

Fundamentos Estatísticos de Ciência dos Dados

voltado para aplicações

Renato Assunção

Copyright © 2017 Renato Assunção

PUBLISHED BY PUBLISHER

BOOK-WEBSITE.COM

Licensed under the Creative Commons Attribution-NonCommercial 3.0 Unported License (the "License"). You may not use this file except in compliance with the License. You may obtain a copy of the License at <http://creativecommons.org/licenses/by-nc/3.0>. Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

First printing, March 2017



Contents

I

Part One

1	Introdução	13
1.1	Cálculo de probabilidades	13
1.2	Estatística: dados, dados, dados ...	17
1.2.1	Usando estatística para distinguir os modelos	19
1.3	Probabilidade e estatística: resumo	21
1.4	Risco de crédito: dados e modelo probabilístico	22
1.4.1	The unreasonable effectiveness of data	23
1.5	Modelos probabilísticos para a análise estatística de dados	24
2	Dados e sua descrição estatística	27
2.1	Dados Estatísticos	27
2.2	Tipologia de variáveis	28
2.3	R, uma linguagem para análise de dados	29
2.4	Dados tabulares em R	29
2.5	Vetores e resumos numéricos	31
2.6	Visualizando dados numéricos	32
2.6.1	Histograma	32
2.6.2	Ramo-e-folhas	38
2.6.3	Boxplot	39
2.6.4	Scatterplot	45
2.6.5	Scatterplot 3-dim	49

2.7	Vetores ou colunas para dados categóricos	51
2.8	Objetos em R	53
2.8.1	Escalares	53
2.8.2	Vetores	54
2.8.3	Matrizes	54
2.8.4	Dataframes	56
2.8.5	Listas	57
2.8.6	Funções	58
2.8.7	Vetorizar sempre que possível	59
2.8.8	Comando <code>apply</code>	60
2.9	Material inicial sobre R	61
2.9.1	Material mais avançado em R	61
2.9.2	Cursos online gratuitos	61
3	Probabilidade Básica	63
3.1	Espaço de probabilidade	63
3.2	O espaço amostral Ω	63
3.2.1	Exemplos de Ω	64
3.3	A σ-álgebra \mathcal{A}	68
3.4	A função de probabilidade \mathbb{P}	69
3.4.1	Consequências	70
3.4.2	Como estabelecer uma função \mathbb{P} ?	71
3.4.3	Visão frequentista	73
3.4.4	\mathbb{P} quando Ω é infinito enumerável	76
3.4.5	\mathbb{P} quando Ω é infinito não-enumerável	76
3.4.6	Quando Ω é complicado	78
3.4.7	Solução a vista	80
3.4.8	Função densidade de probabilidade	80
3.5	Detalhes	81
4	Probabilidade Condicional	85
4.1	Probabilidade Condicional	85
4.1.1	O que é uma probabilidade condicional	85
4.1.2	Probabilidade Condicional e Ciência dos Dados	85
4.1.3	Definindo Probabilidade Condicional	86
4.1.4	Ciência dos dados e condicional	87
4.1.5	Intuição para a definição	87
4.2	Diagramas de Venn	89
4.2.1	Probabilidade condicional no diagrama de Venn	89
4.2.2	$\mathbb{P}(B A)$ e $\mathbb{P}(B)$	90
4.3	Independência de eventos	90
4.3.1	Como surge a independência?	91
4.3.2	Independência no diagrama de Venn	92
4.4	Regra de Bayes	92
4.4.1	Bayes e teste diagnóstico	93
4.4.2	Sensitividade e Especificidade	93
4.4.3	Regra da probabilidade total	95

4.4.4	Extensão da Regra de Bayes	96
4.5	Condisional como nova medida de probabilidade	97
4.6	Independência mútua	98
4.7	Paradoxos com probabilidade condicional	99
5	Introdução a classificação	103
5.1	Introdução	103
5.2	Árvores de classificação	105
5.3	Árvores de classificação no R	110
5.4	Alguns exemplos de árvores de classificação	110
5.4.1	Predicting myocardial infarctions	110
5.4.2	Predictive models for outcome after severe head injury	110
5.4.3	Autism Distinguished from Controls Using Classification Tree Analysis	111
6	Variáveis Aleatórias Discretas	113
6.1	Variáveis aleatórias: formalismo	113
6.2	Variáveis Aletórias e tabelas de dados	115
6.3	Tipos de variáveis aleatórias	115
6.4	Variáveis Aleatórias Discretas	116
6.5	V.A.s discretas: exemplos	116
6.6	A σ-álgebra e a função de probabilidade	118
6.7	Função acumulada	120
6.8	Valor Esperado $\mathbb{E}(X)$	123
6.9	Interpretando $\mathbb{E}(X)$	123
6.10	Principais Distribuições Discretas	124
6.11	Bernoulli	124
6.12	Binomial	125
6.13	Distribuição Multinomial	128
6.14	Distribuição de Poisson	131
6.14.1	A gênese de uma Poisson	134
6.15	Geométrica	139
6.16	Distribuição de Pareto ou Zipf	140
6.16.1	Exemplos de distribuições de Pareto	143
6.16.2	Distribuição de Zipf	143
6.16.3	Como verificar se a distribuição é Pareto?	144
6.17	Comparação entre as distribuições	145
7	Variáveis Aleatórias Contínuas	147
7.1	Introdução	147
7.2	$\mathbb{F}(X)$ no caso contínuo	151
7.3	$\mathbb{E}(X)$ no caso contínuo	153
7.4	Distribuição Uniforme	154

7.5	Distribuição Beta	157
7.6	Distribuição exponencial	162
7.7	Distribuição normal ou gaussiana	165
7.8	Distribuição gama	171
7.9	Distribuição Weibull	174
7.10	Distribuição de Pareto	177
7.10.1	Simulando uma Pareto	180
7.10.2	Ajustando e visualizando uma Pareto	180
8	Independência e Transformações de v.a.'s	183
9	Variância e Desigualdades	185
9.1	Variabilidade e desvio-padrão	185
9.2	Desigualdade de Tchebyshev	192
9.2.1	Opcional: A otimização de Tchebyshev	193
9.2.2	Força e fraqueza de Tchebyshev	194
9.3	Outras desigualdades	194
10	Ajuste de distribuição	195
10.1	Teste qui-quadrado	195
10.2	A estatística Qui-quadrado	198
10.2.1	A distribuição de χ^2	199
10.3	Como usar este resultado de Pearson? O p-valor	200
10.4	Ajustando os graus de liberdade	202
10.5	Teste quando a v.a. é contínua	203
10.6	A função acumulada empírica	205
10.7	Distância de Kolmogorov	207
10.8	Convergência de D_n	209
10.9	Resumo da ópera	210
10.10	Kolmogorov versus Qui-quadrado	210
11	Simulação Monte Carlo	213
11.1	O que é uma simulação Monte Carlo	213
11.2	Geradores de números aleatórios $U(0,1)$	214
11.2.1	Gerador congruencial misto	214
11.3	Simulação de v.a.'s binomiais	215
11.4	Simulação de v.a.'s discretas arbitrárias	216
11.5	Gerando Poisson	218
11.6	Método da transformada inversa	219
11.7	Gerando v.a. com distribuição Gomperz	220
11.8	Gerando v.a. com distribuição de Pareto	221
11.9	Gerando v.a. gaussiana ou normal	224

11.10 Monte Carlo para estimar integrais	224
11.10.1 Integrais com limites genéricos	225
11.11 Método da rejeição	226
11.11.1 Dois teoremas	230
11.11.2 Sobre o impacto de M	230
11.12 História do método Monte Carlo	231
11.13 Aplicação em seguros: valor presente atuarial	233
11.14 Simulando um fundo de pensão	235
11.15 Processo de Poisson: sinistros no tempo	240
11.15.1 Outra abordagem	241
11.15.2 Processo de Poisson não-homogêneo	241
11.16 Provas dos teoremas: opcional	243
12 Vetores Aleatórios	247
12.1 Introdução	247
12.2 Conjunta discreta	250
12.3 Marginal discreta	250
12.4 Independência de duas v.a.'s	251
12.5 Marginal discreta com várias v.a.'s	253
12.6 Simulação de X discreto	256
12.7 Um outro arranjo no caso bi-dimensional	257
12.8 Um longo exemplo: Mobilidade social no Brasil em 1988	258
12.9 Condisional discreta	259
12.10 De volta à mobilidade social	263
12.11 Distribuição condicional de X	264
12.12 Exemplos de distribuições condicionais discretas	265
12.13 Esperança condicional discreta	271
12.14 Variância condicional discreta	272
12.15 Distribuição conjunta contínua	272
12.16 Definição formal de densidade	274
12.16.1 Jointly distributed random variables	277
12.16.2 Geometric probability	279
12.17 Marginal contínua	280
12.18 Condisional contínua	281
12.19 Esperança condicional	284
12.20 Variância condicional	284
12.21 Simulação de um vetor contínuo	285
13 Distribuição Normal Multivariada	289
13.1 Normal bivariada: introdução	289
13.2 O índice ρ de correlação de Pearson	293
13.3 Propriedades de ρ	295

13.4	Matriz de correlação	295
13.5	Propriedades de ρ	299
13.6	Estimando ρ	300
13.7	Distância Estatística de Mahalanobis	300
13.8	Autovetor e autovalor de Σ	305
13.8.1	Formas quadráticas	305
13.8.2	Matrizes positivas definidas	306
13.8.3	Autovetores e autovalores	307
13.8.4	Teorema Espectral	309
13.9	Densidade da normal multivariada	309
14	Análise de Componente Principal e Fatorial	311
14.1	Introdução	311
15	Classificação: Análise Discriminante Linear	313
15.0.1	Bayes' Theorem for random vectors is analogous to Bayes' Theorem for events.	328
15.0.2	Bayes classifiers are optimal.	330
16	Teoremas Limite	337
16.1	Introdução	337
16.2	Convergence	337
16.3	Lei dos Grandes Números	338
16.3.1	Uma forma mais geral	344
16.3.2	A versão forte	344
16.4	Moment generating functions	345
16.4.1	Caso Normal	346
16.5	Teorema Central do Limite	347
16.5.1	Historical remarks	351
16.5.2	Accuracy of the Approximation by the Central Limit Theorem	351
16.6	A concentration inequality: Hoeffding's inequality	353
17	Esperança Condicional e Aproximação	355
17.1	Introdução	355
17.1.1	The conditional expectation $\mathbb{E}(Y X = x)$ is called the regression of Y on X	355
17.2	Optimal Prediction of $(Y X = x)$	359
17.2.1	Inspection paradox	360
17.2.2	Inspection paradox	361
II	Part Two	
18	Régressão Linear	365
18.1	Introdução	365

19	Regressão Logística	367
19.1	Introdução	367
20	Regularização e Fatores Latentes	369
20.1	Introdução	369
21	Estimador de Máxima Verossimilhança	371
21.1	Motivação	371
21.2	Estimando o tempo médio de sobrevida	372
21.3	Quais valores do parâmetro são verossímeis?	374
21.4	Resumo informal da máxima verossimilhança	378
21.5	Por quê a máxima verossimilhança?	378
21.6	Modelos paramétricos	379
21.7	Estimador de máxima verossimilhança	380
21.8	Obtendo o EMV	383
21.9	EMV: soluções analíticas	384
21.9.1	Verossimilhança Bernoulli	384
21.9.2	Verossimilhança Binomial	386
21.9.3	Verossimilhança Poisson	387
21.9.4	Verossimilhança Poisson truncada	387
22	GLM	391
22.1	Introdução	391
23	Teoria de Estimação	393
23.1	Outros métodos de estimação	393
23.1.1	Método de Momentos	393
23.1.2	Um método sem nome	395
23.1.3	Mais um método sem nome	396
23.2	Estimadores são variáveis aleatórias	398
23.3	Comparando dois estimadores	399
23.4	Estimação Pontual	401
24	Algoritmo EM	403
24.1	Misturas de distribuições: introdução	403
24.2	Misturas de distribuições: formalismo	405
24.2.1	Misturas de normais multivariadas	407
24.3	Estimando uma distribuição de mistura	409
24.4	Dados e rótulos	409
24.4.1	A verossimilhança completa	411
24.4.2	A verossimilhança incompleta	412

24.5 O algoritmo EM	412
24.5.1 A distribuição de $\mathbf{Z} \mathbf{Y} = \mathbf{y}$	412
24.5.2 Escolhendo $\theta^{(0)}$	413
24.5.3 $\mathbb{E}(Z_i \mathbf{Y} = \mathbf{y}, \theta^{(0)})$	413
24.5.4 Os passos do algoritmo EM	414
24.5.5 Resumo do algoritmo EM	415
24.6 Exemplos de uso do algoritmo EM	415
24.7 Convergência d algoritmo EM	415
24.7.1 Definições preliminares	415
24.7.2 Desigualdade de Jensen	416
25 Testes de Hipótese	419
25.1 Introdução	419
26 Seleção de Modelos	421
26.1 Introdução	421
26.1.1 The dependence of two random vectors may be quantified by mutual information.	
421	
bibliography	427
Books	427
Articles	427

11.8	Gerando v.a. com distribuição de Pareto
11.9	Gerando v.a. gaussiana ou normal
11.10	Monte Carlo para estimar integrais
11.11	Método da rejeição
11.12	História do método Monte Carlo
11.13	Aplicação em seguros: valor presente atuarial
11.14	Simulando um fundo de pensão
11.15	Processo de Poisson: sinistros no tempo
11.16	Provas dos teoremas: opcional

Part One

12 Vetores Aleatórios 247

12.1	Introdução
12.2	Conjunta discreta
12.3	Marginal discreta
12.4	Independência de duas v.a.'s
12.5	Marginal discreta com várias v.a.'s
12.6	Simulação de \mathbf{X} discreto
12.7	Um outro arranjo no caso bi-dimensional
12.8	Um longo exemplo: Mobilidade social no Brasil em 1988
12.9	Condisional discreta
12.10	De volta à mobilidade social
12.11	Distribuição condicional de X
12.12	Exemplos de distribuições condicionais discretas
12.13	Esperança condicional discreta
12.14	Variância condicional discreta
12.15	Distribuição conjunta contínua
12.16	Definição formal de densidade
12.17	Marginal contínua
12.18	Condisional contínua
12.19	Esperança condicional
12.20	Variância condicional
12.21	Simulação de um vetor contínuo

13 Distribuição Normal Multivariada 289

13.1	Normal bivariada: introdução
13.2	O índice ρ de correlação de Pearson
13.3	Propriedades de ρ
13.4	Matriz de correlação
13.5	Propriedades de ρ
13.6	Estimando ρ
13.7	Distância Estatística de Mahalanobis
13.8	Autovetor e autovalor de Σ
13.9	Densidade da normal multivariada

14 Análise de Componente Principal e Fatorial 311

14.1	Introdução
------	------------

15 Classificação: Análise Discriminante Linear 313

16 Teoremas Limite 337

16.1	Introdução
16.2	Convergence
16.3	Lei dos Grandes Números
16.4	Moment generating functions
16.5	Teorema Central do Limite
16.6	A concentration inequality: Hoeffding's inequality

17 Esperança Condicional e Aproximação 355

17.1	Introdução
17.2	Optimal Prediction of $(Y X = x)$



1. Introdução

1.1 Cálculo de probabilidades

Vamos começar entendendo a diferença entre probabilidade e estatística. Probabilidade é um ramo da matemática pura. Ela permite fazer cálculos matemáticos sobre fenômenos aleatórios. Não precisa de dados estatísticos coletados no mundo real. Probabilidade é uma atividade teórica, uma teoria matemática que não exige dados empíricos.

Podemos resumir o funcionamento desta atividade teórica do probabilista da seguinte forma. Primeiro, estabeleça um modelo probabilístico que descreva o fenômeno de interesse. A seguir, calcule matematicamente a probabilidade de eventos de interesse.

■ **Example 1.1 — Probabilidade: a sequência mais longa de caras.** Uma moeda é jogada repetidamente para cima e o resultado, Cara (K) ou Coroa (C), é anotado. Qual a chance de observarmos uma sequência de pelo menos 8 caras sucessivas em algum momento ao longo de 100 lançamentos da moeda? Este é evento tão raro que deveríamos ficar surpresos caso ele ocorra? Ou, pelo contrário, a chance de termos 8 caras em seguida é grande de modo que, se observarmos 8 caras sucessivas em algum momento ao longo dos 100 lançamentos, isto seria considerado um acontecimento sem novidade.

No cálculo de probabilidades não precisamos jogar a moeda equilibrada nem mesmo uma única vez para obtermos a chance de que tenhamos uma sequência ininterrupta de 8 ou mais caras ao longo de 200 lançamentos. Este cálculo é feito matematicamente, a partir de regras simples de manipulação de probabilidades de eventos elementares.

Quando o número de lançamentos de uma moeda equilibrada é pequeno, este cálculo matemático é obtido a partir da listagem todas as configurações possíveis. Por exemplo, imagine que queremos saber a chance de observar uma sequência de pelo menos três caras sucessivas ao longo de 4 lançamentos de uma moeda equilibrada. A lista de todas as possibilidades de resultados para os quatro lançamentos é composta por 16 sequências:

KKKK	KKKC	KKCK	KCKK
CKKK	KKCC	KCKC	CKKC
KCCC	CKCK	CCKK	KCCC
CKCC	CCKC	CCCC	CCCC

Existem 5 sequências com três ou quatro caras sucessivas: KKK, KKC, KCK, CKC, CKK . Como vamos aprender no próximo capítulo, o cálculo de probabilidades nos leva a concluir que probabilidade de observarmos 3 ou mais caras sucessivas numa sequência de 4 lançamentos é igual a $5/16 \approx 0.31$, ou 31% de chance, um valor considerável.

Infelizmente o cálculo não pode ser feito desta forma simples quando o número de lançamentos da moeda começa a crescer. Com 100 lançamentos, o número de configurações possíveis é igual a 2^{100} e encontrar neste conjunto aquelas configurações com 8 ou mais caras sucessivas é muito difícil.

Existem técnicas especiais para fazer esse cálculo. Veremos no próximo capítulo como ele é feito mas agora importa apenas dizer que ele é feito com papel e lápis (ou um computador), exatamente como uma operação matemática de adição, multiplicação e divisão é efetuada. Trata-se de um cálculo mental, sem necessidade de realizar o experimento físico de jogar a moeda e colecionar os resultados dos lançamentos sucessivos.

Qual a utilidade deste cálculo com moedas? Podemos usá-lo para verificar se a hipótese do *hot hand* em esportes é plausível. O *hot hand* representa a crença mantida por torcedores e profissionais do esporte de que durante um jogo alguns jogadores ou o time inteiro ficam quentes, conseguindo momentaneamente um maior nível de concentração, destreza e habilidade a ponto de conseguirem fazer vários pontos em sequência numa partida. Por exemplo, suponha que um time de basquete ou de vôlei joga contra outro e ambos possuem habilidade compatíveis. Isto significa que, a cada novo ponto, a chance de que ele venha para o time A é igual a 1/2. Assim, a sequência de pontos de um jogo indexada pelo rótulo do time que ficou com aquele ponto seria similar ao resultado de lançar uma moeda honesta sucessivamente. Na verdade, queremos checar se esta similaridade é válida ou se, pelo contrário, a “moeda” tem uma probabilidade de sucesso que flutua em torno de 1/2. Esta flutuação ao longo do jogo seria tal que durante certos períodos prolongados ela é muito maior que 1/2 (momento em que o time entra na fase *hot hand*) e períodos em que fica abaixo de 1/2. Se isto for verdade, deveria haver na sequência de pontos períodos prolongados em que vemos muitas “caras” sucessivas, muito mais do que o que esperamos se for realmente o mecanismo da moeda em funcionamento.

Entretanto, vários estudos tem mostrado que isto não passa de uma ilusão, que os padrões de pontos sucessivos que parecem excepcionais podem acontecer com probabilidade razoável se o cálculo for feito corretamente. Para entender isto completamente, precisamos falar da coleta de dados estatísticos, assunto da seção 1.2.

■ **Example 1.2 — Probabilidade: modelo espacial** 1. Imagine que n pontos são jogados completamente ao acaso num quadrado de área 1 centrado na origem $(0,0)$. Veja três realizações independentes deste experimento na Figura 1.2. Queremos saber a probabilidade de que não exista nenhum ponto num raio r em torno da origem $(0,0)$. Vamos denotar esta probabilidade por $\mathbb{P}_1(r)$.

Toda probabilidade é um número entre 0 e 1. Eventos pouco prováveis possuem probabilidade próxima de zero, enquanto eventos muito prováveis possuem probabilidade próxima de 1. A probabilidade $\mathbb{P}_1(r)$ depende de alguns aspectos do problema. Por exemplo, é natural esperar que $\mathbb{P}_1(r)$ dependa do número n de pontos jogados ao acaso no quadrado de área 1. Se n for muito grande, será difícil que nem ao menos um dos muitos pontos jogados não acabe caindo dentro do círculo r . Por outro lado, se n for muito pequeno, será relativamente fácil que o disco de raio r termine vazio.

Outro aspecto que impacta o valor da probabilidade $\mathbb{P}_1(r)$ é o raio r . Vamos manter o número n de pontos fixo em algum valor. Se r for bastante pequeno (isto é, $r \approx 0$), será difícil que os pontos aleatórios caiam dentro do círculo. Assim, a probabilidade $\mathbb{P}_1(r)$ de que não exista nenhum ponto num raio r em torno da origem $(0,0)$ deve ser um valor próximo de 1, uma probabilidade alta. Quando r aumenta, a probabilidade $\mathbb{P}_1(r)$ deve decrescer para zero. Através de cálculos

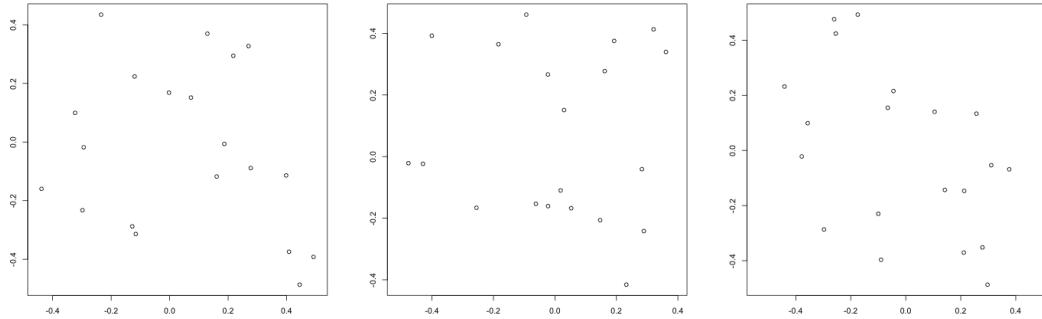


Figure 1.1: Três experimentos independentes de jogar n pontos ao acaso num quadrado

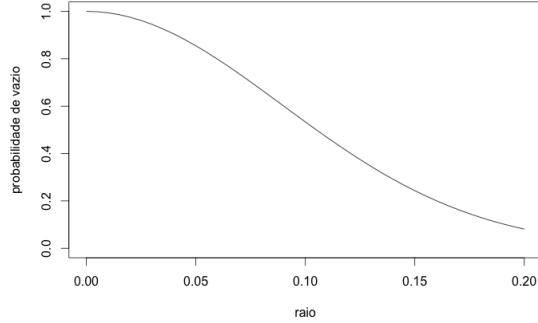


Figure 1.2: Gráfico $f(r) = \exp(-n\pi r^2)$, que é aproximadamente $\mathbb{P}_1(r)$, a probabilidade de que todos os n pontos aleatórios num quadrado unitário estejam pelo menos a uma distância r da origem $(0,0)$

matemáticos rigorosos, pode-se mostrar que $\mathbb{P}_1(r)$ é aproximadamente igual a $\exp(-n\pi r^2)$. A Figura 1.2 mostra o gráfico desta função com $n = 20$ pontos.

A probabilidade $\mathbb{P}_1(r) \approx \exp(-n\pi r^2)$ é obtida *sem dados estatísticos*, nem simulação computacional. É um cálculo puramente matemático. A Figura 1.3 com as três configurações de pontos aleatórios é apenas ilustrativa. Ela não foi usada no cálculo de $\mathbb{P}_1(r)$. Para este modelo de pontos aleatórios, várias outras probabilidades podem ser calculadas. Por exemplo, qual a probabilidade de que existam pelo menos 2 pontos numa certa região de área α ? É aproximadamente igual a

$$1 - e^{-n\alpha} (1 + n\alpha)$$

Não é preciso coletar nenhum dado estatístico para fazer este cálculo.

■ **Example 1.3 — Probabilidade: modelo espacial 2.** Outro modelo probabilístico para a geração de pontos no quadrado leva a resultados bem diferentes no cálculo das probabilidades. Por exemplo, suponha que apenas 5 pontos-pais são jogados completamente ao acaso no quadrado de área 1 centrado na origem $(0,0)$. A seguir, cada ponto-pai gera 4 pontos-filhos de forma que temos 20 pontos-filhos no final. Os filhos espalham-se ao acaso em torno dos pais até uma distância máxima de 0.1. Considere o padrão espacial dos pontos compostos apenas pelos filhos. Veja três realizações independentes deste novo experimento na Figura 1.3.

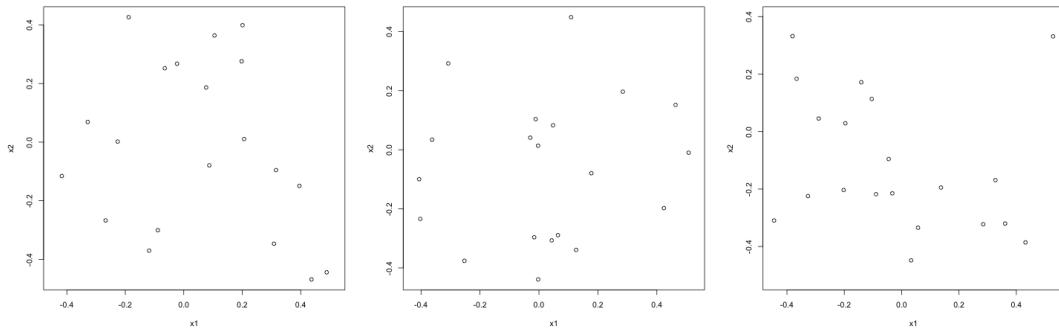


Figure 1.3: Três experimentos independentes do segundo modelo de jogar n pontos ao acaso num quadrado

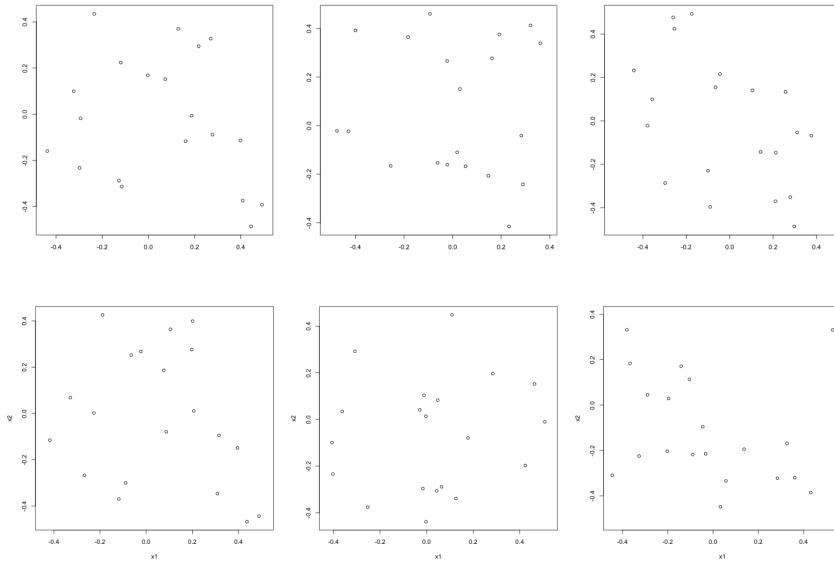


Figure 1.4: Contrastando realizações do primeiro modelo espacial (linha superior) com o segundo modelo espacial (linha inferior)

A Figura 1.4 contrasta três realizações aleatórias do primeiro modelo espacial (linha superior) com outras três realizações do segundo modelo espacial (linha inferior). Veja que não existem diferenças muito óbvias entre estes plots.

No caso do modelo 2, podemos calcular também a probabilidade \mathbb{P}_2 do mesmo evento anterior, de que não exista nenhum ponto num raio r em torno da origem $(0, 0)$. Como no modelo 1, temos $\mathbb{P}_2 \approx 1$ e diminuindo para zero quando o raio r aumenta. Mas esta probabilidade decai de forma bem diferente nos dois modelos, como pode ser visto na Figura 1.3. O modelo 2 tem um decaimento muito mais lento da sua probabilidade do que o modelo 1. Com um raio $r = 0.15$, a probabilidade de não haver ponto dentro do disco é aproximadamente 0.20 no caso do modelo 1 mas aproximadamente igual a 0.70 no caso do modelo 2.

Para explicar este grande diferença, veja que os pontos-filhos aleatórios do modelo 2 são fortemente influenciados pelas posições dos poucos (apenas cinco) pontos-pais. Se estes cinco pontos-pais cairem longe do centro $(0, 0)$, será grande a chance de não haver pontos-filhos dentro do disco em torno da origem. Já no modelo 1, todos os n pontos são jogados de forma independente

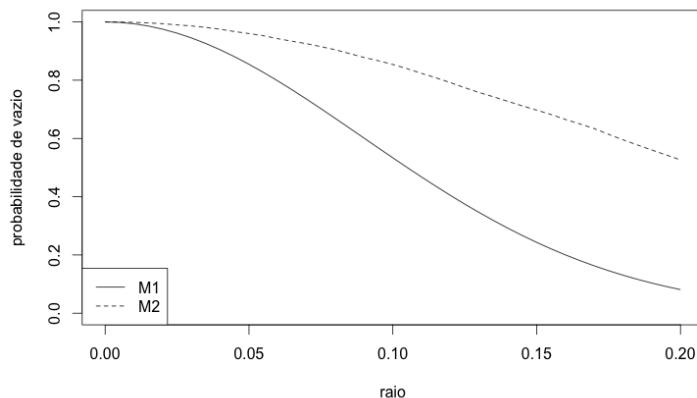


Figure 1.5: Probabilidade de que não exista nenhum ponto num raio r em torno da origem $(0,0)$ pelo modelo 1 (M1) e pelo modelo 2 (M2).

uns dos outros e assim será pequena a chance de que todos eles se afastem ao mesmo tempo do centro. ■

1.2 Estatística: dados, dados, dados ...

“Data!data!data!” he cried impatiently. “I can’t make bricks without clay.”

— Arthur Conan Doyle, *The Adventure of the Copper Beeches*

Em contraste com probabilidade, em estatística estamos sempre lidando com dados obtidos experimentalmente. Estatística é um ramo da matemática aplicada. Ela precisa de dados estatísticos, coletados através de processos de amostragem. O que fazemos com esses dados? Nós procuramos inferir qual foi o modelo probabilístico que gerou os dados observados. Assim, em probabilidade nós estabelecemos um modelo probabilístico e nos perguntamos pelas probabilidades de vários eventos. Em estatística, a natureza ou mundo gera dados aleatórios e o objetivo é fazer inferência sobre o modelo que gerou estes dados. Esta inferência usa o cálculo de probabilidade como uma ferramenta para auxiliar na identificação do modelo gerador dos dados estatísticos. Após identificar o modelo probabilístico que aparentemente gerou os dados observados, desejamos usar este modelo para calcular certas probabilidades. O interesse costuma estar concentrado em calcular a chance da ocorrência de eventos que são possíveis mas que ainda não foram observados. Outro tipo de evento de interesse para os quais gostaríamos de obter uma estimativa precisa são aqueles eventos que não são muito frequentes mas que podem carregar um risco substancial.

■ **Example 1.4 — De volta às sequências longas.** Num jogo de basquete, a soma do número de pontos de cada time é da ordem de 150. Se imaginarmos que uma cesta do time A representa *cara* e uma cesta do time B representa *coroa*, podemos conceber a sequência de pontos do jogo como o resultado dos lançamentos sucessivos de uma moeda. Assim, a probabilidade de *cara* é a probabilidade de que uma cesta ocorrida durante o jogo seja do time A . Como o time A pode ser melhor que o time B devemos imaginar uma moeda não balanceada, em que a probabilidade de *cara* possa ser maior que a probabilidade de *coroa*. Por exemplo, se o time A faz 100 das 150 cestas de um jogo, podemos supor que a probabilidade de *cara* seja duas vezes maior que a probabilidade de *coroa*.

Sequence #1

```
T H H H H T T T H H H H T H H H H H H H T T H H T T H H H H H T T T H H T H H H T
T T H T T H H H H T H T T H T T H H T T T H H H H H T T H H T T H H T H H H H T T T
T H T T T H H T T H T T H H T T T H H T H H T H H T T T H H T H H H H H T H H T T
H T H T T H H H H T H H T H H H H H T H H T T H H T H H H H H T H H T T H H H H T H H
```

Sequence #2

```
T H T H T T T H T T T H T H T T H H H T H H T H T H T H T T T H H T T H H T T H H H T
H H H T T H H H T T H H H T H H H T T H H H T H T H H H T H H H T H H H T H H H T
H H T H H H H T H T H H T H H H T T H T H H H T H H H T T H H H T H H H T H H H H T H
T H H T T T T H T H T H T H H T T H T T H H H H T H H H T H H H H T H H H H T H H H H T H
```

Figure 1.6: Duas sequências de 200 lançamentos de uma moeda equilibrada. Uma delas tem uma probabilidade que se altera ao longo dos lançamentos. Qual delas?

O modelo da moeda não-balanceada para representar a sequência de cestas tem várias consequências. Se este modelo é uma boa representação, devemos então concluir que os pontos aparecem como resultados de um cara ou coroa. Isto significa que, como a moeda não possui memória das jogadas anteriores, o resultado da próxima cesta não depende do que aconteceu antes. Assim, a probabilidade da próxima cesta é a mesma desde o início, não muda em decorrência dos resultados anteriores. Mais do que a influência dos resultados anteriores, a probabilidade fica estática ao importa o que aconteça. Isto é, não existe nenhum mecanismo fazendo a probabilidade de *cara* variar ao longo das jogadas.

Se a hipótese do *hot hand* é verdadeira então este modelo da moeda não é válido para representar os resultados de um jogo. Isto é, com o *hot hand* operando a ocorrência de cestas do time *A* não são como os resultados *cara* em lançamentos sucessivo de uma moeda não-balanceada. A hipótese do *hot hand* leva a pensar em outro modelo para representar o jogo, um modelo em que a probabilidade de *cara* oscilasse durante o jogo, não permanecendo constante. Ela poderia oscilar em decorrência dos resultados anteriores do jogo ou devido a fatores externos ao jogo. De qualquer modo, ela não permaneceria constante, destruindo a proximidade com o modelo da moeda.

Qual é o correto? Não é tão fácil de adivinhar como você poderia pensar inicialmente. As duas sequências mostradas na Figura 1.4 representam o resultado de 200 lançamentos de uma moeda bem equilibrada, com a chance de *cara* (ou heads, *H*) sendo igual a de *coroa* (ou tails, *T*). Entretanto, enquanto uma delas realmente representa a moeda equilibrada, a outra foi gerada por um mecanismo em que a probabilidade do próximo lançamento muda em função do número de caras nos lançamentos prévios. Tente adivinhar rapidamente qual é esta sequência “falsa”.

Este exemplo é inspirado em Schilling (ref ??). Ele explica que Révész, um probabilista, costumava dividir seus alunos em dois blocos. Cada estudante do bloco 1 deveria jogar uma moeda para cima 200 vezes e anotar a sequência obtida. O segundo bloco deveria tentar gerar, de sua própria cabeça, uma sequência que se parecesse, da melhor forma possível, com a que seria obtida jogando-se uma moeda. As sequências dos dois grupos eram embaralhadas e uma chave com a identificação do grupo era guardada. Ele então observava as sequências geradas procurando verificar se elas se conformavam com o que seria tipicamente observado com o lançamento de moedas equilibradas. Quase sempre ele conseguia identificar corretamente qual grupo havia gerado cada uma das sequências.

A sequência “falsa”, aquela que não foi obtida com uma moeda equilibrada sendo lançada sucessivamente, é a segunda. Como isto poderia ser identificado? Vamos ver isto mais tarde neste curso.

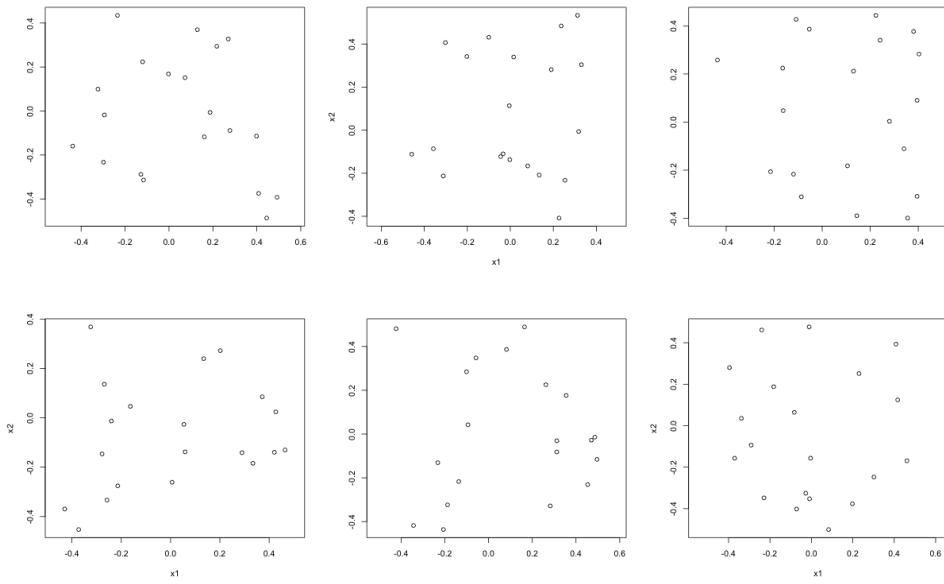


Figure 1.7: Seis realizações aleatórias, algumas obtidas com o primeiro modelo espacial e algumas com segundo modelo espacial. Você consegue identificar qual é qual?

■ **Example 1.5 — De volta aos modelos espaciais.** Na Figura 1.5, temos seis realizações aleatórias de pontos jogados no quadrado unitário. Algumas delas foram obtidas com o primeiro modelo espacial descrito anteriormente enquanto as demais foram obtidas com o segundo modelo espacial. Qual dos dois modelos gerou cada um dos seis plots nesta figura?

Esta não é uma tarefa simples pois as diferenças entre os padrões pontuais gerados pelos dois modelos não são muito grandes. Como cada figura foi gerada por mim, eu sei qual modelo foi usado em cada caso. A Figura 1.5 revela qual é o modelo por trás de cada um dos plots da Figura 1.5.

■

1.2.1 Usando estatística para distinguir os modelos

O objetivo desta seção é ilustrar a diferença entre o cálculo matemático de probabilidades e a análise de dados estatísticos. Vamos usar um teste estatístico para discriminar entre os dois modelos por trás de cada plot da Figura 1.5.

Para cada ponto aleatório, achei a distância até o seu ponto vizinho mais próximo. Fixando um raio r , contei a proporção de pontos de um plot que tiveram distância menor que r . Por exemplo, para $r = 0.10$, obtive a proporção G de pontos observados que tiveram seu vizinho mais próximo a uma distância menor que 0.10. Considerando vários raios r distintos em cada plot, obtive a proporção de pontos que tiveram distância menor que r .

Então, por meio do cálculo de probabilidades, sem usar dados estatísticos, com matemática pura, obtive limites (m, M) tais que, se os dados vierem de fato do modelo 1, o valor da proporção G deveria estar entre m e M com probabilidade muito alta. Se estiver fora dos limites, o modelo 2 deve ser o correto.

A Figura 1.9 mostra como o teste estatístico é realizado. No eixo horizontal temos o raio r . O eixo vertical representa os valores da probabilidade da distância ao vizinho mais próximo ser menor que r . As duas linhas tracejadas foram obtidas com cálculos probabilísticos, puramente matemáticos. Não vamos nos preocupar em descrever estes cálculos neste instante. As linhas tracejadas representam limites para a curva empírica $G(r)$, representada pela linha sólida. Esta curva

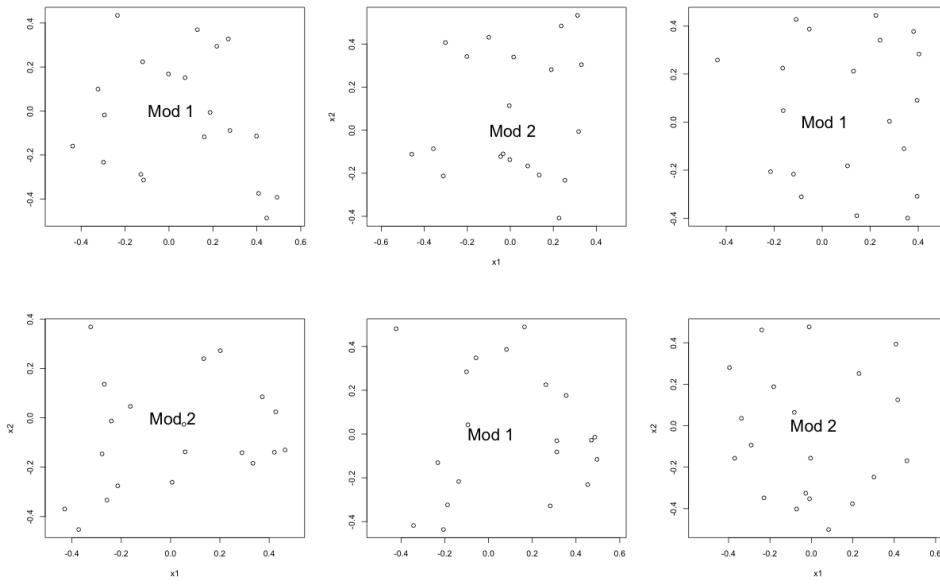


Figure 1.8: Identificando modelo por trás de cada plot da Figura 1.5.

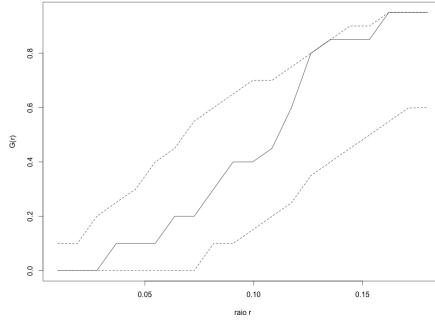


Figure 1.9: Curva empírica $G(r)$ (linha sólida) e seus limites sob o modelo 1 (linhas tracejadas) versus o raio r .

empírica foi obtida calculando a proporção de G de pontos observados que tiveram seu vizinho mais próximo a uma distância menor que r . Fiz estes cálculo usando alguns valores $r_1 < \dots < r_k$ de r obtendo as proporções $G(r_1) < \dots < G(r_k)$. A curva empírica $G(r)$ é obtida conectando-se com segmentos de reta os pontos $(r_j, G(r_j))$.

Na Figura 1.9 vemos a curva empírica $G(r)$ dentro dos limites determinados pelas linhas tracejadas. O teste estatístico recomenda então inferir que o Modelo 1 foi usado para gerar os dados espaciais correspondentes. O raciocínio é que, caso o modelo 1 seja o correto, a curva empírica $G(r)$ deveria estar entre as linhas tracejadas com alta probabilidade. Por outro lado, caso o modelo 2 seja o modelo que esteja gerando os dados, a curva empírica $G(r)$ tenderia a sair dos limites tracejados. O teste pode levar a erros: o modelo 1 pode ser o correto e ainda assim a curva empírica $G(r)$ sair dos seus limites. O contrário também pode ocorrer: o modelo 2 é o correto mas a curva empírica $G(r)$ fica dentro dos limites tracejados. Entretanto, as linhas tracejadas são calculadas para que isto não ocorra com muita frequência. Isto é, as linhas tracejadas são obtidas de tal forma que, com probabilidade muito alta, a curva empírica $G(r)$ fica dentro deses limite caso o modelo 1 seja o correto.

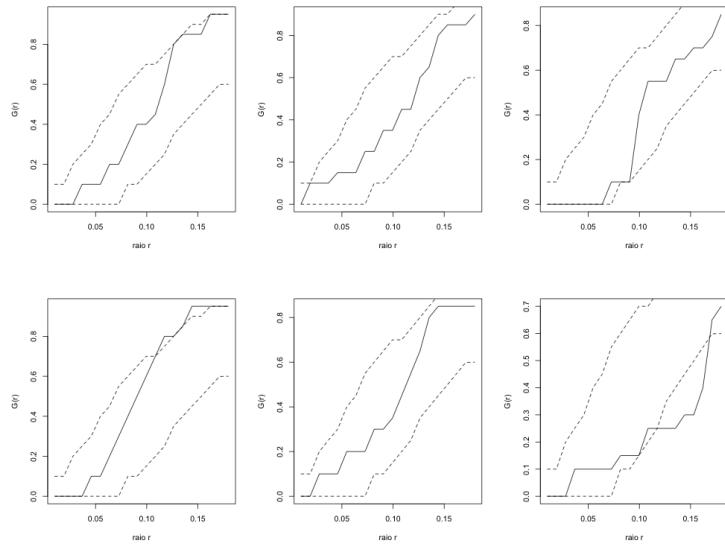


Figure 1.10: O resultado de aplicar um teste estatístico a cada um dos plots da Figura 1.5.

