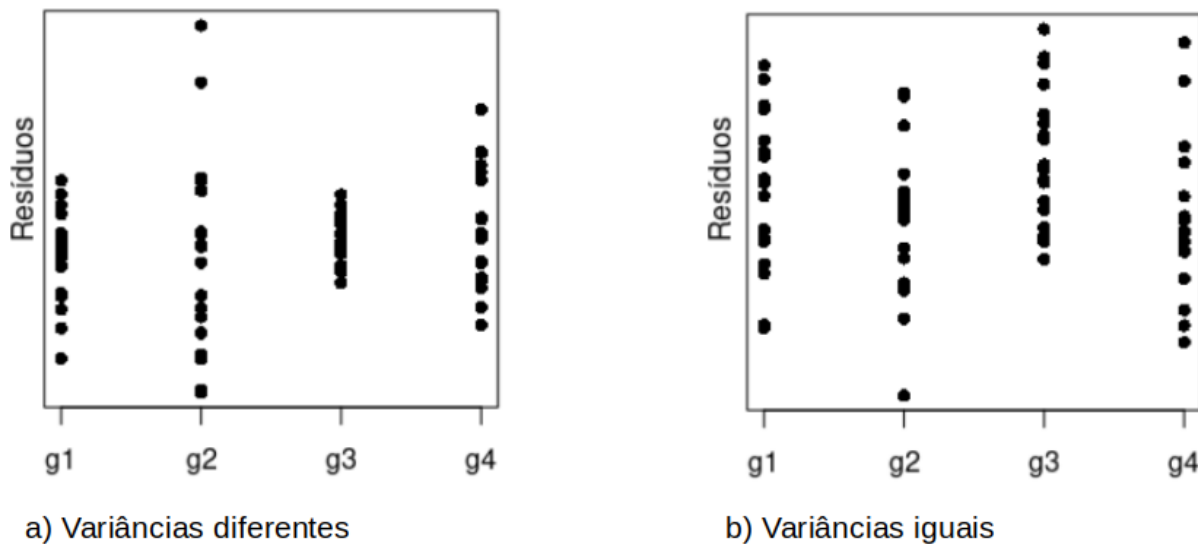


Figura 18.8: Variância de uma variável em quatro grupos diferentes e duas situações: a) variâncias diferentes e b) variâncias iguais.

A igualdade de variâncias também pode ser testada por meio de diversos testes estatísticos como Bartlett, Levene, etc. Não vamos entrar em detalhes desses testes neste texto, mas vamos ver como realizá-los no R na seção 18.3.6.

#### 18.3.4.2 Avaliação da normalidade dos dados

A avaliação da normalidade dos dados é realizada, por exemplo, plotando o gráfico de comparação dos quantis da normal para os resíduos obtidos a partir da expressão (18.18) (capítulo 16, seção 16.2.3), e/ou realizando um teste de normalidade dos resíduos (capítulo 16, seção 16.2.4).

Deve-se levar em conta que a ANOVA é robusta para desvios não muito acentuados da normalidade e da igualdade de variâncias.

Mais adiante, na seção 18.3.6.1, será mostrado como avaliar a normalidade dos resíduos de um modelo de ANOVA com um fator.

#### 18.3.4.3 Independência dos erros

Em relação ao terceiro item, independência dos resíduos, quando os dados são obtidos em uma sequência temporal ou quando os dados estão ordenados em uma outra lógica sequencial, como, por exemplo, uma sequência espacial, deve-se verificar se os resíduos não estão relacionados de alguma forma. Um diagrama dos resíduos versus ordenação dos mesmos para cada nível do fator em estudo é útil para identificar algum padrão de relacionamento. A figura 18.9 mostra três possíveis diagramas que relacionam os resíduos com uma possível sequência lógica em que as observações foram medidas, sendo que os dois primeiros diagramas mostram claramente um padrão linear e não linear entre os resíduos e a ordem em que foram obtidos, respectivamente. O diagrama c mostra o caso em que os resíduos não são relacionados uns com os outros.

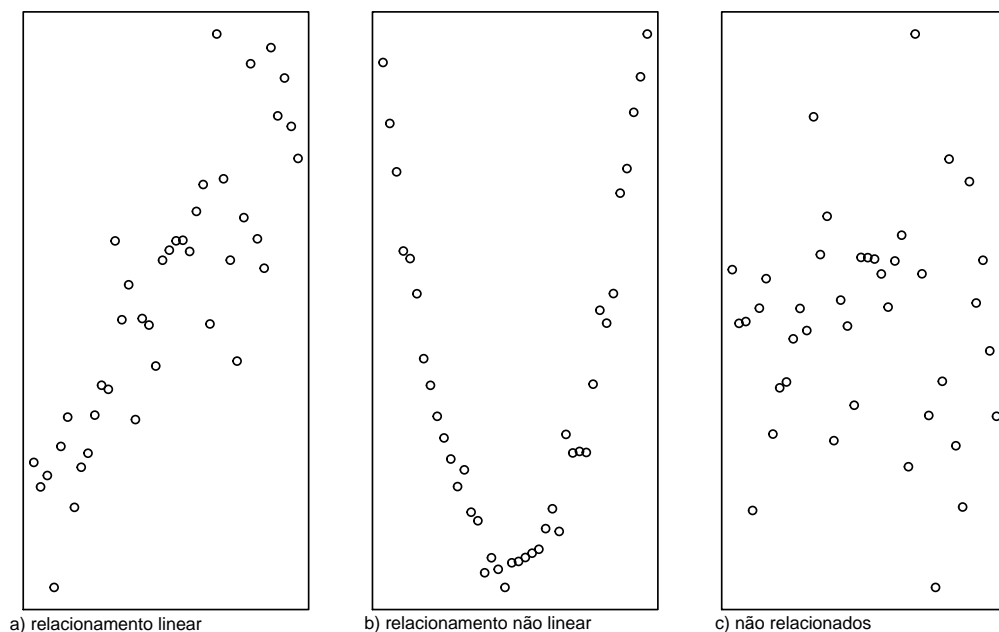


Figura 18.9: Diagramas que relacionam os resíduos com uma possível sequência lógica em que as observações foram medidas. O diagrama c mostra o caso em que os resíduos não são relacionados uns com os outros, enquanto que os diagramas a e b mostram relacionamentos linear e não linear, respectivamente..

### 18.3.5 Teste não paramétrico de Kruskal-Wallis

Se o modelo da análise de variância com um fator apresenta grandes desvios da normalidade ou da igualdade das variâncias dos grupos, um teste não paramétrico pode ser empregado para testar a hipótese nula de igualdade de médias entre os diversos grupos de tratamento. Um teste bastante utilizado é o teste de Kruskal-Wallis, que parte das seguintes suposições:

- 1) a distribuição da variável resposta em cada um dos grupos são contínuas e com a mesma forma, ou seja, possuem a mesma variabilidade, simetria, etc., mas podem diferir na localização da média;
- 2) as amostras dos diferentes grupos são aleatórias e independentes.

O teste de Kruskal-Wallis parte de todas as  $n_T$  observações, que são então ordenadas de 1 a  $n_T$ . Seja  $\bar{R}_i$  a média dos postos para o nível  $i$  do fator em estudo. A estatística para o teste de Kruskal-Wallis é então:

$$\chi_{kw}^2 = \left( \frac{12}{n_T(n_T + 1)} \sum_{i=1}^k n_i \bar{R}_i^2 \right) - 3(n_T + 1) \quad (18.19)$$

Se os  $n_i$  forem razoavelmente grandes (5 ou mais é usualmente recomendado), a estatística (18.19) é uma variável aleatória com aproximadamente uma distribuição  $\chi^2$  com  $k-1$  graus de liberdade quando a hipótese nula de que todas as médias são iguais é verdadeira. Valores suficientemente grandes de  $\chi_{kw}^2$  levam à rejeição de  $H_0$ , ou seja, se  $\chi_{kw}^2 > \chi_{1-\alpha, k-1}^2$ , então  $H_0$  é rejeitada.

O teste de Kruskal-Wallis é mais poderoso do que a análise de variância paramétrica, quando os desvios da normalidade são grandes, mas possui um menor poder estatístico quando as condições para a aplicação do modelo de análise de variância forem satisfeitas.

Múltiplas comparações dos postos pós-teste podem ser realizadas, com diversos tipos de correções, analogamente ao usado para a ANOVA.

### 18.3.6 Análise de variância com um fator no R

Os conteúdos desta seção e de suas subseções podem ser visualizados neste [vídeo](#).

Vamos realizar uma análise de variância para o problema colocado na introdução deste capítulo. Queremos saber se as médias de ácido fólico para os três métodos de ventilação no conjunto de dados *red.cell.folate* são diferentes ou não e quais são os intervalos de confiança para as diferenças de médias entre os tratamentos.

Após termos carregado o conjunto de dados *red.cell.folate*, para realizarmos a ANOVA para um fator no *R Commander*, é preciso selecionar a opção:

Estatísticas  $\Rightarrow$  Médias  $\Rightarrow$  ANOVA para um fator (one-way)...



Ao selecionarmos o item ANOVA para um fator, devemos escolher a variável que define os grupos de estudo e a variável resposta (figura 18.10). Nessa figura, também foi marcada a opção *comparação de médias 1 a 1* para obtermos os intervalos de confiança para as diferenças de médias entre os grupos. Não marcamos a opção para realizar o teste com a suposição de que as variâncias fossem diferentes. Os resultados da análise de variância serão armazenados no objeto *AnovaModel.1*, que será utilizado depois para gerar os diagnósticos gráficos e numéricos do modelo. Esse nome pode ser alterado pelo usuário.

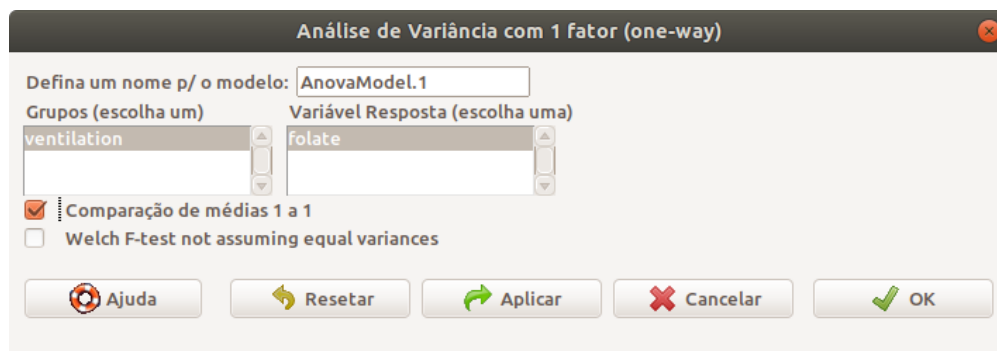


Figura 18.10: Seleção da variável que define os grupos e da variável resposta. Foi selecionada a opção para realizar a comparação de todos os pares de médias.

Os resultados são apresentados nas figuras 18.11 a 18.13. A figura 18.11 mostra o resumo do teste de hipótese, com o valor de  $p$  igual a 0,0436, estatisticamente significativo, quando o nível de significância escolhido é 5%. Logo abaixo são apresentados as médias, os desvios padrões e o número de elementos de cada método de ventilação.

A figura 18.12 mostra na porção inferior as diferenças de médias para cada par de métodos, assim como os limites inferior ( $lwr$ ) e superior ( $upp$ ) dos intervalos de confiança, usando o método de Tukey. A figura 18.13 mostra as três comparações de médias entre os métodos de modo gráfico. A única comparação que foi significativa ao nível de 5% foi a do N2O-O2, op e N2O-O2, 24h. As amplitudes dos intervalos de confiança são bastante elevadas, devido ao tamanho das amostras bastante pequeno.

```
> AnovaModel.1 <- aov(folate ~ ventilation, data=red.cell.folate)
> summary(AnovaModel.1)
              Df Sum Sq Mean Sq F value Pr(>F)
ventilation    2  15516    7758   3.711 0.0436 *
Residuals     19  39716    2090
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> with(red.cell.folate, numSummary(folate, groups=ventilation, statistics=c("mean", "sd")))
      mean      sd data:n
N2O+O2,24h 316.6250 58.71709      8
N2O+O2,op   256.4444 37.12180      9
O2,24h      278.0000 33.75648      5
```

Figura 18.11: Resultados da análise de variância para a configuração da figura 18.10. O valor de  $p$  é igual 0,0436, indicando uma significância estatística limítrofe ao nível de 5%.

```

Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts

Fit: aov(formula = folate ~ ventilation, data = red.cell.folate)

Linear Hypotheses:
              Estimate Std. Error t value Pr(>|t|)
N20+O2,op - N20+O2,24h == 0   -60.18      22.22  -2.709   0.0351 *
O2,24h - N20+O2,24h == 0    -38.62      26.06  -1.482   0.3202
O2,24h - N20+O2,op == 0      21.56      25.50   0.845   0.6792
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)


Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts

Fit: aov(formula = folate ~ ventilation, data = red.cell.folate)

Quantile = 2.5365
95% family-wise confidence level

Linear Hypotheses:
              Estimate   lwr       upr
N20+O2,op - N20+O2,24h == 0  -60.1806 -116.5309  -3.8302
O2,24h - N20+O2,24h == 0   -38.6250 -104.7369  27.4869
O2,24h - N20+O2,op == 0     21.5556  -43.1283  86.2394

N20+O2,24h  N20+O2,op   O2,24h
    "b"         "a"      "ab"

```

Figura 18.12: Continuação dos resultados da análise de variância para a configuração da figura 18.10. Na parte inferior, são apresentados os intervalos de confiança para cada par de grupos, usando o método de Tukey.

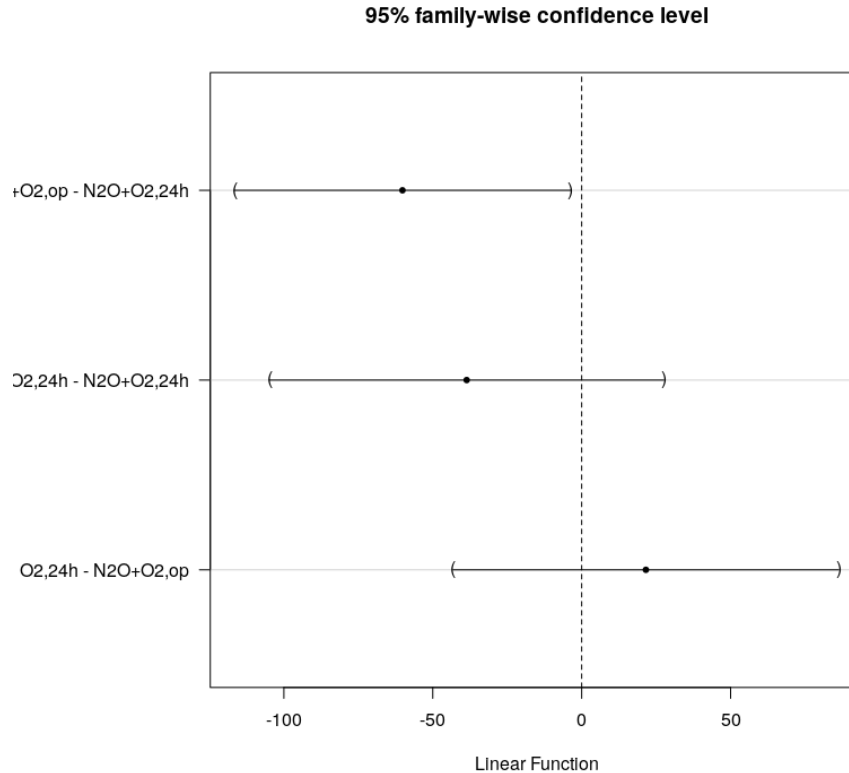


Figura 18.13: Diagrama que mostra os intervalos de confiança para as diferenças entre cada par de médias do ácido fólico, usando o método de Tukey.

Observem que o modelo *AnovaModel.1* gerado por essa análise aparece selecionado ao lado do rótulo *modelo*, conforme indicado na figura 18.14.

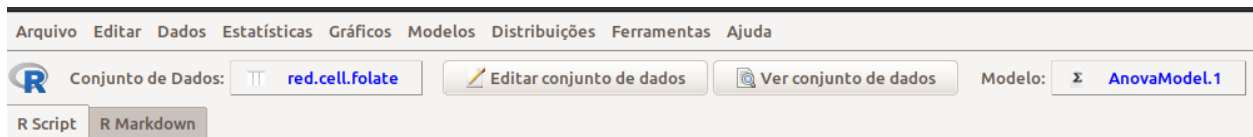


Figura 18.14: *R Commander* com o modelo *Anova.1* selecionado.

### 18.3.6.1 Avaliação da normalidade da variável resposta

Nesta seção, vamos construir o diagrama de comparação de quantis da normal e realizar um teste de normalidade para os resíduos do modelo de ANOVA gerado na seção anterior. Para isso, é necessário criar uma variável no conjunto de dados que irá conter os valores dos resíduos do modelo.

Certificando-nos que o modelo *AnovaModel.1* esteja selecionado, os resíduos podem ser gerados por meio da opção:

Modelos  $\Rightarrow$  Adicionar estatísticas calculadas aos dados

De todas as opções que aparecem na tela da figura 18.15, vamos selecionar somente a opção *Resíduos*. Ao clicarmos em OK, os resíduos serão calculados e aparecem no conjunto de dados *red.cell.folate* como mais uma variável (figura 18.16).

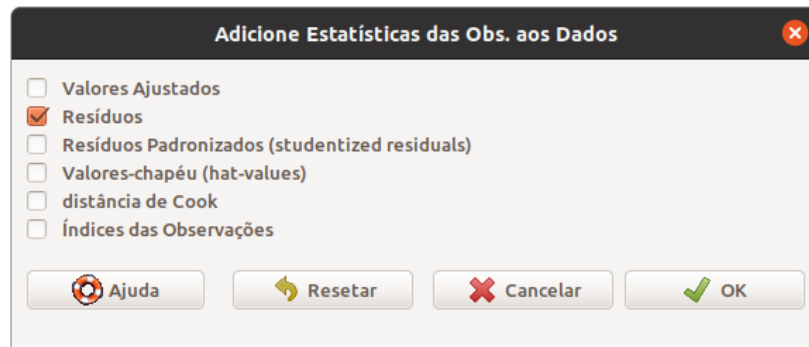


Figura 18.15: Seleção das estatísticas que serão agregadas aos dados de *red.cell.folate*.

red.cell.folate			
	folate	ventilation	residuals.AnovaModel.1
1	243	N20+02, 24h	-73.625000
2	251	N20+02, 24h	-65.625000
3	275	N20+02, 24h	-41.625000
4	291	N20+02, 24h	-25.625000
5	347	N20+02, 24h	30.375000
6	354	N20+02, 24h	37.375000
7	380	N20+02, 24h	63.375000
8	392	N20+02, 24h	75.375000
9	206	N20+02, op	-50.444444
10	210	N20+02, op	-46.444444
11	226	N20+02, op	-30.444444
12	249	N20+02, op	-7.444444
13	255	N20+02, op	-1.444444
14	273	N20+02, op	16.555556
15	285	N20+02, op	28.555556
16	295	N20+02, op	38.555556
17	309	N20+02, op	52.555556
18	241	02, 24h	-37.000000
19	258	02, 24h	-20.000000
20	270	02, 24h	-8.000000
21	293	02, 24h	15.000000
22	328	02, 24h	50.000000

Figura 18.16: Conjunto de dados *red.cell.folate* após o cálculo dos resíduos.

Para gerar o gráfico de comparação de quantis no *R Commander*, usamos a seguinte opção:

Gráficos ⇒ Gráfico de comparação de quantis...

Em seguida, selecionamos a variável *residuals.AnovaModel.1* (figura 18.17). Ao clicarmos em OK, o gráfico é mostrado na figura 18.18, o qual indica que não há desvios significativos da hipótese de normalidade.

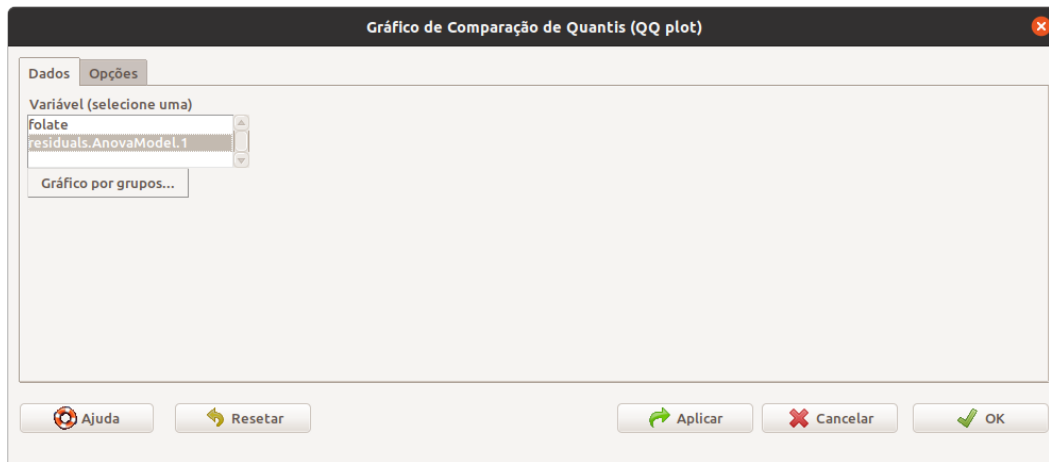


Figura 18.17: Seleção da variável para a construção do gráfico de comparação de quantis da normal.

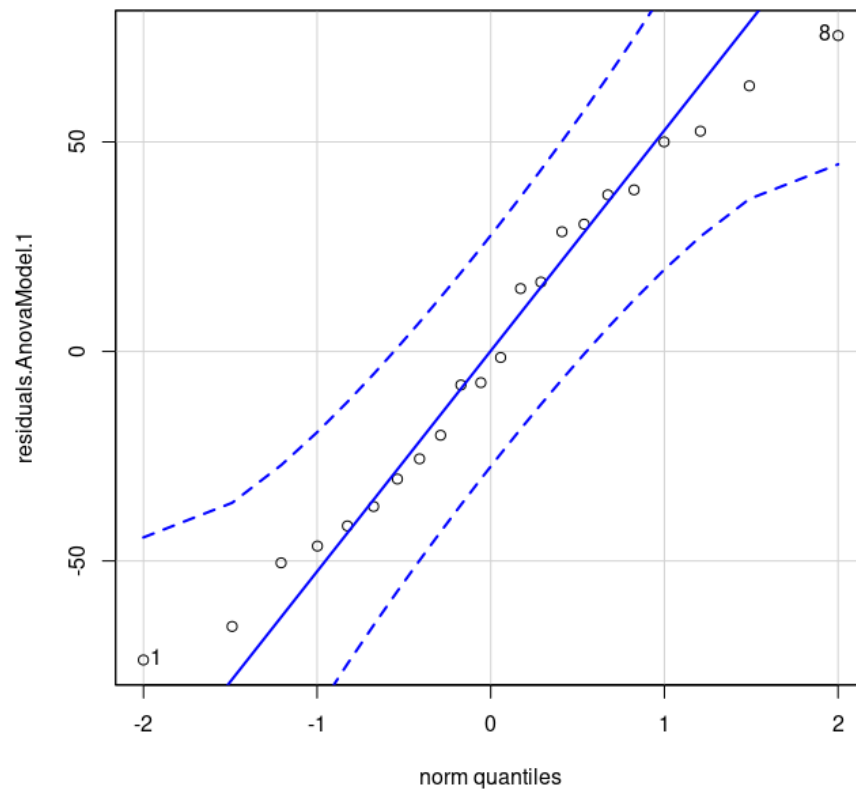


Figura 18.18: Diagrama de comparação dos quantis da normal para os resíduos do modelo.

Vamos agora realizar o teste de hipótese de normalidade de Shapiro-Wilk, acessando a opção:

Estatísticas  $\Rightarrow$  Resumos  $\Rightarrow$  Test of normality...

Em seguida, selecionamos a variável que desejamos testar, o teste de normalidade a ser realizado (figura 18.19) e clicamos em OK.

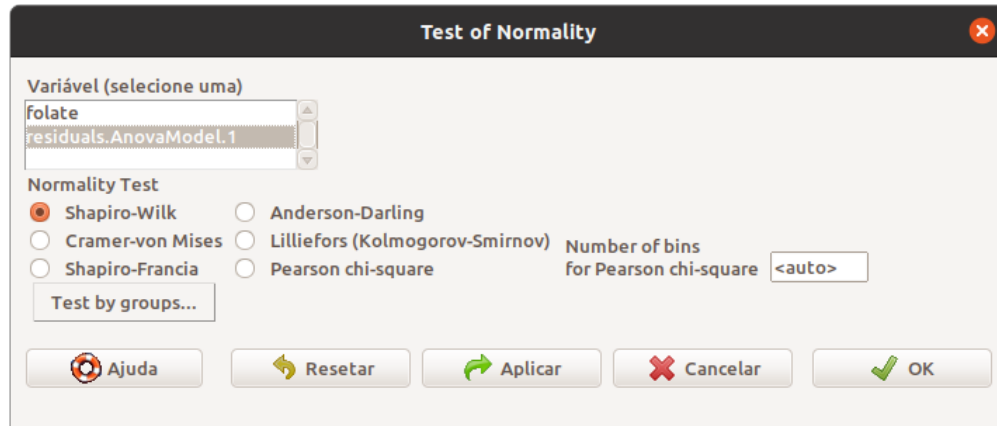


Figura 18.19: Seleção da variável e do teste de normalidade que será realizado.

O resultado é mostrado a seguir. O teste não rejeita a hipótese de normalidade ao nível de 5% ( $p = 0,62$ ), em concordância com a inspeção visual do gráfico de comparação de quantis da normal.

```
normalityTest(~residuals.AnovaModel.1, test="shapiro.test",
              data=red.cell.folate)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals.AnovaModel.1
## W = 0.966, p-value = 0.6188
```

### 18.3.6.2 Diagnósticos gráficos do modelo

A figura 18.20 mostra o menu do *R Commander* para obter uma série de diagnósticos gráficos. Esses diagramas serão gerados para o modelo selecionado no item *Modelo*, na parte superior direita da tela (seta vermelha). Nesse exemplo, o modelo *AnovaModel.1* foi gerado ao realizarmos a análise de variância para o conjunto de dados *red.cell.folate*.

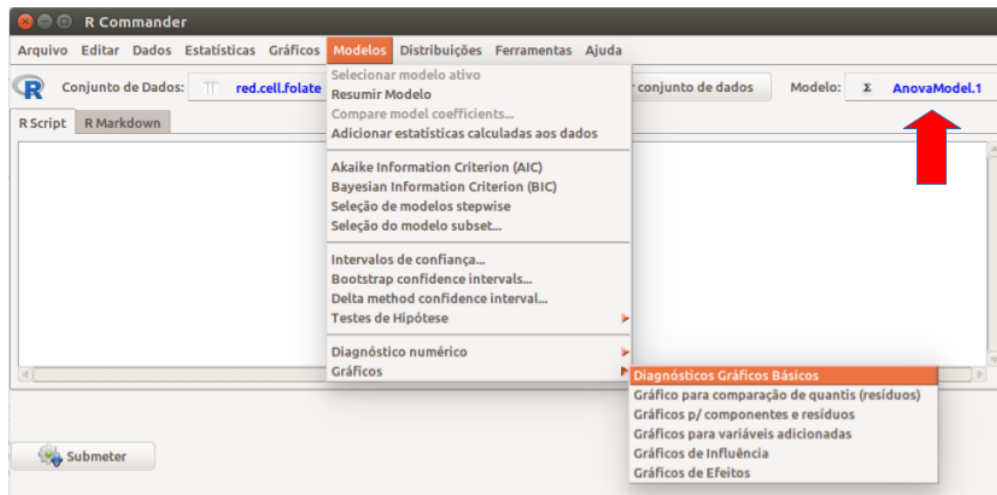


Figura 18.20: Menu do *R Commander* para selecionar os diagnósticos gráficos básicos para a análise de variância.

Ao selecionarmos a opção diagnósticos gráficos básicos na figura 18.20 para o modelo selecionado, os resultados são apresentados na figura 18.21. Ao inspecionarmos os gráficos, particularmente o gráfico de resíduos (superior à esquerda) e o gráfico de comparação de quantis para os resíduos padronizados (superior à direita), não vemos grandes desvios da normalidade, mas as variâncias não parecem ser iguais. Resíduos padronizados são os resíduos divididos pela raiz quadrada do erro quadrático médio residual.

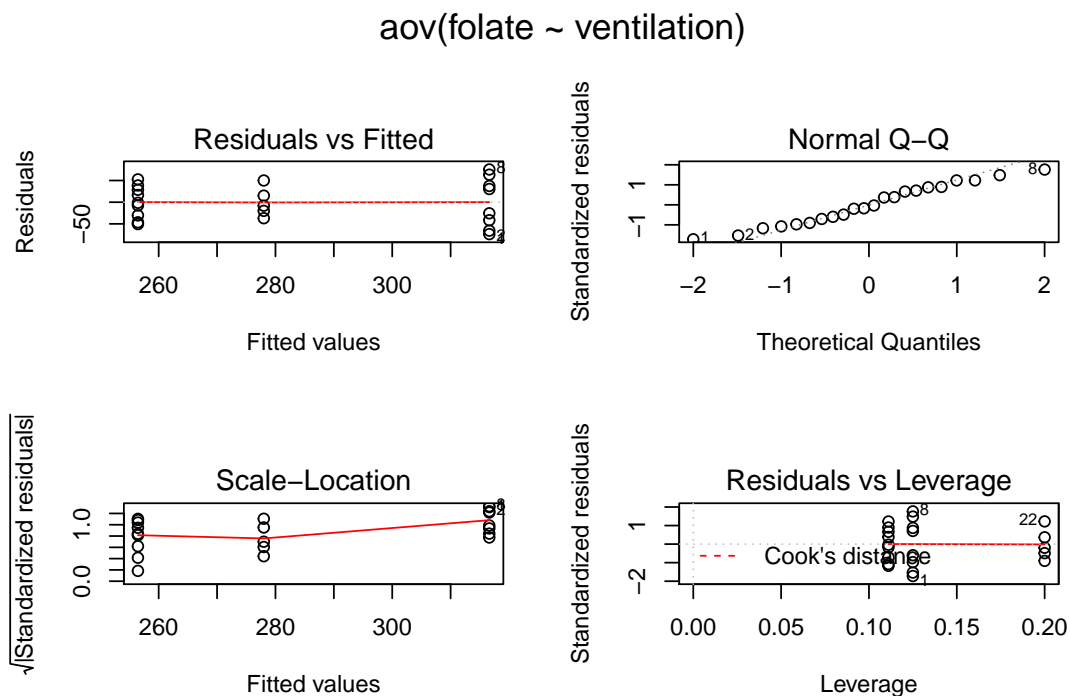


Figura 18.21: Diagnósticos gráficos básicos para a análise de variância do conjunto de dados *red.cell.folate*.

### 18.3.6.3 Testes para a igualdade de variâncias

Vamos realizar o teste de Levene para verificar a igualdade de variâncias, que pode ser acessado no *R Commander* por meio da opção:

Estatísticas  $\Rightarrow$  Variâncias  $\Rightarrow$  Teste de Levene...

Ao selecionarmos o teste, é preciso escolher a variável correspondente ao fator, a variável resposta e se o teste será baseado na média ou mediana (figura 18.22).



Figura 18.22: Configuração das variáveis para o teste de Levene para a comparação das variâncias da variável ácido fólico para os diversos métodos de ventilação.

O resultado é mostrado a seguir, indicando a rejeição da hipótese de igualdade de variâncias ao nível de 5% ( $p = 0,04$ ).

```
with(red.cell.folate, tapply(folate, ventilation, var, na.rm=TRUE))

## N20+O2,24h  N20+O2,op      O2,24h
##   3447.696   1378.028   1139.500

leveneTest(folate ~ ventilation, data=red.cell.folate, center="mean")

## Levene's Test for Homogeneity of Variance (center = "mean")
##      Df F value  Pr(>F)
## group  2   3.823 0.04024 *
##      19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

O teste de Bartlett pode ser acessado no *R Commander* por meio da opção:

Estatísticas  $\Rightarrow$  Variâncias  $\Rightarrow$  Teste de Bartlett

Também precisamos selecionar a variável correspondente ao fator e a variável resposta (figura 18.23).





Figura 18.23: Seleção das variáveis para o teste de Bartlett para a comparação das variâncias da variável ácido fólico para os diversos métodos de ventilação.

O resultado é mostrado a seguir, não indicando a rejeição da hipótese de igualdade de variâncias ao nível de 5% ( $p = 0,35$ ).

```
with(red.cell.folate, tapply(folate, ventilation, var, na.rm=TRUE))
```

```
## N20+02,24h N20+02,op 02,24h
## 3447.696 1378.028 1139.500
```

```
bartlett.test(folate ~ ventilation, data=red.cell.folate)
```

```
##
## Bartlett test of homogeneity of variances
##
## data: folate by ventilation
## Bartlett's K-squared = 2.0951, df = 2, p-value = 0.3508
```

Os testes de Levene e Bartlett produziram resultados conflitantes! Como as amostras são pequenas e a inspeção dos gráficos de diagnóstico sugere que as variâncias são diferentes, vamos realizar a ANOVA, supondo que as variâncias são diferentes.

#### 18.3.6.4 Análise de variância quando as variâncias são diferentes

Ao marcarmos a opção *Welch F-test not assuming equal variances* na figura 18.10 (figura 18.24), a análise de variância será realizada de acordo com o método de Welch (Welch, 1951).

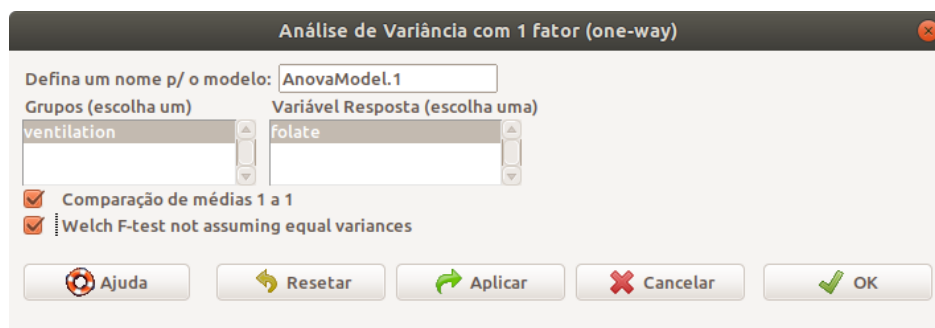


Figura 18.24: Seleção da variável que define os grupos e da variável resposta para uma análise de variância, supondo que as variâncias sejam diferentes nos diferentes grupos.

Os resultados são mostrados na figura 18.25. Os intervalos de confiança para as diferenças de médias essencialmente não são alterados, mas a hipótese nula de igualdade de médias não é rejeitada ao nível de 5% ( $p = 0,09$ ).

```
Multiple Comparisons of Means: Tukey Contrasts

Fit: aov(formula = folate ~ ventilation, data = red.cell.folate)

Linear Hypotheses:
              Estimate Std. Error t value Pr(>|t|)
N20+O2,op - N20+O2,24h == 0   -60.18     22.22  -2.709   0.0352 *
O2,24h - N20+O2,24h == 0     -38.62     26.06  -1.482   0.3203
O2,24h - N20+O2,op == 0       21.56     25.50   0.845   0.6792
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)

Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts

Fit: aov(formula = folate ~ ventilation, data = red.cell.folate)

Quantile = 2.537
95% family-wise confidence level

Linear Hypotheses:
              Estimate   lwr      upr
N20+O2,op - N20+O2,24h == 0  -60.1806 -116.5415  -3.8196
O2,24h - N20+O2,24h == 0    -38.6250 -104.7494  27.4994
O2,24h - N20+O2,op == 0      21.5556  -43.1406  86.2517

N20+O2,24h  N20+O2,op    O2,24h
   "b"         "a"       "ab"

> oneway.test(folate ~ ventilation, data=red.cell.folate) # Welch test

One-way analysis of means (not assuming equal variances)

data: folate and ventilation
F = 2.9704, num df = 2.000, denom df = 11.065, p-value = 0.09277
```

Figura 18.25: Resultados da análise de variância do ácido fólico para os diferentes métodos de ventilação, supondo que as variâncias sejam diferentes nos diferentes grupos.

### 18.3.6.5 Obtenção dos intervalos de confiança para a diferença de médias por outros métodos

A seção 18.3.6 calculou os intervalos de confiança para a comparação de médias duas a duas por meio do método de Tukey. A listagem abaixo mostra os comandos apresentados na janela de script do *R Commander* para gerar os resultados apresentados naquela seção.

```

library(mvtnorm, pos=17)
library(survival, pos=17)
library(MASS, pos=17)
library(TH.data, pos=17)
library(multcomp, pos=17)
library(abind, pos=22)
AnovaModel.1 <- aov(folate ~ ventilation, data=red.cell.folate)
summary(AnovaModel.1)
with(red.cell.folate, numSummary(folate, groups=ventilation,
                                statistics=c("mean", "sd")))

local({
  .Pairs <- glht(AnovaModel.1, linfct = mcp(ventilation = "Tukey"))
  print(summary(.Pairs)) # pairwise tests
  print(confint(.Pairs)) # confidence intervals
  print(cld(.Pairs)) # compact letter display
  old.oma <- par(oma=c(0,5,0,0))
  plot(confint(.Pairs))
  par(old.oma)
})

```

Podemos alterar o comando que cria o objeto *.Pairs* para obtermos os intervalos de confiança segundo o método de Dunnett, por exemplo, que usa um grupo como referência e compara as médias dos demais grupos com a média do grupo de referência. Se alterarmos o comando, substituindo “Tukey” por “Dunnett”, conforme abaixo, e executarmos os dois comandos em sequência, obteremos dois intervalos de confiança, tomando o grupo  $N_2O+O_2, 24h$  como referência. Observem que os dois intervalos são mais estreitos do que os obtidos pelo método de Tukey.

```

.Pairs <- glht(AnovaModel.1,
              linfct = mcp(ventilation = "Dunnett")) # trocando Tukey por Dunnett
print(confint(.Pairs)) # intervalos de confiança

```

```

##
## Simultaneous Confidence Intervals
##
## Multiple Comparisons of Means: Dunnett Contrasts
##
##
## Fit: aov(formula = folate ~ ventilation, data = red.cell.folate)
##
## Quantile = 2.3948
## 95% family-wise confidence level
##
##

```

```
## Linear Hypotheses:
##
##              Estimate   lwr      upr
## N20+02,op - N20+02,24h == 0  -60.1806 -113.3837  -6.9774
## 02,24h - N20+02,24h == 0    -38.6250 -101.0446  23.7946
```

O pacote *DescTools* permite a utilização de outros métodos para lidar com o problema de múltiplas comparações. Os métodos disponíveis na função *PostHocTest* desse pacote são: *hsd* - Tukey, *bonferroni*, *lsd*, *scheffe*, *newmankeuls*, *duncan*. Apesar de constar nessa lista, muitos autores não recomendam o método de Duncan.

Para utilizarmos a função *PostHocTest*, primeiramente instalamos o pacote *DescTools*, caso não esteja instalado, e depois o carregamos.

```
install.packages('DescTools')
library(DescTools)
```

Em seguida, para obtermos intervalos ao nível de confiança de 95% com a correção de Bonferroni para a comparação de médias duas a duas, usamos a função *PostHocTest*, lembrando que o objeto *AnovaModel.1* foi gerado durante a realização da análise de variância. O argumento *conf.level* especifica o nível de confiança.

