CÂMARA DOS DEPUTADOS
DIRETORIA DE RECURSOS HUMANOS
CENTRO DE FORMAÇÃO, TREINAMENTO E APERFEIÇOAMENTO

ANÁLISE ESTATÍSTICA DE DADOS COM ÊNFASE EM PESQUISAS LEGISLATIVAS

Fabiano Peruzzo Schwartz
João Luiz Pereira Marciano

BRASÍLIA/2020

FICHA CATALOGRÁFICA

SCHWARTZ e MARCIANO

ANÁLISE ESTATÍSTICA DE DADOS COM ÊNFASE EM PESQUISAS LEGISLATIVAS [Distrito Federal] 2020.

xviii, 273p., 210 x 297 mm (CEFOR/DRH/CD, Doutor, Mestrado Profissional em Poder Legislativo, 2020).

Grupo de Pesquisa – Câmara dos Deputados. Diretoria de Recursos Humanos.

Centro de Formação, Treinamento e Aperfeiçoamento.

1. Poder Legislativo
2. Ciência de Dados
3. Métodos Quantitativos
4. Modelos Estatísticos
I. CEFOR/DRH/CD
II. Título (série)

REFERÊNCIA BIBLIOGRÁFICA

SCHWARTZ, F. P., MARCIANO, J. L. P. (2020). Análise estatística de dados com ênfase em Pesquisas Legislativas ...., Brasília, DF, 273p.

CESSÃO DE DIREITOS

AUTOR:

TÍTULO:.

ANO: 2020

DEDICATÓRIA

De nada serve a Ciência, se não para a evolução da humanidade e do planeta.
Estejam atentos os cientistas, pois carregam essa responsabilidade.

À minha esposa Flávia, ao meu filho Diogo, aos meus pais, ...

"A arte da vida está na habilidade de vivermos o máximo que pudermos, com alegria. Lutemos pela vida até o último segundo. Após a morte, aceitemos a morte."
Alfredo José Procaci Ferreira

AGRADECIMENTOS

Agradeço primeiramente ...

RESUMO

ANÁLISE ESTATÍSTICA DE DADOS COM ÊNFASE EM PESQUISAS LEGISLATIVAS

Autores:

Brasília, fevereiro de 2020

Este trabalho ....

SUMÁRIO

LISTA DE TABELAS XI

LISTA DE QUADROS XII

LISTA DE FIGURAS XIV

LISTA DE ABREVIAÇÕES XVII

1 INTRODUÇÃO 1

2 PESQUISA ESTATÍSTICA 4

2.1 CIÊNCIA DE DADOS – MÉTODOS DE ANÁLISE 4

2.2 FORMULANDO HIPÓTESE DE PESQUISA 5

2.3 VARIÁVEIS E DADOS 7

2.4 MÉTODO CIENTÍFICO 11

2.5 PRINCÍPIO DA COLETA DE DADOS 11

2.5.1 Método Observacional 12

2.5.2 Método Experimental 14

2.6 ESTATÍSTICA DESCRITIVA 15

2.6.1 Distribuição de Frequências 15

2.6.2 Medidas de Tendência Central 19

2.6.3 Medidas de Dispersão 22

2.6.4 Simetria de Dados 25

2.6.5 Medidas de Achatamento ou Curtose 28

2.7 R 29

2.7.1 O ambiente R 30

2.8 LABORATÓRIO 1 31

2.8.1 Explorando um arquivo de dados 31

2.8.2 Tabelas de distribuição de frequências 32

2.8.3 Histogramas 32

2.8.4 Medidas de tendência central e de dispersão 33

3 DISTRIBUIÇÃO DE PROBABILIDADE 34

3.1 PROBABILIDADE 34

3.2 LABORATÓRIO 2 46

3.2.1 Curva normal padronizada 46

3.2.2 Curva normal 48

3.2.3 Curva normal – variação do achatamento 49

4 INFERÊNCIA ESTATÍSTICA 51

4.1 INFERÊNCIA ESTATÍSTICA 51

4.2 ASSINTOTISMO 52

4.3 INTERVALOS DE CONFIANÇA 56

4.4 LABORATÓRIO 3 60

4.4.1 Teorema do Limite Central 60

5 TESTES ESTATÍSTICOS 62

5.1 MODELOS ESTATÍSTICOS 62

5.2 TESTE DE HIPÓTESE 63

5.2.1 Formulação das hipóteses nula (H_0) e alternativa (H_1) 64

5.2.2 Ajustamento dos dados a um modelo estatístico 65

5.2.3 Avaliação do modelo por meio de um teste estatístico 65

5.2.4 Região crítica 68

5.2.5 Tipos de erros 69

5.2.6 Força ou poder de um teste estatístico 69

5.3 LABORATÓRIO 4 72

5.3.1 Comparação entre médias 72

5.3.1.1 Exercício 1 73

5.3.1.2 Exercício 2 73

5.4 DISTRIBUIÇÃO T DE STUDENT 75

5.4.1 Intervalo de confiança em distribuições *t* 76

5.4.2 Grupos pareados 77

5.4.3 Grupos independentes 80

5.4.3.1 Variâncias iguais 80

5.4.3.2 Variâncias diferentes 82

5.5 LABORATÓRIO 5 84
5.5.1 Simulação 84
5.5.2 Comparação entre a média da população e a média de uma amostra . 84
5.5.3 Comparação entre as médias de duas amostras independentes 85
5.5.4 Comparação entre as médias de duas amostras pareadas 85
5.6 TESTES PARAMÉTRICOS 85
5.6.1 Shapiro-Wilk 85
5.6.2 Kolmogorov-Smirnov 87
5.6.3 Gráfico Q-Q Plot 87
6 CORRELAÇÃO LINEAR E REGRESSÃO LINEAR SIMPLES 90
6.1 CORRELAÇÃO LINEAR 90
6.2 REGRESSÃO LINEAR SIMPLES 93
6.2.1 Formalização matemática da regressão linear simples 99
6.2.2 Modelo de regressão linear simples com adição de erro Gaussiano ... 103
6.2.3 Interpretação dos coeficientes de regressão 105
6.2.4 Resíduos e variações residuais 110
6.2.5 Coeficiente de determinação R^2 115
6.2.6 Estimadores e o processo de inferência na regressão linear 120
6.2.7 Desenvolvendo os conceitos em R 123
6.2.8 Fatores de incerteza e predição de valores 126
6.3 LABORATÓRIO 6 131
6.3.1 Exercício 1 131
6.3.2 Exercício 2 131
6.3.3 Exercício 3 131
6.3.4 Exercício 4 132
7 REGRESSÃO LINEAR MÚLTIPLA 133
7.1 CONTEXTUALIZAÇÃO DO PROBLEMA 133
7.2 O MODELO LINEAR GERAL 134
7.2.1 Variáveis “burras” são espertas 143
7.2.2 Ajuste pelo efeito de grupo (ou do fenômeno provocado) 146
7.2.3 Regressão múltipla com regressores não relacionados 150
7.2.4 Análise dos Resíduos 152

7.2.4.1 Outliers: influência e alavancagem 156
7.2.4.2 Exemplos de diagnósticos 159

7.2.5 Construindo um modelo multivariado 163
7.2.5.1 Regras gerais para a escolha do modelo 164
7.2.5.2 Variação de R^2 165
7.2.5.3 Inflação da Variância 166
7.2.5.4 Selecionado um modelo multivariado 168

7.2.6 Interações em Regressão 170

7.3 LABORATÓRIO 7 171

8 REGRESSÃO LOGÍSTICA 172

8.1 AVALIAÇÃO DO MODELO DE REGRESSÃO LOGÍSTICA 184

8.1.1 Estatística *log-likelihood* 184

8.1.2 Estatística *deviance* 185

8.1.3 Critério de informação 188

8.2 LABORATÓRIO 8 189

9 ANÁLISE DE VARIÂNCIA 190

9.1 ESTATÍSTICA F – COMPARAÇÃO DE DUAS VARIÂNCIAS 190

9.2 ANOVA 192

9.2.1 Múltiplos grupos e propagação do erro 196

9.2.2 Testes *Post Hoc* 198
9.2.2.1 Correção de Bonferroni 198
9.2.2.2 Método de Bonferroni-Holm 199
9.2.2.3 Método de Tukey 199
9.2.2.4 Que método usar? 200

9.3 ANOVA E REGRESSÃO MÚLTIPLA SÃO A MESMA COISA 200

9.4 ANOVA ONE-WAY NO AMBIENTE R 203

9.5 NOVA TWO-WAY NO AMBIENTE R 205

9.6 LABORATÓRIO 9 208

10 QUI-QUIDRADO 209

10.1 DISTRIBUIÇÃO QUI-QUADRADO 210

10.2 TABELA DE CONTINGÊNCIA 212

10.3 LABORATÓRIO 10 215

REFERÊNCIAS BIBLIOGRÁFICAS 216

APÊNDICE I – INTEGRAL 218

APÊNDICE II – REGRESSÃO COM TRÊS REGRESSORES – RESÍDUOS 221

APÊNDICE III – RESÍDUO ESTUDENTIZADO..... 224

APÊNDICE IV – CÓDIGOS EM R..... 225

LISTA DE TABELAS

Tabela 2.1 – Quantidade de eleitores por unidade da federação separados por sexo.® 16

Tabela 2.2 – Estado civil dos candidatos à eleição no estado do Paraná em 2012.® 17

Tabela 2.3 – Quantidade de eleitores por faixa etária.® 18

Tabela 3.1– Seções eleitorais em que ocorreram votos nulos no estado do Paraná, no primeiro turno das eleições de 2012.® 37

Tabela 3.2 – Probabilidades da Curva Normal Padronizada.® 42

Tabela 4.1 – Características da população e da amostra. 51

Tabela 6.1 – Vídeos de propaganda assistidos versus nota de aprovação do governo. 90

Tabela 8.1 – Arrecadação de campanha por candidato. 172

Tabela 9.1 – Soma dos quadrados dentro dos grupos e geral. ® 193

Tabela 9.2 – ANOVA amostras independentes. ® 196

Tabela 10.1 – Tabela de Contingência: pesquisa de opinião sobre preferência partidária classificada por sexo. 213

Tabela 10.2: Tabela de Contingência e valores esperados: pesquisa de opinião sobre preferência partidária classificada por sexo 214

LISTA DE QUADROS

Quadro 2.1 – Métodos de análise de dados 4

Quadro 2.2 – Classificação de variáveis quanto ao nível de medição. 9

Quadro 2.3 – Determinação da amplitude interquartil para os dados da Tabela 2.1. 23

Quadro 3.1 – Código R que estima as probabilidades dos candidatos às eleições do Paraná, com $N \sim (44{,}9;10{,}98)$, terem 60 anos ($x = 60$) ou mais, e 60 anos ou menos: aplicação direta da PDF. 44

Quadro 3.2 – Código R que estima as probabilidades dos candidatos às eleições do Paraná, com $N \sim (44{,}9;10{,}98)$, terem 60 anos ($x = 60$) ou mais, e 60 anos ou menos: aplicação curva normal padronizada. 45

Quadro 5.1 – Probabilidades dos erros do tipo I e II. 69

Quadro 5.2 – Diferença entre as estatísticas *t* e *z*. 75

Quadro 5.3 – Quantis *t* para intervalo bicaudal com 95% de confiança (n=11). 77

Quadro 5.4 – Análise de grupos pareados, assumindo igual variância. 79

Quadro 5.5 – Grupos independentes: variâncias iguais. 81

Quadro 5.6 – Engano ao tratar grupos independentes como pareados. 82

Quadro 5.7 – Grupos independentes: variâncias diferentes. 83

Quadro 5.8 – Teste de Shapiro-Wilk em uma distribuição normal. 86

Quadro 5.9 – Teste de Shapiro-Wilk em uma distribuição uniforme. 86

Quadro 5.10 – Teste de Kolmogorov-Smirnov em uma distribuição normal. 87

Quadro 6.1 – Coeficiente de correlação de Pearson. 93

Quadro 6.2 – Soma dos erros quadráticos. 95

Quadro 6.3 – Função lm (*linear model*) com variáveis centralizadas. 98

Quadro 6.4 – Inclinação da reta das variáveis centralizadas e SEQ. 98

Quadro 6.5 – Regressão linear para os dados de diamantes. 107

Quadro 6.6 – Deslocamento da variável *carat* por um fator igual à média. 108

Quadro 6.7 – Escalonamento da variável *carat*. 109

Quadro 6.8 – Deslocamento e escalonamento combinados. 109

Quadro 6.9 – Predição de preços por meio do modelo de regressão. 109

Quadro 6.10 – Preditor com distribuição uniforme e resultado com variação senoidal. 113

Quadro 6.11 – Preditor com distribuição uniforme e resultado com desvio padrão dependente do preditor. 114

Quadro 6.12 – Estimativa da variância do resíduo. 115

Quadro 6.13 – Parâmetros da regressão linear calculados de forma expressa. 124

Quadro 6.14 – Parâmetros da regressão linear calculados pela função lm. 125

Quadro 6.15 – Obtendo intervalos de confiança para os coeficientes da regressão. 125

Quadro 7.1 – Estimativa de coeficientes pelo método da obtenção dos resíduos para dois regressores. 138
Quadro 7.2 – Estimativa de coeficientes pelo método da obtenção dos resíduos para três regressores. 139
Quadro 7.3 – Regressão Linear Múltipla: exemplos. 142
Quadro 7.4 – Regressão fatorial com múltiplos níveis. 145
Quadro 7.5 – Dados sem regressão linear: inluência e alavancagem - Caso 1. 160
Quadro 7.6 – Dados com regressão linear: inluência e alavancagem - Caso 2. 161
Quadro 7.7 – Análise de resíduos: identificação de padrões. 162
Quadro 7.8 – Inflação da variância. 166
Quadro 7.9 – Cálculo do fator de inflação da variância. 168
Quadro 7.10 – Análise do modelo de regressão com ANOVA. 169
Quadro 8.1 – Comportamento do sigmoide com a variação de β_0 e β_1. 177
Quadro 9.1 – Simulação da distribuição *F* com $gl_1 = 39$ e $gl_2 = 29$. 191
Quadro 9.2 – ANOVA *versus* Regressão Múltipla. 201
Quadro 9.3 – ANOVA *one way* e testes *post hoc*. 203
Quadro 9.4 – 206
Quadro 10.1 – Simulação da distribuição qui-quadrado para $k = 2$ ($gl = 1$). 210

LISTA DE FIGURAS

Figura 2.1 – Histograma: quantidade de eleitores nas UFs (repositório de dados do TSE de abril de 2013).® 18

Figura 2.2 – Distribuição simétrica.® 25

Figura 2.3 – Curvas normais com diferentes médias e desvios-padrões. 26

Figura 2.4 – Distribuições assimétricas.® 27

Figura 2.5 – Curtose da distribuição. 28

Figura 2.6 – Ambiente R com a tela de console pronta para receber comandos. 30

Figura 3.1 – Relação entre a quantidade de votos nulos e a quantidade de seções eleitorais em que esses votos ocorreram no estado do Paraná, no primeiro turno das eleições de 2012.® 35

Figura 3.2 – Distribuição das idades completas dos candidatos do Paraná no dia do primeiro turno das eleições de 2012, com a frequência relativa (f_r) de cada classe destacada sobre a barra. A probabilidade de um candidato ter idade entre 20 e 40 anos corresponde à soma das f_r das barras em vermelho.® 38

Figura 3.3 – Histogramas com intervalos de classe distintos para o mesmo conjunto de dados: (a) 5, (b) 2 e (c) 1.® 39

Figura 3.4 – Distribuição de probabilidades: aproximação do caso discreto (barras vermelhas) ao caso contínuo (linha preta) com intervalo de classe tendendo a zero e quantidade de dados tendendo a infinito.® 40

Figura 3.5 – Proporções da área da curva normal padronizada ($\mu = 0$ e $\sigma = 1$).® 43

Figura 3.6 – FDA, idade dos candidatos do Paraná – Eleições 2012.® 46

Figura 4.1 – Lei dos Grandes Números em ação.® 53

Figura 4.2 – Médias populacional e amostral. 54

Figura 4.3 – Distribuição amostral das médias.® 55

Figura 4.4 – Caixa de Galton. 55

Figura 4.5 – Intervalos de confiança para 20 amostras com tamanho 30 e média populacional igual a 44,9, considerando-se nível de confiança de 95%.® 56

Figura 4.6 – Intervalos de confiança com nível de confiança de 95% para diferentes tamanhos da amostra.® 59

Figura 4.7 – Comparação de ICs a 95% de confiança: a) amostras da mesma população; b) amostras de populações distintas.® 59

Figura 5.1 – $p(z_{calc} > 2{,}54) = 0{,}0055$.® 66

Figura 5.2 – Região crítica: (a) $H_0: \mu = \mu_0$ e $H_1: \mu < \mu_0$; (b) $H_0: \mu = \mu_0$ e $H_1: \mu > \mu_0$; (c) $H_0: \mu = \mu_0$ e $H_1: \mu \neq \mu_0$. ® 68

Figura 5.3 – Erro do tipo II. 70

Figura 5.4 – Força do teste. 71

Figura 5.5 – Distribuições t de Student. ® 76
Figura 5.6 – Boxplot para grupos pareados. ® 78
Figura 5.7 – Dados de Gosset, teste *t*. ® 80
Figura 5.8 – Q-Q Plot. ® 88
Figura 6.1 – Desvio dos dados em relação à média. ® 91
Figura 6.2 – Histograma da altura dos filhos e dos pais. ® 94
Figura 6.3 – Gráfico de dispersão das variáveis *x* e *Y*: a) vários pontos estão sobrepostos; b) o tamanho do ponto é proporcional à quantidade de pontos na coordenada; o ponto vermelho é o centro de massa. ® 96
Figura 6.4 – Distâncias verticais dos pontos à reta. ® 97
Figura 6.5 – Centro de massa deslocado para a origem. ® 97
Figura 6.6 – Identificação do intercepto *y* dada a inclinação β da reta. ® 99
Figura 6.7 – Modelo estatístico da regressão linear simples: $y_i \sim N(\mu_i, \sigma^2)$. ® 104
Figura 6.8 – Regressão linear: preço do diamante em função da massa. ® 108
Figura 6.9 – Resíduos: (a) gráfico de dispersão; (b) resíduos *vs*. x (massa). ® 112
Figura 6.10 – Variação sistemática verificada no gráfico do resíduo *vs*. x. ® 113
Figura 6.11 – Heterocedasticidade verificada no gráfico do resíduo *vs*. x. ® 114
Figura 6.12 – Regressões lineares com parâmetros quase idênticos: $\hat{\beta}_0 \approx 3$, $\hat{\beta}_1 \approx 0.5$, $\bar{x} \approx 9$, $\bar{y} \approx 7.5$ e $R^2 \approx 0.67$. ® 119
Figura 6.13 – (a) estimador não viciado; (b) estimador viciado. 120
Figura 6.14 – Estimador não viciado: (a) com variância pequena; (b) com variância grande. 121
Figura 6.15 – Intervalos de confiança de valores ajustados (linhas vermelhas) e intervalos de predição (linhas azuis). ® 130
Figura 7.1 – Simulação 1: o regressor não está relacionado ao estado do grupo. ® 147
Figura 7.2 – Simulação 2: o regressor está relacionado ao estado do grupo. ® 148
Figura 7.3 – Simulação 3: o regressor está relacionado ao estado do grupo. ® 149
Figura 7.4 – Simulação 4: não há associação marginal entre o estado de grupo e *Y*. ® ... 150
Figura 7.5 – Regressores independentes. ® 151
Figura 7.6 – Representação 3D do modelo com dois regressores independentes. ® 151
Figura 7.7 – Relação entre resíduos desconsiderando-se o regressor x_2. ® 152
Figura 7.8 – Gráficos produzidos pela função **plot**. ® 155
Figura 7.9 – Influência e alavancagem. ® 157
Figura 7.10 – Dados sem regressão linear: inluência e alavancagem - Caso 1. ® 159
Figura 7.11 – Dados com regressão linear: inluência e alavancagem - Caso 2. ® 161
Figura 7.12 – Análise de resíduos: Orly. ® 163

Figura 7.13 – Crescimento monotônico de R^2. ® 166
Figura 8.1 – Classificação com regressão linear: caso 1. ® 173
Figura 8.2 – Classificação com regressão linear: caso 2. ® 174
Figura 8.3 – Curva sigmoide para $\mu(x) = x$. ® 176
Figura 8.4 – Classificação: observações X possuem determinada característica e as observações ○ não a possuem. ® 178
Figura 8.5 – Fronteira de decisão. ® 179
Figura 9.1 – Distribuição *F* para $gl_1 = 3$ e $gl_2 = 2$. ® 190
Figura 9.2 – distribuição F com $gl_1 = 39$ e $gl_2 = 29$ 191
Figura 9.3 – Variação explicada pelo modelo (segmentos azuis): diferença entre a variação total (segmentos em preto) e variação referente ao modelo (segmentos em vermelho e verde). ® 194
Figura 10.1 – Distribuição X^2 com $k = 2$ ($gl = 1$) 211
Figura 10.2 – Distribuição X^2 para distintos valores de *k*. ® 211

LISTA DE ABREVIAÇÕES

PDF	*probability density function* ou função de densidade de probabilidade
SEQ	soma dos erros quadrados
TSE	Tribunal Superior Eleitoral

1 INTRODUÇÃO

Cientistas são pessoas curiosas. Contudo, para se responder a questões interessantes oriundas da curiosidade científica, duas coisas são necessárias: dados e uma boa explicação sobre eles.

Qualquer que seja o fenômeno que se deseja explicar, faz-se necessária a coleta de dados no mundo real para que sejam traçadas as conclusões sobre o que se está estudando. Como nem sempre é possível acessar diretamente processos do mundo real, modelos são construídos com base em observações (dados), na tentativa de se prever como esses processos operam sob determinadas condições.

A Estatística é a área do conhecimento que congrega um conjunto de métodos especialmente apropriados à coleta, à apresentação (organização, resumo e descrição), à análise e à interpretação de dados de observação, tendo como objetivo a compreensão de uma realidade específica para a tomada de decisão.

A Estatística trabalha com dois tipos de conjuntos de dados: a população (ou universo) e a amostra. A **população** é o maior conjunto de entidades sobre o qual se deseja realizar uma investigação. A **amostra** é uma fração, ou parte, dessa população.

Apesar de a Estatística estar focada na obtenção de informações sobre a população, dificilmente os componentes desta são observados em sua totalidade (censo), e a análise, em geral, ocorre sobre os dados da amostra para, então, inferir-se algo sobre a população, seguindo-se o trâmite do método científico.

Atualmente, vivemos a era do “Big Data”, bancos de dados de tamanho bem maior do que os que em geral conhecemos. Chris Anderson, editor da revista *Wired*, escreveu, em 2008, que o mero volume de dados eliminaria a necessidade de teoria e até de método científico. Entretanto, na visão de Silver (2012), “os números, em si, nada dizem. Nós é que falamos por eles. Nós os imbuímos de significado. Mas há que se ter cuidado para não os interpretar de maneira que sirvam aos nossos interesses, de forma desvinculada da sua realidade objetiva. Previsões baseadas em dados podem se concretizar, ou falhar. Quando

negamos nosso papel no processo, as chances de fracasso se elevam. Antes de exigir mais de nossos dados, precisamos exigir mais de nós mesmos".

De acordo com John Tukey (http://en.wikipedia.org/wiki/John_Tukey), um dos mais famosos analistas de dados, a resposta ao que se procura pode não estar contida no conjunto de dados investigado, independentemente deste ser grande ou pequeno. A combinação de um conjunto de dados com o desejo ardente por uma resposta não garantem que esta possa ser obtida de forma razoável. Portanto, pensar cuidadosamente sobre se o conjunto de dados responderá ao que se deseja descobrir é o primeiro e mais importante passo.

Conclusões equivocadas decorrentes da inadequação dos dados analisados não são incomuns. Por exemplo, os modelos de previsão usados na eleição presidencial americana de 2000 e publicados pelos cientistas políticos antes das votações previram a vitória esmagadora de Al Gore por onze pontos (Silver, 2012). Quem ganhou as eleições foi George W. Bush. Fracassos de previsão como esses são relativamente comuns na política. Um estudo de longo prazo realizado por Tetlock (2006), da Universidade da Pensilvânia, revelou que resultados considerados impossíveis aconteciam em aproximadamente 15% dos casos.

Casos semelhantes são encontrados no sistema eleitoral brasileiro (Revista Época, 11/11/2012 - http://revistaepoca.globo.com/ideias/noticia/2012/11/por-que-previsoes-na-politica-e-na-economia-falham-tanto.html):

> A festa da vitória de Gustavo Fruet (PDT) na eleição para prefeito de Curitiba, em 2012, não estava no radar do Ibope. Terceiro colocado nas intenções de voto durante toda a primeira fase da campanha, Fruet apareceu na pesquisa de boca de urna no primeiro turno de votação 5 pontos percentuais atrás do segundo colocado, Luciano Ducci (PSB). Abertas as urnas, Fruet teve 4 mil votos a mais do que Ducci e passou para o segundo turno. No último domingo, ele consolidou a surpresa e conquistou a prefeitura ao vencer Ratinho Jr. (PSC), com 60,65% dos votos válidos. Em entrevista após a eleição, Fruet ironizou as pesquisas. "Não fiquei surpreso com o resultado", disse. "Quem ficou foi o Ibope".

> Após a apuração do primeiro turno, o Ibope tentou explicar o erro. Em nota, reconheceu que as pesquisas não foram capazes de captar a velocidade do crescimento das intenções de voto em Fruet. “As pesquisas buscam medir a opinião das pessoas, e elas mudam conforme as experiências que elas têm”, afirma a diretora do Ibope, Márcia Cavallari. “A opinião pública é dinâmica e responde a estímulos externos”. O erro em Curitiba não chega a ser escandaloso. Fruet chegou a 27,22% dos votos, em relação aos 24% apontados pelas pesquisas – pouco mais de 1 ponto percentual acima da margem de erro, de 2%. Mas a imprecisão demonstra uma limitação do modelo de pesquisas: sozinhas, elas não são suficientes para prever os resultados das eleições.

Por que, então, aprender a Ciência de Dados? Ao longo dos últimos anos, a coleta e o armazenamento de dados tornaram-se muito mais baratos e fáceis. Contudo, enquanto a quantidade de informações está crescendo em 2,5 quintilhões de bytes por dia, segundo estimativas da IBM, o mesmo não pode ser dito sobre a quantidade de informações úteis. A maior parte é apenas ruído, que está aumentando de forma mais rápida que o sinal. Há inúmeras hipóteses a testar e muitos conjuntos de dados a garimpar, mas apenas uma quantidade relativamente constante de verdades objetivas. Em ciência política, por exemplo, podemos testar modelos usados para prever o resultado das eleições, mas uma teoria sobre a influência de mudanças em entidades públicas pode levar décadas para ser testada (Silver, 2012).

Atualmente existem diversas ferramentas de computação disponíveis, e livres, com as quais qualquer pessoa interessada pode realmente fazer algo com todo esse dilúvio de dados nos diferentes domínios da ciência e dos negócios. O Big Data é uma espécie de nova fronteira. Hoje dispomos de dados em áreas em que não costumávamos dispor, como informações sobre as coordenadas GPS de carros de todo mundo e o sequenciamento do genoma. Agora isso é possível. Portanto, somos capazes de responder a perguntas que não podíamos antes. Então, vivemos um tempo incrivelmente emocionante.

...

...

2 PESQUISA ESTATÍSTICA

Neste módulo são apresentados os conceitos básicos e o uso da estatística ...

2.1 CIÊNCIA DE DADOS – MÉTODOS DE ANÁLISE

Há diversos métodos pelos quais a Ciência de Dados provê respostas a questões: descritivo, exploratório, inferencial, preditivo, causal e mecanicista.

Quadro 2.1 – Métodos de análise de dados

Método	**Objetivo**	**Características**
Descritivo	Descrever um conjunto de dados.	É a primeira forma de análise. Geralmente aplicada em censo. Descrição e interpretação são etapas distintas. Descrições não podem ser generalizadas sem modelagem estatística adicional. IBGE USA Census 2010 N-grams
Exploratório	Encontrar relações desconhecidas.	Útil na descoberta de novas conexões de dados, sem necessariamente confirmá-las. Útil na definição de estudos futuros. Não é a palavra final da análise. Não pode ser usado para generalização ou predição. Correlação não implica causa. Basômetro Transparency International
Inferencial	Dizer algo sobre uma grande população a partir de uma amostra relativamente pequena.	É geralmente o objetivo dos modelos estatísticos. Envolve estimativas tanto da quantidade em estudo como da incerteza da medida. Depende fortemente da população e do esquema de amostragem. Utilizada para generalização. Air Pollution Control

Preditivo	Utilizar dados sobre alguns objetos para predizer os valores de outros objetos.	Mais desafiador que o método inferencial. Se *X* prediz *Y* não significa dizer que *X* causa *Y*. A predição acurada depende fortemente da medição das variáveis corretas. Modelos simples e grande quantidade de dados funcionam bem. “Predições são muito difíceis, especialmente as sobre o futuro”. Netflix
Causal	Descobrir o que acontece a uma variável quando outra estiver sofrendo alteração.	Estudos aleatórios são geralmente necessários para a identificação de causa. Efeitos causais são observados em relação a médias e podem não ser aplicáveis a todos os indivíduos. Modelos causais são o “padrão ouro” da análise de dados. Nature
Mecanicista	Entender as exatas mudanças em uma variável que levam a alterações em outras variáveis para cada objeto, individualmente.	Difícil de inferir, raramente utilizado. Geralmente modelado por um conjunto de equações determinísticas (física/engenharia). O componente aleatório é o erro. Conhecidas as equações, mas não os parâmetros, estes podem ser inferidos por análise de dados. Teoria de Erros

2.2 FORMULANDO HIPÓTESE DE PESQUISA

Uma hipótese de pesquisa é qualquer conjectura sobre aspectos desconhecidos de uma população. É uma declaração de opinião testável, criada a partir de uma pergunta (ou questão) de pesquisa.

Há basicamente dois tipos de questões de pesquisa: as que podem ser testadas (testáveis) e as que não podem (não testáveis). Não há uma melhor do que a outra e ambas têm lugar na pesquisa aplicada.

Exemplos de questões que não podem ser testadas:

> Como os cidadãos brasileiros se sentem sobre a realização de um plebiscito?
>
> Quais os principais problemas enfrentados pela sociedade segundo os cidadãos brasileiros?

As respostas a essas questões poderiam ser sintetizadas em tabelas descritivas e os resultados seriam extremamente valiosos para cientistas políticos e marqueteiros de campanhas. Pesquisadores das áreas de administração e ciências sociais frequentemente utilizam questões dessa natureza. A desvantagem das questões não testáveis é que elas não proveem pontos de corte objetivos para os tomadores de decisão.

Para superar este problema, os pesquisadores muitas vezes procuram responder a uma ou mais questões de pesquisa testáveis. Quase todas as questões de pesquisa testáveis começam de uma das seguintes formas:

> Existe diferença significativa entre... ?
>
> Existe relação significativa entre... ?

Exemplos:

> Existe relação significativa entre a idade dos eleitores e a sua intensão de apoiar a realização de plebiscito?
>
> Existe diferença significativa entre a percepção de cidadãos negros e brancos quanto à discriminação social?

Encontrada a pergunta de pesquisa, a hipótese é formulada como uma declaração afirmativa. Para as duas questões anteriores, as hipóteses seriam:

> Há relação significativa entre a idade dos eleitores e a sua intensão de apoiar a realização de plebiscito.

> Há diferença significativa entre a percepção de cidadãos negros e brancos quanto à discriminação social.

Não é possível testar uma hipótese diretamente, ou melhor, não é possível provar uma hipótese ou teoria visto que, de tempos em tempos, surgem novas evidências no mundo científico que rejeitam provas anteriores e conduzem a novas e diferentes explicações para um dado fenômeno. Embora não se possa provar uma hipótese, pode-se rejeitar a sua negação, ou seja, a **hipótese nula**. A hipótese nula é criada adicionando-se uma negativa à declaração original. Para as duas hipóteses anteriores, as hipóteses nulas seriam:

> Não há relação significativa entre a idade dos eleitores e a sua intensão de apoiar a realização de plebiscito.

> Não há diferença significativa entre a percepção de cidadãos negros e brancos quanto à discriminação social.

Todo teste estatístico é feito sobre a hipótese nula. O resultado de um teste estatístico habilitará o pesquisador a (1) rejeitar a hipótese nula, ou (2) não rejeitar a hipótese nula. Não se deve usar a expressão “aceitar a hipótese nula” porque, ao aceitar, assume-se a hipótese nula como verdadeira quando, na verdade, só é possível dizer que não temos evidências suficientes para rejeitá-la.

2.3 VARIÁVEIS E DADOS

Para que hipóteses sejam testadas nós precisamos medir variáveis (Rauen, 2012). **Variável** é um atributo mensurável que pode mudar (ou variar) conforme mudam, por exemplo, o tempo, a localização geográfica ou o grupo de pessoas em que se está sendo realizada uma medida. **Dado** é o valor resultante da mensuração de uma variável em um indivíduo, ou em um caso particular.

Variáveis são classificadas em **quantitativas** (numéricas) e **qualitativas** (categóricas). Variáveis quantitativas são aquelas medidas em uma escala numérica. Por exemplo, em um

estudo sobre o processo eleitoral brasileiro, variáveis quantitativas podem ser a quantidade de candidatos a deputado federal na eleição de 2010, o número de votos nominais de um candidato e a quantidade de zonas eleitorais no Distrito Federal. Variáveis qualitativas são aquelas definidas por um rótulo. São exemplos: a base eleitoral do candidato, o estado de origem do eleitor e o sexo do eleitor.

O resultado numérico decorrente de um experimento é uma **variável aleatória**. Variáveis aleatórias podem ser **discretas** – quando assumem apenas um número contável de possibilidades e podemos determinar a probabilidade de valores específicos – ou **contínuas** – quando, conceitualmente, podem assumir qualquer valor no domínio real e podemos determinar a probabilidade de uma faixa de valores nesse domínio.

Variáveis também são classificadas quanto ao seu **nível de medição** (Quadro 2.2), ou seja, quanto à relação entre o que está sendo medido e a escala de valores que o representa.

No contexto de uma hipótese estatística, variáveis podem ser expressas sob a ótica da causalidade[1], em dois tipos: as de causa e as de efeito. Por exemplo, se considerarmos a declaração "o acréscimo nos recursos financeiros de campanha de um candidato repercute no aumento do número de votos recebidos por ele", a variável de causa (ou preditor) é "recurso financeiro" e a de efeito (ou desfecho) é "número de votos". A variável de causa é também chamada de **variável independente**, pois não depende de qualquer outra variável, ao menos no modelo considerado. A variável de efeito é chamada de **variável dependente** porque seu valor depende do valor da variável de causa.

[1] Causalidade é a relação entre um evento (a causa) e um segundo evento (o efeito), sendo que o segundo evento é uma consequência do primeiro (http://pt.wikipedia.org/wiki/Causalidade).

Quadro 2.2 – Classificação de variáveis quanto ao nível de medição.

Variável	Escala	Descrição
Categórica (qualitativa)	Nominal	Baseada na classificação de elementos, conforme um ou mais atributos, para a formação de grupos distintos. Podem ser comparadas apenas por relações de igualdade ou diferença. Os registros são qualitativos, referentes à categoria do sujeito, objeto ou acontecimento. Todos os elementos têm uma classificação. Números atribuídos servem apenas para identificar se o elemento pertence ou não a uma categoria. Não existe ordenação. Exemplo: raça, localização geográfica, tipo de residência, cor do cabelo, estado civil e cor dos olhos.
	Binária	Quando as características em estudo têm apenas duas categorias, a escala nominal é dita binária ou dicotômica. Exemplo: sexo (masculino, feminino); responder sim ou não a uma pergunta (Reside em Brasília? → sim, não).
	Ordinal	Distingue-se da nominal pela possibilidade de se estabelecer uma ordenação das categorias nas quais os dados são classificados de acordo com uma sequência com significado, sem, no entanto, quantificar a magnitude da diferença face aos outros indivíduos. Exemplos: - Desempenho do candidato da oposição no debate eleitoral: 1 – medíocre 2 – regular 3 – ótimo 4 – empolgante - As respostas podem ser obtidas numa escala Likert (escala ordinal de cinco categorias): 1 – discordo totalmente 2 – discordo parcialmente 3 – indiferente 4 – concordo parcialmente 5 – concordo totalmente
Numérica[2] (quantitativa)	Intervalar	Caracteriza-se pelo fato de que distâncias iguais entre pontos da escala correspondem a quantidades iguais da propriedade sendo medida. Nesse tipo de escala, o valor zero não significa ausência da propriedade medida. Por exemplo, 0º C não significa ausência de temperatura, mas, apenas, que o ponto de congelamento da água foi atingido (convenção). Em escalas intervalares há proporcionalidade entre intervalos, mas não entre valores. Qual seria o dobro de 0º C ou de -2º C? No entanto, pode-se dizer que a mudança de 5º C para 10º C equivale à passagem de 10º C para 15º C.
	Razão	Também chamada de proporcional, esta escala possui um zero absoluto como valor mínimo e não admite valores negativos. A razão entre dois valores da escala corresponde à razão entre dois valores de atributos. Exemplos: idade, altura, peso...

[2] Variáveis quantitativas são ainda: (1) **contínuas**, quando puderem assumir qualquer valor numérico dentro do intervalo definido para a escala (p.e. altura, peso...); (2) **discretas**, quando forem expressas por um conjunto limitado de valores, em geral inteiros, mesmo que o intervalo definido para a escala compreenda um continuum de valores (p.e. número de votos, número de filhos...).

Voltando ao exemplo da campanha, percebe-se que existem outros fatores, além dos recursos financeiros, que podem afetar o número de votos recebidos pelo candidato. Alguns destes fatores incluem uma declaração a favor ou contra determinado tema, a divulgação de um indicador econômico, uma denúncia de corrupção e outros. Todos são fatores que o pesquisador não levou em consideração, mas que podem influenciar o resultado de uma eleição. Esses fatores são denominados **variáveis de confusão ou intervenientes**. Em qualquer situação de pesquisa, deve-se levar em conta a influência destas variáveis. Se elas forem negligenciadas, as conclusões obtidas sobre o estudo podem não ser confiáveis.

Dados são os valores, ou medidas, de variáveis quantitativas e qualitativas pertencentes a um grupo de itens ou objetos. Os dados podem ser classificados em brutos e processados.

Dado bruto é o oriundo diretamente da fonte de que se origina, sem quaisquer modificações feitas pelo analista de dados. São difíceis de ser usados em análises, em geral, pela grande quantidade e por problemas ou inconsistências que precisam ser detectados e corrigidos. A análise de dados inclui a etapa de pré-processamento, que consiste em deixá-los na forma apropriada para a aplicação em modelos estatísticos. Todos os passos referentes ao pré-processamento devem ser documentados.

Dado processado ou tratado, por outro lado, é o dado pronto para a análise. O processamento pode incluir atividades como a união de bases isoladas, a divisão da base original em subconjuntos de variáveis, a transformação de alguns dados e a remoção de ***outliers*** (observações numericamente distantes do restante dos dados). É conveniente que padrões sejam estabelecidos para o processamento ou tratamento dos dados, de acordo com o tipo de dado utilizado. Assim, torna-se possível a obtenção de vantagens decorrentes do uso desses padrões no processamento de cada novo conjunto de dados. Aqui, também, é de suma importância o registro de todos os passos do processamento ou tratamento para que outros analistas possam replicar as análises.

Dados limpos (*tidy datasets* - http://vita.had.co.nz/papers/tidy-data.pdf) são fáceis de manipular, modelar e visualizar, e possuem uma estrutura específica: as variáveis são armazenadas em colunas, cada observação corresponde a uma linha e cada tabela, ou arquivo, armazena dados sobre um único tipo de observação.

Diz-se, frequentemente, que 80% do esforço de análise de dados é dispendido na limpeza dos dados (Dasu and Johnson, 2003), ou seja, o processo de tornar os dados prontos para a análise.

2.4 MÉTODO CIENTÍFICO

Para se medir variáveis com vistas a testar hipóteses científicas, precisamos dispor do método científico. Método científico é um conjunto de regras básicas de como devemos proceder a fim de produzirmos conhecimento dito científico, quer seja este um novo conhecimento, quer seja fruto de uma integração, correção ou evolução (expansão) de conhecimentos pré-existentes de uma determinada área.
(http://pt.wikipedia.org/wiki/M%C3%A9todo_cient%C3%ADfico)

Quando a pesquisa requer tratamento de números para o alcance dos objetivos, o **método quantitativo** deve ser usado. Quando a geração e o teste de uma teoria envolverem a obtenção de dados descritivos, usamos o **método qualitativo**.

Estudos quantitativos, em geral, seguem com rigor um plano estabelecido, baseado em hipóteses claramente indicadas e variáveis que são objeto de definição operacional.

A pesquisa qualitativa é normalmente direcionada ao longo do seu desenvolvimento e não tem por objetivo enumerar ou medir eventos. Geralmente, a pesquisa qualitativa não emprega instrumental estatístico, tem foco de interesse amplo e a obtenção de dados se dá mediante contato direto e interativo do pesquisador com a situação objeto de estudo. Nesse caso, o pesquisador apresenta conclusões com base na sua própria interpretação do fenômeno estudado e, quando aplicável, na perspectiva apresentada pelos participantes do estudo.

2.5 PRINCÍPIO DA COLETA DE DADOS

O primeiro passo na condução de uma pesquisa é a identificação de temas ou questões a serem investigados. Uma pergunta de pesquisa bem elaborada é aquela em que podem ser

identificados **quais** sujeitos ou casos devem ser estudados e **quais** variáveis são importantes. Também é importante considerar **como** os dados serão coletados de forma acurada, a fim de que sejam úteis ao alcance dos objetivos da pesquisa.

Basicamente, existem duas maneiras de se testar uma hipótese: na primeira, denominada **método observacional** (ou correlacional), os fenômenos são observados na medida em que acontecem naturalmente; na segunda, denominada **método experimental**, algumas variáveis ou aspectos do ambiente são manipulados e os respectivos efeitos, observados.

Em ambos os métodos, é desejável que as medições sejam calibradas de tal forma que preservem uniformidade ao longo do tempo e das situações. A apuração de uma urna de votação, por exemplo, deve computar o mesmo número de votos independentemente de quem fizer a leitura/contagem. Contudo, frequentemente existem discrepâncias entre as medições realizadas sobre o mesmo objeto, ao que chamamos de **erro de medição**.

Uma forma de minimizar a probabilidade de ocorrência de um erro de medição é por meio da determinação de propriedades de medida que garantam que a medição será feita de forma apropriada. A primeira propriedade é a **validade**, que avalia se um dado instrumento (aparelho de medição, questionário...) realmente mede o que foi projetado para medir. A segunda é a **confiabilidade**, que avalia se um instrumento pode ser interpretado de forma consistente em situações diferentes. Validade é uma característica necessária, mas não suficiente. Para ser válido, um instrumento deve ser antes confiável.

2.5.1 Método Observacional

O método observacional provê uma visão natural a uma pergunta de pesquisa, uma vez que não há influência externa sobre os acontecimentos e sobre a medição das variáveis (as variáveis não são enviesadas pelo pesquisador). São tipos de estudos observacionais:

> Estudo de **coorte** (estudo de seguimento, *cohort study*): um grupo definido de pessoas (coorte) é acompanhado ao longo de um período de tempo. O investigador observa os subgrupos da coorte expostos e não expostos a um determinado fator de

interesse, com objetivo de comparar os desfechos (p.e.: financiamento em diferentes campanhas, resultados eleitorais) que ocorrem nesses subgrupos.

Estudo **transversal**: é semelhante ao estudo de coorte, no entanto, nos estudos transversais todas as medições são feitas num único "momento", não existindo, portanto, período de seguimento dos indivíduos. Para levar a cabo um estudo transversal o investigador tem que, primeiro, definir a questão a responder, depois, definir a população a estudar e um método de escolha da amostra e, por último, definir os fenómenos a estudar e os métodos de medição das variáveis de interesse. (http://stat2.med.up.pt/cursop/print_script.php3?capitulo=desenhos_estudo&numero=6&titulo=Desenhos%20de%20estudo)

Estudo de **caso**: abordagem metodológica de investigação especialmente adequada quando se procura compreender, explorar ou descrever acontecimentos e contextos complexos, nos quais estão simultaneamente envolvidos diversos fatores e quando o investigador procura respostas para “como?” e “por quê?”.
(http://grupo4te.com.sapo.pt/mie2.html)

Estudo de **caso-controle**: o investigador compara um grupo de pessoas com um desfecho de interesse (caso) com outro grupo de pessoas da mesma população sem aquele desfecho (controle), com o objetivo de encontrar associações entre o desfecho e uma exposição prévia a um determinado fator. É um estudo particularmente útil no caso de desfechos raros, cujas medidas de exposição prévia ao fator de interesse são confiáveis.

No método observacional não se pode afirmar a causalidade entre duas variáveis, ou seja, que uma variável causa mudanças em outra. Como nesse método não existe manipulação da variável “causal” para a mensuração do efeito, apenas se pode dizer que as variáveis se correlacionam de alguma forma.

2.5.2 Método Experimental

O método experimental está associado à causalidade[3] dos eventos. Para se inferir causa e efeito, (1) ambos devem ocorrer próximos no tempo (contiguidade), (2) a causa deve ocorrer antes do efeito e (3) o efeito nunca deve ocorrer na ausência da causa. Estas condições implicam que a causalidade pode ser inferida por confirmação de evidência, ou seja, pela avaliação do grau de correlação entre eventos contíguos. A forma de se inferir causalidade é por meio da comparação de duas situações controladas: uma na qual a causa está presente e outra na qual a causa está ausente. Isto é exatamente o que o método experimental se propõe a fazer, provendo uma comparação de situações na qual a causa proposta está presente ou ausente.

A coleta de dados em um experimento pode ser feita de duas formas: na primeira, a variável independente é manipulada em grupos distintos de participantes; na segunda, a variável independente é manipulada no mesmo grupo de participantes.

Um **projeto de medidas independentes** é aquele em que a variável independente é manipulada em grupos distintos de participantes de modo que um grupo participa de uma condição experimental, enquanto o outro grupo participa de condição diferente. Os resultados são comparados ao final do experimento.

Um **projeto de medidas repetidas** ou **relacionadas** é aquele em que a variável independente é manipulada sobre o mesmo grupo de participantes, em condições distintas: experimental e de controle. Uma vez que a medição efetuada sobre um sujeito é comparada com outra efetuada sobre ele mesmo, esta técnica possibilita ao experimentador controlar as diferenças individuais dos participantes (p.e., intenção de voto ou motivação).

Em ambos os projetos de medidas, há sempre duas formas de variação: a **variação sistemática** ocorre quando o experimentador age sobre todos os participantes em uma condição, mas não em outra; a **variação não sistemática** resulta de fatores aleatórios que ocorrem entre as condições experimentais.

[3] O pensamento filosófico da causalidade, que descreve as relações de causa e efeito, foi tema explorado por pensadores como David Hume (http://www.utilitarian.net/hume/) e John Stuart Mill (http://plato.stanford.edu/entries/mill/).

Em um projeto de medidas repetidas, diferenças entre duas condições experimentais podem ser ocasionadas por apenas dois fatores: (1) o fator de manipulação imposto aos sujeitos em virtude do objetivo da pesquisa; (2) qualquer outro fator que possa afetar a forma na qual a pessoa executa o experimento quando são considerados dois momentos distintos. Para um projeto de medidas independentes, o primeiro fator também se aplica. Contudo, o segundo fator se deve a diferenças entre as características das pessoas alocadas em cada grupo.

Tanto em projetos de medidas repetidas quanto em projetos de medidas independentes é fundamental que se mantenham as variações não sistemáticas em um patamar mínimo. Uma forma de se conseguir isso é por meio da **aleatoriedade** ou **radomização**. A aleatoriedade é um processo importante porque elimina a maioria das fontes de variação não sistemáticas. Entretanto, isso significa que os efeitos de ordem podem se tornar importantes no estudo, sendo necessário o emprego do contrabalanço. Dessa forma, pode-se considerar, com maior grau de certeza, que as variações ocorridas entre as condições experimentais são devidas exclusivamente à manipulação da variável independente.

2.6 ESTATÍSTICA DESCRITIVA

A estatística descritiva é um ramo da estatística que aplica várias técnicas para descrever e sintetizar um conjunto de dados. É também vista como um conjunto de métodos para organizar, apresentar e descrever aspectos representativos do comportamento de uma variável, por meio de tabelas, gráficos e medidas que resumem a forma como a variável está distribuída.

2.6.1 Distribuição de Frequências

A **distribuição de frequências** é uma técnica estatística usada para apresentar uma coleção de objetos classificados, de modo a mostrar o número existente em cada classe.

A **tabela de distribuição de frequências** é uma forma de representação da frequência de cada valor distinto de uma variável. É um recurso estatístico aplicado especialmente a variáveis categóricas (nominais e ordinais), ou variáveis numéricas com pouca variação,

tais como, o número de filhos (0,1,2,3,..) ou o número de divisões de uma casa (1,2,3…). São exemplos as tabelas 2.1, 2.2 e 2.3.

Tabela 2.1 – Quantidade de eleitores por unidade da federação separados por sexo.®

Estado	Quantidade de Eleitores			
	Feminino	**Masculino**	**Não Informado**	**Total**
AC	253.853	249.733	16	503.602
AL	1.013.007	865.676	1	1.878.684
AM	1.104.814	1.081.939	308	2.187.061
AP	227.680	223.835	92	451.607
BA	5.260.736	4.878.808	7564	10.147.108
CE	3.244.114	2.952.802	7912	6.204.828
DF	1.006.079	869.715	542	1.876.336
ES	1.361.066	1.277.673	2337	2.641.076
GO	2.202.145	2.063.305	471	4.265.921
MA	2.322.972	2.229.024	2864	4.554.860
MG	7.741.335	7.303.874	15378	15.060.587
MS	917.289	866.974	1	1.784.264
MT	1.082.356	1.103.050	760	2.186.166
PA	2.564.588	2.572.262	2737	5.139.587
PB	1.504.017	1.356.438	226	2.860.681
PE	3.453.783	3.053.132	6482	6.513.397
PI	1.210.503	1.151.380	1243	2.363.126
PR	4.025.018	3.758.450	6211	7.789.679
RJ	6.385.017	5.562.326	18862	11.966.205
RN	1.230.195	1.124.997	1492	2.356.684
RO	559.338	557.573	0	1.116.911
RR	147.357	147.816	27	295.200
RS	4.343.620	4.002.031	0	8.345.651
SC	2.436.732	2.324.613	5	4.761.350
SE	739.498	654.890	0	1.394.388
SP	16.452.367	14.995.085	55365	31.502.817
TO	491.578	503.490	1	995.069

FONTE: Elaborado pelos autores com base no repositório de dados do TSE em abril de 2013.

Tabela 2.2 – Estado civil dos candidatos à eleição no estado do Paraná em 2012.®

Estado Civil	Frequências			
	f	f_r	A	A_r
CASADO(A)	20.082	0,682	20.082	0,682
DIVORCIADO(A)	1.626	0,055	21.708	0,737
SEPARADO(A) JUDICIALMENTE	752	0,026	22.460	0,763
SOLTEIRO(A)	6.255	0,212	28.715	0,975
VIÚVO(A)	738	0,025	29.453	1,000
Σ	29.453	1,000	–	–

FONTE: Elaborado pelos autores com base no repositório de dados do TSE em abril de 2013.

Frequências absoluta (f), relativa (f_r), acumulada (A) e relativa acumulada (A_r).

Para além destas variáveis, pode-se ainda determinar a tabela de distribuição de frequências de variáveis numéricas recodificadas em variáveis categóricas, como por exemplo, a idade recodificada em grupos etários, como ilustra a Tabela 2.3.

Outro importante recurso utilizado para se conhecer a distribuição de frequências de um conjunto de dados é o **histograma**, também conhecido como **diagrama de frequências**. É uma representação gráfica na qual um conjunto de dados é agrupado em **classes** uniformes, representadas por um retângulo cuja base horizontal é o **intervalo de classe**, e a altura vertical representa a **frequência** com que os valores desta classe estão presentes no conjunto de dados. A Figura 2.1 ilustra um histograma que apresenta a distribuição do quantitativo de eleitores por estado. É possível dizer, a partir da Figura 2.1, que nove estados possuem dois milhões de eleitores ou menos, e que apenas um estado possui mais de trinta milhões de eleitores.

Tabela 2.3 – Quantidade de eleitores por faixa etária.®

Faixa Etária	Quantidade de Eleitores			
	Feminino	Masculino	Não Informado	Total
16 ANOS	19.151	19.548	0	38.699
17 ANOS	27.564	28.356	0	55.920
18 A 20 ANOS	39.183	40.201	0	79.384
21 A 24 ANOS	45.326	45.060	0	90.386
25 A 34 ANOS	50.435	50.442	0	100.877
35 A 44 ANOS	51.783	51.585	2.520	105.888
45 A 59 ANOS	53.699	53.675	12.095	119.451
60 A 69 ANOS	49.957	50.272	8.139	108.368
70 A 79 ANOS	43.479	43.963	6.461	93.903
SUPERIOR A 79 ANOS	35.622	36.355	5.280	77.257
INVÁLIDA	41	30	2	73

FONTE: Elaborado pelos autores com base no repositório de dados do TSE em abril de 2013.

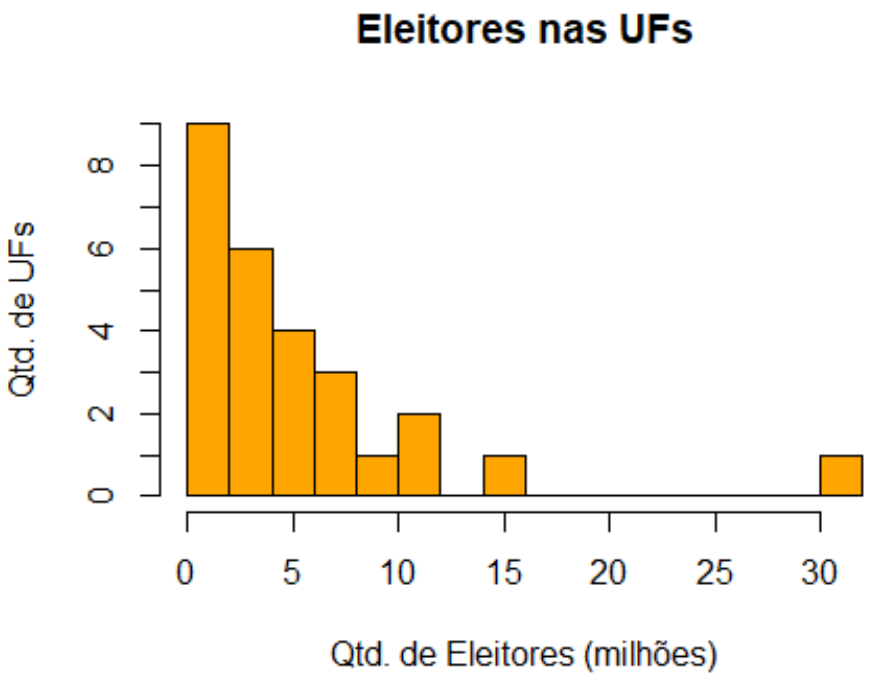


Figura 2.1 – Histograma: quantidade de eleitores nas UFs (repositório de dados do TSE de abril de 2013).®

2.6.2 Medidas de Tendência Central

Para se descrever um conjunto de dados por meio de uma variável quantitativa, faz-se necessário encontrar um indicador ou medida que, de algum modo, represente o conjunto como um todo. Em outras palavras, faz-se necessário dispor de um número que nos indique onde está o centro da distribuição, ou, então, o valor mais capaz de substituir todos os outros, também chamado de **tendência central** ou **centralidade**. Há três medidas comumente utilizadas: a média, a moda e a mediana.

A **média** é a medida de tendência central mais conhecida, é onde a distribuição está centrada, é o valor mais provável de uma distribuição. Segundo o Dicionário Aurélio, "é o valor que se determina segundo uma regra estabelecida a priori e que se utiliza para representar todos os valores da distribuição". Representa o "centro de massa" da distribuição, ou seja, o ponto em torno do qual se equilibram as discrepâncias positivas e negativas. Portanto, pode vir a ser um valor não presente na distribuição. Uma distribuição de frequências será deslocada para a direita ou para a esquerda, conforme a média variar para mais ou para menos. A média é calculada de acordo com a Equação (2.1)

$$média = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2.1}$$

onde Σ é somatório de um conjunto de valores, x_i são valores individuais dos dados, N é a quantidade de valores sobre os quais se deseja determinar a média.

Para efeitos de notação, este texto adota os símbolos utilizados nas equações (2.2) e (2.3): N – quantidade de valores da população; n – quantidade de valores da amostra; μ – média dos valores da população; $\bar{x}$ – média dos valores da amostra.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2.2}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.3}$$

Com base na Tabela 2.1, podemos determinar a média de eleitores por estado:

(503.602 + 1.878.684 + 2.187.061 + 451.607 + 10.147.108 + 6.204.828 + 1.876.336 + 2.641.076 + 4.265.921 + 4.554.860 + 15.060.587 + 1.784.264 + 2.186.166 + 5.139.587 + 2.860.681 + 6.513.397 + 2.363.126 + 7.789.679 + 11.966.205 + 2.356.684 + 1.116.911 + 295.200 + 8.345.651 + 4.761.350 + 1.394.388 + 31.502.817 + 995.069) / 27 = **5.227.512,78**

Neste exemplo se pode constatar que a média é bastante sensível a valores extremos. No caso da Tabela 2.1, os valores variam de 295.200 a 31.502.817. Portanto, parece inadequado afirmar que a média de 5.227.512,78 representa apropriadamente "todos os valores da distribuição", quando mais de nove estados possuem menos de dois milhões de eleitores. Esse aspecto será discutido no tópico medidas de dispersão.

A média é também conhecida como o **valor esperado**, **esperança matemática** ou **expectância** de uma variável aleatória e é definida como a soma das probabilidades de cada possibilidade de saída da experiência multiplicada pelo seu valor. Isso representa o valor médio ou esperado de uma experiência se ela for repetida muitas vezes. Se todos os eventos tiverem igual probabilidade, o valor esperado é a média aritmética definda pelas equações (2.2) e (2.3).

Para uma variável aleatória discreta X com valores possíveis x_1, x_2, x_3, ... e com as suas probabilidades representadas pela função $p(x_i)$, o valor esperado calcula-se pela série

$$E[X] = \sum_{i=1}^{\infty} x_i \, p(x_i) \qquad (2.4)$$

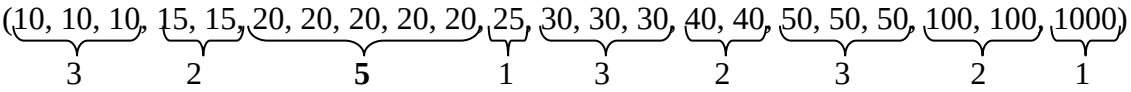
desde que a série seja convergente[4].

A **moda** é o valor que ocorre com maior frequência no conjunto de dados. Por exemplo, suponha que os valores de doações de campanha, em reais, recebidos por um candidato durante um jantar político correspondam ao conjunto a seguir:

(20, 10, 15, 100, 50, 20, 20, 30, 40, 20, 10, 15, 20, 50, 100, 1000, 40, 50, 10, 30, 25, 30)

[4] Para uma variável aleatória contínua X, o valor esperado é calculado mediante a integral de todos os valores da função de densidade de probabilidade $f(x)$: $E[X] = \int_{-\infty}^{\infty} x f(x) dx$

Para se determinar a moda, organizamos os dados em ordem crescente e contamos o número de ocorrências de cada valor. Neste caso, o valor com maior número de ocorrências é 20, que corresponde à moda.



Uma distribuição pode conter uma, duas (bimodal) ou mais modas (multimodal).

A **mediana** ocupa a posição central em um conjunto de dados ordenados e tem a propriedade de dividi-lo em duas partes iguais quanto ao número de seus elementos. Pode-se afirmar que 50% das observações são menores ou iguais à mediana e os 50% restantes são maiores ou iguais a essa medida.

Para se determinar a mediana, os valores dos dados também são ordenados de forma crescente e, então, é identificado o valor na posição central do conjunto. Para uma quantidade ímpar de dados, esse valor central é atribuído diretamente à mediana. Para uma quantidade par, haverá dois valores ao centro e a mediana será calculada pela média aritmética desses dois valores.

No exemplo das doações de campanha há um número de elementos (n) par e o valor da posição central é determinada por $(n+1)/2=(22+1)/2=11,5$. Nesse caso, a mediana deve ser determinada pela média aritmética dos valores nas posições 11 e 12, ou seja, $(25+30)/2=27,5$.

A mediana é mais robusta que a média na presença de *outliers* ou medidas com grande variabilidade. Observe que para o conjunto em questão, os valores da média e da mediana são, respectivamente, 77,5 e 27,5. Contudo, se retirarmos a doação de R$ 1000,00, cuja magnitude se distancia acentuadamente dos demais, verificamos que os novos valores da média e da mediana se alteram, respectivamente, para 33,57 e 25. Verifica-se que a média caiu para a metade (50%) da primeira estimativa, enquanto a mediana variou apenas 10%.

2.6.3 Medidas de Dispersão

Quantificar o espalhamento, ou dispersão, dos dados é importante para compreendê-los. Em qualquer grupo de dados, os valores numéricos são distintos e apresentam desvios em relação à tendência central. As medidas de dispersão servem, então, para avaliar o quanto os dados se assemelham, com base no quanto estão distantes do valor central.

É fácil mostrar que a média é insuficiente para descrever um grupo de dados. Dois grupos podem ter a mesma média e, no entanto, serem muito diferentes na amplitude (ou faixa) de variação de seus dados. Por exemplo, considere os grupos A(5, 5, 5), B(4, 5, 6) e C(0, 5, 10). A média dos três grupos é a mesma (5), contudo, no grupo A não há variação entre os dados (5-5=0), enquanto no grupo B a variação (6-4=2) é menor do que no grupo C (10-0=10). Dessa forma, uma maneira mais completa de se apresentar os dados é por meio da associação da média a uma medida de dispersão.

São medidas de dispersão a amplitude, a amplitude interquartil, a variância, o desvio padrão e o coeficiente de variação.

A **amplitude** é a diferença entre o maior e o menor valor do conjunto de dados. Considerando a Tabela 2.1, a amplitude é dada por 31.502.817-295.200=31.207.617. Por utilizar os valores extremos, a amplitude é dramaticamente afetada por eles. Se desconsiderarmos, na Tabela 2.1, o valor referente ao estado de SP, a amplitude passa a ser 15.060.587-295.200=14.765.387, ou seja, menos da metade da calculada anteriormente.

Uma forma de contornar esse problema é utilizar a **amplitude interquartil** em que são desconsiderados os valores do topo e da base, na razão de 25% em cada extremo. Assim, a amplitude é calculada sobre os demais 50% de elementos que estão ao centro, sem ser afetada pelos extremos. Para tanto, é preciso que os quartis sejam determinados. Segundo o Dicionário Aurélio, quartil é “qualquer das separatrizes que dividem a área de uma distribuição de frequências em domínios de área igual a múltiplos inteiros de um quarto da área total”. Para determiná-los, inicialmente calculamos a mediana, também chamada de **segundo quartil**, que divide o conjunto de dados em duas partes iguais. Em seguida, determinamos a mediana da primeira parte, denominada **primeiro quartil**, e a mediana da segunda parte, ou **terceiro quartil**. O Quadro 2.3 ilustra esse processo.

Quadro 2.3 – Determinação da amplitude interquartil para os dados da Tabela 2.1.

UF	Qtd. Eleitores	Passo 1	Passo 2	Quartil	Amplitude Interquartil
RR	295.200				
AP	451.607				
AC	503.602				
TO	995.069				
RO	1.116.911				
SE	1.394.388				
MS	1.784.264		1.784.264	1º Quartil	4.729.133
DF	1.876.336				
AL	1.878.684				
MT	2.186.166				
AM	2.187.061				
RN	2.356.684				
PI	2.363.126				
ES	2.641.076	2.641.076		2º Quartil	
PB	2.860.681				
GO	4.265.921				
MA	4.554.860				
SC	4.761.350				
PA	5.139.587				
CE	6.204.828				
PE	6.513.397		6.513.397	3º Quartil	
PR	7.789.679				
RS	8.345.651				
BA	10.147.108				
RJ	11.966.205				
MG	15.060.587				
SP	31.502.817				

As medidas de dispersão servem, também, para avaliar o grau de representação da média, ou seja, o quanto a média pode ser considerada um modelo para o conjunto de dados em estudo. Por exemplo, ao verificarmos a distância entre a quantidade de eleitores do estado de Roraima e a média calculada, obtemos 295.200-5.227.512,78=-4.932.312,78. Observa-se que o desvio é negativo, o que significa que a média superestimou a quantidade de eleitores para Roraima. Portanto, o tamanho do desvio é um indicador do grau de representação da média. Considerando-se todo o conjunto de dados, uma forma de se medir o desvio poderia ser dada pela soma dos desvios individuais de cada elemento, ou **erro total**, que para uma população é calculada segundo a Equação (2.5).

$$erro\_total = \sum_{i=1}^{N}(x_i - \mu) \tag{2.5}$$

O problema do erro total é que o resultado da soma das diferenças dos elementos em relação à média é sempre zero, o que nos induziria a afirmar que a média é uma representação perfeita do conjunto de dados. Desenvolvendo matematicamente, temos

$$\begin{aligned} erro\_total &= \sum_{i=1}^{N}(x_i - \mu) \\ &= \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \mu \\ &= N\mu - N\mu = 0 \end{aligned}$$

Uma forma de contornar esse fato é considerar a soma dos quadrados das diferenças, ou **soma dos erros quadrados** (*SEQ*).

$$SEQ = \sum_{i=1}^{N}(x_i - \mu)^2 \tag{2.6}$$

No caso da Tabela 2.1, o *SEQ* é da ordem de um quatrilhão. Fica evidente que o *SEQ* depende da quantidade e da magnitude dos dados do conjunto. Para contornar esse problema, dividimos o *SEQ* pelo número de observações. A essa medida, dá-se o nome de **variância**, representada por $\boldsymbol{\sigma^2}$.

$$\sigma^2 = \frac{SEQ}{N} \tag{2.7}$$

Quando estimada sobre uma amostra, a variância é dada por

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{2.8}$$

O problema de se utilizar a variância é que ela provê medidas em unidades quadradas. Não faz sentido dizer, por exemplo, que o erro quadrático médio é de 1.000 eleitores ao quadrado. Por essa razão, a medida mais utilizada é o **desvio padrão** (***σ*** **–** população; ***s*** **-** amostra), que consiste na raiz quadrada da variância.

O **coeficiente de variação** (*CV*) é uma medida de dispersão empregada para estimar a precisão da média, e é representado pelo desvio-padrão expresso como porcentagem da média. Sua principal qualidade é a capacidade de comparação de distribuições diferentes.

$$CV_{população} = \frac{\sigma}{\mu} \text{ e } CV_{amostra} = \frac{s}{\bar{x}} \tag{2.9}$$

2.6.4 Simetria de Dados

Uma distribuição de frequências é uma **distribuição simétrica** quando as medidas de tendência central (média, moda e mediana) são coincidentes. Isso significa que a média é o ponto de simetria pela qual passa um eixo imaginário que divide o gráfico de distribuição em intervalos de mesma magnitude à esquerda e à direita, com as mesmas concentrações de valores (ver Figura 2.2).

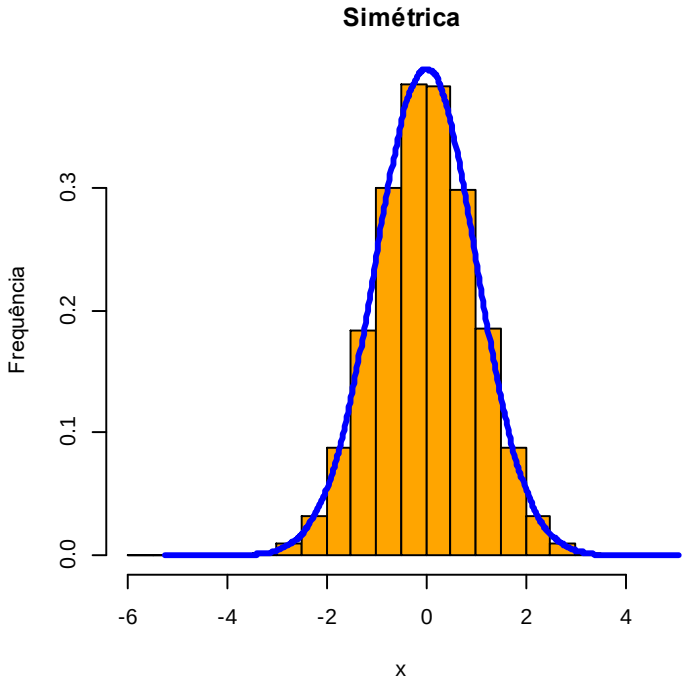


Figura 2.2 – Distribuição simétrica.®

A Figura 2.2 também ilustra uma curva, destacada pela linha azul (linha de base), que é considerada a forma limite do histograma de distribuição de frequências, admitindo-se que o intervalo de classe seja cada vez menor à medida que aumenta o tamanho da amostra. Essa curva é conhecida pelo nome de **curva normal**, gaussiana, ou curva em forma de sino, e é completamente descrita pelos parâmetros da média e do desvio padrão (Figura 2.3), em geral representada pela notação $N\sim(\mu;\sigma)$. A distribuição normal é a mais conhecida e utilizada na ciência estatística. Muitas variáveis aleatórias[5] de ocorrência natural ou de processos práticos obedecem a essa distribuição.

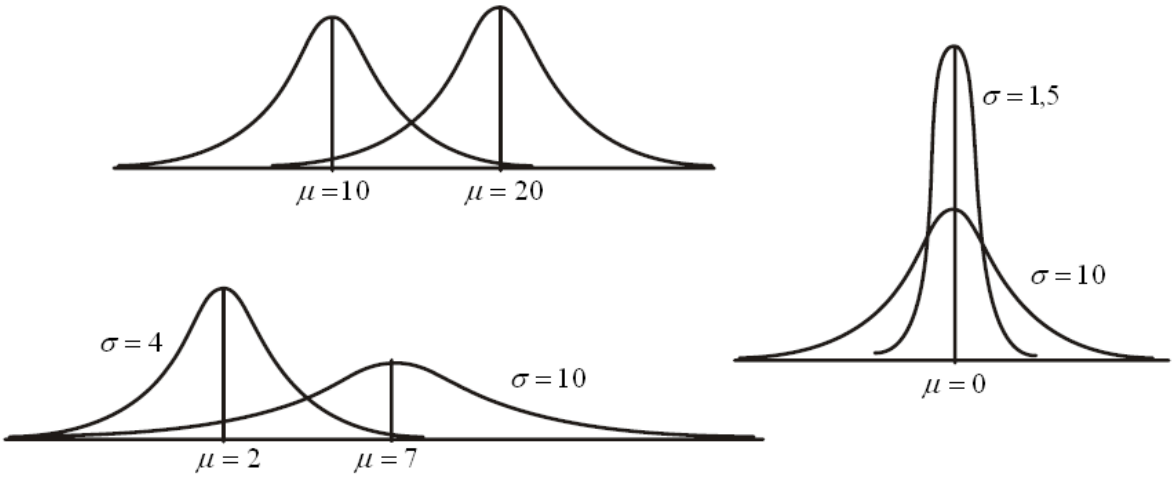


Figura 2.3 – Curvas normais com diferentes médias e desvios-padrões.