Vamos enfatizar uma vez mais: as linhas tracejadas foram obtidas supondo que o modelo 1 seja o correto e isto foi feito por meio do cálculo de probabilidades, *sem dados*. Já a curva contínua $G(r)$ representa um cálculo empírico, baseado nos dados experimentais. O valor de $G(r)$ é uma proporção calculada com os dados estatísticos, a proporção de pontos cuja distância ao ponto vizinho mais próximo é menor que r . Se a curva contínua $G(r)$ calculada com os dados ficar dentro dos limites tracejados obtidos pela teoria do modelo 1, nós então apostamos que o modelo 1 gerou os dados espaciais observados. Se sair fora dos limites, apostamos no modelo 2 para ser o gerador dos dados.

A Figura 1.2.1 mostra o resultado de aplicar este teste estatístico a cada um dos plots da Figura 1.5. Lembre-se que identidade do modelo probabilístico que realmente gerou este padrões espaciais foi revelada na Figura 1.5. Considerando a decisão recomendada pelo teste estatístico na Figura 1.2.1, vemos que tomaríamos uma decisão errada apenas no plot (1,2), cujos dados são gerados pelo modelo 2.

1.3 Probabilidade e estatística: resumo

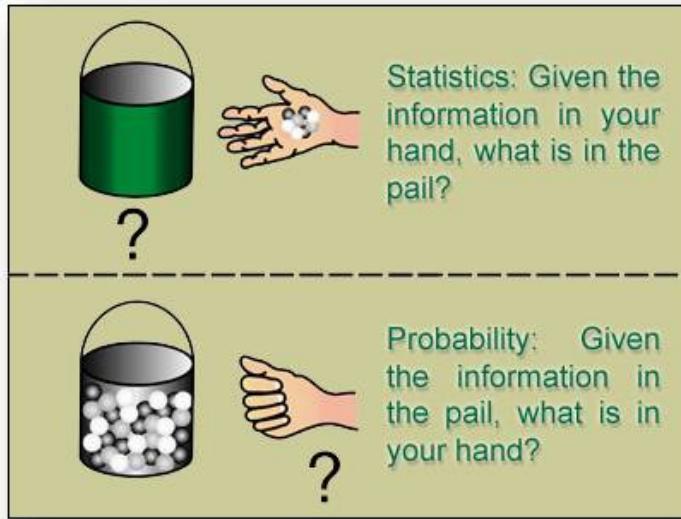
Probabilidade: a partir de um modelo matemático de um fenômeno que gera resultados não-determinísticos, o cálculo de probabilidades permite deduzir matematicamente a probabilidade dacorrência de diversos eventos. Não é preciso ter nenhum dado estatístico para isto.

Estatística: de posse de dados coletados no mundo real, construimos uma tabela de números. Deseja-se descobrir qual foi o modelo probabilístico que gerou estes dados. Identificado o modelo, deseja-se usá-lo para calcular a probabilidade da ocorrência de eventos que possíveis que ainda não foram observados ou que não são muito frequentes.

A imagem na Figura 1.3 mostra as diferenças entre estatística e probabilidade. Esta imagem foi extraída de http://herdingcats.typepad.com/my_weblog/.

Embora diferentes, estatística e probabilidades alimentam-se mutuamente. Uma não poderia existir saudavelmente sem a outra. Os problemas do mundo real para os quais queremos calcular probabilidades inspiram o desenvolvimento de teorias e conceitos probabilísticos, bem como a obtenção de teoremas matemáticos sobre estas probabilidades, muitas vezes contradizendo a nossa intuição. Por outro lado, a capacidade de criar modelos e fazer cálculos probabilísticos cada vez

Figure 1.11: Diferença entre estatística e probabilidade.



mais complexos, nos leva a coletar dados com estruturas sofisticadas e então ser capaz de inferir quais desses modelos estão operando na realidade. Na próxima seção apresentamos um exemplo consagrado do uso de dados para estabelecer um modelo sofisticado de predição para a concessão de crédito financeiro.

1.4 Risco de crédito: dados e modelo probabilístico

Clientes solicitam crédito ou tomam empréstimo com agentes financeiros. Esses agentes querem saber, para cada cliente, se ele vai pagar de volta dentro do prazo o empréstimo feito. Um modelo de *risco de crédito* avalia a probabilidade disso ocorrer dado que o cliente possui certos atributos. Se a probabilidade for baixa, ele é um risco potencial e o crédito deveria ser negado. Se a probabilidade for alta, o crédito deveria ser concedido.

De fato, esta probabilidade é altamente personalizada e deve depender de vários aspectos ligados ao cliente e ao ambiente de negócios. Por exemplo, a probabilidade de pagar de volta dentro do prazo o empréstimo feito deve depender do saldo médio da conta do cliente em relação ao valor do empréstimo. Se o cliente deseja um empréstimo que representa 25% do que ele em geral possui de saldo na sua conta ou se ele deseja um empréstimo que é 10 vezes maior que seu saldo médio, o risco de não pagamento parece ser maior no segundo caso.

Muito outros fatores devem afetar o cálculo dessa probabilidade. Há quanto anos o tomador de empréstimo é cliente da instituição? Qual sua história pregressa em termos de empréstimos e seus pagamentos? O ambiente econômico é de crescimento e portanto favorável a novos investimentos ou é um ambiente de recessão?

Precisamos de um modelo de probabilidade para fazer estes cálculos. Existem muitos modelos possíveis sendo usado atualmente pelas instituições financeiras. Alguns são melhores que outros pois conseguem prever melhor o que os clientes vão fazer no futuro.

Quais os dados necessários para identificar um modelo de probabilidade desses? Busca-se dentre os clientes recentes do banco uma amostra estatística dentre aqueles que pegaram algum empréstimo. Para cada um desses clientes, anota-se uma resposta binária Y :

- $Y = 1$ se o cliente pagou de volta no devido tempo.

- $Y = 0$ case contrário.

Além disso, temos um conjunto de atributos que podem influenciar o comportamento desses clientes. Para cada cliente na amostra, anota-se as seguintes características que potencialmente podem afetar a probabilidade de pagar de volta dentro do prazo o empréstimo tomado:

- Balance of current account
- For how long has been a client (in months)
- Payment of previous credits: *no previous credits/paid back all previous credits; hesitant payment of previous credits; problematic running account.*
- Purpose of credit: *new car; used car; items of furniture; vacation; etc.*
- Amount of credit.
- Value of savings or stocks.
- For how has been employed by current employer (in years).
- Installment in % of available income
- Marital Status, Sex, Age, etc.

1.4.1 The unreasonable effectiveness of data

Precisamos mesmo de um modelo probabilístico? Nos dias de big data, os dados não respondem tudo? Afinal, podemos fazer cálculos diretos e simples a partir dos dados diretamente.

Por exemplo, qual a probabilidade de um cliente com mais de 60 anos e saldo médio maior que 5 mil reais não pagar o crédito? Separe a sub-amostra de clientes com mais de 60 anos e saldo maior que 5 mil. Se esta sub-amostra não for muito pequena ... (digamos, maior que 1000 indivíduos) ... Dentre os indivíduos dessa sub-amostra, obtenha a proporção dos que não pagaram o crédito. Esta proporção é aproximadamente a probabilidade de não-pagamento. Muito simples, apenas contagem no banco de dados.

Nem sempre é tão simples. O cliente tem muitos atributos, não apenas idade e saldo médio. Para cada cliente, temos mais de 15 atributos. Se cada atributo possui apenas dois valores possíveis, temos $2^{15} = 32768$ configurações de clientes. Em cada uma dessas configurações possíveis, queremos a probabilidade de não pagamento. Precisamos de pelo menos uns 100 indivíduos em cada configuração para estimar a probabilidade. Isto dá 32768000, ou mais de 32 milhões de indivíduos na base de dados.

Simplesmente, não existe base com clientes recentes deste tamanho para este problema. Suponha que não exista na base de dados nenhum indivíduo com idade x , saldo y , etc. Ou quem sabe existam apenas 3 indivíduos com estes atributos. Como estimar bem a probabilidade de não pagamento de um novo cliente com estes atributos?

Uma outra situação em que as coisas não são tão simples para um estatístico é quando o evento de interesse é relativamente raro. Considere, por exemplo, as perdas financeiras associadas com tufões em Taiwan. Qual a probabilidade de ocorrer um tufão causando perda maior que 4 milhões nos próximos 10 anos? Como não existe nenhum tufão que, até agora tenha causada uma perda maior que 4 milhões, devemos estimar esta probabilidade como sendo zero?

■ **Example 1.6 — Mais um exemplo.** Dados T_1, T_2, \dots, T_n : o tempo de sobrevida de n pacientes submetidos a um novo tratamento médico. Deseja-se estimar o tempo esperado $\mathbb{E}(T)$ de sobrevida após o tratamento. Simples: tire a média aritmética dos n tempos observados.

Suponha que o experimento precisa fornecer uma estimativa um ano após o início do estudo. Um ano após o estudo, 50% dos pacientes faleceram (e portanto sabe-se o valor de T para estes indivíduos). Mas 50% ainda não faleceram e não se conhece T para estes outros indivíduos. A média dos valores conhecidos vai tender a subestimar o valor esperado de sobrevida. Como fazer neste caso?

■

Figure 1.12: Perdas financeiras em plantações de arroz causadas por tufões em Taiwan.

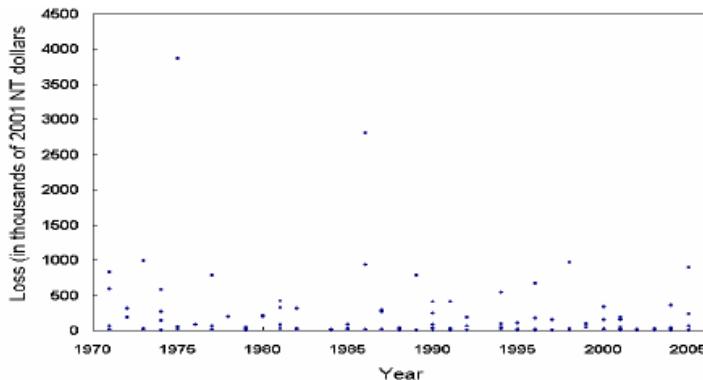


Figure 1. Scatter plot of Taiwan typhoon rice loss

1.5 Modelos probabilísticos para a análise estatística de dados

Precisamos de um *modelo estatístico conceitual*.

Definition 1.5.1 — Modelo Estatístico Conceitual. Uma distribuição de probabilidade hipotética descrevendo como os dados observados poderiam ter sido gerados.

A modelagem é a concepção de um arcabouço matemático capaz de gerar os dados. Os dados que nos interessam não são determinísticos. Assim esse modelo matemático geralmente é um modelo probabilístico ou estocástico. Vamos listar algumas das propriedades desejadas de um bom modelo estatístico.

O modelo probabilístico deve ser capaz de simular dados com características estatísticas semelhantes a aquelas observadas na realidade. Por exemplo, deve ser capaz de prever mais ou menos bem eventos que realmente ocorrem na realidade. O modelo propõe um mecanismo plausível, que corresponde em algum sentido ao que realmente acontece na realidade. Um mecanismo plausível pode sugerir intervenções ou ações que alterem a realidade de alguma maneira desejada (prevendo doenças e fraudes, por exemplo). Finalmente, o modelo deve ser facilmente manipulável matematicamente e conceitualmente. Precisamos fazer cálculos de probabilidade com o modelo. Se ele for muito complexo, não seremos capazes disso.

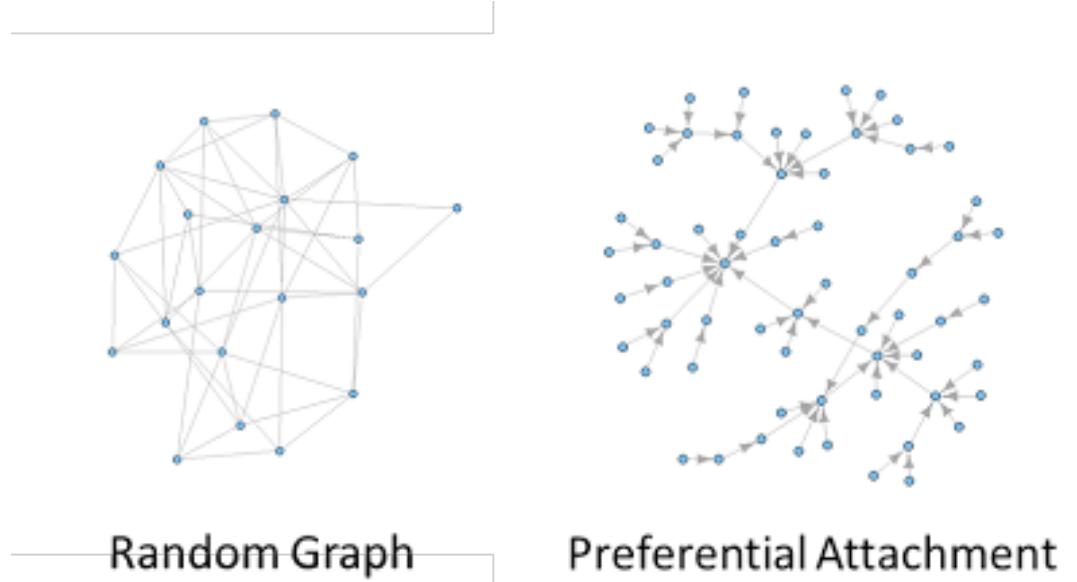
As propriedades costumam ser conflitantes

Muitas vezes, não é possível ter todas as três propriedades simultaneamente. Por exemplo, um modelo para gerar dados que sejam bem realistas talvez tenha que se tornar muito complicado. Isto significa que ele provavelmente vai ser difícil de analisar matematicamente. Por isto, pode ser razoável considerar modelos que reproduzem apenas algumas das características dos dados subjacentes. Queremos reproduzir no modelo as principais características em que estamos mais interessados no momento. O processo de modelagem é geralmente difícil, exige experiência, e muitas vezes é uma ciência e uma arte.

Modelos para quê?

Por que estamos interessados em elaborar modelos matemáticos para os nossos dados ob-

Figure 1.13: Os dois modelos de redes sociais: exemplo de realizações. Modelo de Pólya-Erdös (esquerda) e de Barabási-Albert (direita).



servados? Um bom modelo dá certo significado aos nossos dados e ajuda a entender de forma aproximada o mecanismo por meio do qual os dados são criados. Muitas vezes, o modelo é apenas uma caricatura da situação real. Caricatura é um desenho de um personagem da vida real que enfatiza e exagera algumas das características físicas ou comportamentais da pessoa de uma forma humorística. Em geral, nem de longe, a caricatura é um retrato fiel do indivíduo. Entretanto, ela representa-o de tal modo que, ao vermos a caricatura, imediatamente reconhecemos de quem se trata. Modelos são como caricaturas: capturam o essencial para representar uma situação de forma que propriedades do modelo podem ser depois aplicadas à situação real.

■ **Example 1.7 — Modelos para rede complexa.** Redes complexas possuem a maioria dos seus vértices com poucas arestas. Entretanto, alguns poucos vértices possuem muitas arestas (são os hubs da rede). Seja $\mathbb{P}(K)$ a probabilidade de um vértice possuir k arestas. Quase sempre, encontramos em redes complexas que $\mathbb{P}(K) \approx c/k^\gamma$ onde c e γ são constantes. Isto é chamado uma distribuição de probabilidade na forma power-law (potência inversa de k). Como isto pode acontecer na prática?

Modelo de Pólya-Erdös

Suponha que cada par de vértices joga uma moeda para o alto. Se der cara, um link é estabelecido entre eles. Se der coroa, eles não se ligam. Por mero acaso, alguns vértices terão um número de links maior que outros.

Entretanto, este modelo não é capaz de gerar a característica power-law da realidade. O número de links tem pouca variação em torno da média, nunca gerando os hubs dominantes que vemos nos casos reais. Este não é um bom modelo para as redes complexas da realidade.

Modelo de preferential attachment

O modelo de rede social *preferential-attachment* de Barabási-Albert é uma alternativa bem melhor que o modelo de Pólya-Erdös. Comece com poucos vértices ligados ao acaso entre si pelo modelo anterior. Produza novos vértices sequencialmente. Um novo vértice conecta-se a um nó já existente com uma probabilidade proporcional ao número de arestas que o nó antigo já possui.

A Figura 1.13 mostra dois exemplos de redes sociais geradas a partir dos dois modelos, o modelo de Pólya-Erdös (à esquerda) e de Barabási-Albert (à direita).

O modelo de preferential attachment de Barabási-Albert não é um modelo perfeito para as redes complexas reais. Mas ele induz uma distribuição nos graus dos vértices de redes complexas que possui uma forma de power-law, com cauda pesada. Temos em mãos então um mecanismo hipotético que produz um aspecto muito visível e característico das redes complexas. Temos uma caricatura do processo gerador real das redes complexas.

■

Assim, a intenção do analista de dados é formular uma estrutura matemática simples, mas não trivial, que represente os aspectos essenciais e mais relevantes do fenômeno aleatório de interesse. Semelhante a uma caricatura, um bom modelo probabilístico não é um retrato fiel e perfeito de uma situação real, mas um esboço que reproduz e até amplifica ou exagera os seu traços mais marcantes de forma a torná-lo facilmente reconhecível.

Outro uso de um bom modelo é fazer previsões. Um bom modelo de classificação de risco de crédito serve para isto. Com base em várias características (ou *features*, em inglês) de um usuário, conseguimos prever se ele vai pagar ou não na data combinada um eventual empréstimo. Isto é feito com dados históricos: temos uma enorme coleção de indivíduos que tomaram empréstimo e qual foi o resultado ($Y = 1$, pagou; $Y = 0$, não pagou). Para cada indivíduo, temos também suas características coletadas como um vetor \mathbf{x} . Algumas das características são: sexo, idade, tempo como correntista, saldo médio, etc.

Com estes dados estatísticos, encontramos um modelo para $\mathbb{P}(Y = 1|\mathbf{x})$. Isto é, um modelo para a probabilidade de pagar dado que possui as características \mathbf{x} . Este modelo é usado para predizer o comportamento de futuros tomadores de empréstimo. Um cliente com as características \mathbf{x} chega e pede um empréstimo. Calcule $\mathbb{P}(Y = 1|\mathbf{x})$ usando o modelo. Se a probabilidade é baixa, não conceda o empréstimo.

A tomada de decisões com base em previsões aparece o tempo todo. Devemos conceder o empréstimo? Oferecer desconto a cliente se é grande a chance dele comprar um item muito caro. Cortar a conexão a uma rede se a chance de que certas atividades na rede sejam ação de hackers. Construir uma nova estação metereológica numa localização (x,y) se esta posição minimiza a incerteza de previsões para a região como um todo a partir da rede existente mais a nova estação.



2. Dados e sua descrição estatística

2.1 Dados Estatísticos

Coletamos regularmente dados dos mais variados tipos: numéricos, strings, imagens, sons, vídeos. Estes dados podem estar estruturados de forma complexa. Por exemplo, atributos individuais de usuários do Facebook estruturados na forma de uma grafo de amizade com arestas conectando seguidores e seguidos. Ou podemos ter dados genéticos de indivíduos organizados como árvores genealógicas em estudos de DNA. Todos estes dados podem (e são) analisados estatisticamente.

Entretanto, o tipo de dado mais comum nas análises estatísticas são aqueles organizados de forma tabular. Um exemplo está na Tabela 2.1 que mostra as quatro primeiras linhas de uma tabela com características extraídas de mensagens eletrônicas. Cada linha da tabela corresponde a um email.

As colunas da Tabela 2.1 correspondem a diferentes variáveis extraídas dos emails. Elas são definidas da seguinte forma:

- `spam`: Specifies whether the message was spam.
- `num_char`: The number (#) of characters in the email.
- `line_breaks`: # line breaks in the email (not including text wrapping).
- `format`: Indicates if the email contained special formatting, such as bolding, tables, or links, which would indicate the message is in HTML format.
- `number`: Indicates whether the email contained no number, a small number (< 1 million), or a large number.
- `ratioti`: ratio of image area to text: a message using images instead of words in order to sidestep text-based filtering.
- `%obs`: % HTML with obfuscated text, such as unnecessary hex-encoding of ASCII characters in an attempt to avoid text-based filters.

Esta variáveis foram escolhidas e passaram a ser medidas pois acredita-se que elas podem ser úteis para discriminar emails válidos daqueles que são spam. A primeira coluna foi criada manualmente, com pessoas classificando as mensagens como spam ou não spam. As demais variáveis foram extraídas automaticamente das mensagens, sem intervenção humana. O objetivo é criar uma regra a ser usada em futuras mensagens. Nestas mensagens futuras, apenas as variáveis

spam	num_char	line_breaks	format	number
no	21,705	551	html	small
no	7,011	183	html	big
yes	631	28	text	none
:	:	:	:	:
no	15,829	242	html	small

Table 2.1: Quatro primeiras linhas da tabela `spam`. Fonte: OpenIntro Statistics Project, <https://www.openintro.org/stat/textbook.php>.

das colunas 2 a 6 serão coletadas automaticamente. O objetivo é predizer o valor da primeira variável, `spam`, baseada nas demais variáveis da tabela. Se o modelo for capaz de fazer boas previsões, poderemos descartar com segurança as mensagens de spam sem a necessidade de verificar manualmente cada uma delas. Um modelo estatístico baseado nestes dados é capaz de determinar quais dessas variáveis são relevantes para esta tarefa e como usá-las para predizer se uma mensagem é spam ou não.

Por exemplo, o resultado de uma análise estatística dos dados poderia concluir que apenas duas variáveis são úteis, `ratioti` e `%obs`. Além disso, elas devem ser usadas da seguinte forma: se uma mensagem possui `ratioti` acima de 2 e se `%obs` é maior que 50%, a probabilidade de que a mensagem seja um spam é muito alta e ele deve ser retido. Geralmente, os modelos que são realmente suados para esta tarefa utilizam tabelas com muitas linhas (milhões delas) e com muitas colunas (uma centena ou mais). As regras finais costumam ser mais complexas do que a que apresentamos acima mas a idéia geral é a mesma, apenas a escala do problema fica maior.

2.2 Tipologia de variáveis

Cada linha da tabela corresponde a um *caso*. Casos também são chamados de *observações*, *instâncias*, ou *exemplos*. Cada coluna corresponde a uma *variável*. Uma variável também é chamada de *atributo*, ou *característica* (feature, em inglês). A tabela de dados coletados é chamada de *amostra* (sample, em inglês).

As variáveis podem ser divididas em 4 tipos básicos:

- Variáveis numéricas
 - discreta
 - contínua
- Variáveis categóricas
 - nominal
 - ordinal

Variáveis numéricas

Com um variável *numérica* numa tabela faz sentido somar seus valores (para obter um total geral, por exemplo), subtrair (para medir a diferença entre dois casos, por exemplo) ou tomar médias de seus valores. Exemplos de variáveis numéricas na tabela 2.1 são `num_char`, `line_breaks`, `ratioti` e `%obs`.

Variáveis numéricas *discretas* assumem apenas alguns valores possíveis. Estas valores podem ser colocados numa lista enumerável. Na tabela 2.1, `num_char` e `line_breaks` são variáveis numéricas *discretas*. Elas assumem apenas alguns valores com saltos entre eles (inteiros, neste caso). A lista de valores possíveis não precisa ser finita, como no caso dos inteiros nestas variáveis. Ela precisa ser *enumerável*. Outros exemplos possíveis

No caso de variáveis numéricas *contínuas*, seus valores podem assumir qualquer valor num intervalo da reta real. `ratiot` e `%obs` são exemplos de variáveis contínuas na Tabela 2.1.

■ **Example 2.1 — RRR.** The R statistical language comes with many data sets. Type `data()` to see what they are.

Variáveis categóricas

Como o nome está dizendo, os valores possíveis de variáveis *categóricas* são categorias. Os valores são apenas rótulos indicando diferentes categorias em que os casos podem se classificados. Com estas variáveis categóricas, não faz sentido fazer operações aritméticas com seus valores. Assim, em princípio, nós não somamos, subtraímos ou tiramos médias de colunas na tabela que sejam variáveis categóricas.

No caso de variáveis categóricas *ordinais*, o valor é um rótulo para uma categoria dentre k possíveis e as categorias *podem ser ordenadas*. Existe uma ordem natural nos valores possíveis. Na tabela 2.1, a variável `number` é um exemplo de variável categórica ordinal. Existe uma ordem natural nos valores possíveis: `none < small < big`.

No caso de variáveis categóricas *nominais*, os seus valores possíveis são rótulos de categorias que não podem ser ordenadas. Na tabela 2.1, as variáveis `spam` e `format` são exemplos deste tipo de variável.

■ **Example 2.2** Em pesquisa amostrais usando questionários, é comum que os respondentes (os casos, linhas da tabela), respondam Numa pesquisa, a resposta (*pouco*, *médio*, *muito*) para uma pergunta.

2.3 R, uma linguagem para análise de dados

R é uma linguagem de script interpretada, open-source. Ela é voltada para:

- manipulação de dados,
- análise estatística
- visualização de dados

Elá foi inspirada na linguagem S desenvolvida na AT & T no anos 80. R foi escrita por Ross Ihaka e Robert Gentleman, no Depto de Estatística da Univ de Auckland, NZ.

2.4 Dados tabulares em R

Dados tabulares usualmente são organizados em `data.frames`: são matrizes em que as variáveis (ou colunas) podem ser de tipos diferentes. Alguns dos comandos para ler dados em dataframes são: `read.table`, `read.csv`, e `read.delim`. Vamos usar abaixo `read.csv` para ler um arquivo no formato csv e fazer algumas operações explicadas a seguir.

```
> pressao = read.csv("T1.dat", header = T, row.names = NULL)
> dim(pressao)
[1] 500 501
> pressao = pressao[, 1:18] # selec. 1as. 18 colunas
> colnames(pressao)
[1] "sbp"      "gender"    "married"   "smoke"     "exercise"  "age"
[7] "weight"    "height"    "overwt"    "race"      "alcohol"   "trt"
[13] "bmi"       "stress"    "salt"      "chldbear"  "income"    "educatn"
```

Estes dados são o produto de uma pesquisa conduzida pela empresa farmacêutica GlaxoSmithKline em Toronto, Canadá. Eles foram obtidos em <http://www.math.yorku.ca/Who/Faculty/>

variável	descrição
sbp	Systolic Blood Pressure, Continuous Numerical Variable
gender	Binary Nominal Variable: M = Male, F = Female
married	Binary Nominal Variable: Y = Married, N = Not Married
smoke	Smoking Status, Binary Nominal variable: Y = Smoker, N = Non-Smoker
exercise	Exercise level, Categorical Ordinal variable: 1 = Low, 2 = Medium, 3 = High
age	Continuous Numerical variable (years)
weight	Weight, Continuous Numerical variable (lbs)
height	Height, Continuous Numerical variable (inches)
overwt	Overweight, Categorical ordinal variable: 1 = Normal, 2 = Overweight, 3 = Obese.
race	Race, Categorical nominal variable taking values 1, 2, 3, or 4.
alcohol	Alcohol Use, Categorical ordinal variable: 1 = Low, 2 = Medium, 3 = High
trt	Treatment for hypertension, Binary nominal Variable: Y = Treated, N = Untreated
bmi	Body Mass Index (BMI), Continuous Numerical variable: Weight / Height ² *703
stress	Stress Level, Categorical ordinal variable: 1 = Low, 2 = Medium, 3 = High
salt	Salt (NaCl) Intake Level, Categorical ordinal variable: 1 = Low, 2 = Medium, 3 = High
chldbear	Childbearing Potential, Categorical nominal variable: 1 = Male, 2 = Able Female, 3 = Unable Female
income	Income Level, Categorical ordinal Variable: 1 = Low, 2 = Medium, 3 = High
educatn	Education Level, Categorical ordinal Variable: 1 = Low, 2 = Medium, 3 = High

Table 2.2: Variáveis de uma tabela com os dados coletados em uma pesquisa conduzida pela empresa farmacêutica GlaxoSmithKline em Toronto, Canadá.

Ng/ssc2003/BPMain.htm. O dataframe pressao contém 500 pacientes, cada um deles numa linha da tabela. Estas são os casos ou instâncias ou observações da análise estatística. A tabela possui 501 variáveis ou atributos, as colunas da tabela. As 501 variáveis (ou colunas) consistem de:

- pressão sistólica do paciente
- 17 variáveis clínicas potencialmente preditoras de hipertensão,
- 483 marcadores genéticos

No terceiro comando acima, a tabela inicial é reduzida, ficando apenas com as suas primeiras 18 colunas. Eliminamos os 483 marcadores genéticos ficando apenas com as variáveis clínicas. O último comando pede a listagem dos nomes das colunas do dataframe pressao.

Dos 500 pacientes, metade tinha pressão arterial baixa e metade, elevada (hipertensão). A definição das variáveis, junto com seu tipo, é dada na tabela abaixo.

2.5 Vetores e resumos numéricos

Para entrar rapidamente com pequenos conjuntos de dados em R podemos usar a função `c`, que combina ou concatena elementos num vetor. Depois de armazenar os dados num vetor, aplicamos uma série de funções estatísticas tais como calcular o seu valor máximo, a média, etc.

```
> # gols marcados no brasileirao de 2014, por time
> x = c(67,59,53,49,51,61,36,43,42,46,36,38,37,42,39,34,37,31,31,28)

> max(x)    # uma funcao aplicada ao vetor
[1] 67

> mean(x)  # funcao estatistica
[1] 43

> median(x); sum(x)  # ok ter mais de um comando por linha usando ";"
[1] 40.5
[1] 860

> summary(x)  # resumo basico com 5 numeros
   Min. 1st Qu. Median Mean 3rd Qu. Max.
28.0    36.0    40.5   43.0   49.5   67.0

> sort(x)
[1] 28 31 31 34 36 36 37 37 38 39 42 42 43 46 49 51 53 59 61 67

> x[1]-x[2]    # acessando elementos do vetor
[1] 8

> x > 40 & x < 50  # vetor logico: quem atende 'a condicao?
[1] FALSE FALSE FALSE TRUE FALSE FALSE FALSE TRUE ..... FALSE

> mean( x[ x > 40 & x < 50] )  # aninhando: media dos x's que sao > 40 e < 50
[1] 44.4

> which(x == max(x))  # quais posicoes do vetor sao T
[1] 1
```

Mais comandos auto-explicativos aplicados em um vetor numérico `x`.

```
> x[c(3, 5, 8:11)]  # selecionando elementos de x
[1] 53 51 43 42 46 36

> y = log(x/2) - 3    # se alguma vez precisar disso ...

> y
[1] 0.51154544 0.38439026 0.27714473 .....

> round(y, 3)
[1] 0.512 0.384 0.277 .....
```

```

> sum( log(x) + x^2 )    # e' claro que queremos calcular isto com os gols, certo?
[1] 39226.67

> sum(x > 50)    # operacao numerica com vetor logico
[1] 5

> c(x, c(20, 39, 45))  # acrescentando gols de 3 times adicionais
[1] 67 59 53 49 .... 31 28 20 39 45

> x = c(x, c(20, 39, 45))    # salvando em x

> x[ (length(x) - 20) : (min(x) - 12) ]  # funcoes dentro de indexadores
[1] 53 49 51 61 36 43

> cumsum(x)      # soma acumulada de gols, na ordem do vetor x
[1] 67 126 179 228 279 340 376 ...

> rev(cumsum(x))  # revertendo a soma acumulada de gols
[1] 964 919 880 860 832...

```

Mais comandos em R usando um vetor numérico:

```

> 1:9
[1] 1 2 3 4 5 6 7 8 9

> seq(0, 1, by=0.1)
[1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

> seq(0, 1, length=11)
[1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

> rep(-1, 5)
[1] -1 -1 -1 -1 -1

> rep(c(-1, 0), 5)
[1] -1 0 -1 0 -1 0 -1 0 -1 0

> rep(5, c(-1, 0))
Erro em rep(5, c(-1, 0)) : argumento 'times' invalido

> rep(c(-1, 0), c(5, 3))
[1] -1 -1 -1 -1 -1 0 0 0

> rep(-1:2, rep(3, 4))
[1] -1 -1 -1 0 0 0 1 1 1 2 2 2

```

2.6 Visualizando dados numéricos

2.6.1 Histograma

A maneira de visualizar os dados de uma tabela depende do tipo de variável. Para variáveis numéricas, o histograma é uma excelente opção. Ele permite ver como os dados de uma variável

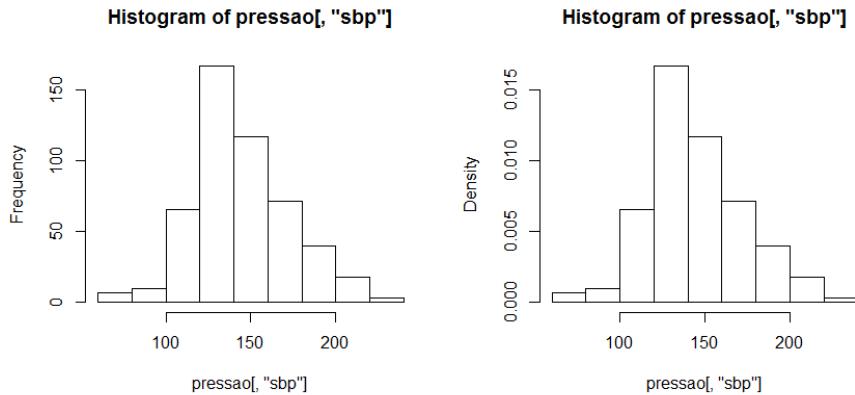


Figure 2.1: Histograma dos 500 valores da variável `sbp`, pressão sistólica, vindos da tabela 2.2.

numérica (tipicamente contínua) espalham-se no intervalo formado pelo menor e pelo maior valor da amostra. Simplesmente olhando o gráfico podemos perceber se os dados tendem a estar acumulados numa pequena região dentro do intervalo delimitado pelos extremos (máximo e mínimo). Ao invés disso, os dados podem estar igualmente bem espalhados dentro daquele intervalo ou pode ter duas pequenas regiões de grande concentração.

Digitando `hist(pressao[, "sbp"])` produz o gráfico na esquerda da Figura 2.1. O gráfico da direita foi feito acrescentando o parâmetro opcional `prob=T` ao comando anterior. Isto é, digitando `hist(pressao[, "sbp"], prob=T)` temos o gráfico da direita. Observe que os dois gráficos são idênticos exceto pela escala vertical. No gráfico da direita a soma das áreas dos retângulos é igual a 1. Vamos discutir estes histogarma com área total igual a 1 mais a frente.

Considerando o gráfico da esquerda, vemos que o intervalo [100, 200] possui cinco retângulos e portanto cada retângulo tem uma base de comprimento aproximadamente igual a 20. Veremos mais tarde como ter controle do tamanho do intervalo bem como de outros aspectos do histograma. Usando 20 como comprimento, os dados estão, grosso modo, espalhados no intervalo [60, 240]. Como os indivíduos se distribuem dentro deste intervalo? Aqui o histograma é útil. A altura de cada retângulo no gráfico da esquerda é a contagem do número de indivíduos da amostra que caíram dentro do intervalo. A regra fundamental para olhar um histograma é a seguinte:

Num histograma, as áreas dos retângulos relativas à área total representam porcentagens

Qual a porcentagem dos indivíduos da amostra que possuem pressão entre 180 e 200? Como a altura do retângulo cuja base é o intervalo [180, 200] tem uma altura menor que 50, um valor próximo de 40. Na verdade, a contagem neste intervalo é exatamente igual a 43, mas isto não importa. Queremos apenas ter uma idéia qualitativa da distribuição. Como existem 500 pacientes na amostra, o valor aproximado de 40 pessoas no intervalo diz que aproximadamente 8% dos indivíduos caíram entre 180 e 200. Se perguntarmos qual a proporção que tem pressão acima de 180, vamos encontrar aproximadamente $(40 + 20 + 5)/500 = 0.13$ ou 13% da amostra com pressão alta. Se quisermos a mediana dos dados, o valor que deixa aproximadamente metade da amostra abaixo dele e a outra metade acima, podemos tentar obter-lo apenas olhando o histograma. É preciso encontrar um ponto no eixo horizontal que deixa a área total dos retângulos à sua esquerda igual a 50% da área total e, claro, a área à esquerda também igual a 50%. De forma aproximada por causa das alturas diferentes e de forma puramente visual, verificar que esta mediana não deve estar nem abaixo de 120 nem acima de 160. Podemos estimar que a mediana deve ser algum valor entre 140 e 160.

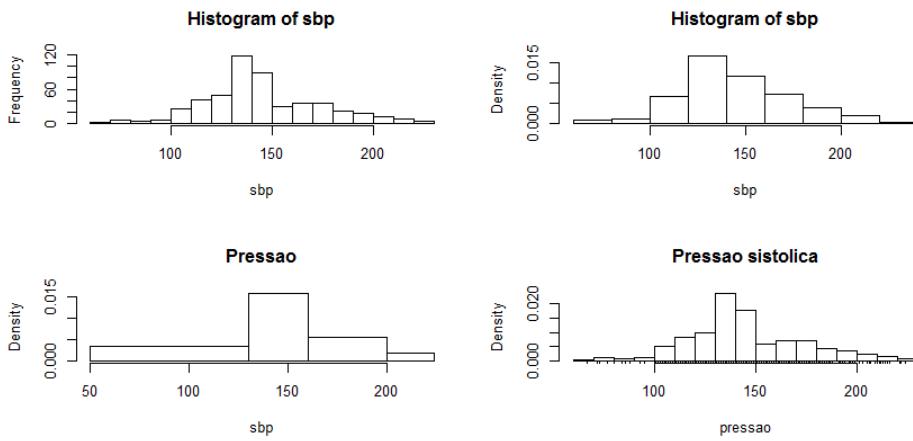


Figure 2.2: Plots mostrando algumas opções ao usar o comando `hist`.

O algoritmo para criar um histograma é muito simples. Temos N casos (ou exemplos ou instâncias de uma variável numérica. Forme uma grade quebrando o eixo horizontal em pequenos intervalos de comprimento Δ . Conte o número de casos em cada um dos intervalos: n_1 casos no intervalo 1, n_2 casos no intervalo 2, etc. de forma que $N = \sum_i n_i$. Faça um retângulo usando o intervalo da grade como base. A altura do i -ésimo retângulo é igual à:

- (A) contagem n_i de dados que caem no intervalo i (gráfico à esquerda na Figura 2.1).
- (B) ou igual à proporção n_i/N que cai no intervalo i dividida por Δ . Isto é, altura é $n_i/(N\Delta)$ (gráfico à direita na Figura 2.1)

No caso (A), a soma das áreas dos retângulos do histograma varia com o tamanho da amostra. No caso (B), a soma das áreas dos retângulos é sempre igual a 1. Esta propriedade é importante pois, como veremos no capítulo ??, ela permite comparar graficamente os histogramas com curvas chamadas densidades de probabilidade. A maneira mais útil de usar um histograma, seja do tipo (A) ou do tipo (B), é calculando as áreas dos retângulos relativamente à área total. A soma (relativa) das áreas dos retângulos de um intervalo do eixo horizontal fornece a proporção dos elementos da amostra que caem naquele intervalo.

O comando no R para criar um histograma é `hist(x)` onde `x` é um vetor numérico ou uma coluna numérica de um dataframe. Este comando usa a contagem n_i descrita acima em (A). Existem várias opções para alterar o histograma básico, incluindo o argumento `prob=T` para criar um histograma do tipo (B):

```
attach(pressao) # com isto, podemos nos referir 'as colunas pelo nome
par(mfrow=c(2,2)) # tela grafica para 4 graficos num formato 2 x 2
hist(sbp, breaks = 20) # controlando numero de breaks
hist(sbp, prob = T) # histograma possui area total 1, opcao (B) acima
# A seguir, controle da grade com breaks e titulo para o grafico
hist(sbp, breaks = c(50, 130, 160, 200, max(sbp)), prob=T, main="Pressao")
# Controlando o rotulo para o eixo horizontal
hist(sbp, breaks = 15, prob=T, xlab="pressao", main="Pressao sistolica")
rug(sbp) # acrescenta um tapete com os dados originais
```

O resultado do uso destes comandos está nos plots da Figura 2.2.

■ **Example 2.3 — Pirâmides etárias são histogramas.** Um exemplo interessante de uso do histograma é ao visualizar a evolução da população brasileira nas décadas mais recentes por

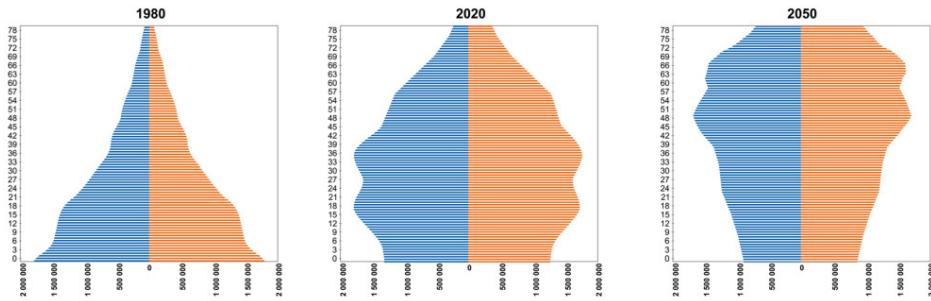


Figure 2.3: Pirâmides etárias do Brasil em 1980 e projeções em 2020 e 2050. Fonte: IBGE.

meio das pirâmides etárias, que são simplesmente histogramas dispostos verticalmente (girando o histograma usual em 90 graus). A Figura 2.3 mostra as pirâmides etárias da população brasileira em 1980 e a sua projeção para os anos de 2020 e 2050. Em cada pirâmide, a população masculina é pintada de azul e a feminina de vermelho. Cada barra horizontal representa um ano de idade. As idades são lidas no eixo vertical. No eixo horizontal, temos a contagem do número de pessoas que possuem aquela idade no ano em questão. Assim, a pirâmide de cada sexo é simplesmente um histograma da distribuição por idade dos indivíduos daquele grupo mas rotacionado de 90º. O histograma masculino é colocado junto ao histograma feminino, o que facilita a comparação entre eles.

É chocante a mudança prevista na estrutura etária do Brasil em apenas 80 anos. Em 1980 a estrutura tinha realmente uma forma de pirâmide com os jovens dominando a população. A parcela que requer aposentadorias, pensões e cuidados maiores e mais caros com a saúde são aqueles acima de 60 anos. Eles representam uma pequena proporção da população total. Visualmente, e de forma muito aproximada, os histogramas nos dizem que a proporção de idosos em 1980 seria menos de 5%, por volta de 15% em 2020 e 25% em 2050. Como os custos de um sistema de previdência social cotumam ser cobertos com contribuições dos mais jovens que ainda estão ativos, temos uma parcela cada vez menor de pessoas sustentando um grupo que cresce relativamente ao total populacional. Se em 2017 a previdência é altamente deficitária, a situação pode ficar insustentável num futuro próximo a menos que haja aumento de impostos (e diminuição do crescimento da economia) ou redução de benefícios (com impacto político negativo para quem implementar a mudança).

■ **Example 2.4** Outro exemplo mostra a ocasional necessidade de transformar os dados para compreender os melhores. A Figura 2.4 mostra histogramas da população residente nos 5564 municípios brasileiros em 2006 e foi gerada com código abaixo:

```
> pop = read.csv("POP2006.csv", header = T, row.names = NULL)
> colnames(pop)
[1] "ESTADO"      "MUNICIPIO"    "POP2006"

> par(mfrow=c(2,2), oma=c(0,0,0,0), mar=c(2, 3, 2, 1))
> hist(pop[,3], main="populacao municipal em 2006")

> sum(pop[,3] > 10^6) # quantas cidades com mais de 1 milhao?
[1] 14
> # Histograma apenas das cidades menores que 1 milhao
> hist(pop[pop[,3] < 10^6,3], main="pop < 1 milhao")
```

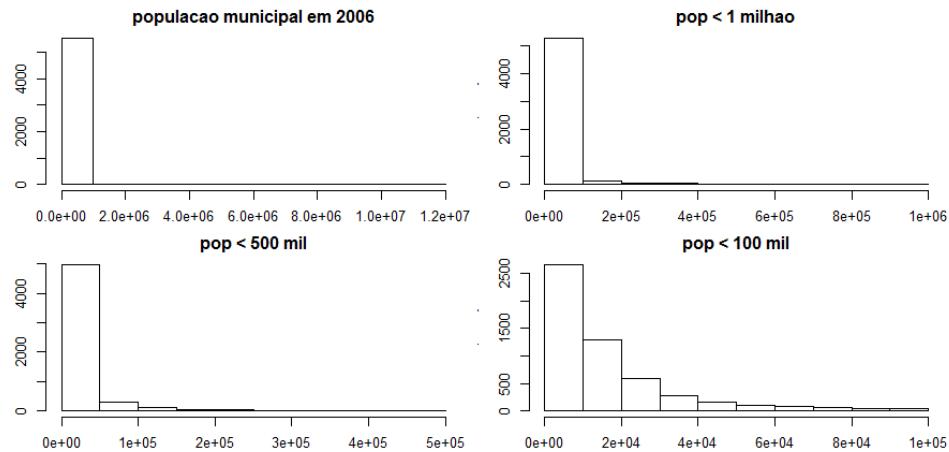


Figure 2.4: População dos 5564 municípios brasileiros em 2006. Fonte: IBGE.

```
> sum(pop[,3] > 5*10^5) # 36 cidades maiores que 500 mil
> hist(pop[pop[,3] < 5*10^5,3], main="pop < 500 mil")

> sum(pop[,3] > 10^5) # 267 cidades maiores que 100 mil
> hist(pop[pop[,3] < 10^5,3], main="pop < 100 mil")
```

O parâmetro `oma` controla o espaço das margens externas da janela gráfica e o parâmetro `mar` controla as margens internas de cada plot. Veja o excelente site Quick-R em <http://www.statmethods.net/advgraphs/axes.html>.