```
PostHocTest(AnovaModel.1, method = "bonferroni", conf.level = 0.95)
```

```
##
##   Posthoc multiple comparisons of means : Bonferroni
##     95% family-wise confidence level
##
## $ventilation
##              diff      lwr.ci   upr.ci   pval
## N20+02,op-N20+02,24h -60.18056 -118.49975 -1.86136 0.0418 *
## 02,24h-N20+02,24h    -38.62500 -107.04688 29.79688 0.4643
## 02,24h-N20+02,op      21.55556  -45.38835 88.49947 1.0000
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Alterando o argumento *method* para *scheffe* na função *PostHocTest*, obtemos os intervalos de confiança para a comparação das médias de acordo com o método de Scheffé. A figura 18.26 mostra o gráfico dos intervalos de confiança para as diferenças de médias de ácido fólico entre os três métodos de ventilação, calculados de acordo com o método de Scheffé.

```
plot(PostHocTest(AnovaModel.1, method = "scheffe", conf.level = 0.95))
```

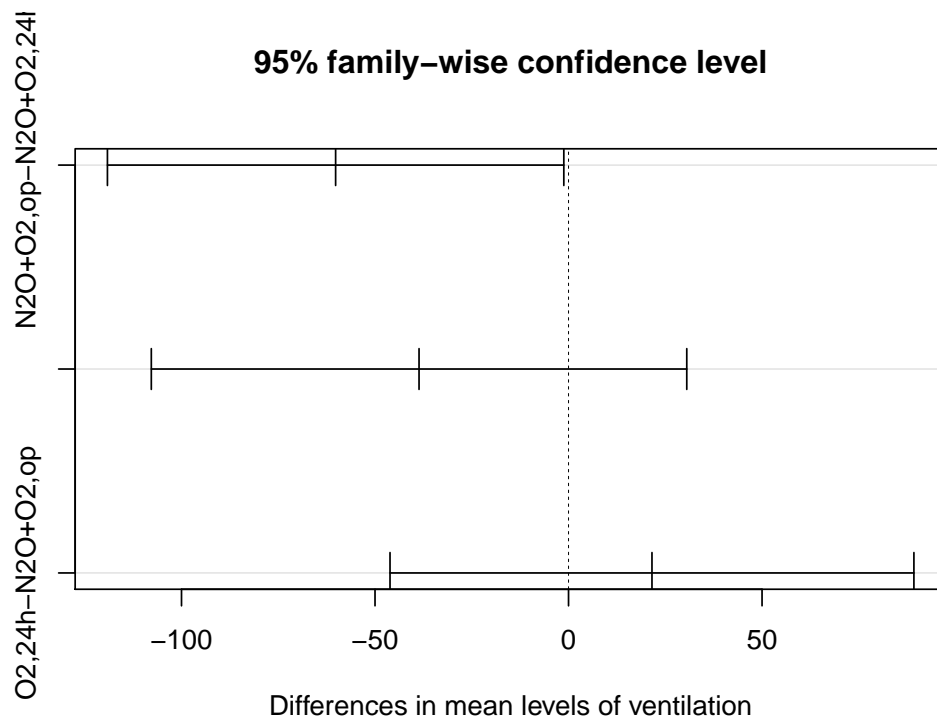


Figura 18.26: Intervalos de confiança da diferença de médias de ácido fólico entre os grupos de ventilação de acordo com o método de Scheffé.

Comparem os intervalos de confiança obtidos pelos diversos métodos mostrados nesta seção.

#### 18.3.6.6 Teste de Kruskal-Wallis no *R Commander*

Vamos utilizar novamente o conjunto de dados *red.cell.folate* para realizar o teste de Kruskal-Wallis no *R Commander*. Dessa vez, vamos utilizar o *RcmdrPlugin.EZR* para realizarmos o teste de Kruskal-Wallis, porque ele oferece o recurso de realizar múltiplas comparações entre os diferentes grupos. Para carregarmos o *plug-in RcmdrPlugin.EZR*, selecionamos no menu *Ferramentas* a opção *Carregar plug-in(s) do Rcmdr...* (seção 17.4.1). Caso ele não apareça na lista de *plug-ins*, é preciso instalá-lo.

Após carregarmos o *plug-in*, selecionamos novamente o conjunto de dados *red.cell.folate* como o conjunto de dados ativo. Para realizarmos o teste de Kruskal-Wallis no *R Commander*, selecionamos a opção:

Statistical analysis  $\Rightarrow$  Testes não paramétricos  $\Rightarrow$  Kruskal-Wallis test

A figura 18.27 mostra a seleção das variáveis para o teste.



Figura 18.27: Seleção da variável que define os grupos e da variável resposta para o teste de Kruskal-Wallis.

Ao realizarmos a seleção das variáveis, o método de múltiplas comparações e pressionarmos o botão OK, os resultados são mostrados a seguir.

```
library(RcmdrPlugin.EZR)
tapply(red.cell.folate$folate, red.cell.folate$ventilation, median,
       na.rm=TRUE)
```

```
## N20+02,24h  N20+02,op    02,24h
##           319          255      270
```

```
res <- NULL
(res <- kruskal.test(folate ~ factor(ventilation), data=red.cell.folate))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  folate by factor(ventilation)
## Kruskal-Wallis chi-squared = 4.1852, df = 2, p-value = 0.1234
```

```
pairwise.kruskal.test(red.cell.folate$folate, red.cell.folate$ventilation,
                     data.name="red.cell.folate", p.adjust.method="holm")
```

```
##
##  Pairwise comparisons using Mann-Whitney U test
##
## data:  red.cell.folate
##
```

```
##          N20+02,24h N20+02,op
## N20+02,op 0.18      -
## 02,24h    0.57      0.57
##
## P value adjustment method: holm
```

O teste de Kruskal-Wallis não rejeita a hipótese nula de igualdade de médias ( $p=0,12$ ). Observem como ele é menos poderoso do que o teste que utiliza o modelo da análise de variância. Nesse exemplo, as suposições para esse modelo são razoavelmente satisfeitas.

## 18.4 Análise de variância com medidas repetidas

Os conteúdos das subseções seguintes (subseções 18.4.1, 18.4.2, 18.4.3, 18.4.4 e 18.4.5) podem ser visualizados neste [vídeo](#).

### 18.4.1 Modelo

Na análise de variância com um fator apresentada nas seções anteriores, observações independentes são realizadas para cada nível do fator em estudo. Ela é uma generalização do teste  $t$  para amostras independentes. Também existe uma generalização para o teste  $t$  pareado (amostras dependentes). No teste  $t$  pareado, por exemplo, para comparar as medidas de peso antes e depois de uma dieta para emagrecimento, uma única amostra de indivíduos pode ser coletada e medidas de peso são realizadas antes e depois do tratamento.

Em vez de apenas duas medidas de peso, poderíamos ter várias medidas realizadas antes do tratamento e após 30, 60 e 90 dias. Uma outra situação é quando três tratamentos diferentes são aplicados ao mesmo paciente em instantes diferentes e a variável resposta é medida após cada tratamento.

Nesses exemplos, o interesse está em saber se existe diferenças entre as médias após cada medida realizada. As unidades de observação (pacientes por exemplo) são chamadas de **blocos**, sendo as medidas realizadas em cada bloco e se supõe que as unidades de observação sejam extraídas aleatoriamente da população de interesse.

Da maneira usual, vamos chamar a variável que indica os instantes das medidas ou os tratamentos oferecidos de **fator**. Se os níveis do fator são diferentes instantes de tempo em que a variável resposta é medida, então o estudo é chamado de **ANOVA com medidas repetidas**.

Os comandos a seguir carregam e mostram o conjunto de dados *heart.rate*, do pacote *ISwR*, que contém dados de frequência cardíaca (*hr*) de 9 pacientes com insuficiência cardíaca antes (*time = 0*) e após a administração de enalaprilato (*time = 30, 60 e 120 min*). Cada um dos 9 pacientes apresenta valores de *hr* em cada um dos instantes de tempo. Nesse caso, diz-se que o estudo é **balanceado**. Se algum paciente não apresentasse todas as medidas, ou se apresentasse mais medidas do que os demais pacientes em algum instante de tempo, o estudo seria **não balanceado**.

```
data(heart.rate, package="ISwR")
heart.rate
```

```
##      hr subj time
## 1   96    1    0
## 2  110    2    0
## 3   89    3    0
## 4   95    4    0
## 5  128    5    0
## 6  100    6    0
## 7   72    7    0
## 8   79    8    0
## 9  100    9    0
## 10  92    1   30
## 11 106    2   30
## 12  86    3   30
## 13  78    4   30
## 14 124    5   30
## 15  98    6   30
## 16  68    7   30
## 17  75    8   30
## 18 106    9   30
## 19  86    1   60
## 20 108    2   60
## 21  85    3   60
## 22  78    4   60
## 23 118    5   60
## 24 100    6   60
## 25  67    7   60
## 26  74    8   60
## 27 104    9   60
## 28  92    1  120
## 29 114    2  120
## 30  83    3  120
## 31  83    4  120
## 32 118    5  120
## 33  94    6  120
## 34  71    7  120
## 35  74    8  120
## 36 102    9  120
```

Vamos supor que temos  $k$  níveis do fator em estudo e que  $n$  unidades de observação serão avaliadas em cada nível do fator. O modelo para a análise de variância com medidas repetidas expressa a variável resposta nos seguintes componentes:



$$X_{ij} = \mu + \rho_i + \alpha_j + \epsilon_{ij} \quad (18.20)$$

onde:

j - indica cada nível do fator em estudo ( $j = 1, 2, \dots, k$ ).

i - indica cada uma das unidades de observação (por exemplo pacientes) ( $i = 1, 2, \dots, n$ ).

$X_{ij}$  - valor da variável resposta correspondente ao nível j do fator para a unidade i.

$\mu$  - média geral da variável X em todas os níveis do fator.

$\alpha_j$  - constante que representa a diferença entre a média da variável X na população correspondente ao nível j do fator de estudo e a média geral. As constantes  $\alpha_j$  satisfazem a condição  $\sum_{j=1}^k \alpha_j = 0$ .

$\rho_i$  - componente que segue a distribuição  $N(0, \sigma_\rho^2)$  e reflete as características específicas de cada unidade de observação.

$\epsilon_{ij}$  - erro associado ao valor j da unidade de observação i. Os erros seguem a distribuição  $N(0, \sigma^2)$ .

$\rho_i$  e  $\epsilon_{ij}$  são independentes.

Esse modelo é chamado de **modelo III (níveis dos fatores misturados)**, porque as unidades de observação (sujeitos) são aleatórias (um subconjunto de todas as unidades possíveis) e os níveis do fator em estudo são fixos.

### 18.4.2 Teste de hipótese

Na análise de variância de modelos fixos com um fator, a hipótese nula para um teste bilateral é que as médias da variável resposta correspondentes a cada nível do fator são iguais e a hipótese alternativa é que pelo menos uma das médias é diferente das demais:

$H_0: \alpha_j = 0, j = 1, 2, \dots, k$

$H_1: \text{nem todos os } \alpha_j = 0$

Para verificar a hipótese  $H_0$ , vamos considerar a situação em que uma variável aleatória X foi medida em k níveis do fator em n indivíduos diferentes, sendo cada indivíduo submetido a todos os níveis do fator em estudo.

A média aritmética de todos os valores de X (média geral) é dada por:

$$\bar{X} = \frac{\sum_{j=1}^k \sum_{i=1}^n X_{ij}}{nk} \quad (18.21)$$

A média aritmética dos valores de X para cada nível do fator é dada por:

$$\bar{X}_{.j} = \frac{\sum_{i=1}^n X_{ij}}{n}, \quad \text{onde } j = 1, 2, \dots, k \quad (18.22)$$

A média aritmética dos valores de X para cada unidade de observação é dada por:

$$\bar{X}_i = \frac{\sum_{j=1}^k X_{ij}}{k}, \quad \text{onde } i = 1, 2, \dots, n \quad (18.23)$$

### Partição da soma dos resíduos em relação à média geral

A expressão a seguir mostra que o desvio de cada valor da variável aleatória em relação à média geral é igual ao desvio da média do respectivo fator em relação à média geral somado ao desvio em relação à média geral da média da unidade de observação correspondente e um termo que representa a interação entre unidades de observação e os níveis do fator em estudo:

$$\underbrace{X_{ij} - \bar{X}}_{\text{desvio em relação à média geral}} = \underbrace{(\bar{X}_{.j} - \bar{X})}_{\text{desvio da média do nível do fator em relação à média geral}} + \underbrace{(\bar{X}_i - \bar{X})}_{\text{desvio da média da unidade de observação em relação à média geral}} + \underbrace{(X_{ij} - \bar{X}_{.j} - \bar{X}_i + \bar{X})}_{\text{termo que descreve a interação entre unidades de observação e níveis do fator}}$$

Elevando cada desvio ao quadrado e somando os quadrados de todos os desvios, pode-se mostrar que o resultado é a expressão abaixo:

$$\sum_{j=1}^k \sum_{i=1}^n (X_{ij} - \bar{X})^2 = \sum_{j=1}^k \sum_{i=1}^n (\bar{X}_{.j} - \bar{X})^2 + \sum_{j=1}^k \sum_{i=1}^n (\bar{X}_i - \bar{X})^2 + \sum_{j=1}^k \sum_{i=1}^n (X_{ij} - \bar{X}_{.j} - \bar{X}_i + \bar{X})^2$$

Essa expressão pode ser escrita como:

$$SQTot = SQS + SQF + SQF.S$$

onde:

$SQTot = \sum_{j=1}^k \sum_{i=1}^n (X_{ij} - \bar{X})^2$  representa a soma total dos quadrados, com nk-1 graus de liberdade.

$SQF = n \sum_{j=1}^k (\bar{X}_{.j} - \bar{X})^2$  representa a soma dos quadrados para o fator, com  $k-1$  graus de liberdade.

$SQS = k \sum_{i=1}^n (\bar{X}_{i.} - \bar{X})^2$  representa a soma dos quadrados para as unidades de análise, com  $n-1$  graus de liberdade.

$SQF.S = \sum_{j=1}^k \sum_{i=1}^n (X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X})^2$  representa a soma dos quadrados da interação entre unidades de análise e níveis do fator, com  $(k-1)(n-1)$  graus de liberdade

O erro quadrático médio da interação (EQMF.S) é obtido dividindo-se SQF.S pelo correspondente número de graus de liberdade:

$$EQMF.S = \frac{\sum_{j=1}^k \sum_{i=1}^n (X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X})^2}{(n-1)(k-1)}$$

O erro quadrático médio do fator (EQMF) é obtido dividindo-se SQF pelo correspondente número de graus de liberdade:

$$EQMF = \frac{n \sum_{j=1}^k (\bar{X}_{.j} - \bar{X})^2}{k-1}$$

Pode-se mostrar que o valor esperado do erro quadrático médio da interação é igual à variância de  $\epsilon_{ij}$ :

$$E[EQMF.S] = \sigma^2$$

e que o valor esperado do erro quadrático médio do fator é igual à expressão abaixo:

$$E[EQMF] = \sigma^2 + \frac{n \sum_{j=1}^k (\alpha_j - \mu)^2}{k-1}$$

## Teste F

Quando a hipótese nula é verdadeira, tanto o erro quadrático médio da interação quanto o erro quadrático médio do fator são estimadores não tendenciosos da variância dos erros do modelo da ANOVA com medidas repetidas. Quando a hipótese nula não é verdadeira, o valor esperado do erro quadrático médio do fator é maior do que a variância de  $\epsilon_{ij}$ , aumentando à medida que as diferenças entre as médias dos níveis dos fatores aumenta.

Assim a divisão do valor do erro quadrático médio do fator (EQMF) pelo erro quadrático médio da interação (EQMF.S) dá uma indicação de quanto a hipótese nula é compatível com os dados. O valor dessa divisão é representado por  $F^*$ :

$$F^* = \frac{EQMR}{EQMF.S} \quad (18.24)$$

Pode-se mostrar que a razão EQMR/EQMF.S, se a hipótese nula for verdadeira, segue a distribuição  $F(k-1, (k-1)(n-1))$ . Dado um nível de significância  $\alpha$ , quando o valor de  $F^*$ , obtido da expressão (18.24) for maior que o quantil  $(1 - \alpha)$  da distribuição  $F(k-1, (k-1)(n-1))$ , então a hipótese nula é rejeitada.

### 18.4.3 Diagnósticos para verificar o modelo de medidas repetidas

Uma condição para que a estatística  $F^*$  siga a distribuição  $F$  é que a variância da diferença entre as médias estimadas para dois níveis quaisquer do fator em estudo seja constante, ou seja:

$$var(\bar{X}_{i.} - \bar{X}_{.j}) = \text{constante}, i \neq j$$

Essa condição é conhecida como **esfericidade**. Alguns pacotes do R realizam um teste estatístico para verificar a condição de esfericidade e realizam ajustes no teste da ANOVA com medidas repetidas, quando a condição não é satisfeita.

Além disso, assim como na análise de variância com um fator, os erros devem possuir uma variância constante, serem independentes e normalmente distribuídos.

Diversos diagnósticos podem ser realizados para verificar se as condições para o modelo de análise de variância com medidas repetidas são satisfeitas, por exemplo:

- 1) um diagrama de interação, que mostra os valores da variável resposta por unidade de análise, pode ser examinado para verificar a interação entre unidades de observação e os níveis do fator em estudo;
- 2) um diagrama de comparação de quantis da normal dos resíduos. Os resíduos para o modelo de medidas repetidas apresentado nesta seção são dados por:

$$r_{ij} = X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}$$

- 3) um diagrama de pontos dos resíduos por níveis do fator em estudo para avaliar a constância da variância do erro;
- 4) diagramas da sequência de resíduos por unidade de análise podem ser úteis para verificar a constância da variância do erro e interferência de outros fatores. Esses gráficos são diagramas de dispersão dos resíduos x níveis do fator (se pudermos considerar que as medições para cada nível foram realizadas numa certa ordem), plotados separadamente para cada indivíduo;
- 5) um diagrama de comparação de quantis dos efeitos principais devido às unidades de análise,  $\bar{X}_{i.} - \bar{X}$ , para avaliar se os efeitos devidos a  $\rho_i$  estão normalmente distribuídos. Os efeitos devido às unidades de análise são estimados pelos desvios das médias de cada indivíduo em relação à média geral.

Na seção 18.4.6, será mostrado como obter e interpretar os diagnósticos acima.

### 18.4.4 Intervalos de confiança

Para obtermos intervalos de confiança para contrastes entre as médias dos diversos níveis do fator em estudo, podemos utilizar os mesmos métodos apresentados na seção 18.3.3, utilizando o valor de  $EQMF.S$  como estimativa da variância dos erros do modelo.

Vamos considerar as diferenças entre dois pares de médias:

$$\bar{D}_{lm} = \bar{X}_{.l} - \bar{X}_{.m}, \quad l, m = 1, \dots, k$$

A variância da diferença entre duas médias  $l$  e  $m$ , por exemplo, pode ser estimada por:

$$var(\bar{D}_{lm}) = var(\bar{X}_{.l}) + var(\bar{X}_{.m}) = \frac{2}{n}EQMF.S$$

Logo o erro padrão da diferença entre duas médias  $l$  e  $m$  é:

$$EP(\bar{D}_{lm}) = \sqrt{\frac{2}{n}EQMF.S}$$

### 18.4.5 Teste de Friedman

Quando as suposições para o modelo de análise de variância com medidas repetidas são seriamente violadas, um teste estatístico que pode ser realizado é o teste de Friedman. Nesse teste, os valores da variável resposta em cada um dos blocos (sujeitos ou unidades de observação) são ordenados e a cada um deles é atribuído um posto. Os postos são então somados para cada nível do fator, sendo  $R_j$  a soma dos postos para o fator  $j$ . A estatística

$$\chi_F^2 = \frac{12n}{k(k+1)} \sum_{j=1}^k (\bar{R}_j - \frac{k+1}{2})^2, \quad \text{sendo } \bar{R}_j = \frac{\sum_{i=1}^n R_{ij}}{n}$$

segue aproximadamente a distribuição qui ao quadrado com  $k-1$  graus de liberdade, desde que o número de blocos não seja muito pequeno.

### 18.4.6 Análise de variância com medidas repetidas no R

O conteúdo desta subseção até a subseção seguinte (seção 18.4.6.1) pode ser visualizado neste [vídeo](#).

Vamos utilizar o *RcmdrPlugin.EZR* para realizar a ANOVA com medidas repetidas para o conjunto de dados *heart.rate*.

Para usar o *RcmdrPlugin.EZR* para a análise com medidas repetidas, é preciso criar um *data.frame* onde cada uma das medidas repetidas apareçam em uma coluna separada (formato largo). Assim precisamos de 4 colunas, contendo as medidas de frequência cardíaca para cada indivíduo nos instantes 0, 20, 60 e 120 minutos. Para converter o conjunto de dados *heart.rate* no formato desejado, usamos a seguinte opção:

Dados  $\Rightarrow$  Conjunto dados ativo  $\Rightarrow$  Reshape dataset from long to wide format...

Na tela de configuração da transformação do conjunto de dados (figura 18.28), damos um nome para o conjunto de dados que será gerado (*heart.rateWide*), e selecionamos a variável que identifica cada indivíduo (*subj*), a variável resposta (*hr* - *variables that vary by occasion*) e o fator (*time* - *within subject factors*).

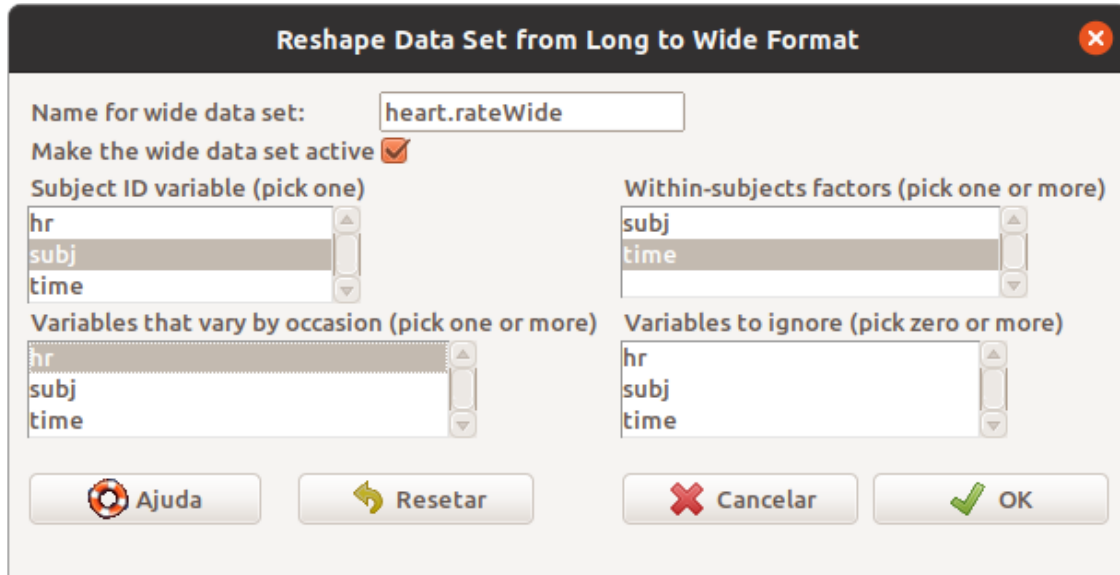


Figura 18.28: Diálogo para converter um *data.frame* do formato longo para o formato largo.

Ao clicarmos em OK, o comando a seguir é executado e o conjunto de dados *heart.rateWide* será criado a partir de *heart.rate*, com quatro variáveis (*X0*, *X30*, *X60*, *X120*) representando as medidas de frequência cardíaca para cada indivíduo nos instantes 0, 20, 60 e 120 minutos, respectivamente.

```
heart.rateWide <- reshapeL2W(heart.rate, within="time", id="subj",  
                             varying="hr")
```

Podemos, se for desejado, alterar os nomes das variáveis *X0*, *X30*, *X60*, *X120* por meio da opção:

Dados ⇒ Modificação de variáveis no conjunto de dados... ⇒ Renomear variáveis...

Na tela mostrada na figura 18.29, selecionamos as variáveis que desejamos renomear. Ao clicarmos em OK, damos os novos nomes para as variáveis selecionadas (figura 18.30).

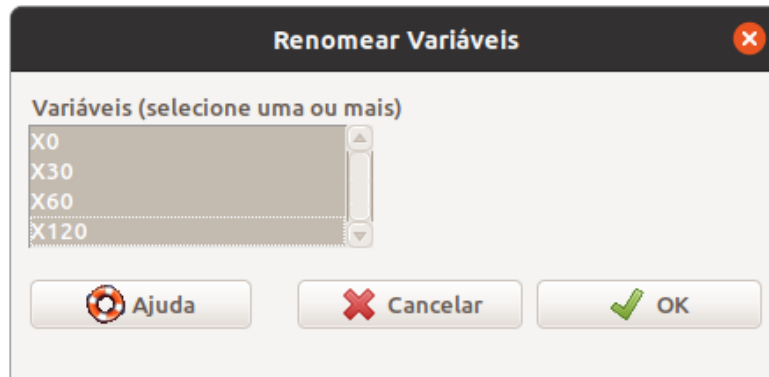


Figura 18.29: Seleção das variáveis a serem renomeadas no conjunto de dados *heart.rateWide*.

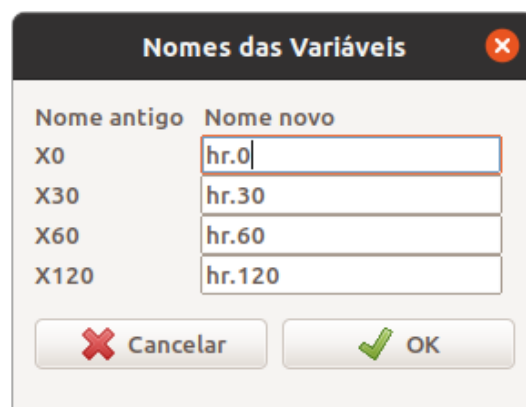


Figura 18.30: Fornecendo os novos nomes das variáveis selecionadas na figura 18.29.

Ao clicarmos em OK, o comando a seguir é executado, atribuindo os novos nomes das variáveis.

```
names(heart.rateWide)[c(1,2,3,4)] <- c("hr.0","hr.30","hr.60","hr.120")
```

O comando a seguir exibe o conjunto de dados *heart.rateWide*. Agora os valores da frequência cardíaca de cada paciente são exibidos numa única linha.

```
heart.rateWide
```

```
##   hr.0 hr.30 hr.60 hr.120
## 1   96   92   86   92
## 2  110  106  108  114
## 3   89   86   85   83
## 4   95   78   78   83
## 5  128  124  118  118
## 6  100   98  100   94
## 7   72   68   67   71
## 8   79   75   74   74
## 9  100  106  104  102
```

Para realizarmos a ANOVA com medidas repetidas no plugin *RcmdrPlugin.EZR*, acessamos a opção:

Statistical analysis  $\Rightarrow$  Continuous variables  $\Rightarrow$  Repeated-measures ANOVA

Na tela de configuração do teste (figura 18.31), selecionamos as variáveis que contêm as medidas de frequência cardíaca. Nesse exemplo, não há variável que separa os pacientes em grupos. Poderíamos realizar comparações par a par entre as médias de frequência cardíaca em cada instante, com correções de Holm ou Bonferroni, mas vamos realizar essas comparações de outra maneira, que mostra as diferenças e erros-padrão de cada comparação.

Os resultados do teste de hipótese são mostrados na figura 18.32, indicando que o valor de  $p$  ( $Pr(>F$  na figura)) é 0,01802 e rejeitando a hipótese nula de igualdade de médias para um nível de significância de 5%.

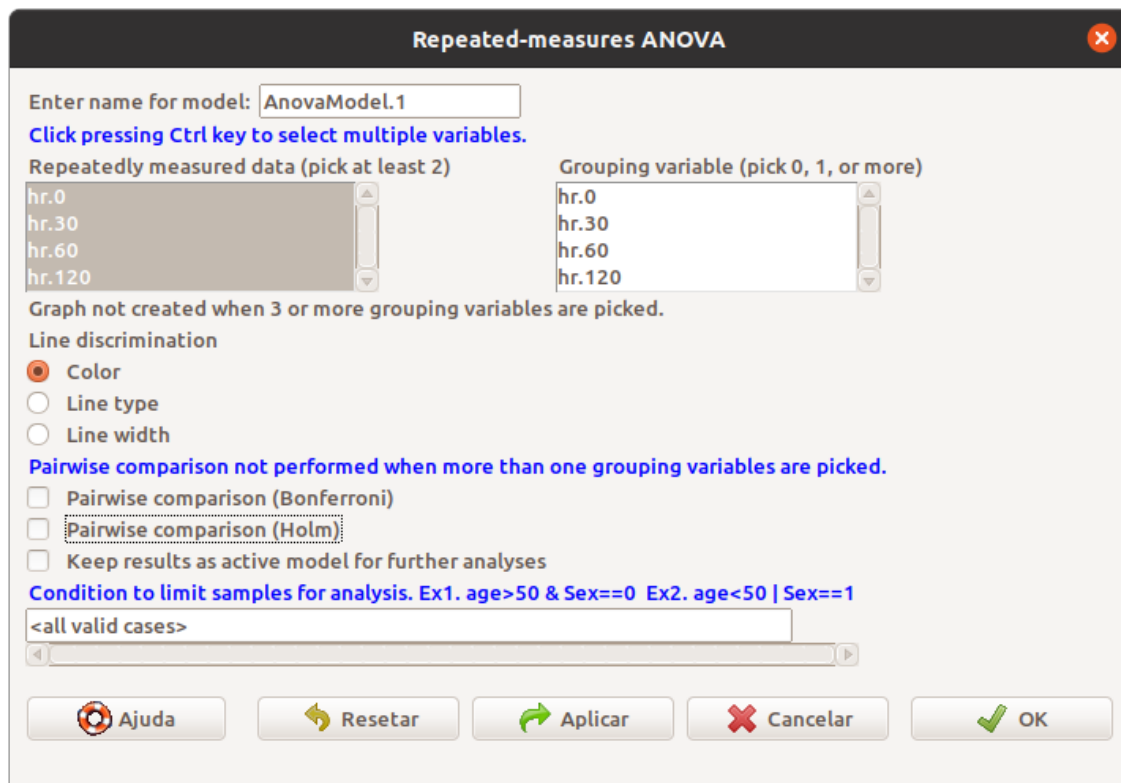


Figura 18.31: Seleção das medidas repetidas, configuração do gráfico de médias e comparação de pares de médias.



```

> summary(res, multivariate=FALSE)

Univariate Type III Repeated-Measures ANOVA Assuming Sphericity

              Sum Sq num Df Error SS den Df  F value    Pr(>F)
(Intercept) 312295      1   8966.6      8 278.6307 0.0000001678 ***
Time          151      3    296.8     24   4.0696   0.01802 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

|

Mauchly Tests for Sphericity

      Test statistic p-value
Time      0.47063  0.4122

Greenhouse-Geisser and Huynh-Feldt Corrections
for Departure from Sphericity

      GG eps Pr(>F[GG])
Time 0.70654   0.03412 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

      HF eps Pr(>F[HF])
Time 0.9679998 0.01930513

```

Figura 18.32: Resultados da análise de variância com medidas repetidas do conjunto de dados *heart.rate*.

Um teste de esfericidade (teste de Mauchly) não rejeitou a suposição de esfericidade. Mesmo assim, são apresentados os resultados de duas correções quando a suposição de esfericidade não é satisfeita: Greenhouse-Geisser epsilon (GGe), e Huynh-Feldt epsilon (HFe). As duas correções não alteram a significância estatística dos resultados ( $p = 0,034$  e  $0,019$ , respectivamente), caso tenha sido adotado o nível de significância de 5%.

A figura 18.33 mostra o gráfico de linhas das médias da frequência cardíaca em cada instante, indicando uma ligeira redução da média da frequência cardíaca após a administração do enalaprilato, ficando razoavelmente estável nos demais instantes.

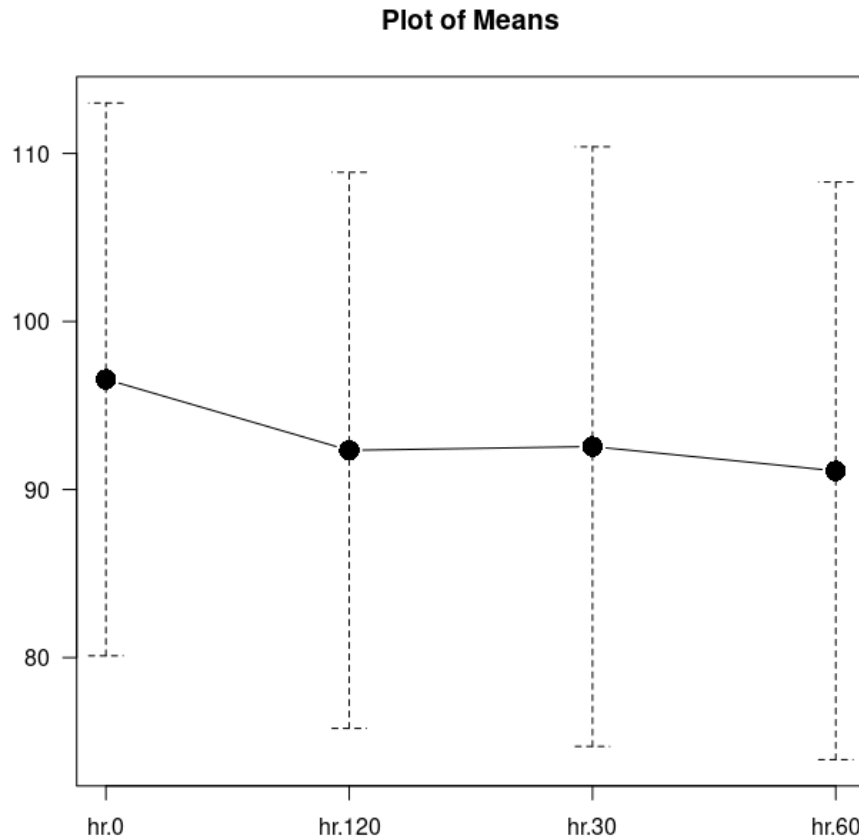


Figura 18.33: Diagrama de linha das médias das medidas de frequência cardíaca nos instantes 0, 30, 60 e 120 min.

Para verificarmos a interação entre as unidades de análise (pacientes) e o tempo de medição, ou seja, para verificarmos se os pacientes apresentam comportamentos diferentes de frequência cardíaca em relação ao tempo de medição, podemos gerar um diagrama de interação. Esse gráfico pode ser gerado por meio da seguinte opção:

Graphs and tables  $\Rightarrow$  Line graph (Repeated measures)

Na tela mostrada na figura 18.34, selecionamos as quatro variáveis que representam as medidas de frequência cardíaca nos quatro instantes diferentes e clicamos em OK. O diagrama de interação é mostrado na figura 18.35.

**Line graph(Repeated measures)**

Click pressing Ctrl key to select multiple variables

Repeatedly measured data (pick at least 2)

hr.0  
hr.30  
hr.60  
hr.120

Grouping variable(pick 0 or 1)

hr.0  
hr.30  
hr.60  
hr.120

Label for y-axis :

☐ Log y-axis

☐ Show different groups in separate graphs

Condition to limit samples for analysis. Ex1. age>50 & Sex==0 Ex2. age<50 | Sex==1

<all valid cases>

Ajuda Resetar Aplicar Cancelar OK

Figura 18.34: Seleção das variáveis que comporão o diagrama de interação entre pacientes e tempo de medição para a frequência cardíaca.

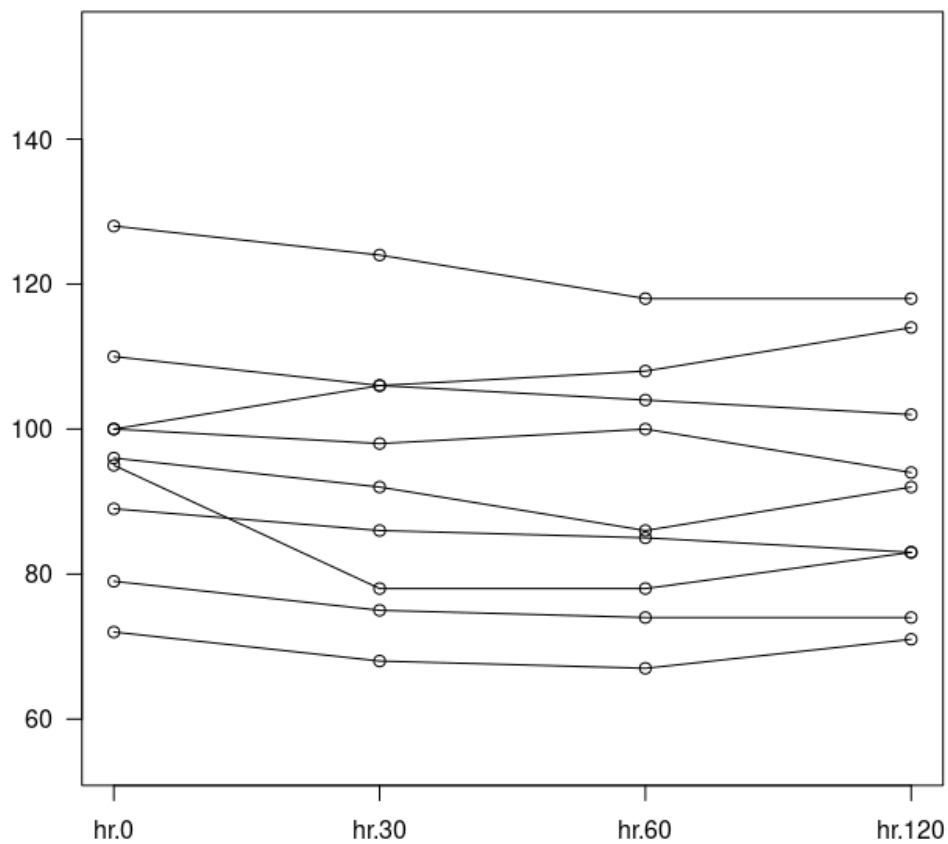


Figura 18.35: Diagrama de interação entre pacientes e tempo de medição para a frequência cardíaca, no conjunto de dados *heart.rate*.

O diagrama de interação une as medidas de frequência cardíaca para cada paciente por linhas. Nesse exemplo, temos 9 linhas correspondendo aos 9 pacientes. Podemos observar que, com uma única exceção, as linhas seguem um padrão de redução da frequência cardíaca do instante 0 até 120 min para cada paciente, com linhas razoavelmente paralelas, o que indica pouca interação entre os pacientes e o tempo. Caso as linhas dos pacientes seguissem padrões diferentes, isto indicaria que a frequência cardíaca apresentaria padrões diferentes de paciente para paciente.

Para realizarmos a comparação par a par das médias de frequência cardíaca em cada instante, podemos utilizar o script a seguir:

```
library(nlme)
library(multcomp)
am2 = lme(hr ~ time, random = ~1|subj, data=heart.rate)
.pairs = glht(am2, linfct=mcp(time="Tukey")) # múltiplas comparações
print(confint(.pairs)) # confidence intervals
old.oma <- par(oma=c(0,5,0,0))
plot(confint(.pairs), xlab = "Frequência cardíaca")
par(old.oma)
```

Primeiramente, carregamos dois pacotes (*nlme* e *multcomp*). Pode ser necessário instalá-los antes de carregá-los. Em seguida, realizamos novamente a análise de variância com medidas repetidas por meio da função *lme* do pacote *nlme*. Para essa análise, usamos o conjunto de dados *heart.rate* original.