Em situações práticas, muitas vezes não se obtém uma distribuição normal. Sempre que a curva da distribuição se afastar do eixo imaginário, será considerada assimétrica, com certo grau de afastamento. Esse afastamento pode acontecer do lado direito (**assimetria positiva**), quando a média for maior que a mediana, ou do lado esquerdo (**assimetria negativa**) da distribuição, quando a média for menor que a mediana (ver Figura 2.4). A assimetria causa imenso efeito sobre a forma de se representar os dados e sobre os testes estatísticos que podem ser realizados sobre eles. Em distribuições assimétricas, a média não deve ser utilizada como representação da distribuição. Lembre-se: a média é sensível a valores extremos. Assim, outras medidas, como a mediana, são mais apropriadas. A assimetria de uma distribuição pode ser calculada por

[5] Em estatística precisamos atribuir uma descrição numérica aos resultados dos experimentos. Variáveis aleatórias associam números reais aos eventos. A partir disso, a probabilidade do evento será definida por funções de probabilidade ou de densidade de probabilidade da variável aleatória, as quais ligam um particular valor à probabilidade daquele evento. Esses conceitos serão abordados na seção ...

$$b_1 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{x_i - \bar{x}}{s}\right]^3 \tag{2.10}$$

ou pelas equações simplificadas de Karl Pearson

$$b_1 = \frac{\mu - mo}{\sigma} = 3\frac{\mu - m}{\sigma} \tag{2.11}$$

onde μ é a média, *m* é a mediana, *mo* é a moda e σ é o desvio padrão.

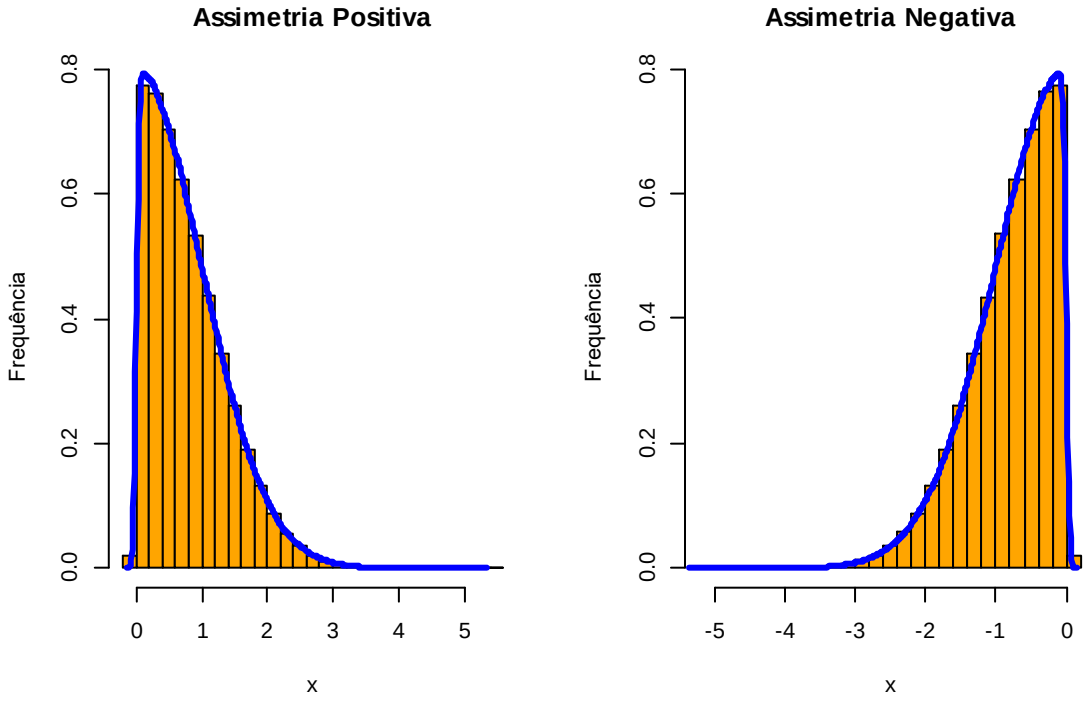


Figura 2.4 – Distribuições assimétricas.®

Diversas transformações podem ser aplicadas a distribuições assimétricas para dar-lhes uma conformação mais próxima à da normal, tais como a aplicação de logaritmos decimais ou naturais, aplicação da função inversa (1/*x*) ou da raiz quadrada ($\sqrt{x}$), entre outras, a depender da forma da assimetria (positiva, quando a média é maior que a mediana; negativa, quando a média é menor que a mediana). Outras técnicas podem envolver a remoção de valores extremos ou a padronização dos dados em torno da média.

2.6.5 Medidas de Achatamento ou Curtose

Curtose é o grau de achatamento da distribuição, ou o quanto uma curva de frequência será achatada em relação a uma curva normal de referência. A forma da curva de distribuição em relação à curtose pode ser leptocúrtica, mesocúrtica ou platicúrtica (Figura 2.5). Quando a distribuição apresenta uma curva de frequência mais fechada que a normal (ou mais aguda em sua parte superior), ela recebe o nome de **leptocúrtica**. Quando a distribuição apresenta uma curva de frequência mais aberta que a normal (ou mais achatada na sua parte superior), ela é chamada de **platicúrtica**. A curva normal, que é a referencial, recebe o nome de **mesocúrtica**.

O grau de curtose de uma distribuição é determinado pelo coeficiente de curtose (k), ou coeficiente percentílico de curtose.

$$k = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})} \tag{2.12}$$

onde Q_3 é o terceiro quartil, Q_1 é o primeiro quartil, P_{90} é o nonagésimo percentil e P_{10} é o décimo percentil.

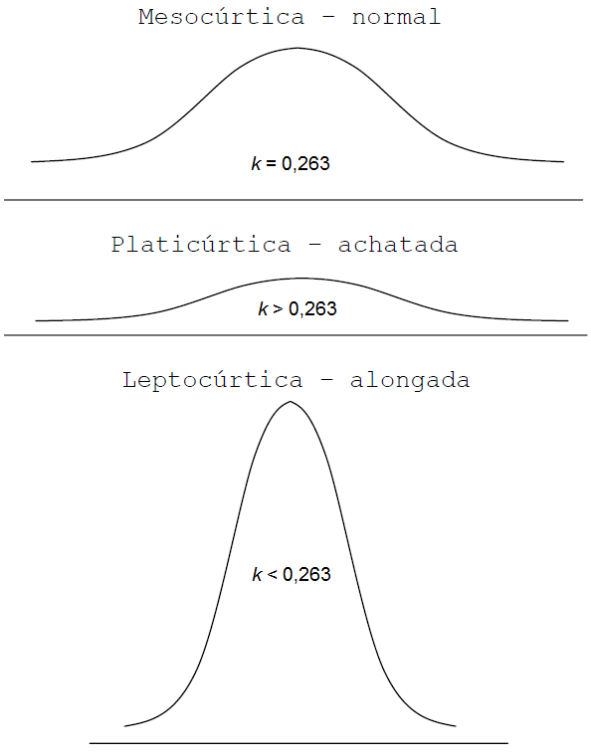


Figura 2.5 – Curtose da distribuição.

2.7 R

R é uma linguagem e um ambiente de desenvolvimento integrado para cálculos estatísticos e gráficos. Foi criada originalmente por Ross Ihaka e por Robert Gentleman no Departamento de Estatística da Universidade de Auckland, Nova Zelândia, e vem sendo desenvolvido por um esforço colaborativo de pessoas em vários locais do mundo. O nome R provém em parte das iniciais dos criadores e também de um jogo figurado com a linguagem S (da Bell Laboratories, antiga AT&T). R é uma linguagem e ambiente similar ao S - pode ser considerado uma implementação distinta do S; embora com importantes diferenças, muitos códigos escritos para o S rodam inalterados no R.

R é também altamente expansível com o uso de **pacotes**, que são bibliotecas para funções específicas ou áreas de estudo específicas. Um conjunto de pacotes é incluído com a instalação de R, com muitos outros disponíveis na rede de distribuição do R (em inglês CRAN – *The Comprehensive R Archive Network* - cran.r-project.org). A linguagem R é largamente usada entre estatísticos e *data miners* para desenvolver rotinas de estatística e análise de dados. A popularidade do R aumentou substancialmente nos últimos anos.

2.7.1 O ambiente R

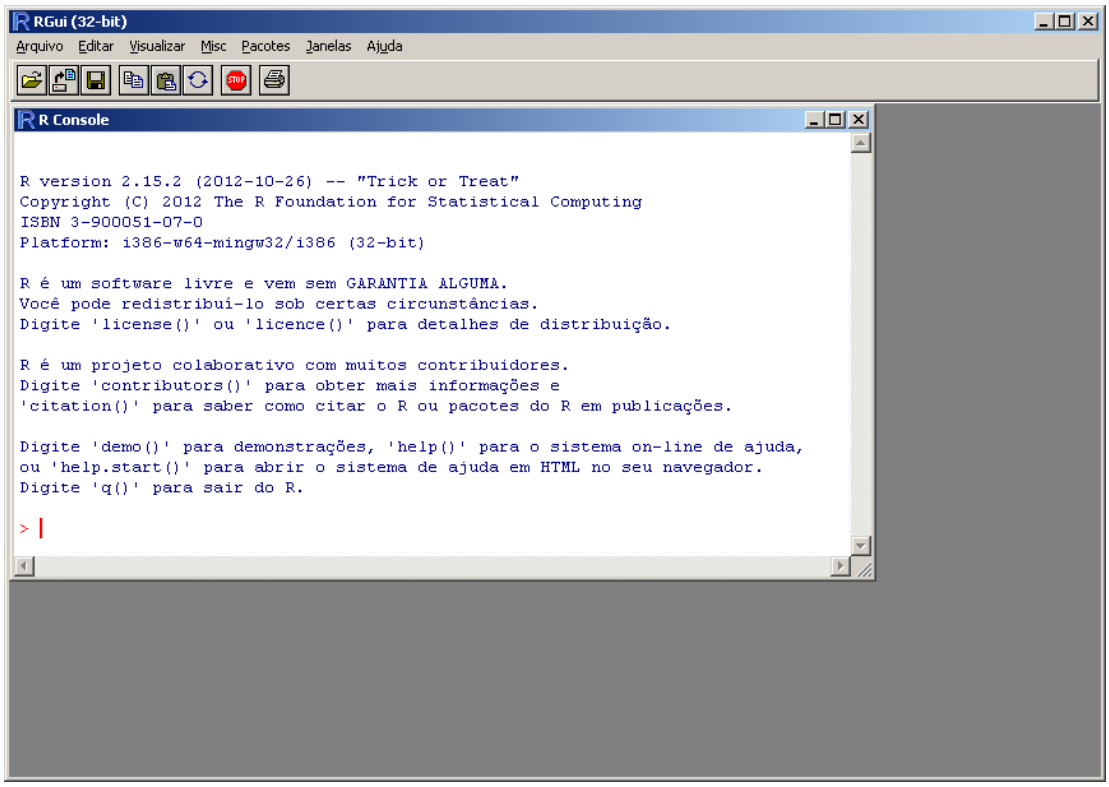


Figura 2.6 – Ambiente R com a tela de console pronta para receber comandos.

Ao iniciar o R em seu computador, você verá algo semelhante à Figura 2.6, com a tela de console aberta, pronta para receber comandos. O símbolo > é o sinal de pronto (“*prompt”)* do R e indica o local onde os comandos devem ser digitados.

Por convenção, este texto mostrará os comandos do R em fonte `Courier New`. As linhas indicadas pelo símbolo # são comentários do código e serão ignoradas pelo R.

Tutoriais:

http://www.stats4stem.org/r-tutorials
http://omanuscritodetunbridgewells.blogspot.com.br/2012/
https://sites.google.com/site/marcosfs2006/material_r
http://www.harding.edu/fmccown/r/
http://vincentarelbundock.github.io/Rdatasets/datasets.html

2.8 LABORATÓRIO 1

2.8.1 Explorando um arquivo de dados

R

```
# remove todas as variáveis da memória
rm(list = ls(all = TRUE))
# leitura do arquivo
tse <- read.table("consulta_cand_2012_PR.txt",sep=";") # com fatores
# informações de resumo
nrow(tse)
ncol(tse)
dim(tse)
summary(tse)
str(tse)
# nominando as variaveis
names(tse)<-
c("DATA_GERACAO","HORA_GERACAO","ANO_ELEICAO","NUM_TURNO",
  "DESCRICAO_ELEICAO","SIGLA_UF","SIGLA_UE","DESCRICAO_UE",
  "CODIGO_CARGO","DESCRICAO_CARGO","NOME_CANDIDATO",
  "SEQUENCIAL_CANDIDATO","NUMERO_CANDIDATO","NOME_URNA_CANDIDATO",
  "COD_SITUACAO_CANDIDATURA","DES_SITUACAO_CANDIDATURA","NUMERO_PARTIDO",
  "SIGLA_PARTIDO","NOME_PARTIDO","CODIGO_LEGENDA","SIGLA_LEGENDA",
  "COMPOSICAO_LEGENDA","NOME_LEGENDA","CODIGO_OCUPACAO",
  "DESCRICAO_OCUPACAO","DATA_NASCIMENTO",
  "NUM_TITULO_ELEITORAL_CANDIDATO","IDADE_DATA_ELEICAO",
  "CODIGO_SEXO","DESCRICAO_SEXO","COD_GRAU_INSTRUCAO",
  "DESCRICAO_GRAU_INSTRUCAO","CODIGO_ESTADO_CIVIL",
  "DESCRICAO_ESTADO_CIVIL","CODIGO_NACIONALIDADE",
  "DESCRICAO_NACIONALIDADE","SIGLA_UF_NASCIMENTO",
  "CODIGO_MUNICIPIO_NASCIMENTO","NOME_MUNICIPIO_NASCIMENTO",
  "DESPESA_MAX_CAMPANHA","COD_SIT_TOT_TURNO","DESC_SIT_TOT_TURNO")
# names(tse)[3]="ANO_ELEICAO"
# names(tse)[34]="DESCRICAO_ESTADO_CIVIL"
summary(tse)
# domínio da variável
t<-table(tse$DESCRICAO_ESTADO_CIVIL,tse$ANO_ELEICAO)
t2<-t
for (i in 2:length(t2)) { t2[i]=t2[i]+t2[i-1] }
cbind(t,t/sum(t),t2,t2/sum(t))
```

2.8.2 Tabelas de distribuição de frequências

R

```
table(tse$DESCRICAO_ESTADO_CIVIL,tse$ANO_ELEICAO)
table(tse$DESCRICAO_ESTADO_CIVIL,tse$DESCRICAO_SEXO)
cbind(table(tse$DESCRICAO_ESTADO_CIVIL,tse$ANO_ELEICAO),
table(tse$DESCRICAO_ESTADO_CIVIL,tse$DESCRICAO_SEXO))
table(tse$DESCRICAO_GRAU_INSTRUCAO,tse$DESCRICAO_SEXO)
table(tse$DESCRICAO_GRAU_INSTRUCAO,tse$ANO_ELEICAO)
cbind(table(tse$DESCRICAO_GRAU_INSTRUCAO,tse$DESCRICAO_SEXO),
      table(tse$DESCRICAO_GRAU_INSTRUCAO,tse$ANO_ELEICAO))
```

2.8.3 Histogramas

R

```
# remove todas as variáveis da memória
rm(list = ls(all = TRUE))
# leitura do arquivo
tse <- read.table("consulta_cand_2012_PR.txt",sep=";") # com fatores
# renomeia a variável V26
names(tse)[26] <- "DATA_NASCIMENTO"
# converte o campo DATA_NASCIMENTO para o formato data
dtnasc <- as.Date(tse$DATA_NASCIMENTO, format="%d/%m/%Y")
# determina a idade dos candidatos na data da eleição de 2012 (7/10/2012)
dteleicao <- as.Date("7/10/2012", format="%d/%m/%Y")
idade2012 <- difftime(dteleicao,dtnasc,units="days")
idade2012
# converte o resultado para a idade em anos
idade2012 <- floor(as.numeric(idade2012)/365.25)
# histograma
hist(idade2012, breaks=20, main="Candidatos do Paraná",
 xlab="Idade",ylab="Frequência", col="orange")
```

O que deu errado? Tente descobrir com o código a seguir.

```
max(idade2012)
idade2012[idade2012 > 150]
tse$DATA_NASCIMENTO[idade2012 > 150]
hist(idade2012[idade2012 < 150], breaks=20, main="Candidatos do Paraná",
 xlab="Idade",ylab="Frequência", col="orange")
```

2.8.4 Medidas de tendência central e de dispersão

R

```
# utilize a variável IDADE_DATA_ELEICAO calculada no exercício anterior
#
# Tendência central
#
# Média
mean(tse$IDADE_DATA_ELEICAO)
# Mediana
median(tse$IDADE_DATA_ELEICAO)
# Moda
mod<-table(tse$IDADE_DATA_ELEICAO)
as.numeric(names(mod[mod==max(mod)]))
#
# Dispersão
#
# Amplitude
max(tse$IDADE_DATA_ELEICAO)-min(tse$IDADE_DATA_ELEICAO)
# Amplitude interquartil
q<-quantile(tse$IDADE_DATA_ELEICAO)
q[[4]]-q[[2]]
# Variância
var(tse$IDADE_DATA_ELEICAO)
# Desvio padrão
sd(tse$IDADE_DATA_ELEICAO)
# coeficiente de variação
sd(tse$IDADE_DATA_ELEICAO)/mean(tse$IDADE_DATA_ELEICAO)
# Quartis
quantile(tse$IDADE_DATA_ELEICAO)
# Percentis
quantile(tse$IDADE_DATA_ELEICAO, probs = seq(0, 1, 0.10), na.rm = FALSE)
```

3 DISTRIBUIÇÃO DE PROBABILIDADE

Neste módulo serão apresentados os ...

3.1 PROBABILIDADE

Segundo o Dicionário Aurélio, **probabilidade** é o "motivo ou indício que deixa presumir a verdade ou a possibilidade dum fato; verossimilhança". Em outras palavras, probabilidade é: a medida da ocorrência de um evento; a quantificação do conhecimento sobre um determinado fato ou evento; uma medida da informação ou crença sobre a ocorrência de um determinado evento; uma medida baseada em registros de experiências passadas sobre a ocorrência de um evento. Matematicamente, a probabilidade ***p*** de um evento[6] ocorrer é expressa em termos do espaço amostral[7].

$$p = \frac{tamanho\_do\_evento}{tamanho\_do\_espaço\_amostral} \tag{3.1}$$

Distribuições de frequência podem ser pensadas em termos de probabilidade. Quando se observa atentamente um histograma, podem ser identificados indícios que permitem conjecturar sobre o comportamento da variável em estudo. Considere, por exemplo, o histograma da Figura 3.1, que ilustra a relação entre a quantidade de votos nulos e a quantidade de seções eleitorais em que esses votos ocorreram no estado do Paraná, no primeiro turno das eleições de 2012.

[6] Evento é o conjunto de resultados de um experimento, um subconjunto do espaço amostral.
[7] Espaço amostral é o conjunto de todos os possíveis resultados de um experimento.

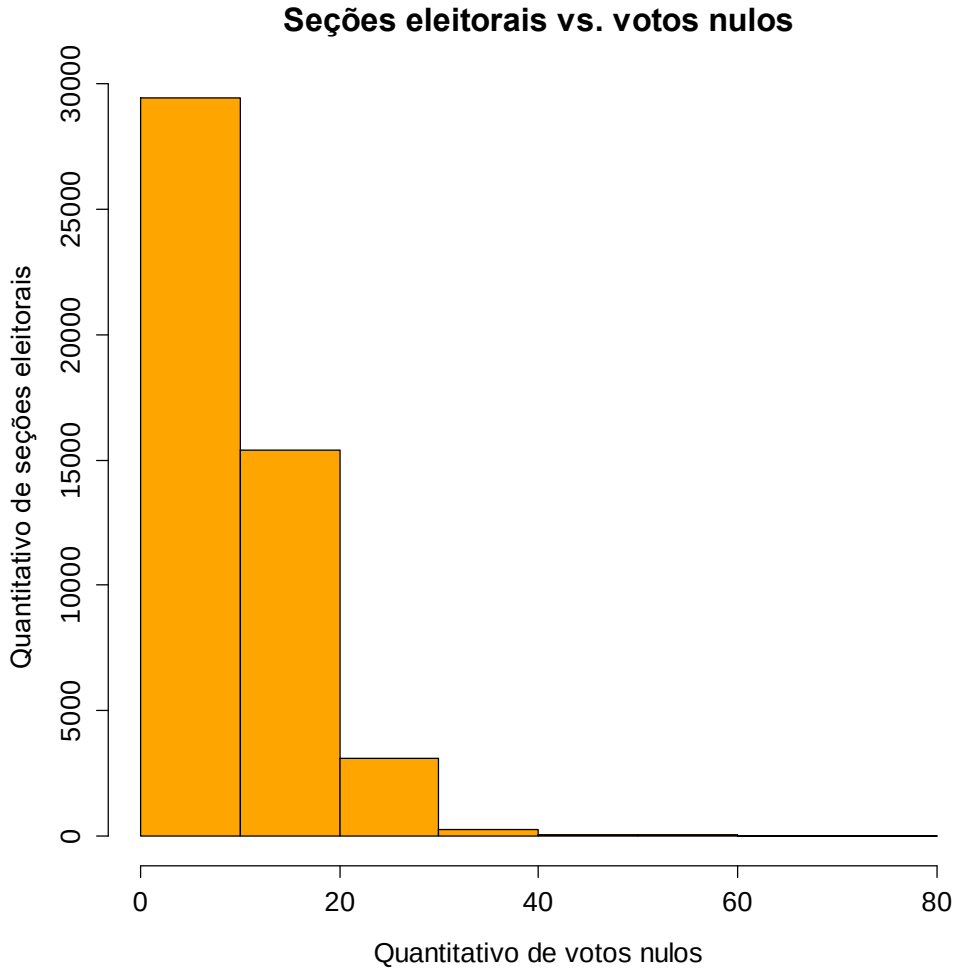


Figura 3.1 – Relação entre a quantidade de votos nulos e a quantidade de seções eleitorais em que esses votos ocorreram no estado do Paraná, no primeiro turno das eleições de 2012.®

É possível afirmar que a quantidade de votos nulos[8] variou de um a dez em cerca de 30.000 seções eleitorais do estado do Paraná. Com base nas quatro barras visíveis do gráfico, esse número pode ser estimado em termos percentuais aproximados[9]: 30.000 / (30.000+15.000+3.000+200) * 100 ≈ 62%. Ou seja, em 62% das seções eleitorais, a quantidade de votos nulos variou de um a dez.

[8] A anulação do voto ocorre quando o eleitor tecla um número inválido, que não está associado a nenhum candidato, e confirma a sua opção. A urna eletrônica emite um aviso indicando que o número é inválido antes da confirmação.

[9] A probabilidade de um evento também pode ser entendida como a razão entre a área do evento e a área total do histograma. Tomado-se por referência a Figura 3.1, cujo intervalo de classe é igual a 10, temos que a probabilidade de ocorrência de 1 a 10 votos nulos nas seções eleitorais do Paraná, sob as mesmas condições da eleição de 2012, é: $p = \frac{área\_do\_evento}{área\_total} = \frac{10*30000}{10*(30000+15000+3000+200)} \approx 0.62$

Assumindo-se, por hipótese, que os fatores que levam à anulação de votos não se alteram muito de um período eleitoral a outro e considerando as informações da Figura 3.1, seria razoável presumir que, mantidas as condições de 2012, a possibilidade de que a quantidade de votos nulos estivesse entre um e dez nas seções eleitorais do Paraná na próxima eleição fosse de aproximadamente 62%. Para a faixa de dez a vinte votos nulos, essa possibilidade mudaria para 31% (≈ 15.000 / 48.200 * 100).

Em outras palavras, pode-se supor que, nas próximas eleições, a probabilidade de que a quantidade de votos nulos esteja entre um e dez nas seções eleitorais do Paraná é de 0,62. Valores de probabilidade variam de 0 (quando não há probabilidade de um dado evento ocorrer) a 1 (quando o evento definitivamente ocorre). Assim, para o evento *quantidade de votos nulos entre um e dez*, há aproximadamente 6 chances (ou 6,2 chances) do evento acontecer em cada 10 seções apuradas.

Para qualquer distribuição de frequências é possível calcular, em tese, a probabilidade de ocorrência, ou **distribuição de probabilidade**, dos dados ou valores. Em estatística, uma distribuição de probabilidade descreve a probabilidade que uma variável tem de assumir determinado valor ao longo de um espaço de valores. Pode ser discreta (p.ex., a quantidade de votos nulos) ou contínua (p.ex., a precipitação pluviométrica durante o mês).

No caso discreto, a distribuição de probabilidade é chamada **função de massa**, ou **função de probabilidade**[10], e associa uma probabilidade à ocorrência dos valores contidos em uma classe da variável. A frequência relativa, comumente encontrada nas tabelas de frequência, é uma estimativa de probabilidade. A Tabela 3.1, por exemplo, apresenta com exatidão as frequências relativas das colunas do histograma da Figura 3.1. Pode-se constatar que, na verdade, a probabilidade de ocorrerem votos nulos na quantidade de um a dez nas seções eleitorais do Paraná é de 0,6103.

[10] Regras da função de probabilidade P de uma variável aleatória *X*: $P(X = x_i) \geq 0$; $\sum_{i=0}^{n} P(X = x_i) = 1$, onde *n* é o número de elementos do espaço amostral. Funções de probabilidade são utilizadas para variáveis aleatória discretas: as probabilidades são concentradas em alguns pontos possíveis; em geral, derivam de um processo de contagem.

Tabela 3.1– Seções eleitorais em que ocorreram votos nulos no estado do Paraná, no primeiro turno das eleições de 2012.®

Quantidade de	**Seções Eleitorais**			
Votos Nulos	f	f_r	A	A_r
1 – 10	29.442	**0,6103**	29.442	0,6103
11 – 20	15.412	**0,3195**	44.854	0,9298
21 – 30	3.085	**0,0640**	47.939	0,9938
31 – 40	238	**0,0049**	48.177	0,9987
41 – 50	22	**0,0005**	48.199	0,9992
51 – 60	23	**0,0005**	48.222	0,9997
61 – 70	13	**0,0003**	48.235	0,9999
71 – 80	3	**0,0001**	48.238	1,0000
Σ	48.238	1,0000	–	–

FONTE: Elaborado pelos autores com base no repositório de dados do TSE em abril de 2013.

Frequências absoluta (f), relativa (f_r), acumulada (A) e relativa acumulada (A_r).

Quando se deseja conhecer a probabilidade de ocorrência de duas ou mais classes, basta somar as probabilidades de cada classe. Assim, da Tabela 3.1, a probabilidade de ocorrência de votos nulos entre 1 e 20 corresponde a 0,93 (0,6103 + 0,3195 = 0,9298). Lembre-se que a soma das probabilidades de todas as classes (ou a soma das frequências relativas) é sempre igual a 1.

Exercício: qual a probabilidade de ocorrerem 31 votos nulos ou mais?

Considere, agora, os dados referentes às idades completas dos candidatos do Paraná no dia do primeiro turno das eleições de 2012 (7 de outubro). A distribuição está ilustrada na Figura 3.2. Observe que, da forma como o intervalo de classe foi definido, com tamanho igual a cinco, 15 classes (ou barras do histograma) podem ser identificadas nessa figura. As frequências relativas são exibidas sobre as barras. Então, seguindo o raciocínio anterior, pode-se dizer, por exemplo, que na data do primeiro turno das eleições de 2012 a probabilidade de um candidato do Paraná ter idade completa entre 20 e 40 anos (barras em vermelho da Figura 3.2) era de 0,336 (0,034+0,064+0,096+0,136).

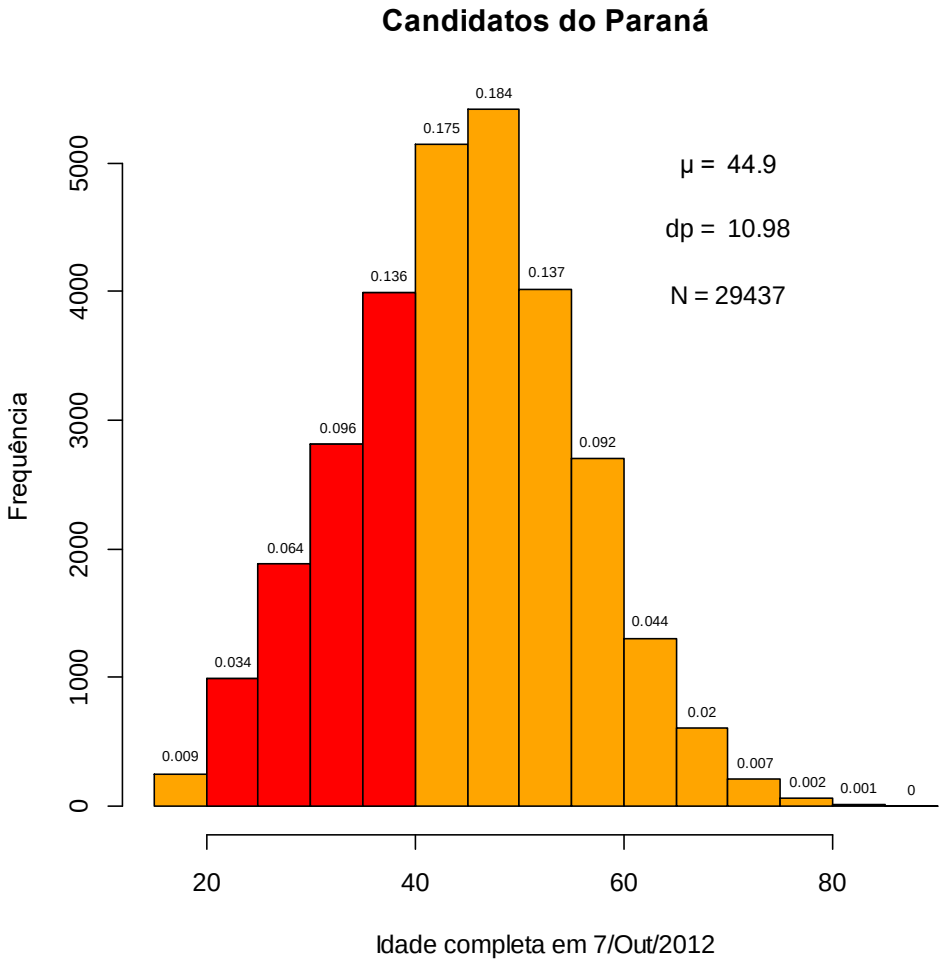


Figura 3.2 – Distribuição das idades completas dos candidatos do Paraná no dia do primeiro turno das eleições de 2012, com a frequência relativa (f_r) de cada classe destacada sobre a barra. A probabilidade de um candidato ter idade entre 20 e 40 anos corresponde à soma das f_r das barras em vermelho.®

Tente imaginar, agora, o que acontece quando o tamanho do intervalo de classe é reduzido. Suponha, por exemplo, que a classe de 45 a 50 anos, correspondente à maior barra da Figura 3.2, com aproximadamente 5.500 ocorrências, seja dividida em dois novos intervalos: [45, 47,5) e [47,5, 50). Considere, então, apenas os valores inteiros compreendidos nesses dois novos intervalos, visto que a variável idade completa é discreta e não faria sentido dizer, neste caso, que alguém tem 45,3 anos. Suponha, ainda, que das 5.500 ocorrências da classe original, 2.500 tenham se ajustado ao primeiro novo intervalo e 3.000, ao segundo. Então, um novo histograma baseado no mesmo conjunto de dados e no intervalo de classe igual a 2,5 passaria a ter duas barras ao invés de uma, e o ponto mais alto do gráfico assumiria o valor 3.000. A Figura 3.3 ilustra esse efeito para os dados da Figura 3.2, com intervalos de classe de tamanhos 5, 2 e 1.

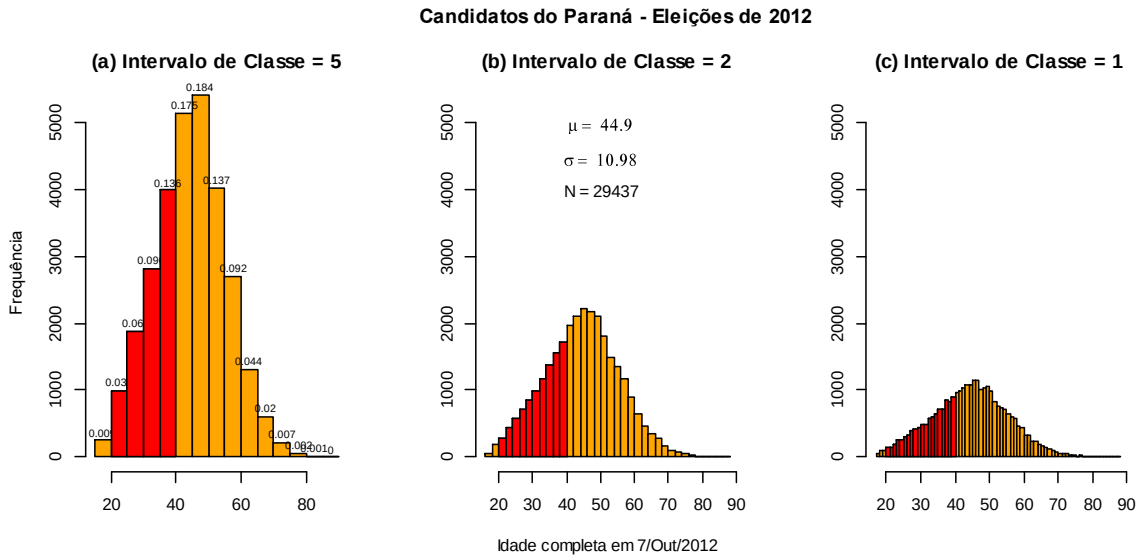


Figura 3.3 – Histogramas com intervalos de classe distintos para o mesmo conjunto de dados: (a) 5, (b) 2 e (c) 1.®

Observe na Figura 3.3 que à medida que o intervalo de classe se torna menor, o contorno do histograma se torna mais suave, amenizando o aspecto de escada ou degraus entre as barras.

O tamanho 1 é o menor limite possível a um intervalo de classe da variável discreta idade completa. Suponha que fossem admitidas idades **in**completas e os valores dos dados dos candidatos do Paraná pudessem ser representados por uma casa decimal. Então, um intervalo (20,0, 30,0], com tamanho menor do que 1, passaria a fazer sentido e o efeito sobre o histograma da Figura 3.3c seria a redistribuição dos dados existentes, entre cada idade completa, em dez novas barras (...dez entre 16 e 17, ...dez entre 20 e 21, dez entre 21 e 22..., dez entre 51 e 52...). Essas novas barras teriam alturas menores que as barras da Figura 3.3c, uma vez que a quantidade de dados permaneceria a mesma.

Agora, imagine um conjunto hipotético cuja quantidade de dados crescesse a cada encolhimento do intervalo de classe, de forma a preencher as novas barras respeitando a tendência (ou distribuição) das barras originais. Para esse conjunto hipotético, cada nova casa decimal acrescida para representar os dados faria o intervalo de classe reduzir em dez vezes o seu tamanho, contudo, novos dados apareceriam para preencher as barras em quantidade suficiente para respeitar a distribuição original. Assim, na medida em que

caminhássemos para infinitas casas decimais, o intervalo de classe tenderia a zero (a um ponto) e cada barra tenderia a uma linha reta vertical, sem espessura, cuja altura corresponderia à densidade de probabilidade naquele ponto. O contorno do histograma seria uma curva suave, contínua. A Figura 3.4 ilustra essa situação.

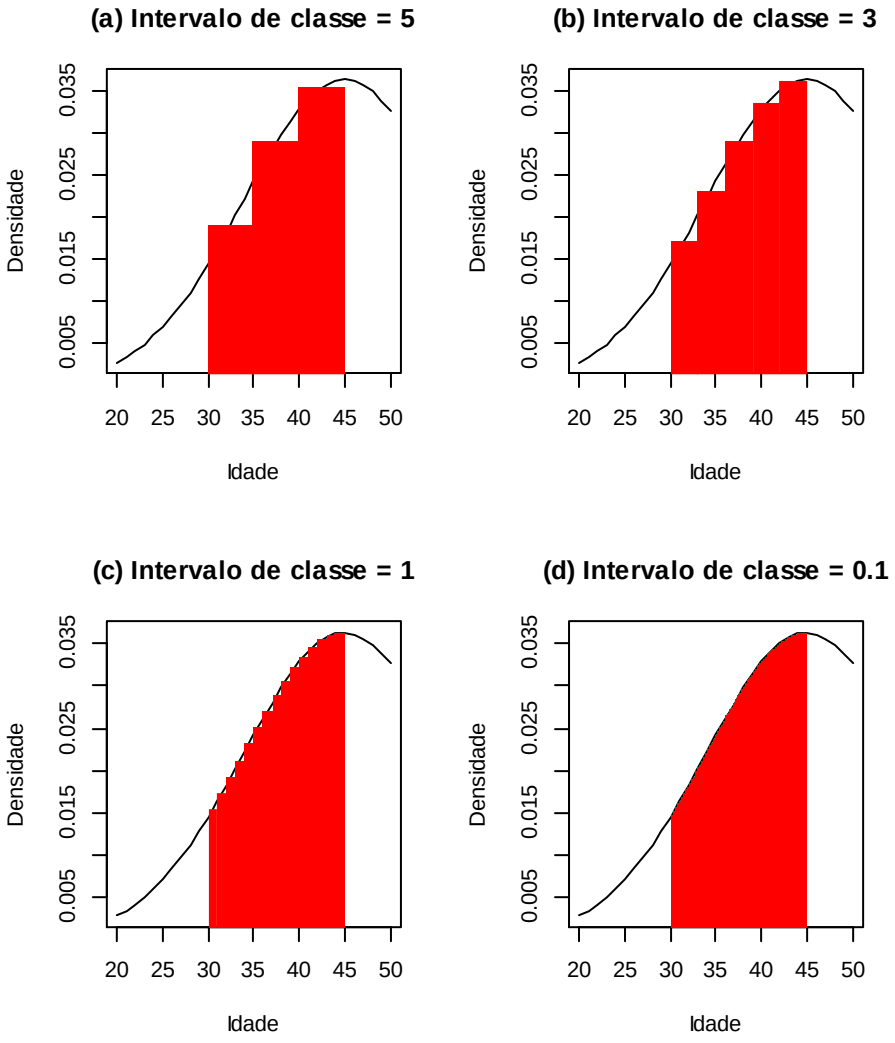


Figura 3.4 – Distribuição de probabilidades: aproximação do caso discreto (barras vermelhas) ao caso contínuo (linha preta) com intervalo de classe tendendo a zero e quantidade de dados tendendo a infinito.®

Esse conjunto hipotético, na verdade, é o conjunto dos números reais $\Re$ e corresponde ao caso contínuo. No caso contínuo, a distribuição de probabilidade é uma curva contínua, isto é, uma função cujo domínio (eixo *x*) são os valores da variável observada e cuja imagem (eixo *y*) são as densidades de probabilidades associadas a cada valor do domínio. Densidade de probabilidade é o limite da frequência relativa quando a quantidade de dados tende a infinito. Não se visualizam as barras de um histograma, mas sim um contorno

suave, ou a linha de probabilidade de ocorrência de cada valor do eixo *x*. Essa linha contínua forma a curva de densidade de probabilidade ou **função de densidade de probabilidade**[11] (PDF – *probability density function*). A PDF pode ser compreendida como uma extensão natural de um histograma, ou a forma limite do histograma, admitindo-se que o intervalo de classe tende a zero (ou tende a um ponto) e que a quantidade de dados tende a infinito. Nesse caso, a probabilidade do ponto é desprezível ($1/\infty = 0$).

Então, a PDF é a função que nos permite calcular a probabilidade para uma variável que pode assumir infinitos valores e cuja probabilidade no ponto é zero. A PDF, em si, não é probabilidade, mas é construída de tal forma que a sua integral defina a probabilidade em um intervalo contínuo. Em geral se constroem bibliotecas dessas funções que são utilizadas de acordo com o evento que se está estudando. A PDF que se ajusta à curva normal é descrita pela Equação (3.2).

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \tag{3.2}$$

Seguindo na analogia de que as barras do histograma (caso discreto) tendem a linhas verticais justapostas (caso contínuo) correspondentes às probabilidades da variável observada, é possível identificar, na Figura 3.4d, que a probabilidade de um candidato ter idade entre 30 e 45 anos é igual à soma das probabilidades de cada valor entre 30 e 45, ou seja, é igual à área sob a curva compreendida entre os valores 30 e 45.

Então, de forma genérica, para uma variável contínua *x*, a probabilidade de $a \leq x \leq b$ é igual à área sob a curva da PDF entre os pontos *a* e *b*. Do Cálculo Numérico (), a área sob uma curva entre dois pontos é igual à integral da curva entre os dois pontos. Na curva normal, as áreas são divididas pelo desvio padrão em torno da média.

[11] Regras da função de densidade de probabilidade f(x) de uma variável aleatória X: $f(x) \geq 0$; $\int_a^b f(x)dx = P(a < x \leq b), b > a$; $\int_{-\infty}^{+\infty} f(x)dx = 1$, onde *n* é o número de elementos do espaço amostral. Funções de densidade de probabilidade são utilizadas para variáveis aleatórias contínuas. Variáveis aleatórias contínuas podem assumir qualquer valor dentro de um intervalo contínuo e tem precisão infinita. As funções FDP são construídas de forma que a sua integral defina as probabilidades dentro de um intervalo contínuo.

Integrais matemáticas são, em geral, conhecidas e temidas por sua complexidade. Contudo, para distribuições de uso mais comum, e sob certas condições ditas ideais, o trabalho de se calcular as probabilidades já foi efetuado por estudiosos da estatística que as organizaram no formato de **tabelas de probabilidades**, encontradas nos principais livros que abordam o assunto (Field *et al*., 2012). A mais conhecida dessas tabelas (ver Tabela 3.2) refere-se à distribuição normal de média 0 e desvio padrão 1, ou **curva normal padronizada** (Figura 3.5). Para este caso ($\mu = 0$ e $\sigma = 1$), a Equação (3.2) se reduz para

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{3.3}$$

Tabela 3.2 – Probabilidades da Curva Normal Padronizada.®

z	*Área maior*	*Área menor*	*Y*
0,00	0,50000	0,50000	0,3989
0,01	0,50399	0,49601	0,3989
0,02	0,50798	0,49202	0,3989
0,03	0,51197	0,48803	0,3988
0,04	0,51595	0,48405	0,3986
0,05	0,51994	0,48006	0,3984
0,06	0,52392	0,47608	0,3982
0,07	0,52790	0,47210	0,3980
0,08	0,53188	0,46812	0,3977
...	...	...	...
1,38	0,91621	0,08379	0,1539
...	...	...	...
1.65	0,95053	0,04947	0.1023
...	...	...	...
1.96	0,97500	0,02500	0.0584

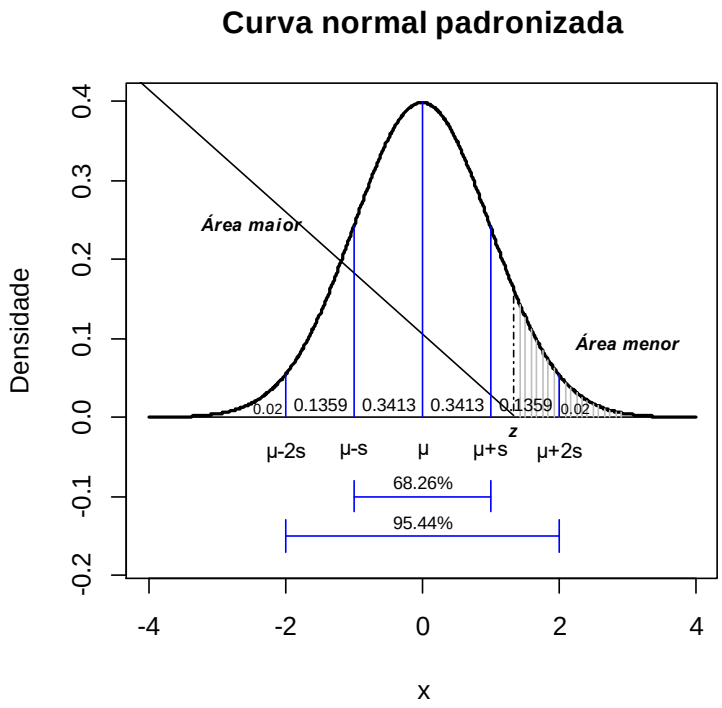


Figura 3.5 – Proporções da área da curva normal padronizada ($\mu = 0$ e $\sigma = 1$).®

A partir das distribuições padronizadas (ideais), é possível calcular a probabilidade de ocorrência dos dados de qualquer outro conjunto de dados cuja distribuição tenha a mesma forma. Logo, se um novo conjunto de dados for descrito por uma distribuição normal de média 0 e desvio padrão 1, basta recorrer à tabela de distribuição de probabilidade e verificar o quão provável é um determinado valor. Contudo, nem toda distribuição normal é descrita por média 0 e desvio padrão 1. Felizmente, qualquer conjunto de dados pode ser convertido para outro com essas características. Essa conversão gera um novo conjunto de valores padronizados denominados ***valores z*** (do inglês, *z-scores*). Inicialmente, é necessário centrar a média dos valores *z* em zero (centralização), subtraindo-se de cada valor original a média do conjunto original. Em seguida, divide-se o resultado pelo desvio padrão do conjunto original (escalonamento), garantindo, assim, um desvio padrão igual a 1 para o novo conjunto de valores *z* (Equação (3.4)).

$$z = \frac{x - \mu}{\sigma} \tag{3.4}$$

Note que o valor padronizado representa o número de desvios padrão pelo qual um valor *x* se distancia da média, para mais ou para menos.

Por exemplo, considere a Figura 3.3 cujos histogramas são referentes à distribuição da idade dos candidatos às eleições do Paraná em 2012. Por inspeção visual é razoável dizer que se trata de uma distribuição normal ($\mu = 44{,}90$ e $\sigma = 10{,}98$). Logo, podemos utilizar a tabela da curva normal padronizada para estimarmos as probabilidades das idades dos candidatos. Suponha que desejássemos saber qual a probabilidade de um candidato ter 60 anos ou mais (x = 60). Transformando para valores *z*, temos z = (60-44,9)/10,98 = 1,38. Localizando *z* na Tabela 3.2, encontramos na coluna "*Área menor*" o valor 0,08379, o que significa dizer que existe uma probabilidade de 8,38% de um candidato às eleições do Paraná ter 60 anos ou mais. Por outro lado, há uma probabilidade de 91,62% de o candidato ter 60 anos ou menos.

As tabelas de probabilidade foram construídas em uma época em que o uso de ferramentas computacionais era escasso ou difícil. Atualmente, com as facilidades proporcionadas pela tecnologia, torna-se fácil estimar as probabilidades por meio de métodos algorítmicos codificados em linguagens de programação.

Os quadros 3.1 e 3.2 contêm blocos de código em R que exemplificam como as probabilidades podem ser estimadas utilizando-se diretamente a Equação (3.2) da curva normal, ou por meio da curva normal padronizada (Equação (3.3)) associada ao valor *z* respectivo.

Quadro 3.1 – Código R que estima as probabilidades dos candidatos às eleições do Paraná, com $N \sim (44{,}9;10{,}98)$, terem 60 anos (x = 60) ou mais, e 60 anos ou menos: aplicação direta da PDF.

```
# área da curva normal para x = 60, N~(44,90;10,98)
x<-60
media<-44.90
dp<-10.98
y <- function(x) {(1/sqrt(2*pi*dp^2)) * exp(-((x - media)^2)/(2*dp^2))}
i<-integrate(y, media, x)
# área maior – 60 anos ou menos
round(0.5+as.numeric(i[1]),5)
# área menor – 60 anos ou mais
round(0.5-as.numeric(i[1]),5)
```

Quadro 3.2 – Código R que estima as probabilidades dos candidatos às eleições do Paraná, com $N \sim (44{,}9;10{,}98)$, terem 60 anos ($x = 60$) ou mais, e 60 anos ou menos: aplicação curva normal padronizada.

```
# curva normal para x = 60, N~(44,90;10,98)
# área da curva normal padronizada
x<-60
media<-44.90
dp<-10.98
z=(x-media)/dp
round(z,2)
i<-integrate(dnorm, 0, z)
# área maior – 60 anos ou menos
round(0.5+as.numeric(i[1]),5)
# área menor – 60 anos ou mais
round(0.5-as.numeric(i[1]),5)
```

FDP e FP descrevem o comportamento de uma variável aleatória a partir de um modelo matemático.

A **função de distribuição acumulada** (FDA), representada por *F*, de uma variável aleatória X é definida por

$$F(a) = P(X \leq a), a \in \Re \tag{3.5}$$

e representa a probabilidade acumulada de um evento até um ponto determinado. No caso discreto, a FDA é dada por

$$F(a) = \sum_{x_i \leq a} P(x_i) \tag{3.6}$$

e, no caso contínuo, por

$$F(a) = \int_{-\infty}^{a} f(x)dx \tag{3.7}$$

As propriedades da FDA são

a. $F(x) =\geq 0$

b. $F(-\infty) = 0$

c. $F(+\infty) = 1$

d. $F(x)$ é sempre ascendente

e. $F(b) - F(a) = P(a < X \leq b), b > a$

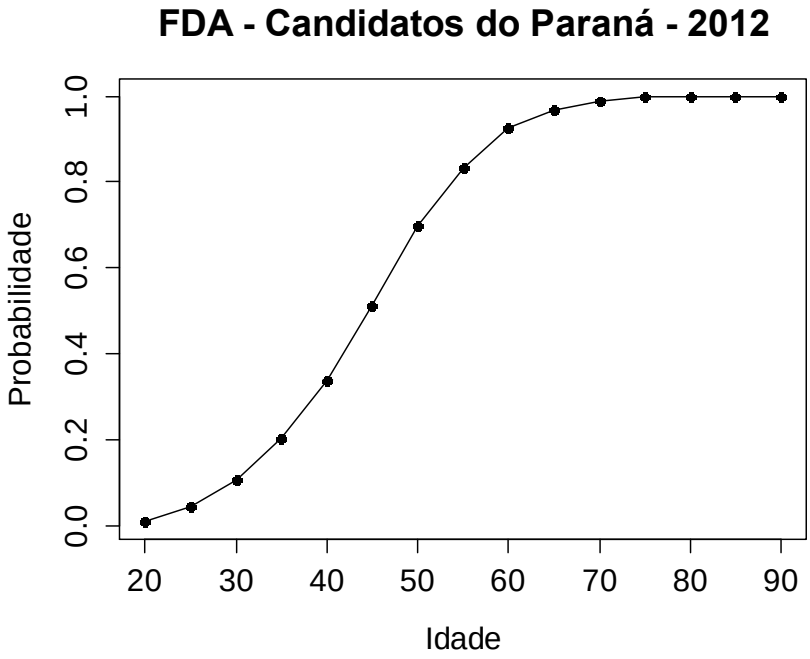


Figura 3.6 – FDA, idade dos candidatos do Paraná – Eleições 2012.®

3.2 LABORATÓRIO 2

Execute os scripts R a seguir e verifique os respectivos comportamentos, alterando as variáveis **media**, **dp**, **x1** e **x2**.

3.2.1 Curva normal padronizada

R

```
#####################################################
# Lab 3.2.1
#####################################################
# remove todas as variáveis da memória
rm(list = ls(all = TRUE))

#####################################
# INFORME ANTES DA EXECUÇÃO DO SCRIPT

# 1. os parâmetros da população
media <- 60
dp <- 10

# 2. o intervalo da análise
#    Obs: quando não for utilizado o intervalo, faça x2<-NULL
x1 <- 50
```

```
x2 <- NULL

########################
# Script principal
########################
if(!is.null(x1)) {
    z1<-(x1-media)/dp
    if(!is.null(x2) ) {
        z2<-(x2-media)/dp
        if(z1 > z2) {aux<-z1; z1<-z2; z2<-aux; aux<-x1; x1<-x2; x2<-aux}
    }
    else z2 <- NULL
}

# eixo x da curva normal padronizada
lim <- c(-4,4)
x <- seq(lim[1], lim[2], by = 0.01)

# traça a curva normal padronizada
plot(x,dnorm(x,0,1),ylab="Densidade",xlab="x",
     main="Curva normal padronizada",type="l",lwd=2,ylim=c(-0.05,0.4))
# linha horzontal em zero
lines(c(lim[1],lim[2]),c(0,0))

# curva normal padronizada
cnp <- function(x) {dnorm(x,0,1)} # curva normal padronizada

# valores de z
if(!is.null(z1)) {
    text(z1,-0.02,paste("z1=",round(z1,2)),cex=0.7,font=4)
    lines(c(z1,z1),c(0,cnp(z1)),lty=4,type="l")
}
if(!is.null(z2)) {
    text(z2,-0.02,paste("z2=",round(z2,2)),cex=0.7,font=4)
    lines(c(z2,z2),c(0,cnp(z2)),lty=4,type="l")
}

# probabilidades
integral <- function(f,a,b) {i<-integrate(f,a,b); as.numeric(i[1])}

# hachura da área
if(!is.null(z1)) {
    if(is.null(z2)) {
        z2 <- lim[2]
    }
    else {
        text(-4.3,0.29,paste("x2 =",x2),cex=0.8,pos=4)
    }
    inc <- (z2-z1)/20
    i<-z1+inc
    while(i < z2){
        lines(c(i,i),c(0,cnp(i)),col="red",lwd=0.5)
        i<-i+inc
    }
    phachura<-round(integral(cnp,z1,z2),4)
    text(-4.3,0.38,paste("Área hachurada =",phachura),cex=0.8,pos=4)
    text(-4.3,0.35,paste("Área branca =",1-phachura),cex=0.8,pos=4)
    text(-4.3,0.32,paste("x1 =",x1),cex=0.8,pos=4)
}
```

3.2.2 Curva normal

R

```
#####################################################
# Lab 3.2.2
#####################################################
# remove todas as variáveis da memória
rm(list = ls(all = TRUE))

#####################################
# INFORME ANTES DA EXECUÇÃO DO SCRIPT

# 1. os parâmetros da população
media <- 60
dp <- 10

# 2. o intervalo da análise
#    Obs: quando não for utilizado o intervalo, faça x2<-NULL
x1 <- 50
x2 <- NULL

########################
# Script principal
########################

# eixo x da curva normal padronizada
lim <- media+c(-4,4)*dp
x <- seq(lim[1], lim[2], by = 0.01)

# curva normal
cnp <- function(x) {dnorm(x,media,dp)} # curva normal

# traça a curva normal
plot(x,cnp(x),ylab="Densidade",xlab="x",
     main="Curva Normal",type="l",lwd=2)
# linha horzontal em zero
lines(lim,c(0,0))

# valores de x
if(!is.null(x1)) {
    text(x1,-0.02,paste("x1=",round(x1,2)),cex=0.7,font=4)
    lines(c(x1,x1),c(0,cnp(x1)),lty=4,type="l")
}
if(!is.null(x2)) {
    text(x2,-0.02,paste("x2=",round(x2,2)),cex=0.7,font=4)
    lines(c(x2,x2),c(0,cnp(x2)),lty=4,type="l")
}

# probabilidades
integral <- function(f,a,b) {i<-integrate(f,a,b); as.numeric(i[1])}

# hachura da área depois de x1
if(!is.null(x1)) {
    if(is.null(x2)) {
        x2 <- lim[2]
    }
    else {
        text(x2,-dnorm(media,media,dp)/70,paste("x2=",x2),cex=0.7)
        text(lim[1],cnp(media)*.85,paste("x2 =",x2),cex=0.8,pos=4)
```

```
    }
    inc <- (x2-x1)/20
    i<-x1+inc
    while(i < x2){
        lines(c(i,i),c(0,cnp(i)),col="gray",lwd=0.5)
        i<-i+inc
    }
    phachura<-round(integral(cnp,x1,x2),4)
    text(lim[1],cnp(media),paste("Área hachurada =",phachura), cex=0.8,
pos=4)
    text(lim[1],cnp(media)*.95,paste("Área branca =",1-phachura),
cex=0.8, pos=4)
    text(lim[1],cnp(media)*.90,paste("x1 =",x1),cex=0.8,pos=4)
    text(x1,-dnorm(media,media,dp)/70,paste("x1=",x1),cex=0.7)
}
```

3.2.3 Curva normal – variação do achatamento

Utilize outros valores para o desvio padrão (5, 15, 3 ...) e verifique a variação do achatamento.

R

```
################################################
# Lab 3.2.3
################################################
# remove todas as variáveis da memória
rm(list = ls(all = TRUE))

#####################################
# INFORME ANTES DA EXECUÇÃO DO SCRIPT

# 1. os parâmetros da população
media <- 60
dp <- 10 # use outros valores de dp (5,15, 3) e verifique o achatamento

# 2. o intervalo da análise
#    Obs: quando não for utilizado o intervalo, faça x2<-NULL
x1 <- 50
x2 <- NULL

########################
# Script principal
########################

# eixo x da curva normal padronizada
lim <- media+c(-4,4)*dp
x <- seq(lim[1], lim[2], by = 0.01)

# curva normal
cnp <- function(x) {dnorm(x,media,dp)} # curva normal

# traça a curva normal
plot(x,cnp(x),ylab="Densidade",xlab="x",
     main="Curva Normal",type="l",lwd=2,ylim=c(-0.05,0.4))
# linha horzontal em zero
lines(lim,c(0,0))
```

```
# valores de x
if(!is.null(x1)) {
    text(x1,-0.02,paste("x1=",round(x1,2)),cex=0.7,font=4)
    lines(c(x1,x1),c(0,cnp(x1)),lty=4,type="l")
}
if(!is.null(x2)) {
    text(x2,-0.02,paste("x2=",round(x2,2)),cex=0.7,font=4)
    lines(c(x2,x2),c(0,cnp(x2)),lty=4,type="l")
}

# probabilidades
integral <- function(f,a,b) {i<-integrate(f,a,b); as.numeric(i[1])}

# hachura da área
if(!is.null(x1)) {
    if(is.null(x2)) {
        x2 <- lim[2]
    }
    else {
        text(lim[1],0.29,paste("x2 =",x2),cex=0.8,pos=4)
    }
    inc <- (x2-x1)/20
    i<-x1+inc
    while(i < x2){
        lines(c(i,i),c(0,cnp(i)),col="red",lwd=0.5)
        i<-i+inc
    }
    phachura<-round(integral(cnp,x1,x2),4)
    text(lim[1],0.38,paste("Área hachurada =",phachura),cex=0.8,pos=4)
    text(lim[1],0.35,paste("Área branca =",1-phachura),cex=0.8,pos=4)
    text(lim[1],0.32,paste("x1 =",x1),cex=0.8,pos=4)
}
```

4 INFERÊNCIA ESTATÍSTICA

Neste módulo serão apresentados os ...

4.1 INFERÊNCIA ESTATÍSTICA

O conceito de **inferência estatística** consiste em descrever uma realidade que está presente na população a partir de uma amostra extraída dessa população.

A população e a amostra são diferentes em suas características. Uma informação populacional é denominada **parâmetro**. Uma informação da amostra é denominada **estimativa**. A Tabela 4.1 relaciona algumas dessas diferenças.

Tabela 4.1 – Características da população e da amostra.

Característica	**População**	**Amostra**
Informação	parâmetro	estimativa
Número de elementos	N	n
Média	μ	$\overline{x}$
Variância	σ^2	s^2
Símbolos	letras gregas	letras latinas
Valor	verdadeiro	Aproximado
Erro	livre de	propenso a

É importante ressaltar a diferença entre essas características para se compreender, por exemplo, que a média estimada em uma amostra (**média amostral**) não é necessariamente igual à média da população (**média populacional**) ou à média de outra amostra. Contudo, uma boa estimativa da média populacional pode ser conseguida a partir da obtenção de várias médias amostrais. As próximas seções exploram esse fenômeno.

4.2 ASSINTOTISMO

Assintotismo é a propriedade de duas linhas assintóticas, que se tangenciam no infinito. É também o termo utilizado para caracterizar o comportamento de estimativas na medida em que o tamanho da amostra (ou outra quantidade relevante) tende ao infinito.

Assintotismos são extremamente úteis para inferências estatísticas simples e aproximações, contudo, não têm eficácia garantida para amostras de tamanho reduzido. Assintotismos formam a base para a interpretação de frequências e probabilidades nos casos em que um evento se repete grande quantidade de vezes.

Para médias amostrais, há importantes constatações. Imagine uma variável aleatória que registre os resultados de um dado experimento que se repete inúmeras vezes. Se desse conjunto de resultados extrairmos diversas amostras e determinarmos a média de cada uma dessas amostras, teremos um novo conjunto de observações, ou seja, o conjunto das médias amostrais. A primeira importante constatação afirma que a média do conjunto de médias amostrais, no limite, tende ao valor que se está tentando estimar: a média da população.

Esse fenômeno é denominado **lei dos grandes números**. Por exemplo, considere a variável aleatória X que registra a média dos resultados de n lançamentos de uma moeda. Na medida em que a moeda é lançada repetidas vezes, X converge para a probabilidade real de caras e coroas (ver Figura 4.1).

Um estimador é consistente se ele converge para aquilo que se está tentando estimar. A lei dos grandes números diz que a média amostral de uma amostra independente e identicamente distribuída é consistente com a média da população. A variância e o desvio padrão da amostra também são consistentes.

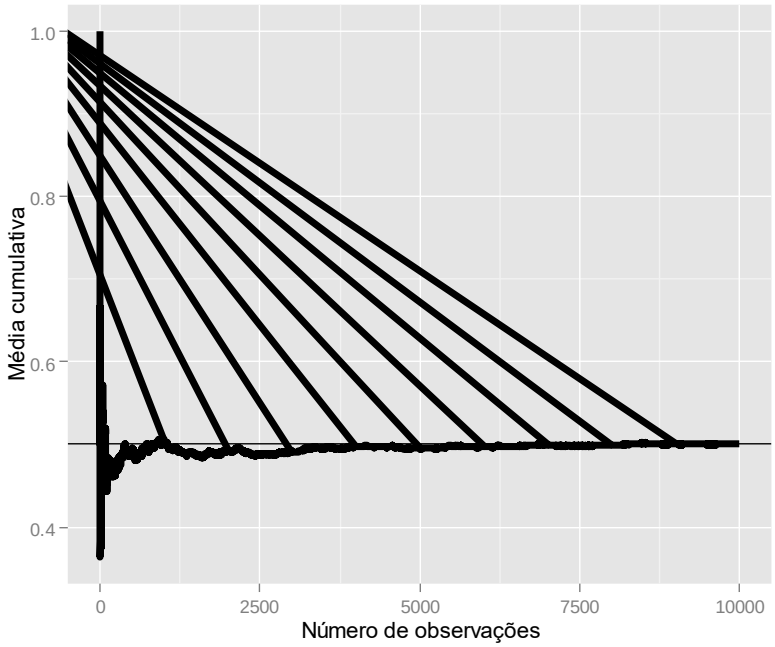


Figura 4.1 – Lei dos Grandes Números em ação.®

O conjunto das médias amostrais forma uma série numérica de médias cujos valores têm sua própria distribuição de frequências, denominada **distribuição amostral das médias**. Por consequência, podem ser determinados a média (das médias amostrais) e o respectivo desvio padrão da distribuição amostral (**erro padrão**).

O erro padrão é uma medida da precisão (e não da dispersão) da média amostral calculada. Isso quer dizer que,

> se de uma população com média μ e desvio padrão σ forem retiradas várias (30 ou mais) amostras com o mesmo tamanho *n*, e para cada amostra se calcular a respectiva média, a distribuição do conjunto dessas médias é normal com média $\mu_{\bar{x}} = \mu$ e desvio padrão

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{4.1}$$

Esse fenômeno é denominado **teorema do limite central** (figuras 4.1 e 4.2). Assim, o erro padrão é o desvio padrão da distribuição das médias ($\sigma_{\bar{x}}$) das amostras de uma população.

Quando não se conhece o desvio padrão da população, usa-se o desvio padrão da amostra para estimá-lo (para n>30) (Equação (4.2)).

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}} \tag{4.2}$$

Quando n é pequeno, mas a variável tem distribuição normal na população, a distribuição amostral das médias aproxima-se de uma curva normal. Quando n é grande, a distribuição amostral das médias será próxima da normal, independentemente da forma da distribuição populacional.

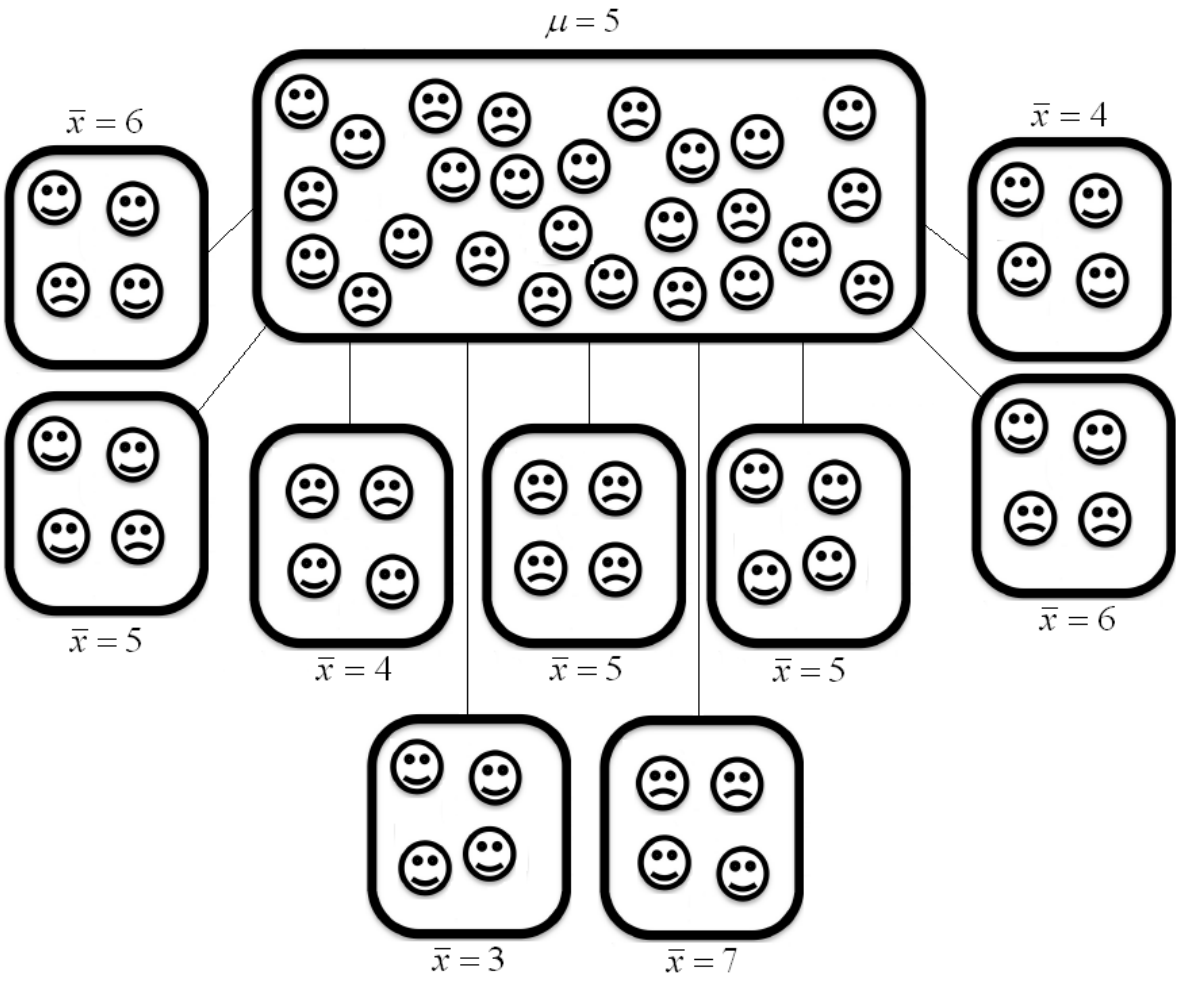


Figura 4.2 – Médias populacional e amostral.

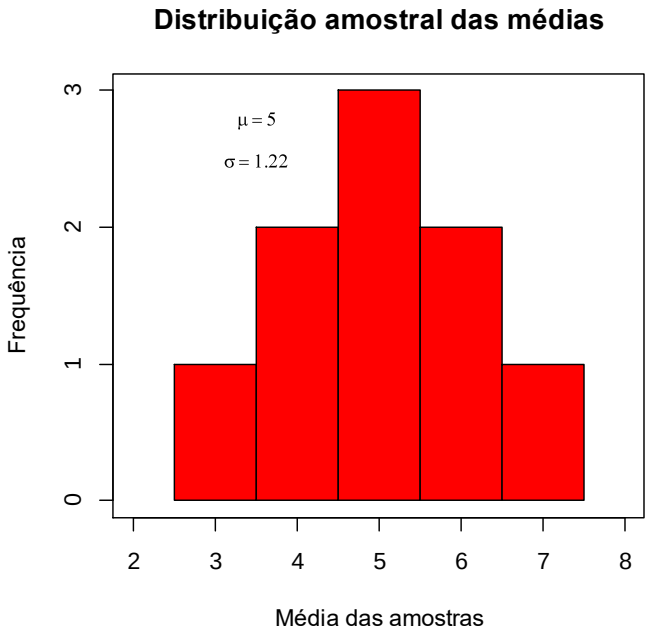


Figura 4.3 – Distribuição amostral das médias.®

A caixa de Galton é um dispositivo inventado por Sir Francis Galton (Galton, 1894) para demonstrar o teorema do limite central, em particular, quando a distribuição normal é aproximada a partir da distribuição binomial.

A máquina é constituída por uma placa vertical com fileiras intercaladas de pinos. Bolas são jogadas a partir do topo e saltam para a esquerda ou para a direita (binômio), sempre que atingem um pino, sendo coletadas em caixas na parte inferior. A altura das colunas de bolas nas caixas se aproxima de uma curva de sino (Figura 4.4).

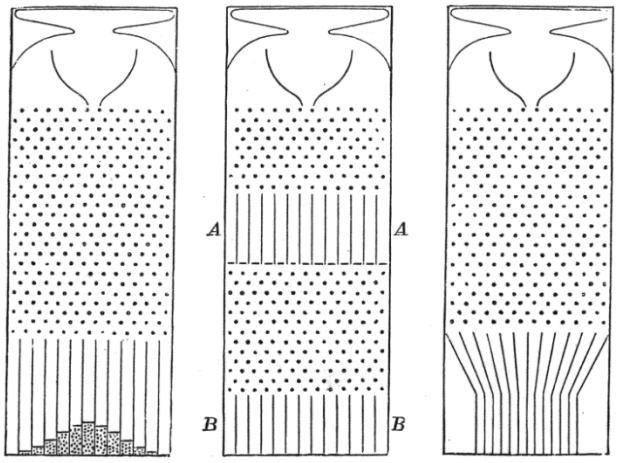


Figura 4.4 – Caixa de Galton.

4.3 INTERVALOS DE CONFIANÇA

O erro padrão é um indicador do quanto as médias amostrais são diferentes. Uma abordagem diferente para avaliar a precisão da média amostral como uma estimativa da média da população é calcular os limites dentro dos quais se supõe que o verdadeiro valor da média estará. Esses limites são chamados de **intervalo de confiança** (IC).

Em estatística, um IC é um intervalo estimado de um parâmetro estatístico. Em vez de estimar o parâmetro por um único valor, é dado um intervalo de estimativas prováveis. Quanto maior a probabilidade de o intervalo conter o parâmetro, maior será o intervalo.

Os ICs devem ser definidos de maneira que possam informar a probabilidade de conterem o verdadeiro valor do objeto estimado (por exemplo, a média). A essa probabilidade se dá o nome de **nível de confiança**. Tipicamente são utilizados níveis de confiança de 95% e 99% (probabilidades 0,95 e 0,99). Quando um IC é determinado com base em um nível de confiança de 95%, isso significa dizer que se forem extraídas 100 amostras de uma população, 95 dessas amostras conterão o parâmetro estimado dentro do IC calculado.