O gráfico na posição (1,1) na figura tem os dados de todos os 5564 municípios. Enquanto no histograma de pressão sistólica tínhamos os dados espalhando-se de maneira simétrica para cada dos lados em torno de um ponto central, aqui os dados distribuem-se no eixo horizontal de forma muito diferente. Existe uma imensa desigualdade nos tamanhos de população dos municípios, com a maioria deles tendo uma população relativamente pequena. Esta grande maioria é a responsável pela primeira barra à esquerda, de altura maior que 5000. De fato, o tamanho de cada intervalo da grade é igual a 10^6 , ou 1 milhão de residentes. O comando

```
sum(pop[,3] > 10^6)
```

retorna 14 cidades com mais de 1 milhão. Assim, um punhado de 14 municípios distribuem-se na maior parte do espaço do eixo horizontal enquanto todos os demais municípios têm menos de 1 milhão de habitantes e estão empilhados na primeira barra do histograma. É impossível ver como esta maioria dos municípios se distribui na pequena faixa de 0 a 1 milhão.

Às vezes, esse problema se resolve eliminando estes poucos valores muito extremos e refazendo o histograma apenas com os restantes. Neste caso, a escala horizontal iria apenas até 1 milhão de habitantes e costuma ser possível visualizar melhor as populações da maioria dos municípios. Mas este não é o caso desses dados. Os gráficos nas posições (1,2) e (2,1) mostram o histograma das populações de cidades com menos de 1 milhão e com menos de 500 mil habitantes, respectivamente. O mesmo tipo de gráfico com extrema desigualdade e dificuldade de enxergar os valores menores se repete.

Na posição (2,2), temos o gráfico com as 5297 cidades com menos de 100 mil residentes. Apenas aqui, eliminando as 267 cidades com mais de 100 mil habitantes, conseguimos visualizar um pouco melhor como os municípios se distribuem em termos de seus tamanhos. Cada intervalo na grade do eixo horizontal possui tamanho igual a 10 mil habitantes. Visualmente podemos estimar que por volta de 80% deles possuem menos de 40 mil habitantes. De fato, a proporção exata é dada por

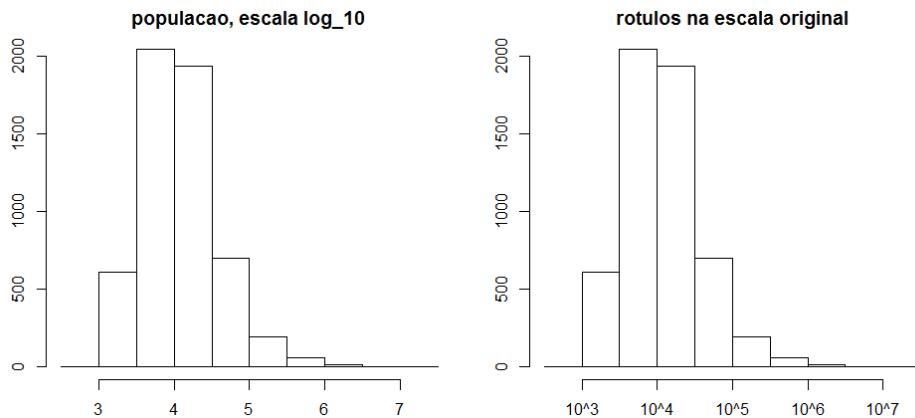


Figure 2.5: População na escala logarítmica (base 10) dos 5564 municípios brasileiros em 2006.
Fonte: IBGE.

```
> sum(pop[,3] < 40000)/length(pop[,3])
[1] 0.8666427
```

Entretanto, mesmo sendo capaz de enxergar a maioria dos municípios neste último gráfico, ele deixa a desejar. Primeiro, nós não conseguimos enxergar ao mesmo tempo onde estão as populações dos 267 municípios maiores que costumam ser os mais importantes em termos econômicos, políticos e culturais. Em segundo lugar, vemos que os municípios possuem uma distribuição de tamanho que decresce a medida que o tamanho aumenta. Podemos tentar estudar como se dá este decrescimento do número de municípios com o aumento de seu tamanho. Será que existe um regra simples para isto? Note que a segunda barra parece ter a metade da altura da primeira e que a terceira barra parece ter também a metade da altura de segunda. Será que a regra é: ao passar de uma categoria de tamanho (digamos, entre 40 e 50 mil habitantes) para a seguinte, o número de cidades se reduz pela metade? Podemos checar isto considerando as razões sucessivas entre as contagens das barras fica em torno de 1/2.

```
> aux = hist(pop[pop[,3] < 10^5,3], main="pop < 100 mil")
> aux$counts # vetor com as contagens das barras do histograma
[1] 2662 1291 585 284 164 93 76 62 42 38
> round( aux$counts[-1]/aux$counts[-10], 3)
[1] 0.485 0.453 0.485 0.577 0.567 0.817 0.816 0.677 0.905
```

Assim, realmente nas primeiras barras esta razão fica em torno de 1/2 mas depois ela vai se elevando de forma que nas últimas categorias o número cai muito pouco. E não sabemos o que acontece com as categorias acima de 100 habitantes.

Uma outra forma de visualizar estes dados, de todos os municípios de uma única vez, é olhá-los na escala logarítmica. Um novo vetor de dados foi obtido tomando-se o logaritmo (base 10) da população de cada cidade. O histograma destes novos valores transformados pelo \log_{10} está no lado esquerdo da Figura 2.5. A escala mostrada no eixo horizontal corresponde aos valores de $\log_{10}(\text{pop})$. Assim, o valor 4 na escala significa que $\log_{10}(\text{pop}) = 4$, ou seja, $\text{pop} = 10^4$, ou 10 mil habitantes. O gráfico da direita na Figura 2.5 é o mesmo que o gráfico da esquerda exceto que, na escala horizontal, os rótulos $3, 4, \dots, 7$ (e apenas esses rótulos) foram substituídos pelos rótulos $10^3, 10^4, \dots, 10^7$ para ajudar a entender melhor o que as barras representam na escala original de populacional.

O que significa olhar um gráfico na escala logarítmica? Ao passar de 3 para 4 na escala \log_{10} , a população aumenta de 10 vezes o seu tamanho. Ao aumentar mais um grau nesta escala, passando

de 4 para 5, novamente o tamanho é multiplicado por 10. Isto é, uma cidade que tem uma distância de n unidades a mais que outra cidade na escala log-da-população possui uma população 10^n vezes maior. Assim, diferenças na escala log traduzem-se por incrementos multiplicativos na escala original. De outro modo: cada salto de tamanho 1 na escala log significa multiplicar por 10 na escala original.

Qual a vantagem de se usar a escala logarítmica? Uma das razões é que esta escala pode ser a mais natural para estudar a variação de tamanho e, em particular, de tamanho de cidades. Imagine que você mora numa cidade A com 20 mil habitantes e muda-se para a cidade B com 100 mil habitantes. O impacto que esta mudança vai causar será grande. Depois de algum tempo, você muda novamente para uma cidade C , maior ainda que B . Caso C tenha 180 mil habitantes, haverá um impacto mas possivelmente não tão grande quanto primeiro, $A \rightarrow B$. Para ter um impacto similar a este primeiro, talvez C tenha de ter um tamanho de 500 mil habitantes para que a vida urbana no novo local seja suficientemente diferente daquele em B . Isto é, ao comparar diferentes tamanhos de lugares, parece ser útil considerar diferenças numa escala multiplicativa e não puramente aditiva. Somar 5 habitantes numa cidade que possui apenas 10 mil terá um impacto enorme enquanto que os mesmos 5 mil adicionados a uma cidade com 500 mil não farão diferença significativa.

A outra razão, mais empírica, é que nos gráficos da Figura 2.5 vemos uma distribuição mais fácil de ser entendida. Ela está distribuída de forma mais balanceada em torno de um valor central. Novamente olhando as áreas debaixo dos retângulos, a população mediana (o valor que divide a amostra em 50% acima e 50% abaixo de si) parece ser por volta de 10000 (isto é, 4 no gráfico da esquerda na Figura 2.5 ou 10^4 no da direita). De fato, esta intuição está correta: `median(pop[, 3])` produz 10687. Os dados não são simétricos em torno desta mediana mas não se estendem para cada um dos dois lados de forma muito desigual.

■

2.6.2 Ramo-e-folhas

Se a quantidade de dados, é pequena, o gráfico de ramo-e-folhas (*stem-and-leaf*, em inglês) é bem útil. O ramo-e-folhas pode ser feito à mão rapidamente e permite visualizar toda a distribuição dos na sua faixa de variação. A ideia básica é usar os próprios dígitos dos valores que queremos visualizar para construir um histograma. Por exemplo, vamos olhar os dados dos gols que cada time fez ao longo do campeonato brasileiro de futebol em 2014.

```
> bras = read.csv("CampeonatoBrasileiro2014.txt", header=T, row.names=NULL)
> head(bras)
      Time Pts Jogos Vit Emp Der Gols GolsSofr SaldoGols Aprov
1   Cruzeiro  80    38  24   8   6   67      38        29     70
2 Sao Paulo  70    38  20  10   8   59      40        19     61
3 Internacion 69    38  21   6  11   53      41        12     60
4 Corinthians 69    38  19  12   7   49      31        18     60
5 Atletico Mineiro 62    38  17  11  10   51      38        13     54
6 Fluminense  61    38  17  10  11   61      42        19     53
> bras[, "Gols"]
[1] 67 59 53 49 51 61 36 43 42 46 36 38 37 42 39 34 37 31 31 28
```

O número de gols por time varia de 28 a 67. Tomando os possíveis primeiros dígitos, 2, 3, 4, 5 ou 6, como os ramos, nós os dispomos numa coluna mais a esquerda. A seguir, empilhamos os segundos dígitos no ramo correspondente, como se fossem folhas brotando do ramo. Veja a saída do comando `stem` abaixo:

```
> stem(bras[, "Gols"])
```

```
The decimal point is 1 digit(s) to the right of the |
```

```
2 | 8
3 | 114667789
4 | 22369
5 | 139
6 | 17
```

Veja a segunda coluna de dados no gráfico: 3 | 114667789. Ela representa todos os valores com o primeiro dígito igual a 3. Isto é, empilhamos nesta coluna os valores 31, 31, 34, 36, ..., 39. Veja outros exemplos de ramo-e-folhas:

```
> sort(bras[, "SaldoGols"])
[1] -28 -25 -17 -17 -12 -10 -10 -5 -3 -2 -1 1 ... 19 19 29

> stem(bras[, "SaldoGols"])
-2 | 85
-0 | 772005321
0 | 17223899
2 | 9
```

Se quiser quebrar cada categoria-dígito em grupos de 5, use "scale"

```
> stem(bras[, "Gols"], scale=2)
```

```
The decimal point is 1 digit(s) to the right of the |
```

```
2 | 8
3 | 114
3 | 667789
4 | 223
4 | 69
5 | 13
5 | 9
6 | 1
6 | 7
```

2.6.3 Boxplot

O boxplot é um resumo gráfico com dos dados com alta compressão: usa 5 números apenas. Ele mostra rapidamente se os dados são simétricos, onde estão concentrados e se existem outliers (valores extremos). A Figura 2.6 mostra o boxplot usando os dados da variável `sbp`, pressão sistólica, vindos da tabela 2.2. A caixa (box) central tem extremidades laterais essencialmente em Q1 e Q3. O valor de Q1 é o primeiro quartil: 25% dos dados ficam abaixo dele, os outros 75%, acima. O valor de Q3 deixa 25% dos dados acima e 75% abaixo. Em inglês, no contexto do boxplot, eles são chamados de lower hinge (Q1) and upper hinge (Q3). A linha que divide a caixa central fica na altura de Q2: a mediana, que deixa 50% dos dados abaixo e 50% acima.

Duas linhas, chamadas de bigodes de gato (whiskers), estendem-se a partir da caixa. A linha superior usualmente tem comprimento igual a 1.5 vezes o comprimento da caixa. Isto é, 1.5 vezes distância interquartílica. Na verdade, ela tem este comprimento se existirem dados maiores que o

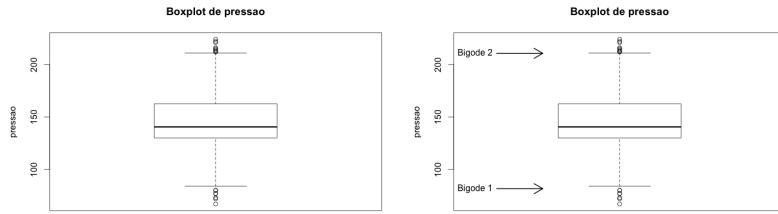


Figure 2.6: Boxplot usando os dados da variável `sbp`, pressão sistólica, vindos da tabela 2.2.

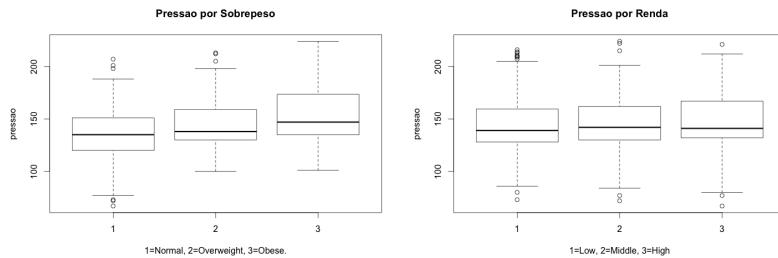


Figure 2.7: Boxplot de pressão sistólica versus a categoria de peso do indivíduo (esquerda) e versus o seu nível de renda (direita).

bigode. Caso o máximo dos dados seja menor que o limiar do bigode superior, o bigode vai apenas até o próprio máximo dos dados. As mesmas definições são usadas para estabelecer o bigode inferior. Num boxplot, os dados além dos bigodes são mostrados individualmente como pontos. Eles são chamados de dados *outliers*, valores extremos que *potencialmente* podem representar valores errados, anomalias ou dados estranhos.

O boxplot é um tipo de visualização muito útil para comparar como uma distribuição muda a medida em que mudamos o valor de *outra* variável categórica. Por exemplo, imagine que queremos estudar se a distribuição de valores da pressão sistólica entre pessoas com peso numa faixa normal é diferente da distribuição de valores da pressão sistólica entre pessoas com sobrepeso e entre pessoas obesas. Se fossemos usar histogramas para isto, teríamos de olhar simultaneamente três histogramas. Ao invés de três categorias, se tivéssemos mais (uma dezena de categorias, digamos) a tarefa ficaria muito difícil. Com boxplots, a visualização escala com facilidade.

A Figura 2.7 mostra no lado esquerdo três boxplots, um para cada grupo de observações (ou casos) de acordo com sua categoria de peso: normal, sobrepeso, obeso. O eixo vertical é comum aos três boxplots e torna possível compará-los. Vemos que as caixas se deslocam verticalmente a medida que o peso aumenta. Isto mostra que a pressão da maioria dos indivíduos obesos está numa faixa de valores um pouco superior que a dos indivíduos com peso normal. Isto não quer dizer que todo indivíduo obeso tenha pressão maior que a de qualquer indivíduo de peso normal. Claramente, existe uma sobreposição razoável dos valores de pressão entre os três grupos de peso. Entretanto, o grupo como um todo (como uma população ou uma distribuição) desloca-se verticalmente ao mudarmos do grupo normal para o obeso.

Em contraste, este deslocamento não ocorre no gráfico do lado direito da Figura 2.7. Ela mostra os boxplots da pressão sistólica particionando a amostra de acordo com o nível de renda do paciente (baixa, média ou alta). Neste novo gráfico, os boxplots são praticamente os mesmos ao longo do eixo vertical. Isto significa que, ao mudarmos de nível de renda, a distribuição dos valores de pressão fica inalterada. A renda não parece ser um fator capaz de afetar a distribuição de pressão. A Figura 2.7 foi feita com os seguintes comandos:

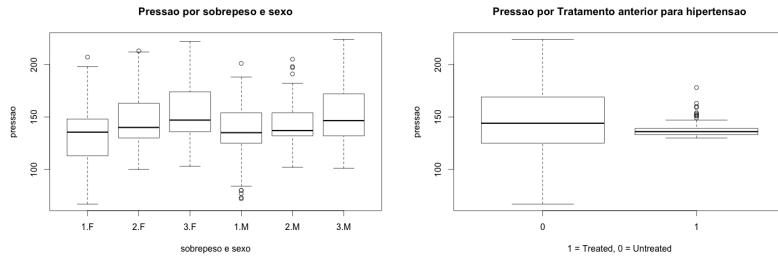


Figure 2.8: Esquerda: Pressão sistólica para as 6 categorias criadas cruzando as variáveis categóricas sobre peso e sexo. Direita: Pressão sistólica para dois grupos de pacientes, aqueles que receberam um tratamento anterior para hipertensão e aqueles não tratados.

```
> par(mfrow=c(1,2))
> boxplot(sbp ~ overwt, ylab="pressao", xlab="1 = Normal, 2 = Overweight, 3 = Obese.")
> title("Pressao por Sobre peso")
> boxplot(sbp ~ income, ylab="pressao", xlab="1=Low, 2=Middle, 3=High")
> title("Pressao por renda")
```

Podemos cruzar duas variáveis categóricas para criar um novo conjunto de categorias. Por exemplo, a Figura 2.8 mostra no lado esquerdo a distribuição de pressão sistólica versus as 6 categorias criadas cruzando as variáveis categóricas sobre peso e sexo. Este gráfico foi criado com os comandos seguintes:

```
> par(mfrow=c(1,1))
> boxplot(sbp ~ overwt*gender, ylab="pressao", xlab="sobre peso e sexo")
> title("Pressao por sobre peso e sexo")
```

Os quatro principais gases ligados ao efeito estufa são o dióxido de carbono (CO₂), metano (CH₄), óxido nitroso (N₂O) e os halocarbonos ou CFC (gases contendo flúor). O gráfico na Figura 2.6.3 mostra uma comparação dos níveis de dois desses gases, CO₂ e CH₄, em três grandes cidades: Londres, Nova York e Los Angeles.

O código para este gráfico foi feito por Eric Cai e foi extraído de <http://bit.ly/2np7ikU>. Ele encontra-se abaixo e assume que existe um dataframe, chamado all.data, com 3 colunas contendo os dados. A primeira coluna é value e contém o nível de poluição. A segunda é location e é uma variável categórica com o nome da cidade. A terceira é pollutant e armazena o tipo de poluente.

```
boxplots.triplet = boxplot(value ~ location + pollutant, data = all.data,
  at = c(1, 1.8, 2.6, 6, 6.8, 7.6), xaxt='n',
  ylim = c(0,27), col = c('white', 'white', 'gray'))
axis(side=1, at=c(1.8, 6.8),
  labels=c('Methane (ppb)\nNumber of Collections = 100',
  'Carbon Dioxide (ppb)\nNumber of Collections = 120'), line=0.5, lwd=0)
title('Comparing Pollution in London, Los Angeles, and New York')
```

O gráfico na Figura 2.6.3 mostra um caso mais interessante, em que muitos box-plots são mostrados em grupos de 3. O código, extraído de <http://bit.ly/2nGyg3G>, está abaixo. As categorias são obtidas pelo cruzamento de duas variáveis. A primeira tem três níveis e é indicada pelas três cores. A segunda possui 20 níveis indicados pelas marcas no eixo horizontal. Assim, este

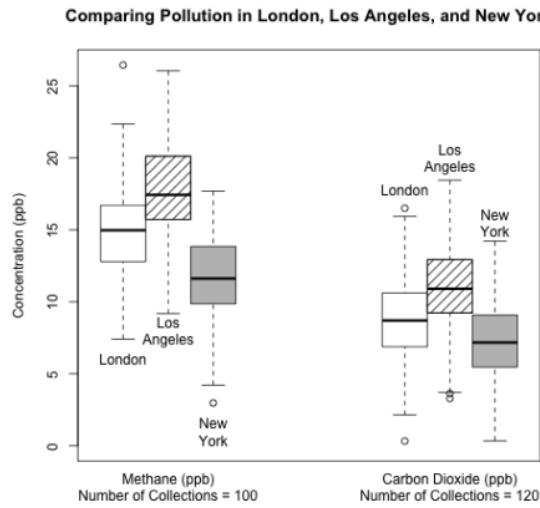


Figure 2.9: Níveis de poluição de dois gases ligados ao efeito estufa, CO₂ e CH₄, em três grandes cidades: Londres, Nova York e Los Angeles. Fonte: <http://bit.ly/2np7ikU>.

gráfico exibe simultaneamente $20 \times 3 = 60$ diferentes distribuições de dados. Veja que podemos acompanhar o valor central (a mediana) de cada um dos 60 grupos de dados, a caixa de cada um deles (que representa a região onde 50% dos dados de cada grupo está localizada), bem como a extensão completa dos dados, incluindo possíveis outliers. Além disso, eles podem ser comparados de forma simples e efetiva, sem muita ginástica mental. É muita informação condensada num espaço físico pequeno mas que é facilmente visualizada. Seria muito difícil ter tudo isto com outro tipo de resumo, tal como 60 histogramas, por exemplo.

```
d = data.frame(x=rnorm(1500),f1=rep(seq(1:20),75),f2=rep(letters[1:3],500))
# first factor has 20+ levels
d$f1 = factor(d$f1)
# second factor a,b,c
d$f2 = factor(d$f2)
```

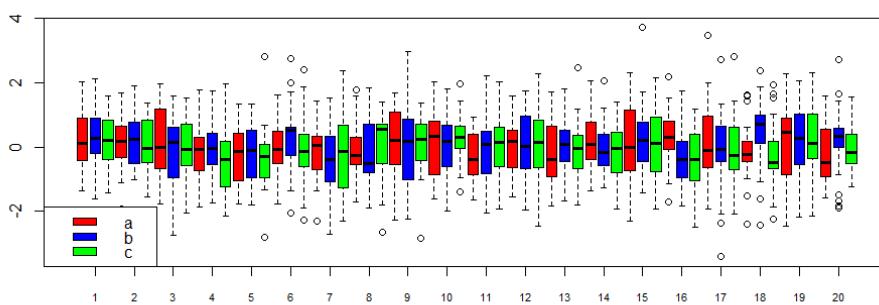


Figure 2.10: Box-plots para cada uma de muitas categorias. Cada grupo de 3 boxplots representa são mostrados em grupos de 3. Fonte: <http://bit.ly/2nGyg3G>

```
boxplot(x~f2*f1,data=d,col=c("red","blue","green"),frame.plot=TRUE,axes=FALSE)

# y axis is numeric and works fine
yts = pretty(d$x,n=5)
axis(2,yts)

# a label at the middle of each group of 3 boxes:
axis(1,at=seq(2,60,3),labels=1:20,cex.axis=0.7)

# Use the legend to handle the f2 factor labels
legend("bottomleft", max(d$x), c("a", "b","c"),fill = c("red", "blue","green"))
```

A distribuição de uma variável pode mudar de forma complexa. No lado direito da Figura 2.8 temos a variável pressão sistólica para dois grupos de pacientes de acordo com o status de `trt`, uma variável categórica binária. A variável `trt` indica se o paciente recebeu ou não um tratamento anterior para hipertensão. O que observamos é que o valor mediano (a linha horizontal no centro da caixa) praticamente não se alterou, mostrando que o tratamento recebido não modificou muito o valor médio da pressão. Entretanto, a dispersão dos valores em torno desta média mudou substancialmente. Os valores dos indivíduos tratados quase não variam em torno de seu valor médio. Em contraste, o grupo não tratado tem grande variabilidade em torno de seu valor médio, com alguns indivíduos possuindo pressão sistólica muito maior ou muito menor que o valor médio do grupo. Os comandos para este segundo gráfico da Figura 2.8 são:

```
> par(mfrow=c(1,1))
> boxplot(sbp ~ trt, ylab="pressao", xlab="1 = Treated, 0 = Untreated")
> title("Pressao versus tratamento anterior para hipertensao")
```

■ **Example 2.5 — Diga-me seu nome que direi sua idade.** A Figura 2.11 usa o boxplot para mostrar a distribuição de idades de mulheres americanas em 2014 de acordo com o seu nome. A imagem vem do site FiveThirtyEighth, <http://53eig.ht/2mHEmAE>. Os gráficos são feitos com dados do *Social Security Administration* americano, que registra os nomes de batismo nos EUA desde 1880 (ver <https://www.ssa.gov/oact/babynames/>). As idades são lidas na primeira horizontal no alto da imagem. Os 25 nomes femininos mais comuns formam as linhas do gráfico. Em cada nome é mostrada apenas a caixa do boxplot (os limites interquartílicos Q_1 e Q_3) e a mediana dentro da caixa. Mostrando apenas a caixa para cada nome, sem os bigodes e outliers, podemos nos concentrar nas idades que compõem os 50% centrais da distribuição de idade de cada nome e observar algumas características interessantes.

Primeiro, nomes possuem histórias: eles nascem e morrem no interesse e no gosto da população. Os nomes no gráfico da esquerda na Figura 2.11 estão ordenados de cima para baixo de acordo com a idade mediana. As mulheres com os nomes mais no alto da imagem tendem a ser bem mais jovens que as mulheres usando os nomes da parte de baixo da imagem. A idade mediana da *Emily* é aproximadamente 17 anos enquanto que as *Dorothy* vivas em 2014 possuem idade mediana igual a 75 anos. A faixa central contendo 50% das mulheres com nome *Emily* vai de 10 anos a 26 anos, aproximadamente, enquanto as *Dorothy* variam entre 63 e 80 anos de idade. Não há como escapar da constatação de que *Dorothy* foi um nome popular no passado enquanto *Emily* é uma das preferidas há 10 anos atrás.

Exceto por alguns nomes, as caixas possuem comprimentos variando de 10 a 20 anos. Assim, para a maioria dos nomes mais populares, o auge de sua popularidade dura de 10 a 20 anos. As exceções claras são *Anna* e *Elizabeth*, nomes que possuem uma popularidade longeva. *Anna* foi popular no passado assim como é popular hoje.

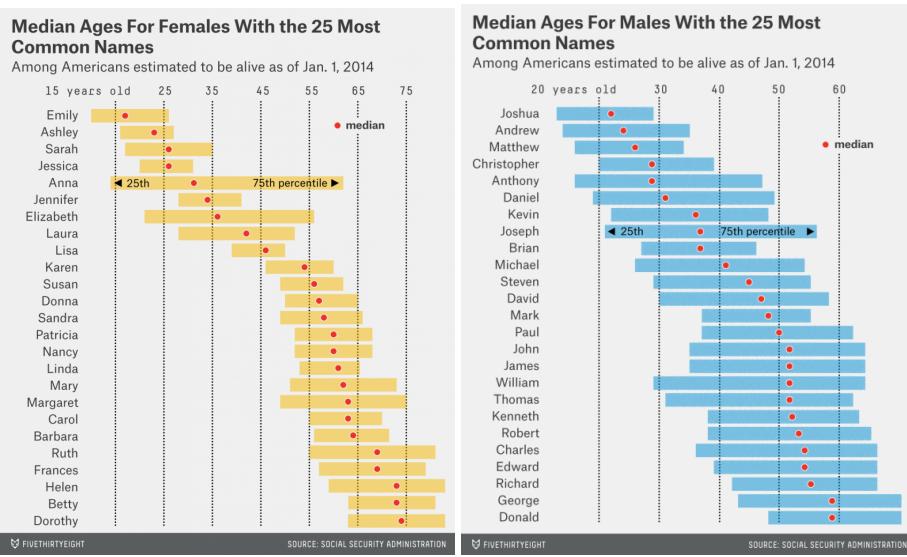


Figure 2.11: Boxplot das idades em 2014 das mulheres (esquerda) e homens (direita) que possuíam um dos 25 nomes femininos ou masculinos mais populares nos EUA. Fonte: site *FiveThirtyEight*, <http://fivethirtyeight.com/2014/01/the-most-common-first-names-in-the-u-s/>.

O mesmo gráfico para os homens está à direita na Figura 2.11. Eles contam uma história com muito menos dinamismo. Os nomes masculinos parecem oscilar menos no gosto das pessoas ao longo do tempo. Joseph, por exemplo, é um dos nomes americanos mais duradouros, nunca tendo saído de moda. Portanto, saber que um homem se chama Joseph não ajuda muito para adivinhar sua idade. A idade mediana dos Joseph que estão vivos em 2014 é 37 anos, e 50% deles se espalham entre 21 e 56 anos, uma larga faixa.

Os boxplots da Figura 2.11 são baseados nas idades das mulheres que estão vivas em 2014. Eles não podem dizer nada sobre os nomes das mulheres que já faleceram. Por exemplo, a extensão da faixa de idade das mulheres que carregam o nome Anna dá a impressão que ele tem uma popularidade constante no tempo. Isto não é verdade. A Figura 2.12 é um belíssimo gráfico mostrando a história dos nomes Anna, Joseph e Brittany nos EUA ao longo do tempo. A curva sólida em cada gráfico mostra o número de pessoas que receberam esses nomes em cada ano. As barras verticais vermelhas representam um histograma. Considere, por exemplo, o caso das Annas. Pegue todas as mulheres que se chamam Anna e estão vivas em 2014. Para cada uma, obtenha seu ano de nascimento. A seguir, faça um histograma dessa variável ano de nascimento. Por exemplo, em 2014, existem aproximadamente 5K Annas vivas nos EUA. Observe que a curva sólida fica praticamente igual ao histograma nos anos mais recentes: praticamente todas as Annas nascidas recentemente ainda estão vivas e portanto a curva e a altura do histograma coincidem praticamente.

Vemos que o nome Anna diminuiu substancialmente sua popularidade de 1900 a 1950. O número de novas Annas adicionadas à população em cada ano passou de 40K em 1900 para aproximadamente 5K em 1950. A maioria das Annas nascidas nas primeiras décadas já faleceram mas o nome permaneceu mais ou menos popular permitindo que um quarto das Annas vivas em 2014 tenham menos de 14 anos (nascidas a partir de 2000). Já o nome Joseph, relativamente a Anna, oscila mas acaba mostrando uma grande estabilidade a longo prazo. Estes dois nomes sempre populares são completamente diferentes do nome Brittany que praticamente nasce em 1970, tem seu auge em 1990 e hoje já não escolhido por quase ninguém. Como esta história é recente, as Brittanys estão praticamente todas vivas em 2014 e assim a curva sólida preta praticamente coincide com o histograma ao longo do tempo.

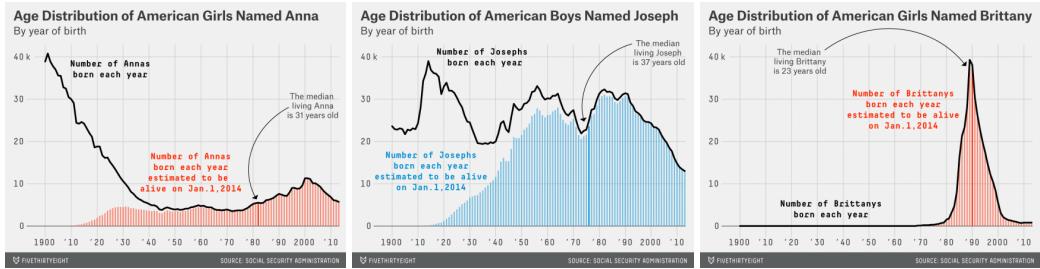


Figure 2.12: Annas, Josephs e Brittany's ao longo do tempo e em 2014. Fonte: site *FiveThirtyEighth*, <http://53eig.ht/2mHEmAe>.

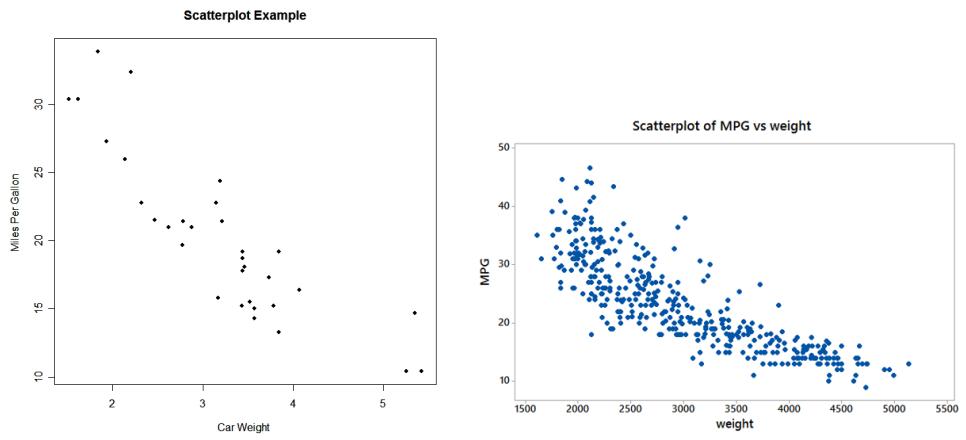


Figure 2.13: Gráfico de dispersão. Cada ponto é um modelo de automóvel. O eixo horizontal mostra seu peso, o eixo vertical mostra o seu desempenho em termos de milhas percorridas por galão de gasolina consumido. O gráfico da direita possui mais modelos de carros que o da esquerda e permite visualizar melhor a relação entre peso e desempenho.

2.6.4 Scatterplot

Scatterplot, ou gráfico de dispersão de pontos ou ainda gráfico de nuvem de pontos, é o campeão dos gráficos estatísticos. Serve para visualizar a relação entre duas variáveis numéricas. O scatterplot mais simples é obtido com o comando `plot(x, y)` em R. A Figura 2.6.4 foi obtida com os seguintes comandos:

```
# Simple Scatterplot, código do site Quick-R
attach(mtcars)
plot(wt, mpg, main="Scatterplot Example",
     xlab="Car Weight ", ylab="Miles Per Gallon ", pch=19)
```

Cada ponto é uma linha da matriz de dados, neste caso reduzida simplesmente às duas colunas `wt` (ou peso do carro) e `mpg` (ou milhas por galão). O gráfico da direita é o mesmo do da esquerda mas com mais automóveis. Vemos uma relação negativa ou inversa entre as variáveis: quando um carro tem seu peso `wt` muito acima do peso médio, seu desempenho `mpg` costuma ser baixo. Vice-versa, quando `wt` é muito baixo, `mpg` tende a ser alto. Isto é o esperado. Em geral, os carros muito pesados precisam queimar mais gasolina para movimentar-se.

A Figura 2.6.4 mostra um desenho esquemático de 7 gráficos de pontos mostrando diferentes graus de associações entre as variáveis x e y . No canto esquerdo temos associações positivas, começando com uma extremamente forte e então diminuindo a força da associação até o gráfico



Figure 2.14: Desenho esquemático de 7 gráficos de pontos mostrando associações lineares entre as variáveis x e y variando de extremamente forte e positiva (esquerda) para extremamente forte e negativa (direita) passando pela completa ausência de associação (gráfico central). Fonte: extraído da wikipedia, artigo *Pearson correlation coefficient*.

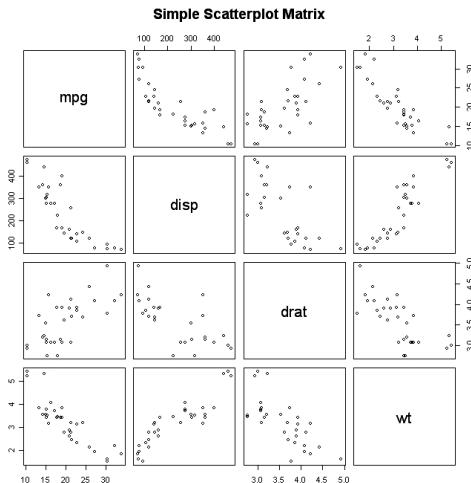


Figure 2.15: Matriz de scatterplots de 4 variáveis de um dataframe usando o comando `pairs()`.

central, que é um exemplo de completa ausência de associação entre x e y . A partir daí, a associação passa a ser negativa e chega a um extremo no canto direito.

Podemos visualizar vários scatterplots simultaneamente com uma matriz de scatterplots. A Figura 2.6.4 mostra uma matriz de scatterplots com quatro variáveis do dataframe `mtcars`. Em cada gráfico, podemos ver que as duas variáveis envolvidas estão associadas positivamente ou negativamente, algumas mais, outras menos fortemente. Basta usar o comando `pairs(~ var1 + var2 + var3, data = nome.do.dataframe)` para exibir uma matriz com as variáveis `var1`, `var2` e `var3` de certo dataframe.

```
# Basic Scatterplot Matrix, código do site Quick-R
pairs(~ mpg + disp + drat + wt, data=mtcars,
      main="Simple Scatterplot Matrix")
```

Na posição (i, j) da matriz de scatterplots encontramos o gráfico de dispersão das variáveis identificadas pelo nome nas posições i e j da diagonal. Por exemplo, no gráfico da posição $(2, 4)$ da Figura 2.6.4 temos o scatterplot da variável 2 (variável `disp`) no eixo vertical e a variável 4 (variável `wt`) no eixo horizontal. Como uma mesma variável é cruzada com todas as demais, para não repetir a escala numérica e assim economizar espaço na saída gráfica, as escalas de cada variável são lidas nas margens mais externas da matriz. Por exemplo, no caso do gráfico $(2, 4)$ a escala vertical da variável `disp` está na margem do gráfico $(2, 1)$ e a escala horizontal da variável `wt` está no topo do gráfico $(1, 4)$.

A situação ideal, do ponto de vista de facilidade de entendimento, é aquela em que a relação entre x e y é de crescimento ou decrescimento aproximadamente linear, como nos exemplos

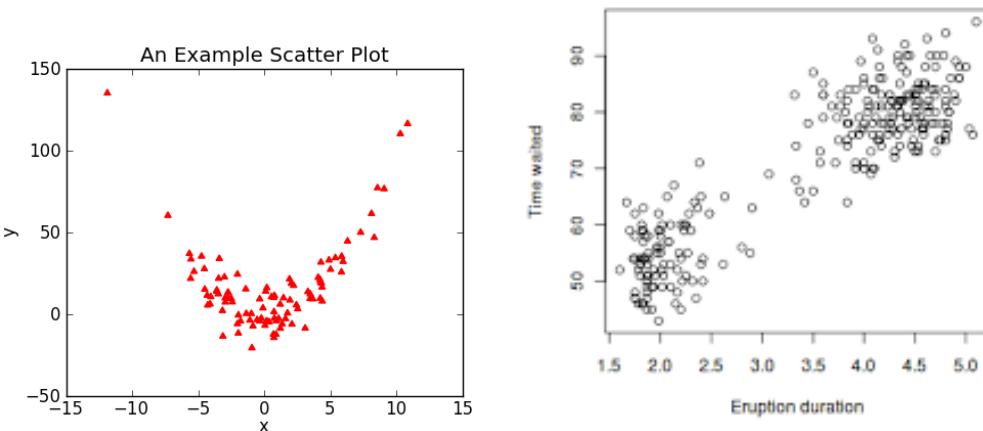


Figure 2.16: Scatterplots com relações mais complexas entre as variáveis.

anteriores. Ou então quando as variáveis possuem pouca associação entre si, como no caso do gráfico central da Figura 2.6.4. Nestes casos ideais (do ponto de vista de facilidade de entendimento da relação), a nuvem de pontos toma uma forma mais ou menos elíptica com o eixo maior da elipse ao longo da linha reta que representa grosseiramente a relação entre x e y , como nos gráficos da Figura 2.6.4.

Nela, cada ponto é um mês onde foram medidas três variáveis numa certa cidade grande: a taxa de mortalidade cardiovascular, a temperatura média no período (em graus Farenheit) e um índice de poluição do ar. Mortalidade parece ter uma associação positiva com a quantidade de partículas em suspensão (mais partícula, mais mortes) e negativa com temperatura (mais quente, menos morte). Na verdade, parece haver uma leve indicação de que talvez com temperatura muito altas a mortalidade recomece a crescer. O gráfico de temperatura versus poluição não mostra nenhuma associação entre as variáveis.

Existem diversas medidas quantitativas do grau de associação entre variáveis tais como correlação linear de Pearson, de Spearman, de Kendall, a informação mútua e o coeficiente de informação maximal. Entretanto, estas medidas são explicadas mais facilmente depois de aprendermos distribuição conjunta de variáveis aleatórias no capítulo 12 e a matriz de correlação no capítulo 13. Por enquanto, vamos julgar o grau de associação de forma subjetiva e com base apenas na visualização dos scatterplots.

Entretanto, a relação entre as variáveis pode ser mais complexa, exigindo mais explicação. Veja os scatterplots da Figura 2.16. O da esquerda mostra y tendo uma relação inicial de decrescimento com o aumento de x e então revertendo para uma relação de crescimento a partir de certo valor de x . O da direita usa dados de um geiser num parque dos EUA que entra em erupção de forma mais ou menos regular. Este geiser é chamado Old Faithful e está localizado às margens de um lago de águas incrivelmente azuis no parque Yellowstone, o parque do Zé Colmeia (https://en.wikipedia.org/wiki/Old_Faithful).

O gráfico mostra o tempo de duração da erupção de um geiser nos EUA (eixo x) versus o tempo de espera para que aquela erupção acontecesse (eixo y). Este tempo de espera começa a ser contado a partir do fim da erupção precedente. Existem duas regiões mais densas com dados. Isto indica a existência de dois regimes de erupção. Olhando a projeção dos pontos ao longo do eixo horizontal, vemos que aproximadamente 70% das erupções tiveram uma duração longa, em torno de 4.5 minutos, enquanto as restantes foram mais curtas, durando em torno de 2.0 minutos. O tempo de espera acompanha de forma positiva ou direta. Para observar as erupções mais longas, os

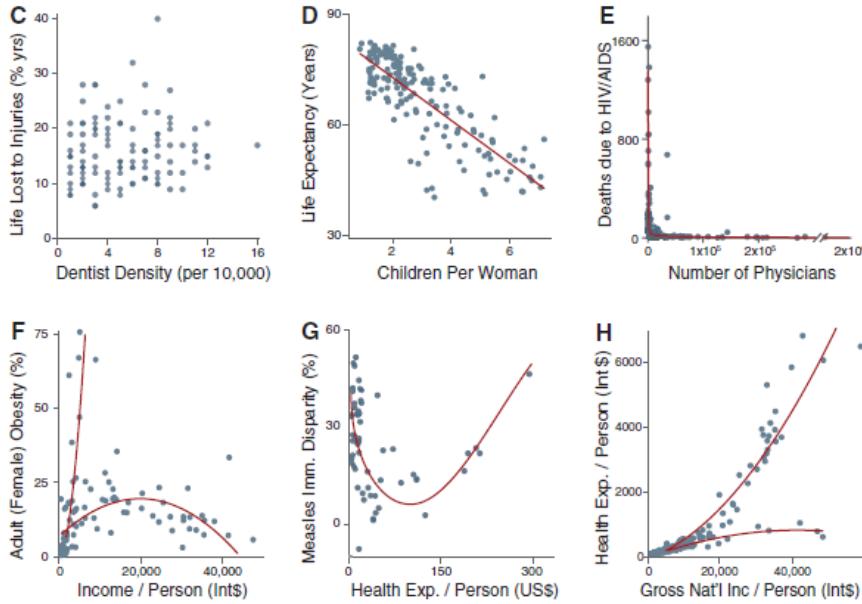


Figure 2.17: Scatterplots onde cada ponto é um país e as variáveis são indicadores sociais ou de saúde coletados pela Organização Mundial de Saúde (OMS). Fonte: [19].

turistas tiveram de esperar por volta de 80 minutos enquanto que ver as erupções mais curtas eles precisaram esperar 55 minutos, em média. Sem nenhum conhecimento do mecanismo envolvido nessas erupções, imagino que uma espera muito longa leva a um acúmulo grande de gases que, para ser liberado, requer uma erupção de maior duração.

Até agora mostramos scatterplots com relações entre x e y relativamente fáceis de se interpretar e entender. Nem sempre assim. Às vezes, as nuvens de pontos se parecem com as nuvens passageiras que assumem formas muito estranhas, mal comportadas do ponto de vista da interpretação e entendimento. Considere, por exemplo, a Figura 2.17 retirada de [19]. Cada gráfico cruza variáveis que são indicadores sociais, econômicos, de saúde e de política. Os itens (ou pontos) são países do mundo e os dados vieram da Organização Mundial de Saúde (OMS).

Os gráficos C e D na primeira linha da Figura 2.17 são do tipo usual, que temos visto até aqui. O primeiro deles cruza o número de dentistas por cada 10 mil habitantes do país com a porcentagem de anos de vida que são perdidos devido a lesões. A nuvem de pontos mostra que existe muito pouca ou nenhuma associação entre x e y neste caso. O segundo gráfico exibe uma clara tendência de decrescimento linear entre x , o número médio de filhos que uma mulher tem ao longo de sua vida reprodutiva, e y , a expectativa de vida (em anos) ao nascer. Claramente, países em que o número de filhos por mulher é alto são também os países que tendem a ter uma expectativa de vida menor que os demais.