Para usarmos a função *lme*, precisamos especificar o conjunto de dados que será analisado (*data = heart.rate*) e uma fórmula que indica como a análise será realizada. A expressão *hr ~ time* indica que *hr* é a variável resposta, o símbolo *~* separa a variável resposta das variáveis independentes, sendo que *time* indica o fator, que é a variável independente. Essa expressão especifica os efeitos fixos do modelo, devido à variável *time*.

O argumento *random = ~1|subj* especifica o componente  $\rho_i$  do modelo. O termo *~1|subj* indica que, a cada indivíduo (variável *subj*), corresponderá um valor aleatório que será adicionado à média geral, de acordo com o modelo (18.20).

O resultado da análise é armazenado em *am2*.

Em seguida, usamos a mesma sequência de comandos que foram usados no exemplo da ANOVA para um fator para o cálculo de intervalos de confiança para todos os pares de médias correspondentes aos níveis dos fator. Nesse caso, a função *glht* usa o objeto *am2*, resultado da análise de variância, e gera todas as diferenças entre as médias de *hr* nos diferentes instantes de tempo, por meio do método de Tukey. O resultado é armazenado no objeto *.pairs*.

O comando *print(confint(.pairs))* imprime os intervalos de confiança para cada comparação de médias (figura 18.36).

##		Estimate	lwr	upr
##	30 - 0 == 0	-4.00000	-8.25894	0.25894
##	60 - 0 == 0	-5.44444	-9.70338	-1.18551
##	120 - 0 == 0	-4.22222	-8.48116	0.03671
##	60 - 30 == 0	-1.44444	-5.70338	2.81449
##	120 - 30 == 0	-0.22222	-4.48116	4.03671
##	120 - 60 == 0	1.22222	-3.03671	5.48116

Figura 18.36: Intervalos de confiança para a comparação de médias de frequência cardíaca para cada par de instantes de tempo, usando o método de Tukey.

Finalmente o comando `plot(confint(.pairs))` gera um gráfico (figura 18.37) com os intervalos de confiança para a comparação de médias de frequência cardíaca para cada par de instantes de tempo.

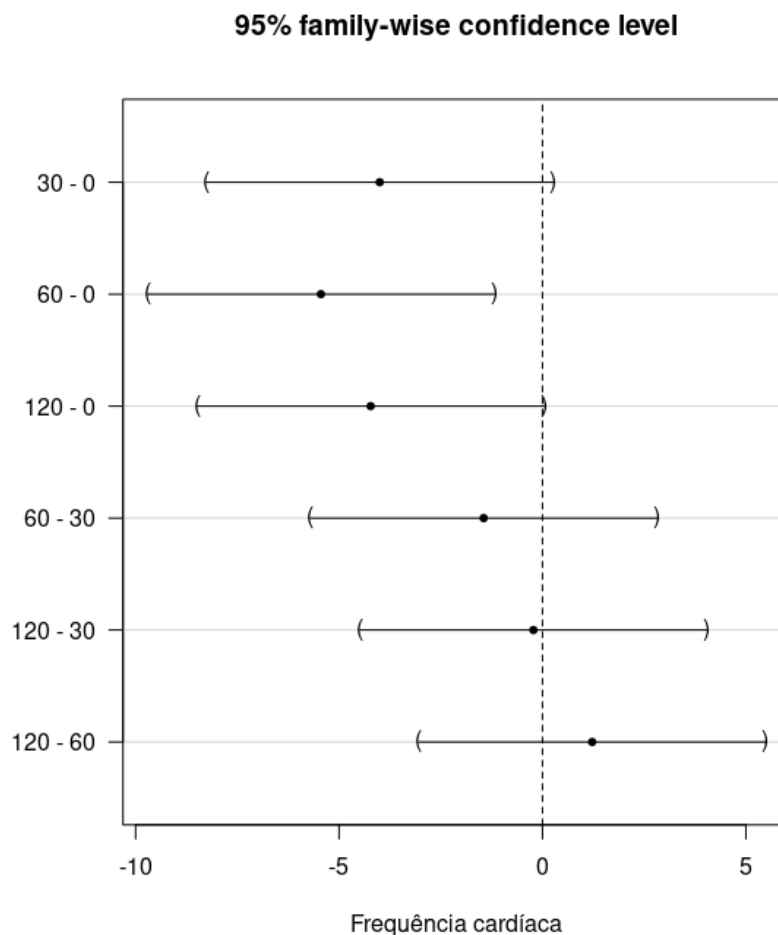


Figura 18.37: Diagrama que mostra os intervalos de confiança para a comparação de médias de frequência cardíaca para cada par de instantes de tempo, usando o método de Tukey.

Nesse exemplo, a diferença entre as medidas nos instantes 60 e 0 são estatisticamente significativas ao nível de 5%, e o intervalo de confiança ao nível de 95% dessa diferença indica uma redução máxima na média da frequência cardíaca entre esses dois instantes de 9,7 bpm.

#### 18.4.6.1 Diagnósticos da ANOVA com medidas repetidas no R

Os conteúdos desta subseção e das subseções seguintes (18.4.6.1.1 a 18.4.6.1.4 e 18.4.6.2) podem ser visualizados neste [vídeo](#).

Nesta seção, vamos avaliar os resíduos do modelo de ANOVA com medidas repetidas, gerado na seção anterior. Recordando, os resíduos para esse modelo são obtidos por meio da expressão

$$r_{ij} = X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}$$

onde:

$X_{ij}$  é a medida da frequência cardíaca no instante  $j$  para o indivíduo  $i$ ;

$\bar{X}_{i.}$  é a média da frequência cardíaca para o indivíduo  $i$ ;

$\bar{X}_{.j}$  é a média da frequência cardíaca para o instante  $j$ ;

$\bar{X}$  é a média geral da frequência cardíaca para todos os indivíduos e todos os instantes de medida.

A sequência de comandos a seguir pode ser executada para gerar os resíduos para a frequência cardíaca e a figura 18.38 pode facilitar a compreensão desse processo:

```
heart.rate$media.hr.subj = ave(heart.rate$hr, heart.rate$subj)
heart.rate$media.hr.time = ave(heart.rate$hr, heart.rate$time)
heart.rate$media.geral = ave(heart.rate$hr)
heart.rate$residuos = heart.rate$hr - heart.rate$media.hr.subj -
                      heart.rate$media.hr.time + heart.rate$media.geral
```

À esquerda, a figura 18.38 mostra o conjunto de dados *heart.rateWide* com as médias da frequência cardíaca em cada instante de medida, as médias da frequência cardíaca para cada indivíduo (médias das linhas) e a média geral da frequência cardíaca. À direita, a figura mostra o conjunto de dados *heart.rate* com o acréscimo das variáveis criadas na sequência de comandos acima.

O primeiro comando cria a variável *media.hr.subj* no conjunto de dados *heart.rate* que irá conter, para cada linha, a média da frequência cardíaca em todos os instantes para o indivíduo indicado na respectiva linha (variável *subj*). Essa média é obtida por meio da função *ave* que gera a média da variável indicada pelo primeiro argumento (*hr*) da função para cada nível da variável indicada pelo segundo argumento (*subj*). Todas as linhas para o mesmo indivíduo terão o mesmo valor de *media.hr.subj*.

$$r_{ij} = X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}$$

Heart.rateWide

	hr.0	hr.30	hr.60	hr.120	médias das linhas
1	96	92	86	92	91,5
2	110	106	108	114	109,5
3	89	86	85	83	85,8
4	95	78	78	83	83,5
5	128	124	118	118	122,0
6	100	98	100	94	98,0
7	72	68	67	71	69,5
8	79	75	74	74	75,5
9	100	106	104	102	103,0
médias das colunas	96,6	92,6	91,1	92,3	93,1
					média geral

$\bar{X}_{i.}$  heart.rate\$media.hr.subj = ave(heart.rate\$hr, heart.rate\$subj)

$\bar{X}_{.j}$  heart.rate\$media.hr.time = ave(heart.rate\$hr, heart.rate\$time)

$\bar{X}$  heart.rate\$media.geral = ave(heart.rate\$hr)

$r_{ij}$  heart.rate\$residuos = heart.rate\$hr - heart.rate\$media.hr.subj - heart.rate\$media.hr.time + heart.rate\$media.geral

	$X_{ij}$	$\bar{X}_{i.}$	$\bar{X}_{.j}$	$\bar{X}$	$r_{ij}$		
heart.rate							
	hr	subj	time	media.hr.subj	media.hr.time	media.geral	resíduos
1	96	1	0	91.50	96.55556	93.13889	1.08333333
2	110	2	0	109.50	96.55556	93.13889	-2.91666667
3	89	3	0	85.75	96.55556	93.13889	-0.16666667
4	95	4	0	83.50	96.55556	93.13889	0.80333333
5	128	5	0	122.00	96.55556	93.13889	2.58333333
6	100	6	0	98.00	96.55556	93.13889	-1.41666667
7	72	7	0	69.50	96.55556	93.13889	-0.91666667
8	79	8	0	75.50	96.55556	93.13889	0.08333333
9	100	9	0	103.00	96.55556	93.13889	-6.41666667
10	92	1	30	91.50	92.55556	93.13889	1.08333333
11	106	2	30	109.50	92.55556	93.13889	-2.91666667
12	86	3	30	85.75	92.55556	93.13889	0.83333333
13	78	4	30	83.50	92.55556	93.13889	-4.91666667
14	124	5	30	122.00	92.55556	93.13889	2.58333333
15	98	6	30	98.00	92.55556	93.13889	0.58333333
16	68	7	30	69.50	92.55556	93.13889	-0.91666667
17	75	8	30	75.50	92.55556	93.13889	0.08333333
18	106	9	30	103.00	92.55556	93.13889	-3.58333333
19	86	1	60	91.50	91.11111	93.13889	-3.47222222
20	108	2	60	109.50	91.11111	93.13889	0.52777778
21	85	3	60	85.75	91.11111	93.13889	1.27777778
22	78	4	60	83.50	91.11111	93.13889	-3.47222222
23	118	5	60	122.00	91.11111	93.13889	-1.97222222
24	100	6	60	98.00	91.11111	93.13889	-4.02777778
25	67	7	60	69.50	91.11111	93.13889	-0.47222222
26	74	8	60	75.50	91.11111	93.13889	0.52777778
27	104	9	60	103.00	91.11111	93.13889	-3.02777778
28	92	1	120	91.50	92.33333	93.13889	-1.30555556
29	114	2	120	109.50	92.33333	93.13889	-5.30555556
30	83	3	120	85.75	92.33333	93.13889	-1.94444444
31	83	4	120	83.50	92.33333	93.13889	-0.30555556
32	118	5	120	122.00	92.33333	93.13889	-3.19444444
33	94	6	120	98.00	92.33333	93.13889	-3.19444444
34	71	7	120	69.50	92.33333	93.13889	-2.30555556
35	74	8	120	75.50	92.33333	93.13889	-0.69444444
36	102	9	120	103.00	92.33333	93.13889	-0.19444444

Figura 18.38: Figura que ilustra a geração dos resíduos para o modelo de ANOVA com medidas repetidas para o conjunto de dados *heart.rate*.

O segundo comando cria a variável *media.hr.time* no conjunto de dados *heart.rate* que irá conter, para cada linha, a média da frequência cardíaca de todos os indivíduos no instante indicado pelo valor da variável *time* na respectiva linha. Essa média também é obtida por meio da função *ave*, onde a variável de agrupamento, segundo argumento da função, é a variável *time*. Todas as linhas com o mesmo instante de medida terão o mesmo valor de *media.hr.time*.

O terceiro comando cria a variável *media.geral* no conjunto de dados *heart.rate* que irá conter em todas as linhas a média geral da frequência cardíaca de todos os indivíduos em todos os instantes. Essa média também é obtida por meio da função *ave*, sem especificação de variável de agrupamento.

Finalmente o quarto comando cria a variável *residuos* no conjunto de dados, aplicando a fórmula para os resíduos, utilizando as variáveis criadas nos comandos anteriores.

Com exceção da variável *residuos*, não é necessário que as demais variáveis sejam acrescentadas ao conjunto de dados. Elas somente foram acrescentadas aqui para fins didáticos.

A partir dos resíduos, vamos nas seções seguintes gerar os seguintes diagramas:

- 1) um diagrama de comparação de quantis da normal dos resíduos;
- 2) um diagrama de pontos dos resíduos por instante de medida para avaliar a constância da variância do erro;
- 3) diagramas da sequência de resíduos por indivíduos para verificar a constância da variância do erro e interferência de outros fatores;

- 4) um diagrama de comparação de quantis dos efeitos principais devido aos indivíduos  $\bar{X}_{i.} - \bar{X}$  pode ser útil para avaliar se os efeitos principais  $\rho_i$  estão normalmente distribuídos com variância constante.

#### 18.4.6.1.1 Diagrama de comparação de quantis da normal para os resíduos

Uma vez obtido os resíduos para o conjunto de dados *heart.rate*, estando esse conjunto de dados como ativo no *R Commander* e supondo que o plugin *RcmdrPlugin.EZR* esteja carregado, para gerarmos o gráfico de comparação de quantis dos resíduos, acessamos a opção:

Original menu  $\Rightarrow$  Gráficos  $\Rightarrow$  Gráfico de comparação de quantis...

Na tela seguinte, selecionamos a variável *residuos* na lista de variáveis e clicamos em OK (figura 18.39).

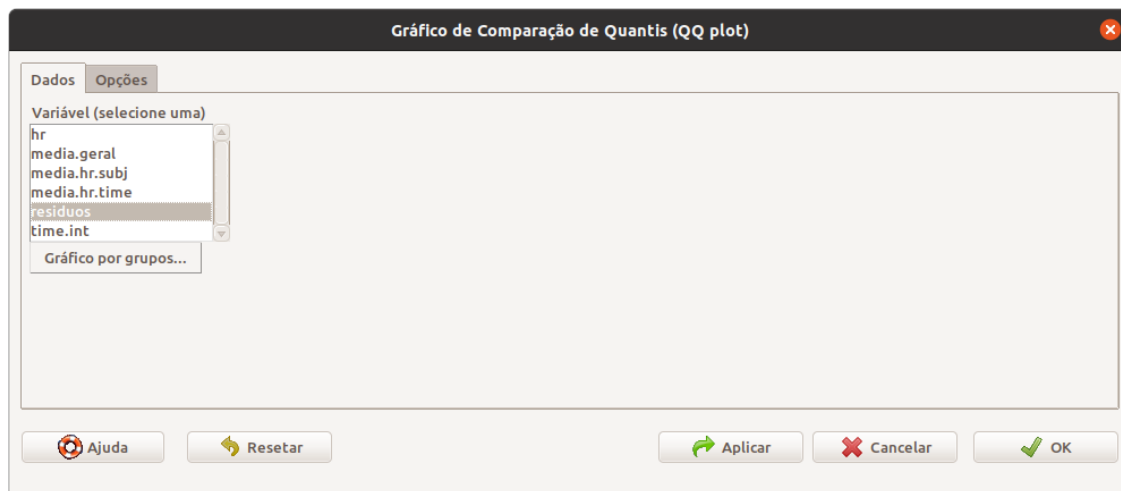


Figura 18.39: Seleção da variável para a construção do gráfico de comparação de quantis da normal.

O comando executado é mostrado a seguir, seguido do diagrama (figura 18.40).

```
with(heart.rate, qqPlot(residuos, dist="norm", id=list(method="y", n=2,
labels=rownames(heart.rate))))
```



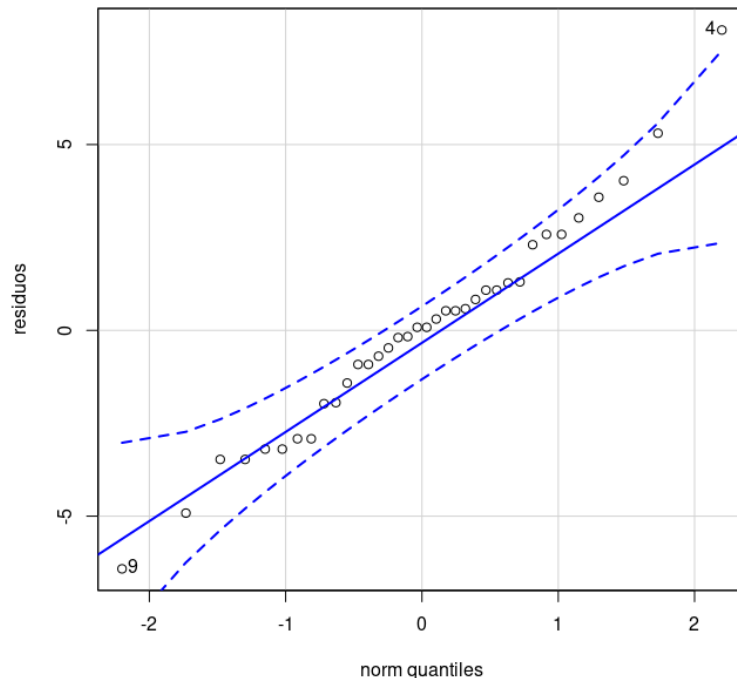


Figura 18.40: Gráfico de comparação de quantis da normal para os resíduos da frequência cardíaca.

Para realizarmos um teste de hipótese de normalidade dos resíduos, acessamos a opção:

Original menu  $\Rightarrow$  Estatísticas  $\Rightarrow$  Resumos  $\Rightarrow$  Test of normality...

Em seguida, selecionamos a variável que desejamos testar, o teste de normalidade a ser realizado (figura 18.41) e clicamos em OK.

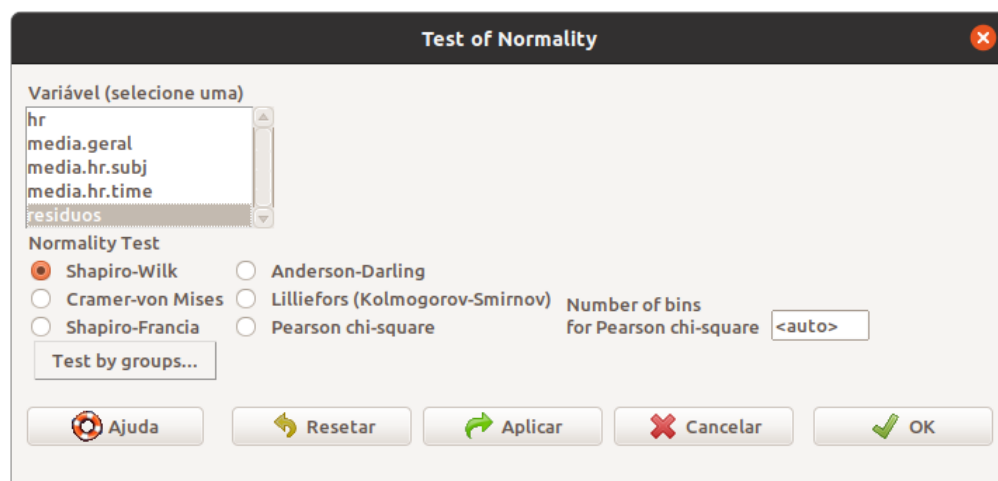


Figura 18.41: Seleção da variável e do teste de normalidade que será realizado.

O resultado é mostrado a seguir. O teste não rejeita a hipótese de normalidade ao nível de 5% ( $p = 0,62$ ), em concordância com a inspeção visual do gráfico de comparação de quantis da normal.

```
normalityTest(~residuals.AnovaModel.1, test="shapiro.test",  
              data=red.cell.folate)
```

```
##  
##  Shapiro-Wilk normality test  
##  
## data:  residuals.AnovaModel.1  
## W = 0.966, p-value = 0.6188
```

**18.4.6.1.2 Diagrama de pontos dos resíduos por instante de medida** Para gerarmos o diagrama de pontos dos resíduos por instante de medida, acessamos a opção:

Original menu  $\Rightarrow$  Gráficos  $\Rightarrow$  Gráfico Strip chart

Na tela seguinte, selecionamos a variável *time* como fator e a variável *resíduos* na lista de variáveis resposta e clicamos em OK (figura 18.42).

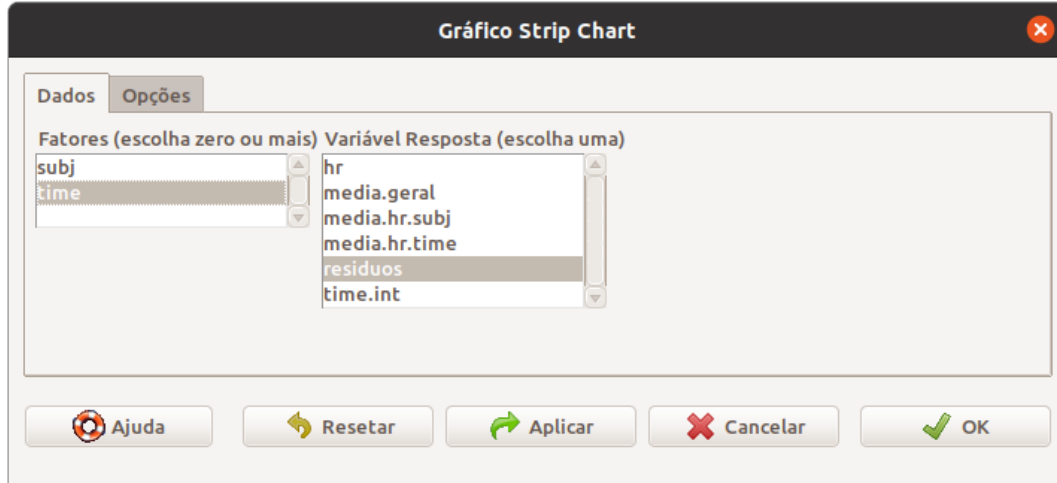


Figura 18.42: Seleção da variável para a construção do gráfico de *strip chart* e da variável de agrupamento.

O comando executado é mostrado a seguir, seguido do diagrama (figura 18.43).

```
stripchart(resíduos ~ time, vertical=TRUE, method="stack", ylab="resíduos",  
           data=heart.rate)
```

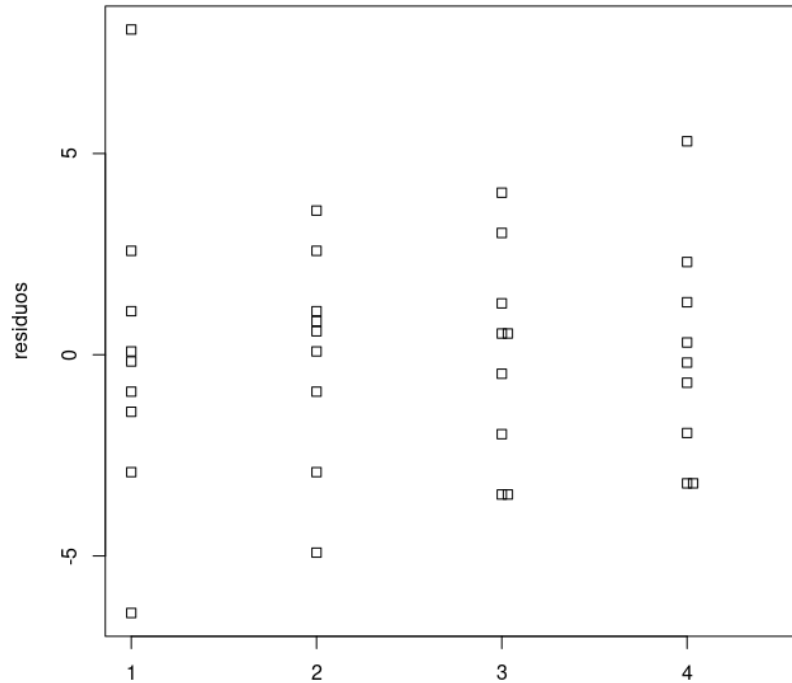


Figura 18.43: Gráfico de *strip chart* dos resíduos para cada instante de medida.

O diagrama de *strip chart* não sugere que haja diferenças das variâncias dos resíduos nos diferentes instantes de medida.

**18.4.6.1.3 Diagramas da sequência dos resíduos por indivíduos** Os diagramas da sequência de resíduos por indivíduos podem ser úteis para verificar a constância da variância do erro e interferência de outros fatores. Eles são diagramas de dispersão dos *resíduos* x *tempo*, plotados separadamente para cada indivíduo. Para gerarmos esses diagramas no *R Commander*, precisamos que a variável *time* seja tratada como numérica. Para isso, o comando abaixo converte a variável *time* em inteira e cria a variável *time.int* no conjunto de dados *heart.rate*.

```
heart.rate$time.int = as.integer(heart.rate$time)
```

Em seguida, para gerarmos os diagramas de sequência dos resíduos por indivíduos, acessamos a seguinte opção no menu:

Original menu ⇒ Gráficos ⇒ Gráfico XY (dispersão) condicionado...

Na tela seguinte (figura 18.44), selecionamos a variável *time.int* como variável explicativa, a variável *resíduos* como variável resposta e a variável *subj* como condição.



Figura 18.44: Seleção das variáveis para a construção dos diagramas de dispersão dos resíduos x tempo para cada indivíduo.

Ao clicarmos em OK, o comando executado é mostrado a seguir, seguido do diagrama que mostra a dispersão dos resíduos x tempo de medida para cada indivíduo separadamente (figura 18.45).

```
xyplot(resíduos ~ time | subj, type="p", pch=16, auto.key=list(border=TRUE),
       par.settings=simpleTheme(pch=16), scales=list(x=list(relation='same'),
       y=list(relation='same')), data=heart.rate)
```

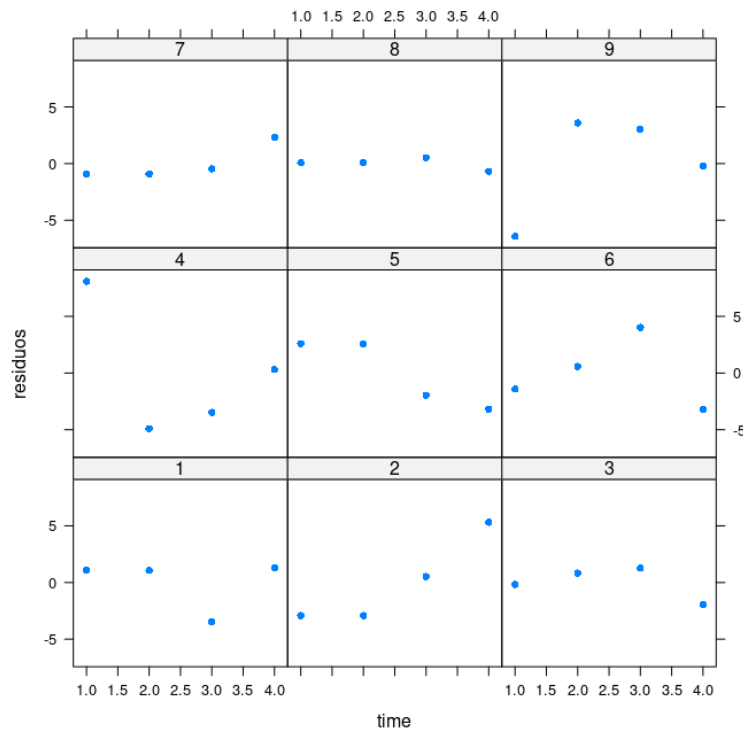


Figura 18.45: Diagrama que mostra a dispersão dos resíduos x tempo de medida para cada indivíduo separadamente.

Os diagramas não sugerem um padrão na distribuição dos resíduos em relação ao tempo, mas é preciso destacar que a amostra é pequena e há apenas 4 pontos por indivíduo.

**18.4.6.1.4 diagrama de comparação de quantis dos efeitos principais devido aos indivíduos** Um diagrama de comparação de quantis dos efeitos principais devido às unidades de análise,  $\bar{X}_{i.} - \bar{X}$ , é utilizado para avaliar se os efeitos principais  $\rho_i$  estão normalmente distribuídos. Para construir esse diagrama, precisamos obter os desvios das médias de cada indivíduo em relação à média geral, que podem ser obtidos por meio do comando a seguir:

```
res.medias.subj = rowMeans(heart.rateWide) - mean(heart.rate$hr)
```

Nesse comando, a função *rowMeans* calcula a média de cada linha do conjunto de dados *heart.rateWide*, fornecendo então a média da frequência cardíaca para cada indivíduo. Em seguida, essas médias são subtraídas da média geral da frequência cardíaca (*mean(heart.rate\$hr)*), gerando então a variável desejada (*res.medias.subj*).

O comando a seguir gera o gráfico de comparação de quantis da normal para a variável *res.medias.subj*, exibido na figura 18.46.

```
qqPlot(res.medias.subj, dist="norm", xlab = "quantis da normal",  
       ylab = "média dos indivíduos - média geral")
```

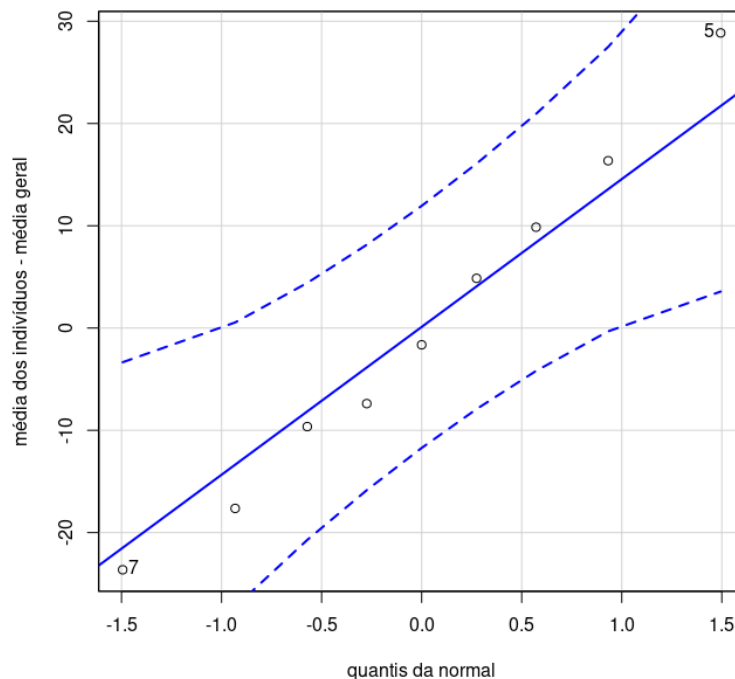


Figura 18.46: Gráfico de comparação dos quantis da normal dos desvios das médias dos indivíduos em relação à média geral.

O comando seguinte realiza o teste de Shapiro-Wilk para a variável *res.medias.subj*. O teste não rejeita a hipótese de normalidade ao nível de 5% ( $p = 0,98$ ), em concordância com a inspeção visual do gráfico de comparação de quantis da normal.

```
normalityTest(~ res.medias.subj, test="shapiro.test")
```

```
##  
##  Shapiro-Wilk normality test  
##  
## data:  res.medias.subj  
## W = 0.98376, p-value = 0.9808
```

### 18.4.6.2 Teste de Friedman no R

Para realizarmos o teste de Friedman, usando o plugin *RcmdrPlugin.EZR*, acessamos a opção:

Statistical analysis  $\Rightarrow$  Testes não paramétricos  $\Rightarrow$  Friedman test

Na tela seguinte, selecionamos as variáveis que indicam as medidas repetidas e selecionamos o método, ou métodos, que serão utilizados para realizar um teste para a comparação par a par das distribuições das medidas de frequência cardíaca em cada instante (figura 18.47).

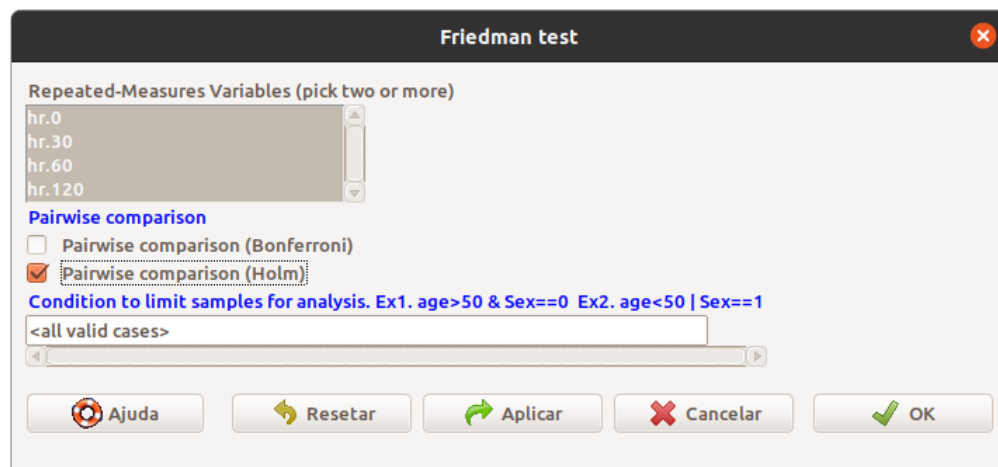


Figura 18.47: Seleção das variáveis que serão analisadas pelo teste de Friedman.

Os comandos executados são mostrados a seguir, seguido dos resultados.

Inicialmente, são mostradas as medianas da frequência cardíaca para cada instante de medida. O teste de Friedman rejeitou a hipótese nula de igualdade das distribuições das medidas de frequência cardíaca em cada instante, mas não teve poder estatístico para apontar diferenças entre os pares de distribuições. Nenhum valor de  $p$  para as comparações par a par foram

abaixo de 5%. Não são mostradas aqui as mensagens geradas pelo R de que a presença de empates ou zeros não permite o cálculo de valores de p exatos.

#### #####Friedman test#####

```
.Responses <- na.omit(with(heart.rateWide,  
                           cbind(hr.0, hr.30, hr.60, hr.120)))  
apply(.Responses, 2, median)
```

```
##   hr.0  hr.30  hr.60 hr.120  
##    96    92    86    92
```

```
res <- NULL  
(res <- friedman.test(.Responses))
```

```
##  
## Friedman rank sum test  
##  
## data:  .Responses  
## Friedman chi-squared = 8.5059, df = 3, p-value = 0.03664
```

```
cat(gettext(domain="R-RcmdrPlugin.EZR", "Friedman test"), " ",  
    gettext(domain="R-RcmdrPlugin.EZR", "p.value"), " = ",  
    signif(res$p.value, digits=3), "", sep="")
```

```
## Friedman test p.value = 0.0366
```

```
pairwise.friedman.test(.Responses, "heart.rateWide", p.adjust.method="holm")
```

```
##  
## Pairwise comparisons using Wilcoxon signed rank test  
##  
## data:  heart.rateWide  
##  
##      hr.0 hr.30 hr.60  
## hr.30  0.37 -      -  
## hr.60  0.21 0.86 -  
## hr.120 0.25 0.89 0.89  
##  
## P value adjustment method: holm
```

## 18.5 Outros tipos de análise de variância

Nas seções anteriores, foram apresentadas duas situações em que a análise de variância pode ser aplicada. A análise de variância pode, entretanto, ser utilizada em diversos outros contextos, por exemplo, com dois ou mais fatores ou variáveis independentes, com ou sem medidas repetidas. No caso de medidas repetidas, o modelo apresentado na seção 18.4 considerava que cada paciente possuía uma e somente uma medida em cada instante ou em cada nível do fator em estudo; entretanto outras situações podem ocorrer como, por exemplo, um paciente não possuir todas as medidas ou possuir mais de uma medida em um determinado nível do fator. Outras formas de amostragem também podem ser utilizadas na análise de variância. Kutner et al. (Kutner et al., 2005) apresentam uma visão bem ampla das diversas situações em que a análise de variância pode ser aplicada.

## 18.6 Exercícios

- 1) Com o conjunto de dados *birthwt* do pacote *MASS* (GPL-2 | GPL-3), faça as atividades abaixo.
  - a) Veja a ajuda para esse conjunto de dados.
  - b) Compare as médias da variável *bwt* (peso ao nascer) entre os três grupos de raças das mães. Obtenha o intervalo de confiança ao nível de 95% para a diferença das médias entre os três grupos, usando o método de Tukey.
  - c) Verifique as suposições para a realização da ANOVA para um fator.
  - d) Realize o teste de Kruskal-Wallis para a comparação do peso ao nascer entre os três grupos de raças das mães.
- 2) O conjunto de dados *WeightLoss* do pacote *carData* (GPL-2 | GPL-3) contém dados artificiais sobre perda de peso e auto-estima ao longo de três meses, para três grupos de indivíduos: Controle, Dieta e Dieta + Exercício.
  - a) Crie um subconjunto de dados de *WeightLoss*, chamado *wlDietEx*, com dados somente dos indivíduos que fizeram dieta + exercício, por meio do comando ao final do exercício 2 do capítulo anterior (capítulo 16).
  - b) Com o conjunto de dados *wlDietEx*, compare as médias de perdas de peso nos três meses. Utilize o nível de significância igual a 5%.
  - c) Faça o gráfico de interação entre as unidades de análise (pacientes) e o tempo de medição e comente o resultado.
  - d) Adapte o script da seção 18.4.6 para obter o intervalo de confiança ao nível de 85% para as diferenças de perda de peso entre as três medições. Comente os resultados.
  - e) Transforme o conjunto de dados *wlDietEx* para o formato longo, usando o primeiro comando ao final do exercício, gerando o conjunto de dados *wlDietExLong*. Para uma compreensão de como essa transformação é realizada, veja este [vídeo](#).
  - f) Use o segundo comando ao final do exercício para calcular os resíduos do modelo, faça o gráfico de comparação de quantis da normal para os resíduos e faça um teste de normalidade dos resíduos. Comente os resultados.



- g) Faça o diagrama de pontos dos resíduos por instantes de medida. Comente o gráfico.
- h) Converta a variável *mes* de *wlDietEx* para inteiro e faça o diagrama da sequência dos resíduos por indivíduos. Comente os gráficos.
- i) Use a terceira sequência de comandos ao final do exercício para verificar a normalidade dos efeitos principais devido aos indivíduos. Comente os resultados.
- j) Faça o teste de Friedman e comente os resultados.

**Convertendo wlDietEx para o formato longo:**

```
wlDietExLong <- reshapeW2L(wlDietEx, within="mes",
                           levels=list(mes=c("1", "2", "3")),
                           varying=list(wl=c("wl1", "wl2", "wl3")), id="id")
```

**Calculando os resíduos para o modelo:**

```
wlDietExLong$residuos = wlDietExLong$wl -
                        ave(wlDietExLong$wl, wlDietExLong$id) -
                        ave(wlDietExLong$wl, wlDietExLong$mes) +
                        ave(wlDietExLong$wl)
```

**Verificando a normalidade dos efeitos principais devido aos indivíduos:**

```
res.medias.subj = rowMeans(wlDietEx) - mean(wlDietExLong$wl)
qqPlot(res.medias.subj, dist="norm", xlab = "quantis da normal",
       ylab = "média dos indivíduos - média geral")
normalityTest(~ res.medias.subj, test="shapiro.test")
```

# Capítulo 19

## Regressão linear

### 19.1 Introdução

Os conteúdos desta seção e das seções 19.2 e 19.3 podem ser visualizados neste [vídeo](#).

A figura 19.1 mostra um diagrama de dispersão relacionando as variáveis *prestígio* (*prestige*) e nível educacional (*education*). Essas variáveis constam do conjunto de dados *Prestige*, disponível no pacote *carData* ([GPL-2](#) | [GPL-3](#)). Esse arquivo contém dados sobre a renda, a escolaridade, o prestígio, etc., relativos a 102 ocupações no Canadá. A variável *prestige* é o escore de prestígio de uma ocupação de acordo com o método de Pineo-Porter. O nível educacional foi medido em anos de escolaridade.