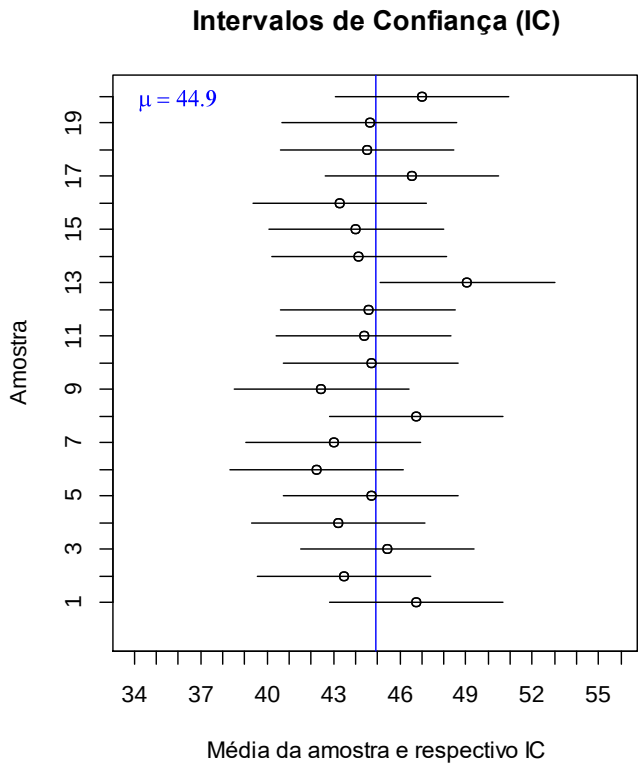


Figura 4.5 – Intervalos de confiança para 20 amostras com tamanho 30 e média populacional igual a 44,9, considerando-se nível de confiança de 95%.®

A Figura 4.5 ilustra ICs com nível de confiança de 95% para 20 amostras de tamanho (*n*) igual a 30. Nesse caso, espera-se que ao menos 19 (0,95 x 20 = 19) dos ICs definidos contenham o valor da média populacional. É o que se observa na Figura 4.5 onde apenas o IC da 13ª amostra não contém a média populacional.

Uma vez definido o nível de confiança com o qual se deseja trabalhar, é preciso se conhecer os valores que delimitam o IC respectivo. Considere, por exemplo, o nível de confiança de 95%. Para um conjunto de dados cuja distribuição tenha $\mu = 0 \text{ e } \sigma = 1$, basta recorrer à curva normal padronizada para se encontrar os limites do IC. Da Figura 3.5, pode-se observar que 95,44% dos valores ao redor da média estão entre -2 e 2. Mais exatamente, da Tabela 3.2, é possível constatar que exatos 95% dos valores que circundam a média estão entre -1.96 e 1.96, o que representaria o IC desejado. Contudo, dificilmente o conjunto de dados estudado terá $\mu = 0 \text{ e } \sigma = 1$. Nesse caso, será preciso converter os valores da curva normal padronizada (valores *z*) para os parâmetros do conjunto em estudo. Da Equação (3.4), tem-se que o IC com nível de confiança de 95% é limitado por

$$-1.96 = \frac{x - \mu}{\sigma} \qquad e \qquad 1.96 = \frac{x - \mu}{\sigma} \tag{4.3}$$

$$x_1 = \mu - 1.96\sigma \qquad e \qquad x_2 = \mu + 1.96\sigma \tag{4.4}$$

A Equação (4.4) pressupõe que os parâmetros populacionais (μ e σ) são conhecidos. Contudo, na prática e para amostras onde *n*>30, utilizam-se a média amostral e o erro padrão (este último estimado com o desvio padrão amostral), visto que, com base no teorema do limite central, o interesse se volta para a variabilidade das médias das amostras (Equação 4.5) e não para a variabilidade das médias das observações na amostra.

$$x_1 = \bar{x} - 1.96\frac{s}{\sqrt{n}} \qquad e \qquad x_2 = \bar{x} + 1.96\frac{s}{\sqrt{n}} \tag{4.5}$$

Observe que a Equação (4.5) é válida apenas para o nível de confiança de 95%. De forma genérica, a Equação (4.5) pode ser escrita conforme a Equação (4.6),

$$x_1 = \bar{x} - z_{\left(\frac{1-p}{2}\right)} \frac{s}{\sqrt{n}} \qquad e \qquad x_2 = \bar{x} + z_{\left(\frac{1-p}{2}\right)} \frac{s}{\sqrt{n}} \tag{4.6}$$

onde *p* é o nível de confiança do IC e $z_{\left(\frac{1-p}{2}\right)}$ é o valor de *z* cuja área menor da curva normal padronizada é igual a $\left(\frac{1-p}{2}\right)$.

Por exemplo, para o nível de confiança de 90% (*p*=0,90) deve ser encontrado o valor de *z* cuja área menor seja $\left(\frac{1-0{,}90}{2}\right) = 0{,}05$. Da Tabela 3.2, verifica-se que $z_{0,05} = 1{,}65$ e os limites do IC são encontrados por substituição direta desse valor na Equação (4.6).

Em sentido estrito, o IC para um parâmetro populacional é um intervalo com uma probabilidade *p* associada, de tal forma que, para várias amostras obtidas, uma proporção igual a $p \times 100\%$ dos intervalos de confiança calculados conterá o parâmetro estatístico em questão.

Intervalos de confiança são a forma predominante de estimativa por intervalo e são usados para indicar a confiabilidade de uma estimativa. Por exemplo, um IC pode ser usado para descrever quão confiáveis são os resultados de uma pesquisa. Sendo o nível de confiança e todas as demais condições iguais, uma pesquisa que resulte num IC pequeno é mais precisa do que outra que resulte num IC maior. Observe, na Figura 4.6, que na medida em que o tamanho da amostra aumenta (ou seja, na medida em que a pesquisa se torna mais precisa), o tamanho do IC diminui.

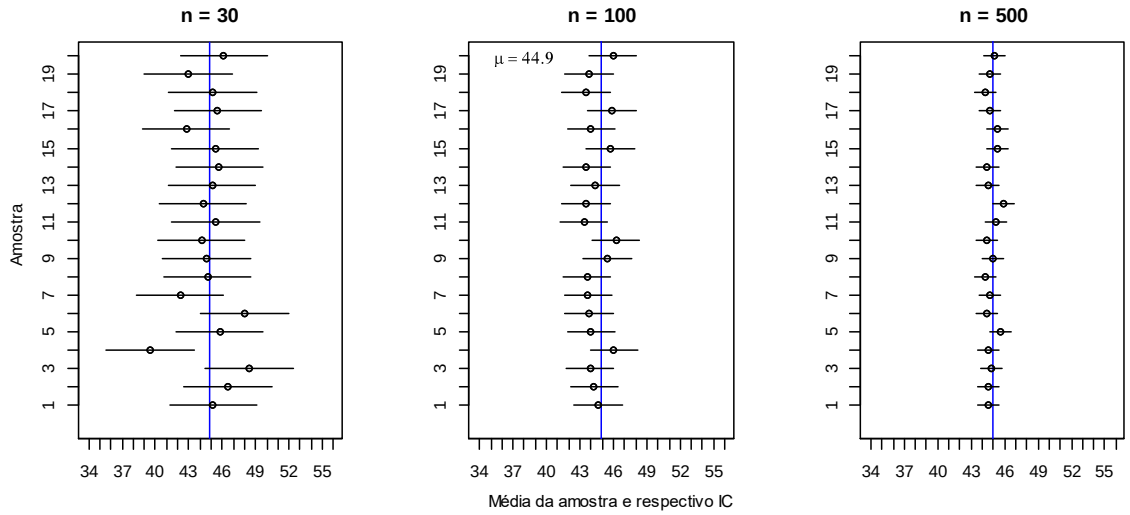


Figura 4.6 – Intervalos de confiança com nível de confiança de 95% para diferentes tamanhos da amostra.®

A comparação de ICs de diferentes amostras pode indicar se as respectivas médias amostrais vêm da mesma população ou de populações diferentes. Sabe-se que as médias de duas ou mais amostras da mesma população podem ser ligeiramente diferentes e que o erro padrão sinaliza o quão diferentes essas médias podem ser. Sabe-se, também, que o IC delimita o intervalo no qual existe grande probabilidade de se encontrar a média populacional.

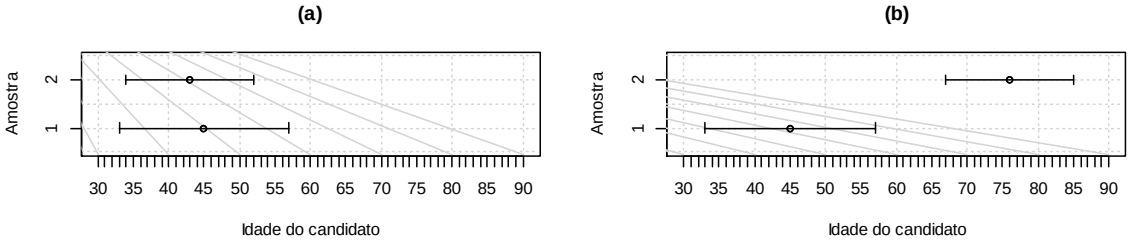


Figura 4.7 – Comparação de ICs a 95% de confiança: a) amostras da mesma população; b) amostras de populações distintas.®

Então, da primeira amostra da Figura 4.7a, pode-se afirmar, com 95% de probabilidade de acerto, que o IC delimitado por 33 e 57 contém a média populacional[12]. A segunda amostra da Figura 4.7a tem a mesma probabilidade (95%) de conter a medida populacional, e seus limites estão entre 34 e 52 anos. Observa-se que há completa superposição dos ICs, além da mesma probabilidade de conterem a média populacional, o que significa que é bastante

[12] Graficamente, a linha que delimita os extremos de um intervalo de confiança é chamada de **barra de erro**.

razoável supor que ambas as amostras (e respectivas médias) são oriundas da mesma população.

No caso da Figura 4.7b, os ICs da primeira e da segunda amostra também têm igual probabilidade de conter a média populacional, contudo, não há sobreposição alguma entre eles. Isso sugere duas possíveis análises: 1) ambos os ICs contêm a média populacional, mas são oriundos de populações diferentes; 2) ambas as amostras são oriundas da mesma população, mas um dos ICs não contém a média populacional. Como os ICs têm nível de confiança de 95%, a segunda análise é pouco provável (apenas 5% de probabilidade), o que remete à conclusão de que a primeira é a mais plausível.

Então, quando os ICs de duas médias amostrais não se sobrepõem, pode-se inferir que essas médias são oriundas de populações diferentes, ou seja, as médias são significativamente diferentes.

4.4 LABORATÓRIO 3

4.4.1 Teorema do Limite Central

Escreva um script R que simule o teorema do limite central. Utilize os dados das idades completas dos candidatos à eleição do Paraná em 2012. Determine as médias de 100 amostras com n = 30. Compare os valores estimados da média e do erro padrão com os valores teóricos. Verifique se a distribuição assemelha-se com a normal. Qual a real cobertura do IC?

R

```
#####################################################
# Simulação
#####################################################
# remove todas as variáveis da memória
rm(list = ls(all = TRUE))
# leitura do arquivo
tse <- read.table("consulta_cand_2012_PR.txt",sep=";") # com fatores
# renomeia a variável V26
names(tse)[26] <- "DATA_NASCIMENTO"
# converte o campo DATA_NASCIMENTO para o formato data
dtnasc <- as.Date(tse$DATA_NASCIMENTO, format="%d/%m/%Y")
# determina a idade dos candidatos na data da eleição de 2012 (7/10/2012)
dteleicao <- as.Date("7/10/2012", format="%d/%m/%Y")
```

```
idade2012 <- difftime(dteleicao,dtnasc,units="days")
# converte o resultado para a idade em anos
idade2012 <- floor(as.numeric(idade2012)/365.25)
idade2012 <- idade2012[idade2012 < 150]

length(idade2012)

media <- mean(idade2012)
dp    <- sd(idade2012)

n <- 30 # tamanho da amostra
SE <- dp/sqrt(n)
nsim <- 100
X  <- NULL # vector of sample averages
S  <- NULL # vector containing the standard deviation of each sample
sn <- NULL # vector of normalized averages
for(i in 1:nsim){
    idade_i <- sample(idade2012, n, replace = TRUE)
    xi    <- mean(idade_i)
    X     <- c(X,xi)
    S     <- c(S,sd(idade_i))
    sn    <- c(sn, (xi-media)/SE)
}

mean(X)
media

sd(X)
SE

library(ggplot2)
dsn <- density(sn)
k <- length(dsn[[1]])
df <- data.frame(y = c(dnorm(dsn[[1]]), dsn[[2]]),
                 x = dsn[[1]],
                 dist   =   factor(rep(c("Standard   Normal",   "Normalized
Averages"), c(k,k))))
g <- ggplot(df, aes(x = x, y = y))
g <- g + geom_line(size = 1.5, aes(colour = dist), fill=T)
g <- g + ylab("density")
g

# cobertura do IC relativa ao tamanho da amostra
coverage <- NULL
for(i in 1:nsim){
    ll <- X[i] - 1.96 * S[i]/sqrt(n)
    ul <- X[i] + 1.96 * S[i]/sqrt(n)
    coverage <- c(coverage,ll < media & ul > media)
}
mean(coverage)
```

5 TESTES ESTATÍSTICOS

Neste módulo ...

5.1 MODELOS ESTATÍSTICOS

Um modelo é uma representação simplificada da realidade. Por exemplo, os mapas disponíveis em um sistema de GPS (*Global Positioning System*) são uma representação mais simples das ruas e avenidas de uma determinada cidade e servem para orientar o caminho e ajudar o motorista a chegar ao seu destino.

Um **modelo estatístico** funciona de maneira similar e representa um fenômeno ou evento de interesse. Em geral, modelos estatísticos descrevem como variáveis aleatórias se relacionam.

Como já mencionado, a média é um modelo estatístico, um dos mais simples. Isso porque a média é um valor hipotético que não precisar ser, necessariamente, um dos valores observados no conjunto de dados (Field et al., 2012). Considere, por exemplo, que a quantidade de votos apurada em cinco urnas corresponda ao conjunto (2, 3, 5, 5, 8) para um determinado candidato. A média é (2+3+5+5+8)/5 = 4,6. Contudo, não é possível dizer que o candidato recebeu 4,6 votos. Logo, a média é um valor hipotético que resume o conjunto de dados.

Um modelo linear é baseado em uma linha reta, ou seja, os dados são resumidos por uma linha reta. A questão central consiste em encontrar a linha reta que melhor representa o conjunto de dados.

Modelos não lineares, ou curvilíneos, podem ser, muitas vezes, a forma mais adequada de se representar um conjunto de dados. Contudo, a formulação e a interpretação desses modelos são consideravelmente mais complexas.

De forma geral, modelos estatísticos resumem-se à Equação (5.1),

$$dado\_observado_i = (modelo) + erro_i \quad (5.1)$$

que significa dizer que o dado observado ($dado\_observado_i$) pode ser predito pelo modelo estatístico escolhido para representá-lo mais uma certa quantidade de erro ($erro_i$).

Em outras palavras, o modelo estatístico será tão melhor quanto menor for o erro. O método de se aferir o quanto um modelo proposto se ajusta aos dados consiste em determinar o **desvio** dos dados observados em relação ao modelo, ou seja, a soma dos erros quadrados.

$$desvio = \sum(erro_i)^2 = \sum(dado\_observado_i - modelo)^2 \quad (5.2)$$

Quando o modelo é a média, a Equação (5.2) coincide com a Equação (2.5) da *SEQ*.

Modelos estatísticos concentram-se, em geral, na análise da amostra, visto que dificilmente são conhecidos todos os elementos de uma população (censo). Saber o quão bem o modelo representa os dados observados (amostra) é um importante passo para se dizer ou inferir algo sobre a população.

5.2 TESTE DE HIPÓTESE

O **teste de hipótese** é o método utilizado para verificar se determinada afirmação sobre um parâmetro populacional é, ou não, apoiada pela evidência obtida a partir dos dados amostrais. O raciocínio consiste em analisar uma amostra e fazer a distinção entre os resultados que podem ocorrer facilmente (muito prováveis) e os que dificilmente ocorrem (pouco prováveis). Por razoabilidade, considera-se, então, que os resultados pouco prováveis **não** são obra do acaso, ou da aleatoriedade, e devem ser decorrentes de algum fenômeno incógnito. A conjectura sobre a existência desse fenômeno é que dá origem à pergunta de pesquisa e à hipótese a ser testada.

De forma geral, o teste de hipótese consiste: 1) na formulação de uma hipótese nula (que nega a existência do fenômeno) e de uma hipótese experimental ou alternativa (que

pressupõe a existência do fenômeno); 2) no ajustamento dos dados experimentais coletados a um modelo estatístico; 3) na avaliação do modelo por meio de um teste estatístico.

5.2.1 Formulação das hipóteses nula (H_0) e alternativa (H_1)

O teste de hipótese é um procedimento estatístico que conduz a uma decisão acerca das hipóteses nula (H_0) e alternativa (H_1). A hipótese nula é assumida como verdadeira e evidência estatística é necessária para rejeitá-la em favor da hipótese alternativa. Logo, compreendê-las e formulá-las adequadamente é o primeiro e importante passo.

Suponha que um cientista político deseje investigar se os programas sociais praticados pelo atual governo são capazes de conquistar votos para o candidato da situação a presidente na próxima eleição. Com recursos escassos para a pesquisa, o cientista decide delimitar a população do estudo aos eleitores residentes nas 80 cidades do país com mais de 200 mil eleitores. Ao recorrer à base de dados do TSE, o cientista verifica que o percentual de votos válidos recebido pelo atual presidente nessas 80 cidades obedece a uma distribuição N~(58.29,17.13). Com base nessa informação, o pesquisador delineia a sua pesquisa com o devido rigor científico, percorre algumas dessas cidades e entrevista diversos eleitores que são beneficiados por programas sociais. Ao final do trabalho, dispõe de uma amostra aleatória de dados com distribuição N~(61,18) e com 258 observações (considere, apenas para fins deste exercício, que n=258 atende aos requisitos científicos da pesquisa), a partir da qual constata que, em média, 61% dos entrevistados afirmaram ter intenção de voto no candidato da situação para não perder o benéfico do programa social.

O fenômeno a ser investigado é "o fato de o eleitor ser beneficiado por um programa social". A hipótese experimental, que pressupõe a existência do fenômeno e os seus consequentes efeitos, é a de que "o percentual de votos válidos recebidos pelo candidato do governo é maior nos grupos de eleitores beneficiados por um programa social". A hipótese nula nega a existência (ou a influência) do fenômeno e diz que "não há diferença no percentual de votos válidos recebidos pelo candidato do governo nos grupos de eleitores beneficiados por um programa social".

Em outras palavras, a hipótese nula é uma afirmação sobre o parâmetro populacional tal como este é especificado, é o *status quo*, ou seja, a circunstância testada. Considerando os parâmetros populacionais do exemplo em questão, $\mu_0 = 58.29\%$ e H_0: $\mu = \mu_0$, ou seja, H_0: $\mu = 58.29\%$.

A hipótese alternativa é uma afirmação que contraria a alegação e representa o que se deseja provar ou estabelecer, sendo formulada para contradizer a hipótese nula. Isso significa que o parâmetro (μ) pode ser maior ou menor que o valor especificado (μ_0), *status quo*, ou simplesmente diferente deste. Para o exemplo em questão, H_1: $\mu > 58.29\%$.

De forma geral, o teste de hipótese confronta as estimativas da amostra com os parâmetros da população que a define e pode ser escrito como

H_0: $\mu = \mu_0$

H_1: $\mu \neq \mu_0$, ou $\mu > \mu_0$, ou $\mu < \mu_0$

5.2.2 Ajustamento dos dados a um modelo estatístico

Um modelo estatístico está sempre associado a uma distribuição de probabilidade. Esta, por sua vez, descreve a probabilidade que uma variável tem de assumir determinado valor ao longo de um espaço de valores. Este curso tem se dedicado a apresentar com detalhes situações em que os dados se ajustam à curva normal, como é o caso do exemplo enunciado na seção 3.5.1. No entanto, outras formas de distribuição podem ser exploradas em http://www.itl.nist.gov/div898/handbook/eda/section3/eda366.htm.

5.2.3 Avaliação do modelo por meio de um teste estatístico

O objetivo de todo teste de hipótese é tentar rejeitar a hipótese nula, ou seja, encontrar evidências probabilísticas de que a média amostral não é igual à média populacional. Logo, se a média amostral apresentar uma probabilidade *p* muito pequena de ocorrer dentro da distribuição populacional definida pelo *status quo* (μ_0), então a probabilidade de $\mu = \mu_0$

também será muito pequena e teremos evidências estatísticas suficientes para rejeitar a hipótese H_0. Importante frisar que μ representa a média populacional da amostra coletada para o teste e μ_0 representa a média da população que define o *status quo*.

Considere o exemplo da seção 5.2.1. Sabemos que as distribuições da população e da amostra são normais. Portanto, uma forma de se observar a média amostral no espaço de probabilidades é convertê-la para o respectivo valor *z* da curva normal padronizada. O valor *z* encontrado, correspondente à média da amostra, é o valor a ser considerado na tomada de decisão e é comumente chamado de **estatística do teste**, denotado por $\boldsymbol{z_{calc}}$.

$$z_{calc} = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{61 - 58.29}{17.13 / \sqrt{258}} = 2{,}54$$

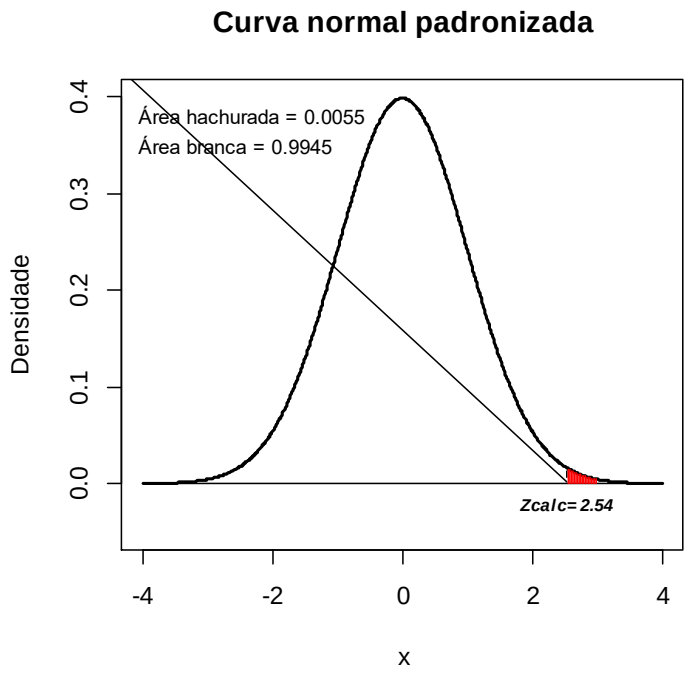


Figura 5.1 – $p(z_{calc} > 2{,}54) = 0{,}0055$.®

Exercício: utilize o script R da seção 3.2.1 para verificar $p(z_{calc} > 2{,}54)$ *e gerar a Figura 5.1.*

Esse valor divide a área da curva normal em duas regiões de probabilidade (Figura 5.1) e é fundamental compreender qual delas sustentará a análise (qual delas é a região crítica). Lembre-se que o objetivo do teste de hipótese é rejeitar H_0. Portanto, o pesquisador do nosso exemplo está interessado nos resultados pouco prováveis, mais distantes da média populacional ($\mu = 0$ na curva normal padronizada), ou seja, naqueles que, neste caso, são os que se encontram na região à direita de z_{calc}. Então, a principal tarefa consiste em determinar $p(z_{calc} > 2{,}54)$, isto é, a probabilidade de z ser maior que 2,54.

Da Figura 5.1, $p(z_{calc} > 2{,}54) = 0{,}0055$. Logo, a probabilidade de ocorrência da média amostral dentro da distribuição de probabilidade da população é "muito pequena" (e, por conseguinte, é muito pequena a chance de $\mu = \mu_0$), o que levaria o pesquisador a rejeitar a hipótese H_0. Contudo, o que significa, exatamente, a expressão "muito pequena" no contexto de um teste de hipótese? Na verdade, esse conceito é arbitrário e depende do tipo de pesquisa que se está realizando.

Dada a necessidade de se comparar resultados estatísticos entre experimentos de diversos pesquisadores, o mundo científico estabeleceu níveis de probabilidade máxima, ou **nível de significância** (α), em que se convencionou rejeitar H_0 (em geral, α é igual a 0,05 ou 0,01). Isso quer dizer que se o valor $\bar{x}$ (estimado a partir da amostra) for tão diferente do valor μ_0 (alegado para o parâmetro populacional) que a probabilidade de ocorrência de $\bar{x}$ seja de no máximo 0,05 (ou 0,01, conforme o caso), então H_0 deve ser rejeitada. Ou seja, rejeita-se H_0 quando $p(\bar{x}) < \alpha$, ou simplesmente, $p < \alpha$.

O uso de α = 0,05 provavelmente vem do fato de que em tempos pretéritos, antes do advento dos computadores, os cientistas validavam seus testes estatísticos confrontando-os com tabelas de "valores críticos". Essas tabelas eram calculadas por pessoas extremamente habilidosas como Ronald Aylmer Fisher (FIELD *et al*., 2012). O teste de hipótese, tal como o conhecemos hoje, é obra das ideias de Fisher.

Voltando ao exemplo, a probabilidade $p(z_{calc} > 2{,}54)$ é de 0,0055. Nesse caso, para qualquer valor de α escolhido (0,05 ou 0,01), há evidências estatísticas para se rejeitar H_0. Então, a hipótese experimental H_1 é aceita como verdadeira, ou seja, há um efeito na

população, em decorrência da participação do eleitor em algum programa social oferecido pelo atual governo, que faz com que a média do percentual de votos válidos aumente.

É quase sempre útil olhar o resultado de um teste conjuntamente com o IC, simplesmente porque o IC preenche a lacuna entre significância estatística e significância prática de forma bastante natural. Isto é, uma vez que o IC é expresso na mesma unidade dos dados em análise, é possível ver se a faixa de valores no IC tem significância prática ou não. Pode-se investigar se a média hipotética é ou não suportada observando se ela está ou não contida no IC.

5.2.4 Região crítica

Conforme visto na seção anterior, **região crítica** é aquela onde os valores da estatística do teste levam à rejeição da hipótese nula. A área da região crítica é igual ao nível de significância. A direção da região crítica é a mesma da hipótese alternativa e está ilustrada na Figura 5.2.

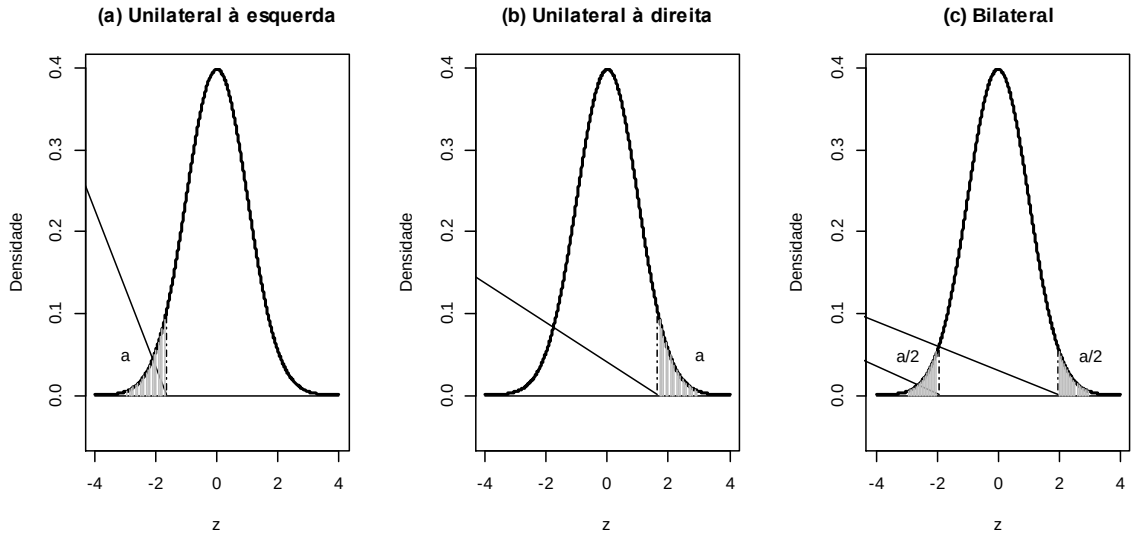


Figura 5.2 – Região crítica: (a) H_0: $\mu = \mu_0$ e H_1: $\mu < \mu_0$; (b) H_0: $\mu = \mu_0$ e H_1: $\mu > \mu_0$; (c) H_0: $\mu = \mu_0$ e H_1: $\mu \neq \mu_0$. ®

Quando se rejeita H_0 é porque existe uma forte evidência de sua falsidade, devido à "muito pequena" probabilidade da média amostral ser igual à populacional de referência. No

entanto, quando falhamos em rejeitar H_0, dizemos que não houve evidência amostral significativa no sentido de permitir a rejeição de H_0.

5.2.5 Tipos de erros

O fato de fazermos inferências sobre a população a partir de resultados amostrais torna o teste de hipótese suscetível a erros. Digamos que exista uma probabilidade de que, mesmo sendo H_0 verdadeira, $\overline{x}$ assuma um valor cujo z_{calc} leve à rejeição de H_0. Nesse caso, estaríamos cometendo um **erro do tipo I**, ou falso positivo, ou seja, quando decidimos que H_0 é falsa, mas na verdade não é.

Por outro lado, suponha que a estatística do teste nos levasse a aceitar H_0 quando esta fosse falsa. Então, estaríamos cometendo um **erro do tipo II**, ou falso negativo, ou seja, quando decidimos que H_0 é verdadeira, mas na verdade não é.

As probabilidades desses erros são denominadas, respectivamente, α e β, onde

$$\alpha = p(\text{erro do tipo I}) = p(\text{rejeitar } H_0 \text{ quando esta for verdadeira})$$
$$\beta = p(\text{erro do tipo II}) = p(\text{falhar em rejeitar } H_0 \text{ quando esta for falsa})$$

Quadro 5.1 – Probabilidades dos erros do tipo I e II.

Decisão	**Realidade**	
	H_0 verdadeira	**H_0 falsa**
Aceitar H_0	decisão correta (1-α)	erro do tipo II (β)
Rejeitar H_0	erro do tipo I (α)	decisão correta (1-β)

5.2.6 Força ou poder de um teste estatístico

Erros do tipo II ocorrem quando o tamanho da amostra é incapaz de detectar o efeito. A força de um teste é a probabilidade de se tomar a decisão correta e rejeitar H_0 quando esta for falsa. Consequentemente, a força é a área da distribuição gerada por um efeito, contida fora do intervalo de confiança da população não submetida ao efeito. Logo, conhecendo-se β, a força do teste é calculada por $1-\beta$.

Suponha que se está estudando o efeito de uma nova campanha publicitária sobre o grau de rejeição de um determinado candidato. Sabe-se que a campanha é efetiva e que sempre deu resultados positivos onde foi aplicada anteriormente, logo, espera-se a diminuição do grau de rejeição. Considere, então, um teste unilateral dado pelas hipóteses:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \end{cases}$$

Neste caso, a região de rejeição é determinada por $\{\overline{X} < X_C\}$, e a interpretação dos erros pode ser vista como:

$$\alpha = \mathbb{P}(\overline{X} < X_C | \mu = \mu_0);$$
$$\beta = \mathbb{P}(\overline{X} > X_C | \mu < \mu_0).$$

A situação ideal é aquela em que ambas as probabilidades, α e β, são próximas de zero. No entanto, é fácil ver que à medida que diminuímos α, β aumenta. A Figura a seguir apresenta esta relação. Considere que as distribuições têm a mesma variabilidade.

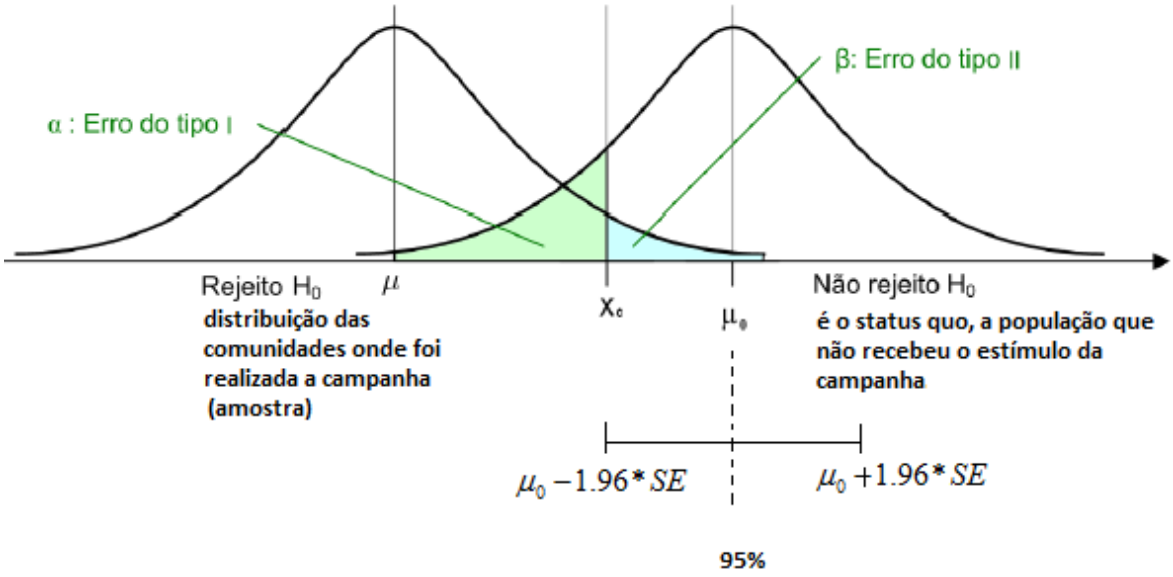


Figura 5.3 – Erro do tipo II.

Para um teste de hipóteses do tipo acima, onde estamos interessados em testar a média de uma população, utilizamos a expressão

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

que é a estatística do teste de hipóteses.

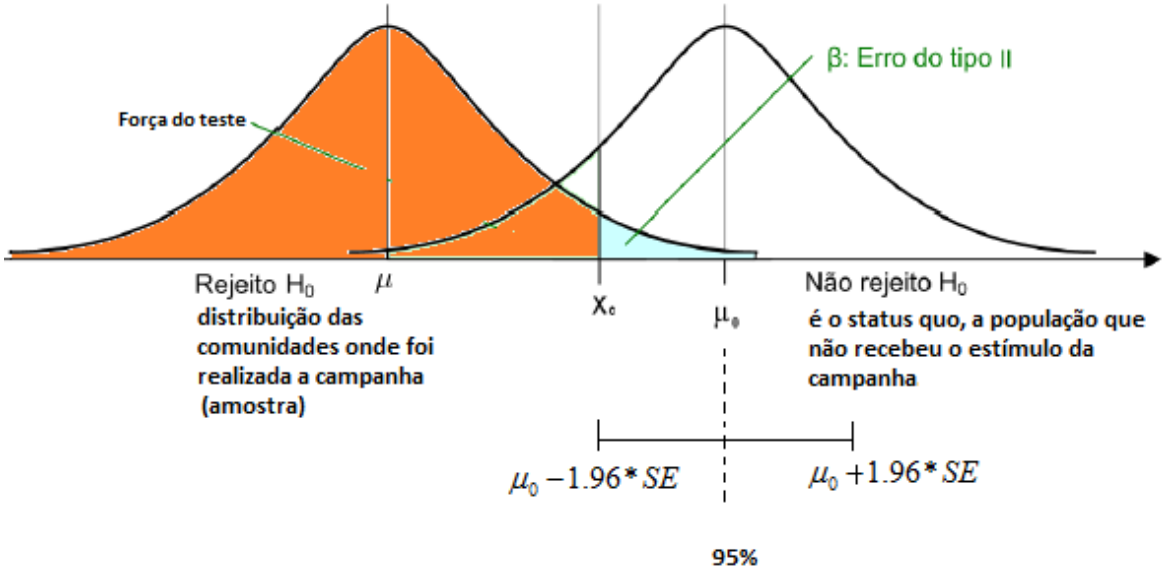


Figura 5.4 – Força do teste.

(Steve McKillup, 2005)

Como seria a hipótese $\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$

5.3 LABORATÓRIO 4

5.3.1 Comparação entre médias

Utilize a função R `dnormalComp`, a seguir, e resolva os exercícios subsequentes.

```
dnormalComp <- function(media1=0, dp1=1, media2=0, dp2=1, nc=.95, rc="=")
{

  ########################
  # Script principal
  ########################

  # eixo x da curva normal
  lim <- c(
           min(c(media1+c(-4,4)*dp1, media2+c(-4,4)*dp2)),
           max(c(media1+c(-4,4)*dp1, media2+c(-4,4)*dp2))
         )
  x <- seq(lim[1], lim[2], by = 0.01)

  # curva normal
  cn1 <- function(x) {dnorm(x,media1,dp1)} # curva normal
  cn2 <- function(x) {dnorm(x,media2,dp2)} # curva normal

  # traça as curvas normais 1 e 2
  if(cn1(media1)>=cn2(media2)){
    plot(x,cn1(x),ylab="Densidade",xlab="x",
         main="Curva Normal",type="l",lwd=2)
    lines(x,cn2(x),lwd=2, col="red")
  } else {
    plot(x,cn2(x),ylab="Densidade",xlab="x",
         main="Curva Normal",type="l",lwd=2,col="red")
    lines(x,cn1(x),lwd=2)
  }


  # linha horzontal em zero
  lines(lim,c(0,0))

  # linhas da média
  lines(c(media1,media1),c(-1,cn1(media1)),lwd=4,type="l")
  lines(c(media2,media2),c(-1,cn2(media2)),lwd=4,type="l",col="red")

  # intervalos de confiaça
  if(rc=="="){
    xI11 <- media1 - qnorm(nc+(1-nc)/2)*dp1
    xI12 <- media1 + qnorm(nc+(1-nc)/2)*dp1
    xI21 <- media2 - qnorm(nc+(1-nc)/2)*dp2
    xI22 <- media2 + qnorm(nc+(1-nc)/2)*dp2
  } else if(rc=="<"){
    xI11 <- media1 - 4*dp1
    xI12 <- media1 + qnorm(1-nc)*dp1
    xI21 <- media2 - 4*dp2
```

```
    xI22 <- media2 + qnorm(1-nc)*dp2
  } else if(rc==">"){
    xI11 <- media1 + qnorm(nc)*dp1
    xI12 <- media1 + 4*dp1
    xI21 <- media2 + qnorm(nc)*dp2
    xI22 <- media2 + 4*dp2
  }


  inc <- (xI12-xI11)/20
  i<-xI11+inc
  lines(c(i,i),c(-1,cn1(i)),col="black",lty=4,lwd=2)
  while(i < xI12){
    lines(c(i,i),c(0,cn1(i)),col="black",lwd=0.5)
    i<-i+inc
  }
  lines(c(i,i),c(-1,cn1(i)),col="black",lty=4,lwd=2)

  inc <- (xI22-xI21)/20
  i<-xI21+inc
  lines(c(i,i),c(-1,cn2(i)),col="red",lty=4,lwd=2)
  while(i < xI22){
    lines(c(i,i),c(0,cn2(i)),col="red",lwd=0.5)
    i<-i+inc
  }
  lines(c(i,i),c(-1,cn2(i)),col="red",lty=4,lwd=2)
}
```

5.3.1.1 Exercício 1

Verifique se as amostras abaixo pertencem à mesma população.

```
a1 <- c(
      18.8,17.591,20.835,19.169,18.755,20.504,
      18.756,17.527,19.29,19.203,18.621,18.977,17.078,22.059,18.419,
      19.919,20.308,17.62,18.585,20.764,21.117,18.899,21.426,17.89,21.055
)

a2 <- c(
      22.284,22.057,22.629,24.62,21.491,21.198,
      21.901,22.881,22.86,22.058,22.699,22.909,
      25.302,17.968,24.515,23.15,24.662,23.327,
      22.447,23.382,22.426,22.787,21.983,24.534,
      22.771,21.043,21.203,24.009,21.917,21.152
)
```

5.3.1.2 Exercício 2

Leia o arquivo “exercicio.csv” com a linha de comando a seguir:

```
df <- read.csv("exercicio.csv",sep=",", header=TRUE)
```

Responda as questões a seguir utilizando nível de confiança de 95%.

1. A amostra da coluna 1 pertence à mesma população da amostra da coluna 2?
   A média da coluna 1 é maior que, menor que ou igual a da coluna 2?

2. A amostra da coluna 2 pertence à mesma população da amostra da coluna 3? A média da coluna 2 é maior que, menor que ou igual a da coluna 3?

3. A amostra da coluna 1 pertence à mesma população da amostra da coluna 3? A média da coluna 1 é maior que, menor que ou igual a da coluna 3?

5.4 DISTRIBUIÇÃO T DE STUDENT

O teste *z* é utilizado quando conhecemos o desvio padrão populacional (σ) e podemos determinar o erro padrão ($\sigma_{\bar{x}}$). Contudo, na maioria das situações práticas, σ é desconhecido. Nos casos em que a amostra possui tamanho razoável (n>30), o desvio padrão amostral *s* é considerado um bom estimador de σ e pode ser usado para calcular $\sigma_{\bar{x}}$. Contudo, para trabalhar com pequenas amostras, é necessário utilizar a distribuição *t* de Student[13] (ou estatística *t*).

A distribuição *t* é uma distribuição de probabilidade teórica. É simétrica, tem forma de sino, média zero e é semelhante à curva normal padrão, porém, é mais achatada e com caudas mais largas. Ou seja, uma simulação de Student pode gerar valores mais extremos que uma simulação normal. O intervalo da variável *t* é [-∞,∞]. O único parâmetro que a define e caracteriza a sua forma é o número de graus de liberdade. Quanto maior for esse parâmetro, mais próxima da normal ela será.

Quadro 5.2 – Diferença entre as estatísticas *t* e *z*.

Estatística	Fórmula	Características e notas
z	$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$	A população segue distribuição normal (ou n > 30) e σ (desvio-padrão) é conhecido. *z* é expresso em quantidades de desvio padrão da média.
t	$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$	A população segue distribuição normal (ou n > 30) e σ (desvio-padrão) é **des**conhecido. Utiliza-se o desvio-padrão amostral.

[13] A estatística *t* foi introduzida em 1908 por William Sealy Gosset, químico da cervejaria Guinness em Dublin, Irlanda ("student" era seu pseudônimo). Gosset havia sido contratado devido à política inovadora de Claude Guinness que consistia em recrutar os melhores graduados de Oxford e Cambridge para os cargos de bioquímico e estatístico da indústria Guinness. Gosset desenvolveu o teste *t* como um modo barato de monitorar a qualidade da cerveja tipo robusta. Ele publicou o teste *t* na revista acadêmica Biometrika em 1908, mas foi forçado a usar seu pseudônimo por Claude Guinness, que acreditava que o uso da estatística no ramo era um segredo industrial. De fato, a identidade de Gosset não foi reconhecida por seus colegas estatísticos.

A diferença fundamental entre *z* e *t* está nos respectivos denominadores. Para *z*, emprega-se σ. Para *t*, emprega-se *s*, que é baseado em graus de liberdade (gl = *n* - 1).

Em estatística, os graus de liberdade são relacionados ao número de observações que estão livres para variar. Em outras palavras, grau de liberdade é o número de determinações independentes (dimensão da amostra) menos o número de parâmetros estatísticos a serem avaliados na população.

A distribuição de *t* é descrita por uma família de distribuições para cada gl (Figura 5.5). A variação de *t* é maior em pequenas amostras; quando $n \rightarrow \infty$, $s \rightarrow \sigma$ e $t \rightarrow z$.

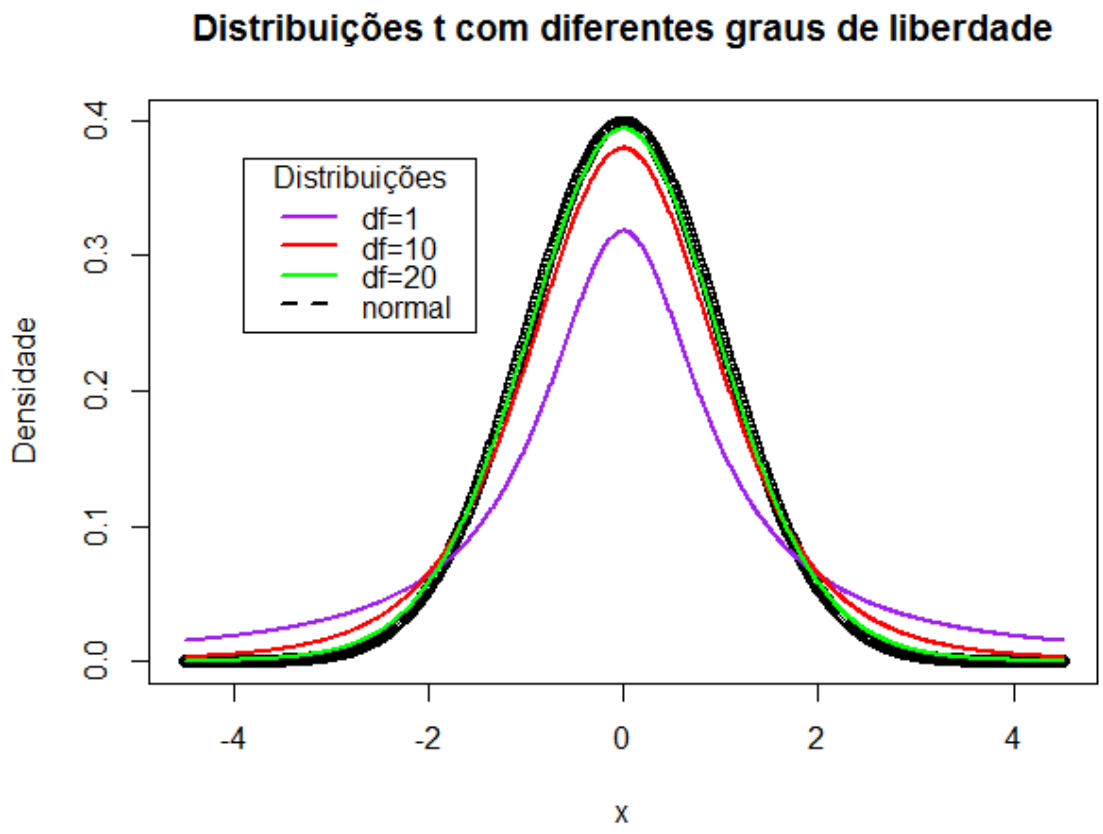


Figura 5.5 – Distribuições t de Student. ®

5.4.1 Intervalo de confiança em distribuições *t*

Intervalos de confiança em distribuições *t* são determinados da mesma forma que para a distribuição normal padrão, contudo, utilizando-se a estatística *t*.

$$x_1 = \bar{x} - t_{n-1}\frac{s}{\sqrt{n}} \qquad e \qquad x_2 = \bar{x} + t_{n-1}\frac{s}{\sqrt{n}} \tag{5.3}$$

Quadro 5.3 – Quantis *t* para intervalo bicaudal com 95% de confiança (*n*=11).

```
n <- 11
gl <- n - 1
t <- qt(.975, gl)
t

[1] 2.228139
```

5.4.2 Grupos pareados

Observações pareadas são frequentemente analisadas por meio de intervalos *t* para a identificação de diferenças. Relembrando conceitos, um **projeto de medidas repetidas**, **relacionadas** ou **pareadas** é aquele em que a variável independente é manipulada sobre o mesmo grupo de participantes, em condições distintas: experimental e de controle. Uma vez que a medição efetuada sobre um sujeito é comparada com outra efetuada sobre ele mesmo, esta técnica possibilita ao experimentador controlar as diferenças individuais dos participantes. Assume-se, então, que os fatores de variabilidade são os mesmos e, portanto, a variância é a mesma antes e depois da manipulação da variável independente.

Por exemplo, considere os dados[14] que Gosset analisou em seu artigo sobre testes *t* no periódico Biometrika, em 1908. Duas drogas diferentes utilizadas para dormir foram ministradas a dois grupos com 10 pacientes cada. A variável "extra" corresponde ao acréscimo em horas de sono proporcionado a cada paciente. A variável "group" rotula a droga ministrada ao paciente.

Embora a documentação (? sleep) declare que o experimento tenha sido realizado em grupos distintos, vamos proceder à análise como se as medicações tivessem sido ministradas sobre um mesmo grupo, em momentos distintos, caracterizando a análise pareada. A Figura 5.6 ilustra o gráfico de caixa (*boxplot*) do grupo para cada droga ministrada.

[14] Os dados da pesquisa de Gosset fazem parte do pacote `datasets` do R.

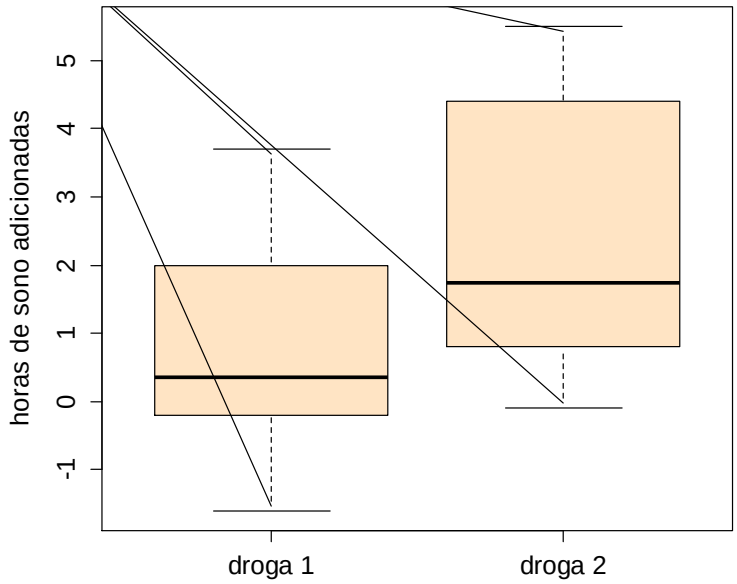


Figura 5.6 – Boxplot para grupos pareados. ®

O objetivo consiste em verificar se as médias das horas de sono adicionadas diferem para as drogas 1 e 2. Por inspeção da Figura 5.6, observando-se as medianas, parece razoável afirmar que diferem, sendo maior para a droga 2. Contudo, precisamos aplicar um teste estatístico, no caso, o teste *t* pareado. A hipótese nula H_0, então, enuncia que as médias são iguais (H_0: $\bar{x}_1 = \bar{x}_2$), ou que a diferença entre elas é nula (H_0: $\bar{x}_2 - \bar{x}_1 = 0$).

O Quadro 5.4 apresenta o código R para a análise pareada dos dados. Na primeira parte do código, os parâmetros do teste *t* são determinados aplicando-se as definições e a Equação (5.4).

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{5.4}$$

Em seguida, a função `t.test` do R é utilizada de três formas distintas, cujos parâmetros são: 1) o vetor de diferenças; 2) os vetores com os resultados medidos para cada droga; 3) a variável dependente “extra” como função da variável independente “group”. Por fim, a Figura 5.7 ilustra a posição relativa de cada indivíduo comparando-se os dois casos.

Quadro 5.4 – Análise de grupos pareados, assumindo igual variância.

```
data(sleep); sleep
g1 <- sleep$extra[1 : 10]; g2 <- sleep$extra[11 : 20]
difference <- g2 - g1
mn <- mean(difference); s <- sd(difference); n <- 10
# estatistica t
t <- mn/(s/sqrt(10))
# graus de liberdade
gl <- n - 1
# p-value bicaudal
p_value <- 2*(1-pt(t, gl))
# intervalo de confiança para nível de confiança de 95%
ic <- mn + c(-1, 1) * qt(.975, n-1) * s / sqrt(n)
# média
media <- mn
cat("t = ",t,", df = ",gl,", p-value = ",p_value,"\nic = ",ic[1]," ",
    ic[2],"\nmédia da diferença = ", media,
    "\ng1: média = ",mean(g1)," SE = ",sd(g1)/sqrt(length(g1)),
    "\ng2: média = ",mean(g2)," SE = ",sd(g2)/sqrt(length(g2)), sep="")

t = 4.062128, df = 9, p-value = 0.00283289
ic = 0.7001142 2.459886
média da diferença = 1.58
g1: média = 0.75 SE = 0.5657345
g2: média = 2.33 SE = 0.6331666

t.test(difference)

        One Sample t-test

data:  difference
t = 4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.7001142 2.4598858
sample estimates:
mean of x
     1.58

t.test(g2, g1, paired = TRUE)

        Paired t-test

data:  g2 and g1
t = 4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.7001142 2.4598858
sample estimates:
mean of the differences
                  1.58

t.test(extra ~ I(relevel(group, 2)), paired = TRUE, data = sleep)

        Paired t-test

data:  extra by I(relevel(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:
 0.7001142 2.4598858
sample estimates:
mean of the differences
                  1.58
```

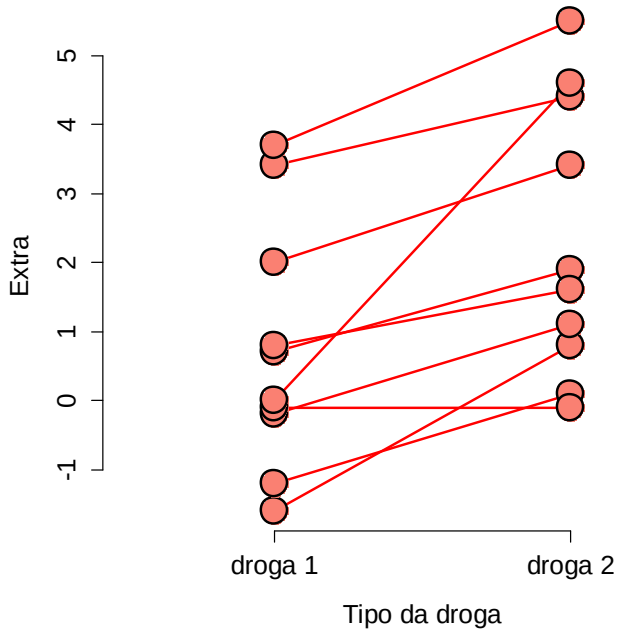


Figura 5.7 – Dados de Gosset, teste *t*. ®

A forma correta de relatar o resultado do teste *t* pareado do Quadro 5.4 é:

> Em média, os participantes experimentaram maior quantidade de horas de sono adicionais ao utilzarem a droga 2 (M = 2.33, SE = 0.57), quando comparado com o uso da droga 1 (M = 0.75, SE = 0.63). Essa diferença é estatisticamente significante: t(9) = 4.06, $p < 0.05$.

5.4.3 Grupos independentes

Um **projeto de medidas independentes** é aquele em que a variável independente é manipulada em grupos distintos de participantes, de modo que um grupo participa de uma condição experimental, enquanto o outro grupo participa de condição diferente.

5.4.3.1 Variâncias iguais

Quando a variância dos dois grupos é a mesma, a estimativa do intervalo de confiança deve levar em conta o tamanho dos grupos. Na verdade, o IC é estimado em função da variância ponderada pelo número de graus de liberdade de cada grupo, conforme a Equação (5.5).

$$s_P^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{5.5}$$

Logo, considerando-se o IC para a diferença das médias, ou seja, para $H_0: \bar{x}_2 - \bar{x}_1 = 0$, temos

$$IC = (\bar{x}_2 - \bar{x}_1) \pm t_{n_1+n_2-2,1-\alpha/2} * s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{5.6}$$

É importante lembrar que o IC da Equação (5.6) só pode ser estimado quando houver certeza de que as variâncias dos grupos são iguais. Em caso de dúvida, deve ser adotada a estimativa para variâncias distintas, discutida na próxima seção.

O Quadro 5.5 ilustra a estimativa do IC, considerando-se variâncias iguais, quando são comparadas as medições da pressão sanguínea sistólica de dois grupos de mulheres: o primeiro faz uso de contraceptivo oral; o segundo é o grupo controle, que não faz uso de contraceptivo. Observa-se que o zero é uma possibilidade do IC e, portanto, não pode ser descartado, ou seja, não há evidências para se rejeitar H_0. Em outras palavras, o uso do contraceptivo não altera a medição da pressão sanguínea sistólica.

Quadro 5.5 – Grupos independentes: variâncias iguais.

```
# Pressão Sanquínea Sistólica
# Teste t para averiguar Ho: m2 – m1 = 0
# grupo controle
n1 <- 21     # tamanho
m1 <- 127.44 # média em mmHg
s1 <- 18.23  # desvio-padrão em mmHg

# grupo que utiliza contraceptivo
n2 <- 8      # tamanho
m2 <- 132.86 # média em mmHg
s2 <- 15.34  # desvio-padrão em mmHg

# variância ponderada
sp <- sqrt(  ((n1-1)*s1^2 + (n2-1)*s2^2) / (n1+n2-2)  )

# intervalo de confiança
(m2 - m1) + c(-1,1) * qt(.975, n1+n2-2) * sp * sqrt(1/n1 + 1/n2)

[1] -9.521097 20.361097
```

Grupos independentes com igual variância podem ser confundidos com grupos pareados. Contudo, são tratamentos distintos uma vez que os independentes levam em conta o tamanho dos grupos (ou gl). O Quadro 5.6 exemplifica a diferença dos tratamentos. Por exemplo, tratar os grupos do experimento de Gosset como pareados nos levaria a rejeitar H_0 (o zero não está contido no IC). Por outro lado, a análise de grupos independentes concluiria pela aceitação de H_0 (o zero está contido no IC).

Quadro 5.6 – Engano ao tratar grupos independentes como pareados.

```
data(sleep); sleep
g1 <- sleep$extra[1 : 10]; g2 <- sleep$extra[11 : 20]
n1 <- length(g1); n2 <- length(g2)

# independente dmesma variância, com variância ponderada por gl
sp <- sqrt(  ((n1-1)*sd(g1)^2 + (n2-1)*sd(g2)^2) / (n1+n2-2)  )
md <- mean(g2-g1)
semd <- sp * sqrt(1/n1 + 1/n2)
rbind(
  # pareado
  t.test(g2, g1, paired=TRUE)$conf
  # mesma variância, independente
  md + c(-1,1) * qt(.975, n1+n2-2) * semd,
  t.test(g2, g1, paired=FALSE, var.equal = TRUE)$conf,
)

           [,1]     [,2]
[1,]  0.7001142 2.459886
[2,] -0.2038740 3.363874
[3,] -0.2038740 3.363874
```

5.4.3.2 Variâncias diferentes

Quando as variâncias forem diferentes, a estimativa do IC é dada por

$$IC = (\bar{x}_2 - \bar{x}_1) \pm t_{df} \sqrt{\frac{S_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{5.7}$$

onde t_{df} é calculado em termos do número de graus de liberdade, como na Equação (5.8) a seguir.

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} \tag{5.8}$$

A técnica das variâncias diferentes funciona bem e deve ser utilizada sempre que houver dúvida sobre a igualdade ou não das variâncias. Repetindo-se o exemplo do Quadro 5.5, agora considerando variâncias diferentes, encontramos um novo IC, ilustrado no Quadro 5.7. Mesmo com o novo critério, não há evidências para se rejeitar H_0.

Quadro 5.7 – Grupos independentes: variâncias diferentes.

```
# Pressão Sanquínea Sistólica
# Teste t para averiguar Ho: m2 – m1 = 0
# grupo controle
n1 <- 21     # tamanho
m1 <- 127.44 # média em mmHg
s1 <- 18.23  # desvio-padrão em mmHg

# grupo que utiliza contraceptivo
n2 <- 8      # tamanho
m2 <- 132.86 # média em mmHg
s2 <- 15.34  # desvio-padrão em mmHg

# graus de liberdade
df <- (s1^2/n1 + s2^2/n2)^2 / ((s1^2/n1)^2/(n1-1) + (s2^2/n2)^2/(n2-1))
# estatística t
t <- qt(.975, df)

# intervalo de confiança
(m2 - m1) + c(-1,1) * t * sqrt( s1^2/n1 + s2^2/n2 )

[1] -8.913327 19.753327
```

5.5 LABORATÓRIO 5

5.5.1 Simulação

```
############################################################
# Lab 3a – Comparação entre normal padrão e curva t
############################################################
library(ggplot2)
library(manipulate)

k <- 1000
xvals <- seq(-5, 5, length = k)
myplot <- function(df){
    d <- data.frame(y = c(dnorm(xvals), dt(xvals, df)),
                    x = xvals,
                    dist = factor(rep(c("Normal", "T"), c(k,k))))
    g <- ggplot(d, aes(x = x, y = y))
    g <- g + geom_line(size = 2, aes(colour = dist))
    g
}
manipulate(myplot(mu), mu = slider(1, 20, step = 1))


############################################################
# Lab 3b
############################################################
library(ggplot2)
library(manipulate)
pvals <- seq(.01, .99, by = .01)
myplot2 <- function(df){
    d <- data.frame(n= qnorm(pvals),t=qt(pvals, df)
                    ,p = pvals)
    g <- ggplot(d, aes(x= n, y = t))
    g <- g + geom_abline(size = 2, col = "lightblue")
    g <- g + geom_line(size = 2, col = "black")
    g <- g + geom_vline(xintercept = qnorm(0.975))
    g <- g + geom_hline(yintercept = qt(0.975, df))
    g
}
manipulate(myplot2(df), df = slider(1, 20, step = 1))
```

5.5.2 Comparação entre a média da população e a média de uma amostra

```
# Temperaturas corporais medidas
x<-c(30.5,35.3,33.2,40.8,42.3,41.5,36.3,43.2,34.6,38.5)

# One Sample t-test
t.test(x,     #amostra a ser testada
       mu=35, #hipótese de nulidade
       alternative="greater") #teste unilateral pela direita

# intervalo de confiança para gl = 9
ic<-mean(x)-1.833*sd(x)/sqrt(length(x))
ic
```

5.5.3 Comparação entre as médias de duas amostras independentes

```
x<-c(30.5,35.3,33.2,40.8,42.3,41.5,36.3,43.2,34.6,38.5)
y<-c(28.2,35.1,33.2,35.6,40.2,37.4,34.2,42.1,30.5,38.4)

#  Welch Two Sample t-test
t.test(x,y,               #amostras a serem testadas
       conf.level = 0.99) #nível de confiança
```

5.5.4 Comparação entre as médias de duas amostras pareadas

```
x<-c(30.5,35.3,33.2,40.8,42.3,41.5,36.3,43.2,34.6,38.5)
y<-c(28.2,35.1,33.2,35.6,40.2,37.4,34.2,42.1,30.5,38.4)

# Paired t-test
t.test(x,y,             #amostras a serem testadas
       conf.level=0.99, #nível de confiança
       paired=T)        #indica dependência entre as amostras
```

5.6 TESTES PARAMÉTRICOS

Os métodos de inferência estatística vistos neste material partem do pressuposto de que os dados são normalmente distribuídos. A qualidade das inferências feitas a partir desses métodos depende de quão próxima a população em estudo está da distribuição normal. Portanto, é necessário testarmos se os dados apresentam desvios da suposição de normalidade.

5.6.1 Shapiro-Wilk

Esse teste, proposto em 1965, calcula uma estatística W que testa se uma amostra aleatória de tamanho *n* provém de uma distribuição normal. A estatística W é calculada de acordo com a Equação (5.4)

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \tag{5.4}$$

em que x_i são os valores amostrais ordenados e a_i são constantes geradas das médias, variâncias e covariâncias das estatísticas de ordem de uma amostra aleatória de tamanho n de uma distribuição normal. Valores pequenos de W são evidência de desvios da normalidade.

Para realizar o teste de Shapiro-Wilk, devemos formular as hipóteses:

H_0: a amostra provém de uma população com distribuição normal.
H_1: a amostra não provém de uma população com distribuição normal.

Então, uma vez determinada a estatística do teste W e o nível de significância α, toma-se a decisão por rejeitar ou aceitar H_0. Os quadros 5.8 e 5.9 ilustram o teste de Shapiro-Wilk para as distribuições normal e uniforme, não normal.

Quadro 5.8 – Teste de Shapiro-Wilk em uma distribuição normal.

```
shapiro.test(rnorm(100, mean = 5, sd = 3))

        Shapiro-Wilk normality test

data:  rnorm(100, mean = 5, sd = 3)
W = 0.9852, p-value = 0.3307
```

Quadro 5.9 – Teste de Shapiro-Wilk em uma distribuição uniforme.

```
shapiro.test(runif(100, min = 2, max = 4))

        Shapiro-Wilk normality test

data:  runif(100, min = 2, max = 4)
W = 0.9522, p-value = 0.001168
```

Observa-se no Quadro 5.8 que o valor de p é maior que 0,05, ou seja, não existem evidências estatísticas significativas para se rejeitar H_0 e, portanto, a distribuição é considerada normal. No Quadro 5.9, $p < 0{,}05$, logo, H_0 é rejeitada e a distribuição é tida como não normal.

A função R para o teste de Shapiro-Wilk tem a limitação de tratar conjuntos de dados com tamanho máximo de 5.000 elementos. Para conjuntos maiores, deve-se usar o teste de Kolmogorov-Smirnov.

5.6.2 Kolmogorov-Smirnov

Este teste observa a máxima diferença absoluta entre a função de distribuição acumulada assumida para os dados, no caso a Normal, e a função de distribuição empírica dos dados. Como critério, comparamos esta diferença com um valor crítico, para um dado nível de significância. A hipótese formulada é a mesma para o teste de Shapiro-Wilk.

Quadro 5.10 – Teste de Kolmogorov-Smirnov em uma distribuição normal.

```
# distribuição normal N~(5,3) de tamanho 100.000
dn=rnorm(100000, mean = 5, sd = 3)
# converter os dados para os valores z respectivos
z=(dn-mean(dn))/sd(dn)
# usar o teste de Kolmogorov-Smirnov para comparar a distribuição com a
# normal padrão
ks.test(z,"pnorm")

   One-sample Kolmogorov-Smirnov test

data:  z
D = 0.0027, p-value = 0.4412
alternative hypothesis: two-sided
```

No Quadro 5.10, $p > 0{,}05$, logo, não evidências estatísticas significativas para se rejeitar H_0. A função R para o teste de Kolmogorov-Smirnov não tem a limitação para o tamanho do conjunto de dados.

5.6.3 Gráfico Q-Q Plot

O gráfico quantil-quantil plot ou Q-Q Plot é utilizado para determinar se dois conjuntos de dados pertencem à mesma distribuição de probabilidades. Em tais gráficos os pontos são formados pelos quantis das duas amostras. Caso os pontos se alinhem numa reta de inclinação 1, as distribuições das duas amostras podem ser consideradas as mesmas.

Os quantis dividem os dados ordenados em q subconjuntos de dados de dimensão essencialmente igual[15]. Dessa forma dão origem a q-quantis; os quantis são estabelecidos a partir de pontos de corte que determinam as fronteiras entre os subconjuntos consecutivos.

[15] Alguns quantis têm nomes especiais: os 100-quantis são chamados percentis (P); os 12-quantis são chamados duo-deciles (Dd); os 10-quantis são chamados decis (D); os 5-quantis são chamados quintis (QU); os 4-quantis são chamados quartis (Q); os 3-quantis são chamados tercis (T).

Visto de outra forma, o k-ésimo quantil é o valor de *x* tal que a probabilidade de um evento da variável aleatória ser inferior a *x* é de no máximo k/q; a probabilidade de a variável aleatória ser superior ou igual a x é pelo menos (q-k)/q. Há q - 1 quantis, sendo k um inteiro satisfazendo 0 < k < q.

O gráfico Q-Q Plot pode ser utilizado para o teste de normalidade. Quando comparado o quantil amostral com o quantil esperado sob normalidade e a configuração de pontos se aproximar de uma reta, a suposição de normalidade é sustentável. A normalidade é suspeita quando houver pontos que se desviam do comportamento linear. A forma como os pontos se desviam do comportamento linear pode fornecer pistas sobre a natureza da não normalidade das observações. Conhecida a razão da não normalidade dos dados, ações corretivas podem ser tomadas (transformações visando normalizar os dados ou uso de técnicas para dados não normais).

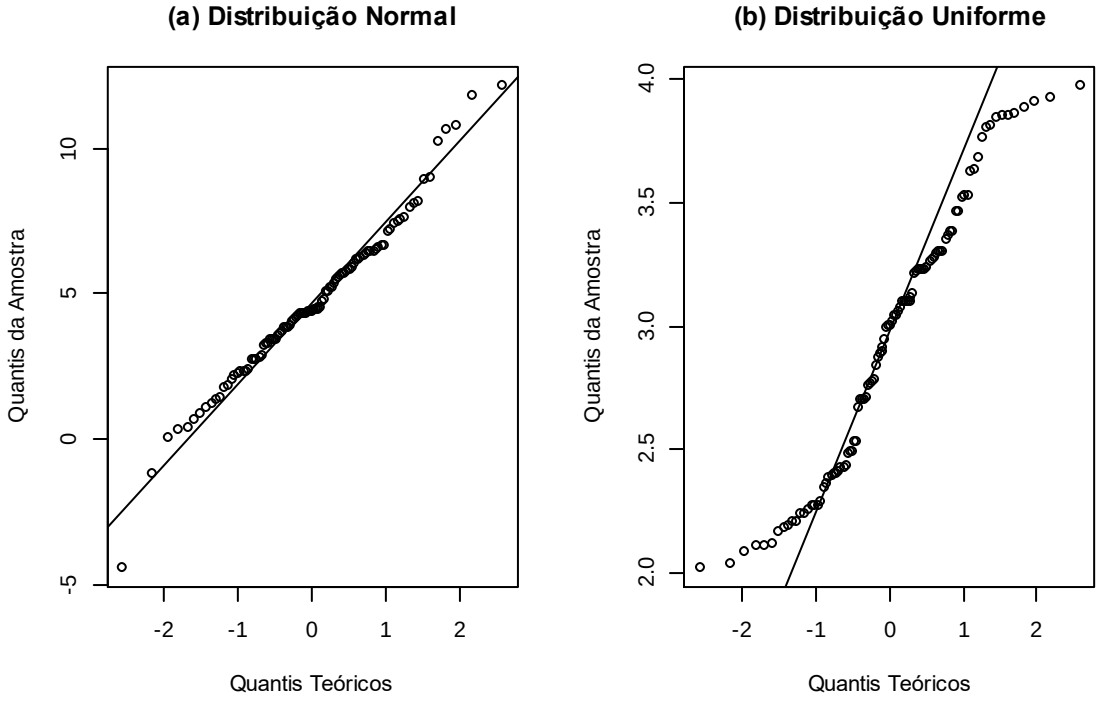


Figura 5.8 – Q-Q Plot. ®

http://www.portalaction.com.br/content/62-teste-de-kolmogorov-smirnov
http://rss.acs.unt.edu/Rdoc/library/fBasics/html/NormalityTests.html
http://ww2.coastal.edu/kingw/statistics/R-tutorials/singlesample.html
http://www.portalaction.com.br/content/64-teste-de-shapiro-wilk

http://www.r-tutor.com/elementary-statistics/non-parametric-methods/wilcoxon-signed-rank-test

6 CORRELAÇÃO LINEAR E REGRESSÃO LINEAR SIMPLES

Neste módulo serão abordados os conceitos de correlação e regressão linear, ambos inventados por Francis Galton, fundador do periódico Biometrika.

6.1 CORRELAÇÃO LINEAR

A **correlação linear** mede o grau de associação linear entre duas variáveis. A maneira mais simples de se saber se duas variáveis são associadas é observar se elas variam conjuntamente (ou covariam). Isso significa dizer que as respectivas variâncias têm comportamentos similares, ou seja, quando uma variável se desvia da sua média, espera-se que a outra se desvie de maneira similar.

Por exemplo, considere que cinco pessoas tenham sido submetidas, individualmente, a seções de vídeos de propaganda sobre as realizações do governo atual. Ao final das seções, cuja quantidade de vídeos era diferente para cada indivíduo, uma nota de zero a quinze era atribuída para o desempenho do atual governo. O resultado está ilustrado na Tabela 6.1.

Tabela 6.1 – Vídeos de propaganda assistidos versus nota de aprovação do governo.

