
INTRODUÇÃO À ESTATÍSTICA

COM



Eric Batista Ferreira
Marcelo Silva de Oliveira

Eric Batista Ferreira
Marcelo Silva de Oliveira

Introdução à **Estatística com R**

Universidade Federal de Alfenas
Editora Unifal-MG
Alfenas - MG
2020

© 2020 Direito de reprodução de acordo com a Lei nº 9.610, de 19 de fevereiro de 1998.

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

Introdução à Estatística com R.

ISBN 978-65-86489-23-1

Dados Internacionais de Catalogação-na-Publicação (CIP)
Sistema de Bibliotecas da Universidade Federal de Alfenas
Biblioteca Central - Campus Sede

Ferreira, Eric Batista

F383i Introdução à Estatística com R. / Eric Batista Ferreira, Marcelo
Silva de Oliveira -- Alfenas -- MG : Editora Universidade Federal de
Alfenas, 2020.

194 f.: il. -

ISBN: 978-65-86489-23-1.

Inclui Bibliografia.

1. Estatística. 2. Probabilidade. 3. Inferência. 4. Software R. I. Ferreira,
Eric Batista. II. Oliveira, Marcelo Silva de.

CDD-519.2

Ficha Catalográfica elaborada por Marlom César da Silva

Bibliotecário-Documentalista CRB6/2735



Universidade Federal de Alfenas - UNIFAL-MG

Endereço: Rua Gabriel Monteiro da Silva, 700 Centro, Alfenas,

Minas Gerais, Brasil, CEP: 37.130-001

Reitor

Sandro Amadeu Cerveira

Vice-Reitor

Alessandro Antônio Costa Pereira

Sistema de Bibliotecas da UNIFAL-MG/SIBI/UNIFAL-MG

Diagramação: Eric Batista Ferreira

Capa: Germano Lopes e Eric Batista Ferreira

Revisão Textual: Renata Chrystina Bianchi de Barros

Eric Batista Ferreira

Departamento de Estatística

Universidade Federal de Alfenas

Alfenas, Minas Gerais, Brasil

E-mail: <eric.ferreira@unifal-mg.edu.br>

Marcelo Silva de Oliveira

Departamento de Estatística

Universidade Federal de Lavras

Lavras, Minas Gerais, Brasil

E-mail: <marcelo.oliveira@ufla.br>



SUMÁRIO

PREFÁCIO	xvii
APRESENTAÇÃO	xix
1 INTRODUÇÃO	19
1.1 Introdução ao R	19
1.1.1 <i>Baixando e instalando</i>	19
1.1.2 <i>Iniciando o R</i>	21
1.1.3 <i>Lendo arquivos</i>	22
1.1.4 <i>Funções básicas</i>	23
1.1.5 <i>Pedindo ajuda</i>	23
1.1.6 <i>Objetos</i>	24
1.2 Alguns conceitos importantes	27
1.2.1 <i>Classificações de uma população</i>	28
1.2.2 <i>Tipos de variáveis</i>	28
1.2.2.1 <i>Qualitativas</i>	28
1.2.2.2 <i>Qualitativas nominais</i>	28
1.2.2.3 <i>Qualitativas ordinais</i>	28
1.2.2.4 <i>Quantitativas</i>	28
1.2.2.5 <i>Quantitativas discretas</i>	29
1.2.2.6 <i>Quantitativas contínuas</i>	29
1.3 Principais aplicações da Estatística	29
1.3.1 <i>Pesquisa científica</i>	29
1.3.2 <i>Processos produtivos</i>	29
1.3.3 <i>Levantamentos em geral</i>	29
2 TÉCNICAS DE SOMATÓRIO	31
2.1 Introdução	31

2.2	Propriedades dos somatórios	32
2.3	Somatórios no R	33
3	ESTATÍSTICA DESCRITIVA	35
3.1	Introdução	35
3.2	Variáveis qualitativas	36
3.2.1	<i>Representações gráficas das distribuições de frequências</i>	37
3.2.1.1	Gráfico de barras e de colunas	37
3.2.1.2	Gráfico de pizza ou Setograma	37
3.3	Variáveis quantitativas discretas	38
3.3.1	<i>Representação gráfica da distribuição de frequências</i>	39
3.3.1.1	Gráfico de agulhas	39
3.4	Variáveis quantitativas contínuas	40
3.4.1	<i>Representação gráfica da distribuição de frequências</i>	42
3.4.1.1	Histograma	42
3.5	Medidas de posição	45
3.5.1	<i>Média (Me)</i>	45
3.5.1.1	Propriedades da Média	45
3.5.2	<i>Mediana (Md)</i>	47
3.5.2.1	Propriedade da Mediana	47
3.5.3	<i>Moda (Mo)</i>	48
3.5.3.1	Método de Czuber	49
3.5.3.2	Propriedades da Moda	49
3.5.4	<i>Influência da assimetria nas medidas de posição</i>	50
3.6	Medidas de dispersão	52
3.6.1	<i>Amplitude (A)</i>	52
3.6.1.1	Propriedades da Amplitude	53
3.6.2	<i>Variância</i>	53
3.6.2.1	Propriedades da Variância	54
3.6.3	<i>Desvio-padrão</i>	55
3.6.3.1	Propriedades do Desvio-Padrão	56
3.6.4	<i>Coefficiente de Variação (CV)</i>	56
3.6.4.1	Propriedades do Coeficiente de Variação	57
4	PROBABILIDADE	59
4.1	Algumas definições úteis	59
4.2	Axiomas da Probabilidade	60
4.2.1	<i>Propriedades derivadas dos axiomas</i>	60
4.3	Definição clássica de probabilidade	61
4.4	Definição frequentista de probabilidade	61

4.5	Regra do “E” e regra do “OU”	62
4.5.1	Regra do “E”	63
4.5.2	Regra do “OU”	63
4.6	Probabilidade condicional	63
4.7	Distribuições de probabilidade discretas	64
4.7.1	Definições úteis	64
4.7.2	Distribuição Binomial	64
4.7.3	Distribuição Poisson	66
4.8	Distribuições de probabilidades contínuas	67
4.9	Função Densidade de Probabilidade (fdp)	69
4.9.1	Propriedades da fdp	69
4.9.2	Distribuição Normal de probabilidades	70
4.9.2.1	Propriedades da Normal	70
4.9.2.2	Distribuição Normal padronizada (padrão ou zero-um)	71
4.9.2.3	Aproximação da Binomial à Normal	74
4.9.2.4	Aproximação da Poisson à Normal	74
5	AMOSTRAGEM	77
5.1	Amostragens não-aleatórias	78
5.2	Amostragens aleatórias	78
5.2.1	Amostragem aleatória simples (AAS)	78
5.2.1.1	Sorteio	79
5.2.1.2	Inconvenientes da AAS	79
5.2.1.3	Modelo Estatístico	79
5.2.2	Amostragem aleatória sistemática (AS)	79
5.2.2.1	Populações finitas	80
5.2.2.2	Populações muito grandes ou infinitas	80
5.2.2.3	Modelo Estatístico	80
5.2.3	Amostragem aleatória estratificada (AAE)	80
5.2.3.1	Modelo Estatístico	81
5.2.4	Amostragem aleatória por conglomerado (AAC)	81
5.2.4.1	Modelo Estatístico	81
6	INFERÊNCIA ESTATÍSTICA	83
6.1	Introdução	83
6.2	Teoria da Estimação	84
6.2.1	Estimação por ponto	84
6.2.2	Propriedades desejadas dos estimadores	85
6.2.2.1	Não-tendenciosidade	85
6.2.2.2	Variância mínima	85

6.2.2.3	Estimadores não-tendenciosos de variância mínima	86
6.3	Distribuições de amostragem	86
6.3.1	<i>Distribuição da média amostral de populações normais</i>	86
6.3.1.1	A distribuição t de Student	87
6.3.1.2	Propriedade da distribuição t de Student	87
6.3.2	<i>Uso das distribuições de amostragem na Inferência Estatística</i>	88
6.3.3	<i>Estimação por intervalo</i>	89
6.3.3.1	Fator de correção para populações finitas	90
6.4	Teoria da Decisão	90
6.4.1	<i>Propriedades desejadas para os testes</i>	91
6.4.2	<i>Estrutura dos testes</i>	91
6.4.2.1	Par de hipóteses	92
6.4.2.2	Estatística de teste	93
6.4.2.3	Regra de decisão	93
6.4.2.4	Conclusão	96
6.4.3	<i>Contingências: tipos de erros e acertos possíveis</i>	97
6.5	Inferência sobre uma população normal	98
6.5.1	<i>Estimação da média populacional (μ)</i>	98
6.5.2	<i>Estimação de uma proporção populacional (p)</i>	103
6.5.3	<i>Estimação do Total populacional (T)</i>	105
6.5.3.1	Estimação de T a partir da média μ	106
6.5.3.2	Estimação de T a partir de uma proporção p	106
6.5.4	<i>Teste para a média populacional (μ)</i>	106
6.5.4.1	Par de hipóteses	106
6.5.4.2	Estatística de teste	106
6.5.4.3	Regra de decisão	107
6.5.5	<i>Teste para uma proporção populacional (p)</i>	109
6.5.5.1	Par de hipóteses	109
6.5.5.2	Estatística de teste	109
6.5.5.3	Regra de decisão	109
6.5.6	<i>Estimação da variância populacional (σ^2)</i>	111
6.5.7	<i>Estimação do desvio-padrão populacional (σ)</i>	112
6.6	Inferência sobre duas populações normais	112
6.6.1	<i>Teste de homogeneidade de variâncias</i>	113
6.6.1.1	Distribuição F	113
6.6.2	<i>Estimação da diferença entre duas medias ($\mu_1 - \mu_2$)</i>	117
6.6.2.1	Variâncias populacionais conhecidas	117
6.6.2.2	Variâncias populacionais desconhecidas	117
6.6.3	<i>Teste sobre a diferença entre duas médias ($\mu_1 - \mu_2$)</i>	118
6.6.3.1	Par de hipóteses	118

6.6.3.2	Estatística de teste	119
6.6.3.3	Regra de decisão	119
6.6.4	<i>O caso dos dados emparelhados</i>	122
6.6.4.1	Estimação do efeito médio da intervenção (μ_d)	122
6.6.4.2	Teste sobre o efeito médio da intervenção (μ_d)	124
7	REGRESSÃO E CORRELAÇÃO	127
7.1	Regressão	127
7.1.1	<i>Modelo Estatístico</i>	128
7.1.2	<i>Esperança Matemática</i>	128
7.1.3	<i>Método dos Quadrados Mínimos</i>	128
7.1.3.1	Par de hipóteses	130
7.1.3.2	Estatística de teste	130
7.1.3.3	Regra de decisão	131
7.2	Correlação	132
	REFERÊNCIAS	139
	APÊNDICE A: EXERCÍCIOS PROPOSTOS	141
	APÊNDICE B: TABELAS	167
	SOBRE OS AUTORES	193



LISTA DE TABELAS

Tabela 1	– Distribuição de frequências absolutas (fa), relativa (fr) e percentual (fp) da atividade em propriedades de uma região.	36
Tabela 2	– Distribuição de frequências absolutas (fa), relativa (fr) e percentual (fp) do número de filhos por casal de uma cidade.	39
Tabela 3	– Distribuição de frequências absoluta (fa), relativa (fr) e percentual (fp) do peso observado em potinhos de canela em pó.	42
Tabela 4	– frequências absolutas acumuladas abaixo ($Fa\downarrow$) e acima de ($Fa\uparrow$).	44
Tabela 5	– Três conjuntos de dados diferentes, que não podem ser diferenciados pelas medidas de posição	52
Tabela 6	– Diferenças básicas entre a amostragem aleatória estratificada (AAE) e a amostragem aleatória por conglomerado (AAC).	81
Tabela 7	– Representação tabular das contingências em um teste de hipóteses: erros e acertos.	97
Tabela 8	– Número diário de espectadores de dois filmes, em milhões de pessoas.	115
Tabela 9	– Volume de polpa (cm^3), volume de água (cm^3) e teor de cálcio ($mg/100ml$) em 20 cocos verdes.	135
Tabela 10	– Tabela auxiliar para cálculo dos coeficientes do modelo linear.	136
Tabela 11	– Distribuição de frequências das idades de pessoas que sofreram acidentes em casa na Inglaterra, no ano de 1977.	144
Tabela 12	– Probabilidades (α) da distribuição normal padrão $N(0,1)$ para valores do quantil Z_t padronizado de acordo com o seguinte evento: $P(0 < Z < Z_t) = \alpha$	168
Tabela 13	– Probabilidades (α) da distribuição normal padrão $N(0,1)$ para valores do quantil Z_t padronizado de acordo com o seguinte evento: $P(Z > Z_t) = \alpha$	169
Tabela 14	– Quantis superiores da distribuição de qui-quadrado (χ^2_α) com ν graus de liberdade, e para diferentes valores da probabilidade (α) de acordo com o seguinte evento: $P(\chi^2 > \chi^2_\alpha) = \alpha$	170

Tabela 15 – Quantis superiores da distribuição de qui-quadrado (χ^2_α) com ν graus de liberdade, e para diferentes valores da probabilidade (α) de acordo com o seguinte evento: $P(\chi^2 > \chi^2_\alpha) = \alpha$	171
Tabela 16 – Quantis superiores da distribuição de F ($F_{0,10}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 10% de acordo com o seguinte evento: $P(F > F_{0,10}) = 0,10$	172
Tabela 17 – Quantis superiores da distribuição de F ($F_{0,10}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 10% de acordo com o seguinte evento: $P(F > F_{0,10}) = 0,10$	173
Tabela 18 – Quantis superiores da distribuição de F ($F_{0,05}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 5% de acordo com o seguinte evento: $P(F > F_{0,05}) = 0,05$	174
Tabela 19 – Quantis superiores da distribuição de F ($F_{0,05}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 5% de acordo com o seguinte evento: $P(F > F_{0,05}) = 0,05$	175
Tabela 20 – Quantis superiores da distribuição de F ($F_{0,025}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 2,5% de acordo com o seguinte evento: $P(F > F_{0,025}) = 0,025$	176
Tabela 21 – Quantis superiores da distribuição de F ($F_{0,025}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 2,5% de acordo com o seguinte evento: $P(F > F_{0,025}) = 0,025$	177
Tabela 22 – Quantis superiores da distribuição de F ($F_{0,01}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 1% de acordo com o seguinte evento: $P(F > F_{0,01}) = 0,01$	178
Tabela 23 – Quantis superiores da distribuição de F ($F_{0,01}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 1% de acordo com o seguinte evento: $P(F > F_{0,01}) = 0,01$	179
Tabela 24 – Quantis superiores da distribuição de F ($F_{0,005}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 0,5% de acordo com o seguinte evento: $P(F > F_{0,005}) = 0,005$	180
Tabela 25 – Quantis superiores da distribuição de F ($F_{0,005}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 0,5% de acordo com o seguinte evento: $P(F > F_{0,005}) = 0,005$	181
Tabela 26 – Quantis superiores da distribuição de F ($F_{0,90}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 90% de acordo com o seguinte evento: $P(F > F_{0,90}) = 0,90$	182
Tabela 27 – Quantis superiores da distribuição de F ($F_{0,90}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 90% de acordo com o seguinte evento: $P(F > F_{0,90}) = 0,90$	183

Tabela 28 – Quantis superiores da distribuição de F ($F_{0,95}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 95% de acordo com o seguinte evento: $P(F > F_{0,95}) = 0,95$	184
Tabela 29 – Quantis superiores da distribuição de F ($F_{0,95}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 95% de acordo com o seguinte evento: $P(F > F_{0,95}) = 0,95$	185
Tabela 30 – Quantis superiores da distribuição de F ($F_{0,975}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 97,5% de acordo com o seguinte evento: $P(F > F_{0,975}) = 0,975$	186
Tabela 31 – Quantis superiores da distribuição de F ($F_{0,975}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 97,5% de acordo com o seguinte evento: $P(F > F_{0,975}) = 0,975$	187
Tabela 32 – Quantis superiores da distribuição de F ($F_{0,99}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 99% de acordo com o seguinte evento: $P(F > F_{0,99}) = 0,99$	188
Tabela 33 – Quantis superiores da distribuição de F ($F_{0,99}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 99% de acordo com o seguinte evento: $P(F > F_{0,99}) = 0,99$	189
Tabela 34 – Quantis superiores da distribuição de F ($F_{0,995}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 99,5% de acordo com o seguinte evento: $P(F > F_{0,995}) = 0,995$	190
Tabela 35 – Quantis superiores da distribuição de F ($F_{0,995}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 99,5% de acordo com o seguinte evento: $P(F > F_{0,995}) = 0,995$	191
Tabela 36 – Quantis superiores da distribuição t de Student (t_α) com ν graus de liberdade e para diferentes valores da probabilidade (α) de acordo com o seguinte evento: $P(t > t_\alpha) = \alpha$	192

LISTA DE FIGURAS

Figura 1	– Foto do CD de lançamento da primeira versão do R, em 29 de fevereiro de 2000, assinado pelos autores.	20
Figura 2	– Tela de boas vindas do R, em sua versão 4.0.2 (“ <i>Taking off Again</i> ”), em ambiente Windows, destacando o prompt de comando.	21
Figura 3	– Mensagem recebida ao se fechar o console do R, em ambiente Windows, perguntando se a área de trabalho deve ser salva.	22
Figura 4	– RStudio: um IDE (ambiente de desenvolvimento integrado) muito útil para o usuário e o programador R.	23
Figura 5	– Esquema de um objeto, estrutura que pode armazenar números, vetores, matrizes, <i>arrays</i> , listas e <i>data frames</i>	25
Figura 6	– Gráfico ilustrativo do comportamento do peso de suínos (kg) em função de sua ingestão diária de ração ao longo do período de engorda.	30
Figura 7	– (a) Gráfico de colunas das principais atividades em propriedades rurais. (b) Gráfico de barras da mesma situação.	37
Figura 8	– Setograma (gráfico de setores ou gráfico de pizza) sobre frequências de ocorrência de café, leite, milho e outras.	38
Figura 9	– (a) gráfico de linhas da variável “número de filhos por casal”. (b) gráfico de colunas da mesma variável.	40
Figura 10	– (a) Histograma do peso de potinhos de canela em pó em uma linha de produção. (b) O mesmo histograma com polígono de frequência.	43
Figura 11	– Ogivas representando as frequências absolutas acumuladas <i>acima de e abaixo de</i> e seu respectivo código em R.	44
Figura 12	– Histograma ilustrando geometricamente como se dá o cálculo da moda por meio do método de Czuber.	49
Figura 13	– Representação de uma distribuição contínua assimétrica à direita, posicionando média, moda e mediana.	51

Figura 14 – Representação de uma distribuição contínua assimétrica à esquerda, posicionando média, moda e mediana.	51
Figura 15 – Representação de uma distribuição contínua simétrica, posicionando média, moda e mediana.	51
Figura 16 – Simulação do lançamento de uma moeda honesta 500 vezes, comportamento de sua frequência relativa.	62
Figura 17 – Gráfico da distribuição de probabilidade binomial, ilustrando os casos de nascerem x machos em 9 ovos.	66
Figura 18 – Gráfico da distribuição de probabilidade Poisson, ilustrando os casos de haverem x chuvas fortes por ano.	67
Figura 19 – Esquema da generalização teórica de histogramas para funções densidade de probabilidade, quando $n \rightarrow \infty$	68
Figura 20 – Representação da integral de uma função densidade de probabilidade, como área total abaixo da curva.	69
Figura 21 – Esquema destacando a área acima de 100km/h em uma distribuição de média 60km/h e variância $400(\text{km/h})^2$	71
Figura 22 – Esquema destacando a area entre 40 e 100km/h em uma distribuição de média 60km/h e variância $400(\text{km/h})^2$	72
Figura 23 – Esquema destacando o intervalo que contem 90% dos veículos em uma distribuição de media 60km/h e variância $400(\text{km/h})^2$	73
Figura 24 – Representação esquemática da retirada de uma amostra da população, como conjuntos.	77
Figura 25 – Demonstração da curva Normal (linha cheia) padrão e da curva t com 5 (linha tracejada) e 30 (linha pontilhada) graus de liberdade.	88
Figura 26 – Representação do quantil superior de 2,5% (área hachurada) na distribuição normal padrão.	90
Figura 27 – Representação, em uma distribuição simétrica, de regiões de aceitação e rejeição de H_0 , definindo testes unilaterais à direita, à esquerda e bilateral.	95
Figura 28 – Representação pictórica da relação negativa entre as probabilidades α e β , para uma amostra de tamanho fixo n	98
Figura 29 – Esquema ilustrativo da amostragem de uma população normal, parâmetros populacionais e estatísticas amostrais.	99
Figura 30 – Esquema ilustrativo da amostragem de duas populações normais, parâmetros populacionais e estatísticas amostrais.	112
Figura 31 – Distribuição F de probabilidade ressaltando a região de aceitação de um teste de homogeneidade de variâncias (de 1 até um F_c qualquer).	113
Figura 32 – Representação de possíveis retas (pontilhadas) e reta estimada por quadrados mínimos (linha cheia) em uma massa de dados fictícia.	129

Figura 33 – Exemplos ilustrativos de quatro possíveis relacionamentos entre variáveis: não correlacionadas (A), positivamente correlacionadas (B), negativamente correlacionadas (C) e relacionamento quadrático (D).	133
Figura 34 – Diagrama de dispersão entre a variável independente “volume de água de coco”, e a variável dependente “volume de polpa de coco”.	135
Figura 35 – Reta de regressão estimada e pontos amostrais, na relação entre a variável independente “volume de água de coco”, e a variável dependente “volume de polpa de coco”.	137



PREFÁCIO

Prefácio deriva dos radicais latinos: *prae* (antes) e *fatio* (dito). Literalmente significa texto introdutório que apresenta o tema desenvolvido pelo(s) autor(es) no conteúdo da obra. Os autores escolhem o prefaciador por um de dois critérios:

1. *É uma pessoa de capacidade conhecida, cujo aval cria boas expectativas sobre o que está por vir nas páginas seguintes.* Impossível aceitar que tenha sido este o motivo que levou os autores a convidar-me para essa honrosa tarefa. Sou do passado, da época do cálculo mental, incapaz de avaliar criteriosamente um trabalho didático que usa recursos computacionais para realizar o aprendizado de Estatística;
2. *Os autores quiseram agradecer o prefaciador por razões de estima pessoal.* Este é o único motivo que sou capaz de admitir sem falsa modéstia.

Enfocando a obra que li com atenção e, muitas vezes, pausadamente para bem entender a linguagem computacional usada no processamento de dados, afirmo que o trabalho desenvolvido pelos autores é digno de respeito, consideração e aplauso. Apreciei muito a exposição didática e virtual da Estatística descritiva. Conceitos claros, objetivos, com sobeja exemplificação e boa quantidade de exercícios de fixação.

O mundo atual presencia um processo acelerado de automação, elevando muito o nível de controle e a produtividade. Estatística e computação consolidam-se como recursos fundamentais, mormente no âmbito da pesquisa. O computador compõe-se de hardware e software. Hardware é o equipamento físico e software é o conjunto de instruções e programas que fazem o aparelho funcionar, ou seja, é o sistema operacional. A quantidade de software que existe atualmente é inestimável, o mesmo ocorrendo com a diversidade de objetivos a que se ajustam. Mesmo familiarizado com computadores, tanto o engenheiro de software, como o mero usuário, todos pasmam ao lembrar que tudo começa com um bit ou dígito binário, 0 e 1, significando não corrente (0) e corrente (1), respectivamente, agrupado em conjuntos de 8 (bytes ou octetos), ou de outras potências de 2.

Dentre os vários softwares usados na Estatística, existentes no mercado (KNIME Analytics Platform, Orange, Trifacta Wrangler, R, OpenRefine, SPSS, SAS, Statistica, S, etc.), os autores optaram sen-

satamente pelo R. É um dos softwares mais difundidos atualmente, tem linguagem própria e acesso livre ao código fonte, permitindo desenvolver pacotes específicos para cada necessidade. Tem grande capacidade gráfica, é compatível com a plataforma Windows e é gratuito.

Eric Batista Ferreira, professor associado III do Departamento de Estatística da UNIFAL-MG, tem experiência em Sensometria, Estatística Multivariada, Estatística Experimental, Probabilidade e Estatística aplicadas, Controle Estatístico de Processo e Cientometria (metrificação da Ciência). Marcelo Silva de Oliveira, aposentado em 2018 como Professor Titular de Estatística da Universidade Federal de Lavras, continua sua vida científica dedicando-se, principalmente, à pesquisa e à produção de livros, além de outras atividades. Tem grande experiência na área de Probabilidade e Estatística aplicadas, e Geoestatística aplicada às Ciências Agrárias.

Esses nomes, simplesmente identificados na capa deste livro, constituem o verdadeiro prefácio dessa importante obra, pois criam a expectativa de conhecê-la toda e também a certeza de um estudo de qualidade.

Parabéns a ambos!

Prof. José Antônio Leite
Professor Aposentado de Estatística da Universidade Federal de Alfenas
Alfenas, 16 de Agosto de 2020.



APRESENTAÇÃO

Segundo a Base Nacional Comum Curricular (BNCC), as crianças entram em contato com os primeiros conceitos de Probabilidade e Estatística ainda no Ensino Fundamental. No Ensino Médio, esses conceitos são reforçados e os jovens percebem a existência de conceitos adicionais e um pouco mais profundos. Até essa fase, todos os conceitos são trabalhados na disciplina de Matemática.

Entretanto, é no curso de graduação que o jovem estudante geralmente tem seu primeiro curso inteiramente dedicado à Estatística e Probabilidade, seja sob o nome de Estatística Básica, seja Introdução à Estatística, Estatística I, ou equivalentes. Vale ressaltar que o curso introdutório de Estatística e Probabilidade é a disciplina mais comumente encontrada em cursos de graduação no Brasil. Ela figura em cursos de Exatas, mas também em Engenharias, cursos da área de Saúde e também de Humanas.

Esta obra é fruto do amadurecimento de um caderno didático, de nossa autoria, publicado pela FAEPE (Fundação de Apoio ao Ensino, Pesquisa e Extensão), utilizado em cursos de graduação e Pós-graduação Lato Sensu na Universidade Federal de Lavras (UFLA), no início dos anos 2000.

Oferecemos este livro aos estudantes e interessados em um primeiro curso universitário sobre Probabilidade e Estatística, num estilo “primer”, como base para um segundo curso mais “encorpado”, ou para aqueles que querem e ou precisam de um material que lhes introduza na Estatística de nível superior, mas não querem ou não precisam acessar um texto extenso. Procuramos dar ao leitor esta opção de introdução essencial à Probabilidade e Estatística *junto* com a linguagem de programação R, própria para Estatística, o que configura uma articulação de grande vantagem para os desafios e oportunidades presentes no mundo atual.

Agradecemos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), pelo aporte financeiro para pesquisas ao longo de nossas carreiras; e à Universidade Federal de Alfenas (Unifal-MG) e à Universidade Federal de Lavras (UFLA), pelo apoio e incentivo para o ensino, pesquisa e extensão.

Prof. Dr. Eric Batista Ferreira
Prof. Dr. Marcelo Silva de Oliveira

INTRODUÇÃO

1.1 Introdução ao R

R é uma linguagem e ambiente para computação estatística e gráfica (R Core Team, 2020). É um projeto GNU similar à linguagem e ambiente S desenvolvida no *Bell Laboratories* por John Chambers e colaboradores. Na década de 1990, o sistema operacional Linux estava em franca expansão entre os pesquisadores, mas este não possuía um software para análises estatísticas nativo. A fim de solucionar essa deficiência, Brian Ripley e Bill Venables, da universidade de Auckland, implementam a linguagem R com aparência do S e a semântica da linguagem *Scheme*, e deixaram de código aberto.

O software disponibiliza uma grande variedade de métodos estatísticos (modelagem linear e não-linear, testes estatísticos, modelos de séries temporais, classificação, métodos multivariados, etc) e técnicas gráficas. Um dos pontos fortes do R é a facilidade com que gráficos bem delineados e de alta qualidade para impressão podem ser produzidos com possibilidade de inclusão de fórmulas e símbolos matemáticos, quando necessário.

Sua primeira versão foi lançada em 29 de fevereiro de 2000 (Figura 1). Escolheram, propositalmente, lançá-lo no dia extra de um ano bissexto. Portanto, em 29 de fevereiro de 2020, o software R completou 20 anos, mais ágil do que nunca!

O software R se presta a diversas funções como, desde uma calculadora científica, passando pela integração e derivação de funções matemáticas, até a realização de complexas análises estatísticas.

Além disso, o R também apresenta uma série de recursos gráficos que permitem a descrição detalhada de todos os aspectos que se pode querer personalizar em um gráfico, como cor, tipo e tamanho de letra, símbolos, títulos e subtítulos, pontos, linhas, legendas, planos de fundo e muito mais.

1.1.1 Baixando e instalando

O R é disponibilizado sob os termos da *GNU General Public License* da *Free Software Foundation* na forma de código aberto. Ele pode ser compilado e roda em um grande número de plataformas UNIX e similares (incluindo FreeBSD e Linux). Também pode ser compilado e roda em Windows 9x/NT/2000 e MacOS.



Figura 1 – Foto do CD de lançamento da primeira versão do R, em 29 de fevereiro de 2000, assinado pelos autores.

Fonte: Tweeter. Acesso em: 09 ago. 2020. Disponível em: <https://twitter.com/_R_Foundation/status/1233671896144793600/photo/1>.

O *download* do R é gratuito de qualquer espelho do site <www.r-project.org>. Após entrar nesse site, clique em *CRAN*, logo abaixo da palavra *Download*. Em seguida, escolha um espelho perto de você, por exemplo, o espelho da Universidade Federal do Paraná: <<http://cran.br.r-project.org/>>. Agora, escolha seu sistema operacional. Por exemplo, *Download R for Windows*. Aqui você opta entre baixar o conjunto de pacotes básicos ou contribuídos. Supondo que você está baixando o R pela primeira vez, escolha a opção *base*. Nesta página o CRAN lhe apresenta uma série de opções como o *readme*, *changes*, etc. Escolha o arquivo executável, por exemplo, *R-4.0.2-win32.exe*. Pronto! É só baixar e executar.

A instalação do R é muito fácil e autodirecionada. Nas versões mais recentes é possível, inclusive, selecionar o idioma *Português (Brasil)* para as barras de menus e mensagens de erro. Porém, vale notar que, pelo menos até a versão 4.0.2, os nomes das funções, dos atributos e os *helps* continuam em Inglês.

A cada ano, três versões do R são disponibilizadas no CRAN. Pelo menos uma a cada semestre. Além disso, sempre existem duas versões disponíveis concomitantemente: uma versão alfa (revisada) e uma versão beta (não revisada, mas mais recente).

Mais que um software que realiza análises estatísticas, R é um ambiente e uma linguagem de programação orientada a objeto. Nele, números, vetores, matrizes, *arrays* e listas podem ficar armazenados em objetos. Pode-se entender objeto como uma caixinha onde você pode guardar o que quiser. A partir daí todas as operações matemáticas podem ser feitas usando esses objetos. Isso torna as coisas mais simples. A Figura 2 mostra a tela inicial do R.

Em ambiente windows, R apresenta diversos botões de atalho para a manipulação de arquivos, pacotes, ajudas, etc. O que se vê na Figura 2, à frente do sinal de “maior que” em vermelho, é o

```
R version 4.0.2 (2020-06-22) -- "Taking Off Again"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R é um software livre e vem sem GARANTIA ALGUMA.
Você pode redistribuí-lo sob certas circunstâncias.
Digite 'license()' ou 'licence()' para detalhes de distribuição.

R é um projeto colaborativo com muitos contribuidores.
Digite 'contributors()' para obter mais informações e
'citation()' para saber como citar o R ou pacotes do R em publicações.

Digite 'demo()' para demonstrações, 'help()' para o sistema on-line de
ajuda,
ou 'help.start()' para abrir o sistema de ajuda em HTML no seu navegador.
Digite 'q()' para sair do R.

[Workspace loaded from ~/.RData]

> |
```

Prompt aguardando comandos

Mensagens de boas vindas

Figura 2 – Tela de boas vindas do R, em sua versão 4.0.2 (“*Taking off Again*”), em ambiente Windows, destacando o prompt de comando.

Fonte: Do autor.

prompt. O *prompt* é herdado de linguagens como o MS-DOS e indica o ponto onde se deve inserir as linhas de comando. Lembre-se: no console do R, tudo o que você *disser* ficará impresso na tela na cor vermelha, e tudo que o R te *responder* ficará impresso em azul.

1.1.2 Iniciando o R

Uma vez que você inicia o R, em ambiente Windows, sua tela é aberta com uma barra de menu, algumas mensagens básicas e um *prompt* vermelho (Figura 2).

As informações básicas se referem ao registro do R, às suas regras de distribuição, seus colaboradores, como citar o R, como pedir uma demonstração, como pedir ajuda, e como sair do R.

A barra de menu traz diversos botões de atalho para a manipulação de arquivos, pacotes, ajudas, etc. O que se vê na Figura 2, à frente do sinal de “maior que”, é o *prompt*. O *prompt* é herdado de linguagens como o MS-DOS e indica o ponto onde se deve inserir as linhas de comando. Lembre-se: tudo o que você disser ao R ficará impresso na tela na cor vermelha, e tudo que o R lhe responder ficará impresso em azul.

Ao tentar sair do R, pela barra de menu ou pelo comando `q()`, é mostrada uma mensagem perguntando se o usuário deseja salvar a *área de trabalho*, ou seja, se os objetos atribuídos devem permanecer com os mesmo valores, ou se tudo que foi feito deve ser ignorado (Figura 3). Quando se inicia novamente o R, após ter salvo a área de trabalho, os objetos anteriormente criados são carregados automaticamente. Aconselha-se que toda a informação desejada seja gravada em outro tipo de arquivo, e a área de trabalho seja raramente salva. Isso evitará confusões quanto ao valor de objetos ao se fazer uma conta.

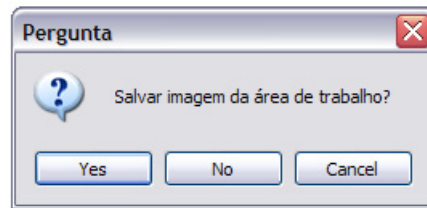


Figura 3 – Mensagem recebida ao se fechar o console do R, em ambiente Windows, perguntando se a área de trabalho deve ser salva.

Fonte: Do autor.

1.1.3 Lendo arquivos

A maneira mais fácil de inserir dados em objetos no R é a leitura de arquivos. O R pode ler arquivos de estruturas simples como as extensões *.txt* e *.r*. Também é possível importar outros tipos de arquivos mais complexos, como os *.xls*, mas os procedimentos de importação serão tratados aqui. O que se aconselha, quando se tem um arquivo *.xls*, é salvá-lo como *.txt* e fazer a leitura normalmente.

Vale lembrar que, quando se salva uma área de trabalho, apenas os valores dos objetos são guardados. Todos os comandos dados e todos os resultados não armazenados em objetos são perdidos. Por esses motivos, é fortemente recomendado que se trabalhe com o R em associação a um editor de texto da sua preferência. Alguns editores de texto muito úteis são: o *script* do R, o RStudio, o Bloco de Notas do Windows, o Tinn-R, o WinEdt e o Emacs. Esses editores são usados tanto para elaborar os arquivos de dados que serão lidos pelo R, quanto para armazenar rotinas (conjuntos de linhas de comando) para a repetição futura da análise.

Aqui, destacamos o [RStudio Team \(2020\)](#). Criado entre 2010 e 2011, esse IDE¹ é o mais famoso, versátil e poderoso gerenciador do R. Nele, podemos visualizar, simultaneamente, o *script*, o console, os objetos da área de trabalho, gráficos e *helps* (Figura 4).

Para ler um arquivo no R, a função mais usada é a `read.table()`. Essa função lê o arquivo e o armazena (se desejado) na forma de *data frame* em um objeto. O primeiro argumento dessa função se refere ao nome do arquivo a ser lido. Esse argumento deve vir entre aspas. Entretanto, o endereço desse arquivo também deve ser passado para o R. Para isso, tem-se duas opções: (1) Na barra de menu, botão Arquivo, mudar diretório para o lugar onde se encontra o arquivo; (2) Escrever todo o endereço do arquivo dentro do primeiro argumento da função `read.table()`. O segundo argumento dessa função se refere ao cabeçalho (nome) das colunas de dados contidas no arquivo. Se as colunas tiverem cabeçalho (*header*), então deve-se digitar `h=TRUE`, caso contrário, `h=FALSE`.

Exemplos de comando de leitura de arquivo quando se muda o diretório de leitura para o lugar onde o arquivo está armazenado, e quando o endereço é informado na função.

```
read.table('nome.txt', h=TRUE)
read.table('C:\\Meus_Documentos\\nome.txt', h=TRUE)
```

¹ IDE: do inglês *Integrated Development Environment* ou Ambiente de Desenvolvimento Integrado, é um programa de computador que reúne características e ferramentas de apoio ao desenvolvimento de software com o objetivo de agilizar este processo.

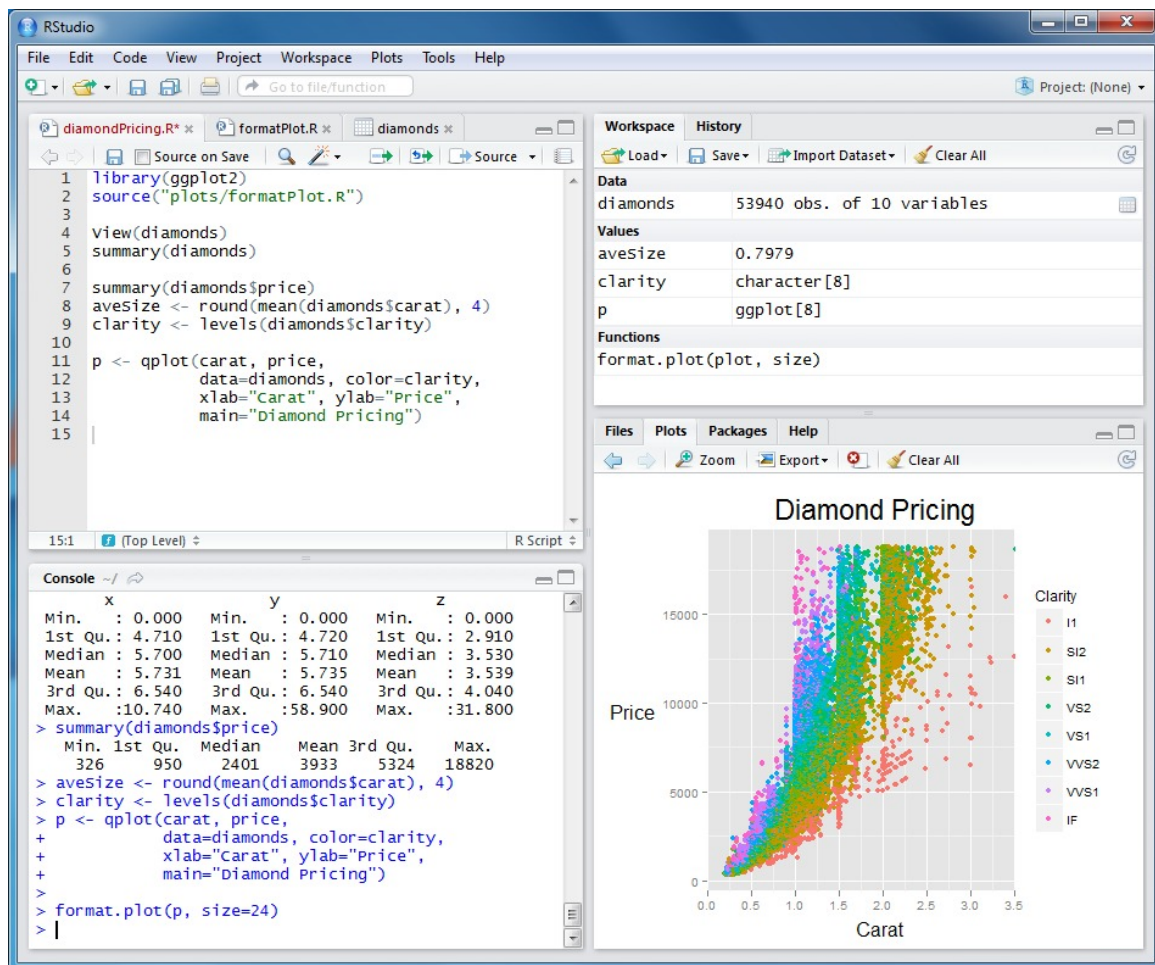


Figura 4 – RStudio: um IDE (ambiente de desenvolvimento integrado) muito útil para o usuário e o programador R.

Fonte: Do autor.

1.1.4 Funções básicas

Aqui são apresentadas algumas funções de uso constante, pertencentes aos pacotes básicos do R:

`sum(x)` : soma todos os elementos de um objeto `x`.

`length(x)` : retorna o comprimento de um objeto `x`.

`rep(x, n)` : repete o número `x`, `n` vezes.

`seq(a, b, by=c)` : gera uma sequência de números contidos entre `a` e `b`, distantes `c` unidades um do outro.

`table(x)` : retorna uma tabela com as frequências absolutas de ocorrência da cada elemento de `x`.

1.1.5 Pedindo ajuda

O jeito mais fácil de se aprender a usar R é consultando constantemente seus tópicos de ajuda. Existem basicamente quatro tipos de ajuda no R:

1. `help('função()')` : Essa ajuda deve ser solicitada quando se sabe da existência de uma

função (sabe-se seu nome exato), mas existem dúvidas em como usá-la. Se o pacote que contém essa função estiver instalado e carregado, será aberta a documentação da mesma para esclarecimentos;

2. `help.search(' ')`: Quando se deseja investigar a existência de uma função, essa ajuda recebe uma palavra-chave (em Inglês) e retorna todas aquelas funções que contêm aquela palavra em sua documentação. A busca é feita nos pacotes existentes no computador em questão, ou seja, se uma busca não retornar nenhum resultado adequado, não significa que a função não exista. Significa que ela não existe, pelo menos, em seu computador;
3. Ajuda Html: Essa ajuda pode ser chamada pela barra de menu, no botão Ajuda (Help). Quando acionada, ela abre um documento em html que contém diversas informações sobre o R, sua linguagem, suas funções básicas, seus pacotes, seus autores, sua licença, perguntas mais frequentes etc;
4. `RSiteSearch(' ')`: Quando conectado à internet, essa ajuda faz a busca de uma palavra-chave em todas as páginas da internet relacionadas com o R, principalmente aquelas páginas publicadas com as perguntas e respostas das listas de discussões do R. Existem diversos tipos de listas de discussões que podem ser encontradas na página do R. Nelas, são tiradas dúvidas mais graves, são dadas sugestões para as novas versões do R, são desvendados e descobertos pequenos erros de programação etc. Elas colocam os usuários do R em contato com os estatísticos que fazem e mantêm o R.

O jeito mais fácil de se aprender a usar R é consultando constantemente seus tópicos de ajuda. Quando desejamos conhecer detalhes de uma função, podemos digitar `help("função")` ou, simplesmente, `?função()`. Caso desejemos saber se um tópico possui função no R, o comando deve ser: `help.search("tópico")`.

Além disso, R também apresenta uma série de recursos gráficos que permitem a descrição de todos os detalhes que se pode querer personalizar em um gráfico, como cor, tipo e tamanho de letra, símbolos, títulos e subtítulos, pontos, linhas, legendas, planos de fundo e muito mais.

1.1.6 Objetos

Mais que um software que realiza análises estatísticas, R é um ambiente e uma linguagem de programação orientada a objeto. Nele, números, vetores, matrizes, *arrays*, *data frames* e listas podem ficar armazenados em objetos (Figura 5). Pode-se entender objeto como uma caixinha onde você pode guardar o que quiser. A partir daí, todas as operações matemáticas podem ser feitas usando esses objetos. Isso torna as coisas mais simples.

Para criar um objeto, basta atribuir um valor a um nome, ou seja, quando se coloca um valor dentro de um objeto este passa a existir automaticamente. Uma atribuição pode ser feita, basicamente, de duas maneiras: usando o sinal de =, ou usando uma seta formada pela junção dos sinais de *menor que* e *menos* <-. Note que essa seta sempre deve levar o valor ao objeto, ou seja, deve sempre apontar para o objeto. Portanto, é possível usar a seta em ambas as direções (<- ou ->).

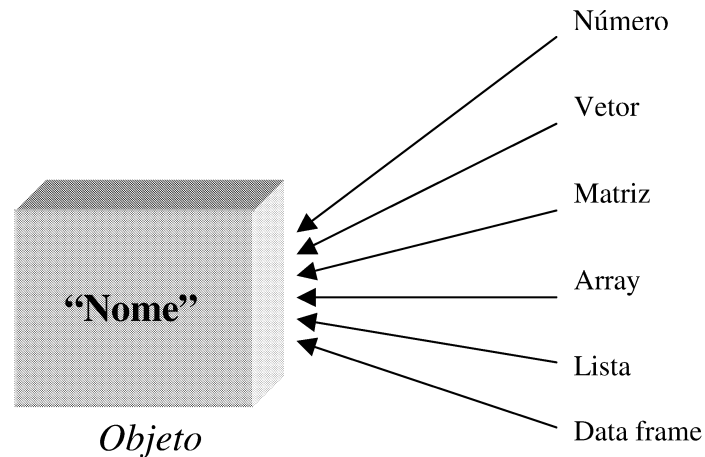


Figura 5 – Esquema de um objeto, estrutura que pode armazenar números, vetores, matrizes, *arrays*, listas e *data frames*.

Fonte: Do autor.

Outro sinal muito útil na linguagem R é o sinal de comentário #, ou seja, sinal a partir do qual o que for escrito não será interpretado como comando.

(a) **Número.** É possível atribuir apenas um número a um objeto. Por exemplo, atribua o número 2 ao objeto *a* e o número 5 ao objeto *x*.

```
a <- 2    #a recebe 2
x <- 5    #x recebe 5
```

Para verificar quanto vale cada objeto, apenas digite seu nome e tecla *enter*. O número 1 entre colchetes, observado abaixo, se refere à primeira posição do vetor, ou seja, um número é entendido com o vetor de uma posição.

```
a
[1] 2
```

Uma vez criados, os objetos podem ser usados em contas, equações, funções, sistemas etc.

```
a+x      #soma
a-x      #subtração
a*x      #produtos de escales
a/x      #divisão
a^x      #potenciação
```

O resultado de uma conta, por sua vez, pode ser guardado dentro de um terceiro objeto.

```
c<-2*a + 300/x
```

(b) **Vetor.** O vetor da linguagem R tem um significado um pouco diferente que o vetor da Matemática. Para o R, um vetor é qualquer conjunto unidimensional de valores. Esses valores podem ser números, *strings* (palavras) ou valores lógicos (F para falso e T para verdadeiro). Em outras palavras, para o R, o vetor tem um significado mais amplo que para a Matemática.

Para se atribuir um conjunto de valores a um objeto pode-se usar o comando `c()`, onde os valores vêm separados por vírgulas, dentro dos parênteses.

```
d<-c(5,8,12,3.5,9,1) #d recebe um vetor
```

É possível se referir especificamente a uma posição do vetor. Imagine que se deseje saber qual o valor que ocupa a quarta posição do vetor d . Essa referência é feita entre colchetes, após o nome do objeto.

```
d[4] #4a posicao do vetor d
```

(c) **Matriz:** Uma matriz é atribuída a um objeto pela função `matrix()`. Essa função tem como argumentos o conjunto de dados, o número de linhas e o número de colunas da matriz, nessa ordem. Note que o conjunto de dados preenchem a matriz coluna por coluna.

Observe o exemplo. Para inserir, no R, a matriz:

$$\begin{pmatrix} 5 & 12 & 9 \\ 8 & 3.5 & 1 \end{pmatrix}$$

Faça:

```
e<-matrix(c(5,8,12,3.5,9,1),2,3)
```

Nas matrizes também é possível referenciar uma linha, uma coluna ou um elemento. Novamente, deve-se usar números entre colchetes, porém respeitando a ordem: primeira posição se refere à linha, e a segunda posição se refere à coluna.

```
e[2,] #linha 2 da matriz e
e[,3] #coluna 3 da matriz e
e[1,3] #elemento da linha 1, coluna 3
```

(d) **Array:** Esse termo em Inglês não possui tradução adequada. Ele representa uma hiper matriz, ou seja, um conjunto de números arranjados em mais de 2 dimensões. Quando tem 3 dimensões, um *array* pode ser entendido como um conjunto de matrizes de mesma dimensão. Por exemplo:

$$\begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} 5 & 6 \\ 6 & 8 \end{pmatrix}$$

Aqui, a referência a linhas e colunas é a mesma das matrizes, e a terceira posição dos colchetes se refere ao valor de interesse na terceira dimensão.

O comando usado é o `array()`. Uma forma de atribuir um *array* a um objeto é inserir um *array* de zeros (nas dimensões desejadas) e depois preenchê-lo com os valores adequados. Outra opção é fazer um vetor que respeite a ordem: por coluna, por matriz; e usá-lo já na construção do *array*.

A primeira posição da função `array()` se refere aos argumentos das matrizes, e a segunda posição se refere às dimensões do mesmo.

```
f<-array(0,c(2,2,2)) #array de zeros
f[, ,1]<-matrix(c(1,2,3,4),2,2) #primeira matriz
f[, ,2]<-matrix(c(5,6,7,8),2,2) #segunda matriz
```

ou

```
f<-array(c(1,2,3,4,5,6,7,8),c(2,2,2)) #de uma só vez
```

Analogamente, pode-se perguntar qual é o valor que ocupa a primeira linha, da segunda coluna, da segunda matriz, do objeto f.

```
f[1,2,2]
```

(e) **Data frame**: Essa estrutura de dados é uma espécie de tabela, de estrutura bidimensional de dados. Podem fazer parte de um mesmo *data frame* números e *strings*. Além disso, podem ser dados nomes às colunas. Sua função é `data.frame()`. Veja o exemplo. Para produzir a tabela abaixo, reproduza o código R que a segue.

Marca	Preço
Volkswagen	32000
Fiat	28000
Ford	29500

```
g<-data.frame('Marca'=c('Volkswagen','Fiat','Ford'),
'Preço'= c(32000,28000,29500))
```

(f) **Lista**: Uma lista é um conjunto de objetos de tamanhos e naturezas diferentes. Ela é regida pela função `list()`. Essa é a estrutura mais geral da linguagem R. Suas posições são designadas por números entre dois parênteses `[[]]`.

Considere o exemplo de lista que contém um número 3 na primeira posição; a matriz

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix},$$

na segunda; a palavra *lista*, na terceira; e, na quarta posição, o vetor

$$(5 \ 6 \ 7 \ 8).$$

```
h<-list(3,matrix(c(1,2,3,4),2,2),'lista',c(5,6,7,8))
```

Suponha que se deseje saber o terceiro elemento do vetor que está alocado na posição quatro da lista *h*, que é o 7. Basta digitar:

```
h[[4]][3]
```

1.2 Alguns conceitos importantes

Estatística é conjunto de técnicas para a coleta, organização, análise e interpretação de *dados* para a descrição de *populações*.

Dado é o valor assumido por uma *variável aleatória* em determinado experimento.

População é o todo que se quer descrever. É o conjunto de elementos com características em comum.

Ex.: Deseja-se saber o teor médio de açúcar (graus Brix) de uma determinada variedade de laranja.

1.2.1 Classificações de uma população

- Finita (ou real): fixa no tempo.
Ex: Conjunto de árvores em um talhão.
- Infinita (ou conceitual): engloba elementos não existentes.
Ex: Plantas de feijão da cultivar *carioca* (que existiram, existem ou virão a existir).

Amostra é o subconjunto com n elementos da população.

Censo é a observação exaustiva de todos os N elementos da população.

Evento é cada possível resultado em um experimento. Ex: A face “cara” cair voltada para cima no lançamento de uma moeda honesta.

Estatística experimental tem por objetivo comparar mais de duas populações simultaneamente (tratamentos).

1.2.2 Tipos de variáveis

Nesta seção são definidos e exemplificados os tipos de variáveis mais comuns e úteis.

1.2.2.1 Qualitativas

São aquelas variáveis que indicam qualidades, atributos, características não numéricas de forma geral.

1.2.2.2 Qualitativas nominais

São aquelas que não permitem uma ordenação natural.

Ex: O conjunto de espécies: Cedro, Cassia e Ipê.

1.2.2.3 Qualitativas ordinais

Por sua vez, são aquelas que admitem uma ordenação natural.

Ex: O ciclo de uma cultura: precoce, médio e tardio.

1.2.2.4 Quantitativas

Resumem-se a medidas, pesagens ou contagens.

1.2.2.5 Quantitativas discretas

São representadas pelas contagens.

Ex: N° de espigas por planta de milho.

1.2.2.6 Quantitativas contínuas

São representadas pelas medições ou pesagens (R).

Ex: Produtividade (t/ha).

1.3 Principais aplicações da Estatística

Nesta seção são apresentadas as duas principais aplicações práticas da Estatística.

1.3.1 Pesquisa científica

A Estatística, em muitos momentos, confunde-se com o próprio conceito de fazer ciência². Ela tem o objetivo de administrar a incerteza acerca de um fenômeno, permitindo uma melhor compreensão do ambiente em que vivemos, e permitindo a tomada de decisões e a realização de previsões.

	Determinísticos	Principalmente usados na Física, Química e Matemática. As relações são exatas, ou seja, possíveis variações casuais são desprezadas. Ex.: $6CO_2 + 6H_2O \rightarrow C_6H_{12}O_6 + 6O_2$
Modelos		
	Probabilísticos	As variações naturais não são desprezadas e são descritas por meio de um componente probabilístico. Ex.: Peso de suínos em função da quantidade de ração ingerida ao longo de sua engorda (Figura 6).

1.3.2 Processos produtivos

A Estatística é usada diariamente no controle de processos produtivos empresariais por meio do Controle de Qualidade.

1.3.3 Levantamentos em geral

Os usos mais populares da Estatística se dá em censos demográficos e pesquisas eleitorais. Porém, há outras participações da estatística, como em pesquisas de mercado, inventários florestais, etc, que são tão importantes para a sociedade quanto a primeira.

Este gráfico foi feito no R. Quer saber como? Dê uma olhada na rotina, apresentada a seguir.

² *Ciência* (do latim *scientia*, traduzido por “conhecimento”) refere-se a qualquer conhecimento ou prática sistemáticos. Em sentido estrito, ciência refere-se ao sistema de adquirir conhecimento baseado no método científico bem como ao corpo organizado de conhecimento conseguido através de tais pesquisas. Além disso, a Ciência se baseia na observação de evidências para refutar (ou afirmar) hipóteses sobre fenômenos naturais.