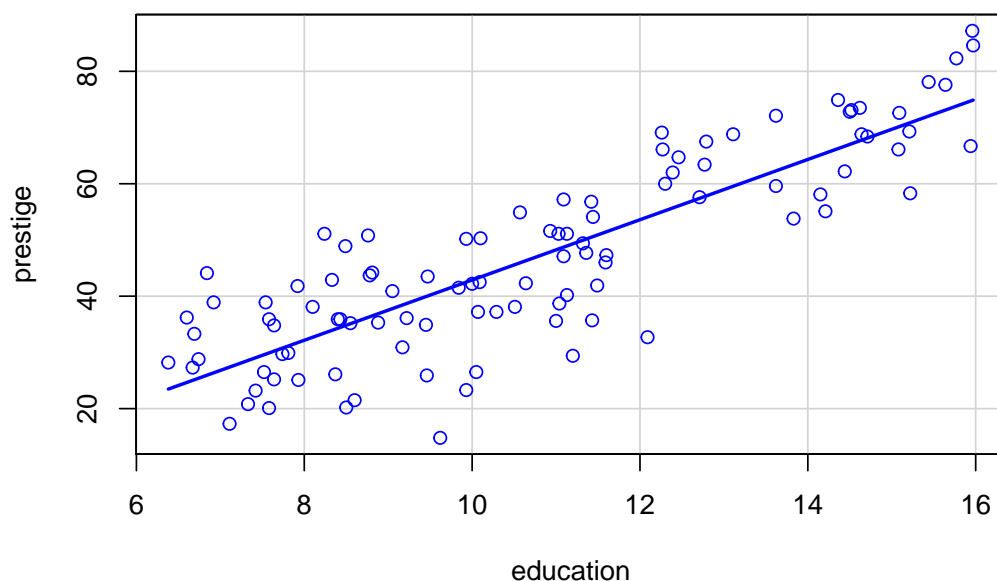


Figura 19.1: Diagrama de dispersão das variáveis *prestige* e *education*.

Cada ponto no gráfico corresponde aos valores do nível educacional (abscissa) e do escore do prestígio (ordenada) para cada ocupação. Observando o gráfico, podemos verificar que há

uma tendência de valores maiores do nível educacional estarem associados a valores maiores de prestígio, e que essa relação é aproximadamente linear, indicada pela reta azul. Essa reta é chamada reta de regressão.

Como outro exemplo, a figura 19.2 mostra um diagrama de dispersão da fração de ejeção do ventrículo esquerdo em função da relação neutrófilo-linfócito, obtido no estudo de Durmus et al. (Durmus et al., 2015). Cada ponto no gráfico corresponde aos valores da relação neutrófilo-linfócito (abscissa) e da fração de ejeção (ordenada) para cada paciente do estudo. Esse gráfico sugere que a fração de ejeção do ventrículo esquerdo tende a diminuir quando a relação neutrófilo-linfócito aumenta. Os autores apresentaram no gráfico a reta que melhor se ajusta a esses dados.

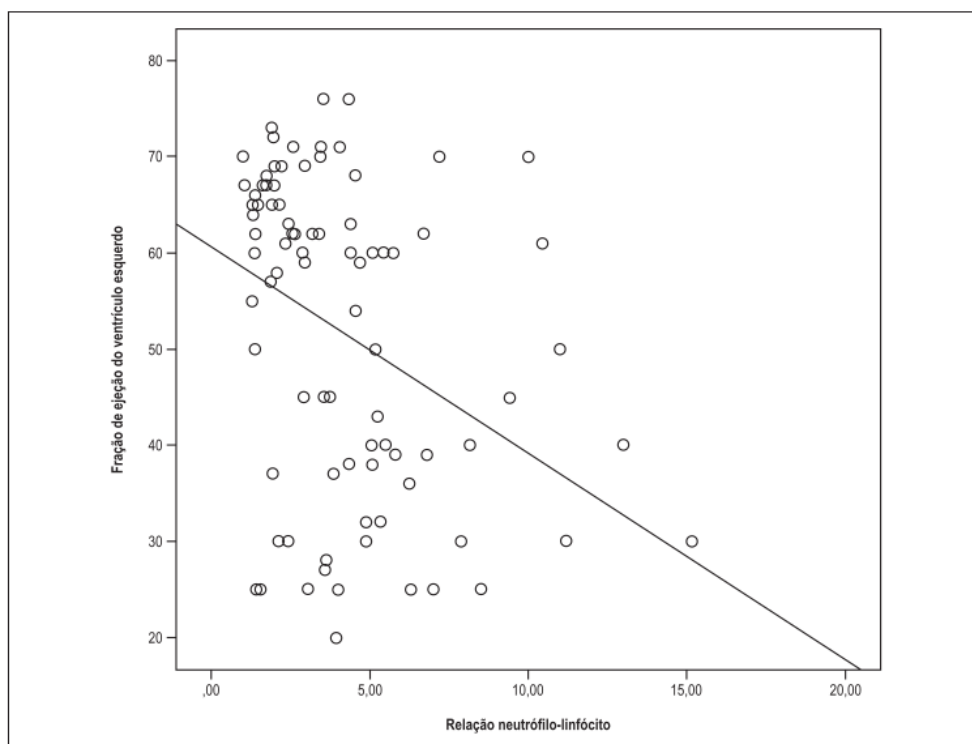


Figura 2 – Análise de correlação da relação neutrófilo-linfócito com a fração de ejeção do ventrículo esquerdo.

Figura 19.2: Diagrama de dispersão com a reta de regressão da fração de ejeção do ventrículo esquerdo em função da relação neutrófilo-linfócito. Fonte: (Durmus et al., 2015) ([CC BY](#)).

Este capítulo irá apresentar a análise de regressão simples, a qual permite a identificação dos coeficientes da reta que melhor se ajusta aos dados, e mostrar como calcular intervalos de confiança desses coeficientes e realizar previsões. Também serão apresentadas técnicas que permitem identificar quando um modelo de regressão é válido. A análise de regressão deve esse nome ao fato de que um dos pioneiros da área, Galton, observou que filhos de pais altos tendiam a ser mais baixos do que os pais e filhos de pais baixos tendiam a ser mais altos que eles, sugerindo que a altura dos filhos tendia a regredir aos valores da altura média da população, daí o nome reta de regressão à relação entre a altura dos pais e dos filhos e de regressão linear a essa técnica estatística.

## 19.2 Equação da reta

A figura 19.3 apresenta uma reta no plano XY, passando por dois pontos  $P_1(X_0, Y_0)$  e  $P_2(X_0 + \Delta X, Y_0 + \Delta Y)$ .

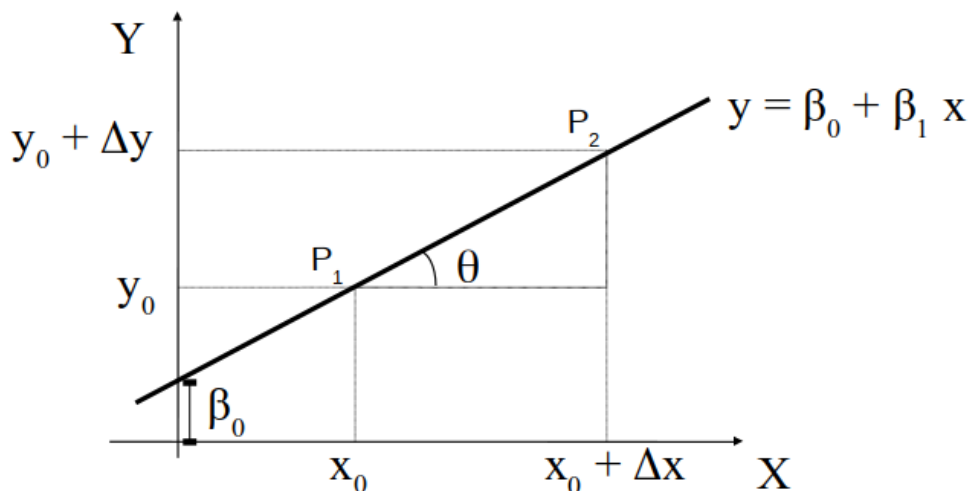


Figura 19.3: Gráfico e equação da reta.

Uma reta é dada pela expressão:

$$y = \beta_1 x + \beta_0,$$

onde  $\beta_1$  é a inclinação da reta e  $\beta_0$  é a interseção da reta com o eixo y.  $\beta_1$  e  $\beta_0$  são os coeficientes da reta. O valor de  $\beta_1$  é dado pela expressão:

$$\beta_1 = \frac{(Y_0 + \Delta Y - Y_0)}{(X_0 + \Delta X - X_0)} = \frac{\Delta Y}{\Delta X} \quad (19.1)$$

e indica o quanto os valores de y variam para cada unidade de aumento na variável x.

## 19.3 Método dos mínimos quadrados

Quando a relação entre duas variáveis numéricas pode ser aproximada por uma linha reta, como na figura 19.1, então devem ser encontrados os coeficientes da reta que melhor expresse essa relação linear. Como obter esses coeficientes? Observem a figura 19.4.

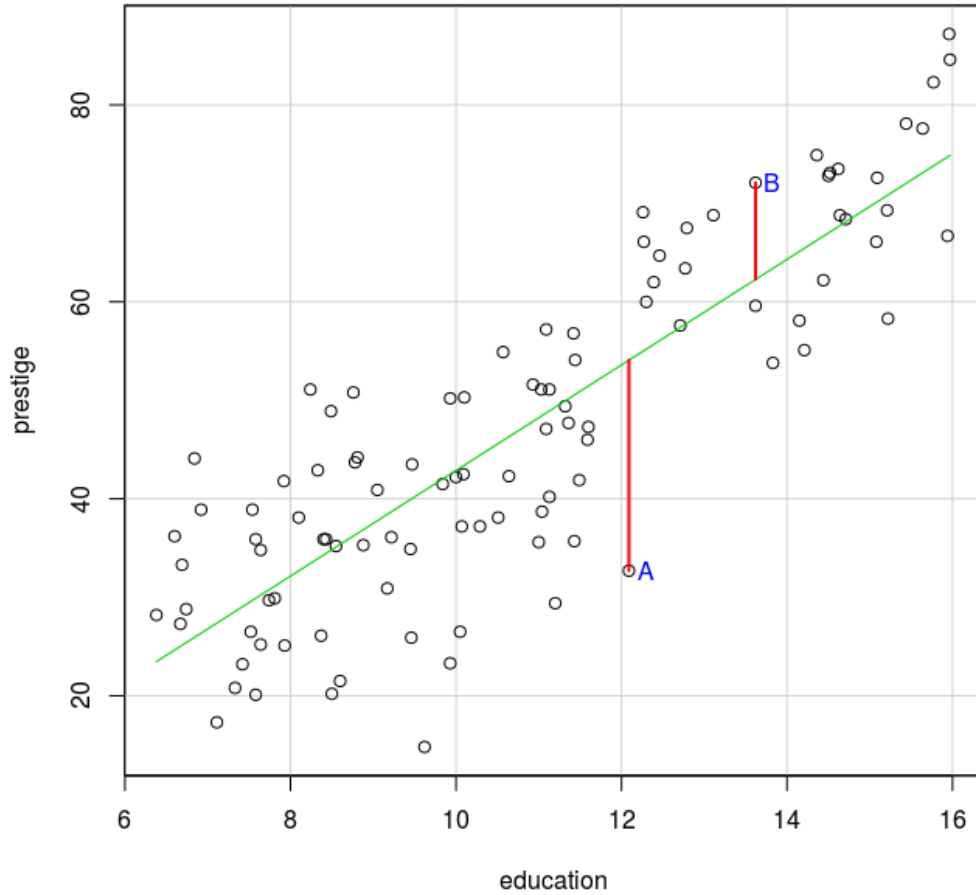


Figura 19.4: Distância vertical entre cada ponto do conjunto de dados à reta de regressão.

A reta que melhor se ajusta aos dados na figura 19.4 deve ser aquela que minimiza a distância vertical entre cada ponto e a reta, mas não basta minimizar a distância vertical de um único ponto, é preciso levar em conta todos os pontos. A figura 19.4 mostra dois pontos A e B e os segmentos que unem esses pontos à reta de regressão e que indicam a distância vertical de cada ponto à reta. Vamos chamar de  $X$  a variável *education* e de  $Y$  a variável *prestige*, e  $n$  o número de pontos no conjunto de dados ( $n = 102$  no conjunto de dados *Prestige*).

Vamos considerar que a reta de regressão é dada pela expressão:

$$Y = \beta_1 X + \beta_0 \quad (19.2)$$

Vamos chamar de  $\hat{y}_i$  o valor de  $y$  na reta de regressão correspondente ao ponto  $x_i$ . Vamos definir a distância vertical do ponto  $(x_i, y_i)$  ( $i = 1, \dots, n$ ) à reta de regressão como o valor absoluto da diferença entre o valor  $y_i$  e o valor  $\hat{y}_i$  do ponto na reta correspondente a  $x_i$ :

$$|y_i - \hat{y}_i|, \text{ onde } \hat{y}_i = \beta_0 + \beta_1 x_i$$

O método mais usado para determinar os coeficientes da reta de regressão é calcular esses coeficientes de tal modo que a soma do quadrado das distâncias de cada ponto à reta de regressão seja mínima, ou seja, calcular os coeficientes  $b_0$  e  $b_1$  de tal modo que a expressão abaixo seja mínima:

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n [y_i - (b_0 + b_1 x_i)]^2 \quad (19.3)$$

Por procurar minimizar o quadrado das distâncias, esse método é conhecido como **método dos mínimos quadrados**. A aplicação [Determinação dos coeficientes da reta de regressão](#) ilustra essa ideia. A figura 19.5 mostra a página inicial dessa aplicação.

Determinação dos coeficientes da reta de regressão

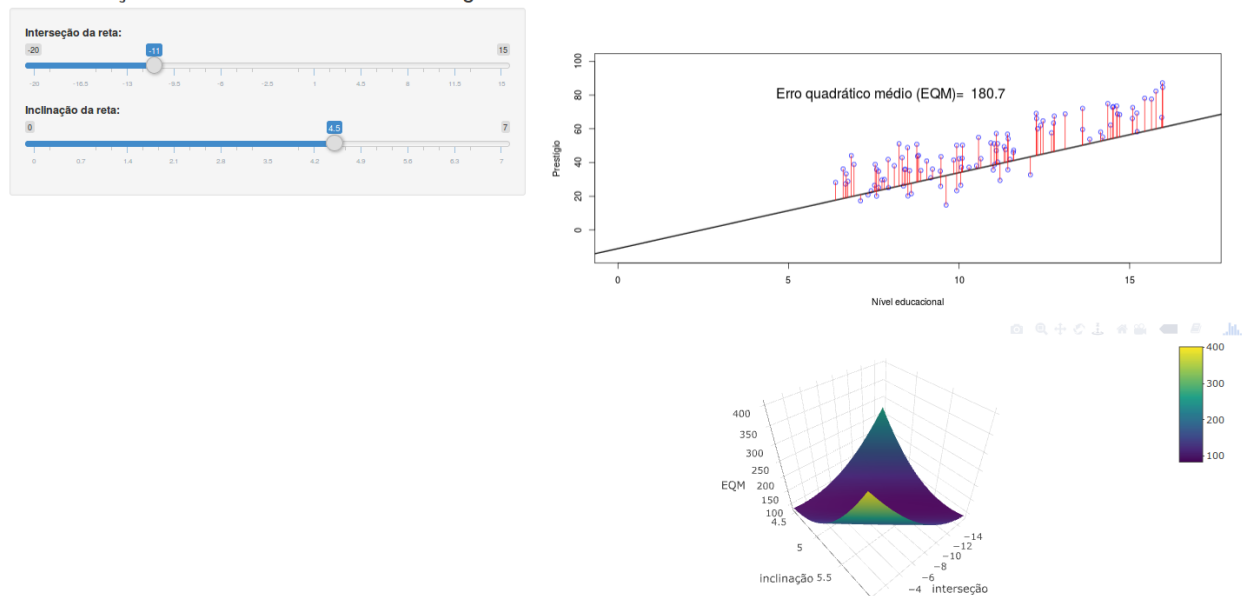


Figura 19.5: Aplicação que ilustra o princípio do método dos mínimos quadrados.

No painel à esquerda, o usuário pode variar a inclinação (coeficiente  $b_1$ ) e a interseção da reta (coeficiente  $b_0$ ), a qual é exibida no gráfico superior do painel principal, juntamente com as distâncias verticais de cada ponto à reta. À medida que o usuário alterar os coeficientes da reta, o gráfico mostra a nova reta e o valor do erro quadrático médio que é a expressão (19.3) dividida pelo número de pontos - 2. Minimizar o erro quadrático médio é equivalente a minimizar a expressão (19.3). Se o usuário alterar os valores da interseção e inclinação da reta, verá que o valor do erro quadrático médio vai variando. Ao escolher valores da interseção próximos a -10,732 e valores da inclinação próximos a 5,361, os valores do erro quadrático médio são próximos ao mínimo valor do erro quadrático médio. O mesmo pode ser observado no gráfico da parte inferior do painel principal (figura 19.6).

O gráfico da figura 19.6 mostra uma superfície que representa o valor do erro quadrático médio (EQM) para cada combinação de valores dos coeficientes  $b_0$  (interseção) e  $b_1$  (inclinação) da reta de regressão. Ao clicarmos com o *mouse* sobre a superfície, o gráfico mostra os correspondentes valores de  $b_0$ ,  $b_1$  e do EQM. Ao movermos o *mouse*, os pontos sobre a superfície vão se alterando. Ao pressionarmos e arrastarmos o *mouse*, podemos girar o gráfico em torno dos eixos coordenados. Ao escolhermos certo ângulo de visualização do gráfico e ao movermos o *mouse*, podemos observar os valores de  $b_0$  e  $b_1$  correspondentes aos menores valores do EQM. Diversos ícones na parte superior direita do gráfico permitem a realização de diversas operações como *zoom*, baixar a imagem no formato png, mover o gráfico, etc.

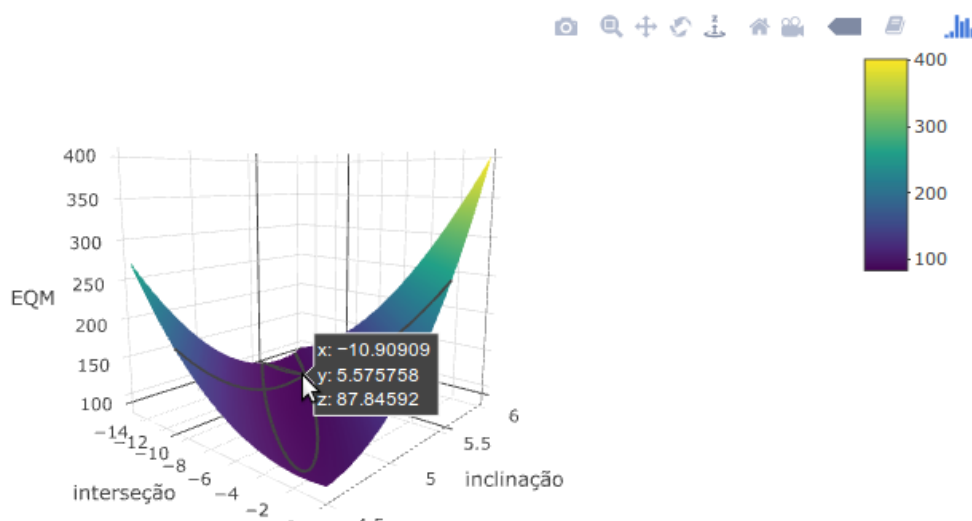


Figura 19.6: Erro quadrático médio em função dos valores dos coeficientes  $b_0$  (interseção) e  $b_1$  (inclinação) da reta de regressão.

Matematicamente existem expressões obtidas a partir das derivadas da função do erro quadrático (19.3) que permitem a obtenção dos valores de  $b_0$  e  $b_1$  que minimizam o erro quadrático, sem a necessidade de usarmos o método de tentativa e erro. Essas expressões são dadas por (Costa Neto, 1977):

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (19.4)$$

$$b_0 = \bar{y} - b_1\bar{x} \quad (19.5)$$

## 19.4 Modelo de regressão linear

O conteúdo desta seção pode ser visualizado neste [vídeo](#).

O modelo de regressão linear simples pode ser escrito como:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad (19.6)$$

onde:

$i$  - indica cada um dos pares de valores  $(x_i, y_i)$  ( $i = 1, 2, \dots, n$ );

$X_i$  -  $i$ -ésimo valor da variável independente  $X$ ;

$\beta_0$  - corresponde à interseção da reta de regressão com o eixo  $Y$ ;

$\beta_1$  - inclinação da reta de regressão;

$\epsilon_i$  - desvio do valor observado  $Y_i$  em relação ao valor previsto pela reta de regressão no ponto  $X_i$ . Os erros  $\epsilon_i$  supostamente seguem uma distribuição normal com média 0 e variância  $\sigma^2$ .

Os coeficientes  $\beta_0$  e  $\beta_1$  da reta de regressão podem ser estimados por meio do método dos mínimos quadrados e as estimativas serão representadas como  $b_0$  e  $b_1$ , respectivamente. Então os valores estimados de  $Y$  para cada valor da variável  $X$  serão dados por:

$$\hat{y}_i = b_0 + b_1 x_i$$

A partir do modelo de regressão, podemos:

- 1) realizar um teste de hipótese de que a inclinação da reta de regressão é nula, ou seja, não existe uma relação linear entre as variáveis  $X$  e  $Y$ ;
- 2) estimar os intervalos de confiança para os parâmetros  $\beta_0$  e  $\beta_1$  da reta de regressão;
- 3) verificar a validade do modelo de regressão;
- 4) estimar o intervalo de confiança para o valor esperado de  $Y$  para um dado valor de  $X$ ;
- 5) estimar o intervalo de confiança para o valor de  $Y$  para um dado valor de  $X$ .

As 3 primeiras questões serão tratadas a seguir.

### 19.4.1 Teste de hipótese

Para verificar se não existe uma relação linear entre as variáveis  $X$  e  $Y$  em um modelo de regressão linear, o seguinte teste de hipótese pode ser realizado:

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$



Para verificar a hipótese  $H_0$ , um raciocínio semelhante ao realizado para a análise de variância para um fator (capítulo 18) pode ser empregado, particionando a soma dos quadrados da variável Y em relação à média geral de Y.

A média aritmética de todos os valores de Y (média geral) é dada por:

$$\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n} \quad (19.7)$$

A média aritmética de todos os valores de X é dada por:

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} \quad (19.8)$$

### Partição da soma dos resíduos em relação à média geral

A expressão a seguir mostra que o desvio de cada valor da variável aleatória Y ( $Y_i$ ) em relação à média geral de Y ( $\bar{Y}$ ) é igual ao desvio de Y em relação ao valor previsto pela reta de regressão ( $\hat{Y}_i$ ) somado ao desvio do valor previsto pela reta de regressão em relação à média geral :

$$\underbrace{Y_i - \bar{Y}}_{\text{desvio em relação à média geral}} = \underbrace{(Y_i - \hat{Y}_i)}_{\text{desvio em relação à reta de regressão}} + \underbrace{(\hat{Y}_i - \bar{Y})}_{\text{desvio da reta de regressão em relação à média geral}}$$

A figura 19.7 ilustra essa fatoração. Os três gráficos mostram o desvio de cada valor da variável Y em relação ao valor previsto pela reta de regressão (gráfico superior à esquerda), o desvio de cada valor previsto pela reta de regressão em relação à média geral de Y (gráfico superior à direita) e o desvio de cada valor de Y em relação à média geral de Y (gráfico inferior à esquerda).

Elevando cada desvio ao quadrado e somando os quadrados de todos os desvios, pode-se mostrar que o resultado é a expressão abaixo:

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (19.9)$$

Essa expressão pode ser escrita como:

$$SQTot = SQR + SQE$$

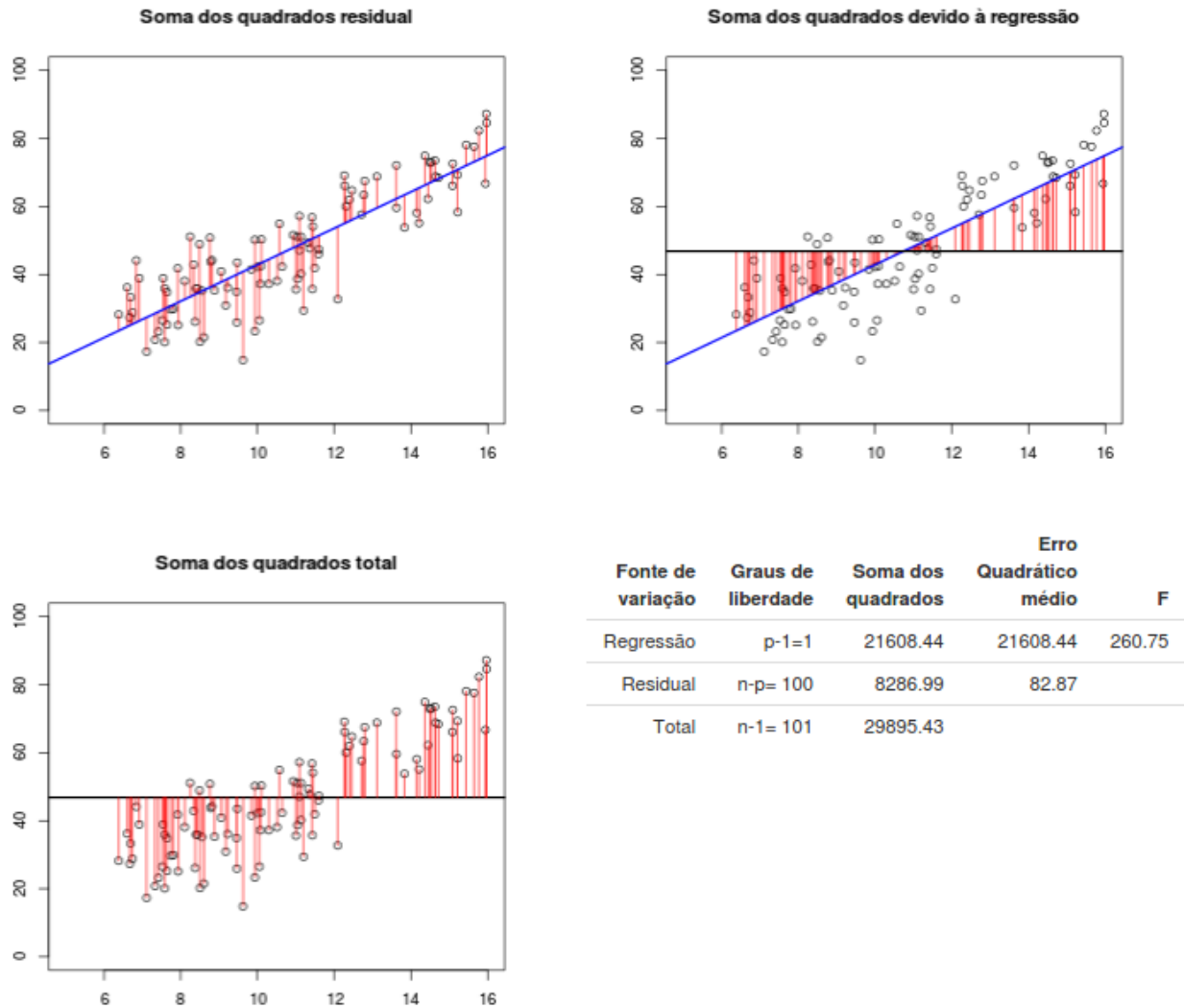


Figura 19.7: Análise de variância aplicada ao modelo de regressão linear.  $p = 2$  no modelo (19.6).

onde:

$$SQTot = \text{Soma total dos quadrados} = \sum_{i=1}^n (Y_i - \bar{Y})^2 \quad (19.10)$$

com  $n-1$  graus de liberdade. Essa soma é mostrada na terceira linha da tabela da figura 19.7.

$$SQE = \text{Soma dos quadrados dos erros} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (19.11)$$

com  $n-2$  graus de liberdade. Essa soma é mostrada na segunda linha da tabela da figura 19.7.

$$SQR = \text{Soma dos quadrados devido à regressão} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \quad (19.12)$$

com 1 grau de liberdade. Essa soma é mostrada na primeira linha da tabela da figura 19.7.

O erro quadrático médio (EQM) é obtido dividindo-se SQE pelo correspondente número de graus de liberdade (segunda linha da tabela da figura 19.7):

$$EQM = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - 2} \quad (19.13)$$

O erro quadrático médio da regressão (EQMR) é obtido dividindo-se SQR pelo correspondente grau de liberdade (primeira linha da tabela da figura 19.7):

$$EQMR = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{1} \quad (19.14)$$

Pode-se mostrar que o valor esperado do erro quadrático médio é igual à variância do erro no modelo de regressão:

$$E[EQM] = \sigma^2 \quad (19.15)$$

e que o valor esperado do erro quadrático médio da regressão é igual à expressão abaixo:

$$E[EQMR] = \sigma^2 + \beta_1^2 \sum_{i=1}^n (X_i - \bar{X})^2 \quad (19.16)$$

## Teste F

Quando a hipótese nula é verdadeira, tanto o erro quadrático médio quanto o erro quadrático médio da regressão são estimadores não tendenciosos da variância do erro do modelo de regressão. Quando a hipótese nula não é verdadeira, o valor esperado do erro quadrático médio da regressão é maior do que a variância do erro, aumentando à medida que o valor absoluto da inclinação da reta de regressão aumenta.

Assim a divisão do valor do erro quadrático médio da regressão (EQMR) pelo erro quadrático médio (EQM) dá uma indicação de quanto a hipótese nula é compatível com os dados. Vamos representar o valor dessa divisão por  $F^*$ :

$$F^* = \frac{EQMR}{EQM} \quad (19.17)$$

Pode-se mostrar que a razão EQMR/EQM, se a hipótese nula for verdadeira, segue a distribuição  $F(1, n-2)$ . Dado um nível de significância  $\alpha$ , quando o valor de  $F^*$ , obtido da expressão (19.17), for maior que o quantil  $1 - \alpha$  da distribuição  $F(1, n-2)$ , então a hipótese nula é rejeitada.

### 19.4.2 Intervalos de confiança para os coeficientes de regressão

Os conteúdos desta seção e da seção 19.4.3 podem ser visualizados neste [vídeo](#).

Vamos supor que uma amostra contendo  $n$  pares de valores  $(x, y)$  sejam extraídas de uma população. Podemos imaginar duas situações distintas:

- 1)  $X$  e  $Y$  possuem uma distribuição conjunta normal e  $n$  unidades de observação sejam extraídas aleatoriamente de uma população e os valores de  $X$  e  $Y$  para cada unidade são medidos;
- 2)  $n$  valores fixos de  $X$  são estabelecidos a priori e, para cada valor de  $X$ , uma unidade de observação é extraída aleatoriamente da população de indivíduos com esse valor de  $X$  e o correspondente valor de  $Y$  é medido.

Em ambos os casos, vamos supor que a relação entre  $X$  e  $Y$  possa ser escrita da seguinte forma:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, 2, \dots, n \quad e \quad \epsilon_i \sim N(0, \sigma^2) \quad (19.18)$$

Como estamos lidando com amostras extraídas de uma população, é de se esperar que, se ajustarmos uma reta de regressão a cada amostra obtida, as estimativas dos coeficientes da reta de regressão irão variar de amostra para amostra.

A aplicação [Variabilidade dos coeficientes da reta de regressão](#) mostra a variabilidade das estimativas dos coeficientes de regressão (figura 19.8). Para isso, partindo de uma relação linear entre as variáveis  $Y$  e  $X$ , dada pela equação (19.18), onde  $\beta_0 = 1$ ,  $\beta_1 = 2$  e  $\sigma^2 = 15$ , obtemos sucessivas amostras de 21 pares  $(x, y)$  onde  $x = 10, 11, 12, \dots, 30$  e os valores de  $y$  são obtidos a partir da equação (19.18). Os valores de  $X$  são fixos, mas os valores de  $Y$  correspondentes a cada valor de  $X$  irão variar de amostra para amostra, devido ao componente aleatório  $\epsilon_i$ .

A aplicação mostra inicialmente duas simulações (ou duas amostras de 21 pares  $(x, y)$ ). Para cada amostra, uma reta de regressão é ajustada aos dados pelo método dos mínimos quadrados (mostrada no painel inferior à direita), gerando as estimativas  $b_1$  e  $b_0$  de  $\beta_1$  e  $\beta_0$ , dadas pela expressões (19.4) e (19.5), respectivamente. É possível observar que as duas retas de regressão correspondentes às duas simulações são diferentes (mostradas no painel superior à direita).

### Variabilidade dos coeficientes da reta de regressão

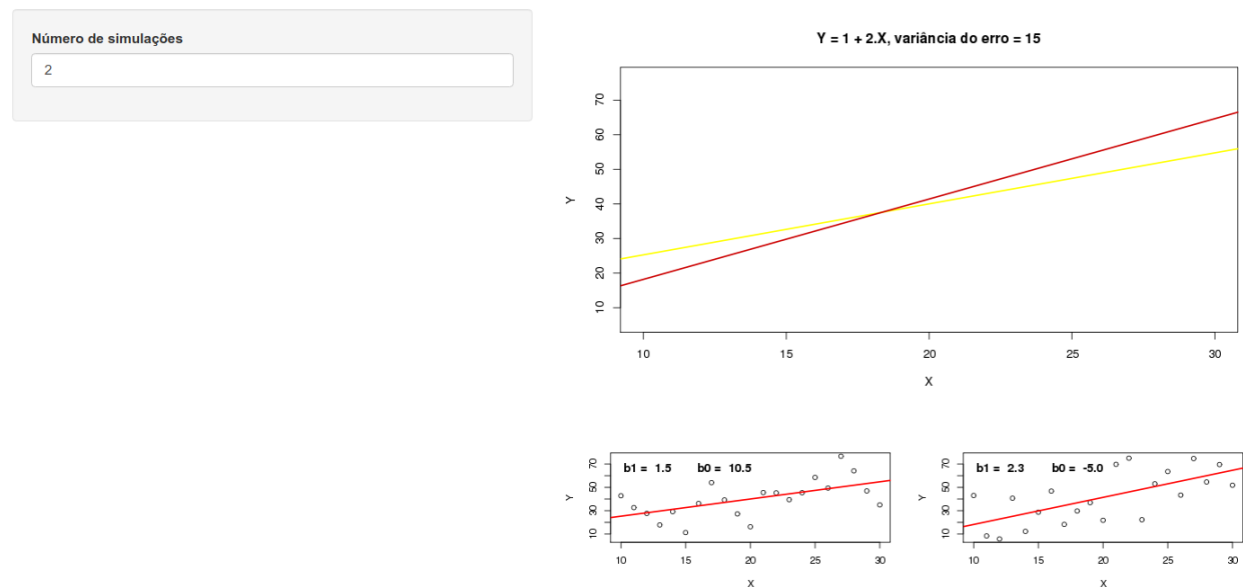


Figura 19.8: Aplicação para ilustrar a variabilidade dos coeficientes da reta de regressão. A parte superior mostra as retas de regressão correspondentes a cada simulação. Na parte inferior, são mostradas as retas de regressão correspondentes às primeiras simulações (máximo de 4). O usuário seleciona o número de simulações, até um máximo de 1000.

Aumentando o número de simulações, iremos obter um perfil das retas de regressão como mostra a figura 19.9. É possível observar uma variabilidade das estimativas dos coeficientes da reta de regressão e uma maior variação dos valores esperados de  $Y$  à medida que nos afastamos do valor médio de  $X$ .

Nas seções seguintes, serão apresentados os intervalos de confiança para os coeficientes da reta de regressão.

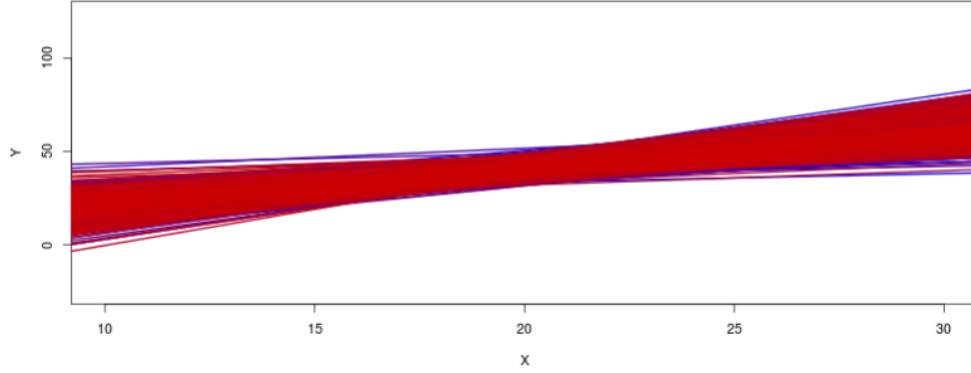


Figura 19.9: Realizando 1000 simulações na aplicação da figura 19.8. É possível observar uma maior variação dos valores esperados de Y à medida que nos afastamos do valor médio de X.

#### 19.4.2.1 Intervalo de confiança para a inclinação da reta de regressão

É possível demonstrar (Kutner et al., 2005) que o valor esperado e a variância da estimativa da inclinação da reta de regressão pelo método dos mínimos quadrados são dados por:

$$E[b_1] = \beta_1$$

$$var[b_1] = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

A variância  $\sigma^2$  pode ser estimada pelo EQM (equação (19.13)), de modo que a variância de  $b_1$  pode ser estimada por:

$$s^2[b_1] = \frac{EQM}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

A estatística  $\frac{b_1 - \beta_1}{s[b_1]}$  segue uma distribuição t de Student, com n-2 graus de liberdade, de modo que o intervalo de confiança de  $\beta_1$  com nível de confiança  $1 - \alpha$  pode ser calculado a partir da expressão abaixo:

$$P \left[ -t_{n-2, 1-\alpha/2} \leq \frac{b_1 - \beta_1}{s[b_1]} \leq t_{n-2, 1-\alpha/2} \right] = 1 - \alpha$$

Portanto o intervalo de confiança de  $\beta_1$  com nível de confiança  $1 - \alpha$  é dado por:

$$b_1 - t_{n-2, 1-\alpha/2} s[b_1] \leq \beta_1 \leq b_1 + t_{n-2, 1-\alpha/2} s[b_1]$$

### 19.4.2.2 Intervalo de confiança para a interseção da reta de regressão com o eixo Y

O valor esperado e a variância da estimativa da interseção da reta de regressão com o eixo Y pelo método dos mínimos quadrados são dados por:

$$E[b_0] = \beta_0$$

e

$$var[b_0] = \sigma^2 \left[ \frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]$$

A variância de  $b_0$  pode ser estimada por:

$$s^2[b_0] = EQM \left[ \frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]$$

A estatística  $\frac{b_0 - \beta_0}{s[b_0]}$  segue uma distribuição t de Student, com n-2 graus de liberdade, de modo que o intervalo de confiança de  $\beta_0$  com nível de confiança  $1 - \alpha$  pode ser calculado a partir da expressão abaixo:

$$P \left[ -t_{n-2, 1-\alpha/2} \leq \frac{b_0 - \beta_0}{s[b_0]} \leq t_{n-2, 1-\alpha/2} \right] = 1 - \alpha$$

Portanto o intervalo de confiança de  $\beta_0$  com nível de confiança  $1 - \alpha$  é dado por:

$$b_0 - t_{n-2, 1-\alpha/2} s[b_0] \leq \beta_0 \leq b_0 + t_{n-2, 1-\alpha/2} s[b_0]$$

### 19.4.3 Coeficiente de determinação

O coeficiente de determinação é uma medida que indica o quanto os pares de pontos (x, y) estão próximos da reta de regressão. Ele é representado por  $R^2$  e é calculado pela expressão:

$$R^2 = \frac{SQTot - SQE}{SQTot} = \frac{SQR}{SQTot} = 1 - \frac{SQE}{SQTot}$$

$0 \leq R^2 \leq 1$  e pode ser interpretado como a proporção da variação de Y que está associada à variável X. Quando os pontos estão alinhados, o valor de  $R^2$  é igual a 1. Quando não há nenhuma relação linear entre X e Y (inclinação da reta igual a 0),  $R^2$  é igual a 0. Em geral  $R^2$  estará entre esses dois limites. Quanto mais próximos de uma linha reta estão os pontos, mais próximo de 1 será o valor de  $R^2$ .