	Pessoa 1	Pessoa 2	Pessoa 3	Pessoa 4	Pessoa 5	Média	Desvio Padrão
Vídeos assistidos	5	4	4	6	8	**5,4**	**1,67**
Nota	9	8	10	13	15	**11,0**	**2,92**

A Figura 6.1 mostra as médias dos vídeos assistidos (linha horizontal em azul) e das notas atribuídas (linha horizontal em vermelho). As linhas verticais representam as diferenças entre os valores observados e as médias das respectivas variáveis. Pode ser identificado um padrão de comportamento similar nas duas variáveis: para os três primeiros participantes, os valores observados estão abaixo da média em ambas as variáveis; para os dois últimos participantes, os valores estão acima. Esse padrão é indicativo de um potencial relacionamento entre as duas variáveis.

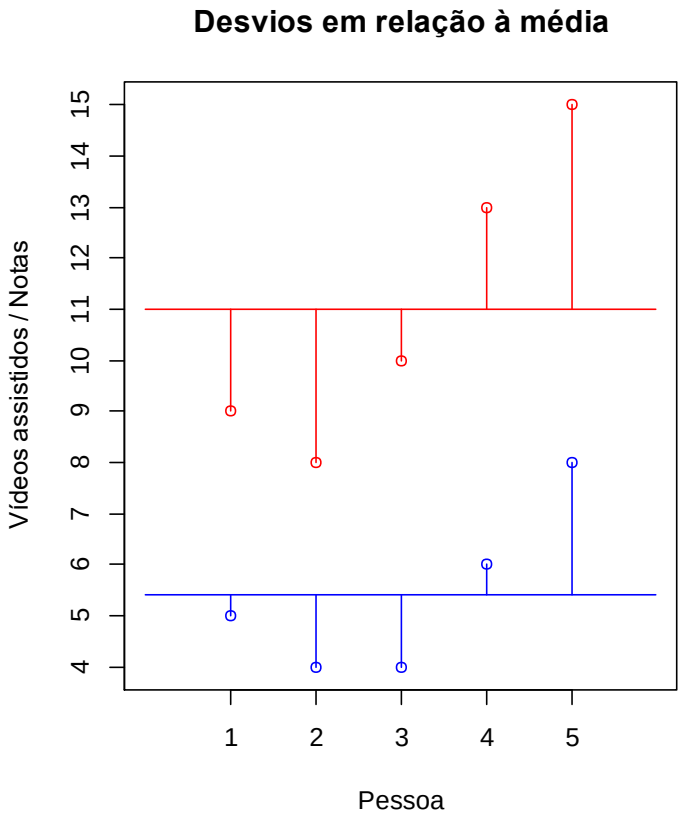


Figura 6.1 – Desvio dos dados em relação à média. ®

Uma forma de se quantificar o grau de similaridade entre duas variáveis é por meio do cálculo da **covariancia** (Equação (6.1)).

$$Cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \tag{6.1}$$

Ao invés de calcular o erro quadrático de cada variável, a covariancia determina o produto dos erros (ou desvios) das variáveis, também conhecido como **produto cruzado dos desvios**. Dessa forma, quando ambos os desvios são positivos ou negativos, o produto será positivo. Caso um desvio seja positivo e o outro negativo, então o resultado será negativo.

O cálculo da covariancia é uma boa maneira de avaliar se duas variáveis são relacionadas. Uma covariancia positiva indica que, na medida em que uma variável se desvia da média, a outra se desvia na mesma direção. A covariancia negativa indica que as variações se dão em sentidos opostos.

Há um problema em se utilizar a covariancia como medida do relacionamento entre variáveis: ela depende da escala das medidas usadas. Dessa forma, covariancias não podem ser comparadas objetivamente porque se torna difícil afirmar se o valor encontrado é grande ou pequeno.

Para converter a covariancia em uma medida padrão, para a qual qualquer escala possa ser convertida, devemos dividi-la pelos desvios padrão das respectivas variáveis (Equação (6.2)).

$$r = \frac{Cov(x,y)}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \tag{6.2}$$

O coeficiente *r* da Equação (6.2) é conhecido como **coeficiente de correlação de Pearson**. O valor de *r* pode variar de -1 a +1. O valor absoluto de *r* mostra a intensidade da relação linear entre as variáveis: 0 indica ausência de relação; 1 indica relação perfeita. Valores entre 0 e 1 podem significar que: os valores de *x* determinam os valores de *y*; os valores de *y* determinam os valores de *x*; *x* e *y* são influenciados por uma terceira variável.

Toda correlação apresenta uma probabilidade de ter ocorrido devido ao acaso. Um teste estatístico é necessário para verificar se a correlação é estatisticamente significativa (para IC = 95%, $p < 0{,}05$):

H_0: não há correlação entre as variáveis, $r = 0$.
H_1: há correlação entre as variáveis, $r \neq 0$.

O Quadro 6.1 ilustra a aplicação de *r* para os dados da Tabela 6.1. Neste caso foi encontrado $r = 0{,}92$, o que indica forte correlação positiva entre as variáveis, ou seja, as variações acontecem na mesma direção. Mas uma forte correlação não basta, é preciso que exista significância estatística. O valor de $p = 0{,}026 < 0{,}05$ revela baixa probabilidade do valor de *r* ser obra do acaso. Logo, H_0 deve ser rejeitada e podemos considerar que *r* é mesmo diferente de zero, isto é, existe correlação entre as variáveis.

Quadro 6.1 – Coeficiente de correlação de Pearson.

```
# variáveis
videos<-c(5,4,4,6,8)
notas<-c(9,8,10,13,15)

cor.test(videos,notas,method="pearson",alternative="two.sided")

    Pearson's product-moment correlation

data:  videos and notas
t = 4.1367, df = 3, p-value = 0.02564
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2156075 0.9949637
sample estimates:
      cor
0.9224101
```

A melhor maneira de interpretar o valor da medida de associação linear de duas variáveis é elevá-lo ao quadrado (r^2). Para o exemplo do Quadro 6.1, $r^2 = 0{,}85$, o que significa dizer que 85% da variação das notas é explicada pela quantidade de vídeos de propaganda assistidos pelos participantes.

6.2 REGRESSÃO LINEAR SIMPLES

Quando adotamos a simplificação de uma realidade por meio de um modelo, a palavra de ordem é parcimônia. Parcimônia no sentido de se encontrar um modelo que seja tão simples quanto possível e interpretável.

Nesse sentido, a **regressão linear** pode ser considerada um dos mais modernos algoritmos de aprendizado de máquina (*machine learning*), pois proporciona a criação de modelos estatísticos amplamente interpretáveis e permite quantificar o que não pode ser explicado pelo modelo. Em geral, modelos estatísticos descrevem como variáveis aleatórias se relacionam.

A **regressão linear simples** é um modelo que nos permite verificar a existência ou não de relacionamento ou dependência entre duas variáveis. Por exemplo, um pesquisador pode supor, por observação ou experimentação, que o tempo de estudo impacta na nota de uma

prova, que a frequência com que se fuma tem influência na idade do primeiro infarto, que a altura de uma pessoa é relacionada ao seu peso.

Então, na verdade, o que quermos é estimar os valores de determinada variável *Y*, e para isso, consideramos os valores de outra variável *x* que acreditamos ter poder de explicação sobre *Y*. Ambas são quantitativas. Para os nossos propósitos, *x* é a variável independente, preditora, não aleatória, sem erro devido ao acaso, representada por letra minúscula.

Alguns tipos de questões são inerentes ao modelo de regressão linear:

- Como usar *x* para predizer *Y*?
- Como encontrar uma relação parcimoniosa, facilmente descrita pela média, entre *x* e *Y*?
- Como investigar a variação em *Y* que aparenta não ser decorrente de *x* (variação residual)?
- Como quantificar o quanto a variação de *x* explica a variação em *Y*?
- Quais as assunções necessárias para a generalização das descobertas decorrentes da relação entre *x* e *Y*?

Uma primeira abordagem para se tentar entender o quanto uma variável explica a outra é conhecer a sua distribuição. Vamos, então, explorar os dados utilizados por Galton em 1885, que consistem em 928 observações das alturas de pais e filhos.

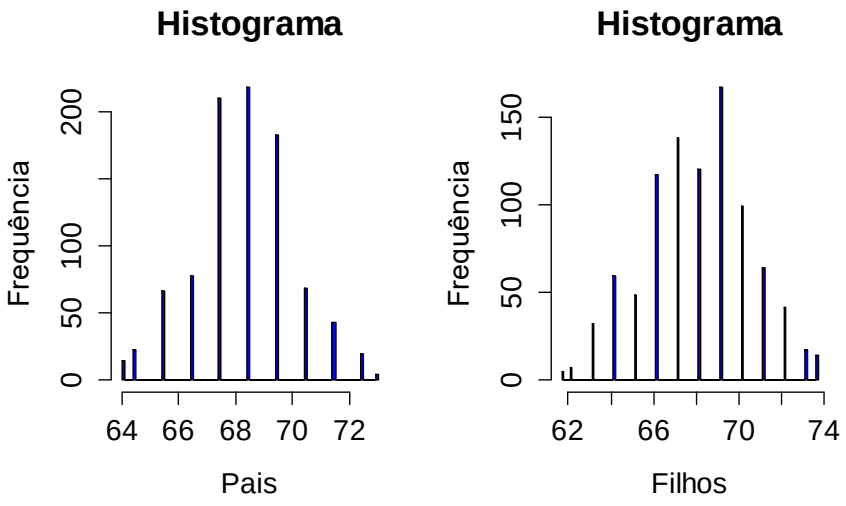


Figura 6.2 – Histograma da altura dos filhos e dos pais. ®

Pode-se observar na Figura 6.2 certa similaridade entre as distribuições, o que indica possível relação entre as variáveis. Mas como determinar um modelo que melhor represente a variação de uma em função da outra? Inicialmente devemos considerar qual a melhor representação de uma variável individualmente, ou seja, qual o ponto central que minimiza a soma dos erros quadráticos (visto na Equação (2.5)) do conjunto de valores da variável, ou qual o centro de massa físico do histograma. Matematicamente falando, queremos saber qual é o valor central ξ que minimiza a Equação (2.6), repetida a seguir.

$$SEQ = \sum_{i=1}^{N} (x_i - \xi)^2 \tag{6.2}$$

Do que foi exposto no Capítulo 2, sabemos que esse valor central é a média.

Exercício: execute o código do quadro a seguir e constate que a média minimiza a SEQ.

Quadro 6.2 – Soma dos erros quadráticos.

```
library(manipulate); library(UsingR); data(galton)
myHist <- function(mu){
    hist(galton$child,col="blue",breaks=100)
    lines(c(mu, mu), c(0, 150),col="red",lwd=5)
    mse <- mean((galton$child - mu)^2)
    text(63, 150, paste("mu = ", mu))
    text(63, 140, paste("SEQ = ", round(mse, 2)))
}
manipulate(myHist(mu), mu = slider(62, 74, step = 0.5))
mean(galton$child)
```

O desenvolvimento matemático a seguir confirma os resultados da simulação.

$$\begin{aligned}
\sum_{i=1}^{N} (y_i - \xi)^2 &= \sum_{i=1}^{n} (y_i - \bar{y} + \bar{y} - \xi)^2 \\
&= \sum_{i=1}^{n} (y_i - \bar{y})^2 + 2\sum_{i=1}^{n} (y_i - \bar{y})(\bar{y} - \xi) + \sum_{i=1}^{n} (\bar{y} - \xi)^2 \\
&= \sum_{i=1}^{n} (y_i - \bar{y})^2 + 2(\bar{y} - \xi)\sum_{i=1}^{n} (y_i - \bar{y}) + \sum_{i=1}^{n} (\bar{y} - \xi)^2 \\
&= \sum_{i=1}^{n} (y_i - \bar{y})^2 + 2(\bar{y} - \xi)(\sum_{i=1}^{n} y_i - n\bar{y}) + \sum_{i=1}^{n} (\bar{y} - \xi)^2 \\
&= \sum_{i=1}^{n} (y_i - \bar{y})^2 + \sum_{i=1}^{n} (\bar{y} - \xi)^2 \\
&\geq \sum_{i=1}^{n} (y_i - \bar{y})^2
\end{aligned}$$

$\sum_{i=1}^{n} y_i - n\bar{y} \rightarrow 0$

Logo, o valor que minimiza a soma dos erros quadráticos é o valor que minimiza o termo $\sum_{i=1}^{n}(\bar{y}-\xi)^2$, ou seja, $\xi = \bar{y}$. Agora, considere as duas variáveis *x* e *Y* conjuntamente, conforme a Figura 6.3.

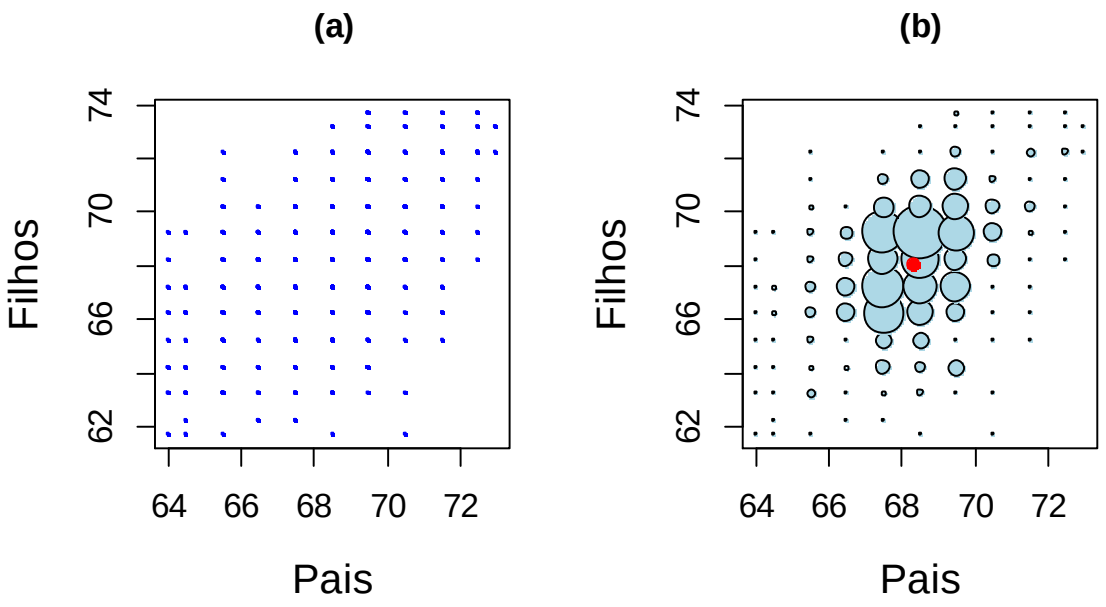


Figura 6.3 – Gráfico de dispersão das variáveis *x* e *Y*: a) vários pontos estão sobrepostos; b) o tamanho do ponto é proporcional à quantidade de pontos na coordenada; o ponto vermelho é o centro de massa. ®

O gráfico de dispersão (Figura 6.3a) provê uma boa visão do comportamento conjunto das variáveis. O desafio do moldelo de regressão linear é encontrar a reta que melhor representa esse comportamento. Iniciemos por determinar um primeiro ponto pelo qual essa reta passaria. Seguindo o mesmo raciocícnio utilizado no caso individual, pode-se determinar o centro de massa da combinação das variáveis aleatórias como o ponto (destacado em vermelho na Figura 6.3b) cujas coordenadas correspondem às médias individuais (ou centro de massa) de cada variável. Então, para completar a caracterização da reta que estamos procurando, precisamos conhecer um segundo ponto ótimo (aquele que intercepta o eixo *y*), ou, alternativamente, descobrir a inclinação que mais bem se ajusta ao conjunto. A inclinação de uma reta (β) é dada pela razão entre a variação ocorrida na dimensão *y* (Δy) e a correspondente variação em *x* (Δx), ou seja, $\beta = \Delta y / \Delta x$. Descobrir a melhor inclinação é equivalente a descobrir a reta que minimiza a soma das distâncias verticais (não ortogonais) dos pontos do conjunto à reta (Figura 6.4).

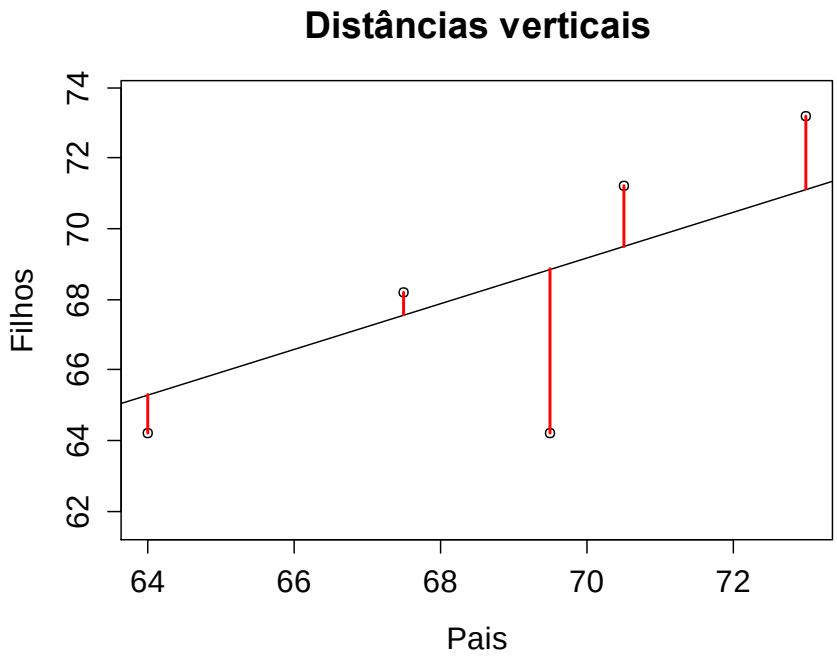


Figura 6.4 – Distâncias verticais dos pontos à reta. ®

Um recurso didático para se determinar a inclinação consiste em deslocar o centro de massa do conjunto para a origem, o que se faz subtraindo de cada variável aleatória a média dos seus valores. Esse processo é também chamado de centralização da variável aleatória. A Figura 6.5 ilustra esse resultado.

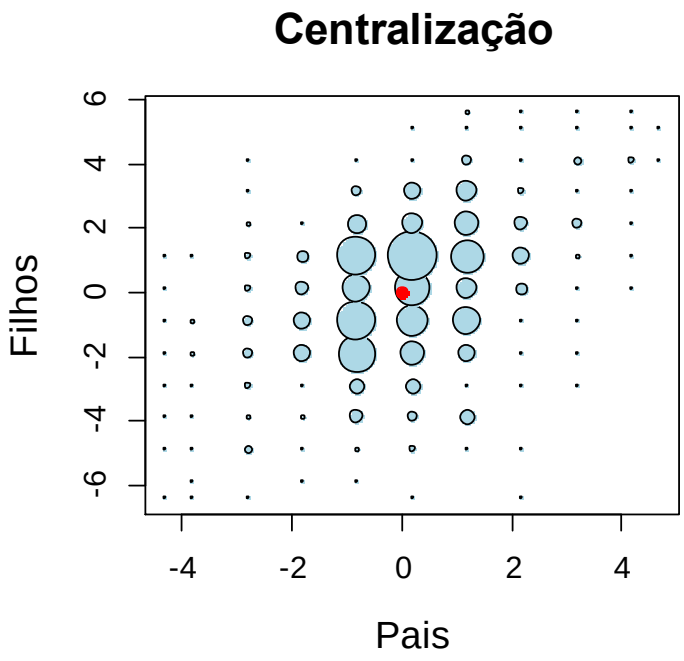


Figura 6.5 – Centro de massa deslocado para a origem. ®

A partir de então, utilizamos o critério dos mínimos quadrados considerando a linha que passa pela origem $y = \beta x$ para estimar β com base nas medições de Galton de x_i (alturas dos pais) e y_i (alturas dos filhos). Representaremos essa estimativa de β por $\hat{\beta}$. Desejamos descobrir a inclinação $\hat{\beta}$ que minimiza a soma das diferenças quadráticas (Equação (6.3)) em relação aos valores estimados $\hat{y}_i = \hat{\beta} x_i$.

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta} x_i)^2 \tag{6.3}$$

Neste ponto, vamos abstrair um pouco da matemática para apresentar a função **lm** (*linear model*), disponível em R, utilizada para a estimação de parâmetros de modelos lineares. O código a seguir exemplifica como determinar o valor da inclinição após o processo de centralização das variáveis. O termo "-1" da expressão instrui o R a suprimir o intercepto do eixo y, apresentando apenas o valor estimado da inclinação.

Quadro 6.3 – Função lm (*linear model*) com variáveis centralizadas.

```
lm(I(child - mean(child)) ~ I(parent - mean(parent)) -1, data = galton)

Call:
lm(formula = I(child - mean(child)) ~ I(parent - mean(parent)) -
    1, data = galton)

Coefficients:
I(parent - mean(parent))
                  0.6463
```

O mesmo resultado pode ser aferido executando-se o código do Quadro 6.4 e identificando a reta que minimiza a soma das distâncias verticais.

Quadro 6.4 – Inclinação da reta das variáveis centralizadas e SEQ.

```
library(UsingR); data(galton); library(manipulate)
myPlot <- function(beta){ y <- galton$child - mean(galton$child)
    x <- galton$parent - mean(galton$parent)
    freqData <- as.data.frame(table(x, y))
    names(freqData) <- c("parent", "child", "freq")
    plot(as.numeric(as.vector(freqData$parent)),
        as.numeric(as.vector(freqData$child)),
        pch = 21, col = "black", bg = "lightblue",
        cex = .15 * freqData$freq, xlab = "parent", ylab = "child")
    abline(0, beta, lwd = 3)
    points(0, 0, cex = 2, pch = 19)
    mse <- mean( (y - beta * x)^2 )
    title(paste("beta = ", beta, "SEQ = ", round(mse, 3)))}
manipulate(myPlot(beta), beta = slider(0.6, 1.2, step = 0.02))
```

Uma vez encontrada a inclinação, podemos retornar o centro de massa ao encontro das médias e determinar o intercepto do eixo *y* (Figura 6.6).

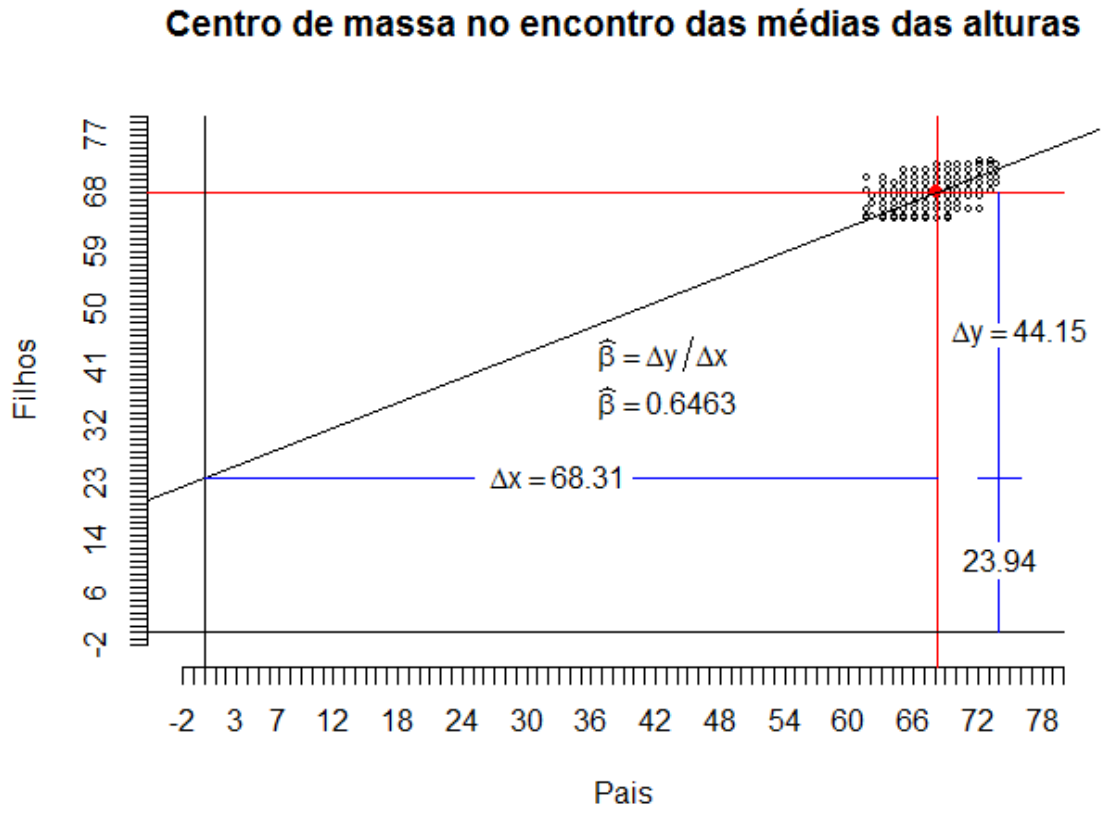


Figura 6.6 – Identificação do intercepto *y* dada a inclinação β da reta. ®

6.2.1 Formalização matemática da regressão linear simples

Conforme mencionado, o método apresentado nas figuras 6.5 e 6.6 tem cunho didático e apresenta uma forma visual de se estimar os parâmetros da reta que procurávamos. Vamos, agora, formalizar matematicamemente o processo de estimação.

Modelos, em geral, descrevem um fenômeno de forma simplificada e são utilizados com o intuito de se compreender o comportamento de uma população. Por serem uma simplificação, possuem um erro ε associado. Então, podemos dizer que a regressão linear é uma função μ de *x* que provê uma estimativa da variável aleatória *Y*, ou seja, $\mu(x) = \mathrm{E}Y$, de forma que

$$\mu(x) = \beta_0 + \beta_1 x \tag{6.4}$$

é a **função da população** e o erro representa a variação de *Y* que não é explicada pelo modelo e pode ser expresso por

$$\varepsilon = Y - \mu \text{ ou } \varepsilon_i = y_i - \mu_i \tag{6.5}$$

O desafio ao construir o modelo consiste em minimizar o erro total, ou a soma dos erros quadráticos *S*, de tal forma a encontrar β_0 e β_1 que trarão a menor diferença entre a previsão μ_i de y_i e o y_i realmente observado, ou seja, devemos minimizar a Equação (6.6).

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon^2 = \sum_{i=1}^{n} (y_i - \mu_i)^2 = \sum_{i=1}^{n} \{y_i - (\beta_0 + \beta_1 x_i)\}^2 \tag{6.6}$$

Contudo, em geral, o recurso de que dispomos para ajustar o modelo é apenas uma amostra da população, ou seja, um conjunto de dados coletados (*training set* – dados de treinamento). Na verdade, não conhecemos os parâmetros β_0 e β_1 que se ajustam à população e tudo o que podemos fazer é estimá-los com base nesses dados coletados. Quando fazemos novas coletas referentes ao mesmo fenômeno investigado, é bastante provável que os parâmetros sofram ajustes. Então, como estamos nos referindo a estimativas dos parâmetros, é conveniente alterarmos a notação para

$$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{6.7}$$

que é a **função da amostra**, onde *i* indica cada uma das *n* observações da base de dados. No caso da amostra, o erro ε passa a ter a notação *e*, que denominamos resíduo.

$$e_i = y_i - \hat{\mu}_i \tag{6.8}$$

Devemos, então, encontrar as estimativas a partir do mínimo da soma dos quadrados[16] dos resíduos $\sum_{i=1}^{n} e_i^2$, onde

[16] Estimadores de mínimos quadrados são utilizados por serem de fácil obtenção. Por essa razão recebem o nome de estimadores de Mínimos Quadrados Ordinários. Além disso, quando os resíduos são elevados ao

$$S(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} \{y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\}^2 \tag{6.9}$$

O método para encontrar os valores mínimos de uma função consiste em igualar a zero as dervivadas parciais em relação aos parêmetros estimados. Iniciando por $\hat{\beta}_0$, temos

$$\frac{\partial S(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \tag{6.10}$$

Igualando a zero e desenvolvendo a equação, temos

$$-2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\beta}_0 - \sum_{i=1}^{n} \hat{\beta}_1 x_i = 0$$

$$n\hat{\beta}_0 = \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i$$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^{n} x_i}{n}$$

e finalmente

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{6.11}$$

Resolvendo para $\hat{\beta}_1$, temos

$$\frac{\partial S(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_1} = 2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i) \tag{6.12}$$

Substituindo a Equação (6.11) em (6.12), igualando a zero e desenvolvendo, temos

---

quadrado, os resíduos maiores são proporcionalmente mais penalizados que os menores: se um resíduo tem o dobro do tamanho de outro, o seu quadrado é quatro vezes maior.

$$-2\sum_{i=1}^{n}(y_i - \bar{y} + \hat{\beta}_1\bar{x} - \hat{\beta}_1 x_i)(x_i) = 0$$

$$\sum_{i=1}^{n}\{x_i(y_i - \bar{y}) - \hat{\beta}_1 x_i(x_i - \bar{x})\} = 0$$

que resulta em

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sum_{i=1}^{n} x_i(x_i - \bar{x})} \tag{6.13}$$

Neste ponto, vamos utilizar um artifício que nos levará a um resultado mais significativo para $\hat{\beta}_1$. Sabemos, do capítulo 2, que a soma das diferenças dos elementos em relação à média é sempre zero. Dessa forma, podemos subtrair do numerador e do denominador da Equação (6.13), respectivamente, os termos $\bar{x}\sum_{i=1}^{n}(y_i - \bar{y})$ e $\bar{x}\sum_{i=1}^{n}(x_i - \bar{x})$. Desenvolvendo um pouco mais, temos

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y}) - \bar{x}\sum_{i=1}^{n}(y_i - \bar{y})}{\sum_{i=1}^{n} x_i(x_i - \bar{x}) - \bar{x}\sum_{i=1}^{n}(x_i - \bar{x})} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})} \tag{6.14}$$

Divindo o numerador e o denominador da Equação (6.14) por *n*-1, e comparando com a Equação (6.2), temos

$$\hat{\beta}_1 = \frac{\frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{n-1}}{\frac{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})}{n-1}} = \frac{r * s_x * s_y}{s_x^2}$$

e finalmente

$$\hat{\beta}_1 = \frac{r * s_y}{s_x} \tag{6.15}$$

onde *r* é o coeficiente de correlação de Pearson. Ou seja, a inclinação de uma regressão linear é a correlação empírica entre as duas variáveis, multiplicada pela razão entre o desvio padrão de *y* e o desvio padrão de *x*.

Algumas conclusões podem ser extraídas a partir das equações (6.11) e (6.15):

- a regressão linear sempre passa pelo encontro das médias $(\bar{x}, \bar{y})$;
- a unidade de medida de $\hat{\beta}_1$ é a mesma da relação y/x;
- a inclinação da reta de regressão é a mesma que se obteria se os dados fossem centrados $(x_i - \bar{x}, y_i - \bar{y})$ e a regressão fosse efetuada em relação à origem;
- se os dados são normalizados, $(\frac{x_i - \bar{x}}{s_x}, \frac{y_i - \bar{y}}{s_y})$, a inclinação da reta de regressão é igual a *r*.

6.2.2 Modelo de regressão linear simples com adição de erro Gaussiano

Na seção anterior, estudamos o método dos mínimos quadrados como uma ferramenta matemática para estimar os parâmetros da reta de regressão linear. No entanto, ainda não o formalizamos como um modelo estatístico, conforme definimos na seção 5.1. Então, vamos definir um modelo probabilístico para a regressão linear, de modo a sustentar inferências estatísticas a partir desse modelo. Reescrevendo a Equação (6.5), temos

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{6.16}$$

onde assumimos que os erros $\varepsilon_1, \varepsilon_2, \varepsilon_3$ ... são independentes e identicamente distribuídos, com distribuição normal, média zero e variância σ^2: $\varepsilon_i \sim N(0, \sigma^2)$.

O modelo de regressão linear simples descreve a variável *Y* como a soma de uma quantidade determinística e outra aleatória. A parte determinística é a reta em função de *x*, que representa a informação esperada de *Y* uma vez conhecida a variável *x*. Em notação matemática, o valor esperado de *Y*, dados os valores de *x*, é escrito como

$$
\begin{aligned}
E[Y \mid x = x_i] &= E[\beta_0 + \beta_1 x_i + \varepsilon_i] \\
&= E[\beta_0 + \beta_1 x_i] + E[\varepsilon_i] \\
&= \beta_0 + \beta_1 x_i + 0 \\
&= \beta_0 + \beta_1 x_i \\
&= \mu_i
\end{aligned}
$$

A parte aleatória, o erro, representa os demais fatores, não previstos, que podem interferir em *Y*. Assim, assumimos que a variância do erro não depende dos valores específicos de *x*.

$$
\begin{aligned}
Var[Y \mid x = x_i] &= Var[\beta_0 + \beta_1 x_i + \varepsilon_i] \\
&= Var[\beta_0] + Var[\beta_1 x_i] + Var[\varepsilon_i] \\
&= 0 + 0 + \sigma^2 \\
&= \sigma^2
\end{aligned}
$$

Sob as assunções de independência e normalidade do erro, o método dos mínimos quadrados é equivalente à estimativa por **máxima verossimilhança**[17]. Nesse sentido, podemos definir nosso modelo estatístico como $y_i \sim N(\mu_i, \sigma^2)$. Note que a variância é a mesma (constante) para todo *i*, ao que denominamos **homocedasticidade**.

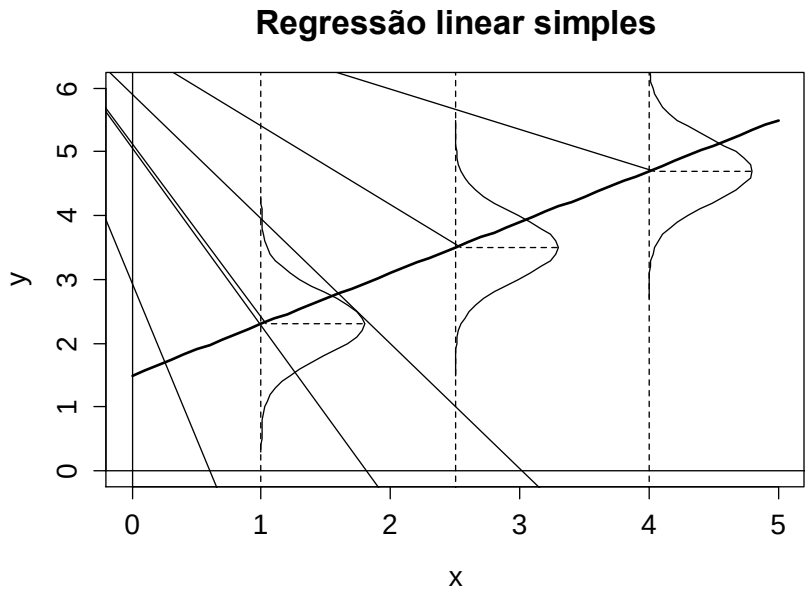


Figura 6.7 – Modelo estatístico da regressão linear simples: $y_i \sim N(\mu_i, \sigma^2)$. ®

[17] A estimativa por máxima verossimilhança (*maximum-likelihood estimation* – MLE) é um método para estimar os parâmetros de um modelo estatístico. De maneira geral, posto um conjunto de dados e um modelo estatístico, o método de máxima verossimilhança estima os valores dos diferentes parâmetros do modelo de maneira a maximizar a probabilidade dos dados observados.
http://www.portalaction.com.br/533-2-estimadores-de-m%C3%A1xima-verossimilhan%C3%A7

6.2.3 Interpretação dos coeficientes de regressão

A grande importância e utilidade dos modelos lineares reside no fato de que são amplamente interpretáveis. No caso da regressão linear (Equação (6.4)), compreender o modelo consiste em entender o significado dos coeficientes β_0 e β_1.

Iniciando por β_0, também chamado de intercepto do eixo *y*, devemos interpretá-lo como o valor esperado quando o preditor *x* é zero.

$$E[\mu(x) \mid x = 0] = E[\beta_0 + \beta_1 \times 0] = \beta_0 \tag{6.17}$$

Contudo, nem sempre esse resultado é de interesse, ou tem algum significado. Há situações em que $x = 0$ é impossível ou está fora do domínio dos dados. Por exemplo, se estivermos analisando dados referentes à altura de pessoas (como no estudo de Galton), o zero não faz sentido e nem a interpretação de β_0. Algumas vezes pode ser útil efetuar o deslocamento da variável *x* com vistas a prover uma interpretação mais apropriada de β_0. Considere a Equação (6.16). Se adicionarmos e subtrairmos o termo $a\beta_1$, o modelo não se altera.

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i = \\ &= \beta_0 + \beta_1 x_i - \beta_1 a + \beta_1 a + \varepsilon_i \\ &= (\beta_0 + \beta_1 a) + \beta_1 (x_i - a) + \varepsilon_i \\ &= \tilde{\beta}_0 + \beta_1 (x_i - a) + \varepsilon_i \end{aligned}$$

O deslocamento dos valores de *x* por uma constante *a* modifica o valor do intercepto do eixo *y*, que passa a ser $\tilde{\beta}_0 = \beta_0 + \beta_1 a$, mas não altera o valor da inclinação. Uma possibilidade é fazer $a = min(x)$, o que significa deslocar o zero para o menor valor possível da variável *x*. Outra é fazer $a = \bar{x}$, de forma que o intercepto seja interpretado como o valor esperado de *Y* para o valor médio de *x*. Esse recurso deve ser utilizado à conveniência da análise.

A inclinação β_1 deve ser interpretada como a mudança esperada em *Y* para cada unidade de mudança da variável *x*. Em outras palavras, para $\beta_1 > 0$, o aumento de uma unidade na variável *x* corresponde ao aumento de β_1 unidades na variável *Y*. Para $\beta_1 < 0$, o aumento de

uma unidade na variável x corresponde à diminuição de β_1 unidades na variável Y. Matematicamente, temos

$$\begin{aligned} E[\mu(x) \mid x = x+1] - E[\mu(x) \mid x = x] &= \beta_0 + \beta_1(x+1) - (\beta_0 + \beta_1 x) \\ &= \beta_0 + \beta_1 x + \beta_1 - \beta_0 - \beta_1 x \\ &= \beta_1 \end{aligned} \tag{6.18}$$

Considere, agora, o impacto de alterarmos a unidade de medida da variável x. Se multiplicarmos e dividirmos x por um valor a, o modelo não se altera.

$$\begin{aligned} y_i = \beta_0 + \beta_1 x_i + \varepsilon_i &= \beta_0 + \frac{\beta_1}{a}(x_i a) + \varepsilon_i \\ &= \beta_0 + \tilde{\beta}_1(x_i a) + \varepsilon_i \end{aligned}$$

Como consequência, a multiplicação de x por um fator a resulta na divisão do coeficiente β_1 pelo mesmo fator a, ou, a divisão de x por um fator a resulta na multiplicação do coeficiente β_1 pelo mesmo fator a.

Com base no exposto, vamos exercitar a aplicação da regressão linear e a interpretação dos seus coeficientes. Considere a situação a seguir:

> Numa porção da floresta Amazônica onde pode estar uma importante jazida de diamantes, índios e garimpeiros refizeram uma lucrativa parceria para extrair e vender as pedras de maneira ilegal. A atividade foi retomada no fim de 2013 na Terra Indígena Roosevelt, uma área que se estende por Rondônia e Mato Grosso. As suspeitas são de que as pedras de Roosevelt acabem chegando às mãos de compradores em Singapura, Bélgica e outros centros de lapidação e comércio de diamantes. É uma longa cadeia ilícita, da qual, em geral, participam doleiros, contrabandistas, empresas de fachada e, por vezes, agentes da lei. O Brasil é participante do Sistema de Certificação do Processo Kimberley, que regulamenta, com a chancela da ONU, o comércio internacional dos diamantes brutos e exige de seus signatários medidas para garantir que suas pedras sejam extraídas somente de áreas legalizadas. Diamantes brutos só podem sair do país com

certificado Kimberley, emitidos pelo Departamento Nacional de Produção Mineral (DNPM). Se forem de áreas não legalizadas, não são, em tese, certificados. Em outubro de 2013, o Senado Federal realizou audiência pública para levar os problemas dos índios Cintas Largas ao Governo Federal e discutir projetos destinados a proporcionar alternativas econômicas e sociais àquela comunidade[18].

Diante disso, suponha que tenhamos sido designados pela Comissão de Minas e Energia para desenvolver estudo sobre o preço de diamantes no mercado internacional, com vistas a propor políticas direcionadas à comunidade dos Cintas Largas. Para esse fim, foi-nos disponibilizada a base de dados *diamond*, da biblioteca "UsingR", que contém os preços praticados em Singapura relativos à massa de pedras utilizadas no comércio de aneis. Nossa missão é apresentar um modelo de regressão linear que ajude a Comissão a fazer predições de preços para uma amostra de pedras extraída de Roosevelt.

Primeiramente, devemos conhecer as informações sobre a base de dados que vamos analisar (http://www.amstat.org/publications/jse/datasets/diamond.txt). Avaliando a documentação, verificamos que os preços são cotados em dólares singapurenses e a unidade de massa é o quilate (ou *carat* – em inglês – que é a medida padrão da massa de diamantes, correspondente a 0.2g). O código a seguir produz a reta de regressão linear ilustrada na Figura 6.8.

Quadro 6.5 – Regressão linear para os dados de diamantes.

```
library(UsingR); data(diamond)
plot(diamond$carat, diamond$price,
     xlab = "Massa (carats)", ylab = "Preço (SIN $)",
     main = "Preço do diamante em função da massa",
     bg = "lightblue",
     col = "black", cex = 1.1, pch = 21, frame = FALSE)
fit <- lm(price ~ carat, data = diamond)
abline(fit, lwd = 2)
fit

Call:
lm(formula = price ~ carat, data = diamond)

Coefficients:
(Intercept)        carat
     -259.6       3721.0
```

[18] Adaptação de http://www.rondoniadinamica.com/arquivo/rondonia-indios-e-garimpeiros-reabrem-garimpo-de-diamantes-na-reserva- roosevelt,66408.shtml
e de http://www.kaninde.org.br/?pag_id=19&p=2351&offset=128

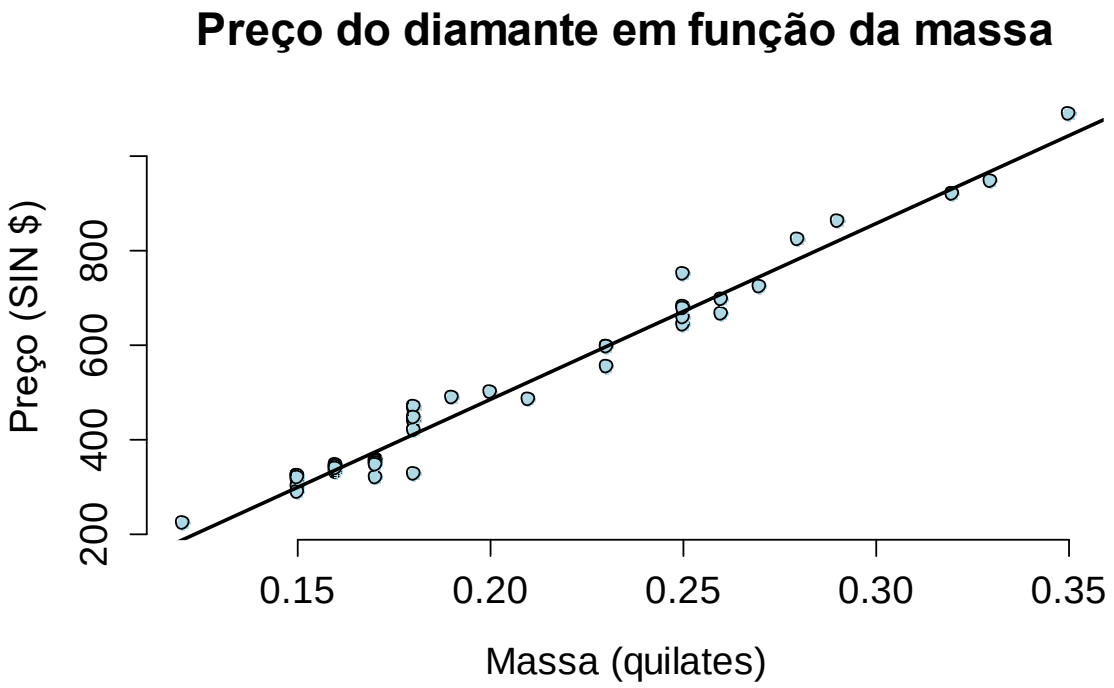


Figura 6.8 – Regressão linear: preço do diamante em função da massa. ®

A função lm, no Quadro 6.5, estima os coeficientes da reta de regressão da variável *price* em função da variável *carat*. Os resultados mostram que é esperado um acréscimo de 3721,02 dólares singapurenses para cada acréscimo de massa de 1 quilate. O intercepto estimado em -259,63 é o valor esperado para o preço de um diamante de 0 quilate, o que não faz nenhum sentido. Portanto, é interessante obtermos um intercepto cuja interpretação tenha um significado mais apropriado. No código a seguir, os valores da variável *carat* são deslocados por um fator igual à média.

Quadro 6.6 – Deslocamento da variável *carat* por um fator igual à média.

```
fit <- lm(price ~ I(carat - mean(carat)), data = diamond)
coef(fit)

           (Intercept) I(carat - mean(carat))
              500.0833              3721.0249
```

A função *I* em lm signfica que a operação dentro dos parêntesis deve ser efetuada antes do ajuste do modelo linear. O novo resultado diz que 500,1 dólares singapurenses é o preço esperado para um diamante que tenha massa igual ao valor médio das massas (0.2042 quilates).

A unidade quilate nem sempre é de fácil entendimento para o leitor não especializado. Poderia ser mais útil à Comissão se utilizássemos gramas como unidade de massa. Isso requer que seja feito o escalonamento da variável *carat*, ou seja, devemos multiplicar *carat* por 0,2 a fim de termos a massa em gramas. Em contrapartida, para que não se altere o modelo, a inclinação deve ser dividida pelo mesmo valor.

$$
\begin{aligned}
price_i &= \beta_0 + \frac{\beta_1}{a}(carat_i \times a) + \varepsilon_i \\
&= -259.6 + \frac{3721.0}{0.2}(carat_i \times 0.2) + \varepsilon_i \\
&= -259.6 + 18605.1(carat_i \times 0.2) + \varepsilon_i
\end{aligned}
$$

Quadro 6.7 – Escalonamento da variável *carat*.

```
fit <- lm(price ~ I(carat * 0.2), data = diamond)
coef(fit)

   (Intercept) I(carat * 0.2)
     -259.6259     18605.1243
```

Quadro 6.8 – Deslocamento e escalonamento combinados.

```
fit <- lm(price ~ I((carat - mean(carat)) * 0.2), data = diamond)
coef(fit)

                    (Intercept) I((carat - mean(carat)) * 0.2)
                       500.0833                     18605.1243
```

Agora que dispomos do modelo ajustado, podemos predizer preços, por exemplo, para diamantes com os seguintes pesos: 0,50g, 1,20g e 0,15g.

Quadro 6.9 – Predição de preços por meio do modelo de regressão.

```
fit <- lm(price ~ I((carat - mean(carat)) * 0.2), data = diamond)
diam <- c(0.50, 1.20, 0.15)
coef(fit)[1]+coef(fit)[2]*diam

9802.645  22826.232  3290.852
```

6.2.4 Resíduos e variações residuais

Conforme discutido na seção 6.2.1, a todo modelo está associado um erro ε que representa a variação não explicada pelo modelo. O erro ε é o erro real, que não conhecemos porque nem mesmo sabemos se o nosso modelo é correto, visto que não atuamos sobre a população e sim sobre uma amostra. O que podemos fazer é tão somente estimar os parâmetros do modelo e verificar o resíduo *e*, que consiste na distância entre o valor amostrado e o valor obtido pela linha de regressão (valor ajustado). Então, podemos considerar que e_i é uma estimativa de ε_i e que o método dos Mínimos Quadrados Ordinários minimiza e_i. As implicações algébricas das estimativas são derivadas exclusivamente da aplicação do método ao modelo de regressão linear.

A análise dos resíduos tem importância fundamental na verificação da qualidade do modelo e fornece evidências sobre possíveis violações nas suposições de independência, normalidade, homocedasticidade e linearidade[19]. Por isso é importante conhecermos as propriedades dos resíduos.

P1. O valor esperado para um resíduo e_i é zero e pode ser obtido a partir das equações (6.7) e (6.8)

$$\begin{aligned} E[e_i] &= E[y_i - \hat{\mu}_i] \\ &= E[y_i] - E[\hat{\mu}_i] \\ &= E[y_i] - E[\hat{\beta}_0 + \hat{\beta}_1 x_i] \\ &= (\beta_0 + \beta_1 x_i) - (\beta_0 + \beta_1 x_i) \\ &= 0 \end{aligned}$$

visto que $\hat{\beta}_0$ e $\hat{\beta}_1$ são estimadores robustos de β_0 e β_1, respectivamente. Alguns valores de *Y* estão acima e outros abaixo da reta (resíduos positivos e negativos).

P2. A soma dos resíduos é zero quando existir um intercepto.

[19] Se os resíduos são independentes entre si, então não podem ser correlacionados entre si e não são correlacionados com *x* e *Y*. Se são identicamente distribuídos com $N(0,\sigma^2)$, então foram gerados a partir da mesma normal. Mesmo que a relação entre *x* e *Y* não seja linear, eles podem estar correlacionados. Nesses casos, a correlação pode ser representada por uma curva quadrática, cúbica (*spline*) etc.

$$\begin{aligned}\sum_{i=1}^{n} e_i &= \sum_{i=1}^{n}(y_i - \widehat{\mu}_i) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\widehat{\beta}_0 - \sum_{i=1}^{n}\widehat{\beta}_1 x_i \\ &= \sum_{i=1}^{n} y_i - n\widehat{\beta}_0 - \widehat{\beta}_1 \sum_{i=1}^{n} x_i \\ &= n\bar{y} - n\widehat{\beta}_0 - \widehat{\beta}_1 n\bar{x} \\ &= n(\bar{y} - \widehat{\beta}_1 \bar{x}) - n\widehat{\beta}_0 \\ &= n(\widehat{\beta}_0) - n\widehat{\beta}_0 \\ &= 0\end{aligned}$$

Quando não existir um intercepto ($\widehat{\beta}_0 = 0$), ou seja, quando tivermos a regressão pela origem, a soma, em geral, não será zero.

$$\begin{aligned}\sum_{i=1}^{n} e_i &= \sum_{i=1}^{n}(y_i - \widehat{\mu}_i) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\widehat{\beta}_1 x_i \\ &= \sum_{i=1}^{n} y_i - \widehat{\beta}_1 \sum_{i=1}^{n} x_i \\ &= n\bar{y} - \widehat{\beta}_1 n\bar{x} \\ &= n(\bar{y} - \widehat{\beta}_1 \bar{x})\end{aligned}$$

P3. A média dos resíduos, quando existir um intercepto, é zero: $\bar{e}_i = \frac{1}{n}\sum_{i=1}^{n} e_i = 0$.

P4. A soma dos valores observados y_i é igual a soma dos valores ajustados $\widehat{\mu}_i$, quando existir um intercepto (ver P2): $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{\mu}_i$.

P5. A média dos valores observados é igual à média dos valores ajustados (dividir ambos os membros de P4 por *n*): $\bar{y} = \bar{\widehat{\mu}}$.

P6. A soma dos resíduos ponderada pelo correspondente valor da variável regressora é sempre zero: $\sum_{i=1}^{n} x_i e_i = 0$. Desenvolvendo, temos

$$\sum_{i=1}^{n} x_i e_i = \sum_{i=1}^{n} x_i (y_i - \widehat{\mu}_i) = \sum_{i=1}^{n} x_i (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)$$

Esse resultado é equivalente ao da Equação (6.12) quando igualada à zero, o que é condição necessária para a obtenção dos mínimos quadrados e, portanto, prova de P6.

P7. A soma dos resíduos ponderada pelo correspondente valor ajustado é sempre zero: $\sum_{i=1}^{n} \hat{\mu}_i e_i = 0$. Desenvolvendo, temos

$$\sum_{i=1}^{n} \hat{\mu}_i e_i = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i) e_i = \hat{\beta}_0 \sum_{i=1}^{n} e_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i e_i = \hat{\beta}_0 \times 0 + \hat{\beta}_1 \times 0 = 0$$

Resíduos podem ser entendidos como o resultado *Y* obtido após a remoção da associação linear com *x*. A variação residual (variação observada após a remoção do preditor) é diferente da variação sistemática (variação explicada pelo modelo de regressão linear). Gráficos de resíduos são bastante úteis na análise dos pressupostos da regressão linear, pois revelam quão bem o modelo se ajusta aos dados coletados, o que nem sempre é possível de se observar em gráficos de dispersão, especialmente quando são adicionadas outras variáveis aos modelos de regressão, situação em que não há recurso equivalente ao gráfico de dispersão. A Figura 6.9 ilustra o gráfico de resíduos dos dados de diamantes.

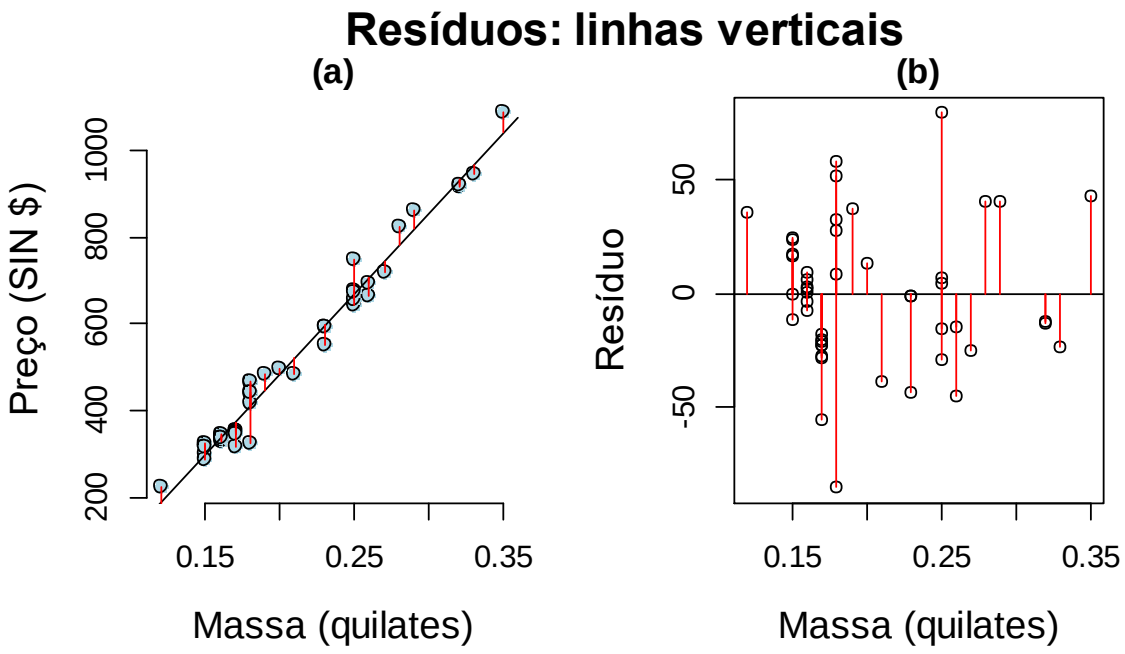


Figura 6.9 – Resíduos: (a) gráfico de dispersão; (b) resíduos *vs*. x (massa). ®

O gráfico de resíduos em função da variável *x* é importante para a identificação de padrões na variação dos resíduos. Quando padrões são encontrados, a suposição de normalidade é quebrada. Muitas vezes o gráfico de dispersão nos faz acreditar que o modelo se ajusta de

forma apropriada ao conjunto de dados. Contudo, variações residuais podem não ser perceptíveis em gráficos de dispersão.

Por suposição, os resíduos são aleatórios com distribuição normal. Quando padrões de comportamento são encontrados nos resíduos, quer dizer que uma variação sistemática não foi captada pelo modelo linear e está sendo inapropriadamente tratada como resíduo.
Considere, por exemplo, que uma variável preditora *x* representa um conjunto de dados com distribuição uniforme, e que a variável aleatória *Y* corresponda à variável *x* acrescida de uma variação senoidal, sistemática, e de um erro aleatório com distribuição normal e média zero (Quadro 6.10).

Quadro 6.10 – Preditor com distribuição uniforme e resultado com variação senoidal.

```
x <- runif(100, -3, 3)
y <- x + sin(x) + rnorm(100, sd = .2)
```

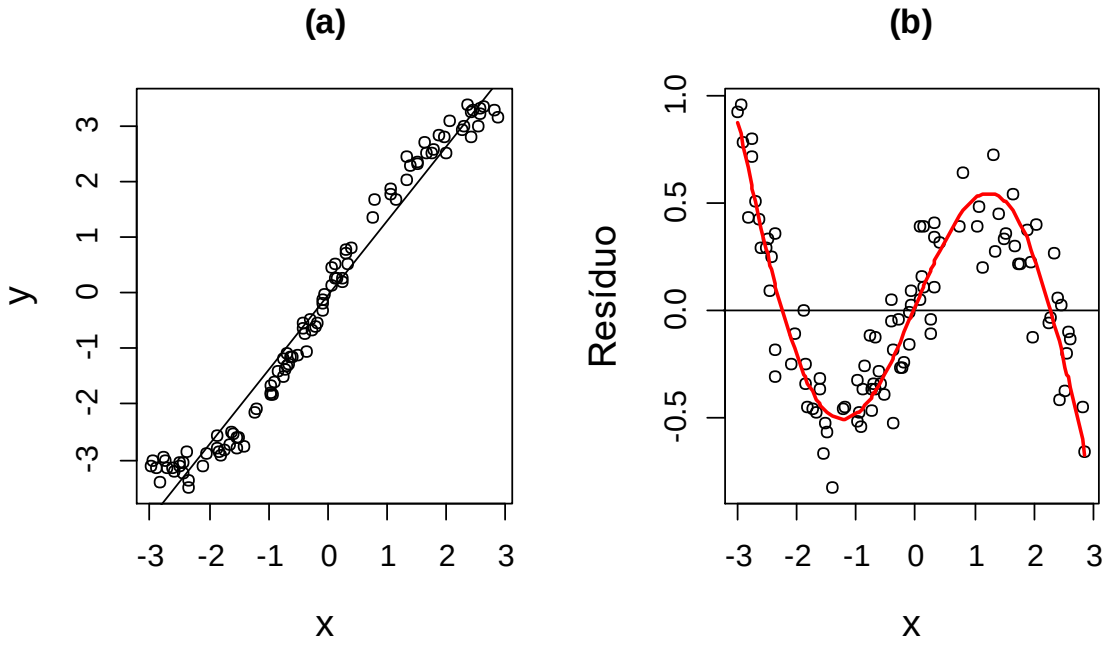


Figura 6.10 – Variação sistemática verificada no gráfico do resíduo *vs.* x. ®

Em uma primeira inspeção do gráfico de dispersão (Figura 6.10a), o modelo de regressão parece se ajustar aos dados de forma satisfatória. Contudo, ao se analisar o gráfico dos

ruídos (Figura 6.10b), percebe-se claramente um padrão senoidal, derrubando assim a suposição da normalidade.

Gráficos de resíduos também são úteis para a verificação da homocedasticidade. No exemplo a seguir, a variável preditora *x* tem distribuição uniforme e a variável aleatória *Y* corresponde à variável *x* acrescida de um erro aleatório com distribuição normal, contudo, o desvio padrão não é constante, mas dependente de *x* (Quadro 6.11).

Quadro 6.11 – Preditor com distribuição uniforme e resultado com desvio padrão dependente do preditor.

```
x <- runif(100, 0, 6)
y <- x + rnorm(100, mean = 0, sd = .001 * x)
```

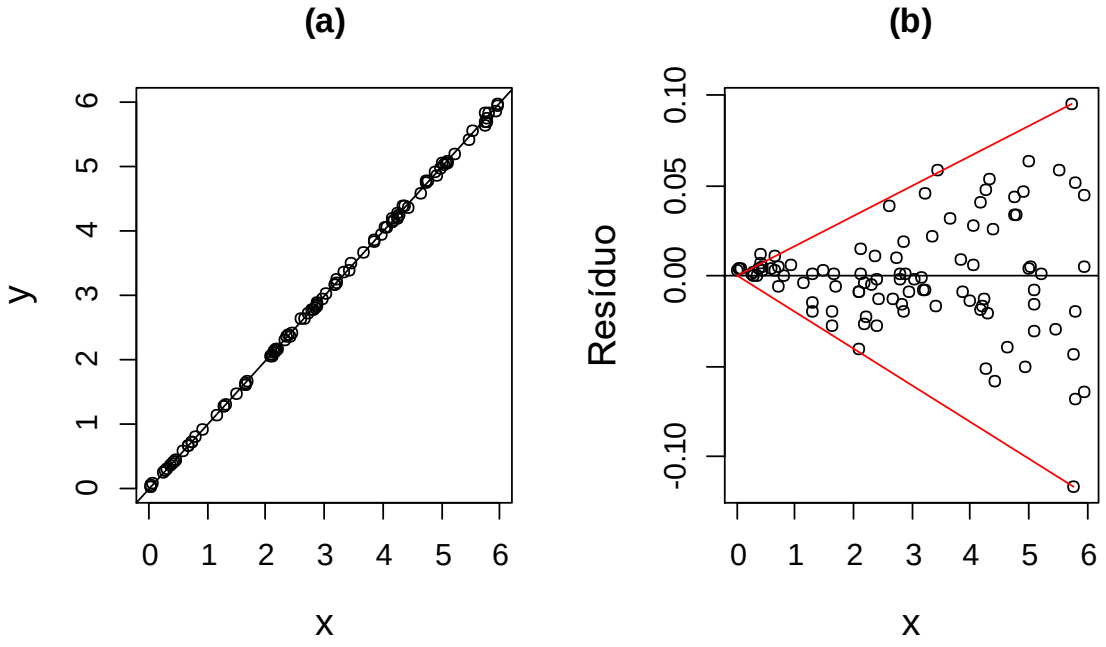


Figura 6.11 – Heterocedasticidade verificada no gráfico do resíduo *vs*. x. ®

A Figura 6.11a nos dá a falsa noção de que o modelo se ajusta perfeitamente aos dados. Contudo, a variância é crescente (Figura 6.11b), dependente de *x*, e o modelo não seria aprorpiado a predições quando *x* fosse muito maior do que 6.

Portanto, para que um modelo linear seja devidamente ajustado a um conjunto de dados, não basta fazermos estimativas da parte determinística (parâmetros β_0 e β_1) do modelo,

mas, também, precisamos estimar a parte aleatória, ou seja, a variância σ^2 dos termos do erro ε_i. Isso é necessário já que inferências a respeito da função de regressão e da predição de *Y* requerem uma estimativa de σ^2.

A estimativa por máxima verossimilhança de σ^2 é $\frac{1}{n}\sum_{i=1}^{n} e_i^2$, ou seja, a média dos resíduos quadráticos. Contudo, para a regressão linear, essa é uma **estimativa viciada**[20]. Para torná-la uma estimativa não viciada, devemos subtrair duas unidades (dois graus de liberdade) do denominador *n*: um referente ao intercepto e outro referente à inclinação. Assim, para

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 \tag{6.19}$$

temos $E[\hat{\sigma}^2] = \sigma^2$. O quadro a seguir contém código R que estima a variância do resíduo, conforme a Equação (6.19), utilizando o estudo dos diamantes.

Quadro 6.12 – Estimativa da variância do resíduo.

```
library(UsingR); data(diamonds)
y <- diamond$price; x <- diamond$carat; n <- length(y)
fit <- lm(y ~ x)
summary(fit)$sigma ^2

[1] 1013.819

sum(resid(fit)^2) / (n - 2) # equação 6.19

[1] 1013.819
```

6.2.5 Coeficiente de determinação R^2

Se desconsiderássemos a dependência que *Y* tem de *x* e baseássemos nossas estimativas apenas nos valores de *Y*, então, como sabemos, o valor central que minimiza os erros

[20]Um estimador *T* é dito não viciado (não enviesado) para algum parâmetro populacional θ se $\mathrm{E}T = \theta$ para todo θ. Se essa igualdade não ocorre, dizemos que *T* é um estimador viciado (enviesado) e a diferença $V(T,\theta) = \mathrm{E}T - \theta$ é chamada de vício (viés) de *T*.
http://www.portalaction.com.br/1447-31-propriedades-dos-estimadores

quadráticos é a média $\bar{y}$. Assim, os erros decorrentes das estimativas seriam dados por $y_i - \bar{y}$ e a variação medida pela variância das amostras (Equação (2.8)) seria proporcional ao termo

$$VT = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{6.20}$$

que denominaremos, por convenção, de **variação total** de *Y*.

Contudo, sabemos que os valores y_i de *Y* estão associados a cada valor x_i do preditor, e que o modelo linear nos permite estimar essa associação por $\hat{\mu}_i$. A partir da Equação (6.8) e da propriedade P5, podemos escrever

$$\begin{aligned}\sum_{i=1}^{n} (y_i - \bar{y})^2 &= \sum_{i=1}^{n} \{(e_i + \hat{\mu}_i) - \bar{y}\}^2 = \sum_{i=1}^{n} \{(\hat{\mu}_i - \bar{y}) + e_i\}^2 \\ &= \sum_{i=1}^{n} (\hat{\mu}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2 + 2\sum_{i=1}^{n} e\ (\hat{\mu}_i - \bar{y})\end{aligned}$$

A partir das propriedades P2, P5 e P7, o terceiro termo acima é igual a zero, isto é,

$$\sum_{i=1}^{n} e\ (\hat{\mu}_i - \bar{y}) = \sum_{i=1}^{n} \hat{\mu}_i e_i - \bar{y}\sum_{i=1}^{n} e_i = 0$$

Consequentemente,

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{\mu}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2 \tag{6.21}$$

Denominaremos o terceiro termo da Equação (6.21) por **variação residual**, ou seja, a variação explicada pelos resíduos, que pode ser expressa por

$$VRE = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 \tag{6.22}$$

O Segundo termo da Equação (6.21) é a **variação explicada pela reta de regressão**, que consiste na diferença entre o valor estimado pelo modelo e o valor quando desconsideramos a influência do preditor.

$$VRR = \sum_{i=1}^{n} (\hat{\mu}_i - \bar{y})^2 \tag{6.23}$$

Reescrevendo a Equação (6.21),

$$VT = VRR + VRE \tag{6.24}$$

Quando iniciamos a construção do modelo linear, fizemos estimativas de forma a minimizar a soma dos quadrados dos resíduos, e esse critério foi garantido matematicamente. Agora, precisamos saber quão bem o modelo se ajusta aos dados por meio de medições epscíficas, denominadas **qualidade de ajuste**. Uma dessas medidas é o **coeficiente de determinação** (***R²***), definido pela relação entre *VRR* e *VT*

$$R^2 = \frac{VRR}{VT} \tag{6.25}$$

ou

$$R^2 = \frac{VT - VRE}{VT} = 1 - \frac{VRE}{VT} \tag{6.26}$$

onde $0 \leq R^2 \leq 1$.

De forma interpretativa, R^2 é a percentagem de variação total explicada pelo modelo de regressão linear. Podemos dizer que quando R^2 se aproxima de 1, o modelo linear se ajusta fortemente ao comportamento dos dados, ou seja, que a variação explicada pela regressão linear se aproxima da variação total (ver Equação (6.25)). Quando R^2 tende a 0, significa dizer que o modelo linear não se ajusta ou fracamente se ajusta aos dados e a variação total tende à variação residual. Em outras palavras, muito pouco da variação de *Y* é explicada pela variação de *x*.

O coeficiente de determinação é estreitamente ligado ao coeficiente de correlção de Pearson. Da geometria da reta de regressão linear, sabemos que

$$\hat{\beta}_1 = \frac{\hat{\mu}_i - \bar{y}}{x_i - \bar{x}} \text{ ou } \hat{\mu}_i - \bar{y} = \hat{\beta}_1 (x_i - \bar{x}) \tag{6.27}$$

Elevando ambos os membros da Equação (6.27) ao quadrado, temos,

$$(\hat{\mu}_i - \bar{y})^2 = \hat{\beta}_1^2 (x_i - \bar{x})^2 \tag{6.28}$$

Aplicando o somatório para todo *i*, temos

$$\sum_{i=1}^{n} (\hat{\mu}_i - \bar{y})^2 = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{6.29}$$

Substituindo a Equação (6.29) em (6.25), e depois, utilizando (6.15)

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{\mu}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{\hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{r^2 \times s_y^2}{s_x^2} \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{r^2 \times s_y^2}{s_x^2} \frac{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}}$$

$$R^2 = r^2 \tag{6.30}$$

Apesar dessa relação é importante entendermos que os conceitos de R^2 e *r* são diferentes e explicam fenômenos distintos:

- *r* é uma medida da associação linear entre duas variáveis;
- R^2 é a medida do quanto a variabilidade em uma variável pode explicar a variabilidade em outra.

Também é importante ter em mente que a simples observação do coeficiente de determinação não é condição suficiente para se afirmar o quanto um modelo de regressão é bom ou ruim. Por exemplo, a Figura 6.12, a partir do conjunto de dados criado por Francis Anscombe (1973), ilustra regressões lineares de variáveis que se relacionam de formas muito diferentes: a Figura 6.12a é o que se poderia esperar de uma regressão; a Figura 6.12b mais se aproxima de uma função quadrática; a Figura 6.12c consiste em uma linha predominantemente sem variação, com um único *out-lier*; a Figura 6.12d apresenta todos

os seus dados em um único *x*, exceto por um ponto mais alto em outro *x* muito distante. Contudo, os parâmetros estimados e R^2 são praticamente idênticos.

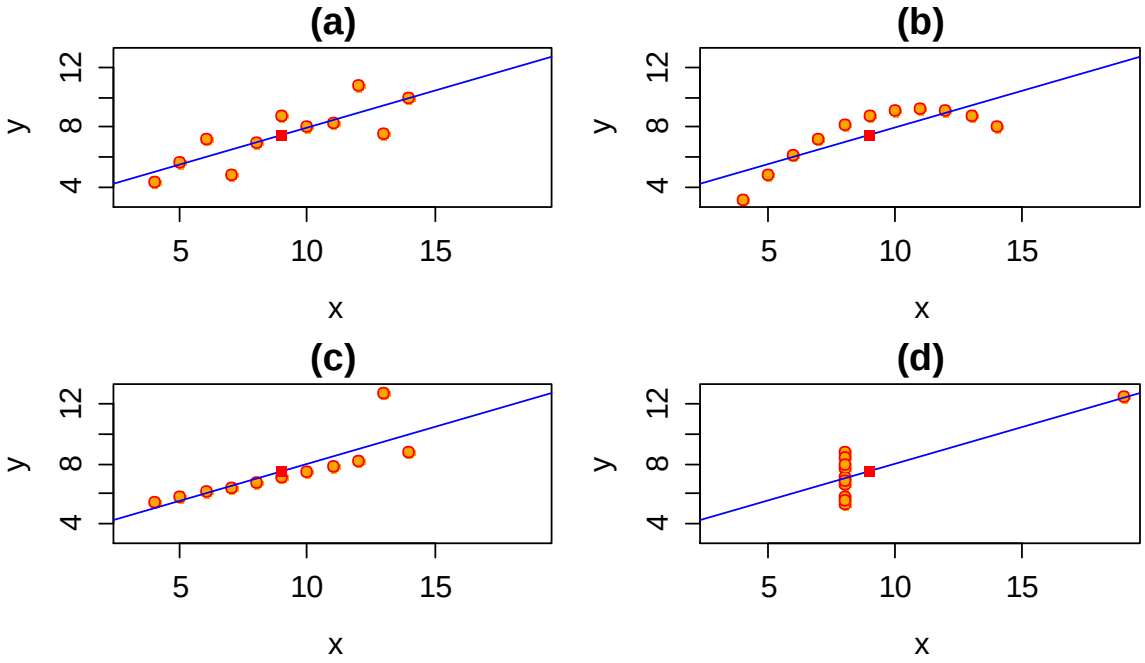


Figura 6.12 – Regressões lineares com parâmetros quase idênticos: $\hat{\beta}_0 \approx 3$, $\hat{\beta}_1 \approx 0.5$, $\bar{x} \approx 9$, $\bar{y} \approx 7.5$ e $R^2 \approx 0.67$. ®

Para evitar dificuldades na interpretação de R^2, alguns estatísticos utilizam o R^2 ajustado, ou R_a^2, que é definido para uma equação com *c* coeficientes como[21]

$$R_a^2 = 1 - \left(\frac{n-1}{n-c}\right)(1-R^2) \qquad (6.31)$$

A interpretação de R_a^2 é a mesma que a de R^2.