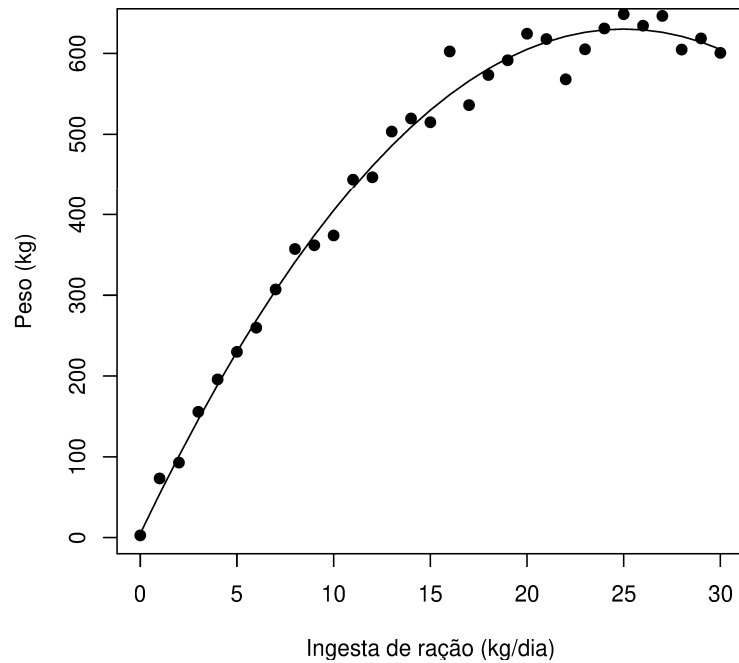


Figura 6 – Gráfico ilustrativo do comportamento do peso de suínos (kg) em função de sua ingestão diária de ração ao longo do período de engorda.

Fonte: Do autor.

```
x<-seq(0, 30)
y<--x^2+50*x+5
plot(x, y, 'l', ylab='Peso_(kg)', xlab='Ingesta_de_ração_(kg/dia)')
e<-rnorm(length(y), 0, 25)
yo<-y+e
points(x, yo, pch=19)
```

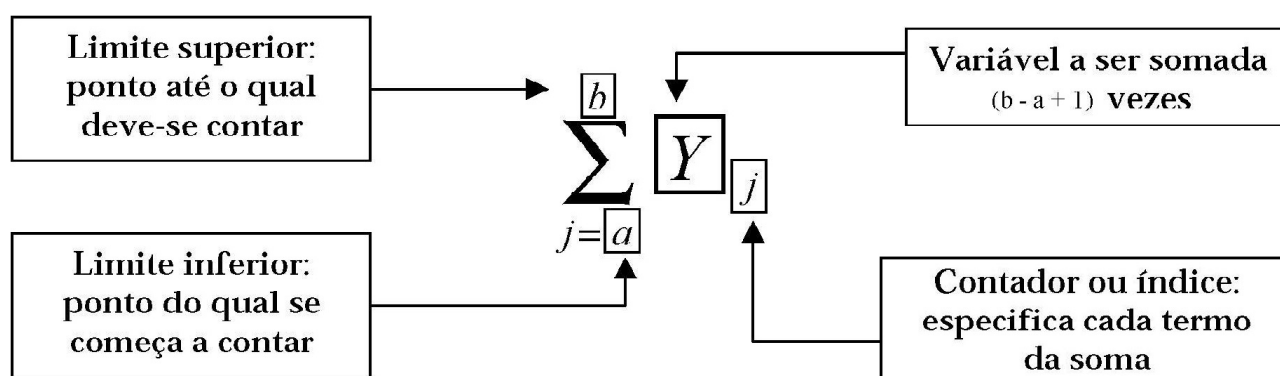

TÉCNICAS DE SOMATÓRIO

2.1 Introdução

As técnicas de *somatório* são um tópico da Matemática, mas, é tão útil na Estatística que vale a pena ser tratado aqui. Se você já tem esse conhecimento, passe para o próximo tópico.

Os somatórios são notações compactas e resumidas para representar somas, de dois ou mais termos, com algum elemento em comum. O somatório (ou a somatória) tem por objetivo simplificar a notação de uma soma de termos, ou seja, de um polinômio.

O *somando*, que vem à direita da letra grega Sigma maiúscula (Σ), é aquilo que está sendo somado. O somando é repetido a cada termo da soma, enquanto o *contador* ou *índice*, varia de um limite inferior até um limite superior. Geralmente, utilizam-se as letras X , Y , e Z como variável e i , j , k , ..., como contador. Entretanto, essa é uma prática que pode ser dispensada, se necessário.



$$\text{Ex}_1: 4 + 4 + 4 = \sum_{i=1}^3 4.$$

$$\text{Ex}_2: y_1 + y_2 + y_3 + y_4 = \sum_{i=1}^4 y_i.$$

$$\text{Ex}_3: (x_1 + y_1)^2 + (x_2 + y_2)^2 + \dots + (x_k + y_k)^2 = \sum_{i=1}^k (x_i + y_i)^2.$$

Ex₄: Seja $A = \{a_1, a_2, a_3, a_4, a_5\}$ um conjunto de dados. Então, o somatório dos elementos de A pode ser escrito como: $a_1 + a_2 + a_3 + a_4 + a_5 = \sum_{i=1}^5 a_i$.

2.2 Propriedades dos somatórios

Propriedade 1: Seja a uma constante, e X uma variável aleatória, então

$$\boxed{\sum_{i=1}^n ax_i = a \sum_{i=1}^n x_i.} \quad (2.1)$$

Prova algébrica:

$$\sum_{i=1}^n ax_i = ax_1 + ax_2 + \dots + ax_n = a(x_1 + x_2 + \dots + x_n) = a \sum_{i=1}^n x_i$$

□

Exemplo numérico:

$$\sum_{i=1}^3 2y_i = 2y_1 + 2y_2 + 2y_3 = 2(y_1 + y_2 + y_3) = 2 \sum_{i=1}^3 y_i.$$

Propriedade 2: Sejam X e Y variáveis aleatórias, então

$$\boxed{\sum_{i=1}^n x_i y_i \neq \left(\sum_{i=1}^n x_i \right) \times \left(\sum_{i=1}^n y_i \right).} \quad (2.2)$$

Prova algébrica:

$$\begin{aligned} \sum_{i=1}^n x_i y_i &\neq \left(\sum_{i=1}^n x_i \right) \times \left(\sum_{i=1}^n y_i \right) \\ x_1 y_1 + x_2 y_2 + \dots + x_n y_n &\neq (x_1 + x_2 + \dots + x_n) \times (y_1 + y_2 + \dots + y_n) \\ x_1 y_1 + x_2 y_2 + \dots + x_n y_n &\neq x_1 y_1 + x_1 y_2 + \dots + x_1 y_n + \dots + x_2 y_1 + x_2 y_2 + \\ &\quad + \dots + x_2 y_n + \dots + x_n y_1 + x_n y_2 + \dots + x_n y_n \end{aligned}$$

□

Exemplo numérico:

$$\begin{aligned} \sum_{i=1}^2 2x_i 3y_i &\neq \left(\sum_{i=1}^2 2x_i \right) \left(\sum_{i=1}^2 3y_i \right) \\ 2x_1 3y_1 + 2x_2 3y_2 &\neq (2x_1 + 2x_2) (3y_1 + 3y_2) \\ 2x_1 3y_1 + 2x_2 3y_2 &\neq 2x_1 3y_1 + 2x_2 3y_1 + 2x_1 3y_2 + 2x_2 3y_2. \end{aligned}$$

Propriedade 3: Sejam a e b constantes, e X e Y variáveis aleatórias, então

$$\boxed{\sum_{i=1}^n ax_i \pm by_i = a \sum_{i=1}^n x_i \pm b \sum_{i=1}^n y_i.} \quad (2.3)$$

Prova algébrica:

$$\begin{aligned}\sum_{i=1}^n ax_i + by_i &= ax_1 + ax_2 + \dots + ax_n + by_1 + by_2 + \dots + by_n \\ &= a(x_1 + x_2 + \dots + x_n) + b(y_1 + y_2 + \dots + y_n) \\ &= a \sum_{i=1}^n x_i + b \sum_{i=1}^n y_i.\end{aligned}$$

□

Exemplo numérico:

$$\begin{aligned}\sum_{i=1}^2 3x_i + 4y_i &= 3x_1 + 3x_2 + 4y_1 + 4y_2 \\ &= 3(x_1 + x_2) + 4(y_1 + y_2) \\ &= 3 \sum_{i=1}^2 x_i + 4 \sum_{i=1}^2 y_i.\end{aligned}$$

Propriedade 4: Seja k uma constante, então

$$\boxed{\sum_{i=1}^n k = nk} \quad (2.4)$$

Prova algébrica:

$$\sum_{i=1}^n k = \underbrace{k + k + \dots + k}_n = nk.$$

□

Ou então,

$$\begin{aligned}\sum_{i=1}^n k &= k \sum_{i=1}^n 1 \\ &= k \underbrace{(1 + 1 + \dots + 1)}_n = nk.\end{aligned}$$

□

Exemplo numérico:

$$\sum_{i=1}^7 5 = \underbrace{5 + 5 + \dots + 5}_{7 \text{ vezes}} = 7 \times 5 = 35$$

2.3 Somatórios no R

O R possui uma função básica chamada `sum()`. Essa função faz a soma de todos os elementos de um objeto que for seu argumento: `sum(objeto)`. Por exemplo:

```
x<-c(4, 3, 6, 2, 1, 3, 2, 4, 5)
sum(x)
```

Se você desejar somar apenas alguns valores de seu objeto (por exemplo, os valores da posição a até a posição b , em um vetor), é só indicar o intervalo desejado da seguinte maneira:

```
sum(objeto[a:b])
sum(x[2:5])
```

Com muita facilidade, podemos somar os quadrados dos elementos de um vetor, no R. Para calcular $\sum_{n=1}^9 x_i^2$, faça

```
sum(x^2)
```

Ainda, se definirmos um outro vetor, digamos y , e desejamos calcular a soma dos produtos, ou seja, $\sum_{n=1}^9 x_i y_i$, basta fazermos o produto interno entre os dois vetores (que é exatamente a soma do produto de suas entradas). Para fazer o produto entre vetores ou matrizes, utilizamos o sinal `%*%`.

```
y<-c(8,7,3,4,6,2,3,7,1)
x%*%y
```

Em outro exemplo, considerando os mesmos vetores de dados x e y , suponha que você precisa resolver o seguinte somatório no R: $\sum_{n=1}^9 5 \ln(x_i) + 3y^i = 17.325.556$.

Então, reproduza o script:

```
sum(5*log(x)+3*y^(1:9))
```

Por fim, caso você tenha uma matriz de dados, por exemplo:

$$A = \begin{pmatrix} 2 & 3 \\ 5 & 1 \end{pmatrix}$$

e precise da soma das linhas, soma das colunas e soma de todas as entradas da matriz, basta fazer:

```
A<-matrix(c(2,5,3,1),2,2)
apply(A,1,sum) #Somadas linhas
apply(A,2,sum) #Somadas colunas
sum(A)        #Soma de todas as entradas
```

ESTATÍSTICA DESCRITIVA

3.1 Introdução

Um bom trabalho de coleta de dados experimentais pode render uma massa de dados confiável, porém desordenados, isto é, brutos. Na sua forma bruta os dados não querem dizer muita coisa, ou seja, não são considerados *informação*. Por isso, o objetivo da Estatística Descritiva é apresentar uma série de técnicas de descrição de dados válidas para censos e amostras.

ALGUMAS DEFINIÇÕES PERTINENTES

Frequência. Representa o número de vezes que um valor (ou intervalo de valores) ocorre em uma massa de dados.

Distribuição de frequências. Geralmente apresentada em forma de tabela, associa os valores da variável com suas frequências de ocorrência.

Tipos de frequência

- frequência absoluta (fa): representa o número de vezes que um valor (ou intervalo) ocorre nos dados.
- frequência relativa (fr): representa, em forma decimal, a proporção de ocorrências de um valor em relação ao tamanho da massa de dados,

$$fr = \frac{fa}{n}. \quad (3.1)$$

- frequência percentual (fp): representa, em forma percentual, a proporção de ocorrências de um valor em relação ao tamanho da massa de dados,

$$fp = fr \times 100. \quad (3.2)$$

3.2 Variáveis qualitativas

Experimentos ou pesquisas que possuem, como foco, variáveis qualitativas, podem ser descritos por meio de distribuições de frequência e suas representações gráficas. A seguir, um exemplo ilustra o procedimento.

Exemplo. Um engenheiro agrônomo faz um levantamento das principais atividades agrícolas em uma amostra contendo 20 propriedades de certa região. O croqui a seguir representa esquematicamente o resultado da pesquisa.

C	L	L	C	S	LA	C	C	L	M
C	M	So	M	L	C	C	M	C	L

- Massa de dados: amostra.
- População: finita: conjunto de todas as propriedades rurais desta região que atualmente apresentam atividades agrícolas.
- Variável aleatória: qualitativa nominal: atividade agrícola.
- Valores assumidos pela variável aleatória na pesquisa: café (C), leite (L), silvicultura (S), milho (M), soja (So), laranja (LA).
- Distribuição de frequência¹:

Tabela 1 – Distribuição de frequências absolutas (fa), relativa (fr) e percentual (fp) da atividade em propriedades de uma região.

Atividade	fa	fr	fp (%)
Café	8	0,40	40,00
Leite	5	0,25	25,00
Milho	4	0,20	20,00
Outras	3	0,15	15,00
Total	20	1,00	100,00

Fonte: Dados fictícios.

Pode-se facilmente fazer uma distribuição de frequências no R compondo-se um objeto (df) de forma conveniente. Veja o exemplo:

```
at<-c('C','L','L','M','C','M','So','L','L','C','M','C','S','L','C','LA',
      'C','M','C','C')
tab.at<-table(at)
df<-matrix(0,5,3)
colnames(df)<-c("fa","fr","fp")
rownames(df)<-c("Café","Leite","Milho","Outras","Total")
df[1,1]<-tab.at["C"]
df[2,1]<-tab.at["L"]
df[3,1]<-tab.at["M"]
```

¹ Classes pouco frequentes podem ser agrupadas em categoria “outras”, em último lugar.

```
df[4,1]<-sum(tab.at["So"], tab.at["S"], tab.at["LA"])
df[5,1]<-length(at)
for(i in 1:5) {df[i,2]<-df[i,1]/length(at)}
for(i in 1:5) {df[i,3]<-df[i,2]*100}
```

3.2.1 Representações gráficas das distribuições de frequências

3.2.1.1 Gráfico de barras e de colunas

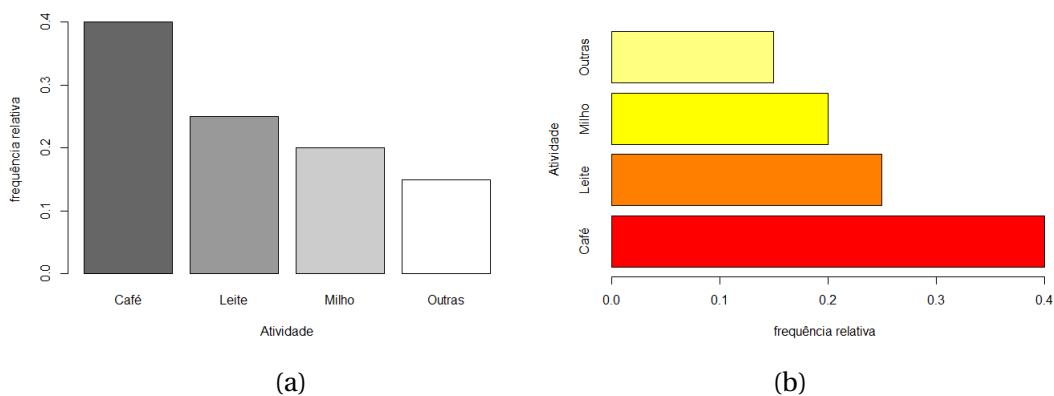


Figura 7 – (a) Gráfico de colunas das principais atividades em propriedades rurais. (b) Gráfico de barras da mesma situação.

Fonte: Do autor.

Reproduza em seu computador a rotina do gráfico de barras...

```
gc<-barplot(df[1:4,2],xlab="Atividade",ylab="frequência_relativa",
col = gray(seq(0.4,1.0,length=4)))
```

... e do gráfico de colunas

```
gc<-barplot(df[1:4,2],horiz=TRUE,ylab="Atividade",
xlab="frequência_relativa",col=heat.colors(4))
```

Note que os gráficos de barras e colunas são feitos com a mesma função (`barplot`). A única diferença é o argumento `horiz`, que deve ser verdadeiro no caso das barras. Mas, lembre-se: mude o nome dos eixos ao inverter o gráfico ou eles ficarão trocados.

3.2.1.2 Gráfico de pizza ou Setograma

O gráfico de pizza, torta ou setograma é um círculo com setores de área proporcional às frequências de ocorrência de cada valor da variável aleatória.

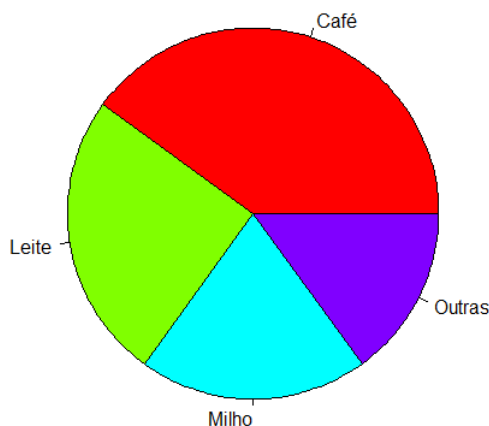


Figura 8 – Setograma (gráfico de setores ou gráfico de pizza) sobre frequências de ocorrência de café, leite, milho e outras.

Fonte: Do autor.

Confira como fazer um setograma no R:

```
pie(df[1:4,2], col=raibow(4), radius=1.05)
```

A função `pie` exige que haja um objeto contendo números decimais que somem 1, ou seja, frequências relativas.

3.3 Variáveis quantitativas discretas

Variáveis quantitativas discretas podem ser vistas como casos particulares de variáveis quantitativas contínuas. Pode-se tratar uma massa de dados de variáveis quantitativas discretas como se fosse de variáveis qualitativas, ou seja, cada valor assumido pela variável pode ser visto como uma classe. Porém, quando a variável, apesar de assumir valores discretos, puder assumir uma quantidade muito grande de valores, ela pode ser tratada como uma variável quantitativa contínua, ou seja, construindo-se classes. Os procedimentos indicados para a manipulação de variáveis quantitativas contínuas serão apresentados a seguir.

A representação gráfica das variáveis quantitativas discretas se dá de forma semelhante à das qualitativas ordinais.

Veja o seguinte exemplo: uma pesquisa da Secretaria de Saúde Pública de um município investigou o número de filhos por casal. Esses são uma parte dos resultados obtidos:

3	4	3	1	3	2	1	1	2	2	4	4	1	3	2	2	4	4	3	3
1	0	2	1	3	2	2	4	2	1	1	4	1	0	1	3	3	0	3	3

A Tabela 2 apresenta a distribuição de frequência do número de filhos por casal em um determinado município.

Tabela 2 – Distribuição de frequências absolutas (fa), relativa (fr) e percentual (fp) do número de filhos por casal de uma cidade.

Classes	fa	fr	fp (%)
0	3	0,075	7,50
1	10	0,250	25,00
2	9	0,225	22,50
3	11	0,275	27,50
4	7	0,175	17,50
Total	40	1,000	100,00

Fonte: Dados fictícios.

De forma semelhante, pode-se fazer uma distribuição de frequências no R compondo-se um objeto (df). No exemplo:

```
filhos<-c(3,4,3,1,3,2,1,1,2,2,4,4,1,3,2,2,4,4,3,3,1,0,2,1,3,2,2,4,
2,1,1,4,1,0,1,3,3,0,3,3)
tab.filhos<-table(filhos)
df<-matrix(0,6,3)
colnames(df)<-c("fa","fr","fp")
rownames(df)<-c(0,1,2,3,4,"Total")
df[1,1]<-tab.filhos["0"]
df[2,1]<-tab.filhos["1"]
df[3,1]<-tab.filhos["2"]
df[4,1]<-tab.filhos["3"]
df[5,1]<-tab.filhos["4"]
df[6,1]<-length(filhos)
for(i in 1:6) {df[i,2]<-df[i,1]/length(filhos)}
for(i in 1:6) {df[i,3]<-df[i,2]*100}
```

3.3.1 Representação gráfica da distribuição de frequências

3.3.1.1 Gráfico de agulhas

Uma das formas de representar graficamente a distribuição de frequências de variáveis quantitativas discretas é o gráfico de linhas. Ao contrário do que se costuma chamar de gráfico de linhas, esse é um gráfico que representa as alturas de cada ocorrência da variável por meio de linhas (Figura 9). Esse é o limite do gráfico de colunas quando a largura da coluna tende a zero. Isso faz sentido já que, nesse caso, a classe se resume a um ponto, ou seja, a amplitude de classe (c) é zero.

Apesar de não fazer muito sentido, o gráfico de colunas tem sido muito utilizado para representar variáveis quantitativas discretas devido a seu apelo visual. A Figura 9 também apresenta o gráfico de colunas para esse exemplo.

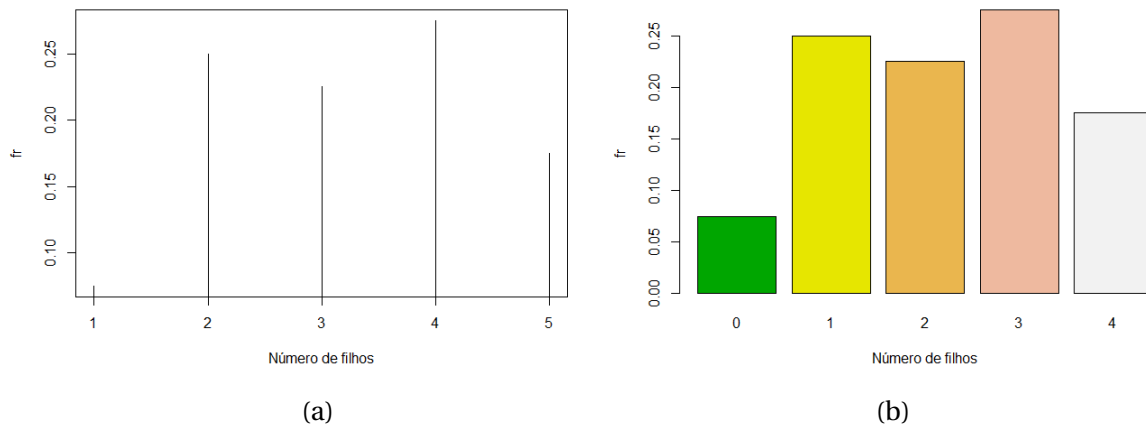


Figura 9 – (a) gráfico de linhas da variável “número de filhos por casal”. (b) gráfico de colunas da mesma variável.

Fonte: Do autor.

Representação da distribuição de frequências de uma variável qualitativa:

```
gl<-plot(df[1:5,2], type="h", xlab="Número_de_filhos", ylab="fr") #linhas
gb<-barplot(df[1:5,2], col=terrain.colors(5),
xlab="Número_de_filhos", ylab="fr") #colunas
```

3.4 Variáveis quantitativas contínuas

Aqui se descreve uma sequência de passos indicados para a construção de uma distribuição de frequências para variáveis quantitativas contínuas. No entanto, é importante ressaltar que essa é apenas uma de uma infinidade de maneiras com as quais se poderia construir uma distribuição de frequência eficiente e compreensível. Na literatura especializada facilmente pode-se encontrar diferentes sugestões de procedimento.

1. Determinar o número de classes (k):

Critério empírico	Critério de Scott (1979)
$k \simeq \begin{cases} \sqrt{n}, & \text{se } n < 100 \\ 5 \log n, & \text{se } n > 100 \end{cases}$	$k \simeq 1 + \frac{An^{1/3}}{3,495}$

n : número de elementos da amostra.

2. Cálculo da Amplitude Total (A):

$$A = MVO - mvo, \quad (3.3)$$

em que MVO é o maior valor observado; e mvo é o menor valor observado.

3. Cálculo da amplitude de classe (c):

Sejam as seguintes fórmulas, se a massa de dados em questão se tratar de censos

$$c = \frac{A}{k} \quad (3.4)$$

ou amostras

$$c = \frac{A}{k-1}. \quad (3.5)$$

4. Limite inferior da primeira classe (LI_1):

Em censos:

$$LI_1 = mvo \quad (3.6)$$

ou amostras

$$LI_1 = mvo - \frac{c}{2}. \quad (3.7)$$

5. Demais limites:

$$LS_i = LI_i + c \quad (3.8)$$

e

$$LS_i = LI_{i+1}, \quad (3.9)$$

para todo $i = 1, \dots, k$.

Para ilustrar a sequência de passos descrita, considere o seguinte exemplo: em uma linha de envasamento de potinhos de canela em pó, a especificação é enchê-los com 50g do produto. Se a envasadora colocar mais que o especificado a empresa estará sendo lesada. Caso contrário, o consumidor será enganado. Por essa razão, é conveniente fazer o acompanhamento dos potinhos envasados. Coletou-se uma amostra de 50 potinhos dessa linha de produção, que aqui são dispostos em ordem crescente, em gramas (g).

45,2	45,3	45,4	45,7	45,9	46,1	46,1	46,2	46,5	46,6	46,9	47,9	48,1	48,1	48,3
48,5	48,8	48,8	49,1	49,2	49,3	49,7	49,8	49,9	50,1	50,2	50,3	50,4	50,5	50,5
50,5	50,6	50,8	51,0	51,1	51,4	51,6	51,6	51,7	51,9	52,5	52,7	52,8	53,0	54,9
55,0	55,2	55,3	55,7	55,7										

Portanto,

$$1. n < 100 \Rightarrow k = \sqrt{n} = \sqrt{50} = 7,07 \sim 7 \text{ classes.}$$

$$2. A = 55,7 - 45,2 = 10,5g$$

$$3. c = \frac{A}{k-1} = \frac{10,5}{7-1} = 1,75g$$

$$4. LI_1 = 45,2 - \frac{1,75}{2} \sim 44,33g$$

Veja como construir uma distribuição de frequências de uma variável quantitativa contínua:

Tabela 3 – Distribuição de frequências absoluta (fa), relativa (fr) e percentual (fp) do peso observado em potinhos de canela em pó.

Classes	fa	fr	fp (%)
[44,33;46,08)	5	0,10	10,0
[46,08;47,83)	6	0,12	12,00
[47,83;49,58)	10	0,20	20,00
[49,58;51,33)	14	0,28	28,00
[51,33;53,08)	9	0,18	18,00
[53,08;54,83)	0	0,00	00,00
[54,83;56,58)	6	0,12	12,00
Total	50	1,000	100,00

Fonte: Dados fictícios.

```
canela<-c(45.2, 45.3, 45.4, 45.7, 45.9, 46.1, 46.1, 46.2, 46.5, 46.6, 46.9, 47.9,
48.1, 48.1, 48.3, 48.5, 48.8, 48.8, 49.1, 49.2, 49.3, 49.7, 49.8, 49.9, 50.1, 50.2,
50.3, 50.4, 50.5, 50.5, 50.5, 50.6, 50.8, 51.0, 51.1, 51.4, 51.4, 51.6, 51.7, 51.9,
52.5, 52.7, 52.8, 53.0, 54.9, 55.0, 55.2, 55.3, 55.7, 55.7)
df<-matrix(0, 8, 3)
colnames(df) <-c("fa", "fr", "fp")
rownames(df) <-c('[44,33;46,08)', '[46,08;47,83)', '[47,83;49,58)',
'[49,58;51,33)', '[51,33;53,08)', '[53,08;54,83)', '[54,83;56,58)', 'Total')
tab.canela<-table(cut(canela,breaks=c(44.33, 46.08, 47.83, 49.58, 51.33, 53.08,
54.83, 56.58)))
df[1:7,1]<-tab.canela
df[8,1]<-length(canela)
for(i in 1:8) {df[i,2]<-df[i,1]/length(canela)}
for(i in 1:8) {df[i,3]<-df[i,2]*100}
```

Note que, mesmo utilizando o R, de acordo com a rotina apresentada é necessário calcular o número e os limites das classes seguindo um método de interesse.

3.4.1 Representação gráfica da distribuição de frequências

3.4.1.1 Histograma

A representação gráfica mais usada para representar variáveis quantitativas contínuas é o histograma. Histogramas são gráficos de barras verticais justapostas em um eixo contínuo. Neles, o eixo X recebe a variável em estudo, ou seja, abriga as classes. A largura das colunas representa a amplitude das classes. O eixo y recebe as frequências (absolutas, relativas, percentuais ou densidades de frequência).

Densidades de frequência são razões entre as frequências de ocorrência e as amplitudes de classe. Elas *traduzem* o que realmente acontece nas classes quando essas possuem amplitudes diferentes entre si. Com i variando da classe 1 à classe k , podem ser calculadas como:

$$df_i = \frac{f_i}{c_i} \quad (3.10)$$

Podem ser calculadas densidades de frequência absolutas, relativas ou percentuais, de acordo com o interesse, dividindo-se as respectivas frequências pelas amplitudes. Contudo, aconselha-se o uso das densidades de frequência relativas (dfr). No histograma em que a altura das colunas representa a dfr , a área corresponde à frequência relativa (probabilidade).

Nota: Aconselha-se evitar construir classes vazias, pois elas são pouco informativas. No exemplo da canela em pó, as classes 6 e 7 podem ser fundidas em uma só classe. Dessa forma, a frequência relativa passa a valer 0,12 e a amplitude de classe vale 3,5g. Então, por exemplo, calculando-se a dfr , tem-se:

$$dfr_6 = \frac{0,00 + 0,12}{3,5} = 0,0343.$$

Dessa maneira pode-se construir o histograma referente ao exemplo (Figura 10).

Outro dispositivo visual comumente usado é o *polígono de frequência*, que nada mais é do que a união, por meio de segmentos de reta, dos pontos médios das classes (Figura 10).

Podem ser úteis também as frequências absolutas acumuladas *para cima* (ou *acima de*) e *para baixo* (ou *abaixo de*). Podem informar, por exemplo, quantos potinhos de canela contêm menos de 48g. Uma tabela pode ser construída para explicitar os limites das classes e quantos elementos da amostra estão abaixo ou acima daquele valor (Tabela 4). Os dispositivos gráficos usados para representá-las chamam-se *ogivas* (Figura 11).

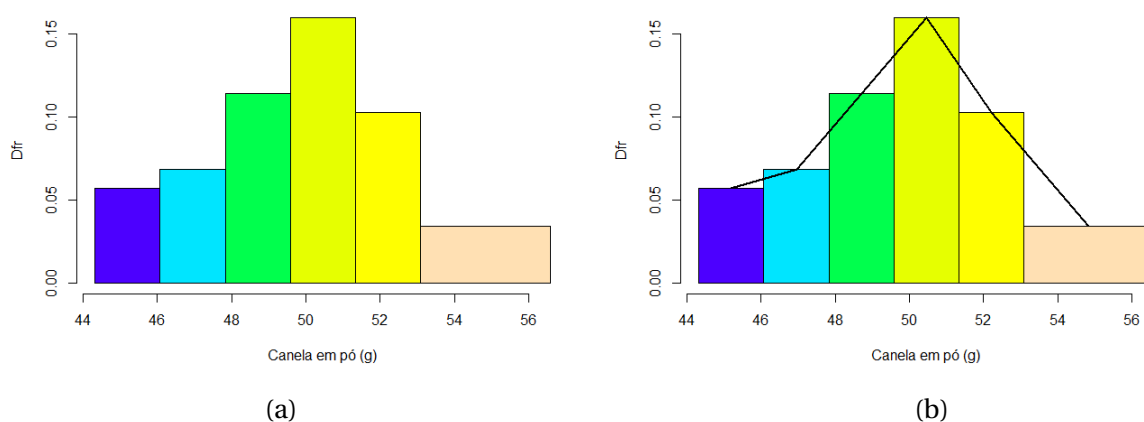


Figura 10 – (a) Histograma do peso de potinhos de canela em pó em uma linha de produção. (b) O mesmo histograma com polígono de frequência.

Fonte: Do autor.

A rotina para construir histogramas usa a função `hist()` do R. Nela, atributos como a densidade de frequência relativa e cores das colunas podem ser facilmente modificados.

Figura 10(a):

```
h<-hist(canela, breaks=c(44.33,46.08,47.83,49.58,51.33,53.08,56.58),
freq=FALSE, ylab="Dfr", xlab="Canela_em_pó_(g)", main="",
col=topo.colors(6))
```

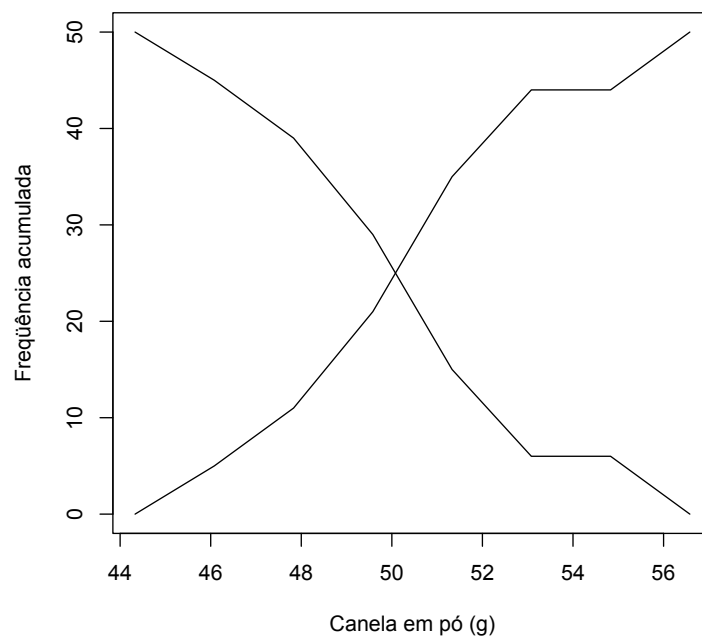
Figura 10(b) (além da rotina para a Figura 10(a)):

```
h<-hist(canela, breaks=c(44.33,46.08,47.83,49.58,51.33,53.08,56.58),
freq=FALSE, ylab="Dfr", xlab="Canela_em_pó_(g)", main="",
col=topo.colors(6))
points(h$mids, h$density, "l", lwd=2)
```

Tabela 4 – frequências absolutas acumuladas abaixo (Fa_{\downarrow}) e acima de (Fa_{\uparrow}).

Limite de classe (g)	Fa_{\downarrow}	Fa_{\uparrow}
44,33	0	50
46,08	5	45
47,83	11	39
49,58	21	29
51,33	35	15
53,08	44	6
54,83	44	6
56,58	50	0

Fonte: Dados fictícios.

Figura 11 – Ogivas representando as frequências absolutas acumuladas *acima de* e *abaixo de* e seu respectivo código em R.

Fonte: Do autor.

A seguir, a rotina usada para construir a Figura 11.

```
lim<-c(44.33,46.08,47.83,49.58,51.33,53.08,54.83,56.58)
ab<-c(0, 5, 11, 21, 35, 44, 44, 50)
ac<-c(50, 45, 39, 29, 15, 6, 6, 0)
plot(lim, ab, "l", ylab='frequência_acumulada', xlab='Canela_em_pó_(g)')
points(lim, ac, "l")
```

3.5 Medidas de posição

Quando se tratam de variáveis quantitativas, os dados podem ser resumidos sob a forma de *distribuições de frequência* ou por *medidas descritivas*. Medidas descritivas são formas de, em um único número, tentar expressar a informação trazida pelos dados.

As duas categorias de medidas descritivas são: *medidas de posição* e *medidas de dispersão*.

As medidas de posição indicam a posição global dos dados na escala de valores possíveis.

3.5.1 Média (Me)

A média aritmética simples, que se trata de somar todas as observações e dividir o resultado por quantos elementos você somou, é um conceito conhecido de todos nós desde a infância. Em Estatística, a média é bastante utilizada. Afinal, é o estimador da média populacional. A média amostral é geralmente denotada por \bar{X} ou \bar{Y} , e a média populacional, denotada por μ .

Dados não agrupados	Dados agrupados
---------------------	-----------------

$$\bar{Y} = \sum_{i=1}^n \frac{Y_i}{n} \qquad \bar{Y} = \sum_{i=1}^k f r_i m_i$$

sendo m_i o ponto central da classe i ,

$$m_i = \frac{LS_i + LI_i}{2}. \tag{3.11}$$

3.5.1.1 Propriedades da Média

Sejam X e Y variáveis aleatórias, e k uma constante.

Propriedade 1: Se $X = Y + k$, então $\bar{X} = \bar{Y} + k$.

Prova algébrica:

$$\begin{aligned} \bar{X} &= \frac{\sum_{i=1}^n X_i}{n} \\ &= \frac{x_1 + x_2 + \dots + x_n}{n} \\ &= \frac{(y_1 + k) + (y_2 + k) + \dots + (y_n + k)}{n} \end{aligned}$$

$$\begin{aligned}
&= \frac{(y_1 + y_2 + \dots + y_n) + (k + \dots + k)}{n} \\
&= \frac{(y_1 + y_2 + \dots + y_n) + nk}{n} \\
&= \frac{\sum_{i=1}^n Y_i}{n} + k \\
&= \bar{Y} + k
\end{aligned}$$

□

Propriedade 2: Se $X = Y \times k$, então $\bar{X} = \bar{Y} \times k$

Prova algébrica:

$$\begin{aligned}
\bar{X} &= \frac{\sum_{i=1}^n X_i}{n} \\
&= \frac{x_1 + x_2 + \dots + x_n}{n} \\
&= \frac{(y_1 k) + (y_2 k) + \dots + (y_n k)}{n} \\
&= \frac{y_1 + y_2 + \dots + y_n}{n} \times k \\
&= \frac{\sum_{i=1}^n Y_i}{n} \times k \\
&= \bar{Y} \times k
\end{aligned}$$

□

Propriedade 3: Seja $e_i = y_i - \bar{y}$ o i -ésimo desvio, então, para $i = 1, \dots, n$, $\sum_{i=1}^n e_i = 0$.

Prova algébrica:

$$\begin{aligned}
\sum_{i=1}^n e_i &= \sum_{i=1}^n (y_i - \bar{y}) \\
&= \sum_{i=1}^n y_i - \sum_{i=1}^n \bar{y} \\
&= \sum_{i=1}^n y_i - \sum_{i=1}^n \left(\sum_{i=1}^n \frac{y_i}{n} \right) \\
&= \sum_{i=1}^n y_i - n \sum_{i=1}^n \frac{y_i}{n} \\
&= \sum_{i=1}^n y_i - \sum_{i=1}^n y_i = 0
\end{aligned}$$

□

3.5.2 Mediana (Md)

É aquele elemento que ocupa a posição central, ou seja, divide a massa de dados em duas partes iguais.

Dados não agrupados	Dados agrupados
$Md(Y) = \begin{cases} Y_{\left(\frac{n+1}{2}\right)}, & \text{se } n \text{ ímpar} \\ \frac{Y_{\left(\frac{n}{2}\right)} + Y_{\left(\frac{n}{2}+1\right)}}{2}, & \text{se } n \text{ par} \end{cases}$	É o valor que separa a área do gráfico em duas partes iguais.

Para calcularmos a mediana em dados não agrupados, esses precisam estar ordenados. Isto é fundamental. Para denotar isso, escrevemos os índices entre parênteses, $Y_{(i)}$, o que significa que aquela é a posição que é ocupada pelo dado na amostra ordenada.

No caso de dados agrupados, a mediana é encontrada somando-se as áreas das colunas até se obter 50%. Caso seja necessário, deve, por geometria, determinar o ponto que, dentro da coluna, garante essa mesma área de cada lado.

Exemplo. Seja $Y = \{3, 5, 6, 8, 9\}$. Então, $Md(Y) = 6$ e $\bar{y} = 6,2$.

3.5.2.1 Propriedade da Mediana

Propriedade 1: Se $X = Y + k$, então, $Md(X) = Md(Y) + k$.

Prova algébrica:

- Se n é ímpar:

$$\begin{aligned} Md(X) &= x_{\frac{n+1}{2}} \\ &= y_{\frac{n+1}{2}} + k \\ &= Md(Y) + k \end{aligned}$$

□

- Se n é par:

$$\begin{aligned}
 Md(X) &= \frac{y_{\frac{n}{2}} + y_{\frac{n}{2}+1}}{2} \\
 &= \frac{x_{\frac{n}{2}} + k + x_{\frac{n}{2}+1} + k}{2} \\
 &= \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1} + 2k}{2} \\
 &= \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} + k \\
 &= Md(Y) + k
 \end{aligned}$$

□

Propriedade 2: Se $X = Y \times k$, então, $Md(X) = Md(Y) \times k$.

Prova algébrica:

- Se n é ímpar:

$$\begin{aligned}
 Md(X) &= x_{\frac{n+1}{2}} \\
 &= y_{\frac{n+1}{2}} \times k \\
 &= Md(Y) \times k
 \end{aligned}$$

□

- Se n é par:

$$\begin{aligned}
 Md(X) &= \frac{y_{\frac{n}{2}} + y_{\frac{n}{2}+1}}{2} \\
 &= \frac{\left(x_{\frac{n}{2}} \times k\right) + \left(x_{\frac{n}{2}+1} \times k\right)}{2} \\
 &= \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} \times k \\
 &= Md(Y) \times k
 \end{aligned}$$

□

3.5.3 Moda (Mo)

É o valor mais frequente, aquele que mais se repete.

Variáveis discretas

Verifica-se o valor que mais se repete.

Variáveis contínuas

Aconselha-se trabalhar com dados agrupados, pelo método de Czuber.

3.5.3.1 Método de Czuber

O método de Czuber permite calcular a moda amostral em dados agrupados. Como era de se esperar, a moda estará contida na classe mais frequente ou, no histograma, a coluna mais alta. Essa classe recebe o nome de classe modal. Dentro da classe modal a moda se situará mais próxima àquela classe adjacente que for, consecutivamente, mais alta. Analise a fórmula e entenda sua lógica no histograma ilustrativo da Figura 12.

$$Mo(Y) = LI_M + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times c_M, \quad (3.12)$$

sendo $\Delta_1 = Dfr_M - Dfr_{M-1}$ e $\Delta_2 = Dfr_M - Dfr_{M+1}$; LI_M é o limite inferior da classe modal; Dfr_M é a densidade de frequência relativa da classe modal; Dfr_{M-1} é a densidade de frequência relativa da classe anterior à modal; Dfr_{M+1} é a densidade de frequência relativa da classe posterior à modal; c_M : amplitude da classe modal.

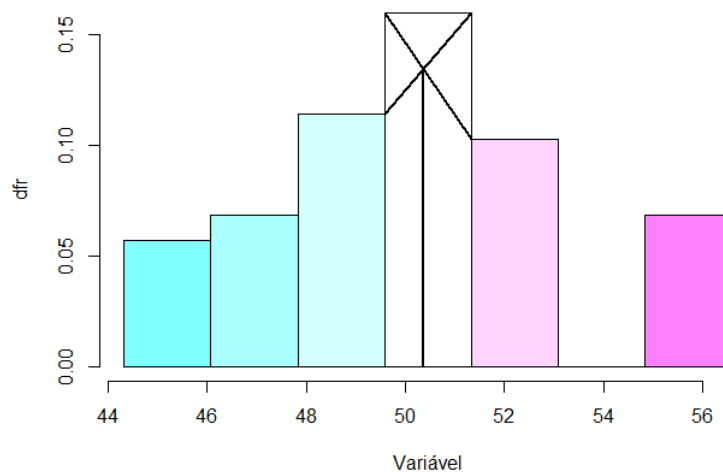


Figura 12 – Histograma ilustrando geometricamente como se dá o cálculo da moda por meio do método de Czuber.

Fonte: Do autor.

Esta é a rotina usada para fazer o gráfico da Figura 12:

```
h<-hist(canela, breaks=c(44.33,46.08,47.83,49.58,51.33,53.08,54.83,
56.58), main='', freq=FALSE, ylab='dfr', xlab='Variável',
col=cm.colors(7))
points(c(49.58,51.33), c(h$density[3], h$density[4]), "l")
points(c(49.58,51.33), c(h$density[4], h$density[5]), "l")
points(c(50.357,50.357), c(0.134,0), "l")
```

3.5.3.2 Propriedades da Moda

Propriedade 1: Se $X = Y + k$, então $Mo(X) = Mo(Y) + k$.

Prova algébrica:

$$\begin{aligned} Mo(X) &= x \text{ mais freqüente.} \\ &= y \text{ mais freqüente} + k \\ &= Mo(Y) + k \end{aligned}$$

Propriedade 2: Se $X = Y \times k$, então, $Mo(X) = Mo(Y) \times k$.

Prova algébrica:

$$\begin{aligned} Mo(X) &= x \text{ mais freqüente.} \\ &= y \text{ mais freqüente} \times k \\ &= Mo(Y) \times k \end{aligned}$$

Das medidas de posição apresentadas, apenas a moda não se encontra implementada nos comandos básicos do R.

```
ex<-c(3,5,6,8,9)
mean(ex)           #média
median(ex)         #mediana
```

3.5.4 Influência da assimetria nas medidas de posição

As medidas de posição podem ser influenciadas pelo grau de assimetria da distribuição populacional dos dados, e conseqüente assimetria verificada na amostra. Essa influência pode não ser boa, pois pode deixar a medida de posição pouco representativa do conjunto de dados.

Imagine um exemplo extremo e artificial em que os dados disponíveis são: $x = \{1, 2, 2, 1, 1, 1, 2, 99\}$. Temos que $\bar{x} = 13,6$. Uma pessoa que não está vendo o conjunto de dados, mas foi informada que a média vale 13,6, terá uma impressão completamente deturpada. Afinal, nessa amostra existem vários uns e dois, um noventa e nove, mas nenhum número parecido com 13. Essa amostra não está “errada”. O fato de apresentar um dado discrepante (o valor 99) representa a característica de um fenômeno com distribuição acentuadamente assimétrica à direita. Isto é, um fenômeno em que valores muito discrepantes (neste caso, valores muito altos) podem ocorrer, mesmo que raramente. Um deles (o valor 99) aconteceu no exemplo apresentado.

Em distribuições assimétricas à direita, observamos que a média apresenta valor maior que a mediana, que por sua vez apresenta valor maior que a moda (Figura 14). O contrário acontece quando a assimetria é negativa (Figura 13). Mas note que a média é sempre a mais sensível. Por outro lado, em distribuições simétricas, média, moda e mediana geralmente são equivalentes (Figura 15).

Portanto, dizemos que essa é a ordem da influência sofrida pelas medidas de posição na presença de assimetria: a *média* é mais influenciada que a *mediana*, que é mais influenciada que a *moda*.

Essa informação pode ser levada em conta na hora de escolher a medida de posição mais adequada para cada caso. Objetivamente, quando não há valores discrepantes recomenda-se que a média seja escolhida por ser mais intuitiva. Em caso de haver valores extremos ou discrepantes, devemos

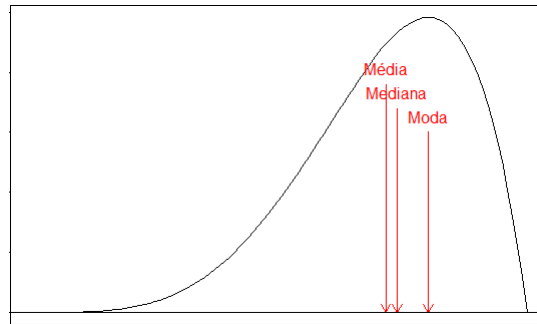


Figura 13 – Representação de uma distribuição contínua assimétrica à direita, posicionando média, moda e mediana.

Fonte: Do autor.

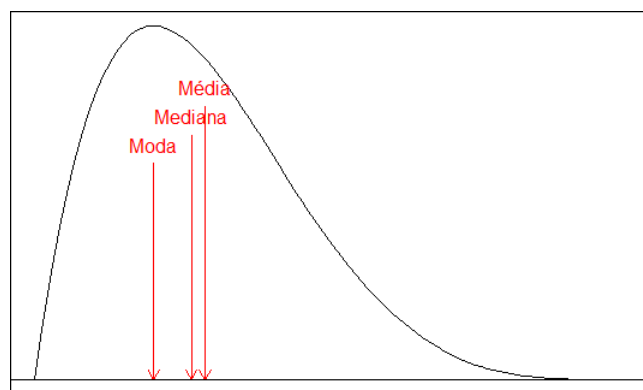


Figura 14 – Representação de uma distribuição contínua assimétrica à esquerda, posicionando média, moda e mediana.

Fonte: Do autor.

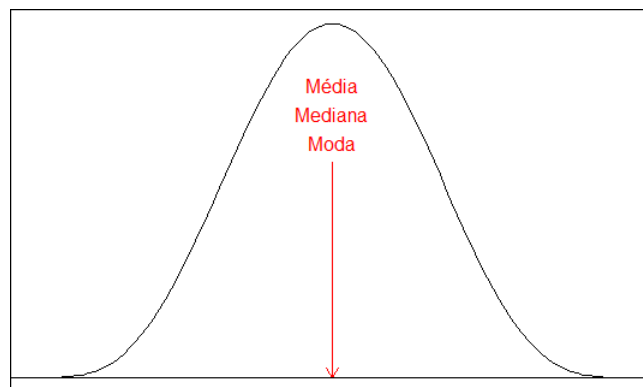


Figura 15 – Representação de uma distribuição contínua simétrica, posicionando média, moda e mediana.

Fonte: Do autor.

dar preferência à mediana ou à moda. Como a moda nem sempre existe em dados não agrupados, e nem sempre é única, a mediana deve ser escolhida, nesses casos.

3.6 Medidas de dispersão

As medidas de dispersão indicam quanto os dados variam. Considere o seguinte exemplo: imagine três situações distintas em que certa variável X é medida em quatro observações (quatro elementos em cada amostra). Os três conjuntos de dados resultantes estão explicitados na tabela a seguir. Caso você aplique, nesses conjuntos de dados, as medidas de posição conhecidas até o momento, o resultado será exatamente o mesmo, sugerindo que as três massas de dados são iguais. Mas isso claramente não é verdade! Nessa situação, o que fazer?

Tabela 5 – Três conjuntos de dados diferentes, que não podem ser diferenciados pelas medidas de posição

Observação	I	II	III
1	100	80	10
2	100	100	100
3	100	100	100
4	100	120	190
\bar{x}	100	100	100
$Md(x)$	100	100	100
$Mo(x)$	100	100	100

Fica claro que as medidas de posição, por si só, não são suficientes para descrever um conjunto de dados. No exemplo, os três conjuntos de dados diferem quanto à *variabilidade*. Por exemplo, o conjunto III varia muito mais que os outros dois. Aí está evidenciada a importância das medidas de dispersão ou variabilidade.

3.6.1 Amplitude (A)

Amplitude total (ou simplesmente Amplitude), como já mencionando na construção de histogramas, é o intervalo total de variação dos dados.

$$A = MVO - mvo$$

No exemplo,

	I	II	III
Amplitude	$100 - 100 = 0$	$120 - 80 = 40$	$190 - 10 = 180$

Os conjuntos já começam a mostrar suas diferenças. Mas há uma desvantagem: *amplitudes só podem ser comparadas se os conjuntos tiverem o mesmo número de dados*. É intuitivo que se dois conjuntos apresentam números de elementos diferentes, o conjunto maior tem mais chance de ter uma amplitude também maior. Nesse caso, a diferença entre as amplitudes dos conjuntos refletiria a diferença no número de elementos, e não a variabilidade dos dados. Além disso, essa é uma medida de dispersão limitada, pois só leva em conta os valores extremos.

Considere agora os conjuntos de dados:

I	5	15	15	15	40	$A_I = 35$
II	5	10	20	30	40	$A_{II} = 35$

Os conjuntos são diferentes, apresentam variabilidades diferentes, porém a amplitude não conseguiu detectar esse fato.

3.6.1.1 Propriedades da Amplitude

Propriedade 1: Se $X = Y + k$, então, $A(X) = A(Y)$.

Prova algébrica:

$$\begin{aligned}
 A(X) &= MVO_x - mvo_x \\
 &= (MVO_y + k) - (mvo_y + k) \\
 &= MVO_y - mvo_y = A(Y)
 \end{aligned}$$

□

Propriedade 2: Se $X = Y \times k$, então, $A(X) = A(Y) \times k$.

Prova algébrica:

$$\begin{aligned}
 A(X) &= MVO_x - mvo_x \\
 &= (MVO_y \times k) - (mvo_y \times k) \\
 &= (MVO_y - mvo_y) \times k \\
 &= A(Y) \times k
 \end{aligned}$$

□

3.6.2 Variância

A variância e o desvio padrão são as duas medidas de dispersão mais usadas. Elas são grandezas proporcionais e por isso serão tratadas em um mesmo tópico.

Ambas se valem de todas as observações para calcular suas quantidades e se baseiam no desvio em relação à média

$$e_i = y_i - \bar{y}.$$

Notação: normalmente, a variância da população é designada pela letra grega sigma minúsculo ao quadrado (σ^2); e a variância da amostra, pela letra s maiúsculo ao quadrado (S^2), quando se tratar da variável aleatória; e s minúsculo ao quadrado (s^2), quando se tratar de uma estimativa.

Se X é uma variável aleatória, $V(X)$ também denota a Variância de X .

Censos	Amostras	Dados agrupados
$\sigma^2 = \frac{\sum_{i=1}^n e_i^2}{N}$	$S^2 = \frac{\sum_{i=1}^n e_i^2}{n-1}$	$\sigma^2 = \frac{\sum_{i=1}^k f r_i (m_i - \bar{y})^2}{n-1}$

sendo m_i o ponto médio da classe i ($i = 1, 2, \dots, k$).

No exemplo,

	I	II	III
Variância	0	266,67	5400

Nesse caso, a variância identificou a diferença na variabilidade dos conjuntos. Porém, seu valor absoluto não é interpretável de forma prática porque ela se expressa no quadrado da unidade dos dados. Por exemplo, o peso de um grupo de bovinos alimentados com certa ração varia $266,67kg^2$.

3.6.2.1 Propriedades da Variância

Propriedade 1: Se $X = Y + k$, então, $V(X) = V(Y)$.

Prova algébrica:

$$\begin{aligned}
 V(X) &= \frac{\sum_{i=1}^n e_i^2}{n-1} \\
 &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \\
 &= \frac{\sum_{i=1}^n (y_i + k - (\bar{y} + k))^2}{n-1}
 \end{aligned}$$

$$= \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1} = V(Y)$$

□

Propriedade 2: Se $X = Y \times k$, então, $V(X) = V(Y) \times k^2$.

Prova algébrica:

$$\begin{aligned}
 V(X) &= \frac{\sum_{i=1}^n e_i^2}{n-1} \\
 &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \\
 &= \frac{\sum_{i=1}^n (y_i \times k - (\bar{y} \times k))^2}{n-1} \\
 &= \frac{\sum_{i=1}^n ((y_i - \bar{y}) \times k)^2}{n-1} \\
 &= \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1} \times k^2 \\
 &= V(Y) \times k^2
 \end{aligned}$$

□

3.6.3 Desvio-padrão

Notação: O desvio padrão recebe a mesma notação que a variância, porém, sem o quadrado. Ou seja, o desvio padrão da população é designado por σ ; e o desvio padrão da amostra, por S (variável aleatória) ou s (estimativa de S).

Se X é uma variável aleatória, $DP(X)$ também denota o Desvio Padrão de X .

$$\begin{array}{cc}
 \text{Censos} & \text{Amostras} \\
 \sigma = \sqrt{\sigma^2} & S = \sqrt{S^2}
 \end{array}$$

No exemplo,

	I	II	III
Desvio Padrão	0	16,33	73,48

Além de utilizar todos os dados para computar sua medida de variabilidade, o desvio padrão ainda retorna um valor expresso na unidade dos dados, o que o torna mais facilmente interpretável. Por exemplo, quando se diz que o peso de bovinos alimentados com certa ração costuma variar 16,33kg ao redor da sua média de peso, o leitor consegue ter uma ideia prática da variação.