Já os demais gráficos da Figura 2.17 são um pouco mais complicados de analisar. O gráfico E possui a imensa maioria dos seus pontos-países concentrados em torno da origem $(0,0)$ dificultando um entendimento melhor do que acontece com a maioria dos países. Apesar disso, podemos observar que um número elevado de mortes por HIV/AIDS só acontece em países com poucos médicos enquanto, ao mesmo tempo, países com muitos médicos tem muito poucas ou zero mortes por HIV/AIDS. Os gráficos da segunda linha de plots têm uma linha vermelha sobreposta para indicar a relação entre x e y . Os gráficos F e G mostram uma tendência não-linear, aproximadamente parabólica, entre x e y exceto que, em F , um pequeno grupo de países parece escapar desta relação geral criando uma tendência linear entre x e y . O gráfico em H também mostra esta existência

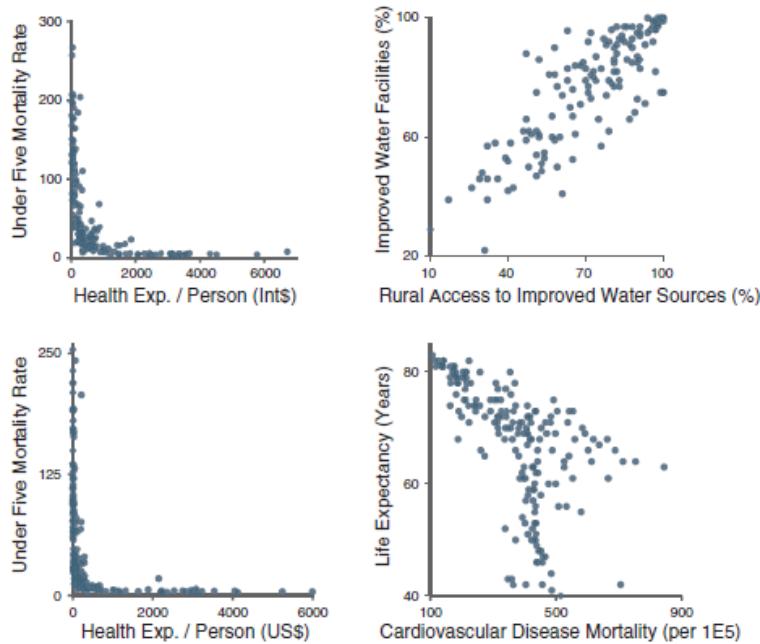


Figure 2.18: Mais scatterplots com os dados de indicadores sociais de países coletados pela Organização Mundial de Saúde (OMS). Fonte: [19].

de uma relação de associação positiva para a maioria dos países entre x , o PIB *per capita*, e y , gasto em saúde *per capita*. Entretanto, alguns poucos países parecem seguir outra tendência onde, apesar de bem ricos em termos *per capita*, mostram uma saturação no gasto em saúde num nível relativamente baixo.

A Figura 2.18 mostra mais scatterplots vindos do mesmo artigo [19], como na Figura 2.17. São mais gráficos mostrando quão diversa e complicada ou quão direta e simples pode ser a relação estatística entre duas variáveis.

Como vimos anteriormente com os boxplots, podemos também aqui introduzir mais uma variável para entender melhor a relação entre x e y . A Figura 2.19, também extraída de [19], usa dados de abundância de espécies de bactérias que colonizam o intestino de humanos e outros mamíferos. Camundongos foram utilizados neste experimento, sob dois tipos de dieta, uma com baixo teor de gordura e açúcar (LF/PP) e outra chamada de ocidental (*western*), com alto teor de gordura e açúcar. Eles tiveram seus intestinos colonizados com bactérias de amostras fecais humanas. Os gráficos da Figura 2.19 mostram os níveis de prevalência de diferentes pares de bactérias em cada eixo, cada ponto representando um camundongo do experimento.

Em todos eles, vemos um tipo de associação de não-coexistência entre as bactérias: quando uma espécie é abundante, a outra é menos abundante. Várias dessas associações de não-coexistência são parcialmente explicadas pela dieta, como no gráfico A da Figura 2.19. Sob a dieta LF/PP a espécie *Bacteroidaceae OTU* domina, enquanto que sob a dieta ocidental é a espécie *Erysipelotrichaceae* que domina. Em B, o sexo do camundongo adicionou um nível de explicação ao gráfico: fêmeas tinham apenas uma das bactérias. Em C, é uma terceira variável, associada com a origem humana da amostra fecal, que ajuda a explicar a ocorrência de pontos em diferentes regiões do gráfico.

2.6.5 Scatterplot 3-dim

Scatterplots tri-dimensionais, como os da Figura 2.6.5, são bonitos mas não são muito úteis para analisar dados pois é difícil visualizar exatamente onde estão os pontos. Ancorando os dados no plano $x - y$, como no lado direito da Figura 2.6.5, ajuda nesta tarefa mas, ainda assim, pessoalmente,

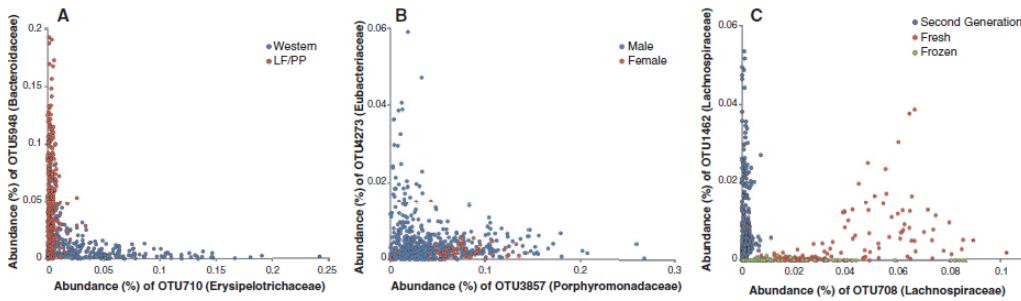
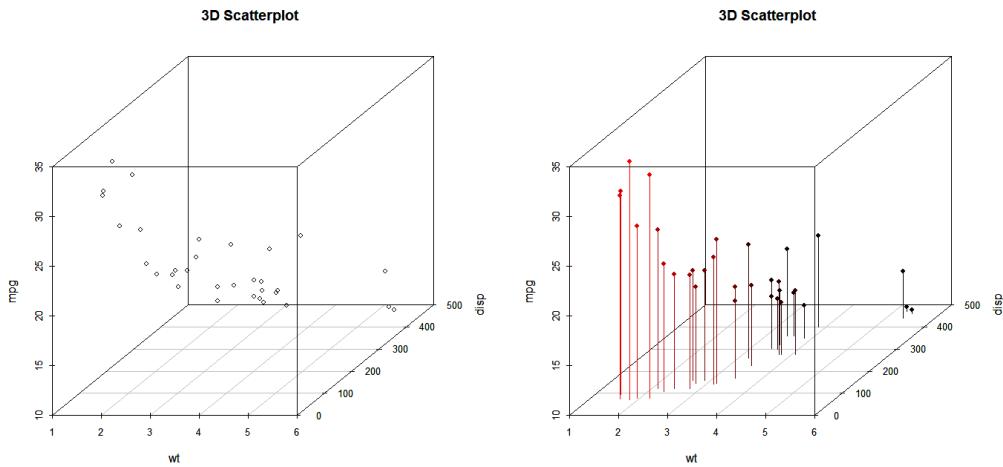


Figure 2.19: Abundância de diferentes espécies de bactérias comuns no intestino humano. Fonte: [19].



eu não acho estes gráficos muito úteis para análise.

```
# 3D Scatterplot, código do site Quick-R
library(scatterplot3d)
attach(mtcars)
scatterplot3d(wt, disp, mpg, main="3D Scatterplot")
#
# 3D Scatterplot with Coloring and Vertical Drop Lines
scatterplot3d(wt, disp, mpg, pch=16, highlight.3d=TRUE,
type="h", main="3D Scatterplot")
```

Existe uma library em R, rgl, que permite visualizar dinamicamente estas nuvens tri-dimensionais. Isto ajuda muito na visualização dos dados e eu gosto muito de usá-la quando analiso três variáveis simultaneamente. Como numa superfície bi-dimensional, como desta página, não é possível apreender a utilidade desta ferramenta, use o código abaixo no R para experimentar, após instalar a library rgl.

```
# Spinning 3d Scatterplot, código do site Quick-R
install.packages("rgl") # ou use a interface grafica no RStudio
library(rgl)
attach(mtcars)
plot3d(wt, disp, mpg, col="red", size=3)
```

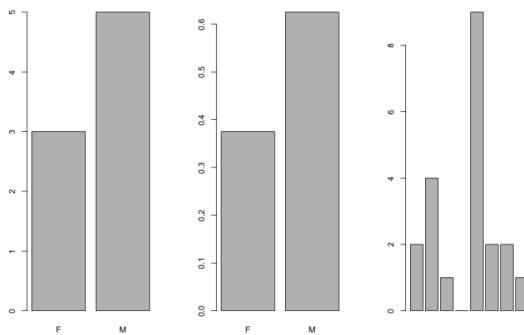


Figure 2.20: Barplot.

2.7 Vetores ou colunas para dados categóricos

Dados categóricos em R possuem algumas funções próprias.

```
> y = c("M", "F", "M", "M", "M", "F", "M", "F")           # vetor de caracteres

> sum(y)          # caracteres nao podem ser somados
Erro em sum(y) : 'type' invalido (character) do argumento

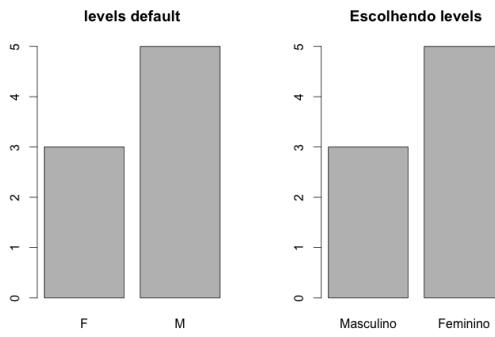
> table(y)        # eles podem ser tabelados
y
F M
3 5

> # valor de retorno de table() e' um vetor numerico de dimensao = numero de strings distintas
> yc = table(y)
> yc[2]      # o vetor tem nomes (os strings distintos) para as suas entradas
M
5

> yc/length(y)   # podemos operar numericamente
y
F      M
0.375 0.625

> # ver figura abaixo para o resultado destes comandos
> par(mfrow=c(1,3))  # janela grafica dividida em 1 x 3 celulas
> barplot(table(y), main="frequencias")      # grafico das contagens da tabela
> barplot(table(y)/length(y), main="proporcoes")    # plotando as proporcoes
> x = c(2, 4, 1, 0, 9, 2, 2, 1)
> barplot(x, main="barplot de vetor")
```

A questão não é apenas ser um vetor de caracteres. Vetores com números podem representar categorias. Por exemplo, poderíamos ter um vetor com os números 0 e 1 onde 0 representaria um caso “Masculino” e 1 representaria um caso “Feminino”. R tem uma classe de objetos para trabalhar com variáveis categóricas: `factor`. R adapta-se automaticamente em resposta dos comandos quando o objeto é um fator. Para criar um fator, use o comando `factor` ou `as.factor`.



```

> y = c("M", "F", "M", "M", "M", "F", "M", "F")
> y
[1] "M" "F" "M" "M" "M" "F" "M" "F"

> plot(y)
Erro em plot.window(...) : valores finitos sao necessarios para 'ylim'
Alem disso: Mensagens de aviso perdidas:
1: In xy.coords(x, y, xlabel, ylabel, log) : NAs introduzidos por coercao
2: In min(x) : nenhum argumento nao faltante para min; retornando Inf
3: In max(x) : nenhum argumento nao faltante para max; retornando -Inf

> y = factor(y)
> y
[1] M F M M M F M F
Levels: F M
> # armazena y como 3 1's e 5 2's e associa
> # 1="F" e 2="M" internamente (alfabeticamente)
> # y agora e' uma variavel nominal

> plot(y, main="levels default")

> levels(y) = c("M" = "Masculino", "F" = "Feminino")
> y
[1] Feminino Masculino Feminino Feminino Feminino Masculino Feminino Masculino
Levels: Masculino Feminino

> plot(y, main = "Escolhendo levels")

> summary(y)
Masculino  Feminino
      3          5

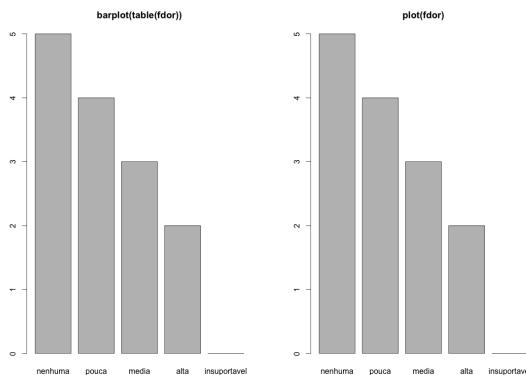
```

Não é só a questão de ser caracter para que um vetor seja uma variável categórica. Números podem representar categorias. Transformando em fator: escolha os níveis do fator.

```

> dor = c(0, 3, 5, 5, 1, 1, 0, 1, 0, 0, 1, 3, 3, 0)  # niveis de dor
> fdor = factor(dor, levels=c(0, 1, 3, 5, 1000))    # transformando num objeto tipo "factor"
> levels(fdor) = c("nenhuma", "pouca", "media", "alta", "insuportavel") # expressando os níveis

```



```
> fdor      # veja que nao existem caso de dor insuportavel
[1] nenhuma media alta alta pouca pouca nenhuma pouca nenhuma ...
Levels: nenhuma pouca media alta insuportavel
```

Os comandos acima armazenam fdor de forma que temos o seguinte mapeamento: 0 → 1, 1 → 2, 3 → 3, 5 → 4, e 1000 → 5. e associa 1 a nenhuma, 2 a pouca, 3 a media, 4 a alta, e 5 a insuportavel. O vetor fdor agora é um fator com estes níveis. Os comandos a seguir mostram como fazer visualizar dados categóricos armazenados em fatores.

```
> par(mfrow=c(1,2))
> barplot(fdor)
Erro em barplot.default(fdor) : 'height' deve ser um vetor ou uma matriz
> barplot(table(fdor), main="barplot(table(fdor))")
> plot(fdor, main="plot(fdor)")  # comando generico "plot" responde com "barplot" quando a
```

2.8 Objetos em R

Tipos de dados/objetos em R:

- scalars
- vetores: numerical, logical, character
- matrizes,
- data-frames,
- listas,
- funções.

2.8.1 Escalares

```
> x = -1.3
> x
[1] -1.3
> x = 2
> x
[1] 2
> x = pi
> x
[1] 3.141593
```

```
> x = "Pedro Paulo"
> x
[1] "Pedro Paulo"
```

2.8.2 Vetores

```
# VETORES NUMERICOS
> x = c(1, 4, -1, 4)
> x
[1] 1 4 -1 4
> 1:10
[1] 1 2 3 4 5 6 7 8 9 10
# VETORES LOGICOS
> x = c(T, T, F, T)
> x
[1] TRUE TRUE FALSE TRUE
> 1:6 > 2
[1] FALSE FALSE TRUE TRUE TRUE TRUE
> sum(1:6 > 2)      # transforma em numerico: T=1 e F=0
[1] 4
# VETORES DE CARACTERES
> x = c("Pedro Paulo", "Pedro", "P")
> x
[1] "Pedro Paulo" "Pedro"       "P"
> letters[1:6]
[1] "a" "b" "c" "d" "e" "f"
```

Uma função muito útil ao lidar com caracteres é `paste`:

```
> x = c("Pedro", "Paulo", "Pedro", "Manoel")

> paste(x, 1:4)
[1] "Pedro 1" "Paulo 2" "Pedro 3" "Manoel 4"

> paste(x, 1:4, sep = "")
[1] "Pedro1" "Paulo2" "Pedro3" "Manoel4"

> rep(paste("T", 1:3, sep = ""), c(4, 4, 3))
[1] "T1" "T1" "T1" "T1" "T2" "T2" "T2" "T2" "T3" "T3" "T3"
```

2.8.3 Matrizes

Matrizes são dados tabulares de um único tipo em toda a matriz: ou toda numérica, ou toda lógica, ou toda de caracteres. Se quiser ter dados de tipos diferentes precisa usar dataframes (a seguir) ou listas (mais a frente).

```
> x = matrix(1:6, ncol=3, byrow=T)
> x
[,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
```

```
> matrix(letters[1:6], ncol=3, byrow=T)
     [,1] [,2] [,3]
[1,] "a"  "b"  "c"
[2,] "d"  "e"  "f"

> cbind(1:3, 10:12) # concatena vetores como colunas de uma matriz
     [,1] [,2]
[1,]    1   10
[2,]    2   11
[3,]    3   12

> cbind(1:3, c("a", "b", "c")) # tipos diferentes sao coagidos a um tipo unico
     [,1] [,2]
[1,] "1"  "a"
[2,] "2"  "b"
[3,] "3"  "c"
```

Vamos ver o operador de seleção de elementos em uma matriz.

```
> x = matrix(1:12, ncol =4, byrow=T)
> x
     [,1] [,2] [,3] [,4]
[1,]    1    2    3    4
[2,]    5    6    7    8
[3,]    9   10   11   12

> x[2, 4]    # elemento (2,4)
[1] 8

> x[, 1:2]  # selecionado as duas 1as colunas
     [,1] [,2]
[1,]    1    2
[2,]    5    6
[3,]    9   10

> x[2:3 , 3:4] # sub-matriz bloco
     [,1] [,2]
[1,]    7    8
[2,]   11   12

> x[, c(1,3)] # seleciona colunas 1 e 3
     [,1] [,2]
[1,]    1    3
[2,]    5    7
[3,]    9   11
```

Vamos ver agora as principais operações matriciais.

```
> x = matrix(1:4, ncol=2, byrow=T)
> x + t(x) # x + sua transposta
```

```

[,1] [,2]
[1,]    2     5
[2,]    5     8

> x/x # operacao elemento a elemento
[,1] [,2]
[1,]    1     1
[2,]    1     1

> x %*% t(x) # multiplicacao matricial
[,1] [,2]
[1,]    5    11
[2,]   11    25

> solve(x) # inversa de x
[,1] [,2]
[1,] -2.0  1.0
[2,]  1.5 -0.5

```

Algumas operações para fazer decomposições matriciais:

```

> eigen(x)      # autovalores e autovetores, saida e' lista com 2 elementos
$values        # 1o elemento e' um vetor com os dois autovalores
[1]  5.3722813 -0.3722813

$vectors       # 2o. elemento e' uma matriz onde cada coluna e' um autovetor
[,1]      [,2]
[1,] -0.4159736 -0.8245648
[2,] -0.9093767  0.5657675

> svd(x)        # decomposicao do valor singular, saida e' lista
$d
[1] 5.4649857 0.3659662

$u
[,1]      [,2]
[1,] -0.4045536 -0.9145143
[2,] -0.9145143  0.4045536

$v
[,1]      [,2]
[1,] -0.5760484  0.8174156
[2,] -0.8174156 -0.5760484

```

2.8.4 Dataframes

Data frames: são dados tabulares como as matrizes, mas as colunas podem ter tipos diferentes.

```

> x = c(2, 3, 5)
> y = c("a", "bb", "ccc")

```

```

> z = c(TRUE, FALSE, TRUE)
> df = data.frame(x, y, z)
> df
   x   y   z
1 2  a  TRUE
2 3  bb FALSE
3 5 ccc TRUE

> df[1:2, 2:3]  # operador [...] funciona como em matriz
   y   z
1  a  TRUE
2 bb FALSE

> df$x  # acessando colunas nomeadas de data frames
[1] 2 3 5
> df$x[1:2]  # colunas sao como vetores
[1] 2 3

```

2.8.5 Listas

São estruturas genéricas para coletar objetos diversos num único objeto.

```

> x = 1:10
> y = c("a", "b", "c")
> z = matrix(1:4, ncol=2)
> w = list(x, y, z)
> w
[[1]]
[1] 1 2 3 4 5 6 7 8 9 10

[[2]]
[1] "a" "b" "c"

[[3]]
 [,1] [,2]
[1,]    1    3
[2,]    2    4

> w[[3]]      # acessando o 3º elemento da lista
 [,1] [,2]
[1,]    1    3
[2,]    2    4

```

É importante diferenciar [...] e [[...]] em listas:

- [...] retorna um ELEMENTO da lista: um vetor, uma matriz etc.
- [...] retorna uma sub-lista da lista original.

```

> w[c(1, 3)]      # sub-lista contendo apenas o 1º e 3º elementos de w
[[1]]
[1] 1 2 3 4 5 6 7 8 9 10

```

```

[[2]]
 [,1] [,2]
[1,]    1    3
[2,]    2    4

> is.list(w[c(1, 3)])    # testa se e' uma lista
[1] TRUE

> w[[2]]      # retorna o elemento 2 da lista
[1] "a" "b" "c"

> w[[c(2,3)]]  # retorna o 3o. elemento do elemento 2 da lista w
[1] "c"

```

2.8.6 Funções

A linguagem R permite extensões com a criação de funções. A estrutura geral para criação de uma função é a seguinte:

```

myfun = function(arg1, arg2,...)
{
  ....corpo da funcao...
  return(x)  # x e' qualquer objeto mas, em geral, e' uma lista
}

```

R possui as estruturas de controle usuais: `for`, `while`, `if`, `if else`. Permite também chamar funções em C, C++, FORTRAN, java etc. Um exemplo simples de função:

```

myfun <- function(x1, x2) {
  pint = sum(x1 * x2) # produto interno dos vetores
  s1 = sqrt(sum(x1*x1)) # comprimento do vetor 1
  s2 = sqrt(sum(x2*x2)) # comprimento do vetor 2
  z = pint/(s1*s2)
  x = list(prod.int = pint, coseno = z, dados = cbind(x1, x2))
  return(x)
}

myfun # imprime na tela a definicao da funcao

# aplicando myfun a c(1,2,3) e c(3, 4, 7)
myfun(x1=c(1,2,3), x2=c(3, 4, 7))

$prod.int
[1] 32

$coseno
[1] 0.9941916

$dados
  x1 x2
[1,]  1  3

```

```
[2,] 2 4
[3,] 3 7
```

2.8.7 Votorizar sempre que possível

Vetorizar significa transformar loops em operações vetoriais. O código R fica muito mais eficiente.

```
> x = runif(100000) # 100 mil numeros aleatorios em x
> y = runif(100000) # idem em y
> z = numeric()      # criando objeto numerico z

> start <- Sys.time()
> for(i in 1:100000){ z[i] = x[i] + y[i] }
> end <- Sys.time()
> end - start
Time difference of 23.09923 secs

> start <- Sys.time()
> z = x + y
> end <- Sys.time()
> end - start
Time difference of 1.352752 secs
```

Este próximo exemplo foi copiado de <http://www.r-bloggers.com/how-to-use-vectorization-to-streamline-your-r-code/>. A tarefa é escrever um programa que jogue uma moeda honesta n vezes. A cada 100 lançamentos, imprima a proporção de caras menos 1/2. Imprima também o número de caras menos a metade do número de lançamentos até o momento. Vamos escrever um programa em R com “sabor c”(cheio de loops, sem vetorizar).

```
coin_toss1 = function(n){
  result = c()
  for(i in c(1:n)) {
    if(i == 1){
      ## the optional outputs are 0 and 1. I am assigning 1 to heads
      tosses = sample(c(0,1),1)
    }
    else{
      ## creating a vector that has history of all tosses
      tosses = c(tosses,sample(c(0,1),1))
    }
    ## when we reach a toss number that a multiple of 100 we output the status
    if(i %% 100 == 0){
      ## output the percent of heads away from 50%
      percent = (sum(tosses) / length(tosses)) - 0.5
      ## output the number of heads away from half of all tosses
      number = sum(tosses) - (length(tosses) / 2)
      result = rbind(result, c(percent, number))
    }
  }
  result
}
```

Agora outro código, com sabor R, onde os loops foram vetorizados:

```
coin_toss2 = function(n, step=100) {
  # Record number of heads at each step
  tosses = cumsum(sample(c(0,1), n, replace=TRUE))
  # Define steps for summaries
  steps = seq(step,n, by=step)
  # Compute summaries
  cbind(tosses[steps] / steps - .5, tosses[steps] - steps/2)
}
```

Vamos agora comparar a eficiência dos dois códigos:

```
> start <- Sys.time()
> x = coin_toss1(100000)
> end <- Sys.time()
> end - start
Time difference of 24.23292 secs

> start <- Sys.time()
> x = coin_toss2(100000)
> end <- Sys.time()
> end - start
Time difference of 1.098358 secs
```

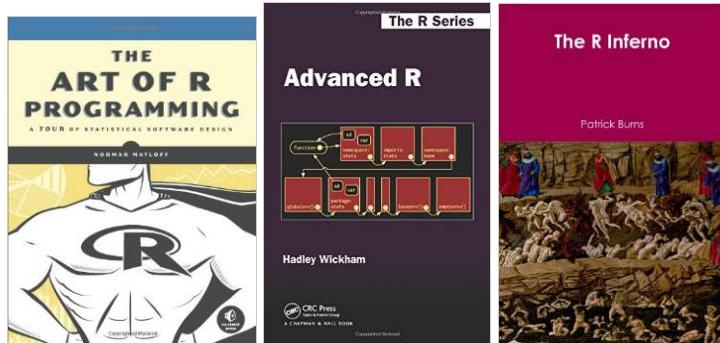
Ver capítulos 3 e 4 do livro *The R Inferno*, free pdf em <http://www.burns-stat.com/documents/books/the-r-inferno/> Abstract: If you are using R and you think you're in hell, this is a map for you. Even if it doesn't help you with your problem, it might amuse you (and hence distract you from your sorrow).

2.8.8 Comando apply

Operar repetidamente sobre as colunas ou linhas de uma matriz ou dataframe é uma operação tão comum que tem um comando especial em R: `apply`. A sintaxe mais simples de uso deste comando é: `apply(mat, index, FUN)`. Ele aplica a função `FUN` na matriz `mat` ao longo das suas linhas ou colunas: se `index=1`, aplica `FUN` em cada linha da matriz `mat`. Se `index=2`, aplica `FUN` nas colunas. Exemplo:

```
> mat = matrix(1:12, ncol =4)
> mat
     [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
> apply(mat, 2, sum)  # valor de retorno e' um vetor de dimensao = no de cols de mat
[1]  6 15 24 33
```

O comando `lapply` opera em listas ao invés de matrizes. Ver também `tapply`, `mapply`, `rapply`, `eapply`.



2.9 Material inicial sobre R

- Comece lendo na wikipedia: [http://en.wikipedia.org/wiki/R_\(programming_language\)](http://en.wikipedia.org/wiki/R_(programming_language))
- Download R para Linux, Mac, Windows em <http://cran.r-project.org/>
- Rstudio, free IDE para R: <http://www.rstudio.com/>
- Muitos tutoriais disponíveis no CRAN e na web. Escolha o seu sabor...
- R-tutorial: excelente - <http://www.r-tutor.com/>
- Outro: <http://mran.revolutionanalytics.com/documents/getting-started/>
- Em português: <http://www.leg.ufpr.br/~paulojus/embrapa/Rembrapa/>
- Quick-R: <http://www.statmethods.net/>
- Curso: <http://www.pitt.edu/~njc23/> (apenas os slides, excelente)
- Lista brasileira de discussão do R: <http://www.leg.ufpr.br/doku.php/software:rbr>

2.9.1 Material mais avançado em R

- An Introduction to R: <http://cran.r-project.org/doc/manuals/r-release/R-intro.html>
- Livro The Art of R Programming, de Norman Matloff Disponível no site <https://www.safaribooksonline.com/library/view/the-art-of/9781593273842/>
- Livro Advanced R, de Hadley Wickham
- R-bloggers: <http://www.r-bloggers.com/>
- Livro The R Inferno, free pdf: <http://cran.r-project.org/doc/manuals/r-release/R-intro.html> Free pdf em <http://www.burns-stat.com/documents/books/the-r-inferno/>
- Tem também o tutorial Impatient R em <http://www.burns-stat.com/documents/tutorials/impatient-r/>

2.9.2 Cursos online gratuitos

- Lista de cursos online de estatística: <https://www.class-central.com/search?q=statistics>
- Coursera: Data Analysis and Statistical Inference, começando em Março, 2015 <https://www.coursera.org/course/statistics>
- Specialization in Data Science em Coursera: nove cursos. Um dos nove cursos: R Programming <https://www.coursera.org/course/rprog>
- O campeão dos MOOCs: Machine Learning , com Andrew Ng, Na plataforma coursera: <https://www.coursera.org/course/ml>
- Statistical Learning : Stanford professors Trevor Hastie and Rob Tibshirani <https://www.youtube.com/channel/UC40WDcPB1peiBXDfCSZ3h-w/feed>



3. Probabilidade Básica

3.1 Espaço de probabilidade

Vamos lidar com fenômenos não determinísticos, probabilísticos, aleatórios. O modelo matemático para **qualquer** fenômeno probabilístico é o espaço de probabilidade. Espaço de probabilidade é uma 3-upla constituída por três elementos que satisfazem os três axiomas de Kolmogorov (1903-1987).

Definition 3.1.1 — Espaço de probabilidade. Um espaço amostral é uma 3-upla $(\Omega, \mathcal{A}, \mathbb{P})$ onde:

Ω é um conjunto com **todos** os resultados possíveis do fenômeno.

\mathcal{A} é uma σ -álgebra de sub-conjuntos $A \subseteq \Omega$, os sub-conjuntos aos quais vamos atribuir probabilidades.

\mathbb{P} é uma função matemática atribuindo probabilidades aos sub-conjuntos de \mathcal{A} :

$$\mathbb{P} : \mathcal{A} \longrightarrow [0, 1]$$

$$A \longrightarrow \mathbb{P}(A)$$

Esta função \mathbb{P} deve satisfazer aos três axiomas de Kolmogorov.

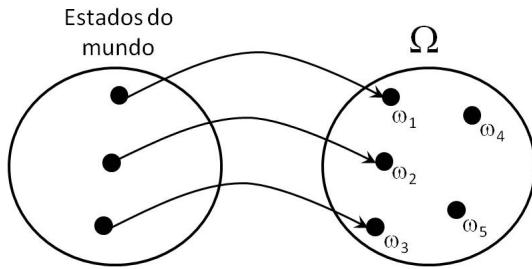
Vamos ver cada um desses elementos com mais detalhes a seguir. Depois de entender exatamente o que eles querem dizer, precisamos aprender o que são os três axiomas de Kolmogorov que eles devem satisfazer e por quê, intuitivamente, eles devem satisfazê-los.

3.2 O espaço amostral Ω

Ω é um conjunto representando **todos** os resultados possíveis do fenômeno. Em inteligência artificial, falamos de todos os possíveis “estados do mundo” ao invés de resultados possíveis. Cada resultado possível deve ser completamente especificado e único em Ω . Não pode haver dois elementos em Ω representando o mesmo resultado possível. Ver Figura 3.1.

O “mundo” representado em Ω é limitado. Diz respeito ao mundo que você observa ou estuda no momento. A todo estado do mundo corresponde um, e somente um, elemento $\omega \in \Omega$.

Um ponto curioso é que Ω pode ter mais elementos que estados do mundo. Isto é, pode ter

Figure 3.1: O espaço amostral Ω

elementos que representam resultados impossíveis. Veremos a utilidade prática disso daqui a pouco. Antes, vamos ver alguns exemplos de Ω .

3.2.1 Exemplos de Ω

■ **Example 3.1 — O exemplo canônico: lançar a moeda uma vez.** Observa-se o lançamento de uma moeda. Neste caso, podemos tomar

- $\Omega = \{\text{cara, coroa}\}$
- ou $\Omega = \{c, \tilde{c}\}$
- ou $\Omega = \{0, 1\}$
- ou $\Omega = \{T, F\}$

Qualquer uma dessas opções é válida, dois símbolos para representar os dois resultados possíveis do experimento probabilístico de jogar uma moeda. ■

■ **Example 3.2 — Mais moedas.** Observa-se três lançamentos sucessivos de uma moeda. Então

$$\Omega = \{ccc, cc\tilde{c}, c\tilde{c}c, \dots, \tilde{c}\tilde{c}\tilde{c}\}$$

Ω tem 8 elementos. O mundo deste segundo exemplo é mais amplo que aquele do primeiro observador-exemplo. Neste novo mundo podemos calcular a probabilidade do segundo lançamento da moeda ser *cara*. No mundo do primeiro observador não podemos calcular a probabilidades referentes ao segundo ou terceiro lançamentos da moeda pois eles não pertencem ao Ω daquele mundo. ■

■ **Example 3.3 — Links da Web.** Observa-se o número de out-links de uma página da web escolhida ao acaso.

$$\Omega = \{0, 1, 2, 3, \dots\} = \mathbb{N}$$

Ω é o conjunto infinito de números naturais. Isto é estranho pois em qualquer instante existe um número máximo de páginas da web. Faz sentido deixar que as páginas tenham qualquer número de links, um número infinito de possibilidades? Faz sentido? Queremos deixar que uma página tenha a possibilidade de ter qualquer número de links. Um número máximo de links deve existir mas é desconhecido. Além disso, número máximo muda com o tempo.

Poderíamos colocar um número máximo que claramente deve superar o o máximo (digamos, um bilhão de links) mas, curiosamente isto tornar a manipulação matemática muito mais difícil e menos produtiva do que se assumirmos que $\Omega = \mathbb{N}$. A razão é que temos vários modelos de probabilidade quando $\Omega = \mathbb{N}$ mas não quando $\Omega = \{0, 1, 2, 3, \dots, 10^9\}$. Estes modelos com $\Omega = \mathbb{N}$ são bem estudados e sabemos extrair deles muita informação útil.

Assim, usamos \mathbb{N} pela conveniência matemática de se trabalhar com distribuições de probabilidade definidas sobre \mathbb{N} . Vamos atribuir uma probabilidade estritamente maior que zero a cada um dos infinitos elementos de Ω . Nós “corrigimos” o excesso de elementos em Ω atribuindo

probabilidades muito próximas de zero aos elementos de \mathbb{N} que são números grandes demais, absurdamente grandes para representarem in-links. Uma distribuição de probabilidade power-law (lei de potência) será uma das maneiras de fazer isto.

Adiantando o que veremos mais tarde, poderíamos fazer:

$$\mathbb{P}(\omega = k) = C \frac{1}{(k+1)^2}$$

para $k = 0, 1, 2, \dots$ onde $C = 6/\pi^2$ (a constante foi obtida por Euler (Basel problem)). Neste caso, teremos

- $\mathbb{P}(\omega = 100) \approx 6.110^{-5}$
- $\mathbb{P}(\omega = 10000) \approx 6.110^{-9}$
- e probabilidades menores ainda para números maiores que estes.

■ **Example 3.4 — In-links e Out-links.** O experimento consiste em observar o número de in-links e out-links de n páginas da web. Podemos então definir

$$\Omega = \{(i_1, o_1, i_2, o_2, \dots, i_n, o_n) \in \mathbb{N}^{2n}\}$$

■ **Example 3.5 — Itens num supermercado.** Um supermercado possui 15000 produtos. Registra-se o número de itens de cada produto comprado por um cliente. Então

$$\Omega = \{(n_1, n_2, \dots, n_{15000}) \in \mathbb{N}^{15000}\}$$

■ **Example 3.6 — Espaço de Grafos.** Temos um conjunto finito V com n vértices. O interesse reside nas relações não direcionais entre os vértices. Quem conecta com quem? Podemos definir

$$\Omega = \{\text{Grafos não-direcionais em } V\},$$

um conjunto com 2^p elementos, onde $p = n(n-1)/2$, o número de pares não-ordenados de vértices.

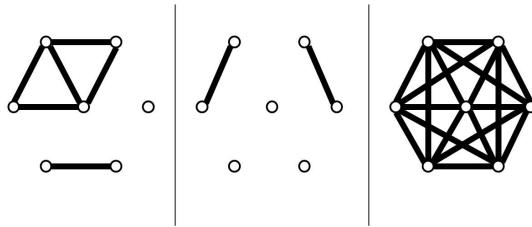


Figure 3.2: Três estados do mundo do conjunto de grafos, três elementos de Ω .

■ **Example 3.7 — Espaço de imagens .** Uma imagem com 512×512 pixels, cada pixel tem um tom de cinza. O tom de cinza de cada pixel é codificado com um inteiro entre 0 e 255. 8 bits, $2^8 = 256$ tons possíveis: 0 é preto e 255 é branco.

Ω é o conjunto de todas as matrizes M de dimensão 512×512 e com $M(i, j) \in \{0, 1, \dots, 255\}$. Assim, Ω é um conjunto finito mas com um número gigantesco de elementos. De fato, a cardinalidade do conjunto Ω é igual a 256^{512^2} . Dois elementos de Ω são duas imagens 512×512 em tons de cinza.

Dois exemplos estão na Figura 3.7. A imagem da esquerda é uma imagem “estruturada”, possuindo vários objetos. A imagem da direita é uma em que cada pixel é um número aleatório entre 0 e 255, completamente desestruturada. Como atribuir probabilidades neste conjunto? Vamos aprender mais tarde a atribuir probabilidades de forma que algumas imagens sejam mais prováveis que outras.



■ **Example 3.8 — Outro exemplo canônico: escolher um número real ao acaso.** Imagine o experimento de escolher um número real $x \in [0, 1]$ ao acaso. Este experimento não pode de fato ser feito no computador onde os números possuem representação binária (e decimal) finita. Entretanto, podemos imaginar conceitualmente este experimento. Neste caso, $\Omega = [0, 1]$. Teremos mais a falar sobre este exemplo quando quisermos entender melhor a atribuição de probabilidades em espaços contínuos ou não-enumeráveis. ■

■ **Example 3.9 — Altura.** Selecionar um habitante adulto de Belo Horizonte ao acaso e medir sua altura em metros. $\Omega = (1.30, 3.0)$ é uma boa escolha? Quem sabe $(1.0, 4.0)$? Ou então $\Omega = (0, \infty)$? Ou ainda $\Omega = (-\infty, \infty) = \mathbb{R}$

Para este exemplo, na prática, $\Omega = (0, \infty)$ ou $\Omega = \mathbb{R}$ são as escolhas preferidas embora obviamente não exista altura negativa ou altura maior que, digamos, 5 metros. Como no caso do número de links numa página da web, estamos criando um conjunto Ω que, com certeza, possui todos os resultados possíveis mas que também possui uma série de resultados impossíveis. Vamos corrigir este “excesso” por meio do terceiro elemento da 3-upla do espaço de probabilidade, a função de probabilidade \mathbb{P} . Qualquer que seja a escolha, Ω é um conjunto infinito não-enumerável. ■

■ **Example 3.10 — Tweets, um caso mais complexo.** Colete todos os tweets emitidos em Belo Horizonte a partir do instante $t=0$. Cada tweet é registrado. Nossa “Mundo” está interessado em calcular a probabilidade de que:

- um tweet fale sobre música,
- o tempo de espera entre dois tweets sobre música seja estável no tempo,
- o número de caracteres usados no tweet tende a ser maior que a média no caso dos Tweets sobre música.

Ω pode ser representado pelo conjunto de todas as funções-escada do tipo mostrado na Figura 3.3. Os tempos $t_1 < t_2 < t_3 < \dots$ são os instantes em que os tweets chegam. Um degrau cinza da função-escada representa um tweet sobre música. Um degrau preto, um sobre não-música. A largura do degrau é proporcional ao número de caracteres.

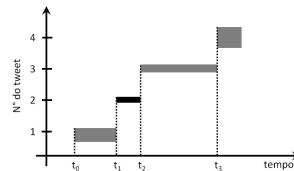


Figure 3.3: Um elemento de Ω no caso dos tweets.

O gráfico na Figura 3.3 acima é um resultado possível, um elemento $\omega \in \Omega$. Ω é formado pelas infinitas funções desse tipo variando os t_i 's, as cores e larguras dos degraus. É um conjunto infinito não-enumerável. Como atribuir probabilidades neste conjunto? ■

■ **Example 3.11 — Movimento Browniano.** Este exemplo é importante para motivar a necessária abstração e complicação matemática dos espaços de probabilidade. Não veremos este exemplo no resto do curso mas ele é um exemplo típico de processo estocástico, um assunto crucial em probabilidade mais avançada. Este exemplo tem importância histórica. Einstein publicou em 1905 um paper fundamental explicando o movimento browniano como efeito da movimentação atômica e ganhou um prêmio Nobel por isto alguns anos mais tarde.

É um tanto complicado apresentar a definição precisa de um movimento browniano neste altura mas ele não é difícil de entender intuitivamente. Seja (X_t, Y_t) a posição de uma partícula no instante de tempo t . Ela parte da origem de modo que $(X_0, Y_0) = (0, 0)$. A movimentação da partícula depende apenas de onde ela está num dado momento. Supondo que ela está em (X_t, Y_t) no tempo t , no tempo $t + \Delta$ ela passa a ter a posição $(X_{t+\Delta}, Y_{t+\Delta})$ onde $X_{t+\Delta} = X_t + \mathcal{N}_x(0, \Delta)$ e $Y_{t+\Delta} = Y_t + \mathcal{N}_y(0, \Delta)$. O incremento $\mathcal{N}_x(0, \Delta)$ possui distribuição gaussiana ou normal com média zero e variância igual a Δ , e analogamente para $\mathcal{N}_y(0, \Delta)$. Veremos no capítulo ?? a definição da distribuição gaussiana, mas neste momento você pode imaginar que o incremento em cada eixo de coordenadas tem igual chance de ocorrer para a frente ou para trás e tem comprimento médio aproximadamente igual a $0.80\sqrt{\Delta}$. O movimento browniano é o resultado de tomar $\Delta \rightarrow 0$, o que gera uma curva no plano similar à da Figura 3.4.

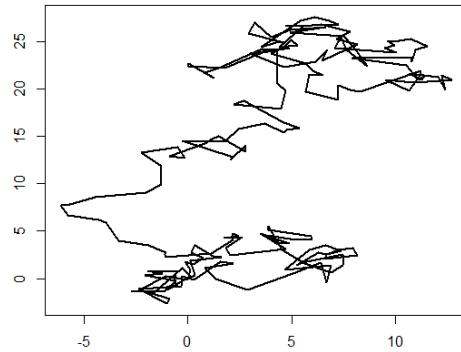


Figure 3.4: Movimento errático de um grão de pólen na superfície da água observado a cada 1 segundo.

Esta curva é aleatória. Repetindo o experimento nas mesmas condições gera curvas diferentes tais como as quatro mostradas na Figura 3.5.

Ω é o conjunto de todas as curvas do plano que podem aparecer como resultado do experimento, quatro das quais aparecem na Figura 3.5. Ω é um conjunto com infinitas curvas. Estas curvas possuem propriedades matemáticas muito curiosas. Por exemplo, elas são contínuas mas não possuem derivada em ponto nenhum e, além disso, o comprimento da curva em qualquer intervalo de tempo é infinito. Estas e várias outras propriedades não-intuitivas são rigorosamente derivadas com o ferramental de probabilidade avançada. A questão que não vamos responder mas que justifica a parafernália da 3-upla do espaço de probabilidades é: como atribuir probabilidades a este espaço amostral Ω composto dessas funções? ■

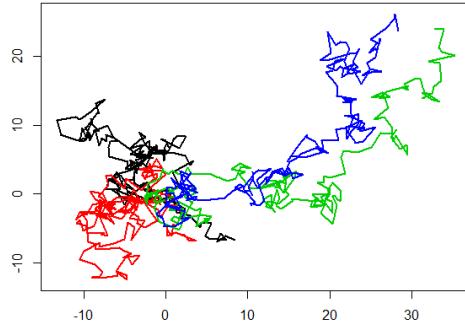


Figure 3.5: Quatro realizações do movimento browniano.

■ **Example 3.12 —** $\{0,1\}^\infty$. Um último exemplo não-trivial (e que também não será estudado no resto do curso). Joga-se uma moeda para cima independentemente *indefinidamente*. Isto é, o número de lançamentos é infinito. Vamos representar por 0 e 1 os resultados de um lançamento da moeda. Como não existe um limite para o número de lançamentos da moeda, o espaço amostral terá elementos da forma

$$\omega = (\omega_1, \omega_2, \omega_3, \dots)$$

onde cada ω_i será igual a 0 ou 1. Isto é, o elemento $\omega \in \Omega$ será um vetor de comprimento infinito onde cada entrada é 0 ou 1. O espaço amostral Ω é composto portodos os infinitos elementos ω desta forma. Curiosidade: $\Omega = [0,1]$ pois a expansão de um número real entre 0 e 1 na base 2 é da forma ω acima. ■

3.3 A σ -álgebra \mathcal{A}

O segundo elemento do espaço de probabilidade $(\Omega, \mathcal{A}, \mathbb{P})$ é a σ -álgebra \mathcal{A} . Queremos atribuir probabilidades a subconjuntos de elementos de Ω . Se $A \subseteq \Omega$ queremos calcular a sua probabilidade $\mathbb{P}(A)$ de alguma forma. O significado da notação $\mathbb{P}(A)$, a ser estabelecida na próxima seção, é

$$\mathbb{P}(A) = \mathbb{P}(\{\text{ocorrer } \omega \in \Omega \text{ tal que } \omega \in A\})$$

Por exemplo, no nosso exemplo simples do lançamento de uma dado equilibrado temos $\Omega = \{1, 2, 3, 4, 5, 6\}$. Além de querermos calcular probabilidades tais como $\mathbb{P}(\text{ocorrer face 4})$, vamos querer também calcular probabilidades tais como $\mathbb{P}(\text{ocorrer face maior que 3})$ ou $\mathbb{P}(\text{ocorrer face par})$. Estas probabilidades são equivalentes a calcular a probabilidade de que o resultado ω do experimento pertença ao subconjunto $\{4, 5, 6\}$ no primeiro caso e ao subconjunto $\{2, 4, 6\}$ no segundo caso. Assim, vamos querer calcular probabilidades de um resultado específico $\omega \in \Omega$ mas também probabilidades de que o resultado venha de um subconjunto $A \subseteq \Omega$.