[21] http://www.portalaction.com.br/analise-de-regressao/16-coeficiente-de-determinacao

6.2.6 Estimadores e o processo de inferência na regressão linear

Quando utilizamos estimadores, duas propriedades são desejadas: (1) que o estimador seja não viciado e (2) que sua variância seja tão pequena quanto possível. Se isso ocorrer, o processo de inferência ocorrerá em condições ótimas.

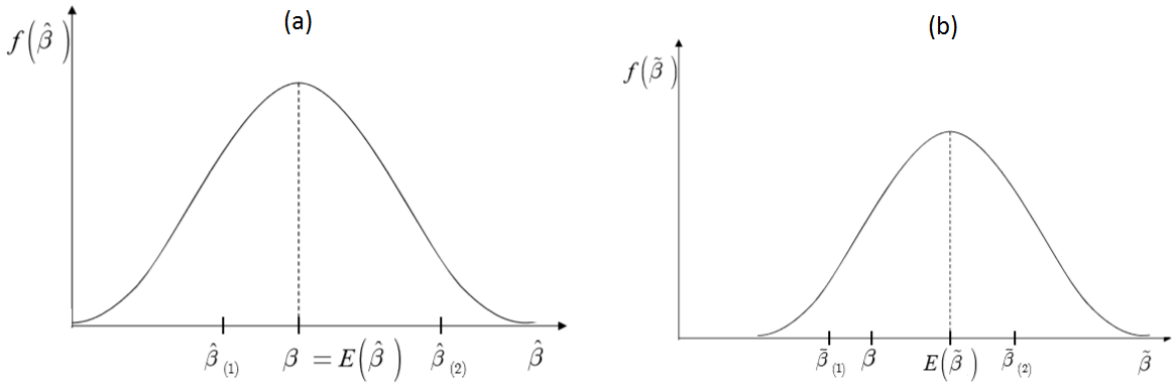


Figura 6.13 – (a) estimador não viciado; (b) estimador viciado.

A Figura 6.13a ilustra o caso de um estimador $\hat{\beta}$ não viciado, cujo valor esperado é igual ao do parâmetro β estimado. O estimador $\hat{\beta}$ é uma variável aleatória. Para cada nova amostra de *Y*, os valores de *x* são mantidos fixos[22] e $\hat{\beta}$ assume um valor diferente (podendo assumir infinita quantidade de valores), porém, na média, igual ao parâmetro β. Na Figura 6.13a, duas estimativas de β ($\hat{\beta}_{(1)}$ e $\hat{\beta}_{(2)}$) são obtidas. A primeira é relativamente próxima de β, enquanto que a segunda é mais distante. Em qualquer caso, o não vício, ou não enviesamento, é a propriedade desejada porque garante que, na média, o estimador é centrado no valor do parâmetro estimado.

Já o estimador $\tilde{\beta}$ da Figura 6.13b é viciado, uma vez que sua média é diferente de β. O viés é dado por $E(\tilde{\beta}) - \beta$. Como se pode notar, a estimativa $\tilde{\beta}_{(1)}$ (Figura 6.13b) é mais próxima de β que a estimativa $\hat{\beta}_{(2)}$ (Figura 6.13a), contudo, isso é obra do acaso. Um

[22] Os valores do regressor *x* são matidos fixos a cada nova amostra de *Y*, ou seja, para mesmos valores de *x*, diferentes amostras de *Y* são obtidas. Este é um forte pressuposto no caso das ciências sociais, onde, em geral, os dados são obtidos por observação e não por experimentação. É importante ressaltar que os resultados obtidos usando esse pressuposto permaneceriam virtualmente idênticos se assumirmos que os regressores são estocásticos, isso somado à condição de independência entre os regressores e o erro aleatório. Esse pressuposto pode ser enunciado como: o regressor *x* é distribuído independentemente do erro aleatório.

estimador não viciado sempre será preferível a um viciado, independentemente do que ocorrer em uma amostra específica, visto que não possui nenhum desvio sistematizado em relação ao valor do parâmetro estimado.

Outra propriedade desejada, denominada **eficiência do estimador**, diz respeito à variância dos estimadores (Figura 6.14).

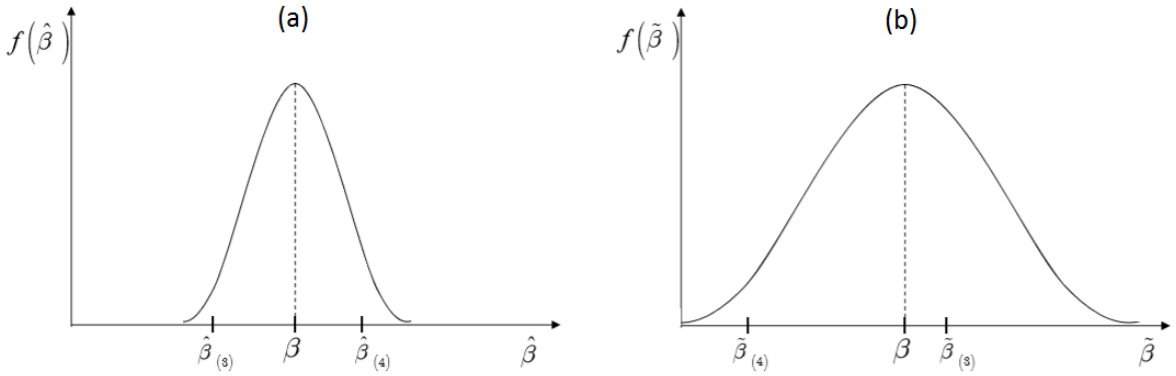


Figura 6.14 – Estimador não viciado: (a) com variância pequena; (b) com variância grande.

Embora a estimativa $\tilde{\beta}_{(3)}$ (Figura 6.14b) seja a mais próxima de β, isso é obra do acaso. Sempre deve ser preferível o estimador com a menor variância possível. Por exemplo, quando utilizamos $\hat{\beta}$ é praticamente improvável que uma estimativa esteja tão distante de β quanto a estimativa $\tilde{\beta}_{(4)}$ (Figura 6.14b), visto que a faixa de variação de $\hat{\beta}$ é muito menor que a de $\tilde{\beta}$.

Sob as condições de não viés e de variância pequena, os estimadores de mínimos quadrados ordinários possuem algumas propriedades ideais e podem ser considerados os melhores estimadores lineares não viciados (*best linear unbiased estiamtors* - *BLUE*).

Por exemplo, considere que o estimador $\hat{\theta}$ de θ é não viciado e possui variância pequena. Então, estatísticas na forma $\frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}}$ frequentemente apresentam as seguintes propriedades:

1. São normalmente distribuídas *N*~(0,1) para grandes amostras e têm uma distribuição *t-Student* para amostras menores (finitas), caso em que a variância estimada ($\hat{\sigma}_{\hat{\theta}}$) é substituída na estatística: $\frac{\hat{\theta}-\theta}{\hat{\sigma}_{\hat{\theta}}}$.
2. Podem ser usadas para testes de hipótese do tipo $H_0:\theta=\theta_0$ versus $H_1:\theta>,<,\neq\theta_0$.
3. Podem ser usadas para criar intervalos de confiança para θ na forma $\hat{\theta}\pm Q_{1-\alpha/2}\hat{\sigma}_{\hat{\theta}}$ onde $Q_{1-\alpha/2}$ é o quantil relevante tanto para a distribuição normal quanto para a *t-Student*, conforme o caso.

Sabendo que a distribuição amostral de estimadores de mínimos quadrados é centrada em torno do valor verdadeiro do parâmetro, precisamos conhecer o quanto a distribuição é espalhada. A variância de um estimador é um indicador da acurácia do estimador.

Nesse sentido, é importante determinarmos a variância dos estimadores do intercepto e da inclinação da reta de regressão. Começando pela inclinação, a partir da Equação (6.14), podemos escrever

$$Var(\hat{\beta}_1)=Var\left(\frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$$

$$=Var\left(\frac{\sum_{i=1}^{n}y_i(x_i-\bar{x})-\bar{y}\sum_{i=1}^{n}(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)=Var\left(\frac{\sum_{i=1}^{n}y_i(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$$

$$=\frac{Var(\sum_{i=1}^{n}y_i(x_i-\bar{x}))}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2}=\frac{\sum_{i=1}^{n}\sigma^2(x_i-\bar{x})^2}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2}=\frac{\sigma^2\sum_{i=1}^{n}(x_i-\bar{x})^2}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2}$$

Logo,

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{6.32}$$

A variância de $\hat{\beta}_0$ é dada por[23]

$$Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right) \tag{6.33}$$

Na prática, o σ das equações (6.32) e (6.33) é substituído por sua estimativa.

Dado o exposto nesta seção, também não é surpresa considerar que sob a condição de erros gaussianos, independentes e identicamente distribuídos, a estatística

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}}$$

segue uma distribuição *t-Student* com *n*-2 graus de liberdade e uma distribuição normal para grandes valores de *n*. Esse fato pode ser usado para criar intervalos de confiança e para realizar testes de hipótese.

6.2.7 Desenvolvendo os conceitos em R

O Quadro 6.13 apresenta código R que calcula expressamente os parâmetros da regressão linear para a base de dados de diamantes, para fins didáticos.

[23] Demonstração em http://www.portalaction.com.br/analise-de-regressao/13-propriedades-dos-estimadores

Quadro 6.13 – Parâmetros da regressão linear calculados de forma expressa.

```
library(UsingR); data(diamond)
y <- diamond$price; x <- diamond$carat; n <- length(y)

beta1 <- cor(y, x) * sd(y) / sd(x) # Equação 6.15
beta0 <- mean(y) - beta1 * mean(x) # Equação 6.11
mu <- beta0 + beta1 * x            # Equação 6.7
e <- y -  mu                       # Equação 6.8
sigma <- sqrt(sum(e^2) / (n-2))    # Equação 6.19
ssx <- sum((x - mean(x))^2)
seBeta0 <- (1 / n + mean(x) ^ 2 / ssx) ^ .5 * sigma  # Equação 6.33
seBeta1 <- sigma / sqrt(ssx)                         # Equação 6.32

# Estatística t para o teste de hipótese
# H0: β0 = 0 ou H1: β0 >,<,≠ 0
tBeta0 <- beta0 / seBeta0;
# H0: β1 = 0 ou H1: β1 >,<,≠ 0
tBeta1 <- beta1 / seBeta1

# probabilidade p da estatístcia t com df = n-2
pBeta0 <- 2 * pt(abs(tBeta0), df = n - 2, lower.tail = FALSE)
pBeta1 <- 2 * pt(abs(tBeta1), df = n - 2, lower.tail = FALSE)

# monta um data frame com os resultados
coefTable <- rbind(c(beta0, seBeta0, tBeta0, pBeta0),
                   c(beta1, seBeta1, tBeta1, pBeta1))
colnames(coefTable) <- c("Estimate", "Std. Error", "t value","Pr(>|t|)")
rownames(coefTable) <- c("(Intercept)", "x")



# coeficiente de determinação
VRR <- sum((mu - mean(y))^2)      # Equação 6.23
VRE <- sum((y - mu)^2)            # Equação 6.22
VT  <- VRR + VRE                  # Equação 6.24
R2  <- VRR / VT                   # Equação 6.25

# coeficiente de determinação ajustado
R2a <- 1 - ((n-1)/(n-2)) * (1-R2) # Equação 6.26

coefTable
```

```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -259.6259   17.31886 -14.99094 2.523271e-19
x           3721.0249   81.78588  45.49715 6.751260e-40
```

```
paste("Residual standard error:",round(sigma,2),"on",n-2,"degrees of
freedom")
```

```
"Residual standard error: 31.84 on 46 degrees of freedom"
```

```
paste0("Multiple R-squared: ",round(R2,4),", Adjusted R-squared: ",
round(R2a,4))
```

```
"Multiple R-squared: 0.9783, Adjusted R-squared: 0.9778"
```

Na prática, a função lm é utilizada para a obtenção desses resultados, conforme apresentado no Quadro 6.14.

Quadro 6.14 – Parâmetros da regressão linear calculados pela função lm.

```
library(UsingR); data(diamond)
y <- diamond$price; x <- diamond$carat;

fit <- lm(y ~ x)
summary(fit)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-85.159 -21.448  -0.869  18.972  79.370

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -259.63      17.32  -14.99   <2e-16 ***
x            3721.02      81.79   45.50   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.84 on 46 degrees of freedom
Multiple R-squared:  0.9783,   Adjusted R-squared:  0.9778
F-statistic:  2070 on 1 and 46 DF,  p-value: < 2.2e-16
```

Intervalos de confiança para os estimadores também podem ser calculados a partir dos coeficientes estimados, conforme apresentado no Quadro 6.15.

Quadro 6.15 – Obtendo intervalos de confiança para os coeficientes da regressão.

```
library(UsingR); data(diamond)
y <- diamond$price; x <- diamond$carat;

fit <- lm(y ~ x)
coefs <- summary(fit)$coefficients

# interval de confiança para β0
coefs[1,1] + c(-1, 1) * qt(.975, df=fit$df) * coefs[1,2]

[1] -294.4870 -224.7649

# interval de confiança para β1
coefs[2,1] + c(-1, 1) * qt(.975, df=fit$df) * coefs[2,2]

[1] 3556.398 3885.651
```

Interpretando os resultados para β_1, podemos dizer que, com 95% de confiança, estima-se que o crescimento de 1 carat no peso do diamante resulta em um acréscimo de preço que pode variar de 3556,4 a 3888,7 dólares singaporenses.

6.2.8 Fatores de incerteza e predição de valores

Já sabemos que o modelo de regressão linear simples é composto por uma parte determinísitica, esperada, e outra aleatória, atrelada a um erro, que produz uma incerteza sobre o fenômeno investigado. Quando se relata o resultado de medição de uma grandeza, é necessário que alguma indicação quantitativa da qualidade do resultado seja apresentada, de forma tal que se possa avaliar a confiabilidade do resultado.

A **incerteza** é um parâmetro que indica a qualidade de uma medida de forma quantitativa. No modelo de regressão linear, a incerteza pode ser aferida por meio da definição de intervalos de confiança para os valores ajustados (IC_{VA}). Os IC_{VA}s indicam a precisão das estimativas: quanto menor a amplitude, maior a precisão.

Como vimos na seção 6.2.6, os estimadores de mínimos quadrados ordinários podem ser considerados os melhores estimadores lineares não viciados (*BLUE*). Isso significa dizer que dado um valor x_0 da variável regressora x, o valor esperado de $\hat{\beta}_0 + \hat{\beta}_1 x_0$ é $\beta_0 + \beta_1 x_0$, ou seja, $\hat{\beta}_0 + \hat{\beta}_1 x$ é um estimador pontual para a esperança de Y. Em termos matemáticos, a esperança desse estimador é dada por

$$E[\hat{\beta}_0 + \hat{\beta}_1 x] = E[\hat{\beta}_0] + E[\hat{\beta}_1 x] = \beta_0 + \beta_1 x \tag{6.34}$$

e a variância por

$$Var[\hat{\beta}_0 + \hat{\beta}_1 x] = Var[\hat{\beta}_0] + Var[\hat{\beta}_1 x] + 2xCov[\hat{\beta}_0, \hat{\beta}_1] \tag{6.35}$$

Substituindo as equações (6.32) e (6.33) em (6.35), temos

$$
\begin{aligned}
Var[\hat{\beta}_0 + \hat{\beta}_1 x] &= \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] + x^2 \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} + 2x \frac{-\bar{x}\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\
&= \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]
\end{aligned}
\tag{6.35}
$$

Então

$$
\hat{\beta}_0 + \hat{\beta}_1 x \sim N\left(\beta_0 + \beta_1 x; \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] \right) \tag{6.36}
$$

A partir da Equação (6.36), podemos definir um intervalo de confiança para $E[Y|x]$ (esperança de Y dado x) por meio da estatística

$$
\frac{(\hat{\beta}_0 + \hat{\beta}_1 x) - (\beta_0 + \beta_1 x)}{\sqrt{\sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)}} \tag{6.37}
$$

a qual tem distribuição t-*Student* com n-2 graus de liberdade.

Como σ^2 não é conhecido, devemos usar o estimador não-viciado da (Equação (6.19)). O IC_{VA}, então, é dado por

$$
\hat{\beta}_0 + \hat{\beta}_1 x \pm t_{(\frac{\alpha}{2}, n-2)} \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)} \tag{6.38}
$$

onde a amplitude do intervalo é

$$2t_{(\frac{\alpha}{2},n-2)}\sqrt{\widehat{\sigma}^2\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)}$$

e o erro padrão dos valores ajustados (EP_{VA}) no ponto x_0, por exemplo, é dado por

$$EP_{VA}=\widehat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \tag{6.39}$$

A região delimitada pelos IC_{VA}s dos valores de *x* é denominda **banda de confiança** (representada pelas linhas vermelhas da Figura 6.15). As amplitudes dos intervalos crescem na medida em que os valores de *x* se afastam dos valores centrais (em torno de $\bar{x}$).

Considere, agora, que desejamos fazer a predição de um valor y_0 de *Y* em um dado valor x_0 para o qual não existem medições, considerando as incertezas da predição. Da Equação (6.7), podemos dizer que $\widehat{\mu}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$ é uma predição de y_0. Então, o erro de predição é $y_0 - \widehat{\mu}_0$, onde y_0 e $\widehat{\mu}_0$ são independentes, visto que y_0 não pertence à amostra utilizada para o ajuste do modelo. Da definição do modelo de regressão linear,

$$y_0 \sim N(\beta_0 + \beta_1 x_0; \sigma^2)$$

e a partir da esperança e da variança do erro de predição, podemos definir o **intervalo de predição** (IP). Matematicamente,

$$E[y_0 - \widehat{\mu}_0] = E[y_0] - E[\widehat{\mu}_0] = y_0 - y_0 = 0 \tag{6.40}$$

e

$$
\begin{aligned}
Var[y_0 - \hat{\mu}_0] &= Var[y_0] - Var[\hat{\mu}_0] \\
&= \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) \\
&= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)
\end{aligned}
\tag{6.41}
$$

Então,

$$
y_0 - \hat{\mu}_0 \sim N\left[0; \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) \right] \tag{6.42}
$$

Novamente, a partir da Equação (6.36), podemos definir um intervalo de predição para $E[(y_0 - \hat{\mu}_0) \mid x]$ (esperança de $y_0 - \hat{\mu}_0$ dado *x*) por meio da estatística

$$
\frac{(y_0 - \hat{\mu}_0) - 0}{\sqrt{\sigma^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)}} \tag{6.43}
$$

a qual tem distribuição t-*Student* com *n*-2 graus de liberdade. Daí,

$$
\pm t_{(\frac{\alpha}{2}, n-2)} = \frac{(y_0 - \hat{\mu}_0) - 0}{\sqrt{\sigma^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)}} \tag{6.44}
$$

e substituindo pelo estimador não-viciado da (Equação (6.19)), temos

$$
y_0 = \hat{\mu}_0 \pm t_{(\frac{\alpha}{2}, n-2)} \sqrt{\hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)} \tag{6.45}
$$

onde a amplitude do intervalo de predição é

$$2t_{(\frac{\alpha}{2},n-2)}\sqrt{\hat{\sigma}^2\left(1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)}$$

e o erro padrão em relação aos valores preditos (EP_{PRED}) no ponto x_0, por exemplo, é dado por

$$EP_{PRED}=\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \tag{6.46}$$

A região delimitada pelos IPs é denominda **banda de predição** (representada pelas linhas azuis da Figura 6.15). As amplitudes dos intervalos de predição também crescem na medida em que os valores de *x* se afastam dos valores centrais (em torno de $\bar{x}$). Isso revela o risco de fazer predições fora do intervalo observado.

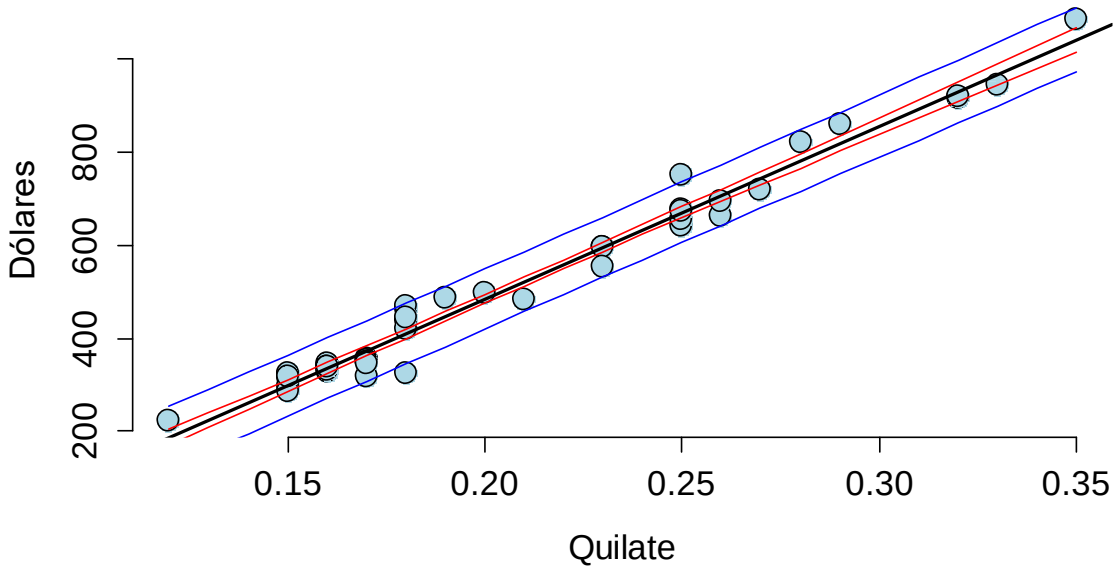


Figura 6.15 – Intervalos de confiança de valores ajustados (linhas vermelhas) e intervalos de predição (linhas azuis). ®

6.3 LABORATÓRIO 6

Para os exercícios a seguir: (1) efetue a regressão linear; (2) encontre os coeficientes de regressão; (3) plote o gráfico de dispersão e a reta de regressão; (4) plote o histograma dos resíduos; (5) plote o gráfico dos resíduos em função do preditor; (6) determine o valor esperado para os preditores especificados; e (7) utilize a função `summary` para verificar o valor de R^2.

6.3.1 Exercício 1

O conjunto de dados `cars`, do pacote `datasets`, registra a distância (`distance`) percorrida até a parada total de um determinado carro e a velocidade (`speed`) em que o freio foi pressionado. Qual o valor esperado da distância frenagem para as velocidades (mph): 5, 10, 14, 23 e 30.

6.3.2 Exercício 2

O conjunto de dados `airquality`, do pacote `datasets`, registra leituras diárias da qualidade do ar em Nova York no ano de 1973. Verifique, por meio da regressão linear, a variação da concentração de Ozônio em função da radiação solar. Estime a concentração de Ozônio esperada para os seguintes níveis de radiação (lang): 10, 100, 200, 300. Verifique, também, a variação da concentração de Ozônio em função da velocidade do vento. Estime a concentração de Ozônio esperada para as seguintes velocidades do vento (mph): 1.5, 3.7, 10, 15.3, 22.
Obs.: desconsidere os registros em que não há valor observado (NA).

6.3.3 Exercício 3

O arquivo `Eleicoes2014`, disponível na pasta do laboratório, contém dados da eleição de 2014 para Deputado Federal. Verifique, por meio da regressão linear, a variação do quantitativo de votos nominais (`Nominais`) em função da receita de campanha (`Receitas.em.2014`).

6.3.4 Exercício 4

Utilize a função R `lmIC`, a seguir, e repita os exercícios anteriores considerando a banda de confiança e a banda de predição.

```
lmIC <- function(x, Y, vx=NULL, nc=.95){

  n <- length(x)
  tc <- nc + (1-nc)/2 # tc = t crítico

  plot(x, Y , frame=FALSE, pch=21, col="black", bg="lightblue", cex=2)
  fit <- lm(Y ~ x)
  abline(fit, lwd=2)
  beta0 <- coef(fit)[1]
  beta1 <- coef(fit)[2]
  xVals <- seq(min(x,na.rm=TRUE),
               max(x,na.rm=TRUE),
               by=(max(x,na.rm=TRUE)-min(x,na.rm=TRUE))/200)
  yVals <- beta0 + beta1 * xVals
  sigma <- sqrt(sum(fit$residuals^2) / (n-2))
  ssx <- sum((x - mean(x))^2)
  se1 <- sigma * sqrt(1 / n + (xVals - mean(x))^2 / ssx)
  se2 <- sigma * sqrt(1 + 1 / n + (xVals - mean(x))^2 / ssx)
  lines(xVals, yVals + qt(tc,n-2) * se1, col="red")
  lines(xVals, yVals - qt(tc,n-2) * se1, col="red")
  lines(xVals, yVals + qt(tc,n-2) * se2, col="blue")
  lines(xVals, yVals - qt(tc,n-2) * se2, col="blue")

  if( !is.null(vx) ){
    ret <- NULL
    yVals <- beta0 + beta1 * vx
    se2 <- sigma * sqrt(1 + 1 / n + (vx - mean(x))^2 / ssx)
    for(i in 1:length(vx)){
      ret <- c(ret, yVals[i],
               yVals[i] - qt(tc,n-2) * se2[i],
               yVals[i] + qt(tc,n-2) * se2[i])
    }
    return(matrix(ret, ncol=3, byrow = TRUE,
                  dimnames=list(c(),c("esperado","Icmin","ICmax"))))
  }
}
```

7 REGRESSÃO LINEAR MÚLTIPLA

Neste capítulo, os conceitos apresentados para a regressão linear simples serão estendidos para o caso em que o comportamento de uma variável é explicado por mais de um preditor. Aqui, tentaremos expor os resultados com o menor desenvolvimeno possível de álgebra linear.

7.1 CONTEXTUALIZAÇÃO DO PROBLEMA

Suponha que um pesquisador deseje lhe convencer sobre a hipótese de que existe relação entre o consumo de drops de hortelã e a perda da função pulmonar medida em termos do volume expiratório forçado – VEF (volume de ar que pode ser expirado forçosamente em um segundo, após inspiração profunda). Num primeiro momento, você poderia reagir de forma cética e dizer que "é sabido que o fumo está relacionado à perda da função pulmonar e que fumantes tendem a consumir mais drops de hortelã que os não fumantes, o que, provavelmente, deve ser a causa". Perguntado pelo pesquisador como ele poderia convencê-lo de que está certo, você responde: "Se não fumantes que consomem drops de hortelã apresentarem menor capacidade respiratória do que não fumantes que não consomem drops de hortelã e, ainda, se fumantes que consomem drops de hortelã também apresentarem menor capacidade respiratória do que fumantes que não consomem drops de hortelã, então eu estarei mais inclinado a acreditar na sua pesquisa". Ou seja, para considerar os resultados do cientista, ele precisa demonstrar que o fenômeno "consumo de drops de hortelã" produz algum efeito mesmo quando o fenômeno "fumar" permanece fixo ou neutralizado.

Considere outro exemplo. Em Ciência Política, o conceito de eficácia política[24] (ou atuação política) é definido pela fé e confiança dos cidadãos no governo e pela crença de que cada cidadão pode compreender e influenciar os assuntos políticos. É comumente medida por meio de questionários (*surveys*) e é utilizada como um indicador da saúde (ou maturidade) da sociedade civil. Quando cidadãos têm baixa eficácia política, eles não têm fé no governo e não acreditam que possam realizar ações capazes de impactar nas ações do governo e dos seus líderes políticos. Isso geralmente leva a atos de violência porque os

[24] Political efficacy: http://en.wikipedia.org/wiki/Political_efficacy

cidadãos sentem como se não tivessem outra opção para serem escutados e terem suas reivindicações atendidas. O senso comum diz que a eficácia política está fortemente correlacionada com a participação na vida social e política. Contudo, estudos da American National Election Studies[25] mostram que a diferença na taxa de participação em atividades sócio-políticas é fortemente influenciada pelos diferentes níveis de educação dos cidadãos. Uma interpretação cuidadosa não contraria o senso comum, apenas aponta que aqueles que apresentam baixa eficácia política possuem menores níveis de educação e essa é a razão pela qual participam menos.

Em um último exemplo, considere que uma companhia de seguros está interessada no quanto os sinistros dos últimos anos podem prever o tempo médio em que os automóveis segurados permanecerão nas oficinas para reparo no presente ano. Pretende-se utilizar uma enorme quantidade de dados contida nos sinistros para predizer um único número. O modelo de regressão linear simples não é equipado para manipular mais de um preditor. Então, como generalizar esse modelo de forma a incorporar diversos preditores? Quais são as consequências de se adicionar vários preditores? Certamente, deve haver consequências decorrentes do uso de preditores que não são relacionados com *Y*, assim como deve haver consequências decorrentes da omissão de preditores que são relacionados com *Y*.

A próxima seção tratará da generalização do modelo linear simples para o modelo com múltiplos preditores.

7.2 O MODELO LINEAR GERAL

O modelo linear geral estende a regressão linear simples pela adição de termos ao modelo.

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \varepsilon_i = \sum_{k=1}^{p} \beta_k x_{ki} + \varepsilon_i \qquad (7.1)$$

onde $\varepsilon_i \sim N(0, \sigma^2)$. Na Equação (7.1), geralmente x_{1i} é igual a 1 ($x_{1i} = 1$), de forma que o intercepto seja incluído no modelo.

[25] http://electionstudies.org/index.htm

A estimativa dos coeficientes do modelo geral se dá pelo mesmo critério dos mínimos quadrados adotado para a regressão simples. Sob a condição de erros gaussianos independentes e identicamente distribuídos, o método dos mínimos quadrados é equivalente à estimativa por máxima verossimilhança e minimiza a expressão

$$S(\beta_1, \beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \sum_{k=1}^{p} \beta_k x_{ki} \right)^2 \tag{7.2}$$

No modelo geral também são observadas as suposições de independência, normalidade, homocedasticidade e linearidade. Importante compreender que o conceito da linearidade se trata da linearidade dos coeficientes, não dos preditores. A denominação "linear" se deve ao fato de que a esperança de *Y* (parte sistemática do modelo) é função linear dos parâmetros β_k, considerando-se valores fixos das variáveis regressoras. Por exemplo,

$$y_i = \beta_1 x_{1i}^2 + \beta_2 x_{2i}^2 + \ldots + \beta_p x_{pi}^2 + \varepsilon_i \tag{7.3}$$

ainda é um modelo linear visto que x_{ki}^2 é apenas um outro *x* que sofreu uma transformação *T*, tal que $T\{x_{ki}\} = x_{ki}^2 = y_{ki}$.

Um **sistema linear** é definido pelo princípio da superposição. Se y_{1i} e y_{2i} são o resultado das transformações *T* de x_{1i} e x_{2i}, então o sistema é linear se e somente se

$$T\{x_{1i} + x_{2i}\} = T\{x_{1i}\} + T\{x_{2i}\} = y_{1i} + y_{2i} \tag{7.4}$$

e

$$T\{ax_{ki}\} = aT\{x_{ki}\} = ay_{ki} \tag{7.5}$$

onde *a* é uma constante arbitrária. A primeira propriedade (7.4) é chamada de **aditividade** e a segunda (7.5), de **homogeneidade**. De forma combinada, temos o **princípio da superposição**

$$T\{ax_{1i} + bx_{2i}\} = aT\{x_{1i}\} + bT\{x_{2i}\} = ay_{1i} + by_{2i} \tag{7.6}$$

Logo, um sistema como o da Equação (7.3) é linear, assim como o sistema da Equação (7.7) também o é.

$$y_i = \beta_1 sen(x_{1i}) + \beta_2 sen(x_{2i}) + \ldots + \beta_p sen(x_{pi}) + \varepsilon_i \tag{7.7}$$

Como na regressão linear simples, o desafio no modelo geral é estimar os coeficientes de forma que o erro/resíduo seja o menor possível. Nesse sentido e com a finalidade de fundamentar a interpretação da **regressão linear múltipla** (ou multivariada), vamos voltar à **regressão pela origem** apresentada na seção 6.2.

Conforme discutido, a aplicação do método dos mínimos quadrados tem por objetivo encontrar os valores mínimos da função do resíduo, o que consiste em igualar a zero as dervivadas parciais da função do resíduo em relação cada parêmetro estimado. Tomado-se, portanto, a Equação (6.12) – derivada parcial em relação a $\widehat{\beta}_1$ –, podemos estimar a inclinação $\widehat{\beta}_1$ da reta de regressão pela origem, fazendo o intercepto $\widehat{\beta}_0 = 0$, ou seja

$$2\sum_{i=1}^{n}\left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)(-x_i) = 0$$

$$\sum_{i=1}^{n}\left(y_i - 0 - \widehat{\beta}_1 x_i\right)(-x_i) = \sum_{i=1}^{n}\left(y_i - 0 - \widehat{\beta}_1 x_i\right)(-x_i) = 0$$

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \tag{7.8}$$

O resíduo da regressão pela origem é dado pela diferença entre o valor observado y_i e o valor esperado $\widehat{\beta}_1 x_i$. De forma geral, dizemos que o resíduo resultante da regressão pela origem de *a* (variável dependente) pelo regressor *b* (variável independente) é dado por

$$e_{i,a|b} \equiv a_i - \frac{\sum_{j=1}^{n} b_j a_j}{\sum_{j=1}^{n} b_j^2} b_j \tag{7.9}$$

Considere, agora, o modelo com dois regressores, x_1 e x_2, representados no sistema de equações (7.10).

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})x_{1i} = 0 \\ \sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})x_{2i} = 0\end{aligned} \tag{7.10}$$

É possível expressar os coeficientes de regressão em termos do resíduo da regressão pela origem (Equação (7.9)). Isolando-se $\hat{\beta}_2$ na segunda equação de (7.10), temos

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_{1i})x_{2i} - \sum_{i=1}^{n}\hat{\beta}_2 x_{2i}^2 = 0$$

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_{1i})x_{2i}}{\sum_{i=1}^{n} x_{2i}^2} \tag{7.11}$$

Substituindo-se (7.11) na primeira equação de (7.10)

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_{1i} - \frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_{1i})x_{2i}}{\sum_{i=1}^{n} x_{2i}^2} x_{2i})x_{1i} = 0$$

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_{1i} - \frac{\sum_{i=1}^{n} y_i x_{2i}}{\sum_{i=1}^{n} x_{2i}^2} x_{2i} + \hat{\beta}_1 \frac{\sum_{i=1}^{n} x_{1i} x_{2i}}{\sum_{i=1}^{n} x_{2i}^2} x_{2i})x_{1i} = 0$$

$$\sum_{i=1}^{n}(y_i - \frac{\sum_{i=1}^{n} y_i x_{2i}}{\sum_{i=1}^{n} x_{2i}^2} x_{2i} - \hat{\beta}_1 x_{1i} + \hat{\beta}_1 \frac{\sum_{i=1}^{n} x_{1i} x_{2i}}{\sum_{i=1}^{n} x_{2i}^2} x_{2i})x_{1i} = 0$$

$$\sum_{i=1}^{n}(y_i - \frac{\sum_{i=1}^{n} y_i x_{2i}}{\sum_{i=1}^{n} x_{2i}^2} x_{2i} - \left[x_{1i} - \frac{\sum_{i=1}^{n} x_{1i} x_{2i}}{\sum_{i=1}^{n} x_{2i}^2} x_{2i} \right] \hat{\beta}_1)x_{1i} = 0$$

e aplicando-se a definição (7.9), temos

$$\sum_{i=1}^{n}(e_{i,Y|x_2} - \widehat{\beta}_1 e_{i,x_1|x_2})x_{1i} = 0 \qquad (7.12)$$

Isolando-se $\widehat{\beta}_1$ na primeira equação de (7.10) e repetindo-se o mesmo procedimento, encontramos

$$\sum_{i=1}^{n}(e_{i,Y|x_1} - \widehat{\beta}_2 e_{i,x_2|x_1})x_{2i} = 0 \qquad (7.13)$$

Das equações (7.12) e (7.13) podemos extrair a combinação dos resíduos necessários para a estimativa dos coeficientes. Por exemplo, verifica-se em (7.12) que $\widehat{\beta}_1$ é função de $e_{i,Y|x_2}$ e $e_{i,x_1|x_2}$, que podem ser tratados, respectivamente, como as novas variáveis dependente *x* e independente *Y*. Por raciocínio análogo, a partir de (7.13), $\widehat{\beta}_2$ é função de $e_{i,Y|x_1}$ e $e_{i,x_2|x_1}$. O Quadro 7.1 apresenta uma simulação para o caso de dois regressores.

Quadro 7.1 – Estimativa de coeficientes pelo método da obtenção dos resíduos para dois regressores.

```
n <- 100; x1 <- rnorm(n); x2 <- rnorm(n)

# resíduo da regressão pela origem de a por b – Equação (7.9)
e <- function(a, b) a -  sum( a * b ) / sum( b ^ 2) * b

# modelo linear para 2 regressores com e~N(0,0.1)
y <- x1 + x2 + rnorm(n, sd = .1)

# estimativa dos coeficientes pela função lm
coef(lm(y ~ x1 + x2 - 1))

       x1        x2
1.0155701 0.9791699

# estimativa dos coeficientes pelo método dos resíduos
# beta 1 – Equação (7.12)
ey <- e(y, x2)
ex <- e(x1, x2)
sum(ey * ex) / sum(ex ^ 2)

[1] 1.01557

# beta 2 – Equação (7.13)
ey <- e(y, x1)
ex <- e(x2, x1)
sum(ey * ex) / sum(ex ^ 2)
```

```
[1] 0.9791699
```

Para o caso de três regressores, o sistema de equações em função dos resíduos é representado em (7.14). O Apêndice II ilustra o desenvolvimento matemático o Quadro 7.2 contém a simulação para três regressores.

$$\begin{aligned}
\sum_{i=1}^{n}(e_{i,Y|x_3} - \widehat{\beta}_1 e_{i,x_1|x_3} - \widehat{\beta}_2 e_{i,x_2|x_3})x_{1i} = 0 \\
\sum_{i=1}^{n}(e_{i,Y|x_1} - \widehat{\beta}_2 e_{i,x_2|x_1} - \widehat{\beta}_3 e_{i,x_3|x_1})x_{2i} = 0 \\
\sum_{i=1}^{n}(e_{i,Y|x_2} - \widehat{\beta}_1 e_{i,x_1|x_2} - \widehat{\beta}_3 e_{i,x_3|x_2})x_{3i} = 0
\end{aligned} \tag{7.14}$$

Quadro 7.2 – Estimativa de coeficientes pelo método da obtenção dos resíduos para três regressores.

```
n <- 100; x1 <- rnorm(n); x2 <- rnorm(n); x3 <- rnorm(n)

# resíduo da regressão pela origem de a por b – Equação (7.9)
e <- function(a, b) a -  sum( a * b ) / sum( b ^ 2) * b

# modelo linear para 3 regressores com e~N(0,0.1)
y <- x1 + x2 + x3 + rnorm(n, sd = .1)

# estimativa dos coeficientes pela função lm
coef(lm(y ~ x1 + x2 + x3 - 1))

       x1        x2        x3
1.0006805 1.0094248 0.9890997

# estimativa dos coeficientes pela análise de resíduos – possibilidade 1
# beta1 – Equação (7.14a)
ey <- e(e(y, x3), e(x2, x3))
ex <- e(e(x1, x3), e(x2, x3))
sum(ey * ex) / sum(ex ^ 2)
[1] 1.000681
# beta2 – Equação (7.14a)
ey <- e(e(y, x3), e(x1, x3))
ex <- e(e(x2, x3), e(x1, x3))
sum(ey * ex) / sum(ex ^ 2)
[1] 1.009425
# beta3 – Equação (7.14c)
ey <- e(e(y, x2), e(x1, x2))
ex <- e(e(x3, x2), e(x1, x2))
sum(ey * ex) / sum(ex ^ 2)
[1] 0.9890997

# estimativa dos coeficientes pela análise de resíduos – possibilidade 2
# beta1 – Equação (7.14c)
ey <- e(e(y, x2), e(x3, x2))
ex <- e(e(x1, x2), e(x3, x2))
sum(ey * ex) / sum(ex ^ 2)
[1] 1.000681
# beta2 – Equação (7.14b)
```

```
ey <- e(e(y, x1), e(x3, x1))
ex <- e(e(x2, x1), e(x3, x1))
sum(ey * ex) / sum(ex ^ 2)
[1] 1.009425
# beta3 - Equação (7.14b)
ey <- e(e(y, x1), e(x2, x1))
ex <- e(e(x3, x1), e(x2, x1))
sum(ey * ex) / sum(ex ^ 2)
[1] 0.9890997
```

Generalizando para o modelo de *p* regressores, chegamos a um sistema de *p* equações

$$\begin{aligned}
&\sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \ldots - \hat{\beta}_p x_{pi})x_{1i} = 0 \\
&\sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \ldots - \hat{\beta}_p x_{pi})x_{2i} = 0 \\
&\quad\vdots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots \\
&\sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \ldots - \hat{\beta}_p x_{pi})x_{pi} = 0
\end{aligned} \tag{7.15}$$

onde, por exemplo, o coeficiente $\hat{\beta}_p$ pode ser estimado como em (7.11)

$$\hat{\beta}_p = \frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \ldots - \hat{\beta}_{p-1} x_{p-1,i})x_{pi}}{\sum_{i=1}^{n} x_{pi}^2} \tag{7.16}$$

e as equações podem ser expressas em termos dos resíduos da regressão pela origem.

$$\begin{aligned}
&\sum_{i=1}^{n}(e_{i,Y|x_p} - e_{i,x_1|x_p}\hat{\beta}_1 - e_{i,x_2|x_p}\hat{\beta}_2 - \ldots - e_{i,x_{p-1}|x_p}\hat{\beta}_{p-1})x_{1i} = 0 \\
&\sum_{i=1}^{n}(e_{i,Y|x_p} - e_{i,x_1|x_p}\hat{\beta}_1 - e_{i,x_2|x_p}\hat{\beta}_2 - \ldots - e_{i,x_{p-1}|x_p}\hat{\beta}_{p-1})x_{2i} = 0 \\
&\quad\vdots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots \\
&\sum_{i=1}^{n}(e_{i,Y|x_p} - e_{i,x_1|x_p}\hat{\beta}_1 - e_{i,x_2|x_p}\hat{\beta}_2 - \ldots - e_{i,x_{p-1}|x_p}\hat{\beta}_{p-1})x_{p-1,i} = 0
\end{aligned} \tag{7.17}$$

No contexto semântico, significa dizer que

> a estimativa de mínimos quadrados para um coeficiente $\hat{\beta}_k$ do modelo de regressão multivariada corresponde à estimativa dos resíduos da regressão

pela origem de cada um dos demais regressores x_j, $j \neq k$ e da saída *Y* com o regressor x_k .

Avançando um pouco mais, chegamos a

$$\begin{aligned}
&\sum_{i=1}^{n}(e_{i,Y|x_p} - e_{i,x_1|x_p}\widehat{\beta}_1 - e_{i,x_2|x_p}\widehat{\beta}_2 - \ldots - e_{i,x_{p-1}|x_p}\widehat{\beta}_{p-1})e_{i,x_1|x_p} = 0 \\
&\sum_{i=1}^{n}(e_{i,Y|x_p} - e_{i,x_1|x_p}\widehat{\beta}_1 - e_{i,x_2|x_p}\widehat{\beta}_2 - \ldots - e_{i,x_{p-1}|x_p}\widehat{\beta}_{p-1})e_{i,x_2|x_p} = 0 \\
&\quad\vdots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots \\
&\sum_{i=1}^{n}(e_{i,Y|x_p} - e_{i,x_1|x_p}\widehat{\beta}_1 - e_{i,x_2|x_p}\widehat{\beta}_2 - \ldots - e_{i,x_{p-1}|x_p}\widehat{\beta}_{p-1})e_{i,x_{p-1}|x_p} = 0
\end{aligned} \tag{7.18}$$

As equações (7.17) e (7.18) mostram que o sistema de *p* equações e *p* variáveis de (7.15) foi reduzido para um sistema de *p-1* equações e *p-1* variáveis. Cada variável foi substituída pelo respectivo resíduo da regressão pela origem com x_p. Esse processo é repetido até que somente *Y* e uma única variável *x* permaneça. Então, o coeficiente pode ser estimado como em (7.8). Na prática, a obtenção dos coeficientes no modelo geral também é efetuada por meio da função **lm** (*linear model*) do R.

A interpretação da regressão multivariada consiste em dizer que

um coeficiente de regressão β_k representa a mudança produzida na resposta *Y* por uma unidade de variação em x_k , mantendo-se os demais regressores fixos.

Todos os demais conceitos aplicados à regressão linear simples são válidos para a regressão múltipla:

- Valores ajustados: $\widehat{\mu}_i = \sum_{k=1}^{p} x_{ik}\widehat{\beta}_k$ (7.19)

- Resíduos: $e_i = y_i - \hat{\mu}_i$

  (7.20)

- Variância estimada: $\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2$ (7.21)

- Valor esperado da variância: $E[\hat{\sigma}^2] = \sigma^2$

- Predições para novos valores $x_1,...,x_p$ são efetuadas pela simples substituição dos valores em (7.1).

- Cada coeficiente possui um erro padrão $\hat{\sigma}_{\hat{\beta}_k}$ e $\frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}_{\hat{\beta}_k}}$ segue uma distribuição *t*-Student com *n-p* graus de liberdade.

- Valores preditos possuem erro padrão e podemos calcular a banda de confiança e a banda de predição.

Quadro 7.3 – Regressão Linear Múltipla: exemplos.

```
library(datasets)
data(swiss)
require(stats)
require(graphics)

# conhecendo os dados
? swiss

# verificando as relações marginais
pairs(swiss, panel = panel.smooth, main = "Swiss data", col = 3 +
(swiss$Catholic > 50))

# regressão múltipla - ver estimativa ajustada de Agriculture
summary(lm(Fertility ~ . , data = swiss))$coefficients
summary(lm(Fertility ~ . , data = swiss))

# estimativa ajustada de Agriculture
summary(lm(Fertility ~ Agriculture, data = swiss))$coefficients

# inversao de sinal - exemplo 1
y = swiss$Fertility
x1 = swiss$Agriculture
x2 = swiss$Examination
summary(lm(y ~ x1))$coef
summary(lm(y ~ x1 + x2))$coef

library(ggplot2)
dat = data.frame(y = swiss$Fertility, x1 = swiss$Agriculture, x2 =
swiss$Examination, ey = resid(lm(y ~ x2)), ex1 = resid(lm(x1 ~ x2)))
g = ggplot(dat, aes(y = y, x = x1, colour = x2))
g = g + geom_point(colour="grey50", size = 5) + geom_smooth(method = lm,
se = FALSE, colour = "black")
```

```
g = g + geom_point(size = 4)
g
g2 = ggplot(dat, aes(y = ey, x = ex1, colour = x2))
g2 = g2 + geom_point(colour="grey50", size = 5) + geom_smooth(method =
lm, se = FALSE, colour = "black") + geom_point(size = 4)
g2

# inversão de sinal - exemplo 2
n <- 100; x2 <- 1 : n; x1 <- .01 * x2 + runif(n, -.1, .1);
y = -x1 + x2 + rnorm(n, sd = .01)
summary(lm(y ~ x1))$coef
summary(lm(y ~ x1 + x2))$coef


dat = data.frame(y = y, x1 = x1, x2 = x2, ey = resid(lm(y ~ x2)), ex1 =
resid(lm(x1 ~ x2)))
g = ggplot(dat, aes(y = y, x = x1, colour = x2))
g = g + geom_point(colour="grey50", size = 5) + geom_smooth(method = lm,
se = FALSE, colour = "black")
g = g + geom_point(size = 4)
g
g2 = ggplot(dat, aes(y = ey, x = ex1, colour = x2))
g2 = g2 + geom_point(colour="grey50", size = 5) + geom_smooth(method =
lm, se = FALSE, colour = "black") + geom_point(size = 4)
g2


# quanto o modelo explica o índice de fertilidade
summary(lm(Fertility ~ Agriculture
           + Examination
           + Education
           + Catholic
           + Infant.Mortality
           , data = swiss))


# o sinal inverte com a inclus?o de Examination + Education
# ambos são inversamente correlacionados com Agriculture
summary(lm(Fertility ~ Agriculture
           # + Examination
           # + Education
           + Catholic
           + Infant.Mortality
           , data = swiss))

cor(swiss$Agriculture, swiss$Examination)
cor(swiss$Agriculture, swiss$Education)
cor(swiss$Education, swiss$Examination)

# e se incluirmos uma variável desnecessária
z <- swiss$Agriculture + swiss$Education
lm(Fertility ~ . + z, data = swiss)
```

7.2.1 Variáveis “burras” são espertas

Considere o modelo linear

$$y_i = \beta_0 + x_{1i}\beta_1 + \varepsilon_i \tag{7.22}$$

onde cada x_{1i} é binário, de forma que recebe o valor 1, se um indivíduo (ou observação) pertencer a um determinado grupo, e o valor 0 caso contrário.

Então, para indivíduos pertencentes ao grupo, o valor esperado é

$$E[y_i] = \beta_0 + \beta_1 \tag{7.23}$$

Para indivíduos não pertencentes ao grupo, o valor esperado é

$$E[y_i] = \beta_0 \tag{7.24}$$

Aplicando os mínimos quadrados, podemos dizer que $\widehat{\beta}_0 + \widehat{\beta}_1$ é o valor médio para aqueles pertencentes ao grupo e $\widehat{\beta}_0$ é a média para os não pertencentes.

O coeficiente β_1 é interpretado como o acréscimo/decréscimo na média quando comparamos aqueles pertencentes ao grupo com os não pertencentes.

Considere, agora, múltiplos fatores. Por exemplo, considere a filiação político-partidária em três níveis: PT, PSDB e OUTROS. Então,

- $y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i$
- x_{1i} é 1 para petistas e 0 caso contrário.
- x_{2i} é 1 para tucanos e 0 caso contrário.
- Para petistas, $E[y_i] = \beta_0 + \beta_1$

- Para tucanos, $E[y_i] = \beta_0 + \beta_2$
- Para outros, $E[y_i] = \beta_0$
- β_1 compara PT com OUTROS.
- β_2 compara PSDB com OUTROS.
- $\beta_1 - \beta_2$ compara PT com PSDB.
- A escolha da categoria de referência altera a interpretação dos coeficientes.

Quadro 7.4 – Regressão fatorial com múltiplos níveis.

```
library(datasets)
require(stats)
require(graphics)
data(InsectSprays); head(InsectSprays)
? InsectSprays

boxplot(count ~ spray, data = InsectSprays, col = "lightgray",
        xlab="Tipo do spray", ylab="Quantidade de insetos")

# modelo linear
summary(lm(count ~ spray, data = InsectSprays))$coef

# explicitando o modelo
summary(lm(count ~
             I(1 * (spray == 'B')) + I(1 * (spray == 'C')) +
             I(1 * (spray == 'D')) + I(1 * (spray == 'E')) +
             I(1 * (spray == 'F'))
           , data = InsectSprays))$coef

# e se incluirmos as 6 variáveis
summary(lm(count ~
             I(1 * (spray == 'B')) + I(1 * (spray == 'C')) +
             I(1 * (spray == 'D')) + I(1 * (spray == 'E')) +
             I(1 * (spray == 'F')) + I(1 * (spray == 'A')), data =
InsectSprays))$coef

# e se omitirmos o intercepto
summary(lm(count ~ spray - 1, data = InsectSprays))$coef
unique(ave(InsectSprays$count, InsectSprays$spray))

# reordenando os níveis
spray2 <- relevel(InsectSprays$spray, "C")
summary(lm(count ~ spray2, data = InsectSprays))$coef
```

Em resumo, se tratarmos a variável **spray** como um fator, o R inclui um intercepto e omite o primeiro nível de fator, considerando a ordem alfabética.

- Todos os testes *t* são efetuados para comparar o primeiro nível do fator com os demais níveis.
- O valor esperado (média empírica) do primeiro fator é o intercepto.
- O valor esperado dos demais fatores consistem da soma do respectivo coeficiente com o itercepto.

Se omitirmos o intercepto, então todos os níveis do fator são incluídos.

- Os coeficientes são as médias dos respectivos grupos.
- Os testes *t* testam se a média de cada grupo é diferente de zero.

7.2.2 Ajuste pelo efeito de grupo (ou do fenômeno provocado)

Considere que sobre um determinado comportamento observado de duas variáveis em uma amostra se deseje testar o efeito de um fenômeno qualquer adicionado. Divide-se, então, a amostra em dois grupos (controle e tratamento), aplica-se o estímulo sobre um dos grupos e observam-se os efeitos.

A seguir, são apresentadas quatro simulações de modelos multivariados considerando o regressor mais a variável de grupo.

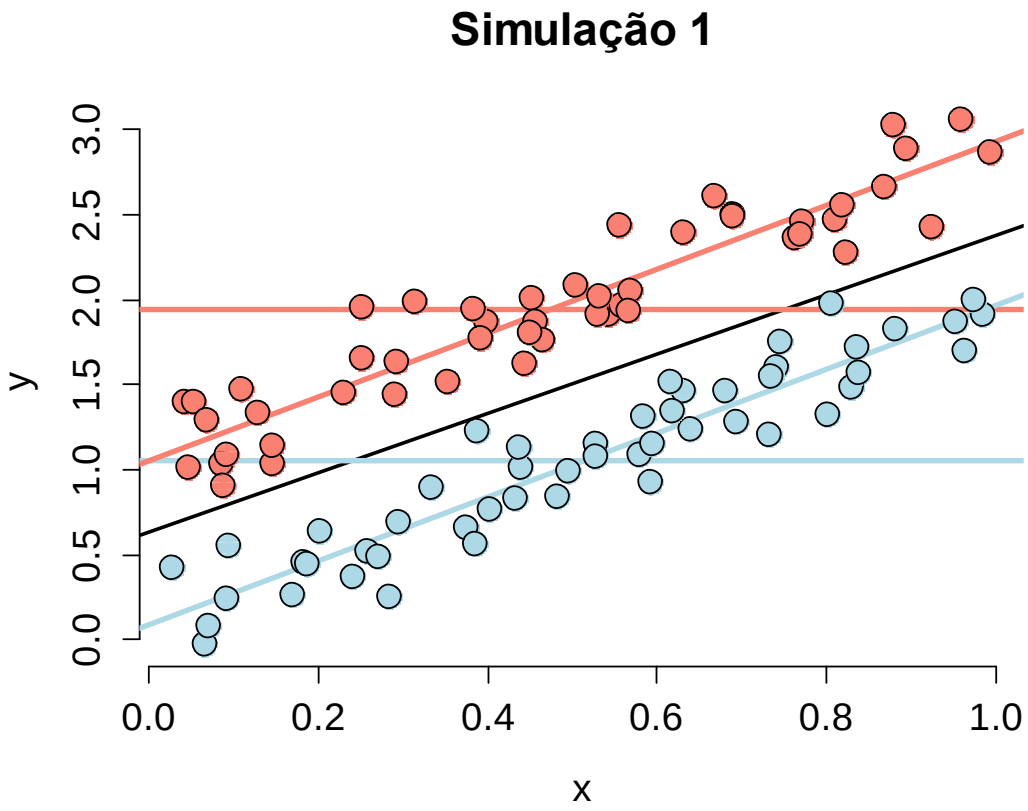


Figura 7.1 – Simulação 1: o regressor não está relacionado ao estado do grupo. ®

Considerações:

- A variável *x* não está relacionada ao status de grupo.
- A variável *x* está relacionada com *Y*, mas o intercepto depende do estado de grupo.
- A variável de grupo está relacionada *Y*.
  - ✓ A relação entre o estado do grupo e *Y* é constante em relação a *x*.
  - ✓ A relação entre o grupo e *Y* desconsiderando *x* é aproximadamente a mesma que se observa mantendo-se *x* constante.
- O modelo detecta ajustes em função do grupo.

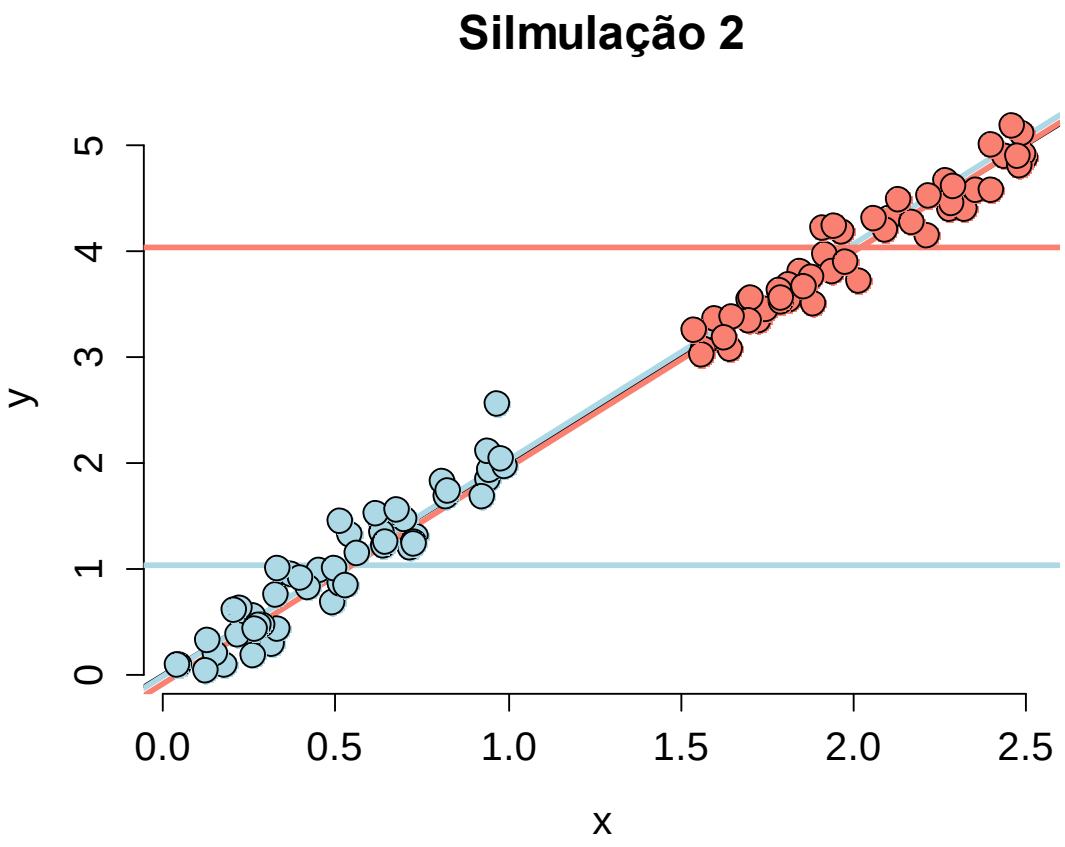


Figura 7.2 – Simulação 2: o regressor está relacionado ao estado do grupo. ®

Considerações:

- A variável *x* é altamente relacionada com o estado do grupo.
- A variável *x* está relacionada a *Y* e o intercepto não depende do grupo.
  - ✓ A variável *x* permanece relacionada a *Y* mantendo-se o estado do grupo constante.
- A variável de grupo é marginalmente relacionada com *Y*, desconsiderando-se *x*.
- O modelo não estima ajustes devido ao efeito de grupo.
  - ✓ Não há dados capazes de informar o relacionamento entre o grupo e *Y*.

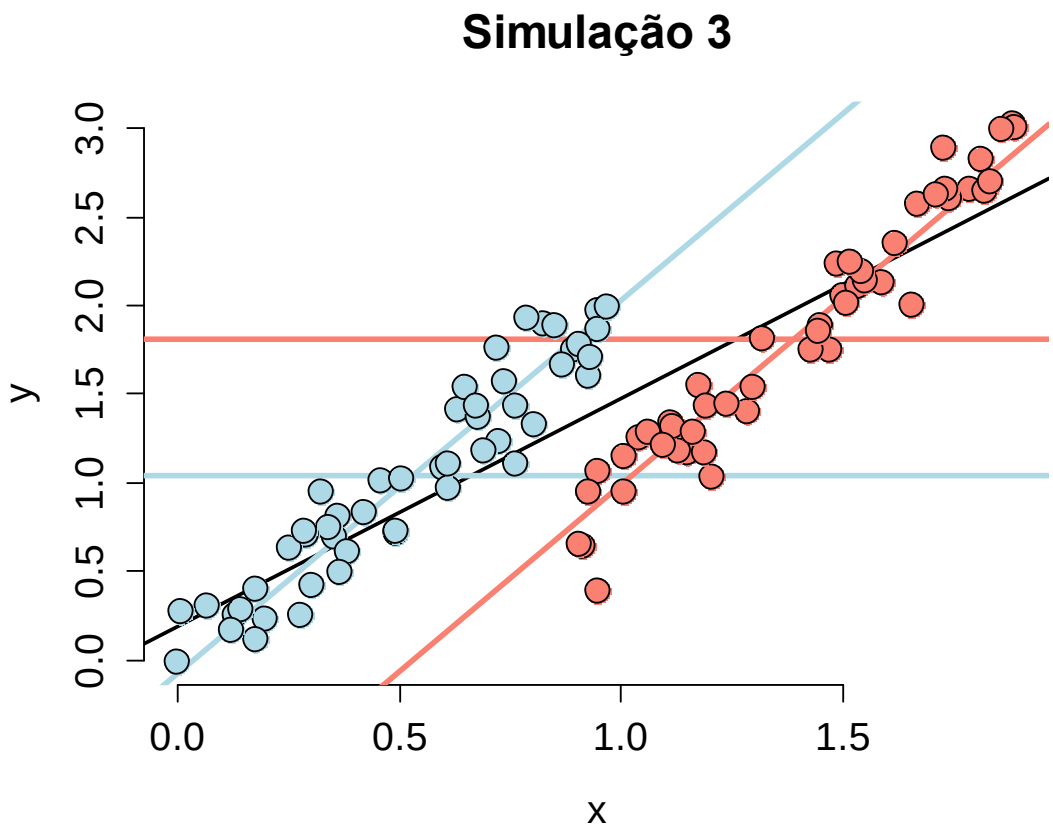


Figura 7.3 – Simulação 3: o regressor está relacionado ao estado do grupo. ®

Considerações:

- A associação marginal mostra o grupo rosa maior do que o grupo azul.
- O relacionamento ajustado mostra o grupo azul maior do que o rosa.
- A variável *x* está relacionada ao status de grupo.
    - ✓ Há evidência direta para a comparação dos grupos mantendo-se *x* fixo.

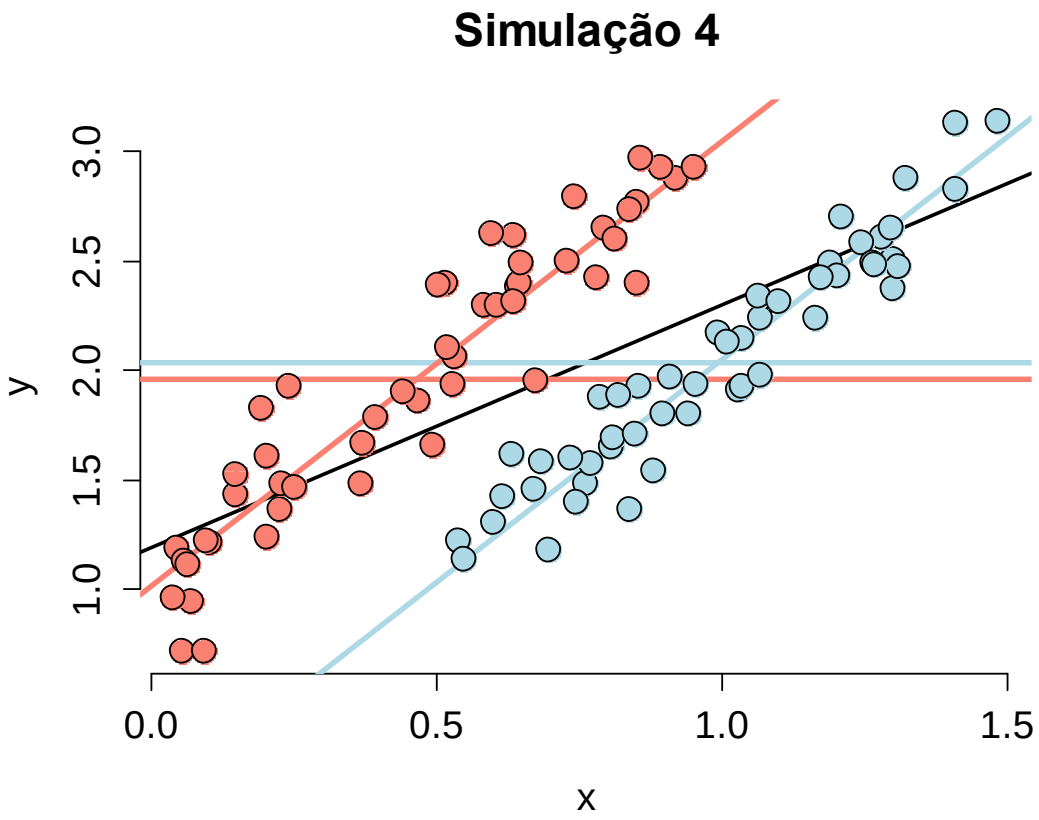


Figura 7.4 – Simulação 4: não há associação marginal entre o estado de grupo e *Y*. ®

Considerações:

- Nenhuma associação marginal entre o estado de grupo e *Y*.
- Forte relacionamento ajustado.
- Há evidência direta para comparar os grupos.

7.2.3 Regressão múltipla com regressores não relacionados

A adição de um novo regressor ao um modelo multivariado praticamente não provoca ajustes nas relações marginais quando o novo regressor não se relaciona com os regressores presentes no modelo. Esse fato é exemplificado na Figura 7.5, a qual mostra a relação entre *Y* e x_1, e utiliza um gradiente de cores para representar um segundo regressor x_2, que não se relaciona com x_1. Pode-se constatar que as cores variam com *Y*, isto é, quanto menor *Y*, mais escura a cor, e vice-versa. Entretanto, não se observa modificação das cores na direção da variação de x_1.

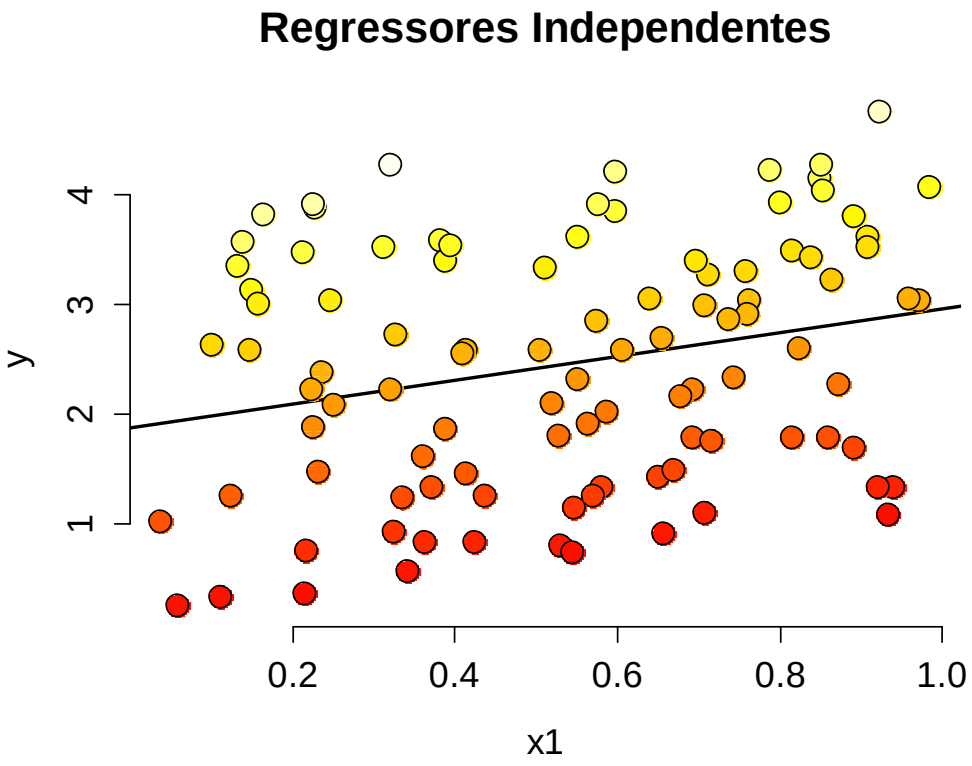


Figura 7.5 – Regressores independentes. ®

Em representação 3D (Figura 7.6), identifica-se relação linear perfeita entre de Y e x_2.

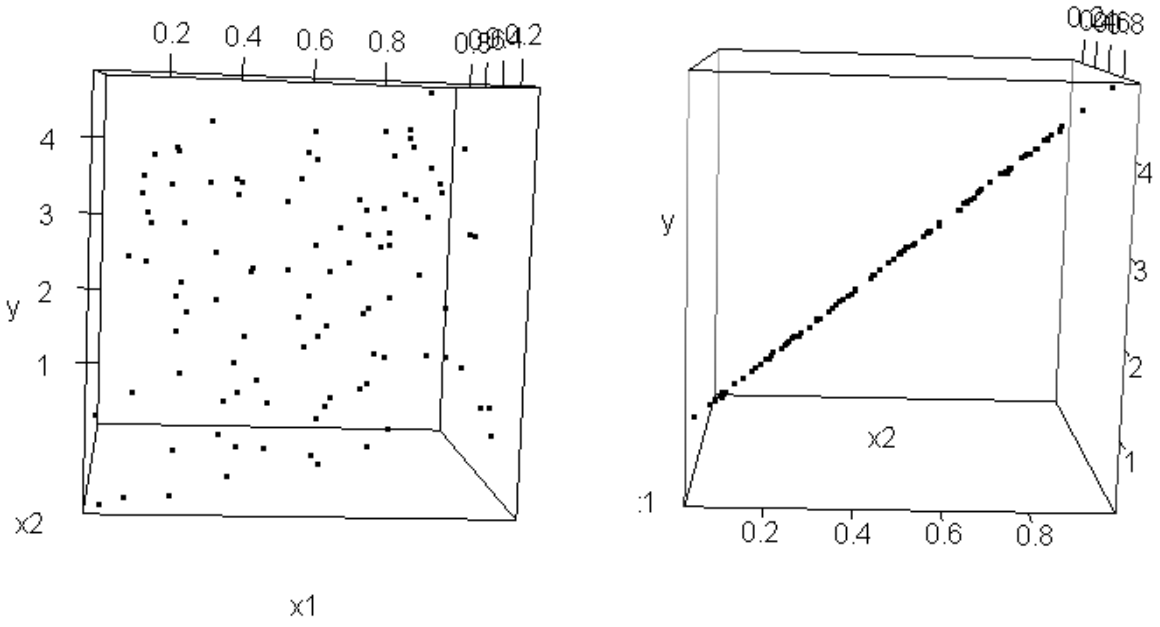


Figura 7.6 – Representação 3D do modelo com dois regressores independentes. ®

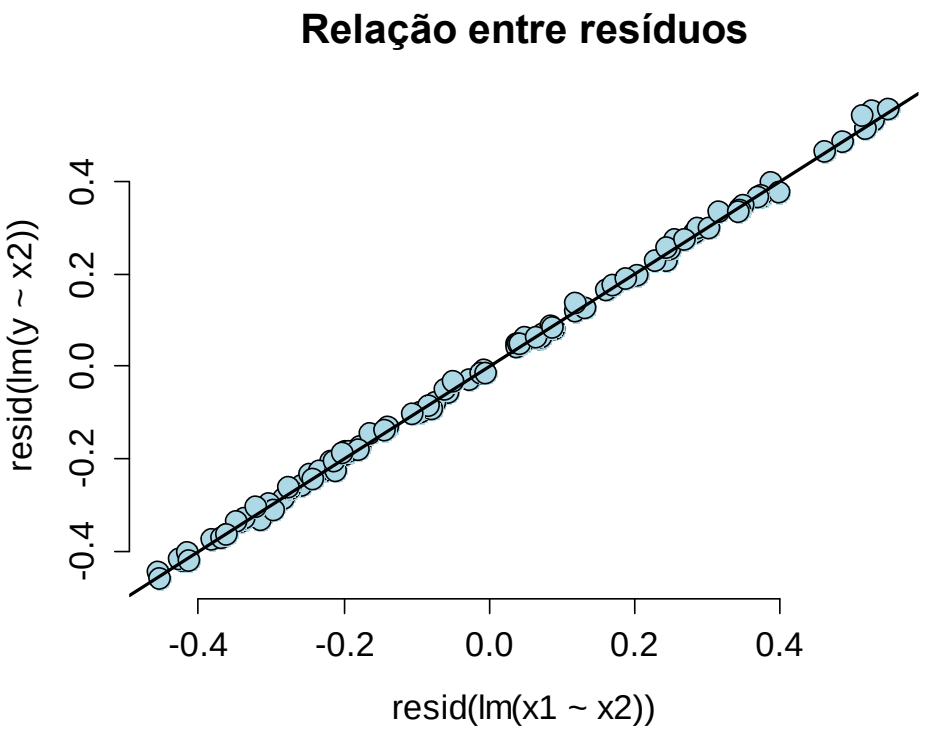


Figura 7.7 – Relação entre resíduos desconsiderando-se o regressor x_2. ®

Considerações:

- x_1 não é relacionado com x_2.
- x2 é fortemente relacionado com Y.
- O relacionamento ajustado entre x_1 e Y é praticamente inalterado quando se considera x_2.
- Quase não há variabilidade residual após desconsiderarmos x_2.