3.6.3.1 Propriedades do Desvio-Padrão

Propriedade 1: Se $X = Y + k$, então, $DP(X) = DP(Y)$.

Prova algébrica:

$$\begin{aligned} DP(X) &= \sqrt{V(X)} \\ &= \sqrt{V(Y)} \\ &= DP(Y) \end{aligned}$$

□

Propriedade 2: Se $X = Y \times k$, então, $DP(X) = DP(Y) \times k$.

Prova algébrica:

$$\begin{aligned} DP(X) &= \sqrt{V(X)} \\ &= \sqrt{V(Y) \times k^2} \\ &= DP(Y) \times k \end{aligned}$$

□

3.6.4 Coeficiente de Variação (CV)

O Coeficiente de variação é uma medida de variabilidade padronizada, ou seja, expressa percentualmente a variação dos dados em relação à média.

Censos	Amostras
$CV(\%) = \frac{\sigma}{\mu}$	$CV(\%) = \frac{S}{\bar{X}}$

Sua grande vantagem é permitir a comparação de grandezas diferentes, que estão em unidades diferentes (por exemplo: o que é mais variável, o ganho de peso de suínos ou a altura de plantas de milho?).

Por outro lado, ele possui sérias restrições de uso e inspira cuidados. Primeiro, quando a média da variável aleatória em questão tende a zero, o CV tende ao infinito (o que não faz sentido prático). Segundo, de acordo com as propriedades da média e do desvio padrão, a adição de uma constante às observações altera a média da nova variável aleatória, mas não altera seu desvio padrão, ou seja, por meio de algumas transformações de variáveis o CV pode ser criminosamente manipulado.

No exemplo,

	I	II	III
CV(%)	0	16,33	73,48

Nesse caso a interpretação se torna ainda mais imediata, porém não podemos nos esquecer das ressalvas feitas anteriormente.

3.6.4.1 Propriedades do Coeficiente de Variação

Propriedade 1: Se $X = Y + k$, então, $\begin{cases} CV(X) < CV(Y), & \text{se } k > 0 \\ CV(X) > CV(Y), & \text{se } k < 0 \end{cases}$

Prova algébrica:

$$\begin{aligned} CV(X) &= \frac{DP(X)}{\bar{X}} \\ &= \frac{DP(Y+k)}{\bar{Y+k}} \\ &= \frac{DP(Y)}{\bar{Y}+k} \end{aligned}$$

Se $k > 0$, então $\frac{DP(Y)}{\bar{Y}+k} < \frac{DP(Y)}{\bar{Y}}$
 Se $k < 0$, então $\frac{DP(Y)}{\bar{Y}+k} > \frac{DP(Y)}{\bar{Y}}$

□

Propriedade 2: Se $X = Y \times k$, então, $CV(X) = CV(Y)$.

Prova algébrica:

$$\begin{aligned} CV(X) &= \frac{DP(X)}{\bar{X}} \\ &= \frac{DP(Y \times k)}{\bar{Y \times k}} \\ &= \frac{DP(Y) \times k}{\bar{Y} \times k} \\ &= \frac{DP(Y)}{\bar{Y}} = CV(Y) \end{aligned}$$

□

As medidas de dispersão encontradas de forma direta no R são a variância e o desvio padrão, porém a amplitude e o CV podem ser implementados facilmente, como segue:

```
c1<-c(100,100,100,100)
c2<-c(80,100,100,120)
c3<-c(10,100,100,190)
Ac1<-range(c1)[2] - range(c1)[1] #Amplitude
var(c2)                          #Variância
sd(c2)                            #Desvio padrão
CVc3<-sd(c3)/mean(c3)*100        #Coeficiente de variação
```

PROBABILIDADE

Para todos nós, noções de probabilidade ou respostas intuitivas a questões de probabilidade são comuns desde a mais tenra idade. Qualquer pessoa, por menos conhecimento estatístico que tenha, é capaz de responder à pergunta: qual a probabilidade de se retirar uma carta de ouros de um baralho honesto?

Fazemos intuitivamente: “se um baralho honesto tem 13 cartas do naipe ouros e o total de cartas é 52, então a chance de uma carta de ouros ser tirada ao acaso é $\frac{13}{52} = \frac{1}{4} = 1 : 4 = 0,25$ ou 25%”.

4.1 Algumas definições úteis

Probabilidade é um valor entre 0 e 1 que representa a frequência relativa de ocorrência de um evento, ao longo de infinitas repetições de um experimento aleatório.

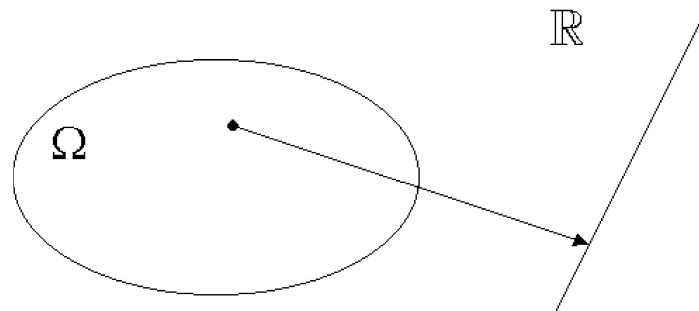
Evento é um resultado ou conjunto de resultados possíveis em um experimento aleatório.


Experimento aleatório é o procedimento para o qual não se consegue prever, exatamente, o resultado. Ex: Vai chover amanhã, aqui, às 16h?

Evento certo é o evento que possui probabilidade 100% de ocorrer. Ex: Sair cara ou coroa num lançamento de uma moeda.

Evento impossível é aquele que possui probabilidade 0 de ocorrer. Ex: Sair 7 no lançamento de um dado honesto.

Variável aleatória é a função que associa um valor na reta real a cada ponto do espaço amostral.



Ex.: No lançamento de um dado honesto, o espaço amostral Ω é composto por suas seis diferentes faces. A cada uma delas está associado um número natural de 1 a 6. Suponha que a variável aleatória X descreve o resultado desse lançamento e o dado cai com face  virada para cima. Nesse caso, associamos um valor a este evento, ou seja, $x = 4$.

Distribuição de probabilidade é uma função que descreve a probabilidade de ocorrência dos possíveis resultados de uma variável aleatória.

Parâmetro de uma distribuição é a constante que determina (estabelece) completamente uma distribuição de probabilidade.

Espaço amostral é o conjunto de todos os resultados possíveis de serem observados num dado fenômeno, ou seja, todos os resultados possíveis de um experimento. É representado simbolicamente pela letra grega Omega maiúscula Ω . Ex: $\Omega = \{\text{cara, coroa}\}$.

4.2 Axiomas da Probabilidade

1. Se A é um evento pertencente a Ω , então $P(A) > 0$.
2. $P(\Omega) = 1$.
3. Sejam A_1, A_2, A_3, \dots eventos disjuntos pertencentes a Ω (intersecção nula), então $P(A_1 \cup A_2 \cup A_3 \cup \dots) = P(A_1) + P(A_2) + P(A_3) + \dots$

4.2.1 Propriedades derivadas dos axiomas

1. $P(\bar{A}) = 1 - P(A)$.
2. $P(\emptyset) = 0$.
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
4. $0 \leq P(A) \leq 1$.

4.3 Definição clássica de probabilidade

Formalizando o que foi feito intuitivamente, temos

$$P(A) = \frac{|A|}{|\Omega|} \quad (4.1)$$

em que $|\cdot|$ é a cardinalidade de um conjunto¹; A é o conjunto que contém os resultados de nosso interesse; e Ω é o espaço amostral.

Essa divisão intuitiva, do número de resultados de interesse pelo número de resultados possíveis, deve ser feita com cuidado. Antes de mais nada, cheque se seu espaço amostral é finito e equiprovável.

No exemplo do baralho, essas condições estão satisfeitas. Existem 52 resultados possíveis, pois $\Omega = \{x|x \text{ é uma carta do baralho}\}$. Além disso, cada um deles tem probabilidade $1/52$ de acontecer. O conjunto que estamos interessados é $A = \{a|a \text{ é uma carta do naipe de ouros}\}$. Esse conjunto tem cardinalidade 13, pois ele contém 13 elementos². Sendo assim, a divisão pode ser feita, e a resposta 25% está correta.

4.4 Definição frequentista de probabilidade

A probabilidade de um evento A acontecer é o resultado da razão entre sua frequência de ocorrência e o número de ensaios (n), quando o número de ensaios cresce indefinidamente:

$$P(A) = \frac{f_A}{n}. \quad (4.2)$$

Em essência, uma probabilidade é uma frequência relativa, em qualquer caso. Até mesmo nos casos onde a definição clássica pode ser aplicada, a definição frequentista poderia ser examinada apenas com o prejuízo de ser mais trabalhosa. Por exemplo, poderíamos retirar uma carta do baralho, examinar se era de ouros, anotar o resultado e retorná-la ao baralho. Esse processo poderia ser repetido um número grande de vezes. O resultado seria, sucessivamente, cada vez mais parecido com 25%.

A Lei dos Grandes Números é um conceito fundamental em Estatística e Probabilidade que descreve como a média de uma amostra, suficientemente grande e selecionada aleatoriamente, se torna provável de estar perto da média da população.

Jacok Bernoulli, matemático suíço que viveu entre 1654 e 1705, afirmou em seu livro *Ars Conjectandi* que, se um evento de probabilidade p for observado repetidamente ao longo de realizações independentes, a relação da frequência observada desse evento ao número total das repetições converge para p enquanto o número das repetições se torna arbitrariamente grande (BERNOULLI, 1713).

Dizendo com outras palavras, e colocando no contexto da construção de histogramas, pode-se entender que quando $n \rightarrow \infty$, as frequências das classes tendem a se estabilizar.

¹ Cardinalidade é o número de elementos de um conjunto.

² Para uma perfeita compreensão deste capítulo é necessária uma revisão de Teoria dos Conjuntos.

Considere o seguinte exemplo: no lançamento de uma moeda honesta, qual a probabilidade de sair cara? Resp.: 50%. Isto é intuitivo. Não apenas devido à definição clássica, mas também devido à lei dos grandes números.

A Figura 16 traz a simulação de 500 lançamentos de uma moeda honesta, a contagem do número de caras obtidas em cada lançamento, e a plotagem da frequência relativa de caras (número de caras/números de lançamentos). Note que, quanto mais o n (número de lançamentos) aumenta, mais a frequência relativa tende a se estabilizar em 50%, o que corrobora com a afirmação intuitiva anterior.

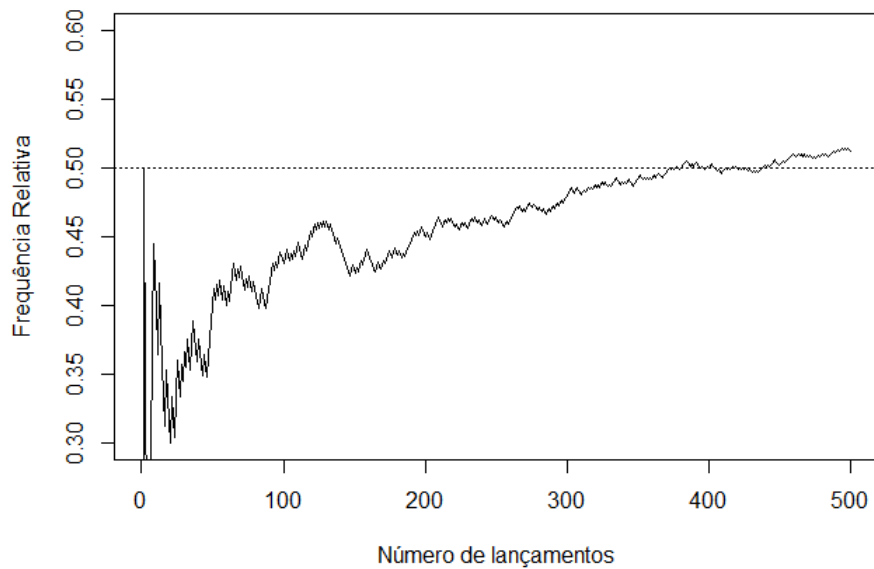


Figura 16 – Simulação do lançamento de uma moeda honesta 500 vezes, comportamento de sua frequência relativa.

Fonte: Do autor.

Rotina R da simulação do lançamento de uma moeda honesta 500 vezes.

```

cara<-0
fr<-vector("numeric",500)
for (i in 1:500) {
moeda<-runif(1,0,1)
if (moeda>0.5) {cara<-cara+1}
fr[i]<-cara/i
}
x<-seq(1:500)
plot(x,fr,"l",xlab="Número_de_lançamentos",ylab="Frequência_Relativa",
ylim=c(0.3,0.6))
abline(h=.5,lty=3)

```

4.5 Regra do “E” e regra do “OU”

Aqui serão apresentadas as regras do “e” e do “ou” apenas para eventos independentes.

4.5.1 Regra do “E”

A probabilidade de ocorrerem dois eventos A e B simultaneamente é

$$P(A \text{ e } B) = P(A) \times P(B) = P(A \cap B). \quad (4.3)$$

Exemplo. No lançamento de 2 dados honestos, qual a probabilidade de se tirar 3 e 5?

$$P(3) \times P(5) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36} = 0,0278 = 2,78\%$$

Utilizando o R como calculado, você pode digitar:

```
1/6 * 1/6
```

4.5.2 Regra do “OU”

A probabilidade de ocorrer o evento A ou o evento B em um experimento é

$$P(A \text{ ou } B) = P(A) + P(B) = P(A \cup B). \quad (4.4)$$

Exemplo. no lançamento de um dado honesto qual a probabilidade de se tirar 3 ou 5?

$$P(3) + P(5) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = 0,3333 = 33,33\%$$

Utilizando o R como calculado, você pode digitar:

```
1/6 + 1/6
```

4.6 Probabilidade condicional

Suponha dois eventos A e B *dependentes*. Como A e B são dependentes, a probabilidade de A depende da ocorrência (ou não ocorrência) de B . Para calcularmos a probabilidade de A ocorrer usaremos a fórmula de Bayes:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad (4.5)$$

em que $P(A|B)$ é a probabilidade de A ocorrer *dado* que B ocorreu; e $P(A \cap B)$ é a probabilidade de A e B ocorrerem ao mesmo tempo. Note que, aqui, pelo fato de A e B não serem independentes,

$$P(A \cap B) \neq P(A)P(B).$$

Exemplo. Carlos, quando almoça fora, tem 40% de chance de escolher uma churrascaria. Quando escolhe ir a uma churrascaria, tem 85% de chance de sair da dieta. Entretanto, mesmo quando Carlos não vai à churrascaria, tem 45% de chance de sair da dieta. Carlos quebrou a dieta. Qual é a probabilidade de ele ter ido à churrascaria.

Resolução. A primeira coisa a fazer é atribuir quais são os eventos A e B . O que sabemos, chamamos de B . O que queremos saber, chamamos de A . Portanto, A : ir à churrascaria, e B : sair da dieta.

Em segundo lugar, aconselhamos a fazer a árvore decisória, que contém os possíveis resultados e suas probabilidades de ocorrência:

	Sair da dieta (0,85)	$\Rightarrow 0,4 \times 0,85 =$	0,34
Ir à churrascaria (0,40)			
	Não sair da dieta (0,15)	$\Rightarrow 0,4 \times 0,15 =$	0,06
<hr/>			
	Sair da dieta (0,45)	$\Rightarrow 0,6 \times 0,45 =$	0,27
Não ir à churrascaria (0,60)			
	Não sair da dieta (0,15)	$\Rightarrow 0,6 \times 0,55 =$	0,33
			1,00

Agora fica bem mais claro quais as probabilidades precisamos para realizar o cálculo. A e B acontecem simultaneamente com 34% de probabilidade. Já a probabilidade de B é dada por $0,34 + 0,27$, pois Carlos pode sair da dieta quando vai OU quando não vai à churrascaria.

Portanto, a probabilidade de Carlos ter ido à churrascaria é:

$$P(A|B) = \frac{0,34}{0,34 + 0,27} = \frac{0,34}{0,61} = 0,5574 = 55,74\%.$$

4.7 Distribuições de probabilidade discretas

Se caracteriza pela função $f(X)$, em que X é uma variável aleatória discreta.

Análogo à variável aleatória contínua, se X assume k valores, então:

$$\sum_{i=1}^k P[X = x_i] = 1.$$

4.7.1 Definições úteis

Esperança matemática de X é o valor médio esperado para infinitas realizações da variável aleatória discreta X .

$$E[X] = Me(X) = \sum_{i=1}^k x_i P[X = x_i]$$

Variância de X é uma medida da variabilidade das infinitas realizações da variável aleatória discreta X .

$$\sigma_X^2 = \sum_{i=1}^k [(X - E[X])^2 \times P[X = x_i]]$$

4.7.2 Distribuição Binomial

A distribuição Binomial de probabilidades é uma distribuição discreta que se caracteriza pela seguinte função densidade:

$$P[X = x] = C_{n,x} p^x q^{n-x}, \quad (4.6)$$

em que,

$$C_{n,x} = \frac{n!}{x!(n-x)!}. \quad (4.7)$$

Os parâmetros da Binomial são: n (número de eventos) e p (probabilidade de sucesso). q não é considerado um parâmetro porque ele é função de p , $q = 1 - p$.

A Binomial possui quatro características marcantes:

- (a) Ser uma soma de ensaios de Bernoulli, ou seja, de ensaios que possuem apenas dois resultados possíveis (sucesso ou fracasso).
- (b) As realizações desses ensaios são eventos independentes.
- (c) A probabilidade de sucesso (p) é constante ao longo dos ensaios.
- (d) O número de ensaios é finito.

A esperança matemática (média) e a variância da distribuição Binomial são funções de seus dois parâmetros, n e p , a saber:

$$E[X] = np; \quad (4.8)$$

$$V(X) = npq. \quad (4.9)$$

Exemplo. Em uma ninhada de aves, nove ovos foram chocados. Qual a probabilidade de nascerem sete machos?

Resolução. Sabe-se que a probabilidade de um filhote ser macho é 50% ($p = 0,5$); o número de ovos nessa ninhada é 9 ($n = 9$); e a variável aleatória (X), desse experimento, representa o número de filhotes machos. Além disso, queremos descobrir a probabilidade dessa variável assumir o valor sete, ou seja, de nascerem sete machos nessa ninhada ($x = 7$); então

$$P[X = 7] = C_{9,7} p^7 q^{9-7} = 36 \times 0,5^7 \times 0,5^2 \cong 7\%.$$

Para resolver esse problema no R, basta digitar:

```
dbinom(7, 9, 0.5)
```

De forma semelhante, toda a distribuição de probabilidades dessa situação pode ser calculada e plotada em um gráfico de agulhas:

```
x<-0:9
p<-dbinom(x, 9, 0.5)
plot(x, p, "h", ylab="P[X=x]")
axis(1, 0:9, 0:9)
```

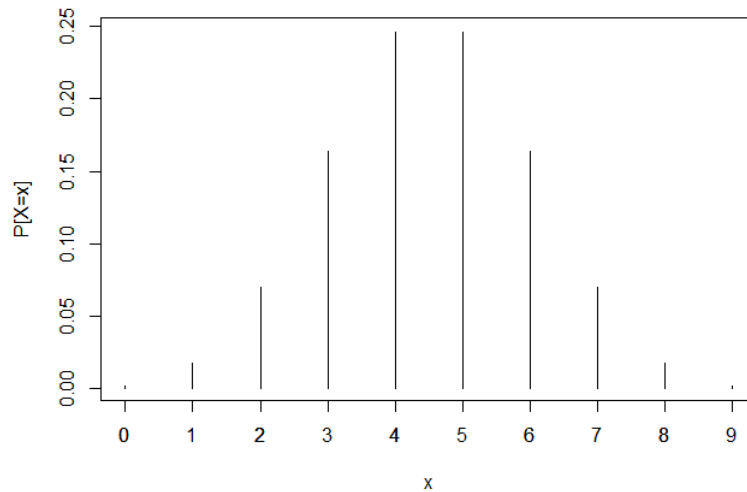


Figura 17 – Gráfico da distribuição de probabilidade binomial, ilustrando os casos de nascerem x machos em 9 ovos.

Fonte: Do autor.

4.7.3 Distribuição Poisson

Distribuição discreta de probabilidades caracterizada pela seguinte função de probabilidade:

$$P[X = x] = \frac{e^{-\lambda} \lambda^x}{x!}. \quad (4.10)$$

Neste livro, a distribuição de Poisson será usada apenas como uma aproximação a distribuição Binomial, quando n é muito grande e p é muito pequeno. Portanto, vale ressaltar que, além das características da distribuição Binomial, a Poisson apresenta as seguintes particularidades:

- Pode descrever eventos raros.
- A variável discreta assume apenas valores inteiros positivos ($X = 0, 1, \dots$).
- $n > 50$ e $p < 0,1$.

O parâmetro da distribuição Poisson é o λ . Note, pelas fórmulas a seguir, que λ também é a média dessa distribuição.

$$E[X] = V[X] = \lambda; \quad (4.11)$$

$$\lambda = np. \quad (4.12)$$

Exemplo. Supondo que o número médio de ocorrências de chuvas fortes seja de 1,5 chuvas por ano, qual a probabilidade de, em um dado ano, não haver nenhuma chuva forte? E apenas uma? E duas? Faça a distribuição de probabilidade dessa variável até 6 chuvas intensas.

Resolução.

$$P[X = 0] = \frac{e^{-1,5} 1,5^0}{0!} = 0,2231.$$

$$P[X = 1] = \frac{e^{-1,5} 1,5^1}{1!} = 0,3347.$$

$$\vdots$$

$$P[X = 5] = \frac{e^{-1,5} 1,5^5}{5!} = 0,0141.$$

$$P[X = 6] = \frac{e^{-1,5} 1,5^6}{6!} = 0,0035.$$

Portanto, temos a seguinte distribuição de frequências:

X	0	1	2	3	4	5	6
$P[X = x]$	0,2231	0,3347	0,2510	0,1255	0,0421	0,0142	0,0035

No R é ainda mais fácil calcular essas probabilidades e plotar o gráfico de agulhas:

```
x<-0:6
p<-dpois(x, 1.5)
plot(x, p, "h", ylab="P[X=x]")
```

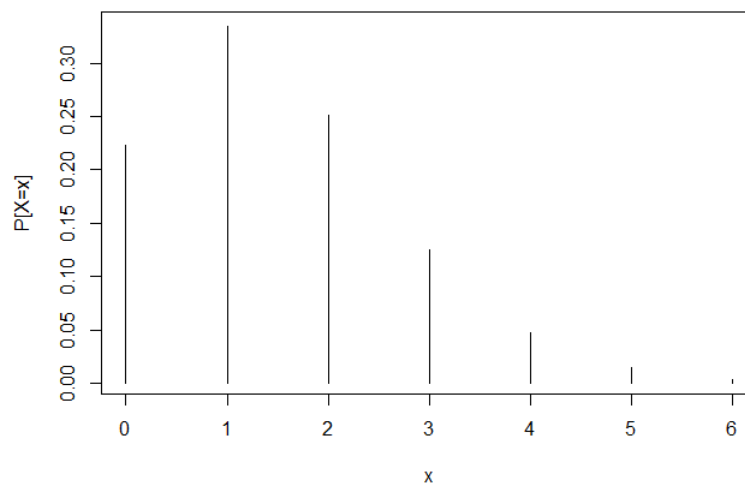


Figura 18 – Gráfico da distribuição de probabilidade Poisson, ilustrando os casos de haverem x chuvas fortes por ano.

Fonte: Do autor.

4.8 Distribuições de probabilidades contínuas

Pode-se entender distribuições contínuas de probabilidade como generalizações de histogramas construídos para grandes tamanhos amostrais ($n \rightarrow \infty$).

Vamos entender o que acontece quando o tamanho da amostra cresce. Sejam as seguintes regras de determinação da amplitude e número de classes de um histograma convencional:

$$k = 5 \log n \quad c = \frac{A}{k-1}.$$

Ilustrativamente, a Figura 19 sugere como podemos entender uma distribuição contínua. A figura sugere que, quando o tamanho da amostra (n) aumenta, aumenta também o número de classes (colunas do histograma) e diminui a amplitude de classe (largura da base da coluna). No limite, temos então uma distribuição “suavizada” ou contínua. É como se o número de colunas fosse tão grande, e as colunas tão finas, que apenas um *continuum* pudesse ser visto.

É essa a representação que fazemos de uma população infinita com distribuição contínua subjacente, ou simplesmente, distribuição de probabilidade de uma variável aleatória contínua.

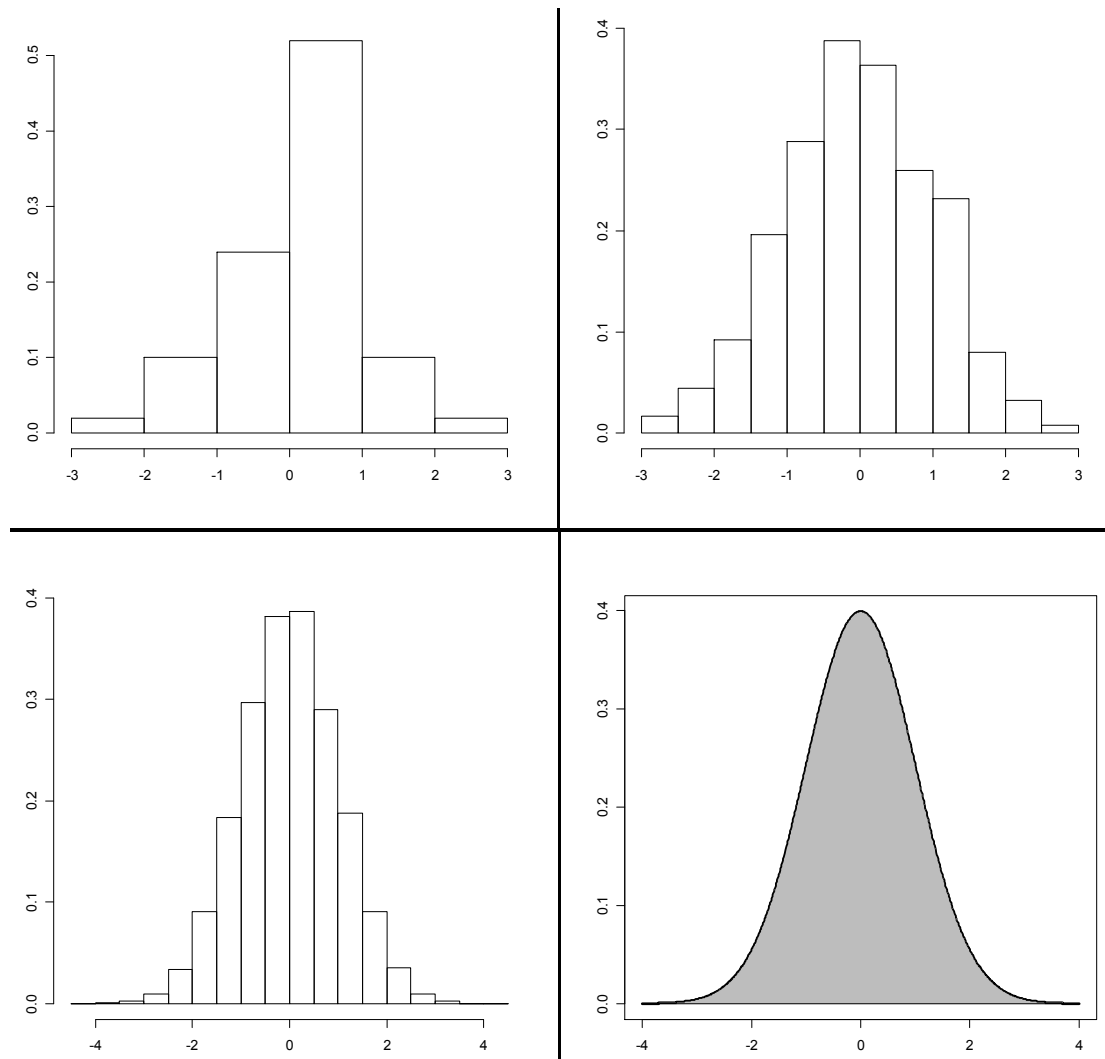


Figura 19 – Esquema da generalização teórica de histogramas para funções densidade de probabilidade, quando $n \rightarrow \infty$.

Fonte: Do autor.

Veja a rotina utilizada para a construção dos gráficos da Figura 19.

```
x1 <- rnorm(50, 0, 1)
hist(x1, freq=FALSE, xlab="", ylab="", main="")
x2 <- rnorm(500, 0, 1)
hist(x2, freq=FALSE, xlab="", ylab="", main="")
x3 <- rnorm(50000, 0, 1)
```

```
hist(x3, freq=FALSE, xlab="", ylab="", main="")
x4<-seq(-4, 4, by=.01)
y<-dnorm(x4, mean=0, sd=1, log = FALSE)
plot(x4, y, "l", lwd=3, xlab="", ylab="", main="")
polygon(x4, y, col = "gray")
```

4.9 Função Densidade de Probabilidade (fdp)

Função densidade de probabilidade (fdp) é a função que descreve a probabilidade de uma variável aleatória associada a uma população infinita, onde a área representa a probabilidade e, a altura, a densidade de probabilidade.

4.9.1 Propriedades da fdp

Propriedade 1: A área total abaixo da curva é igual a 1,

$$\int_{-\infty}^{+\infty} f(x)dx = 1.$$

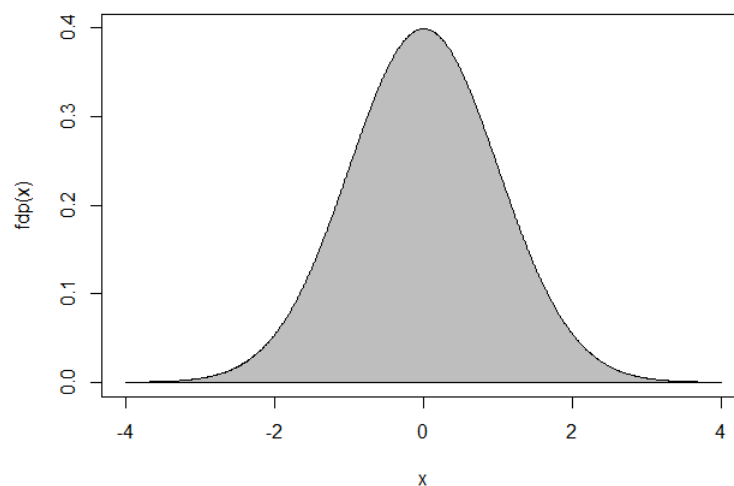


Figura 20 – Representação da integral de uma função densidade de probabilidade, como área total abaixo da curva.

Fonte: Do autor.

A rotina a seguir mostra como foi feita a Figura 20.

```
x <- seq(-4, 4, by=.01)
y<-dnorm(x, mean=0, sd=1, log = FALSE)
rx<-rev(x)
ry<-vector("numeric", length(rx))
x<-c(x, rx)
y<-c(y, ry)
plot(x, y, "l", xlab="x", ylab="fdp(x)")
polygon(x, y, col = "gray")
```

Propriedade 2: Não existe probabilidade negativa.

$$f(x_0) \geq 0 \quad \forall x_0 \in D(X)$$

4.9.2 Distribuição Normal de probabilidades

Definição: uma variável aleatória contínua tem distribuição Normal se ela segue a seguinte função densidade de probabilidade:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}. \quad (4.13)$$

Notação³:

$$N(\mu, \sigma^2).$$

Portanto, a média e a variância são os dois parâmetros da distribuição Normal, sendo μ o parâmetro de locação e σ^2 o parâmetro de forma.

$$Me(X) = \mu; \quad (4.14)$$

$$V(X) = \sigma^2. \quad (4.15)$$

4.9.2.1 Propriedades da Normal

1. É simétrica.
2. Tem forma de sino.
3. Está definida de $-\infty$ a $+\infty$, com caudas assintóticas ao eixo X.
4. Apresenta a particularidade: $Me = Mo = Md$.
5. Dois parâmetros: De forma: σ^2 e de locação μ .

Exemplo. A velocidade de veículos em uma rodovia segue uma distribuição Normal com média 60km/h e variância $400(\text{km/h})^2$.

(a) Qual a probabilidade de um veículo ser flagrado a mais de 100km/h ?

Solução:

$$\int_{100}^{\infty} f(x) dx = \int_{100}^{\infty} \frac{1}{20\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-60}{20}\right)^2} dx.$$

Essa integral não é trivial!

Veja a rotina utilizada para fazer a Figura 21.

³ Lê-se: distribuição Normal com média μ e variância σ^2 .

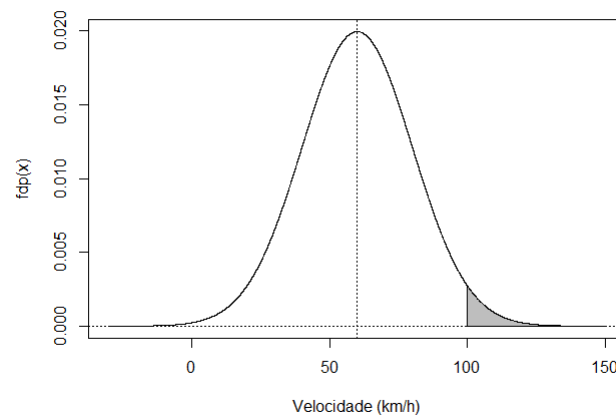


Figura 21 – Esquema destacando a área acima de 100 km/h em uma distribuição de média 60 km/h e variância $400(\text{km/h})^2$.

Fonte: Do autor.

```
x<-seq(-30, 150, by =.01)
y<-dnorm(x, mean=60, sd=20, log = FALSE)
rx<-seq(100, 150, by =.1)
ry<-numeric(2*length(rx))
ry[1:length(rx)]<-dnorm(rx, mean=60, sd=20, log=FALSE)
rx<-c(rx, rev(rx))
plot(x, y, 'l', xlab='Velocidade_(km/h)', ylab='fdp(x)')
polygon(rx, ry, col = "gray")
abline(v=60, h=0, lty=3)
```

4.9.2.2 Distribuição Normal padronizada (padrão ou zero-um)

Uma solução simples e eficiente para o problema apresentado acima é a definição da distribuição Normal padronizada. Essa distribuição apresenta, como principais características, a média igual a zero e a variância (e o desvio padrão) igual a 1.

Mediante as seguintes propriedades, uma distribuição Normal qualquer (média e variância quaisquer) pode ser temporariamente transformada em uma Normal padrão, por conveniência. Devido a essa correspondência, resultados de integrais calculadas (por processos numéricos) para atender à Normal padrão servem para gerar resultados de quaisquer integrais que se deseje em outras Normais.

Seja X uma variável aleatória que segue uma distribuição $N(\mu, \sigma^2)$. As seguintes afirmações se verificam:

1. $(X - \mu) \sim N(0, \sigma^2)$;
2. $\frac{X}{\sigma} \sim N\left(\frac{\mu}{\sigma}, 1\right)$; e, portanto,
3. $\left(\frac{X - \mu}{\sigma}\right) \sim N(0, 1)$

Devido a definição de seus parâmetros, ou seja, sabendo que $\mu = 0$ e $\sigma^2 = 1$, a função densidade de probabilidade da Normal Padronizada se reduz a

$$f(X) = \frac{1}{\sqrt{2\pi}} e^{-\frac{X^2}{2}}. \quad (4.16)$$

(a) Então, retomando o problema, podemos definir uma variável aleatória Z , tal que

$$Z = \frac{X - \mu}{\sigma}.$$

Daí,

$$z = \frac{100 - 60}{20} = 2$$

e utilizando a Tabela 13 do Apêndice B, temos que

$$P(X > 100) = P(Z > 2) = 0,5 - 0,4772 = 0,0228 = 2,28\%$$

No R:

```
pnorm(100, 60, 20, lower.tail=FALSE)
```

(b) E qual a chance de um automóvel estar trafegando entre 40 e 70km/h?

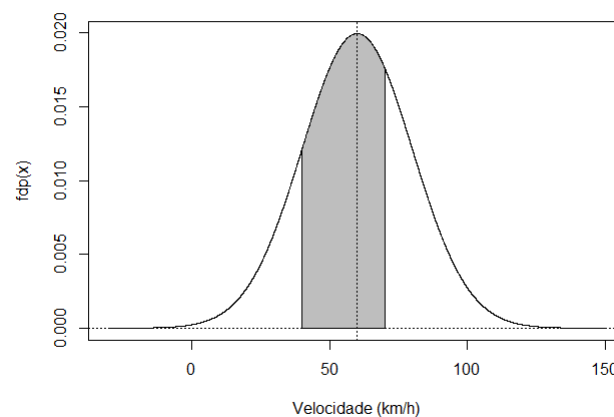


Figura 22 – Esquema destacando a área entre 40 e 100km/h em uma distribuição de média 60km/h e variância 400(km/h)².

Fonte: Do autor.

Veja a rotina utilizada para fazer a Figura 22.

```
x<-seq(-30, 150, by = .01)
y<-dnorm(x, mean=60, sd=20, log = FALSE)
rx<-seq(40, 70, by =.1)
ry<-numeric(2*length(rx))
ry[1:length(rx)]<-dnorm(rx, mean=60, sd=20, log = FALSE)
rx<-c(rx, rev(rx))
plot(x, y, 'l', xlab='Velocidade_(km/h)', ylab='fdp(x)')
polygon(rx, ry, col = 'gray')
abline(v=60, h=0, lty=3)
```

Aqui, devemos transformar dois pontos:

$$z_1 = \frac{40 - 60}{20} = -1$$

e

$$z_2 = \frac{70 - 60}{20} = 0,5$$

E, olhando na tabela,

$$P(40 < X < 70) = P(-1 < z < 0,5) = 0,3413 + 0,1915 = 53,28\%$$

No R, teríamos apenas que digitar:

```
pnorm(70, 60, 20) - pnorm(40, 60, 20)
```

(c) Qual intervalo contém 90% dos veículos?

O processo aqui é o inverso. Eu tenho a probabilidade (ou porcentagem de veículos) e desejo saber os pontos, ou seja, as velocidades que delimitam essa área.

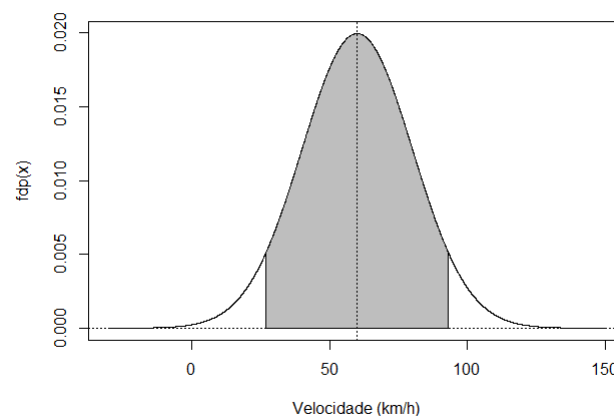


Figura 23 – Esquema destacando o intervalo que contém 90% dos veículos em uma distribuição de media 60km/h e variância $400(\text{km/h})^2$.

Fonte: Do autor.

Veja a rotina utilizada para fazer a Figura 23.

```
x<-seq(-30, 150, by = .01)
y<-dnorm(x, mean=60, sd=20, log = FALSE)
rx<-seq(27.1, 92.9, by=.1)
ry<-numeric(2*length(rx))
ry[1:length(rx)]<-dnorm(rx, mean=60, sd=20, log = FALSE)
rx<-c(rx, rev(rx))
plot(x, y, 'l', xlab='Velocidade_(km/h)', ylab='fdp(x)')
polygon(rx, ry, col = 'gray')
abline(v=60, h=0, lty=3)
```

$$1,645 = \frac{X_2 - 60}{20} \implies X_2 = 92,9 \text{ km/h}$$

$$-1,645 = \frac{X_1 - 60}{20} \implies X_1 = 27,1 \text{ km/h}$$

$$P(27,1 < X < 92,9) = 90\%$$

No R seria, simplesmente,

```
qnorm(0.05, 60, 20) ; qnorm(0.95, 60, 20)
```

4.9.2.3 Aproximação da Binomial à Normal

Quando nos deparamos com uma situação em que uma variável aleatória é Binomial, mas n e p são muito grandes,

$$np > 5 \text{ e } npq > 5$$

as contas usuais da distribuição Binomial se tornam mais difíceis de executar. Em tais situações podemos utilizar uma aproximação pela distribuição Normal para fazer os cálculos de interesse. Devemos seguir as seguintes relações de transformação:

$$\mu = np \text{ e } \sigma^2 = npq.$$

Exemplo. Um Engenheiro Agrônomo faz um teste de germinação com $n = 500$ sementes, sabendo que seu poder nominal de germinação é de $p = 83\%$. Se seu poder de germinação estiver correto, qual é a probabilidade de que, pelo menos, 430 sementes germinem?

$$\mu = 500 \times 0,83 = 415 \text{ sementes.}$$

$$\sigma^2 = 500 \times 0,83 \times 0,17 = 70,55 \implies \sigma = 8,4 \text{ sementes.}$$

$$z = \frac{430 - 415}{8,4} \approx 1,79.$$

$$P(X > 430) = P(Z > 1,79) = 0,0367 = 3,67\%.$$

4.9.2.4 Aproximação da Poisson à Normal

Em situação semelhante, porém, dessa vez se tratando de uma variável Poisson, uma média muito grande,

$$\lambda > 15,$$

também podemos aproximar à uma distribuição Normal seguindo as transformações:

$$\mu = \lambda \text{ e } \sigma^2 = \lambda.$$

Exemplo. Suponha que a média de chuvas fracas por ano em certa região seja de $\lambda = 30$. Qual a probabilidade de ocorrerem mais de 45 chuvas desse tipo no próximo ano?

$$\lambda = \mu = \sigma^2 = 30$$

$$\sigma = 5,48 \text{ chuvas fracas}$$

$$z = \frac{45 - 30}{5,8} = 2,74$$

$$P(X < 45) = P(Z < 2,74) = 0,0031 = 0,31\%$$

AMOSTRAGEM

Na grande maioria das situações práticas, o exame exaustivo de todos os elementos de uma população (censo) não é possível por ser cara, demorada ou mesmo impossível (populações infinitas).

Nesses casos, é imperativo o uso de técnicas de amostragem. O uso de amostras traz diversas vantagens como:

- menor custo;
- maior rapidez;
- boa acurácia (se coletada corretamente);
- torna viável o exame em análises destrutivas, etc.

As amostras devem ser representativas, ou seja, guardar semelhanças com a população. Para isso, é necessário, sempre que possível, a aleatorização, casualização ou sorteio.

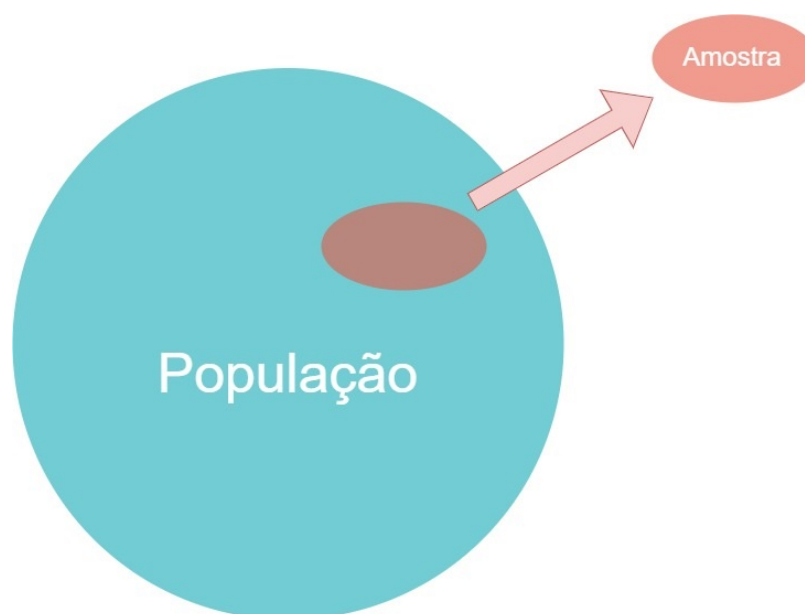


Figura 24 – Representação esquemática da retirada de uma amostra da população, como conjuntos.

Fonte: Do autor.

5.1 Amostragens não-aleatórias

São utilizadas quando o sorteio não é possível, mas se faz um esforço para garantir a representatividade. Algumas delas:

a) Difícil acesso:

Exemplo: Amostragem de minério em vagões. Embora seja praticamente impossível coletarmos minérios em todas as partes do vagão, essa tende a ser uma situação homogênea e a coleta em pontos da superfície pode ser suficiente.

b) Coleta a esmo:

Exemplo: Amostragem de solo. Nesse caso, costuma-se fazer um caminhar em ziguezague, que é um esforço de aleatorização.

c) Por conveniência:

Exemplo: Pesquisa em que se fazem perguntas a pessoas próximas ao pesquisador. Em casos em que se pode considerar que as pessoas que estão passando ao lado do pesquisador naquele momento estão ali por puro acaso.

d) Intencional:

Exemplo: Escolha de cada indivíduo. O pesquisador determina exatamente o que será observado, mas se esforça para garantir a representatividade de sua escolha.

e) Autoescola:

Exemplo: Voluntários para uma pesquisa médica. A área médica frequentemente se enquadra nesta amostragem, na qual não se pode escolher as pessoas que ficam doentes, por exemplo. Neste caso, pode-se considerar que as pessoas têm a mesma probabilidade de adoecer.

5.2 Amostragens aleatórias

Esse é o conjunto de técnicas de amostragem mais utilizado, uma vez que são nas amostragens desse tipo que se baseiam as ferramentas inferenciais que desejamos utilizar na maioria dos casos.

Você vai notar que, aqui, a palavra *aleatória* é sinônimo de *sorteio*. Ou seja, em uma amostragem aleatória há um sorteio em alguma etapa do processo.¹

5.2.1 Amostragem aleatória simples (AAS)

- Deve ser realizada em populações estritamente homogêneas.
- Ela se caracteriza pelo sorteio de n elementos de uma população.

¹ Entenda como *sorteio* a escolha de elementos populacionais, em que todos têm a mesma probabilidade de serem escolhidos.

- Pode ser feita com ou sem reposição dos elementos amostrados à população.
- Os elementos da população devem estar numerados, por exemplo, de 1 a N .

5.2.1.1 Sorteio

(a) Em calculadoras científicas:

1. Pressione a tecla $RAN\# \times N$ (tamanho da população);
2. Repita o processo n vezes.

(b) Sorteio no software R: Suponha uma população com 200 elementos da qual se deseje sortear 15 elementos, de forma inteiramente ao acaso, e sem reposição. Então

```
N<-200
n<-15
sample(1:N, n, replace=FALSE)
```

5.2.1.2 Inconvenientes da AAS

- Populações estritamente homogêneas são pouco comuns.
- É muito trabalhosa em populações grandes, porque todos os elementos devem ser numerados (ou, pelo menos, identificados), e impossível em populações infinitas.

5.2.1.3 Modelo Estatístico

$$Y_i = \mu + e_i, \quad (5.1)$$

em que Y_i é a observação do indivíduo i da amostra; μ é a média da população; e e_i é o desvio aleatório referente ao indivíduo i .

5.2.2 Amostragem aleatória sistemática (AS)

- É utilizada em situações em que elementos da população estão dispostos em série.
- Apenas o 1º elemento é sorteado. Os demais são tomados sistematicamente e se distanciam a um *passo de amostragem* (k).
- **Passo de amostragem:** inteiro mais próximo da divisão do tamanho da população (N) pelo tamanho da amostra desejada (n).
- Objetivo: facilitar o processo.

5.2.2.1 Populações finitas

- (a) Defini-se o passo de amostragem: $k = \frac{N}{n}$.
- (b) Sorteia-se o primeiro elemento dentre os k primeiros.
- (c) Tornam-se os demais de k em k .

5.2.2.2 Populações muito grandes ou infinitas

- (a) Toma-se o 1° elemento a esmo.
- (b) Tomam-se os demais elementos de maneira igualmente espaçada.
- (c) No R, suponhamos o mesmo exemplo onde se desejam 15 elementos em uma população de tamanho 200. Então,

```
N<-200
n<-15
k<-round(N/n)
amostra<-numeric(n)
amostra[1]<-sample(k,1)
for(i in 2:n) amostra[i]<-amostra[i-1]+k
amostra
```

5.2.2.3 Modelo Estatístico

$$Y_i = \mu + u_i, \quad (5.2)$$

em que²

$$u_i = \rho u_{ij} + e_i. \quad (5.3)$$

Geralmente admiti-se ρ (parâmetro de auto correlação³) igual a zero, daí,

$$Y_i = \mu + e_i. \quad (5.4)$$

5.2.3 Amostragem aleatória estratificada (AAE)

- É utilizada em populações heterogêneas.
- Deve-se dividir a população em estratos homogêneos (dentro de si), mas heterogêneos entre si. Daí sorteiam-se elementos dentro de cada estrato, proporcionalmente a seu tamanho.
- No R, o sorteio deve ser feito conforme foi mostrado na AAS, porém, dentro de cada estrato.

² Esse é apenas um dos modelos possíveis para esse caso. Você pode encontrar outros modelos estatísticos para a AS em livros de Amostragem, como (BOLFARINE; BUSSAB, 2005).

³ Note que, quando $\rho = 0$, o modelo estatístico é igual ao da amostragem aleatória simples.

5.2.3.1 Modelo Estatístico

$$Y_{ij} = \mu + t_i + e_{ij} \quad (5.5)$$

em que Y_{ij} é o valor do indivíduo j do estrato i ; μ é a média populacional; t_i é o efeito do estrato i ; e_{ij} é o efeito aleatório (desvio) do indivíduo j do estrato i .

Obs.: Média do estrato i : $\mu_i = \mu + t_i$.

5.2.4 Amostragem aleatória por conglomerado (AAC)

Conglomerados são a subdivisão da população objetivando economia de recursos, pois somente alguns serão sorteados.

Importante considerar, que:

- Principal objetivo: economia de tempo e recursos.
- Há homogeneidade *entre* conglomerados e espera-se que a variabilidade de população esteja representada *dentro* de cada um deles.
- No R, o sorteio deve ser feito conforme foi mostrado na AAS, porém, dentre o número de conglomerados. Dentro de cada conglomerado, outra(s) técnicas de amostragem devem ser utilizadas até se chegar no elemento.

Exemplo: Em uma pesquisa dentre os domicílios de Lavras, se uma AAS fosse feita, provavelmente os sorteados ficariam muito espalhados, sendo difícil de serem observados. Daí sorteia-se 7 bairros e 10 domicílios por bairro para facilitar o processo.

5.2.4.1 Modelo Estatístico

$$Y_{ij} = \mu + c_i + e_{ij}, \quad (5.6)$$

em que Y_{ij} é o valor do indivíduo j do conglomerado i ; μ é a média populacional; t_i é o efeito (aleatório) do conglomerado i ; e_{ij} é o efeito aleatório (desvio) do indivíduo j do conglomerado i .

Tabela 6 – Diferenças básicas entre a amostragem aleatória estratificada (AAE) e a amostragem aleatória por conglomerado (AAC).

AAE	AAC
Todo estado é observado	Alguns conglomerados são sorteados
Efeito fixo	Efeito aleatório
Estratos são diferentes entre si	Conglomerados são semelhantes entre si
Objetivo: > representatividade	Objetivo: < custo

INFERÊNCIA ESTATÍSTICA

6.1 Introdução

A obtenção de conclusões a respeito da população estará sempre limitada pela inerente incompletude da amostra, acarretando um certo grau de incerteza nessas conclusões. Lidar com essa incerteza, controlando-a e medindo-a, é a tarefa da Inferência Estatística. Para início, devemos fazer algumas definições a fim de munir o leitor com termos básicos que serão utilizados ao longo deste capítulo.

Inferência Estatística é o conjunto de técnicas que generalizam informações amostrais para toda a população.

Grandes áreas A Inferência Estatística pode ser didaticamente dividida em Teoria da Estimação (onde estão os intervalos de confiança) e Teoria da Decisão (onde residem os testes de hipóteses).

Parâmetro populacional Constante numérica que descreve uma população, em geral de valor desconhecido.

Os principais parâmetros de interesse são:

1. Proporção p de um atributo dos elementos de uma dada população.
2. Média μ , variância σ^2 , e desvio-padrão σ .

Os valores aproximados para os parâmetros desconhecidos, obtidos através das amostras, são conhecidos pelo nome de *estimativas*. Assim, definem-se os conceitos a seguir.

Estimativa Valor aproximado de um parâmetro populacional desconhecido calculado a partir de uma amostra.

Estimação O ato de obter uma estimativa.

Estimador Corresponde à expressão algébrica que permite obter uma estimativa, ou ainda, a variável aleatória que é usada no processo de estimação.

6.2 Teoria da Estimação

Há dois modos de se fazer estimação na Estatística: por ponto e por intervalo. A *estimação por ponto* consiste em atribuir apenas um valor numérico que aproxima (ou estima) o verdadeiro valor do parâmetro, desconhecido. A estimação por ponto, mesmo sendo feita por meio de um estimador não-tendencioso de pequena variância (isto é, exato e preciso), não resolve completamente o problema da estimação. Para a estimação é necessário, ainda, responder a duas questões:

- (i) Qual é o tamanho da confiança (probabilidade de estarmos certos) que podemos ter no valor estimado quanto a ele ser igual ao valor do parâmetro? 90%? 10%? 95%? 99%? Quanto?
- (ii) Qual é o tamanho do erro cometido na estimação? (Esse erro é medido por $\hat{\theta} - \theta$).

Esses dois problemas acima são resolvidos, na Estatística, por meio do conceito de *intervalos de confiança* (IC), ou, equivalentemente, *estimação por intervalo*. A estimação por intervalo consiste em atribuir não apenas um único valor para o desconhecido parâmetro em questão, mas, também, em atribuir todo um conjunto de valores possíveis, normalmente por meio de um intervalo de valores possíveis, haja vista que os parâmetros muito frequentemente são reais ao único valor desconhecido do parâmetro, anexando ao intervalo dado também uma medida de sua probabilidade de acerto.

6.2.1 Estimação por ponto

Consiste na obtenção de uma estimativa do valor paramétrico com base em informações vindas da amostra. Os estimadores mais comuns são \bar{X} , S , S^2 , \hat{p} , entre outros.

Uma maneira equivalente de denotar um estimador é utilizando o acento circunflexo sobre o parâmetro que se deseja estimar. Por exemplo, se eu quero escrever que *o estimador da média populacional é a média amostral*, então escrevo $\hat{\mu} = \bar{X}$.

Agora, podemos escrever que os principais estimadores pontuais são:

$$\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

$$\hat{\sigma}^2 = S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

$$\hat{\sigma} = S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$$

$$\hat{p} = \frac{x}{n}, \text{ sendo } x \text{ o número de sucessos}$$

$$\widehat{\mu_1 - \mu_2} = \hat{\mu}_1 - \hat{\mu}_2 = \bar{X}_1 - \bar{X}_2$$

$$\widehat{\left(\frac{\sigma_1}{\sigma_2}\right)} = \frac{\hat{\sigma}_1}{\hat{\sigma}_2} = \frac{S_1^2}{S_2^2}$$

6.2.2 Propriedades desejadas dos estimadores

Um estimador vai calcular, ou produzir, um valor aproximado para o correspondente parâmetro. Logo, é altamente desejável que um estimador tenha propriedades que garantam que esta aproximação é suficientemente boa. Dentre as várias propriedades que um estimador pode ter serão vistas aqui apenas duas que são muito importantes. De fato, estas são altamente desejadas.