### 19.4.4 Validação do modelo de regressão linear

Os conteúdos desta seção e da seção 19.5 podem ser visualizados neste [vídeo](#).

O modelo de regressão linear possui as seguintes suposições:

- 1) a função de regressão é linear;
- 2) os erros são independentes e seguem uma distribuição normal com média igual a 0 e variância constante,  $\sigma^2$ .

A análise dos resíduos é um dos principais recursos para verificarmos a validade de um modelo de regressão linear. O resíduo  $e_i$  é igual à diferença entre o valor observado e o valor previsto pela reta de regressão:

$$e_i = Y_i - \hat{Y}_i$$

Os resíduos podem ser considerados como erros observados e não como os erros reais que seriam dados pela expressão  $\epsilon_i = Y_i - E[Y_i]$ , onde  $E[Y_i] = \beta_0 + \beta_1 x_i$

Se o modelo de regressão linear for adequado, os resíduos devem refletir as suposições do modelo de regressão linear.

Frequentemente a análise de resíduos é realizada com os resíduos padronizados, que são obtidos pela divisão de cada resíduo pela seu desvio padrão,  $\sigma$ , que é estimado pela raiz quadrada do EQM. Logo:

$$\text{resíduo padronizado} = \frac{e_i}{\sqrt{EQM}}$$

Diagramas de dispersão dos resíduos (ou resíduos padronizados) em relação à variável independente ou valores previstos pela reta de regressão podem mostrar visualmente violações das suposições do modelo de regressão linear.

A figura 19.10 mostra um diagrama de dispersão dos resíduos padronizados x valores previstos pela reta de regressão para cada um dos pontos da amostra do conjunto de dados *Prestige*.

Nesse diagrama, podemos observar que os resíduos não parecem seguir algum padrão que pudesse indicar um outro tipo de relacionamento entre a variável dependente e a variável independente, ou alguma dependência entre os resíduos. Também não há evidência de que a dispersão dos resíduos se altera para os diversos valores esperados pela reta de regressão. Um diagrama de comparação de quantis (*normal probability plot*) nos permitirá verificar se os resíduos padronizados seguem uma distribuição normal ou não.



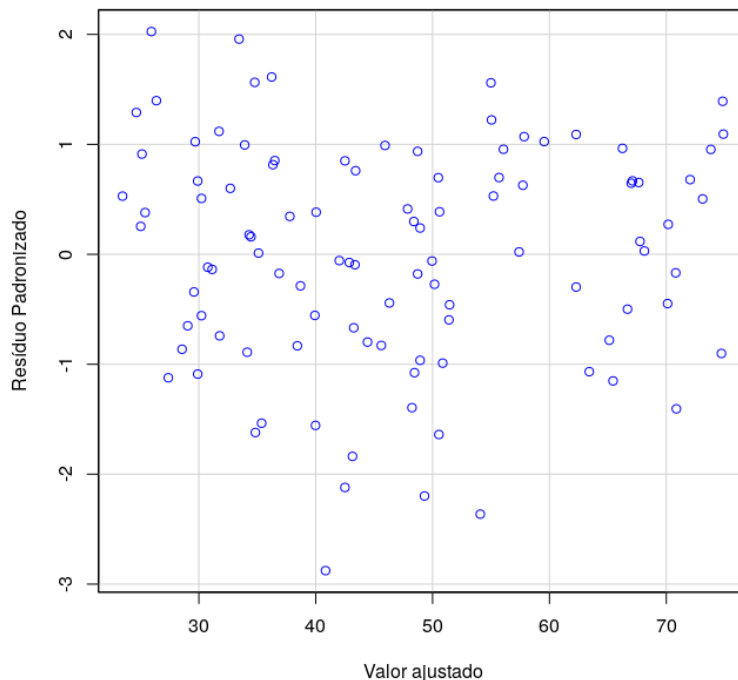


Figura 19.10: Diagrama de dispersão dos resíduos padronizados x valores previstos pela reta de regressão para cada um dos pontos da amostra do conjunto de dados *Prestige*.

Quando os resíduos seguem uma distribuição normal, espera-se que 5% dos resíduos padronizados sejam menores que -1,96 ou maiores do que 1,96. No gráfico acima, verifica-se que alguns resíduos padronizados são menores do que -2 e um resíduo é maior do que 2, mas a grande maioria dos resíduos padronizados estão situados entre -1,96 e 1,96, não indicando que haja pontos que estejam muito fora dos valores previstos pelo modelo de regressão linear.

Outros diagnósticos para o modelo de regressão linear como identificação de casos influentes e omissão de variáveis importantes estão além do escopo deste texto.

## 19.5 Análise de Regressão no *R Commander*

Vamos realizar a análise da relação entre as variáveis *prestige* e *education* do conjunto de dados *Prestige*. Após carregarmos o conjunto de dados, selecionamos no menu do *R Commander* a opção:

Estatísticas  $\Rightarrow$  Ajuste de Modelos  $\Rightarrow$  Regressão Linear

Na tela *Regressão Linear* (figura 19.11), selecionamos a variável resposta ou variável dependente (*prestige*) e a variável explicativa ou independente (*education*). É preciso atribuir um nome ao modelo que será construído (seta verde na figura 19.11). Esse nome será usado para referenciar os componentes desse modelo mais adiante.



Figura 19.11: Configuração para a análise de regressão linear de *prestige* x *education* no conjunto de dados *Prestige*.

Ao clicarmos no botão OK, os comandos abaixo serão executados, com os resultados apresentados a seguir.

```
RegModel.1 <- lm(prestige~education, data=Prestige)
summary(RegModel.1)
```

```
##
## Call:
## lm(formula = prestige ~ education, data = Prestige)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.0397  -6.5228   0.6611   6.7430  18.1636
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -10.732      3.677  -2.919  0.00434 **
## education      5.361      0.332  16.148 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.103 on 100 degrees of freedom
## Multiple R-squared:  0.7228, Adjusted R-squared:  0.72
## F-statistic: 260.8 on 1 and 100 DF, p-value: < 2.2e-16
```

Inicialmente, o resumo do modelo mostra a função usada para gerar o modelo de regressão:

```
lm(formula = prestige ~ education, data = Prestige)
```

*lm* significa *linear model* e é o nome da função que deve ser usada para criar um modelo de

regressão linear. O primeiro argumento da função *lm* é *formula*, a qual deve iniciar com uma variável dependente seguida do sinal  $\sim$  e da variável independente. O outro argumento, *data*, define o conjunto de dados a ser utilizado.

Em seguida, são mostrados alguns quantis dos resíduos do modelo, seguidos dos coeficientes da reta de regressão, onde são mostrados a estimativa pelo método dos mínimos quadrados, o erro padrão, o valor da estatística t e o valor de p para cada um dos dois coeficientes. Verificamos que os dois coeficientes são estatisticamente diferentes de zero. Na parte final, o resumo do modelo mostra o valor de  $R^2$ , o valor da estatística F para o modelo de regressão e o valor de p para o modelo.

Para obtermos os intervalos de confiança para os coeficientes do modelo, utilizamos a seguinte opção do menu:

Modelos  $\Rightarrow$  Intervalos de confiança

Em seguida, digitamos o nível de confiança desejado e clicamos em OK (figura 19.12).

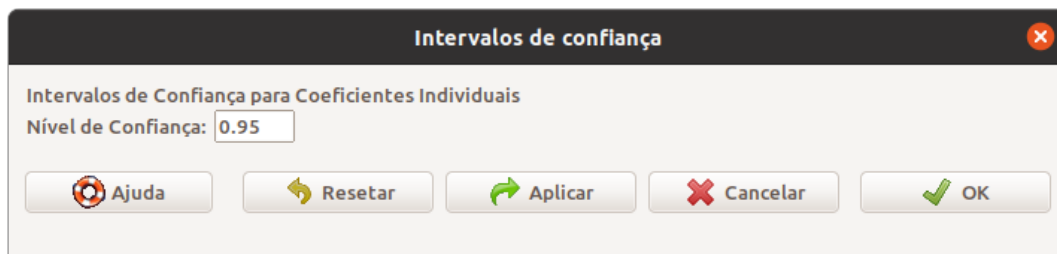


Figura 19.12: Especificação do nível de confiança dos intervalos de confiança dos coeficientes do modelo de regressão linear.

O comando a seguir é executado e gera os intervalos de confiança para a interseção e a inclinação do modelo de regressão linear.

```
Confint(RegModel.1, level=0.95)
```

```
##           Estimate      2.5 %    97.5 %
## (Intercept) -10.731982 -18.027220 -3.436744
## education    5.360878   4.702223  6.019533
```

Para obtermos alguns gráficos diagnósticos do modelo gerado, selecionamos a opção:

Modelos  $\Rightarrow$  Gráficos  $\Rightarrow$  Diagnósticos Gráficos Básicos

Os gráficos gerados são mostrados na figura 19.13.

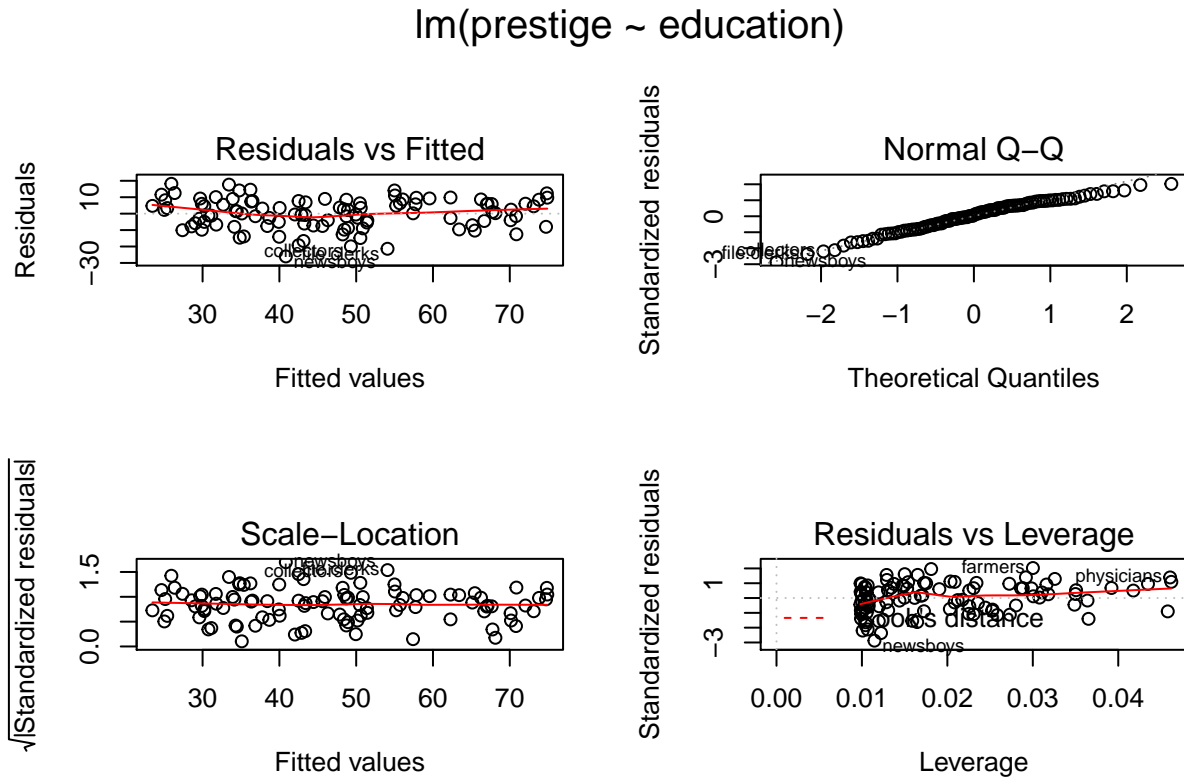


Figura 19.13: Diagnósticos para o modelo de análise de regressão linear de *prestige* x *education* para o conjunto de dados *Prestige*.

O primeiro gráfico na parte superior à esquerda mostra os resíduos x valores ajustados pela reta de regressão. O gráfico da parte superior à direita é o gráfico de comparação de quantis (*normal probability plot*) dos resíduos padronizados e o gráfico da parte inferior à esquerda é o diagrama de dispersão da raiz quadrada dos resíduos padronizados x valores ajustados pela reta de regressão. Os três gráficos não indicam desvios importantes da hipótese de normalidade, indicam que a variância é uniforme para cada valor ajustado e que os resíduos são independentes. A explicação para o último gráfico está além do escopo deste texto.

## 19.6 Coeficiente de correlação linear

Os conteúdos desta seção e de suas subseções podem ser visualizados neste [vídeo](#).

Voltando à figura 19.1 que mostra o diagrama de dispersão relacionando as variáveis *prestígio* (*prestige*) e nível educacional (*education*), uma medida do relacionamento linear entre as duas variáveis é dada pelo coeficiente de correlação linear de Pearson, que é calculado pela fórmula:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (19.19)$$

O coeficiente de correlação pode assumir um valor entre -1 e +1. Se maiores valores de uma variável  $x$ , em geral, correspondem a valores maiores de outra variável  $y$  e valores menores de  $x$ , em geral, correspondem a menores valores de  $y$ , então o coeficiente de correlação é maior que zero (figura 19.14a), sendo igual a 1 quando os valores de  $x$  e  $y$  estão perfeitamente alinhados em uma reta com inclinação positiva. Se maiores valores de  $x$ , em geral, correspondem a valores menores de  $y$  e valores menores de  $x$ , em geral, correspondem a maiores valores de  $y$ , então o coeficiente de correlação é menor que zero (figura 19.14b), sendo igual a -1 quando os valores de  $x$  e  $y$  estão perfeitamente alinhados em uma reta com inclinação negativa. Quando não há uma tendência de valores de  $y$  variarem de acordo com os valores de  $x$ , o coeficiente de correlação é zero (figura 19.14c).

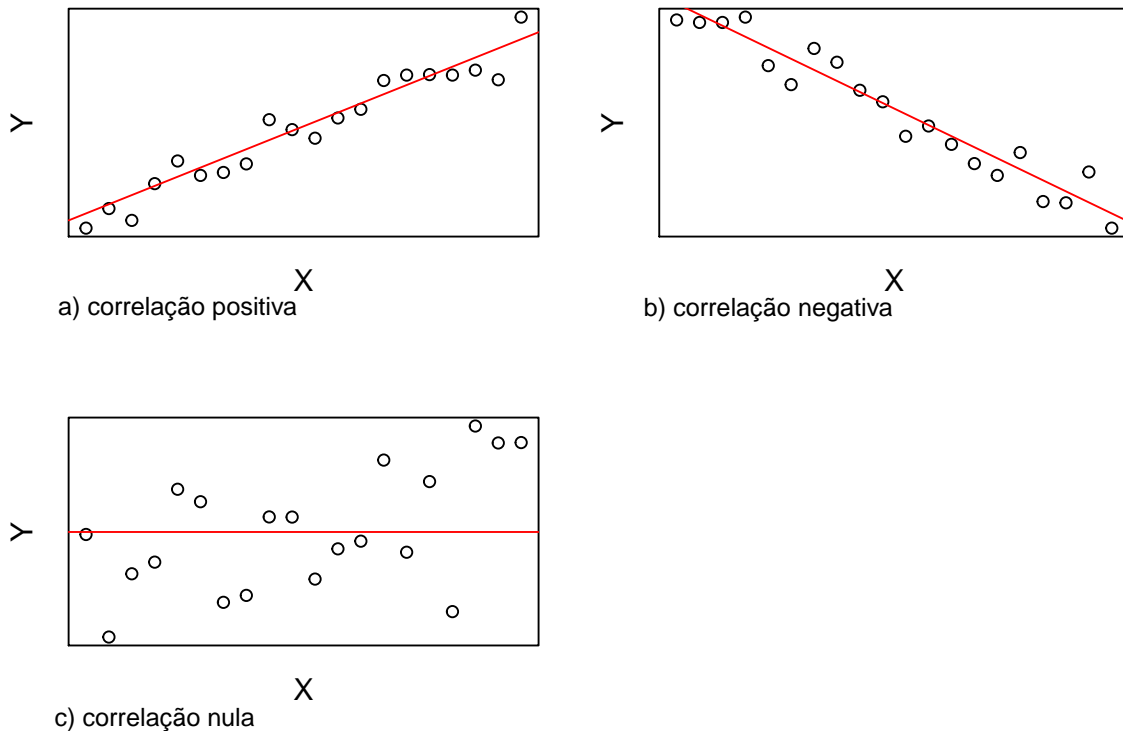


Figura 19.14: Relações lineares entre as variáveis  $x$  e  $y$ : a) correlação linear positiva, b) correlação linear negativa, c) correlação nula.

O quadrado do coeficiente de correlação linear é igual ao coeficiente de determinação do modelo de regressão linear simples.

### 19.6.1 Teste de hipótese bilateral e intervalos de confiança para o coeficiente de correlação

O teste de hipótese bilateral e o cálculo do intervalo de confiança para o coeficiente de correlação linear partem da suposição de que os valores das variáveis correlacionadas (vamos chamá-las de X e Y) proveem aleatoriamente de uma *distribuição normal bivariada*. Vamos chamar o coeficiente de correlação real de  $\rho$ . Vamos considerar duas situações:

$$1) H_0 : \rho = 0; \quad H_1 : \rho \neq 0$$

Para testarmos a hipótese nula de que o coeficiente de correlação é zero, pode-se mostrar que a estatística apropriada é:

$$t = r \sqrt{\frac{n-2}{1-r^2}} \quad (19.20)$$

Quando  $H_0$  for verdadeira,  $t$  segue uma distribuição  $t$  de Student, com  $n-2$  graus de liberdade, e o teste de hipótese é realizado da forma conhecida.

$$2) H_0 : \rho = \rho_0; \quad H_1 : \rho \neq \rho_0$$

Um outro enfoque deve ser utilizado para realizar um teste de hipótese quando a hipótese nula se referir a um valor de coeficiente de correlação diferente de zero. O coeficiente de correlação,  $r$ , deve ser transformado para  $z_r$  como se segue (transformação de Fisher):

$$z_r = \frac{1}{2} \ln \frac{1+r}{1-r} \quad (19.21)$$

onde  $\ln$  é o logaritmo natural (base  $e$ ). Pode-se mostrar que, sob a hipótese nula e para amostras não muito pequenas,  $z_r$  é aproximadamente normalmente distribuído com média igual a  $z_{\rho_0} = \frac{1}{2} \ln \frac{1+\rho_0}{1-\rho_0}$ , e desvio padrão estimado igual a  $\sqrt{\frac{1}{n-3}}$ . Para testarmos então a hipótese nula de que  $\rho = \rho_0 \neq 0$ , a estatística a ser utilizada é:

$$z = \frac{z_r - z_{\rho_0}}{\sqrt{1/(n-3)}} \quad (19.22)$$

que segue uma distribuição normal padrão.

Para calcularmos o intervalo de confiança  $(1 - \alpha)$  para  $\rho$ , inicialmente calculamos o intervalo de confiança para  $z_\rho$ , dado por:

$$\left[ z_r - z_{1-\alpha/2} \sqrt{\frac{1}{n-3}}, z_r + z_{1-\alpha/2} \sqrt{\frac{1}{n-3}} \right] \quad (19.23)$$

Os limites do intervalo de confiança para  $\rho$  são obtidos a partir dos limites de  $z_\rho$  acima, usando a inversa da relação (19.21):

$$r = \frac{e^{2z_r} - 1}{e^{2z_r} + 1}$$

e substituindo  $z_r$  sucessivamente pelos dois limites obtidos na expressão (19.23).

### 19.6.2 Cálculo do coeficiente de correlação no *R Commander*

Para calcularmos o valor, o intervalo de confiança e realizarmos um teste de hipótese para o coeficiente de correlação entre duas variáveis numéricas no *R Commander*, utilizamos a opção:

Estatísticas  $\Rightarrow$  Resumos  $\Rightarrow$  Teste de Correlação...

Na tela de configuração do teste, o usuário deve seleccionar as variáveis que serão correlacionadas, o tipo de correlação e o tipo de teste de hipótese (figura 19.15).

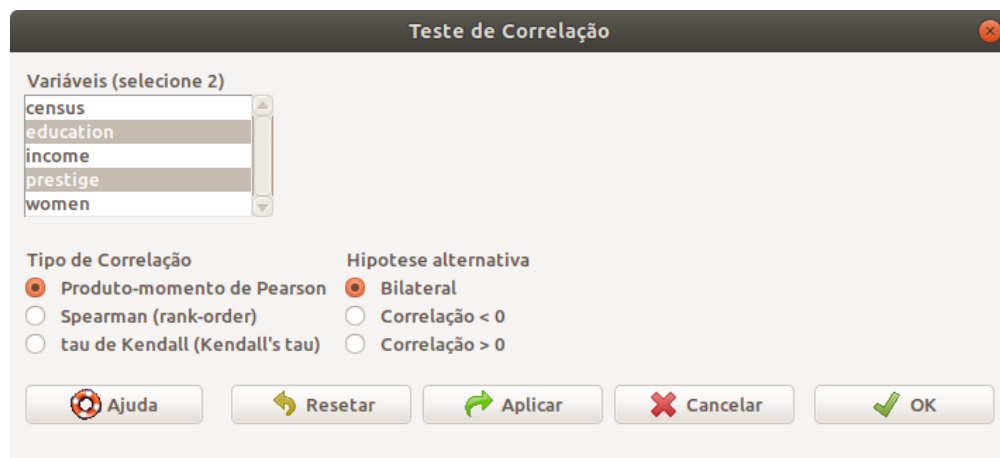


Figura 19.15: Seleção das duas variáveis que serão correlacionadas, tipo de coeficiente de correlação (Pearson neste exemplo) e se o teste é bilateral ou unilateral.

A função executada é mostrada a seguir. Os resultados mostram o valor de  $p$  ( $< 2,2 \cdot 10^{-16}$ ), o valor do coeficiente de correlação (0,85) e o intervalo de confiança ao nível de 95%.

```
with(Prestige, cor.test(education, prestige, alternative="two.sided",
                        method="pearson"))
```

```
##
## Pearson's product-moment correlation
##
## data: education and prestige
## t = 16.148, df = 100, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## 0.7855899 0.8964367
## sample estimates:
## cor
## 0.8501769
```

Caso desejássemos outro nível de confiança, bastaria alterarmos a função acima, acrescentando o parâmetro *conf.level*. Por exemplo, se desejássemos que o nível de significância do teste fosse 10% (nível de confiança = 90%), utilizaríamos a função *cor.test* da seguinte forma:

```
with(Prestige, cor.test(education, prestige, alternative="two.sided",
                        method="pearson", conf.level = .90))
```

```
##
## Pearson's product-moment correlation
##
## data: education and prestige
## t = 16.148, df = 100, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 90 percent confidence interval:
## 0.7974164 0.8900372
## sample estimates:
## cor
## 0.8501769
```

O coeficiente de correlação linear indica o grau de relação linear entre duas variáveis numéricas. Quando ele for zero, significa que uma variável não é linearmente relacionada à outra, mas isso não quer dizer que não exista nenhum relacionamento entre essas variáveis. Há muitos outros tipos de relações entre duas variáveis que não são lineares. A figura 19.16 mostra três possíveis relações: senoidal, quadrática e exponencial.



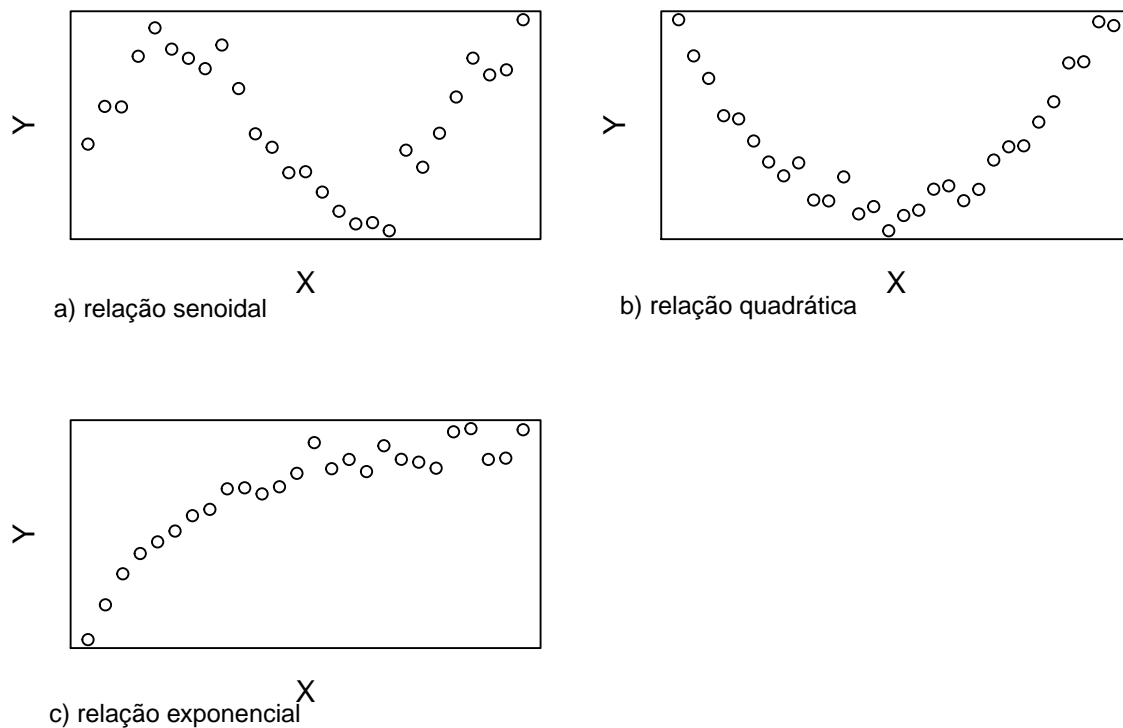


Figura 19.16: Exemplos de relações não lineares entre duas variáveis.

### 19.6.3 Coeficiente de correlação de Spearman

Caso as variáveis não possuam uma distribuição conjunta normal e não haja alguma transformação das variáveis que torne a distribuição conjunta das variáveis transformadas normal, então pode-se recorrer a métodos não paramétricos para verificar o relacionamento linear entre as variáveis. Um dos métodos mais frequentemente utilizados é o coeficiente de correlação de postos de Spearman. Para o cálculo do coeficiente de Spearman, postos são atribuídos aos valores das variáveis, de maneira análoga ao teste não paramétrico de Wilcoxon para duas amostras (seção 16.2.5) e, então, o coeficiente de correlação é calculado de acordo com a equação (19.19), utilizando os postos das variáveis e não os valores originais.

Pacotes estatísticos realizam testes de hipótese para a hipótese de correlação nula entre os postos das variáveis.

Para realizarmos o teste de correlação de Spearman para as variáveis *prestige* e *education*, basta selecionarmos as variáveis, o tipo de teste e se é bilateral ou não, conforme a figura 19.17, acessada por meio da opção.

Estatísticas  $\Rightarrow$  Resumos  $\Rightarrow$  Teste de Correlação...



Figura 19.17: Configuração para o cálculo do coeficiente de correlação de Spearman.

A função executada (*cor.test*) é mostrada abaixo seguida dos resultados.

```
with(Prestige, cor.test(education, prestige, alternative="two.sided",
                        method="spearman"))
```

```
##
## Spearman's rank correlation rho
##
## data: education and prestige
## S = 32100, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
## rho
## 0.8184923
```

## 19.7 Exercícios

- 1) Com o conjunto de dados *CrohnD* do pacote *robustbase* ([GPL-2](#) | [GPL-3](#)), faça as atividades abaixo.
  - a) Veja a ajuda do conjunto de dados.
  - b) Faça um diagrama de dispersão do índice de massa corporal (IMC) por peso, com a reta de regressão superposta.
  - c) Construa um modelo de regressão linear simples do IMC em função do peso. Identifique os valores dos coeficientes de regressão, o coeficiente de determinação e o valor de p do modelo. Comente sobre o poder explicativo do modelo.
  - d) Obtenha os diagnósticos gráficos básicos do modelo e comente os gráficos dos resíduos x valores ajustados pela reta de regressão, o gráfico de comparação de quantis dos resíduos padronizados e o diagrama de dispersão da raiz quadrada dos resíduos padronizados x valores ajustados.

- e) Obtenha os intervalos de confiança para os coeficientes do modelo de regressão.
- 2) Com o conjunto de dados *Pima.te* do pacote [MASS](#) (GPL-2 | GPL-3), faça as atividades abaixo.
- a) Veja a ajuda do conjunto de dados.
  - b) Faça um diagrama de dispersão da glicose por índice de massa corporal (IMC), com a reta de regressão superposta.
  - c) Construa um modelo de regressão linear simples da glicose em função do IMC. Identifique os valores dos coeficientes de regressão, o coeficiente de determinação e o valor de  $p$  do modelo. Comente sobre o poder explicativo do modelo.
  - d) Comparando com o modelo do exercício anterior e da conhecida fórmula para o cálculo do IMC, comente o fato de o coeficiente de determinação do modelo do exercício 1 ser bem maior do que o deste exercício.
  - e) Obtenha os diagnósticos gráficos básicos do modelo e comente os gráficos dos resíduos x valores ajustados pela reta de regressão, o gráfico de comparação de quantis dos resíduos padronizados e o diagrama de dispersão da raiz quadrada dos resíduos padronizados x valores ajustados.
  - f) Obtenha os intervalos de confiança para os coeficientes do modelo de regressão.

# Capítulo 20

## Análise de sobrevida

Os conteúdos das seções 20.1, 20.2 e 20.3 podem ser visualizados neste [vídeo](#).

### 20.1 Introdução

A análise de sobrevida visa a fornecer estimativas da probabilidade de um indivíduo experimentar um dado evento de interesse e eventualmente comparar curvas de sobrevida entre diferentes estratos de pacientes. O nome “análise de sobrevida” deriva da utilização desse tipo de análise quando o evento de interesse é a ocorrência ou não de morte, mas ela pode ser utilizada para outros eventos, como recidiva de câncer, infecção, etc.

Essa é uma área bastante complexa e há inúmeras publicações dedicadas ao assunto. Este capítulo mostra como construir curvas de sobrevida no R para dados com censura à direita, como estimar probabilidades de sobrevida por meio do método de Kaplan-Meier e como comparar curvas de sobrevida por meio do teste *log-rank*.

### 20.2 Conjunto de dados utilizado neste capítulo

Vamos utilizar o conjunto de dados *cancer* do pacote [survival](#) (LGPL-2 | LGPL-2.1 | LGPL-3).

Essa base contém dados relativos à sobrevida de 228 indivíduos com câncer avançado do pulmão do *North Central Cancer Treatment Group*. Neste capítulo, será utilizado o plugin *RcmdrPlugin.survival*. Ao instalar esse plugin com as dependências, o pacote *survival* também será instalado.

Os passos para a instalação desse pacote são os mesmos utilizados para a instalação do *R Commander*, seção A.6.

Os comandos abaixo carregam o plugin *RcmdrPlugin.survival* e o conjunto de dados *cancer* no *R Commander*. Após a execução desses comandos, selecione o conjunto de dados *cancer* como conjunto de dados ativo.

```
library(RcmdrPlugin.survival)
data(cancer, package="survival")
```

A figura 20.1 mostra a estrutura desse conjunto de dados.

lung {survival}	R Documentation
NCCTG Lung Cancer Data	
<b>Description</b>	
Survival in patients with advanced lung cancer from the North Central Cancer Treatment Group. Performance scores rate how well the patient can perform usual daily activities.	
<b>Usage</b>	
lung cancer	
<b>Format</b>	
inst:	Institution code
time:	Survival time in days
status:	censoring status 1=censored, 2=dead
age:	Age in years
sex:	Male=1 Female=2
ph.ecog:	ECOG performance score as rated by the physician. 0=asymptomatic, 1= symptomatic but completely ambulatory, 2= in bed <50% of the day, 3= in bed > 50% of the day but not bedbound, 4 = bedbound
ph.karno:	Karnofsky performance score (bad=0-good=100) rated by physician
pat.karno:	Karnofsky performance score as rated by patient
meal.cal:	Calories consumed at meals
wt.loss:	Weight loss in last six months
<b>Note</b>	
The use of 1/2 for alive/dead instead of the usual 0/1 is a historical footnote. For data contained on punch cards, IBM 360 Fortran treated blank as a zero, which led to a policy within the section of Biostatistics to never use "0" as a data value since one could not distinguish it from a missing value. The policy became a habit, as is often the case; and the 1/2 coding endured long beyond the demise of punch cards and Fortran.	
<b>Source</b>	
Terry Therneau	
<b>References</b>	
Loprinzi CL. Laurie JA. Wieand HS. Krook JE. Novotny PJ. Kugler JW. Bartel J. Law M. Bateman M. Klatt NE. et al. Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. Journal of Clinical Oncology. 12(3):601-7, 1994.	
[Package survival version 3.1-12 <a href="#">Index</a> ]	

Figura 20.1: Estrutura do conjunto de dados *cancer* do pacote *survival*.

O comando abaixo lista os cinco primeiros registros desse conjunto de dados.

```
head(cancer, 5)
```

```
##   inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss
## 1    3  306      2  74   1        1        90        100    1175      NA
## 2    3  455      2  68   1         0        90         90    1225      15
## 3    3 1010      1  56   1         0        90         90       NA      15
## 4    5  210      2  57   1         1        90         60    1150      11
## 5    1  883      2  60   1         0       100         90       NA       0
```

Para construirmos uma curva de sobrevida para os pacientes com câncer de pulmão, precisamos trabalhar com as seguintes variáveis:

*time*: tempo de sobrevida em dias

*status*: status da censura (1=censura, 2=morte)

Em geral, em estudos em que se busca identificar se um indivíduo experimentou um determinado evento (desfecho) de interesse durante a duração do estudo e o tempo entre a entrada no estudo e a ocorrência do evento, é preciso incluir uma variável que indica se o indivíduo experimentou ou não o evento durante o período em que ele(a) participou do estudo. Caso o indivíduo não tenha experimentado o desfecho no período em que ele(a) foi observado, dizemos que houve uma **censura**, e o indivíduo foi *censurado*.

Um sujeito pode ser censurado devido às seguintes razões:

- perda de acompanhamento;
- retirada do estudo;
- não ocorreu o evento de interesse ao final do período de estudo.

Essas três razões se aplicam a censuras conhecidas como **censuras à direita**, porque ocorrem após o paciente entrar no estudo. Censuras à esquerda e censuras de intervalo também são possíveis, mas este texto irá tratar somente de censuras à direita.

No conjunto de dados *cancer*, a variável *status* indica se ocorreu ou não uma “censura” para o respectivo paciente, sendo o evento de interesse a morte. O *status* igual a 1 significa que houve censura, ou seja, até o momento em que o paciente foi acompanhado no estudo, ele(a) ainda estava vivo(a) e o valor do tempo de sobrevida corresponde ao tempo decorrido desde a entrada do paciente no estudo até o momento em que ele(a) foi retirado(a) do estudo. O tempo de sobrevida desse paciente é maior do que o indicado no estudo, mas não sabemos o que aconteceu com ele(a) após o momento de saída do estudo. O *status* igual a 2 indica que o paciente morreu durante o acompanhamento no estudo e o valor do tempo de sobrevida corresponde ao tempo desde a entrada do paciente no estudo até a sua morte.

Assim, se observarmos os três primeiros registros do conjunto de dados *cancer*, iremos verificar que os dois primeiros pacientes morreram nos dias 306 e 455 após o início do acompanhamento de cada um deles, respectivamente. Nesses dois casos, não houve censura (*status* = 2). O terceiro paciente foi censurado (*status* = 1) após 1010 dias de acompanhamento. Assim a sobrevida do terceiro paciente é superior a 1010 dias, mas não sabemos exatamente qual o seu valor.

**Observação:** apesar de o conjunto de dados *cancer* codificar os pacientes que foram censurados com o valor 1 e os que morreram com o valor 2, em geral, são utilizados os valores 0 para indicar a censura e o valor 1 para indicar os indivíduos que experimentaram o evento estudado.

## 20.3 Obtendo a curva de sobrevida no R

A curva de sobrevida mostra, para cada instante de tempo, a probabilidade de um indivíduo sobreviver além daquele instante de tempo.

Nesta seção, vamos construir uma curva de sobrevida para os pacientes do conjunto de dados *cancer* a partir do *R Commander*. A seção seguinte irá explicar como essa curva é obtida.

Uma vez carregados o plugin *RcmdrPlugin.survival* e tendo o conjunto de dados *cancer* como conjunto ativo, para gerar a curva de sobrevida, selecionamos a seguinte opção do menu do *R Commander*:

Estatísticas ⇒ Análise de Sobrevida ⇒ Função estimada de sobrevida

Na tela *Função de Sobrevida* (figura 20.2), selecionamos na primeira lista a variável que indica o tempo até a ocorrência do evento ou censura (*time* no conjunto de dados *cancer*). Na segunda lista, selecionamos a variável que indica se o indivíduo foi censurado ou não (nesse exemplo, a variável é *status*). Em tipo de censura, vamos selecionar a opção *Right*.

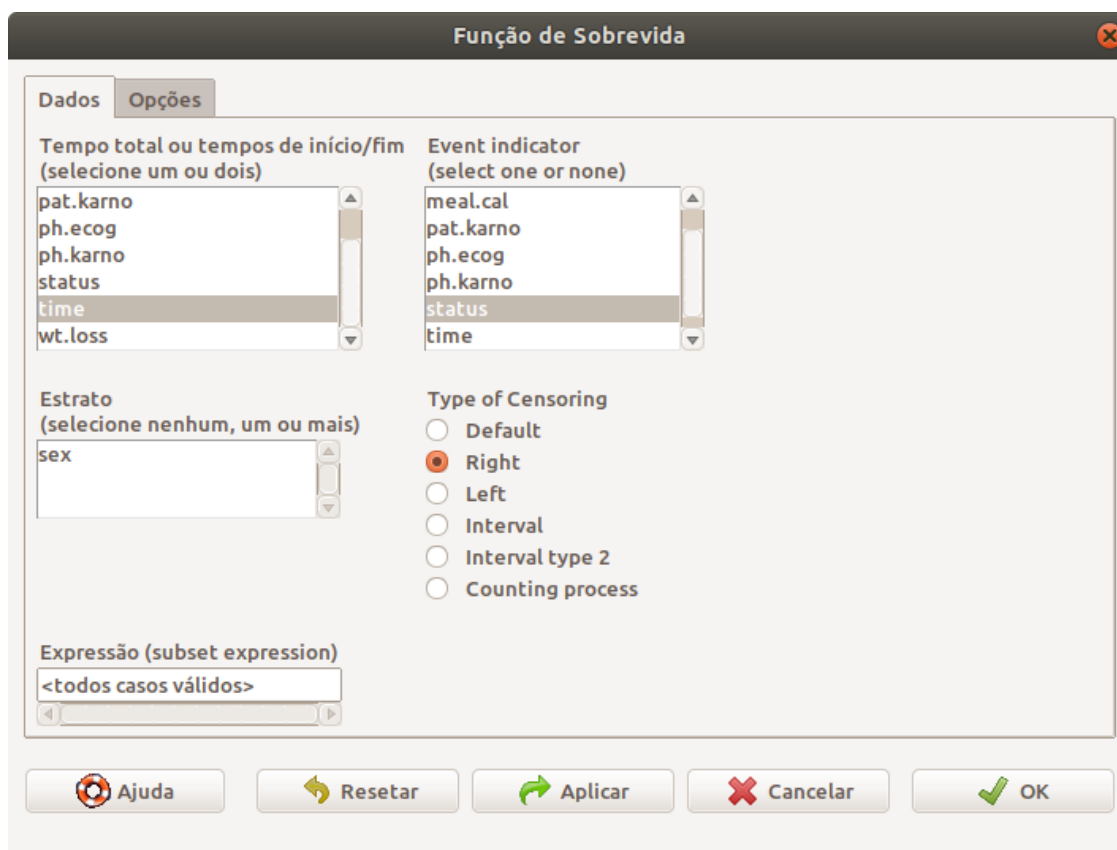


Figura 20.2: Tela para especificar as variáveis que serão utilizadas para construir a curva de sobrevida.