7.2.4 Análise dos Resíduos

Tanto na Regressão Linear Simples quanto na Regressão Múltipla, as suposições do modelo ajustado precisam ser validadas para que os resultados sejam confiáveis. Chamamos de Análise dos Resíduos o conjunto de técnicas utilizadas para investigar a adequabilidade de um modelo de regressão com base nos resíduos. Como visto anteriormente, o resíduo é dado pela diferença entre o dado observado e a resposta estimada do modelo (Equação (7.15)).

A ideia básica da análise dos resíduos é que, se o modelo for apropriado, os resíduos devem refletir as propriedades impostas pelo termo de erro do modelo. Tais suposições já foram abordadas na regressão simples e repetimos a seguir, tomando por referência a Equação (7.1) do modelo multivariado:

- ε_i e ε_j são independentes para $i \neq j$;
- $Var(\varepsilon_i) = \sigma^2$ deve ser constante (homocedasticidade);
- $\varepsilon_i \sim N(0, \sigma^2)$ (normalidade);
- O modelo é linear (Eaquação (7.3)).

Na Regressão Múltipla, além das suposições listadas acima, precisamos diagnosticar colinearidade e multicolinearidade entre os regressores. Quando não há relacionamento entre os regressores, dizemos que são ortogonais. Na prática, é muito difícil que as variáveis de entrada sejam ortogonais, o que, em princípio, não constitui um problema. Mas se as variáveis forem muito correlacionadas, as inferências baseadas no modelo de regressão podem ser equivocadas ou pouco confiáveis.

Portanto, é necessário identificar quais variáveis independentes apresentam forte correlação. Importante, também, é a identificação de pontos atípicos, ou *outliers*, que são observações numericamente distantes do restante dos dados e podem exercer influência significativa sobre os resultados da regressão.

As técnicas utilizadas para verificar as suposições mencionadas podem ser informais (como gráficos) ou formais (como testes). É importante que ambas as técnicas sejam consideradas na tomada de decisão visto que são complementares.

Exemplos de técnicas gráficas:

- Gráfico dos resíduos versus valores ajustados: verifica a homoscedasticidade do modelo, isto é, se σ^2 é constante.
- Gráfico dos resíduos padronizados versus quantis teóricos (QQ-plot): verifica a normalidade dos dados.

- Gráfico dos resíduos padronizados versus valores ajustados: verifica se existem *outliers* em *Y*.

  O resíduo padronizado é uma medida do quão forte é a diferença entre os valores observados e esperados. Para amostras suficientemente grandes, o resíduo padronizado pode ser comparado ao z-*score*. A padronização pode funcionar mesmo se suas variáveis não são normalmente distribuídas.

  De forma geral, interpretamos o gráfico procurando pelos residuos inferiores a -2 (em que a frequência observada da célula é menor do que a frequência esperada) e pelos residuos superiores a 2 (em que a frequência observada é maior do que a frequência esperada). Quando os resíduos estão além de +/- 3, significa que algo extremamente incomum está acontecendo. É como fazer a associação com as curvas de densidade de probabilidade: quando os dados são normalmente distribuídos, 95% destá dentro de 2 desvios-padrão. Algo maior do que isso provavelmente é um *outlier*.

- Gráfico de Alavancagem (Leverage): auxilia na identificação de pontos de influência.
- Gráfico dos resíduos versus a ordem de coleta dos dados: avalia a hipótese de independência dos dados.

A Figura 7.8 ilustra quatro gráficos produzidos pela função **plot**, do R, quando o resultado da regressão linear é passado como parâmetro. O termo *standardized residual* (resíduo padronizado – em tradução livre) é a razão entre o resíduo e o respectivo desvio padrão.

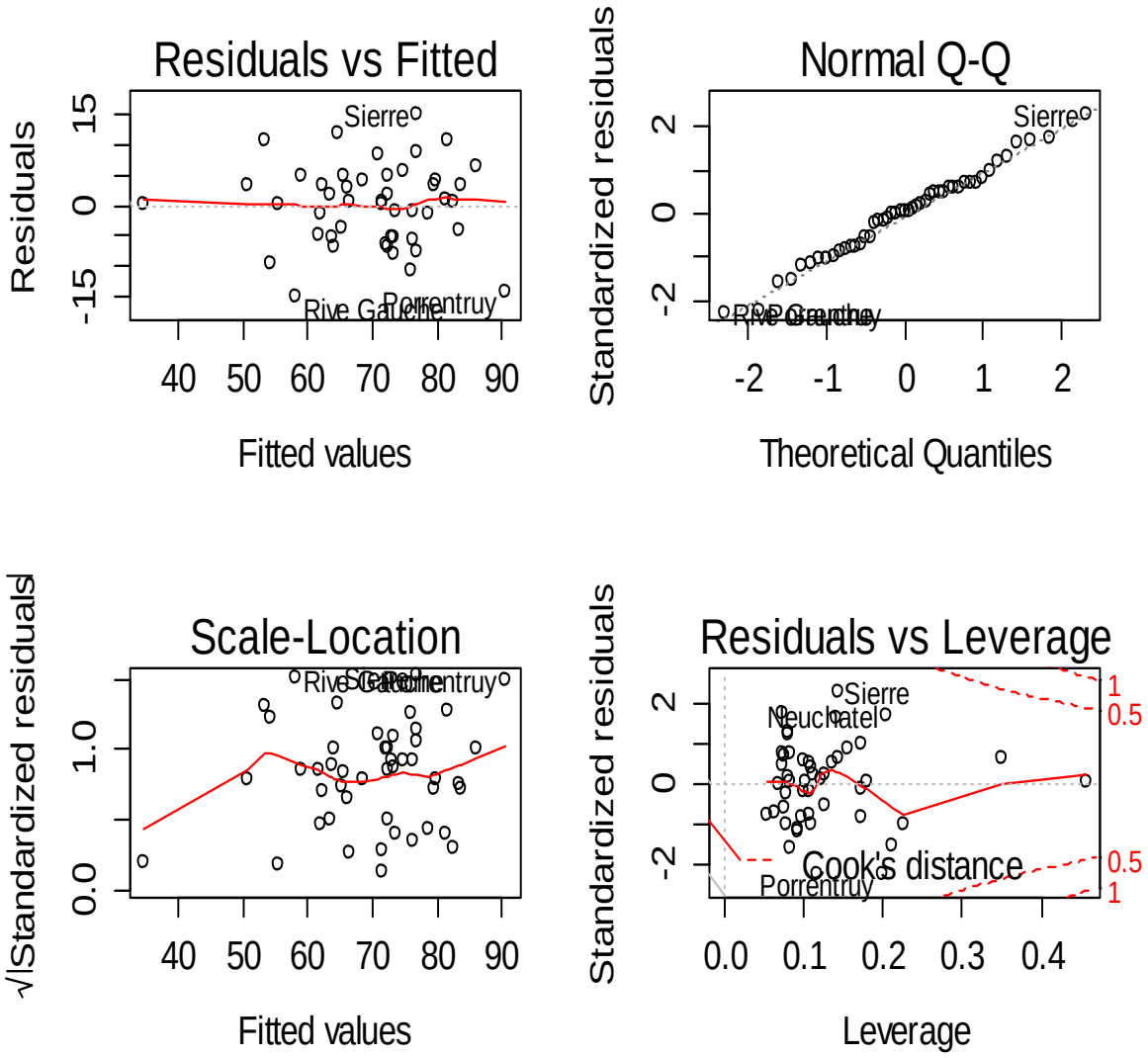


Figura 7.8 – Gráficos produzidos pela função **plot**. ®

Exemplos de técnicas formais:

- Testes de normalidade como Kolmogorov-Smirnov e Shapiro-Wilk.
- Teste de Durbin-Watson para testar independência dos resíduos. ®
- Teste de Breusch-Pagan e Goldfeld-Quandt para testar se os resíduos são homoscedásticos. ®
- Teste de falta de ajuste para verificar se o modelo ajustado é realmente linear.

Algumas técnicas formais serão exploradas nas seções seguintes.

7.2.4.1 Outliers: influência e alavancagem

Já definimos *outliers* como observações numericamente distantes do restante dos dados, resultantes de processos espúrios ou reais. No entanto, esse conceito é insuficiente e dois outros são necessários para caracterizar uma determinada observação como um *outlier*: influência e alavancagem.

A **influência** de uma observação pode ser pensada em termos do quanto os valores preditos das outras observações do conjunto poderiam ser diferentes caso a observação em questão não fosse incluída. A distância de Cook é uma boa medida da influência de uma observação e é proporcional à soma das diferenças ao quadrado entre as previsões feitas com todas as observações na análise e as previsões feitas deixando-se de fora a observação em questão. Se as previsões forem as mesmas com ou sem a observação em questão, então essa observação não tem qualquer influência sobre o modelo de regressão. Se as previsões diferirem bastante quando a observação não for incluída na análise, então essa observação é influente.

A **alavancagem** de uma observação é baseada em quanto o valor da observação na variável preditora difere da média (tendência central) da variável preditora. Quanto maior o poder de alavancagem de uma observação (quanto mais distante da média $\bar{x}$ da variável preditora), maior o potencial de ser uma observação influente. Por exemplo, uma observação em $x_i = \bar{x}$ não tem qualquer efeito sobre a inclinação da linha de regressão, independentemente da grandeza do seu valor na variável predita y_i. Por outro lado, uma observação em x_i distante de $\bar{x}$ tem potencial para afetar a inclinação.

A **distância** de uma observação baseia-se no erro de predição para a observação: quanto maior o erro de predição, maior a distância. A medida mais comumente utilizada para a distância é o *studentized residual* (resíduo estudentizado – em tradução livre; ver Apêndice III).

Mesmo uma observação com uma grande distância não terá tanta influência se a sua alavancagem for baixa. É a combinação do efeito de alavanca com a distância de observação que determina a sua influência.

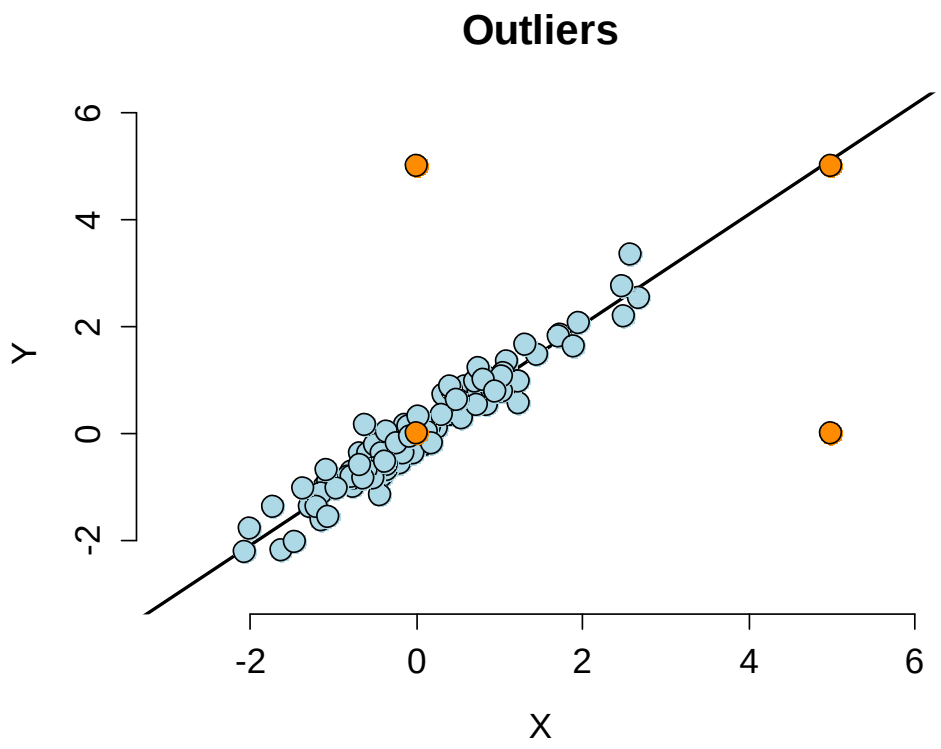


Figura 7.9 – Influência e alavancagem. ®

A Figura 7.9 destaca quatro pontos na cor laranja, cujas considerações sobre influência e alavancagem são traçadas a seguir:

- Ponto superior à esquerda: tem baixa alavancagem, pois se desvia na direção vertical em ponto próximo à média de *x*; tem baixa influência visto que sua inclusão não altera de forma significativa a reta de regressão (valores preditos); em um conjunto de dados com menor quantidade de elementos, a influência do ponto cresceria.
- Ponto inferior à esquerda: tem baixa alavancagem, baixa influência e não deve ser um *outlier* em sentido algum visto que se encontra onde há grande concentração de dados, ou, sobre o valor esperado.
- Ponto superior à direita: tem alta alavancagem, mas opta por não a exercer, visto que está sobre a reta de regressão, e, portanto, teria pouca influência real, estando em conformidade com a relação de regressão.
- Ponto inferior à direita: tem alto poder de alavancagem e exerceria grande influência se fosse incluído no ajuste.

A linguagem R provê funções para a medição da influência e alavancagem:

- rstandard – resíduos padronizados: resíduos divididos por seus desvios-padrão;
- rstudent – resíduos estudentizados onde os resíduos são divididos pelo respectivo desvio padrão e o i-ésimo ponto do conjunto de dados foi suprimido do cálculo do desvio padrão para produzir uma distribuição *t*.
- hatvalues – medidas de alavancagem.
- dffits – mudança na resposta predita quando o i-ésimo ponto é suprimido ao se fazer o ajuste do modelo.
- dfbetas – mudança nos coeficientes individuais quando o i-ésimo ponto é suprimido ao se fazer o ajuste do modelo.
- cooks.distance – alteração global nos coeficientes quando o i-ésimo ponto é suprimido ao se fazer o ajuste do modelo.
- resid – retorna os resíduos comuns.
- resid (fit) / (1 - hatvalues (fit)) – a diferença entre a resposta e a resposta prevista no i-ésimo dado, quando este não for incluído no ajuste do modelo.

Como utilizar esses recursos:

- Desconfie de regras simplistas para o diagnóstico de *outliers*. O uso dessas ferramentas é específico do contexto. Você deve sondar seus dados de maneiras diferentes para diagnosticar problemas diferentes.
- Nem todas as medidas têm escalas absolutas significativas. Você deve considera-las ao avaliar o comportamento dos dados sob análise.
- Padrões nos gráficos de resíduos geralmente indicam aspectos pobres de ajuste do modelo. Estes podem incluir:
    - ✓ Heteroskedasticity (variância não constante).
    - ✓ Termos faltantes no modelo.
    - ✓ Padrões temporais.
- Medidas de alavancagem (valores chapéu) podem ser úteis para diagnosticar erros de entrada de dados.
- Medidas de influência respondem a "como a exclusão ou inclusão de um ponto impacta em determinado aspecto particular do modelo".

7.2.4.2 Exemplos de diagnósticos

Caso 1

Considere o conjunto de pontos azuis da Figura 7.10, que tem formato aproximadamente circular. Nesses casos, a regressão linear não é bem definida e os coeficientes não são estatisticamente significativos (Quadro 7.4 – BLOCO 1). Adicionando-se o ponto laranja, a nova regressão linear passa a apresentar inclinação significativa (Quadro 7.4 – BLOCO 2). No terceiro bloco de código do Quadro 7.4, as funções **hatvalues** e **dfbetas** são utilizadas respectivamente, para medida de alavancagem e de influência. Podemos observar que, em comparação aos demais pontos (para exemplificar foram mostrados apenas os dez primeiros), o ponto (10, 10), que foi colocado na primeira posição do vetor, apresenta alto poder de alavancagem e exerce grande influência sobre o conjunto, criando uma relação de regressão forte onde não deveria haver uma.

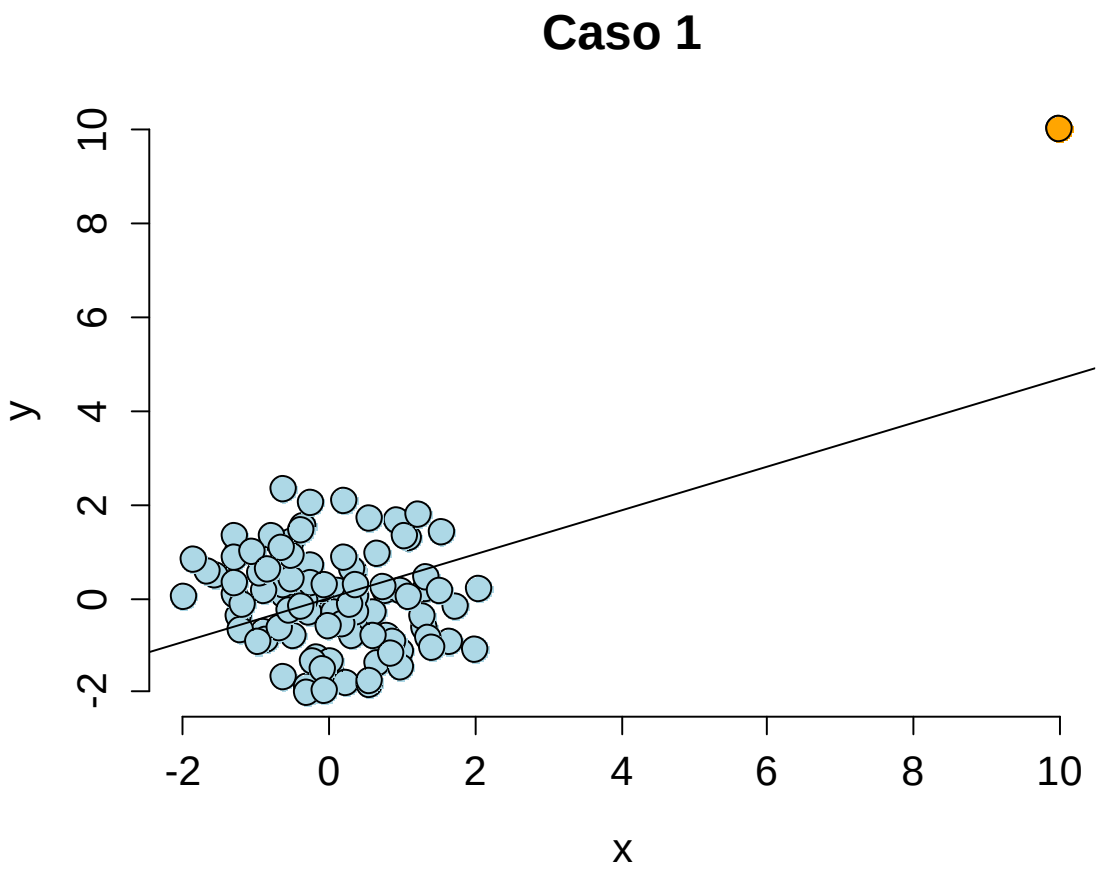


Figura 7.10 – Dados sem regressão linear: inluência e alavancagem - Caso 1. ®

Quadro 7.5 – Dados sem regressão linear: inluência e alavancagem - Caso 1.

```
# BLOCO 1: conjunto de pontos azuis - regressão linear
n <- 100; x <- c(rnorm(n)); y <- c(c(rnorm(n)))
summary(lm(y ~ x))$coefficients

              Estimate Std. Error   t value  Pr(>|t|)
(Intercept) 0.14200337 0.09728097 1.4597241 0.1475641
x           0.05013278 0.09584699 0.5230501 0.6021200

# BLOCO 2: inserção do ponto laranja - regressão linear
x <- c(10, x); y <- c(10, y)
fit <- lm(y ~ x)
summary(fit)$coefficients

              Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 0.1862674 0.11736374 1.587095 1.156801e-01
x           0.5086961 0.08278106 6.145078 1.680747e-08

# BLOCO 3: diagnóstico
# alavancagem
round(hatvalues(fit)[1 : 10], 3)

    1     2     3     4     5     6     7     8     9    10
0.495 0.010 0.029 0.011 0.015 0.013 0.011 0.023 0.011 0.015

# influência
round(dfbetas(fit)[1 : 10, 2], 3)

     1      2      3      4      5      6      7      8      9     10
 6.698  0.000 -0.116 -0.021  0.082  0.021  0.036  0.118 -0.052 -0.041
```

Caso 2

Considere, agora, o conjunto de pontos azuis da Figura 7.11, que claramente apresenta regressão linear bem definida, com inclinação estatisticamente significante (Quadro 7.5 – BLOCO 1). A adição do ponto laranja praticamente não altera a inclinação comparando-se com a nova regressão linear (Quadro 7.5 – BLOCO 2). Muito embora o ponto laranja tenha alto poder de alavancagem em comparação aos demais, essa influência não é exercida.

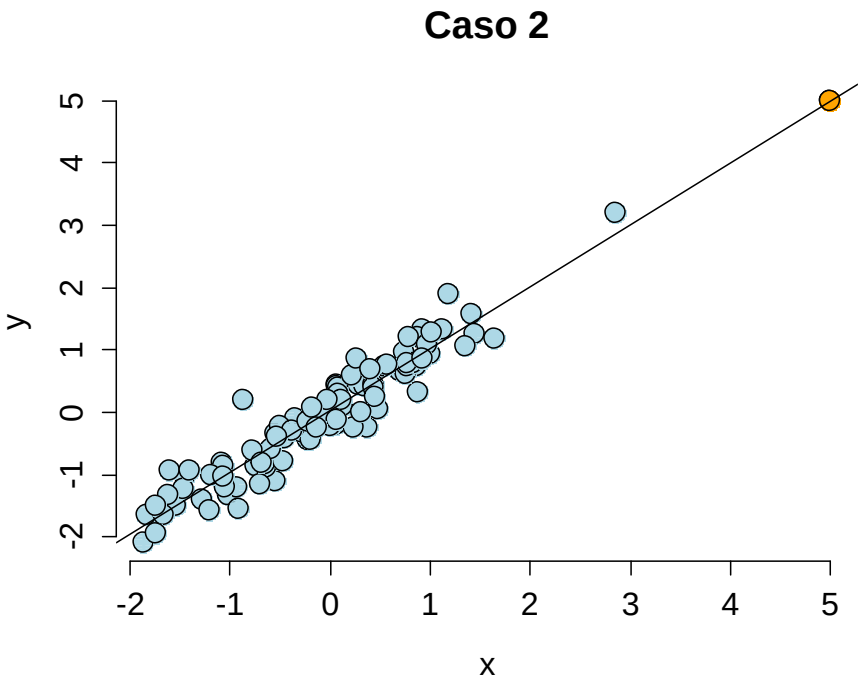


Figura 7.11 – Dados com regressão linear: inluência e alavancagem - Caso 2. ®

Quadro 7.6 – Dados com regressão linear: inluência e alavancagem - Caso 2.

```
# BLOCO 1: conjunto de pontos azuis - regressão linear
x <- rnorm(n); y <- x + rnorm(n, sd = .3)
summary(lm(y ~ x))$coefficients

             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 0.05517047 0.03018277  1.82788 7.061016e-02
x           0.99358364 0.02969068 33.46450 1.989048e-55

# BLOCO 2: inserção do ponto laranja - regressão linear
x <- c(5, x); y <- c(5, y)
fit <- lm(y ~ x)
summary(fit)$coefficients

             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 0.0549143 0.02979808  1.842881 6.833869e-02
x           0.9926780 0.02643408 37.552964 2.254631e-60

# BLOCO 3: diagnóstico
# alavancagem
round(hatvalues(fit)[1 : 10], 3)

    1     2     3     4     5     6     7     8     9    10
0.207 0.010 0.016 0.010 0.012 0.034 0.026 0.011 0.010 0.011

# influência
round(dfbetas(fit)[1 : 10, 2], 3)

     1      2      3      4      5      6      7      8      9     10
-0.034  0.014  0.134 -0.026  0.000  0.105  0.016  0.000  0.005  0.050
```

Caso 3

Considere o modelo linear multivariado gerado a partir do conjunto de dados disponível no site do Prof. Leonard A. Stefanski, do Departamento de Estatística da Universidade Estadual da Carolina do Norte. A investigação inicial dos dados mostra cinco variáveis quantitativas (Quadro 7.6 – BLOCO 1), cuja regressão múltipla de V1 como função das demais variáveis (Quadro 7.6 – BLOCO 2) apresenta coeficientes bastante significativos. Em princípio isso indicaria um bom modelo, contudo, ao se analisar o gráfico dos resíduos em função dos valores ajustados, observa-se claramente um padrão, o que fere o pressuposto de normalidade dos resíduos.

Quadro 7.7 – Análise de resíduos: identificação de padrões.

```
# BLOCO 1: leitura e exploração dos dados
dat <- read.table('orly_owl_Lin_4p_5_flat.txt', header = FALSE)
head(dat,4)

        V1         V2         V3         V4       V5
1 -0.75052 -0.282230  0.228190 -0.084136 -0.24748
2 -0.39380 -0.074787 -0.013689  0.072776 -0.36026
3 -0.15599  0.358390 -0.118070  0.013815 -0.65672
4 -0.68392 -0.059086 -0.060048 -0.231480 -0.03806

# BLOCO 2: regressão multivariada
summary(lm(V1 ~ . -1, data = dat))$coef

    Estimate Std. Error   t value     Pr(>|t|)
V2 0.9856157 0.12798121  7.701253 1.989126e-14
V3 0.9714707 0.12663829  7.671225 2.500259e-14
V4 0.8606368 0.11958267  7.197003 8.301184e-13
V5 0.9266981 0.08328434 11.126919 4.778110e-28

# BLOCO 3: valores ajustados versus resíduos
fit <- lm(V1 ~ . - 1, data = dat);
plot(predict(fit), resid(fit), pch = '.')
```

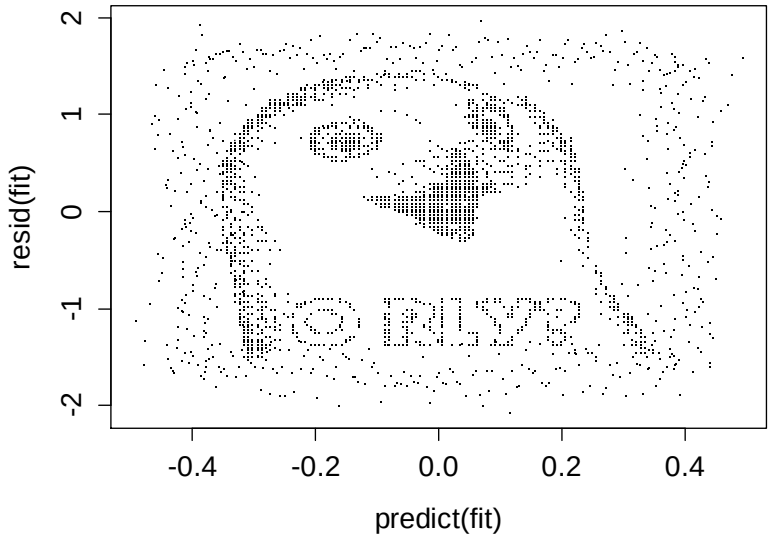


Figura 7.12 – Análise de resíduos: Orly. ®

7.2.5 Construindo um modelo multivariado

A escolha ou definição de um modelo linear multivariado é um processo difícil que consiste em avaliar as consequências de se incluir variáveis que não deveriam ser incluídas, ou de se omitir variáveis que deveriam estar no modelo.

Depende, também, do contexto em que a modelagem está inserida. Por exemplo, um modelo de predição tem um conjunto próprio de critérios necessários à interpretação e à padronização dos resultados, que pode diferir do conjunto de critérios de um modelo para estudar mecanismos específicos ou para estabelecer efeitos de causa.

Em geral, o interesse na modelagem recai sobre a parcimônia (ou simplicidade), em que se espera um modelo com representações de dados de fácil interpretação e que aumentem nosso entendimento sobre o fenômeno estudado. Em metáfora, um modelo é uma lente pela qual olhamos nossos dados.

Nesse sentido, qual é, então, o modelo certo? Há incontáveis formas pelas quais um modelo pode ser considerado inadequado. Aqui focaremos apenas na inclusão ou exclusão de variáveis.

Donald Rumsfeld, ex-Secretário de Defesa dos Estados Unidos, ficou lembrado por suas declarações durantre seu tempo no governo. Uma de suas frases mais famosas ilustra bem o desafio de se construir um bom modelo:

Há coisas conhecidas que conhecemos. São coisas que sabemos que sabemos.
Há coisas conhecidas que desconhecemos, isto é, coisas que sabemos que não sabemos.
Mas há, também, coisas desconhecidas que desconhecemos, ou seja,
coisas que não sabemos que não sabemos.
Donald Rumsfeld

No contexto da análise multivariada:

- conhecidos que conhecemos são regressores de que dispomos e sabemos que devemos considerar a inclusão no modelo;
- conhecidos que desconhecemos são regressores que gostaríamos de incluir no modelo, mas não dispomos;
- desconhecidos que desconhecemos são regressores sobre os quais nada sabemos e nem imaginamos que deveriam ser incluídos no modelo.

7.2.5.1 Regras gerais para a escolha do modelo

Como regras gerais, podemos considerar o seguinte:

1. A omissão de variáveis resulta no enviesamento dos coeficientes, a menos que a variável de estudo em que estamos interessados não seja correlacionada com os regressores omitidos.

   Por essa razão, a coleta dos dados em que se aplica um determinado tratamento (efeito que se deseja estudar) deve ser feita de forma aleatória, como uma tentativa de não correlacionar o indicador de tratamento com variáveis que não devem ser inseridas no modelo. Contudo, se existirem muitas variáveis espúrias no contexto, mesmo a coleta aleatória não será suficiente para neutralizar efeitos indesejados.

2. A inclusão de variáveis que não deveriam ser incluídas faz crescer o erro padrão dos coeficientes dos demais regressores.

   Na verdade, a inclusão de qualquer nova variável faz crescer o erro padrão dos outros regressores. Portanto, devemos evitar a inclusão de variáveis desnecessárias.

3. O modelo tende ao ajustamento perfeito quando a quantidade de regressores não redundantes se aproxima do tamanho da amostra.

4. R^2 cresce monotonicamente na medida em que novos regressores são incluídos, independentemente da importância do regressor.

5. A soma dos erros quadráticos decresce monotonocamente na medida em que novos regressores são incluídos.

7.2.5.2 Variação de R^2

A Figura 7.13 ilustra a variação de R^2 na medida em que novos regressores são incluídos. As simulações são independentes uma da outra e não existem relacionamentos entre os regressores dentro de cada simulação. Observa-se a conformidade com as regras 3 e 4, com R^2 crescendo monotonicamente e alcançando 100% de ajuste quando o quantitativo de regressores se iguala ao tamanho da amostra.

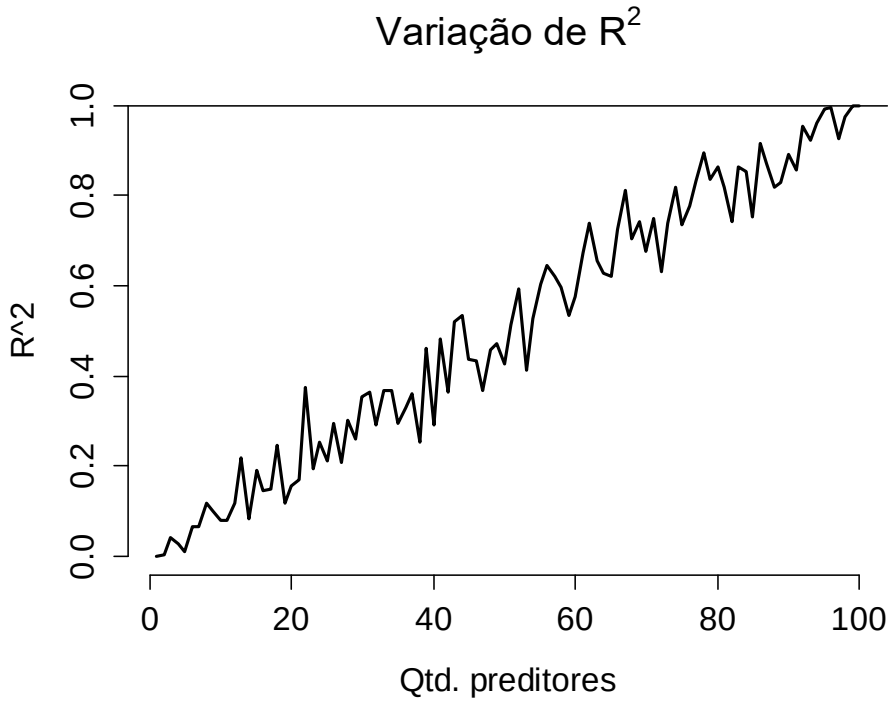


Figura 7.13 – Crescimento monotônico de R^2. ®

7.2.5.3 Inflação da Variância

O Quadro 7.7, BLOCO 1, ilustra o efeito de crescimento do erro padrão dos coeficientes de regressão, ou inflação da variância, mencionado na regra 2, quando as variáveis não são correlacionadas. Observa-se, nesse caso, um crescimento discreto do desvio padrão (ou da variância). No BLOCO 2 do Quadro 7.7, os regressores são correlacionados e a inflação da variância se dá de forma bem mais acentuada.

Quadro 7.8 – Inflação da variância.

```
# BLOCO 1: regressores não correlacionados
n <- 100; nosim <- 1000
x1 <- rnorm(n); x2 <- rnorm(n); x3 <- rnorm(n);
betas <- sapply(1 : nosim, function(i){
  y <- x1 + rnorm(n, sd = .3)
  c(coef(lm(y ~ x1))[2],
    coef(lm(y ~ x1 + x2))[2],
    coef(lm(y ~ x1 + x2 + x3))[2])
})
round(apply(betas, 1, sd), 5)

     x1      x1      x1
0.02718 0.02804 0.02851
```

```
# BLOCO 2: regressores correlacionados
n <- 100; nosim <- 1000
x1 <- rnorm(n); x2 <- x1/sqrt(2) + rnorm(n) /sqrt(2)
x3 <- x1 * 0.95 + rnorm(n) * sqrt(1 - 0.95^2);
betas <- sapply(1 : nosim, function(i){
  y <- x1 + rnorm(n, sd = .3)
  c(coef(lm(y ~ x1))[2],
    coef(lm(y ~ x1 + x2))[2],
    coef(lm(y ~ x1 + x2 + x3))[2])
})
round(apply(betas, 1, sd), 5)
```

```
     x1      x1      x1
0.03008 0.04406 0.09731
```

```
# BLOCO 3: exemplo da inflação na base de dados swiss
data(swiss);
fit <- lm(Fertility ~ Agriculture, data = swiss)
a <- summary(fit)$cov.unscaled[2,2]
fit2 <- update(fit, Fertility ~ Agriculture + Examination, data = swiss)
fit3 <- update(fit, Fertility ~ Agriculture + Examination + Education,
data = swiss)
c(summary(fit2)$cov.unscaled[2,2],
  summary(fit3)$cov.unscaled[2,2]) / a # normalização
```

```
[1] 1.891576 2.089159
```

No Quadro 7.7, BLOCO 3, a base de dados **swiss** é utilizada. Quando ao modelo inicial (`Fertility ~ Agriculture`) acrescenta-se a variável Examination (`Fertility ~ Agriculture + Examination`), verifica-se um crescimento de 89% na variância. Quando são adicionadas as variáveis Examination e Education (`Fertility ~ Agriculture + Examination + Education`), constata-se um acréscimo de mais de 100%.

Quando regressores adicionais inseridos ao modelo são ortogonais aos regressores de interesse, então não ocorre inflação da variância. Denomina-se **fator de inflação da variância** (*verified inflation factor* – VIF) o crescimento na variância decorrente da inclusão de um novo regressor, comparando-se com a situação ideal onde as variáveis de regressão não são linearmente relacionadas. É usado para descrever o quanto de multicolinearidade (correlação entre preditores) existe em uma análise de regressão. A multicolinearidade é um problema porque pode aumentar a variação dos coeficientes de regressão, tornando-os instáveis e difíceis de interpretar.

É importante lembrar que a inflação da variância é apenas parte da história quando estamos construindo um modelo. Por vezes, desejamos incluir certas variáveis mesmo que dramaticamente inflacionem a variância.

A linguagem R dispõe da função **vif** para o cálculo do fator de inflação da variância. O Quadro 7.8 determina os VIFs para as variáveis da base de dados **swiss**. Observam-se valores altos para as variáveis Examination e Education visto que são correlacionadas entre si.

Quadro 7.9 – Cálculo do fator de inflação da variância.

```
# BLOCO 1: Cálculo dos VIFs
library(car)
fit <- lm(Fertility ~ . , data = swiss)
vif(fit)

    Agriculture  Examination    Education   Catholic   Infant.Mortality
       2.284129  3.675420       2.774943    1.937160   1.107542
```

Em síntese, supondo-se que um modelo é linear com erros aditivos independentes e identicamente distribuídos, com variância finita, podemos descrever matematicamente o impacto de omitirmos variáveis necessárias ou incluirmos variáveis desnecessárias:

- Se o modelo é subdimensionado, a estimativa de variância é tendenciosa.
- Se o modelo é superdimensionado, incluindo todas as covariantes necessárias e/ou desnecessárias, a estimativa de variância é não tendenciosa, no entanto, a inflação da variação é maior quando incluirmos variáveis desnecessárias.

7.2.5.4 Selecionado um modelo multivariado

A seleção automática das covariantes de um modelo é uma tarefa difícil. Depende muito do que se deseja explorar.

Quando a parcimônia é o mais importante, um bom desenho deve prever a aleatoriedade e a extratificação. Um bom desenho pode muitas vezes eliminar a necessidade de modelos complexos, embora muitas vezes o controle sobre o desenho seja limitado.

Também podem ser utilizados métodos de redução dimensional para a redução de espaços complexos de covariáveis, ou testes da razão de verossimilhança, para avaliar se um preditor é importante. Na regressão linear o interesse está no valor de VRR (Equação (6.23)), que é a variação explicada pela regressão, ou soma dos quadrados da diferença entre o valor estimado pelo modelo e o valor quando desconsideramos a influência do preditor. Um valor alto de VRR sugere que a variável independente é importante, caso contrário, a variável independente não é útil na predição da variável resposta.

Para avaliarmos a significância do modelo como um todo, utilizamos a análise de variância (ANOVA). A análise de variância é baseada na decomposição da soma de quadrados e nos graus de liberdade associados a variável resposta *Y*. Em outras palavras, o desvio de uma observação em relação à média pode ser decomposto como o desvio da observação em relação ao valor ajustado pela regressão mais o desvio do valor ajustado em relação à média.

$$y_i - \bar{y} = (y_i - \bar{y} + \hat{\mu}_i - \hat{\mu}_i) = (y_i - \hat{\mu}_i) + (\hat{\mu}_i - \bar{y}) \tag{7.20}$$

Elevando cada componente ao quadrado, chegamos à mesma Equação (6.24). O Quadro 7.9 ilustra a utilização da ANOVA para análise do modelo de regressão.

Quadro 7.10 – Análise do modelo de regressão com ANOVA.

```
# ANOVA
fit1 <- lm(Fertility ~ Agriculture, data = swiss)
fit3 <- update(fit1, Fertility ~ Agriculture + Examination + Education,
data = swiss)
fit5 <- update(fit1, Fertility ~ Agriculture + Examination + Education +
Catholic + Infant.Mortality, data = swiss)
anova(fit1, fit3, fit5)

Analysis of Variance Table

Model 1: Fertility ~ Agriculture
Model 2: Fertility ~ Agriculture + Examination + Education
Model 3: Fertility ~ Agriculture + Examination + Education + Catholic +
    Infant.Mortality
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     45 6283.1
2     43 3180.9  2    3102.2 30.211 8.638e-09 ***
3     41 2105.0  2    1075.9 10.477 0.0002111 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

No Quadro 7.9 observamos que a variação na soma dos quadrados dos resíduos (RSS – *residual sum of squares*) é significativa quando comparamos os modelos com diferentes graus de liberdade, o que significa que devemos, sim, adicionar as variáveis consideradas.

Quando os modelos de interesse são aninhados e sem muitos parâmetros para diferenciação, a utilização de testes de razão de verossimilhança é bastante controversa. Nesses casos, devemos adotar uma abordagem investigativa, onde adicionamos regressores um a um, de acordo com o objeto de estudo, verificamos sua influência e tomamos uma decisão de acordo com o contexto da pesquisa. Esta abordagem nem sempre é sistemática, mas tende a nos ensinar muito sobre os dados (ver exemplo).

7.2.6 Interações em Regressão

A adição de termos de interação a um modelo de regressão pode expandir muito a compreensão das relações entre as variáveis no modelo e permitir que mais hipóteses sejam testadas.

https://www.theanalysisfactor.com/interpreting-interactions-in-regression/
https://www.theanalysisfactor.com/clarifications-on-interpreting-interactions-in-regression/
https://www.theanalysisfactor.com/interactions-categorical-and-continuous-variables/

https://www.theanalysisfactor.com/interactions-categorical-and-continuous-variables/
https://www.theanalysisfactor.com/interactions-categorical-and-continuous-variables/

https://gssdataexplorer.norc.org/projects/502634

http://gss.norc.org/get-the-data/spss
http://www.milanor.net/blog/how-to-open-an-spss-file-into-r/

7.3 LABORATÓRIO 7

Carregue a base de dados **Prestige** do pacote **car**. Utilizando técnicas gráficas e formais:

a) Conheça a estrutura dos dados e explore as variáveis quantitativas e qualitativas.
b) Faça análises marginais e multivariadas.
c) Verifique as premissas do modelo linear.
d) Com base nas análises, proponha um ou mais modelos lineares multivariados. Explique a sua escolha.
e) Utilize o(s) modelo(s) proposto(s) para fazer pelo menos uma predição.

8 REGRESSÃO LOGÍSTICA

Frequentemente estamos interessados em resultados binários, que têm dois valores, ou chamados de resultados de Bernoulli (binomial):

- eleito / não eleito
- vivo / morto
- venceu / perdeu
- maligno / benigno
- sucesso / insucesso
- ...

Estudos dessa natureza são denominados problemas de classificação. Podemos tentar resolver problemas de classificação com as técnicas de regressão que vimos estudando. Por exemplo, considere os dados fictícios da Tabela 8.1.

Tabela 8.1 – Arrecadação de campanha por candidato.

Candidato	**Partido**	**Arrecadação** **(x 1000 Reais)**	**Status**
Candidato 1	PV	40	Não Eleito
Candidato 2	PCdoB	30	Não Eleito
Candidato 3	PL	20	Não Eleito
Candidato 4	PSB	10	Não Eleito
Candidato 5	PT	95	Eleito
Candidato 6	PMDB	80	Eleito
Candidato 7	PSDB	70	Eleito
Candidato 8	PP	60	Eleito

Suponha que desejamos prever a possibilidade de um determinado candidato ser eleito com base no valor arrecadado em campanha. A partir da Tabela 8.1, podemos traçar o gráfico de dispersão do resultado da eleição em função do valor arrecadado e, em seguida, ajustar

uma reta de regressão linear. Uma vez que estamos diante de uma resposta dicotômica (não eleito – 0; eleito – 1), podemos estabelecer um limiar de classificação no meio da margem (0,5) acima do qual classificamos a resposta como 1 (eleito), abaixo do qual classificamos a resposta como 0 (não eleito). A Figura 8.1 ilustra essa situação e aponta que o valor limiar da arrecadação de campanha é de R$ 50.620,00. Isso significa dizer que valores arrecadados em campanha acima desse limiar são, em princípio, suficientes para eleger o candidato.

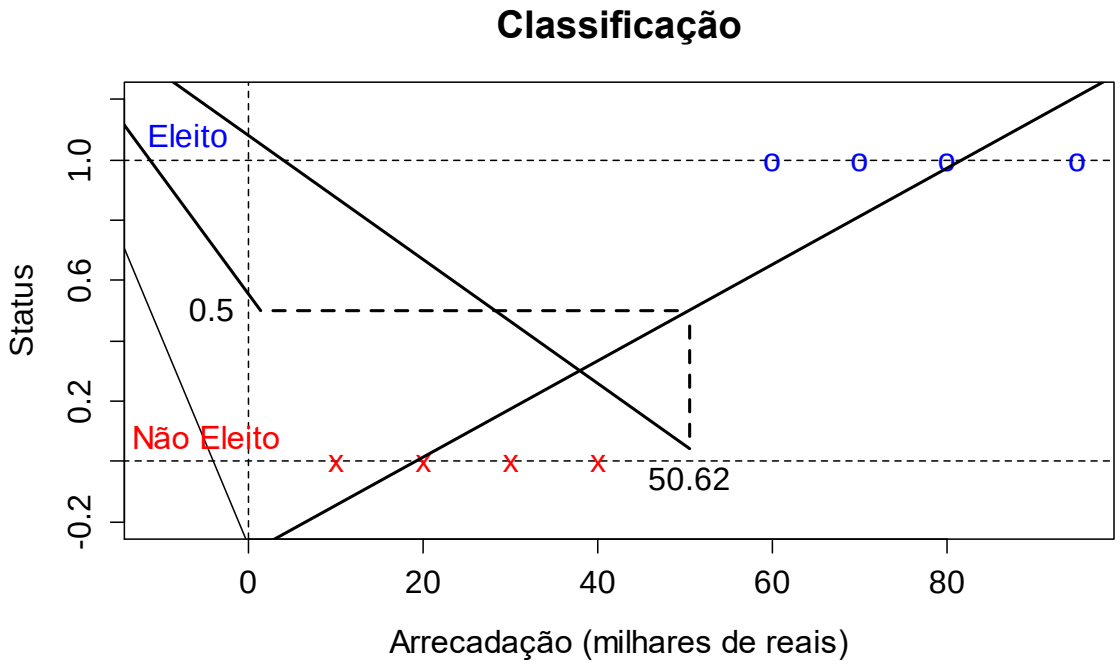


Figura 8.1 – Classificação com regressão linear: caso 1. ®

Considere, agora, que ao conjunto de dados da Tabela 8.1 foram acrescentadas duas novas observações nas quais as arrecadações foram, respectivamente, R$ 185.000,00 e R$ 190.000,00. A nova reta de regressão (Figura 8.2) ajusta o nosso modelo classificador para um novo limiar de R$ 60.830,00 e alguns efeitos indesejados começam a aparecer. Por exemplo, o candidato que arrecadou R$ 60.000,00, e que foi eleito, passaria a ser classificado como não eleito. Outro problema do modelo linear é que a faixa de respostas possíveis pela regressão pode retornar valores no intervalo [-∞,∞], o que não se mostra adequado quando o valor esperado é dicotômico [0, 1]. Também, a regressão linear assume que a variância do erro é constante e independente dos valores assumidos pelos preditores, o que não ocorre quando a variável resposta é binomial. Por fim, os resíduos não podem

ser normalmente distribuídos, já que a variável resposta assume apenas dois valores possíveis.

Logo, estamos diante de um problema que na verdade é não linear e precisamos de um algoritmo que se ajuste à resposta dicotômica e que garanta maior estabilidade ao limiar de classificação. Contudo, é desejável que o modelo seja simples, interpretável e que represente bem os dados.

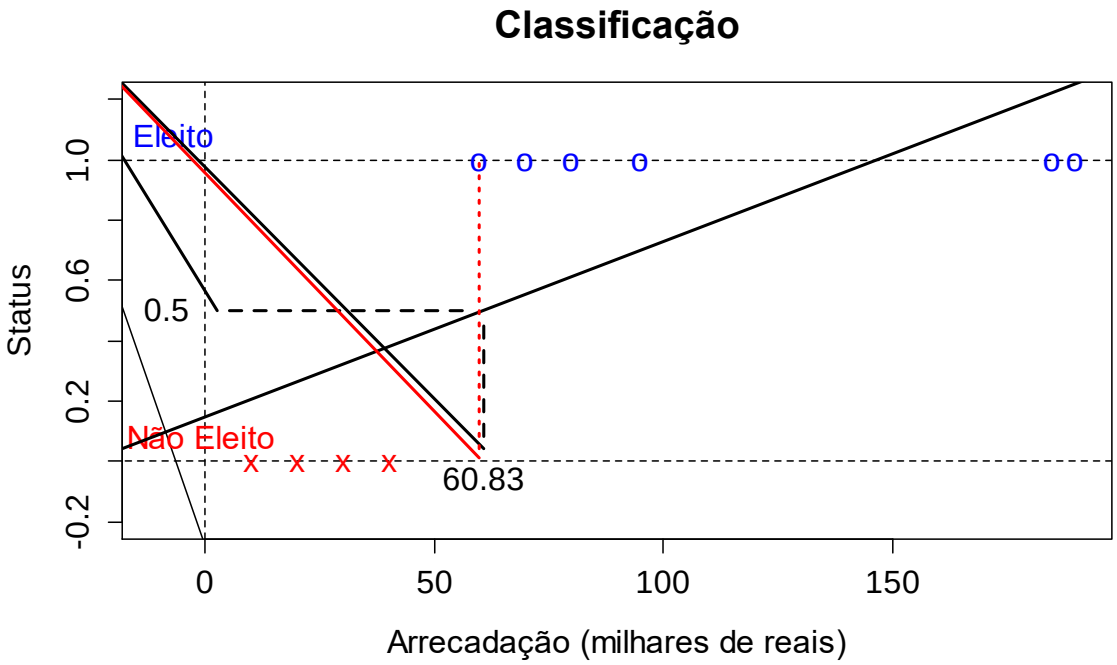


Figura 8.2 – Classificação com regressão linear: caso 2. ®

A técnica de **modelos lineares generalizados** (MLG), desenvolvida por Nelder e Wedderburn (1972), permite a generalização ou flexibilização dos modelos lineares clássicos, de forma que toda a estrutura para a estimação e predição em modelos lineares possa ser estendida para modelos não lineares. Os modelos lineares clássicos são, na verdade, casos especiais de modelos lineares generalizados.

Os MLGs envolvem uma variável resposta univariada, variáveis explanatórias e uma amostra aleatória de *n* observações independentes, de forma que:

1. a variável resposta, dependente, ou **componente aleatório** do modelo, tem uma distribuição de probabilidade pertencente à família de distribuições exponenciais

canônicas que engloba a normal, gama e normal inversa, para dados contínuos; a binomial para proporções; Poisson e binomial negativa para contagens;

2. as variáveis explanatórias especificam uma função linear entre as variáveis independentes, constituindo o **componente sistemático** (ou determinístico) do modelo, denominado **preditor linear**;
3. a ligação entre os componentes aleatório e sistemático é feita por meio de uma função matemática, denominada **função de ligação**.

O componente aleatório consiste nas observações $(y_1,\ldots,y_n)$ da variável aleatória *Y*.

O componente sistemático do MLG é definido por um vetor de preditores lineares $\eta=(\eta_1,\ldots,\eta_n)$ que está associado ao conjunto das variáveis independentes por meio de um modelo linear $\eta=X\beta$, onde *X* é a matriz de *p* variáveis independentes x_k, com $k\in\{1,\ldots,p\}$, onde cada variável tem *n* observações, e $\beta=(\beta_1,\ldots,\beta_p)$ é o vetor de parâmetros do modelo.

$$\eta_i=\sum_{k=1}^{p}\beta_k x_{ki} \tag{8.1}$$

A função de ligação descreve a relação entre o valor esperado quando os dados são modelados pelo componente sistemático e o valor esperado do conjunto de observações efetuadas (componente aleatório). Seja $\mu_i=\mathrm{E}(y_i\mid x_i)$ o valor esperado pelo componente sistemático, com $i\in\{1,\ldots,n\}$, então a função de ligação *g* é função de μ, monotônica e diferenciável, definida por $g(\mu)$, e o valor esperado da *i*-ésima observação é $g(\mu_i)$. Dessa forma, a função de ligação conecta os valores esperados das observações às variáveis explanatórias. As regressões linerares simples e com múltiplos regressores são MLGs em que $g(\mu)=\mu$.

Considerações teóricas e práticas sugerem que quando a variável resposta é binária, devemos utilizar a função de ligação **sigmoide**, ou **função logística**.

$$g(\mu(x)) = \frac{e^{\mu(x)}}{1+e^{\mu(x)}} = \frac{1}{1+e^{-\mu(x)}} \tag{8.2}$$

A curva sigmoide tem formato em S e suas extremidades convergem assintoticamente para os valores 0 e 1, o que é desejável para um classificador binário. A Figura 8.3 ilustra a sigmoide da Equação (8.2) para $\mu(x) = x$ (ou seja, o preditor linear tem um único regressor x e $\beta_0 = 0$ e $\beta_1 = 1$).

Interpretando a curva sigmoide, podemos dizer que $g(\mu(x))$ é a **probalidade estimada *p* de que a variável resposta y_i seja igual a 1**, dado que x é igual a x_i e é parametrizado por $\beta_0 = 0$ e $\beta_1 = 1$ (em notação formal, $p(Y = 1 \mid x, \beta)$).

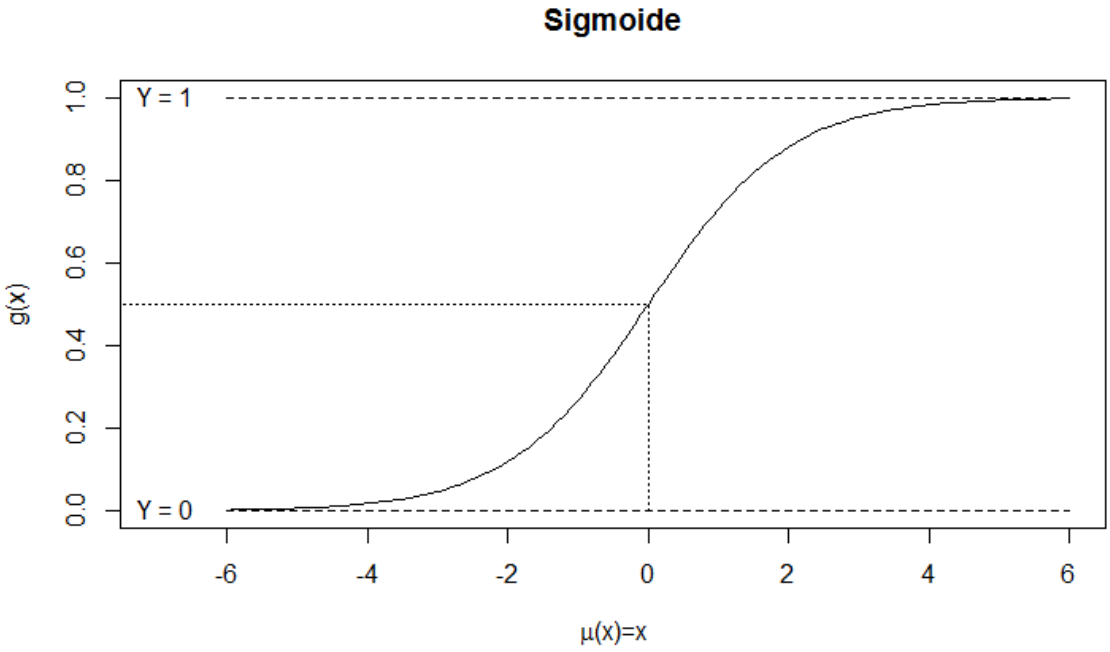


Figura 8.3 – Curva sigmoide para $\mu(x) = x$. ®

A função *g* se comporta como a função de distribuição de probabilidade de densidade simétrica, com ponto médio zero; como $\mu(x)$ se move no eixo dos números reais, *g* cresce monotonicamente entre os limites de 0 e 1. O significado de *g* varia de acordo com a definição das variáveis. Na versão eleito/não eleito, *g* é a probabilidade de um resultado

binário (eleição ou não de um candidato); o componente sistemático é $\mu(x) = \beta_0 + \beta_1 x$, sendo *x* uma variável de estímulo ou de exposição contínua (como a quantidade de verba arrecadada); β_0 determina a localização da curva no eixo *x* e β_1, a sua inclinação. Execute o código do Quadro 8.1 para verificar o comportamento do sigmoide na medida em que β_0 e β_1 variam.

Quadro 8.1 – Comportamento do sigmoide com a variação de β_0 e β_1.

```
# Sigmoide
library(manipulate)
x <- seq(-10, 10, length = 1000)
manipulate(
  plot(x, exp(beta0 + beta1 * x) / (1 + exp(beta0 + beta1 * x)),
       type = "l", lwd = 3, frame = FALSE),
  beta1 = slider(-4, 4, step = .1, initial = 2),
  beta0 = slider(-4, 4, step = .1, initial = 0)
)
```

No entanto, como a função de ligação é contínua (monotônica e diferenciável), precisamos definir um limiar para fazermos a classificação adequadamente. Em outras palavras, precisamos definir uma **fronteira de decisão**.

Para a função logística, é razoável predizermos $Y = 1$ quando $g(\mu(x)) \geq 0{,}5$, e $Y = 0$ quando $g(\mu(x)) < 0{,}5$, considerando que ambos são equiprováveis. Partindo desse pressuposto e olhando a Figura 8.3, podemos também predizer a variável resposta a partir do eixo *x* (domínio), pois sabemos que $g(\mu(x)) \geq 0{,}5$ sempre que $\mu(x) \geq 0$, e que $g(\mu(x)) < 0{,}5$ sempre que $\mu(x) < 0$. Logo, a fronteira de decisão deve ser tal que atenda as condições

- prediga $y = 1$ se $\mu(x) \geq 0$;
- prediga $y = 0$ se $\mu(x) < 0$.

e é com base nessas condições que os parâmetros β são estimados.

A esse algoritmo de estimativa dos parâmetros β damos o nome de **regressão logística**, que é utilizada quando a variável resposta é qualitativa com dois (ou mais) resultados

possíveis. Esses parâmetros são estimados por meio do ajuste do modelo baseado nos preditores disponíveis e nos dados observados. O modelo a ser escolhido consiste naquele em que, inseridos os valores dos preditores, os valores de *Y* mais se aproximam dos observados. Especificamente, os parâmetros são estimados pelo algoritmo de máxima verossimilhança (*maximum-likelihood estimation*), o qual seleciona coeficientes que fazem com que os dados observados sejam os mais prováveis de acontecer.

Apenas para ilustrar, considere a Figura 8.4. Trata-se de um conjunto de dados em que as observações representadas por X possuem determinada característica e as observações representadas por ○ não a possuem. A variável resposta *Y* é uma função de dois regressores, x_1 e x_2. Logo, o preditor linear (componente sistemático) é o valor esperado $E[Y] = \mu(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ da regressão múltipla $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ e é definido por β_0, β_1 e β_2.

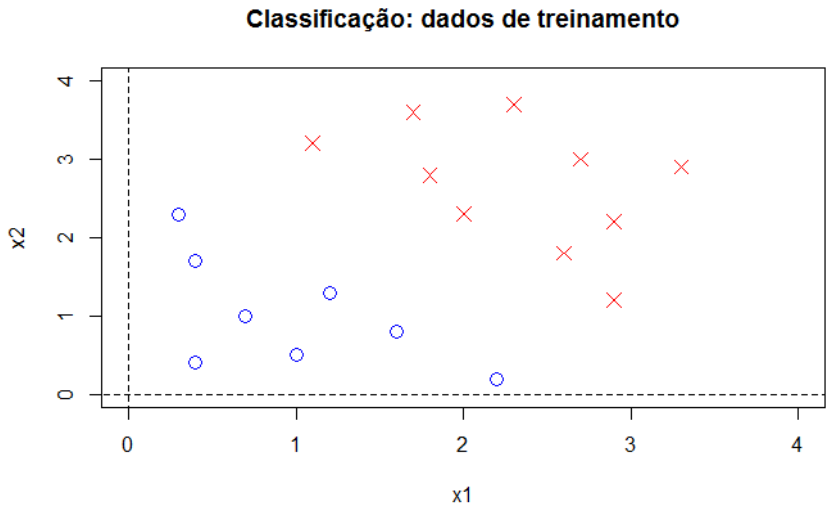


Figura 8.4 – Classificação: observações X possuem determinada característica e as observações ○ não a possuem. ®

Agora, considere que a regressão logística tenha estimado $\hat{\beta}_0 = -3$, $\hat{\beta}_1 = 1$ e $\hat{\beta}_2 = 1$: essas estimativas fazem sentido? Das condições necessárias à definição da fronteria de decisão, sabemos que para predizer $Y = 1$, temos que $-3 + 1x_1 + 1x_2 \geq 0$ ou $x_1 + x_2 \geq 3$, sendo a

fornteira definida por $x_1 + x_2 = 3$. Ao traçarmos essa reta, temos como resultado a linha verde na Figura 8.5, que certamente faz todo sentido.

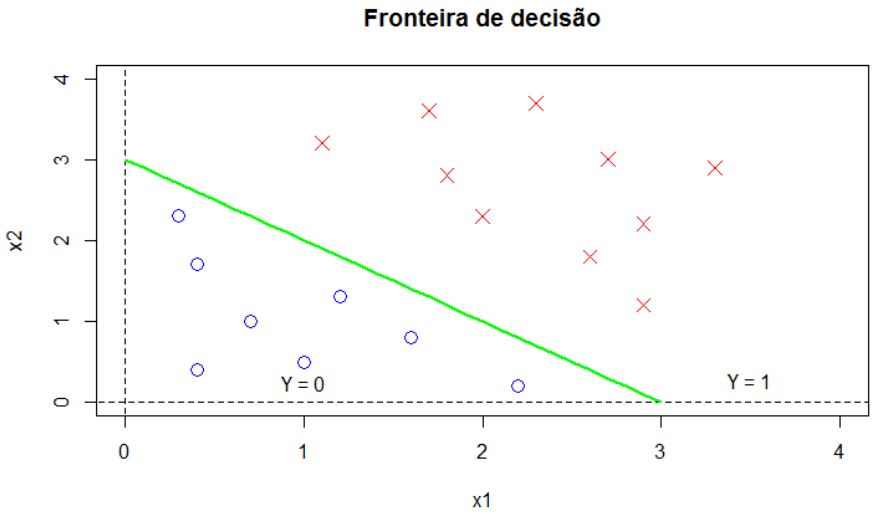


Figura 8.5 – Fronteira de decisão. ®

Os dados de treinamento são utilizados para definir os parâmetros β e esses parâmetros são utilizados para definir a fronteira de decisão. A fronteira de decisão é uma propriedade dos parâmetros, não dos dados.

[*em breve vou inserir aqui texto sobre funções de custo e estimativas dos parâmetros com gradient descent, que para a regressão linear é o algoritmo dos mínimos quadrados e cuja sistemática é a mesma para a regressão logística, apenas alterando-se as funções de custo para funções logarítmicas:*

$$\text{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$]

Como vimos, o que distingue um modelo de regressão logístico e binário de um modelo de regressão linear é a variável aleatória dependente *Y*, que no modelo logístico se apresenta na escala nominal (ver definição de escala niominal no Quadro 2.2). Enquanto na regressão linear o modelo tenta prever o valor de *Y* a partir de um ou mais preditores, na regressão logística o modelo prevê a probabilidade de *Y* ocorrer, conhecidos os valores dos preditores. Essa diferença se reflete na escolha do modelo paramétrico e nas hipóteses envolvidas. Ao trabalharmos com dados de natureza binária, a esperança condicional

$\mathrm{E}(Y \mid x)$ deve ser menor ou igual a 1 e maior ou igual a 0 ($0 \leq \mathrm{E}(Y \mid x) \leq 1$) e os resultados são interpretados em termos de probabilidades.

Uma forma de ganharmos interpretabilidade para o modelo logístico é trabalharmos com o conceito de chance. Chances são bastante interpretáveis e parcimoniosas.

Chance (ou *odds* – chances) é definida como a razão entre duas probabilidades: a probabilidade de sucesso *p* e a probabilidade de fracasso *q*, onde *q* = 1 - *p*

$$odds = \frac{p}{1-p} = \frac{p}{q} \tag{8.3}$$

Por exemplo, se a probabilidade de sucesso for 0,75, então a chance é de 3:1 (três para um, ou seja, para cada três ocorrências de sucesso, há uma de fracasso). Com base na definição, as propriedades das chances são:

- Se *p* é igual *q*, então *odds*(successo) = 1 (ou 1:1).
- Se *p* < *q*, então *odds*(successo) < 1.
- Se *p* > *q*, então *odds*(successo) > 1.
- Ao contrário da probabilidade, que não pode exceder 1, não há limite superior para *odds*.

Repare que se aplicarmos a definição de chances à variáveil resposta *Y*, ou *odds*(*Y*), estaremos limitados ao intervalo [0,∞], o que ainda nos impede de aplicar a regressão linear ao modelo. Então, por que trabalharmos com chances?

Uma razão para se utilizar chances está relacionada à transformação central para o estudo da regressão logística, ou transformação *logit* de *p*, definida como o logaritmo natural das chances.

$$logit(p) = \ln\left[\frac{p}{1-p}\right] = \ln\left(\frac{p}{q}\right) \tag{8.4}$$

A transformação *logit* é também chamada de log das chances (ou *log odds*) e suas proprieadades são derivadas das propriedades das chances.

- Se *odds*(sucesso) = 1, então *logit*(*p*) = 0.
- Se *odds*(sucesso) < 1, então *logit*(*p*) < 0.
- Se *odds*(sucesso) > 1, então *logit*(*p*) > 0.
- A transformação *logit* falha se *p* = 0.

A transformação *logit*(*p*) está definida no intervalo [-∞,∞], logo, podemos assumir um modelo linear nos preditores. Portanto, ao aplicar a transformação *logit* à regressão logística, ganhamos em parcimônia, ou seja, ganhamos poder de interpretação do modelo.

Considere o modelo em que o componente sistemático é definido por um único regressor. A função *g* é dada por

$$g(\mu(x)) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \tag{8.5}$$

Lembrando que *g* é a probabilidade estimada *p* de que *Y* =1, ou seja, *g* é a probabilidade *p* de sucesso de *Y*, então a transformação *logit* é dada por

$$logit(p) = \ln\left[\frac{p}{1-p}\right] = \ln\left[\frac{g(\mu(x))}{1-g(\mu(x))}\right] = \mu(x) = \beta_0 + \beta_1 x \tag{8.6}$$

A função *logit* tem muitas das propriedades da função $\mu(x)$ do modelo linear pois é linear em seus parâmetros e contínua. Os parâmetros β_0 e β_1 têm significados similares aos análogos da regressão linear.

Segue-se, de (8.6), que β_0 corresponde a *log odds* (ou *logit(p)*) quando *x* = 0, ou seja, corresponde às chances no domínio logarítmico quando $x = 0$. Para se encontrar as chances no domínio $\Re$, utiliza-se a função e^{β_0}, que é a função inversa de *logit(p)*.

Efetuando-se o mesmo procedimento para o caso em que o preditor varia de uma unidade, ou seja, varia de $x = x_i$ a $x = x_i + 1$, temos

$$\begin{aligned} \ln\left[\frac{g(\mu(x_i+1))}{1-g(\mu(x_i+1))}\right] - \ln\left[\frac{g(\mu(x_i))}{1-g(\mu(x_i))}\right] &= \mu(x_i+1) - \mu(x_i) \\ &= \beta_0 + \beta_1(x_i+1) - (\beta_0 + \beta_1 x_i) \\ &= \beta_1 \end{aligned} \tag{8.7}$$

Então, β_1 é o incremento no valor de *log odds* (8.6), ou taxa de variação de *log odds* decorrente do aumento de uma unidade em *x*.

A Equação (8.7) também pode ser escrita como

$$\ln\left[\frac{\dfrac{g(\mu(x_i+1))}{1-g(\mu(x_i+1))}}{\dfrac{g(\mu(x_i))}{1-g(\mu(x_i))}}\right] = \beta_1 \tag{8.8}$$

que, no domínio $\Re$ é

$$e^{\beta_1} = \frac{\dfrac{g(\mu(x_i+1))}{1-g(\mu(x_i+1))}}{\dfrac{g(\mu(x_i))}{1-g(\mu(x_i))}} \tag{8.9}$$

À razão à direita de (8.9) é conhecida como **razão de chances** (*odds ratio*) e compara as chances de sucesso na observação $x = x_i + 1$ com as chances de sucesso na observação $x = x_i$.

Exemplo

```
# Jogos do Ravens
# https://dl.dropboxusercontent.com/u/7710864/data/ravensData.rda

load("ravensData.rda")
head(ravensData)

  ravenWinNum ravenWin ravenScore opponentScore
1           1        W         24             9
2           1        W         38            35
```

```
3           1        W         28             13
4           1        W         34             31
5           1        W         44             13
6           0        L         23             24
```

```
logRegRavens <- glm(ravensData$ravenWinNum ~ ravensData$ravenScore,
                    family = "binomial")
summary(logRegRavens)
```

```
Call:
glm(formula = ravensData$ravenWinNum ~ ravensData$ravenScore,
    family = "binomial")

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.758  -1.100   0.530   0.806   1.495

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           -1.6800     1.5541   -1.08     0.28
ravensData$ravenScore   0.1066     0.0667    1.60     0.11

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 24.435  on 19  degrees of freedom
Residual deviance: 20.895  on 18  degrees of freedom
AIC: 24.89

Number of Fisher Scoring iterations: 5
```

Interpretando β_1:

- Para números *x* próximos de zero: $e^x \approx 1 + x$
- exp(0.1066) = 1.112489
- As chances estimadas do Ravens vencer crescem 11% para cada ponto marcado.
- β_1 é uma *odds ratio*, uma razão, uma taxa de variação.

Interpretando β_0:

- exp(-1.6800) = 0.186374
- Seria bom poder interpretar β_0 como as chances do Ravens vencer quando marcar zero pontos. A questão é: quantos jogos há na base de dados nos quais o Ravens marcou zero ou perto de zero? O intercepto é uma extrapolação, não faz sentido.

8.1 AVALIAÇÃO DO MODELO DE REGRESSÃO LOGÍSTICA

De forma geral, a regressão logística prevê a probabilidade de um evento ocorrer, baseando-se em observações prévias e na frequência com que o evento já ocorreu. Na regressão múltipla, para avaliarmos o quanto um modelo se ajusta aos dados, comparamos os valores observados com os valores ajustados. Da mesma forma, na regressão logística, podemos recorrer aos valores observados e aos valores ajustados para avaliar o modelo.

8.1.1 Estatística *log-likelihood*

Em estatística, uma **função de verossimilhança** é uma função dos parâmetros de um modelo estatístico. Funções de verossimilhança desempenham um papel-chave na inferência estatística, particularmente, em métodos de estimativa de um parâmetro de um conjunto de estatísticas.

A verossimilhança de um conjunto de valores de parâmetros θ, dado um conjunto de observações *x*, é igual à probabilidade dessas observações dado esse conjunto de valores de parâmetros θ.

$$\mathcal{L}(\theta \mid x) = p(x \mid \theta) \tag{8.10}$$

No caso discreto, seja *X* uma variável aleatória com distribuição de probabilidade discreta *p*, dependente de um parâmetro θ. Então, a função

$$\mathcal{L}(\theta \mid x) = p_\theta(x) = P_\theta(X = x) \tag{8.11}$$

é a função de verossimilhança de θ, dadas as saídas *x* de *X*. A equação (8.11) pode ser escrita como $P(X = x \mid \theta)$ e, frequentemente, é escrita como $P(X = x; \theta)$ para enfatizar que não é uma probabilidade condicional visto que θ é um parâmetro e não uma variável aleatória.

No caso contínuo, seja X uma variável aleatória com distribuição de probabilidade contínua e função de densidade de probabilidade f dependente de um parâmetro θ. Então, a função

$$\mathcal{L}(\theta \mid x) = f_\theta(x) \tag{8.12}$$

é a função de verossimilhança de θ, dadas as saídas x de X. A equação (8.12) pode ser escrita como $f(x \mid \theta)$, mas não deve ser considerada uma função de densidade de probabilidade condicional.

Para muitas aplicações, o logaritmo natural da função de verossimilhança, chamado de ***log-likelihood***, é mais conveniente para se trabalhar, porque o logaritmo é uma função monotonamente crescente, atinge o seu valor máximo nos mesmos pontos que a função original e, portanto, pode ser usado no lugar da estimativa de máxima probabilidade. Encontrar o máximo de uma função muitas vezes envolve tomar a derivada dessa função e resolvê-la para o parâmetro a ser estimado. Para a regressão logística, a medida *log-likelihood* é dada por

$$log\text{-}likelihood = \sum_{i=1}^{N} \left[y_i \ln(P(y_i)) + (1 - y_i) \ln(1 - P(y_i)) \right] \tag{8.13}$$

O *log-likelihood* é baseado na soma de probabilidades associadas aos valores preditos e observados. A estatística *log-likelihood* é análoga à soma dos quadrados dos resíduos da regressão múltipla, no sentido de que é um indicador do quantitativo da informação que permanece inexplicado após o ajuste do modelo. Consequentemente, grandes valores do *log-likelihood* indicam modelos fracamente ajustados: quanto maior o *log-likelihood*, menos explicadas as observações.

8.1.2 Estatística *deviance*

A estatística *deviance* (*deviance statistic*), ou estatística do desvio, é uma medida do quão bem o modelo linear generalizado se ajusta aos dados: quanto maior o desvio, pior o ajuste.

A estatística *deviance* é estreitamente relacionada com a estatística *log-likelihood*, e é dada por

$$deviance = -2 * log\text{-}likelihood \tag{8.14}$$

É também referida como **-2LL** por causa da forma como é calculada e tem distribuição chi-quadrado (χ^2), o que simplifica o cálculo de significância dos valores.

Utiliza-se a estatística do desvio para comparar modelos. Assim como na regressão múltipla, na regressão logística é estabelecido um modelo de base que fornece a melhor predição, com base na saída que acontece com maior frequência, ou seja, quando o modelo inclui apenas uma constante (intercepto). Se a partir daí adicionamos um ou mais preditores ao modelo, podemos mensurar as melhorias como

$$\chi^2 = \left(-2LL_{baseline} - (-2LL_{new})\right) = 2LL_{new} - 2LL_{baseline} \tag{8.15}$$

com *df* graus de liberdade, onde

$$df = k_{new} - k_{baseline} \tag{8.16}$$

e *k* é a quantidade de parâmetros. Como o modelo de base inclui apenas uma constante, $k_{baseline}$ é sempre 1. A equação (8.15) é conhecida como taxa de verossimilhança e tem distribuição chi-quadrado.

A função **glm**, do R, reporta duas formas de desvio - o **desvio nulo** (*Null deviance*) e o **desvio residual** (*Residual deviance*).

O desvio nulo mostra o quão bem a variável resposta é predita por um modelo que inclui apenas o intercepto (grande média). É determinado pela aplicação da estatística *deviance* sobre o modelo saturado (assume que cada valor observado tem seu prório parâmetro) e o model nulo (inclui apenas o intercepto), onde

$$Null\ deviance = 2LL_{Modelo\_Saturado} - 2LL_{Modelo\_Nulo} \tag{8.17}$$

com

$$df = df_{Modelo\_Saturado} - df_{Modelo\_Nulo} \tag{8.18}$$

graus de liberdade.

O desvio residual, como o próprio nome indica, é o desvio decorrente dos resíduos do modelo proposto (assume que os dados podem ser explicados por uma quantidade de parâmetros menor que a quantidade de observações), ou seja, da diferença entre o valor predito e o valor observado, onde

$$Residual\ deviance = 2LL_{Modelo\_Saturado} - 2LL_{Modelo\_Proposto} \tag{8.19}$$

com

$$df = df_{Modelo\_Saturado} - df_{Modelo\_Proposto} \tag{8.20}$$

graus de liberdade.

De forma geral, podemos dizer que o modelo proposto é considerado bom quando:

1. o desvio residual for tal que não represente diferença estatisticamente significante entre o modelo saturado (observações) e o modelo proposto;

2. o desvio nulo e o desvio residual forem estatisticamente diferentes, tal que

$$Null\ deviance - Residual\ deviance = 2LL_{Modelo\_Proposto} - 2LL_{Modelo\_Nulo} \tag{8.21}$$

se aproxima de uma distribuição chi-quadrado com

$$df = df_{Modelo\_Proposto} - df_{Modelo\_Nulo} \tag{8.22}$$

graus de liberdade.

8.1.3 Critério de informação

O coeficiente de determinação R^2 que utilizamos na regressão linear também poderia ser adequado à análise de modelos na regressão logística. O problema com R^2 é que quanto mais variáveis adicionamos ao modelo, maior é o R^2.

O indicador denominado **critério de informação de Akaike** (*Akaike information criterion* – AIC) é uma medida que penaliza o modelo por possuir mais variáveis, assim como o faz o R^2 ajustado. O AIC é definido como

$$AIC = n\ln\left(\frac{SEQ}{n}\right) + 2k \tag{8.17}$$

onde n é a quantidade de observações e k a quantidade de preditores. A cada nova variável adicionada ao modelo, AIC será acrescido de 2.

Quanto maior o valor de AIC, pior o ajuste do modelo. Não é possível dizer que AIC = 10 é pequeno ou que AIC = 500 é grande. Apenas podemos comparar AICs entre modelos definidos sobre os mesmos dados.

Outro indicador, denominado **critério de informação de baiesiano** (*Bayesian information criterion* – BIC), também penaliza o modelo pela quantidade de preditores e é baseado, em parte, na função de verossimilhança, bem como e é estreitamente relacionado com AIC. A penalidade em BIC é maior que em AIC.

$$BIC = -2\ln\hat{L} + k\ln(n) \tag{8.18}$$

onde $\hat{L}$ é o valor maximizado da função de verossimilhança $\mathcal{L}$ do modelo.

Quanto menor o valor de BIC, melhor o modelo. O BIC só é válido para n muito maior que k.

8.2 LABORATÓRIO 8

9 ANÁLISE DE VARIÂNCIA

Este capítulo introduz o conceito de análise de variância e apresenta uma discussão sobre os temas gerais das técnicas de análise.

9.1 ESTATÍSTICA F – COMPARAÇÃO DE DUAS VARIÂNCIAS

A comparação das variâncias de amostras de duas populações independentes e distribuídas normalmente pode ser efetuada por meio de teste estatístico.

Quando as populações são normalmente distribuídas e possuem variâncias iguais ($\sigma_1^2 = \sigma_2^2$), a distribuição da razão das variâncias amostrais, conhecida como **estatística *F***, é dada pela **distribuição *F***, mostrada na Figura 9.1.

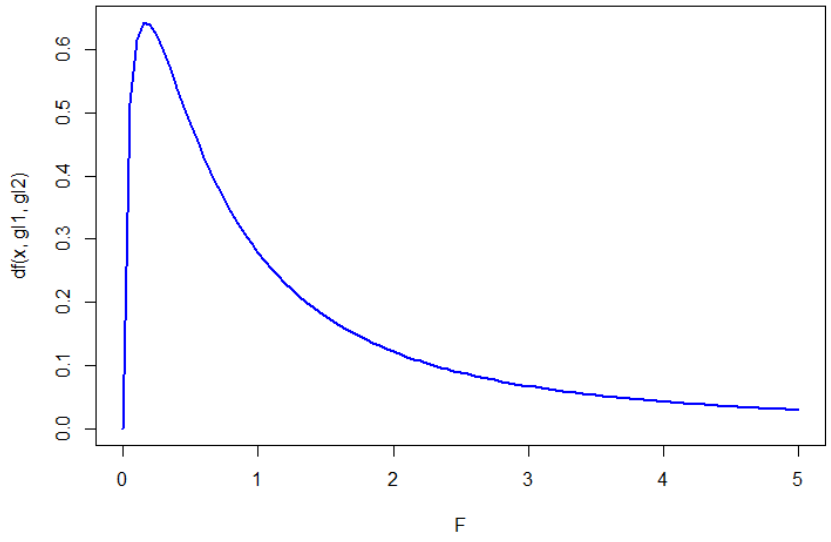


Figura 9.1 – Distribuição *F* para $gl_1 = 3$ e $gl_2 = 2$. ®

Isso quer dizer que, se repetirmos continuadamente o experimento que consiste em selecionar aleatoriamente amostras de duas populações normalmente distribuídas com variâncias iguais, a distribuição da estatística *F*, onde

$$F = \frac{s_1^2}{s_2^2} \tag{9.1}$$

é a distribuição *F*.

O Quadro 9.1 contém uma simulação para o cômputo de *F* com base em amostras de duas distribuições normais com variância $\sigma^2 = 1$, uma com tamanho $n_1 = 40$ ($gl_1 = 39$) e outra com tamanho $n_2 = 30$ ($gl_2 = 29$). A Figura 9.2 ilustra o resultado.

Quadro 9.1 – Simulação da distribuição *F* com $gl_1 = 39$ e $gl_2 = 29$.