6.2.2.1 Não-tendenciosidade

Para a compreensão dessa propriedade, suponha um parâmetro qualquer θ . Este θ pode ser a média μ da população, a variância σ^2 , o desvio padrão σ , uma proporção populacional p , entre outros. Um estimador $\hat{\theta}$ desse parâmetro θ é chamado não tendencioso se, ao se tomar infinitas amostras de uma população, o valor médio de $\hat{\theta}$ é igual a θ . Ou seja, $\hat{\theta}$ é não tendencioso se sua esperança matemática for igual a θ :

$$E[\hat{\theta}] = \theta \quad (6.1)$$

Um estimador não-tendencioso também é chamado de não viciado, ou ainda, não viesado. Todas essas nomenclaturas são equivalentes.

Outra maneira de definir não-tendenciosidade seria dizer que um estimador não-tendencioso não inclina-se nem a subestimar nem a superestimar o valor θ populacional.

Esse critério é bom e desejável, mas pode ainda não permitir discriminar entre estimadores. Como exemplo, podemos verificar que três estimadores da média populacional, digamos \bar{X} , \tilde{X} e X^* sejam não viesados para μ , isto é,

$$E[\bar{X}] = \mu, E[\tilde{X}] = \mu, E[X^*] = \mu$$

Então, como escolher entre eles? Por causa disso, é necessário usar outro critério: o critério de variância mínima, que será estudado à seguir.

6.2.2.2 Variância mínima

Considere outra vez um estimador qualquer $\hat{\theta}$. Se infinitas amostras forem coletadas em uma dada população, os valores de $\hat{\theta}$ irão variar de amostra para amostra, ou seja, esse conjunto de valores do estimador $\hat{\theta}$ apresentará uma certa *variância*, dada por

$$Var[\hat{\theta}] = \sigma_{\hat{\theta}}^2 \quad (6.2)$$

Essa variância nos fala sobre o conceito de precisão. Esse conceito é relativo, pois, se a variância de um estimador $\hat{\theta}_1$ é menor que a de um outro estimador $\hat{\theta}_2$, então $\hat{\theta}_1$ é mais preciso que $\hat{\theta}_2$, isto é, sempre precisamos de, pelo menos, dois estimadores para dizer qual é mais preciso do que outro. Poderíamos, se quiséssemos, definir precisão como $\frac{1}{\sigma_{\hat{\theta}}^2}$. A escolha do estimador que tem a menor variância, entre vários possíveis, é o critério denominada de *variância mínima*. A “variância mínima” leva-nos à estimadores “muito precisos”.

6.2.2.3 Estimadores não-tendenciosos de variância mínima

A junção das duas propriedades, a de não-tendenciosidade e de variância mínima, gera o ideal. Estimadores $\hat{\theta}$ com $E[\hat{\theta}] = \theta$ e $Var[\hat{\theta}]$ pequena, a menor possível dentre os não viesados, são os ideais. Esses são estimadores não-tendenciosos com variância mínima.

Quando encontramos um estimador não-tendencioso que tem a menor variância possível dentre todos os estimadores não-tendenciosos, estes são chamados MVUE, sigla em inglês para *minimum variance unbiased estimator*, estimadores não-tendenciosos de variância mínima.

Dada a média populacional μ , pode-se demonstrar que a média amostral é MVUE para μ . Em outras palavras, nada é melhor do que a média amostral para se estimar a média populacional μ , mesmo que existam outros concorrentes (tais como a mediana ou a moda).

6.3 Distribuições de amostragem

Como ficou evidente nas definições anteriores das propriedades de não-tendenciosidade e mínima variância, o comportamento de um estimador em infinitas amostragens em uma dada população é muito importante na Inferência Estatística. Por causa disto, vamos agora estudar melhor este comportamento que nós denominamos de *Distribuições de Amostragem*.

Distribuição de Amostragem É a distribuição de probabilidade de estimadores $\hat{\theta}$ ao longo de infinitas amostras aleatórias extraídas de uma mesma população dada.

O principal resultado sobre distribuições de amostragem é sobre a média amostral. A seguir, vamos anunciar este resultado conhecido como *Teorema Central do Limite*.

Seja uma população qualquer com média μ e variância σ^2 . Se infinitas amostras de tamanho n são coletadas dessa população, então as médias (\bar{X}) das amostras terão distribuição aproximadamente Normal com média μ e variância $\frac{\sigma^2}{n}$ à medida que n tende ao infinito.

A partir deste teorema poderemos construir vários resultados importantes para a Inferência Estatística.

6.3.1 Distribuição da média amostral de populações normais

Sabemos que se X tem distribuição Normal, $Z = \frac{X - \mu}{\sigma}$ tem distribuição Normal padronizada, $N(0, 1)$. Agora, como saber qual a distribuição da média \bar{X} ?

Se $X \sim N(\mu, \sigma^2)$, então as seguintes afirmações se verificam:

1. $\sum_{i=1}^n X \sim N(n\mu, n\sigma^2)$; e

2. $\frac{\sum_{i=1}^n X}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$; portanto,

3. $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

Se padronizarmos a média, ou seja, subtraírmos da média populacional e dividirmos por seu desvio padrão, conseguimos encontrar a quantidade que segue uma Normal padrão:

$$f(\bar{X}) = Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1). \quad (6.3)$$

Mas como σ é geralmente desconhecido, tem-se que utilizar uma quantidade que mais se assemelhe. Essa quantidade é o seu estimador S . Porém, quando substituimos σ por S , devemos pagar um preço. A nova quantidade não segue mais uma distribuição Normal padrão (Z), mas uma aproximação da Normal padrão, a distribuição *t de Student*, com $n - 1$ *graus de liberdade*, t_{n-1} .

$$f(\bar{X}) = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1} \quad (6.4)$$

Graus de liberdade: O número de graus de liberdade para um conjunto de dados corresponde ao número de valores que podem variar após terem sido impostas certas restrições a todos os valores.

6.3.1.1 A distribuição t de Student

A distribuição t de Student é uma distribuição de probabilidades muito parecida com a distribuição Normal padrão. Ela é simétrica, tem forma de sino, é centrada em zero, mas é mais “larga” e sua forma varia em função do tamanho da amostra (n). Quanto maior a amostra mais a distribuição t se assemelha a uma distribuição z, ou seja,

$$n \rightarrow \infty, \quad t_{n-1} \rightarrow N(0, 1).$$

6.3.1.2 Propriedade da distribuição t de Student

1. A distribuição t de Student é diferente, conforme o tamanho da amostra.
2. Ela tem a mesma forma geral simétrica (forma de sino) que a distribuição Normal, mas reflete a maior variabilidade (com distribuições mais amplas) que é esperada em pequenas amostras.
3. Tem média $t = 0$.
4. O desvio padrão da distribuição t de Student varia com o tamanho da amostra, mas é superior a 1.
5. Na medida em que aumenta o tamanho n da amostra, a distribuição t de Student se aproxima mais e mais da distribuição Normal padronizada. Para valores $n > 30$, as diferenças são tão pequenas que podemos utilizar os valores críticos Z em lugar de valores críticos t .

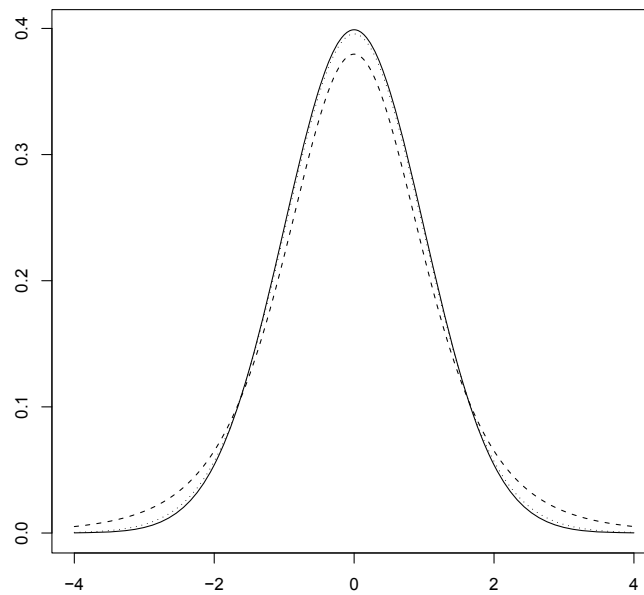


Figura 25 – Demonstração da curva Normal (linha cheia) padrão e da curva t com 5 (linha tracejada) e 30 (linha pontilhada) graus de liberdade.

Fonte: Do autor.

Veja a rotina utilizada para fazer a Figura 25.

```
x<-seq(-4, 4, by=.01)
yn<-dnorm(x, mean=0, sd=1)
yt1<-dt(x, df=5)
yt2<-dt(x, df=30)
plot(x, yn, 'l', lty=1, xlab='', ylab='')
points(x, yt1, 'l', lty=2)
points(x, yt2, 'l', lty=3)
```

6.3.2 Uso das distribuições de amostragem na Inferência Estatística

Uma vez que passamos a conhecer o Teorema Central do Limite, e a distribuição *t de Student*, que são dois dos mais importantes resultados sobre a distribuição de amostragem de médias amostrais, podemos, agora, passar a entender como podemos estimar o parâmetro média populacional, e também o parâmetro proporção populacional. Outros estimadores, tais como os estimadores da variância e do desvio-padrão, também podem ser estudados através de suas respectivas distribuições de amostragem, mas este texto não abordará tais complementações por tratar-se de um texto introdutório.

6.3.3 Estimação por intervalo

Como dito anteriormente, o estimador intervalar é aquele capaz de retornar um intervalo de possíveis valores para o parâmetro que nos interessa, e ainda, você pode saber qual é a confiança (ou probabilidade) de esse intervalo conter o verdadeiro valor paramétrico.

Intervalo de confiança (IC): é um intervalo que contém o real valor do parâmetro com probabilidade $\gamma = 1 - \alpha$.

Coefficiente de confiança: $\gamma = 1 - \alpha$

Significância: α

A maioria dos intervalos de confiança segue a seguinte forma geral:

$$IC_{\gamma}(\theta) = \hat{\theta} \pm e \quad (6.5)$$

em que γ é o coeficiente de confiança do intervalo; θ é o parâmetro de interesse ou uma função dele; $\hat{\theta}$ é o estimador de θ ; e e é o erro de estimação, dado geralmente por

$$e = q_{\alpha/2} EP_{\hat{\theta}} \quad (6.6)$$

em que $q_{\alpha/2}$ é o quantil superior $\alpha/2$ da distribuição de $\hat{\theta}$ e; $EP_{\hat{\theta}}$ é o erro padrão de $\hat{\theta}$.

Erro-padrão: é o desvio-padrão de uma distribuição de amostragem, ou seja, o desvio-padrão da distribuição de um estimador.

Quantil superior $\alpha/2$: é um valor real, que a variável aleatória tem probabilidade $\alpha/2$ de assumir valores maiores que ele.

Por exemplo, na distribuição Normal Padrão, o quantil superior de 2,5% é 1,96. Ou seja, a variável Z tem 2,5% de probabilidade de valer mais de 1,96.

Algebricamente,

$$P[Z > 1,96] = 2,5\%$$

Graficamente, como na Figura 26.

Quer saber como foi feita a Figura 26? Olhe e reproduza o script!

```
x<-seq(-4, 4, by = .01)
y<-dnorm(x, mean=0, sd=1, log = FALSE)
rx<-seq(1.96, 4, by = .1)
ry<-numeric(2*length(rx))
ry[1:length(rx)]<-dnorm(rx, mean=0, sd=1, log = FALSE)
rx<-c(rx, rev(rx))
plot(x, y, 'l', xlab='', ylab='', xaxt="n")
```

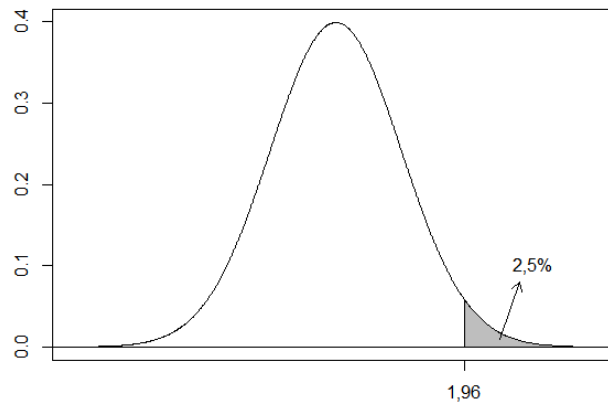


Figura 26 – Representação do quantil superior de 2,5% (área hachurada) na distribuição normal padrão.

Fonte: Do autor.

```
axis(1, at=1.96, labels="1,96")
polygon(rx, ry, col = "gray")
abline(h=0, lty=1)
text(3, 0.1, "2,5%")
arrows(2.5, 0.01, 2.8, 0.08, length = 0.1)
```

6.3.3.1 Fator de correção para populações finitas

Em populações finitas o erro (e) do IC deve ser corrigido pelo fator de correção multiplicativo:

$$\sqrt{\frac{N-n}{N-1}}. \quad (6.7)$$

Ou seja, o intervalo de confiança mais comum passa a ser

$$IC_{\gamma}(\theta) = \left[\hat{\theta} \pm e \sqrt{\frac{N-n}{N-1}} \right] \quad (6.8)$$

6.4 Teoria da Decisão

Como já vimos, informações acerca de uma população de interesse pode ser obtida por meio da *amostragem*. O passo seguinte é o de *generalizar* essas informações para a população. Essa generalização, como já vimos, é a *inferência*. Vimos também que a Inferência Estatística pode ser dividida em duas grandes áreas, sendo a primeira a Estimação de parâmetros desconhecidos da população, e a segunda, os Testes de Hipóteses.

Os Testes de Hipóteses são utilizados quando o interesse do usuário reside não na estimação de parâmetros, mas, sim, na verificação da validade, ou não, de uma determinada *hipótese* frequentemente com a finalidade de tomar alguma decisão acerca da população estudada. A diferença entre estimar um parâmetro e testar uma hipótese é que, na estimação, o objetivo é produzir números que servirão para se fazer cálculos que dimensionarão sistemas, enquanto que, na testagem, o objetivo é tomar uma decisão num processo de gestão.

A verificação de uma hipótese de interesse acerca da população é o que chamamos de *teste de hipótese*, ou, também, *teste estatístico*.

Teste de hipótese: é uma ferramenta que permite testar se um valor é pertinente de ser o valor real de parâmetro em questão.

A teoria de testes faz parte de um conjunto de conceitos e métodos chamado de *teoria da decisão*, pois tanto a rejeição, quanto a não-rejeição de hipóteses já são decisões em si mesmas (*rejeitar* é uma decisão e, *aceitar*, isto é, *não rejeitar*, também é uma decisão).

Os testes podem se referir ao modelo utilizado para descrever a população de interesse, ou ainda, admitindo que o modelo seja satisfatório, podem se referir aos parâmetros do modelo. Por outro lado, as hipóteses podem se referir ao(s) parâmetro(s) do modelo probabilístico, por sua vez tido como satisfatório.

6.4.1 Propriedades desejadas para os testes

Um teste estatístico deve ser construído e avaliado segundo dois critérios de desempenho, critérios estes que servem para dimensionar os testes, e também classificá-los. Esses critérios são:

- (i) Riscos (ou probabilidades) de decisões erradas.
- (ii) Custo para a tomada de decisão.

Um terceiro critério poderia ser aventado, a saber, o da *utilidade* da decisão tomada. Neste livro será admitido que toda e qualquer decisão tomada a partir de um teste estatístico é previamente considerada útil para o analista. Por fim, um quarto critério poderia ser aventado, que é a *facilidade* ou *viabilidade* de aplicação do teste. Apesar de que podemos discernir estes quatro critérios para a qualidade de um teste de hipótese, comumente se consideram apenas os dois primeiros critérios, (i) e (ii) acima descritos. Conforme se delineiam esses dois critérios de desempenho acima, (i) e (ii), podemos ter três tipos de testes estatísticos:

1. Testes de significância.
2. Testes mais poderosos.
3. Testes sequenciais.

Os dois primeiros tipos acima (1 e 2) são comuns e, às vezes, indistintamente, chamados de testes de hipótese. São alguns desses testes que estudaremos aqui nesse livro. O terceiro tipo de teste (o de número 3 acima), têm teoria e método um pouco mais sofisticados e não serão objeto de enfoque extenso neste livro, sendo deixados para estudos futuros pelo leitor interessado.

6.4.2 Estrutura dos testes

Uma estrutura geral, didaticamente estabelecida, de todo teste de hipóteses, consiste de quatro partes: um *par de hipóteses*, uma *estatística de teste*, uma *regra de decisão* e uma *conclusão*.

6.4.2.1 Par de hipóteses

Os testes apresentam uma hipótese principal sob julgamento, chamada de *hipótese de nulidade* ou *hipótese nula*, representada pela notação H_0 . Se rejeitada, então uma outra hipótese candidata é considerada como verdadeira: a chamada *hipótese alternativa*, representada por H_1 ou H_a .

Considere o exemplo do estudo da proporção de pessoas favoráveis a uma determinada medida governamental. Pode-se supor que esta proporção, na população toda, esteja acima ou abaixo de um nível considerado satisfatório:

$$\begin{cases} H_0 : & \text{a proporção } p \text{ de pessoas favoráveis é igual ou menor a } p_0. \\ H_1 : & \text{a proporção } p \text{ de pessoas favoráveis é superior a } p_0. \end{cases}$$

Ou, simplesmente

$$\begin{cases} H_0 : & p \leq p_0 \\ H_1 : & p > p_0 \end{cases}$$

O estabelecimento do par de hipóteses é feito pelo pesquisador (por você!) e, por isso, é fonte de inseguranças, muitas vezes. Por esse motivo, aqui vão algumas dicas gerais que vão te ajudar a estabelecer corretamente as hipóteses:

- O sinal de igualdade ($=$, \geq , \leq) deve sempre estar em H_0 .
- H_0 e H_1 são disjuntas e muitas vezes complementares. Isso quer dizer que nunca haverá interseção entre elas. Por exemplo, uma média é igual a 5 ou é diferente de 5. Não dá para ser as duas coisas ao mesmo tempo!
- H_0 é uma hipótese de *não ação* e H_1 é uma hipótese de *ação*. Isso quer dizer que a hipótese nula é aquilo que “você já tem”, que já vale antes do teste acontecer. Por exemplo, um médico vai mudar o antibiótico que costuma prescrever se o novo antibiótico no mercado se mostrar mais eficiente. Para isso vai considerar, digamos, o tempo até a recuperação. Ele sabe que o atual antibiótico tem um tempo médio de recuperação de 10 dias. Se o novo antibiótico tiver tempo médio de recuperação significativamente inferior a 10 dias ele decidirá por mudar suas prescrições. Note que H_0 é o que já acontece, ou seja, *tempo médio de 10 dias*. Automaticamente, H_1 será aquilo que fará a mudança (ação) acontecer, *tempo médio inferior a 10 dias*. Daí, o par de hipóteses deve ser estabelecido assim

$$\begin{cases} H_0 : & \mu = 10 \\ H_1 : & \mu < 10 \end{cases}$$

Na prática, a aceitação ou rejeição de H_0 (e, conseqüentemente, a rejeição ou aceitação de H_1) são feitas mediante uma amostra aleatória da qual estimativas apropriadas são calculadas.

Se a distribuição de amostragem dos estimadores correspondentes for conhecida, então pode-se calcular a probabilidade da estimativa observada ter ocorrido, *admitindo a hipótese de nulidade H_0 como verdadeira*. Se essa probabilidade for baixa, então existem bons motivos para rejeitar essa hipótese e aceitar H_1 .

6.4.2.2 Estatística de teste

A estatística de teste é a “conta” que fazemos com os dados amostrais para verificar qual é a evidência que temos a favor ou contra H_0 .

Devemos lembrar que a estatística de teste é uma estatística (é função apenas de quantidades amostrais e valores conhecidos) e é uma variável aleatória (possui distribuição de probabilidade conhecida). É conhecendo a estatística de teste que saberemos qual distribuição de probabilidade (tabela) devemos utilizar. Por exemplo, se você calcular um t_c , digamos

$$t_c = \frac{\bar{X} - \mu_0}{S/\sqrt{n}},$$

vai certamente precisar comparar esse valor com um quantil da distribuição *t de Student*.

Neste livro não entraremos no mérito de como obter as melhores estatísticas de teste. Apresentaremos para você as estatísticas com as propriedades mais interessantes para que seu teste seja eficiente.

6.4.2.3 Regra de decisão

Em um teste de hipóteses, a regra de decisão é o teste em si. De posse do valor da estatística de teste aplicada na amostra, que representa a evidência que você tem contra ou a favor de H_0 , você deve compará-lo com um quantil da distribuição mais apropriada e tomar a decisão que o teste recomendar.

Fazemos testes (regras de decisão) o tempo todo em nossa vida. Veja só: quando você diz “vou comprar carne nesse açougue se estiver mais barata que no açougue perto da minha casa”, ou ainda, “se a maioria de vocês for andar de bicicleta, também vou”, você está propondo uma regra de decisão!

No caso de um teste t, por exemplo, dizemos “se o t calculado for superior a um quantil superior da distribuição t, rejeitarei H_0 ”. Algebricamente, se $t_c > t_{\alpha/2}$, rejeita-se H_0 . Desta maneira, acabamos estabelecendo uma *região de rejeição de H_0* (R.R.) e uma *região de aceitação de H_0* (R.A.).¹

Dessa forma, ao estabelecermos regiões de aceitação e rejeição de H_0 na distribuição de amostragem da estatística de teste, acabamos criando o conceito de *lateralidade de um teste*. Assim, um teste pode ser dito bilateral, unilateral à direita ou unilateral à esquerda, de acordo com o posicionamento da sua região de rejeição de H_0 .

Veja a Figura 27 (página 74) para mais informações. Ela foi feita em uma distribuição simétrica para fins de ilustração mas, note, nem sempre procede dessa maneira. Muitas vezes a distribuição de amostragem da estatística de teste é assimétrica, como a F ou a χ^2 .

Note ainda que as regiões de rejeição tem tamanho α nos testes unilaterais e tamanho $\alpha/2$ nos testes bilaterais. Isso acontece porque a significância do teste (α) é dividida em partes iguais para as duas regiões de rejeição, no teste bilateral.

¹ Existe o receio de alguns estatísticos em dizer *aceitar H_0* , uma vez que H_0 é algo que já acontece. Ou seja, não precisamos aceitar algo que já existe. Precisamos sim *rejeitar* ou *não rejeitar H_0* . Mesmo conscientes e respeitando este argumento, os autores deste livro se reservam o direito de dizer *aceitar H_0* , quando ela não for rejeitada. Comportamento semelhante pode ser encontrado em (MOOD; GRAYBILL; BOES, 1974).

Por fim, vale a pena ressaltar que a lateralidade do teste é vista pelo sinal de H_1 . Quando temos o sinal de \neq em H_1 , o teste será bilateral; quando o sinal for $<$, unilateral à esquerda; e quando for \geq , unilateral à direita.

Suponha um parâmetro qualquer, digamos θ , para o qual pretendemos testar se θ_0 é um valor plausível. Então, podemos estabelecer três tipos de pares de hipóteses, a depender da lateralidade desejada. Veja:

$$\left\{ \begin{array}{l} H_0: \theta \geq \theta_0 \\ H_1: \theta < \theta_0 \end{array} \right. \quad \left\{ \begin{array}{l} H_0: \theta \leq \theta_0 \\ H_1: \theta > \theta_0 \end{array} \right. \quad \left\{ \begin{array}{l} H_0: \theta = \theta_0 \\ H_1: \theta \neq \theta_0 \end{array} \right.$$

Unilateral à esquerda

Unilateral à direita

Bilateral

Da mesma forma que o sinal de H_1 define a lateralidade do teste, é a necessidade prática do contexto que demanda o sinal de H_1 . Por exemplo, considere o seguinte problema: *Um município deve multar as indústrias que poluem seus rios se o nível médio de poluentes na água for superior a 10ppm*. Note que este teste é intrinsecamente unilateral à direita. Afinal, nada é feito se o nível de poluentes for menor ou igual a 10ppm (não ação) e uma multa é aplicada somente se a média for superior a 10ppm. Ou seja, o par de hipóteses é

$$\left\{ \begin{array}{l} H_0: \mu \leq 10ppm \\ H_1: \mu > 10ppm \end{array} \right.$$

Quer reproduzir a Figura 27 no R?

```
# Teste unilateral à direita
x<-seq(-4, 4, by =.01)
y<-dnorm(x, mean=0, sd=1, log = FALSE)
rx<-seq(1.64, 4, by =.1)
ry<-numeric(2*length(rx))
ry[1:length(rx)]<-dnorm(rx, mean=0, sd=1, log = FALSE)
rx<-c(rx, rev(rx))
plot(x, y, 'l', xlab='', ylab='', xaxt="n", yaxt="n",
main='Teste_unilateral_à_direita')
axis(1, at=1.64, labels=expression(paste(q[alpha])), cex.axis=2)
polygon(rx, ry, col = "gray")
abline(h=0, lty=1)
text(3, 0.1, labels=expression(alpha), cex=2)
text(0, 0.2, "R.A.", cex=2)
text(3, 0.2, "R.R.", cex=2)
arrows(2.5, 0.01, 2.8, 0.08, length = 0.1)

# Teste unilateral à esquerda
x<-seq(-4, 4, by =.01)
y<-dnorm(x, mean=0, sd=1, log = FALSE)
rx<-seq(-4, -1.64, by =.1)
```

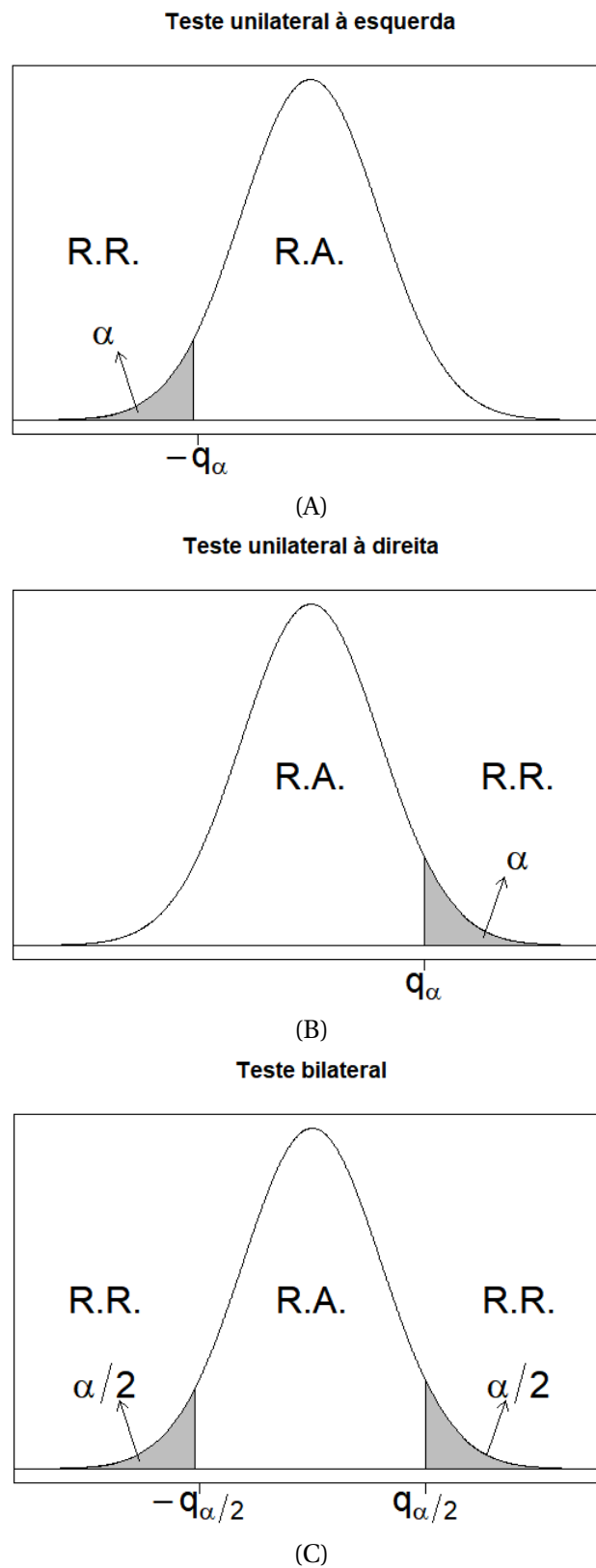



Figura 27 – Representação, em uma distribuição simétrica, de regiões de aceitação e rejeição de H_0 , definindo testes unilaterais à direita, à esquerda e bilateral.

Fonte: Do autor.

```

ry<-numeric(2*length(rx))
ry[1:length(rx)]<-dnorm(rx, mean=0, sd=1, log = FALSE)
rx<-c(rx, rev(rx))
plot(x, y, 'l', xlab='', ylab='', xaxt="n", yaxt="n",
main='Teste_unilateral_à_esquerda')
axis(1,at=-1.64,labels=expression(paste(-q[alpha])),cex.axis=2)
polygon(rx, ry, col = "gray")
abline(h=0, lty=1)
text(-3, 0.1, labels=expression(alpha),cex=2)
text(0,0.2,"R.A.",cex=2)
text(-3,0.2,"R.R.",cex=2)
arrows(-2.5,0.01,-2.8,0.08,length = 0.1)

# Teste bilateral
x<-seq(-4, 4, by =.01)
y<-dnorm(x, mean=0, sd=1, log = FALSE)
rx<-seq(-4, -1.64, by =.1)
ry<-numeric(2*length(rx))
ry[1:length(rx)]<-dnorm(rx, mean=0, sd=1, log = FALSE)
rx<-c(rx, rev(rx))
plot(x, y, 'l', xlab='', ylab='', xaxt="n", yaxt="n",
main='Teste_bilateral')
axis(1,at=-1.64,labels=expression(paste(-q[alpha/2])),cex.axis=2)
polygon(rx, ry, col = "gray")
abline(h=0, lty=1)
text(-3, 0.1, labels=expression(alpha/2),cex=2)
text(0,0.2,"R.A.",cex=2)
text(-3,0.2,"R.R.",cex=2)
arrows(-2.5,0.01,-2.8,0.08,length = 0.1)
text(3,0.2,"R.R.",cex=2)
arrows(2.5,0.01,2.8,0.08,length = 0.1)
rx<-seq(1.64, 4, by =.1)
ry<-numeric(2*length(rx))
ry[1:length(rx)]<-dnorm(rx, mean=0, sd=1, log = FALSE)
rx<-c(rx, rev(rx))
polygon(rx, ry, col = "gray")
axis(1,at=1.64,labels=expression(paste(q[alpha/2])),cex.axis=2)
text(3, 0.1, labels=expression(alpha/2),cex=2)

```

6.4.2.4 Conclusão

A conclusão de teste de hipótese é de fundamental importância. Nela, expressamos qual foi a decisão tomada, reafirmamos o nível de significância adotado e informamos o tipo de teste utilizado. Essas afirmações são importantes para deixarmos claro que aquela decisão tomada é passível de estar

errada, mas temos controle sobre a probabilidade de cometer cada tipo de erro. Leia a próxima seção para conhecer os tipos de erros existentes em um teste.

6.4.3 Contingências: tipos de erros e acertos possíveis

Em todo teste de hipóteses, quando uma decisão é tomada quatro resultados são possíveis como decorrência da combinação de duas decisões possíveis (aceitar ou rejeitar H_0), com duas “verdades” possíveis (H_0 é verdadeira ou H_0 é falsa) (Tabela 7).

É claro que rejeitar uma hipótese falsa e aceitar uma hipótese verdadeira são acertos, e não nos é útil diferenciá-los nesse momento. Contudo, *rejeitar uma hipótese verdadeira* constitui o que chamamos de erro tipo I. Já, *aceitar uma hipótese falsa* é chamado de erro tipo II.

Por exemplo: se uma pessoa está no banco dos réus ela pode ser, na situação, inocente ou culpada. Nesta mesma situação, o Juiz, por sua vez, precisa decidir se essa pessoa é inocente ou culpada. Mas o juiz nem sempre acerta! Ele pode errar, então, de duas formas: prender o inocente, ou soltar o culpado. Se aceitarmos que todos somos inocentes até que se prove o contrário, temos que H_0 : ser inocente. Daí, prender o inocente seria cometer o erro tipo I e, soltar o culpado, o erro tipo II.

Tabela 7 – Representação tabular das contingências em um teste de hipóteses: erros e acertos.

Decisão	Verdade	
	H_0 é verdadeira	H_0 é falsa
Aceita-se H_0	Acerto	Erro Tipo II
Rejeita-se H_0	Erro Tipo I	Acerto

Como diferenciamos os dois tipos de erros possíveis, temos que nos concentrar em como controlá-los². Então, vamos nomear a probabilidade de cada tipo de erro acontecer.

$$P[\text{Erro Tipo I}] = \alpha \quad P[\text{Erro Tipo II}] = \beta$$

Como nos testes de hipóteses o tamanho da amostra n é considerado fixo (não é uma variável aleatória), temos que quando α é conhecido, β é desconhecido. Ou seja, só conhecemos a probabilidade de cometer um dos erros.

Erro Tipo I: é o erro que se comete ao rejeitar H_0 , sendo ela verdadeira.

Erro Tipo II: é o erro que se comete ao aceitar H_0 , sendo ela falsa.

Nível de significância do teste: é o valor da probabilidade de se cometer o erro tipo I.

Poder do teste: é a probabilidade de se rejeitar H_0 , quando ela é realmente falsa. Em outras palavras, é $1 - \beta$, onde β é a probabilidade de se cometer o erro tipo II.

² A palavra “controlar” aqui se refere a como devemos interferir para escolher ou modular seus valores.

Duas características desejadas em um teste são: pequena probabilidade de se cometer o erro tipo I (α), e alto poder (β pequeno). Porém, temos que α e β são *inversamente proporcionais* (Figura 28). Contudo, desejamos que ambos sejam pequenos por serem probabilidades de cometermos erros. A saída empírica para esse dilema é escolhermos o valor de α entre 1% e 10%, de acordo com a avaliação de qual é o pior tipo de erro para nós.³

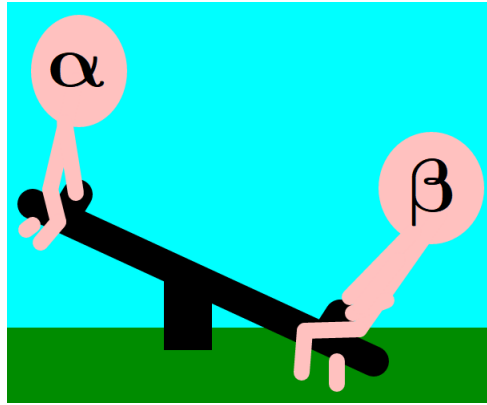


Figura 28 – Representação pictórica da relação negativa entre as probabilidades α e β , para uma amostra de tamanho fixo n .

Fonte: Do autor.

Se em uma situação o pior tipo de erro for o erro tipo I, aconselha-se $\alpha \rightarrow 1\%$. Se o erro tipo II for pior, aconselha-se $\alpha \rightarrow 10\%$.⁴ Contudo, se ambos erros forem igualmente graves, o usual é tomar $\alpha = 5\%$.

Veremos mais pormenorizadamente o desenvolvimento destes critérios de desempenho nos casos a seguir, pelos quais vamos apresentar alguns dos testes de hipótese mais úteis, nas aplicações da Estatística.

6.5 Inferência sobre uma população normal

Suponha uma população com distribuição normal. Essa população tem média μ e variância σ^2 . Dessa população é retirada uma amostra aleatória de tamanho n , com a qual podemos calcular estatísticas amostrais como média amostral (\bar{x}), variância amostral (s^2), proporção amostral (\hat{p}), dentre outras (Figura 29).

6.5.1 Estimação da média populacional (μ)

A primeira coisa que você vai notar é que o estimador intervalar da média depende do fato de conhecermos ou não a variância populacional σ^2 .

³ É claro que é possível nos depararmos com situações onde precisamos escolher $\alpha < 1\%$ ou $\alpha > 10\%$. Contudo, por serem menos comuns, vamos nos concentrar nessa faixa de variação geralmente adotada para α em trabalhos científicos.

⁴ A ideia nesse caso é que, já que não podemos escolher um β pequeno, usamos o conhecimento do fato de que são inversamente proporcionais e escolhemos um α “grande”. Com isso, diminuímos o valor de β , mesmo sem saber quanto ele vale.

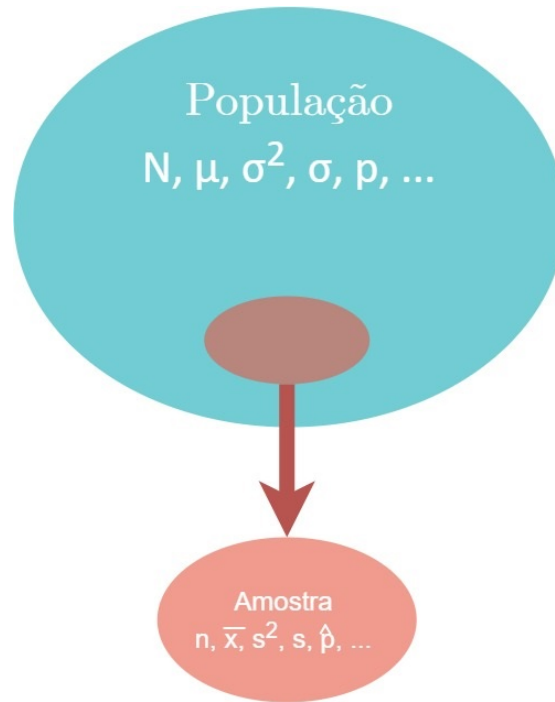


Figura 29 – Esquema ilustrativo da amostragem de uma população normal, parâmetros populacionais e estatísticas amostrais.

Fonte: Do autor.

Estimar a média por ponto não varia, mas por intervalo sim. Quando conhecemos a variância populacional, usamos a distribuição normal padrão (Z), mas quando não conhecemos, precisamos usar a variância amostral e a distribuição *t de Student*.

- **Por ponto:**

$$\hat{\mu} = \bar{X}$$

- **Por intervalo:**

Para populações infinitas

σ^2 conhecido

σ^2 desconhecido

$$IC_{\gamma}(\mu) = \left[\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

$$IC_{\gamma}(\mu) = \left[\bar{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n}} \right]$$

 Para populações finitas

 σ^2 conhecido σ^2 desconhecido

$$IC_{\gamma}(\mu) = \left[\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \right]$$

$$IC_{\gamma}(\mu) = \left[\bar{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \right]$$

em que \bar{X} é a média amostral; n é o tamanho da amostra; S é o desvio-padrão amostral; σ é o desvio-padrão populacional; e $z_{\alpha/2}$ e $t_{\alpha/2}$ são quantis superiores das distribuições Z e t , respectivamente.

Note que, no caso da estimação da média, as duas possíveis expressões para o erro de estimação são

 Para populações infinitas

 σ^2 conhecido σ^2 desconhecido

$$e = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$e = t_{\alpha/2} \frac{S}{\sqrt{n}}$$

 Para populações finitas

 σ^2 conhecido σ^2 desconhecido

$$e = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

$$e = t_{\alpha/2} \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Essas expressões vêm dos intervalos de confiança para a média, com variância populacional conhecida e desconhecida. A partir delas, podemos escrever uma expressão para o tamanho mínimo da amostra. Escrevemos uma expressão para o erro máximo que desejamos cometer: na prática, basta trocar o sinal de = pelo sinal de \geq . Em seguida, isolamos o n no primeiro membro com uma manipulação algébrica simples. Ficamos então com

 σ^2 conhecido σ^2 desconhecido

$$n \geq \frac{z_{\alpha/2}^2 \sigma^2}{e^2}$$

$$n \geq \frac{t_{\alpha/2}^2 S^2}{e^2}$$

E, considerando o fator de correção para populações finitas, temos

σ^2 conhecido	σ^2 desconhecido
$n \geq \frac{z_{\alpha/2}^2 \sigma^2 N}{e^2 (N-1) + z_{\alpha/2}^2 \sigma^2}$	$n \geq \frac{t_{\alpha/2}^2 S^2 N}{e^2 (N-1) + t_{\alpha/2}^2 S^2}$

Com essas expressões, podemos planejar o tamanho amostral de um experimento que tem, como principal objetivo, estimar a média de uma população normal. Pense como isso pode ser útil! Aqui cabem algumas observações interessantes:

1. Note que para calcularmos o tamanho da amostra precisamos estipular o erro máximo que admitimos cometer (lembre-se que em um processo de amostragem não é possível cometer erro zero!), e que esse erro deve estar na mesma unidade da variável que você está estudando. Por exemplo, se está estudando o comprimento em metros, o erro deve ser estipulado também em metros.
2. Precisamos conhecer a variância populacional (o que é raro) ou sua estimativa (S^2). Essa estimativa pode ser retirada da literatura, ou ainda, de uma amostra-piloto⁵.
3. Para o tamanho amostral que exige um quantil da distribuição t também precisamos conhecer os graus de liberdade, dado aqui por $\nu = n - 1$. Mas se tivermos uma amostra-piloto, essa questão está resolvida. Note que há, aqui, uma recursão. A amostra-piloto tem um tamanho $n^{(1)}$, que abastece a fórmula e gera um tamanho da próxima amostra, digamos, $n^{(2)}$. Esse tamanho pode voltar a abastecer os graus de liberdade e gerar outro tamanho amostral, $n^{(3)}$. Isso não acontece indefinidamente. O processo converge para um certo valor de n , que deve ser o tamanho amostral adotado.
4. Note, por fim, que o coeficiente de confiança (γ) do intervalo que será construído no futuro precisa ser agora definido. Com ele, calcularemos o $\alpha/2$, necessário para consultar o quantil das distribuições Z e t .

Exemplo. O tomate é um item que sofre muita variação de preço ao consumidor final. Sua produção está sujeita a intemperes, pragas, doenças e flutuações nos preços de sementes e insumos. Uma pesquisa foi feita no município para estimar o preço médio do tomate no mês de julho. Sabe-se que existem 63 pontos de vendas de tomate na cidade, entre supermercados, mercadinhos, feiras-livres e outros. Contudo, foram amostrados 12 locais de venda, o que revelou um preço médio de R\$3,57 por quilograma e desvio-padrão de R\$0,66. Sabendo disso, responda:

- a) Estime, por ponto e por intervalo, o preço médio do tomate no município.
- b) Qual foi o erro cometido na estimação da letra (a)?
- c) Qual deveria ser o tamanho da amostra para que o erro cometido fosse 60% do erro anterior?

⁵ Uma *amostra-piloto* é uma amostra preliminar, uma amostra antes da amostra. Estimativas vindas de amostras-piloto são comumente usadas em Estatística, como pode ser visto em (BOLFARINE; BUSSAB, 2005).

Resolução. (a) Como a confiança do intervalo não foi fixada, a primeira coisa a fazermos é escolher um coeficiente de confiança *alto o suficiente*. Escolho $\gamma = 0,95$.

O enunciado deixa claro que a população é finita e de tamanho $N = 63$. Portanto, os $n = 12$ locais amostrados formam uma amostra. Logo, deduzimos que $\bar{x} = 3,57$ reais e $s = 0,66$ reais. Então, $s^2 = 0,4356$ reais². Também entendemos que a variância populacional é desconhecida e, portanto, devemos usar o IC baseado na distribuição t , com $\nu = 12 - 1 = 11$ graus de liberdade. Também precisamos do quantil superior da distribuição t , que considerando a confiança escolhida e os graus de liberdade, consultando a Tabela 36 do Apêndice B, vale $t_{2,5\%} = 2,201$.

- **Por ponto:** $\hat{\mu} = \bar{x} = 3,57$ reais.

- **Por intervalo:**

$$\begin{aligned} IC_{95\%}(\mu) &= \left[\bar{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \right] \\ &= \left[3,57 \pm 2,201 \frac{0,66}{\sqrt{12}} \sqrt{\frac{63-12}{63-1}} \right] \\ &= [3,57 \pm 0,38] \\ &= [3,19; 3,95] \end{aligned}$$

Conclusão: Com 95% de confiança, pode-se afirmar que o preço médio do quilograma do tomate está entre R\$3,19 e R\$3,95.

(b) O erro cometido na estimação foi de R\$0,38.

(c) Vamos chamar o novo erro de e' . Então, $e' = 0,60e = 0,60(0,38) = 0,23$ reais.

Assim,

$$\begin{aligned} n &\geq \frac{t_{\alpha/2}^2 S^2 N}{e^2(N-1) + t_{\alpha/2}^2 S^2} \\ &\geq \frac{2,201^2(0,4356)(63)}{0,23^2(63-1) + 2,201^2(0,4356)} \\ &\geq 24,66 \\ &\approx 25 \text{ pontos de coleta} \end{aligned}$$

Tudo bem parar o exercício por aqui. Mas se você quiser continuar, vamos ver o que acontece se alimentarmos o processo recursivo?

Com esse novo tamanho de amostra, vamos obter um novo grau de liberdade para a distribuição t : $\nu = 25 - 1 = 24$. Logo, de acordo com a Tabela 36 do Apêndice B, $t_{2,5\%} = 2,064$. Com esse novo valor de t obtemos um novo tamanho amostral, $n = 23$.

Novamente, podemos realimentar o processo e obter $\nu = 22 \Rightarrow t_{2,5\%} = 2,074 \Rightarrow n = 23$.

Note que o processo já convergiu, com apenas 3 iterações. Então, o tamanho de amostra recomendado é 23 pontos de venda (o que é muito parecido com os 25 que havíamos encontrado sem o

processo iterativo.)

No R esse problema teria sido resolvido facilmente. Entretanto, não foram fornecidos todos os dados, mas apenas as estatística amostrais. Ainda assim, poderíamos estimar e dimensionar a próxima amostra, da seguinte maneira:

```
N<-63
xb<-3.57
s<-0.66
n<-12
gama<-0.95
alfa<-1-gama
talfa2<-qt(alfa/2,n-1,lower.tail=F)
e<-round(talfa2*s/sqrt(n)*sqrt((N-n)/(N-1)),2)
#IC(media)
round(xb-e,2) ; round(xb+e,2)
#Dimensionamento da amostra
el<-0.23
n<-c(n, ceiling((talfa2^2*s^2*N)/(el^2*(N-1) + talfa2^2*s^2)))
while(n[length(n)]!=n[length(n)-1]){
t1<-qt(alfa/2, n[length(n)]-1, lower.tail=F)
n<-c(n, ceiling((t1^2*s^2*N)/(el^2*(N-1) + t1^2*s^2)))
}
```

6.5.2 Estimação de uma proporção populacional (p)

Uma proporção é definida como o número de sucessos sobre o número de ensaios e , portanto, vem de uma variável Binomial. Isso define o estimador pontual da proporção, ou seja, o estimador da proporção populacional é a proporção amostral.

Existem diversas aproximações utilizadas para estimar uma proporção por intervalo, porém muito mais trabalhosas. Por isso podemos utilizar a aproximação binomial à normal. Entretanto, este procedimento não é recomendado quando

$$n\hat{p} \leq 5 \text{ ou } n(1 - \hat{p}) \leq 5.$$

- **Por ponto:**

$$\hat{p} = \frac{x}{n},$$

em que x é o número de sucessos e n o tamanho da amostra (número de ensaios).

- **Por intervalo:** $IC_{\gamma}(p) = \left[\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} \right]$

em que \hat{p} e \hat{q} são proporções amostrais de sucesso e fracasso, respectivamente; n é o tamanho da amostra; $z_{\alpha/2}$ é um quantil superior da distribuição Z .

Sob essa aproximação, a expressão para o erro de estimação, com e sem correção para populações finitas, fica

Para populações infinitas	Para populações finitas
$e = z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$	$e = z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} \sqrt{\frac{N-n}{N-1}}$

que geram essas expressões para calcular o tamanho amostral

Para populações infinitas	Para populações finitas
$n \geq \frac{z_{\alpha/2}^2 \hat{p}\hat{q}}{e^2}$	$n \geq \frac{z_{\alpha/2}^2 \hat{p}\hat{q} N}{e^2(N-1) + z_{\alpha/2}^2 \hat{p}\hat{q}}$

Como essas expressões para o tamanho amostral só dependem de quantis da distribuição Z - e essa distribuição não depende do tamanho amostral - aqui não temos processo recursivo.

Exemplo. A busca por uma vacina contra a COVID-19 está marcando o século como uma das mais incisivas e velozes. Um grupo de pesquisadores desenvolveu uma possível vacina e aplicou em 2000 pessoas, das quais 1289 apresentaram anticorpos após dias à sua aplicação. Sabendo disso, responda corretamente.

- Estime a proporção de indivíduos imunizados por ponto e por intervalo.
- Qual foi o erro de estimação cometido no item anterior?
- Qual deveria ser o tamanho da amostra para que fosse cometido metade desse erro?

Resolução. (a) Nesse exemplo, compreendemos que as 2000 pessoas estudadas são uma amostra das pessoas que precisam se imunizar contra a COVID-19. Porém, não foi dito qual é o tamanho da população. Possivelmente, todas as pessoas que existem no mundo, ou virão existir. Sendo assim, estamos diante de uma população infinita.

Novamente, não foi fixada a confiança desejada para o intervalo. Então, escolho $\gamma = 0,90$. Assim, o quantil superior $\alpha/2$ da distribuição Z , segundo a Tabela 13 do Apêndice B, é $z_{5\%} = 1,64$.

Dito isso, temos a estimação:

- **Por ponto:** $\hat{p} = \frac{1289}{2000} = 0,6445 = 64,45\%$.

- **Por intervalo:**

$$\begin{aligned} IC_{90\%}(p) &= \left[\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} \right] \\ &= \left[0,6445 \pm 1,64 \sqrt{\frac{0,6445(1-0,6445)}{2000}} \right] \\ &= [0,6445 \pm 0,0176] \\ &= [0,6269; 0,6621] \end{aligned}$$

Conclusão: Com 90% de confiança, pode-se afirmar que a proporção de pessoas imunizadas e com anticorpos está entre 62,69% e 66,21%.

(b) O erro cometido na estimação foi de 1,76%.

(c) Para calcular o novo tamanho amostral, devemos considerar que a população é infinita e que o novo erro deve ser $e' = 0,5(0,0176) = 0,0088$. Assim,

$$\begin{aligned} n &\geq \frac{z_{\alpha/2}^2 \hat{p} \hat{q}}{e^2} \\ &\geq \frac{1,64^2(0,6445)(1 - 0,6445)}{0,0088^2} \\ &\geq 7.957,651 \\ &\approx 7.958 \text{ pessoas.} \end{aligned}$$

No R, esse problema seria bem fácil de se resolver, ainda mais por não haver processo recursivo no dimensionamento da amostra. Novamente, apenas as estatísticas amostrais foram informadas. Então, temos o *script*:

```
x<-1289
n<-2000
p<-x/n
gama<-0.90
alfa<-1-gama
zalfa2<-round(qnorm(alfa/2,0,1,lower.tail=F),2)
e<-round(zalfa2*sqrt(p*(1-p)/n),4)
#IC(p)
round(p-e,4) ; round(p+e,4)
#Dimensionamento da amostra
e1<-0.0088
n1<-ceiling((zalfa2^2*p*(1-p))/e1^2)
```

6.5.3 Estimação do Total populacional (T)

Neste tópico vamos apresentar um parâmetro derivado de outros parâmetros existente somente em populações finitas. Afinal, uma função de um parâmetro, que não envolva variáveis aleatórias, é também um parâmetro.

Sendo N o tamanho, μ a média, e p uma proporção populacionais, $N\mu$ e Np são totais. Por exemplo, total de café vendido por dia em uma padaria; total de carne consumida em um restaurante; total de brasileiros contaminados pela COVID-19; etc. Sendo assim, definimos:

$$T = N\mu \quad \text{ou} \quad T = Np$$

A estimação do total, definido de qualquer uma das duas formas, é trivial, agora que sabemos como estimar a média e a proporção.

6.5.3.1 Estimação de T a partir da média μ

- **Por ponto:**

$$\hat{T} = N\bar{X}$$

- **Por intervalo:**

$$IC_{\gamma}(T) = N \times IC_{\gamma}(\mu) = \left[N\bar{X} \pm t_{\alpha/2} N \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \right]$$

6.5.3.2 Estimação de T a partir de uma proporção p

- **Por ponto:**

$$\hat{T} = N\hat{p}$$

em que x é o número de sucessos e n o tamanho da amostra (número de ensaios).

- **Por intervalo:**

$$IC_{\gamma}(T) = N \times IC_{\gamma}(p) = \left[N\hat{p} \pm z_{\alpha/2} N \sqrt{\frac{\hat{p}\hat{q}}{n}} \sqrt{\frac{N-n}{N-1}} \right]$$

6.5.4 Teste para a média populacional (μ)

Aqui também temos a mesma consideração feita sobre o conhecimento da variância populacional feita anteriormente. Vale lembrar ainda que conhecer a variância populacional é uma suposição muito artificial, ou seja, dificilmente acontece na “vida real”. Por isso mesmo, apresentaremos apenas o teste t para uma média.

6.5.4.1 Par de hipóteses

$$\left\{ \begin{array}{l} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{array} \right. \quad \text{ou} \quad \left\{ \begin{array}{l} H_0: \mu = \mu_0 \\ H_1: \mu > \mu_0 \end{array} \right. \quad \text{ou} \quad \left\{ \begin{array}{l} H_0: \mu = \mu_0 \\ H_1: \mu < \mu_0 \end{array} \right.$$

Bilateral

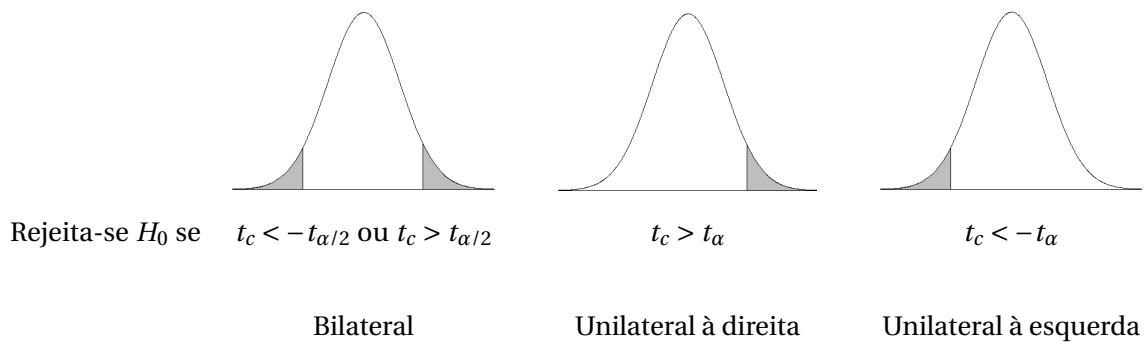
Unilateral à direita

Unilateral à esquerda

6.5.4.2 Estatística de teste

$$t_c = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \quad (6.9)$$

6.5.4.3 Regra de decisão



Exemplo. Já foi dito neste capítulo que um município deve multar toda indústria que poluir a água de rios e córregos com com nível médio de poluentes maior que 10ppm. Suponha que uma amostra de 30 alíquotas fora coletada nos arredores de uma indústria, revelando nível médio de poluente de 10,2ppm e desvio-padrão de 2ppm. Responda corretamente.