Definition 3.3.1 A σ -álgebra \mathcal{A} é o conjunto dos subconjuntos $A \subseteq \Omega$ para os quais podemos calcular $\mathbb{P}(A)$. Os subconjuntos $A \in \mathcal{A}$ são chamados de *eventos*. Os subconjuntos $A \subseteq \Omega$ da forma $A = \{\omega\}$, compostos por um único elemento de Ω , são chamados de *eventos atômicos*.

Idealmente, queremos calcular $\mathbb{P}(A)$ para todo e qualquer subconjunto $A \subset \Omega$. Infelizmente, *em alguns casos*, não podemos calcular $\mathbb{P}(A)$ para todo e qualquer subconjunto $A \subseteq \Omega$. A σ -álgebra \mathcal{A} é simplesmente a classe dos sub-conjuntos de Ω para os quais podemos calcular $\mathbb{P}(A)$.

A σ -álgebra \mathcal{A} é muito simples se Ω for

- um conjunto finito de elementos (tal como $\Omega = \{0, 1, 2, 3\}$)
- ou um conjunto infinito mas enumerável (tal como $\Omega = \mathbb{N} = \{0, 1, 2, 3, \dots\}$)

Nestes dois casos $\mathcal{A} = 2^\Omega$, igual ao conjunto das partes de Ω , o conjunto de todos os sub-conjuntos de Ω . Se Ω for finito ou infinito enumerável, \mathcal{A} contém todos os subconjuntos de Ω . Neste caso não temos de nos preocupar: poderemos calcular $\mathbb{P}(A)$ para todo e qualquer subconjunto $A \subset \Omega$. Todo subconjunto de Ω é um evento para o qual teremos uma probabilidade.

A σ -álgebra \mathcal{A} é um conceito mais complicado se Ω for um conjunto não-enumerável como, por exemplo, o intervalo $[0, 1]$ ou a reta real. Neste caso, *não* poderemos calcular $\mathbb{P}(A)$ para todo e qualquer subconjunto $A \subset \Omega$. A σ -álgebra \mathcal{A} não conterá todos os subconjuntos de Ω . Nem todo subconjunto de Ω será um evento.

Por quê não é possível ter uma probabilidade $\mathbb{P}(A)$ para todo e qualquer subconjunto $A \subset \Omega$ nestes casos? Em teoria da medida (ou tamanho) de conjuntos, prova-se que não existe uma maneira *matematicamente consistente* de calcular o tamanho (ou a medida) de todos os subconjuntos de um conjunto não-enumerável. Podemos calcular a medida (ou o tamanho) de conjuntos muuuuuito estranhos mas não podemos calcular a medida de todos os subconjuntos.

A consequência teórica disso é que a σ -álgebra \mathcal{A} não contém todos os subconjuntos de Ω caso ele seja um conjunto não-enumerável (como o intervalo $[0, 1]$ ou a reta real). Entretanto, na prática da análise de dados, podemos ignorar isto e seguir trabalhando como se a σ -álgebra \mathcal{A} contivesse todos os subconjuntos. A razão para isto é que todos os subconjuntos A que podemos conceber na análise de dados, mesmo que muito complicados, pertencem a \mathcal{A} . Mas se é assim, o que não pertence a \mathcal{A} ? Os conjuntos que não estão em \mathcal{A} são tão estranhos que não podem ser exibidos, não temos uma fórmula para obtê-los.

Por exemplo, suponha que $\Omega = [0, 1]$ ou $\Omega = \mathbb{R}$. Pode-se mostrar que na menor σ -álgebra de alguma relevância prática (chamada de σ -álgebra de Borel) todo evento pode ser escrito com um número finito ou um número infinito enumerável de operações de \cup , \cap , e c de intervalos da reta. Veja que é muito difícil que algo diferente disso seja de utilidade prática na linha reta. Discutimos um pouco mais desses problemas no apêndice, o qual pode ser omitido sem prejuízo do entendimento do restante do livro.

3.4 A função de probabilidade \mathbb{P}

O terceiro elemento do espaço de probabilidade $(\Omega, \mathcal{A}, \mathbb{P})$ é a função de probabilidade \mathbb{P} que se encarrega de fazer a atribuição de probabilidades aos eventos $A \subset \mathcal{A}$. Já definimos quais são os resultados possíveis do fenômeno aleatório: são os elementos de Ω . Já definimos também quais são os eventos, os sub-conjuntos A de resultados em Ω para os quais podemos calcular uma probabilidade $\mathbb{P}(A)$: os eventos são os subconjunto de Ω que pertencem à σ -álgebra \mathcal{A} . Na prática, os eventos são todos o sub-conjuntos de Ω que podemos conceber.

Precisamos agora definir $\mathbb{P}(A)$ para todo A de forma consistente. Quais as propriedades que esta atribuição deve ter para que a gente não chegue a resultados inconsistentes ou contraditórios? Quais os requisitos mínimos que esta atribuição de probabilidades deve satisfazer, não importa o quanto complicados sejam Ω e \mathcal{A} ? Por exemplo, não queremos deduzir que, a partir de certa atribuição de probabilidade, acabamos por encontrar probabilidades negativas ou maiores que 1. Ou obtermos $\mathbb{P}(A) < \mathbb{P}(B)$ apesar de sabermos que A contém todos os resultados que pertencem a B e, portanto, deveríamos ter $\mathbb{P}(A) \geq \mathbb{P}(B)$.

É um tanto surpreendente que precisa-se de muito pouco para que uma função \mathbb{P} seja uma atribuição de probabilidade válida, que com certeza não vai gerar resultados inconsistentes. Basta que \mathbb{P} seja qualquer função como definida abaixo.

Definition 3.4.1 — Função de probabilidade \mathbb{P} . Dados um espaço amostral Ω e uma σ -álgebra de eventos \mathcal{A} , uma função de probabilidade \mathbb{P} é qualquer função tal que

$$\mathbb{P} : \mathcal{A} \longrightarrow [0, 1]$$

$$A \longrightarrow \mathbb{P}(A)$$

e que obedece aos três *axiomas de Kolmogorov*:

Axioma 1 $\mathbb{P}(A) \geq 0 \quad \forall A \in \mathcal{A}$

Axioma 2 $\mathbb{P}(\Omega) = 1$

Axioma 3 $\mathbb{P}(A_1 \cup A_2 \cup A_3 \cup \dots) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) + \dots$

se os eventos A_1, A_2, \dots forem disjuntos (isto é, mutuamente exclusivos).

Os dois primeiros axiomas estão apenas fixando uma escala para a probabilidade. Isto é, a probabilidade $\mathbb{P}(A)$ de um evento A qualquer deve ser um número entre 0 e 1, sendo que algum evento ocorre com certeza pois $\mathbb{P}(\Omega) = 1$.

O importante mesmo é o terceiro axioma:

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$$

se os A_i 's são disjuntos. No jargão matemático, dizemos que a função de probabilidade é uma função σ -aditiva ao satisfazer esta propriedade.

Para entender um pouco melhor o que esta propriedade significa, vamos considerar um caso particular do Axioma 3. Vamos tomar dois eventos A e B quaisquer e fazer $A_1 = A$, $A_2 = B$, e $A_n = \emptyset$, o conjunto vazio, para $n \geq 3$. Vamos também assumir que $A \cap B = \emptyset$. Assim, neste caso particular,

$$A_1 \cup A_2 \cup A_3 \cup A_4 \dots = A \cup B \cup \emptyset \cup \emptyset \dots = A \cup B$$

e o Axioma 3 permite concluir que

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(\emptyset) + \mathbb{P}(\emptyset) + \dots = \mathbb{P}(A) + \mathbb{P}(B).$$

Isto é, a função \mathbb{P} é aditiva sobre conjuntos disjuntos. A probabilidade de ocorrer um evento em A ou B é a soma das probabilidades de A e B .

Qualquer função que possua estas três propriedades, não importa o que sejam Ω e \mathcal{A} , será uma função de probabilidade válida. Não quer dizer que toda e qualquer função seja boa ou de utilidade prática. A regra acima apenas diz que funções \mathbb{P} serão incorretas como atribuições de probabilidade. Basta que \mathbb{P} satisfaça aos três axiomas de Kolmogorov para que \mathbb{P} seja uma atribuição de probabilidades válida.

3.4.1 Consequências

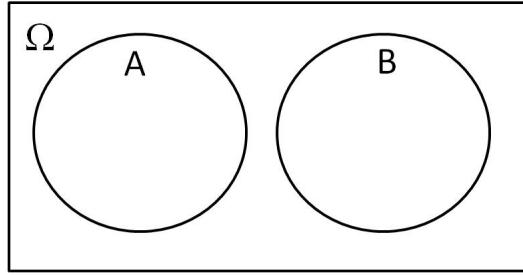
Em particular, do axioma 3, se $A \cap B = \emptyset$ então

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) \tag{3.1}$$

Todo o restante do cálculo de propriedades é decorrente destes três axiomas de Kolmogorov. Por exemplo, as seguintes propriedades podem ser derivadas imediatamente se \mathbb{P} satisfaz aos três axiomas de Kolmogorov.

Proposition 3.4.1 — Propriedades de \mathbb{P} . **(P1)** $\mathbb{P}(A^C) = 1 - \mathbb{P}(A)$

(P2) $0 \leq \mathbb{P}(A) \leq 1$ para todo evento $A \in \mathcal{A}$.

Figure 3.6: Diagrama de Venn com $A \cap B = \emptyset$.

(P3) se $A_1 \subseteq A_2 \implies \mathbb{P}(A_1) \leq \mathbb{P}(A_2)$

(P4) $\mathbb{P}(\bigcup_{n=1}^{\infty} A_i) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_i)$

(P5) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

Esta última propriedade é o caso geral de $\mathbb{P}(A \cup B)$, quando $A \cap B$ pode ser diferente do conjunto vazio.

Vamos provar a propriedade **P1**, que $\mathbb{P}(A^C) = 1 - \mathbb{P}(A)$ supondo que \mathbb{P} satisfaz aos três axiomas de Kolmogorov. Temos $\mathbb{P}(\Omega) = 1$ e $\Omega = A \cup A^c$. Como $A \cap A^c = \emptyset$, podemos usar (3.1). Assim,

$$1 = \mathbb{P}(\Omega) = \mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c)$$

e portanto

$$\mathbb{P}(A^C) = 1 - \mathbb{P}(A)$$

Para provar a propriedade **P2**, veja que o Axioma 1 da definição 3.4.1 já estabelece que $0 \leq \mathbb{P}(A)$ para qualquer $A \subseteq \Omega$. Pelo mesmo Axioma 1 da definição 3.4.1, temos $0 \leq \mathbb{P}(A^c)$ já que A^c é também um evento. Para verificar que $\mathbb{P}(A) \leq 1$, basta usar a prova de **P1**, em que concluímos que

$$1 = \mathbb{P}(A) + \mathbb{P}(A^c) \geq \mathbb{P}(A)$$

já que $\mathbb{P}(A^c) \geq 0$. Assim, $0 \leq \mathbb{P}(A)$ e $\mathbb{P}(A) \leq 1$. Ou seja, $0 \leq \mathbb{P}(A) \leq 1$.

Vamos provar agora **P3**: se $A_1 \subseteq A_2 \implies \mathbb{P}(A_1) \leq \mathbb{P}(A_2)$. Como $A_1 \subseteq A_2$, podemos escrever $A_2 = A_1 \cup (A_2 - A_1)$. Como $A_1 \cap (A_2 - A_1) = \emptyset$, pelo axioma 3 temos

$$\mathbb{P}(A_2) = \mathbb{P}(A_1 \cup (A_2 - A_1)) = \mathbb{P}(A_1) + \mathbb{P}(A_2 - A_1) \geq \mathbb{P}(A_1)$$

pois, pelo axioma 1, $\mathbb{P}(A_2 - A_1)$ tem de ser maior ou igual a zero.

Não vamos provar as outras duas propriedades. Ficam como exercício.

3.4.2 Como estabelecer uma função \mathbb{P} ?

OK, vimos que qualquer função \mathbb{P} que satisfaça aos axiomas de Kolmogorov é válida como uma função de atribuição de probabilidades. Mas como escolhemos uma dessas funções válidas num caso prático? Usamos uma combinação de conveniência matemática (facilidade de manuseio) com boa aproximação da realidade. Existe um trade-off entre estes dois aspectos. Se focarmos apenas no uso de modelos matematicamente muito simples vamos acabar com modelos que são muito distantes da realidade do fenômeno, que não o representam bem. Se insistirmos em incorporar todos os aspectos que podem afetar um fenômeno, teremos um modelo probabilístico inviável do ponto de vista matemático e computacional.

Para definir a função de probabilidade \mathbb{P} devemos considerar três casos:

- Ω é finito: $\Omega = \{\omega_1, \omega_2, \dots, \omega_N\}$

- Ω é infinito enumerável: $\Omega = \{\omega_1, \omega_2, \dots\}$
- Ω é não-enumerável, tal como $\Omega = (0, 1)$ ou $\Omega = \mathbb{R}^2$.

Como podemos esperar, o terceiro caso tem algumas complicações a mais em relação aos outros dois.

Vamos começar com o caso em que Ω é finito. Seja $\Omega = \{\omega_1, \omega_2, \dots, \omega_N\}$ onde os ω_i são eventos atômicos distintos, indivisíveis.

Notation 3.1. $\mathbb{P}(\{\omega_i\}) = \mathbb{P}(\omega_i)$.

Você pode atribuir valores $\mathbb{P}(\omega_i) \geq 0$ de forma completamente arbitrária desde que elas satisfaçam a restrição de que sua soma seja igual a 1:

$$\mathbb{P}(\omega_1) + \dots + \mathbb{P}(\omega_N) = \sum_{i=1}^N \mathbb{P}(\omega_i) = 1$$

■ **Example 3.13 — Mineração de dados num micromercado.** Suponha que existam apenas três produtos num micromercado: A, B , e C . Ω é composto pelas possíveis 8 cestas de produtos:

- 0 (ou nenhum produto),
- apenas A , apenas B , apenas C ,
- apenas os produtos AB juntos, apenas AC juntos, apenas BC juntos,
- os 3 produtos ABC juntos.

Vamos representar $\Omega = \{0, A, B, C, AB, AC, BC, ABC\}$.

Fez-se uma análise estatística do padrão de compras de vários clientes. Observou-se, por exemplo, que aproximadamente 17% dos clientes saíram com a cesta A e 3% saíram com a cesta AB . Isto permitiu obter aproximadamente as probabilidades as possibilidades de cada $\omega \in \Omega$. Por exemplo, $\mathbb{P}(A) \approx 0.17$ e $\mathbb{P}(AB) \approx 0.03$

Assim, podemos atribuir probabilidades aos elementos atômicos de Ω usando estas probabilidades aproximadas:

ω	0	A	B	C	AB	BC	AC	ABC	soma
$\mathbb{P}(\omega)$	0.02	0.17	0.19	0.09	0.03	0.21	0.18	0.11	1.00

Veja que elas são maiores ou iguais a zero e somam 1. ■

Uma estratégia simples para definir \mathbb{P} quando Ω é finito é atribuir probabilidades aos elementos ω de Ω . As probabilidades para os eventos $A \subseteq \Omega$ são derivadas dessa atribuição usando o Axioma 3 da definição 3.4.1. Suponha que $\Omega = \{\omega_1, \omega_2, \dots, \omega_N\}$ e que tenhamos atribuído probabilidades $\mathbb{P}(\omega_i) \geq 0$ para todo $i = 1, \dots, N$ tal que elas somam 1.

Como fica a probabilidade $\mathbb{P}(A)$ de um evento $A = \{\omega_{i_1}, \dots, \omega_{i_n}\}$? A é um conjunto finito, sub-conjunto de Ω , com n elementos, que pode ser visto como uma união de n conjuntos atômicos:

$$A = \{\omega_{i_1}, \dots, \omega_{i_n}\} = \bigcup_j \{\omega_{i_j}\}$$

Pelo o Axioma 3 de 3.4.1 temos

$$\mathbb{P}(A) = \mathbb{P}\left(\bigcup_j \{\omega_{i_j}\}\right) = \sum_{j=1}^k \mathbb{P}(\{\omega_{i_j}\}) = \sum_{\omega_i \in A} \mathbb{P}(\omega_i)$$

Repetindo: atribua probabilidades (somando 1) aos eventos atômicos $\omega \in \Omega$. Para qualquer evento $A \subset \Omega$:

- identifique quais os elementos ω_i que pertencem a A

- some suas probabilidades $\mathbb{P}(\omega_i)$

$$\mathbb{P}(A) = \mathbb{P}(\{\omega_{i_1}, \dots, \omega_{i_n}\}) = \sum_{\omega_i \in A} \mathbb{P}(\omega_i)$$

■ **Example 3.14 — micromercado de novo.** Voltando ao exemplo do micromercado com três produtos, A, B , e C , a cesta de compras de um cliente pode ser representada por $\Omega = \{0, A, B, C, AB, AC, BC, ABC\}$ com as probabilidades obtidas empiricamente (com dados)

ω	0	A	B	C	AB	BC	AC	ABC	soma
$\mathbb{P}(\omega)$	0.02	0.17	0.19	0.09	0.03	0.21	0.18	0.11	1.00

Aqui estão alguns eventos $E \subseteq \Omega$ compostos e suas probabilidades:

- E significa levar o produto A na cesta, ou $E = \{A, AB, AC, ABC\}$ e portanto $\mathbb{P}(E) = 0.17 + 0.03 + 0.18 + 0.12 = 0.49$.
- $E =$ levar o produto A mas não o produto C . Ou seja, $E = \{A, AB\}$ e $\mathbb{P}(E) = 0.17 + 0.03 = 0.20$.
- $E =$ uma cesta vazia, ou $E = \{0\}$ e portanto $\mathbb{P}(E) = \mathbb{P}(0) = 0.02$.
- $E =$ uma cesta vazia ou com pelo menos um produto. Então $E = \Omega$ e portanto $\mathbb{P}(E) = \mathbb{P}(\Omega) = 1$.
- $E =$ cesta com 4 produtos distintos. Ops, este evento não existe. Portanto $E = \emptyset$, o conjunto vazio, com $\mathbb{P}(\emptyset) = 1 - \mathbb{P}(\emptyset^c) = 1 - \mathbb{P}(\Omega) = 1 - 1 = 0$.

■

3.4.3 Visão frequentista

Se o fenômeno sob análise puder ser repetido:

- indefinidamente,
- nas mesmas condições
- de forma independente (sem que uma repetição afete a seguinte),

então $\mathbb{P}(\omega) \approx \frac{m}{N}$ onde m é o número de vezes que o resultado ω ocorreu nas suas N repetições.

Podemos tomar $\mathbb{P}(\omega) = \frac{m}{N}$, ignorando a aproximação amostral embutida. Esta é chamada a visão *frequentista* de probabilidade.

E $\mathbb{P}(A)$ para $A = \{\omega_{i_1}, \dots, \omega_{i_n}\}$? Temos duas possibilidades. Tome $\mathbb{P}(\omega_{i_j}) = \frac{m_{i_j}}{N}$ para cada $\omega_{i_j} \in A$ e some estas probabilidades:

$$\mathbb{P}(A) = \sum_j \mathbb{P}(\omega_{i_j}) = \sum_j \frac{m_{i_j}}{N}$$

A outra opção é simplesmente verificar quantas vezes o evento A ocorreu nas N repetições independentes e tomar

$$\mathbb{P}(A) = \frac{m}{N}$$

onde m é o número de vezes que o evento A ocorreu nas N repetições.

A segunda opção produz o mesmo resultado que a primeira opção.

■ **Example 3.15 — Frequentista no micromercado.** No exemplo, temos três produtos: A, B , e C . Uma análise estatística permitiu obter aproximadamente as probabilidades:

ω	0	A	B	C	AB	BC	AC	ABC	soma
$\mathbb{P}(\omega)$	0.02	0.17	0.19	0.09	0.03	0.21	0.18	0.11	1.00

Como elas foram obtidas? Um grande número N de clientes foram observados: são as repetições. Contou-se o número de vezes m em que a cesta foi BC . Finalmente tivemos $\mathbb{P}(BC) = m/N = 0.21$ ou 21% dos clientes. ■

■ **Example 3.16 — Dados de Weldon.** Walter Raphael Weldon (1860-1906) foi biólogo inglês fascinado pela teoria da evolução, como muitos brilhantes cientistas na primeira metade do século XX (ver Figura fig:WeldonEDados). Darwin publicou *The Origin of Species* em 1859 com sucesso e impacto imediato. Pelas décadas seguintes biólogos procuraram acumular evidências da evolução das espécies, mostrando que a diversidade biológica é o resultado de um processo de descendência com variabilidade (filhos não são idênticos aos pais) e com uma força (seleção natural) atuando para que alguns dos indivíduos tenham mais probabilidade de deixar descendentes. Foi um dos mais importantes biólogos defendendo a teoria da evolução e junto com Francis Galton e Karl Pearson fundou o periódico *Biometrika*, um dos mais importantes da estatística até hoje. Welson escreve num dos seus artigos que “... the questions raised by the Darwinian hypothesis are purely statistical, and the statistical method is the only one at present obvious by which that hypothesis can be experimentally checked”. A redescoberta do trabalho de Mendel em 1900 acendeu um debate envolvendo Welson sobre como seria possível combinar a teoria da evolução e a teoria genética. Esta junção foi realizada na década de 30 Sir Ronald Fisher, o maior estatístico que já existiu e que encontraremos várias vezes neste livro. Ter uma suficiente variabilidade é crucial para criar flexibilidade evolutiva em face das mudanças ambientais que criam as forças da seleção natural. A biologia evolutiva e a estatística moderna surgiram juntas, formando um par inextrincável.

Em uma carta de 1894 para Francis Galton, Weldon escreve que lançou 26306 vezes um conjunto de 12 dados. Sua motivação era “to judge whether the differences between a series of group frequencies and a theoretical law, taken as a whole, were or were not more than might be attributed to the chance fluctuation of random sampling”.



Figure 3.7: Walter Raphael Weldon (1860-1906), biólogo evolutivo e pioneiro da aplicação da estatística a problemas biométricos. Ele escreveu numa carta sobre o experimento de lançar 12 dados 26306 vezes e este experimento foi analisado por Karl Pearson em 1900 no artigo que apresentou o teste qui-quadrado.

A cada lançamento dos 12 dados, ele registrou o número de dados que exibiram a face 5 ou a face 6. Chamando de sucesso ao aparecimento de um 5 ou 6 num dos 12 dados, em cada lançamento dos 12 dados ele podia ter de zero sucessos (nenhum dos 12 dados com um 5 ou 6) até 12 sucessos (todos os 12 dados mostrando um 5 ou um 6). Os resultados obtidos estão na tabela abaixo representados como N_k onde $k = 0, \dots, 12$. A linha f_k mostra a frequência relativa do evento *obter k sucessos* (multiplicadas por 10000, para evitar muitos zeros depois do ponto decimal). Pela visão frequentista, a probabilidade de obter k sucessos ao lançar os 12 dados deveria ser aproximadamente igual a esses valores f_k .

k	0	1	2	3	4	5	6	7	8	9	10	11	12	Total
N_k	185	1149	3265	5475	6114	5194	3067	1331	403	105	14	4	0	26306
f_k	70.33	436.78	1241.16	2081.27	2324.18	1974.45	1165.89	505.97	153.20	39.91	5.32	1.52	0.00	
p_k	77.07	462.44	1271.71	2119.52	2384.46	1907.57	1112.75	476.89	149.03	33.12	4.97	0.45	0.02	
d_k	-6.74	-25.66	-30.55	-38.25	-60.28	66.88	53.14	29.08	4.17	6.79	0.35	1.07	-0.02	

Em 1900, Karl Pearson[18] publica um artigo fundamental na história da estatística, apresentando o teste qui-quadrado, capaz de medir, num certo sentido, a distância entre dados empíricos e previsões baseadas em modelos probabilísticos. Este teste será estudado no capítulo 10, onde voltaremos a este experimento de Weldon. Nesse artigo, Karl Pearson ele analisa as contagens de Weldon procurando ver se eles são compatíveis com a hipótese ou modelo de que os dados são bem balanceados, de que todas as 6 faces dos dados são igualmente prováveis. Se isto for verdade, a probabilidade de termos k sucessos ao lançar 12 dados é o que se vê na linha p_k da tabela acima (as contagens seguiriam um modelo binomial $\text{Bin}(12, 1/3)$, ver o próximo capítulo). Estas probabilidades também estão multiplicadas por 10 mil.

Para a maioria das pessoas as duas linhas de números, f_k e p_k , são bem próximas. Entretanto, Pearson mostrou que, se os dados fossem bem balanceados, seria altamente improvável, quase impossível, termos diferenças entre f_k e p_k do tamanho que vemos na tabela acima. A probabilidade de termos diferenças do tamanho das que vemos na tabela (que parecem pequenas) é 0.000016 (este é o p-valor do teste qui-quadrado, ver capítulo 10). Ou seja, os dados não são bem balanceados. A última linha da tabelamostra as diferenças $f_k - p_k$. Elas são negativas para k pequeno e positivas para k grande. Isto quer dizer que existe um “excesso” de eventos com muitos dados mostrando 5 ou 6. Os dados do experimento tendem a mostrar mais faces 5 ou 6 do que o esperado sob o modelo de dados bem balanceados. Pearson sugeriu que isto era devido ao processo de construção dos dados. Dados de cassino são cuidadosamente construídos. Dados de madeira baratos, do tipo que Weldon usou, são desbalanceados. Ao cavar a madeira para fazer as reentrâncias e formar as faces dos dados criamos um desequilíbrio. A face 6, a mais escavada, fica num lado do dado e a face 1, a menos escavada, fica do lado oposto. Assim, face 6 fica mais leve que sua face oposta (face 1) e costuma sair com mais frequência. Em conclusão, usando a frequência relativa como aproximação para a probabilidades reais que governam o lançamento dos dados baratos, o experimento de Weldon mostrou que eles não eram bem平衡ados. ■

R Vamos exercitar um olhar crítico no exemplo do minimercado. Pelo argumento frequentista, para que $\mathbb{P}(BC) \approx m/N = 0.21$, deveríamos ter repetições nas mesmas condições. Talvez isto não seja razoável. Alguns clientes são velhos, outros são jovens. Alguns compram no inverno e outros no verão. As condições em que as repetições estão ocorrendo não parecem ser idênticas. Se as condições não são idênticas pode ser que as probabilidades não se mantenham constantes. Por exemplo, a chance do produto A estar na cesta do cliente é alta se ele for jovem ou se for uma compra no verão mas a probabilidade é baixa caso contrário.

Uma outra suposição é que as repetições são independentes. Vamos formalizar este conceito de independência probabilística a seguir mas basicamente ele significa que o resultado de uma repetição não afeta as probabilidades de nenhuma outra. Isto também pode ser questionado. Alguns clientes podem influenciar outros via telefone ou comentários. Outro motivo é que, se os clientes não forem todos distintos, as compras de um mesmo cliente podem ser muito semelhantes. Para pensar numa situação limite, imagine que apenas um único cliente que compra sempre a mesma cesta tenha sido observado. Estimar as probabilidades baseado nos dados deste único cliente não é uma boa idéia.

Outra suposição é que as repetições podem ser feitas indefinidamente. Suponha que estejamos interessados em $\Omega = \{TGG, \overline{TGG}\}$ onde TGG significa a chance de uma terceira grande mundial nos próximos 5 anos e \overline{TGG} a sua não-ocorrência. Não parece razoável querer estabelecer probabilidades invocando frequências em repetições prolongadas nas mesmas condições deste tipo de evento.

A abordagem bayesiana assume que probabilidades são subjetivas e podem ser manipuladas com as regras do cálculo de probabilidade (ver na disciplina PGM: Modelos Gráficos Probabilísticos). Veremos ao longo do curso que existem várias maneiras de adaptar a versão básica

da abordagem frequentista para situações mais realistas, com as repetições não precisando ser nas mesmas condições e também com dependência entre elas.

3.4.4 \mathbb{P} quando Ω é infinito enumerável

Suponha que $\Omega = \{\omega_1, \omega_2, \dots\}$. Este caso é idêntico ao caso finito. Basta atribuir $\mathbb{P}(\omega_i) \geq 0$ aos elementos atômicos de Ω tal que as probabilidades somem 1. Para obter $\mathbb{P}(A)$ para algum evento composto A , some os valores $\mathbb{P}(\omega)$ de todos os elementos $\omega \in A$.

Por exemplo, uma moeda honesta é lançada repetidamente até observarmos a primeira coroa \tilde{c} . O espaço amostral é infinito e composto pelos elementos atômicos representando a sequência observada

$$\Omega = \{\tilde{c}, c\tilde{c}, cc\tilde{c}, \dots\}$$

Uma atribuição válida de probabilidades é a seguinte:

$$\mathbb{P}(\omega_i) = \begin{cases} \mathbb{P}(\tilde{c}) = 1/2 \\ \mathbb{P}(c\tilde{c}) = (1/2)(1/2) \\ \mathbb{P}(cc\tilde{c}) = (1/2)(1/2)(1/2) \\ \vdots \end{cases}$$

Ela é válida pois $\mathbb{P}(\omega_i) \geq 0$ para todo $\omega_i \in \Omega$ e temos

$$\sum_i \mathbb{P}(\omega_i) = \sum_i \mathbb{P}(\underbrace{cc\dots c}_{i \text{ terms}}\tilde{c}) = \sum_{i=1}^{\infty} \left(\frac{1}{2}\right)^i = 1$$

Seja A o evento composto em que a moeda é lançada um número par de vezes:

$$A = \{c\tilde{c}, ccc\tilde{c}, ccccc\tilde{c}, \dots\}$$

Temos

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \left(\frac{1}{2}\right)^{2i} = \sum_{i=1}^{\infty} \left(\frac{1}{4}\right)^i = \frac{1}{3}$$

Considere o evento B em que a moeda é lançada menos de 5 vezes. Então

$$\mathbb{P}(B) = \mathbb{P}(\{\tilde{c}, c\tilde{c}, cc\tilde{c}, ccc\tilde{c}\}) = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16}.$$

3.4.5 \mathbb{P} quando Ω é infinito não-enumerável

Os conjuntos não-enumeráveis de novo. Os conjuntos infinitos *enumeráveis* são infinitinhos. Os conjuntos infinitos não-enumeráveis são infinitões. Existem várias dificuldades em lidar rigorosamente com eles. Daremos apenas um exemplo para estes conjuntos.

Selecione um número completamente ao acaso no intervalo $[0,1]$. $\Omega = [0,1]$. Como atribuir probabilidades aos eventos $A \subseteq [0,1]$? Vamos tentar o mesmo procedimento do caso em que Ω é finito ou enumerável. Isto é, atribua um valor $\mathbb{P}(\omega)$ para cada número real $\omega \in [0,1]$ e defina

$$\mathbb{P}(A) = \sum_{\omega_i \in A} \mathbb{P}(\omega_i)$$

Isto não vai dar certo. Como nenhum ponto é favorecido, escolhemos um ponto “completamente ao acaso”. Por isto, deveremos fazer $\mathbb{P}(\omega) = \xi > 0$ para todo $\omega \in \Omega$. Isto é, nenhum ponto tem mais chance que outro de ser escolhido, a probabilidade é mesma para todos eles.

E então o que é $\mathbb{P}(A)$? Suponha que $A = \{1/2, 1/4, 1/8, 1/16, \dots\}$. Pelo Axioma 3 de 3.4.1 temos

$$\mathbb{P}(A) = \sum_{\omega_i \in A} \mathbb{P}(\omega_i) = \sum_{\omega_i \in A} \xi = \xi \cdot \infty = \infty$$

se $\xi > 0$. Portanto, algo está errado.

O erro é assumir $\mathbb{P}(\omega) = \xi > 0$. Todo ponto em $[0, 1]$ tem probabilidade 0! O correto é assumir que $\mathbb{P}(\omega) = 0$ para todo ponto $\omega \in [0, 1]$. Mas se todo número em $[0, 1]$ tem probabilidade zero, como poderemos ter $\mathbb{P}([0, 1/2]) = 1/2$?

Este paradoxo sempre aparece quando representamos a realidade com os números da reta real. Por exemplo, em física, a representação da realidade com números reais gera paradoxos. Suponha que o intervalo $[0, 1]$ represente um segmento de fio com massa de 1 grama. Suponha que o fio tem sua massa perfeitamente e regularmente distribuída no fio. Dizemos que ele tem uma densidade de massa constante.

Qual a massa de um ponto $x \in [0, 1]$? Suponha que o ponto tem uma massa $\xi > 0$. Como a densidade é constante, todos os pontos devem ter a mesma massa $\xi > 0$. Como existem infinitos pontos, a massa total deveria ser $\xi \times \infty = \infty$ e não, 1 grama.

O modelo que representa o fio por um segmento de reta é incorreto do ponto de vista físico, é apenas uma aproximação idealizada da realidade. O fio possui unidades atômicas que possuem massa. Sua representação como uma linha contínua leva a paradoxos.

A solução matemática para tornar a representação útil é assumir que:

- Todo ponto isolado do fio possui massa zero.
- A massa associada com um segmento $[a, b]$ é diretamente proporcional ao seu comprimento.
- Como a massa total de $[0, 1]$ é 1 grama, a massa de $[a, b] \in [0, 1]$ é $b - a$.
- Por exemplo, $[0, 1/2]$ tem massa 1/2 grama, $[1/2, 3/4]$ tem massa 1/4 grama, etc.
- Note que o ponto x é também o intervalo $[x, x]$, que possui massa 0 pois tem comprimento 0.

Uma maneira um pouco mais complicada é usar uma função densidade de massa. Esta função será muito útil quando a massa não estiver distribuída de maneira uniforme. A função densidade de massa é uma função $f(x)$ definida para cada x no segmento $[0, 1]$. Esta função é tal que a massa no segmento $[a, b]$ é a sua integral entre a e b :

$$\text{massa em } [a, b] = \int_a^b f(x) dx$$

Se tomarmos $f(x) = 1$ para todo $x \in [0, 1]$ teremos

$$\text{massa em } [a, b] = \int_a^b f(x) dx = \int_a^b 1 dx = b - a$$

Esta é a função densidade $f(x)$ para o fio com massa uniformemente distribuída em $[0, 1]$.

A idéia de uma função mais geral que a função constante é espalhar a massa total do objeto por meio da função $f(x)$. A massa pode não ser uniformemente distribuída. Por exemplo, o material é uma liga com dois elementos (cobre e zinco). Em certas regiões, existe mais cobre que zinco. Em outras, o zinco domina. A densidade do fio vai variar de acordo com a proporção de zinco no local. Ela pode estar mais concentrada em algumas regiões do fio que em outras. Isto fica refletido imediatamente na função densidade $f(x)$. Nas regiões onde a massa é mais concentrada, $f(x)$ será maior.

A Figura 3.8 mostra exemplos de diferentes densidade de massa $f(x)$ para $x \in [0, 1]$. A massa de qualquer subconjunto é obtida por integração no subconjunto.

Com conjuntos Ω não-enumeráveis tais como $\Omega \subseteq \mathbb{R}^n$ adotamos o mesmo procedimento. Massa total de probabilidade de Ω é 1 pois $\mathbb{P}(\Omega) = 1$. Espalhe em Ω a massa total de probabilidade usando

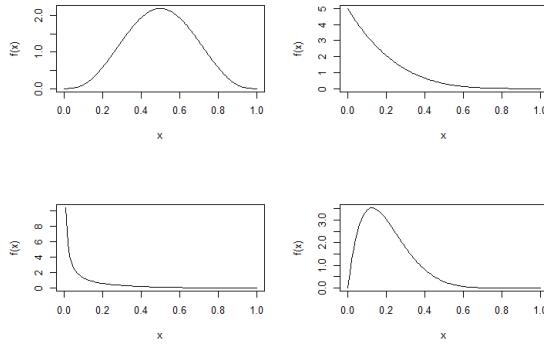


Figure 3.8: Quatro diferentes funções densidade de massa $f(x)$.

$f(x)$ para $x \in \Omega$. A massa de probabilidade de qualquer evento $A \subset \Omega$ é obtida por integração:

$$\mathbb{P}(A) = \int_A f(x)dx$$

■ **Example 3.17** Pense no experimento de escolher um ponto completamente ao acaso em $[0, 1]$. Tome $f(x) = 1$ para $x \in [0, 1]$. Vamos considerar o evento $A = [a, b]$, um intervalo. O experimento escolhe um único ponto em $[0, 1]$, e não intervalos. Ocorrer o evento A significa que o ponto escolhido pertence ao intervalo $A = [a, b]$.

$$\mathbb{P}(\text{Intervalo}[a,b]) = \text{Comprimento do intervalo } [a,b] \quad (3.2)$$

Todo ponto isolado do intervalo $[0, 1]$ possui probabilidade zero. ■



Em \mathcal{A} , todo evento pode ser aproximado com um número finito ou infinito enumerável de operações \cup , \cap , e C em intervalos da reta. Sabemos calcular probabilidades de intervalos. Como todo evento é obtido a partir de operações de conjunto em cima de intervalos, a probabilidade $\mathbb{P}(A)$ pode ser estabelecida para todo evento $A \in \mathcal{A}$. Este é essencialmente o resultado do Teorema de extensão de Caratheodory. Começamos com probabilidades sobre conjuntos básicos e o teorema nos garante que, por extensão, a probabilidade fica determinada para todos os demais eventos.

3.4.6 Quando Ω é complicado

Definir densidades para Ω complicados pode ser difícil. Pior: pode ser impossível pois não sabemos explicitar uma densidade em vários casos. Novamente, os conjuntos infinitos vêm nos assombrar. É sempre a dificuldade de lidar matematicamente com o infinito “excessivo”. Existem situações práticas que exigem trabalhar com estes conjuntos Ω e temos de solucionar isto de alguma forma.

Por exemplo, no caso do movimento browniano, onde observamos o movimento errático de um grão de pólen na superfície da água observado a cada 1 segundo (ver Figura 3.5). Ω é o conjunto de todas as curvas erráticas do movimento browniano. Como definir eventos (sub-conjuntos de Ω) aqui? Queremos calcular, por exemplo, a probabilidade da trajetória da partícula não se auto-interseccar nos primeiros 10 minutos. Este evento corresponde a uma imenso conjunto de curvas de Ω . Qual a probabilidade de sua ocorrência? Como atribuir probabilidades de forma consistente a todos os eventos?

Num segundo exemplo, considere o lançamento de uma moeda honesta indefinidamente e que tem $\Omega = \{0, 1\}^\infty$. Como definir probabilidades de forma consistente para todos os eventos

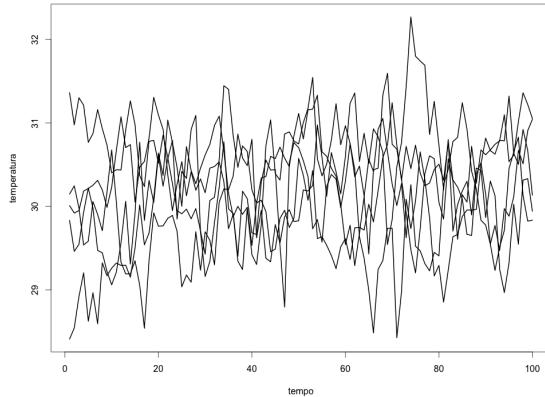


Figure 3.9: Ω é formado pelo conjunto infinito de funções contínuas. A figura mostra um pequeno número de exemplares desse conjunto Ω .

neste caso? Eventos devem levar em conta os infinitos lançamentos. Seja f_n a proporção de 1's nos primeiros n lançamentos. Monitore f_n ao longo de uma sequência $\omega \in \Omega$ tal como $(0, 1, 0, 0, 0, 1, 0, \dots)$. O que acontece com f_n quando n cresce? Com base na nossa experiência, esperamos ver $f_n \rightarrow 1/2$. Mas isto acontece com certeza? Com probabilidade 1? Ou existe alguma chance, por mínima que seja, de que f_n não convirja para $1/2$?

Ou, quem sabe, f_n não convirja para valor algum, oscilando no intervalo $(0, 1)$ sem estabilizar-se permanentemente em torno de nenhum valor. Afinal, podemos pensar em muitas (infinitas!) sequências $\omega \in \Omega$ tais que $f_n \not\rightarrow 1/2$. Por exemplo, $\omega = (0, 1, 0, 1, 1, 1, 1, 1, 1, \dots)$ Ou $\omega = (1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, \dots)$ Qual a probabilidade de que ocorra uma dessas infinitas seqüências com $f_n \not\rightarrow 1/2$? A resposta é: a probabilidade é igual a zero (é o da teorema Lei Forte dos Grandes Números, ver capítulo ??). Numa sequência infinita de lançamentos de uma moeda equilibrada a probabilidade de que f_n não convirja para $1/2$ é zero. Mas existem infinitas sequências desse tipo em Ω , que não convergem. E no entanto ela possuem probabilidade zero de ocorrência.

E se a moeda tiver a probabilidade de cara bem pequena, digamos $\theta \approx 0$. Qual a probabilidade de que f_n não convirja para este θ bem próximo de zero? É zero novamente. Isto está divertido. Vamos ver um outro evento. Pegue o comprimento da mais longa sequência de 1's ininterruptos na série infinita de lançamentos da moeda honesta. Qual a probabilidade de que este comprimento seja pelo menos 2000? Curioso? A probabilidade é igual a 1, vai acontecer com certeza ao longo da sequência infinita de lançamentos da moeda. Esta conclusão é um resultado direto do Teorema de Borel-Cantelli, assunto do Capítulo ??.

Um caso mais curioso ainda. Suponha que Ω seja o conjunto formado por todas as funções contínuas. Este Ω poderia ser o resultado do experimento de observar a curva contínua de temperatura durante 24 horas num certo local. A Figura 3.9 mostra alguns resultados desse experimento. Isto é, o elemento $\omega \in \Omega$ é uma das infinitas curvas possíveis. Eventos são sub-conjuntos de curvas deste conjunto Ω .

Como atribuir probabilidades de forma consistente a todos os eventos possíveis? Por exemplo, se A e B são dois eventos (dois subconjuntos de curvas) tais que $A \subset B$ então devemos ter $\mathbb{P}(A) \leq \mathbb{P}(B)$. O que poderia ser uma densidade de probabilidade neste conjunto Ω de curvas contínuas? Como integrar neste conjunto? Precisamos de uma noção de integral mais complexa que a integral de Riemann, uma noção de medida ou tamanho de conjuntos. Isto é assunto de cursos avançados de probabilidade e processos estocásticos.

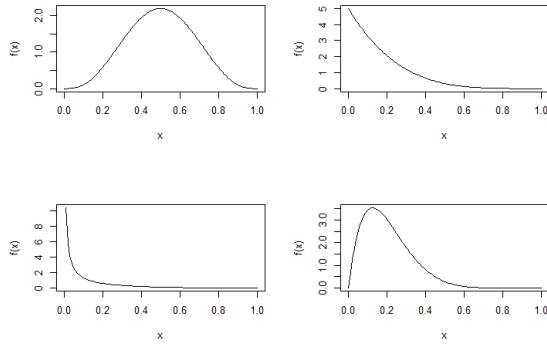


Figure 3.10: Quatro diferentes funções densidade de massa de probabilidade $f(x)$.

3.4.7 Solução a vista

Vamos evitar todas estas complicações. Na prática da análise de dados não trabalhamos diretamente com Ω . Reduzimos o fenômeno estocástico a algumas poucas características numéricas com as quais descrevemos o experimento aleatório. Estas características são chamadas *variáveis aleatórias*, assunto dos capítulos 6 e 7. Na prática, isto vai significar que, no “pior caso”, teremos Ω equivalente a subconjuntos de \mathbb{R}^n , para os quais podemos definir densidades de probabilidade.

3.4.8 Função densidade de probabilidade

Variáveis aleatórias, assunto do próximo capítulo, fazem com que $\Omega \subset \mathbb{R}^n$ na prática. E este caso é muito fácil pois então a densidade $f(\omega)$ pode ser qualquer função

$$\begin{aligned} f : \Omega \subset \mathbb{R}^n &\longrightarrow \mathbb{R} \\ \omega &\longrightarrow f(\omega) \end{aligned}$$

tal que:

- $f(\omega) \geq 0$, (para não obtermos uma probabilidade negativa).
- e $\int_{\Omega} f(\omega) d\omega = 1$

■ **Example 3.18 — Quando Omega é um intervalo .**



[Observações] Devemos ter $f(\omega) \geq 0$, um limite inferior. Mas podemos ter $f(\omega) > 1$: não há um limite superior. A restrição fundamental é que a integral sobre todo Ω deve ser 1. Não se exige que o valor $f(\omega)$ em cada $\omega \in \Omega$ seja menor que 1. Para obter a probabilidade de um evento $A \subset \Omega$ basta integrar $f(x)$ sobre a região A :

$$\mathbb{P}(A) = \int_A f(x) dx$$

Assim, uma probabilidade $\mathbb{P}(A)$ é área (ou volume) sob uma curva (ou superfície) de densidade.

■ **Example 3.19 — Dardos em alvo circular.** Suponha que dardos são atirados num alvo circular de raio 1. Um jogador possui uma habilidade que faz com que a chance de acertar numa região A próxima do centro é maior que se esta mesma região estiver próxima da borda. Esta habilidade está representada pela densidade

$$f(x, y) = c \left(\sqrt{x^2 + y^2} - 1 \right)^2$$

para x, y no disco unitário c é uma constante para garantir que $\int_{\Omega} f(x, y) dx dy = 1$. Pode-se mostrar que então $c = 14\pi/12$.

Seja $r = \sqrt{x^2 + y^2}$, a distância de (x, y) até a origem. Podemos reescrever a densidade anterior da seguinte forma:

$$f(x, y) = c \left(\sqrt{x^2 + y^2} - 1 \right)^2 = c(r - 1)^2$$

Isto torna mais simples a visualização da densidade com um mapa de calor ou curvas de nível.