```
# Estatísitica F
Fs <- c()
for(i in 1:100000){
  s1 <- sd(rnorm(40, mean = 0, sd = 1))^2
  s2 <- sd(rnorm(30, mean = 0, sd = 1))^2
  Fs <- c(Fs,s1/s2)
}
# Distribuição F
hist(Fs, probability =TRUE, breaks=200, xlim=c(0,3), xlab="F",
     main="Distribuição F")
curve(df(x,39,29),0,3,add=T,col="red",lwd=2)
```

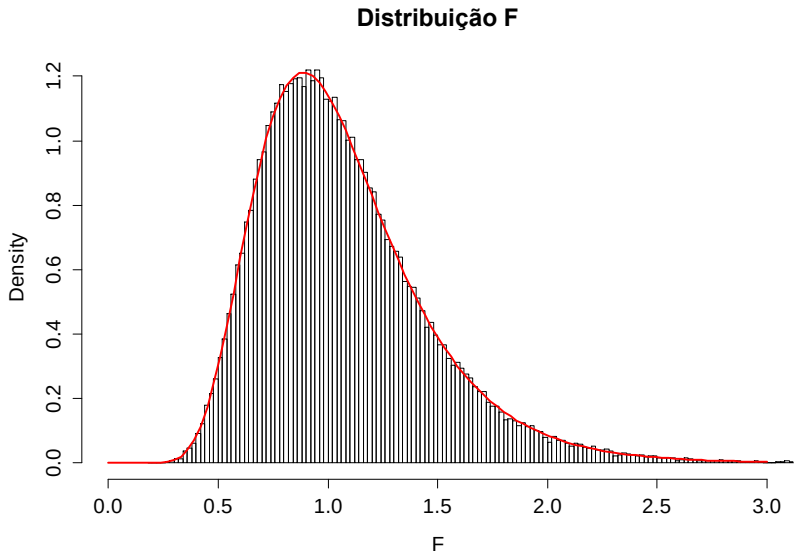


Figura 9.2 – distribuição F com $gl_1 = 39$ e $gl_2 = 29$.

A distribuição *F* possui as seguintes propriedades:

- não é simétrica; é assimétrica à direita;
- os valores de *F* são sempre positivos;
- há uma curva da distribuição *F* diferente para cada par de graus de liberdade de cada amostra (numerador e denominador);
- a área sob a curva *F* é igual a 1.

Se as populações têm realmente variâncias iguais, *F* tende para o valor 1. Logo, um valor de *F* próximo de 1 favorece a conclusão de que $\sigma_1^2 = \sigma_2^2$. Por outro lado, um valor grande de *F* constitui evidência de que as variâncias populacionais são diferentes.

Como a área sob a curva da distribuição *F* é igual a 1, o teste de hipótese pode ser feito como na curva normal, ou seja, estabelecendo-se a probabilidade da região crítica (ou nível de significância α) em que se rejeita H_0.

9.2 ANOVA

Em geral, o propósito da **análise de variância** (*analysis of variance* – ANOVA) é testar se existem diferenças significativas entre médias de grupos independentes. Quando comparamos duas médias, a ANOVA produzirá os mesmos resultados que o teste *t* para amostras independentes (se estivermos comparando dois grupos diferentes de observações) ou o teste *t* para amostras dependentes (se estamos comparando duas variáveis em um mesmo conjunto de observações).

Por que o nome análise de variância? Pode parecer estranho que um procedimento que compare médias tenha esse nome. No entanto, o nome é derivado do fato de que, quando se deseja testar diferenças estatísticas entre médias, na verdade o que efetivamente se compara são variâncias.

O coração da ANOVA reside no fato de que as variações podem ser divididas, isto é, particionadas. Como se sabe da Equação (2.8), a variação é uma função da soma dos quadrados (desvios), ou *SS* (*sum of squares*) para abreviar. Assim, vamos ilustrar o particionamento da variância identificando os conceitos fundamentais a partir de um exemplo prático.

Considere o conjunto de observações da Tabela 9.1. As médias para os dois grupos são bastante diferentes (2 e 6, respectivamente). A soma dos quadrados para cada grupo é igual a 2.

Tabela 9.1 – Soma dos quadrados dentro dos grupos e geral. ®

	Grupo 1	**Grupo 2**
Observação 1	2	6
Observação 2	3	7
Observação 3	1	5
Média	2	6
SS grupo	2	2
Média global	4	
SS global	28	

Considerando ambos os grupos de forma conjunta, obtemos a média global ($\bar{x}_{global}$) igual a 4 e *SS* total (SS_T) de 28, com base no conjunto total de observações ($n_T = n_1 + n_2 = 6$), calculado por

$$SS_T = \sum_{i=1}^{n_T} \left(x_i - \bar{x}_{global}\right)^2 \tag{9.1}$$

Nesse caso, o número de graus de liberdade gl_T é dado por $n_T - 1$.

O problema, então, consiste em descobrirmos quanto da variação total pode ser explicada pelo modelo estatístico que, no cenário da ANOVA, é baseado na média de cada grupo. Em outras palavras, a média de um grupo é o modelo estatístico desse grupo. Então, a variação referente a um dado ou observação consiste na diferença entre a distância dessa observação à média do grupo e a distância dessa observação à média geral. Logo, a variação referente ao modelo é

$$SS_M = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left[\left(x_j - \bar{x}_i\right) - \left(x_j - \bar{x}_{global}\right)\right]^2 \tag{9.2}$$

onde k é a quantidade de grupos e n_i o tamanho do i-ésimo grupo. A Figura 9.3

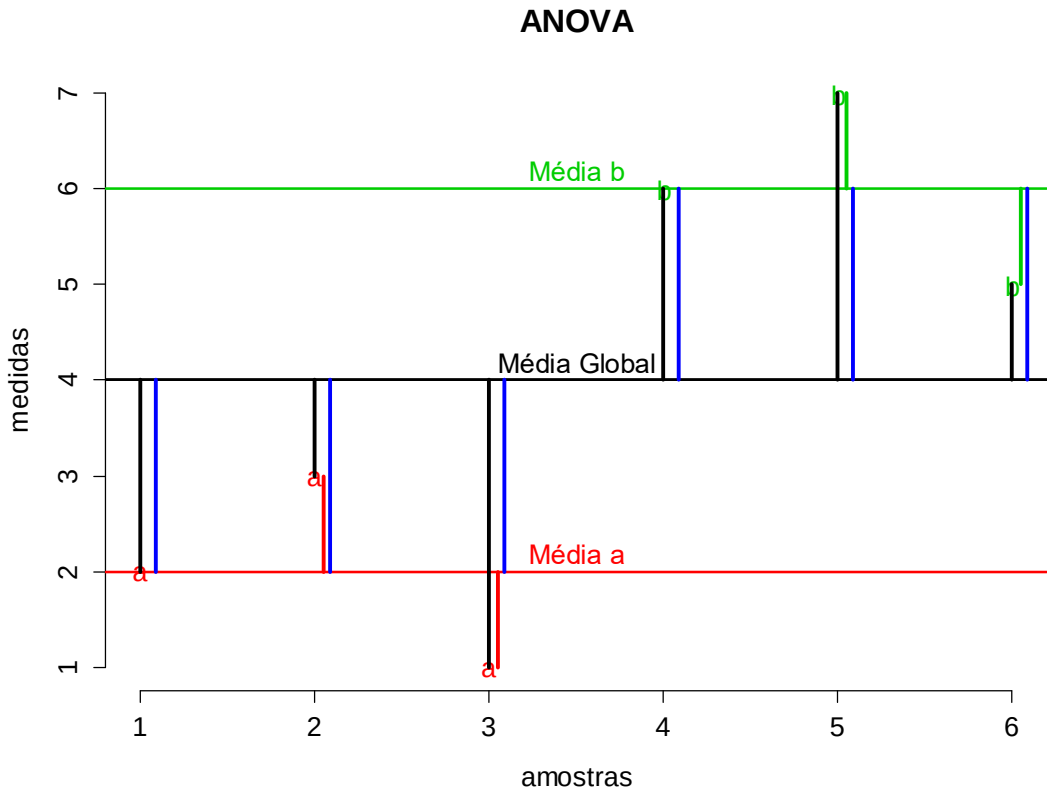


Figura 9.3 – Variação explicada pelo modelo (segmentos azuis): diferença entre a variação total (segmentos em preto) e variação referente ao modelo (segmentos em vermelho e verde). ®

Da Figura 9.3, observando os segmentos azuis, podemos reescrever a variação explicada pelo modelo de forma simplificada, como

$$SS_M = \sum_{i=1}^{k} n_i \left(\bar{x}_i - \bar{x}_{global} \right)^2 \tag{9.3}$$

onde k é a quantidade de grupos e n_i o tamanho do i-ésimo grupo. O número de graus de liberdade gl_M é sempre o número de parâmetros estimados menos um. Em suma, este valor será o número de grupos menos um ($k-1$).

Por fim, a variação não explicada pelo modelo, decorrente de fatores externos não previstos ou mensurados, é variação residual, determinada pela soma dos quadrados dos resíduos (SS_R). Conhecendo-se SS_T e SS_M, SS_R poderia ser facilmente calculado pela diferença entre ambos ($SS_R = SS_T - SS_M$). No entanto, esse raciocínio proporciona pouco

conhecimento sobre o significado dessa grandeza e, caso alguma das outras quantidades tenha sido equivocadamente calculada, também será equivocada a estimativa de SS_R.

Por definição, resíduo é a diferença entre o que prevê o modelo e o que foi realmente observado. Então, determinando o resíduo para cada grupo e somando todos os grupos, temos

$$SS_R = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left[\left(x_{ji} - \bar{x}_i \right) \right]^2 \tag{9.4}$$

onde k é a quantidade de grupos e n_i o tamanho do i-ésimo grupo.

Reescrevendo a Equação (9.4) em termos da variância, temos

$$SS_R = \sum_{i=1}^{k} s_k^2 \left(n_k - 1 \right) \tag{9.5}$$

onde k é a quantidade de grupos e n_k o tamanho do k-ésimo grupo. O número de graus de liberdade gl_R para SS_R é dado pela diferença $gl_R = gl_T - gl_M$.

Como vimos, a variação SS_M mede a variação explicada pelo modelo estatístico (por exemplo, decorrente de manipulação experimental) e SS_R mede a variação devida a fatores externos. No entanto, como ambos os valores são somas quadráticas, ambos são influenciados pela quantidade de itens somados. Para eliminar esse efeito, ANOVA compara a variação quadrática média (*mean square* – MS) dessas grandezas, o que se consegue pela simples divisão pelos respectivos graus de liberdade. Logo,

$$MS_M = \frac{SS_M}{gl_M} = \frac{SS_M}{k-1} \tag{9.6}$$

e

$$MS_R = \frac{SS_R}{gl_R} = \frac{SS_R}{n_T - k} \tag{9.7}$$

onde n_T é o número total de observações e k a quantidade de grupos.

Podemos comparar essas duas estimativas de variância por meio do teste *F* (ver distribuição *F*) que testa se a razão entre as duas é significativamente diferente de 1.

$$F = \frac{MS_M}{MS_R} \tag{9.8}$$

Sob a hipótese nula de que não existem diferenças entre as médias dos grupos da população, a variância explicada pelo modelo deve ser aproximadamente a mesma decorrente dos desvios.

O teste ANOVA para os dados da Tabela 9.1 resulta nos dados da Tabela 9.2.

Tabela 9.2 – ANOVA amostras independentes. ®

	SS	**gl**	***MS***	***F***	***p***
Modelo	24	1	24	24	0,008
Resíduo	4	4	1		

9.2.1 Múltiplos grupos e propagação do erro

Em substituição à Tabela 9.2, poderíamos ter simplesmente calculado a diferença entre médias pelo teste *t* de amostras independentes e chegariamos à mesma conclusão, e, de fato, teríamos um resultado idêntico. No entanto, a técnica de análise de variância é muito mais flexível e poderosa, e pode ser aplicada a problemas de investigação mais complexos.

A ANOVA é um teste paramétrico (baseada em estimativas de parâmetros) utilizado quando o pesquisador deseja verificar se existem diferenças entre as médias de uma determinada variável (variável resposta) em relação a um tratamento com dois ou mais níveis categóricos (variável preditora). O teste *t* também é utilizado nesse tipo de procedimento (com no máximo dois níveis), porém a ANOVA é indicada para análises com *n* amostral superior a 30.

Por que não utilizar o teste *t* quando temos mais de dois níveis categóricos? O teste *t* poderia ser utilizado combinando-se matematicamente ($C_{n,p} = \frac{n!}{p!(n-p)!}$) os fatores dois-a-dois. Como primeiro problema, na medida em que a quantidade de fatores aumenta, cresce, também, a quantidade de testes. O segundo problema, e mais importante, trata-se da propagação do erro amostral. Por exemplo, se tivéssemos três fatores A, B e C, teríamos que fazer três testes *t*: AB, BC e AC. Cada teste *t* seria efetuado com base em um nível de confiança determinado, em geral, 95%. Ou seja, para concluirmos pela não rejeição de H_0, esse fenômeno teria que ocorrer simultaneamente nos três testes. A probabilidade de fenômenos que ocorrem simultaneamente é igual ao produto da probabilidade de cada um, portanto: 0,95 x 0,95 x 0,95 = 0,857. Isso significa que nosso nível de confiança real é de apenas 85,7%. Em outra ótica, significa que o erro do Tipo I (0,05) foi praticamente triplicado (0,143).

ANOVA cria um modo para testar várias hipóteses nulas ao mesmo tempo. Mais do que testar os fatores individualmente, ANOVA é um teste global que considera todos os fatores ao mesmo tempo e as seguintes hipóteses amplas:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k$$
$$H_1: pelo\ menos\ um\ \beta_j \neq \beta_k\ para\ j \neq k$$

A lógica por trás desse processo tem a ver com a quantidade de variância que existe na população. É provável que o pesquisador não conheça a variação real da população, mas esta pode ser estimada, por amostragem, com base na variação da amostra. Então, basta comparar as diferenças entre as amostras para verificar se são estatisticamente significantes, ou não, mantendo-se o erro de amostragem original.

Antes de realizar a ANOVA é importante verificar os pressupostos da análise. As premissas segundo as quais a estatística *F* é confiável são as mesmas adoatadas para testes paramétricos com base em distribuição normal, a saber: (1) as variações em cada condição experimental devem ser iguais (homocedasticidade) ou, pelo menos, aproximadamente iguais; (2) as observações devem ser independentes; (3) e a variável dependente deve ser

medida em escala intervalar (ver Seção 2.3, Quadro 2.2). Em termos de normalidade, o que importa é que as distribuições dentro dos grupos sejam normais.

9.2.2 Testes *Post Hoc*

Conforme visto, o teste ANOVA não informa quais grupos diferem ou onde está a diferença, somente que existe uma diferença.

Testes *post hoc* consistem de comparações pareadas para identificar todas as combinações dos grupos testados em ANOVA. É como se cada par de grupos fosse submetido a um teste *t*. Contudo, é sabido que a sucessão de testes *t* ocasiona a inflação do erro do Tipo I, o que compromete a análise estatística. Portanto, testes *post hoc* foram pensados de forma a garantir que as comparações pareadas permanecessem com a probabilidade de erro especificada pelo nível de significância α. Existem diversos testes *post hoc* desenvolvidos para ANOVA, comentados a seguir. Exemplos de aplicação dos métodos com recursos de R serão apresentados nas seções ANOVA One Way e ANOVA Two Way.

9.2.2.1 Correção de Bonferroni

A Correção de Bonferroni é a forma mais popular (e mais fácil) de teste *post hoc* e consiste em dividir α pelo número de comparações *k*, garantindo assim que o erro cumulativo Tipo I seja inferior a α.

$$p_{crit} = \frac{\alpha}{k} \tag{9.9}$$

Por exemplo, se efeturamos 10 testes para α = 0,05, devemos usar 0,005 como o nosso critério de significância. Na prática, isso significa que a probabilidade de rejeitar um efeito que existe realmente (diferença verdadeira nos dados) é aumentada, ou seja, aumenta-se a probabilidade de um erro Tipo II.

9.2.2.2 Método de Bonferroni-Holm

Diversos aperfeiçoamentos foram feitos à Correção de Bonferroni. Uma delas é o Método de Bonferroni-Holm (Holm, 1979), que funciona conforme o algoritmo a seguir:

1. Calcula-se, inicialmente, o valor de *p* para cada par de grupos nos dados.
2. Em seguida, os *m* valores *p* calculados são ordenados do menor para o maior.
3. Verifica-se, então, se o primeiro valor *p* é maior que ou igual a α/m: em caso afirmativo, o procedimento é parado e não há valores de *p* significativos; caso contrário, segue-se ao próximo passo.
4. O segundo valor *p* é comparado a $\alpha/(m-1)$: caso *p* seja maior que ou igual a $\alpha/(m-1)$, o processo se encerra e não há outros valores *p* significativos; caso contrário, segue-se ao próximo passo.
5. O procedimento se repete até o último *p* calculado.

A ideia-chave por trás do método é a verificação passo-a-passo. Isso significa que, enquanto uma comparação for significativa, passamos para o próximo, mas no ponto em que nos deparamos com uma comparação não significativa, paramos e assumimos que todos as demais comparações também não são significativas.

9.2.2.3 Método de Tukey

O Teste proposto por Tukey (1953) é também conhecido como teste de Tukey da diferença honestamente significativa (*honestly significant difference*) e teste de Tukey da diferença totalmente significativa (*wholly significant difference*). É um teste exato em que, para a família de todas as comparações duas a duas, a taxa de erro da família de testes é exatamente α (e o intervalo de confiança é exatamente 1- α). Métodos de comparações múltiplas exatos são raros. O teste de Tukey tem sido analiticamente mostrado como ótimo, no sentido de que, dentre todos os procedimentos que resultam em intervalos de confiança com mesmo tamanho para todas diferenças duas a duas, com coeficiente de confiança da família de pelo menos 1- α, o teste de Tukey é o que resulta em intervalos menores. Isso quer dizer que, se a família consiste em todas comparações duas a duas e o

teste de Tukey pode ser usado, ele resultará em intervalos menores do que qualquer outro método de comparação múltipla de uma etapa.

A estratégia de Tukey consiste em definir a menor diferença significativa. Tal procedimento utiliza a amplitude da distribuição studentizada.

9.2.2.4 Que método usar?

Depois de uma ANOVA, é necessária uma análise mais específica para se descobrir quais grupos diferem:

- Quando não há hipóteses específicas antes do experimento, testes *post hoc* devem ser utilizados.
- Quando as amostras têm tamanhos iguais e as variações de grupo são semelhantes, Tukey é a melhor opção.
- Quando se deseja garantir o controle da taxa de erro Tipo I, Bonferroni ou Bonferroni-Homes são mais apropriados.

A aplicação dos métodos *post hoc* é ilustrada no Quadro 9.3 da Seção 9.4.

9.3 ANOVA E REGRESSÃO MÚLTIPLA SÃO A MESMA COISA

ANOVA e Regressão Linear não são apenas relacionadas, são a mesma coisa. ANOVA é apenas um caso especial de regressão. Compreender esse conceito é fundamental para o entendimento do modelo linear como um todo e suas aplicações são de longo alcance.

Considere um modelo composto por uma única variável independente categórica com três categorias, representando os três partidos supostamente mais influentes na Câmara dos Deputados: PMDB, PSDB e PT. A variável dependente consiste no tempo em minutos de exposição na mídia, nos últimos 100 dias, dos respectivos líderes, concedendo entrevistas.

O Bloco 1 do Quadro 9.2 ilustra o cáclulo de F (0,055) conforme as definições da Seção 9.2, donde se conclui que não há diferença entre as médias de tempo de exposição dos líderes (p = 0,946).

Quadro 9.2 – ANOVA *versus* Regressão Múltipla.

```
##########
# BLOCO 1
##########
set.seed(1456728)
g1 <- rnorm(100,5,2)  # PMDB
g2 <- rnorm(100,5,2)  # PSDB
g3 <- rnorm(100,5,2)  # PT

m1 <- mean(g1); sd1 <- sd(g1)
m2 <- mean(g2); sd2 <- sd(g2)
m3 <- mean(g3); sd3 <- sd(g3)

k <- 3 # quantidade de fatores ou grupos

n1 <- length(g1)
n2 <- length(g2)
n3 <- length(g3)
nT <- n1 + n2 + n3 # quantidade de observações

g <- c(g1,g2,g3)  # grupão
mg <- mean(g)  # média grupão

SST <- sum((g - mg)^2)

SSM <- sum(((g1 - m1) - (g1 - mg))^2) +
       sum(((g2 - m2) - (g2 - mg))^2) +
       sum(((g3 - m3) - (g3 - mg))^2)
# ou
SSM <- n1*(m1-mg)^2 + n2*(m2-mg)^2 + n3*(m3-mg)^2

SSR <- sum((g1-m1)^2) + sum((g2-m2)^2) + sum((g3-m3)^2)
#ou
SSR = SST - SSM

glM <- k - 1 # graus de liberdade do modelo
glR <- nT - k # graus de liberdade do resíduo

MSM <- SSM / glM # média quadrática do modelo
MSR <- SSR / glR # média quadrática do resíduo

Fs <- MSM / MSR # estatística F
p.value <- 1 - pf(Fs, glM, glR)
sprintf("F = %7.5f    p.vaule = %7.5f", Fs, p.value)

[1] "F = 0.05530    p.vaule = 0.94621"

##########
# BLOCO 2
##########
# associar categorias a cada
# conjunto dos tempos coletados
# 1- PMDB,  2 – PSDB e 3 - PT
partido <- replicate(100,1:3)
partido <- as.factor(c(partido[1,],partido[2,],partido[3,]))
dados <- data.frame(tp=g,partido)

# análise exploratória
boxplot(tp ~ partido, data=dados)
```

```
# usando ANOVA - aov()
res <- aov(tp ~ partido, data=dados)
summary(res)

             Df Sum Sq Mean Sq F value Pr(>F)
partido       2    0.4   0.214   0.055  0.946
Residuals   297 1148.2   3.866

##########
# BLOCO 3
##########
# usando regressão múltipla – lm()
s <- summary(  lm(tp ~ partido, data=dados)  )
s

Call:
lm(formula = tp ~ partido, data = dados)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7909 -1.4662 -0.2446  1.3371  6.7985

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.036440   0.196622  25.615   <2e-16 ***
partido2    -0.003545   0.278066  -0.013    0.990
partido3     0.078254   0.278066   0.281    0.779
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.966 on 297 degrees of freedom
Multiple R-squared:  0.0003722,       Adjusted R-squared:  -0.006359
F-statistic: 0.0553 on 2 and 297 DF,  p-value: 0.9462

# verificando o valor de p
1 - pf(s$fstatistic[1], s$fstatistic[2], s$fstatistic[3])

    value
0.9462114
```

No Bloco 2 do Quadro 9.2, uma variável categórica (*partido*) é associada ao conjunto global dos tempos aferidos (*tp*), de forma a identificar cada líder por meio do respectivo partido. Em seguida, executa-se a ANOVA por meio da função **aov**, tomando-se *tp* como função de *partido* (*tp ~ partido*), cujos resultados confirmam os previamente encotrados no Bloco 1 (F = 0,055 e p = 0,946).

No Bloco 3 do Quadro 9.2, executa-se o modelo linear **lm**, com base na mesma função *tp ~ partido*, a partir do qual se chega à mesma conclusão de que as médias são diferentes (F = 0,055 e p = 0,946). Contudo, a estimativa dos coeficientes nos permite análises adicionais. Inicialmente, podemos afirmar que a média de tempo de exposição do líder do PMDB é de 5.036440 minutos (intercepto), e que essa estimativa é estatisticamente diferente de zero. O segundo coeficiente nos provê a interpretação de que a diferença entre as médias de tempo

de exposição dos líderes do PSDB e do PMDB não é estatisiticamente significativa. De forma análoga, do terceiro coeficiente, conclui-se que não há diferença significativa entre a média dos tempos de exposição dos líderes do PT e do PMDB. Portanto, confirma-se o resultado da ANOVA que falha em rejeitar a hipótese nula de que as médias são iguais.

*Exercício: Altere a média do conjunto referente aos tempos do líder do PSDB para **7** e repita a análise utilizando ANOVA (aov) e Regressão Linear (lm). Verifique se é possível detectar a diferença de forma global e de forma individualizada.*

Como regra geral, quando nada se sabe sobre os dados em estudo, a abordagem de análise deve recair sobre o método da regressão linear. Quando o pesquisador tem conhecimento prévio sobre o conjunto de dados e/ou provoca efeito específico sobre determinado conjunto, ANOVA deve ser a escolha.

9.4 ANOVA ONE-WAY NO AMBIENTE R

Quando comparamos mais de dois grupos com base em apenas uma variável independente, utilizamos o teste ANOVA *one way*. Nesse caso, o seguinte procedimento deve ser observado:

1. Análise exploratória: como em qualquer análise, é uma boa idéia para começar por representar graficamente os dados e computar algumas estatísticas descritivas.
2. Verificação de homocedasticidade: usar o teste de Levene para verificar a homogeneidade da variância.
3. ANOVA básica: executar a análise principal da variância. Dependendo do que for encontrado na etapa anterior, uma versão robusta do teste deve ser executada.
4. Testes *post hoc*: executar teste *post hoc.*

Quadro 9.3 – ANOVA *one way* e testes *post hoc*.

```
# ANOVA One Way
##########
# BLOCO 1
##########
# carregar dados
set.seed(1456728)
tp1 <- rnorm(100,5.4,2) # PMDB
```

```
tp2 <- rnorm(100,9,2)    # PSDB
tp3 <- rnorm(100,5,2)    # PT
tp <- c(tp1, tp2, tp3)
partido <- replicate(100,c("PMDB","PSDB","PT"))
partido <- as.factor(c(partido[1,], partido[2,], partido[3,]))
dados <- data.frame(tp,partido)

# Análise exploratória
boxplot(tp ~ partido, data=dados)
##########
# BLOCO 2
##########
# teste de homocedasticidade
# leveneTest(outcome variable, group, center = median/mean)
# diferença -> p < 0.05
library(car)
leveneTest(dados$tp, dados$partido, center=mean)   # média

Levene's Test for Homogeneity of Variance (center = mean)
      Df F value Pr(>F)
group   2  1.5196 0.2205
      297

leveneTest(dados$tp, dados$partido, center=median) # mediana

Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group   2  1.4177 0.2439
      297

##########
# BLOCO 3
##########
# ANOVA lm() ou aov()
# newModel<-aov(outcome ~ predictor(s), data = dataFrame, na.action = an
action))
modelo <- aov(tp ~ partido, data=dados)
summary(modelo)

             Df Sum Sq Mean Sq F value Pr(>F)
partido       2  946.3   473.2   122.4 <2e-16 ***
Residuals   297 1148.2     3.9

##########
# BLOCO 4
##########
# post hoc
# pairwise.t.test(outcome, predictor, paired = FALSE, p.adjust.method =
"method=c("bonferroni", "BH")"
pairwise.t.test(dados$tp, dados$partido, paired = FALSE, p.adjust.method
= "bonferroni")

        Pairwise comparisons using t tests with pooled SD

data:  dados$tp and dados$partido

     PMDB   PSDB
PSDB <2e-16 -
PT   0.74   <2e-16

P value adjustment method: bonferroni

pairwise.t.test(dados$tp, dados$partido, paired = FALSE, p.adjust.method
= "BH")
```

```
        Pairwise comparisons using t tests with pooled SD

data:  dados$tp and dados$partido

     PMDB   PSDB
PSDB <2e-16 -
PT   0.25   <2e-16

P value adjustment method: BH
# method="Tukey"
#  newModel<-glht(aov.Model, linfct = mcp(predictor = "method"), base =
x)
library(multcomp)
newModel<-glht(modelo, linfct = mcp(partido = "Tukey"), base = dados)
summary(newModel)


         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = tp ~ partido, data = dados)

Linear Hypotheses:
                  Estimate Std. Error t value Pr(>|t|)
PSDB - PMDB == 0   3.5965     0.2781  12.934   <1e-04 ***
PT - PMDB == 0    -0.3217     0.2781  -1.157     0.48
PT - PSDB == 0    -3.9182     0.2781 -14.091   <1e-04 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

No Bloco 1 do Quadro 9.3, os dados são carregados e análise exploratória preliminar é efetuada. No Bloco 2, o teste de homocedasticidade falha em rejeitar a hipótese nula (variâncias iguais) tanto para a média (p = 0,22) quanto para a mediana (p = 0,24). No Bloco 3, a ANOVA aponta que existe diferença estatisticamente significativa entre as médias (p = 2e-16). Por fim, no Bloco 4, testes *post hoc* são executados e as diferenças par-a-par identificadas. Conclui-se que não há diferença significativa entre as médias do PT e PMDB, contudo, PSDB difere significativamente do PMDB e do PT.

9.5 NOVA TWO-WAY NO AMBIENTE R

Quando comparamos mais de dois grupos com base em duas variáveis independentes, utilizamos o teste ANOVA *two way*.

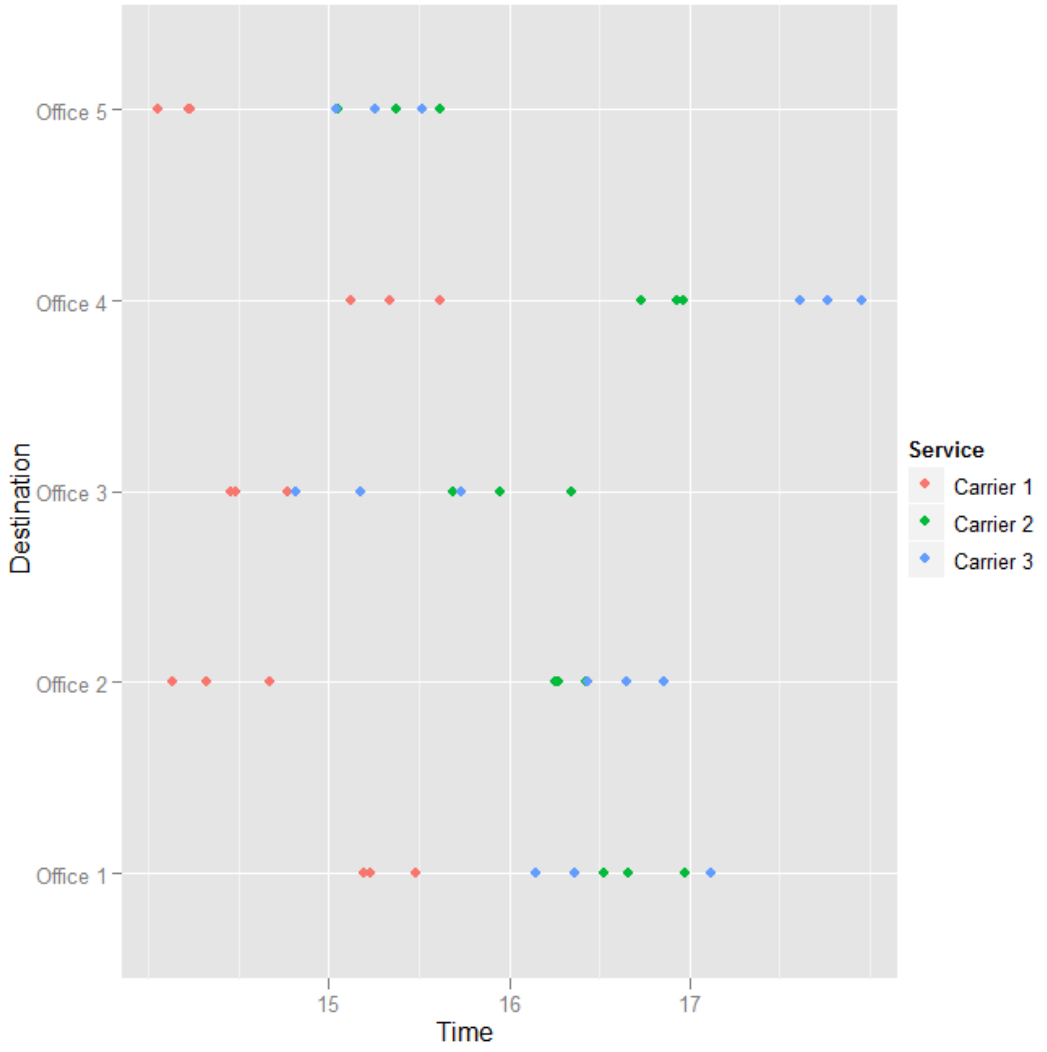
Como exemplo, considere uma empresa que regularmente tem que enviar pacotes entre os seus cinco diferentes subescritórios e tem a opção de usar três concorrentes serviços de

entrega de encomendas, os quais cobram quantidades aproximadamente semelhantes para cada entrega. Para determinar qual serviço usar, a empresa decide executar uma experiência de navegação de três pacotes de sua sede para cada um dos cinco subescritórios. O prazo de entrega para cada pacote é gravada e os dados processados no R.

Quadro 9.4 –

```
# carga dos dados
delivery.df = data.frame(
  Service = c(rep("Carrier 1", 15), rep("Carrier 2", 15),
    rep("Carrier 3", 15)),
  Destination = c(rep(c("Office 1", "Office 2", "Office 3",
    "Office 4", "Office 5"), 9)),
  Time = c(15.23, 14.32, 14.77, 15.12, 14.05,
  15.48, 14.13, 14.46, 15.62, 14.23, 15.19, 14.67, 14.48, 15.34, 14.22,
  16.66, 16.27, 16.35, 16.93, 15.05, 16.98, 16.43, 15.95, 16.73, 15.62,
  16.53, 16.26, 15.69, 16.97, 15.37, 17.12, 16.65, 15.73, 17.77, 15.52,
  16.15, 16.86, 15.18, 17.96, 15.26, 16.36, 16.44, 14.82, 17.62, 15.04)
)

# análise exploratória para investigação de tendências
ggplot(delivery.df, aes(Time, Destination, colour = Service)) +
geom_point()
```



O gráfico mostra um padrão para todos os subescritórios em que a operadora de serviços 1 apresenta prazos de entrega mais curtos do que os outros dois serviços. Há também uma indicação de que as diferenças entre os serviços varia de acordo com os cinco subescritórios e podemos esperar que o termo de interação entre a empresa de entrega e os subescritórios possa ser significativo no modelo de variância de dois fatores. Para ajustar o modelo de variância de dois fatores que usamos o código:

```
# two way ANOVA
delivery.mod1 = aov(Time ~ Destination*Service, data = delivery.df)
summary(delivery.mod1)

                    Df Sum Sq Mean Sq F value   Pr(>F)
Destination          4 17.542   4.385  61.155 5.41e-14 ***
Service              2 23.171  11.585 161.560  < 2e-16 ***
Destination:Service  8  4.189   0.524   7.302 2.36e-05 ***
Residuals           30  2.151   0.072
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Temos fortes evidências de que existem diferenças entre os três serviços de entrega, entre os cinco destinos dos subescritórios, e que existe uma interação entre destino e serviço em linha com o que vimos na trama original dos dados. Agora que temos montado o modelo e identificou os fatores importantes que precisamos para investigar os diagnósticos modelo para garantir que as várias hipóteses são, em geral válido.

Podemos traçar os resíduos do modelo contra valores ajustados para olhar para as tendências óbvias que não são coerentes com os pressupostos do modelo sobre a independência e variância comum. O primeiro passo é o de criar uma estrutura de dados com os valores ajustados e resíduos do modelo acima:

https://www.r-bloggers.com/two-way-analysis-of-variance-anova/

9.6 LABORATÓRIO 9

10 QUI-QUIDRADO

Qui-quadrado, simbolizado por X^2, é um teste não paramétrico[26], cujo princípio básico é comparar proporções, isto é, as possíveis divergências entre as frequências observadas e esperadas para um certo evento.

Podemos contar os incidentes de um evento e comparar os resultados obtidos, ou dados observados, com os resultados esperados para o evento. Suponha que 27 pessoas fossem perguntadas sobre a preferência partidária entre PT, PSDB ou PMDB. Caso não existisse preferência definida, esperaríamos que 9 respondessem PT, 9 PSDB, e 9 PMDB. Usamos, então, um teste qui-quadrado para comparar o que observamos (real) com o que esperamos. Se nossa amostra indicar que 2 gostam do PT, 20 do PSDB, e 5 do PMDB, podemos afirmar, com algum grau de confiança, que mais pessoas preferem o PSDB. Se nossa amostra indicar que 8 preferem o PT, 10 o PSDB, e 9 o PMDB, então não é adequado afirmar que o PSDB é o preferido.

Em outras palavras, o teste qui-quadrado nos diz com quanta certeza os valores observados podem ser explicados pela teoria em estudo.

A aplicação do teste qui-quadrado pressupõe o atendimento às seguintes condições:

- os grupos devem ser independentes;
- os itens de cada grupo devem ser selecionados aleatoriamente;
- as observações devem ser frequências ou contagens;
- cada observação deve pertencer a uma e somente uma categoria (ou classe); e
- a amostra deve ser relativamente grande (ao menos 5 observações em cada célula e, no caso de poucos grupos, pelo menos 10).

A forma de cálculo do teste qui-quadrado foi proposta por Karl Pearson, em 1900, e consiste na aferição do desvio entre proporções observadas (resultado de um experimento) e esperadas (distribuição esperada para o fenômeno), representada por

[26] Teste não paramétrico é aquele que não depende de parâmetros populacionais como, por exemplo, média e variância.

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \quad (10.1)$$

onde O_i é a frequência observada e E_i a frequência esperada para cada uma das k classes.

As frequências observadas são obtidas diretamente dos dados das amostras, enquanto que as frequências esperadas são estimadas a partir de conceitos teóricos estabelecidos ou de conhecimento adquirido previamente sobre os eventos.

10.1 DISTRIBUIÇÃO QUI-QUADRADO

Se repetirmos continuadamente o experimento que consiste em selecionar aleatoriamente amostras de uma população, efetuarmos a contagem das observações respectivas a cada uma das k classes contidas na amostra e calcularmos a estatística X^2, conforme a Equação (10.1), encontramos a distribuição qui-quadrado.

Quadro 10.1 – Simulação da distribuição qui-quadrado para $k = 2$ ($gl = 1$).

```
# Estatística X2
k <- 2          # classes
gl <- k - 1     # graus de liberdade
p <- rep(1/k,k) # vetor de probabilidades para classes equiprováveis
n <- 160        # quantidade de observações
E <- p * n      # frequência esperada para cada classe
rept <- 10000   # número de repetições
X2 <- vector("double", rept) # vetor de estatísticas qui-quadrdado
for(i in 1:rept){
    # dados observados
    s <- sample(1:k, n, replace=TRUE)

    # frequência das classes observadas
    O <- matrix(table(s))
    O <- O[1:length(O)]

    # qui-quadrado
    X2i <- sum(((O-E)^2)/E)
    X2[i] <- X2i
}
# Distribuição X2
X2 <- sort(X2)
hist(X2, probability =TRUE, breaks=30, ylim=c(0,1), xlim=c(0,10),
     xlab=expression(paste("X"^2)),
     main=expression(paste("Distribuição X"^2)))
lines(X2, dchisq(X2, df=gl), col="red", lwd=3)
```

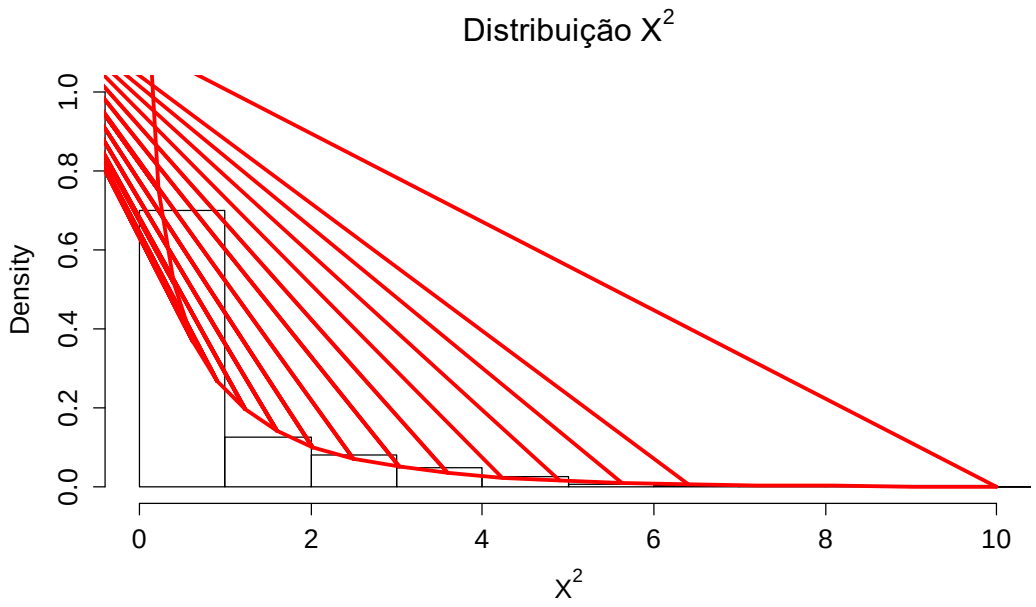


Figura 10.1 – Distribuição X^2 com $k = 2$ ($gl = 1$).

A curva de densidade de probabilidade para a distribuição X^2 depende do número de graus de liberdade, onde $gl = k - 1$. A Figura 10.2 ilustra a distribuição para distintos valores de k.

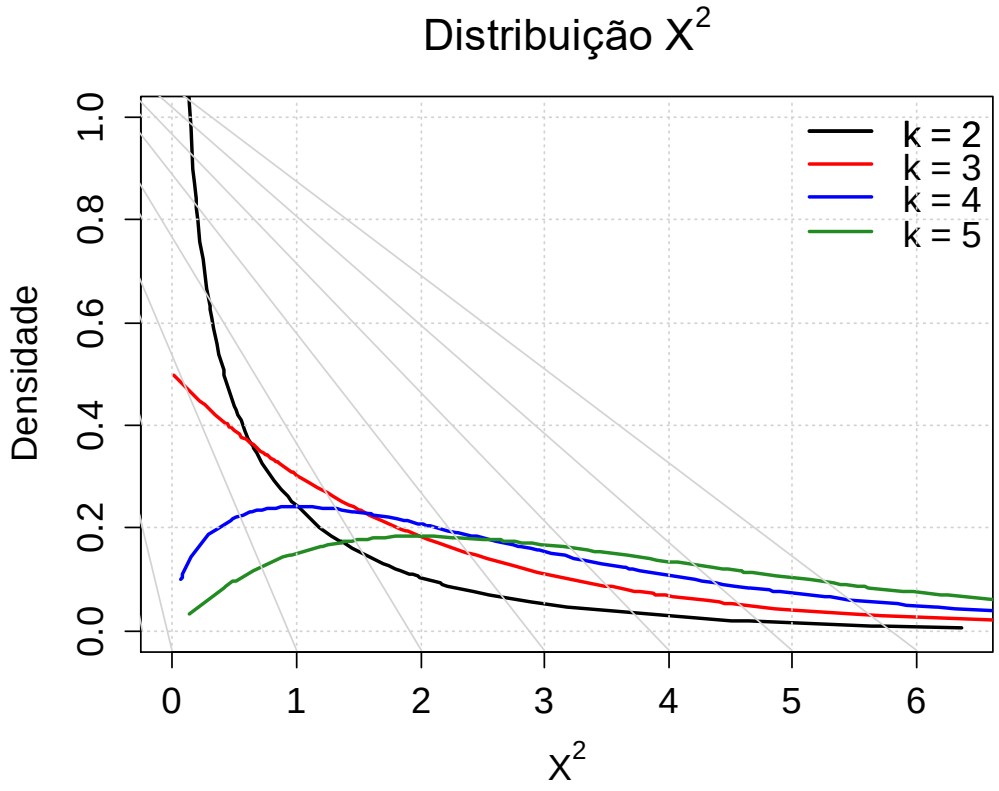


Figura 10.2 – Distribuição X^2 para distintos valores de k. ®

A partir das curvas de densidade de probabilidade da Figura 10.2, as hipóteses nula e alternativa podem ser formuladas:

> H_0: As frequências observadas não diferem das frequências esperadas, isto é, não existe diferença entre as frequências (contagens) dos grupos.
>
> H_1: As frequências observadas são diferentes das frequências esperadas, portanto existe diferença entre as frequências dos grupos.

Como a área sob a curva da distribuição X^2 é igual a 1, o teste de hipótese pode ser feito como na curva normal, ou seja, estabelecendo-se a probabilidade da região crítica (ou nível de significância α) em que se rejeita H_0. No ambiente R, utilizamos a função **pchisq** para determinar a probabilidade de um dado valor X^2 para *k* classes.

10.2 TABELA DE CONTINGÊNCIA

E quando não há hipótese baseada em alguma terioa existente, podemos aplicar o teste qui-quadrado? A resposta é sim.

O teste X^2 pode ser aplicado nos casos em que não se dispõe de uma teoria que permita efetuar o cálculo das classes esperadas. Nesses casos, a essência do teste consiste em comparar as frequências observadas com as frequências esperadas quando se observa a condição de independência de grupos. Se a diferença entre as frequências observadas e esperadas for suficientemente grande, então a hipótese nula de independência é rejeitada:

> H_0: Não há associação entre os grupos, ou seja, as variáveis categóricas são independentes.
>
> H_1: Há associação entre os grupos, ou seja, as variáveis categóricas são dependentes.

Esse tipo de situação ocorre quando desejamos verificar se uma determinada característica se distribui igualmente entre grupos, como sexo, classe social, grupo racial, faixa etária,

localização geográfica, etc. Portanto, não há como prever um valor esperado teórico, mas há como estimar um valor esperado a partir da relação cruzada dos grupos. Esse método é denominado **tabela de contingência**.

A estatística X^2 é a soma das diferenças entre valores observados e esperados em todas as células da tabela, escaladas pela magnitude dos valores esperados, da seguinte forma:

$$X^2 = \sum_{i,j} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \tag{10.2}$$

onde *i* representa as linhas e *j* as colunas da tabela, O_{ij} é o valor observado na célula (*i*, *j*) e E_{ij} é o valor esperado. Os valores de E_{ij} são computados a partir das probabilidades marginais, isto é, a partir dos totais de linhas e colunas convertidos em proporções.

Suponha que 150 pessoas tenham sido consultadas na pesquisa de preferência por partidos e que queiramos testar a dependência da variável partido com o grupo sexo. A tabela de contingência está representada na Tabela 10.1.

Tabela 10.1 – Tabela de Contingência: pesquisa de opinião sobre preferência partidária classificada por sexo.

Partido	**Sexo**		**Total**
	Masculino	**Feminino**	
PT	25	45	70
PSDB	15	25	40
PMDB	10	30	40
Total	50	100	**150**

FONTE: Elaborado pelos autores.

Por definição, dois eventos são independentes quando a probabilidade de ambos acontecerem juntos é igual ao produto das probabilidades de cada evento acontecer individualmente:

$$P(AB) = P(A)P(B) \tag{10.3}$$

Então, o valor esperado para o acontecimento conjunto de dois eventos independentes é dado por

$$E(AB) = P(A)P(B) * N \tag{10.4}$$

onde N é o número total de eventos.

Por exemplo, o valor esperado para a célula (1, 1) (preferência pelo PT, sexo masculino) é a probabilidade marginal do partido escolhido ser o PT, multiplicada pela probabilidade marginal da escolha ter sido feita pelo sexo masculino, multiplicada por N, ou seja

$$\frac{25+45}{150} * \frac{25+15+10}{150} * 150 = \frac{70 * 50}{150} \approx 23{,}3$$

Portanto, se a preferência pelo PT independe de o sexo do entrevistado ser masculino, então o valor esperado é de aproximadamente 23,3 ocorrências para essa combinação. A Tabela 10.2 mostra a tabela de contingência com os respectivos valores esperados.

Tabela 10.2: Tabela de Contingência e valores esperados: pesquisa de opinião sobre preferência partidária classificada por sexo.

Partido	**Sexo**		**Total**	**Valor esperado**	
	Masculino	**Feminino**		**Masculino**	**Feminino**
PT	25	45	70	(70*50)/150 ≈ **23,3**	(70*100)/150 ≈**46,7**
PSDB	15	25	40	(40*50)/150 ≈ **13,3**	(40*100)/150 ≈**26,7**
PMDB	10	30	40	(40*50)/150 ≈ **13,3**	(40*100)/150 ≈**26,7**
Total	50	100	**150**		

FONTE: Elaborado pelo autor.

A partir da Equação 10.2, o valor de X^2 é dado por

$$X^2 = \frac{(25-23{,}3)^2}{23{,}3} + \frac{(45-46{,}7)^2}{46{,}7} + \frac{(15-13{,}3)^2}{13{,}3} + \frac{(25-26{,}7)^2}{26{,}7} + \frac{(10-13{,}3)^2}{13{,}3} + \frac{(30-26{,}7)^2}{26{,}7} \approx 1{,}74$$

A probabilidade de X^2 ser aproximadamente 1,74 pode ser determinada pela função **pchisq**, do R, onde o número de graus de liberdade é dado por

$$gl = (número\ de\ linhas - 1) * (número\ de\ colunas - 1) \qquad (10.5)$$

donde se encontra p ≈ 0,42. Logo, o teste falha em rejeitar H_0 e, portanto, as variáveis partido e sexo são independentes.

Exercício: Construa um script R para determinar X^2, a partir dos dados da Tabela 10.2, e execute o teste qui-quadrado para testar a hipótese nula.

10.3 LABORATÓRIO 10

REFERÊNCIAS BIBLIOGRÁFICAS

BARBETTA, Pedro Alberto. Estatística Aplicada às Ciências Sociais. Florianópolis: Ed. Da UFSC, 2008.

FIELD, A.; MILES, J.; FIELD, Z. Discovering Statistics Using R. London: Sage, 2012.

GOMES, F. de B. C. Interações entre o Legislativo e o Executivo federal do Brasil na de_nição de políticas de interesse amplo: uma abordagem sistêmica, com aplicação na saúde. Tese (Doutorado em Ciência Política) _ Universidade do Estado do Rio de Janeiro - Instituto de Estudos Sociais e Políticos, Rio de Janeiro - RJ, 2011. Disponível em: <http://www.iesp.uerj.br/teses-online>. Acesso em: 30 jun. 2011.

MOTTA, V. T. Bioestatística. Caxias do Sul: Educs, 2006.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. J. R. Stat. Soc., Series A, v.135, p.370-384, 1972.

BRASIL. Ministério do Planejamento. Melhoria da gestão pública por meio da definição de um guia referencial para a medição do desempenho da gestão, e controle para o gerenciamento dos indicadores de eficiência, eficácia e de resultados do Programa Nacional de Gestão Pública e Desburocratização. Brasília: Ministério do Planejamento, 2009.

FIELD, A. Descobrindo a Estatística Utilizando o SPSS. São Paulo: Artmed, 2009.

Galton, Sir Francis (1894). Natural Inheritance. Macmillan. pp. 63f.

Holm, S. (1979) A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics. 6, 65-70.

PAREONLINE - Practical Assessment, Research and Evaluation. [S.l.], 2013. Disponível em: <http://www.pareonline.net>. Acesso em: 12 jun. 2013.

THE R Project for Statistical Computing. [S.l.], 2013. Disponível em: <http://www.r-project.org>. Acesso em: 12 jun. 2013.

Tetlock, P. E. Expert Political Judgment: How Good Is It? How Can We Know? 2006

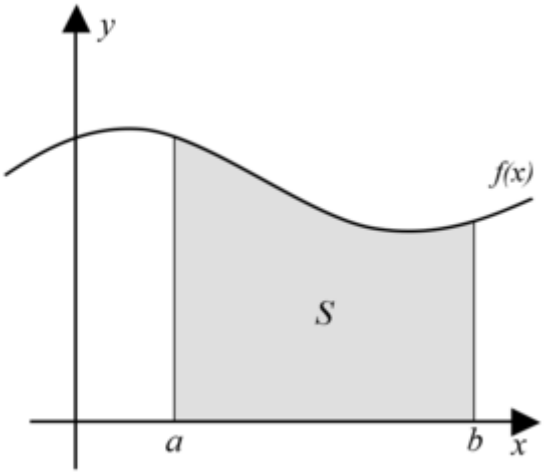
TUKEY, J.W. The problem of multiple comparisons. Mimeographs Princeton University, Princeton, N.J., 1953.

Rauen, 2012, http://linguagem.unisul.br/paginas/ensino/pos/linguagem/eventos/simfop/artigos_IV%20sfp/_F%C3%A1bio_Rauen.pdf

Steve McKillup, 2005. statistics-explained-an-introductory-guide-for-life-scientists (pg 97 a 101)

APÊNDICE I – INTEGRAL

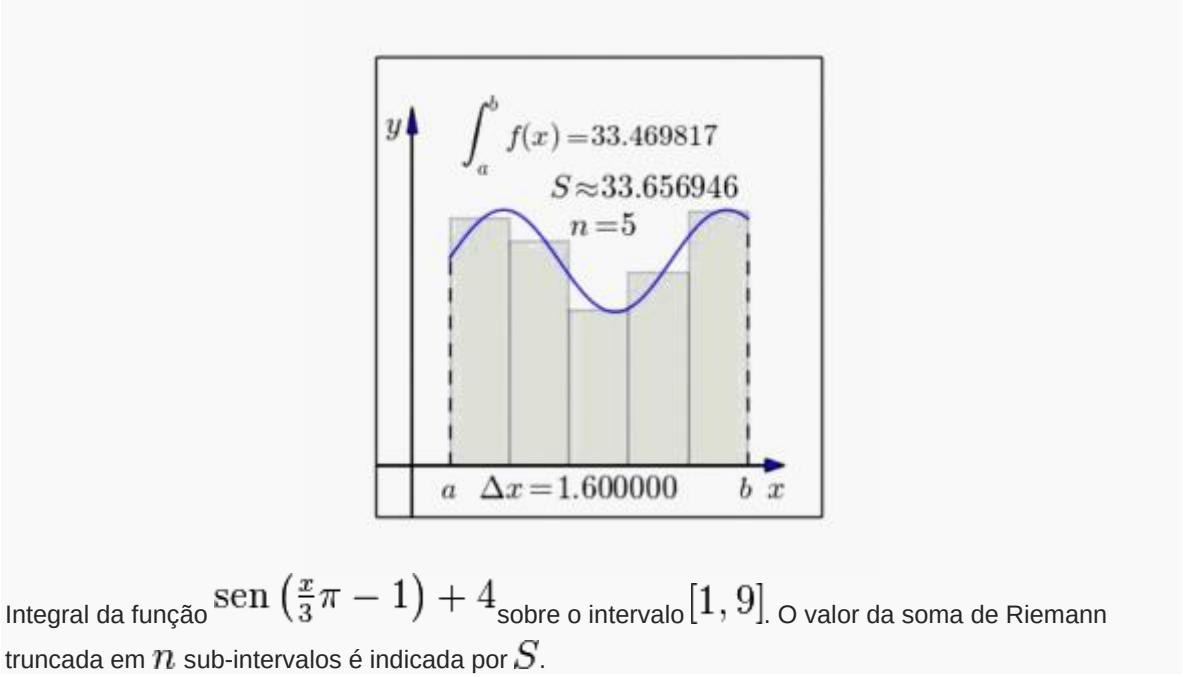
Integral definida - DEFINIÇÃO FORMAL E NOTAÇÃO



Integrando a área de uma função abaixo de uma curva.

Seja f uma função contínua definida no intervalo $[a,\ b]$. A **integral definida** desta função é denotada como:

Em linguagemmatemática	**Em português**
$S = \int_a^b f(x)dx$	S é a integral da função $f(x)$, no intervalo entre a e b. $\int$ é o sinal da integral, $f(x)$ é o integrando e os pontos a e b são os limites (inferior e superior, respectivamente) de integração.
Onde $f : [a, b] \to \mathbb{R}$	f é uma função com domínio no espaço fechado [a,b] (com $a \leq x \leq b$) e com imagem no conjunto dos números reais



Integral da função $\operatorname{sen}\left(\frac{x}{3}\pi - 1\right) + 4$ sobre o intervalo $[1, 9]$. O valor da soma de Riemann truncada em n sub-intervalos é indicada por S.

A ideia desta notação utilizando um *S* comprido é generalizar a noção de somatório[4] . Isto porque, intuitivamente, a integral de $f(x)$ sobre o intervalo $[a, b]$ pode ser entendida como a soma de pequenos retângulos de base Δx tendendo a zero e altura $f(x_i^*)$, onde o produto $f(x_i^*)\Delta x$ é a área deste retângulo. A soma de todas estas pequenas áreas (áreas infinitesimais), fornece a área entre a curva $y = f(x)$ e o eixo das abscissas. Mais precisamente, pode-se dizer que a integral acima é o valor limite da soma:[3]

Em linguagem matemática	Em Português
$\int_a^b f(x)dx = \lim_{\Delta x \to 0} \sum_{i=0}^{n} f(x_i^*)\Delta x$	A integral de $f(x)$ no intervalo [a,b] é igual ao limite do somatório de cada um dos valores que a função f(x) assume, de 0 a n, multiplicados por Δx. O que se espera é que quando n for muito grande o valor da soma acima se aproxime do valor da área abaixo da curva e, portanto, da integral de $f(x)$ no intervalo. Ou seja, que o limite esteja definido. A definição de integral aqui apresentada é chamada de soma de Riemann, mas há outras formas (equivalentes).
onde $\Delta x = \frac{b-a}{n}$	Comprimento dos pequenos subintervalos nos quais se divide o intervalo [a,b]. Os extremos destes intervalos são

	os números $x_0\ (=a)\ , x_1, ...x_n\ (=b)$.
onde $x_i^* = \lim_{\Delta x \to 0} i \cdot \Delta x + a$	Equivale a um ponto num intervalo de a até b da função quando o valor do número de termos n tende a infinito ou equivalentemente quando o valor de Δx tende a 0,nesse caso a letra i define o enésimo termo de uma sequência infinita ligada aos valores que cada x_i^* assumirá.
onde $f(x_i^*)$	Valor ("altura") da função $f(x)$ quando x é igual ao ponto amostral x_i^*, definido como um ponto que está no subintervalo $[x_{i-1}, x_i]$ (podendo até mesmo ser um destes pontos extremos do subintervalo).

Uma integral definida pode ser própria ou imprópria, convergente ou divergente. Neste último caso, ela representa uma área infinita.

Integral Indefinida

A integral indefinida de $f(x)$ é a função (ou família de funções) definida por [5] [6] :

$$\int f(x)dx = F(x) + C$$

em que C é uma constante indeterminada e $F(x)$ é uma **antiderivada** de $f(x)$, i.e. $F'(x) = f(x)$. A notação $\int f(x)dx$ é lida como: a integral de $f(x)$ em relação a x.

APÊNDICE II – REGRESSÃO COM TRÊS REGRESSORES – RESÍDUOS

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i})x_{1i} = 0$$

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i})x_{2i} = 0$$

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i})x_{3i} = 0$$

Substituindo $\hat{\beta}_1$ na terceira/segunda equação, temos

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i})x_{1i}}{\sum_{i=1}^{n} x_{1i}^2}$$

$$\sum_{i=1}^{n}(y_i - \frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i})x_{1i}}{\sum_{i=1}^{n} x_{1i}^2} x_{1i} - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i})x_{3i} = 0$$

$$\sum_{i=1}^{n}(y_i - \frac{\sum_{i=1}^{n} y_i x_{1i}}{\sum_{i=1}^{n} x_{1i}^2} x_{1i} + \hat{\beta}_2 \frac{\sum_{i=1}^{n} x_{2i} x_{1i}}{\sum_{i=1}^{n} x_{1i}^2} x_{1i} + \hat{\beta}_3 \frac{\sum_{i=1}^{n} x_{3i} x_{1i}}{\sum_{i=1}^{n} x_{1i}^2} x_{1i} - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i})x_{3i} = 0$$

$$\sum_{i=1}^{n}(y_i - \frac{\sum_{i=1}^{n} y_i x_{1i}}{\sum_{i=1}^{n} x_{1i}^2} x_{1i} - \hat{\beta}_2 \left[x_{2i} - \frac{\sum_{i=1}^{n} x_{2i} x_{1i}}{\sum_{i=1}^{n} x_{1i}^2} x_{1i} \right] - \hat{\beta}_3 \left[x_{3i} - \frac{\sum_{i=1}^{n} x_{3i} x_{1i}}{\sum_{i=1}^{n} x_{1i}^2} x_{1i} \right])x_{3i} = 0$$

$$\sum_{i=1}^{n}(e_{i,Y|x_1} - \hat{\beta}_2 e_{i,x_2|x_1} - \hat{\beta}_3 e_{i,x_3|x_1})x_{3i} = 0$$

$$\sum_{i=1}^{n}(e_{i,Y|x_1} - \hat{\beta}_2 e_{i,x_2|x_1} - \hat{\beta}_3 e_{i,x_3|x_1})x_{2i} = 0$$

Substituindo $\hat{\beta}_2$ na primeira/terceira equação, temos

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_3 x_{3i})x_{2i}}{\sum_{i=1}^{n} x_{2i}^2}$$

$$\sum_{i=1}^{n}(y_i-\widehat{\beta}_1x_{1i}-\frac{\sum_{i=1}^{n}(y_i-\widehat{\beta}_1x_{1i}-\widehat{\beta}_3x_{3i})x_{2i}}{\sum_{i=1}^{n}x_{2i}^2}x_{2i}-\widehat{\beta}_3x_{3i})x_{1i}=0$$

$$\sum_{i=1}^{n}(y_i-\widehat{\beta}_1x_{1i}-\frac{\sum_{i=1}^{n}y_ix_{2i}}{\sum_{i=1}^{n}x_{2i}^2}x_{2i}+\widehat{\beta}_1\frac{\sum_{i=1}^{n}x_{1i}x_{2i}}{\sum_{i=1}^{n}x_{2i}^2}x_{2i}+\widehat{\beta}_3\frac{\sum_{i=1}^{n}x_{3i}x_{2i}}{\sum_{i=1}^{n}x_{2i}^2}x_{2i}-\widehat{\beta}_3x_{3i})x_{1i}=0$$

$$\sum_{i=1}^{n}(y_i-\frac{\sum_{i=1}^{n}y_ix_{2i}}{\sum_{i=1}^{n}x_{2i}^2}x_{2i}-\left[x_{1i}-\frac{\sum_{i=1}^{n}x_{1i}x_{2i}}{\sum_{i=1}^{n}x_{2i}^2}x_{2i}\right]\widehat{\beta}_1-\left[x_{3i}-\frac{\sum_{i=1}^{n}x_{3i}x_{2i}}{\sum_{i=1}^{n}x_{2i}^2}x_{2i}\right]\widehat{\beta}_3)x_{1i}=0$$

$$\sum_{i=1}^{n}(e_{i,Y|x_2}-\widehat{\beta}_1e_{i,x_1|x_2}-\widehat{\beta}_3e_{i,x_3|x_2})x_{1i}=0$$

$$\sum_{i=1}^{n}(e_{i,Y|x_2}-\widehat{\beta}_1e_{i,x_1|x_2}-\widehat{\beta}_3e_{i,x_3|x_2})x_{3i}=0$$

Substituindo $\widehat{\beta}_3$ na primeira/segunda equação, temos

$$\widehat{\beta}_3=\frac{\sum_{i=1}^{n}(y_i-\widehat{\beta}_1x_{1i}-\widehat{\beta}_2x_{2i})x_{3i}}{\sum_{i=1}^{n}x_{3i}^2}$$

$$\sum_{i=1}^{n}(y_i-\widehat{\beta}_1x_{1i}-\widehat{\beta}_2x_{2i}-\frac{\sum_{i=1}^{n}(y_i-\widehat{\beta}_1x_{1i}-\widehat{\beta}_2x_{2i})x_{3i}}{\sum_{i=1}^{n}x_{3i}^2}x_{3i})x_{1i}=0$$

$$\sum_{i=1}^{n}(y_i-\widehat{\beta}_1x_{1i}-\widehat{\beta}_2x_{2i}-\frac{\sum_{i=1}^{n}y_ix_{3i}}{\sum_{i=1}^{n}x_{3i}^2}x_{3i}+\widehat{\beta}_1\frac{\sum_{i=1}^{n}x_{1i}x_{3i}}{\sum_{i=1}^{n}x_{3i}^2}x_{3i}+\widehat{\beta}_2\frac{\sum_{i=1}^{n}x_{2i}x_{3i}}{\sum_{i=1}^{n}x_{3i}^2}x_{3i})x_{1i}=0$$

$$\sum_{i=1}^{n}(y_i-\frac{\sum_{i=1}^{n}y_ix_{3i}}{\sum_{i=1}^{n}x_{3i}^2}x_{3i}-\widehat{\beta}_1\left[x_{1i}-\frac{\sum_{i=1}^{n}x_{1i}x_{3i}}{\sum_{i=1}^{n}x_{3i}^2}x_{3i}\right]-\widehat{\beta}_2\left[x_{2i}-\frac{\sum_{i=1}^{n}x_{2i}x_{3i}}{\sum_{i=1}^{n}x_{3i}^2}x_{3i}\right])x_{1i}=0$$

$$\sum_{i=1}^{n}(e_{i,Y|x_3}-\widehat{\beta}_1e_{i,x_1|x_3}-\widehat{\beta}_2e_{i,x_2|x_3})x_{1i}=0$$

$$\sum_{i=1}^{n}(e_{i,Y|x_3}-\widehat{\beta}_1e_{i,x_1|x_3}-\widehat{\beta}_2e_{i,x_2|x_3})x_{2i}=0$$

$$\sum_{i=1}^{n}(e_{i,Y|x_3} - \hat{\beta}_1 e_{i,x_1|x_3} - \hat{\beta}_2 e_{i,x_2|x_3})x_{1i} = 0$$

$$\sum_{i=1}^{n}(e_{i,Y|x_1} - \hat{\beta}_2 e_{i,x_2|x_1} - \hat{\beta}_3 e_{i,x_3|x_1})x_{2i} = 0$$

$$\sum_{i=1}^{n}(e_{i,Y|x_2} - \hat{\beta}_1 e_{i,x_1|x_2} - \hat{\beta}_3 e_{i,x_3|x_2})x_{3i} = 0$$

APÊNDICE III – RESÍDUO ESTUDENTIZADO

The population variance σ^2 is unknown and can be estimated by MSE, the mean square error.

Semistudentized residuals are defined as

$$e_i^* = \frac{e_i}{\sqrt{MSE}}$$

but, since the variance of residuals depends on both σ^2 and X, their estimated variance is:

$$\widehat{V}(e_i) = MSE(1 - h_{ii})$$

where h_{ii} is the ith diagonal element of the hat matrix.

Standardized residuals, also called **internally studentized residuals**, are:

$$r_i = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}}$$

However the single e_i and MSE are non independent, so r_i can't have a t distribution. The procedure is then to delete the ith observation, fit the regression function to the remaining $n - 1$ observations, and get new $\hat{y}$'s which can be denoted by $\hat{y}_{i(i)}$. The difference:

$$d_i = y_i - \hat{y}_{i(i)}$$

is called *deleted residual*. An equivalent expression that does not require a recomputation is:

$$d_i = \frac{e_i}{1 - h_{ii}}$$

Denoting the new X and MSE by $X_{(i)}$ and $MSE_{(i)}$, since they do not depend on the ith observation, we get:

$$t_i = \frac{d_i}{\sqrt{\frac{MSE_{(i)}}{1-h_{ii}}}} = \frac{e_i}{\sqrt{MSE_{(i)}(1 - h_{ii})}} \sim t_{n-p-1}$$

The t_i's are called **studentized** (deleted) **residuals**, or **externally studentized residuals**.

See Kutner et al., Applied Linear Statistical Models, Chapter 10.

APÊNDICE IV – CÓDIGOS EM R

PESQUISA ESTATÍSTICA