- Escreva o par de hipóteses para essa situação utilizando o parâmetro μ .
- Quais são as duas formas de acertar e as duas formas de errar, neste caso.
- Para você, qual é o pior tipo de erro? Sendo assim, qual significância você escolheria para o teste?
- Essa indústria deve ser multada? Termine de fazer o teste para responder essa pergunta de maneira embasada.

Resolução. (a) Vamos lembrar que já identificamos esse como um problema tipicamente unilateral à direita. Afinal, a indústria só será multada se o valor médio for *superior* a 10ppm. Outra observação. Não caia na tentação de pensar que a indústria deve ser multada porque a contaminação média *amostral* foi 10,2ppm! A regra é multar a indústria se a média populacional for maior que 10ppm, e não a amostral. Daí, precisamos fazer o teste de qualquer maneira. Isso posto, podemos escrever o par de hipóteses:

$$\begin{cases} H_0: \mu \leq 10 \\ H_1: \mu > 10 \end{cases}$$

(b) **Acertar:**

- * A poluição média ser realmente $\mu > 10$ e a indústria ser multada;
- * A poluição média ser $\mu \leq 10$ e a indústria não ser multada.

Errar:

- * Erro tipo I: Multar uma empresa inocente.
- * Erro tipo II: Não multar uma empresa culpada.

(c) Para mim, os dois erros são igualmente graves⁶. Portanto, eu escolho assumir $\alpha = 0,05$.

(d) Como o par de hipóteses já foi estabelecido, sigamos:

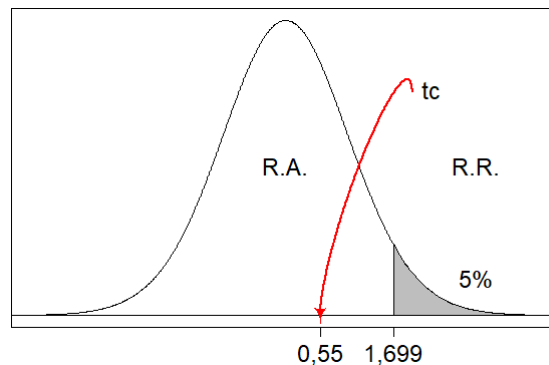
⁶ Você pode ter uma opinião diferente da nossa, ok?

Estatística de teste:

$$t_c = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{10,2 - 10}{2/\sqrt{30}} \approx 0,55 \quad (6.10)$$

Regra de decisão: como este é um teste unilateral à direita, devemos rejeitar H_0 se $t_c > t_\alpha$. Como já conhecemos t_c , vamos consultar o valor de $t_{0,05}$. Para isso, vamos lembrar que os graus de liberdade são $\nu = 30 - 1 = 29$ e, de acordo com a Tabela 36 do Apêndice B, $t_{0,05} = 1,699$. No R, seria:

```
qt(0.05, 29, lower.tail = F)
```



Dessa forma, temos que $t_c < t_\alpha$, e t_c está na região de aceitação de H_0 .

Conclusão: Com 5% de significância, pode-se afirmar que o nível médio de poluentes na água dos rios e córregos ao redor da indústria não é superior a 10ppm.

Código para reproduzir essa figura no R. Note que é necessário ter instalado o pacote `diagram` antes de rodar o script.

```
require(diagram)

x<-seq(-4, 4, by =.01)
y<-dt(x, df=29)
rx<-seq(1.699, 4, by =.1)
ry<-numeric(2*length(rx))
ry[1:length(rx)]<-dt(rx, df=29)
rx<-c(rx, rev(rx))
plot(x, y, 'l', xlab='', ylab='', xaxt="n", yaxt="n")
axis(1, at=c(0.55, 1.699), labels=c("0,55", "1,699"), cex.axis=1.5)
polygon(rx, ry, col = "gray")
abline(h=0, lty=1)
text(3, 0.05, labels="5%", cex=1.5)
text(0, 0.2, "R.A.", cex=1.5)
text(3, 0.2, "R.R.", cex=1.5)
text(2, 0.3, "tc", cex=1.5, pos=4)
curvedarrow(from = c(2, 0.3), to = c(0.55, 0),
curve = 0.05, arr.pos = 1, arr.col="red",
lcol='red')
```

Uma outra forma de fazermos esse teste é utilizando o próprio R. Observe que, neste problema, os dados amostrais não foram fornecidos. Foram informadas apenas as estatísticas amostrais. Sendo assim, podemos fazer como no script a seguir e obter o mesmo resultado.

```

alfa<-0.05
gl<-29
tc<-round((10.2 - 10) / (2/sqrt(30)), 2)
ta<-round(qt(alfa, gl, lower.tail=FALSE), 3)

```

6.5.5 Teste para uma proporção populacional (p)

6.5.5.1 Par de hipóteses

$$\left\{ \begin{array}{l} H_0: p = p_0 \\ H_1: p \neq p_0 \end{array} \right. \quad \text{ou} \quad \left\{ \begin{array}{l} H_0: p \leq p_0 \\ H_1: p > p_0 \end{array} \right. \quad \text{ou} \quad \left\{ \begin{array}{l} H_0: p \geq p_0 \\ H_1: p < p_0 \end{array} \right.$$

Bilateral

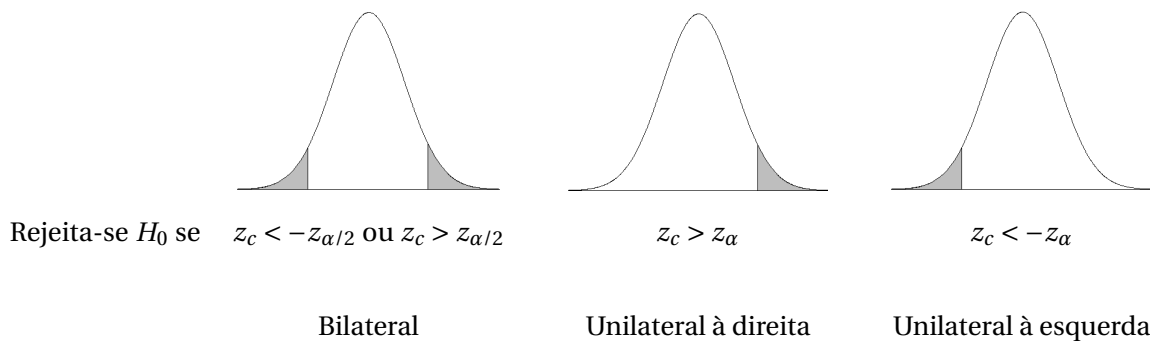
Unilateral à direita

Unilateral à esquerda

6.5.5.2 Estatística de teste

$$z_c = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}\hat{q}}{n}}} \quad (6.11)$$

6.5.5.3 Regra de decisão



Exemplo. Devido ao aumento da criminalidade, os moradores de um grande bairro estão cogitando contratar uma empresa de segurança particular. Entretanto, essa ação gerará custos. Uma assembleia foi feita, onde 40 pessoas compareceram. Dessas, 10 foram favoráveis. A contratação só acontecerá se mais de 50% dos moradores concordarem. Se considerarmos que os 40 moradores que compareceram na assembleia é uma amostra aleatória e representativa do bairro, a contratação deve ser feita?

Resolução. Primeiramente, vamos escrever o par de hipóteses desse problema. Se *mais* de 50% dos moradores do bairro concordarem uma ação será tomada: a contratação da empresa de vigilância. Caso contrário, não há contratação (não ação).

Par de hipóteses: Diante do contexto, este deve ser um teste unilateral à direita.

$$\begin{cases} H_0: p \leq p_0 \\ H_1: p > p_0 \end{cases}$$

Aqui, o erro tipo I (rejeitar H_0 verdadeira), é contratar sem que a maioria deseje. Por sua vez, o erro tipo II (aceitar H_0 falsa), é não contratar, sendo que a maioria deseja.

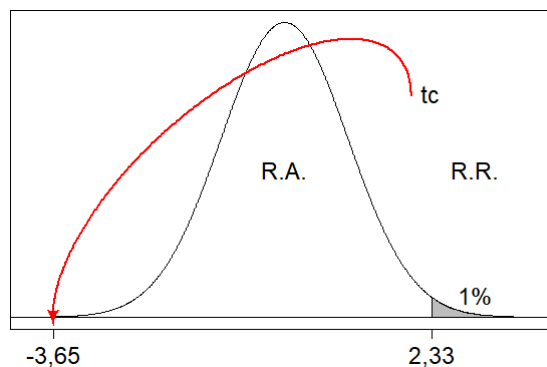
Considero que o erro tipo I é mais grave, e por isso desejo fazer o teste a $\alpha = 1\%$ de significância. Além disso, precisaremos calcular $\hat{p} = 10/40 = 0,25$. Logo, $\hat{q} = 0,75$.

Estatística de teste:

$$z_c = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}\hat{q}}{n}}} = \frac{0,25 - 0,50}{\sqrt{\frac{0,25(0,75)}{40}}} = -3,65$$

Regra de decisão: como este é um teste unilateral à direita, devemos rejeitar H_0 se $t_c > t_\alpha$. Pela Tabela 13 do Apêndice B, $z_{0,01} = 2,33$. Nesse caso, a probabilidade foi vista dentro da tabela, onde encontramos 0,0099 que é o mais próximo do que desejamos. Daí, vemos que ela está na linha 2,3 e coluna 3. Portanto, $z_{0,01} = 2,33$. No R, seria ainda mais simples:

```
qnorm(0.01, 0, 1, lower.tail = F)
```



Dessa forma, temos que $t_c < t_\alpha$, e t_c está na região de aceitação de H_0 .

Conclusão: Com 1% de significância, pode-se afirmar que a empresa de vigilância não deve ser contratada, uma vez que menos de 50% dos moradores concordam com isso.

Código para reproduzir essa figura no R.

```
require(diagram)
x<-seq(-4, 4, by = .01)
y<-dnorm(x, 0, 1)
rx<-seq(2.33, 4, by = .1)
```



```

ry<-numeric(2*length(rx))
ry[1:length(rx)]<-dnorm(rx, 0, 1)
rx<-c(rx, rev(rx))
plot(x, y, 'l', xlab='', ylab='', xaxt="n", yaxt="n")
axis(1,at=c(-3.65,2.33),labels=c("-3,65","2,33"),cex.axis=1.5)
polygon(rx, ry, col = "gray")
abline(h=0, lty=1)
text(3, 0.03, labels="1%",cex=1.5)
text(0,0.2,"R.A.",cex=1.5)
text(3,0.2,"R.R.",cex=1.5)
text(2,0.3,"tc",cex=1.5,pos=4)
curvedarrow(from = c(2,0.3),to = c(-3.65,0),
curve = 0.03, arr.pos = 1, arr.col="red",
lcol='red')

```

Fazer esse teste no R é também muito fácil. Usaremos, pela primeira vez, a função `prop.test()`. Essa função apresenta no seu *output* um intervalo de confiança e um teste de hipótese. Como aqui estamos interessados no teste, vamos fixar $p = 0.5$ e `alternative = "greater"` para designar um teste unilateral à direita.

```
prop.test(10, 40, p=0.5, alternative="greater")
```

Você verá que uma das informações do *output* é $p\text{-value} = 0.9987$. É disso que precisamos para concluir nosso teste. Equivalentemente a comparar t_c com t_α , o *valor-p* tem interpretação fácil e direta. Ele representa a área acima de t_c . Ora, se 99,87% da área está acima de t_c , isso quer dizer que t_c está bem à esquerda da nossa normal, ou seja, na região de aceitação. Logo, aceitamos H_0 . Na prática, apenas comparamos o valor-p com nosso nível de significância. Como valor-p é maior que α , aceitamos H_0 .

6.5.6 Estimação da variância populacional (σ^2)

Como era de se esperar, o estimador pontual da variância populacional (σ^2) é a variância amostral (S^2). Por outro lado, pela primeira vez veremos uma estimação intervalar baseada em uma distribuição assimétrica: a distribuição qui-quadrado χ^2 .

- **Por ponto:** $\hat{\sigma}^2 = S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$
- **Por intervalo:** $IC_\gamma(\sigma^2) = \left[\frac{(n-1)S^2}{\chi_{\alpha/2,v}^2}; \frac{(n-1)S^2}{\chi_{1-\alpha/2,v}^2} \right]$

em que σ^2 é variância populacional; S^2 é a variância amostral; n é o tamanho da amostra; $\chi_{\alpha/2,v}^2$ e $\chi_{1-\alpha/2,v}^2$ são quantis da distribuição χ^2 com v graus de liberdade.

6.5.7 Estimação do desvio-padrão populacional (σ)

Além do estimador pontual do desvio-padrão populacional ser o desvio-padrão amostral, o seu estimador intervalar é muito intuitivo, para quem já conhece o estimador intervalar da variância. Basta tomarmos sua raiz quadrada.

- **Por ponto:** $\hat{\sigma} = S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$
- **Por intervalo:** $IC_{\gamma}(\sigma^2) = \left[\sqrt{\frac{(n-1)S^2}{\chi_{\alpha/2, \nu}^2}}; \sqrt{\frac{(n-1)S^2}{\chi_{1-\alpha/2, \nu}^2}} \right]$

em que σ^2 é variância populacional; S^2 é a variância amostral; n é o tamanho da amostra; $\chi_{\alpha/2, \nu}^2$ e $\chi_{1-\alpha/2, \nu}^2$ são quantis da distribuição χ^2 com ν graus de liberdade.

6.6 Inferência sobre duas populações normais

Suponha duas distribuições normais. A população 1 tem média μ_1 e variância σ_1^2 . A população 2 tem média μ_2 e variância σ_2^2 . Dessas populações retiram-se amostras de tamanho n_1 e n_2 , respectivamente, com as quais podemos calcular estatísticas amostrais como média amostral (\bar{x}_1 e \bar{x}_2), variância amostral (s_1^2 e s_2^2), proporções amostrais (\hat{p}_1 e \hat{p}_2), dentre outras (Figura 30).

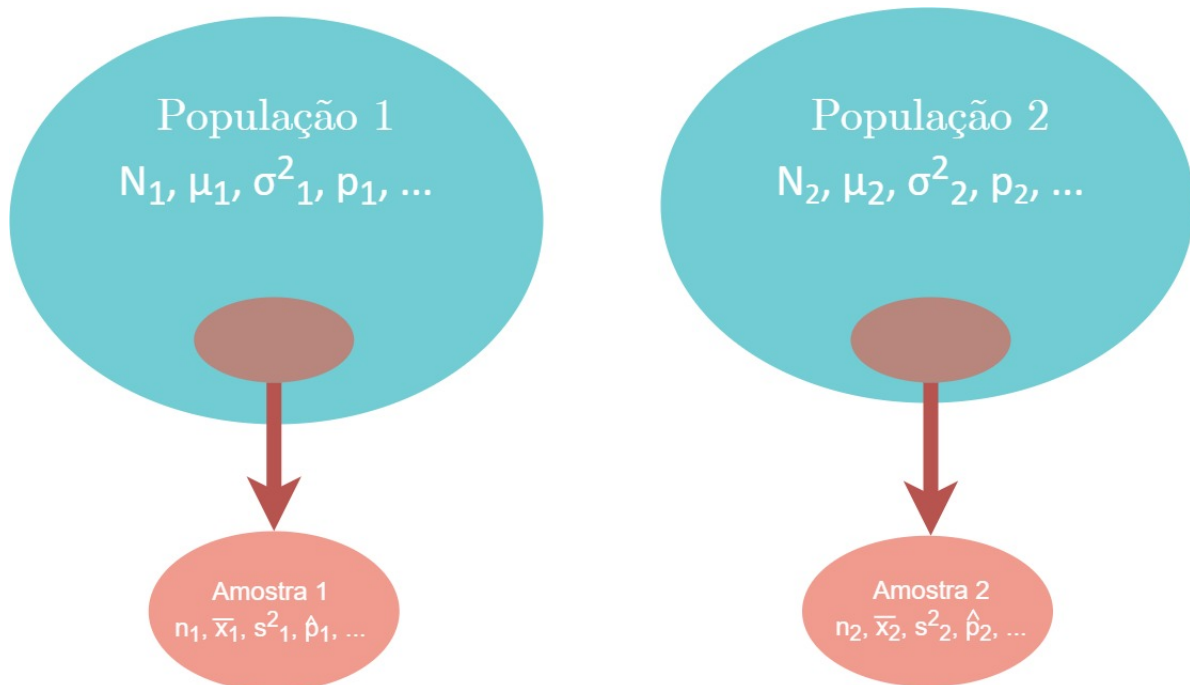


Figura 30 – Esquema ilustrativo da amostragem de duas populações normais, parâmetros populacionais e estatísticas amostrais.

Fonte: Do autor.

6.6.1 Teste de homogeneidade de variâncias

No jargão estatístico, dizemos que duas médias são iguais e que duas variâncias são *homogêneas*. Então, um teste de homogeneidade de variâncias testa, tão somente, a igualdade entre variâncias. No campo da Estatística Experimental temos a necessidade de testar a homogeneidade de mais de duas variâncias. Aqui, fora do contexto experimental, vamos nos concentrar em comparar apenas duas variâncias.

Uma vez que estamos diante de duas populações normais, temos um resultado importante. A razão entre as variâncias amostrais segue distribuição F de Snedecor, com $n_1 - 1$ e $n_2 - 1$ graus de liberdade.

6.6.1.1 Distribuição F

A distribuição de F de Snedecor é uma distribuição contínua, que recebeu este nome em homenagem ao biólogo e estatístico britânico Ronald Fisher e ao matemático norte-americano George Waddel Snedecor.

Trata-se de uma distribuição assimétrica à direita, ou seja, tem uma longa cauda à direita (Figura 31). O domínio da função é não negativo. Só está definida para $X \geq 0$. Como a razão entre variâncias nunca será um valor negativo, é adequada para modelar esse fenômeno.

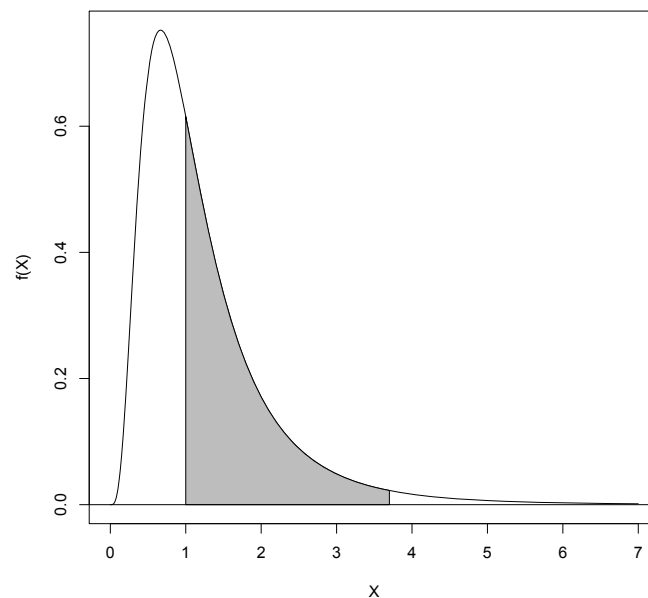


Figura 31 – Distribuição F de probabilidade ressaltando a região de aceitação de um teste de homogeneidade de variâncias (de 1 até um F_c qualquer).

Fonte: Do autor.

Veja a rotina utilizada para fazer a Figura 31.

```
x<-seq(0, 7, by=.01)
y<-df(x, 10, 10)
plot(x, y, "l", xlab= 'X', ylab='f(X)')
```

```

rx<-seq(1, 3.7, by =.1)
ry<-vector('numeric', 2*length(rx))
ry[1:length(rx)]<-df(rx, 10, 10)
rx<-c(rx, rev(rx))
plot(x, y, 'l', xlab='X', ylab='f(X)')
polygon(rx, ry, col = "gray")
abline(h=0)

```

Como todo teste, o teste F para homogeneidade de variâncias apresenta as quatro partes: par de hipóteses, estatística de teste, regra de decisão e conclusão.

Par de hipóteses

Considerando as definições das hipóteses H_0 e H_1 , o teste F deseja testar

$$\left\{ \begin{array}{l} H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1 \\ H_1: \frac{\sigma_1^2}{\sigma_2^2} \neq 1 \end{array} \right. \quad \text{ou} \quad \left\{ \begin{array}{l} H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1 \\ H_1: \frac{\sigma_1^2}{\sigma_2^2} > 1 \end{array} \right. \quad \text{ou} \quad \left\{ \begin{array}{l} H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1 \\ H_1: \frac{\sigma_1^2}{\sigma_2^2} < 1 \end{array} \right.$$

Bilateral

Unilateral à direita

Unilateral à esquerda

O que corresponde a

$$\left\{ \begin{array}{l} H_0: \sigma_1^2 = \sigma_2^2 \\ H_1: \sigma_1^2 \neq \sigma_2^2 \end{array} \right. \quad \text{ou} \quad \left\{ \begin{array}{l} H_0: \sigma_1^2 = \sigma_2^2 \\ H_1: \sigma_1^2 > \sigma_2^2 \end{array} \right. \quad \text{ou} \quad \left\{ \begin{array}{l} H_0: \sigma_1^2 = \sigma_2^2 \\ H_1: \sigma_1^2 < \sigma_2^2 \end{array} \right.$$

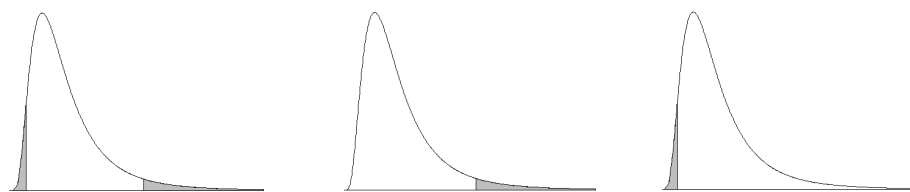
Estatística de teste

$$F_c = \frac{S_1^2}{S_2^2} \quad (6.12)$$

Regra de decisão

Os quantis da distribuição F, com os quais devemos comparar nossa estatística de teste F_c , podem ser encontrados nas tabelas de 16 a 35 do Apêndice B. Para consultá-las, devemos ter conhecimento do valor:

- (1) da probabilidade, informação essa que designa qual tabela olhar;
- (2) dos graus de liberdade do numerador, dados por $\nu_1 = n_1 - 1$, que indicam a coluna que deve ser consultada; e
- (3) dos graus de liberdade do denominador, dados por $\nu_2 = n_2 - 1$, que indicam a linha que devemos olhar.



Rejeita-se H_0 se $F_c < F_{1-\alpha/2}$ ou $F_c > F_{\alpha/2}$

$F_c > F_\alpha$

$F_c < F_{1-\alpha}$

Bilateral

Unilateral à direita

Unilateral à esquerda

Exemplo. Produtores de cinema de Hollywood monitoraram o número de espectadores de dois filmes lançados nos Estados Unidos, nos dez primeiros dias de exibição. Veja o número diário de espectadores, no mundo todo, em milhões de pessoas:

Tabela 8 – Número diário de espectadores de dois filmes, em milhões de pessoas.

Filme A	Filme B
2,06	2,20
2,12	2,76
2,35	2,85
2,13	3,07
2,22	3,12
2,11	3,27
1,81	3,59
2,02	2,74
2,19	3,28
2,28	2,51

De posse desta amostra, responda corretamente:

- (a) As variâncias, do número de espectadores dos filmes A e B, podem ser ditas homogêneas (iguais)?
 (b) E o número médio de espectadores por dia, é o mesmo para os dois filmes?

Resolução. Nesse problema, estamos considerando os dez primeiros dias de exibição como uma amostra aleatória e significativa de todos os dias que o filme está sendo ou será exibido. Neste primeiro momento, responderemos apenas a pergunta (a). A pergunta (b) será respondida posteriormente.

(a) Como não há indicação, no contexto ou da pergunta, de direção para H_1 , trata-se de um teste bilateral. Assim, nossa população 1 será o filme A, e a população 2 será o filme B.

Par de hipóteses:

$$\left\{ \begin{array}{l} H_0: \frac{\sigma_A^2}{\sigma_B^2} = 1 \\ H_1: \frac{\sigma_A^2}{\sigma_B^2} \neq 1 \end{array} \right.$$

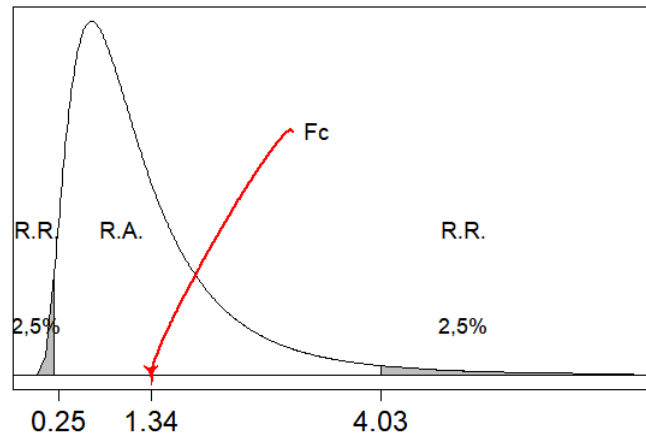
Estatística de teste:

$$F_c = \frac{0,0225}{0,1677} = 1,34$$

Regra de decisão: temos que escolher um nível de significância para esse teste. Como os dois tipos são, nossa opinião, igualmente graves, escolhemos $\alpha = 5\%$. Sabemos que os graus de liberdade do numerador e do denominador são iguais a 9, ou seja, $\nu_A = \nu_B = 9$. Então, vamos consultar as tabelas F para encontrar os quantis superiores $F_{\alpha/2}$ e $F_{1-\alpha/2}$. Nesse caso, consultando a Tabela 20 do Apêndice B, temos que $F_{0,025} = 4,03$; e consultando a Tabela 30 do Apêndice B, temos que $F_{0,975} = 0,2484$. No R, também seria muito fácil encontrar esses quantis:

```
qf(0.025, 9, 9, lower.tail = F)
```

```
qf(0.975, 9, 9, lower.tail = F)
```



Dessa forma, temos que $F_{1-\alpha/2} < F_c < F_{\alpha/2}$, e está na região de aceitação de H_0 .

Conclusão: Com 5% de significância, pode-se afirmar que as variâncias populacionais, do número diário de espectadores dos dois filmes, são iguais.

Código para reproduzir essa figura no R.

```
require(diagram)
x<-seq(0, 7, by =.01)
y<-df(x, 9, 9)
rx<-seq(4.03, 7, by =.1)
ry<-numeric(2*length(rx))
ry[1:length(rx)]<-df(rx, 9, 9)
rx<-c(rx, rev(rx))
plot(x, y, 'l', xlab='', ylab='', xaxt="n", yaxt="n")
axis(1, at=c(0.2484,1.34,4.03),
labels=c("0.25", "1.34", "4.03"),
cex.axis=1.5)
polygon(rx, ry, col = "gray")
rx<-seq(0, 0.25, by =.1)
ry<-numeric(2*length(rx))
ry[1:length(rx)]<-df(rx, 9, 9)
rx<-c(rx, rev(rx))
polygon(rx, ry, col = "gray")
abline(h=0, lty=1)
text(5, 0.1, labels="2,5%",cex=1.1)
text(0, 0.1, labels="2,5%",cex=1.1)
text(1,0.3,"R.A.",cex=1.2)
text(0,0.3,"R.R.",cex=1.2)
text(5,0.3,"R.R.",cex=1.2)
text(3,0.5,"Fc",cex=1.2,pos=4)
```

```

curvedarrow(from = c(3, 0.5), to = c(1.34, 0),
curve = 0.03, arr.pos = 1, arr.col="red",
lcol='red')

```

6.6.2 Estimação da diferença entre duas médias ($\mu_1 - \mu_2$)

Ao contrário da inferência sobre duas variâncias, pela qual consideramos a razão, na inferência sobre duas médias consideramos a diferença entre elas. Essa diferença se deve à distribuição de probabilidade mais adequada. Sendo as populações normais, a diferença entre médias segue distribuição normal ou t , a julgar pelo conhecimento das variâncias populacionais.

Mais uma vez, ressaltamos que assumir as variâncias populacionais conhecidas pode ser muito extremista e artificial. Porém, o apresentaremos pois este é um resultado importante.

6.6.2.1 Variâncias populacionais conhecidas

Uma vez que são conhecidas σ_1^2 e σ_2^2 , avaliamos se elas são iguais ou diferentes, pois isso interfere na estimação intervalar da diferença entre as médias. Em ambos os casos a estimação por ponto é a mesma.

- **Por ponto:**

$$\widehat{\mu_1 - \mu_2} = \hat{\mu}_1 - \hat{\mu}_2 = \bar{X}_1 - \bar{X}_2$$

- **Por intervalo:**

$$\sigma_1^2 \neq \sigma_2^2$$

$$IC_\gamma(\mu_1 - \mu_2) = \left[(\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right].$$

$$\sigma_1^2 = \sigma_2^2 = \sigma^2$$

$$IC_\gamma(\mu_1 - \mu_2) = \left[(\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2} \sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \right].$$

6.6.2.2 Variâncias populacionais desconhecidas

No caso em que não são conhecidas as variâncias populacionais, a tendência natural é utilizar seus estimadores. S^2 é o estimador de σ^2 . Porém, a distribuição de $\bar{X}_1 - \bar{X}_2$ não é mais normal, mas sim uma t . Mesmo quando as variâncias populacionais são desconhecidas, deve-se decidir por considerá-las iguais ou diferentes: fazemos isso pelo teste F. Novamente, a estimação pontual permanece inalterada, e temos diferença na estimação intervalar.

- **Por ponto:**

$$\widehat{\mu_1 - \mu_2} = \hat{\mu}_1 - \hat{\mu}_2 = \bar{X}_1 - \bar{X}_2$$

• **Por intervalo:**

Quando as variâncias populacionais são diferentes, as variâncias amostrais não podem ser combinadas e os graus de liberdade da distribuição t precisam de correção.

$$\sigma_1^2 \neq \sigma_2^2$$

$IC_\gamma(\mu_1 - \mu_2)$	ν
----------------------------	-------

$\left[(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right].$	$\frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)^2}{\frac{\left(\frac{S_1^2}{n_1} \right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2} \right)^2}{n_2 - 1}}$
---	--

Por sua vez, quando as variâncias populacionais são homogêneas, a correção nos graus de liberdade da distribuição t não precisa ser feita, e as variâncias amostrais podem ser combinadas em uma variância chamada *pooled* (S_p^2).

$$\sigma_1^2 = \sigma_2^2 = \sigma^2$$

$IC_\gamma(\mu_1 - \mu_2)$	ν	S_p^2
----------------------------	-------	---------

$\left[(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2} \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \right].$	$\nu = n_1 + n_2 - 2.$	$\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$
--	------------------------	---

6.6.3 Teste sobre a diferença entre duas médias ($\mu_1 - \mu_2$)

6.6.3.1 Par de hipóteses

$$\left\{ \begin{array}{l} H_0: \mu_1 - \mu_2 = 0 \\ H_1: \mu_1 - \mu_2 \neq 0 \end{array} \right. \quad \text{ou} \quad \left\{ \begin{array}{l} H_0: \mu_1 - \mu_2 = 0 \\ H_1: \mu_1 - \mu_2 > 0 \end{array} \right. \quad \text{ou} \quad \left\{ \begin{array}{l} H_0: \mu_1 - \mu_2 = 0 \\ H_1: \mu_1 - \mu_2 < 0 \end{array} \right.$$

Bilateral

Unilateral à direita

Unilateral à esquerda

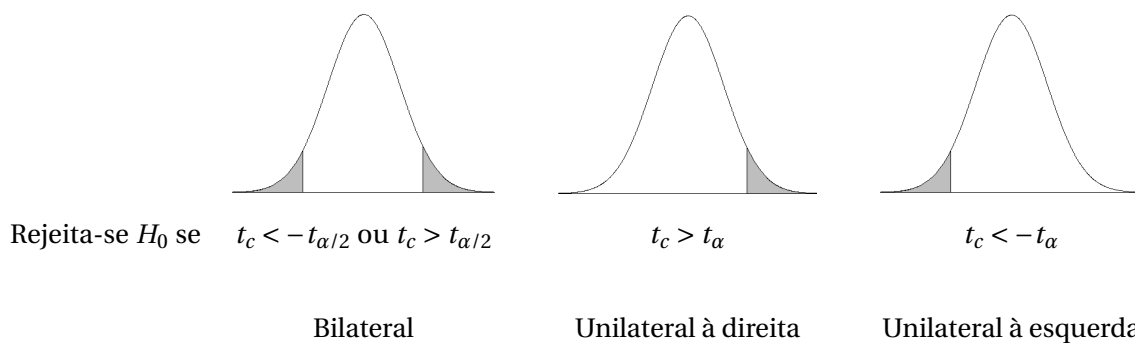
O que corresponde a

$$\left\{ \begin{array}{l} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 \neq \mu_2 \end{array} \right. \quad \text{ou} \quad \left\{ \begin{array}{l} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 > \mu_2 \end{array} \right. \quad \text{ou} \quad \left\{ \begin{array}{l} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 < \mu_2 \end{array} \right.$$

6.6.3.2 Estatística de teste

Variâncias conhecidas		
	$\sigma_1^2 = \sigma_2^2 = \sigma^2$	$\sigma_1^2 \neq \sigma_2^2$
Estatística	$z_c = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$	$z_c = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
Variâncias desconhecidas		
	$\sigma_1^2 = \sigma_2^2 = \sigma^2$	$\sigma_1^2 \neq \sigma_2^2$
Estatística	$t_c = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$	$t_c = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$
S_p^2	$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$	
ν	$n_1 + n_2 - 2$	$\frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)^2}{\frac{\left(\frac{S_1^2}{n_1} \right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2} \right)^2}{n_2 - 1}}$

6.6.3.3 Regra de decisão



Exemplo. Vamos retomar o exemplo anterior, dos filmes de Hollywood, para respondermos a letra (b).

Resolução. (b) Sabemos que as variâncias populacionais não são conhecidas, mas podem ser ditas homogêneas, pelo teste F que fizemos na letra (a). Sendo assim, sabemos qual estatística de teste utilizar. Além disso, não há indicação de direção para construir o par de hipóteses. Logo, será um teste bilateral.

Par de hipóteses:

$$\begin{cases} H_0: \mu_1 - \mu_2 = 0 \\ H_1: \mu_1 - \mu_2 \neq 0 \end{cases}$$

Estatística de teste: Precisaremos da variância combinada S_p^2 mas, para isso, precisamos das duas variâncias amostrais. Temos que $S_A^2 = 0,0225$ e $S_B^2 = 0,1677$. Portanto,

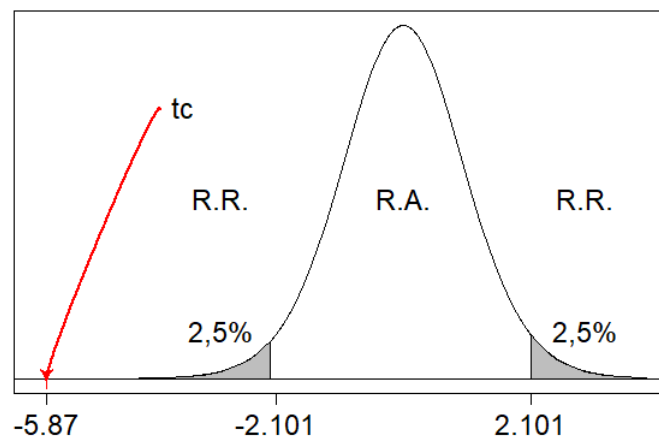
$$S_p^2 = \frac{9(0,0225) + 9(0,1677)}{18} = 0,0951$$

Daí,

$$t_c = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{2,129 - 2,939}{\sqrt{0,0951 \left(\frac{1}{10} + \frac{1}{10} \right)}} = -5,87$$

Regra de decisão: Adotaremos o nível de significância de $\alpha = 0,05$, pois os dois erros são igualmente graves. Sendo assim, podemos obter os quantis da Tabela 36 do Apêndice B ou do R. Temos que os graus de liberdade são $\nu = n_A + n_B - 2 = 18$, e $t_{0,025} = 2,101$ e $-t_{0,025} = -2,101$.

```
qt(0.025, 18, lower.tail = F)
```



Dessa forma, temos que $t_c < -t_{\alpha/2}$, e está na região de rejeição de H_0 .

Conclusão: Com 5% de significância, pode-se afirmar que o número médio de espectadores diários dos filmes A e B não são iguais.

Código para reproduzir essa figura no R.

```
require(diagram)
x<-seq(-6, 4, by =.01)
y<-dt(x, 18)
```

```

rx<-seq(2.101, 4, by =.1)
ry<-numeric(2*length(rx))
ry[1:length(rx)]<-dt(rx, 18)
rx<-c(rx, rev(rx))
plot(x, y, 'l', xlab='', ylab='', xaxt="n", yaxt="n")
axis(1,at=c(-5.87,-2.101,2.101),
labels=c("-5.87","-2.101","2.101"),
cex.axis=1.5)
polygon(rx, ry, col = "gray")
rx<-seq(-6, -2.101, by =.1)
ry<-numeric(2*length(rx))
ry[1:length(rx)]<-dt(rx, 18)
rx<-c(rx, rev(rx))
polygon(rx, ry, col = "gray")
abline(h=0, lty=1)
text(3, 0.05, labels="2,5%",cex=1.5)
text(-3, 0.05, labels="2,5%",cex=1.5)
text(0,0.2,"R.A.",cex=1.5)
text(3,0.2,"R.R.",cex=1.5)
text(-3,0.2,"R.R.",cex=1.5)
text(-4,0.3,"tc",cex=1.5,pos=4)
curvedarrow(from = c(-4,0.3),to = c(-5.87,0),
curve = 0.01, arr.pos = 1, arr.col="red",
lcol='red')

```

E se quisermos fazer esse teste no R? Existe uma forma ainda mais fácil e rápida para isso! Primeiro, vamos atribuir os dados a vetores `a` e `b`, em seguida, usar a função `t.test()`. Veja:

```

require(diagram)
a<-c(2.06,2.12,2.35,2.13,2.22,2.11,1.81,2.02,2.19,2.28)
b<-c(2.20,2.76,2.85,3.07,3.12,3.27,3.59,2.74,3.28,2.51)
t.test(a,b)

```

A função `t.test()` também tem a característica de realizar o teste `t` e o intervalo de confiança, nesse caso, para a diferença entre médias. Veja a saída da função:

```

Welch Two Sample t-test

data:  a and b
t = -5.8733, df = 11.372, p-value = 9.382e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-1.112335 -0.507665
sample estimates:
mean of x mean of y
2.129      2.939

```

Para interpretá-la, vamos nos concentrar no valor-p e no intervalo de confiança. O valor-p ($9,382 \times 10^{-5}$), menor que o nível de significância adotado ($\alpha/2 = 0,025$), indica que devemos rejeitar H_0 . Portanto, o número médio de espectadores diários não é o mesmo para cada filme.

Pelas médias amostrais, vemos que o filme B tem mais espectadores diários (2,939 milhões) do que o filme A (2,129 milhões), na amostra. Mas, e na população? Olhe o intervalo de 95% de confiança produzido: [-1,112 ; -0,508]. Ele foi feito considerando a diferença paramétrica $\mu_A - \mu_B$. Isso quer dizer que, com 95% de confiança, a diferença entre as duas médias populacionais está entre 1,112 e 0,508. Em outras palavras, o filme B tem de 508 mil à 1,112 milhão de espectadores diários *a mais* que o filme A.

6.6.4 O caso dos dados emparelhados

Suponha que tenhamos dados sabidamente normais mas, ao contrário de duas populações independentes, tratam-se de dados vindos das mesmas unidades amostrais *antes* e *após* uma intervenção acontecer. Esses dados são chamados *emparelhados*. Portanto, temos duas populações *dependentes*, mas ainda assim, normais.

Dados emparelhados São dados tomados no mesmo elemento amostral, antes e após uma intervenção.

Como exemplo de dados emparelhados, podemos pensar em medidas do peso de pacientes obesos antes e depois de uma dieta; número de insetos por planta antes e após a aplicação de um inseticida; a intenção de votos em um candidato antes e após um debate, entre outros.

Nessas situações, podemos estimar a magnitude do efeito da intervenção, estimando a média das diferenças $X_{\text{antes}} - X_{\text{depois}}$ entre o *antes* e o *depois*. Por outro lado, se tivermos uma suspeita ou hipótese sobre a magnitude desse efeito, podemos testar se a diferença média tem certo valor hipotético.

Como tratam-se de dados normais, a diferença entre eles também segue distribuição normal. Nesse caso, é como se voltássemos a ter uma só população, a *população das diferenças*. Então, para ambas as abordagens de estimação e decisão, precisamos novamente utilizar as distribuições normal e t de Student.

6.6.4.1 Estimação do efeito médio da intervenção (μ_d)

Diante do que compreendemos, suponha que as diferenças sigam distribuição normal, para a qual desejamos inferir sobre a média. Portanto, desconhecemos o valor da média populacional. Entretanto, a variância dessa distribuição normal, σ_d^2 , pode ser conhecida ou desconhecida. Assim, temos uma mesma forma de estimar a média por ponto, mas duas formas de estimá-la por intervalo.

- **Por ponto:** $\hat{\mu} = \bar{d} = \frac{\sum_{i=1}^n d_i}{n}$,
em que $d_i = X_{i,\text{antes}} - X_{i,\text{depois}}$, para $i = 1, 2, \dots, n$.

- **Por intervalo:**

σ_d^2 conhecida	σ_d^2 desconhecida
$IC_{\gamma}(\mu_d) = \left[\bar{d} \pm z_{\alpha/2} \frac{\sigma_d}{\sqrt{n}} \right]$	$IC_{\gamma}(\mu_d) = \left[\bar{d} \pm t_{\alpha/2} \frac{S_d}{\sqrt{n}} \right]$

Exemplo. Pacientes obesos, 1 ano após cirurgia bariátrica do tipo *sleeve*, costumam estabilizar em um peso muito inferior ao inicial. Ainda assim, há casos em que o paciente recupera boa parte ou todo o peso perdido no período pós-operatório. Um grupo de 15 pacientes foram acompanhados, e tiveram seu peso registrado antes da cirurgia e após 1 ano. Dessa forma podemos calcular as diferenças no peso. Todos os dados estão em quilogramas:

Antes	141	197	205	167	198	174	225	164	213	194	164	225	132	169	169
Após	96	74	98	109	74	99	89	100	87	117	96	91	97	113	95
Diferença	45	123	107	58	124	75	136	64	126	77	68	134	35	56	74

Sabendo disto, estime o efeito médio da cirurgia bariátrica por ponto e por intervalo com 95% de confiança.

Resolução. Primeiramente, calculamos as estatísticas amostrais, nos concentrando na amostra de diferenças (terceira linha da tabela de dados). Temos que $\bar{d} = 86,8\text{kg}$, $s_d = 34,61\text{kg}$, $n = 15$ e, para 95% de confiança, $t_{\alpha/2} = 2,145$.

- **Por ponto:** $\hat{\mu}_d = 86,8\text{kg}$

- **Por intervalo:**

$$\begin{aligned}
 IC_{95\%}(\mu_d) &= 86,8 \pm 2,145 \frac{34,61}{\sqrt{15}} \\
 &= 86,8 \pm 19,16741 \\
 &= [67,6; 106,0]
 \end{aligned}$$

Conclusão. Com 95% de confiança, podemos afirmar que a perda de peso média, devido à técnica *sleeve*, varia de 67,6 a 106kg.

Para resolver esse mesmo exercício no R, insira os dados fornecidos, calcule as estatísticas e proceda a estimação. Essa última pode ser feita da maneira mais detalhada ou mais direta, utilizando a função `t.test()`.

```

a<-c(141,197,205,167,198,174,225,164,213,194,164,225,132,169,169)
b<-c(96,74,98,109,74,99,89,100,87,117,96,91,97,113,95)
d<-a-b
#Estimação por ponto
(md<-mean(d))
#Estimação por intervalo
(Sd<-sd(d))

```

```

ta2<-qt(0.025, 14, lower.tail=F)
md=ta2*Sd/sqrt(15)
md+ta2*Sd/sqrt(15)
#Ou
t.test(d)

```

6.6.4.2 Teste sobre o efeito médio da intervenção (μ_d)

Nesta situação, há uma suspeita sobre o valor do efeito médio da intervenção. Essa suspeita ($\mu_{d,0}$) pode ser testada por meio de um teste t ou teste Z, a depender do conhecimento da variância populacional das diferenças. Sendo assim, seguimos os passos:

Par de hipóteses. Uma suspeita muito comum para ser usada neste teste é o $\mu_{d,0} = 0$. Ou seja, é útil testar se a intervenção foi sem efeito (já que a média das diferenças poderia ser zero). Porém, aqui expressaremos o caso mais geral, onde a suspeita é um valor qualquer $\mu_{d,0}$.

$$\left\{ \begin{array}{l} H_0: \mu_d = \mu_{d,0} \\ H_1: \mu_d \neq \mu_{d,0} \end{array} \right. \quad \text{ou} \quad \left\{ \begin{array}{l} H_0: \mu_d \leq \mu_{d,0} \\ H_1: \mu_d > \mu_{d,0} \end{array} \right. \quad \text{ou} \quad \left\{ \begin{array}{l} H_0: \mu_d \geq \mu_{d,0} \\ H_1: \mu_d < \mu_{d,0} \end{array} \right.$$

Bilateral

Unilateral à direita

Unilateral à esquerda

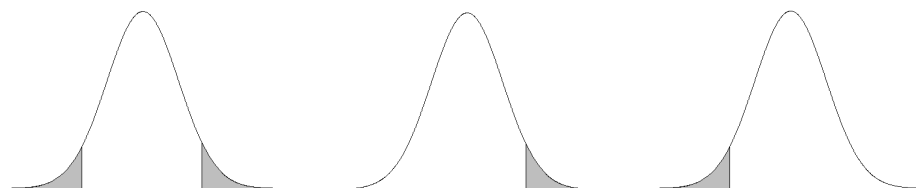
Estatística de teste. A julgar pelo conhecimento da variância populacional σ_d^2 , a estatística de teste pode assumir as seguintes formas:

 σ_d^2 conhecida σ_d^2 desconhecida

$$z_c = \frac{\bar{d} - \mu_{d,0}}{\sigma_d / \sqrt{n}}$$

$$t_c = \frac{\bar{d} - \mu_{d,0}}{S_d / \sqrt{n}}$$

Regra de decisão. Como é muito mais comum e realista o não conhecimento da variância populacional, a regra de decisão apresentada aqui será aquela baseada na distribuição t.

Rejeita-se H_0 se $t_c < -t_{\alpha/2}$ ou $t_c > t_{\alpha/2}$ $t_c > t_\alpha$ $t_c < -t_\alpha$

Bilateral

Unilateral à direita

Unilateral à esquerda

Exemplo. Vamos voltar ao exemplo da cirurgia bariátrica. Um artigo científico divulgou que o método *sleeve* proporciona, em média, uma perda de peso de *pelo menos* 90kg. Se houver evidência de que essa afirmação é falsa, será enviada uma carta ao editor da revista.

Teste, com 5% de significância, se o dado divulgado pelo artigo está correto.

Resolução. Lembremos que as estatísticas amostrais desse conjunto de dados são: $\bar{d} = 86,8\text{kg}$, $s_d = 34,61\text{kg}$, $n = 15$ e, para 5% de significância em um teste unilateral, $t_\alpha = 1,761$. Esse valor pode ser verificado na Tabela 36 ou ainda no R.

```
qt(0.05, 14)
```

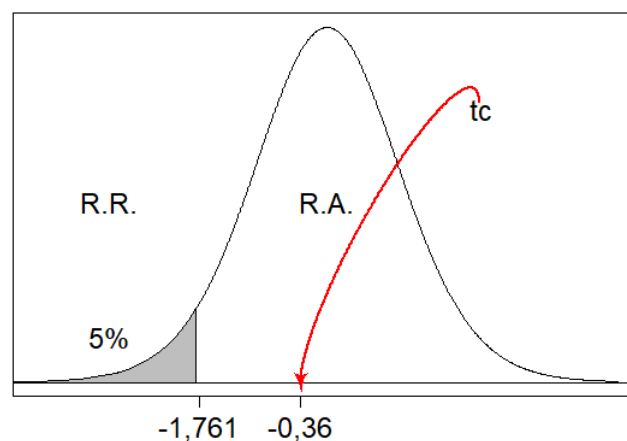
Par de hipóteses: Aqui, vale chamar atenção para o fato de que será tomada uma atitude somente se $\mu_d < 90\text{kg}$. Se $\mu_d \geq 90\text{kg}$, o artigo terá feito uma afirmação correta, e nada será feito. Como H_0 é a hipótese de *não ação*, o teste deve ser unilateral e deve ser, claramente:

$$\begin{cases} H_0: \mu_d \geq 90 \\ H_1: \mu_d < 90 \end{cases}$$

Estatística de teste: Neste exemplo conhecemos apenas a variância amostral, logo, faremos um teste t . Dessa forma, temos a seguinte estatística de teste.

$$t_c = \frac{86,8 - 50}{34,61/\sqrt{15}} = -0,36$$

Regra de decisão: Na distribuição t de Student, posicionamos o valor crítico $-t_\alpha$ e a estatística de teste t_c . Assim, vemos que não há evidências para rejeitar a hipótese nula, pois $t_c > -t_\alpha$.



Conclusão: Com 5% de significância, pode-se afirmar que não evidências para afirmar que o artigo está errado em sua afirmação, ou seja, a perda de peso média proporcionada pela técnica *sleeve* é maior ou igual a 90kg.

Código para reproduzir essa figura no R.

```
require(diagram)
x<-seq(-4, 4, by =.01)
y<-dt(x, 18)
rx<-seq(-5, -1.761, by =.1)
ry<-numeric(2*length(rx))
ry[1:length(rx)]<-dt(rx, 18)
rx<-c(rx, rev(rx))
plot(x, y, 'l', xlab='', ylab='', xaxt="n", yaxt="n")
axis(1, at=c(-1.761, -0.36),
labels=c("-1,761", "-0,36"),
```

```

cex.axis=1.5)
polygon(rx, ry, col = "gray")
abline(h=0, lty=1)
text(-3, 0.05, labels="5%", cex=1.5)
text(0, 0.2, "R.A.", cex=1.5)
text(-3, 0.2, "R.R.", cex=1.5)
text(1.8, 0.3, "tc", cex=1.5, pos=4)
curvedarrow(from = c(2.1, 0.31), to = c(-0.36, 0),
            curve = 0.03, arr.pos = 1, arr.col="red",
            lcol='red')

```

Para resolver este teste no R, podemos fazê-lo pela via mais demorada e detalhada ou pela via rápida, utilizando novamente a função `t.test`. Veja que, por esta função, devemos analisar o *valor-p*. Como ele foi maior que o nível de significância do teste ($p\text{-value} = 0,3628 > 0,05$) não devemos rejeitar H_0 .

```

a<-c(141, 197, 205, 167, 198, 174, 225, 164, 213, 194, 164, 225, 132, 169, 169)
b<-c(96, 74, 98, 109, 74, 99, 89, 100, 87, 117, 96, 91, 97, 113, 95)
d<-a-b
#Teste de H0: mu_d >= 50
md<-mean(d)
Sd<-sd(d)
(ta<-qt(0.05, 14))
(tc<-(md-90)/(Sd/sqrt(15)))
# tc > tc
#Ou, simplesmente,
t.test(d, mu=90, alternative="less")

```

REGRESSÃO E CORRELAÇÃO

Regressão: é o estudo que busca ajustar uma equação a um conjunto de dados de forma que a relação entre variáveis possa ser descrita por uma função. Essa função busca explicar a variação de uma variável-resposta por meio de variáveis explicativas ou covariáveis.

Correlação: reflete a associação ou relacionamento entre duas variáveis.

7.1 Regressão

Um estudo de regressão busca essencialmente associar uma variável Y (denominada variável-resposta) a um conjunto de outras p variáveis X_1, X_2, \dots, X_p (denominadas covariáveis ou variáveis explicadoras). Esta associação se dá segundo uma forma funcional do tipo

$$Y = f(X_1, X_2, \dots, X_p),$$

na qual a função f pode ser, à princípio, qualquer uma. Quando f assume a forma funcional linear (isto é, f é uma combinação linear das covariáveis

$$f(X_1, X_2, \dots, X_p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p,$$

em que os coeficientes β_i ($i = 1, 2, \dots, p$) são números fixos chamados parâmetros), a regressão é chamada linear. Caso contrário, é uma regressão não-linear. Numa regressão linear, quando $p = 1$, denominamos o estudo como regressão linear simples. Caso contrário, denominamos regressão linear múltipla.

Agora, vamos discutir sobre a *regressão linear simples*. Neste contexto, a ideia da palavra *linear* remete ao modelo candidato a explicar o relacionamento entre a variável resposta e as covariáveis, que é um modelo linear nos parâmetros, ou seja, os parâmetros são coeficientes multiplicativos de um modelo aditivo. Não há logaritmos ou parâmetros no expoente das covariáveis, por exemplo.

Por sua vez, a palavra *simples* significa que pretendemos, com uma reta, explicar a relação entre duas variáveis.

7.1.1 Modelo Estatístico

$$Y_i = \beta_0 + \beta_1 X_{1i} + e_i, \quad (7.1)$$

em que Y_i é o i -ésimo valor da variável Y ; β_0 é o intercepto; β_1 é o coeficiente angular; X_{1i} é o valor da (co)variável X para o indivíduo i ; e_i é o erro aleatório associado a Y_i .

7.1.2 Esperança Matemática

$$E[Y_i] = \beta_0 + \beta_1 X_{1i} \quad (7.2)$$

7.1.3 Método dos Quadrados Mínimos

Existem diversos métodos que permitem estimar os parâmetros de interesse no contexto de Regressão Linear Simples (β_0 e β_1), mas aqui trataremos apenas do Método dos Quadrados Mínimos.

O Método dos Quadrados Mínimos fornece aqueles valores de β_0 e β_1 que minimizam a soma de quadrados dos resíduos, ou seja, que minimizam a distância entre os valores observados e os estimados pelo modelo (reta), ao longo de todas as observações, ao mesmo tempo.

Por exemplo, considere a Figura 32. Ela mostra a representação de uma massa de dados fictícia. Duas variáveis quaisquer X e Y se relacionam de maneira diretamente proporcional, ou seja, são *positivamente correlacionadas*. Esse tipo de relacionamento sugere que uma reta pode ser o modelo ideal para descrever o comportamento, por exemplo, de Y , de acordo com o comportamento de X . A Figura 32 destaca que é possível haver infinitas retas passando por entre os pontos, porém, apenas uma delas possui a propriedade de estar à menor distância quadrática de todos os pontos, ao mesmo tempo.