Para uma região A qualquer dentro do disco, temos

$$\mathbb{P}(A) = \int_A f(x, y) dx dy$$

Indivíduos com habilidades diferentes terão sua densidade diferente. A densidade deverá expressar quais as regiões mais prováveis de serem atingidas. Como seria um mapa de calor da densidade $f(x, y)$ de um jogador “cego”? E de um jogador extremamente habilidoso? E um jogador que tem um viés par a direita, que tende a jogar o dardo deslocado para a direita? ■

■ **Example 3.20** Interesse no tempo de espera pelo primeiro comentário após a postagem de um vídeo do YouTube do canal de Whindersson Nunes.

Espaço amostral Ω ? $\Omega = (0, \infty)$

Densidade $f(x)$? Várias alternativas para $f(x)$ - ver no gráfico a seguir.

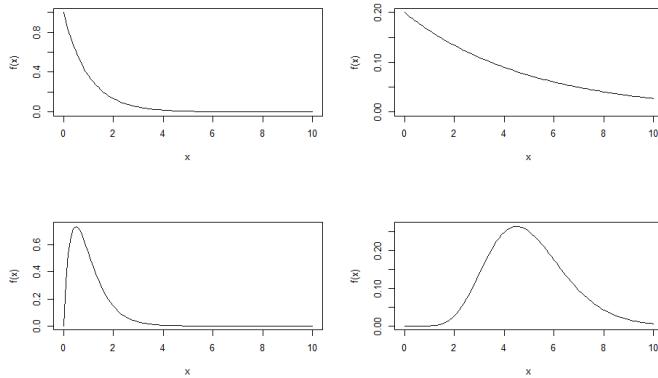


Figure 3.11: Quatro possíveis modelos de densidade de probabilidade $f(x)$ para tempo de espera por comentário num canal do You Tube. Visualmente, veja como são diferentes as probabilidades $\mathbb{P}(A) = \mathbb{P}((2, 4))$. ■

3.5 Detalhes

Como dissemos na seção 3.3, quando Ω é um conjunto não-enumerável, tal como $\Omega = [0, 1]$ ou $\Omega = (0, \infty)$, a σ -álgebra \mathcal{A} não contém todos os subconjuntos de Ω .

Para entender isto um pouco melhor, considere o caso em que escolhemos um número real ao acaso no intervalo $[0, 1]$. Assim, $\Omega = [0, 1]$. Vamos dizer que a medida (ou tamanho) de $\Omega = [0, 1]$ seja 1. Notação: $\mu([0, 1]) = 1$.

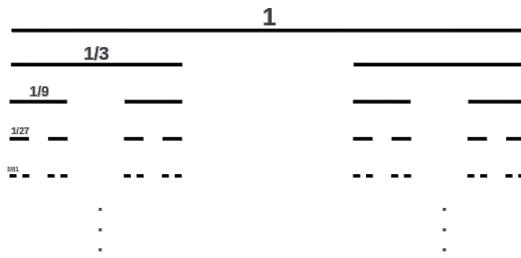


Figure 3.12: O conjunto de Cantor.

A medida $\mu(I)$ de um sub-intervalo $I \subset [0, 1]$ será igual a seu comprimento. Por exemplo, $\mu((0, 1/2)) = 1/2$ e $\mu((1/2, 3/4)) = 1/4$.

E se tivermos um conjunto formado pela união de k sub-intervalos disjuntos? Como eles não se sobrepõem, a medida da união será a soma das medidas dos intervalos componentes. Por exemplo, $\mu((0, 1/2) \cup (3/4, 1)) = \mu((0, 1/2)) + \mu((3/4, 1)) = 3/4$.

E se $A = \{1/2\}$, o sub-conjunto formado apenas pelo ponto $1/2$? Para sermos *consistentes* com a atribuição da medida de um intervalo como sendo igual a seu comprimento, teremos de fazer a medida (ou tamanho) igual a zero: $\mu(\{1/2\}) = 0$

E se tivermos vários pontos em $[0, 1]$? Por exemplo, $A = \{1/4, 2/4, 3/4\}$? Para sermos consistentes com a noção de medida igual a um comprimento, vamos precisar tomar $\mu(\{1/4, 2/4, 3/4\}) = 0$

E se tivermos infinitos pontos em $[0, 1]$? A resposta depende do tipo de infinito. Sem ter a intenção de demonstrar formalmente as afirmações a seguir, vamos apenas dar uma ideia da necessidade de considerar coisas tão complicadas quanto uma σ -álgebra.

Suponha que A seja um conjunto infinito mas enumerável. Por exemplo, suponha que $A = \{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \dots, \frac{1}{n}, \dots\}$. Neste caso, pode-se mostrar que, para sermos consistentes com o que já assumimos até agora, teremos de tomar sua medida ou tamanho como $\mu(A) = 0$.

Em geral, se A for um conjunto infinito mas enumerável deveremos ter $\mathbb{P}(A) = 0$. Por exemplo, considere o sub-conjunto A formado por *todos os números racionais* de $[0, 1]$. Isto é, todas as frações em $[0, 1]$. Este sub-conjunto tem medida (ou tamanho) 0, apesar de A ser um conjunto denso em $[0, 1]$. Isto é surpreendente pois, para quaisquer dois pontos $a < b$ de $[0, 1]$, por menor que seja a distância entre a e b , existem infinitos pontos racionais entre eles. Ainda assim, apesar dos racionais serem densos, apesar de estarem em qualquer mínimo segmento de reta em quantidade infinita, se quisermos uma medida consistente de tamanho, teremos de dizer que a medida de todos os racionais em $[0, 1]$ é zero. Por outro lado, o conjunto dos irracionais em $[0, 1]$ possui medida (ou tamanho) igual a 1, a mesma medida que o conjunto total $[0, 1]$.

■ **Example 3.21 — O conjunto de Cantor K .** Matemáticos pensaram em conjuntos muito mais estranhos que os racionais, como o conjunto de Cantor K , por exemplo. Este conjunto está representado na Figura 3.12 e é construído da seguinte forma.

- Delete o terço central do intervalo $[0, 1]$.
- A seguir, delete o terço central de cada um dos dois sub-intervalos.
- Itere ad infinitum (portanto isto não é um algoritmo).
- O resultado “final” é o conjunto de Cantor.

O conjunto de Cantor K tem propriedades notáveis (ver https://pt.wikipedia.org/wiki/Conjunto_de_Cantor). Sobra alguma coisa no final? Sim sobra muito: K possui um número não-enumerável de pontos. K é equivalente ao conjunto dos números que usam apenas os dígitos 0 e 2 na sua representação em base 3.

Para todo ponto $x \in K$, temos uma sequência de pontos distintos $x_n \rightarrow x$. Assim, K não possui

pontos isolados. Em torno de todo ponto de K existem infinitos outros pontos de K , não importa quanto pequena seja a vizinhança. E no entanto K é totalmente desconexo (não contém intervalos)!! Pois bem, pode-se provar rigorosamente que a medida (ou tamanho) deste conjunto K é igual a zero.

Subconjuntos de conjuntos não-enumeráveis podem ser muito estranhos. Tão estranhos que alguns não podem ser medidos. Prova-se que não existe uma maneira *consistente* de estender o conceito de medida (ou tamanho) para *todos* os subconjuntos de $[0, 1]$. A consequência é que nem todo subconjunto de $[0, 1]$ pode ter um tamanho associado a ele.

Quem são estes conjuntos estranhos, tão estranhos que não podemos associar um medida ao seu tamanho? Eles são chamados de conjuntos não-mensuráveis. Ver https://en.wikipedia.org/wiki/Non-measurable_set para mais detalhes. Ninguém nunca “viu” um desses conjuntos não-mensuráveis. Não existe uma maneira construtiva de gerar estes conjuntos não-mensuráveis. O que se prova é que *existem* conjuntos não mensuráveis mas não conseguimos *construir* (e exibir) um deles.

Todos os exemplos conhecidos de conjuntos não-mensuráveis usam o Axioma da Escolha (ver https://pt.wikipedia.org/wiki/Axioma_da_escolha) e assim não podemos mostrar um deles *explicitamente*. Um exemplo de conjunto não-mensurável é o conjunto de Vitali (ver http://en.wikipedia.org/wiki/Vitali_set)

A σ -álgebra de Borel

As σ -álgebras mais populares são as de Borel. Se $\Omega = \mathbb{R}$ ou $\Omega = (0, 1)$, tome todos os intervalos e gere *todos* os conjuntos possíveis usando \cup , \cap e C em número *enumerável*. \mathcal{A} é fechado para \cup , \cap e C em número enumerável: se $A_n \in \mathcal{A}$ então $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$. Se $\Omega = \mathbb{R}^n$, comece com os “cubos” em \mathbb{R}^n e gere todos e gere todos os conjuntos possíveis usando \cup , \cap e C em número enumerável.

A σ -álgebra \mathcal{A} de Borel é grande o suficiente para conter todos os casos de interesse prático. Se você for capaz de “pensar” num conjunto, ele provavelmente estará na σ -álgebra de Borel.

Conjuntos não-mensuráveis são tão estranhos que podem ser ignorados na prática da análise de dados. Os conjuntos mensuráveis são tantos e tão diversos que incluem todos os sub-conjuntos de Ω que podem aparecer em problemas de engenharia e de matemática aplicada. Portanto, vamos ignorar esta complicação de que a σ -álgebra \mathcal{A} não inclui todos os subconjuntos de Ω . Todo sub-conjunto que você conseguir imaginar, acredite, será um conjunto mensurável e vai fazer parte da σ -álgebra \mathcal{A} .

Em resumo...

- A teoria de σ -álgebras e de conjuntos mensuráveis é muito importante no estudo rigoroso dos fundamentos de probabilidade.
- No estudo dos processos estocáticos, onde Ω é bastante complicado, estas questões matemáticas são importantes.
- Por exemplo, no caso do movimento browniano, Ω é um conjunto infinito de curvas contínuas do plano. No caso do lançamento da moeda *indefinidamente*, $\Omega = \{0, 1\}^{\infty}$.
- A σ -álgebra também aparece ao definir as distribuições condicionais de forma rigorosa.
- Na nossa disciplina e na prática da análise de dados e aprendizado de máquina probabilístico, este assunto não é muito relevante.
- Ele é um tópico importante para o estudo rigoroso da *teoria* de estatística e de aprendizagem de máquina. Especialmente nos tempos atuais, quando muitas ferramentas estão sendo desenvolvidas para buscar aproximações em espaços de funções, este assunto é de interesse para vários teóricos. Árvores de regressão, por exemplo, é um problema em que Ω é bem complicado.
- Vamos ignorar as complicações de σ -álgebra no restante desse livro. Vamos supor que todo sub-conjunto que formos capazes de conceber estará na σ -álgebra \mathcal{A} .



4. Probabilidade Condicional

4.1 Probabilidade Condisional

4.1.1 O que é uma probabilidade condicional

Seja B um evento em Ω e $\mathbb{P}(B)$ sua probabilidade de ocorrência. Sem poder ver o resultado do experimento diretamente, somos informados apenas que outro evento A ocorreu. Isto muda a probabilidade de B ocorrer? Por exemplo, dois dados bem equilibrados são lançados em sequência. Você apostaria na ocorrência de B : o primeiro dado vai resultar num 6. Se você souber que a soma dos dois dados foi menor que 8 (evento A) e pudesse rever sua aposta, você colocaria mais fichas na ocorrência de B ? Ou menos fichas?

De posse da informação de que certo evento A ocorreu, queremos recalcular as chances de outros eventos B_1, B_2, \dots . Chamamos a isto de probabilidade de B condicionada à ocorrência do evento A , ou de probabilidade de B dado que A ocorreu, ou, mais curto ainda, probabilidade de B dado A .

Notation 4.1. Notação: $\mathbb{P}(B|A)$

A imensa maioria das técnicas de ciência dos dados são algoritmos para fazer cálculos de probabilidade condicional.

4.1.2 Probabilidade Condisional e Ciência dos Dados

Este o momento do diagnóstico de um câncer de estômago para um paciente qualquer. B é o evento em que o paciente terá pelo menos mais 1 ano de vida. Suponha que, baseado em vários casos similares, sabemos que aproximadamente 70% dos pacientes nas condições sobrevivem por mais de um ano. Estamos usando a idéia frequentista. Dentre todos os pacientes observados em situação semelhante no passado recente, 70% deles viveu mais de um ano a partir do diagnóstico. Portanto, assumir que $\mathbb{P}(B) = 0.70$ é razoável.

Seja A o evento em que um paciente de câncer de estômago tenha uma biópsia confirmando que o tumor é benigno. Imaginamos que $\mathbb{P}(B|A)$ seja maior que $\mathbb{P}(B) = 0.70$. Como recalcular a probabilidade da ocorrência de B ? Se tivermos um grande número de pacientes inicialmente diag-

nosticados e com biópsia posterior indicando benigno, contamos a proporção dos que sobrevivem mais de um ano dentro desse subgrupo de indivíduos. Isto será uma boa aproximação para $\mathbb{P}(B|A)$.

O problema fica mais complicado se o evento A representar a seguinte informação:

- biópsia indica um tumor benigno,
- paciente tem 45 anos de idade,
- é homem,
- sempre morou em Santa Catarina,
- é fumante
- e sempre come salames e salsichas defumadas.

Não haverá uma amostra muito grande de pacientes nestas condições exatas. Talvez apenas 2, 1 ou até zero pessoas tenham sido observadas nestas condições. Isto impede usar a simples frequência ocorrida nestes pouquíssimos casos para aproximar $\mathbb{P}(B|A)$. Ferramentas de ciência dos dados calculam estas probabilidades usando vários truques. Elas procuram extrair o máximo de informação dos dados.

De maneira geral, dadas as características representadas por A , como fica a chance de ocorrer B ? Dado que os sensores do robô dizem que ocorreu A , qual a chance de que ele esteja na região B ? Dado que o usuário comprou o conjunto A de itens nas últimas visitas, qual é o item B para o qual a probabilidade de compra é máxima? Dado certo comportamento de uma ação no mercado financeiro nos últimos 3 anos, qual a probabilidade de que ela suba 10% ou mais dentro de 30 dias? Dadas certas características A de um e-mail, qual a chance dele ser um spam?

Probabilidade condicional é extremamente importante em teoria mas é mais importante ainda na prática da análise de dados. Ela pode ser difícil de calcular e é a fonte de quase todos os paradoxos no cálculo de probabilidade.

4.1.3 Definindo Probabilidade Condisional

Primeira questão: como passar de $\mathbb{P}(B)$ para $\mathbb{P}(B|A)$? Qual a relação entre $\mathbb{P}(B)$ e $\mathbb{P}(B|A)$? Podemos ter $\mathbb{P}(B) = \mathbb{P}(B|A)$? Veremos que, em alguns casos sim. *Nestes casos* em que $\mathbb{P}(B) = \mathbb{P}(B|A)$ a ocorrência de A não afeta as chances da ocorrência de B . Entretanto, na maioria das vezes, teremos $\mathbb{P}(B) \neq \mathbb{P}(B|A)$. Vamos querer identificar estes casos e saber, por exemplo, quando temos $\mathbb{P}(B) < \mathbb{P}(B|A)$. Mais do que isto, queremos uma fórmula que nos permita calcular de maneira exata $\mathbb{P}(B|A)$ em qualquer situação.

Alguns casos óbvios

Alguns casos são fáceis de calcular pois eles são casos extremos. Por exemplo, lançar um dado bem equilibrado e anotar a face: $\Omega = \{1, 2, \dots, 6\}$. Seja $B = \{4, 5, 6\}$ com $\mathbb{P}(B) = 3/6$. Vamos considerar um evento $A \subset B$. Por exemplo, $A = \{5, 6\}$. Intuitivamente, o que deveria ser $\mathbb{P}(B|A)$? Qual a probabilidade de que a face seja 4, 5 ou 6 sabendo que saiu 5 ou 6? Ao saber que um resultado $\omega \in A$ ocorreu, automaticamente inferimos que B também ocorreu pois $A \subset B$. Assim, devemos ter $\mathbb{P}(B|A) = 1$. Observe que, neste caso, $\mathbb{P}(B|A) = 1 > \mathbb{P}(B) = 3/6$. De qualquer modo, o caso óbvio é que, se $A \subset B$, então $\mathbb{P}(B|A) = 1$.

Outro caso óbvio: $A \cap B = \emptyset$. Intuitivamente, o que deveria ser $\mathbb{P}(B|A)$? Se o evento que ocorreu está em A , ele não pode estar em B pois A e B são disjuntos. Assim, ao saber que um evento $\omega \in A$ ocorreu, automaticamente inferimos que B não ocorreu. Assim, devemos ter $\mathbb{P}(B|A) = 0 \leq \mathbb{P}(B)$.

■ **Example 4.1** Exemplo do dado equilibrado com $\Omega = \{1, 2, \dots, 6\}$. Seja B o evento FACE PAR. Isto é, $B = \{2, 4, 6\}$ e $\mathbb{P}(B) = \frac{1}{2}$. Seja $A = \{5\}$. É claro que $A \cap B = \emptyset$. Intuitivamente, se ocorreu a face 5, qual a chance de ocorrer uma face par? Esta chance é zero. Ou você apostaria na ocorrência de B neste caso? ■

Assim, dois casos intuitivamente óbvios são:

- Se $A \subset B$ então $\mathbb{P}(B|A) = 1$.

- Se $A \cap B = \emptyset$ então $\mathbb{P}(B|A) = 0$.

E o caso geral? Sejam A e B dois eventos com $\mathbb{P}(A) > 0$ e com $A \cap B \neq \emptyset$. Como calcular $\mathbb{P}(B|A)$?

Definição 4.1.1 — Probabilidade Condisional. Sejam A e B dois eventos com $\mathbb{P}(A) > 0$. Então, por definição,

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} \quad (4.1)$$

Assim, para calcular a probabilidade de que B ocorreu *dado que* A ocorreu:

- Calcule a probabilidade $\mathbb{P}(A \cap B)$ de que A e B tenham ambos ocorrido
- Aumente esta probabilidade multiplicando-a por $1/\mathbb{P}(A) > 1$.

■ **Example 4.2** Considere o lançamento de uma moeda honesta 5 vezes seguidas:

$$\begin{aligned} \Omega &= \{CCCCC, CCCCC\tilde{C}, \dots, \tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}\} \\ &\hookrightarrow 32 \text{ elementos} \end{aligned}$$

Temos $\mathbb{P}(\omega) = 1/32$, igualmente prováveis. Seja $B = \{\omega \in \Omega ; 1^{\text{o}} \text{ elemento é } C\}$ Temos $\mathbb{P}(B) = 1/2$ pois 16 elementos em Ω têm C na 1^a posição. Como são igualmente prováveis, $\mathbb{P}(B) = 16/32 = 1/2$.

É fornecida a seguinte informação: ocorreu $A = \{ \text{Houve apenas uma coroa nos 5 lançamentos} \}$. Intuitivamente, $\mathbb{P}(B|A) > \mathbb{P}(B) = 1/2$. De fato, calculando $\mathbb{P}(B|A)$ pela definição:

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{4/32}{5/32} = \frac{4}{5}$$

Assim a probabilidade mudou bastante ao sabermos que A ocorreu:

$$\mathbb{P}(B) = \frac{1}{2} \rightsquigarrow \mathbb{P}(B|A) = \frac{4}{5}$$

Saber que ocorreu apenas uma coroa em cinco lançamentos torna altamente provável que a 1^a posição seja cara. ■

4.1.4 Ciência dos dados e condicional

Este é um dos grandes objetivos gerais de ciência dos dados. Quando tivermos um sistema complexo, envolvendo vários fatores, obtemos a um custo baixo algumas informações. Estas informações são representadas por A . Usamos estas informações de baixo custo para recalcular as probabilidades de eventos B que não sabemos se ocorreram: $\mathbb{P}(B|A)$. Com estas probabilidades recalculadas podemos tomar decisões. Não sabemos se B ocorreu por várias possíveis razões:

- porque estão no futuro,
- porque são caros demais para se observar,
- porque são impossíveis de se observar (saber o sentimento de uma pessoa),
- ou é anti-ético conhecer.

4.1.5 Intuição para a definição

Vimos a definição $\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$. Por quê usamos a fórmula acima? Por quê esta definição, e não outra tal como $\frac{\mathbb{P}(A \cup B)}{\mathbb{P}(A)}$ ou $\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$? A resposta é: para ser consistente com a experiência, com o conhecimento empírico.

Para ver isto, vamos encontrar $\mathbb{P}(B|A)$ de duas formas distintas num caso simples de simular no computador. Uma das formas será através da contagem do evento B dentre aqueles casos em que A ocorre. Esta é a forma natural de estimar probabilidades: pela frequência relativa.

A segunda forma será pela definição $\mathbb{P}(B|A) = \mathbb{P}(A \cap B)/\mathbb{P}(A)$. Veremos que as duas coincidem e portanto que a definição (4.1) é o que tem de ser se quisermos ser consistentes com nossa experiência cotidiana.

Role um dado bem balanceado duas vezes. Então $\Omega = \{(1,1), (1,2), \dots, (6,6)\}$ e $\mathbb{P}(\omega) = \frac{1}{36}$ para todo $\omega \in \Omega$. Seja $B = [1^{\circ} \text{ dado é um } 6]$. Então $B = \{(6,1), (6,2), (6,3), (6,4), (6,5), (6,6)\}$ e $\mathbb{P}(B) = 6/36 = 1/6$.

Seja $A = [\text{Soma das faces é maior que } 8]$. Temos

$$A = \{(3,6), (4,5), (4,6), (5,4), (5,5), (5,6), (6,3), (6,4), (6,5), (6,6)\}$$

$$\text{e } \mathbb{P}(A) = 10/36 = 0.28.$$

Quanto é $\mathbb{P}(B|A)$? Devemos esperar que seja maior ou menor que $\mathbb{P}(B)$? A soma das faces varia de 2 a 12. Ser maior que 8 quer dizer que é um valor alto e que podemos esperar que as duas faces sejam pelo menos moderadamente alta. Com certeza, se uma delas for 1 ou 2 não poderemos ter a soma das faces maior que 8.

De fato, usando a fórmula,

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} = \frac{4/36}{10/36} = 0.4 > 1/6 = 0.17 = \mathbb{P}(B)$$

Vamos agora calcular $\mathbb{P}(B|A)$ simulando os dados num computador. Não usaremos a fórmula (4.1) mas frequências relativas e veremos que vamos obter o mesmo resultado dados por (4.1).

Replique os lançamentos duplos um grande número N de vezes (por exemplo, $N = 100$ mil). Vamos apresentar os resultados na Tabela 4.1.

Repetição	1	2	3	4	5	6	7	8	9	10	...
Dado 1	2	5	5	2	6	4	2	1	6	6	...
Dado 2	1	5	1	3	1	5	3	6	4	3	...
B ocorreu?	N	N	N	N	Y	N	N	N	Y	Y	...
A ocorreu?	N	Y	N	N	N	Y	N	N	Y	Y	...

Table 4.1: Lançamentos duplos de dados

Vamos considerar apenas as vezes em que A ocorreu. Na minha simulação eu obtive 13886 vezes. *Dentre estas 13886 ocorrências*, verifique quantas vezes o evento B ocorreu. Eu obtive 5623 vezes. É natural esperarmos $\mathbb{P}(B|A) \approx 5623/13886 = 0.405$. Por quê? Considerando apenas as 13886 vezes em que A ocorreu, verificamos qual a proporção de vezes que ocorreu B . Esta é a maneira de estimar empiricamente, apenas com dados, o valor de $\mathbb{P}(B|A)$. Pela frequência relativa da ocorrência do evento B *dado que o evento A ocorreu*.

Vamos agora estimar $\mathbb{P}(B|A)$ de outra forma: considerando o numerador e o denominador da definição. $\mathbb{P}(B|A) = \mathbb{P}(A \cap B)/\mathbb{P}(A)$. Sabemos que

$$\mathbb{P}(A) \approx \frac{\text{nº de vezes que } A \text{ ocorreu}}{N}.$$

Assim,

$$\mathbb{P}(A) \approx \frac{13886}{N} \quad \text{ou} \quad 13886 \approx N \mathbb{P}(A)$$

Do mesmo modo, pela interpretação de probabilidade como frequência em longas repetições,

$$\mathbb{P}(A \cap B) \approx \frac{\text{nº de vezes em que } A \text{ e } B \text{ ocorrem}}{N}$$

Mas A e B ocorrem 5623 vezes em N . Lembre-se nós separamos os 13886 casos em que A ocorreu e depois contamos os casos em que B ocorreu entre estes 13886 casos)

Então $\mathbb{P}(A \cap B) \approx \frac{5623}{N} \Rightarrow N \mathbb{P}(A \cap B) \approx 5623$

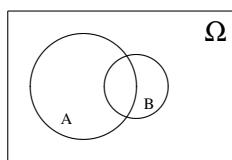
Desse modo,

$$\mathbb{P}(B|A) \approx \frac{5623}{13886} \approx \frac{N \mathbb{P}(A \cap B)}{N \mathbb{P}(A)} = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$$

Nossa conclusão é que: se quisermos manter intacta nossa idéia de que a probabilidade de um evento é aproximadamente igual à sua frequência relativa numa longa série de repetições independentes, então a definição da probabilidade condicional $\mathbb{P}(B|A)$ tem de ser $\mathbb{P}(A \cap B)/\mathbb{P}(A)$. Nenhuma outra definição vai gerar resultados consistentes com os experimentos que fizemos.

4.2 Diagramas de Venn

É comum representarmos eventos usando diagramas de conjuntos de Venn.



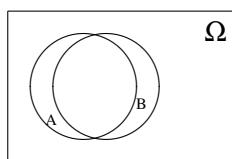
Ω é o retângulo maior envolvente. Os eventos são figuras com tamanhos proporcionais à sua probabilidade. Qual o valor aproximado de $\mathbb{P}(A)$?

- $\mathbb{P}(A) \approx 0.90?$
- $\mathbb{P}(A) \approx 1/4?$
- $\mathbb{P}(A) \approx 1/8?$
- $\mathbb{P}(A) \approx 0.01?$

Com as mesmas opções, qual o valor aproximado de $\mathbb{P}(B)$?

4.2.1 Probabilidade condicional no diagrama de Venn

Como enxergar a definição $\mathbb{P}(B|A) = \mathbb{P}(A \cap B)/\mathbb{P}(A)$ neste diagrama? $\mathbb{P}(B|A)$ é o tamanho de $A \cap B$ relativamente ao tamanho de A .

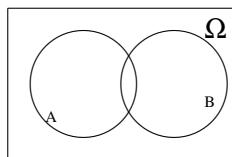


Responda: qual o valor aproximado de $\mathbb{P}(B|A)$?

- $\mathbb{P}(B|A) \approx 0.85?$
- $\mathbb{P}(B|A) \approx 1/3?$
- $\mathbb{P}(B|A) \approx 1/8?$
- $\mathbb{P}(B|A) \approx 0.05?$

Temos $\mathbb{P}(B|A)$ bem maior que $\mathbb{P}(B) \approx 1/3$. Assim, a resposta correta seria $\mathbb{P}(B|A) \approx 0.85$.

Considere agora a seguinte situação:

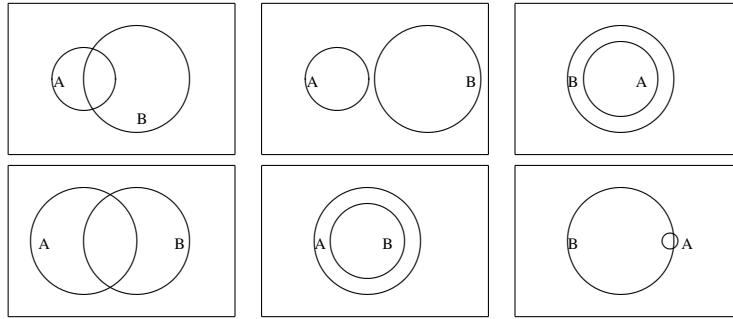


Qual o valor aproximado de $\mathbb{P}(B|A)$?

- $\mathbb{P}(B|A) \approx 0.85?$
- $\mathbb{P}(B|A) \approx 1/3?$
- $\mathbb{P}(B|A) \approx 1/8?$
- $\mathbb{P}(B|A) \approx 0.05?$

Temos $\mathbb{P}(B|A)$ bem menor que $\mathbb{P}(B) \approx 1/3$.

Em todos os casos abaixo temos $\mathbb{P}(B) \approx 1/5$. Obtenha $\mathbb{P}(B|A)$ aproximadamente em cada caso:

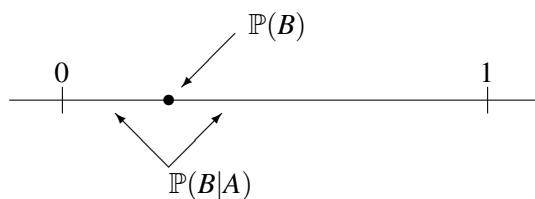


O diagrama (1,1) tem $\mathbb{P}(B|A) \approx \dots 0.40$. Em (1,2) : $\mathbb{P}(B|A) \dots = 0$ (1,3) : $\mathbb{P}(B|A) \dots = 1.0$. O diagrama (2,1) tem $\mathbb{P}(B|A) \approx \dots 0.40$, como no diagrama (1,1). E no diagrama (2,2), temos $\mathbb{P}(B|A) \dots 0.85$ (2,3) : $\mathbb{P}(B|A) \dots 0.9$.

4.2.2 $\mathbb{P}(B|A)$ e $\mathbb{P}(B)$

Se $A \subset B$ então $\mathbb{P}(B|A) = 1$. A informação de que A ocorreu torna *certa* a ocorrência de um resultado $\omega \in B$. Se $A \cap B = \emptyset$ então $\mathbb{P}(B|A) = 0$. A informação de que A ocorreu torna *impossível* a ocorrência de qualquer $\omega \in B$. Estas são *situações extremas*: saber que A ocorreu leva a um conhecimento sem incerteza sobre a ocorrência de B .

Na maioria das vezes, saber que A ocorreu não vai eliminar a incerteza sobre a ocorrência de B . Teremos $0 < \mathbb{P}(B|A) < 1$



Podemos ter $\mathbb{P}(B|A) > \mathbb{P}(B)$ ou $\mathbb{P}(B|A) < \mathbb{P}(B)$.

4.3 Independência de eventos

Há um outro caso importante: quando saber que A ocorreu não tem qualquer influência na incerteza sobre a ocorrência de B . Isto é, existem casos em que $\mathbb{P}(B|A) = \mathbb{P}(B)$

Definition 4.3.1 Dizemos que A e B são *eventos independentes* se $\mathbb{P}(B|A) = \mathbb{P}(B)$.

Esta definição de independência é equivalente a esta outra: A e B são eventos independentes se, e somente se, $\mathbb{P}(A \cap B) = \mathbb{P}(A) \mathbb{P}(B)$ pois

$$\begin{array}{ccc}
 \overbrace{\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}} & = & \mathbb{P}(B|A) \\
 & \downarrow & \\
 \text{Definição de} & & \text{Se forem} \\
 \text{Probabilidade} & & \text{independentes} \\
 \text{Condisional} & & \\
 & \downarrow & \\
 & \mathbb{P}(A \cap B) = \mathbb{P}(A) * \mathbb{P}(B) &
 \end{array}$$

4.3.1 Como surge a independência?

A independência de eventos pode surgir de duas formas distintas. Ela pode surgir porque nós *supomos* que os eventos são independentes. Por exemplo, pensando sobre o mecanismo físico envolvido, supomos que lançamentos sucessivos de uma moeda são independentes: a moeda não tem memória do que aconteceu. Assim $\mathbb{P}(\text{Cara no 2o.} | \text{Cara no 1o.}) = \mathbb{P}(\text{Cara no 2o.}) = 1/2$. Não deduzimos matematicamente que os eventos são independentes. Nós assumimos que eles são independentes pensando sobre o mecanismo envolvido no lançamento da moeda.

A outra forma pela qual a independência de eventos pode surgir é quando *verificamos* matematicamente que $\mathbb{P}(A \cap B) = \mathbb{P}(A) \mathbb{P}(B)$ ou que $\mathbb{P}(B|A) = \mathbb{P}(B)$. As vezes, não podemos intuir facilmente que A e B são independentes. Nestes casos, calculamos $\mathbb{P}(B|A)$ e $\mathbb{P}(B)$ e, voilá: se as probabilidades forem iguais, os eventos são independentes.

■ **Example 4.3 — Independência de eventos.** Alguns exemplos óbvios de independência: em repetições de certos experimentos como rolar um dado duas vezes e anotar o resultado. Eventos relacionados apenas ao primeiro lançamento devem ser independentes de eventos relacionados apenas ao segundo lançamento do dado. Isto é intuitivo a partir de nossa experiência com este tipo de situação. As probabilidades devem se manter as mesmas: rolar um dado não o modifica fisicamente a ponto de afetar as probabilidades das 6 faces. Além disso, o dado não tem memória do que saiu antes de forma que um lançamento não afeta o seguinte.

Mas podemos verificar matematicamente esta intuição checando a validade da condição $\mathbb{P}(A \cap B) = \mathbb{P}(A) \mathbb{P}(B)$. Temos $\Omega = \{(1,1), \dots, (6,6)\}$ com $\mathbb{P}(\omega) = 1/36$. Se A é o evento de que a primeira rolagem é par então

$$A = \{(2,1), \dots, (2,6), (4,1), \dots, (4,6), (6,1), \dots, (6,6)\}$$

e $\mathbb{P}(A) = 18/36 = 1/2$. Seja agora B o evento de que a segunda rolagem é divisível por 3. Assim, $B = \{(1,3), \dots, (6,3), (1,6), \dots, (6,6)\}$. e $\mathbb{P}(B) = 12/36 = 1/3$.

Temos $A \cap B = \{(2,3), (4,3), (6,3), (2,6), (4,6), (6,6)\}$ e, como esperado, A e B são independentes:

$$\mathbb{P}(A \cap B) = 6/36 = 1/6 = \mathbb{P}(A) \mathbb{P}(B)$$

■

■ **Example 4.4 — Exemplo menos óbvio.** Rola-se um dado bem equilibrado uma única vez. Seja $A = \{2, 4, 6\}$ e $B = \{1, 2, 3, 4\}$. Temos $\mathbb{P}(A) = 1/2$ e $\mathbb{P}(B) = 2/3$. Além disso, $A \cap B = \{2, 4\}$ e $\mathbb{P}(A \cap B) = 1/3$.

Como

$$\mathbb{P}(A \cap B) = \frac{1}{3} = \frac{1}{2} \frac{2}{3} = \mathbb{P}(A) \mathbb{P}(B),$$

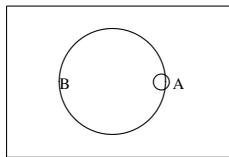
concluímos que os eventos A e B são independentes. É muito difícil alguém conseguir enxergar isto intuitivamente. ■

4.3.2 Independência no diagrama de Venn

Se A e B são eventos disjuntos num diagrama de Venn, eles não são independentes. Se são disjuntos então $A \cap B = \emptyset$ e portanto $\mathbb{P}(A \cap B) = 0$. Se $\mathbb{P}(A)$ e $\mathbb{P}(B)$ são > 0 então $0 = \mathbb{P}(A \cap B) \neq \mathbb{P}(A) \mathbb{P}(B)$. Portanto, eles não podem ser eventos independentes.

Se $A \subset B$ então $\mathbb{P}(A \cap B) = \mathbb{P}(A) \neq \mathbb{P}(A) \mathbb{P}(B)$ e portanto A e B não são independentes.

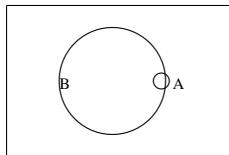
Exceto nestes dois casos, é difícil verificar visualmente se A e B são independentes num diagrama de Venn. Teríamos de ser capazes de ver que o tamanho de $A \cap B$ relativamente a Ω é igual ao produto das proporções dos tamanhos de A e de B . Entretanto, se $\mathbb{P}(B|A)$ for muito diferente de $\mathbb{P}(B)$ poderemos dizer com segurança que A e B não são independentes. Por exemplo, sem fazer nenhuma conta podemos dizer que A e B não são independentes neste caso:



Visualmente é óbvio que $\mathbb{P}(B) \approx 1/3$ mas que $\mathbb{P}(B|A) \approx 1$. Portanto, a ocorrência de A aumenta as chances da ocorrência de B . Explicação: A é um evento raro pois $\mathbb{P}(A) \approx 0$. Entretanto, a maior parte de A está em B . Se o raro evento A ocorrer, é altamente provável que seja um dos $\omega \in A \cap B$.

4.4 Regra de Bayes

As probabilidades $\mathbb{P}(A|B)$ e $\mathbb{P}(B|A)$ podem ser completamente diferentes. Por exemplo, veja o diagrama de Venn abaixo:



Temos $\mathbb{P}(B|A) \approx 1$ mas $\mathbb{P}(A|B) \approx 1/25$.

Uma outra situação meio ridícula mas que mostra como elas podem ser completamente diferentes:

$$\mathbb{P}(\text{ser Drácula} \mid \text{não dorme a noite}) \approx 0$$

mas

$$\mathbb{P}(\text{não dorme a noite} \mid \text{ser Drácula}) \approx 1.$$

Existe uma relação matemática muito simples entre $\mathbb{P}(A|B)$ e $\mathbb{P}(B|A)$. Temos

$$\begin{aligned}\mathbb{P}(B|A) &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} \Rightarrow \mathbb{P}(A \cap B) = \mathbb{P}(B|A) \mathbb{P}(A) \\ \mathbb{P}(A|B) &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \Rightarrow \mathbb{P}(A \cap B) = \mathbb{P}(A|B) \mathbb{P}(B)\end{aligned}$$

Igualando as duas expressões para $\mathbb{P}(A \cap B)$ temos

$$\mathbb{P}(B|A) \mathbb{P}(A) = \mathbb{P}(A|B) \mathbb{P}(B)$$

Definition 4.4.1 — Regra de Bayes.

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B) \mathbb{P}(B)}{\mathbb{P}(A)}$$

O principal uso da regra de Bayes é quando temos uma das probabilidades condicionais, digamos $\mathbb{P}(A|B)$, e queremos calcular a inversa: $\mathbb{P}(B|A)$. Neste caso,

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B) \mathbb{P}(B)}{\mathbb{P}(A)}$$

Para isto precisamos também das probabilidades não-condicionais $\mathbb{P}(A)$ e $\mathbb{P}(B)$.

4.4.1 Bayes e teste diagnóstico

A população brasileira deve ser rastreada com um teste para detecção do vírus HIV? Acredita-se que 0.1% são HIV+ $\Rightarrow \approx 200$ mil dentre 200 milhões. Um teste diagnóstico para testar a presença do vírus não é perfeito. Por causa disso, podemos definir a tabela de confusão considerando a situação real o resultado indicado pelo teste diagnóstico:

Vírus?	Resultado do Teste	
	$T+$	$T-$
$V+$	ok	erro
$V-$	erro	ok

Um paciente recebe o resultado do teste e ele é positivo ou $T+$. A questão é: ele tem de fato o vírus (é $V+$) ou aconteceu um erro? O principal problema médico é calcular $\mathbb{P}(V+|T+)$. Como obter isto? Pela regra de Bayes pois temos $\mathbb{P}(T+|V+)$.

4.4.2 Sensitividade e Especificidade

Aplica-se o teste diagnóstico em dois grandes grupos de indivíduos:

- um em que sabidamente todos possuem o vírus HIV
- outro em que sabidamente eles não possuem o vírus HIV

A probabilidade de qualquer evento A ocorrer no primeiro grupo será $\mathbb{P}(A|V+)$. No segundo grupo será $\mathbb{P}(A|V-)$.

Definition 4.4.2 — Sensitividade e Especificidade. Com base na frequência daqueles que respondem $T+$ em cada grupo o laboratório que produz o teste estima as seguintes quantidades:

- **Sensitividade** : $\mathbb{P}(T+|V+) \approx 0.99\%$
- **Especificidade** : $\mathbb{P}(T-|V-) \approx 0.95\%$

Quanto maior estas duas probabilidades, melhor. Idealmente, gostaríamos que ambos fossem iguais a 1. Mas na prática, os testes de diagnósticos podem cometer erros.

Motivo para os nomes:

- O teste é sensível à presença do vírus: se o vírus estiver presente, o teste é +?
- O teste é específico para o vírus HIV: se o paciente tiver qualquer outra coisa que não seja o vírus, o teste não deveria dar positivo.

Suponha que $\mathbb{P}(T+|V+) = 0.99$ (sensibilidade) e $\mathbb{P}(T-|V-) = 0.95$ (especificidade). As duas probabilidades são relativamente altas, o que é bom. Mas veremos que isto não é suficiente para um teste em massa.

Definition 4.4.3 As probabilidades complementares estão associadas a erros de diagnóstico e os médicos usam dois termos para eles:

- *Falso positivo* (FP): $T+$ para um paciente que é $V-$
- *Falso negativo* (FN): $T-$ para um paciente que é $V+$

As probabilidades de FP e FN são obtidas diretamente da sensibilidade e especificidade:

$$\mathbb{P}(FP) = \mathbb{P}(T+|V-) = 0.05 = 1 - 0.95 = 1 - \text{especificidade}$$

$$\mathbb{P}(FN) = \mathbb{P}(T-|V+) = 0.01 = 1 - 0.99 = 1 - \text{sensibilidade}$$

A partir das frequências na tabela de confusão, podemos estimar essas probabilidades:

Vírus?	Resultado do Teste		Total
	$T+$	$T-$	
$V+$	sens $\mathbb{P}(T+ V+)$	$1 - \text{sens}$ $\mathbb{P}(FN) = \mathbb{P}(T- V+)$	1.0
$V-$	1-esp $\mathbb{P}(FP) = \mathbb{P}(T+ V-)$	esp $\mathbb{P}(T- V-)$	1.0

Mas não queremos apenas $\mathbb{P}(FP) = \mathbb{P}(T+|V-)$ e $\mathbb{P}(FN) = \mathbb{P}(T-|V+)$. Mais importante é calcular as probabilidades inversas. O médico tem em mãos o resultado $T+$ do exame. Dado que ele tem este resultado $T+$, qual a probabilidade de que o paciente tenha o vírus? Isto é, qual o valor de $\mathbb{P}(V+|T+)$? Do mesmo modo, queremos saber qual o valor da probabilidade $\mathbb{P}(V-|T-)$. De posse de uma estimativa de $\mathbb{P}(V+)$, usamos a regra de Bayes para obter estas probabilidades inversas.

Temos $\mathbb{P}(V+) = 0.001$, uma estimativa grosseira. Esta é a estimativa da prevalência do vírus na população em geral. Se não soubermos este valor, podemos calcular as probabilidades com diversos cenários plausíveis para $\mathbb{P}(V+)$ e ver como as probabilidades se modificam (talvez elas não mudem muito).

Pela regra de Bayes, temos:

$$\mathbb{P}(V+|T+) = \frac{\mathbb{P}(T+|V+)\mathbb{P}(V+)}{\mathbb{P}(T+)} = \frac{0.99 * 0.001}{\mathbb{P}(T+)}$$

Para obter $\mathbb{P}(T+)$ usamos um truque muito útil baseado em interseção de conjuntos. O lado esquerdo da Figura 4.1 mostra o evento $T+$ e a decomposição de Ω obtida pela partição $\Omega = [V+] \cup [V-]$.

No lado direito da Figura 4.1, o evento $T+$ e a decomposição $\Omega = [V+] \cup [V-]$ são misturados. Escrevemos o evento $[T+]$ como $[T+] = ([T+] \cap [V+]) \cup ([T+] \cap V-)$.

Agora podemos calcular $\mathbb{P}(T+)$ facilmente:

$$\begin{aligned} \mathbb{P}(T+) &= \mathbb{P}(T+ \cap (V+ \cup V-)) \\ &= \mathbb{P}((T+ \cap V+) \cup (T+ \cap V-)) \\ &= \mathbb{P}(T+ \cap V+) + \mathbb{P}(T+ \cap V-) \\ &= \mathbb{P}(T+|V+) \times \mathbb{P}(V+) + \mathbb{P}(T+|V-) \times \mathbb{P}(V-) \\ &= 0.99 \times 0.001 + 0.05 \times 0.999 \\ &= 0.05094 \end{aligned}$$

Finalizando o cálculo com a regra de Bayes, temos:

$$\mathbb{P}(V+|T+) = \frac{0.99 \times 0.001}{0.05094} = 0.019$$

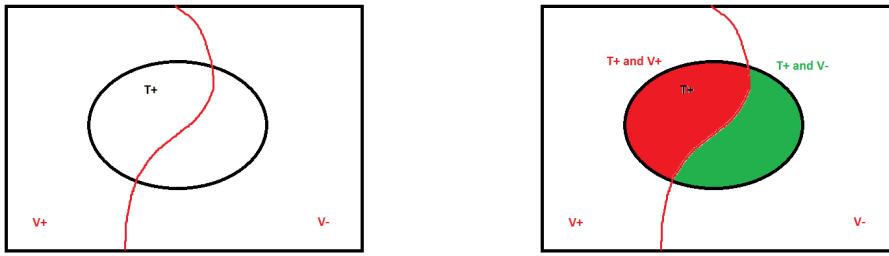


Figure 4.1: Esquerda: Evento $[T+]$ e Ω decomposto como $\Omega = [V+] \cup [V-]$. Direita: Decompondo o evento $[T+] = ([T+] \cap [V+]) \cup ([T+] \cap V-)$.

Assim, se tivermos um $T+$, será muito alta a chance do paciente ser $V-$ ou não ter o vírus. Apenas 2% dos indivíduos com teste positivo ($T+$) possuem o vírus de fato.