Na aba *Opções* (figura 20.3), há diversas opções que podem ser configuradas para construir a curva de sobrevida. Vamos manter as opções padrão:

- intervalos de confiança: logaritmo;
- gráfico com intervalos de confiança: default, correspondente à opção sim;
- nível de confiança : 95%;
- método: Kaplan-Meier;
- método de variância: Greenwood;
- quantis para estimar:  $P_{25}$ , Mediana,  $P_{75}$ ;
- indicar os tempos de censura: sim;
- resumo: padrão.

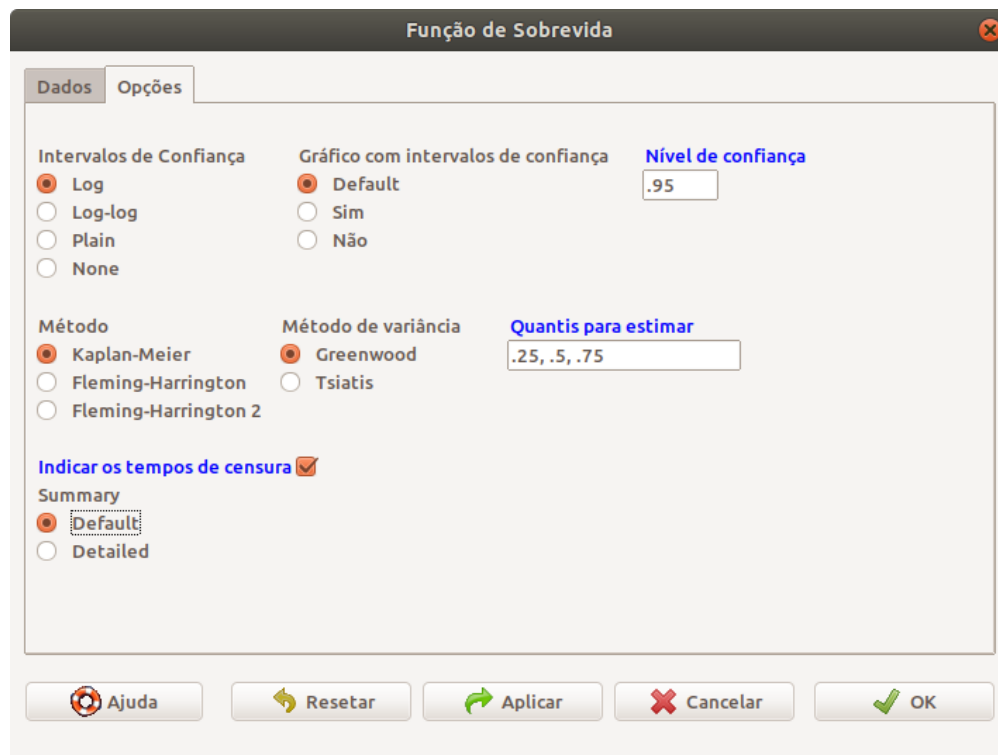


Figura 20.3: Tela para definir as opções para construir a curva de sobrevida.

Ao clicarmos em OK, a seguinte sequência de comandos é executada:

- 1) A função *survfit* gera uma série de componentes utilizados para construir a curva de sobrevida. Esses componentes são armazenados no objeto *.Survfit*. Um resumo do objeto *.Survfit* é exibido na tela, indicando a função chamada, o número de pessoas observadas ( $n$ ), o número de mortes (*events*), a mediana do tempo de sobrevida (*median*), que é o tempo no qual a probabilidade de sobrevida é de 50%, e os limites inferior ( $0.95LCL$ ) e superior ( $0.95UCL$ ) do intervalo de confiança para a mediana do tempo de sobrevida.

```
.Survfit <- survfit(Surv(time, status, type="right") ~ 1, conf.type="log",
                  conf.int=0.95, type="kaplan-meier",
                  error="greenwood", data=cancer)
```



```
.Survfit
```

```
## Call: survfit(formula = Surv(time, status, type = "right") ~ 1, data = cancer,  
##      error = "greenwood", conf.type = "log", conf.int = 0.95,  
##      type = "kaplan-meier")  
##  
##          n events median 0.95LCL 0.95UCL  
## [1,] 228    165    310    285    363
```

- 2) A curva de sobrevida é exibida por meio da função *plot*, aplicada sobre o objeto *.Survfit* (figura 20.4). Ela mostra, como dito antes, para cada instante de tempo, a probabilidade de um indivíduo sobreviver além daquele instante de tempo. Essa curva é decrescente com o tempo.

```
par(mar=c(3,3,2,1))  
plot(.Survfit, mark.time=TRUE)
```

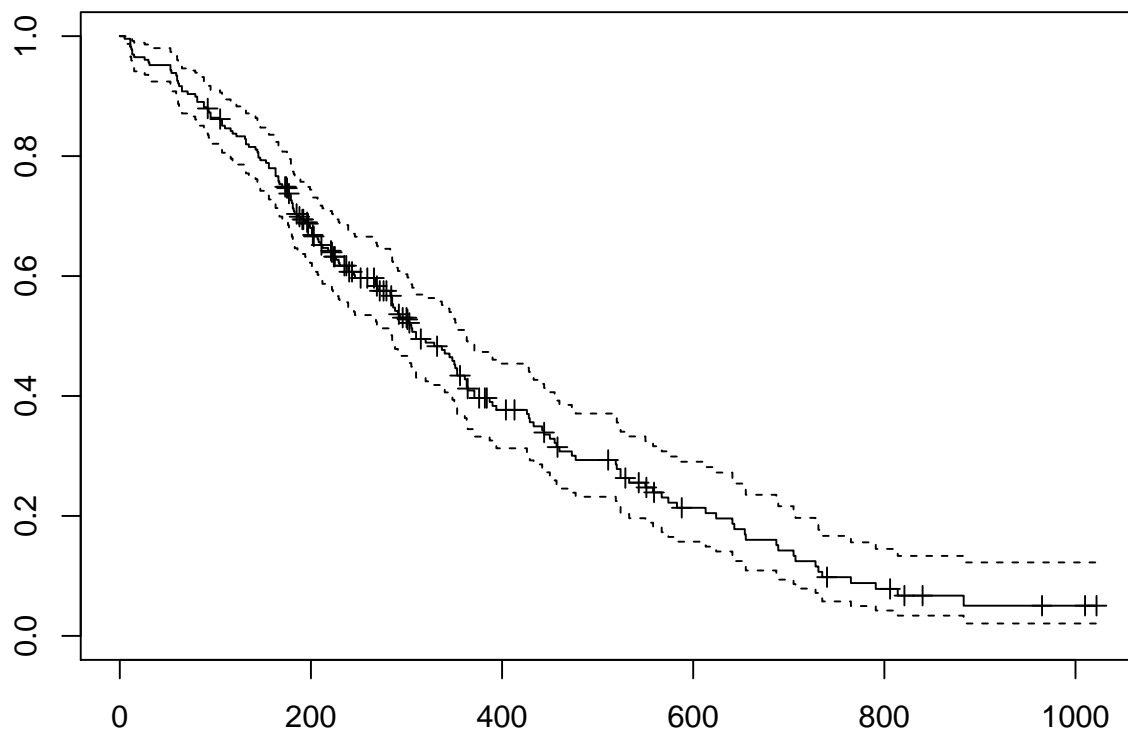


Figura 20.4: Curva de sobrevida pelo método de Kaplan-Meier para o conjunto de dados cancer.

- 3) São exibidos os valores das estimativas e respectivos intervalos de confiança para o  $P_{25}$ , a mediana e o  $P_{75}$  do tempo de sobrevida por meio da função *quantile*.

```
quantile(.Survfit, quantiles=c(.25,.5,.75))
```

```
## $quantile
##  25  50  75
## 170 310 550
##
## $lower
##  25  50  75
## 145 285 460
##
## $upper
##  25  50  75
## 197 363 654
```

4) O objeto *.Survfit* é removido da área de trabalho.

```
remove(.Survfit)
```

## 20.4 Estimando a probabilidade de sobrevida

Os conteúdos desta seção e da seção seguinte (seção 20.5) podem ser visualizados neste [vídeo](#).

O método de **Kaplan-Meier** é um método não paramétrico, sendo o mais utilizado para estimar tempos de sobrevida e probabilidades de sobrevida. Ele resulta em uma função degrau, onde há um degrau para baixo a cada instante em que um evento ocorre.

A **probabilidade de sobrevida** em um certo instante de tempo é a probabilidade de um indivíduo experimentar o evento de interesse após o instante  $t$ .

A função *Surv* do pacote *survival* gera um valor para cada indivíduo que é o seu tempo de sobrevida, seguido por um sinal de + se o indivíduo foi censurado. Vamos ver os primeiros 10 valores:

```
Surv(cancer$time, cancer$status)[1:10]
```

```
## [1] 306 455 1010+ 210 883 1022+ 310 361 218 166
```

Observamos que esses valores correspondem ao tempo de sobrevida dos 10 primeiros indivíduos do conjunto de dados *cancer*, indicando que os indivíduos 3 e 6 foram censurados ( $\text{status} = 1$ ).

Para estimar as probabilidades de sobrevida, os tempos de sobrevida são ordenados em ordem crescente. A função a seguir mostra os 15 menores tempos de sobrevida do conjunto de dados *cancer*, com a censura indicada pelo sinal + após o valor do tempo. Podemos ver que um indivíduo sobreviveu somente 5 dias, três indivíduos sobreviveram 11 dias, um indivíduo sobreviveu 12 dias, dois indivíduos sobreviveram 13 dias e assim por diante.

```
sort(Surv(cancer$time, cancer$status))[1:15]
```

```
## [1] 5 11 11 11 12 13 13 15 26 30 31 53 53 54 59
```

Os valores gerados pela função *Surv* são utilizados pela função *survfit* para gerar o objeto *.Survfit* na seção anterior. Vamos gerar novamente esse objeto e compreender como são estimadas as probabilidades de sobrevida após um certo tempo *t* e como é construída a curva de sobrevida.

O objeto *.Survfit* possui diversos componentes, cujos nomes podem ser obtidos por meio da função *names* aplicada a *.Survfit*, que nos permite construir uma tabela que gera as probabilidades de sobrevida de pacientes com câncer de pulmão em cada instante de tempo.

```
.Survfit <- survfit(Surv(time, status, type="right") ~ 1, conf.type="log",
                  conf.int=0.95, type="kaplan-meier",
                  error="greenwood", data=cancer)
names(.Survfit)[1:6]
```

```
## [1] "n"          "time"       "n.risk"     "n.event"    "n.censor"   "surv"
```

```
names(.Survfit)[7:12]
```

```
## [1] "std.err"    "cumhaz"     "std.chaz"   "type"       "logse"      "conf.int"
```

```
names(.Survfit)[13:16]
```

```
## [1] "conf.type" "lower"      "upper"      "call"
```

Utilizando os componentes *time*, *n.risk*, *n.event*, *n.censor* e *surv* de *.Survfit*, podemos montar uma tabela que dá a estimativa da probabilidade de sobrevida até pelo menos cada um dos tempos de sobrevida dos pacientes do conjunto de dados *cancer*. O comando abaixo mostra as 10 primeiras linhas desta tabela.

```
head(with(.Survfit, data.frame(time, n.risk, n.event, n.censor, surv)), 10)
```

```
##      time n.risk n.event n.censor      surv
## 1      5     228       1         0 0.9956140
## 2     11     227       3         0 0.9824561
## 3     12     224       1         0 0.9780702
## 4     13     223       2         0 0.9692982
## 5     15     221       1         0 0.9649123
## 6     26     220       1         0 0.9605263
## 7     30     219       1         0 0.9561404
## 8     31     218       1         0 0.9517544
## 9     53     217       2         0 0.9429825
## 10    54     215       1         0 0.9385965
```

Para cada linha da tabela acima, *time* indica o tempo de sobrevida, *n.risk* o número de pacientes em risco de experimentar o evento antes do valor indicado por *time*, *n.event* indica o número de indivíduos que experimentaram o evento no tempo mostrado na linha corrente, *n.censor* indica o número de indivíduos que foram censurados no tempo mostrado na linha corrente, *surv* é a probabilidade de sobrevida até pelo menos o instante indicado pela linha corrente.

Assim a primeira linha indica que 1 paciente morreu no 5º dia e nenhum foi censurado nesse dia. 228 pacientes entraram no estudo. Logo 227 pacientes sobreviveram até pelo menos o quinto dia de acompanhamento. Então a probabilidade de um paciente sobreviver ao instante 5 é estimada pelo número de pacientes que sobreviveram ao 5º dia de acompanhamento dividido pelo número de pacientes em risco de morte no início do acompanhamento ( $227/228 = 0,9956$ ).

A segunda linha indica que 3 pacientes morreram no 11º dia e nenhum foi censurado nesse dia. 227 pacientes estavam em risco de morrer após o 5º dia. Logo 224 pacientes sobreviveram ao 11º quinto dia de acompanhamento. Então a probabilidade de um paciente sobreviver ao 11º dia, tendo sobrevivido ao 5º dia, é estimada pelo número de pacientes que sobreviveram ao 11º dia de acompanhamento dividido pelo número de pacientes em risco do evento após o 5º dia ( $224/227 = 0,9868$ ). Finalmente a probabilidade de o paciente sobreviver a 11 dias é igual ao produto da probabilidade de ele sobreviver a 5 dias (0,9956) pela probabilidade de ele sobreviver a 11 dias, tendo sobrevivido a 5 dias (0,9868). Logo a probabilidade de um indivíduo sobreviver a 11 dias é 0,9825 ( $0,9956 \times 0,9868$ ).

A figura 20.5 mostra o passo a passo para os cálculos da probabilidade de sobrevida em cada instante pelo método de Kaplan-Meier. As colunas *n.risk*, *tempo*, *n.event* e *n.censor* possuem o mesmo significado dos componentes *n.risk*, *time*, *n.event* e *n.censor* do objeto *.Survfit*.

A coluna *probabilidade de sobrevida no período  $t_{i-1}$  a  $t_i$*  mostra a probabilidade de um indivíduo sobreviver ao instante  $t_i$ , caso ele esteja vivo no instante  $t_{i-1}$ ,  $p(t_i) = P(t > t_i | t > t_{i-1})$ , e é calculada, em cada linha, pela expressão:

$$p(t_i) = P(t > t_i | t > t_{i-1}) = \frac{n.risk_i - n.event_i}{n.risk_i}$$

O número de pacientes em risco na linha *i* é calculado subtraindo do número de pacientes em risco na linha anterior (*i-1*) a soma do número de pacientes que morreram com o número de pacientes censurados, ambos na linha (*i-1*):

$$n.risk_i = n.risk_{i-1} - n.event_{i-1} - n.censor_{i-1}$$

O número de pacientes em risco na linha *i* é calculado subtraindo do número de pacientes em risco na linha *A* coluna *probabilidade de sobrevida no instante  $t_i$*  mostra a probabilidade de um indivíduo sobreviver ao instante  $t_i$ ,  $P(t_i) = P(t > t_i)$ , e é calculada, na linha *i*, pelo produto da *probabilidade de sobrevida no instante  $t_{i-1}$*  (linha anterior) pela *probabilidade de sobrevida no período  $t_{i-1}$  a  $t_i$*  na mesma linha.

$$P(t_i) = P(t > t_i) = P(t_{i-1}) \cdot p(t_i)$$

A linha destacada pelo retângulo vermelho na figura 20.5 mostra esse cálculo para a probabilidade de sobrevida no instante  $t_2 = 11$  (0,9825), que é igual ao produto da probabilidade de sobrevida no instante  $t_1 = 5$  (0,9956) pela probabilidade de sobrevida no período  $t_1$  a  $t_2$  (0,9868). Seguindo esse raciocínio, calcula-se as probabilidades de sobrevida nos demais instantes.

Linha i	n.risk	tempo	n.event	n.censor	probabilidade de sobrevida no período $t_{i-1}$ a $t_i$ - $p(t_i)$	probabilidade de sobrevida no instante $t_i$ - $P(t_i)$
1	228	5	1	0	0,99561	0,99561
2	227	11	3	0	0,98678	0,98246
3	224	12	1	0	0,99554	0,97807
4	223	13	2	0	0,99103	0,96930
5	221	15	1	0	0,99548	0,96491
6	220	26	1	0	0,99545	0,96053
7	219	30	1	0	0,99543	0,95614
8	218	31	1	0	0,99541	0,95175
...	...	...	...	...	...	...
20	201	92	1	1	0,99502	0,87719
21	199	93	1	0	0,99497	0,87278
22	198	95	2	0	0,98990	0,86397
23	196	105	1	1	0,99490	0,85956
24	194	107	2	0	0,98969	0,85070
...	...	...	...	...	...	...
41	170	170	1	0	0,99412	0,74879
42	169	173	0	1	1,00000	0,74879
43	168	174	0	1	1,00000	0,74879
44	167	175	1	1	0,99401	0,74431
45	165	176	1	0	0,99394	0,73980
...	...	...	...	...	...	...

Figura 20.5: Tabela para obter as estimativas da probabilidade de sobrevida pelo método de Kaplan-Meier.

O retângulo verde na figura 20.5 mostra uma situação onde ocorre uma censura. Os pacientes censurados em um instante de tempo causam uma redução no número de pacientes sob risco do evento nos instantes posteriores.

O retângulo alaranjado na figura 20.5 mostra duas situações onde ocorre uma censura no instante  $t_{42} = 173$  e outra censura no instante  $t_{43} = 174$ , mas não ocorreu nenhuma morte entre o instante  $t_{41} = 170$  e  $t_{43} = 174$ . Podemos observar que as probabilidades de sobrevida nos instantes  $t_{42}$  e  $t_{43}$  continuam iguais à probabilidade de sobrevida no instante  $t_{41}$ . Portanto precisamos atualizar as estimativas das probabilidades de sobrevida somente nos instantes em que ocorrem o evento de interesse.

A partir da tabela da figura 20.5, a curva de sobrevida é construída, plotando para cada instante em que ocorreu o evento de interesse a probabilidade de sobrevida nesse instante. Os pontos assim obtidos são unidos por uma função em degrau, da seguinte forma (figura 20.6): dados dois pontos  $(t_{i-1}, P[t_{i-1}])$  e  $(t_i, P[t_i])$ , eles são unidos por uma linha horizontal que vai de  $(t_{i-1}, P[t_{i-1}])$  a  $(t_i, P[t_{i-1}])$  e, em seguida, por uma linha vertical que vai de  $(t_i, P[t_{i-1}])$  a  $(t_i, P[t_i])$ .

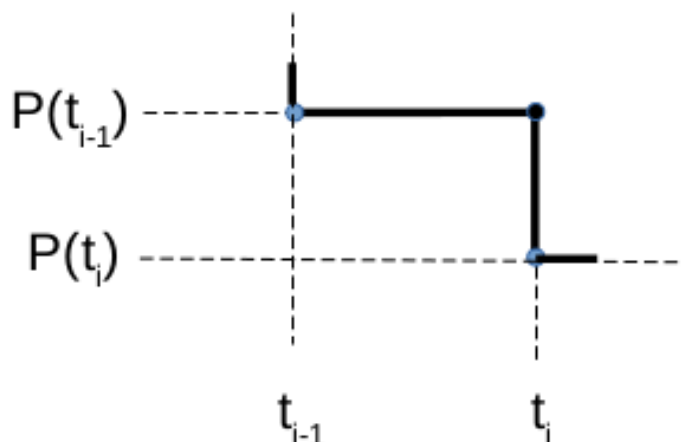


Figura 20.6: Como os pontos de coordenadas  $(t_{i-1}, P[t_{i-1}])$  e  $(t_i, P[t_i])$  são unidos para construir a curva de sobrevida pelo método de Kaplan-Meier.

## 20.5 Obtendo as probabilidades de sobrevida em instantes específicos

Uma quantidade de interesse na análise de sobrevida é a probabilidade de sobreviver além de um certo instante de tempo. O comando abaixo mostra como estimar a probabilidade de sobreviver além de 1 e 2 anos para o conjunto de dados *cancer*. No argumento *times*, especificamos um vetor com dois valores (365,25 e 730,5), que correspondem, respectivamente, a 1 e a 2 anos, já que estamos medindo o tempo em dias e, em média, um ano contém 365,25 dias.

```
summary(survfit(Surv(time, status) ~ 1, data=cancer), times=c(365.25, 730.5))

## Call: survfit(formula = Surv(time, status) ~ 1, data = cancer)
##
##   time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   365     65     121   0.409  0.0358    0.3447    0.486
##   730     13      38   0.116  0.0283    0.0716    0.187
```

Nos resultados acima, a coluna *survival* fornece as probabilidades de sobrevida, que são iguais a 0,41 e 0,12 para um e dois anos, respectivamente. As colunas *std.err*, *lower 95% CI* e *upper 95% CI* fornecem os valores do erro padrão, o limite inferior do intervalo de confiança e o limite superior do intervalo de confiança para a probabilidade de sobrevida além do primeiro e segundo ano, respectivamente.

Nesse exemplo, o erro padrão da probabilidade de sobrevida além de um certo instante  $t_i$ ,  $EP[P(t_i)]$ , é calculado de acordo com a proposta de Greenwood (Greenwood, 1926):

$$EP[P(t_i)] = P(t_i) \cdot \sqrt{\sum_{j=1}^i \frac{n.event_j}{n.risk_j - n.event_j}}$$

Os limites inferior e superior para o intervalo com nível de confiança  $(100 - \alpha)\%$  para a probabilidade de sobrevida além de um certo instante  $t_i$  são dados por:

$$[max(P(t_i) - z_{1-\alpha/2} \cdot EP[P(t_i)]), 0), min(P(t_i) + z_{1-\alpha/2} \cdot EP[P(t_i)], 1)]$$

ou seja, o limite inferior do intervalo de confiança para  $P(t_i)$  é o máximo entre 0 e  $P(t_i) - z_{1-\alpha/2} \cdot EP[P(t_i)]$  e o limite superior do intervalo de confiança para  $P(t_i)$  é o mínimo entre 1 e  $P(t_i) + z_{1-\alpha/2} \cdot EP[P(t_i)]$ .

A fórmula de Greenwood é acurada somente assintoticamente, ou seja, para valores suficientemente grandes para  $n.risk_i$ . Há diversas outras propostas que visam a tornar mais acuradas as estimativas para amostras não muito grandes.

## 20.6 Obtendo a curva de sobrevida para diferentes estratos

Os conteúdos desta seção e da seção seguinte (seção 20.7) podem ser visualizados neste [vídeo](#).

Frequentemente, deseja-se comparar curvas de sobrevida para diferentes tratamentos para uma determinada condição clínica, ou comparar curvas de sobrevida para diferentes estratos

de pacientes. Vamos ilustrar como construir curvas de sobrevida para diferentes níveis de uma variável categórica, usando novamente o conjunto de dados *cancer*, o qual possui a variável categórica *sex*.

Como a variável *sex* está como um vetor numérico, vamos convertê-la para fator por meio do comando:

```
cancer <- within(cancer, {  
  sex <- factor(sex, labels = c('masculino', 'feminino'))  
})
```

Para construir curvas de sobrevida para diferentes níveis de uma variável categórica, escolhemos novamente a opção:

Estatísticas ⇒ Análise de Sobrevida ⇒ Função estimada de sobrevida

Na aba *Dados* (figura 20.7), selecionamos na primeira lista a variável *time* e na segunda lista, a variável *status*. Vamos selecionar a opção *Right* em tipo de censura e a variável *sex* no item estrato.

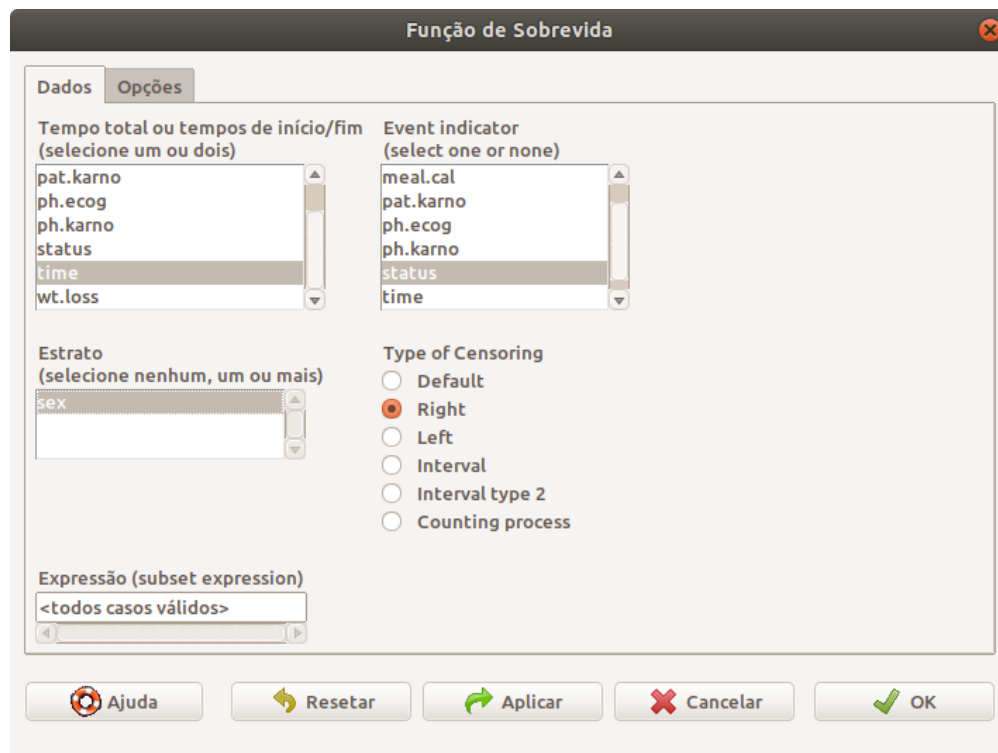


Figura 20.7: Tela para configurar as variáveis que serão utilizadas para construir a curva de sobrevida para diferentes estratos.



Ao clicamos em OK, os comandos a seguir serão executados, seguidos dos resultados e das curvas de sobrevida para cada estrato (figura 20.8).

```
.Survfit <- survfit(Surv(time, status, type="right") ~ sex, conf.type="log",
                    conf.int=0.95, type="kaplan-meier",
                    error="greenwood", data=cancer)
```

```
.Survfit
```

```
## Call: survfit(formula = Surv(time, status, type = "right") ~ sex, data = cancer,
##      error = "greenwood", conf.type = "log", conf.int = 0.95,
##      type = "kaplan-meier")
##
```

```
##              n events median 0.95LCL 0.95UCL
## sex=masculino 138    112    270    212    310
## sex=feminino  90     53    426    348    550
```

```
plot(.Survfit, col=1:2, lty=1:2, mark.time=TRUE)
legend("bottomleft", legend=c("sex=masculino", "sex=feminino"), col=1:2,
      lty=1:2, bty="n")
```

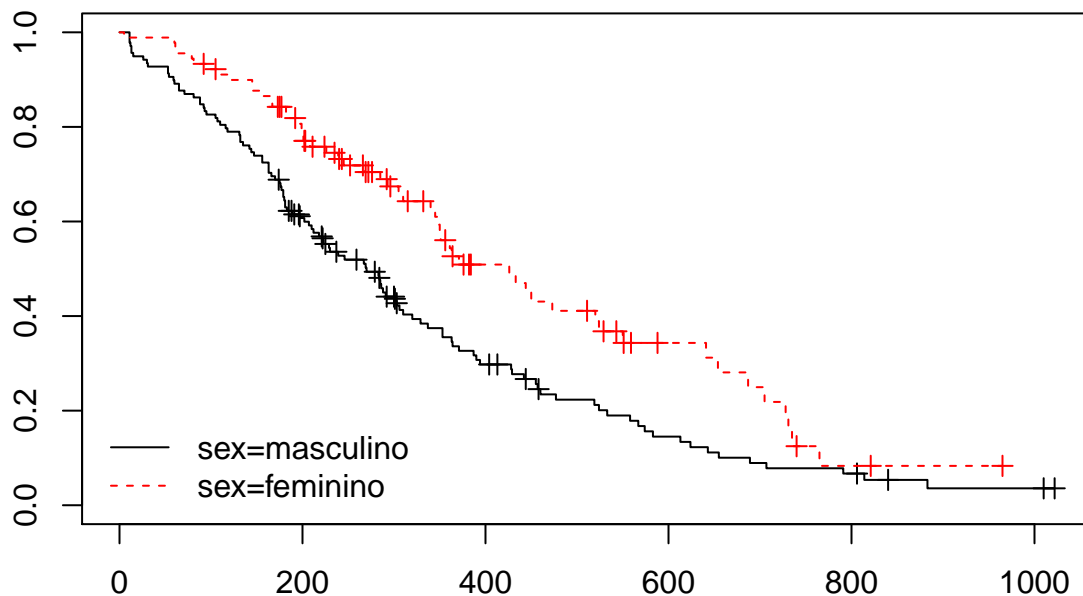


Figura 20.8: Curvas de sobrevida para cada sexo no conjunto de dados *cancer*.

```
quantile(.Survfit, quantiles=c(.25,.5,.75))
```

```
## $quantile
##           25  50  75
## sex=masculino 144 270 457
## sex=feminino  226 426 687
##
## $lower
##           25  50  75
## sex=masculino 107 212 387
## sex=feminino  186 348 550
##
## $upper
##           25  50  75
## sex=masculino 177 310 574
## sex=feminino  340 550  NA
remove(.Survfit)
```

Podemos ver que a sobrevida é maior entre as mulheres do que nos homens, sendo a mediana do tempo de sobrevida igual a 270 e 426 dias para os homens e as mulheres, respectivamente.

## 20.7 Comparação de funções de sobrevida em diferentes estratos

Existem vários testes estatísticos que avaliam a hipótese nula de que não há diferença entre as funções de sobrevida entre diferentes estratos. Um método frequentemente usado é o teste *log-rank*. Diversas formas desta estatística foram publicadas por diferentes estatísticos, por isso esse teste é conhecido por diversos nomes. Esse teste compara o número de eventos observados em cada grupo com o número de eventos esperados se os dois grupos são combinados em um só. Vamos ilustrar uma das formas como esse teste é realizado na função *survdiff* do pacote *survival*, conhecida como **teste log-rank de Mantel-Haenszel**.

Para realizar o teste log-rank de Mantel-Haenszel, uma tabela como a tabela 20.1 é construída. Nessa tabela, vamos supor que temos dois estratos ou grupos, onde A representa um estrato e B o outro. Cada linha  $i$  da tabela contém os tempos de ocorrência do evento de interesse em ordem crescente ( $t_i$ ), o número de indivíduos em risco em cada estrato ( $nA_i$  e  $nB_i$ ), o número total de indivíduos em risco ( $nT_i$ ), o número de eventos observados no tempo  $t_i$  em cada estrato ( $nAo_i$  e  $nBo_i$ ) e nos dois estratos em conjunto ( $nABo_i = nAo_i + nBo_i$ ), e o número de eventos esperados de ocorrer em cada estrato no tempo  $t_i$  ( $nAe_i$  e  $nBe_i$ ).

Sob a hipótese nula de que não há diferença entre as curvas de sobrevida entre os dois estratos, o número de eventos esperados em cada estrato e em cada instante  $t_i$  é igual ao produto do número total de eventos no instante  $t_i$  pela proporção de indivíduos em risco em cada estrato:

Tabela 20.1: Tabela simplificada para realizar o teste *log-rank*.

tempo	Em Risco			Eventos Observados			Eventos Esperados	
	Grupo A	Grupo B	Total	Grupo A	Grupo B	Total	Grupo A	Grupo B
t <sub>1</sub>	nA <sub>1</sub>	nB <sub>1</sub>	nT <sub>1</sub>	nAo <sub>1</sub>	nBo <sub>1</sub>	nABo <sub>1</sub>	nAe <sub>1</sub>	nBe <sub>1</sub>
...	...	...	...	...	...	...	...	...
t <sub>i</sub>	nA <sub>i</sub>	nB <sub>i</sub>	nT <sub>i</sub>	nAo <sub>i</sub>	nBo <sub>i</sub>	nABo <sub>i</sub>	nAe <sub>i</sub>	nBe <sub>i</sub>
...	...	...	...	...	...	...	...	...
t <sub>n</sub>	nA <sub>n</sub>	nB <sub>n</sub>	nT <sub>n</sub>	nAo <sub>n</sub>	nBo <sub>n</sub>	nABo <sub>n</sub>	nAe <sub>n</sub>	nBe <sub>n</sub>
				$\sum n_{Ao}$	$\sum n_{Bo}$		$\sum n_{Ae}$	$\sum n_{Be}$

$$nAe_i = nABo_i \cdot \frac{nA_i}{nT_i}$$

$$nBe_i = nABo_i \cdot \frac{nB_i}{nT_i}$$

A última linha da tabela mostra as somas dos valores observados ( $\sum n_{Ao}$  e  $\sum n_{Bo}$ ) e esperados ( $\sum n_{Ae}$  e  $\sum n_{Be}$ ) em cada estrato. A estatística para o teste *log-rank* é calculada pela fórmula a seguir:

$$\chi_{lr}^2 = \frac{(\sum n_{Ao} - \sum n_{Ae})^2}{\sum n_{Ae}} + \frac{(\sum n_{Bo} - \sum n_{Be})^2}{\sum n_{Be}} \quad (20.1)$$

A estatística (20.1) é uma variável aleatória com aproximadamente uma distribuição  $\chi^2$  com 1 grau de liberdade, quando a hipótese nula de que não há diferença entre as funções de sobrevida de cada estrato é verdadeira, desde que os nA<sub>i</sub> e nB<sub>i</sub> não sejam muito pequenos.

Valores suficientemente grandes de  $\chi_{lr}^2$  levam à rejeição de H<sub>0</sub>, ou seja, se  $\chi_{lr}^2 > \chi_{1-\alpha,1}^2$ , então H<sub>0</sub> é rejeitada.

Vamos ver como realizaríamos esse teste para um pequeno número de registros do conjunto de dados *cancer*. O resultado dos dois comandos abaixo são os 5 primeiros instantes de tempo em que ocorreram mortes entre as mulheres no conjunto de dados *cancer*, o número de mulheres em risco de morrer e o número de mulheres que morreram em cada instante, respectivamente.

```
.Survfit <- survfit(Surv(time, status, type="right") ~ 1, conf.type="log",
                    conf.int=0.95, type="kaplan-meier", error="greenwood",
                    data=cancer, subset=sex == "feminino")
head(with(.Survfit, data.frame(time, n.risk, n.event)), 5)
```

```
##   time n.risk n.event
## 1    5     90      1
## 2   60     89      1
## 3   61     88      1
## 4   62     87      1
## 5   79     86      1
```

De modo análogo, o resultado dos dois comandos a seguir são os 5 primeiros instantes de tempo em que ocorreram mortes entre os homens no conjunto de dados *cancer*, o número de homens em risco de morrer e o número de homens que morreram em cada instante, respectivamente.

```
.Survfit <- survfit(Surv(time, status, type="right") ~ 1, conf.type="log",
                    conf.int=0.95, type="kaplan-meier", error="greenwood",
                    data=cancer, subset=sex == "masculino")
head(with(.Survfit, data.frame(time, n.risk, n.event)), 5)
```

```
##   time n.risk n.event
## 1   11    138      3
## 2   12    135      1
## 3   13    134      2
## 4   15    132      1
## 5   26    131      1
```

A tabela 20.2 mostra como a tabela 20.1 é construída para os 5 menores tempos nos quais houve pelo menos uma morte no conjunto de dados. Os dois estratos são: A - homens, B - mulheres.

Tabela 20.2: Tabela simplificada, com os 5 primeiros eventos de morte no conjunto de dados *cancer*, para ilustrar a aplicação do teste *log-rank*.

tempo	Em Risco			Observados			Esperados	
	Homens	Mulheres	Total	Homens	Mulheres	Total	Homens	Mulheres
5	138	90	228	0	1	1	0,61	0,39
11	138	89	227	3	0	3	1,82	1,18
12	135	89	224	1	0	1	0,60	0,40
13	134	89	223	2	0	2	1,20	0,80
15	132	89	221	1	0	1	0,60	0,40
				<b>7</b>	<b>1</b>		<b>4,83</b>	<b>3,17</b>

A primeira linha da tabela 20.2 corresponde ao tempo  $t_1 = 5$  dias. Nesse instante, 138 homens ( $nA_1 = 138$ ) e 90 mulheres ( $nB_1 = 90$ ) estavam em risco, logo 228 pessoas estavam em risco de morte. Em  $t_1 = 5$ , houve uma morte entre as mulheres ( $nB_1 = 1$ ) e nenhuma entre os homens ( $nA_1 = 0$ ), logo o número total de mortes no instante 5 é igual a 1

( $nAB_1 = 1$ ). O número de mortes esperadas entre os homens em  $t_1 = 5$  é dado então por  $nAe_1 = 1 \times 138/228 = 0,61$ . O número de mortes esperadas entre as mulheres em  $t_1 = 5$  é dado então por  $nBe_1 = 1 \times 90/228 = 0,39$ . De modo análogo, se obtém as outras 4 linhas da tabela, correspondentes aos tempos 11, 12, 13 e 15 dias.

O número de mortes observadas e esperadas entre os homens até o instante  $t = 15$  dias foi de 7 ( $\sum nAo = 7$ ) e 4,83 ( $\sum nAe = 4,83$ ), respectivamente. Entre as mulheres, o número de mortes observadas e esperadas até o instante  $t = 15$  dias foi de 1 ( $\sum nBo = 1$ ) e 3,17 ( $\sum nBe = 3,17$ ), respectivamente. Substituindo esses valores na expressão (20.1), obtemos:

$$\chi_{lr}^2 = \frac{(7-4,83)^2}{4,83} + \frac{(1-3,17)^2}{3,17} = 2,46$$

Se adotarmos o nível de significância de 5%, esse valor de  $\chi_{lr}^2 < \chi_{95,1}^2 = 3,84$ , e não rejeitaríamos a hipótese nula. Nós usamos, porém, somente uma pequena porção do conjunto de dados *cancer*. Para fazer o teste para todos os registros do conjunto de dados *cancer*, usamos a seguinte opção no *R Commander*:

Estatísticas  $\Rightarrow$  Análise de Sobrevida  $\Rightarrow$  Compare as funções de sobrevida

Na tela *Compare Funções de Sobrevida* (figura 20.9), selecionamos a variável de tempo (*time*) e censura (*status*) e a variável de estratificação (*sex*). O valor do parâmetro  $\rho = 0$  implica que o teste será realizado como mostrado mais acima. Diferentes variantes do teste de log-rank são especificadas por outros valores para  $\rho$ . Para maiores informações, vide a ajuda da função *survdiff* do pacote *survival*.