```
#########################################################
# Tabela 2.1 Quantidade de eleitores por unidade da federação
# separados por sexo
#########################################################
# remove todas as variáveis da memória
rm(list = ls(all = TRUE))
# leitura do arquivo
tse <- read.table("perfil_eleitorado_2012.txt",sep=";") # com fatores
# nominando as variaveis
names(tse)<-
c("PERIODO","UF","MUNICIPIO","COD_MUNICIPIO_TSE","NR_ZONA","SEXO","FAIXA_
ETARIA","GRAU_DE_ESCOLARIDADE","QTD_ELEITORES_NO_PERFIL")
# domínio da variável
uf <- levels(tse$UF)
sexo <- levels(tse$SEXO)
# construindo a tabela de frequências
fem<-c()
masc<-c()
ninf<-c()
total<-c()
i=1
for (i in 1:length(uf)) {
   fem<-c(fem,sum(tse$QTD_ELEITORES_NO_PERFIL[tse$UF==uf[i] &
          tse$SEXO==sexo[1]]))
   masc<-c(masc,sum(tse$QTD_ELEITORES_NO_PERFIL[tse$UF==uf[i] &
           tse$SEXO==sexo[2]]))
   ninf<-c(ninf,sum(tse$QTD_ELEITORES_NO_PERFIL[tse$UF==uf[i] &
           tse$SEXO==sexo[3]]))
   total<-c(total,sum(tse$QTD_ELEITORES_NO_PERFIL[tse$UF==uf[i]]))
}
tfreq<-data.frame(UF=uf, Feminino=fem, Masculino=masc, NaoInformado=ninf,
   Total=total)
tfreq
write.csv(tfreq,"tabela21.csv")
```

```
#########################################################
# Tabela 2.2 - Estado civil dos eleitores do estado do Paraná em 2012
#########################################################
# remove todas as variáveis da memória
rm(list = ls(all = TRUE))
# leitura do arquivo
tse <- read.table("consulta_cand_2012_PR.txt",sep=";") # com fatores
names(tse)[3]="ANO_ELEICAO"
names(tse)[34]="DESCRICAO_ESTADO_CIVIL"
# domínio da variável
t<-table(tse$DESCRICAO_ESTADO_CIVIL,tse$ANO_ELEICAO)
t2<-t
for (i in 2:length(t2)) { t2[i]=t2[i]+t2[i-1] }
t <- cbind(t,t/sum(t),t2,t2/sum(t))
t
write.csv(t,"tabela22.csv",quote=FALSE)
```

```
#########################################################
# Tabela 2.3 - Quantidade de eleitores por faixa etária
#########################################################
# remove todas as variáveis da memória
rm(list = ls(all = TRUE))
# leitura do arquivo
tse <- read.table("perfil_eleitorado_2012.txt",sep=";") # com fatores
names(tse)[1]="PERIODO"
names(tse)[6]="SEXO"
names(tse)[7]="FAIXA_ETARIA"
# tabela de frequências
cbind(table(tse$FAIXA_ETARIA,tse$SEXO),
      table(tse$FAIXA_ETARIA,tse$PERIODO))


#########################################################
# Figura 2.1 - Histograma: quantidade de eleitores vs. UF
# (repositório de dados do TSE de abril de 2013).
#########################################################
# remove todas as variáveis da memória
rm(list = ls(all = TRUE))
# leitura do arquivo
tse <- read.table("perfil_eleitorado_2012.txt",sep=";") # com fatores
# nominando as variaveis
names(tse)[2]="UF"
names(tse)[9]="QTD_ELEITORES_NO_PERFIL"
# domínio da variável
uf <- levels(tse$UF)
# construindo a tabela de frequências
total<-c()
i=1
for (i in 1:length(uf)) {
  total<-c(total,sum(tse$QTD_ELEITORES_NO_PERFIL[tse$UF==uf[i]]))
}
hist(total[uf!="ZZ"], breaks=20, main="Eleitores nas UFs",
     xlab="Qtd. de Eleitores (milhões)",ylab="Qtd. de UFs",
     col="orange", axes = FALSE)
axis(1,c(0, 5e6, 10e6, 15e6, 20e6, 25e6, 30e6),c(0,5,10,15,20,25,30))
axis(2,0:9,0:9)


#########################################################
# Figura 2.2 - Distribuição simétrica
#########################################################
# remove todas as variáveis da memória
library(ESPRESSO)
rm(list = ls(all = TRUE))
N=1300000
windows(4,4)
par(mfrow=c(1,1),no.readonly=FALSE, cex = 0.7)
hist(skew.rnorm(num.obs  =  N,  mean  =  0,  sd  =  1,  skewness  =
         0),freq=FALSE,col="orange",
     main="Simétrica", xlab="x",ylab="Frequência")
lines(density(skew.rnorm(num.obs  =  N,  mean  =  0,  sd  =  1,  skewness  =
         0)),lwd=3,col="blue")


#########################################################
# Figura 2.4 - Distribuições assimétricas
```

```
#########################################################
# remove todas as variáveis da memória
library(ESPRESSO)
rm(list = ls(all = TRUE))
N=1300000
windows(6,4)
par(mfrow=c(1,2),no.readonly=FALSE, cex = 0.7)
hist(skew.rnorm(num.obs  =  N,  mean  =  0,  sd  =  1,  skewness  =
         80),freq=FALSE,col="orange",
     main="Assimetria Positiva", xlab="x",ylab="Frequência")
lines(density(skew.rnorm(num.obs  =  N,  mean  =  0,  sd  =  1,  skewness  =
         80)),lwd=3,col="blue")
hist(skew.rnorm(num.obs  =  N,  mean  =  0,  sd  =  1,  skewness  =  -
         80),freq=FALSE,col="orange",
     main="Assimetria Negativa", xlab="x",ylab="Frequência")
lines(density(skew.rnorm(num.obs  =  N,  mean  =  0,  sd  =  1,  skewness  =  -
         80)),lwd=3,col="blue")
```

DISTRIBUIÇÃO DE PROBABILIDADE

```
#########################################################
# Figura 3.1 - Relação entre a quantidade de votos nulos e a
# quantidade de seções eleitorais em que esses votos ocorreram
# no estado do Paraná, no primeiro turno das eleições de 2012.
#########################################################
# remove todas as variáveis da memória
rm(list = ls(all = TRUE))
# leitura do arquivo
tse <- read.table("votacao_secao_2012_PR.txt",sep=";") # com fatores
# nominando as variaveis
names(tse)[14]="NUM_VOTAVEL"
names(tse)[15]="QTDE_VOTOS"
# domínio da variável
hist(tse$QTDE_VOTOS[tse$NUM_VOTAVEL=="96"],    breaks=10,    main="Seções
         eleitorais vs. votos nulos",
 xlab="Quantidade de votos nulos",ylab="Seções Eleitorais", col="orange")


#########################################################
# Tabela 3.1 - Seções eleitorais em que ocorreram votos nulos
# no estado do Paraná, no primeiro turno das eleições de 2012
#########################################################
# remove todas as variáveis da memória
rm(list = ls(all = TRUE))
# leitura do arquivo
tse <- read.table("votacao_secao_2012_PR.txt",sep=";") # com fatores
# nominando as variaveis
names(tse)[14]="NUM_VOTAVEL"
names(tse)[15]="QTDE_VOTOS"
# intervalo de classe
ic<-10
# votos nulos
vn<-tse$QTDE_VOTOS[tse$NUM_VOTAVEL=="96"]
# faixa de valores
mini <- min(vn)
maxi <- max(vn)
faixa <-(maxi - mini + 1)
# qtd de intervalos
qint <- ceiling(faixa/ic)
# vetores de rótulo e frequência
rotulo<-NULL
freq<-NULL
# monta os vetores
for(i in 1:qint){
    rotulo <- c(rotulo,paste(mini+(i-1)*ic,"-",mini+i*ic-1))
    freq <- c(freq,length(vn[vn>=mini+(i-1)*ic & vn<=mini+i*ic-1]))
}
df<-data.frame(rotulo,freq)
fA<-df$freq
for (i in 2:length(fA)) { fA[i]=fA[i]+fA[i-1] }
df<-data.frame(df$rotulo,   df$freq,   round(df$freq/sum(df$freq),4),   fA,
         round(fA/fA[length(fA)],4))
colnames(df)<-c("rótulo","f","fr","A","Ar")
```

```
df


#################################################
# Figura 3.2 - Distribuição das idades completas dos
# candidatos do Paraná ...
#################################################
# remove todas as variáveis da memória
rm(list = ls(all = TRUE))
# leitura do arquivo
tse <- read.table("consulta_cand_2012_PR.txt",sep=";") # com fatores
# renomeia a variável V26
names(tse)[26] <- "DATA_NASCIMENTO"
# converte o campo DATA_NASCIMENTO para o formato data
dtnasc <- as.Date(tse$DATA_NASCIMENTO, format="%d/%m/%Y")
# determina a idade dos candidatos na data da eleição de 2012 (7/10/2012)
dteleicao <- as.Date("7/10/2012", format="%d/%m/%Y")
idade2012 <- difftime(dteleicao,dtnasc,units="days")
# converte o resultado para a idade em anos
idade2012 <- floor(as.numeric(idade2012)/365.25)
idade2012 <- idade2012[idade2012 < 150]
# histograma
hist(idade2012, breaks="Sturges", main="Candidatos do Paraná",
 xlab="Idade completa em 7/Out/2012",ylab="Frequência", col="orange")
# informações da população
N <- length(idade2012)
text(x=70,y=4000,paste("N =",N))
par(font=5)
text(70,5000,paste("m = ",round(mean(idade2012),2)))
text(70,4500,paste("s = ",round(sd(idade2012),2)))
par(font=1)
# determina as frequências relativas das classes e
# mostra sobre as barras
i <- 15
inc <- 5
while(i < 90){
   f <- length(idade2012[idade2012 > i & idade2012 <= i+inc])
   fr <- f/N
   x <- (i + (i + inc))/2
   y <- f + 150
   text(x,y, round(fr,3), cex = 0.6)
   # barras em vermelho
   if(i >= 20 & i < 40) {
     rect(i, 0, i+inc, length(idade2012[idade2012 > i & idade2012 <=
          i+inc]),
          col="red", border=par("fg"), lty=NULL, xpd=FALSE)
   }
   i <- i + inc
}


#################################################
# Figura 3.3 - Histogramas com intervalos de classe
# distintos para o mesmo conjunto de dados
#################################################
# remove todas as variáveis da memória
rm(list = ls(all = TRUE))
# leitura do arquivo
tse <- read.table("consulta_cand_2012_PR.txt",sep=";") # com fatores
# renomeia a variável V26
```

```
names(tse)[26] <- "DATA_NASCIMENTO"
# converte o campo DATA_NASCIMENTO para o formato data
dtnasc <- as.Date(tse$DATA_NASCIMENTO, format="%d/%m/%Y")
# determina a idade dos candidatos na data da eleição de 2012 (7/10/2012)
dteleicao <- as.Date("7/10/2012", format="%d/%m/%Y")
idade2012 <- difftime(dteleicao,dtnasc,units="days")
# converte o resultado para a idade em anos
idade2012 <- floor(as.numeric(idade2012)/365.25)
idade2012 <- idade2012[idade2012 < 150]

windows(8,4)
par(mfrow=c(1,3),no.readonly=FALSE, cex = 0.7)

# histograma 1
hist(idade2012, breaks="Sturges", main="\n\n(a) Intervalo de Classe = 5",
 xlab="",ylab="Frequência", col="orange", ylim = c(0,5500))
# informações da população
N <- length(idade2012)
# determina as frequências relativas das classes e
# mostra sobre as barras
i <- 15
inc <- 5
while(i < 90){
   f <- length(idade2012[idade2012 > i & idade2012 <= i+inc])
   fr <- f/N
   x <- (i + (i + inc))/2
   y <- f + 150
   text(x,y, round(fr,3), cex = 0.6)
   # barras em vermelho
   if(i >= 20 & i < 40) {
     rect(i, 0, i+inc, length(idade2012[idade2012 > i & idade2012 <=
          i+inc]),
          col="red", border=par("fg"), lty=NULL, xpd=FALSE)
   }
   i <- i + inc
}

# histograma 2
hist(idade2012, breaks=30, main="\n\n(b) Intervalo de Classe = 2",
 xlab="Idade completa em 7/Out/2012",ylab="", col="orange", ylim =
         c(0,5500))
# informações da população
text(x=50,y=4000,paste("N =",N))
par(font=5)
text(50,5000,paste("m = ",round(mean(idade2012),2)))
text(50,4500,paste("s = ",round(sd(idade2012),2)))
par(font=1)
# mostra sobre as barras
i <- 20
inc <- 2
while(i >= 20 & i < 40){
   # barras em vermelho
   rect(i, 0, i+inc, length(idade2012[idade2012 > i & idade2012 <=
         i+inc]),
          col="red", border=par("fg"), lty=NULL, xpd=FALSE)
   i <- i + inc
}

# histograma 3
hist(idade2012, breaks=60, main="\n\n(c) Intervalo de Classe = 1",
 xlab="",ylab="", col="orange", ylim = c(0,5500))
```

```
# mostra sobre as barras
i <- 20
inc <- 1
while(i >= 20 & i < 40){
   # barras em vermelho
   rect(i,  0,  i+inc,  length(idade2012[idade2012  >  i  &  idade2012  <=
         i+inc]),
          col="red", border=par("fg"), lty=NULL, xpd=FALSE)
   i <- i + inc
}

par(mfrow=c(1,1),no.readonly=FALSE, cex = 0.7)
title("Candidatos do Paraná - Eleições de 2012\n\n\n")


#####################################################
# Figura 3.4 - Distribuição de probabilidades: aproximação
# do caso discreto (barras vermelhas) ao caso contínuo (linha preta)
#####################################################
# remove todas as variáveis da memória
rm(list = ls(all = TRUE))
# parâmetros da curva normal
media<-44.90
dp<-10.98
# função de densidade de probabilidade da curva normal
y <- function(x) {(1/sqrt(2*pi*dp^2)) * exp(-((x - media)^2)/(2*dp^2))}
# faixa considerada da curva normal
x<-20:50
# limites de idade em estudo
x1<-30
x2<-45

windows(5,6)
par(mfrow=c(2,2),no.readonly=FALSE, cex = 0.7)
# intervalos de classe
i<-c(5,3,1,0.1)
g<-c("(a)","(b)","(c)","(d)")
for(j in 1:length(i)){
   plot(x,sapply(x,y),col="white",ylab="Densidade",xlab="Idade",
        main=paste(g[j],"Intervalo de classe =",i[j]))
   lines(x,sapply(x,y))
   bar<-(x2-x1)/i[j]
   for(b in 1:bar){
      rect(x1+(b-1)*i[j], 0, x1+b*i[j], sapply((2*x1+2*b*i[j]-i[j])/2,y),
           col="red", border=NA, lty=NULL, xpd=FALSE)
   }
}


#########################################################
# Tabela 3.2 - Probabilidades da Curva Normal Padronizada
#########################################################
# remove todas as variáveis da memória
rm(list = ls(all = TRUE))

lpor<-NULL # larger portiion - área maior
spor<-NULL # smaller portion - área menor
y<-NULL
z<-NULL
# constroi o eixo x
x<-NULL
```

```
i<-0
inc<-0.01
while(i<=0.08){
  x<-c(x,i)
  i<-i+inc
}
x<-c(x,1.38,1.65,1.96) # acrescenta dois novos valores 1.38 e 1.96
# calcula as áreas
for(j in x){
  i<-integrate(dnorm, 0, j) # área da curva normal padronizada
  spor<-c(spor,round(0.5-as.numeric(i[1]),5))
  lpor<-c(lpor,1-round(0.5-as.numeric(i[1]),5))
  z<-c(z,j)
  y<-c(y,round((1/sqrt(2*pi)) * exp(-(j^2)/2),4))
}
cbind(z,lpor,spor,y)


#################################################
# Figura 3.5 - Proporções da área da curva normal padronizada
#################################################
# remove todas as variáveis da memória
rm(list = ls(all = TRUE))
# parâmetros da curva normal padronizada
media<-0
dp<-1
# função de densidade de probabilidade da curva normal
y <- function(x) {(1/sqrt(2*pi*dp^2)) * exp(-((x - media)^2)/(2*dp^2))}
# faixa considerada da curva normal
x<-NULL
lim<-media+4*dp
i<--lim
while(i < lim){
    x<-c(x,i)
    i<-i+0.01
}
# traça o gráfico
windows(5,5)
plot(x,sapply(x,y),ylab="Densidade",xlab="x",
       main="Curva normal padronizada",type="l",lwd=2,ylim=c(-0.18,0.4))
# linha horzontal em zero
lines(c(-lim,lim),c(0,0))
# texto
par(font=5)
yt <- -0.04 # posição y do texto
text(media-2*dp,yt,"m-2s",cex=0.8)
text(media-dp,yt,"m-s",cex=0.8)
text(media,yt,"m",cex=0.8)
text(media+dp,yt,"m+s",cex=0.8)
text(media+2*dp,yt,"m+2s",cex=0.8)
par(font=1)
# valor z
z <- dp+0.35
text(z,yt/2.8,"z",cex=0.7,font=4)
lines(c(z,z),c(0,sapply(z,y)),lty=4,type="l")
# hachura da área menor
inc <- (3*dp-z)/20
i<-z+inc
while(i <= 3*dp){
    lines(c(i,i),c(0,sapply(i,y)),col="gray",lwd=0.5)
    i<-i+inc
```

```
}
# indicação das áreas
text(3,0.1,"Área menor",cex=0.7,font=4)
text(-2.5,0.25,"Área maior",cex=0.7,font=4)
# linhas verticais sobre o desvio padrão
cor="blue"
lines(c(media-2*dp,media-2*dp),c(0,sapply(media-2*dp,y)),col=cor)
lines(c(media-dp,media-dp),c(0,sapply(media-dp,y)),col=cor)
lines(c(media,media),c(0,sapply(media,y)),col=cor)
lines(c(media+dp,media+dp),c(0,sapply(media+dp,y)),col=cor)
lines(c(media+2*dp,media+2*dp),c(0,sapply(media+2*dp,y)),col=cor)
# eixos percentuais
lines(c(media-dp,media+dp),c(-0.1,-0.1),col=cor)
lines(c(media-dp,media-dp),c(-0.1-0.02,-0.1+0.02),col=cor)
lines(c(media+dp,media+dp),c(-0.1-0.02,-0.1+0.02),col=cor)
text(media,-0.1+0.02,"68.26%",cex=0.7)
lines(c(media-2*dp,media+2*dp),c(-0.15,-0.15),col=cor)
lines(c(media-2*dp,media-2*dp),c(-0.15-0.02,-0.15+0.02),col=cor)
lines(c(media+2*dp,media+2*dp),c(-0.15-0.02,-0.15+0.02),col=cor)
text(media,-0.15+0.02,"95.44%",cex=0.7)
# probabilidades
integral <- function(f,a,b) {i<-integrate(f, a,b); as.numeric(i[1])}
yt <- 0.02 # posição y do texto
text(0.5*dp,yt,round(integral(y, media, media+dp),4),cex=0.7)
text(1.5*dp,yt,round(integral(y, media+dp, media+2*dp),4),cex=0.7)
text(2.25*dp,yt-0.006,round(integral(y,                        media+2*dp,
         media+4*dp),2),cex=0.6)
text(-0.5*dp,yt,round(integral(y, media-dp, media),4),cex=0.7)
text(-1.5*dp,yt,round(integral(y, media-2*dp, media-dp),4),cex=0.7)
text(-2.25*dp,yt-0.006,round(integral(y,        media-4*dp,        media-
         2*dp),2),cex=0.6)


#####################################################
# Figura 3.6 - FDA, idade dos candidatos do Paraná – Eleições 2012
#####################################################
# remove todas as variáveis da memória
rm(list = ls(all = TRUE))
# leitura do arquivo
tse <- read.table("consulta_cand_2012_PR.txt",sep=";") # com fatores
# renomeia a variável V26
names(tse)[26] <- "DATA_NASCIMENTO"
# converte o campo DATA_NASCIMENTO para o formato data
dtnasc <- as.Date(tse$DATA_NASCIMENTO, format="%d/%m/%Y")
# determina a idade dos candidatos na data da eleição de 2012 (7/10/2012)
dteleicao <- as.Date("7/10/2012", format="%d/%m/%Y")
idade2012 <- difftime(dteleicao,dtnasc,units="days")
# converte o resultado para a idade em anos
idade2012 <- floor(as.numeric(idade2012)/365.25)
idade2012 <- idade2012[idade2012 < 150]

# histograma acumulado
h <- hist(idade2012,freq = TRUE)
h.cum <- cumsum(h$counts/length(idade2012))
x <- h$breaks[-1]
plot(x, h.cum, pch=16,main="FDA - Candidatos do Paraná - Eleições de
         2012",
     xlab="Idade", ylab="Probabilidade")
lines(x, h.cum)
```

INFERÊNCIA ESTATÍSTICA

```
#####################################################
# Figura 4.1 - Lei dos Grandes Números em ação
#####################################################
library(ggplot2)
n <- 10000
media <- cumsum(sample(0:1, n, replace = TRUE))/(1:n)
g <- ggplot(data.frame(x = 1:n, y = media), aes(x = x, y = y))
g <- g + geom_hline(yintercept = 0.5) + geom_line(size = 2)
g <- g + labs(x = "Número de observações", y = "Média cumulativa")
g


#####################################################
# Figura 4.3 - Distribuição amostral das médias
#####################################################
# remove todas as variáveis da memória
rm(list = ls(all = TRUE))
windows(5,5)
plot.new()
plot.window(xlim=c(2,8), ylim=c(0,3))
title("Distribuição amostral das médias",xlab="Média das amostras",
       ylab="Frequência")
axis(1,2:8,c(2,3,4,5,6,7,8))
axis(2,0:3,c(0,1,2,3))
rect(2.5, 0, 3.5, 1, col="red", lty=NULL, xpd=FALSE)
rect(3.5, 0, 4.5, 2, col="red", lty=NULL, xpd=FALSE)
rect(4.5, 0, 5.5, 3, col="red", lty=NULL, xpd=FALSE)
rect(5.5, 0, 6.5, 2, col="red", lty=NULL, xpd=FALSE)
rect(6.5, 0, 7.5, 1, col="red", lty=NULL, xpd=FALSE)
box()
pop<-c(6,5,4,3,5,7,5,6,4)
par(font=5)
text(3.5,2.8,paste("m =",round(mean(pop),2)),cex=0.8)
text(3.5,2.5,paste("s =",round(sd(pop),2)),cex=0.8)
par(font=1)


#####################################################
# Figura 4.5 - Intervalos de confiança para 20 amostras com
# tamanho 30 e média populacional igual a 44,9, considerando-se
# nível de confiança de 95%
#####################################################
# remove todas as variáveis da memória
rm(list = ls(all = TRUE))
# leitura do arquivo
tse <- read.table("consulta_cand_2012_PR.txt",sep=";") # com fatores
# renomeia a variável V26
names(tse)[26] <- "DATA_NASCIMENTO"
# converte o campo DATA_NASCIMENTO para o formato data
dtnasc <- as.Date(tse$DATA_NASCIMENTO, format="%d/%m/%Y")
# determina a idade dos candidatos na data da eleição de 2012 (7/10/2012)
dteleicao <- as.Date("7/10/2012", format="%d/%m/%Y")
idade2012 <- difftime(dteleicao,dtnasc,units="days")
# converte o resultado para a idade em anos
```

```
idade2012 <- floor(as.numeric(idade2012)/365.25)
idade2012 <- idade2012[idade2012 < 150]

media<-mean(idade2012)
dp<-sd(idade2012)

# qtd. amostras
qta<-20

# plota o box
windows(5,6)
plot.new()
plot.window(xlim=c(media-1*dp,media+1*dp), ylim=c(0,qta))
title("Intervalos de Confiança (IC)", ylab="Amostra", xlab="Média da
         amostra e respectivo IC")
axis(1,round(media-5*dp):round(media+5*dp),round(media-
         5*dp):round(media+5*dp))
axis(2,1:qta,1:qta)
box()
# linha vertical na média
lines(c(media,media),c(-1,qta+1),col="blue")
# tamanho da amostra
n<-30
ep<-dp/sqrt(n) # erro padrao com base no dp populacional
# calcula os ICs e plota
for(i in 1:qta){
  amostra<-sample(idade2012,n)
  mediaa<-mean(amostra)
  #dpa<-sd(amostra)
  #ep<-dpa/sqrt(n) # erro padrao

  i1<-mediaa-1.96*ep
  i2<-mediaa+1.96*ep
  lines(c(i1,i2),c(i,i))
  lines(c(mediaa,mediaa),c(i,i),type="o")
}
par(font=5)
text(36,20,paste("m =",round(media,2)),col="blue")
par(font=1)


################################################
# Figura 4.6 - Intervalos de confiança com nível de
# confiança de 95% para diferentes tamanhos da amostra
################################################
# remove todas as variáveis da memória
rm(list = ls(all = TRUE))
# leitura do arquivo
tse <- read.table("consulta_cand_2012_PR.txt",sep=";") # com fatores
# renomeia a variável V26
names(tse)[26] <- "DATA_NASCIMENTO"
# converte o campo DATA_NASCIMENTO para o formato data
dtnasc <- as.Date(tse$DATA_NASCIMENTO, format="%d/%m/%Y")
# determina a idade dos candidatos na data da eleição de 2012 (7/10/2012)
dteleicao <- as.Date("7/10/2012", format="%d/%m/%Y")
idade2012 <- difftime(dteleicao,dtnasc,units="days")
# converte o resultado para a idade em anos
idade2012 <- floor(as.numeric(idade2012)/365.25)
idade2012 <- idade2012[idade2012 < 150]

media<-mean(idade2012)
```

```
dp<-sd(idade2012)

# qtd. amostras
qta<-20

windows(8,4)
par(mfrow=c(1,3),no.readonly=FALSE, cex = 0.7)
for(n in c(30,100,500)){ # n eh o tamanho da amostra

  # plota o box
  plot.new()
  plot.window(xlim=c(media-1*dp,media+1*dp), ylim=c(0,qta))
  title(paste("\nn =",n))
  axis(1,round(media-5*dp):round(media+5*dp),round(media-
         5*dp):round(media+5*dp))
  axis(2,1:qta,1:qta)
  box()
  # linha vertical na média
  lines(c(media,media),c(1,qta))
  if(n==100){
    par(font=5)
    text(38,20,paste("m =",round(media,2)))
    par(font=1)
  }

  ep<-dp/sqrt(n) # erro padrao com base no dp populacional

  for(i in 1:qta){
     amostra<-sample(idade2012,n)
     mediaa<-mean(amostra)
     # quando não se têm os parâmetros da população,
     # utilizamos as estimativas da amostra
     #dpa<-sd(amostra)
     #ep<-dpa/sqrt(n) # erro padrao

     i1<-mediaa-1.96*ep
     i2<-mediaa+1.96*ep
     lines(c(i1,i2),c(i,i))
     lines(c(mediaa,mediaa),c(i,i),type="o")
  }
}

par(mfrow=c(1,1),no.readonly=FALSE, cex = 0.7)
title(ylab="Amostra", xlab="Média da amostra e respectivo IC")


################################################
# Figura 4.7 - Comparação de ICs a 95% de confiança:
# a) amostras da mesma população; b) amostras de populações distintas
################################################
# remove todas as variáveis da memória
rm(list = ls(all = TRUE))

windows(8,2)
par(mfrow=c(1,2),no.readonly=FALSE, cex = 0.7)
# Figura 3.10a - médias pertencentes à mesma população
plot.new()
plot.window(xlim=c(30,90), ylim=c(0.5,2.5))
axis(1,30:90,30:90)
axis(2,0:2,0:2)
box()
```

```
grid()
media<-45
ep<-12
lines(c(media-ep,media+ep),c(1,1))
lines(c(media-ep,media-ep),c(0.9,1.1))
lines(c(media+ep,media+ep),c(0.9,1.1))
lines(c(media,media),c(1,1),type="o")
media<-43
ep<-9
lines(c(media-ep,media+ep),c(2,2))
lines(c(media-ep,media-ep),c(1.9,2.1))
lines(c(media+ep,media+ep),c(1.9,2.1))
lines(c(media,media),c(2,2),type="o")
title("(a)",ylab="Amostra", xlab="Média da amostra e respectivo IC")

# Figura 3.10b - médias de populações diferentes
plot.new()
plot.window(xlim=c(30,90), ylim=c(0.5,2.5))
axis(1,30:90,30:90)
axis(2,0:2,0:2)
box()
grid()
media<-45
ep<-12
lines(c(media-ep,media+ep),c(1,1))
lines(c(media-ep,media-ep),c(0.9,1.1))
lines(c(media+ep,media+ep),c(0.9,1.1))
lines(c(media,media),c(1,1),type="o")
media<-76
ep<-9
lines(c(media-ep,media+ep),c(2,2))
lines(c(media-ep,media-ep),c(1.9,2.1))
lines(c(media+ep,media+ep),c(1.9,2.1))
lines(c(media,media),c(2,2),type="o")
title("(b)",ylab="Amostra", xlab="Média da amostra e respectivo IC")
```

TESTES ESTATÍSTICOS

```
################################################
# Figura 5.1 - Normal padronizaza Zcalc = 2.54
################################################
# remove todas as variáveis da memória
rm(list = ls(all = TRUE))

z1 <- 2.54

# eixo x da curva normal padronizada
lim <- c(-4,4)
x <- seq(lim[1], lim[2], by = 0.01)

# traça a curva normal padronizada
plot(x,dnorm(x,0,1),ylab="Densidade",xlab="x",
     main="Curva normal padronizada",type="l",lwd=2,ylim=c(-0.05,0.4))
# linha horzontal em zero
lines(c(lim[1],lim[2]),c(0,0))

# curva normal padronizada
cnp <- function(x) {dnorm(x,0,1)} # curva normal padronizada

# valores de z
if(!is.null(z1)) {
    text(z1,-0.02,paste("Zcalc=",round(z1,2)),cex=0.7,font=4)
    lines(c(z1,z1),c(0,cnp(z1)),lty=4,type="l")
}

# probabilidades
integral <- function(f,a,b) {i<-integrate(f,a,b); as.numeric(i[1])}

# hachura da área
if(!is.null(z1)) {
    z2 <- lim[2]
    inc <- (z2-z1)/50
    i<-z1+inc
    while(i < z2){
        lines(c(i,i),c(0,cnp(i)),col="red",lwd=0.5)
        i<-i+inc
    }
    phachura<-round(integral(cnp,z1,z2),4)
    text(-4.3,0.38,paste("Área hachurada =",phachura),cex=0.8,pos=4)
    text(-4.3,0.35,paste("Área branca =",1-phachura),cex=0.8,pos=4)
}



################################################
# Figura 5.2 - Região crítica
################################################
# remove todas as variáveis da memória
rm(list = ls(all = TRUE))

# eixo x da curva normal padronizada
x<-NULL
lim<-4
```

```
i<--lim
while(i < lim){
    x<-c(x,i)
    i<-i+0.01
}
# cria a janela
windows(8,4)
par(mfrow=c(1,3),no.readonly=FALSE, cex = 0.7)
# curva normal padronizada
cnp <- function(x) {dnorm(x,0,1)} # curva normal padronizada
par(font=5)

# Unilateral à esquerda
# traça a curva normal padronizada
plot(x,cnp(x),ylab="Densidade",xlab="z",
        main="Unilateral à esquerda",type="l",lwd=2,ylim=c(-0.05,0.4))
# linha horzontal em zero
lines(c(-lim,lim),c(0,0))

# valores de z = -1.645 -> alpha = 0.05
z1<--1.645
lines(c(z1,z1),c(0,cnp(z1)),lty=4,type="l")

# hachura da área antes de z
inc <- (z1+3)/20
i<--3+inc
while(i <= z1){
   lines(c(i,i),c(0,cnp(i)),col="gray",lwd=0.5)
   i<-i+inc
}
text(-3,0.05,"a")

# Unilateral à direita
# traça a curva normal padronizada
plot(x,cnp(x),ylab="Densidade",xlab="z",
        main="Unilateral à direita",type="l",lwd=2,ylim=c(-0.05,0.4))
# linha horzontal em zero
lines(c(-lim,lim),c(0,0))

# valores de z = 1.645 -> alpha = 0.05
z1<-1.645
lines(c(z1,z1),c(0,cnp(z1)),lty=4,type="l")

# hachura da área depois de z
inc <- (3-z1)/20
i<-z1+inc
while(i <= 3){
   lines(c(i,i),c(0,cnp(i)),col="gray",lwd=0.5)
   i<-i+inc
}
text(3,0.05,"a")


# Bilateral
# traça a curva normal padronizada
plot(x,cnp(x),ylab="Densidade",xlab="z",
        main="Bilateral",type="l",lwd=2,ylim=c(-0.05,0.4))
# linha horzontal em zero
lines(c(-lim,lim),c(0,0))

# valores de z = 1.96 -> alpha/2 = 0.025
```

```
z1<--1.96
z2<-1.96
lines(c(z1,z1),c(0,cnp(z1)),lty=4,type="l")
lines(c(z2,z2),c(0,cnp(z2)),lty=4,type="l")

# hachura da área antes de z1
inc <- (z1+3)/20
i<--3+inc
while(i <= z1){
   lines(c(i,i),c(0,cnp(i)),col="gray",lwd=0.5)
   i<-i+inc
}
text(-3,0.05,"a/2")

# hachura da área depois de z2
inc <- (3-z2)/20
i<-z2+inc
while(i <= 3){
   lines(c(i,i),c(0,cnp(i)),col="gray",lwd=0.5)
   i<-i+inc
}
text(3,0.05,"a/2")


##########################################################
# Figura 5.5 - Distribuições t de Student
##########################################################
# Distribuições t-Student
# com vários graus de liberdade
# comparadas com a distribuição normal
x <- seq(-4.5, 4.5, .01)
normal.curve <- dnorm(x)

df <- c(1, 5, 20)
colors <- c("purple", "red", "green", "black")
labels <- c("df=1", "df=10", "df=20", "normal")

plot(x, normal.curve, , lty=2, xlab="x",
     ylab="Densidade", main="Distribuições t com diferentes graus de
          liberdade")

for (i in 1:3){
    lines(x, dt(x,df[i]), lwd=2, col=colors[i])
}

legend("topleft", inset=.1, title="Distribuições",
       labels, lwd=2, lty=c(1, 1, 1, 2), col=colors)


##########################################################
# Figura 5.6 - Dados de Gosset, teste t
##########################################################
data(sleep)

boxplot(sleep$extra ~ sleep$group, names=c("droga 1", "droga 2"),
        col = "bisque", ylab="horas de sono adicionadas")
```

```
#########################################################
# Figura 5.7 - Dados de Gosset, teste t
#########################################################
data(sleep)
g1 <- sleep$extra[1 : 10]; g2 <- sleep$extra[11 : 20]; n <- 10
plot(c(0.5, 2.5), range(g1, g2), type = "n", frame = FALSE, xlab = "Tipo
da droga", ylab = "Extra", axes = FALSE)
axis(2)
axis(1, at = 1 : 2, labels = c("droga 1", "droga 2"))
for (i in 1 : n) lines(c(1, 2), c(g1[i], g2[i]), lwd = 2, col = "red")
for (i in 1 : n) points(c(1, 2), c(g1[i], g2[i]), lwd = 2, col = "black",
         bg = "salmon", pch = 21, cex = 3)



#########################################################
# Figura 5.8 - QQ Plot
#########################################################
# remove todas as variáveis da memória
rm(list = ls(all = TRUE))
windows(6,4)
par(mfrow=c(1,2),no.readonly=FALSE, cex = 0.7)

qqnorm(rnorm(100, mean = 5, sd = 3),main="(a) Distribuição Normal",xlab =
         "Quantis Teóricos", ylab = "Quantis da Amostra")
qqline(rnorm(100, mean = 5, sd = 3))
qqnorm(runif(100, min = 2, max = 4),main="(b) Distribuição Uniforme",xlab
         = "Quantis Teóricos", ylab = "Quantis da Amostra")
qqline(runif(100, min = 2, max = 4))
```

CORRELAÇÃO LINEAR E REGRESSÃO LINEAR SIMPLES

```
#####################################################
# Figura 6.1 - Desvio dos dados em relação à média
#####################################################
# remove todas as variáveis da memória
rm(list = ls(all = TRUE))

pessoas<-c(1,2,3,4,5)
videos<-c(5,4,4,6,8)
notas<-c(9,8,10,13,15)

# médias
mv<-mean(videos)
mn<-mean(notas)

# limites do eixo y
ymin=min(c(notas,videos))
ymax=max(c(notas,videos))

# plota o box
windows(5,6)
plot.new()
plot.window(xlim=c(0,6), ylim=c(ymin,ymax))
title("Desvios em relação à média", ylab="Vídeos assistidos / Notas",
         xlab="Pessoa")
axis(1,1:5,1:5)
axis(2,0:ymax,0:ymax)
box()

# média dos vídeos
lines(c(0,6),c(mv,mv),col="blue")
points(pessoas,videos,col="blue")
for(i in 1:length(pessoas)){
  lines(c(i,i),c(mv,videos[i]),col="blue")
}

# média das notas
lines(c(0,6),c(mn,mn),col="red")
points(pessoas,notas,col="red")
for(i in 1:length(pessoas)){
  lines(c(i,i),c(mn,notas[i]),col="red")
}


#############################################################
# Figura 6.2 - Histograma da altura dos filhos e dos pais.
#############################################################
library(UsingR); data(galton)
par(mfrow=c(1,2))
hist(galton$parent, col="blue", breaks=100, xlab="Pais",
     ylab="Frequência", main="Histograma")
hist(galton$child, col="blue", breaks=100, xlab="Filhos",
     ylab="Frequência", main="Histograma")
```

```
#############################################################
# Figura 6.3 - Gráfico de dispersão das variáveis X e Y.
#############################################################
library(UsingR); data(galton)
par(mfrow=c(1,2))
par(cex.axis=0.8, cex.main=0.8)
plot(galton$parent, galton$child, pch=19, cex=0.15, col="blue",
     xlab="Pais", ylab="Filhos", main="(a)")
x <- galton$parent
y <- galton$child
freqData <- as.data.frame(table(x, y))
names(freqData) <- c("parent","child","freq")
plot(as.numeric(as.vector(freqData$parent)),
     as.numeric(as.vector(freqData$child)),
     pch = 21, col = "black", bg = "lightblue",
     cex = .08 * freqData$freq, xlab = "Pais", ylab = "Filhos",
          main="(b)")
points(mean(galton$parent),mean(galton$child),col="red",pch=19)


#############################################################
# Figura 6.4 - Distâncias verticais
#############################################################
library(UsingR); data(galton)
v <- c(60,100,500,800,900)
plot(galton$parent, galton$child, pch=19, cex=0.7, col="blue",
     xlab="Pais", ylab="Filhos", main="Distâncias verticais", type="n")
points(galton$parent[v], galton$child[v])
galton.lm <- lm(child ~ parent, data=galton)
abline(coef(galton.lm))
for(i in 1:length(v)){
    lines(c(galton$parent[v[i]],galton$parent[v[i]]),
          c(galton$child[v[i]],galton.lm$fitted.values[v[i]]),
          col="red", lwd=2)
}


#############################################################
# Figura 6.5 - Centro de massa deslocado para a origem
#############################################################
library(UsingR); data(galton)
par(cex.axis=0.8, cex.main=1.1)
x <- galton$parent - mean(galton$parent)
y <- galton$child - mean(galton$child)
freqData <- as.data.frame(table(x, y))
names(freqData) <- c("parent","child","freq")
plot(as.numeric(as.vector(freqData$parent)),
     as.numeric(as.vector(freqData$child)),
     main = "Centralização",
     pch = 21, col = "black", bg = "lightblue",
     cex = .08 * freqData$freq, xlab = "Pais", ylab = "Filhos")
points(mean(x),mean(x),col="red",pch=19)
```

```
#########################################################################
# Figura 6.6 - Identificação do intercepto y dada a inclinação β da reta.
#########################################################################
library(UsingR); data(galton)
m <- lm(child ~ parent, data = galton)
y <- function(x) { coef(m)[1] + coef(m)[2]*x }

plot.new()
plot.window(xlim=c(-2,80), ylim=c(-2,80))
title("Centro de massa no encontro das médias das alturas",xlab="Pais",
      ylab="Filhos")
axis(1,-2:80,-2:80)
axis(2,-2:80,-2:80)
points(galton$child, galton$parent, cex=0.5)
points(mean(galton$child), mean(galton$parent), pch=19, col="red")
lines(c(mean(galton$parent),mean(galton$parent)), c(-9,80), col="red")
lines(c(-9,80), c(mean(galton$child),mean(galton$child)), col="red")
lines(c(0,0),c(-9,80))
lines(c(-9,80),c(0,0))
abline(m)
# text(3,45,substitute(y==Psi*z-Sigma(beta^gamma)), pos=4)
text(35,42,substitute(beta==Delta*y/Delta*x), pos=4)
text(35,35,substitute(beta==0.6463), pos=4)

# linha azul da variação delta x
dx <- mean(mean(galton$parent))
lines(c(0,25),c(coef(m)[1],coef(m)[1]), col="blue")
lines(c(40,dx),c(coef(m)[1],coef(m)[1]), col="blue")
text(25,coef(m)[1],substitute(Delta*x==68.31), pos=4)

# linha azul da variação delta y
dy <- dx * coef(m)[2]
lines(c(72,76),c(coef(m)[1],coef(m)[1]), col="blue")
lines(c(74,74),c(coef(m)[1],44), col="blue")
lines(c(74,74),c(48,mean(galton$child)), col="blue")
text(68,46,substitute(Delta*y==44.15), pos=4)

# linha azul do intercepto y
lines(c(74,74),c(0,8), col="blue")
lines(c(74,74),c(14,coef(m)[1]), col="blue")
text(69,11,round(mean(galton$child)-dy,2), pos=4)




#############################################################
# Figura 6.7 - Modelo estatístico da regressão linear simples:
#############################################################
x <- seq(0,5,0.1)
rl <- function(x) { 1.5 + 0.8*x }
y <- sapply(x,rl)

plot.new()
plot.window(xlim=c(0,5), ylim=c(0,6))
axis(1, 0:5, 0:5)
axis(2, 0:6, 0:6)
box()
title("Regressão linear simples", xlab="x", ylab="y")
lines(x, y, lwd=2)
lines(c(-2,7), c(0,0))
lines(c(0,0), c(-2,7))
```

```
lines(c(1,1), c(-2,7), lty="dashed")
lines(c(2.5,2.5), c(-2,7), lty="dashed")
lines(c(4,4), c(-2,7), lty="dashed")

for(i in c(1,2.5,4)){
  xi <- i
  mu <- rl(xi)
  x <- seq(mu-2,mu+2,0.1)
  lines(dnorm(x,mu,0.5)+xi,x)
  lines(c(xi,max(dnorm(x,mu,0.5)+xi)), c(mu,mu), lty="dashed")
}


#########################################################################
# Figura 6.8 - Regressão linear: preço do diamante em função da massa
#########################################################################
library(UsingR); data(diamond)
plot(diamond$carat, diamond$price,
     xlab = "Massa (quilates)", ylab = "Preço (SIN $)",
     main = "Preço do diamante em função da massa",
     bg = "lightblue",
     col = "black", cex = 1.1, pch = 21, frame = FALSE)
fit <- lm(price ~ carat, data = diamond)
abline(fit, lwd = 2)



#########################################################################
# Figura 6.9 - Resíduos: (a) gráfico de dispersão e (b) resíduos vs. x
#########################################################################
library(UsingR); data(diamond)
x <- diamond$carat
y <- diamond$price
par(mfrow=c(1,2))
par(cex.axis=0.8, cex.main=0.9)
## regressão linear price ~ carat com indicação dos resíduos
plot(x, y, main="\n\n(a)",
     xlab = "Massa (quilates)", ylab = "Preço (SIN $)",
     bg = "lightblue",
     col = "black", cex = 1.1, pch = 21, frame = FALSE)
fit <- lm(y ~ x)
abline(fit, lwd=1.5)
# linhas dos resíduos
for(i in 1:length(x)){
    lines(c(x[i],x[i]),
          c(y[i],fit$fitted.values[i]),
          col="red")
}
## gráfico dos resíduos versus a massa (x)
plot(x,fit$residuals, main="\n\n(b)",
     xlab = "Massa (quilates)", ylab = "Resíduo")
lines(c(0,1),c(0,0))
# linhas dos resíduos
for(i in 1:length(x)){
    lines(c(x[i],x[i]),
          c(0,fit$residuals[i]),
          col="red")
}
par(mfrow=c(1,1), cex.main=1.2)
title("Resíduos: linhas verticais")
```

```
#########################################################################
# Figura 6.10 - Resíduos: (a) gráfico de dispersão e (b) resíduos vs. x
#########################################################################
x <- runif(100, -3, 3)
y <- x + sin(x) + rnorm(100, sd = .2)
par(mfrow=c(1,2))
par(cex.axis=0.8, cex.main=0.9)
## gráfico de dispersão
plot(x, y, main="(a)")
fit <- lm(y ~ x)
abline(fit)
## gráfico do resíduo em função de x
plot(x, fit$residuals, main="(b)", ylab = "Resíduo")
lines(c(-4,4),c(0,0))
# variação senoidal
y <- x + sin(x)
fit <- lm(y ~ x)
lines(x[order(x)],fit$residuals[order(x)], col="red", lwd=2)


#########################################################################
# Figura 6.11 - Heterocedasticidade verificada no gráfico do resíduo vs x
#########################################################################
x <- runif(100, 0, 6)
y <- x + rnorm(100, mean = 0, sd = .01 * x)
par(mfrow=c(1,2))
par(cex.axis=0.8, cex.main=0.9)
## gráfico de dispersão
plot(x, y, main="(a)")
fit <- lm(y ~ x)
abline(fit)
## gráfico do resíduo em função de x
plot(x, fit$residuals, main="(b)", ylab = "Resíduo")
lines(c(-1,7),c(0,0))
lines(c(0,x[order(fit$residuals)[1]]),c(0,min(fit$residuals)), col="red")
lines(c(0,x[order(fit$residuals)[length(x)]]),c(0,max(fit$residuals)),
         col="red")



#########################################################################
# Figura 6.12 - Regressões lineares com parâmetros quase idênticos
#########################################################################
data(anscombe)
par(mfrow = c(2, 2), mar = 0.1+c(4,4,2,1), oma =  c(0, 0, 2, 0))
t <- c("(a)","(b)","(c)","(d)")
## y ~ x
for(i in 1:4){
    x <- anscombe[,i]
    y <- anscombe[,i+4]
    plot(x, y, col = "red", pch = 21, bg = "orange", cex = 1.2,
         xlim = c(3, 19), ylim = c(3, 13), main=t[i])
    abline(lm(y ~ x), col = "blue")
    points(mean(x), mean(y), pch = 15, col="red")
}
```

```
#########################################################################
# Figura 6.15 - Regressões lineares com parâmetros quase idênticos
#########################################################################
library(UsingR); data(diamond)
x <- diamond$carat; y <- diamond$price; n <- nrow(diamond)

plot(x, y , frame=FALSE, xlab="Quilate", ylab="Dólares", pch=21,
     col="black", bg="lightblue", cex=2)
fit <- lm(y ~ x)
abline(fit, lwd=2)
beta0 <- fit$coefficients[1]
beta1 <- fit$coefficients[2]
xVals <- seq(min(x), max(x), by=.01)
yVals <- beta0 + beta1 * xVals
sigma <- sqrt(sum(fit$residuals^2) / (n-2))
ssx <- sum((x - mean(x))^2)
se1 <- sigma * sqrt(1 / n + (xVals - mean(x))^2 / ssx)
se2 <- sigma * sqrt(1 + 1 / n + (xVals - mean(x))^2 / ssx)
lines(xVals, yVals + qt(.975,n-2) * se1, col="red")
lines(xVals, yVals - qt(.975,n-2) * se1, col="red")
lines(xVals, yVals + qt(.975,n-2) * se2, col="blue")
lines(xVals, yVals - qt(.975,n-2) * se2, col="blue")
```

REGRESSÃO LINEAR MÚLTIPLA

```
#####################################################
# Figura 7.1 - Efeito de grupo: simulação 1.
#####################################################
n <- 100; t <- rep(c(0, 1), c(n/2, n/2));
x <- c(runif(n/2), runif(n/2));
beta0 <- 0; beta1 <- 2; tau <- 1; sigma <- .2

# modelo
y <- beta0 + x * beta1 + t * tau + rnorm(n, sd = sigma)
plot(x, y, type = "n", frame = FALSE,
     main="Simulação 1")

abline(lm(y ~ x), lwd = 2)

abline(h = mean(y[1 : (n/2)]), lwd = 3, col = "lightblue")
abline(h = mean(y[(n/2 + 1) : n]), lwd = 3, col = "salmon")

fit <- lm(y ~ x + t)
abline(coef(fit)[1], coef(fit)[2], lwd = 3, col = "lightblue")
abline(coef(fit)[1]  +  coef(fit)[3],  coef(fit)[2],  lwd  =  3,  col  =
         "salmon")

points(x[1  :  (n/2)],  y[1  :  (n/2)],  pch  =  21,  col  =  "black",  bg  =
         "lightblue", cex = 2)
points(x[(n/2 + 1) : n], y[(n/2 + 1) : n], pch = 21, col = "black", bg =
         "salmon", cex = 2)


#####################################################
# Figura 7.2 - Efeito de grupo: simulação 2.
#####################################################
n <- 100;
t <- rep(c(0, 1), c(n/2, n/2));
x <- c(runif(n/2), 1.5 + runif(n/2));
beta0 <- 0; beta1 <- 2; tau <- 0; sigma <- .2

y <- beta0 + x * beta1 + t * tau + rnorm(n, sd = sigma)
plot(x, y, type = "n", frame = FALSE,
     main="Silmulação 2")

abline(lm(y ~ x), lwd = 2)

abline(h = mean(y[1 : (n/2)]), lwd = 3, col = "lightblue")
abline(h = mean(y[(n/2 + 1) : n]), lwd = 3, col = "salmon")

fit <- lm(y ~ x + t)
abline(coef(fit)[1], coef(fit)[2], lwd = 3, col = "lightblue")
abline(coef(fit)[1]  +  coef(fit)[3],  coef(fit)[2],  lwd  =  3,  col  =
         "salmon")

points(x[1  :  (n/2)],  y[1  :  (n/2)],  pch  =  21,  col  =  "black",  bg  =
         "lightblue", cex = 2)
points(x[(n/2 + 1) : n], y[(n/2 + 1) : n], pch = 21, col = "black", bg =
         "salmon", cex = 2)
```

```
#####################################################
# Figura 7.3 - Efeito de grupo: simulação 3.
#####################################################
n <- 100;
t <- rep(c(0, 1), c(n/2, n/2));
x <- c(runif(n/2), .9 + runif(n/2));
beta0 <- 0; beta1 <- 2; tau <- -1; sigma <- .2

y <- beta0 + x * beta1 + t * tau + rnorm(n, sd = sigma)
plot(x, y, type = "n", frame = FALSE,
     main="Simulação 3")

abline(lm(y ~ x), lwd = 2)

abline(h = mean(y[1 : (n/2)]), lwd = 3, col="lightblue")
abline(h = mean(y[(n/2 + 1) : n]), lwd = 3, col="salmon")

fit <- lm(y ~ x + t)
abline(coef(fit)[1], coef(fit)[2], lwd = 3, col="lightblue")
abline(coef(fit)[1] + coef(fit)[3], coef(fit)[2], lwd = 3, col="salmon")

points(x[1 : (n/2)], y[1 : (n/2)], pch = 21, col = "black", bg =
         "lightblue", cex = 2)
points(x[(n/2 + 1) : n], y[(n/2 + 1) : n], pch = 21, col = "black", bg =
         "salmon", cex = 2)


#####################################################
# Figura 7.4 - Efeito de grupo: simulação 4.
#####################################################
n <- 100;
t <- rep(c(0, 1), c(n/2, n/2));
x <- c(.5 + runif(n/2), runif(n/2));
beta0 <- 0; beta1 <- 2; tau <- 1; sigma <- .2

y <- beta0 + x * beta1 + t * tau + rnorm(n, sd = sigma)
plot(x, y, type = "n", frame = FALSE,
     main="Simulação 4")

abline(lm(y ~ x), lwd = 2)

abline(h = mean(y[1 : (n/2)]), lwd = 3, col = "lightblue")
abline(h = mean(y[(n/2 + 1) : n]), lwd = 3, col = "salmon")

fit <- lm(y ~ x + t)
abline(coef(fit)[1], coef(fit)[2], lwd = 3, col = "lightblue")
abline(coef(fit)[1] + coef(fit)[3], coef(fit)[2], lwd = 3, col =
         "salmon")

points(x[1 : (n/2)], y[1 : (n/2)], pch = 21, col = "black", bg =
         "lightblue", cex = 2)
points(x[(n/2 + 1) : n], y[(n/2 + 1) : n], pch = 21, col = "black", bg =
         "salmon", cex = 2)
```

```
#####################################################
# Figura 7.5 - Regressores não relacionados.
#####################################################
n <- 100;
x2 <- runif(n)
x1 <- runif(n)
beta0 <- 0; beta1 <- 1; tau <- 4 ; sigma <- .01

y <- beta0 + x1 * beta1 + tau * x2 + rnorm(n, sd = sigma)
plot(x1, y, type = "n", frame = FALSE,
     main="Regressores Independentes")

abline(lm(y ~ x1), lwd = 2)

co.pal <- heat.colors(n)
points(x1, y, pch = 21, col = "black", bg = co.pal[round((n - 1) * x2 +
         1)], cex = 2)


#####################################################
# Figura 7.6 - Representação 3D - dois regressores.
#####################################################
# primeiramente, executar código da Figura 7.5
library(rgl)
plot3d(x1, x2, y)



#####################################################
# Figura 7.7 - Relação entre os resíduos.
#####################################################
# primeiramente, executar código da Figura 7.5
plot(resid(lm(x1 ~ x2)), resid(lm(y ~ x2)), frame = FALSE,
     col = "black", bg = "lightblue", pch = 21, cex = 2,
     main="Relação entre resíduos")
abline(lm(I(resid(lm(x1 ~ x2))) ~ I(resid(lm(y ~ x2)))), lwd = 2)



#####################################################################
# Figura 7.8 - Resultado da função plot para a regressão linear.
#####################################################################
data(swiss); par(mfrow = c(2, 2))
fit <- lm(Fertility ~ . , data = swiss)
plot(fit)



#####################################################################
# Figura 7.9 - Outliers.
#####################################################################
n <- 100; x <- rnorm(n); y <- x + rnorm(n, sd = .3)
plot(c(-3, 6), c(-3, 6), type = "n", frame = FALSE,
     xlab = "X", ylab = "Y", main = "Outliers")
abline(lm(y ~ x), lwd = 2)
points(x, y, cex = 2, bg = "lightblue", col = "black", pch = 21)
points(0, 0, cex = 2, bg = "darkorange", col = "black", pch = 21)
points(0, 5, cex = 2, bg = "darkorange", col = "black", pch = 21)
points(5, 5, cex = 2, bg = "darkorange", col = "black", pch = 21)
points(5, 0, cex = 2, bg = "darkorange", col = "black", pch = 21)
```

```
#####################################################################
# Figura 7.10 - Diagnóstico: caso 1
#####################################################################
n <- 100; x <- c(10, rnorm(n)); y <- c(10, c(rnorm(n)))
plot(x, y, frame = FALSE, cex = 2, pch = 21,
     main="Caso 1", bg = "lightblue", col = "black")
points(10, 10, cex=2, pch=21 , bg="orange")
abline(lm(y ~ x))



#####################################################################
# Figura 7.11 - Diagnóstico: caso 2
#####################################################################
x <- rnorm(n); y <- x + rnorm(n, sd = .3)
x <- c(5, x); y <- c(5, y)
plot(x, y, frame = FALSE, cex = 2, pch = 21,
     main="Caso 2", bg = "lightblue", col = "black")
points(5, 5, cex=2, pch=21 , bg="orange")
fit2 <- lm(y ~ x)
abline(fit2)


#####################################################################
# Figura 7.12 - Orly
#####################################################################
dat<-
          read.table('http://www4.stat.ncsu.edu/~stefanski/NSF_Supported/
          Hidden_Images/orly_owl_files/orly_owl_Lin_4p_5_flat.txt',
          header = FALSE)
fit <- lm(V1 ~ . - 1, data = dat);
plot(predict(fit), resid(fit), pch = '.')


#####################################################################
# Figura 7.13 - Simulação de R^2
#####################################################################
n <- 100
plot(c(1, n), 0 : 1, type = "n", frame = FALSE,
     xlab = "Qtd. preditores",
     ylab = "R^2",
     main = expression(paste("Variação de R"^2)))
r <- sapply(1 : n, function(p) {
                       y <- rnorm(n); x <- matrix(rnorm(n * p), n, p)
                       summary(lm(y ~ x))$r.squared
                   }
)
lines(1 : n, r, lwd = 2)
abline(h = 1)
```

```
#################################################
# Figura 8.1 - Classificação por regressão linear
#################################################
y <- c(0,0,0,0,1,1,1,1)
x <- c(10,20,30,40,60,70,80,95)

plot(x, y, type="n",
     xlab="Arrecadação (milhares de reais)", ylab="Status",
     main = "Classificação",
     xlim=c(-10,max(x)), ylim=c(-0.2,1.2))
lines(c(0,0),c(-0.3,1.3), lty=2)
lines(c(-20,max(x)+10),c(1,1), lty=2)
lines(c(-20,max(x)+10),c(0,0), lty=2)

text(-5,0,"Não Eleito", pos = 3, col="red")
text(-7,1,"Eleito", pos = 3, col="blue")
points(x[1:4],y[1:4], pch="x", col="red")
points(x[5:length(x)],y[5:length(x)], pch="o", col="blue" )
fit <- lm(y ~x)
abline(fit, lwd=2 )

ylim <- 0.5
xlim <- (ylim - coef(fit)[1])/coef(fit)[2]

lines(c(xlim,xlim),c(0,ylim), lty=2, lwd=2)
lines(c(0,xlim),c(ylim,ylim), lty=2, lwd=2)
text(0,0.5,"0.5", pos = 2)
text(xlim,-0.14,round(xlim,2), pos = 3)


#########################################################
# Figura 8.2 - Classificação por regressão linear: caso 2
#########################################################
y <- c(0,0,0,0,1,1,1,1,1,1)
x <- c(10,20,30,40,60,70,80,95,185,190)

plot(x, y, type="n",
     xlab="Arrecadação (milhares de reais)", ylab="Status",
     main = "Classificação",
     xlim=c(-10,max(x)), ylim=c(-0.2,1.2))
lines(c(0,0),c(-0.3,1.3), lty=2)
lines(c(-20,max(x)+10),c(1,1), lty=2)
lines(c(-20,max(x)+10),c(0,0), lty=2)

text(-1,0,"Não Eleito", pos = 3, col="red")
text(-7,1,"Eleito", pos = 3, col="blue")
points(x[1:4],y[1:4], pch="x", col="red")
points(x[5:length(x)],y[5:length(x)], pch="o", col="blue" )
fit <- lm(y ~x)
abline(fit, lwd=2 )

ylim <- 0.5
xlim <- (ylim - coef(fit)[1])/coef(fit)[2]

lines(c(xlim,xlim),c(0,ylim), lty=2, lwd=2)
lines(c(0,xlim),c(ylim,ylim), lty=2, lwd=2)
lines(c(60,60),c(0,1), lty=3, lwd=2, col="red")
text(0,0.5,"0.5", pos = 2)
text(xlim,-0.14,round(xlim,2), pos = 3)
#########################################################
# Figura 8.3 - Sigmoide
```

```
#########################################################
x <- seq(-6,6,0.1)
sigmoide <- function(z){
  1/(1+exp(-z))
}
plot(c(-7,-6,-3,0,3,6), c(0,0.2,0.4,0.6,0.8,1.0),
     type= "n", main="Sigmoide", ylab="g(x)",
     xlab=expression(paste(mu,"(x)=x")))

lines(x,sigmoide(x))

lines(c(0,0),c(0,sigmoide(0)), lty=3)
lines(c(-15,0),c(sigmoide(0),sigmoide(0)), lty=3)

lines(c(-6,6), c(0,0), lty=2)
lines(c(-6,6), c(1,1), lty=2)

text(-7.5, 0, "Y = 0", pos = 4)
text(-7.5, 1, "Y = 1", pos = 4)



#########################################################
# Figura 8.4 - Classificação: dados de treinamento
#########################################################
x2 <- c(0.4, 1.7, 2.3, 1.0, 0.5, 1.3, 0.8, 0.2,
        3.2, 3.6, 2.8, 2.3, 3.7, 1.8, 3.0, 1.2, 2.2, 2.9)
x1 <- c(0.4, 0.4, 0.3, 0.7, 1.0, 1.2, 1.6, 2.2,
        1.1, 1.7, 1.8, 2.0, 2.3, 2.6, 2.7, 2.9, 2.9, 3.3)

plot(c(0:4), c(0:4), type= "n",
     main="(a) Dados de treinamento",
     xlab="x1", ylab="x2")
points(x1[1:8],x2[1:8], pch=1, cex=1.5, col = "blue")
points(x1[9:18],x2[9:18], pch=4, cex=1.5, col = "red")
lines(c(-1,5), c(0,0), lty=2)
lines(c(0,0), c(-1,5), lty=2)

text(1, 0, "y = 0", pos = 3)
text(3.5, 0.5, "y = 1", pos = 1)


#########################################################
# Figura 8.5 - Fronteira de decisão
#########################################################
x2 <- c(0.4, 1.7, 2.3, 1.0, 0.5, 1.3, 0.8, 0.2,
        3.2, 3.6, 2.8, 2.3, 3.7, 1.8, 3.0, 1.2, 2.2, 2.9)
x1 <- c(0.4, 0.4, 0.3, 0.7, 1.0, 1.2, 1.6, 2.2,
        1.1, 1.7, 1.8, 2.0, 2.3, 2.6, 2.7, 2.9, 2.9, 3.3)

plot(c(0:4), c(0:4), type= "n",
     main="Fronteira de decisão",
     xlab="x1", ylab="x2")
points(x1[1:8],x2[1:8], pch=1, cex=1.5, col = "blue")
points(x1[9:18],x2[9:18], pch=4, cex=1.5, col = "red")
lines(c(-1,5), c(0,0), lty=2)
lines(c(0,0), c(-1,5), lty=2)

x1 <- seq(0,3,0.1)
x2 <- 3 - x1
lines(x1, x2, col="green", lwd=2)
```

```
text(1, 0, "Y = 0", pos = 3)
text(3.5, 0.5, "Y = 1", pos = 1)



############################################################
# Figura 9.1 - Análise de Variâncias
############################################################
gl1 <- 3
gl2 <- 2
curve(df(x,gl1,gl2),0,5,col="blue",lwd=2,xlab="F")



############################################################
# Tabela 9.1 e Tabela 9.2 - Análise de Variâncias
############################################################
g1 <- c(2,3,1)
g2 <- c(6,7,5)
k <- 2 # quantidade de fatores ou grupos

n1 <- length(g1)
n2 <- length(g2)
nT <- n1 + n2 # quantidade de observações

m1 <- mean(g1); v1 <- sd(g1)^2
m2 <- mean(g2); v2 <- sd(g2)^2

g <- c(g1,g2)  # grupão
mg <- mean(g)  # média grupão

SST <- sum((g - mg)^2)

SSM <- sum(((g1 - m1) - (g1 - mg))^2) + sum(((g2 - m2) - (g2 - mg))^2)
# ou
SSM <- n1*(m1-mg)^2 + n2*(m2-mg)^2

SSR <- sum((g1-m1)^2) + sum((g2-m2)^2)
#ou
SSR = SST - SSM

glM <- k - 1 # graus de liberdade do modelo
glR <- nT - k # graus de liberdade do resíduo

MSM <- SSM / glM # média quadrática do modelo
MSR <- SSR / glR # média quadrática do resíduo

Fs <- MSM / MSR # estatística F

p.value <- 1 - pf(Fs, glM, glR)

# teste t
t.test(g1, g2, paired=FALSE, var.equal = FALSE)

# Tabela 9.2
# usando aov
dados <- data.frame(grupo=c(1,1,1,2,2,2), obs=c(g1,g2))
boxplot(obs ~ grupo, data=dados)
res <- aov(obs ~ grupo, data=dados)
summary(res)
```

```
#########################################################
# Figura 9.3 - Variança explicada pelo modelo
#########################################################
g1 <- c(2,3,1)
g2 <- c(6,7,5)

local<-factor(rep(letters[1:2],each=3))
medidas<- c(g1,g2)

boxplot(medidas~local,frame.plot=F,ylim=c(0,10),ylab="Medidas",xlab="Loca
l")

plot(medidas,pch=as.character(local),col=as.numeric(local)+1,frame.plot=F
, main="ANOVA", xlab="amostras")
abline(h=mean(medidas),lwd=2)
text(3.5,mean(medidas)+0.2,"Média Global")

medias<-aggregate(medidas,list(local),mean)

abline(h=medias[1,2],lwd=2,col=2)
text(3.5,medias[1,2]+0.2,"Média a",col=2)
abline(h=medias[2,2],lwd=2,col=3)
text(3.5,medias[2,2]+0.2,"Média b",col=3)

abline(h=mean(medidas),lwd=2)
text(25,mean(medidas)-0.2,"Média Geral")
segments(1:6, mean(medidas),1:6,medidas,lwd=3)

segments((1:6)+0.05, rep(medias[,2],each=3),(1:6)+0.05, medidas,lwd=3,
         col=as.numeric(local)+1)

segments((1:6)+0.09, rep(c(mean(g1),mean(g2)),each=3),(1:6)+0.09,
         rep(mean(medidas),each=6),lwd=3, col="blue")



#########################################################
# Figura 10.2 - Distribuição X² para distintos valores de k
#########################################################
n <- 100
cor <- c("black","red","blue","forestgreen")
k <- 2:5

for(i in k) {
  X2 <- sort(rchisq(n, df=i-1))
  if(i == 2){
      plot(X2, dchisq(X2, df=i-1), lwd=2, type="l", ylim=c(0, 1),
           ylab="Densidade", xlab=expression(paste("X"^2)),
           main=expression(paste("Distribuição X"^2)))
      grid()
  } else{
      lines(X2, dchisq(X2, df=i-1), col=cor[i-1], lwd=2)
  }
}

legend("topright", legend=c("k = 2", "k = 3", "k = 4", "k = 5"),
       col=cor, lty=1, lwd=2, bty="n")
```