Idem, calculando a outra probabilidade inversa:

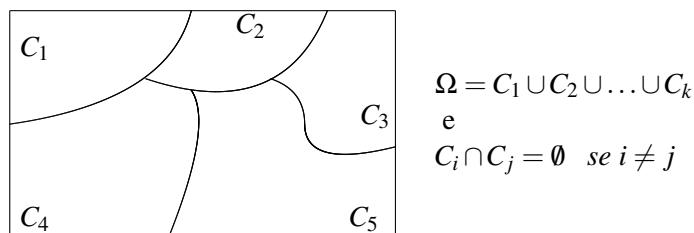
$$\begin{aligned}\mathbb{P}(V-|T-) &= \frac{\mathbb{P}(T-|V-)\mathbb{P}(V-)}{\mathbb{P}(T-)} \\ &= \frac{0.95 \times (1 - 0.001)}{1 - 0.05094} = 0.9999895\end{aligned}$$

Se o teste for negativo, é praticamente certo que o indivíduo será $V-$.

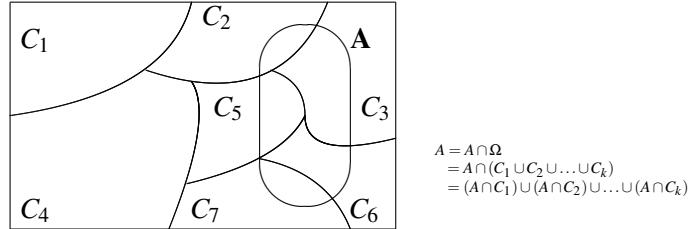
Estes cálculos mostram por que não fazemos um rastreamento em massa na população brasileira. Como $\mathbb{P}(V+|T-) \approx 0$ então se o teste não detecta, a chance de estar infectado é baixa. OK, isto é bom. Mas $\mathbb{P}(V-|T+) \approx 1$. Isto é, quase todos detectados pelo teste não estão infectados. Quantas pessoas dariam positivas (falsamente ou corretamente) pelo teste? Isto é, quantos teriam $T+?$ Aproximadamente $200 \text{ milhões} \times \mathbb{P}(T+) \approx 10 \text{ milhões}$, um número enorme. Destes, 98% (ou 9.8 milhões) não têm HIV: a imensa maioria de um número enorme de pessoas. As dificuldades de garantir um teste em todos e o custo envolvido levam a outra estratégia: fazer uma busca ativa entre pessoas de grupos de risco (que teriam $\mathbb{P}(V+)$ bem maior).

4.4.3 Regra da probabilidade total

Na regra de Bayes, derivamos uma fórmula muito útil, chamada fórmula da probabilidade total. Vamos ver o caso geral. O espaço amostral Ω é partitionado nos eventos C_1, C_2, \dots, C_k .



Para qualquer evento A temos

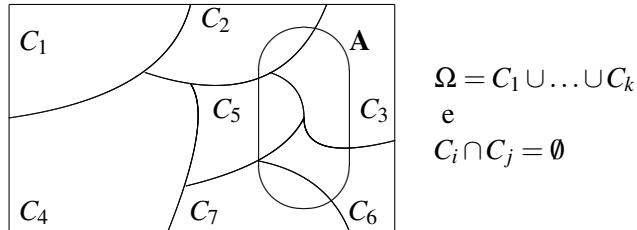


Temos então

$$\begin{aligned} \mathbb{P}(A) &= \mathbb{P}(A \cap C_1) + \dots + \mathbb{P}(A \cap C_k) \\ &= \mathbb{P}(A|C_1)\mathbb{P}(C_1) + \dots + \mathbb{P}(A|C_k)\mathbb{P}(C_k) \end{aligned}$$

4.4.4 Extensão da Regra de Bayes

O espaço amostral Ω é particionado nos eventos C_1, \dots, C_k :



Temos

$$\begin{aligned} \mathbb{P}(C_i|A) &= \frac{\mathbb{P}(A|C_i)\mathbb{P}(C_i)}{\mathbb{P}(A)} \\ &= \frac{\mathbb{P}(A|C_i)\mathbb{P}(C_i)}{\mathbb{P}(A \cap C_1) + \dots + \mathbb{P}(A \cap C_k)} \\ &= \frac{\mathbb{P}(A|C_i)\mathbb{P}(C_i)}{\mathbb{P}(A|C_1)\mathbb{P}(C_1) + \dots + \mathbb{P}(A|C_k)\mathbb{P}(C_k)} \end{aligned}$$

Isto é,

$$\mathbb{P}(C_i|A) = \frac{\mathbb{P}(A|C_i)\mathbb{P}(C_i)}{\sum_{j=1}^k \mathbb{P}(A|C_j)\mathbb{P}(C_j)}$$

Esta é a fórmula geral da regra de Bayes.

■ **Example 4.5 — Regra de Bayes geral.** Em uma urna, existem 6 bolas de cores desconhecidas. Três bolas são retiradas sem reposição e são pretas. Ache a probabilidade de que não restam bolas pretas na urna.

Vamos definir $A = 3$ bolas pretas são retiradas. Seja C_i o evento de que existem i bolas pretas na urna com $i = 0, 1, \dots, 6$. A urna tem 6 bolas e retiramos 3 bolas ao acaso.

Quanto é $\mathbb{P}(A|C_i)$? Esta parte é fácil:

$$\left\{ \begin{array}{l} \mathbb{P}(A|C_0) = \frac{\mathbb{P}(A \cap C_0)}{\mathbb{P}(C_0)} = \frac{0}{\mathbb{P}(C_0)} = 0 \\ \mathbb{P}(A|C_1) = 0 \\ \mathbb{P}(A|C_2) = 0 \\ \mathbb{P}(A|C_3) = 1/20 \text{ (Hipergeométrica)} \\ \mathbb{P}(A|C_4) = 1/5 \\ \mathbb{P}(A|C_5) = 1/2 \\ \mathbb{P}(A|C_6) = 1 \end{array} \right.$$

Queremos calcular $\mathbb{P}(C_3|A)$. Temos

$$\begin{aligned} \mathbb{P}(C_3|A) &= \frac{\mathbb{P}(A|C_3) * \mathbb{P}(C_3)}{\sum_{j=0}^6 \mathbb{P}(A|C_j) * \mathbb{P}(C_j)} \\ &= \frac{\frac{1}{20} * \mathbb{P}(C_3)}{0 + 0 + 0 + \frac{1}{20}\mathbb{P}(C_3) + \frac{1}{5}\mathbb{P}(C_4) + \frac{1}{2}\mathbb{P}(C_5) + 1\mathbb{P}(C_6)} \\ &= ?? \end{aligned}$$

Precisamos estabelecer o valor de $\mathbb{P}(C_j)$, a probabilidade de que existam j bolas pretas na urna. Isto depende do mecanismo que colocou bolas na urna e isto não foi explicado no problema. Vamos mostrar algumas possibilidades para $\mathbb{P}(C_j)$.

Um primeiro cenário é que qualquer número de bolas pretas entre 0 e 6 tem a mesma probabilidade. Então $\mathbb{P}(C_j) = \frac{1}{7}$ para todo j . Basta substituir este valores agora na fórmula acima para $\mathbb{P}(C_3|A)$.

Como segundo cenário, suponha que a bolas são escolhidas preferencialmente de uma única cor. Então os valores de $\mathbb{P}(C_j)$ para $j = 0$ e $j = 6$ seriam os maiores, com um valor mínimo com $j = 3$. Por exemplo, $\mathbb{P}(C_j) = \frac{1}{28}(j-3)^2$.

Outra opção de cenário é a seguinte: Existem 10 cores distintas e a cor de cada bola é escolhida ao acaso. A chance de colocar uma bola preta na urna é $1/10$. A chance de colocar j bolas pretas na urna de 6 bolas é

$$\mathbb{P}(C_j) = \binom{6}{j} (0.1)^j (0.9)^{6-j}$$

■

4.5 Condicional como nova medida de probabilidade

Seja $(\Omega, \mathcal{A}, \mathbb{P})$ um espaço de probabilidade qualquer. Vamos fixar um evento $A \subseteq \Omega$ com $\mathbb{P}(A) > 0$. Sabemos que, para qualquer evento $B \subseteq \Omega$, calculamos $\mathbb{P}(B|A)$ através da definição $\mathbb{P}(B|A) = \mathbb{P}(A \cap B)/\mathbb{P}(A)$. Ao invés de pensarmos nesta probabilidade condicional apenas para um par de eventos A e B , podemos fixar A e calcular a probabilidade condicional dado A para uma série de eventos B_1, B_2, B_3, \dots . Na verdade, queremos fixar A e recalcula a probabilidade (condicionada à ocorrência de A) do evento B para todo e qualquer $B \subseteq \Omega$. Isto é, queremos atribuir uma nova medida de probabilidade \mathbb{P}_A aos eventos da σ -álgebra \mathcal{A} . Ou ainda de outro modo, queremos criar um novo espaço de probabilidade $(\Omega, \mathcal{A}, \mathbb{P}_A)$ onde Ω e \mathcal{A} são os mesmos do espaço de

probabilidade original mas a função de probabilidade muda de \mathbb{P} para \mathbb{P}_A . Agora, para cada evento B na σ -álgebra \mathcal{A} , calculamos sua probabilidade como

$$\mathbb{P}_A(B) = \mathbb{P}(B|A) = \mathbb{P}(A \cap B)/\mathbb{P}(A).$$

Veja que estamos definindo uma função de probabilidade $\mathbb{P}_A(\cdot)$ com argumento em \mathcal{A} :

$$\begin{aligned}\mathbb{P}_A : \mathcal{A} &\longrightarrow [0, 1] \\ B &\longrightarrow \mathbb{P}_A(B) = \mathbb{P}(B|A)\end{aligned}$$

Para que esta função seja uma atribuição de probabilidade aos eventos $B \subseteq \mathcal{A}$ válida, ela deve obedecer aos três axiomas de Kolmogorov:

Axioma 1 $\mathbb{P}_A(B) \geq 0 \quad \forall B \in \mathcal{A}$

Axioma 2 $\mathbb{P}_A(\Omega) = 1$

Axioma 3 $\mathbb{P}_A(B_1 \cup B_2 \cup B_3 \cup \dots) = \mathbb{P}_A(B_1) + \mathbb{P}_A(B_2) + \mathbb{P}_A(B_3) + \dots$ se os eventos B_1, B_2, \dots forem disjuntos (isto é, mutuamente exclusivos).

Isto realmente acontece pois:

- $\mathbb{P}_A(B) = \mathbb{P}(A \cap B)/\mathbb{P}(A) \geq 0$ pois $\mathbb{P}(A \cap B) \geq 0$ e $\mathbb{P}(A) > 0$.

- $\mathbb{P}_A(\Omega) = \mathbb{P}(\Omega \cap A)/\mathbb{P}(A) = \mathbb{P}(A)/\mathbb{P}(A) = 1$

- Quanto à última propriedade, se B_1, B_2, \dots são disjuntos,

$$\mathbb{P}_A(B_1 \cup B_2 \cup B_3 \cup \dots) = \frac{\mathbb{P}([B_1 \cup B_2 \cup B_3 \cup \dots] \cap A)}{\mathbb{P}(A)} \tag{4.2}$$

$$= \frac{\mathbb{P}((B_1 \cap A) \cup (B_2 \cap A) \cup \dots)}{\mathbb{P}(A)} \tag{4.3}$$

$$= \frac{\mathbb{P}(B_1 \cap A) + \mathbb{P}(B_2 \cap A) + \dots}{\mathbb{P}(A)} \tag{4.4}$$

$$= \frac{\mathbb{P}(B_1 \cap A)}{\mathbb{P}(A)} + \frac{\mathbb{P}(B_2 \cap A)}{\mathbb{P}(A)} + \dots \tag{4.5}$$

$$= \mathbb{P}(B_1|A) + \mathbb{P}(B_2|A) + \dots \tag{4.6}$$

$$= \mathbb{P}_A(B_1) + \mathbb{P}_A(B_2) + \dots \tag{4.7}$$

$$(4.8)$$

Assim, a função de probabilidade $\mathbb{P}_A(\cdot) = \mathbb{P}(\cdot|A)$ reatribui probabilidades aos eventos da σ -álgebra \mathcal{A} criando um novo espaço de probabilidade. Isto significa que *todas* as propriedades válidas para uma função de probabilidade qualquer também são válidas para a função de probabilidade particular \mathbb{P}_A . Por exemplo, as propriedades vistas em 3.4.1 são válidas para $\mathbb{P}_A(\cdot) = \mathbb{P}(\cdot|A)$.

Assim, temos

(P1) $\mathbb{P}(B^C|A) = 1 - \mathbb{P}(B|A)$.

(P2) $0 \leq \mathbb{P}(B|A) \leq 1$ para todo evento $B \in \mathcal{A}$.

(P3) se $B_1 \subseteq B_2 \implies \mathbb{P}(B_1|A) \leq \mathbb{P}(B_2|A)$

(P4) $\mathbb{P}(\bigcup_{n=1}^{\infty} B_i|A) \leq \sum_{n=1}^{\infty} \mathbb{P}(B_i|A)$

(P5) $\mathbb{P}(B \cup C|A) = \mathbb{P}(B|A) + \mathbb{P}(C|A) - \mathbb{P}(B \cap C|A)$

A demonstração de que estas propriedades são válidas seguem o mesmo raciocínio usado acima para mostrar que \mathbb{P}_A atende aos três axiomas de Kolmogorov.

4.6 Independência mútua

Falamos da independência de dois eventos A e B . Eles são eventos independentes se

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \mathbb{P}(B),$$

ou, equivalentemente,

$$\mathbb{P}(A|B) = \mathbb{P}(A).$$

E quando tivermos vários eventos E_1, E_2, \dots, E_n ? Infelizmente, não basta olhar os pares de eventos e verificar a definição acima. Os eventos E_1, E_2, \dots, E_n são eventos independentes se toda combinação de eventos satisfizer a regra do produto:

$$\mathbb{P}(E_{i_1} \cap E_{i_2} \cap \dots \cap E_{i_m}) = \mathbb{P}(E_{i_1}) \dots \mathbb{P}(E_{i_m})$$

para toda seleção de índices i_1, i_2, \dots, i_m e para todo m entre 2 e n . Estes eventos são chamados *mutuamente independentes*.

Podemos deduzir que se A, B , e C são independentes então C é também independente de $A \cap B$, de $A \cap B^c$, de $A \cup B$, de B^c etc. (ver lista de exercícios).

Se os eventos são mutuamente independentes então qualquer par de eventos é independente. Um resultado curioso é que a conversa não é verdade. Podemos ter eventos independentes par a par mas que não são mutuamente independentes. Por exemplo, podemos ter A e B independentes, A e C independentes, e B e C independentes mas A, B, C dependentes. Um uso prático desta distinção aparece numa técnica para compartilhar senhas em criptografia (ver [7], seção 8.9.2).

4.7 Paradoxos com probabilidade condicional

Cálculos com probabilidades condicionais muitas vezes levam a resultados curiosos e pouco intuitivos. Probabilidade condicional é uma fonte constante de resultados paradoxais. Vamos mostrar alguns nesta seção.

■ **Example 4.6 — Sem saber a cor da bola retirada.** Uma urna contém duas bolas rosas e duas bolas marrons (Figura 4.2). Uma delas é escolhida completamente ao acaso. Suponha que esta bola primeira bola retirada seja rosa. Ela *não* é reposta na urna. Defina o evento R_1 como $R_1 = [\text{1a. bola rosa}]$. Uma segunda bola é retirada escolhendo-se uma das três restantes completamente ao acaso. Seja R_2 o evento $R_2 = [\text{2a. bola rosa}]$. Qual a probabilidade de R_2 ocorrer dado que R_1 ocorreu? Dado que R_1 ocorreu, restam na urna uma bola azul e duas marrons. Portanto, $\mathbb{P}(R_2|R_1) = 1/3$. Caso a primeira bola retirada seja marrom, evento representado por M_1 , temos $\mathbb{P}(R_2|M_1) = 2/3$.

O curioso vem agora: suponha que uma bola é retirada mas não sabemos qual a sua cor. Com certeza foi uma bola rosa ou marrom, apenas não sabemos qual foi. Vamos denotar por B o evento de que uma bola de cor desconhecida foi retirada. Observe que $B = R_1 \cup M_1$ e que R_1 e M_1 são disjuntos: $R_1 \cap M_1 = \emptyset$. Vamos calcular $\mathbb{P}(R_2|B) = \mathbb{P}(R_2|R_1 \cup M_1)$ e verificar que esta probabilidade é igual a $1/2$, a mesma probabilidade $\mathbb{P}(R_1)$ de retirar uma bola rosa quando a urna está completa. Isto é, a retirada de uma bola de cor desconhecida não muda a probabilidade de ocorrer R_2 , mas se a cor da bola retirada for revelada, então a probabilidade muda bastante.

Como $B = R_1 \cup M_1$ com $R_1 \cap M_1 = \emptyset$. Temos então

$$\begin{aligned}\mathbb{P}(R_2 \cap B) &= \mathbb{P}(R_2 \cap R_1) + \mathbb{P}(R_2 \cap M_1) \\ &= \mathbb{P}(R_2|R_1)\mathbb{P}(R_1) + \mathbb{P}(R_2|M_1)\mathbb{P}(M_1) \\ &= \frac{1}{3} \frac{1}{2} + \frac{2}{3} \frac{1}{2} = \frac{1}{2}\end{aligned}$$

■

■ **Example 4.7 — Paradoxo dos três prisioneiros.** Três prisioneiros numa masmorra medieval, A, B e C , foram condenados à morte. Sentindo-se abençoado pelo nascimento de seu herdeiro, o senhor feudal decide que um deles será escolhido ao acaso por sorteio e libertado. A fica sabendo

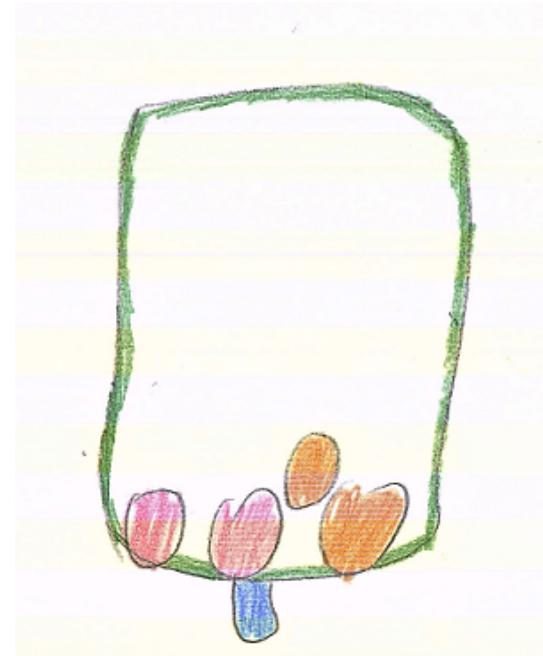


Figure 4.2: Urna com duas bolas rosas e duas bolas marrons, desenhadas por Sosô. Uma bola é retirada ao acaso e não é reposta na urna. Se sua cor não é registrada, qual a probabilidade de que uma segunda bola retirada seja rosa?

da decisão e diz ao seu carcereiro: “Eu já sei que pelo menos um dos outros dois prisioneiros, B ou C , será executado. Você poderia então me informar o nome de um deles que será executado? Como estamos incomunicáveis, esta informação não servirá para nada.” Convencido pelo argumento, o carcereiro diz que C será executado. O prisioneiro A raciocina que sua chance de ser libertado era $1/3$, mas agora que ele sabe que restaram apenas ele e B , sua chance subiu para $1/2$. Ou não?

Inicialmente, existem três resultados possíveis, todos com probabilidades iguais:

- A será libertado, com probabilidade $1/3$.
- B será libertado, com probabilidade $1/3$.
- C será libertado, com probabilidade $1/3$.

Com a informação a ser fornecida pelo carcereiro, existem quatro possibilidades com diferentes probabilidades:

- $[A \text{ lib}, B \text{ exec}]$: A será libertado e A é informado que B será executado, com probabilidade $1/6$.
- $[A \text{ lib}, C \text{ exec}]$: A será libertado e A é informado que C será executado, com probabilidade $1/6$.
- $[B \text{ lib}, C \text{ exec}]$: B será libertado e A é informado que C será executado, com probabilidade $1/3$.
- $[C \text{ lib}, B \text{ exec}]$: C será libertado e A é informado que B será executado, com probabilidade $1/3$.

A probabilidade de que o carcereiro informe que C será executado é igual a

$$\mathbb{P}(C \text{ exec}) = \mathbb{P}([A \text{ lib}, C \text{ exec}] \cup [B \text{ lib}, C \text{ exec}]) = \mathbb{P}(A \text{ lib}, C \text{ exec}) + (B \text{ lib}, C \text{ exec}) = \frac{1}{6} + \frac{1}{3} = \frac{1}{2}$$

Dado que o carcereiro informa que C será executado, a probabilidade condicional de interesse é

$$\mathbb{P}(A \text{ lib} | C \text{ exec}) = \frac{\mathbb{P}([A \text{ lib} \cap C \text{ exec}])}{\mathbb{P}(C \text{ exec})} = \frac{1/6}{1/2} = \frac{1}{3}.$$

Assim, a probabilidade de que A seja libertado não se altera com a informação fornecida. ■



5. Introdução a classificação

5.1 Introdução

Classificação supervisionada é uma tarefa de análise de dados que procura aprender a partir de dados estatísticos como usar os atributos de objetos para separá-los em dois ou mais conjuntos distintos chamados classes. Considere um conjunto de dados estatísticos no formato tabular usual onde os atributos (colunas) são divididos em dois tipos. Um dos atributos é uma variável nominal com o rótulo (*label*, em inglês) identificando a real classe do objeto-linha da tabela. As demais variáveis são atributos que queremos usar para predizer o rótulo ou classe.

De forma esquemática, imagine que temos a seguinte tabela de dados com k atributos e n casos (ver Tabelas em (5.1)). A variáveis (ou colunas) X_1, \dots, X_k representam os atributos dos objetos. A variável (ou coluna) Y contém os dois rótulos ou classes (0 e 1) em que estes objetos estão divididos. Os atributos da linha i da tabela são representados pelo vetor k -dimensional

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ik})$$

enquanto o seu rótulo (ou classe) correspondente é representado por Y_i . O objetivo da classificação supervisionada é aprender uma função h que recebe como entrada os atributos de um indivíduo e fornece como resposta uma boa aproximação para a probabilidade do caso receber um rótulo. No caso de apenas dois rótulos, 0 ou 1, queremos uma função h que forneça a probabilidade de que o rótulo Y ser igual a 1 condicionada aos valores dos atributos. Isto é, dado que o vetor de atributos de certo indivíduo é \mathbf{x} , obter de forma aproximada a probabilidade de que seu rótulo ou classe seja 1:

$$\mathbb{P}(Y = 1 | \mathbf{X} = \mathbf{x}) \approx h(\mathbf{x}).$$

Por exemplo, para o caso da linha $i = 3$ da Tabela ??, queremos que a função forneça uma estimativa ou uma aproximação para

$$\mathbb{P}(Y_3 = 1 | \mathbf{X} = \mathbf{x}_3) = \mathbb{P}(Y_3 = 1 | \mathbf{X} = (5.09, 3, \dots, 27.91)) \approx h(5.09, 3, \dots, 27.91) = h(\mathbf{x}_3)$$

$$\begin{bmatrix}
 X_1 & X_2 & \dots & X_k & Y \\
 \hline
 x_{11} & x_{12} & \dots & x_{1k} & Y_1 \\
 x_{21} & x_{22} & \dots & x_{2k} & Y_2 \\
 x_{31} & x_{32} & \dots & x_{3k} & Y_3 \\
 x_{41} & x_{42} & \dots & x_{4k} & Y_4 \\
 x_{51} & x_{52} & \dots & x_{5k} & Y_5 \\
 \vdots & \vdots & \vdots & \vdots & \vdots \\
 x_{n1} & x_{n2} & \dots & x_{nk} & Y_n
 \end{bmatrix} = \begin{bmatrix}
 X_1 & X_2 & \dots & X_k & Y \\
 \hline
 1.37 & 3 & \dots & -24.97 & 0 \\
 2.75 & 2 & \dots & 39.55 & 1 \\
 5.09 & 3 & \dots & 27.91 & 0 \\
 3.11 & 3 & \dots & 2.36 & 1 \\
 7.36 & 1 & \dots & 12.99 & 1 \\
 \vdots & \vdots & \vdots & \vdots & \vdots \\
 2.22 & 2 & \dots & 65.96 & 0
 \end{bmatrix} \quad (5.1)$$

Observe que, na linha 3 da tabela, o valor Y_3 realmente observado para este caso foi o rótulo ou classe 0 (isto é, tivemos a ocorrência do evento $Y_3 = 0$) mas a função h está fornecendo a probabilidade de que Y_3 seja igual a 1. Entretanto, como existem apenas duas classes neste exemplo, se quisermos $\mathbb{P}(Y_3 = 0 | \mathbf{X} = \mathbf{x}_3)$ basta fazer uma subtração:

$$\mathbb{P}(Y_3 = 0 | \mathbf{X} = \mathbf{x}_3) = 1 - \mathbb{P}(Y_3 = 1 | \mathbf{X} = \mathbf{x}_3) \approx 1 - h(\mathbf{x}_3)$$

■ **Example 5.1** Quando uma instituição financeira concede um empréstimo para um indivíduo ou para uma empresa, ela precisa avaliar o *risco de crédito* associado com aquele empréstimo. O risco de crédito é a chance ou probabilidade da instituição ter uma perda financeira resultante da falha do devedor em pagar o empréstimo ou em cumprir todas as obrigações contratuais (pagando com muito atraso, por exemplo). No momento de decidir se o empréstimo solicitado deve ou não ser concedido para aquele cliente particular, a instituição faz uma estimativa da probabilidade do cliente não honrar os termos do contrato no futuro. Como ela faz isto? Usando dados estatísticos do passado e de outros clientes. Já veremos algoritmos para isto.

Tendo em mãos a estimativa da probabilidade de não-pagamento daquele empréstimo no futuro por aquele cliente (uma estimativa do seu risco de crédito, no jargão da área), a instituição toma suas decisões. Se a probabilidade for muito alta, a instituição se recusa a conceder o empréstimo para aquele cliente. Se ela for moderada, a instituição talvez conceda o empréstimo mas cobrando juros mais altos do que o usual. Se a probabilidade for baixa, ela concede o empréstimo cobrando juros baixo.

No momento da solicitação do crédito, é impossível saber com exatidão qual será o resultado futuro. A decisão é baseada numa aproximação para a probabilidade de que o cliente seja um alto risco e esta aproximação para a probabilidade é aprendida ou estimada a partir dos dados estatísticos.

A aproximação para os riscos de crédito (ou probabilidades de não-pagamento) são calculadas a partir de um algoritmo que fornece a saída $h(\mathbf{x})$ para um indivíduo que possua certos atributos. Entre os atributos mais comuns usados pelas instituições financeiras podemos citar: X_1 = o saldo médio bancário nos últimos meses, X_2 = o valor do empréstimo relativo a este saldo médio, X_3 = idade do cliente, X_5 = há quanto tempo ele é cliente, X_6 = seu sexo (M ou F), etc.

Dado um perfil $bs\mathbf{X} = (x_1, x_2, \dots, x_6) = \mathbf{x}$, qual a probabilidade aproximada de que seu rótulo seja 1 (não-pagamento)? Isto é, queremos $\mathbb{P}(Y = 1 | \mathbf{X} = \mathbf{x}) \approx h(\mathbf{x})$. Veremos nas próximas seções dois algoritmos para obter esta aproximação a partir de dados históricos. ■

A tarefa de classificação supervisionada é uma das mais estudadas em aprendizagem de máquina. Outros exemplos onde Y representa a classe e \mathbf{X} conjunto de atributos são os seguintes:

- Classificar certos insetos (os objetos) em uma de três subespécies (as três classes de Y) usando como atributos \mathbf{x} um conjunto de 12 medições de características morfológicas (de forma).

- Classificar mulheres como portadoras ou não-portadoras de um distúrbio genético (as duas classes de Y) usando como atributos um vetor de medições em proteínas do sangue e histórico familiar.
- Classificar a qualidade de uma nova bateria de celular como boa ou ruim (as classes em Y) com base em algumas medições preliminares.
- Classificar mensagens de email como spam ou não-spam com base em características do seu cabeçalho e do seu conteúdo.

Este problema de classificação supervisionada será estudado em detalhes no capítulo 15. Aqui, nós vamos apenas apresentar de forma superficial alguns algoritmos. O objetivo do capítulo atual é apresentar um exemplo interessante de uso das ideias de probabilidade condicional.

5.2 Árvores de classificação

Árvores de classificação é uma técnica estatística que usa métodos estatísticos e algorítmicos. Ela não requer conhecimentos de probabilidade do usuário. As árvores classificam os objetos selecionando de um grande número de variáveis aquelas que são as mais importantes para predizer os resultados. A análise é baseada numa segmentação recursiva binária. Para entender este algoritmo, vamos começar com um exemplo bem simples. Queremos obter uma aproximação $h(\mathbf{x})$ para o risco de crédito $\mathbb{P}(Y = 1 | \mathbf{X} = \mathbf{x})$ usando apenas dois atributos, a idade X_1 do indivíduo e sua renda média X_2 nos últimos 6 meses. Coletamos dados de 100 clientes que pediram crédito recentemente na instituição. As primeiras 10 linhas da tabela com os dados coletados é a seguinte:

```
> cbind(renda, idade, credito)[1:10,]
   renda idade credito
[1,] 4566   38      0
[2,] 22409   57      0
[3,] 24730   68      0
[4,] 2548    31      0
[5,] 15922   27      1
[6,] 18461   53      0
[7,] 18644   46      0
[8,] 13175   41      0
[9,] 4665    30      0
[10,] 14262   24      1
```

A Figura 5.1 mostra, no seu lado esquerdo, os dados de todos os clientes através de um gráfico de dispersão dos dois atributos (idade e renda). Ao lado, repetimos este gráfico mas agora identificamos cada ums indivíduos como sendo um mau pagador (círculo) ou um bom pagador (estrela). Os indivíduos desses dois diferentes rótulos estão em regiões bem separadas. Podemos tentar particionar ou segmentar o espaço das variáveis (o plano, neste exemplo) de forma que apenas uma classe de pagadores fique dentro de cada região. Uma tentativa ruim de criar regiões que separem as classes pode ser vista no terceiro gráfico da Figura 5.1 onde ladrilhos retangulares são criados pelas linhas vermelhas. Esta tentativa é ruim porque as regiões criadas possuem tanto maus quanto bons pagadores. Uma tentativa bem melhor pode ser vista no último gráfico da Figura 5.1. Veja que agora a separação é excelente, com cada uma das quatro regiões contendo indivíduos de apenas uma das duas classes, ou a de bons pagadores ou a de maus pagadores.

A segmentação do espaço das variáveis no último gráfico da Figura 5.1 pode ser representada por uma árvore binária tal como aquela no lado esquerdo da Figura 5.2. Inicialmente, decidimos se o indivíduo possui renda menor que 13K ou não. Caso positivo, o caso vai para o ramo da esquerda. Se negativo, ele vai para o ramo da direita. Dentro do ramo esquerdo (isto é, para os casos em que

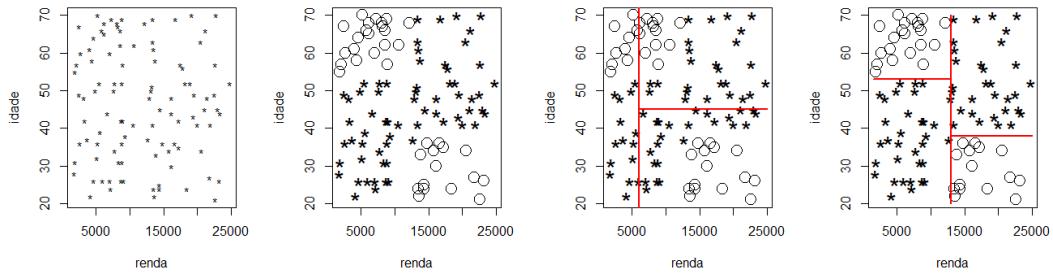


Figure 5.1: Esquerda: Dados de renda e idade de clientes que solicitaram empréstimo. À direita, o mems gráfico identificando os maus (círculo) e os bons pagadores (estrela). A seguir, uma segmentação ruim e uma segmentação boa em termos de separação das classes com bases nos atributos renda e idade.

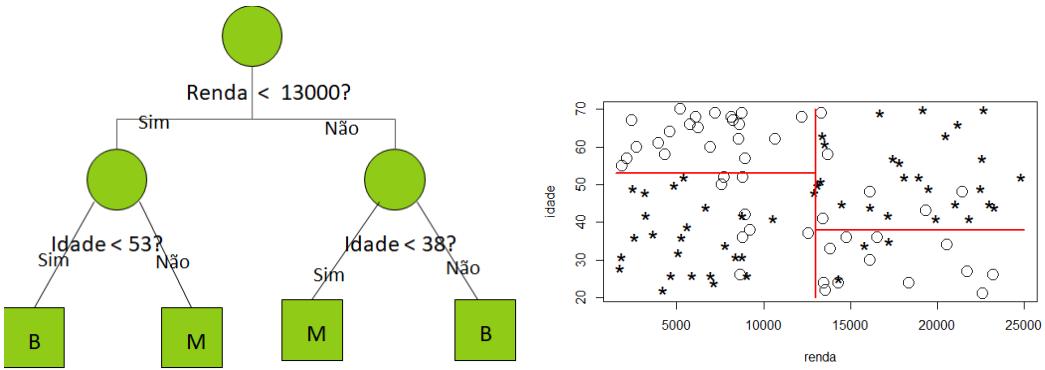


Figure 5.2: Esquerda: Representação sob forma de árvore da segmentação no último gráfico da Figura 5.1. Um gráfico de dispersão um pouco mais realista: nenhuma segmentação em quatro regiões consegue separar completamente os bons e maus pagadores.

a renda é menor que 13k), fazemos uma nova decisão. Se a idade for menor que 53 anos, vá para o ramo da esquerda. Caso contrário, vá para o ramo da direita. Para os casos em que a renda é maior que 13k, a segmentação é diferente. Ao invés de olhar se a idade é menor que 53 anos, verificamos se a idade é menor que 38 anos. Se sim, vá para a esquerda. Senão, vá para a direita. No final, ficamos com quatro folhas terminais na árvore. Em cada uma das folhas temos casos que tiveram, todos eles, uma única classe indicada no gráfico da Figura 5.2. Por um lado, toda segmentação do plano com ladrilhos em forma de retângulos pode ser expressa por uma árvore binária como esta. Por outro lado, toda árvore binária criada com os ramos dividindo-se cada ramo com base em único atributo acima ou abaixo de certo limiar gera uma segmentação do espaço particionado em retângulos com os lados paralelos aos eixos das coordenadas. De agora em diante, vamos usar alternativamente a representação de uma dada segmentação como uma árvore binária ou como ladrilhos retangulares no plano.

A segmentação no último gráfico da Figura 5.1 representada nesta árvore é perfeita mas raramente ela acontece na prática. Em geral, uma segmentação com linhas retas como as que fizemos até agora não será capaz de separar completamente o dados em grupos compostos apenas de uma única classes. Considere, por exemplo, o exemplo no segundo gráfico da Figura 5.2. Neste caso, nenhuma segmentação em quatro regiões consegue separar completamente os bons e maus pagadores. Podemos criar um número maior de regiões que tenha menos variação no rótulo ou

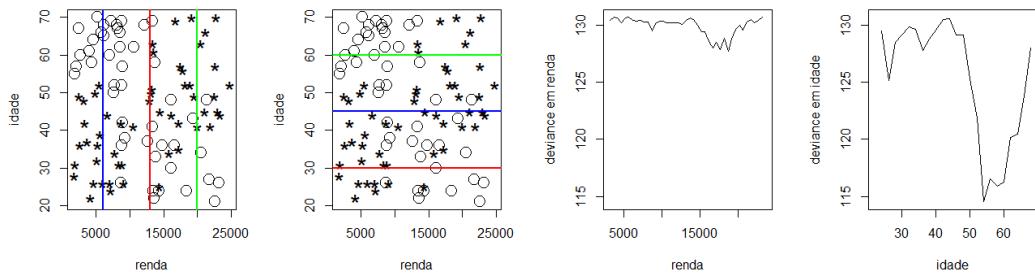


Figure 5.3: Esquerda:

podemos apenas nos dar por satisfeitos com a melhor segmentação possível em quatro regiões. Idealmente, queremos a separação perfeita: em cada segmento, termos apenas uma classe, bom ou mau. Na prática, nós buscamos a porcentagem de bons dentro dos nós terminais (ou regiões da segmentação) próxima de 100% ou de 0% em cada segmento. Calculamos uma medida de impureza em cada segmento criado: quanto mais distante de 100% ou de 0%, mais impuro o nó ou região criada.

Como obter uma boa segmentação. Existe um algoritmo iterativo simples. Comecemos obtendo a primeira separação.

- Percorra o eixo horizontal (renda) parando em cada valor possível e fazendo uma segmentação:
 - Calcule a impureza dos dois segmentos resultantes.
 - Escolha aquele valor de renda que produz a menor impureza nos segmentos.
 - Esta é a melhor separação possível considerando um corte com base na renda.
 - Não faça ainda esta segmentação; aguarde o resultado do próximo passo descrito abaixo.
- Voltando ao gráfico original, percorra o eixo vertical (idade) parando em cada valor possível e fazendo uma segmentação:
 - Calcule a impureza dos dois segmentos resultantes.
 - Escolha aquele valor de idade que produz a menor impureza nos segmentos.
 - Esta é a melhor separação possível considerando um corte com base na idade.

Uma das duas, ou a melhor segmentação com base na idade ou a melhor segmentação com base na renda, deve ser a melhor delas. Escolha a melhor delas e finalmente faça a primeira segmentação.

Na Figura 5.3, mostramos no primeiro gráfico três alternativas de partição (linhas verticais coloridas) com base em três pontos de corte da variável renda (6K, 13K, 20K). No segundo gráfico, temos três alternativas de partição (linhas horizontais coloridas) com base em três pontos de corte da variável idade (30, 45, 60). Considerando esta 6 alternativas de partição, escolhe-se aquela que leva ao menor grau de impureza nos nós resultantes (nas regiões resultantes). Neste caso, o grau de impureza das linhas verticais 6K, 13K e 20K é 130.6, 130.3, 129.9, enquanto que o grau de impureza das linhas horizontais 30, 45 e 60 é 129.1, 130.0 e 116.2. Assim, dentre estas 6 opções, a melhor delas é a linha horizontal em idade igual a 60 anos.

O grau de impureza gerado por uma segmentação pode ser calculado de várias maneiras e aqui nós usamos o valor da *deviance* (um termo em inglês). A fórmula da deviance não é relevante agora mas, para deixar registrado, ela é uma soma sobre todas as folhas da árvore. Neste primeiro passo, é apenas a soma sobre as duas regiões geradas pela linha vertical ou as duas regiões geradas pela linha horizontal. Seja $i = 1, 2$ um índice para a região gerada e $j = 0, 1$ um índice para a classe. Seja n_{ij} o número de objetos da classe j na região i e $n_{i+} = n_{i0} + n_{i1}$. Como $p_{ij} = n_{ij}/n_{i+}$, temos

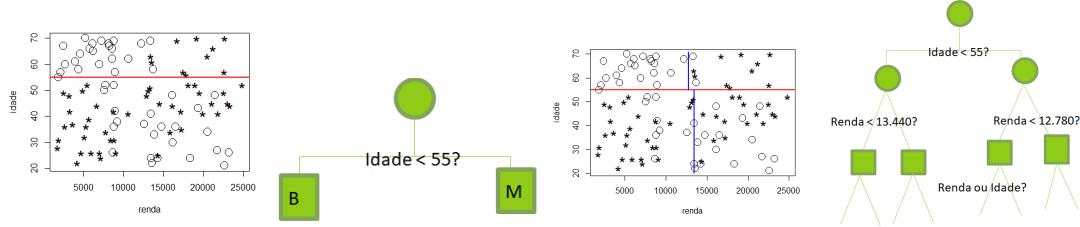


Figure 5.4: Primeira e segunda segmentação com as árvores correspondentes.

então a deviance dada por

$$D = -2 \sum_i \sum_j n_{ij} \log(p_{ij})$$

Esta medida está associada com a famosa medida de entropia de uma distribuição de probabilidade. Se valor $n_{ij} = 0$ teremos $p_{ij} = n_{ij}/n_{+} = 0$ e $n_{ij} \log(p_{ij})$ é uma expressão de valor indefinido. Usando o fato de que o limite $x \log(x) \rightarrow 0$ quando $x \rightarrow 0$, nós definimos que $n_{ij} \log(p_{ij}) = 0$ se $n_{ij} = 0$.

Tendo uma medida de grau de impureza definida, como a medida D acima, não precisamos nos limitar apenas às seis alternativas apresentadas nos dois primeiros gráficos da Figura 5.3 (as três segmentações verticais e as três segmentações horizontais). Vamos varrer cada eixo das coordenadas, primeiro o de renda e em seguida o de idade, calculando o valor de impureza em cada possível valor de uma grade regular imposta no eixo. Os dois últimos gráficos da Figura 5.3 mostram o valor da medida de impureza considerando cada valor possível dos atributos (renda ou idade). Como o idela é a menor impureza possível, a melhor segmentação é uma reta horizontal separando os objetos abaixo ou acima da idade de 55 anos aproximadamente.

Se fizermos esta primeira segmentação, separando os casos em que Idade é menor que 55 anos daqueles casos em que a Idade é maior ou igual a 55 anos, teremos o gráfico de pontos no lado esquerdo da Figura 5.4.

Existem agora dois segmentos: os clientes com idade menor que 55 anos e aqueles com idade maior ou igual a 55 anos. A ideia do algoritmo é iterar dentro de cada segmento criado neste primeiro passo. Assim, no segmento *idade menor que 55 anos*, percorremos cada um dos dois eixos coordenados, o eixo horizontal (renda) e o eixo vertical *entre 20 e 55 anos*, encontrando em cada eixo o ponto que segmenta verticalmente ou horizontalmente com a menor impureza possível o retângulo dos indivíduos com idade menor que 55 anos. Neste caso, a melhor segmentação é possível é usar uma reta vertical cortando este primeiro grupo em dois sub-grupos: aqueles com renda menor que 13.400 e aqueles com renda maior ou igual a 13.440.

Repetimos este procedimento dentro do segundo grupo, aquele dos indivíduos com idade maior que 55 anos. Checando qual corte ao longo de cada dos dois atributos gera a menor impureza, descobrimos que o melhor é partitionar ao longo da renda também quebrando entre aqueles que possuem renda menor que 12.780 e os que possuem renda maior ou igual a 12.780. O resultado é o terceiro gráfico da Figura 5.4.

Neste ponto, os dados segmentados em quatro grupos. Agora, simplesmente repetimos o procedimento de busca dentro de cada um desses quatro segmentos. Em alguns dos subgrupos criados não vale a pena segmentar mais pois eles já estão razoavelmente puros (com a proporção de bons pagadores próxima de 0% ou de 100%). Em outros, pode valer a pena segmentar mais. Neste exemplo, fazemos apenas uma segmentação adicional, ao longo do atributo idade, em um dos quatro segmentos já criados e o resultado final está no gráfico da Figura 5.5. A árvore correspondente está também nesta figura. O critério usado para interromper a segmentação num dado grupo é resultante de um teste estatístico que não será discutido neste momento.

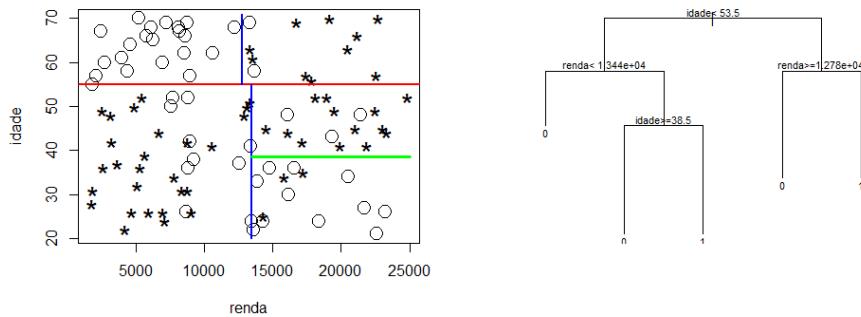


Figure 5.5: Partição final e árvore correspondente.

Ramo 1	Ramo 2
casado	solteiro, divorciado, viúvo
solteiro	casado, divorciado, viúvo
divorciado	solteiro, casado, viúvo
viúvo	solteiro, casado, divorciado
casado, solteiro	divorciado, viúvo
casado, divorciado	solteiro, viúvo
casado, viúvo	divorciado, solteiro

Table 5.1: Tabela com as possíveis segmentações baseadas no atributo categóric *estado civil*.

É fácil ver que, se tivermos mais de dois atributos, o algoritmo funciona do mesmo modo. Simplesmente percorra a faixa de variação de um atributo de cada vez escolhendo aquele ponto de corte e atributo que melhor separa as duas classes de objetos (minimiza a impureza). Em seguida, itere o algoritmo dentro de cada segmento criado. Prossiga dentro de cada segmento que vai sendo criado até que um teste estatístico decida que não vale mais a pena continuar segmentando.