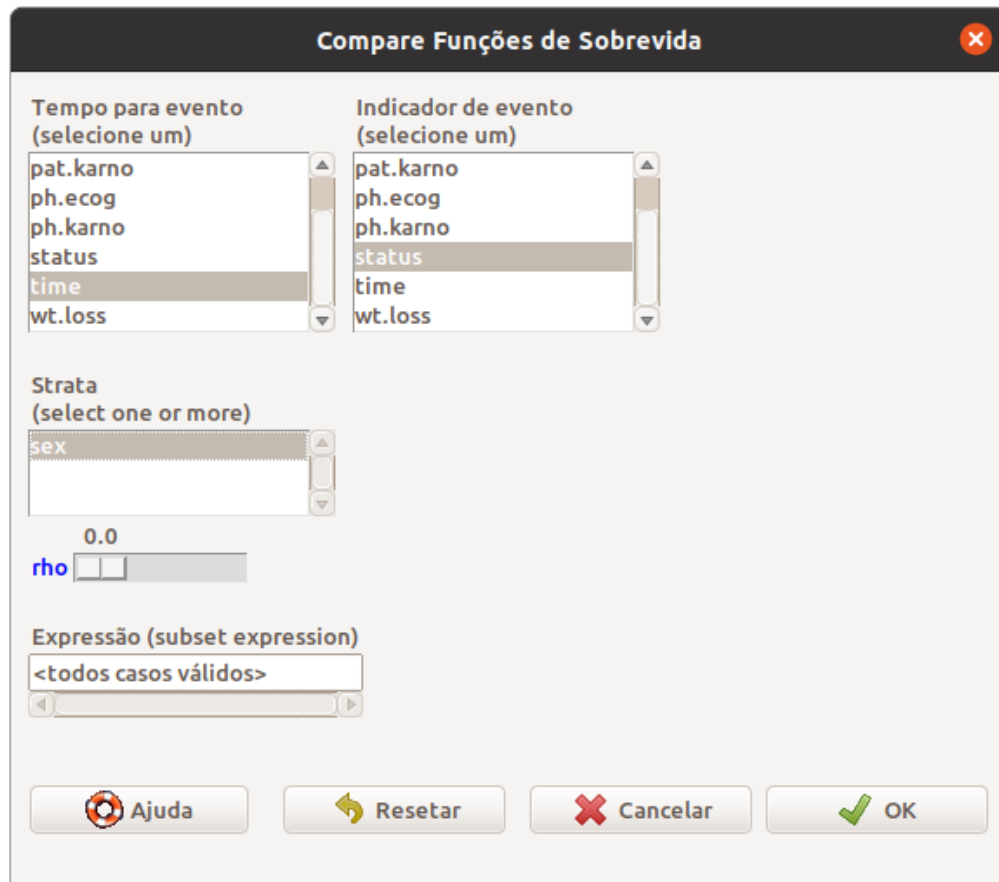


Figura 20.9: Tela para configurar as variáveis que serão utilizadas para realizar o teste *log-rank* para comparar as curvas de sobrevida para diferentes estratos.

A função a seguir é executada e os resultados são mostrados logo após.

```
survdif(Surv(time,status) ~ sex, rho=0.0, data=cancer)
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ sex, data = cancer, rho = 0)
##
##               N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=masculino 138      112    91.6      4.55     10.3
## sex=feminino  90       53    73.4      5.68     10.3
##
## Chisq= 10.3  on 1 degrees of freedom, p= 0.001
```

Houve 112 mortes entre os homens e 53 entre as mulheres no período de observação. O valor de  $\chi^2_{lr} = 10,3$ , e o valor de  $p = 0,001$ . Se adotarmos o nível de significância de 5%, a hipótese nula de igualdade das funções de sobrevida para homens e mulheres é rejeitada.

Este capítulo é apenas uma introdução ao tema. A análise de sobrevida pode ser completada com modelos que levam em conta os efeitos de várias variáveis simultaneamente sobre a

função de sobrevida. Um dos modelos mais utilizados é o modelo de riscos proporcionais de Cox. A teoria e a construção desses modelos estão fora do escopo deste texto.

O tutorial *Survival Analysis in R*, de Emily Zabor (Zabor, 2018) inclui tópicos mais avançados da análise de sobrevida usando o R, incluindo modelos de risco proporcionais de Cox e textos de referência para aqueles que desejam se aprofundar nesse assunto.

## 20.8 Exercício

- 1) Com o conjunto de dados *stagec* do pacote *rpart* ([GPL-2](#) | [GPL-3](#)) do R, faça as atividades abaixo.
  - a) Verifique a ajuda para o conjunto de dados.
  - b) Carregue o conjunto de dados.
  - c) Visualize os registros do conjunto de dados.
  - d) Obtenha a curva de sobrevida, considerando a progressão do tumor como evento de interesse.
  - e) Obtenha as probabilidades de sobrevida em 5 e 10 anos.
  - f) Obtenha as curvas de sobrevida para cada nível do status de ploidia do tumor. Comente os resultados.
  - g) Compare as curvas de sobrevida em “f” pelo teste log-rank.
  - h) Gere um relatório no R Markdown.

# Apêndice A

## Instalação do R, *RStudio* e *R Commander*

Este apêndice descreve o passo a passo para a instalação do R, de um programa que oferece um ambiente integrado de desenvolvimento baseado no R (*RStudio*), e de um pacote que fornece uma interface gráfica para a utilização do R (*R Commander*).

### A.1 O que é o R?

O R é uma linguagem e um ambiente para a realização de análises estatísticas e construção de gráficos e é altamente extensível. R é disponível como software livre sob os termos da Licença Pública Geral GNU da Free Software Foundation.

O R é um dialeto da linguagem S, desenvolvida por John Chambers e outros na empresa Bell Telephone *Laboratories*, originalmente parte da *AT&T Corp.* De acordo com Roger Peng (Peng, 2016b), a filosofia da linguagem S foi assim descrita por John Chambers:

*“[W]e wanted users to be able to begin in an interactive environment, where they did not consciously think of themselves as programming. Then as their needs became clearer and their sophistication increased, they should be able to slide gradually into programming, when the language and system aspects would become more important.”*

Uma importante limitação da linguagem S é que ela estava somente disponível em um pacote comercial, S-PLUS. O R começou a ser desenvolvido por Robert Gentleman e Ross Ihaka (“R & R”), ambos do Departamento de Estatística da Universidade de Auckland, na Nova Zelândia, em 1991.

O primeiro relato da distribuição do R foi em 1993, quando algumas cópias foram disponibilizadas no StatLib, um sistema de distribuição de softwares estatísticos.

Com o incentivo de um dos primeiros usuários deste programa, Martin Mächler (do Instituto Federal de Tecnologia de Zurique, na Suíça), “R & R”, em 1995, lançaram o código fonte do R, disponível por ftp. Em 1997, foi formado um grupo de profissionais que têm acesso ao



código fonte do R, possibilitando, assim, a atualização mais rápida do software. Desde então, o R vem ganhando cada vez mais adeptos em todo o mundo (Melo, 2017).

As seguintes referências foram utilizadas para desenvolver este material: (Peng, 2016b), (Peng, 2016a), (Peng, 2016c), (Dalgaard, 2008), (Melo, 2017), (Melo, 2019a) e (Melo, 2019b). As últimas três referências estão disponíveis neste [endereço](#).

## A.2 Vantagens do R

O R possui as seguintes características:

- além de gratuito, é um programa poderoso, estável e pode ser copiado e distribuído sem nenhum problema;
- é um programa que tem uma longa história, com mais de 25 anos de desenvolvimento;
- é apoiado por uma grande equipe de desenvolvedores em todo o mundo;
- pode ser usado nos sistemas operacionais Windows, Linux e Mac OS;
- amplamente utilizado no meio acadêmico.

As seções seguintes descrevem o passo a passo para a instalação do R, de um programa que oferece um ambiente integrado de desenvolvimento baseado no R (*RStudio*), e de um pacote que fornece uma interface gráfica para a utilização do R (*R Commander* - *Rcmdr*).

## A.3 Instalação do R e do pacote *R Commander*

Uma instalação do R contém uma ou mais bibliotecas ou pacotes. Alguns desses pacotes fazem parte da instalação básica do R. Outros podem ser baixados e instalados, à medida que for necessário.

Ao instalar um pacote, é criada uma pasta no disco do computador com o conteúdo do pacote. Você pode criar o seu próprio pacote.

Um pacote pode conter funções escritas na linguagem R, conjuntos de dados e/ou bibliotecas de códigos compilados em outras linguagens. Eles contêm funções que os usuários não irão utilizar todo o tempo.

Para um usuário iniciante no R, vamos utilizar um pacote que oferece uma interface gráfica para realizar análises estatísticas, criar gráficos, carregar, manipular, importar ou exportar conjuntos de dados. Esse pacote é chamado de *R Commander* (*Rcmdr*).

Este [vídeo](#) mostra como instalar o R e o pacote *R Commander* no sistema operacional *Windows*. Este outro [vídeo](#) fornece um breve tour dos recursos do *R Commander*.

De maneira alternativa, são apresentados a seguir os passos para a instalação do R no *Windows*. Neste exemplo, será utilizada a versão 3.5.0 do R. Utilize a última versão que encontrar.

Para instalar o R, siga os seguintes passos:

- Baixe o programa do sítio <http://cran.r-project.org/bin/windows/base/>.

- Execute o programa de instalação R-3.5.0-win.exe. Para isso, basta dar um duplo clique no arquivo (Figura A.1).

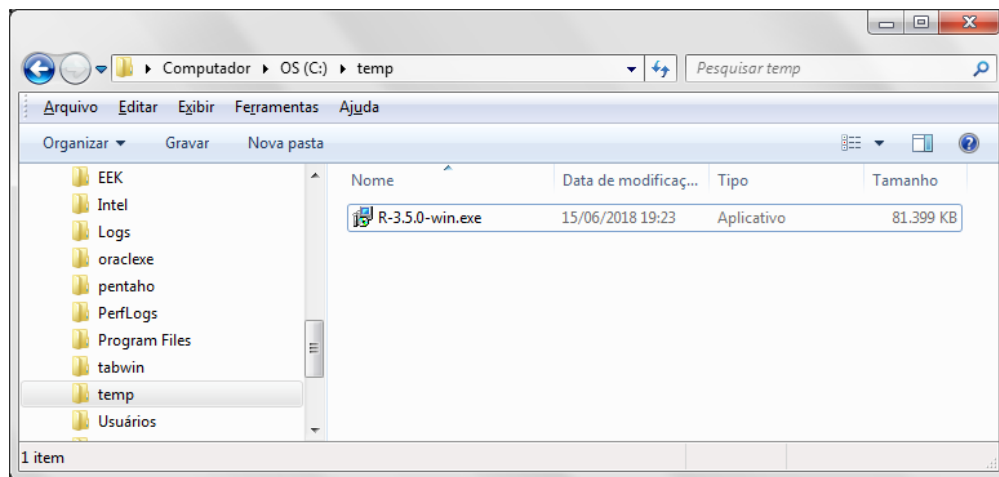


Figura A.1: Programa para a instalação do R.

- Selecione o idioma e clique em avançar nas próximas telas, aceitando as opções padrões. Ao final, será exibida a tela da figura A.2. Clique em concluir para encerrar a instalação.

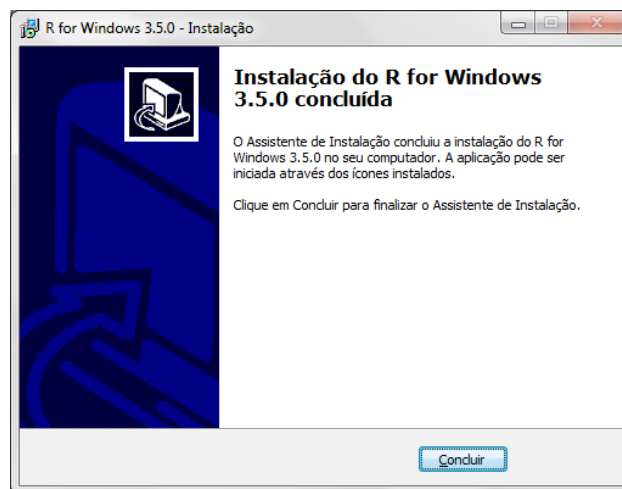


Figura A.2: Tela de encerramento da instalação do R no Windows.

- O ícone do R aparece na área de trabalho em seu computador (figura A.3).



Figura A.3: Ícone do programa R.

- Para executar o R, basta dar um duplo clique neste ícone. Surge então a tela mostrada na figura A.4.

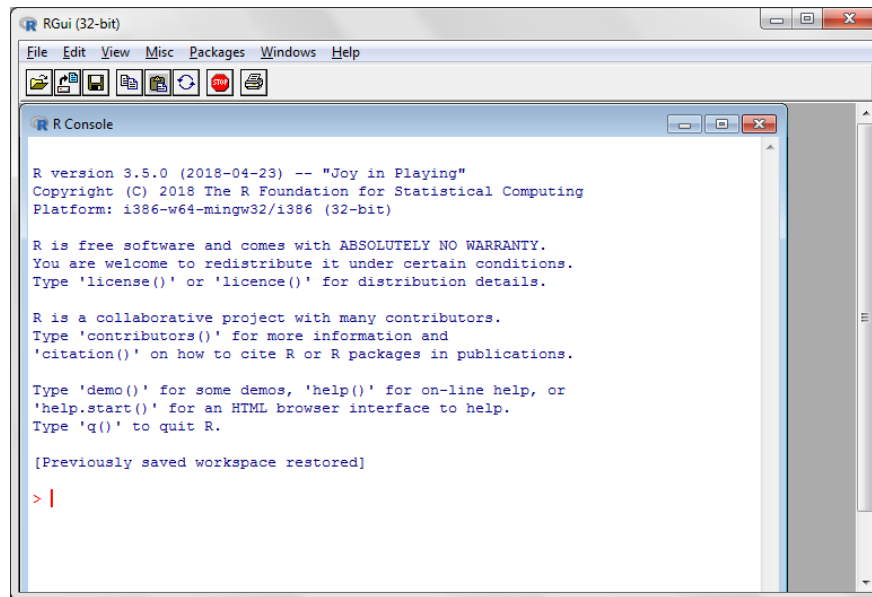


Figura A.4: Tela inicial do R.

Pronto! O R já pode ser utilizado. Apesar de o R poder ser utilizado exclusivamente a partir de sua instalação, neste livro, sempre será utilizado o *R Commander*, ou o *RStudio*, eventualmente acompanhado do R Commander.

**Observação:** Para instalar o *R Commander* no *macOS*, é necessário instalar o *XQuartz* e também o *Tcl/Tk*.

*XQuartz* é uma versão do X11 compatível com o *macOS*. X11 é um sistema gráfico para máquinas Unix.

*Tcl/Tk* é um kit de ferramentas para o desenvolvimento de aplicações *desktop*.

## A.4 Instalação do *RStudio*

O *RStudio* é um ambiente integrado de código aberto para escrever *scripts* no R e utilizar outros recursos baseados no R.

Este [vídeo](#) mostra como instalar e fornece um breve tour dos recursos do *RStudio*.

As figuras a seguir mostram o passo a passo para a instalação do *RStudio*. Existe uma versão gratuita que pode ser instalada a partir do [sítio](#). Nessa página (figura A.5), selecione o botão *Download RStudio*.

Há diversas versões do *RStudio*. Baixe a versão gratuita (figura A.6) e, em seguida, o instalador para o seu sistema operacional (figura A.7).

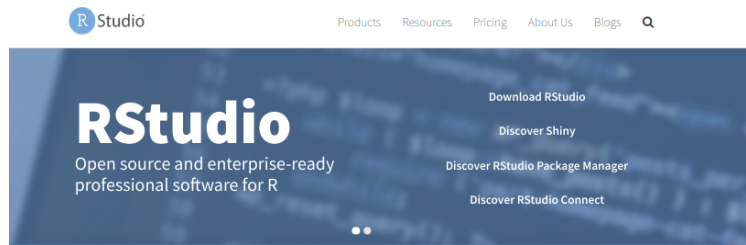


Figura A.5: Sítio do *RStudio*.

Choose Your Version of RStudio

RStudio is a set of integrated tools designed to help you be more productive with R. It includes a console, syntax-highlighting editor that supports direct code execution, and a variety of robust tools for plotting, viewing history, debugging and managing your workspace. [Learn More about RStudio features.](#)

	RStudio Desktop Open Source License	RStudio Desktop Commercial License	RStudio Server Open Source License	RStudio Server Pro Commercial License	RStudio Server Pro + RStudio Connect Commercial License
	FREE	\$995 per year	FREE	\$9,995 per year	\$29,995 per year
	<a href="#">DOWNLOAD</a> Learn More	<a href="#">BUY</a> Learn More	<a href="#">DOWNLOAD</a> Learn More	<a href="#">DOWNLOAD</a> Learn More	<a href="#">TALK</a> Learn More
Integrated Tools for R	●	●	●	●	●
Priority Support		●		●	●
Access via Web Browser			●	●	●
Enterprise Security				●	●
Project Sharing				●	●
Manage Multiple R Sessions & Versions				●	●
Admin Dashboard				●	●
Load Balancing				●	●
One-Click Publishing					●
Self-Managed Content					●
Scheduled Reports					●
License	AGPL	Commercial	AGPL	Commercial	Commercial
Pricing	FREE	\$995/yr	FREE	\$9,995/yr	\$29,995/yr
	RStudio Desktop Open Source	RStudio Desktop Commercial	RStudio Server Open Source	RStudio Server Pro	RStudio Server Pro + RStudio Connect
	<a href="#">DOWNLOAD NOW</a>	<a href="#">BUY NOW</a>	<a href="#">DOWNLOAD NOW</a>	<a href="#">DOWNLOAD NOW</a>	<a href="#">CONTACT SALES</a>

Figura A.6: Página do *RStudio* com as versões disponíveis para instalação.

#### RStudio Desktop 1.2.1335 — Release Notes

RStudio requires R 3.0.1+. If you don't already have R, download it [here](#).

Linux users may need to import RStudio's public code-signing key prior to installation, depending on the operating system's security policy.

#### Installers for Supported Platforms

Installers	Size	Date	MD5
RStudio 1.2.1335 - Windows 7+	126.9 MB	2019-04-08	d0e2470f1f8ef4cd35a669aa323a2136
RStudio 1.2.1335 - Mac OS X 10.12+ (64-bit)	121.1 MB	2019-04-08	6c570b0e2144583f7c48c284ce299eef
RStudio 1.2.1335 - Ubuntu 14/Debian 8 (64-bit)	92.2 MB	2019-04-08	c1b07d0511469abfe582919b183eee83
RStudio 1.2.1335 - Ubuntu 16 (64-bit)	99.3 MB	2019-04-08	c142d69c210257fb10d18c045fff13c7
RStudio 1.2.1335 - Ubuntu 18 (64-bit)	100.4 MB	2019-04-08	71a8d1990c0d97939804b46cfb0aea75
RStudio 1.2.1335 - Fedora 19+/RedHat 7+ (64-bit)	114.1 MB	2019-04-08	296b6ef88969a91297fab6545f256a7a
RStudio 1.2.1335 - Debian 9+ (64-bit)	100.6 MB	2019-04-08	1e32d4d6f6e216f086a81ca82ef65a91
RStudio 1.2.1335 - OpenSUSE 15+ (64-bit)	101.6 MB	2019-04-08	2795a63c7efd8e2aa2dae86ba09a81e5
RStudio 1.2.1335 - SLES/OpenSUSE 12+ (64-bit)	94.4 MB	2019-04-08	c65424b06ef6737279d982db9eefcael

#### Zip/Tarballs

Zip/tar archives	Size	Date	MD5
RStudio 1.2.1335 - Windows 7+	186.6 MB	2019-04-08	f1e013ade0c241969400507cf258e0ad
RStudio 1.2.1335 - Ubuntu 14/Debian 8 (64-bit)	137.6 MB	2019-04-08	e3e1ea2dd113fd9cfd40bc5035effdde
RStudio 1.2.1335 - Ubuntu 18 (64-bit)	147.8 MB	2019-04-08	5ee7dd7b501675f0a631c62d403ea1b6
RStudio 1.2.1335 - Debian 9+ (64-bit)	148.1 MB	2019-04-08	8090451cb7d520633eba80fd355ad4c1
RStudio 1.2.1335 - Fedora 19+/RedHat 7+ (64-bit)	147.2 MB	2019-04-08	34630cd7c66c3429879bd79982349380

#### Source Code

A tarball containing source code for RStudio v1.2.1335 can be downloaded from [here](#)

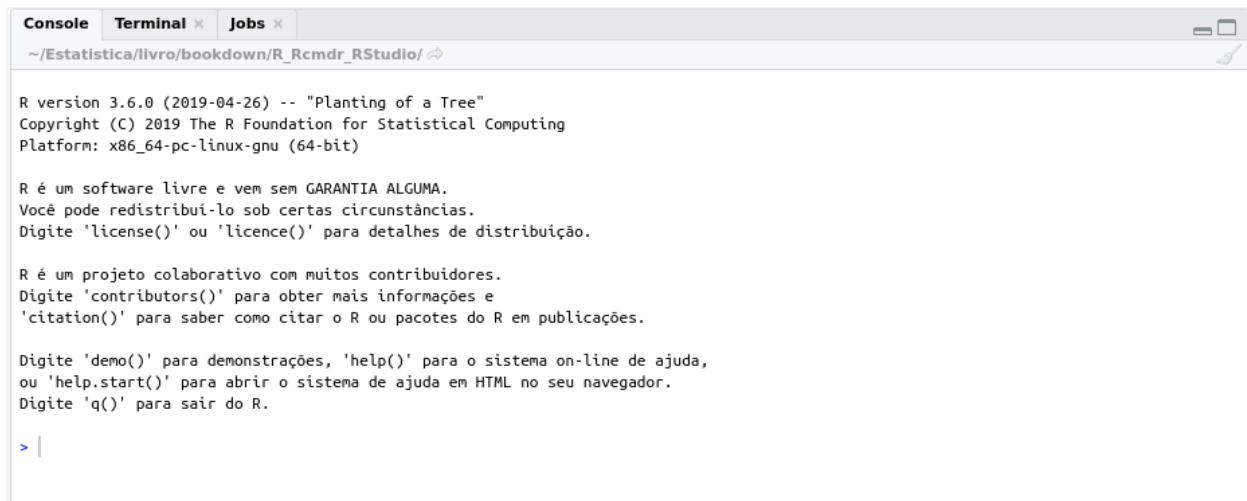
Figura A.7: Instaladores disponíveis para o *RStudio*.

Ao baixar o instalador, basta executá-lo que o programa será instalado. Após a instalação, para executar o *RStudio*, basta selecioná-lo na lista de aplicações ou clicar em seu ícone na área de trabalho. Ao iniciar o *RStudio*, automaticamente os pacotes básicos do R são carregados, não sendo necessário executar o R a partir do menu ou área de trabalho.

O *RStudio* disponibiliza uma série de resumos de como utilizar funções de diversos pacotes do R: [RStudio cheat sheets](#).

## A.5 Console do RStudio

Ao executarmos o *RStudio*, a sua console, primeira aba da janela inferior à esquerda (figura A.8), permite ao usuário digitar, executar os comandos e visualizar os resultados. Na console, são exibidas a versão do R utilizada e algumas informações sobre o R e, na parte inferior, o sinal “>”. Este símbolo é chamado *prompt* de comando e significa que o R está apto a receber um comando nessa linha. Por essa razão, essa linha é chamada linha de comando.



```
Console Terminal Jobs
~/Estatistica/livro/bookdown/R_Rcmdr_RStudio/

R version 3.6.0 (2019-04-26) -- "Planting of a Tree"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R é um software livre e vem sem GARANTIA ALGUMA.
Você pode redistribuí-lo sob certas circunstâncias.
Digite 'license()' ou 'licence()' para detalhes de distribuição.

R é um projeto colaborativo com muitos contribuidores.
Digite 'contributors()' para obter mais informações e
'citation()' para saber como citar o R ou pacotes do R em publicações.

Digite 'demo()' para demonstrações, 'help()' para o sistema on-line de ajuda,
ou 'help.start()' para abrir o sistema de ajuda em HTML no seu navegador.
Digite 'q()' para sair do R.

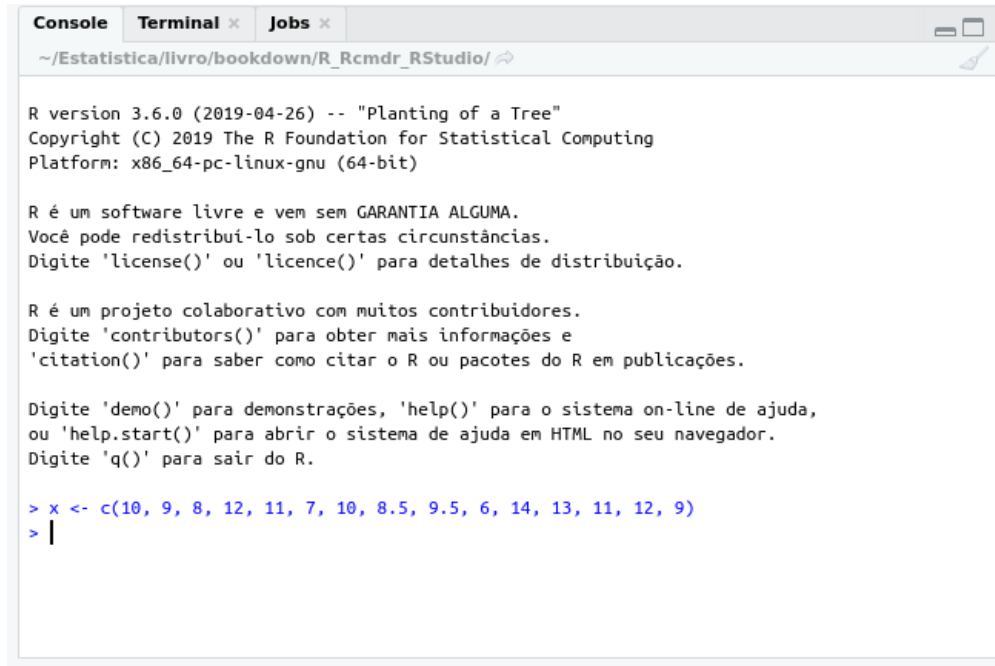
> |
```

Figura A.8: Console do *RStudio* com o *prompt* (`>`) e o cursor aguardando a digitação de um comando.

Para executarmos um comando, basta digitá-lo e, em seguida, apertamos a tecla *Enter*. O resultado do comando é o que chamamos de *output* ou saída. Neste texto, os comandos serão mostrados sem o *prompt* e numa área sombreada e o resultado da execução do comando é mostrado a seguir precedido de “##”, quando houver resultados a serem exibidos.

A figura A.9 mostra o comando `x <- c(10, 9, 8, 12, 11, 7, 10, 8.5, 9.5, 6, 14, 13, 11, 12, 9)` sendo executado na console do *RStudio*. Esse comando cria uma variável `x` (nome escolhido arbitrariamente), com 15 elementos, que são os números colocados entre parênteses e separados por vírgulas. Assim a função “`c`” gera um vetor a partir dos elementos entre parênteses. Observem o uso do “`<-`” para alocar os valores a uma variável. Essa forma é a clássica do R. A versão atual do R aceita também o “`=`”, como usado em outras linguagens de programação.

Para visualizarmos os valores da variável `x`, basta digitarmos `x` na linha de comando e pressionarmos a tecla *Enter* (Figura A.10). O R mostra os valores da variável. O número 1 entre colchetes no início “[1]” indica apenas que o elemento seguinte é o primeiro valor do vetor `x`.



```
Console Terminal x Jobs x
~/Estatistica/livro/bookdown/R_Rcmdr_RStudio/

R version 3.6.0 (2019-04-26) -- "Planting of a Tree"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

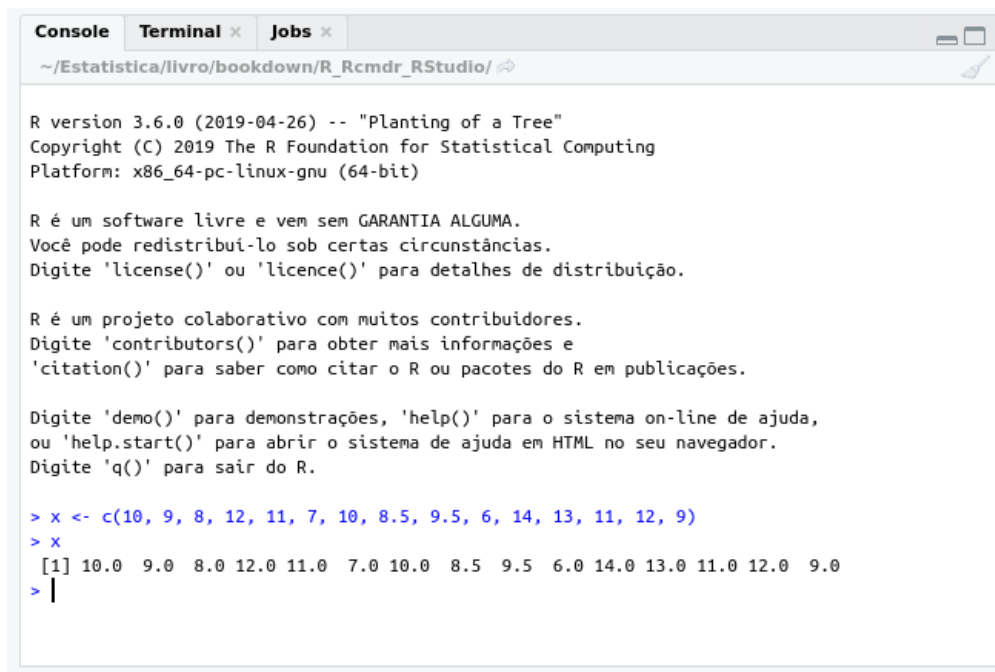
R é um software livre e vem sem GARANTIA ALGUMA.
Você pode redistribuí-lo sob certas circunstâncias.
Digite 'license()' ou 'licence()' para detalhes de distribuição.

R é um projeto colaborativo com muitos contribuidores.
Digite 'contributors()' para obter mais informações e
'citation()' para saber como citar o R ou pacotes do R em publicações.

Digite 'demo()' para demonstrações, 'help()' para o sistema on-line de ajuda,
ou 'help.start()' para abrir o sistema de ajuda em HTML no seu navegador.
Digite 'q()' para sair do R.

> x <- c(10, 9, 8, 12, 11, 7, 10, 8.5, 9.5, 6, 14, 13, 11, 12, 9)
> |
```

Figura A.9: Console do *RStudio*, após a execução do comando  $x \leftarrow c(10, 9, 8, 12, 11, 7, 10, 8.5, 9.5, 6, 14, 13, 11, 12, 9)$ .



```
Console Terminal x Jobs x
~/Estatistica/livro/bookdown/R_Rcmdr_RStudio/

R version 3.6.0 (2019-04-26) -- "Planting of a Tree"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R é um software livre e vem sem GARANTIA ALGUMA.
Você pode redistribuí-lo sob certas circunstâncias.
Digite 'license()' ou 'licence()' para detalhes de distribuição.

R é um projeto colaborativo com muitos contribuidores.
Digite 'contributors()' para obter mais informações e
'citation()' para saber como citar o R ou pacotes do R em publicações.

Digite 'demo()' para demonstrações, 'help()' para o sistema on-line de ajuda,
ou 'help.start()' para abrir o sistema de ajuda em HTML no seu navegador.
Digite 'q()' para sair do R.

> x <- c(10, 9, 8, 12, 11, 7, 10, 8.5, 9.5, 6, 14, 13, 11, 12, 9)
> x
[1] 10.0  9.0  8.0 12.0 11.0  7.0 10.0  8.5  9.5  6.0 14.0 13.0 11.0 12.0  9.0
> |
```

Figura A.10: Visualização dos valores da variável  $x$ .

## A.6 Instalação do pacote do *R Commander* a partir do *RStudio*

É possível instalar o pacote *R Commander*, e qualquer outro pacote do R, a partir do *RStudio*. Caso já tenha instalado o *R Commander* na seção A.3, não é necessário executar os passos mostrados abaixo, mas aconselhamos a leitura para entender como instalar um pacote do R a partir do *RStudio*.

Para instalar o *R Commander*, ou qualquer outro pacote, a partir do *RStudio*, seguimos os passos abaixo:

- Executamos o *RStudio*. A tela de entrada do *RStudio* é mostrada na figura A.11.

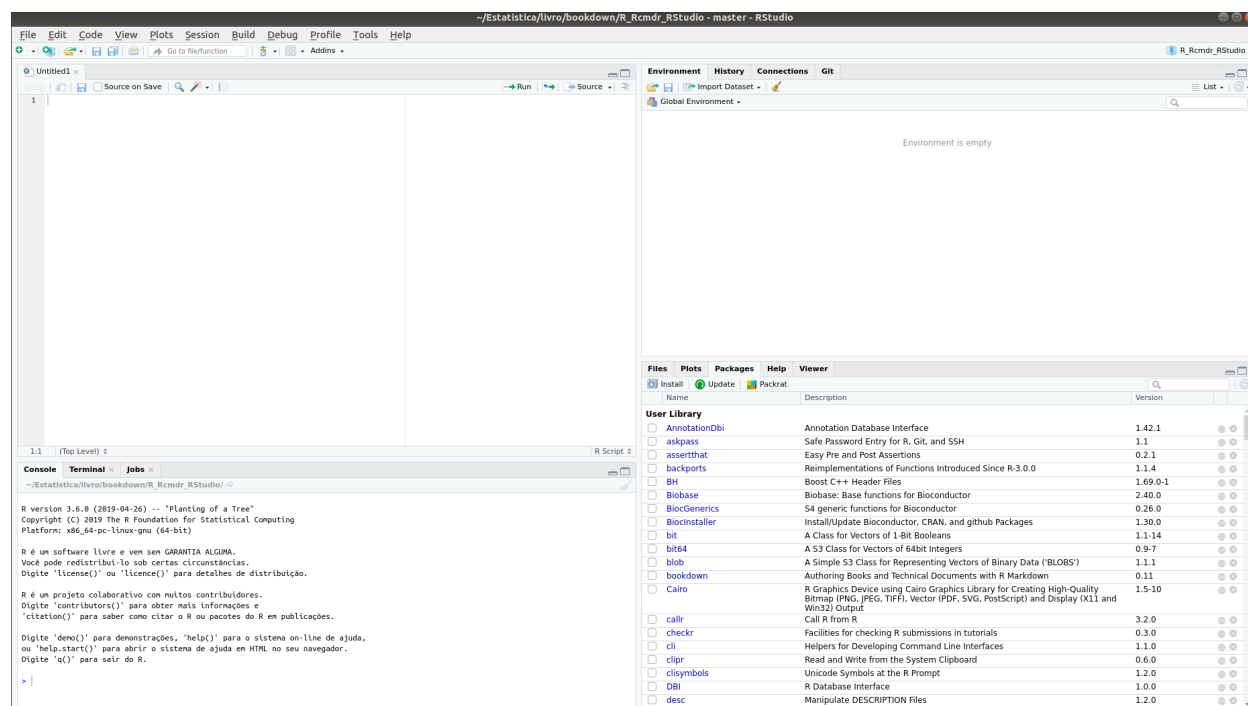


Figura A.11: Tela de entrada do *RStudio*.

- Clicamos na aba *packages* e, em seguida, no botão *Install* (figura A.12).



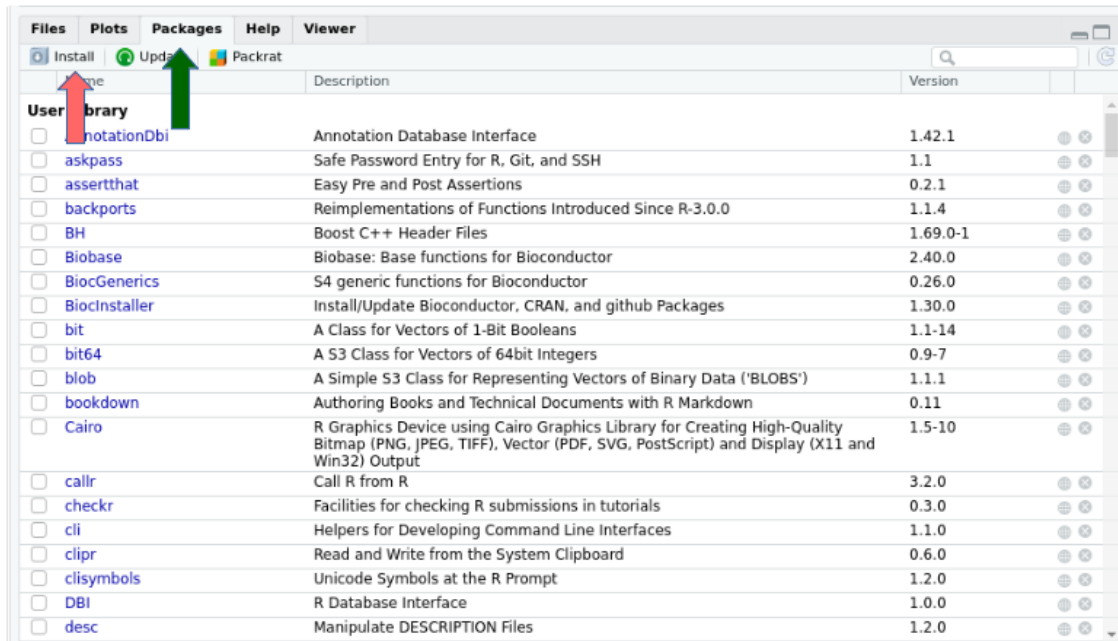


Figura A.12: Para instalar um pacote, clicamos na aba *Packages* (seta verde) e, em seguida, no botão *Install* (seta vermelha).

- Na caixa de diálogo *Install packages*, começamos a digitar *Rcmdr* na caixa de texto *Packages*. Ao iniciarmos a digitação, uma lista suspensa mostra opções de pacotes. Selecionamos *Rcmdr* e clicamos no botão *Install* (figura A.13). A instalação será inicializada e pode demorar um tempo. O progresso da instalação irá sendo exibido na janela da *Console* (canto inferior esquerdo do *RStudio*). Aguardamos até o sinal de *prompt* (>) aparecer na parte inferior da *console*.

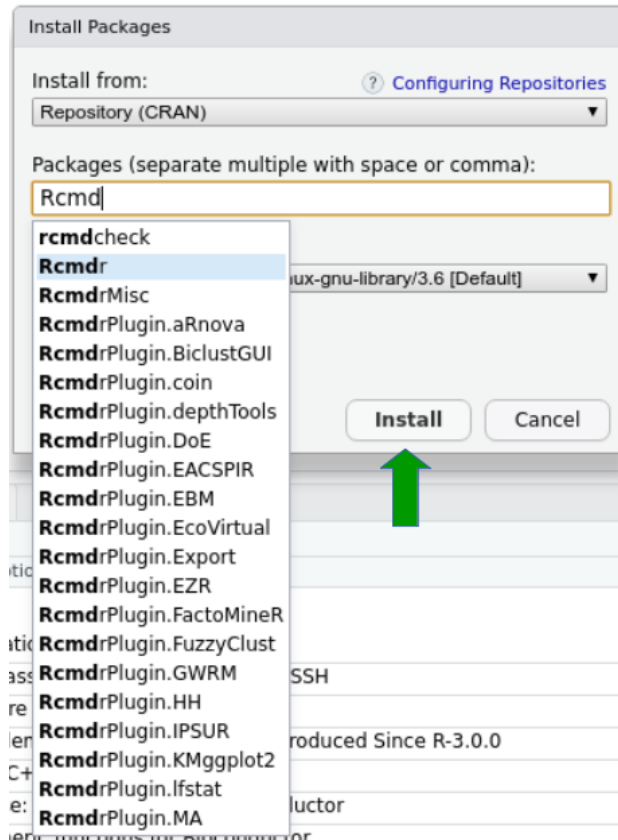


Figura A.13: Para instalar o *R Commander*, digitamos *Rcmdr* na caixa de texto *Packages* e, em seguida, clicamos no botão *Install* (seta verde).