A seguir, a rotina para você reproduzir a Figura 32.

```
x<-seq(0, 100)
y<-2*x + 35
y1<-y + rnorm(101, 0, 50)
reg<-lm(y1~x)
a<-reg$coefficients[1]
b<-reg$coefficients[2]
y2<-a + b*x
y3<-(y2[51] - 50*(b-1))+(b-1)*x
y4<-(y2[51] - 50*(b+1))+(b+1)*x
y5<-(y2[51] - 50*(b+2))+(b+2)*x
plot(x, y1, pch=19, xlab='X', ylab='Y')
lines(x, y2, lwd=2)
lines(x, y3, lty=3)
lines(x, y4, lty=3)
lines(x, y5, lty=3)
```

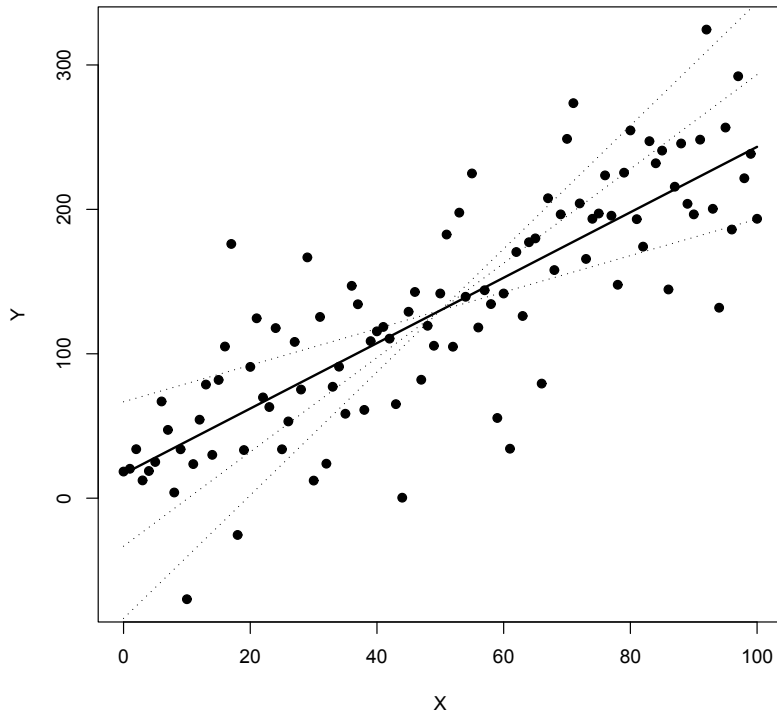


Figura 32 – Representação de possíveis retas (pontilhadas) e reta estimada por quadrados mínimos (linha cheia) em uma massa de dados fictícia.

Fonte: Do autor.

Os coeficientes da reta podem ser estimados pelo Método dos Quadrados Mínimos por meio dos estimadores 7.3 e 7.4.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n X_i\right)\left(\sum_{i=1}^n Y_i\right)}{n}}{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}} \quad (7.3)$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = \frac{\sum_{i=1}^n Y}{n} - \hat{\beta}_1 \left(\frac{\sum_{i=1}^n X}{n} \right) \quad (7.4)$$

Uma forma útil de inferir sobre a qualidade do modelo linear de regressão ajustado é um índice chamado *coeficiente de determinação*.

O coeficiente de determinação indica, percentualmente, quanto da variação da variável dependente (Y) pode ser explicada pelo modelo de regressão linear.

$$r^2 = \frac{\text{Variação explicada pelo modelo}}{\text{Variação total}} \quad (7.5)$$

Vamos nos ater ao caso em que nosso modelo é uma reta, ou seja, estamos lidando com uma regressão linear simples¹. Nesse caso particular o coeficiente de determinação se torna

$$r^2 = \frac{\text{Variação explicada pela reta}}{\text{Variação total}} = \frac{SQRL}{SQT}, \quad (7.6)$$

em que $SQRL$ significa Soma de Quadros de Regressão Linear, e SQT significa Soma de Quadrados Total, e essas quantidades são calculadas pelas expressões:

$$SQRL = \frac{\sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n X_i\right)\left(\sum_{i=1}^n Y_i\right)}{n}}{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}} \quad (7.7)$$

$$SQT = \frac{\left(\sum_{i=1}^n Y_i\right)^2}{n} \quad (7.8)$$

A diferença entre a variação total e a variação explicada pela reta de regressão é chamada de *desvio*, e pode ser calculada pela Soma de Quadrados de Desvios (SQD)

$$SQD = SQT - SQRL. \quad (7.9)$$

Além de estimar os coeficientes da reta de regressão, é possível testar se eles são significativos. Por exemplo, pode-se testar se o coeficiente β_1 é estatisticamente igual a zero, ou não. Se for considerado igual a zero, ou seja, se aceita-se H_0 nesse teste, isso significa que apenas a constante β_0 seria suficiente para explicar os dados, Y não varia com a variação de X . Por outro lado, se β_1 for considerado significativo, esse é um bom indicativo a favor do modelo estimado.

Pode ser feito o seguinte teste F para o ajuste de uma regressão linear.

7.1.3.1 Par de hipóteses

Vale ressaltar que este teste para β_1 foi construído para ser bilateral, mas, não precisa ser sempre assim. Podemos construir um teste unilateral para este parâmetro trocando o sinal de \neq de H_1 por $<$ ou $>$.

$$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$$

7.1.3.2 Estatística de teste

$$F_c = \frac{SQRL}{S^2},$$

¹ Mais detalhes e discussões sobre o coeficiente de determinação pode ser encontrados em (DRAPER; SMITH, 1998).

em que

$$S^2 = \frac{SQD}{n-2}.$$

7.1.3.3 Regra de decisão

$$\begin{cases} F_c < F(\alpha, \nu_1, \nu_2), & \text{Aceita-se } H_0 \\ F_c > F(\alpha, \nu_1, \nu_2), & \text{Rejeita-se } H_0 \end{cases},$$

em que $F(\alpha, \nu_1, \nu_2)$ é o valor tabelado para a distribuição F com ν_1 e ν_2 graus de liberdade, e 100 α % de probabilidade.

Agora, vamos falar brevemente sobre a *regressão linear múltipla*. Quando desejamos explicar a variável dependente (Y) por mais de uma covariável (X_1, X_2, \dots, X_k), mas ainda adotando um modelo linear, em tese deveríamos encontrar estimadores para $k + 1$ parâmetros! Em outras palavras, precisaríamos de um estimador para β_0 , outro para β_1 , outro para β_2 e assim por diante, até o estimador para β_k .

Em razão de o processo de obtenção de tão numerosos estimadores ser muito trabalhoso, para cada k possível utilizamos um artifício matemático fantástico para contornar esse problema: a notação matricial! A seguir vamos escrever um sistema de equações em notação matricial para ilustrar.

O sistema de equações normais é um sistema de equações que nos permite encontrar o estimador de um vetor Θ de parâmetros, de comprimento k qualquer. Com ele, podemos facilmente encontrar estimativas de quadrados mínimos com apenas uma conta matricial².

A ideia, aqui, é somente escrever o sistema de equações (uma para cada observação Y_i), reescrevê-lo em notação matricial, isolar o erro ε e encontrar o estimador de Θ para que a soma dos quadrados dos erros, ou seja $\varepsilon'\varepsilon$, seja mínima.

Sistema de Equações Normais (SEN):

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \dots + \beta_k x_{k1} + e_1 \\ y_2 &= \beta_0 + \beta_1 x_{12} + \beta_2 x_{22} + \dots + \beta_k x_{k2} + e_2 \\ &\vdots \qquad \qquad \qquad \vdots \\ y_n &= \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \dots + \beta_k x_{kn} + e_n \end{aligned}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

² Note que o estimador (7.10) depende da existência da inversa de $X'X$. Logo, uma série de aspectos devem ser observados. Para uma leitura mais aprofundada sobre o assunto, consulte (SEARLE, 1997).

Expressando esse sistema em notação matricial,

$$\begin{aligned} Y &= X\Theta + \varepsilon \\ \varepsilon &= Y - X\Theta \\ \varepsilon'\varepsilon &= (Y - X\Theta)'(Y - X\Theta) \end{aligned}$$

Diferenciando-se $\varepsilon'\varepsilon$ com respeito a Θ e igualando-se a zero, tem-se que o estimador de quadrados mínimos do vetor Θ é

$$\begin{aligned} X'X\Theta &= X'Y \\ \hat{\Theta} &= (X'X)^{-1}X'Y. \end{aligned} \tag{7.10}$$

7.2 Correlação

Uma vez que o relacionamento entre variáveis já foi modelado, é natural perguntar qual é a força ou intensidade desse relacionamento. A *correlação* representa essa associação. A correlação é uma medida padronizada da *covariância*, que é uma variância conjunta entre *duas* variáveis. É por isso que a correlação é definida entre *pares* de variáveis.

Uma das medidas mais usuais da correlação é o *coeficiente de correlação linear de Pearson*. Esse índice mede apenas a correlação *linear*. Neste caso, a palavra linear se refere à *reta*. Em outras palavras, o quanto o relacionamento entre duas variáveis pode ser descrito por uma reta, seja ela ascendente ou descendente. Outra forma de pensar na correlação linear é considerar se duas variáveis são diretamente ou inversamente proporcionais. Ou ainda, se elas não se relacionam.

A ideia é que quando duas variáveis forem *inversamente proporcionais*, seu r se aproxime de -1; quando forem *diretamente proporcionais*, seu r se aproxime de 1; e quando *não se relacionarem de forma linear*, seu r se aproxime de zero. Todos esses tipos de relacionamento podem ser vistos na Figura 33 de A a C e, além deles, na subfigura 33D, vemos que quando o relacionamento existe mas não é linear, o coeficiente de correlação linear não se aplica.

Veja aqui o script utilizado para gerar as subfiguras contidas na Figura 33. Note dois aspectos: (i) na primeira linha de código é carregado um pacote (`mvtnorm`), que antes precisa ser instalado; (ii) devido a serem sorteios, a sua reprodução do script vai gerar gráficos ligeiramente diferentes.

```
require(mvtnorm)

# r->0
sigma <- matrix(c(1,0,0,1), ncol=2)
x <- rmvnorm(n=50, mean=c(0,0), sigma=sigma)
r<-round(cor(x[,1],x[,2]),2)
plot(x,xlab='X',ylab='Y',main=paste('r_=',r),pch=20)

# r->1
sigma <- matrix(c(1,.9,.9,1), ncol=2)
x <- rmvnorm(n=50, mean=c(0,0), sigma=sigma)
```

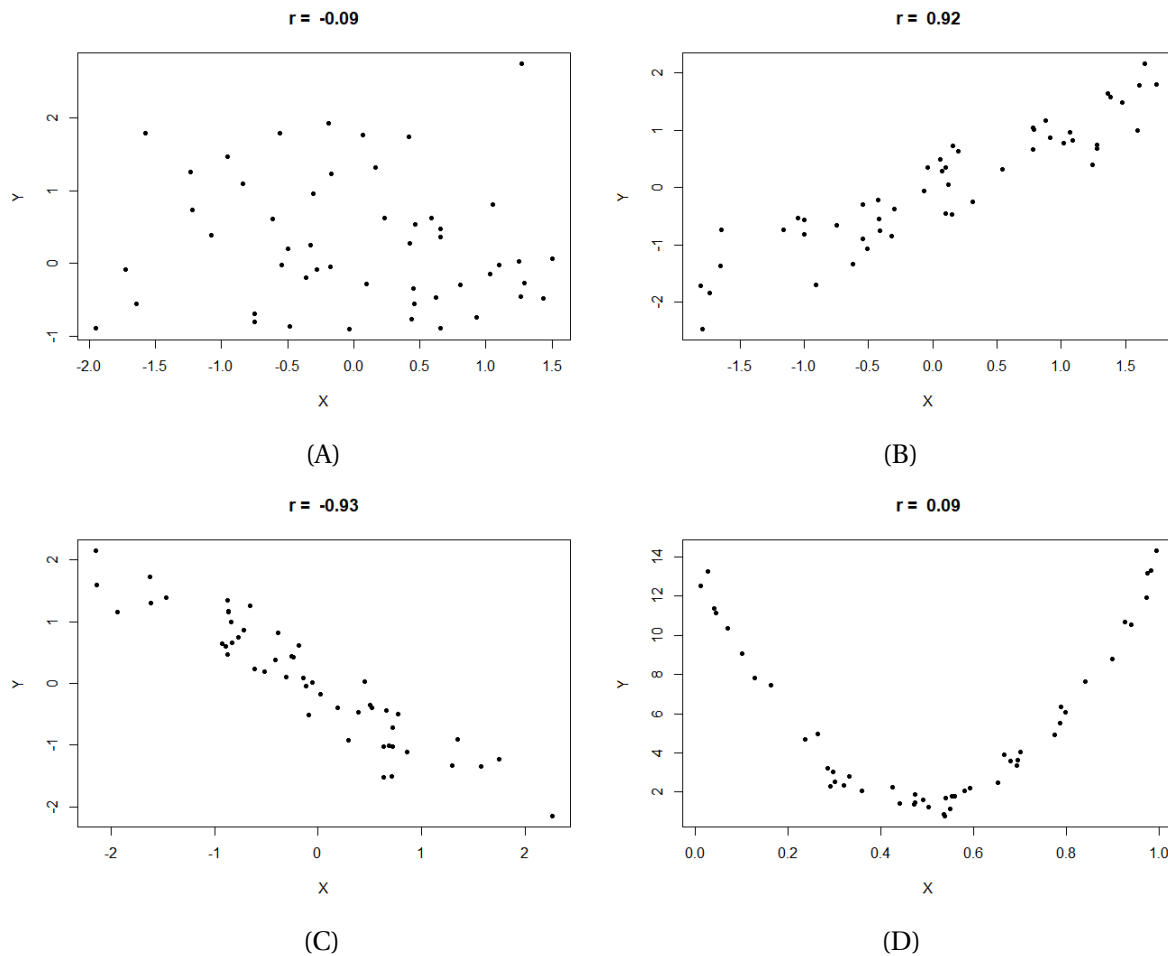


Figura 33 – Exemplos ilustrativos de quatro possíveis relacionamentos entre variáveis: não correlacionadas (A), positivamente correlacionadas (B), negativamente correlacionadas (C) e relacionamento quadrático (D).

Fonte: Do autor.

```
r<-round(cor(x[,1],x[,2]),2)
plot(x,xlab='X',ylab='Y',main=paste('r= $\square$ ',r),pch=20)

# r->-1
sigma <- matrix(c(1,-.9,-.9,1), ncol=2)
x <- rmvnorm(n=50, mean=c(0,0), sigma=sigma)
r<-round(cor(x[,1],x[,2]),2)
plot(x,xlab='X',ylab='Y',main=paste('r= $\square$ ',r),pch=20)

# relacionamento quadrático
x<-runif(50,0,1)
y<-1+x+50*(x-0.5)^2
y<-y+rnorm(50,0,.5)
r<-round(cor(x,y),2)
plot(x,y,xlab='X',ylab='Y',main=paste('r= $\square$ ',r),pch=20)
```

Para uma amostra de pares de variáveis aleatórias (X, Y) , medidas no mesmo indivíduo, o coeficiente de correlação linear de Pearson é dado por:

$$r = \frac{\sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n X_i\right)\left(\sum_{i=1}^n Y_i\right)}{n}}{\sqrt{\left(\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}\right)\left(\sum_{i=1}^n Y_i^2 - \frac{\left(\sum_{i=1}^n Y_i\right)^2}{n}\right)}} \quad (7.11)$$

e, por ser uma medida padronizada, varia de -1 a 1 .

Vale ainda destacar que, assim como uma média ou uma variância, a correção pode ser compreendida como *parâmetro* e como *estatística*. Ou seja, assumimos que exista uma *correlação populacional*, denotada por ρ , que deve ser estimada por uma *correlação amostral*, aqui denotada por r .

Exemplo: As famosas *cocadeiras* baianas costumam produzir suas próprias quitandas. Para isso, elas cumprem a difícil tarefa de quebrar dezenas de cocos por dia.



Fonte: Site de receitas Panelinha.³

Querendo evitar trabalho desnecessário, elas desejam quebrar apenas aqueles frutos que contêm uma grande quantidade de polpa. Portanto, procedem da seguinte maneira: furam o coco, medem sua quantidade de água e, com base em sua experiência, decidem se vale a pena quebrá-lo. Como esse procedimento é impreciso, as trabalhadoras desejam a nossa ajuda. Considerando os dados apresentados na Tabela 9, vamos ajustar um modelo adequado para prever o volume de polpa de frutos de coco (Y) a partir de sua quantidade de água (X).

O primeiro passo para a resolução desse problema é a *análise exploratória* dos dados, por exemplo, por meio de um *diagrama de dispersão* (Figura 34).

A seguir, a rotina para você reproduzir a Figura 34 em R.

```
polpa<-c(9.02,13.10,14.76,21.54,15.62,18.34,20.23,8.88,14.06,23.59,16.62,
21.93,10.56,12.28,20.68,9.53,13.73,5.73,15.08,21.57)
agua<-c(17.87,13.75,12.72,6.98,11.01,10.48,10.19,19.11,12.72,0.45,10.67,
1.59,14.91,14.14,9.40,16.23,12.74,20.64,12.34,6.44)
```

³ Disponível em: <<https://www.panelinha.com.br/receita/Cocada>>.

Tabela 9 – Volume de polpa (cm^3), volume de água (cm^3) e teor de cálcio ($mg/100ml$) em 20 cocos verdes.

Fruto	Polpa	Água	Cálcio
1	9,02	17,87	7,52
2	13,10	13,75	24,77
3	14,76	12,72	30,74
4	21,54	6,98	7,58
5	15,62	11,01	23,33
6	18,34	10,48	20,49
7	20,23	10,19	14,84
8	8,88	19,11	5,51
9	14,06	12,72	31,03
10	23,59	0,45	4,33
11	16,62	10,67	21,75
12	21,93	1,59	4,92
13	10,56	14,91	19,50
14	12,28	14,14	20,16
15	20,68	9,40	12,20
16	9,53	16,23	11,54
17	13,73	12,74	29,39
18	5,73	20,64	3,79
19	15,08	12,34	23,40
20	21,57	6,44	7,47

Fonte: Dados fictícios.

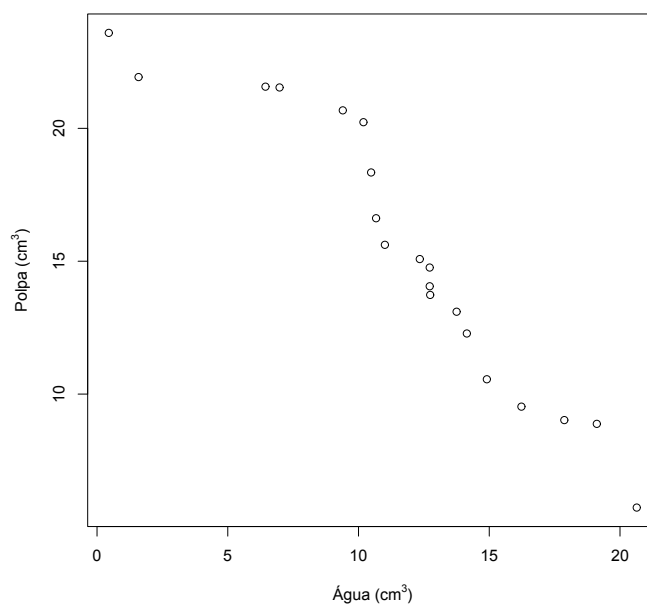


Figura 34 – Diagrama de dispersão entre a variável independente “volume de água de coco”, e a variável dependente “volume de polpa de coco”.

Fonte: Do autor.

```
plot(agua, polpa, xlab=expression(paste('Água_', ' ', cm^3, ' ')),
     ylab=expression(paste('Polpa_', ' ', cm^3, ' ')))
```

Os dados amostrais sugerem que, quanto mais água tem um coco, menos polpa ele possui. Além dessa informação, o diagrama de dispersão nos faz suspeitar que as duas variáveis possuem uma relação linear, ou seja, uma reta pode ser um bom modelo para este caso. Vamos estimar os coeficientes dessa reta de duas formas.

(a) Usando uma calculadora: uma calculadora das mais simples pode nos ajudar a estimar os coeficientes β_0 e β_1 , facilmente. Primeiramente, façamos uma tabela auxiliar com todas as quantidades de interesse (Tabela 10).

Tabela 10 – Tabela auxiliar para cálculo dos coeficientes do modelo linear.

X	Y	X ²	XY
17,87	9,02	319,3369	161,1874
13,75	13,10	189,0625	180,1250
12,72	14,76	161,7984	187,7472
6,98	21,54	48,7204	150,3492
11,01	15,62	121,2201	171,9762
10,48	18,34	109,8304	192,2032
10,19	20,23	103,8361	206,1437
19,11	8,88	365,1921	169,6968
12,72	14,06	161,7984	178,8432
0,45	23,59	0,2025	10,6155
10,67	16,62	113,8489	177,3354
1,59	21,93	2,5281	34,8687
14,91	10,56	222,3081	157,4496
14,14	12,28	199,9396	173,6392
9,40	20,68	88,3600	194,3920
16,23	9,53	263,4129	154,6719
12,74	13,73	162,3076	174,9202
20,64	5,73	426,0096	118,2672
12,34	15,08	152,2756	186,0872
6,44	21,57	41,4736	138,9108
234,38	306,85	3253,462	3119,430

De posse dessa tabela, só precisamos substituir as quantidades nas fórmulas dos coeficientes.

$$\hat{\beta}_1 = \frac{3119,430 - \frac{(234,38)(306,85)}{20}}{3253,462 - \frac{(234,38)^2}{20}} = \frac{3119,430 - 3595,975}{3253,462 - 2746,7} \approx -0,94$$

$$\hat{\beta}_0 = \frac{306,85}{20} + 0,94 \left(\frac{234,38}{20} \right) \approx 26,36$$

(b) Usando o R: é ainda mais fácil resolver tal problema usando o R. Inicialmente, devemos inserir a massa de dados, por exemplo, em vetores. Em seguida, podemos usar a função `lm()`, que ajusta modelos lineares. Para especificarmos que desejamos uma reta, basta dizer que os dados do vetor *polpa* são função dos dados do vetor *água*. Da seguinte maneira:

```
polpa<-c(9.02, 13.10, 14.76, 21.54, 15.62, 18.34, 20.23, 8.88,
14.06, 23.59, 16.62, 21.93, 10.56, 12.28, 20.68, 9.53, 13.73,
```

```
5.73, 15.08, 21.57)
agua<-c(17.87, 13.75, 12.72, 6.98, 11.01, 10.48, 10.19, 19.11,
12.72, 0.45, 10.67, 1.59, 14.91, 14.14, 9.40, 16.23, 12.74,
20.64, 12.34, 6.44)
lm(polpa~agua)
```

O resultado obtido por esse procedimento é exatamente o mesmo. As únicas diferenças se devem a arredondamentos feitos no item anterior. Isso mostra que ambos os procedimentos são equivalentes.

Finalmente, podemos informar para as cocadeiras que o modelo estimado é a reta $\text{Polpa} = 26,36 - 0,94 \times \text{Água}$, e ilustrar o resultado com a reta estimada plotada junto aos pontos amostrais (Figura 35).

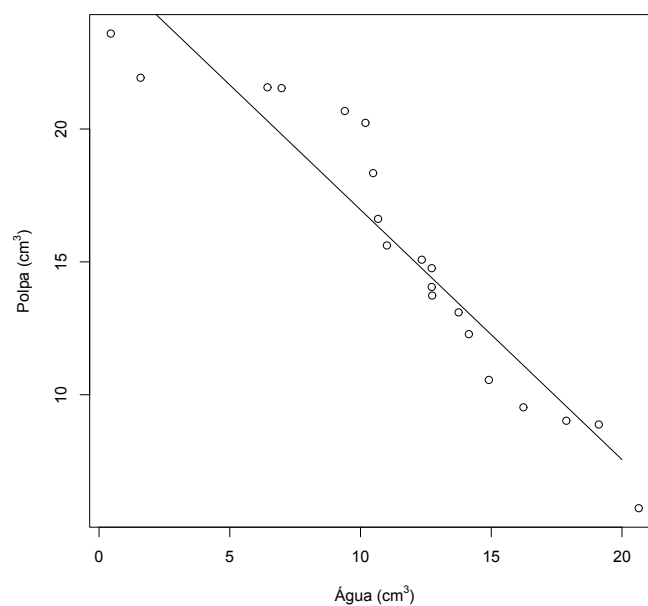


Figura 35 – Reta de regressão estimada e pontos amostrais, na relação entre a variável independente “volume de água de coco”, e a variável dependente “volume de polpa de coco”.

Fonte: Do autor.

Para se reproduzir a Figura 35 deve-se repetir o procedimento do diagrama de dispersão e acrescentar os seguintes comandos:

```
x<-seq(1:20)
reta<-26.36-0.94*x
lines(x,reta)
```




REFERÊNCIAS

BERNOULLI, J. *Usum & Applicationem Praecedentis Doctrinae in Civilibus, Moralibus & Oeconomicis*. Chapter 4 (translated into english by oscar sheynin). [S.l.: s.n.], 1713. Citado na página 61.

BLAND, M. *An Introduction to Medical statistics*. 4th. ed. Oxford: Oxford University Press, 2015. Citado na página 144.

BOLFARINE, H.; BUSSAB, W. O. *Elementos de Amostragem*. 1. ed. [S.l.]: Editora Blucher, 2005. Citado 2 vezes nas páginas 80 e 101.

DRAPER, N. R.; SMITH, H. *Applied Regression Analysis*. 3rd. ed. [S.l.]: Wiley-Interscience, 1998. Citado na página 130.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. *Introduction to the Theory of Statistics*. 3rd. ed. [S.l.]: McGraw-Hill, 1974. Citado na página 93.

R Core Team. *R: A language and environment for statistical computing*. 2020. Disponível em: <<https://www.R-project.org/>>. Acesso em: 22 jun. 2020. Citado na página 19.

RStudio Team. *RStudio: Integrated Development Environment for R*. Boston, MA, 2020. Disponível em: <<http://www.rstudio.com/>>. Acesso em: 22 jun. 2020. Citado na página 22.

SEARLE, S. R. *Linear Models*. 1st. ed. [S.l.]: John Wiley & Sons, 1997. Citado na página 131.



APÊNDICE A: EXERCÍCIOS PROPOSTOS

Neste apêndice, trazemos listas de exercícios para que você possa praticar os conceitos estudados ao longo do livro. Em cada exercício você vai encontrar uma sugestão de dificuldade, que pode ser (*), (**) ou (***). É claro que a dificuldade é subjetiva e varia de pessoa para pessoa em um mesmo exercício. O intuito é apenas dar uma sugestão. De posse dessa informação, você pode escolher quais exercícios resolver, caso não possa fazê-los todos.

Lista de exercícios: Técnicas de somatório

1. (*) Sejam os seguintes conjuntos de dados: $X = \{2, 4, 4, 3, 2\}$ e $Y = \{1, 2, 3, 6, 7\}$.

Obtenha:

$$\begin{array}{lll}
 1.1. \sum_{j=1}^4 X_j & 1.2. \sum_{j=1}^5 Y_j & 1.3. \sum_{j=1}^4 4X_j^2 \\
 1.4. \sum_{j=1}^5 X_j Y_j & 1.5. \sum_{j=1}^5 (3X_j + 2Y_j) & 1.6. \sum_{j=2}^4 X_j Y_j + \sum_{j=1}^5 Y_j^2.
 \end{array}$$

2. (**) Seja \bar{X} a média aritmética e S^2 , a variância:

$$\bar{X} = \frac{\sum_{j=1}^n X_j}{n} \qquad S^2 = \frac{1}{n-1} \left[\sum_{j=1}^n X_j^2 - \frac{\left(\sum_{j=1}^n X_j \right)^2}{n} \right]$$

Dado o conjunto de dados $X = \{2, 4, 5, 6, 1, 8\}$, calcule a sua média e variância.

3. (***) Demonstre numérica e algebricamente que $\sum_{j=1}^n (X_j - \bar{X}) = 0$. Use os dados do exercício anterior para demonstrar numericamente.

4. (***) Demonstre algebricamente as 4 propriedades básicas dos somatórios, sabendo que k e n são números inteiros (constantes) e x_i e y_i são variáveis.

a) $\sum_{i=1}^n kx_i = k \sum_{i=1}^n x_i$

b) $\sum_{i=1}^n k = nk$

c) $\sum_{i=1}^n (x_i + y_i) = \sum_{i=1}^n x_i + \sum_{i=1}^n y_i$

d) $\sum_{i=1}^n x_i y_i \neq \sum_{i=1}^n x_i \sum_{i=1}^n y_i$

Lista de exercícios:

Conceitos, Distribuição de Frequências e Dispositivos Gráficos

1. (***) Considere os seguintes cenários:

I. Uma cooperativa médica deseja realizar uma pesquisa com o objetivo de caracterizar os consultórios de seus cooperados. Como forma de obter estas informações, foram distribuídos questionários para todos os seus associados por meio dos quais procurou-se avaliar o nível tecnológico adotado (baixo, médio ou alto), a especialidade (pediatria, geriatria, etc.), e o número de funcionários de cada consultório.

II. Um pesquisador necessita obter informações a respeito da produção de café no sul de Minas Gerais. Para tanto, visita 50 propriedades e faz uma avaliação referente ao tamanho da área plantada (ha), a produção obtida (Kg), e as principais pragas e doenças.

Para cada um dos cenários, pergunta-se:

- a) Qual é a população em estudo?
- b) Classifique essa população.
- c) Utilizou-se uma amostra para realizar o estudo? Por quê?
- d) Quais foram as variáveis estudadas?
- e) Classifique as variáveis quanto à sua natureza.
- f) Diga qual gráfico você utilizaria para representar cada variável.

2. (**) Um levantamento investigou os pesos (Kg) de 49 crianças com idade entre 8 a 12 anos:

19,1	30,6	34,2	37,2	39,4	44,3	50,1
20,0	31,0	35,2	38,6	39,4	44,4	56,2
23,3	31,1	35,3	38,7	41,0	44,5	57,1
24,2	32,0	36,3	39,0	42,1	45,2	57,2
28,2	32,0	36,5	39,0	42,2	47,5	58,2
28,2	33,1	36,7	39,1	43,0	48,3	60,2
30,5	34,0	37,2	39,2	43,0	49,2	60,3

- a) Construa uma distribuição de frequências de ocorrência.
- b) Construa o dispositivo gráfico mais adequado para representar a distribuição de frequências.

3. (**) Nos cenários a seguir, identifique: (i) a população; (ii) a classificação da população; (iii) a variável de interesse; (iv) a classificação da variável.

a) Foi feito um estudo para compreender a distribuição da renda familiar entre os moradores da cidade de Alfenas (MG), utilizando os dados do Censo 2020 do IBGE.

- b) Uma pesquisa será realizada para saber qual é a escolaridade dos pais e mães de estudantes de uma faculdade privada.
- c) Engenheiros florestais farão um estudo da identificação das espécies vegetais que havia na região que foi alagada pela Hidrelétrica do Funil.
- d) Agrônomos da EMATER fizeram, em 2019, um levantamento da área (ha) das propriedades rurais do sul de Minas Gerais. Como há muitas propriedades, apenas algumas delas foram visitadas.

4. (***) Em um estudo, foram contadas as quantidades de irmãos que cada um dos 50 estudantes de Odontologia da Unifal possuem, encontrando-se o seguinte resultado:

5	1	5	3	1	2	2	1	1	0
4	2	0	4	4	4	4	3	3	2
1	1	3	2	3	1	2	3	4	3
4	0	2	0	5	2	3	4	3	4
0	0	4	3	2	2	4	3	4	3

- a) Estes dados são de uma população ou de uma amostra?
- b) Classifique a variável em questão.
- c) Faça uma distribuição de frequência dos dados e construa o gráfico mais adequado para representá-la.

5. (***) A tabela a seguir, adaptada de (BLAND, 2015), traz apenas as informações publicadas por esse autor. No entanto, deixamos células em branco para que você possa preencher e, com isso, responder as questões a seguir:

Tabela 11 – Distribuição de frequências das idades de pessoas que sofreram acidentes em casa na Inglaterra, no ano de 1977.

Idade (anos)	c	fr	fp (%)	Dfr
0 - 4			25,3	
5 - 14			18,9	
15 - 44			30,5	
45 - 64			13,6	
65+			11,7	
Total				

Fonte: (BLAND, 2015), p.46.

- a) Preencha a distribuição de frequências com especial atenção para qual deve ser a correta amplitude de classe. Note que as classes não tem tamanhos iguais.
- b) Qual limite superior foi provavelmente utilizado pelo autor da tabela para calcular as informações da última classe?
- c) Construa um histograma com a frequência relativa de ocorrência e outro com a densidade de frequência relativa. Compare-os. Em que eles se diferem? Qual dos dois é correto? Por quê?

Lista de exercícios: Medidas de posição e dispersão

1. (*) Um estudo foi conduzido para verificar o consumo energético de mulheres adolescentes que sofriam de bulimia. A seguir, são listados os valores de ingestão calórica diária, em kcal/kg.

15,9 18,9 25,1 16,0 19,6 25,2 16,5 21,5 25,6
21,6 28,0 17,6 22,9 28,7 18,1 23,6 29,2 18,4

a) Calcule as 3 medidas de posição estudadas e escolha uma para melhor representar essa amostra. Justifique sua escolha.

b) Calcule as 4 medidas de dispersão estudadas e escolha uma para melhor representar essa amostra. Justifique sua escolha.

2. (**) O departamento de marketing de uma grande indústria de alimentos está fazendo um estudo para mudar a embalagem de seu macarrão instantâneo. Duas embalagens (A e B) foram avaliadas quanto à aceitação, por 10 consumidores desse produto. Solicitou-se que fossem dadas notas de 1 a 5 para mediar a intensidade da aceitação de cada tipo de embalagem.

	Indivíduo									
	1	2	3	4	5	6	7	8	9	10
Embalagem A	3	3	4	4	4	4	4	5	5	5
Embalagem B	3	3	3	3	2	2	2	1	1	1

a) Para cada embalagem, calcule as medidas de posição estudadas e escolha uma para melhor representar cada amostra. Justifique suas escolhas.

b) Para cada embalagem, calcule as medidas de dispersão estudadas e escolha uma para melhor representar cada amostra. Justifique suas escolhas.

c) Escolha uma medida de posição para comparar a aceitação das embalagens. Conclua qual foi preferida.

d) Escolha uma medida de dispersão para comparar as variabilidades das aceitações. Qual conjunto de notas é mais variável? Justifique.

Lista de exercícios: Probabilidade simples

1. (*) A probabilidade que um homem esteja vivo daqui a 30 anos é $2/5$, e a da sua mulher é $2/3$. Determinar a probabilidade que daqui a 30 anos:

- a) Ambos estejam vivos;
- b) Somente o homem esteja vivo;
- c) Somente a mulher esteja viva;
- d) Nenhum esteja vivo;
- e) Pelo menos um esteja vivo.

2. (**) Uma companhia que fura poços artesianos trabalha numa região escolhendo aleatoriamente o ponto de furo. Não encontrando água nessa tentativa, sorteia outro local e, caso também não obtenha sucesso, faz uma terceira e última tentativa. Admita a probabilidade 0,7 de encontrar água em qualquer ponto dessa região. Determine o espaço amostral e calcule a probabilidade de:

- a) Encontrar água na segunda tentativa.
- b) Encontrar água em até duas tentativas.
- c) Encontrar água.

3. (**) A tabela a seguir apresenta dados dos 1000 ingressantes de uma universidade, com informações sobre área de estudo e classe sócio-econômica de cada um deles.

Área	Classe		
	Alta	Média	Baixa
Exatas	120	156	68
Humanas	72	85	112
Biológicas	169	145	73

Se um aluno ingressante é escolhido ao acaso, determine a probabilidade de:

- a) Ser da classe econômica mais alta.
- b) Estudar na área de Exatas.
- c) Estudar na área de Humanas, sendo de classe média.
- d) Ser da classe baixa, dado que estuda na área de Biológicas.

4. (**) Estatísticas dos últimos anos do departamento estadual de estradas são apresentadas na tabela a seguir contendo o número de acidentes com vítimas (fatais ou não), e as condições do principal motorista envolvido (sóbrio ou alcoolizado).

Motorista	Vítimas	
	Não fatais	Fatais
Sóbrio	1228	275
Alcoolizado	2393	762

Você diria que o fato do motorista estar ou não alcoolizado interfere na ocorrência de vítimas fatais?

Lista de exercícios: Distribuição Binomial

1. (**) Sabe-se que 5% de uma população está, hoje, com febre. Qual a probabilidade de que num grupo de 10 pessoas retirado dessa população tenha-se:
 - a) nenhuma pessoa com febre?
 - b) duas pessoas com febre?
 - c) mais de uma pessoa com febre?

2. (**) Num grupo de cinco bebês, qual a probabilidade de:
 - a) não haver meninas?
 - b) haver duas meninas?
 - c) haver pelo menos duas meninas?

3. (*) Numa criação de coelhos, 40% são machos. Num dia em que nasçam vinte coelhos, qual a probabilidade de nascerem:
 - a) cinco coelhos machos?
 - b) pelo menos dois coelhos machos?

4. (***) Suponha que a percentagem de germinação de uma semente de feijoeiro seja de 60%. Serão semeadas três sementes por cova em um canteiro com vinte e quatro covas.
 - a) Qual a probabilidade de obter-se pelo menos uma cova falhada no canteiro?
 - b) Qual será o número esperado de covas falhadas no canteiro?

5. (**) Um carrinho de picolés contém somente dois sabores: coco e morango. O vendedor abastece, pela manhã, com 40 picolés de coco e 60 de morango. O primeiro cliente pede 10 picolés, e diz que podem ser pegos ao acaso. Sabendo disso, responda:
 - a) Quantos você espera que sejam de coco e quantos de morango?
 - b) Qual é a probabilidade de todos serem de morango?
 - c) Qual é a probabilidade de 5 serem de coco?

6. (**) No lançamento de 5 moedas honestas, responda corretamente:
 - a) Qual o número esperado de caras?
 - b) Qual a probabilidade de sair exatamente 1 cara?
 - c) Qual a probabilidade de sair pelo menos 4 caras?
 - d) Qual a probabilidade de sair no máximo 2 caras?
 - e) Qual é a variância do número de caras?
 - f) Esboce o gráfico da distribuição de probabilidades desse experimento.

7. (***) Em um bairro de Alfenas, o IBGE observou famílias e seus respectivos número de filhos.

Filhos	0	1	2	3	4	>4
Nº de famílias	45	120	44	13	5	2

- a) Calcule a probabilidade de ocorrência de cada número de filhos.
- b) Modele a variável X : *número de filhos* com apenas dois resultados possíveis: *no máximo 1 filho e mais de 1 filho*. Dessa forma, cada família representa um ensaio de Bernoulli. Isso posto, escreva o espaço amostral do experimento.
- c) Encontre as probabilidades de sucesso e fracasso, considerando que você deseja contar o número de famílias com mais de 1 filho.
- d) Qual é o número esperado de famílias com mais de 1 filho?
- e) Qual é a variância do número de famílias com mais de 1 filho?
- f) Qual é a probabilidade de 64 famílias terem mais de 1 filho?
8. (*) Um inspetor de qualidade extrai uma amostra de 10 tubos aleatoriamente de uma carga muito grande de tubos que, se sabe, que contém 20% de tubos defeituosos. Qual é a probabilidade de que não mais que 2 dos tubos extraídos sejam defeituosos?
9. (*) Um engenheiro de inspeção extrai uma amostra de 15 itens aleatoriamente de um processo de fabricação sabido produzir 85% de itens aceitáveis. Qual a probabilidade de que 10 dos itens extraídos sejam aceitáveis?
10. (**) Uma prova de múltipla escolha tem 10 questões. Cada questão tem 5 alternativas, mas apenas 1 é correta. Um aluno que responde todas as questões a esmo (“chutando”) tem qual probabilidade de:
- a) Fechar a prova (tirar 10)?
- b) Zerar a prova (tirar 0)?
- c) Qual é a nota esperada para esse aluno?
- d) Qual é a variância da nota desse aluno?

Lista de exercícios: Distribuição Poisson

1. (***) Numa lâmina verificou-se que existem, em média, 5 bactérias/ cm^2 . A lâmina foi subdividida em 300 quadros de $1cm^2$.

- a) Qual é a probabilidade de encontrar, no máximo, 6 bactérias em um quadrinho?
- b) Em quantos destes quadros em média você espera encontrar, no máximo, 6 bactérias?
- c) Qual é a probabilidade de se encontrar mais de 4 bactérias por centímetro quadrado?

2. (***) Um pesquisador da área de zootecnia conseguiu uma série de dados dos últimos 120 anos com o registro do número de ocorrências de uma doença rara em equinos da localidade em que trabalhava. Os dados obtidos foram:

Ocorrências	0	1	2	3	4	5
Anos	55	40	17	5	2	1

- a) Calcule a frequência relativa (probabilidade) observada para a variável X .
- b) Estime o número médio de doenças/ano com as frequências calculadas em (a).
- c) Calcule a frequência esperada (em anos) para cada valor de X , segundo a Poisson.
- d) Compare os resultados esperados com os observados. Com base nesta comparação, você pode afirmar que a distribuição de Poisson é adequada para explicar a ocorrência desta doença na região de estudo? Justifique.

3. (**) Esse mesmo zootecnista havia comprado um lote de vacinas com deficiência nominal de imunização rotulada como “1 animal vacinado não imunizado em cada 2500 animais vacinados, em média”. Para um teste, a vacina foi aplicada em um lote de 5000 animais. Depois de decorrido algum tempo, constatou-se que 4 animais manifestaram a doença.

- a) Qual era o número de animais que você esperava que não fossem imunizados?
- b) Calcule a probabilidade de ocorrência, segundo a Poisson, de 2 e 4 animais não serem imunizados.
- c) Compare as duas probabilidades da letra (b) e discuta se o rótulo está correto ou errado.

4. (*) Um nutricionista experiente sabe que a probabilidade de atender uma criança diabética é de 1%. Um mutirão é feito no dia do nutricionista, com atendimentos na praça principal da cidade. Esse nutricionista vai atender, gratuitamente, 150 crianças ao longo do dia. Sabendo disso, responda corretamente:

- a) Qual é o número esperado de crianças diabéticas que ele vai atender?
- b) Qual é a probabilidade de ele atender 3 crianças diabéticas?
- c) Qual é a probabilidade de ele não atender nenhuma criança diabética?
- d) Qual é a variância de crianças diabéticas atendidas?

5. (*) O Censo do IBGE de 2010 revelou que, no Brasil, 5 milhões de homens se chamam *José*. Sabemos que o Brasil possui aproximadamente 200 milhões de habitantes, dos quais aproximadamente 100 milhões são homens. Então, se você tomar aleatoriamente 100 homens, responda:
- Qual é a probabilidade de haver 1 José?
 - Qual é a probabilidade de não haver nenhum José?
 - Qual é a probabilidade de todos se chamarem José?
 - Qual é o número esperado de Josés nesse grupo?
6. (**) Quais são as diferenças entre as distribuições de Poisson e Binomial?
7. (*) Um departamento de polícia recebe em média 5 solicitações por hora. Qual a probabilidade de receber 2 solicitações numa hora selecionada aleatoriamente?
8. (*) Em um posto de gasolina, um funcionário experiente sabe que um número médio de 6 clientes por hora param para colocar gasolina.
- Qual é a probabilidade de 3 clientes abastecerem?
 - Qual é a probabilidade de 3 clientes ou menos abastecerem?
 - Qual é o valor esperado e a variância para esta distribuição?
9. (**) Um por cento das lâmpadas incandescentes produzidas numa fábrica são defeituosas. Encontre a probabilidade de haver mais de uma defeituosa numa amostra aleatória de 30 lâmpadas, usando:
- A distribuição Binomial.
 - A distribuição de Poisson.
 - Compare os resultados.
10. (**) O Ministério da Educação afirma que o Brasil possui 7,3 milhões de estudantes universitários. Desses, 73.353 participaram do Programa Ciências Sem Fronteiras (PCSF), de graduação sanduíche em outros países. Se uma amostra de 100 estudantes foram tomados ao acaso, responda:
- Qual é o número esperado de participantes do PCSF nessa amostra?
 - Qual é a variância dessa distribuição?
 - Qual é a probabilidade de haver, nesse grupo, mais de 1 participante do PCSF?
 - E de haver menos de 2 participantes?
 - Esboce o gráfico dessa distribuição de probabilidades.

Lista de exercícios: Distribuição Normal

1. (*) Considerando que os pesos dos coelhos Norfolk ao abate aos 90 dias obedeça uma distribuição Normal, com média de $2,70\text{kg}$ e variância de $0,04\text{kg}^2$. Responda:

- a) Qual é o peso esperado para um coelho tomado ao acaso?
- b) Qual a frequência de coelhos com peso acima de $2,90\text{kg}$?
- b) E entre $2,90$ e $3,00\text{kg}$?
- c) Qual o peso que é superado por apenas 1% dos coelhos?

2. (***) O diâmetro X das esferas de rolamentos fabricadas por certa indústria tem distribuição Normal, com média $0,6140\text{cm}$ e variância $0,00252\text{cm}^2$. A esfera é classificada como *conforme* se seu diâmetro X está no intervalo $[0,5000; 0,7000]$, e *defeituosa* se estiver fora. Sabendo disso, calcule:

- a) a probabilidade de fabricação de esferas boas.
- b) a probabilidade de fabricação de esferas defeituosas.

O estatístico dessa fábrica quer estudar a variável Y : *número de esferas defeituosas* em um lote de 10.000 unidades produzidas. Sabendo disso,

- c) você modelaria Y como Binomial, ou Poisson? Por quê?
- d) quantas esferas defeituosas você espera nesse lote?
- e) qual é a probabilidade de haver mais de 2 defeituosas?

O setor de vendas lhe informou que o lucro L proporcionado por esfera depende de sua classificação: $L = R\$10,00$ se a esfera for vendida como *conforme*, ou $L = R\$1,10$ se for vendida defeituosa. Sabe-se que a esperança da variável L é dada por

$$E[X] = \sum_{i=1}^2 L_i P[L = L_i]$$

- f) Encontre o lucro médio por esfera.
- g) E o lucro total do lote.

3. (***) Uma máquina de empacotar café produz pacotes de café com pesos segundo uma distribuição Normal de frequências, com média de 500g e desvio padrão de 3g .

- a) Num lote de 10.000 pacotes, em quantos você espera encontrar menos de 490g de café?

Considere que é possível ajustar a média com que os pacotes são cheios. Qual deve ser o ajuste da média para que 99% dos pacotes não tenham peso inferior a 500g ?

4. (*) Em uma granja de frangos, um estudo estatístico cuidadoso determinou que a lei de frequências de ocorrência de peso final (em kg) desses frangos, dentro de um período pré-especificado de tempo, é Normal, com 1881g de média e desvio-padrão igual a 210g .

Se um frango é escolhido aleatoriamente, responda:

- a) Qual é o peso esperado para ele?
- b) Qual é a probabilidade de ele ter peso superior a 2000g?
- c) Qual é a probabilidade de ele ter menos de 1800g?
- d) E de estar entre 1700g e 2010g?
- e) Qual é o intervalo de peso, centrado na média, em que se encontram 95% dos frangos?

5. (*) O peso declarado na embalagem de uma marca de iogurte grego é de 100g. Contudo, seu peso real varia, minimamente, de um potinho para outro. Considere que o peso do potinho siga uma Normal com média 103g e desvio padrão 3g. Você compra um desses potinhos, ao acaso, e se pergunta:

- a) Quanto espero que seja o peso real desta unidade?
- b) Qual a probabilidade de, esta unidade, ter menos de 100g?
- c) Qual a probabilidade de, esta unidade, ter mais de 105g?
- d) Qual a probabilidade de, esta unidade, ter entre 102g e 105g?
- e) Qual é o intervalo de peso que contém 85% dos potinhos de iogurte produzidos?

6. (*) O Quociente de inteligência (QI) é uma medida padronizada obtida por meio de testes desenvolvidos para avaliar as capacidades cognitivas (inteligência) de um sujeito. Suponha que o QI dos alunos admitidos em uma certa universidade tenha Distribuição Normal com média 114, e desvio padrão 10. Pergunta-se:

- a) Qual a probabilidade de um indivíduo ter QI menor que 105?
- b) Qual a probabilidade de um indivíduo ter QI maior que 124?
- c) Qual a probabilidade de um indivíduo, tomado ao acaso, tenha QI entre 104 e 114?
- d) Qual é o intervalo de QI, centrado na média, que contém 90% da população?
- e) Qual é o QI esperado para um aluno tomado ao acaso?

7. (*) O tempo gasto para dar uma volta numa pista de Fórmula 1 segue distribuição Normal com média $1,32min$, e desvio padrão $0,42min$. Sabendo disso, responda corretamente:

- a) Qual é a probabilidade de uma volta ser dada em mais de $1,5min$?
- b) Qual é a probabilidade de uma volta ser dada em menos de $1,0min$?
- c) Qual é a probabilidade de uma volta ser dada entre $1,0min$ e $1,5min$?
- d) Qual é o intervalo de tempo, centrado na média, em que 90% das voltas são dadas?
- e) Qual é o tempo esperado para ser gasto em uma volta?

Lista de exercícios: Técnicas de amostragem

1. (*) Uma pessoa retirou três maçãs da superfície de cada uma das caixas de maçãs que estavam em uma quitanda para verificar a sua qualidade. Isto é uma amostra aleatória? Existe algum problema com este método de amostragem? Comente como você faria.

2. (**) Deseja-se testar, durante um mês, um novo tipo de biscoito integral. O objetivo é conhecer o incremento médio do trânsito intestinal, quando utilizado o novo biscoito. Para isso, planejou-se determinar a diferença entre o trânsito intestinal do mês em que foi fornecido o novo biscoito, e o trânsito intestinal do mês anterior. Sabe-se que, em qualquer caso, antes e depois dos biscoitos, o trânsito intestinal de pessoas jovens é superior (ou pelo menos diferente) do trânsito intestinal de adultos, sendo esta diferença significativa. O grupo de possíveis consumidores desses biscoitos é de 653 indivíduos, sendo que, após uma análise estatística e de custos, determinou-se oferecer os biscoitos à 36 pessoas.

a) Qual é a população em estudo?

b) Qual é a amostra?

c) Qual é o tamanho da população, e qual é o tamanho da amostra?

d) A população é finita ou infinita? Porque?

e) Para esse tipo de estudo, qual tipo de amostragem você recomendaria utilizar? Por quê?

3. (**) Planeje uma amostragem aleatória sistemática para amostrar 20 hastes de amortecedores da linha de produção da Magnetti Marelli Cofap, durante um turno de produção de 6300 hastes, aproximadamente. Sorteie as hastes que deverão ser selecionadas.

4. (**) Uma empresa cafeeira do Sul de Minas dispõe de 3.200 funcionários distribuídos nas diversas atividades, conforme o quadro abaixo:

Atividade	Empregados
Campo	1.600
Armazém	720
Indústria	480
Administração	240
Gerência	160

a) Na sua opinião, seria razoável levantar as informações desejadas através de uma amostragem aleatória simples de $n = 160$ funcionários? Justifique.

b) Planeje uma amostragem estratificada de $n = 160$, determinando o tamanho da amostra para cada atividade.

c) Usando sua calculadora científica, sorteie os componentes da amostra para os empregados que trabalham na gerência.

Lista de exercícios:
Distribuição da média (\bar{x})

1. (*) Frangos de corte geralmente são abatidos com $1,8kg$. Um lote de frangos foi vendido com 25 aves. Sabendo que o peso ao abate segue uma distribuição Normal com média $1,8kg$, e variância $0,04kg^2$, encontre a probabilidade de o peso médio do lote vendido:

- a) ser superior a $2,5kg$.
- b) ser inferior a $1,7kg$.
- c) estar entre $1,7kg$ e $2,0kg$.

2. (*) O tempo que os alunos de uma turma levam para fazer uma prova de Estatística segue uma Normal com média $1h$ e variância $15min^2$. Qual é a probabilidade do tempo médio de prova de um grupo de 4 alunos ser:

- a) maior que $1h30min$.
- b) menor que $1h$.
- c) entre $45min$ e $1h15min$.

3. (**) Um elevador possui o seguinte aviso: “O peso médio dos ocupantes não deve ultrapassar $90kg$.” Além disso, você sabe que o peso de adultos segue uma Normal com média $70kg$ e variância desconhecida.

Em uma amostra prévia, você observou que o desvio-padrão amostral era $S = 20kg$.

Qual é a probabilidade de um grupo de 6 pessoas violar o aviso?

4. (**) Em uma comunidade, a “idade da 1ª cárie” segue uma Normal com média 5 anos e variância desconhecida. Porém, você tem uma estimativa dessa variância, $S^2 = 2 anos^2$.

Você atende um grupo de 10 crianças dessa comunidade. Qual é a probabilidade de a idade média da 1ª cárie:

- a) estar entre 4 e 6 anos?
- b) ser maior que 6,5 anos?
- c) ser menor que 10 anos?

Lista de exercícios:

Estimação da média (μ), do total (T) a partir da média e tamanho de amostra (n)

1. (**) Numa Universidade, foi tomada uma amostra de 40 estudantes anotando-se as suas alturas (cm). A soma das alturas e dos quadrados das alturas foram:

$$\sum_{i=1}^{40} x_i = 6.950 \quad \sum_{i=1}^{40} x_i^2 = 1.213.463$$

- Estime a altura média por ponto e por intervalo com 95% de confiança. Interprete-o.
- Qual foi o erro cometido nesse estudo?
- Qual deveria ser o tamanho da amostra para cometer metade desse erro?

2. (*) Em um encontro de 200 jovens estudantes, uma gincana foi feita. O evento foi um sucesso e os organizadores gostariam de divulgar a idade média dos jovens participantes. Como não conseguiriam entrevistar todos eles, entrevistaram 16, tomando-os de forma aleatória. Sabendo que a idade amostral média foi de 15 anos e o desvio padrão amostral foi de 2 anos, responda:

- Estime a idade média populacional dos 200 jovens, por ponto e por intervalo com 90% de confiança. Interprete-o.
- Qual foi o erro cometido?
- Qual deveria ser o tamanho da amostra para cometer 90% desse erro?

3. (*) Um pecuarista se entusiasmou por uma nova ração amplamente divulgada pelos meios de comunicação. Para verificar a eficiência da ração, ele selecionou uma amostra de 49 bois de seu rebanho e os alimentou por 30 dias, obtendo um ganho de peso médio de 31,7kg, com um desvio padrão de 2,6kg.

- Estime o ganho de peso médio no rebanho inteiro, por ponto e por intervalo de confiança de 95%. Interprete-o.
- Qual deveria ser o tamanho da amostra para que o erro não fosse superior a 0,7kg, com probabilidade de 95%.

4. (**) Foi coletada uma amostra aleatória simples de tamanho $n = 30$ do rebanho $N = 201$ do Núcleo dos Criadores de Gado Holandês do Sul de Minas (NCGH) com o objetivo de descrever a produção de leite. Os dados obtidos na amostra foram:

17,7	20,7	19,3	19,3	18,0	16,9	19,7	20,1	21,0	21,2
23,3	15,3	23,7	18,8	25,2	18,0	22,8	21,1	18,8	25,9
19,3	19,6	26,6	14,3	19,7	32,7	14,1	16,8	19,7	19,3

- Estime a produção média de leite do NCGH, por ponto e por intervalo.
- Qual foi o erro cometido?

- c) Calcule o tamanho da amostra para cometer 80% desse erro.
- d) Estime a produção total do NCGH, por ponto e por intervalo.

Lista de exercícios:

Estimação de uma proporção (p), do total (T) a partir de uma proporção e do tamanho da amostra (n)

1. (***) Um levantamento amostral sobre aspectos de higiene e saúde envolvendo bairros periféricos de Alfenas mostrou, entre outros fatos, a seguinte resposta à pergunta: “Com qual frequência você lava sua caixa d’água?”

Resposta	Número de residências
Nunca	13
3 em 3 meses	11
6 em 6 meses	4
Anualmente	22
Poucas vezes	18
Total	68

Obs: Considere 5.000 residências na periferia de Alfenas, quando esta pesquisa foi feita.

a) Estime, por ponto e por intervalo, a proporção de residências da periferia que lavam a caixa d’água *nunca* ou *poucas vezes*. (Se você souber como, incorpore o fator de correção para população finita.)

b) Qual foi o erro cometido na estimação da proporção?

c) Qual deveria ser o número de casas visitadas para que o erro cometido fosse, no máximo, 60% desse?

d) Estime, por ponto e por intervalo, o número total de residências da periferia que lavam a caixa d’água *nunca* ou *poucas vezes*.

e) Qual foi o erro cometido na estimação desse total?

2. (***) Com o objetivo de estudar a criminalidade durante um ano, uma amostra do número de ocorrências policiais em um certo bairro de São Paulo, durante 28 dias, apresentou os seguintes resultados:

7, 11, 8, 9, 10, 14, 6, 8, 8, 7, 8, 10, 10, 14, 12, 9, 11, 13, 13, 8, 6, 8, 13, 10, 14, 5, 14, 10.

a) Estime, por ponto e por intervalo, a proporção de dias violentos (com, pelo menos, 12 ocorrências). Use a confiança de 90%.

b) Qual foi o erro cometido ao estimar essa proporção?

c) Qual deveria ter sido o número de dias observados, caso o pesquisador desejasse cometer, no máximo, 40% desse erro?

d) Em um ano (365 dias) e com a mesma confiança de 90%, quais seriam as estimativas (pontual e intervalar) do número total de dias violentos nesse bairro?

e) Qual foi o erro cometido na estimação desse total?

3. (**) Suponha que estejamos interessados em estudar os consumidores de macarrão instantâneo, dentre os alunos da Nutrição. Uma amostra de tamanho 300 forneceu que 100 indivíduos consomem esse produto. Suponha que, atualmente, existam 600 alunos no curso de Nutrição e responda corretamente:

- a) Estime a proporção de interesse por ponto e por intervalo com 92% de confiança.
- b) Qual foi o erro cometido?
- c) Calcule o tamanho da amostra para cometer 90% desse erro.
- d) Estime, por ponto e por intervalo com a mesma confiança, o número total de alunos que consomem macarrão instantâneo, no curso de Nutrição.
- e) Qual foi o erro cometido na estimação desse total?

4. (*) Em uma linha de produção de um alimento, deve-se medir o diâmetro final da embalagem para que não cause problemas no envazamento. Para verificar a adequação do processo, são selecionadas, aleatoriamente, 12 embalagens, que revelam os diâmetros (cm):

3,01 3,05 2,99 2,99 3,00 3,02 2,98 2,99 2,97 2,97 3,02 3,01

- a) Estime, por ponto e por intervalo, a proporção de diâmetros maiores que 3cm.
- b) Qual foi o erro cometido na estimação dessa proporção?
- c) Qual deve ser o tamanho da amostra para que o erro seja, no máximo, 10% do erro cometido?

Lista de exercícios:

Estimação da variância (σ^2) e do desvio padrão (σ)

1. (*) Um biólogo, especialista em jacarés, quer conhecer quanto varia o comprimento de adultos do jacaré-de-papo-amarelo (*Caiman latirostris*). Uma amostra de 10 animais foi sorteada e forneceu comprimento médio de $1,69m$ e variância de $0,01m^2$.

a) Estime, por ponto e por intervalo de 95% de confiança, a variância e o desvio padrão.

2. (*) Com o objetivo de estudar a criminalidade durante um ano, uma amostra do número de ocorrências policiais em um certo bairro de São Paulo, durante 28 dias, apresentou os seguintes resultados:

7, 11, 8, 9, 10, 14, 6, 8, 8, 7, 8, 10, 10, 14, 12, 9, 11, 13, 13, 8, 6, 8, 13, 10, 14, 5, 14, 10.

a) Para ter uma ideia da variabilidade, estime (por ponto e por intervalo) o desvio padrão do número de ocorrências.

3. (*) Um provedor de acesso à internet monitora a duração das conexões de seus clientes com o objetivo de dimensionar seus equipamentos. Uma amostra de 500 conexões resultou num valor médio observado de $25min$ e variância de $50min^2$.

a) Estime a variância de todas as conexões desse provedor por ponto e por intervalo de 95% de confiança.

b) Faça o mesmo para o desvio padrão, porém, a estimativa intervalar precisa ter coeficiente de confiança (γ) igual a 0,9.

4. (*) Em uma linha de produção de um alimento deve-se medir o diâmetro final da embalagem para que não cause problemas no envazamento. Para verificar a adequação do processo, são selecionadas, aleatoriamente, 12 embalagens, que revelam os diâmetros (cm):

3,01 3,05 2,99 2,99 3,00 3,02 2,98 2,99 2,97 2,97 3,02 3,01

a) Estime, por ponto e por intervalo, o desvio padrão dos diâmetros.

b) Escolha uma confiança diferente e estime, por ponto e por intervalo, a variância dos diâmetros.

Lista de exercícios:

Testes para a média (μ) e para uma proporção (p)

1. (**) Indique claramente com palavras ou, preferencialmente, com parâmetros, as hipóteses H_0 , H_1 , os dois tipos de acertos possíveis e os dois tipos de erros, indicando qual é o erro tipo I, tipo II e qual é o erro mais grave na sua opinião.

a) Um laticínio afirma que o teor médio de gordura do seu creme de leite é de 25%, mas alguns consumidores desconfiam que é menos.

b) Um empresário calcula que a viabilidade econômica de sua infra-estrutura é atingida com uma produtividade média de $8,2t/ha$, e deseja saber se corresponde à produtividade atual.

c) Um ortodontista utiliza medicamentos diferentes em dois grupos de pacientes (A e B) e deseja saber se eles determinam condições sanitárias médias diferentes.

d) O limite de tolerância de uma certa virose é de 10%. Deseja-se saber se determinada cidade deve ser considerada infectada.

e) Laura vai hoje ao clube. Ela gostaria de nadar na piscina, mas o céu está totalmente nublado. Ela precisa decidir se leva ou não seu biquíni.

f) Um juiz da Vara Criminal de Alfenas está diante de um réu e precisa decidir, diante do conjunto de provas, se ele é culpado ou inocente.

g) A diretora de uma escola infantil suspeita que um fornecedor tenha fraudado o leite das crianças.

2. (**) Um biólogo acredita que peixes do gênero Astymax (Lambari), da represa de Camargos, possuem seus pesos distribuídos segundo uma Normal, com média 13,4g.

a) Se uma amostra de 15 lambaris for pescada e apresentar peso médio de 15,2g, e um desvio padrão de 4,8g, a afirmação do biólogo pode ser considerada falsa com 5% de significância?

b) E se 15,2g fosse a média de uma amostra de 35 lambrs, com o mesmo desvio padrão. Sua decisão mudaria com os mesmos 5% de significância?

3. (*) Os produtores de uma certa semente de milho afirmam que seu poder de germinação é de 92%. Um fazendeiro tomou 10 sementes e plantou-as para testar se é verdadeira tal afirmação. Oito sementes germinaram. Este fato pode substituir cientificamente a acusação de que as sementes não têm 92% de germinação?

4. (*) Uma montadora de automóveis anuncia que seus carros consomem, em média, 11 litros a cada 100Km (considere que o consumo possua uma distribuição Normal). Você compra dois automóveis dessa marca e verifica que consumiram 12 litros por 100Km, em média, com desvio padrão de 0,8 litros.

O que você pode concluir sobre o anúncio da montadora a um nível de significância de 5%.

5. (*) Um pesquisador da área de saúde deseja mostrar que os indivíduos portadores de febre amarela apresentam um teor de glicose inferior à média dos indivíduos não portadores, que é de 120mg/dL . Para tanto, coletou uma amostra de sangue em sete indivíduos portadores de febre amarela, e fez a avaliação do teor de glicose. Os resultados obtidos foram: 119, 122, 120, 110, 112, 115 e 116.

À 10% de significância, qual deveria ser a conclusão do pesquisador?

6. (*) O pão de cevada é um dos alimentos mais ricos em ferro de origem vegetal da alimentação ocidental. Um *site* da internet afirma que esse alimento tem, em média, $6,5\text{mg}/100\text{g}$ de ferro. Para testar se essa afirmação procede, uma amostra de 25 pães foi coletada apresentando média de $4,8\text{mg}/100\text{g}$ e desvio padrão de $1,5\text{mg}/100\text{g}$. Sabendo disso, teste a afirmação do site com 5% de significância e decida se o site diz a verdade.

Lista de exercícios:

Teste para a comparação entre duas médias (μ_1 e μ_2)

1. (**) Para comparar 2 procedimentos cirúrgicos foram feitas 16 cirurgias, anotando-se os os dias até a recuperação. Os resultados foram:

Procedimento 1	1,1 1,0 1,4 1,3 1,5 0,9
Procedimento 2	2,0 2,5 2,4 2,1 1,8 1,9 1,9 2,3 2,5 2,6

- Com 5% de significância, infira sobre a igualdade das variâncias populacionais.
- De posse da decisão da letra (a), compare as médias dos procedimentos cirúrgicos, também adotando uma significância de 5%.
- Explícite quais são as duas formas de acertar e os dois erros possíveis nesse contexto.

2. (**) Um administrador colecionou dados sobre o aumento na produtividade no último ano para uma amostra de empresas que produzem equipamentos para mecanização agrícola. As firmas foram classificadas de acordo com seus investimentos em pesquisa e desenvolvimento nos últimos 3 anos. Os resultados do estudo seguem abaixo, pelo qual o aumento na produtividade foi medido numa escala de 0 à 10.

Investimento	Firmas					
	1	2	3	4	5	6
Baixo	7,6	8,2	6,8	5,8	6,9	6,6
Alto	8,5	9,7	10,1	7,8	9,6	9,5

- Escreva o par de hipóteses e explícite quais são as duas formas de acertar e os dois erros possíveis nesse contexto.
- Escolha uma significância ao considerar um dos dois erros o mais grave. Justifique.
- Com o α escolhido, infira sobre a igualdade das variâncias populacionais.
- De posse da decisão da letra (c), compare as médias das produtividades, também adotando o mesmo nível de significância.

3. (*) Num torneio de voleibol, a recuperação energética dos jogadores após uma partida é de suma importância. Um grupo de nutricionistas desenvolveu 2 cardápios com o objetivo de melhorar a recuperação energética (%) e diagnosticar qual dieta foi eficiente nesta recuperação. Foram observados 5 atletas para cada dieta, obtendo-se os seguintes resultados:

Dieta A	45	51	50	62	43
Dieta B	45	35	43	59	48

Decida se as dietas são equivalentes ou se existe uma melhor, ao nível de 5% de significância.

4. (**) O departamento de marketing de uma grande indústria de alimentos está fazendo um estudo para mudar a embalagem de seu macarrão instantâneo. Duas embalagens (A e B) foram avaliadas, quanto à aceitação, por 10 consumidores desse produto. A escala utilizada foi: 1 (péssimo), 2 (ruim), 3 (regular), 4 (bom) e 5 (ótimo). Os dados amostrais seguem.

	Consumidor									
	1	2	3	4	5	6	7	8	9	10
Embalagem A	3	3	4	4	4	4	4	5	5	5
Embalagem B	3	3	3	3	2	2	2	1	1	1

- Escreva o par de hipóteses e explicita quais são as duas formas de acertar, e os dois erros possíveis nesse contexto.
- Escolha uma significância ao considerar um dos dois erros o mais grave. Justifique.
- Com o α escolhido, infira sobre a igualdade das variâncias populacionais.
- De posse da decisão da letra (c), compare as notas médias das embalagens, também adotando o mesmo nível de significância.

5. (*) Dois irmãos gostam da mesma marca de carro. Então, decidiram comprar dois exemplares idênticos, ambos zero quilômetros. Um deles viaja muito e sempre anota o consumo do carro na estrada. Por sua vez, o outro nunca viaja, mas anota o consumo do carro dentro da cidade. Os dados a seguir estão em km/L de gasolina.

Cidade	7,7	7,4	9,3	8,5	10,0	8,9	10,0	9,2	8,8	8,9
Estrada	12,3	10,2	10,9	11,9	10,9	11,5	11,7	9,6	10,7	11,0

Compare o consumo na estrada e na cidade, verificando se os consumos médios são iguais.

6. (**) O responsável geral por um posto de saúde recebe um lote de vacinas contra a gripe comum (influenza), com o prazo de validade apagado. Ele precisa decidir entre vacinar ou não os moradores de 10 bairros da redondeza.

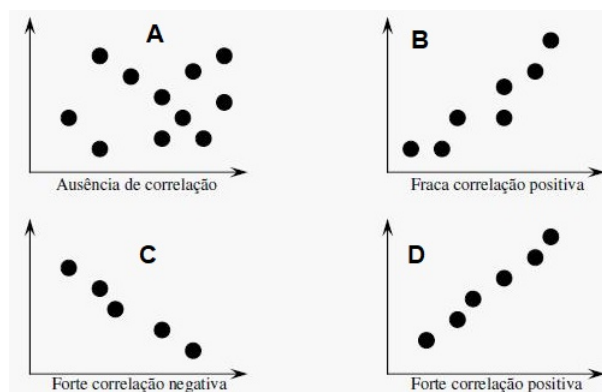
Se ele vacinar e as vacinas estiverem vencidas, não haverá proteção contra a gripe, mas também não haverá dano à saúde. Todo o trabalho será em vão. Se ele decidir descartar o lote, por outro lado, pode estar jogando vacina boa (dentro da validade) fora.

Refleta sobre essa situação e responda corretamente:

- Quais são as duas formas que o gerente tem de acertar?
- Quais são as duas formas que o gerente tem de errar?
- Na sua opinião, qual é o erro mais grave?
- Qual o valor você sugere para α , nesse caso. Por quê?

Lista de exercícios: Correlação e Regressão

1. (*) Dê um palpite para o valor do coeficiente de correlação linear representado em cada figura, de A a D.



2. (**) Foi feito um estudo sobre o peso final (Y) de peixes tratados com doses extras de ração (X). Os resultados obtidos foram os seguintes:

Ração (g/peixe)	0	5	10	15	20
Peso (g)	495	560	590	620	615

- Atribua as notações X e Y para as variáveis independente e dependente desse problema.
- Faça um diagrama de dispersão para os dados.
- Estime o coeficiente de correlação linear de Pearson (ρ) e interprete-o.
- Estime a equação de regressão que melhor se ajusta aos dados.
- Plote a equação estimada no gráfico de dispersão.
- Qual é o peso esperado para um peixe que se alimente com 8g extras de ração?
- Qual é o peso esperado para um peixe que se alimente com 60g extras de ração?

3. (**) Procurou-se realizar um estudo com o objetivo de saber o efeito na produção de massa muscular de um grupo de pessoas alimentadas com diferentes teores de proteína. Foram obtidos os seguintes dados:

Teor de proteína (%)	10	12	14	16	18	20	22
Massa muscular (%)	11,8	10,2	12,1	13,2	12,1	15,4	15,6

- Atribua as notações X e Y para as variáveis independente e dependente desse problema.
- Faça um diagrama de dispersão para os dados.
- Estime o coeficiente de correlação linear de Pearson (ρ) e interprete-o.
- Estime a equação de regressão que melhor se ajusta aos dados.
- Plote a equação estimada no gráfico de dispersão.
- Qual é a massa muscular esperada para uma pessoa que se alimente com 5% de proteína?
- Qual é a massa muscular esperada para uma pessoa que se alimente com 15% de proteína?



APÊNDICE B: TABELAS

Tabela 14 – Quantis superiores da distribuição de qui-quadrado (χ^2_α) com ν graus de liberdade, e para diferentes valores da probabilidade (α) de acordo com o seguinte evento: $P(\chi^2 > \chi^2_\alpha) = \alpha$.

ν	0,995	0,990	0,975	0,950	0,900	0,750	0,500
1	0,000039	0,000157	0,000982	0,003932	0,015791	0,101532	0,455
2	0,010025	0,020101	0,050636	0,102587	0,210721	0,575364	1,386
3	0,071721	0,114831	0,215793	0,351843	0,584369	1,213	2,366
4	0,206989	0,297109	0,484418	0,710723	1,064	1,923	3,357
5	0,411742	0,554298	0,831212	1,145	1,610	2,675	4,351
6	0,675727	0,872090	1,237	1,635	2,204	3,455	5,348
7	0,989256	1,239	1,690	2,167	2,833	4,255	6,346
8	1,344	1,646	2,180	2,733	3,490	5,071	7,344
9	1,735	2,088	2,700	3,325	4,168	5,899	8,343
10	2,156	2,558	3,247	3,940	4,865	6,737	9,342
11	2,603	3,053	3,816	4,575	5,578	7,584	10,341
12	3,074	3,571	4,404	5,226	6,304	8,438	11,340
13	3,565	4,107	5,009	5,892	7,042	9,299	12,340
14	4,075	4,660	5,629	6,571	7,790	10,165	13,339
15	4,601	5,229	6,262	7,261	8,547	11,037	14,339
16	5,142	5,812	6,908	7,962	9,312	11,912	15,338
17	5,697	6,408	7,564	8,672	10,085	12,792	16,338
18	6,265	7,015	8,231	9,390	10,865	13,675	17,338
19	6,844	7,633	8,907	10,117	11,651	14,562	18,338
20	7,434	8,260	9,591	10,851	12,443	15,452	19,337
21	8,034	8,897	10,283	11,591	13,240	16,344	20,337
22	8,643	9,542	10,982	12,338	14,041	17,240	21,337
23	9,260	10,196	11,689	13,091	14,848	18,137	22,337
24	9,886	10,856	12,401	13,848	15,659	19,037	23,337
25	10,520	11,524	13,120	14,611	16,473	19,939	24,337
26	11,160	12,198	13,844	15,379	17,292	20,843	25,336
27	11,808	12,879	14,573	16,151	18,114	21,749	26,336
28	12,461	13,565	15,308	16,928	18,939	22,657	27,336
29	13,121	14,256	16,047	17,708	19,768	23,567	28,336
30	13,787	14,953	16,791	18,493	20,599	24,478	29,336
40	20,707	22,164	24,433	26,509	29,051	33,660	39,335
50	27,991	29,707	32,357	34,764	37,689	42,942	49,335
60	35,534	37,485	40,482	43,188	46,459	52,294	59,335
120	83,852	86,923	91,573	95,705	100,624	109,220	119,334
240	187,324	191,990	198,984	205,135	212,386	224,882	239,334
480	403,949	410,874	421,189	430,198	440,745	458,754	479,334
960	850,891	861,015	876,028	889,081	904,291	930,093	959,333

Tabela 15 – Quantis superiores da distribuição de qui-quadrado (χ^2_α) com ν graus de liberdade, e para diferentes valores da probabilidade (α) de acordo com o seguinte evento: $P(\chi^2 > \chi^2_\alpha) = \alpha$.

ν	0,500	0,250	0,100	0,050	0,025	0,010	0,005
1	0,454940	1,323	2,706	3,841	5,024	6,635	7,879
2	1,386	2,773	4,605	5,991	7,378	9,210	10,597
3	2,366	4,108	6,251	7,815	9,348	11,345	12,838
4	3,357	5,385	7,779	9,488	11,143	13,277	14,860
5	4,351	6,626	9,236	11,070	12,833	15,086	16,750
6	5,348	7,841	10,645	12,592	14,449	16,812	18,548
7	6,346	9,037	12,017	14,067	16,013	18,475	20,278
8	7,344	10,219	13,362	15,507	17,535	20,090	21,955
9	8,343	11,389	14,684	16,919	19,023	21,666	23,589
10	9,342	12,549	15,987	18,307	20,483	23,209	25,188
11	10,341	13,701	17,275	19,675	21,920	24,725	26,757
12	11,340	14,845	18,549	21,026	23,337	26,217	28,300
13	12,340	15,984	19,812	22,362	24,736	27,688	29,819
14	13,339	17,117	21,064	23,685	26,119	29,141	31,319
15	14,339	18,245	22,307	24,996	27,488	30,578	32,801
16	15,338	19,369	23,542	26,296	28,845	32,000	34,267
17	16,338	20,489	24,769	27,587	30,191	33,409	35,718
18	17,338	21,605	25,989	28,869	31,526	34,805	37,156
19	18,338	22,718	27,204	30,144	32,852	36,191	38,582
20	19,337	23,828	28,412	31,410	34,170	37,566	39,997
21	20,337	24,935	29,615	32,671	35,479	38,932	41,401
22	21,337	26,039	30,813	33,924	36,781	40,289	42,796
23	22,337	27,141	32,007	35,172	38,076	41,638	44,181
24	23,337	28,241	33,196	36,415	39,364	42,980	45,559
25	24,337	29,339	34,382	37,652	40,646	44,314	46,928
26	25,336	30,435	35,563	38,885	41,923	45,642	48,290
27	26,336	31,528	36,741	40,113	43,195	46,963	49,645
28	27,336	32,620	37,916	41,337	44,461	48,278	50,993
29	28,336	33,711	39,087	42,557	45,722	49,588	52,336
30	29,336	34,800	40,256	43,773	46,979	50,892	53,672
40	39,335	45,616	51,805	55,758	59,342	63,691	66,766
50	49,335	56,334	63,167	67,505	71,420	76,154	79,490
60	59,335	66,981	74,397	79,082	83,298	88,379	91,952
120	119,334	130,055	140,233	146,567	152,211	158,950	163,648
240	239,334	254,392	268,471	277,138	284,802	293,888	300,182
480	479,334	500,519	520,111	532,075	542,599	555,006	563,561
960	959,333	989,180	1016,566	1033,193	1047,760	1064,867	1076,621

Tabela 16 – Quantis superiores da distribuição de F ($F_{0,10}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador (valor da probabilidade (α) de 10% de acordo com o seguinte evento: $P(F > F_{0,10}) = 0,10$).

ν_2	ν_1										
	1	2	3	4	5	6	7	8	9	10	11
1	39,86	49,50	53,59	55,83	57,24	58,20	58,91	59,44	59,86	60,20	60,47
2	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39	9,40
3	5,54	5,46	5,39	5,34	5,31	5,28	5,26	5,25	5,24	5,23	5,22
4	4,55	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92	3,91
5	4,06	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,30	3,28
6	3,78	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,94	2,92
7	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70	2,68
8	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54	2,52
9	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42	2,40
10	3,29	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32	2,30
11	3,23	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27	2,25	2,23
12	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19	2,17
13	3,14	2,76	2,56	2,43	2,35	2,28	2,23	2,20	2,16	2,14	2,12
14	3,10	2,73	2,52	2,39	2,31	2,24	2,19	2,15	2,12	2,10	2,07
15	3,07	2,70	2,49	2,36	2,27	2,21	2,16	2,12	2,09	2,06	2,04
16	3,05	2,67	2,46	2,33	2,24	2,18	2,13	2,09	2,06	2,03	2,01
17	3,03	2,64	2,44	2,31	2,22	2,15	2,10	2,06	2,03	2,00	1,98
18	3,01	2,62	2,42	2,29	2,20	2,13	2,08	2,04	2,00	1,98	1,95
19	2,99	2,61	2,40	2,27	2,18	2,11	2,06	2,02	1,98	1,96	1,93
20	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,96	1,94	1,91
21	2,96	2,57	2,36	2,23	2,14	2,08	2,02	1,98	1,95	1,92	1,90
22	2,95	2,56	2,35	2,22	2,13	2,06	2,01	1,97	1,93	1,90	1,88
23	2,94	2,55	2,34	2,21	2,11	2,05	1,99	1,95	1,92	1,89	1,87
24	2,93	2,54	2,33	2,19	2,10	2,04	1,98	1,94	1,91	1,88	1,85
25	2,92	2,53	2,32	2,18	2,09	2,02	1,97	1,93	1,89	1,87	1,84
26	2,91	2,52	2,31	2,17	2,08	2,01	1,96	1,92	1,88	1,86	1,83
27	2,90	2,51	2,30	2,17	2,07	2,00	1,95	1,91	1,87	1,85	1,82
28	2,89	2,50	2,29	2,16	2,06	2,00	1,94	1,90	1,87	1,84	1,81
29	2,89	2,50	2,28	2,15	2,06	1,99	1,93	1,89	1,86	1,83	1,80
30	2,88	2,49	2,28	2,14	2,05	1,98	1,93	1,88	1,85	1,82	1,79
40	2,84	2,44	2,23	2,09	2,00	1,93	1,87	1,83	1,79	1,76	1,74
50	2,81	2,41	2,20	2,06	1,97	1,90	1,84	1,80	1,76	1,73	1,70
60	2,79	2,39	2,18	2,04	1,95	1,87	1,82	1,77	1,74	1,71	1,68
120	2,75	2,35	2,13	1,99	1,90	1,82	1,77	1,72	1,68	1,65	1,63
240	2,73	2,32	2,11	1,97	1,87	1,80	1,74	1,70	1,66	1,63	1,60
480	2,72	2,31	2,10	1,96	1,86	1,79	1,73	1,68	1,64	1,61	1,58
960	2,71	2,31	2,09	1,95	1,85	1,78	1,72	1,68	1,64	1,61	1,58
∞	2,71	2,30	2,08	1,94	1,85	1,77	1,72	1,67	1,63	1,60	1,57

Tabela 17 – Quantis superiores da distribuição de F ($F_{0,10}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 10% de acordo com o seguinte evento: $P(F > F_{0,10}) = 0,10$.

ν_2	ν_1										
	12	13	14	15	20	30	40	60	120	240	∞
1	60,71	60,90	61,07	61,22	61,74	62,26	62,53	62,79	63,06	63,19	63,33
2	9,41	9,41	9,42	9,42	9,44	9,46	9,47	9,47	9,48	9,49	9,49
3	5,21	5,21	5,20	5,20	5,18	5,16	5,15	5,14	5,13	5,11	5,13
4	3,90	3,89	3,88	3,87	3,84	3,82	3,80	3,79	3,78	3,77	3,76
5	3,27	3,26	3,25	3,24	3,21	3,17	3,16	3,14	3,12	3,12	3,10
6	2,90	2,89	2,88	2,87	2,84	2,80	2,78	2,76	2,74	2,73	2,72
7	2,67	2,65	2,64	2,63	2,59	2,56	2,54	2,51	2,49	2,48	2,47
8	2,50	2,49	2,48	2,46	2,42	2,38	2,36	2,34	2,32	2,30	2,29
9	2,38	2,36	2,35	2,34	2,30	2,25	2,23	2,21	2,18	2,17	2,16
10	2,28	2,27	2,26	2,24	2,20	2,16	2,13	2,11	2,08	2,07	2,06
11	2,21	2,19	2,18	2,17	2,12	2,08	2,05	2,03	2,00	1,99	1,97
12	2,15	2,13	2,12	2,10	2,06	2,01	1,99	1,96	1,93	1,92	1,90
13	2,10	2,08	2,07	2,05	2,01	1,96	1,93	1,90	1,88	1,86	1,85
14	2,05	2,04	2,02	2,01	1,96	1,91	1,89	1,86	1,83	1,81	1,80
15	2,02	2,00	1,99	1,97	1,92	1,87	1,85	1,82	1,79	1,77	1,76
16	1,99	1,97	1,95	1,94	1,89	1,84	1,81	1,78	1,75	1,73	1,72
17	1,96	1,94	1,93	1,91	1,86	1,81	1,78	1,75	1,72	1,70	1,69
18	1,93	1,92	1,90	1,89	1,84	1,78	1,75	1,72	1,69	1,67	1,66
19	1,91	1,89	1,88	1,86	1,81	1,76	1,73	1,70	1,67	1,65	1,63
20	1,89	1,87	1,86	1,84	1,79	1,74	1,71	1,68	1,64	1,63	1,61
21	1,87	1,86	1,84	1,83	1,78	1,72	1,69	1,66	1,62	1,60	1,59
22	1,86	1,84	1,83	1,81	1,76	1,70	1,67	1,64	1,60	1,59	1,57
23	1,84	1,83	1,81	1,80	1,74	1,69	1,66	1,62	1,59	1,57	1,55
24	1,83	1,81	1,80	1,78	1,73	1,67	1,64	1,61	1,57	1,55	1,53
25	1,82	1,80	1,79	1,77	1,72	1,66	1,63	1,59	1,56	1,54	1,52
26	1,81	1,79	1,77	1,76	1,71	1,65	1,61	1,58	1,54	1,52	1,50
27	1,80	1,78	1,76	1,75	1,70	1,64	1,60	1,57	1,53	1,51	1,49
28	1,79	1,77	1,75	1,74	1,69	1,63	1,59	1,56	1,52	1,50	1,48
29	1,78	1,76	1,75	1,73	1,68	1,62	1,58	1,55	1,51	1,49	1,47
30	1,77	1,75	1,74	1,72	1,67	1,61	1,57	1,54	1,50	1,48	1,46
40	1,71	1,70	1,68	1,66	1,61	1,54	1,51	1,47	1,42	1,40	1,38
50	1,68	1,66	1,64	1,63	1,57	1,50	1,46	1,42	1,38	1,35	1,33
60	1,66	1,64	1,62	1,60	1,54	1,48	1,44	1,40	1,35	1,32	1,29
120	1,60	1,58	1,56	1,55	1,48	1,41	1,37	1,32	1,26	1,23	1,19
240	1,57	1,55	1,53	1,52	1,45	1,38	1,33	1,28	1,22	1,18	1,13
480	1,56	1,54	1,52	1,50	1,44	1,36	1,31	1,26	1,19	1,15	1,09
960	1,55	1,53	1,51	1,49	1,43	1,35	1,30	1,25	1,18	1,14	1,06
∞	1,55	1,52	1,50	1,49	1,42	1,34	1,30	1,24	1,17	1,12	1,00

Tabela 18 – Quantis superiores da distribuição de F ($F_{0,05}$) com v_1 graus de liberdade do numerador, e v_2 graus de liberdade do denominador valor da probabilidade (α) de 5% de acordo com o seguinte evento: $P(F > F_{0,05}) = 0,05$.

v_2	v_1										
	1	2	3	4	5	6	7	8	9	10	11
1	161,45	199,50	215,70	224,58	230,16	234,0	236,8	238,9	240,5	241,9	242,98
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,40
3	10,13	9,55	9,27	9,11	9,01	8,94	8,88	8,84	8,81	8,78	8,76
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,94
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,70
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,60
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,31
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,10
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,94
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,82
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,72
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,63
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,57
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,51
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,46
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,41
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,37
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,34
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,31
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,28
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,26
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,24
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,22
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,20
26	4,23	3,37	2,97	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,18
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,17
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,15
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,14
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,13
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,04
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03	1,99
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,95
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91	1,87
240	3,88	3,03	2,64	2,41	2,25	2,14	2,05	1,98	1,92	1,87	1,83
480	3,86	3,01	2,62	2,39	2,23	2,12	2,03	1,96	1,90	1,85	1,81
960	3,85	3,01	2,61	2,38	2,22	2,11	2,02	1,95	1,89	1,84	1,80
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,79

Tabela 19 – Quantis superiores da distribuição de F ($F_{0,05}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 5% de acordo com o seguinte evento: $P(F > F_{0,05}) = 0,05$.

ν_2	ν_1										
	12	13	14	15	20	30	40	60	120	240	∞
1	243,91	244,69	245,36	245,95	248,0	250,1	251,1	252,2	253,3	253,8	254,31
2	19,41	19,42	19,42	19,43	19,45	19,46	19,47	19,48	19,49	19,49	19,50
3	8,74	8,72	8,71	8,69	8,65	8,60	8,57	8,54	8,49	8,42	8,53
4	5,91	5,89	5,87	5,86	5,80	5,75	5,72	5,69	5,66	5,64	5,63
5	4,68	4,66	4,64	4,62	4,56	4,50	4,46	4,43	4,40	4,39	4,36
6	4,00	3,98	3,96	3,94	3,87	3,81	3,77	3,74	3,70	3,69	3,67
7	3,57	3,55	3,53	3,51	3,44	3,38	3,34	3,30	3,27	3,25	3,23
8	3,28	3,26	3,24	3,22	3,15	3,08	3,04	3,01	2,97	2,95	2,93
9	3,07	3,05	3,03	3,01	2,94	2,86	2,83	2,79	2,75	2,73	2,71
10	2,91	2,89	2,86	2,85	2,77	2,70	2,66	2,62	2,58	2,56	2,54
11	2,79	2,76	2,74	2,72	2,65	2,57	2,53	2,49	2,45	2,43	2,40
12	2,69	2,66	2,64	2,62	2,54	2,47	2,43	2,38	2,34	2,32	2,30
13	2,60	2,58	2,55	2,53	2,46	2,38	2,34	2,30	2,25	2,23	2,21
14	2,53	2,51	2,48	2,46	2,39	2,31	2,27	2,22	2,18	2,15	2,13
15	2,48	2,45	2,42	2,40	2,33	2,25	2,20	2,16	2,11	2,09	2,07
16	2,42	2,40	2,37	2,35	2,28	2,19	2,15	2,11	2,06	2,03	2,01
17	2,38	2,35	2,33	2,31	2,23	2,15	2,10	2,06	2,01	1,99	1,96
18	2,34	2,31	2,29	2,27	2,19	2,11	2,06	2,02	1,97	1,94	1,92
19	2,31	2,28	2,26	2,23	2,16	2,07	2,03	1,98	1,93	1,90	1,88
20	2,28	2,25	2,22	2,20	2,12	2,04	1,99	1,95	1,90	1,87	1,84
21	2,25	2,22	2,20	2,18	2,10	2,01	1,96	1,92	1,87	1,84	1,81
22	2,23	2,20	2,17	2,15	2,07	1,98	1,94	1,89	1,84	1,81	1,78
23	2,20	2,18	2,15	2,13	2,05	1,96	1,91	1,86	1,81	1,79	1,76
24	2,18	2,15	2,13	2,11	2,03	1,94	1,89	1,84	1,79	1,76	1,73
25	2,16	2,14	2,11	2,09	2,01	1,92	1,87	1,82	1,77	1,74	1,71
26	2,15	2,12	2,09	2,07	1,99	1,90	1,85	1,80	1,75	1,72	1,69
27	2,13	2,10	2,08	2,06	1,97	1,88	1,84	1,79	1,73	1,70	1,67
28	2,12	2,09	2,06	2,04	1,96	1,87	1,82	1,77	1,71	1,68	1,65
29	2,10	2,08	2,05	2,03	1,94	1,85	1,81	1,75	1,70	1,67	1,64
30	2,09	2,06	2,04	2,01	1,93	1,84	1,79	1,74	1,68	1,65	1,62
40	2,00	1,97	1,95	1,92	1,84	1,74	1,69	1,64	1,58	1,54	1,51
50	1,95	1,92	1,89	1,87	1,78	1,69	1,63	1,58	1,51	1,48	1,44
60	1,92	1,89	1,86	1,84	1,75	1,65	1,59	1,53	1,47	1,43	1,39
120	1,83	1,80	1,78	1,75	1,66	1,55	1,50	1,43	1,35	1,31	1,25
240	1,79	1,76	1,73	1,71	1,61	1,51	1,44	1,37	1,29	1,24	1,17
480	1,77	1,74	1,71	1,69	1,59	1,48	1,42	1,35	1,26	1,20	1,12
960	1,76	1,73	1,70	1,68	1,58	1,47	1,41	1,33	1,24	1,18	1,08
∞	1,75	1,72	1,69	1,67	1,57	1,46	1,39	1,32	1,22	1,15	1,00

Tabela 20 – Quantis superiores da distribuição de F ($F_{0,025}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 2,5% de acordo com o seguinte evento: $P(F > F_{0,025}) = 0,025$.

ν_2	ν_1										
	1	2	3	4	5	6	7	8	9	10	11
1	647,79	799,50	864,14	899,58	921,85	937,1	948,2	956,7	963,3	968,6	973,03
2	38,51	39,00	39,15	39,25	39,30	39,33	39,36	39,37	39,39	39,40	39,41
3	17,44	16,03	15,42	15,08	14,87	14,71	14,60	14,51	14,44	14,39	14,34
4	12,22	10,65	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84	8,79
5	10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62	6,57
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46	5,41
7	8,07	6,54	5,89	5,52	5,29	5,12	5,00	4,90	4,82	4,76	4,71
8	7,57	6,06	5,41	5,05	4,82	4,65	4,53	4,43	4,36	4,30	4,24
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96	3,91
10	6,94	5,46	4,82	4,47	4,24	4,07	3,95	3,85	3,78	3,72	3,66
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53	3,47
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37	3,32
13	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31	3,25	3,20
14	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21	3,15	3,09
15	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06	3,01
16	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05	2,99	2,93
17	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98	2,92	2,87
18	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93	2,87	2,81
19	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,88	2,82	2,76
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77	2,72
21	5,83	4,42	3,82	3,48	3,25	3,09	2,97	2,87	2,80	2,73	2,68
22	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76	2,70	2,65
23	5,75	4,35	3,75	3,41	3,18	3,02	2,90	2,81	2,73	2,67	2,62
24	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70	2,64	2,59
25	5,69	4,29	3,69	3,35	3,13	2,97	2,85	2,75	2,68	2,61	2,56
26	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,65	2,59	2,54
27	5,63	4,24	3,65	3,31	3,08	2,92	2,80	2,71	2,63	2,57	2,51
28	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,61	2,55	2,49
29	5,59	4,20	3,61	3,27	3,04	2,88	2,76	2,67	2,59	2,53	2,48
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51	2,46
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39	2,33
50	5,34	3,97	3,39	3,05	2,83	2,67	2,55	2,46	2,38	2,32	2,26
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27	2,22
120	5,15	3,80	3,23	2,89	2,67	2,52	2,39	2,30	2,22	2,16	2,10
240	5,09	3,75	3,17	2,84	2,62	2,46	2,34	2,25	2,17	2,10	2,05
480	5,06	3,72	3,14	2,81	2,59	2,43	2,31	2,22	2,14	2,08	2,02
960	5,04	3,70	3,13	2,80	2,58	2,42	2,30	2,21	2,13	2,06	2,01
∞	5,02	3,69	3,12	2,79	2,57	2,41	2,29	2,19	2,11	2,05	1,99

Tabela 21 – Quantis superiores da distribuição de F ($F_{0,025}$) com v_1 graus de liberdade do numerador, e v_2 graus de liberdade do denominador valor da probabilidade (α) de 2,5% de acordo com o seguinte evento: $P(F > F_{0,025}) = 0,025$.

v_2	v_1										
	12	13	14	15	20	30	40	60	120	240	∞
1	976,71	979,84	982,53	984,87	993,1	1001	1006	1010	1014	1016	1018
2	39,41	39,42	39,43	39,43	39,45	39,46	39,47	39,48	39,49	39,49	39,50
3	14,30	14,27	14,24	14,21	14,11	14,00	13,94	13,85	13,68	13,43	13,90
4	8,75	8,71	8,68	8,66	8,56	8,46	8,41	8,36	8,31	8,28	8,26
5	6,53	6,49	6,45	6,43	6,33	6,22	6,17	6,12	6,06	6,03	6,02
6	5,37	5,33	5,30	5,27	5,17	5,07	5,01	4,96	4,90	4,88	4,85
7	4,67	4,63	4,60	4,57	4,47	4,36	4,31	4,26	4,20	4,17	4,14
8	4,20	4,16	4,13	4,10	4,00	3,89	3,84	3,78	3,73	3,70	3,67
9	3,87	3,83	3,80	3,77	3,67	3,56	3,51	3,45	3,39	3,36	3,33
10	3,62	3,58	3,55	3,52	3,42	3,31	3,26	3,20	3,14	3,11	3,08
11	3,43	3,39	3,36	3,33	3,23	3,12	3,06	3,00	2,94	2,91	2,88
12	3,28	3,24	3,21	3,18	3,07	2,96	2,91	2,85	2,79	2,76	2,72
13	3,15	3,12	3,08	3,05	2,95	2,84	2,78	2,72	2,66	2,63	2,60
14	3,05	3,01	2,98	2,95	2,84	2,73	2,67	2,61	2,55	2,52	2,49
15	2,96	2,92	2,89	2,86	2,76	2,64	2,59	2,52	2,46	2,43	2,40
16	2,89	2,85	2,82	2,79	2,68	2,57	2,51	2,45	2,38	2,35	2,32
17	2,82	2,79	2,75	2,72	2,62	2,50	2,44	2,38	2,32	2,28	2,25
18	2,77	2,73	2,70	2,67	2,56	2,44	2,38	2,32	2,26	2,22	2,19
19	2,72	2,68	2,65	2,62	2,51	2,39	2,33	2,27	2,20	2,17	2,13
20	2,68	2,64	2,60	2,57	2,46	2,35	2,29	2,22	2,16	2,12	2,09
21	2,64	2,60	2,56	2,53	2,42	2,31	2,25	2,18	2,11	2,08	2,04
22	2,60	2,56	2,53	2,50	2,39	2,27	2,21	2,14	2,08	2,04	2,00
23	2,57	2,53	2,50	2,47	2,36	2,24	2,18	2,11	2,04	2,01	1,97
24	2,54	2,50	2,47	2,44	2,33	2,21	2,15	2,08	2,01	1,97	1,94
25	2,51	2,48	2,44	2,41	2,30	2,18	2,12	2,05	1,98	1,94	1,91
26	2,49	2,45	2,42	2,39	2,28	2,16	2,09	2,03	1,95	1,92	1,88
27	2,47	2,43	2,39	2,36	2,25	2,13	2,07	2,00	1,93	1,89	1,85
28	2,45	2,41	2,37	2,34	2,23	2,11	2,05	1,98	1,91	1,87	1,83
29	2,43	2,39	2,36	2,32	2,21	2,09	2,03	1,96	1,89	1,85	1,81
30	2,41	2,37	2,34	2,31	2,20	2,07	2,01	1,94	1,87	1,83	1,79
40	2,29	2,25	2,21	2,18	2,07	1,94	1,88	1,80	1,72	1,68	1,64
50	2,22	2,18	2,14	2,11	1,99	1,87	1,80	1,72	1,64	1,59	1,55
60	2,17	2,13	2,09	2,06	1,94	1,82	1,74	1,67	1,58	1,53	1,48
120	2,05	2,01	1,98	1,94	1,82	1,69	1,61	1,53	1,43	1,38	1,31
240	2,00	1,96	1,92	1,89	1,77	1,63	1,55	1,46	1,35	1,29	1,21
480	1,97	1,93	1,89	1,86	1,74	1,60	1,52	1,42	1,31	1,24	1,14
960	1,96	1,92	1,88	1,85	1,72	1,58	1,50	1,41	1,29	1,21	1,10
∞	1,94	1,90	1,87	1,83	1,71	1,57	1,48	1,39	1,27	1,19	1,00

Tabela 22 – Quantis superiores da distribuição de F ($F_{0,01}$) com v_1 graus de liberdade do numerador, e v_2 graus de liberdade do denominador valor da probabilidade (α) de 1% de acordo com o seguinte evento: $P(F > F_{0,01}) = 0,01$.

v_2	v_1										
	1	2	3	4	5	6	7	8	9	10	11
1	4052	4999	5403	5625	5764	5859	5928	5981	6022	6056	6083
2	98,50	99,01	99,05	99,24	99,30	99,33	99,36	99,37	99,39	99,40	99,41
3	34,12	30,74	29,34	28,60	28,11	27,77	27,52	27,32	27,16	27,03	26,92
4	21,20	18,00	16,68	15,98	15,52	15,21	14,97	14,80	14,66	14,55	14,45
5	16,26	13,28	12,05	11,39	10,96	10,67	10,45	10,29	10,15	10,05	9,96
6	13,75	10,92	9,77	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,79
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,54
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,73
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,18
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,77
11	9,65	7,21	6,21	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,46
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,22
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	4,02
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,86
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,73
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,62
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,52
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,43
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,36
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,29
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,24
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,18
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,14
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,09
25	7,77	5,57	4,67	4,18	3,85	3,63	3,46	3,32	3,22	3,13	3,06
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	3,02
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,99
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,96
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,93
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,91
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,73
50	7,17	5,06	4,20	3,72	3,41	3,19	3,02	2,89	2,78	2,70	2,63
60	7,08	4,98	4,12	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,56
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,40
240	6,74	4,69	3,86	3,40	3,09	2,88	2,71	2,59	2,48	2,40	2,32
480	6,69	4,65	3,82	3,36	3,06	2,84	2,68	2,55	2,44	2,36	2,28
960	6,66	4,63	3,80	3,34	3,04	2,82	2,66	2,53	2,43	2,34	2,27
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,25

Tabela 23 – Quantis superiores da distribuição de F ($F_{0,01}$) com v_1 graus de liberdade do numerador, e v_2 graus de liberdade do denominador valor da probabilidade (α) de 1% de acordo com o seguinte evento: $P(F > F_{0,01}) = 0,01$.

v_2	v_1										
	12	13	14	15	20	30	40	60	120	240	∞
1	6106	6126	6143	6157	6209	6261	6287	6313	6339	6353	6366
2	99,42	99,42	99,43	99,43	99,45	99,47	99,47	99,48	99,49	99,49	99,50
3	26,82	26,74	26,67	26,60	26,34	26,02	25,79	25,43	24,62	23,38	26,13
4	14,37	14,31	14,25	14,20	14,02	13,84	13,74	13,65	13,56	13,51	13,46
5	9,88	9,82	9,77	9,72	9,55	9,37	9,28	9,19	9,08	9,01	9,02
6	7,72	7,66	7,60	7,56	7,40	7,23	7,14	7,06	6,97	6,92	6,88
7	6,47	6,41	6,36	6,31	6,16	5,99	5,91	5,83	5,74	5,70	5,65
8	5,67	5,61	5,56	5,52	5,36	5,20	5,12	5,03	4,95	4,90	4,86
9	5,11	5,05	5,01	4,96	4,81	4,65	4,57	4,48	4,40	4,36	4,31
10	4,71	4,65	4,60	4,56	4,41	4,25	4,17	4,08	4,00	3,95	3,91
11	4,40	4,34	4,29	4,25	4,10	3,94	3,86	3,78	3,69	3,65	3,60
12	4,16	4,10	4,05	4,01	3,86	3,70	3,62	3,54	3,45	3,41	3,36
13	3,96	3,91	3,86	3,82	3,66	3,51	3,43	3,34	3,25	3,21	3,17
14	3,80	3,75	3,70	3,66	3,51	3,35	3,27	3,18	3,09	3,05	3,00
15	3,67	3,61	3,56	3,52	3,37	3,21	3,13	3,05	2,96	2,91	2,87
16	3,55	3,50	3,45	3,41	3,26	3,10	3,02	2,93	2,84	2,80	2,75
17	3,46	3,40	3,35	3,31	3,16	3,00	2,92	2,83	2,75	2,70	2,65
18	3,37	3,32	3,27	3,23	3,08	2,92	2,84	2,75	2,66	2,61	2,57
19	3,30	3,24	3,19	3,15	3,00	2,84	2,76	2,67	2,58	2,54	2,49
20	3,23	3,18	3,13	3,09	2,94	2,78	2,69	2,61	2,52	2,47	2,42
21	3,17	3,12	3,07	3,03	2,88	2,72	2,64	2,55	2,46	2,41	2,36
22	3,12	3,07	3,02	2,98	2,83	2,67	2,58	2,50	2,40	2,35	2,31
23	3,07	3,02	2,97	2,93	2,78	2,62	2,54	2,45	2,35	2,31	2,26
24	3,03	2,98	2,93	2,89	2,74	2,58	2,49	2,40	2,31	2,26	2,21
25	2,99	2,94	2,89	2,85	2,70	2,54	2,45	2,36	2,27	2,22	2,17
26	2,96	2,90	2,86	2,81	2,66	2,50	2,42	2,33	2,23	2,18	2,13
27	2,93	2,87	2,82	2,78	2,63	2,47	2,38	2,29	2,20	2,15	2,10
28	2,90	2,84	2,79	2,75	2,60	2,44	2,35	2,26	2,17	2,12	2,06
29	2,87	2,81	2,77	2,73	2,57	2,41	2,33	2,23	2,14	2,09	2,03
30	2,84	2,79	2,74	2,70	2,55	2,39	2,30	2,21	2,11	2,06	2,01
40	2,66	2,61	2,56	2,52	2,37	2,20	2,11	2,02	1,92	1,86	1,80
50	2,56	2,51	2,46	2,42	2,27	2,10	2,01	1,91	1,80	1,74	1,68
60	2,50	2,44	2,39	2,35	2,20	2,03	1,94	1,84	1,73	1,67	1,60
120	2,34	2,28	2,23	2,19	2,03	1,86	1,76	1,66	1,53	1,46	1,38
240	2,26	2,20	2,16	2,11	1,96	1,78	1,68	1,57	1,43	1,35	1,25
480	2,22	2,17	2,12	2,08	1,92	1,74	1,63	1,52	1,38	1,29	1,17
960	2,20	2,15	2,10	2,06	1,90	1,72	1,61	1,50	1,35	1,26	1,11
∞	2,18	2,13	2,08	2,04	1,88	1,70	1,59	1,47	1,32	1,22	1,00

Tabela 24 – Quantis superiores da distribuição de F ($F_{0,005}$) com v_1 graus de liberdade do numerador, e v_2 graus de liberdade do denominador valor da probabilidade (α) de 0,5% de acordo com o seguinte evento: $P(F > F_{0,005}) = 0,005$.

v_2	v_1										
	1	2	3	4	5	6	7	8	9	10	11
1	16211	20000	21614	22500	23056	23437	23715	23925	24091	24224	24334
2	198,50	199,04	198,70	199,21	199,29	199,3	199,4	199,4	199,4	199,4	199,41
3	55,55	49,49	47,03	45,76	44,90	44,29	43,83	43,46	43,16	42,91	42,68
4	31,33	26,28	24,23	23,15	22,45	21,97	21,62	21,35	21,14	20,96	20,82
5	22,78	18,31	16,51	15,55	14,93	14,51	14,19	13,95	13,76	13,61	13,48
6	18,64	14,54	12,90	12,03	11,46	11,07	10,79	10,57	10,39	10,25	10,13
7	16,24	12,40	10,87	10,05	9,52	9,16	8,89	8,68	8,51	8,38	8,27
8	14,69	11,04	9,59	8,80	8,30	7,95	7,69	7,50	7,34	7,21	7,10
9	13,61	10,11	8,71	7,96	7,47	7,13	6,89	6,69	6,54	6,42	6,31
10	12,83	9,43	8,08	7,34	6,87	6,54	6,30	6,12	5,97	5,85	5,75
11	12,23	8,91	7,60	6,88	6,42	6,10	5,86	5,68	5,54	5,42	5,32
12	11,75	8,51	7,22	6,52	6,07	5,76	5,52	5,35	5,20	5,09	4,99
13	11,37	8,19	6,92	6,23	5,79	5,48	5,25	5,08	4,94	4,82	4,72
14	11,06	7,92	6,68	6,00	5,56	5,26	5,03	4,86	4,72	4,60	4,51
15	10,80	7,70	6,47	5,80	5,37	5,07	4,85	4,67	4,54	4,42	4,33
16	10,58	7,51	6,30	5,64	5,21	4,91	4,69	4,52	4,38	4,27	4,18
17	10,38	7,35	6,15	5,50	5,07	4,78	4,56	4,39	4,25	4,14	4,05
18	10,22	7,21	6,02	5,37	4,96	4,66	4,44	4,28	4,14	4,03	3,94
19	10,07	7,09	5,91	5,27	4,85	4,56	4,34	4,18	4,04	3,93	3,84
20	9,94	6,99	5,81	5,17	4,76	4,47	4,26	4,09	3,96	3,85	3,76
21	9,83	6,89	5,73	5,09	4,68	4,39	4,18	4,01	3,88	3,77	3,68
22	9,73	6,81	5,65	5,02	4,61	4,32	4,11	3,94	3,81	3,70	3,61
23	9,63	6,73	5,58	4,95	4,54	4,26	4,05	3,88	3,75	3,64	3,55
24	9,55	6,66	5,52	4,89	4,49	4,20	3,99	3,83	3,69	3,59	3,50
25	9,48	6,60	5,46	4,83	4,43	4,15	3,94	3,78	3,64	3,54	3,45
26	9,41	6,54	5,41	4,79	4,38	4,10	3,89	3,73	3,60	3,49	3,40
27	9,34	6,49	5,36	4,74	4,34	4,06	3,85	3,69	3,56	3,45	3,36
28	9,28	6,44	5,31	4,70	4,30	4,02	3,81	3,65	3,52	3,41	3,32
29	9,23	6,40	5,27	4,66	4,26	3,98	3,77	3,61	3,48	3,38	3,29
30	9,18	6,35	5,24	4,62	4,23	3,95	3,74	3,58	3,45	3,34	3,25
40	8,83	6,07	4,97	4,37	3,99	3,71	3,51	3,35	3,22	3,12	3,03
50	8,63	5,90	4,82	4,23	3,85	3,58	3,38	3,22	3,09	2,99	2,90
60	8,49	5,79	4,73	4,14	3,76	3,49	3,29	3,13	3,01	2,90	2,82
120	8,18	5,54	4,50	3,92	3,55	3,28	3,09	2,93	2,81	2,71	2,62
240	8,03	5,42	4,38	3,82	3,45	3,19	2,99	2,84	2,71	2,61	2,52
480	7,95	5,36	4,33	3,77	3,40	3,14	2,94	2,79	2,67	2,56	2,48
960	7,92	5,33	4,30	3,74	3,37	3,11	2,92	2,77	2,64	2,54	2,46
∞	7,88	5,30	4,28	3,72	3,35	3,09	2,90	2,74	2,62	2,52	2,43

Tabela 25 – Quantis superiores da distribuição de F ($F_{0,005}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 0,5% de acordo com o seguinte evento: $P(F > F_{0,005}) = 0,005$.