- Após a instalação, para carregarmos o *R Commander*, digitamos o comando `library(Rcmdr)` após o sinal de prompt na console do *RStudio* (figura A.14) e pressionamos a tecla *Enter*.

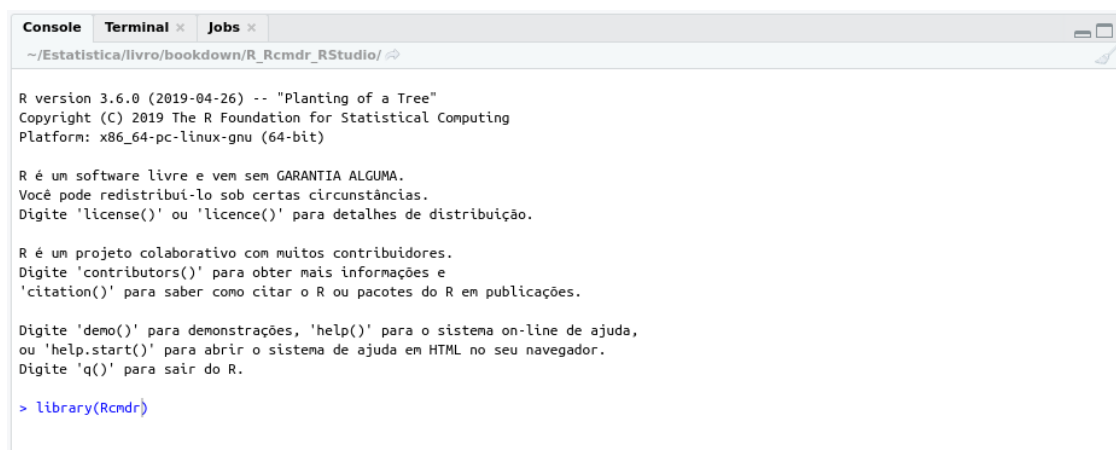


Figura A.14: Comando para o carregamento do *R Commander* a partir do *RStudio*.

- Ao iniciarmos o carregamento do *R Commander*, pode acontecer de aparecer a tela

mostrada na figura A.15, indicando que alguns pacotes estão faltando para carregar o *Rcmdr*. Nesse caso, selecionamos *Sim* e, na tela seguinte (figura A.16), pressionamos OK. Após alguns instantes, os pacotes faltantes estarão instalados e o *R Commander* será inicializado.

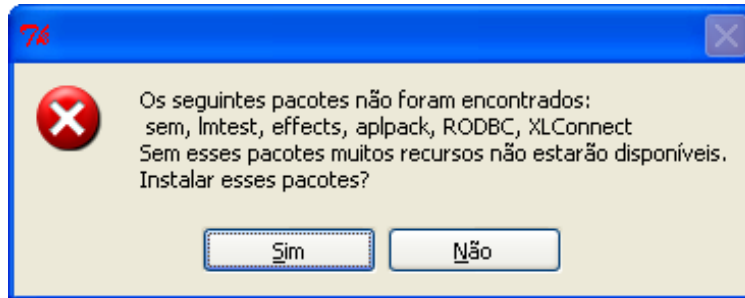


Figura A.15: Mensagem que solicita a instalação de alguns pacotes por ocasião da primeira vez que o *R Commander* é executado.

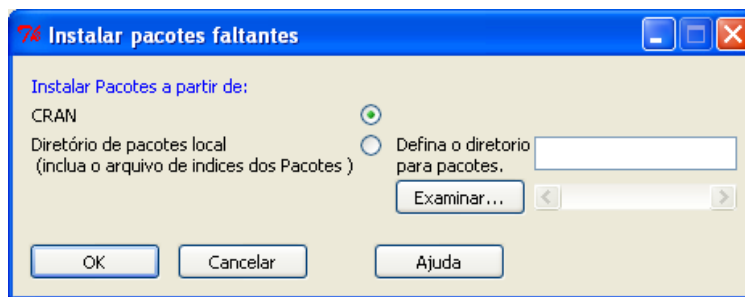


Figura A.16: Tela de definição do local onde os pacotes dos quais o *R Commander* depende precisam ser obtidos. Utilizaremos a opção padrão.

- A figura A.17 mostra a tela principal do *R Commander* quando o mesmo é carregado pelo *RStudio*.

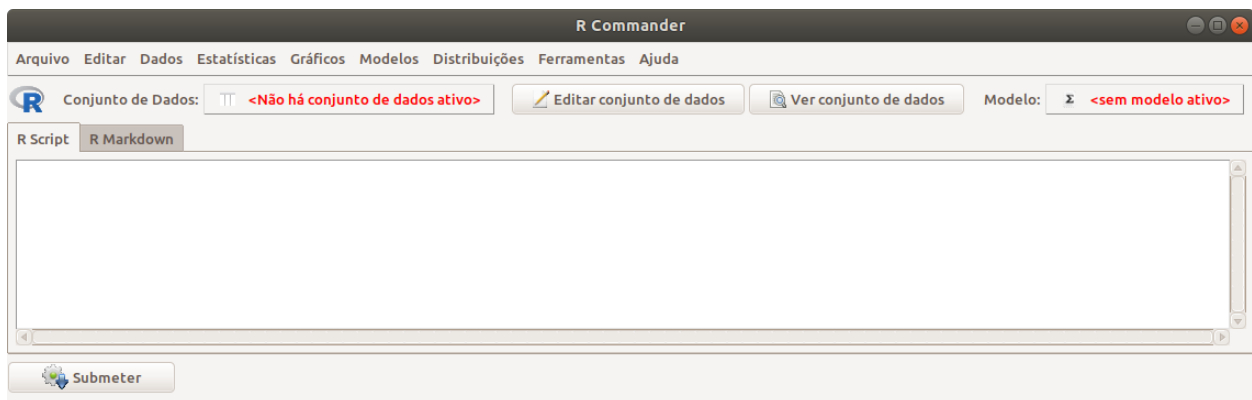


Figura A.17: Tela principal do *R Commander*.

A execução de comandos via *R Commander* é um pouco diferente do que no *RStudio*. Na aba *R Script*, digitamos o comando e clicamos no botão *Submeter* (figura A.18). Os resultados da execução do comando (se houver), e eventuais mensagens de erro irão ser mostradas na console do *RStudio* (figura A.19). Nesse caso, não houve nenhum resultado a ser exibido, nem mensagens de erro.



Figura A.18: Execução de um comando na interface gráfica do *R Commander*. Digitamos o comando e clicamos no botão *Submeter* (seta verde).

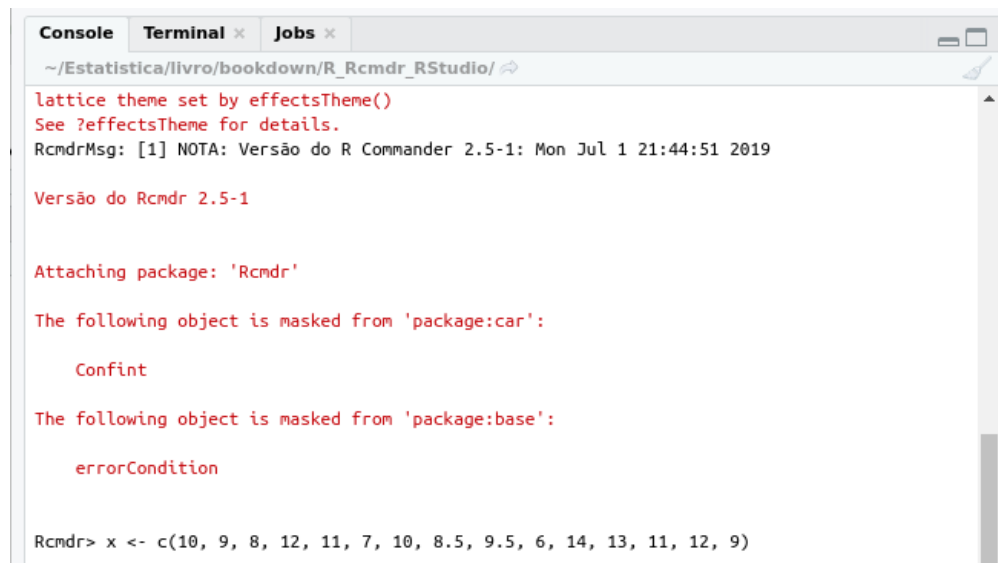


Figura A.19: Exibição do comando executado via *R Commander* e eventuais resultados na console do *RStudio*.

Após digitarmos *x* na janela da aba *R Script* e clicarmos no botão *Submeter* (A.20), os valores da variável *x* são exibidos na console do *RStudio* (figura A.21).

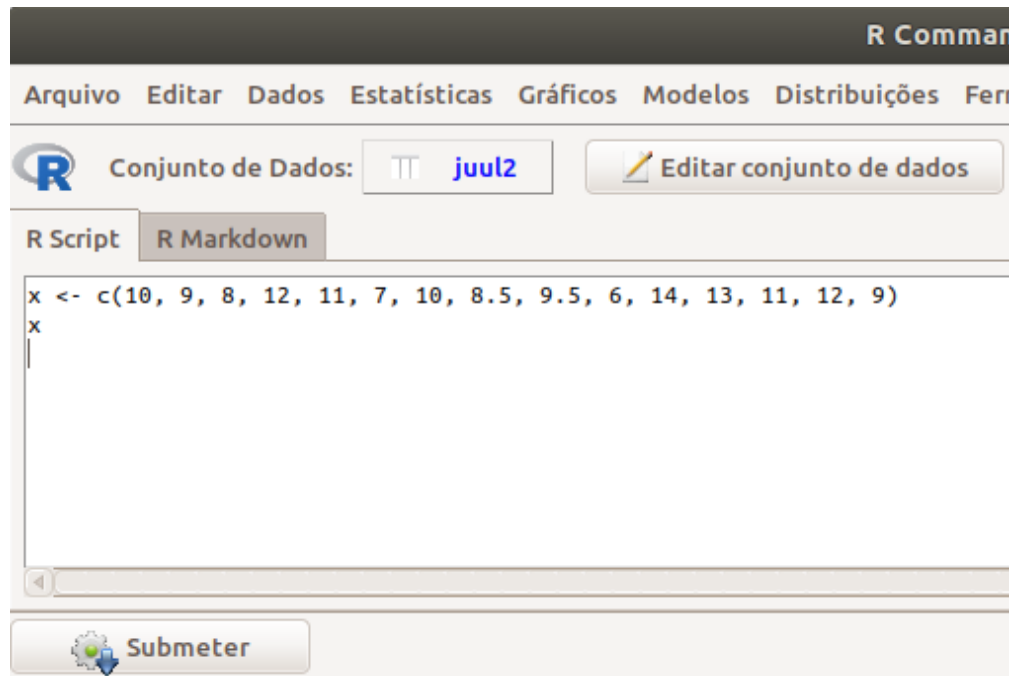


Figura A.20: Comando para exibir o conteúdo da variável *x* no *R Commander*.

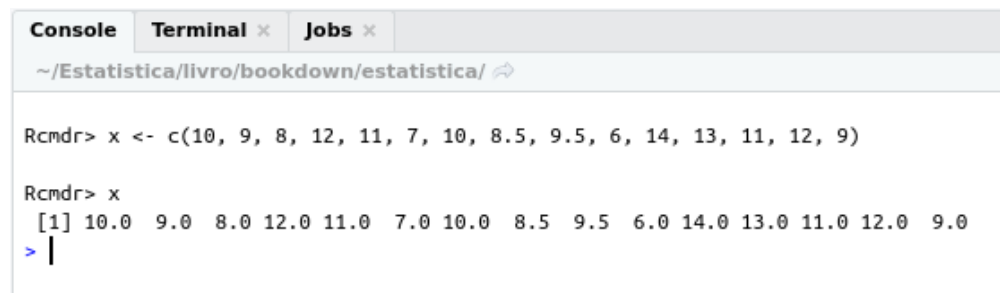


Figura A.21: Exibição na console do *RStudio* dos valores da variável *x*, a partir do comando executado no *R Commander*.

### Observação:

O *R Commander* pode ser também utilizado a partir da tela inicial do R, sem utilizar o *RStudio*. Para isso, digitamos o comando `library(Rcmdr)` após o sinal de *prompt* na tela de entrada do R (figura A.22) e pressionamos a tecla *Enter*. Nesse caso, duas outras seções serão exibidas na tela do *R Commander*: Resultados e Mensagens. Os gráficos serão exibidos na janela do R.

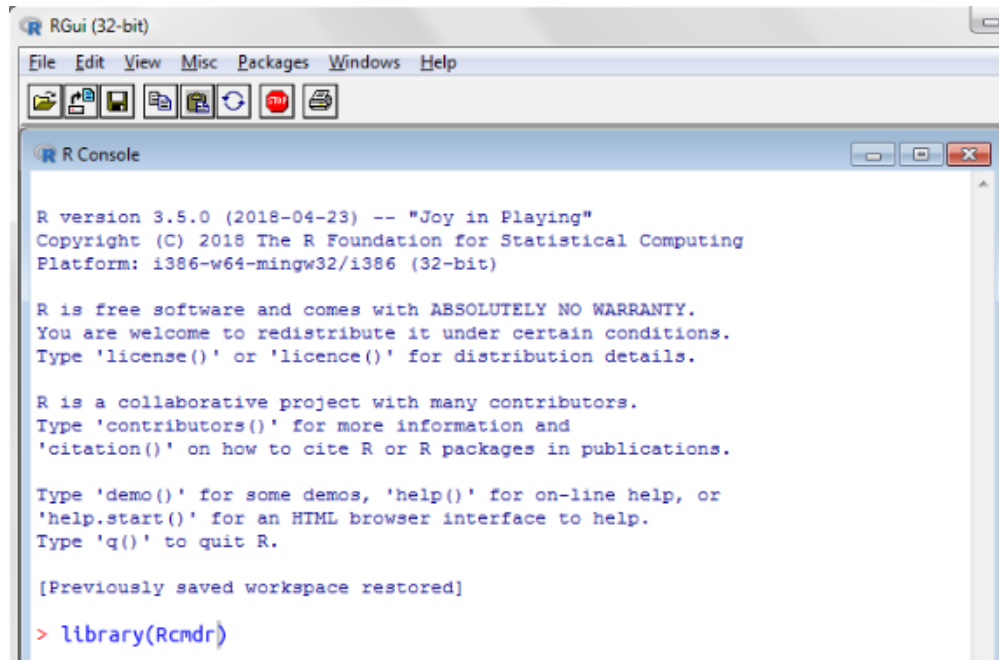


Figura A.22: Carregamento do *R Commander* a partir da tela inicial do R.

Apesar de o R poder ser utilizado exclusivamente a partir de sua instalação, neste livro, sempre será utilizado o *R Commander*, ou o *RStudio*, eventualmente acompanhado do *R Commander*.

Este [vídeo](#) mostra como utilizar o *RStudio* em conjunto com o *R Commander*.

## Apêndice B

### Código da função `paired_proportions`

Abaixo é exibido o código da função `paired_proportions`, utilizada no capítulo 17, seção 17.3.3.

```
library(dplyr)
paired_proportions <- function(data, id, row, col, row_ref, row_trt, col_ref,
                               col_out, case_control = FALSE, alpha = 0.05, ...) {
  id_subj99 = id
  summary = mutate(data,
                    cell = case_when(
                      data[, row] == row_ref & data[, col] == col_ref ~ 1,
                      data[, row] == row_trt & data[, col] == col_ref ~ 2,
                      data[, row] == row_ref & data[, col] == col_out ~ 4,
                      data[, row] == row_trt & data[, col] == col_out ~ 8,
                      TRUE ~ 0)) %>% group_by(data[, id_subj99]) %>%
    summarize(cell_class = sum(cell))
  tab = table(summary$cell_class)
  if (case_control == TRUE) {
    #
    # tab["5"] -> (row_ref in both),
    # tab["6"] -> (row_ref in col_out and row_trt in col_ref),
    # tab["9"] -> (row_trt in col_out and row_ref in col_ref),
    # tab["10"] -> (row_trt in both),
    #
    if (is.na(tab["5"])) tab["5"] = 0
    if (is.na(tab["6"])) tab["6"] = 0
    if (is.na(tab["9"])) tab["9"] = 0
    if (is.na(tab["10"])) tab["10"] = 0
    mat = matrix(c(tab["10"], tab["6"], tab["9"], tab["5"]), nrow = 2)
    colnames(mat) = c(row_trt, row_ref)
    rownames(mat) = c(row_trt, row_ref)
    names(dimnames(mat)) = c(col_out, col_ref)
  } else {
```

```

#
# tab["3"] -> (col_ref in both),
# tab["6"] -> (col_ref in row_trt and col_out in row_ref),
# tab["9"] -> (col_out in row_trt and col_ref in row_ref),
# tab["12"] -> (col_out in both),
#
if (is.na(tab["3"])) tab["3"] = 0
if (is.na(tab["6"])) tab["6"] = 0
if (is.na(tab["9"])) tab["9"] = 0
if (is.na(tab["12"])) tab["12"] = 0
mat = matrix(c(tab["12"], tab["6"], tab["9"], tab["3"]), nrow = 2)
colnames(mat) = c(col_out, col_ref)
rownames(mat) = c(col_out, col_ref)
names(dimnames(mat)) = c(row_trt, row_ref)
}
#
# IC proportion differences - Wald with Bonett-Price Laplace adjustment
#
n = sum(mat)
r = mat[1,1]
t = mat[2,1]
s = mat[1,2]
n1. = r+s
n.1 = r+t
p2 = n1. / n
p1 = n.1 / n
p2_ = (s + 1) / (n + 2)
p1_ = (t + 1) / (n + 2)
z = qnorm(1-alpha/2)
D = p2 - p1
Dinf = max(-1, (p2_-p1_) - z * sqrt((p2_+p1_-(p2_-p1_)^2)/(n+2)))
Dsup = min((p2_-p1_) + z * sqrt((p2_+p1_-(p2_-p1_)^2)/(n+2)), 1)
prop_diff = c(D, Dinf, Dsup)
#
# IC odds ratio - Transformed Wilson score
#
or_est = s/t
nd = t+s
div = 2*(nd+z^2)
num1 = (2*s + z^2)
num2 = z*sqrt(z^2 + 4*s*(1-s/nd))
pu = (num1 + num2) / div
pi = (num1 - num2) / div
or = c(or_est, pi/(1-pi), pu/(1-pu))

```



```

#
# IC relative risk - Bonett-Price hybrid Wilson score
#
rr_est = p2/p1
nn = r + t + s
A = sqrt((nd + 2) / ((n1.+1)*(n.1+1)))
B = sqrt((1-(n1.+1)/(nn+2))/(n1.+1))
C = sqrt((1-(n.1+1)/(nn+2))/(n.1+1))
z = A/(B+C) * z
den = 2*(nn+z^2)
l1 = 2*n1.+z^2 - z*sqrt(z^2+4*n1.*(1-n1./nn))
u1 = 2*n1.+z^2 + z*sqrt(z^2+4*n1.*(1-n1./nn))
l2 = 2*n.1+z^2 - z*sqrt(z^2+4*n.1*(1-n.1/nn))
u2 = 2*n.1+z^2 + z*sqrt(z^2+4*n.1*(1-n.1/nn))
rr = c(rr_est, l1/u2, u1/l2)
print(mat)
print(mcnemar.test(mat, ...))
text_low = paste("Lower ", sprintf("%.0f", (1-alpha)*100), "% CI",
                 sep = ' ')
text_upper = paste("Upper ", sprintf("%.0f", (1-alpha)*100), "% CI",
                   sep = ' ')
if (case_control == FALSE) {
  names(prop_diff) = c("proportion differences", text_low, text_upper)
  print(prop_diff)
  cat("", sep = "\n")
  names(rr) = c("relative risk", text_low, text_upper)
  print(rr)
  cat("", sep = "\n")
}
names(or) = c("odds ratio", text_low, text_upper)
print(or)
cat("", sep = "\n")
}

```

# Referências Bibliográficas

- Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. John Wiley & Sons, Inc, New York.
- Altman, D. G. (1998). Statistical reviewing for medical journals. *Statistics in Medicine*, 17:2661–2674.
- Amess, J. A. L., Burman, J. F., Rees, G. M., Nancekievill, D. G., and Mollin, D. L. (1978). Megaloblastic haemopoiesis in patients receiving nitrous oxide. *Lancet*, 312(8085):339–342.
- Andrade, M. V. S., Andrade, L. A. P., Bispo, A. F., Freitas, L. d. A., Andrade, M. Q. S., Feitosa, G. S., and Feitosa-Filho, G. S. (2018a). Avaliação da Intensidade de Sangramento de Procedimentos Odontológicos em Pacientes Anticoagulados com Varfarina ou Dabigatrana. *Arquivos Brasileiros de Cardiologia*, 111(3):394–399.
- Andrade, V. G., Yamashiro, F. S., Oliveira, C. V., Moreira, A., Winckler, F. C., and Silva, G. F. (2018b). Insulin resistance reduction after sustained virological response with direct acting antiviral: not every population improves. *Arquivos de Gastroenterologia*, 55(3):274–278.
- Barata, C. B. and Valete, C. O. S. (2018). Perfil clínico-epidemiológico de 106 pacientes pediátricos portadores de urolitíase no Rio de Janeiro. *Revista Paulista de Pediatria*, 36(3):261–267.
- Bernstam, E. V., Smith, J. W., and Johnson, T. R. (2010). What is biomedical informatics? *Journal of Biomedical Informatics*, 43(1):104.
- Bero, L. and Rennie, D. (1996). Influences on the quality of published drug studies. *International journal of technology assessment in health care*, 12(2):209–337.
- Bezerra, S. M. F. M. C., Sotto, M. N., Orii, N. M., Alves, C., and Duarte, A. J. S. (2011). Efeitos da radiação solar crônica prolongada sobre o sistema imunológico de pescadores profissionais em Recife, Brasil. *Anais Brasileiros de Dermatologia*, 86(2):222–233.
- Bittencourt, H. S., Reis, H. F. C., Lima, M. S., and Neto, M. G. (2017). Ventilação Não Invasiva em Pacientes com Insuficiência Cardíaca: Revisão Sistemática e Meta-Análise. *Arq Bras Cardiol*, 108(2):161–168.
- Brindle, R., Williams, O. M., Davies, P., Harris, T., Jarman, H., Hay, A. D., and Featherstone,

- P. (2017). Adjunctive clindamycin for cellulitis: a clinical trial comparing flucloxacillin with or without clindamycin for the treatment of limb cellulitis. *BMJ Open*, 7:e013260.
- Campbell, I. (2007). Chi-squared and Fisher–Irwin tests of two-by-two tables with small sample recommendations. *Statistics in Medicine*, 26:3661–3675.
- Costa Neto, P. L. d. O. (1977). *Estatística*. Editora Edgard Blücher LTDA.
- Cruz, R. N., Retzlaff, G., Gomes, R. Z., and Reche, P. M. (2015). Influência do diabetes mellitus sobre a perviedade da fístula arteriovenosa para hemodiálise. *J Vasc Bras*, 14(3):217–223.
- Dalgaard, P. (2008). *Introductory Statistics with R*. Springer.
- Davies, H. T. O., Crombie, I. K., and Tavakoli, M. (1998). When can odds ratio mislead? *BMJ*, 316:989–991.
- Dawson, B. and Trapp, R. (2001). *Bioestatística Básica e Clínica*. McGraw-Hill Interamericana do Brasil Ltd, Rio de Janeiro - Brazil, 3 edition.
- Durmus, E., Kivrak, T., Gerin, F., Sunbul, M., Sari, I., and Erdogan, O. (2015). Relações Neutrófilo-Linfócito e Plaqueta-Linfócito Como Preditores de Insuficiência Cardíaca. *Arquivos Brasileiros de Cardiologia*, 105(6):606–613.
- Fagerland, M., Lydersen, S., and Laake, P. (2014). Recommended tests and confidence intervals for paired binomial proportions. *Statistics in Medicine*, 33:2850–2875.
- Farias da Guarda, S. N., Santos, J. P. S., Reis, M. S. M., Passos, R. d. H., Correia, L. C., Caldas, J. R., Gobatto, A. L. N., Teixeira, M., Oliveira, A., Ribeiro, M. P., Batista, P. B. P., Calderaro, M., Paschoal Junior, F., Pontes-Neto, O. M., and Ramos, J. G. R. (2021). Realistic simulation is associated with healthcare professionals’ increased self-perception of confidence in providing acute stroke care: a before-after controlled study. *Arquivos de Neuro-Psiquiatria*, 79(1):2–7.
- Fernandes-Taylor, S., Hyun, J. K., Reeder, R. N., and Harris, A. H. S. (2011). Common statistical and research design problems in manuscripts submitted to high-impact medical journals. *BMC Research Notes*, 4:304.
- Fleiss, J. (1981). *Statistical Methods for Rates and Proportions*. John Wiley & Sons, New York, 2 edition.
- Fletcher, R. H., Fletcher, S. W., and Fletcher, G. S. (2014). *Epidemiologia Clínica - Elementos Essenciais*. Artmed, Porto Alegre, 5 edition.
- Furuta, S. E., Weckx, L. L. M., and Figueiredo, C. R. (2017). Estudo clínico, duplo-cego, randomizado, em crianças com amigdalites recorrentes submetidas a tratamento homeopático. *Revista de Homeopatia*, 80(1/2):164–173.
- Furuta, S. E., Weckx, L. M., and Figueiredo, C. R. (2003). Estudo clínico, randomizado, duplo-cego, em crianças com adenóide obstrutiva, submetidas a tratamento homeopático. *Revista Brasileira de Otorrinolaringologia*, 69(3 parte 1):343–347.

- Greenwood, M. (1926). *A report on the natural duration of cancer*. Number 33 in Reports on Public Health and Medical Subject. H. M. Stationery Office, London.
- Griffiths, D. (2008). *Head First Statistics*. O'Reilly.
- Grippe, T. C., Allam, N., Brandão, P. R. P., Pereira, D. A., Cardoso, F. E. C., Aguilar, A. C. R., and Kessler, I. M. (2018). Is transcranial sonography useful for diagnosing Parkinson's disease in clinical practice? *Arq Neuro-Psiquiatr*, 76(7):459–466.
- Guyatt, G., Rennie, D., Meade, M. O., and Cook, D. J. (2008). *Users' Guide to the Medical Literature. Essentials of Evidence-Based Clinical Practice*. McGraw-Hill, 2 edition.
- Hadorn, D. C., Baker, D., Hodges, J. S., and Hicks, N. (1996). Rating the Quality of Evidence for Clinical Practice Guidelines. *Journal of Clinical Epidemiology*, 49(7):749–754.
- Haijanen, J., Sippola, S., Tuominen, R., Grönroos, J., Paaianen, H., Rautio, T., Nordström, P., Aarnio, M., Rantanen, T., Hurme, S., and Salminen, P. (2019). Cost analysis of antibiotic therapy versus appendectomy for treatment of uncomplicated acute appendicitis: 5-year results of the APPAC randomized clinical trial. *PLoS ONE*, 14(7):e0220202.
- Harding, D. (1996). The range of a set of data. *Teaching Statistics*, 18(1):81.
- Higgins, J., Altman, D. G., Gotzsche, P., Jüni, P., Moher, D., Oxman, A., Savovic, J., Schulz, K., Weeks, L., and Sterne, J. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*, 343(d5928).
- Hjerkind, K. V., Stenehjem, J. S., and Nilsen, T. I. L. (2017). Adiposity, physical activity and risk of diabetes mellitus: prospective data from the population-based HUNT study, norway. *BMJ Open*, 7:e013142.
- Huang, M., Zhu, L., Jin, Y., Fang, Z., Chen, Y., and Yao, Y. (2021). Associação entre Infecção por *Heicobacter Pylori* e Hipertensão Arterial Sistêmica: Metanálise. *Arquivos Brasileiros de Cardiologia*, 117(4):626–636.
- Höel, P. G. (1971). *Introduction to Mathematical Statistics*. John Wiley & Sons, Inc., 4 edition.
- Kho, M. E., Molloy, A. J., Clarke, F. J., Reid, J. C., Herridge, M. S., Karachi, T., Rochweg, B., Fox-Robichaud, A. E., Seely, A. J., Mathur, S., Lo, V., Burns, K. E., Ball, I. M., Pellizzari, J. R., Tarride, J.-E., Rudkowski, J. C., Koo, K., Heels-Ansdell, D., and Cook, D. J. (2019). Multicentre pilot randomised clinical trial of early in-bed cycle ergometry with ventilated patients. *BMJ Open Res*, 6:e000383.
- Kikenny, C., Parsons, N., Kadyszewski, E., Festing, M. F. W., Cuthill, I. C., Fry, D., Hutton, J., and Altman, D. G. (2009). Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS ONE*, 4(11):e7824.
- Kutner, M., Nachtsheim, C., Neter, J., and Li, W. (2005). *Applied Linear Statistical Models*. McGraw-Hill/Irwin, Boston, 5 edition.
- Leisch, F. and Dimitriadou, E. (2010). mlbench: Machine Learning Benchmark Problems.

- Lopes, J. M., Fernandes, S. G. G., Dantas, F. G., and Medeiros, J. L. A. (2015). Associação da depressão com as características sociodemográficas, qualidade do sono e hábitos de vida em idosos do Nordeste brasileiro: estudo seccional de base populacional. *Revista Brasileira de Geriatria e Gerontologia*, 18(3):521–531.
- Malacarne, J., Heirich, A. S., Cunha, E. A. T., Kolte, I. V., Souza-Santos, R., and Basta, P. C. (2019). Desempenho de testes para o diagnóstico de tuberculose pulmonar em populações indígenas no Brasil: a contribuição do Teste Rápido Molecular. *J Bras Pneumol*, 45(2):e20180185.
- Manly, B. F. J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman & Hall, London, 2 edition.
- Medeiros, J. d. S., Rivera, M. A. A., Benigna, M. J. C., Cardoso, M. A. A., and Costa, M. J. d. C. (2003). Estudo caso-controle sobre exposição precoce ao leite de vaca e ocorrência de Diabetes Mellitus tipo 1 em CampinaGrande, Paraíba. *Rev Bras Saúde Matern Infant*, 3(3):271–280.
- Melo, F. R. R. (2017). Introdução à programação com a linguagem R.
- Melo, F. R. R. (2019a). Introdução ao R Commander.
- Melo, F. R. R. (2019b). R Commander: um pouco além dos menus gráficos.
- Meyer, P. L. (1969). *Introductory Probability and Statistical Applications*. Addison-Wesley Publishing Company, Inc., 2 edition.
- Ministério da Saúde, S. d. V. e. S. (2011). *Manual de instruções para o preenchimento da declaração de nascido vivo*. Editora MS, Brasília, 4 edition.
- Miranda, D. C., Brucki, S. M. D., and Yassuda, M. S. (2018). The Mini-Addenbrooke’s Cognitive Examination (M-ACE) as a brief cognitive screening instrument in Mild Cognitive Impairment and mild Alzheimer’s disease. *Dementia e Neuropsychologia*, 12(4):368–373.
- Novaes Neto, E. M., Araújo, T. M., and Sousa, C. C. (2020). Hipertensão Arterial e Diabetes Mellitus entre trabalhadores da saúde: associação com hábitos de vida e estressores ocupacionais. *Revista Brasileira de Saúde Ocupacional*, 45:e28.
- Nubila, B. C. L. S., Lacerda, G. C., Rey, H. C. V., and Barbosa, R. M. (2021). Remoção Percutânea de Eletrodos de Estimulação Cardíaca Artificial em um Único Centro Sul-Americano. *Arquivos Brasileiros de Cardiologia*, 116(5):908–916.
- Owens, D. K. and Sox, H. C. (2014). Biomedical Decision Making: Probabilistic Clinical Reasoning. In *Biomedical Informatics. Computer Applications in Health Care and Biomedicine*. Shortliffe EH, Cimino JJ, Springer, 4th edition.
- Parsons, Nick R, Price, C. L., Hiskens, R., Achten, J., and Costa, M. L. (2012). An evaluation of the quality of statistical design and analysis of published medical research: results from a systematic survey of general orthopaedic journals. *BMC Medical Research Methodology*, 12:60.

- Pearson, E. (1947). The choice of statistics tests illustrated on the interpretation of data classed in a  $2 \times 2$  table. *Biometrika*, 34(1/2):139–167.
- Pearson, K. (1900). 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. *Philosophical Magazine Series*, 50(302):157–175.
- Peng, R. D. (2016a). *Exploratory Data Analysis with R*. Leanpub.
- Peng, R. D. (2016b). *R Programming for Data Science*. Leanpub.
- Peng, R. D. (2016c). *Report Writing for Data Science*. Leanpub.
- Pereira, H. O., Rezende, E. M., and Couto, B. R. G. M. (2015). Tempo de internação pré-operatório: um fator de risco para reduzir a infecção cirúrgica em fraturas de fêmur. *Rev Bras Ortop*, 50(6):638–646.
- Perez-Gurbindo, I., Álvarez Méndez, A. M., Pérez-García, R., Arribas-Cobo, P., and Angulo-Carrere, M. T. (2021). Fatores associados às quedas em pacientes de hemodiálise: um estudo caso-controle. *Rev Latino-Am Enfermagem*, 29(e3505):1–9.
- Pinto Filho, J. L. O., Nobre, S. B., and Mariano Neto, M. (2020). O perfil socioeconômico e a percepção ambiental dos pescadores da Lagoa do Apodi, Rio Grande do Norte, Brasil. *Interações*, 21(4):721–737.
- Pocock, S. J. (1983). *Clinical Trials: A Practical Approach*. Wiley, Chichester.
- Quadros, T. M. B., Gordia, A. P., Silva, R. C. R., and Silva, L. R. (2015). Capacidade preditiva de indicadores antropométricos para o rastreamento da dislipidemia em crianças e adolescentes. *Jornal de Pediatria*, 91(5):455–463.
- Queiroz, C. F., Lemos, A. C. M., Bastos, M. d. L. S., Neves, M. C. L. C., Camelier, A. A., Carvalho, N. B., and Carvalho, E. M. (106). Perfil inflamatório e imunológico em pacientes com DPOC: relação com a reversibilidade do VEF 1. *Jornal Brasileiro de Pneumologia*, 42(4):241–247.
- Rahman, M. M., Kopec, J. A., Cibere, J., Godsmith, C. H., and Anis, A. H. (2013). The relationship between osteoarthritis and cardiovascular disease in a population health survey: a cross-sectional study. *BMJ Open*, 3:e002624.
- Ribeiro, M. A. S., Fiori, H. H., Luz, J. H., Garcia, P. C. R., and Fiori, R. M. (2019). Diagnóstico rápido da síndrome do desconforto respiratório por aspirado bucal em recém-nascidos prematuros. *Jornal de Pediatria*, 95(4):489–494.
- Rocha, V. S., Aliti, G., Moraes, M. A., and Rabelo, E. R. (2009). Repouso de Três Horas Não Aumenta Complicações Após Cateterismo Cardíaco Diagnóstico com Introdutor Arterial 6 F: Ensaio Clínico Randomizado. *Rev Bras Cardiol Invas*, 17(4):512–517.
- Rothman, K. J., Greenland, S., and Lash, T. L. (2011). *Epidemiologia Moderna*. Artmed, Porto Alegre, 3 edition.

- Salgado, P. O., Souza, C. C., Prado Júnior, P. P., Balbino, P. C., Ribeiro, L., Paiva, L. C., and Brombine, N. L. M. (2018). O uso da simulação no ensino da técnica de aspiração de vias aéreas: ensaio clínico randomizado controlado. *Rev Min Enferm*, 22(e-1090):1–9.
- Serpa, F. S., Piana, M. P., Braga Neto, F., Campinhos, F. L., Silveira, M. G., Chiabai, J., and Zandonade, E. (2014). Eficácia da terapia Anti-IgE no controle da asma. *Braz J Allergy Immunol*, 2(4):147–53.
- Severo, I. M., Kuchenbecker, R. S., Vieira, D. F. V. B., Lucena, A. F., and Almeida, M. A. (2018). Fatores de risco para quedas em pacientes adultos hospitalizados: um estudo caso-controle. *Revista Latino-Americana de Enfermagem*, 26(e3016).
- Shortliffe, E. H. and Cimino, J. J. (2014). *Biomedical Informatics. Computer Applications in Health Care and Biomedicine*. Springer, 4th edition.
- Silva, P. V., Salman, A. A., Cristóvão, S. A. B., Carnieto, N. M., Erudilho, E., Mauro, M. F. Z., Ticky, M. C., Dutra, G., Giordano, B., and Mangione, J. A. (2014). Impacto do Escore SYNTAX no Prognóstico de Pacientes com Doença Multiarterial Tratados por Intervenção Coronária Percutânea. *Revista Brasileira de Cardiologia Invasiva*, 22(3):258–263.
- Souza, D. S., Noblat, L. d. A. C. B., and Santos, P. d. M. (2015). Fatores associados à qualidade de vida sob a perspectiva da terapia medicamentosa em pacientes com asma grave. *J Bras Pneumol*, 41(6):496–501.
- Sperandio, E. F., Arantes, R. L., Matheus, A. C., Silva, R. P., Lauria, V. T., Romiti, M., Gagliardi, A. R. d. T., and Dourado, V. Z. (2016). Distúrbio ventilatório restritivo sugerido por espirometria: associação com risco cardiovascular e nível de atividade física em adultos assintomáticos. *J Bras Pneumol*, 42(1):22–28.
- van Bemmelen, J. H. and Musen, M. A. (1997). *Handbook of Medical Informatics*. Springer.
- Vanin, L. K., Zatti, H., Soncini, T., Nunes, R. D., and Siqueira, L. B. S. (2019). Fatores de risco materno-fetais associados à prematuridade tardia. *Revista Paulista de Pediatria*, 38(e2018136).
- Venables, W. and Ripley, B. (2002). *Modern Applied Statistics with S*. Springer, New York, 4 edition.
- Weckx, L. L. M., Hirata, C. H. W., Abreu, M. A. M. M., Fillizolla, V. C., and Silva, O. M. P. (2009). Levamisol não previne lesões de estomatite aftosa recorrente: um ensaio controlado randomizado, duplo-cego e controlado por placebo. *Revista da Associação Médica Brasileira*, 55(2):132–138.
- Welch, B. (1951). On the Comparison of Several Mean Values: An Alternative Approach. *Biometrika*, 38:330–336.
- Wikipedia (2019). Hodges and Lehmann estimator.
- Yates, F. (1934). Contingency tables involving small numbers and the  $\chi^2$  test. *Journal of the Royal Statistical Society*, 1 Supplement:217–235.

Yi, D., Li, G., Zhou, L., Xiao, Q., Zhang, Y., Liu, X., Chen, H., Pettigrew, J. C., Yi, D., Liu, L., and Wu, Y. (2015). Statistical use in clinical studies: Is there evidence of a methodological shift? *PLoS ONE*, 10(10):e0140159.

Zabor, E. (2018). Survival Analysis in R.