ν_2	ν_1										
	12	13	14	15	20	30	40	60	120	240	∞
1	24426	24505	24572	24630	24836	25044	25148	25253	25359	25411	25464
2	199,42	199,42	199,43	199,43	199,4	199,5	199,5	199,5	199,5	199,5	199,50
3	42,49	42,32	42,16	42,02	41,43	40,59	39,93	38,85	36,37	32,90	41,83
4	20,70	20,60	20,51	20,44	20,16	19,89	19,75	19,61	19,47	19,39	19,32
5	13,37	13,28	13,20	13,13	12,89	12,64	12,50	12,36	12,20	12,07	12,14
6	10,03	9,95	9,88	9,81	9,59	9,36	9,24	9,12	9,00	8,94	8,88
7	8,18	8,10	8,03	7,97	7,76	7,54	7,42	7,31	7,20	7,15	7,08
8	7,01	6,94	6,87	6,81	6,61	6,40	6,29	6,18	6,06	6,01	5,95
9	6,23	6,15	6,09	6,03	5,83	5,63	5,52	5,41	5,30	5,25	5,19
10	5,66	5,59	5,53	5,47	5,27	5,07	4,97	4,86	4,75	4,69	4,64
11	5,24	5,16	5,10	5,05	4,86	4,65	4,55	4,45	4,34	4,28	4,23
12	4,91	4,84	4,77	4,72	4,53	4,33	4,23	4,12	4,01	3,96	3,90
13	4,64	4,57	4,51	4,46	4,27	4,07	3,97	3,87	3,76	3,70	3,65
14	4,43	4,36	4,30	4,25	4,06	3,86	3,76	3,66	3,55	3,49	3,44
15	4,25	4,18	4,12	4,07	3,88	3,69	3,59	3,48	3,37	3,32	3,26
16	4,10	4,03	3,97	3,92	3,73	3,54	3,44	3,33	3,22	3,17	3,11
17	3,97	3,90	3,84	3,79	3,61	3,41	3,31	3,21	3,10	3,04	2,98
18	3,86	3,79	3,73	3,68	3,50	3,30	3,20	3,10	2,99	2,93	2,87
19	3,76	3,70	3,64	3,59	3,40	3,21	3,11	3,00	2,89	2,83	2,78
20	3,68	3,61	3,55	3,50	3,32	3,12	3,02	2,92	2,81	2,75	2,69
21	3,60	3,54	3,48	3,43	3,24	3,05	2,95	2,84	2,73	2,67	2,61
22	3,54	3,47	3,41	3,36	3,18	2,98	2,88	2,77	2,66	2,60	2,55
23	3,47	3,41	3,35	3,30	3,12	2,92	2,82	2,71	2,60	2,54	2,48
24	3,42	3,35	3,30	3,25	3,06	2,87	2,77	2,66	2,55	2,49	2,43
25	3,37	3,30	3,25	3,20	3,01	2,82	2,72	2,61	2,50	2,44	2,38
26	3,33	3,26	3,20	3,15	2,97	2,77	2,67	2,56	2,45	2,39	2,33
27	3,28	3,22	3,16	3,11	2,93	2,73	2,63	2,52	2,41	2,35	2,29
28	3,25	3,18	3,12	3,07	2,89	2,69	2,59	2,48	2,37	2,31	2,25
29	3,21	3,15	3,09	3,04	2,86	2,66	2,56	2,45	2,33	2,27	2,21
30	3,18	3,11	3,06	3,01	2,82	2,63	2,52	2,42	2,30	2,24	2,18
40	2,95	2,89	2,83	2,78	2,60	2,40	2,30	2,18	2,06	2,00	1,93
50	2,82	2,76	2,70	2,65	2,47	2,27	2,16	2,05	1,93	1,86	1,79
60	2,74	2,68	2,62	2,57	2,39	2,19	2,08	1,96	1,83	1,76	1,69
120	2,54	2,48	2,42	2,37	2,19	1,98	1,87	1,75	1,61	1,52	1,43
240	2,45	2,39	2,33	2,28	2,09	1,89	1,77	1,64	1,49	1,40	1,28
480	2,40	2,34	2,28	2,23	2,05	1,84	1,72	1,59	1,43	1,33	1,19
960	2,38	2,32	2,26	2,21	2,02	1,81	1,69	1,56	1,40	1,29	1,13
∞	2,36	2,29	2,24	2,19	2,00	1,79	1,67	1,53	1,36	1,25	1,00

Tabela 27 – Quantis superiores da distribuição de F ($F_{0,90}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 90% de acordo com o seguinte evento: $P(F > F_{0,90}) = 0,90$.

ν_2	ν_1										
	12	13	14	15	20	30	40	60	120	240	∞
1	0,0165	0,0164	0,0164	0,0163	0,0162	0,0161	0,0160	0,0159	0,0159	0,0158	0,0158
2	0,1063	0,1062	0,1062	0,1061	0,1059	0,1057	0,1056	0,1055	0,1055	0,1054	0,1054
3	0,1917	0,1919	0,1921	0,1923	0,1929	0,1935	0,1938	0,1941	0,1945	0,1946	0,1948
4	0,2567	0,2573	0,2579	0,2584	0,2601	0,2620	0,2629	0,2639	0,2649	0,2654	0,2659
5	0,3060	0,3071	0,3080	0,3088	0,3119	0,3151	0,3167	0,3184	0,3202	0,3211	0,3221
6	0,3443	0,3458	0,3471	0,3483	0,3526	0,3571	0,3596	0,3621	0,3647	0,3660	0,3674
7	0,3748	0,3767	0,3784	0,3799	0,3854	0,3913	0,3945	0,3977	0,4012	0,4029	0,4047
8	0,3997	0,4020	0,4040	0,4058	0,4124	0,4196	0,4235	0,4275	0,4317	0,4339	0,4362
9	0,4204	0,4230	0,4253	0,4274	0,4351	0,4435	0,4480	0,4528	0,4578	0,4604	0,4631
10	0,4378	0,4408	0,4434	0,4457	0,4544	0,4639	0,4691	0,4746	0,4804	0,4834	0,4865
11	0,4527	0,4560	0,4589	0,4614	0,4710	0,4816	0,4874	0,4936	0,5001	0,5035	0,5071
12	0,4657	0,4692	0,4723	0,4751	0,4855	0,4971	0,5035	0,5103	0,5175	0,5213	0,5253
13	0,4770	0,4807	0,4841	0,4871	0,4983	0,5108	0,5177	0,5251	0,5331	0,5373	0,5417
14	0,4869	0,4909	0,4945	0,4976	0,5096	0,5230	0,5305	0,5384	0,5470	0,5516	0,5564
15	0,4958	0,5000	0,5037	0,5070	0,5197	0,5340	0,5419	0,5504	0,5597	0,5646	0,5698
16	0,5037	0,5081	0,5120	0,5155	0,5287	0,5438	0,5522	0,5613	0,5712	0,5765	0,5820
17	0,5108	0,5154	0,5194	0,5231	0,5370	0,5528	0,5616	0,5712	0,5817	0,5873	0,5932
18	0,5172	0,5220	0,5262	0,5300	0,5444	0,5610	0,5702	0,5803	0,5914	0,5973	0,6036
19	0,5231	0,5280	0,5323	0,5363	0,5512	0,5684	0,5781	0,5887	0,6003	0,6066	0,6132
20	0,5284	0,5335	0,5380	0,5420	0,5575	0,5753	0,5854	0,5964	0,6085	0,6151	0,6221
21	0,5333	0,5385	0,5431	0,5473	0,5632	0,5816	0,5921	0,6035	0,6162	0,6231	0,6305
22	0,5378	0,5431	0,5479	0,5522	0,5685	0,5875	0,5983	0,6102	0,6234	0,6306	0,6382
23	0,5420	0,5474	0,5523	0,5567	0,5734	0,5930	0,6041	0,6164	0,6301	0,6376	0,6456
24	0,5459	0,5514	0,5564	0,5608	0,5780	0,5980	0,6095	0,6222	0,6364	0,6441	0,6524
25	0,5494	0,5551	0,5601	0,5647	0,5822	0,6028	0,6146	0,6276	0,6422	0,6503	0,6589
26	0,5528	0,5585	0,5637	0,5683	0,5862	0,6072	0,6193	0,6327	0,6478	0,6561	0,6651
27	0,5559	0,5617	0,5670	0,5717	0,5899	0,6114	0,6238	0,6375	0,6530	0,6616	0,6709
28	0,5588	0,5647	0,5701	0,5749	0,5934	0,6153	0,6279	0,6420	0,6580	0,6668	0,6764
29	0,5615	0,5675	0,5729	0,5778	0,5967	0,6190	0,6319	0,6463	0,6627	0,6718	0,6816
30	0,5641	0,5702	0,5757	0,5806	0,5998	0,6225	0,6356	0,6504	0,6672	0,6765	0,6866
40	0,5832	0,5900	0,5960	0,6015	0,6230	0,6489	0,6642	0,6816	0,7019	0,7134	0,7263
50	0,5952	0,6023	0,6088	0,6147	0,6377	0,6659	0,6827	0,7021	0,7252	0,7386	0,7538
60	0,6033	0,6108	0,6175	0,6237	0,6479	0,6777	0,6957	0,7167	0,7421	0,7571	0,7743
120	0,6245	0,6328	0,6403	0,6472	0,6747	0,7095	0,7312	0,7574	0,7908	0,8120	0,8385
240	0,6356	0,6443	0,6523	0,6596	0,6890	0,7268	0,7509	0,7806	0,8203	0,8473	0,8849
480	0,6412	0,6502	0,6584	0,6660	0,6964	0,7359	0,7613	0,7931	0,8371	0,8685	0,9182
960	0,6441	0,6532	0,6615	0,6692	0,7001	0,7405	0,7666	0,7997	0,8461	0,8806	0,9420
∞	0,6469	0,6562	0,6646	0,6724	0,7039	0,7452	0,7721	0,8065	0,8557	0,8940	1,0000

Tabela 28 – Quantis superiores da distribuição de F ($F_{0,95}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 95% de acordo com o seguinte evento: $P(F > F_{0,95}) = 0,95$.

ν_2	ν_1										
	1	2	3	4	5	6	7	8	9	10	11
1	0,0062	0,0050	0,0046	0,0045	0,0043	0,0043	0,0042	0,0042	0,0042	0,0041	0,0041
2	0,0540	0,0526	0,0522	0,0520	0,0518	0,0517	0,0517	0,0516	0,0516	0,0516	0,0515
3	0,0987	0,1047	0,1078	0,1097	0,1109	0,1118	0,1125	0,1131	0,1135	0,1138	0,1141
4	0,1297	0,1440	0,1517	0,1565	0,1598	0,1623	0,1641	0,1655	0,1667	0,1677	0,1685
5	0,1513	0,1728	0,1849	0,1926	0,1980	0,2020	0,2051	0,2075	0,2095	0,2112	0,2126
6	0,1670	0,1944	0,2102	0,2206	0,2279	0,2334	0,2377	0,2411	0,2440	0,2463	0,2483
7	0,1788	0,2111	0,2301	0,2427	0,2518	0,2587	0,2641	0,2684	0,2720	0,2750	0,2775
8	0,1881	0,2243	0,2459	0,2606	0,2712	0,2793	0,2857	0,2909	0,2951	0,2988	0,3018
9	0,1954	0,2349	0,2589	0,2752	0,2872	0,2964	0,3037	0,3096	0,3146	0,3187	0,3223
10	0,2014	0,2437	0,2697	0,2875	0,3007	0,3108	0,3189	0,3256	0,3311	0,3358	0,3398
11	0,2064	0,2511	0,2788	0,2979	0,3121	0,3231	0,3320	0,3392	0,3453	0,3504	0,3549
12	0,2106	0,2574	0,2865	0,3068	0,3220	0,3338	0,3432	0,3511	0,3576	0,3632	0,3680
13	0,2143	0,2628	0,2932	0,3146	0,3305	0,3430	0,3531	0,3614	0,3684	0,3744	0,3796
14	0,2174	0,2675	0,2991	0,3213	0,3380	0,3512	0,3618	0,3706	0,3780	0,3843	0,3898
15	0,2201	0,2716	0,3042	0,3273	0,3447	0,3584	0,3695	0,3787	0,3865	0,3931	0,3989
16	0,2225	0,2752	0,3087	0,3326	0,3506	0,3648	0,3763	0,3859	0,3941	0,4010	0,4071
17	0,2247	0,2784	0,3128	0,3373	0,3559	0,3706	0,3825	0,3925	0,4009	0,4082	0,4145
18	0,2266	0,2813	0,3165	0,3416	0,3606	0,3758	0,3881	0,3984	0,4071	0,4146	0,4212
19	0,2283	0,2839	0,3198	0,3454	0,3650	0,3805	0,3932	0,4038	0,4128	0,4205	0,4273
20	0,2298	0,2863	0,3227	0,3489	0,3689	0,3848	0,3978	0,4087	0,4179	0,4259	0,4329
21	0,2312	0,2885	0,3255	0,3521	0,3725	0,3887	0,4020	0,4131	0,4226	0,4309	0,4380
22	0,2325	0,2904	0,3280	0,3550	0,3758	0,3923	0,4059	0,4173	0,4270	0,4354	0,4428
23	0,2337	0,2922	0,3303	0,3577	0,3788	0,3956	0,4095	0,4211	0,4310	0,4396	0,4471
24	0,2348	0,2939	0,3324	0,3602	0,3816	0,3987	0,4128	0,4246	0,4347	0,4435	0,4512
25	0,2358	0,2954	0,3343	0,3625	0,3842	0,4015	0,4158	0,4279	0,4382	0,4471	0,4550
26	0,2367	0,2968	0,3361	0,3646	0,3866	0,4042	0,4187	0,4309	0,4414	0,4505	0,4585
27	0,2375	0,2981	0,3378	0,3666	0,3888	0,4067	0,4214	0,4338	0,4444	0,4537	0,4618
28	0,2383	0,2994	0,3394	0,3684	0,3909	0,4090	0,4239	0,4364	0,4472	0,4566	0,4649
29	0,2391	0,3005	0,3408	0,3702	0,3929	0,4111	0,4262	0,4389	0,4499	0,4594	0,4677
30	0,2398	0,3016	0,3422	0,3718	0,3947	0,4131	0,4284	0,4413	0,4523	0,4620	0,4705
40	0,2448	0,3094	0,3523	0,3837	0,4083	0,4281	0,4446	0,4587	0,4708	0,4814	0,4908
50	0,2479	0,3142	0,3584	0,3911	0,4166	0,4374	0,4547	0,4695	0,4823	0,4935	0,5035
60	0,2499	0,3174	0,3626	0,3960	0,4222	0,4436	0,4616	0,4769	0,4902	0,5019	0,5122
120	0,2551	0,3255	0,3731	0,4086	0,4367	0,4598	0,4792	0,4959	0,5105	0,5234	0,5350
240	0,2577	0,3297	0,3785	0,4151	0,4441	0,4681	0,4883	0,5058	0,5211	0,5347	0,5468
480	0,2590	0,3317	0,3812	0,4183	0,4479	0,4723	0,4929	0,5108	0,5265	0,5404	0,5529
960	0,2597	0,3328	0,3825	0,4200	0,4498	0,4744	0,4953	0,5133	0,5292	0,5433	0,5560
∞	0,2603	0,3338	0,3839	0,4216	0,4517	0,4765	0,4976	0,5159	0,5319	0,5462	0,5591

Tabela 29 – Quantis superiores da distribuição de F ($F_{0,95}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador (valor da probabilidade (α) de 95% de acordo com o seguinte evento: $P(F > F_{0,95}) = 0,95$).

ν_2	ν_1										
	12	13	14	15	20	30	40	60	120	240	∞
1	0,0041	0,0041	0,0041	0,0041	0,0040	0,0040	0,0040	0,0040	0,0039	0,0039	0,0039
2	0,0515	0,0515	0,0515	0,0515	0,0514	0,0514	0,0514	0,0513	0,0513	0,0513	0,0513
3	0,1144	0,1146	0,1147	0,1149	0,1155	0,1161	0,1164	0,1167	0,1170	0,1171	0,1173
4	0,1692	0,1697	0,1703	0,1707	0,1723	0,1740	0,1749	0,1758	0,1767	0,1772	0,1777
5	0,2138	0,2148	0,2157	0,2165	0,2194	0,2224	0,2240	0,2257	0,2274	0,2282	0,2291
6	0,2500	0,2515	0,2528	0,2539	0,2581	0,2626	0,2650	0,2674	0,2699	0,2712	0,2726
7	0,2797	0,2817	0,2833	0,2848	0,2903	0,2962	0,2994	0,3026	0,3060	0,3078	0,3096
8	0,3045	0,3068	0,3089	0,3107	0,3174	0,3247	0,3286	0,3327	0,3370	0,3393	0,3416
9	0,3254	0,3281	0,3305	0,3327	0,3405	0,3492	0,3539	0,3588	0,3640	0,3667	0,3695
10	0,3433	0,3464	0,3491	0,3515	0,3605	0,3704	0,3758	0,3815	0,3876	0,3908	0,3940
11	0,3587	0,3621	0,3651	0,3678	0,3779	0,3890	0,3951	0,4016	0,4085	0,4121	0,4159
12	0,3722	0,3759	0,3792	0,3821	0,3931	0,4055	0,4122	0,4194	0,4272	0,4313	0,4355
13	0,3841	0,3881	0,3916	0,3948	0,4067	0,4201	0,4275	0,4354	0,4440	0,4485	0,4532
14	0,3946	0,3988	0,4026	0,4060	0,4188	0,4332	0,4412	0,4499	0,4592	0,4641	0,4693
15	0,4040	0,4085	0,4125	0,4161	0,4296	0,4451	0,4537	0,4629	0,4730	0,4784	0,4841
16	0,4124	0,4171	0,4214	0,4251	0,4395	0,4558	0,4650	0,4749	0,4857	0,4915	0,4976
17	0,4201	0,4250	0,4294	0,4333	0,4484	0,4656	0,4753	0,4858	0,4973	0,5036	0,5101
18	0,4270	0,4321	0,4367	0,4408	0,4565	0,4746	0,4848	0,4959	0,5081	0,5147	0,5217
19	0,4333	0,4386	0,4433	0,4476	0,4639	0,4828	0,4935	0,5052	0,5181	0,5251	0,5325
20	0,4391	0,4445	0,4494	0,4539	0,4708	0,4904	0,5016	0,5138	0,5273	0,5347	0,5425
21	0,4444	0,4500	0,4551	0,4596	0,4771	0,4975	0,5090	0,5218	0,5360	0,5437	0,5520
22	0,4493	0,4551	0,4603	0,4649	0,4829	0,5040	0,5160	0,5293	0,5441	0,5522	0,5608
23	0,4538	0,4597	0,4651	0,4699	0,4884	0,5101	0,5225	0,5362	0,5516	0,5601	0,5692
24	0,4580	0,4641	0,4695	0,4745	0,4934	0,5157	0,5286	0,5428	0,5588	0,5676	0,5770
25	0,4619	0,4681	0,4737	0,4787	0,4981	0,5211	0,5342	0,5489	0,5655	0,5746	0,5845
26	0,4656	0,4719	0,4776	0,4827	0,5026	0,5260	0,5396	0,5547	0,5718	0,5813	0,5915
27	0,4690	0,4754	0,4812	0,4864	0,5067	0,5307	0,5446	0,5602	0,5778	0,5876	0,5982
28	0,4722	0,4787	0,4846	0,4899	0,5106	0,5351	0,5494	0,5653	0,5835	0,5936	0,6046
29	0,4752	0,4818	0,4878	0,4932	0,5142	0,5393	0,5539	0,5702	0,5889	0,5993	0,6106
30	0,4780	0,4847	0,4908	0,4963	0,5177	0,5432	0,5581	0,5749	0,5940	0,6047	0,6164
40	0,4991	0,5066	0,5134	0,5196	0,5438	0,5733	0,5907	0,6108	0,6343	0,6477	0,6627
50	0,5124	0,5204	0,5277	0,5344	0,5605	0,5927	0,6121	0,6347	0,6616	0,6774	0,6953
60	0,5215	0,5299	0,5376	0,5445	0,5721	0,6064	0,6272	0,6518	0,6815	0,6993	0,7198
120	0,5453	0,5548	0,5634	0,5713	0,6029	0,6434	0,6688	0,6998	0,7397	0,7653	0,7975
240	0,5578	0,5678	0,5770	0,5854	0,6194	0,6636	0,6920	0,7275	0,7754	0,8083	0,8547
480	0,5642	0,5745	0,5840	0,5927	0,6280	0,6743	0,7044	0,7426	0,7959	0,8345	0,8962
960	0,5675	0,5779	0,5875	0,5964	0,6323	0,6798	0,7108	0,7505	0,8069	0,8494	0,9261
∞	0,5707	0,5813	0,5911	0,6001	0,6367	0,6854	0,7174	0,7587	0,8187	0,8660	1,0000

Tabela 30 – Quantis superiores da distribuição de F ($F_{0,975}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 97,5% de acordo com o seguinte evento: $P(F > F_{0,975}) = 0,975$.

ν_2	ν_1										
	1	2	3	4	5	6	7	8	9	10	11
1	0,0015	0,0013	0,0012	0,0011	0,0011	0,0011	0,0011	0,0010	0,0010	0,0010	0,0010
2	0,0260	0,0256	0,0255	0,0255	0,0254	0,0254	0,0254	0,0254	0,0254	0,0254	0,0254
3	0,0573	0,0623	0,0648	0,0662	0,0672	0,0679	0,0684	0,0688	0,0691	0,0694	0,0696
4	0,0818	0,0939	0,1002	0,1041	0,1068	0,1087	0,1102	0,1114	0,1123	0,1131	0,1137
5	0,0999	0,1186	0,1288	0,1354	0,1399	0,1433	0,1459	0,1480	0,1497	0,1511	0,1523
6	0,1135	0,1377	0,1515	0,1606	0,1670	0,1718	0,1756	0,1786	0,1810	0,1831	0,1849
7	0,1239	0,1529	0,1698	0,1811	0,1892	0,1954	0,2002	0,2041	0,2073	0,2100	0,2123
8	0,1321	0,1650	0,1846	0,1979	0,2076	0,2150	0,2208	0,2256	0,2295	0,2328	0,2357
9	0,1387	0,1750	0,1969	0,2120	0,2230	0,2315	0,2383	0,2438	0,2484	0,2523	0,2556
10	0,1442	0,1833	0,2072	0,2238	0,2361	0,2456	0,2532	0,2594	0,2646	0,2690	0,2729
11	0,1487	0,1903	0,2160	0,2339	0,2473	0,2577	0,2661	0,2729	0,2787	0,2836	0,2879
12	0,1526	0,1962	0,2235	0,2426	0,2570	0,2682	0,2773	0,2848	0,2910	0,2964	0,3011
13	0,1559	0,2014	0,2300	0,2503	0,2655	0,2774	0,2871	0,2952	0,3019	0,3077	0,3127
14	0,1588	0,2059	0,2358	0,2569	0,2730	0,2856	0,2959	0,3044	0,3116	0,3178	0,3231
15	0,1613	0,2099	0,2408	0,2629	0,2796	0,2929	0,3036	0,3126	0,3202	0,3268	0,3325
16	0,1635	0,2134	0,2453	0,2681	0,2855	0,2993	0,3106	0,3200	0,3280	0,3349	0,3409
17	0,1655	0,2165	0,2493	0,2729	0,2909	0,3052	0,3169	0,3267	0,3350	0,3422	0,3485
18	0,1673	0,2193	0,2529	0,2771	0,2957	0,3105	0,3226	0,3327	0,3414	0,3489	0,3554
19	0,1689	0,2219	0,2562	0,2810	0,3001	0,3153	0,3278	0,3383	0,3472	0,3550	0,3617
20	0,1703	0,2242	0,2592	0,2845	0,3040	0,3197	0,3325	0,3433	0,3525	0,3605	0,3675
21	0,1716	0,2262	0,2619	0,2877	0,3077	0,3237	0,3369	0,3479	0,3574	0,3657	0,3729
22	0,1728	0,2282	0,2643	0,2907	0,3110	0,3274	0,3409	0,3522	0,3619	0,3704	0,3778
23	0,1739	0,2299	0,2666	0,2934	0,3141	0,3308	0,3445	0,3562	0,3661	0,3748	0,3824
24	0,1749	0,2315	0,2687	0,2959	0,3170	0,3339	0,3480	0,3598	0,3700	0,3788	0,3866
25	0,1759	0,2330	0,2707	0,2982	0,3196	0,3369	0,3511	0,3632	0,3736	0,3826	0,3906
26	0,1767	0,2344	0,2725	0,3004	0,3221	0,3396	0,3541	0,3664	0,3770	0,3862	0,3943
27	0,1775	0,2357	0,2742	0,3024	0,3244	0,3421	0,3569	0,3694	0,3801	0,3895	0,3977
28	0,1783	0,2369	0,2758	0,3043	0,3265	0,3445	0,3595	0,3721	0,3830	0,3926	0,4010
29	0,1790	0,2381	0,2772	0,3061	0,3285	0,3467	0,3619	0,3747	0,3858	0,3955	0,4040
30	0,1796	0,2391	0,2786	0,3077	0,3304	0,3488	0,3642	0,3772	0,3884	0,3982	0,4069
40	0,1844	0,2469	0,2887	0,3199	0,3444	0,3644	0,3811	0,3954	0,4078	0,4187	0,4284
50	0,1873	0,2516	0,2950	0,3274	0,3530	0,3740	0,3917	0,4068	0,4200	0,4316	0,4420
60	0,1892	0,2548	0,2992	0,3325	0,3589	0,3806	0,3989	0,4147	0,4284	0,4405	0,4513
120	0,1941	0,2628	0,3099	0,3455	0,3740	0,3976	0,4176	0,4349	0,4501	0,4636	0,4757
240	0,1966	0,2669	0,3154	0,3522	0,3817	0,4063	0,4272	0,4454	0,4614	0,4757	0,4885
480	0,1978	0,2690	0,3181	0,3556	0,3857	0,4107	0,4322	0,4508	0,4672	0,4819	0,4951
960	0,1984	0,2700	0,3195	0,3573	0,3876	0,4130	0,4347	0,4535	0,4702	0,4850	0,4985
∞	0,1990	0,2711	0,3209	0,3590	0,3896	0,4152	0,4372	0,4562	0,4731	0,4882	0,5018

Tabela 31 – Quantis superiores da distribuição de F ($F_{0,975}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 97,5% de acordo com o seguinte evento: $P(F > F_{0,975}) = 0,975$.

ν_2	ν_1										
	12	13	14	15	20	30	40	60	120	240	∞
1	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
2	0,0254	0,0254	0,0254	0,0254	0,0253	0,0253	0,0253	0,0253	0,0253	0,0253	0,0253
3	0,0698	0,0699	0,0700	0,0702	0,0706	0,0710	0,0712	0,0715	0,0717	0,0718	0,0719
4	0,1143	0,1147	0,1152	0,1155	0,1168	0,1182	0,1189	0,1196	0,1203	0,1207	0,1211
5	0,1533	0,1541	0,1549	0,1556	0,1580	0,1606	0,1619	0,1633	0,1648	0,1655	0,1662
6	0,1864	0,1877	0,1888	0,1898	0,1935	0,1974	0,1995	0,2017	0,2039	0,2050	0,2062
7	0,2143	0,2161	0,2176	0,2189	0,2239	0,2292	0,2321	0,2351	0,2382	0,2398	0,2414
8	0,2381	0,2403	0,2422	0,2438	0,2500	0,2568	0,2604	0,2642	0,2682	0,2703	0,2725
9	0,2585	0,2611	0,2633	0,2653	0,2727	0,2809	0,2853	0,2899	0,2948	0,2974	0,3000
10	0,2762	0,2791	0,2817	0,2840	0,2925	0,3020	0,3072	0,3127	0,3185	0,3215	0,3247
11	0,2916	0,2948	0,2977	0,3003	0,3100	0,3208	0,3267	0,3329	0,3397	0,3432	0,3469
12	0,3051	0,3087	0,3119	0,3147	0,3254	0,3375	0,3441	0,3512	0,3588	0,3628	0,3670
13	0,3171	0,3210	0,3245	0,3276	0,3393	0,3525	0,3598	0,3676	0,3761	0,3806	0,3853
14	0,3279	0,3320	0,3357	0,3391	0,3517	0,3660	0,3739	0,3825	0,3919	0,3968	0,4021
15	0,3375	0,3419	0,3458	0,3494	0,3629	0,3783	0,3868	0,3962	0,4063	0,4118	0,4175
16	0,3461	0,3508	0,3550	0,3587	0,3730	0,3894	0,3986	0,4087	0,4196	0,4255	0,4317
17	0,3540	0,3589	0,3633	0,3672	0,3823	0,3997	0,4095	0,4201	0,4319	0,4383	0,4450
18	0,3612	0,3663	0,3709	0,3750	0,3908	0,4091	0,4194	0,4308	0,4433	0,4501	0,4573
19	0,3677	0,3730	0,3778	0,3821	0,3986	0,4178	0,4286	0,4406	0,4539	0,4611	0,4688
20	0,3737	0,3792	0,3842	0,3886	0,4058	0,4258	0,4372	0,4498	0,4638	0,4714	0,4795
21	0,3793	0,3849	0,3901	0,3947	0,4124	0,4332	0,4451	0,4583	0,4730	0,4811	0,4897
22	0,3844	0,3902	0,3955	0,4003	0,4186	0,4402	0,4526	0,4663	0,4817	0,4901	0,4992
23	0,3891	0,3951	0,4005	0,4054	0,4243	0,4466	0,4595	0,4738	0,4898	0,4987	0,5082
24	0,3935	0,3997	0,4052	0,4103	0,4297	0,4527	0,4660	0,4808	0,4975	0,5068	0,5167
25	0,3976	0,4039	0,4096	0,4148	0,4347	0,4584	0,4721	0,4874	0,5048	0,5144	0,5248
26	0,4015	0,4079	0,4137	0,4190	0,4394	0,4637	0,4778	0,4937	0,5116	0,5216	0,5325
27	0,4051	0,4116	0,4176	0,4229	0,4438	0,4687	0,4832	0,4996	0,5181	0,5285	0,5398
28	0,4084	0,4151	0,4212	0,4266	0,4479	0,4735	0,4884	0,5051	0,5243	0,5351	0,5467
29	0,4116	0,4184	0,4246	0,4301	0,4519	0,4779	0,4932	0,5104	0,5302	0,5413	0,5533
30	0,4146	0,4215	0,4278	0,4334	0,4555	0,4822	0,4978	0,5155	0,5358	0,5472	0,5597
40	0,4370	0,4448	0,4519	0,4583	0,4836	0,5147	0,5333	0,5547	0,5800	0,5946	0,6108
50	0,4512	0,4596	0,4672	0,4742	0,5017	0,5359	0,5567	0,5810	0,6103	0,6275	0,6471
60	0,4610	0,4698	0,4778	0,4851	0,5143	0,5509	0,5734	0,6000	0,6325	0,6520	0,6747
120	0,4867	0,4966	0,5057	0,5141	0,5480	0,5917	0,6195	0,6536	0,6980	0,7267	0,7631
240	0,5002	0,5108	0,5206	0,5296	0,5661	0,6143	0,6455	0,6849	0,7386	0,7760	0,8291
480	0,5071	0,5181	0,5282	0,5375	0,5756	0,6262	0,6595	0,7020	0,7621	0,8062	0,8775
960	0,5107	0,5218	0,5321	0,5416	0,5804	0,6323	0,6667	0,7110	0,7748	0,8234	0,9125
∞	0,5142	0,5256	0,5360	0,5457	0,5853	0,6386	0,6741	0,7203	0,7884	0,8427	1,0000

Tabela 32 – Quantis superiores da distribuição de F ($F_{0,99}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 99% de acordo com o seguinte evento: $P(F > F_{0,99}) = 0,99$.

ν_2	ν_1										
	1	2	3	4	5	6	7	8	9	10	11
1	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002
2	0,0102	0,0101	0,0101	0,0101	0,0101	0,0101	0,0101	0,0101	0,0101	0,0101	0,0101
3	0,0293	0,0325	0,0339	0,0348	0,0354	0,0358	0,0361	0,0364	0,0366	0,0367	0,0369
4	0,0472	0,0556	0,0599	0,0626	0,0644	0,0658	0,0668	0,0676	0,0682	0,0687	0,0692
5	0,0615	0,0753	0,0829	0,0878	0,0912	0,0937	0,0956	0,0972	0,0984	0,0995	0,1004
6	0,0728	0,0915	0,1023	0,1093	0,1143	0,1181	0,1211	0,1234	0,1254	0,1270	0,1284
7	0,0817	0,1047	0,1183	0,1274	0,1340	0,1391	0,1430	0,1462	0,1488	0,1511	0,1529
8	0,0888	0,1156	0,1317	0,1427	0,1508	0,1570	0,1619	0,1659	0,1692	0,1720	0,1744
9	0,0947	0,1247	0,1430	0,1557	0,1651	0,1724	0,1782	0,1829	0,1869	0,1902	0,1931
10	0,0996	0,1323	0,1526	0,1668	0,1774	0,1857	0,1923	0,1978	0,2023	0,2062	0,2096
11	0,1037	0,1388	0,1609	0,1764	0,1881	0,1973	0,2047	0,2108	0,2159	0,2203	0,2241
12	0,1072	0,1444	0,1680	0,1848	0,1975	0,2074	0,2155	0,2223	0,2279	0,2328	0,2370
13	0,1102	0,1492	0,1742	0,1921	0,2057	0,2164	0,2252	0,2324	0,2386	0,2439	0,2485
14	0,1128	0,1535	0,1797	0,1986	0,2130	0,2244	0,2338	0,2415	0,2482	0,2538	0,2588
15	0,1152	0,1573	0,1846	0,2044	0,2195	0,2316	0,2415	0,2497	0,2568	0,2628	0,2681
16	0,1172	0,1606	0,1890	0,2095	0,2254	0,2380	0,2484	0,2571	0,2645	0,2709	0,2765
17	0,1191	0,1636	0,1929	0,2142	0,2306	0,2438	0,2547	0,2638	0,2716	0,2783	0,2842
18	0,1207	0,1663	0,1964	0,2184	0,2354	0,2491	0,2604	0,2699	0,2780	0,2850	0,2912
19	0,1222	0,1688	0,1996	0,2222	0,2398	0,2539	0,2656	0,2754	0,2839	0,2912	0,2977
20	0,1235	0,1710	0,2025	0,2257	0,2437	0,2583	0,2704	0,2806	0,2893	0,2969	0,3036
21	0,1247	0,1730	0,2052	0,2289	0,2474	0,2623	0,2748	0,2853	0,2943	0,3021	0,3090
22	0,1259	0,1749	0,2076	0,2318	0,2508	0,2661	0,2788	0,2896	0,2989	0,3070	0,3141
23	0,1269	0,1766	0,2099	0,2345	0,2539	0,2695	0,2826	0,2936	0,3032	0,3115	0,3188
24	0,1278	0,1781	0,2120	0,2371	0,2567	0,2727	0,2860	0,2974	0,3071	0,3156	0,3232
25	0,1287	0,1796	0,2139	0,2394	0,2594	0,2757	0,2893	0,3008	0,3108	0,3195	0,3272
26	0,1295	0,1810	0,2157	0,2415	0,2619	0,2785	0,2923	0,3041	0,3143	0,3232	0,3311
27	0,1303	0,1822	0,2173	0,2436	0,2642	0,2811	0,2951	0,3071	0,3175	0,3266	0,3346
28	0,1310	0,1834	0,2189	0,2455	0,2664	0,2835	0,2978	0,3100	0,3206	0,3298	0,3380
29	0,1316	0,1845	0,2204	0,2472	0,2684	0,2858	0,3003	0,3127	0,3234	0,3328	0,3412
30	0,1322	0,1855	0,2217	0,2489	0,2703	0,2879	0,3026	0,3152	0,3261	0,3357	0,3442
40	0,1367	0,1931	0,2319	0,2612	0,2846	0,3039	0,3201	0,3341	0,3463	0,3571	0,3667
50	0,1395	0,1978	0,2381	0,2689	0,2935	0,3138	0,3311	0,3460	0,3591	0,3706	0,3809
60	0,1413	0,2009	0,2424	0,2740	0,2995	0,3206	0,3386	0,3542	0,3679	0,3800	0,3908
120	0,1460	0,2089	0,2532	0,2874	0,3151	0,3383	0,3582	0,3755	0,3908	0,4045	0,4168
240	0,1483	0,2130	0,2588	0,2943	0,3232	0,3475	0,3684	0,3867	0,4029	0,4175	0,4306
480	0,1495	0,2151	0,2616	0,2978	0,3273	0,3522	0,3736	0,3924	0,4091	0,4241	0,4377
960	0,1501	0,2161	0,2630	0,2995	0,3293	0,3545	0,3762	0,3953	0,4122	0,4275	0,4413
∞	0,1507	0,2171	0,2644	0,3013	0,3314	0,3569	0,3789	0,3982	0,4154	0,4309	0,4449

Tabela 33 – Quantis superiores da distribuição de F ($F_{0,99}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 99% de acordo com o seguinte evento: $P(F > F_{0,99}) = 0,99$.

ν_2	ν_1										
	12	13	14	15	20	30	40	60	120	240	∞
1	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002
2	0,0101	0,0101	0,0101	0,0101	0,0101	0,0101	0,0101	0,0101	0,0101	0,0101	0,0101
3	0,0370	0,0371	0,0371	0,0372	0,0375	0,0377	0,0379	0,0380	0,0381	0,0382	0,0383
4	0,0696	0,0699	0,0702	0,0704	0,0713	0,0723	0,0728	0,0732	0,0738	0,0740	0,0743
5	0,1011	0,1018	0,1024	0,1029	0,1047	0,1066	0,1076	0,1087	0,1097	0,1103	0,1109
6	0,1296	0,1306	0,1315	0,1323	0,1352	0,1383	0,1400	0,1417	0,1435	0,1444	0,1453
7	0,1546	0,1560	0,1573	0,1584	0,1625	0,1669	0,1692	0,1717	0,1743	0,1756	0,1770
8	0,1765	0,1783	0,1799	0,1813	0,1866	0,1924	0,1955	0,1987	0,2022	0,2040	0,2058
9	0,1956	0,1978	0,1998	0,2015	0,2080	0,2151	0,2190	0,2231	0,2274	0,2297	0,2320
10	0,2125	0,2151	0,2174	0,2194	0,2270	0,2355	0,2401	0,2450	0,2502	0,2530	0,2558
11	0,2274	0,2303	0,2329	0,2352	0,2440	0,2537	0,2591	0,2648	0,2710	0,2742	0,2776
12	0,2407	0,2439	0,2468	0,2494	0,2592	0,2702	0,2763	0,2828	0,2899	0,2936	0,2975
13	0,2525	0,2561	0,2592	0,2621	0,2729	0,2851	0,2919	0,2993	0,3072	0,3115	0,3159
14	0,2631	0,2670	0,2704	0,2735	0,2853	0,2987	0,3062	0,3143	0,3232	0,3279	0,3329
15	0,2728	0,2769	0,2806	0,2839	0,2966	0,3111	0,3193	0,3282	0,3379	0,3431	0,3486
16	0,2815	0,2859	0,2898	0,2933	0,3069	0,3225	0,3313	0,3409	0,3515	0,3572	0,3633
17	0,2894	0,2941	0,2982	0,3020	0,3163	0,3330	0,3424	0,3528	0,3642	0,3704	0,3769
18	0,2967	0,3015	0,3059	0,3099	0,3250	0,3426	0,3527	0,3637	0,3760	0,3826	0,3897
19	0,3033	0,3084	0,3130	0,3171	0,3330	0,3516	0,3622	0,3739	0,3870	0,3941	0,4017
20	0,3095	0,3148	0,3195	0,3238	0,3404	0,3599	0,3711	0,3835	0,3973	0,4049	0,4130
21	0,3152	0,3206	0,3256	0,3300	0,3473	0,3677	0,3794	0,3924	0,4070	0,4151	0,4237
22	0,3204	0,3261	0,3312	0,3358	0,3537	0,3749	0,3871	0,4008	0,4162	0,4246	0,4337
23	0,3253	0,3311	0,3364	0,3412	0,3596	0,3817	0,3944	0,4087	0,4248	0,4337	0,4433
24	0,3299	0,3359	0,3413	0,3462	0,3652	0,3880	0,4012	0,4161	0,4329	0,4423	0,4523
25	0,3341	0,3403	0,3458	0,3509	0,3705	0,3940	0,4077	0,4231	0,4406	0,4504	0,4610
26	0,3381	0,3444	0,3501	0,3552	0,3754	0,3996	0,4137	0,4297	0,4479	0,4581	0,4692
27	0,3418	0,3483	0,3541	0,3594	0,3800	0,4049	0,4195	0,4360	0,4549	0,4655	0,4770
28	0,3453	0,3519	0,3578	0,3632	0,3844	0,4099	0,4249	0,4419	0,4615	0,4725	0,4845
29	0,3486	0,3553	0,3614	0,3669	0,3885	0,4146	0,4300	0,4476	0,4678	0,4791	0,4916
30	0,3517	0,3585	0,3647	0,3703	0,3924	0,4191	0,4349	0,4529	0,4738	0,4855	0,4984
40	0,3753	0,3830	0,3901	0,3966	0,4221	0,4538	0,4730	0,4952	0,5216	0,5369	0,5541
50	0,3902	0,3987	0,4064	0,4134	0,4415	0,4767	0,4984	0,5238	0,5548	0,5731	0,5941
60	0,4006	0,4095	0,4177	0,4251	0,4550	0,4930	0,5165	0,5446	0,5793	0,6002	0,6247
120	0,4280	0,4382	0,4476	0,4563	0,4915	0,5376	0,5673	0,6040	0,6523	0,6839	0,7244
240	0,4426	0,4535	0,4637	0,4730	0,5113	0,5625	0,5961	0,6390	0,6982	0,7399	0,8000
480	0,4501	0,4614	0,4719	0,4817	0,5217	0,5757	0,6117	0,6582	0,7249	0,7745	0,8560
960	0,4539	0,4655	0,4762	0,4861	0,5270	0,5825	0,6197	0,6684	0,7394	0,7944	0,8969
∞	0,4577	0,4695	0,4804	0,4906	0,5324	0,5895	0,6280	0,6789	0,7550	0,8166	1,0000

Tabela 34 – Quantis superiores da distribuição de F ($F_{0,995}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 99,5% de acordo com o seguinte evento: $P(F > F_{0,995}) = 0,995$.

ν_2	ν_1										
	1	2	3	4	5	6	7	8	9	10	11
1	0,0001	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
2	0,0050	0,0050	0,0050	0,0050	0,0050	0,0050	0,0050	0,0050	0,0050	0,0050	0,0050
3	0,0180	0,0201	0,0211	0,0216	0,0220	0,0223	0,0225	0,0227	0,0228	0,0229	0,0230
4	0,0319	0,0380	0,0412	0,0432	0,0445	0,0455	0,0462	0,0468	0,0473	0,0477	0,0480
5	0,0439	0,0546	0,0605	0,0643	0,0669	0,0689	0,0704	0,0716	0,0726	0,0734	0,0741
6	0,0537	0,0688	0,0774	0,0831	0,0872	0,0903	0,0927	0,0946	0,0962	0,0976	0,0987
7	0,0616	0,0806	0,0919	0,0995	0,1050	0,1092	0,1125	0,1152	0,1175	0,1193	0,1209
8	0,0681	0,0906	0,1042	0,1136	0,1205	0,1258	0,1300	0,1334	0,1363	0,1387	0,1408
9	0,0735	0,0989	0,1147	0,1257	0,1338	0,1402	0,1452	0,1494	0,1529	0,1558	0,1584
10	0,0780	0,1061	0,1238	0,1362	0,1455	0,1528	0,1587	0,1635	0,1676	0,1710	0,1740
11	0,0818	0,1122	0,1316	0,1453	0,1557	0,1639	0,1705	0,1760	0,1806	0,1846	0,1880
12	0,0851	0,1175	0,1384	0,1533	0,1647	0,1737	0,1810	0,1871	0,1922	0,1966	0,2005
13	0,0879	0,1222	0,1444	0,1604	0,1727	0,1824	0,1904	0,1970	0,2026	0,2075	0,2117
14	0,0904	0,1262	0,1497	0,1667	0,1798	0,1902	0,1988	0,2059	0,2120	0,2172	0,2218
15	0,0926	0,1299	0,1544	0,1723	0,1861	0,1972	0,2063	0,2139	0,2204	0,2261	0,2310
16	0,0946	0,1331	0,1586	0,1774	0,1919	0,2035	0,2131	0,2212	0,2281	0,2341	0,2393
17	0,0963	0,1360	0,1625	0,1819	0,1971	0,2093	0,2193	0,2278	0,2351	0,2414	0,2469
18	0,0979	0,1386	0,1659	0,1861	0,2018	0,2145	0,2250	0,2339	0,2415	0,2481	0,2539
19	0,0993	0,1410	0,1690	0,1898	0,2061	0,2192	0,2302	0,2394	0,2474	0,2543	0,2603
20	0,1006	0,1431	0,1719	0,1933	0,2100	0,2236	0,2349	0,2445	0,2528	0,2599	0,2663
21	0,1017	0,1451	0,1745	0,1964	0,2136	0,2276	0,2393	0,2492	0,2577	0,2652	0,2718
22	0,1028	0,1469	0,1769	0,1993	0,2170	0,2313	0,2433	0,2536	0,2624	0,2701	0,2768
23	0,1038	0,1486	0,1791	0,2020	0,2201	0,2348	0,2471	0,2576	0,2666	0,2746	0,2816
24	0,1047	0,1501	0,1812	0,2045	0,2229	0,2380	0,2506	0,2613	0,2706	0,2788	0,2860
25	0,1055	0,1516	0,1831	0,2068	0,2256	0,2410	0,2538	0,2648	0,2744	0,2827	0,2901
26	0,1063	0,1529	0,1849	0,2090	0,2281	0,2437	0,2569	0,2681	0,2779	0,2864	0,2940
27	0,1070	0,1541	0,1865	0,2110	0,2304	0,2463	0,2597	0,2712	0,2811	0,2899	0,2976
28	0,1077	0,1553	0,1881	0,2129	0,2326	0,2488	0,2624	0,2741	0,2842	0,2931	0,3010
29	0,1083	0,1564	0,1895	0,2146	0,2346	0,2511	0,2649	0,2768	0,2871	0,2962	0,3042
30	0,1089	0,1574	0,1909	0,2163	0,2365	0,2532	0,2673	0,2793	0,2898	0,2990	0,3072
40	0,1133	0,1648	0,2010	0,2286	0,2509	0,2693	0,2850	0,2985	0,3104	0,3208	0,3302
50	0,1159	0,1694	0,2072	0,2363	0,2598	0,2794	0,2962	0,3107	0,3234	0,3347	0,3449
60	0,1177	0,1726	0,2115	0,2416	0,2660	0,2864	0,3038	0,3190	0,3324	0,3443	0,3550
120	0,1223	0,1805	0,2224	0,2551	0,2818	0,3044	0,3239	0,3410	0,3561	0,3697	0,3819
240	0,1246	0,1846	0,2280	0,2620	0,2901	0,3138	0,3344	0,3525	0,3686	0,3831	0,3962
480	0,1257	0,1867	0,2308	0,2656	0,2943	0,3186	0,3398	0,3584	0,3750	0,3900	0,4036
960	0,1263	0,1877	0,2322	0,2674	0,2964	0,3210	0,3425	0,3614	0,3782	0,3935	0,4073
∞	0,1269	0,1887	0,2337	0,2692	0,2985	0,3235	0,3452	0,3644	0,3815	0,3970	0,4111

Tabela 35 – Quantis superiores da distribuição de F ($F_{0,995}$) com ν_1 graus de liberdade do numerador, e ν_2 graus de liberdade do denominador valor da probabilidade (α) de 99,5% de acordo com o seguinte evento: $P(F > F_{0,995}) = 0,995$.

ν_2	ν_1										
	12	13	14	15	20	30	40	60	120	240	∞
1	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
2	0,0050	0,0050	0,0050	0,0050	0,0050	0,0050	0,0050	0,0050	0,0050	0,0050	0,0050
3	0,0230	0,0231	0,0232	0,0232	0,0234	0,0235	0,0236	0,0237	0,0238	0,0239	0,0239
4	0,0483	0,0485	0,0487	0,0489	0,0496	0,0503	0,0506	0,0510	0,0514	0,0516	0,0517
5	0,0747	0,0752	0,0757	0,0761	0,0775	0,0790	0,0798	0,0806	0,0815	0,0819	0,0823
6	0,0997	0,1005	0,1012	0,1019	0,1043	0,1069	0,1082	0,1096	0,1111	0,1118	0,1126
7	0,1223	0,1235	0,1246	0,1255	0,1290	0,1327	0,1347	0,1368	0,1390	0,1402	0,1413
8	0,1426	0,1441	0,1455	0,1468	0,1513	0,1563	0,1590	0,1619	0,1649	0,1664	0,1681
9	0,1606	0,1625	0,1642	0,1658	0,1715	0,1778	0,1812	0,1848	0,1887	0,1907	0,1928
10	0,1766	0,1789	0,1810	0,1828	0,1896	0,1972	0,2014	0,2058	0,2105	0,2130	0,2156
11	0,1910	0,1936	0,1960	0,1981	0,2060	0,2149	0,2197	0,2250	0,2306	0,2336	0,2367
12	0,2038	0,2068	0,2094	0,2118	0,2208	0,2309	0,2365	0,2425	0,2491	0,2525	0,2562
13	0,2154	0,2187	0,2216	0,2242	0,2342	0,2455	0,2519	0,2587	0,2661	0,2701	0,2742
14	0,2258	0,2294	0,2326	0,2355	0,2464	0,2589	0,2660	0,2736	0,2819	0,2864	0,2910
15	0,2353	0,2392	0,2426	0,2457	0,2576	0,2712	0,2789	0,2873	0,2965	0,3015	0,3067
16	0,2439	0,2481	0,2517	0,2551	0,2678	0,2826	0,2909	0,3001	0,3102	0,3156	0,3214
17	0,2518	0,2562	0,2601	0,2636	0,2772	0,2930	0,3020	0,3119	0,3229	0,3288	0,3351
18	0,2591	0,2637	0,2678	0,2715	0,2859	0,3028	0,3124	0,3230	0,3348	0,3412	0,3480
19	0,2657	0,2706	0,2749	0,2788	0,2939	0,3118	0,3220	0,3333	0,3459	0,3528	0,3602
20	0,2719	0,2769	0,2815	0,2856	0,3014	0,3202	0,3310	0,3429	0,3564	0,3638	0,3717
21	0,2776	0,2828	0,2875	0,2918	0,3083	0,3280	0,3394	0,3520	0,3663	0,3741	0,3826
22	0,2829	0,2883	0,2932	0,2976	0,3148	0,3353	0,3472	0,3605	0,3756	0,3839	0,3929
23	0,2878	0,2934	0,2985	0,3030	0,3209	0,3422	0,3546	0,3686	0,3844	0,3932	0,4026
24	0,2924	0,2982	0,3034	0,3081	0,3265	0,3487	0,3616	0,3762	0,3927	0,4019	0,4119
25	0,2967	0,3026	0,3080	0,3129	0,3319	0,3548	0,3682	0,3833	0,4006	0,4103	0,4208
26	0,3007	0,3068	0,3123	0,3173	0,3369	0,3605	0,3744	0,3901	0,4081	0,4183	0,4292
27	0,3045	0,3107	0,3164	0,3215	0,3416	0,3659	0,3803	0,3966	0,4153	0,4258	0,4373
28	0,3081	0,3144	0,3202	0,3255	0,3460	0,3711	0,3859	0,4027	0,4221	0,4331	0,4450
29	0,3114	0,3179	0,3238	0,3292	0,3502	0,3759	0,3912	0,4085	0,4286	0,4400	0,4525
30	0,3146	0,3212	0,3272	0,3327	0,3542	0,3805	0,3962	0,4141	0,4348	0,4466	0,4596
40	0,3386	0,3463	0,3532	0,3596	0,3848	0,4164	0,4356	0,4579	0,4846	0,5002	0,5177
50	0,3540	0,3623	0,3699	0,3769	0,4048	0,4402	0,4620	0,4878	0,5194	0,5382	0,5598
60	0,3647	0,3735	0,3816	0,3890	0,4189	0,4572	0,4810	0,5096	0,5452	0,5669	0,5922
120	0,3931	0,4033	0,4127	0,4215	0,4570	0,5040	0,5345	0,5725	0,6229	0,6560	0,6988
240	0,4082	0,4193	0,4295	0,4389	0,4779	0,5303	0,5651	0,6097	0,6721	0,7163	0,7805
480	0,4160	0,4275	0,4381	0,4480	0,4888	0,5443	0,5817	0,6303	0,7008	0,7537	0,8416
960	0,4200	0,4317	0,4425	0,4526	0,4944	0,5516	0,5903	0,6412	0,7165	0,7753	0,8863
∞	0,4240	0,4360	0,4470	0,4573	0,5000	0,5590	0,5991	0,6525	0,7333	0,7995	1,0000

Tabela 36 – Quantis superiores da distribuição t de Student (t_α) com ν graus de liberdade e para diferentes valores da probabilidade (α) de acordo com o seguinte evento: $P(t > t_\alpha) = \alpha$.

ν	0,250	0,200	0,150	0,100	0,050	0,025	0,010	0,005	0,001
1	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,599
3	0,765	0,979	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,500	5,408
8	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,768
24	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,646
40	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	3,551
60	0,679	0,848	1,045	1,296	1,671	2,000	2,390	2,660	3,460
120	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	3,373
240	0,676	0,843	1,039	1,285	1,651	1,970	2,342	2,596	3,332
480	0,675	0,842	1,038	1,283	1,648	1,965	2,334	2,586	3,311
960	0,675	0,842	1,037	1,282	1,646	1,962	2,330	2,581	3,301
1920	0,675	0,842	1,037	1,282	1,646	1,961	2,328	2,578	3,296
3840	0,675	0,842	1,037	1,282	1,645	1,961	2,327	2,577	3,293
∞	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,291

SOBRE OS AUTORES

Eric Batista Ferreira



Nascido em Juiz de Fora-MG, em 5 de outubro de 1979, desde cedo se encantou com a Ciência e o Fazer científico. Quando criança, misturava soluções e extratos de folhas e flores no intuito de criar poções mágicas! Com 18 anos se formou em Laticínios pelo Instituto Cândido Tostes. Em 2002, graduou-se em Engenharia Agrônômica pela Universidade Federal de Lavras (UFLA). Em 2004, foi o primeiro estudante do Programa de Pós-graduação em Estatística e Experimentação Agropecuária (PPG-EEA) da UFLA a passar direto para o doutorado por mudança de nível. Em 2007 recebeu o grau de doutor, tendo sido o primeiro pós-graduando do PPG-EEA a fazer doutorado na modalidade *sanduíche*, na Open University (Milton Keynes, Inglaterra). Teve o prazer de ser orientado por muitos pesquisadores notáveis, destacando-se o Prof. Marcelo Silva de Oliveira, e o Dr. John Clifford Gower. Em 2008, concluiu pós-doutorado em Estatística Multivariada sob a supervisão do Prof. Daniel Furtado Ferreira (UFLA). Em 2012, se graduou em Matemática pela Universidade Federal de Alfenas (Unifal-MG), onde é professor e pesquisador desde 2007, até os dias atuais. Em 2013, concluiu pós-doutorado em Sensometria no NOFIMA (Ås, Noruega), sob a supervisão do Dr. Per Lea, e do Dr. Tormod Næs. Em 2020, se gradua em Física pela Unifal-MG. É filho de Rui e Marilene, irmão da Ester, marido da Lucivane e pai da Laura.

Marcelo Silva de Oliveira

Iniciou seus estudos universitários no curso de Engenharia Elétrica da Universidade Federal de Minas Gerais em 1979. Gradou-se em Engenharia Agrícola pela Universidade Federal de Lavras em 1985, obteve o grau de mestre em Estatística pela Universidade Estadual de Campinas em 1991, e o grau de doutor em Engenharia pela Universidade de São Paulo em 2000. Desde 1986 dedicou-se principalmente a formação de pessoas, através do ensino em diversos cursos de graduação e de pós-graduação, e à pesquisa. Aposentou-se em 2018 como Professor Titular de Estatística da Universidade Federal de Lavras, porém, continua sua vida científica dedicando-se, ainda, a algumas de suas atividades científicas anteriores, principalmente à pesquisa e à produção de livros, além de adicionar maior medida a outras atividades "fora da Ciência"(nem tanto...). É casado com Nilma e tem dois filhos, Hiel e Reuel. O Reuel casou-se com a Fabiane, que é mais uma filha, a primeira filha. O Hiel parece estar a caminho de nos trazer a segunda filha...