Algumas vezes o atributo não é uma variável numérica como renda ou idade. Ela pode ser categórica como, por exemplo, o estado civil de um indivíduo, com os valores *casado*, *solteiro*, *divorciado*, *viúvo* ou a religião, com os valores *católico*, *evangélico*, *ateu*, *outro*. Como fazer no caso de um atributo categórico como esses? Se tivermos m categorias no atributo, existirão $2^m - 1$ possíveis segmentações. Por exemplo, com os quatro estados civis, existem $2^4 - 1 = 7$ segmentações possíveis.

Para cada possível segmentação categórica, calcule a redução da impureza ao fazer a segmentação. Escolha aquela segmentação categórica que produz a maior redução de impureza. Compare com a redução de impureza que os outros atributos (numéricos ou categóricos) proporcionam. Escolha a variável e a segmentação que produzem a máxima redução de impureza.

Prossiga segmentando de forma iterativa. Quando parar? Quando não houver mais evidência estatística de que segmentar adicionalmente produz decréscimo significativo da impureza da árvore. O teste estatístico é realizado a cada nova possível segmentação. Árvore final é a melhor segmentação que se pode fazer com os dados disponíveis. Existem outras maneiras escolher a árvore final. Por exemplo, podemos continuar segmentando até que haja uma pureza completa em cada folha, mesmo que para isto a folha contenha apenas um único indivíduo. Em seguida, podemos voltar podando a árvore deletando os ramos que não são realmente necessários. Uma discussão mais aprofundada desse algoritmo e dessas regras de poda será feita no capítulo ??.

Até agora, a variável alvo é categórica com duas categorias: bons e maus pagadores. Este

método de classificação baseado em segmentações com árvores também funciona se:

- A variável alvo tiver mais de duas classes;
- A variável alvo for numérica, ao invés de classes (número de dias em atraso, por exemplo). Neste caso, precisamos apenas de uma definição apropriada do que é a impureza nos nós da árvore. É comum que a impureza seja medida através do desvio-padrão dentro dos nós.

5.3 Árvores de classificação no R

Em R, recomendamos o uso do pacote `rpart`. O principal comando para criar a árvore é:

```
rpart(formula, data= ) where
```

onde `formula` possui o seguinte formato: `outcome ~ predictor1 + predictor2 + predictor3 + etc` e `data` = especifica o data frame onde estão as variáveis. Se usarmos a fórmula `outcome ..`, com um ponto no lugar de uma lista somando os nomes das variáveis, estaremos dizendo para o comando `rpart` usar todas as variáveis disponíveis no dataset execto a variável `outcome` com as classes dos indivíduos. O dataframe pode ser ignorado se as variáveis existirem como objetos separados.

```
# Classification Tree with rpart
library(rpart)
fit <- rpart(credito ~ renda + idade)
plot(fit)
text(fit, cex=0.7)
summary(fit) # informacao sobre os splits da arvore
```

COMPLETAR COM EXEMPLO MAIS SUBSTANCIAL

5.4 Alguns exemplos de árvores de classificação

Esta seção ilustra alguns exemplos reais de árvores de classificação. *Reescrever completamente. Atualmente, tem apenas copy-paste de abstracts de artigos de medicina.*

5.4.1 Predicting myocardial infarctions

To determine whether data available to physicians in the emergency room can accurately identify which patients with acute chest pain are having myocardial infarctions, [goldman1982computer] analyzed 482 patients at one hospital. Using recursive partitioning analysis, they constructed a decision protocol in the format of a simple flow chart to identify infarction on the basis of nine clinical factors. Figura 5.6 shows the result. In prospective testing on 468 other patients at a second hospital, the protocol performed as well as the physicians. Moreover, an integration of the protocol with the physicians' judgments resulted in a classification system that preserved sensitivity for detecting infarctions, significantly improved the specificity (from 67 per cent to 77 per cent, $P<0.01$) and positive predictive value (from 34 per cent to 42 per cent, $P = 0.016$) of admission to an intensive-care area. The protocol identified a subgroup of 107 patients among whom only 5 per cent had infarctions and for whom admission to non-intensive-care areas might be appropriate.

5.4.2 Predictive models for outcome after severe head injury

Many previous studies have constructed several predictive models for outcome after severe head injury, but these have often used expensive, time consuming, or highly specialized measurements. The goal of this study [rovillas2004classification] was to develop a simple, easy to use a model

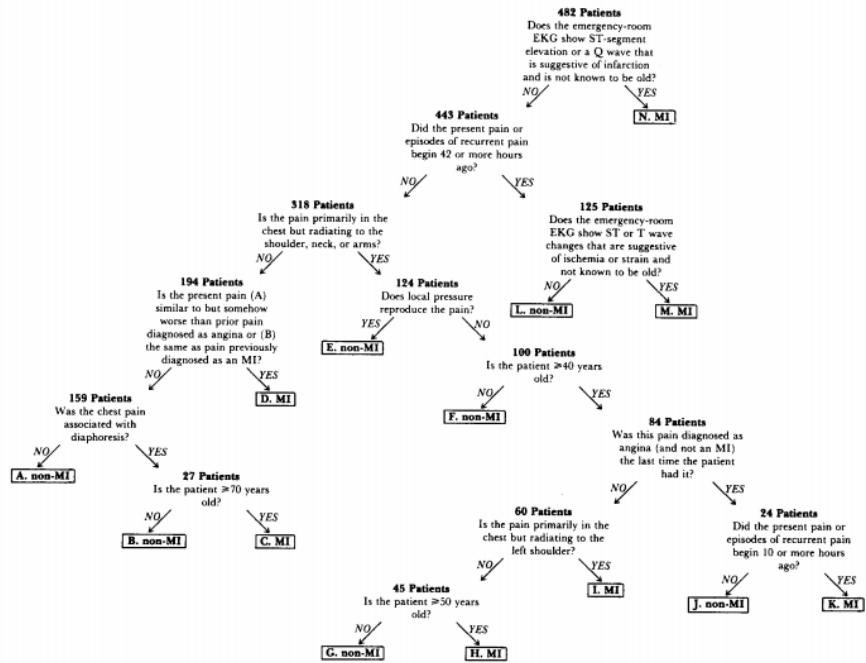


Figure 5.6: Árvore de classificação para diagnosticar pacientes sofrendo infarto do miocárdio dentre aqueles procurando o pronto socorro com dor aguda no peito. Figura extraída de [goldman1982computer]

involving only variables that are rapidly and easily achievable in daily routine practice. To this end, a classification and regression tree (CART) technique was employed in the analysis of data from 345 patients with isolated severe brain injury who were admitted to Asclepeion General Hospital of Athens from January, 1993, to December, 2000. A total of 16 prognostic indicators were examined to predict neurological outcome at 6 months after head injury. Figura 5.7 mostra os resultados obtidos. Our results indicated that Glasgow Coma Scale was the best predictor of outcome. With regard to the other data, not only the most widely examined variables such as age, pupillary reactivity, or computed tomographic findings proved again to be strong predictors, but less commonly applied parameters, indirectly associated with brain damage, such as hyperglycemia and leukocytosis, were found to correlate significantly with prognosis too. The overall cross-validated predictive accuracy of CART model for these data was 87%. All variables included in this tree have been shown previously to be related to outcome. Methodologically, however, CART is quite different from the more commonly used statistical methods, with the primary benefit of illustrating the important prognostic variables as related to outcome. This technique may prove useful in developing new therapeutic strategies and approaches for patients with severe brain injury.

5.4.3 Autism Distinguished from Controls Using Classification Tree Analysis

In the paper [neely2007quantitative], the authors use a classification tree (CART) method to distinguish between individuals with autism and normal controls based on features extracted from structural magnetic resonance images (MRI). The CART method yielded a high specificity in classifying autism subjects from controls based on the relationship between the volume of the left fusiform gyrus (LFG) gray and white matter, the right temporal stem (RTS) and the right inferior temporal gyrus gray matter (RITG-GM). These findings demonstrate different relationships within temporal lobe structures that distinguish subjects with autism from controls. Figure 5.8 shows some of the results.

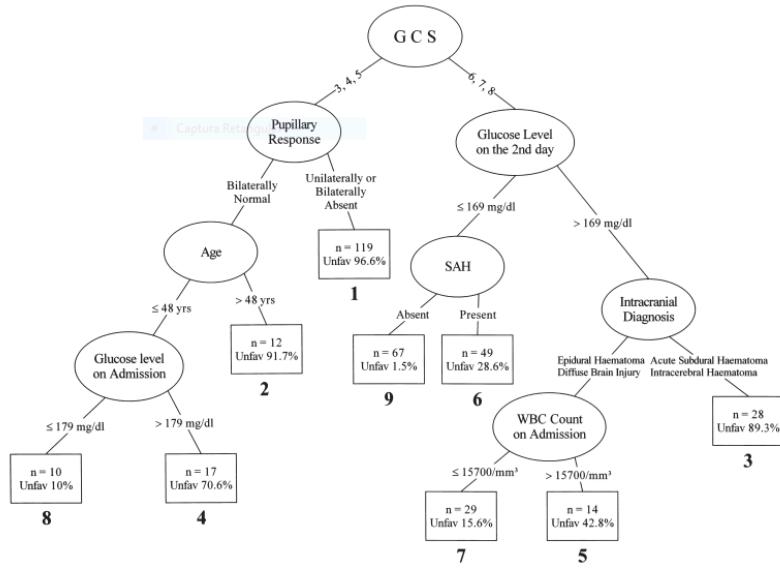


Figure 5.7: Prediction tree based on 345 patients with severe head injury. Ovals denote intermediate subgroups subject to further splitting; squares denote terminal prognostic subgroups. The numbers below the squares represent the prognostic rank of each subgroup based on the proportion of unfavorable outcomes. GCS, Glasgow Coma Scale; SAH, subarachnoid hemorrhage; WBC, white blood cells. Figura extraída de [rovlias2004classification].

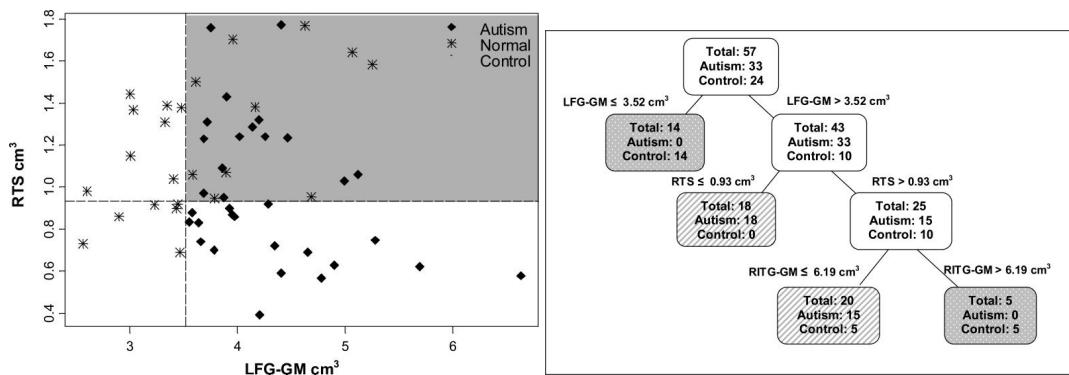
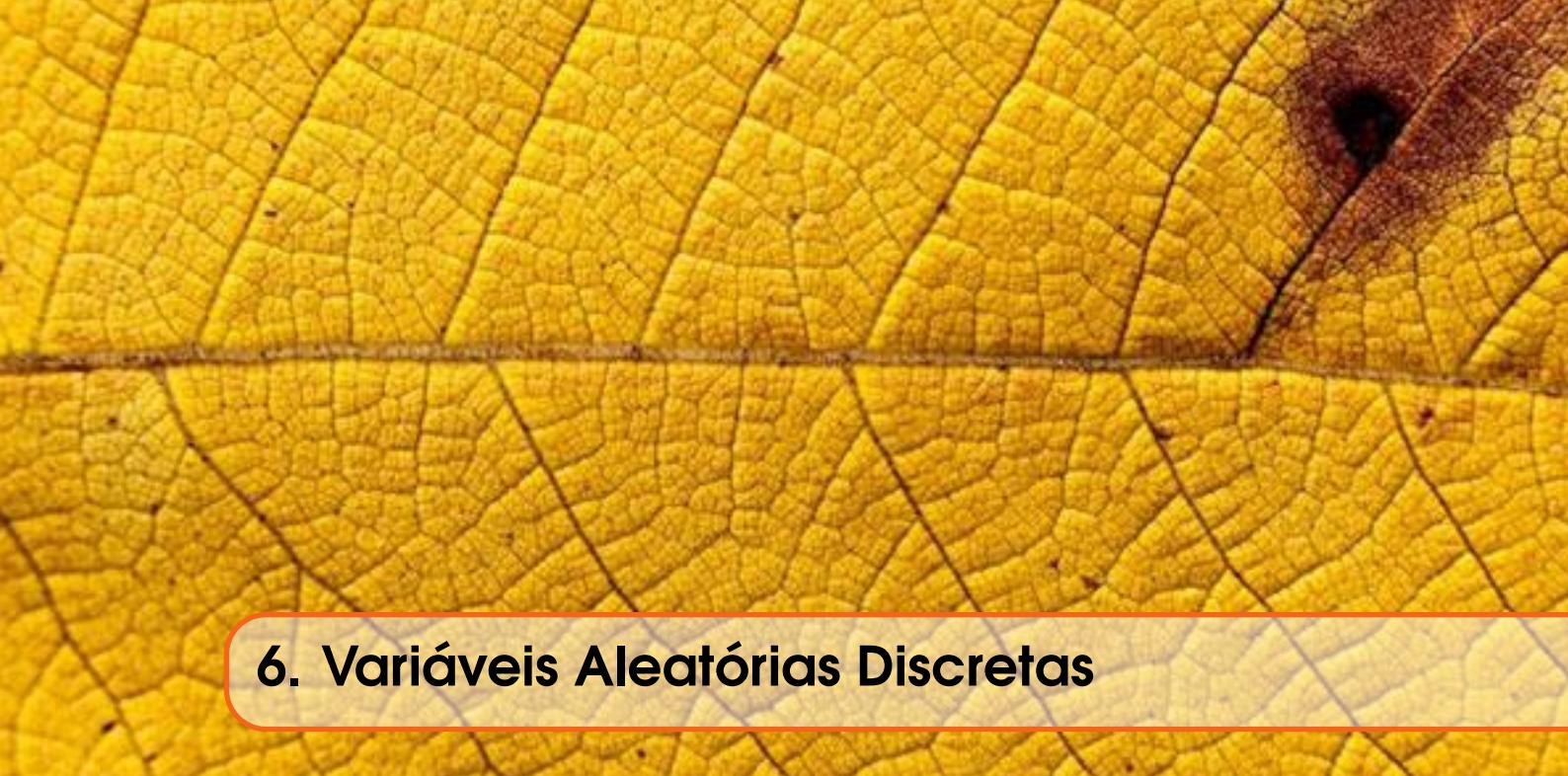


Figure 5.8: **Left:** The two dimensional classification tree shows how the actual observations are separated based on the first two splits of the regression tree. The left fusiform gyrus gray matter (LFG-GM) is on the x-axis. This represents the first split formed by the tree and the 14 red stars on the left side of the orange line are the controls classified by the first split of the tree. **Right:** Classification tree of autism vs. normal control groups Autism vs. Control Groups Classification Tree including left fusiform gyrus gray matter (LFG-GM), right inferior temporal gyrus gray matter (RITG-GM), and the right temporal stem (RTS). White boxes indicate locations on the tree that are still subject to splitting. Lined and checkered boxes are locations on the tree that have completed the splitting. Lined boxes have a higher proportion of autistic individuals and checkered boxes have a higher proportion of normal control individuals.



6. Variáveis Aleatórias Discretas

6.1 Variáveis aleatórias: formalismo

Probabilidade é um assunto de matemática. Estabelece um espaço de probabilidade $(\Omega, \mathcal{A}, \mathbb{P})$ e fazemos cálculos matemáticos de probabilidade. Estatística, data mining e machine learning são assuntos que lidam com dados. No caso mais simples e usual, temos uma tabela cheia de números (ou rótulos para categorias tais como *Masculino* e *Feminino*). Nesta tabela, linhas são itens, colunas são atributos medidos nos itens. Como ligar estes dois assuntos? A ligação é fornecida pelo conceito de *variável aleatória*.

O espaço de probabilidade $(\Omega, \mathcal{A}, \mathbb{P})$ é a base matemática da probabilidade. O espaço de probabilidade precisa atribuir probabilidade a todo evento (ou subconjunto) $A \subset \Omega$. Se Ω for muito complicado, podemos estar interessados apenas em *alguns* aspectos específicos do experimento aleatório. *Variáveis aleatórias* (v.a.) constituem a ferramenta para reduzir o espaço de probabilidade $(\Omega, \mathcal{A}, \mathbb{P})$ ao mínimo necessário na prática. Variáveis aleatórias são características numéricas do fenômeno aleatório com probabilidades associadas.

Definition 6.1.1 — Variáveis aleatórias. Formalmente, uma variável aleatória é uma função matemática (mensurável) X de Ω para \mathbb{R} . Isto é,

$$\begin{aligned} X: \Omega &\rightarrow \mathbb{R} \\ \omega &\rightarrow X(\omega) \end{aligned}$$

Assim, X é uma função que associa o valor real $X(\omega)$ ao resultado ω do experimento. Cada resultado ω tem um valor $X(\omega)$. Claro, diferentes ω podem ter um mesmo valor $X(\omega)$. Comentamos sobre a restrição de que uma variável aleatória seja uma função mensurável na nota 6.1. Daqui por diante, vamos abreviar a expressão *variável aleatória* por v.a.

■ **Example 6.1** Vamos voltar a este exemplo básico e recorrente: o lançamento repetido n vezes de uma moeda desonesta com probabilidade de sair cara igual a $\theta \in (0, 1)$. O espaço amostral é composto por todas as n -uplas com C ou \tilde{C} em cada posição. Isto é, $\Omega = \{\omega = (s_1, s_2, \dots, s_n) ; s_i = C \text{ ou } \tilde{C} \text{ para } i = 1, \dots, n\}$. Vamos definir v.a.'s que vão associar um número real a cada $\omega \in \Omega$.

Considere então as seguintes v.a's:

$$\begin{aligned} X: \Omega &\rightarrow \mathbb{R} \\ \omega &\rightarrow X(\omega) = \text{no. de caras nos } n \text{ lançamentos} \end{aligned}$$

ou a proporção de caras nos n lançamentos da moeda, o que significa que teremos um valor no conjunto $\{0/n, 1/n, \dots, (n-1)/n, 1\}$:

$$\begin{aligned} Y: \Omega &\rightarrow \mathbb{R} \\ \omega &\rightarrow Y(\omega) = \text{proporção de caras nos } n \text{ lançamentos} \end{aligned}$$

A próxima variável aleatória registra a ordem do lançamento associado com a última aparição de uma cara na sequência (se não houver caras, associamos o valor 0):

$$\begin{aligned} Z: \Omega &\rightarrow \mathbb{R} \\ \omega &\rightarrow Z(\omega) = \max \{\{i \text{ tais que } s_i = C \text{ para } i = 1, \dots, n\} \cup \{0\}\} \end{aligned}$$

■

Ao contrário da notação tradicional que usa f, g, h para denotar funções matemáticas, usamos letras maiúsculas do final do alfabeto, tais como X, Y, Z ou W para denotar as variáveis aleatórias. Esta é uma tradição da qual não temos como escapar já esta é a notação usada em todo o mundo. No início, isto é confuso e irritante. Depois, a gente se acostuma.

■ **Example 6.2** Considere o exemplo da Figura 3.9, onde o espaço amostral Ω é o conjunto formado por todas as funções contínuas no período de $[0, 24]$ horas. Isto é, o elemento $\omega \in \Omega$ é uma função contínua f com domínio $[0, 24]$. Eventos são sub-conjuntos de curvas deste conjunto Ω com infinitas curvas f .

Podemos definir uma série de variáveis aleatórias de interesse potencial neste conjunto. Para deixar mais explícito a situação deste exemplo, vamos agora denotar o elemento *omega* por f , a função contínua que é o resultado do experimento de observar a temperatura por 24 horas. Temos então, por exemplo, a temperatura máxima ao longo do dia

$$\begin{aligned} X: \Omega &\rightarrow \mathbb{R} \\ f &\rightarrow X(f) = \max_{x \in [0, 24]} f(x) \end{aligned}$$

a temperatura ao meio dia

$$\begin{aligned} Y: \Omega &\rightarrow \mathbb{R} \\ f &\rightarrow Y(f) = f(12) \end{aligned}$$

a temperatura média ao longo do dia

$$\begin{aligned} Z: \Omega &\rightarrow \mathbb{R} \\ f &\rightarrow Z(f) = \int_0^{24} f(x) dx \end{aligned}$$

■

R

Uma variável aleatória é quase qualquer função matemática X que vai de Ω para \mathbb{R} . A única restrição é que a função precisa ser mensurável, uma condição bastante técnica. Na prática, esta condição de mensurabilidade da função pode ser ignorada e podemos pensar que variável aleatória é qualquer função que tenha alguma relevância para a análise de dados. Toda função “prática” é mensurável. Toda função envolvendo um número finito ou infinito enumerável de operações envolvendo logs, exponenciais, polinômios, funções trigonométricas, funções escada, todas elas são mensuráveis. É difícil pensar numa função praticamente útil que não possa ser escrita dessa forma.

6.2 Variáveis Aletórias e tabelas de dados

Lembre-se da tabela de dados estatísticos. Nas linhas, temos os itens, indivíduos, casos, instâncias ou exemplos (tais como diferentes pacientes com câncer de um hospital ou diferentes clientes de um banco). Nas colunas, temos características ou atributos dos itens. Por exemplo, podemos ter colunas representando sexo, idade e estágio do câncer, ou saldo médio na conta corrente, tempo como correntista. Informalmente, *variáveis aleatórias* (v.a.) são as representações *matemáticas ou probabilísticas* dessas colunas de atributos na tabela de dados estatísticos.

Como é a conexão entre a tabela de dados e o modelo probabilístico? O espaço de probabilidade é formado pelo trio $(\Omega, \mathcal{A}, \mathbb{P})$. Um exemplo de tabela de dados está na tabela abaixo. Dizemos que Ω é o conjunto de todos os e-mails já recebidos e a receber. Note que Ω é um conjunto de tamanho indefinido e provavelmente infinito. A matriz de dados contém apenas uma *amostra* de elementos de Ω . Cada *linha* da tabela corresponde a um elemento distinto de Ω . Cada coluna representa diferente características ou medições sobre os e-mails. Em geral, supomos que os diferentes e-mails da amostra (as diferentes linhas da tabela) representam eventos independentes uns dos outros. Isto é, um e-mail com certas características não influencia as características dos outros emails. Numa mesma linha da tabela, as entradas medem diferentes características do *mesmo* ω (mesmo e-mail). Assim, os elementos *dentro de uma mesma linha da tabela* não costumam ser independentes. Ao contrário, como várias medições tentam captar a mesma coisa (a natureza spam/não-spam do e-mail), elas tendem a estar associadas ou correlacionadas. Vamos ver a definição precisa de correlação mais tarde, mas basta dizer que quando o número de caracteres é baixo temos um indicativo de que a variável *spam* deve ter o valor *yes*.

	spam	num_char	line_breaks	format	number
1	no	21,705	551	html	small
2	no	7,011	183	html	big
3	yes	631	28	text	none
:	:	:	:	:	:
50	no	15,829	242	html	small

Definition 6.2.1 — Visão informal de uma v.a.. Ao invés da sua definição formal, toda v.a. pode ser pensada simplesmente como sendo a combinação de dois componentes:

- um conjunto de valores possíveis na reta real;
- probabilidades associadas a estes valores possíveis.

Os valores possíveis estão associados com o significado específico das variáveis aleatórias. Já as probabilidades vêm de um modelo $(\Omega, \mathcal{A}, \mathbb{P})$ que muitas vezes *não* precisa ser explicitamente apresentado. Isto facilita muito a vida.

6.3 Tipos de variáveis aleatórias

Temos três tipos básicos de *dados estatísticos* nas tabelas de dados estatísticos:

- dados categóricos ou não-numéricos, que podem ser nominais (tais como sexo ou religião) ou ordinais (por exemplo, a resposta a uma pergunta como “Você confia muito, pouco ou nada nos membros do Congresso?”)
 - dados numéricos discretos: número de filhos, número de requisições nas últimas duas horas.
 - dados numéricos contínuos: saldo na conta corrente, temperatura, índice de inflação.
- Estes dados são representados por dois tipos de variáveis aleatórias:
- V.A.s Discretas:** Para os dados categóricos ou numéricos discretos.
 - V.A.s Contínuas:** Para os dados numéricos contínuos.

6.4 Variáveis Aleatórias Discretas

Definition 6.4.1 — V.A.s Discretas. As variáveis aleatórias discretas servem para modelar as colunas da tabela de dados que possuem valores categóricos ou numéricos discretos. Podemos pensar numa v.a. discreta como sendo composta de duas listas enumeráveis.

- Uma lista de valores possíveis para a v.a.: $\{x_1, x_2, \dots\}$.
- Uma lista com a probabilidade associada a cada um desses valores: $\{p(x_1), p(x_2), \dots\}$.

As duas listas em conjunto definem o que chamamos de *distribuição de probabilidade* da v.a. X .

Definition 6.4.2 Em geral a lista de valores possíveis na definição acima possui apenas os valores x em que $p(x)$ é estritamente maior que zero. Isto é, só entram os valores em que $p(x) > 0$. O conjunto dos pontos x tais que $p(x) > 0$ é chamado de *conjunto suporte* da distribuição.

Podemos representar as duas listas numa tabela:

Valores possíveis	x_1	x_2	x_3	...
Probab assoc	$p(x_1)$	$p(x_2)$	$p(x_3)$...

A lista de probabilidades deve ter valores ≥ 0 e eles devem somar 1. Juntas, estas duas tabelas definem uma função $x_i \rightarrow p(x_i)$ do conjunto de valores possíveis x_i para as probabilidades $p(x_i)$. Esta função é chamada *função massa de probabilidade* da v.a. X .

A melhor maneira de visualizar uma variável aleatória discreta é com um gráfico das probabilidades associadas. A Figura ?? mostra as duas listas da definição informal 6.1 no caso de uma v.a. discreta X . A lista ordenada de valores possíveis é $\{0, 1, 2, 3\}$ e as probabilidades associadas está na lista ordenada $\{0.5, 0.3, 0.1, 0.1\}$, respectivamente. O eixo horizontal contém os valores possíveis $\{0, 1, 2, 3\}$ e o eixo vertical mostra as probabilidades $p(x_i)$ onde x_i é um dos valores possíveis de X .

6.5 V.A.s discretas: exemplos

■ **Example 6.3 — V.A. binária.** Uma coluna da tabela de dados indica o sexo de um indivíduo ω escolhido de uma população. Arbitrariamente, associamos o valor 0 a MASC e 1 a FEM. Isto é, $X(\omega) = 0$ se ω for do sexo masculino e $X(\omega) = 1$ se ω for do sexo feminino. Para cada indivíduo ω olhamos apenas seu sexo, representado por $X(\omega) \in \{0, 1\}$. Para acabar a especificação dessa v.a. discreta, precisamos especificar a lista de probabilidades associada. Digamos, $p(0) = 0.35$ e $p(1) = 1 - 0.35 = 0.65$. ■

■ **Example 6.4 — Monitoramento de bombas de gasolina.** Num posto de gasolina, monitora-se a cada 5 minutos durante as horas de pico o uso de suas 4 bombas de abastecimento de veículos. De 5 em 5 minutos, anota-se o número de bombas em uso. Os itens ou instâncias são os diferentes instantes de tempo. Os dados são numéricos discretos e, em cada instante, podem ser 0, 1, 2, 3 ou 4. Seja ω um dos instantes de tempo. $X(\omega)$ é o número de bombas em uso. É preciso também

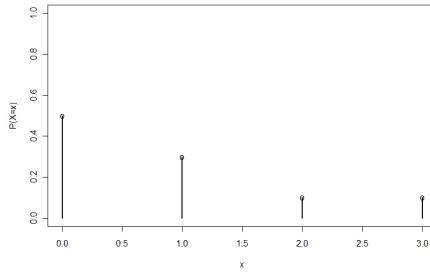


Figure 6.1: Função $p(x_i)$ onde x_i é um dos valores possíveis de X . Também chamada de função massa de probabilidade. X tem valores possíveis $\{0, 1, 2, 3\}$ com probabilidades $\{0.5, 0.3, 0.1, 0.1\}$, respectivamente.

especificar as probabilidades de cada valor possível para X . Por exemplo, a tabela ?? apresenta uma possível especificação para as probabilidades.

Valores possíveis	0	1	2	3	4
Probab assoc	$p(0) = 0.32$	$p(1) = 0.42$	$p(2) = 0.21$	$p(3) = 0.04$	$p(4) = 0.01$

■

■ **Example 6.5 — Número de seguidores numa rede social.** Numa rede social, escolha n usuários-vértices ao acaso e conte o número de arestas incidentes de cada um deles. (seguidores do usuário). Os itens ou instâncias são os diferentes usuários. Os dados são numéricos discretos e podem ser $0, 1, 2, 3, \dots$ sem um limite máximo natural. Seja ω um dos usuários e $X(\omega)$ o seu número de seguidores. $X(\omega) \in \{0, 1, 2, 3, \dots\} = \mathbb{N}$. Especificando as probabilidades (sem explicar de onde tiramos isto):

Val. pos. k	0	1	2	3	...	223	...
Probab $p(k)$	0.001	0.002	0.002	0.04	...	0.002	...

A lista (infinita) de probabilidades deve ter valores ≥ 0 e eles devem somar 1. Isto é, $1 = \sum_{k=0}^{\infty} p(k)$.

■

■ **Example 6.6 — Resposta em pesquisa por amostragem.** Pergunta-se a uma amostra de indivíduos (as instâncias) qual é a sua religião: católica, protestante, sem religião, outras religiões cristãs, espírita, outras. São seis categorias possíveis para cada resposta, claramente não numéricas e sem ordenação. Vamos representar esta coluna de dados com uma variável aleatória X . Como X é uma função de Ω para \mathbb{R} , arbitrariamente nós vamos associar um número a cada categoria da resposta.

Seja $X(\omega)$ uma variável aleatória que, para cada indivíduo ω da população, associe um número da seguinte forma:

$$X(\omega) = \begin{cases} 1, & \text{se } \omega \text{ é católico} \\ 2, & \text{se } \omega \text{ é protestante} \\ 3, & \text{se } \omega \text{ não tem religião} \\ 4, & \text{se } \omega \text{ é de outras religiões cristãs} \\ 5, & \text{se } \omega \text{ é espírita} \\ 6, & \text{se } \omega \text{ é de alguma outra religião} \end{cases}$$

A associação entre as categorias e os números correspondentes é completamente arbitrária. Qualquer outra associação seria válida. Por exemplo, poderíamos ter definido:

$$X(\omega) = \begin{cases} -2, & \text{se } \omega \text{ é católico} \\ -1, & \text{se } \omega \text{ é protestante} \\ 0, & \text{se } \omega \text{ não tem religião} \\ 1, & \text{se } \omega \text{ é de outras religiões cristãs} \\ 5, & \text{se } \omega \text{ é espírita} \\ 999, & \text{se } \omega \text{ é de alguma outra religião} \end{cases}$$

Na prática, com estes atributos não-numéricos, os valores da variável aleatória serão usados apenas como um rótulo (numérico) para a categoria.

Vamos voltar a especificação anterior, em que $X(\omega) \in \{1, \dots, 6\}$. Para completar a especificação da v.a., precisamos também declarar as probabilidades associadas com cada categoria de religião (ou cada valor possível da v.a.). Por exemplo, usando os dados do IBGE, na década de 80, ao escolher um indivíduo ao acaso da população brasileira, temos as seguintes probabilidades:

Val. pos. k	1 (cat)	2 (pro)	3 (s.rel)	4 (out. cr.)	5 (esp)	6 (out)
Probab $p(k)$	0.75	0.15	0.07	0.01	0.01	0.01

■

6.6 A σ -álgebra e a função de probabilidade

A atribuição de probabilidade a cada valor possível de uma v.a. X é consequência das probabilidades definidas no espaço de probabilidade $(\Omega, \mathcal{A}, \mathbb{P})$. Por exemplo, lance uma moeda 6 vezes, com C = cara e \tilde{C} = coroa. $\Omega = \{CCCCCC, \tilde{C}CCCCC, C\tilde{C}CCCC, \dots, \tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}\}$ Ω possui 36 elementos e $\mathbb{P}(\omega) = 1/36$. Se não estivermos interessados na ordem em que os resultados aparecem, mas apenas no número total de caras, podemos focar apenas numa versão reduzida do espaço de probabilidade. Definimos $X(\omega)$ como sendo o número de C 's em ω . Formalmente, temos

$$\begin{aligned} X: \Omega &\rightarrow \mathbb{R} \\ \omega &\rightarrow X(\omega) = \text{número de } C\text{'s em } \omega \end{aligned}$$

Portanto, $X(\omega) \in \{0, 1, \dots, 6\} \subset \mathbb{R}$. Estes são os valores possíveis da v.a. X .

Cada um desses valores possíveis possui uma probabilidade que é induzida pelo espaço de probabilidade original $(\Omega, \mathcal{A}, \mathbb{P})$. Uma proposição acerca do valor de uma v.a. X em \mathbb{R} determina um evento A em Ω . Por exemplo, a proposição $[X = 6]$ é equivalente ao evento em que os 6 lançamentos tiveram 6 caras. A proposição $[X \geq 5]$ é equivalente ao evento em que os 6 lançamentos tiveram pelo menos 5 caras. Alguns exemplos de proposições sobre o valor da v.a. X e os eventos equivalentes são os seguintes:

$$[X = 6] = \{\omega \in \Omega : X(\omega) = 6\} = \{\omega = (CCCCCC)\}$$

Lembre-se que o símbolo “:” deve ser lido como “tais que”. Isto é, o conjunto $\{\omega \in \Omega : X(\omega) = 6\}$ deve ser lido como $\{\omega \in \Omega \text{ tais que } X(\omega) = 6\}$. Seguem outros exemplos:

$$[X = 5] = \{\omega \in \Omega : X(\omega) = 5\} = \{(\tilde{C}CCCCC), (C\tilde{C}CCCC), \dots, (CCCCCC)\}$$

$$[X = 1] = \{\omega \in \Omega : X(\omega) = 1\} = \{(C\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}), (\tilde{C}C\tilde{C}\tilde{C}\tilde{C}\tilde{C}), \dots, (\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C})\}$$

$$[X = 0] = \{\omega \in \Omega : X(\omega) = 0\} = \{\omega = (\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C})\}$$

Mais exemplos e notação:

$$[X \geq 5] = \{\omega \in \Omega : X(\omega) \geq 5\} = \{(CCCCC), (\tilde{C}CCCC), (C\tilde{C}CCCC), \dots, (CCCCC\tilde{C})\}$$

ou então

$$[X \leq 1] = \{\omega \in \Omega : X(\omega) \leq 1\} = \{(\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}), (C\tilde{C}\tilde{C}\tilde{C}\tilde{C}), (\tilde{C}C\tilde{C}\tilde{C}\tilde{C}), \dots, (\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C})\}$$

Sendo subconjuntos, eventos podem ser manipulados com as operações usuais de união, interseção e complementar. De fato, proposições compostas sobre o valor de X são equivalentes a um evento resutante de operações sobre eventos mais simples. Por exemplo:

$$\begin{aligned}[X \leq 5 \text{ and } X > 4] &= \{\omega \in \Omega : X(\omega) \leq 5\} \cap \{\omega \in \Omega : X(\omega) > 4\} \\ &= \{(\tilde{C}CCCC), (C\tilde{C}CCCC), \dots, (CCCCC\tilde{C})\}\end{aligned}$$

Notation 6.1. Para uma v.a. X e um número real x , a notação $[X \leq x]$ significa o evento $\{\omega \in \Omega : X(\omega) \leq x\}$. A condição de que X deve ser uma função mensurável na definição 6.1.1 é a garantia de que este evento realmente pertence à σ -álgebra \mathcal{A} para todo $x \in \mathbb{R}$.

Sejam a e b dois números reais arbitrários e considere o intervalo $(a, b]$. A notação $[a < X \leq b]$ ou $[X \in (a, b]]$ é equivalente ao evento $\{\omega \in \Omega ; X(\omega) \in (a, b]\}$. Outros intervalos tais como (a, b) , $[a, b]$ ou (a, b) tem notações análogas. Note que $[X \leq x] = [X \in [x, \infty))$.

Como $[X \in (a, b)]$ é um evento, e portanto um subconjunto de Ω , podemos manipular vários eventos deste tipo com a operações de conjuntos. Como no exemplo do lançamento da moeda, podemos estar interessados na probabilidade de que X não pertença a um intervalo $(a, b]$. Temos

$$[X \notin (a, b)] = \{\omega \in \Omega : X(\omega) \notin (a, b)\} = \{\omega \in \Omega : X(\omega) \in (a, b]\}^c = [X \in (a, b)]^c.$$

■ **Example 6.7** A variável aleatória K conta o número de anos completos de um indivíduo de certa população no momento de sua morte. Assim, os valores possíveis para K estão no conjunto $\{0, 1, 2, \dots, 95, 96, \dots\}$ sem um limite superior. Se o interesse é obter a probabilidade de que o indivíduo viva pelo menos 60 anos mas faleça antes de completar 64 anos de idade, então estamos interessados no evento

$$[X \geq 60 \cap X < 64] = [X = 60] \cup [X = 61] \cup [X = 62] \cup [X = 63]$$

■

O interesse pode estar concentrado em obter a probabilidade de que uma v.a. X esteja no intervalo $(1, 2)$ ou seja maior que 3. Isto é, temos interesse em calcular a probabilidade de um evento A dado por

$$\begin{aligned}[X \in (1, 2) \text{ ou } X > 3] &= \{\omega \in \Omega ; X(\omega) \in (1, 2) \text{ ou } X(\omega) > 3\} \\ &= \{\omega \in \Omega ; X(\omega) \in (1, 2)\} \cup \{\omega \in \Omega ; X(\omega) > 3\} \\ &= [X \in (1, 2)] \cup [X(\omega) > 3]\end{aligned}$$

Note que os eventos $[X \in (1, 2)]$ e $[X(\omega) > 3]$ são disjuntos pois não podemos ter um valor ω tal que, ao mesmo tempo, $X(\omega) \in (1, 2)$ e $X(\omega) > 3$.

Notation 6.2. Em geral, estamos interessados em calcular probabilidades de eventos definidos por valores de v.a.'s. Seja $B \subseteq \mathbb{R}$ um subconjunto da reta real. Denotamos

$$\mathbb{P}(X \in B) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\})$$

Por exemplo, para o evento $[X \leq 2]$ temos

$$\mathbb{P}([X \leq 2]) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq 2\}) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in [2, \infty)\}).$$

Ou ainda, $\mathbb{P}([X = 2]) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = 2\})$. Em geral, escrevemos apenas $\mathbb{P}(X \leq 2)$ e $\mathbb{P}(X = 2)$ ao invés da notação mais carregada $\mathbb{P}([X \leq 2])$ e $\mathbb{P}([X = 2])$.

6.7 Função acumulada

Definition 6.7.1 A função distribuição acumulada de probabilidade da v.a. X é a função matemática $\mathbb{F}(x)$ definida para todo $x \in \mathbb{R}$ e dada por

$$\begin{aligned}\mathbb{F}: \mathbb{R} &\rightarrow [0, 1] \\ x &\rightarrow \mathbb{F}(x) = \mathbb{P}(X \leq x)\end{aligned}$$

Esta função é simples e não possui *nenhuma* informação adicional além daquela contida na lista de probabilidades da definição informal 6.4.1. No entanto, apesar de não trazer informação adicional a estas duas listas, ela é muito importante tanto na teoria quanto na prática de análise de dados aparecendo em testes estatísticos (como no teste de Kolmogorov) e como ferramenta para obter provas de certos teoremas. Por isto, vamos estudá-la com cuidado. Vamos começar calculando $\mathbb{F}(x)$ num caso particular.

■ **Example 6.8 — Cálculo de $\mathbb{F}(x)$ para v.a. discreta.** Suponha que temos uma v.a. aleatória discreta X com valores possíveis $\{0, 1, 2, 3\}$ e probabilidades associadas $p(k) = \mathbb{P}(X = k)$ dadas por

Valores possíveis k	1	2	3	4
Probab assoc $p(k)$	0.1	0.4	0.2	0.3

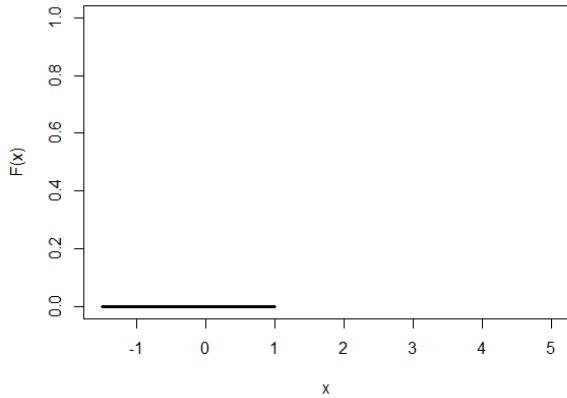
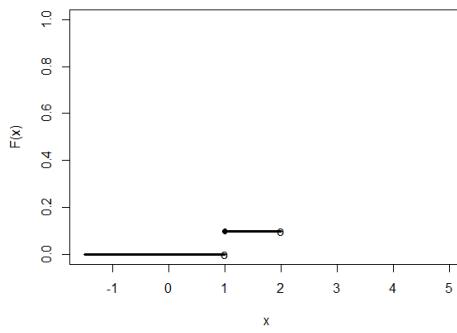
Vamos calcular $\mathbb{F}(x) = \mathbb{P}(X \leq x)$ para alguns dos valores de x . Considere, por exemplo, $x = -1$. Quanto é $\mathbb{F}(-1)$, o valor da função \mathbb{F} avaliada no ponto -1 . Pela definição de \mathbb{F} devemos obter

$$\mathbb{F}(-1) = \mathbb{P}(X \leq -1) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq -1\}).$$

Esta probabilidade é zero pois, para todo ω , temos $X(\omega) \in \{1, 2, 3, 4\}$ e portanto $X(\omega)$ é sempre um valor maior que -1 . Isto é, não existe nenhum $\omega \in \Omega$ tal que $X(\omega) \leq -1$. Assim, $\{\omega \in \Omega : X(\omega) \leq -1\} = \emptyset$ e portanto $\mathbb{F}(-1) = \mathbb{P}(X \leq -1) = \mathbb{P}(\emptyset) = 0$. Pelo mesmo argumento, para qualquer $x < 1$, teremos $\mathbb{F}(x) = \mathbb{P}(X \leq x) = 0$.

Exatamente no ponto $x = 1$, a função $\mathbb{F}(x)$ dá um salto. De fato, o evento $[X \leq 1]$ é idêntico ao evento $[X = 1]$ já que não existe nenhum ω tal que $X(\omega) < 1$:

$$\begin{aligned}\{\omega \in \Omega : X(\omega) \leq 1\} &= \{\omega \in \Omega : X(\omega) < 1\} \cup \{\omega \in \Omega : X(\omega) = 1\} \\ &= \emptyset \cup \{\omega \in \Omega : X(\omega) = 1\} \\ &= \{\omega \in \Omega : X(\omega) = 1\} \\ &= [X = 1]\end{aligned}$$

Figure 6.2: $\mathbb{F}(x) = \mathbb{P}(X \leq x)$ para $x < 1$.Figure 6.3: $\mathbb{F}(x) = \mathbb{P}(X \leq x)$ para $x < 2$.

Dessa forma, temos

$$\mathbb{F}(1) = \mathbb{P}(X \leq 1) = \mathbb{P}(X < 1) + \mathbb{P}(X = 1) = 0 + p(1) = 0.1$$

A função $\mathbb{F}(x)$ salta de 0 para $x < 1$ para 0.1 no ponto $x = 1$. Para $x = 1.5$ temos

$$\mathbb{F}(1.5) = \mathbb{P}(X \leq 1.5) \tag{6.1}$$

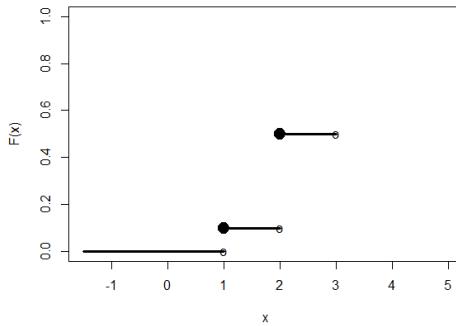
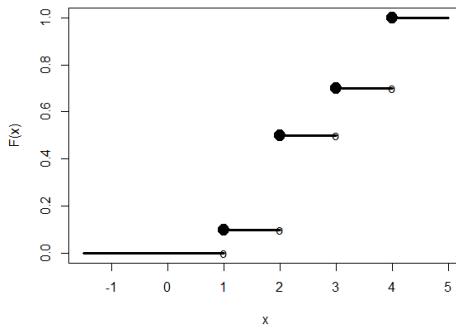
$$= \mathbb{P}([X < 1] \cup [X = 1] \cup [1 < X \leq 1.5]) \tag{6.2}$$

$$= \mathbb{P}(\emptyset \cup [X = 1] \cup \emptyset) \tag{6.3}$$

$$= \mathbb{P}(X = 1) = 0.1 \tag{6.4}$$

Pelo mesmo argumento, para qualquer x tal que $1 < x < 2$ temos $\mathbb{F}(x) = \mathbb{P}(X = 1) = 0.1$. Exatamente no ponto $x = 2$, a função $\mathbb{F}(x)$ dá mais um salto. O evento $[X \leq 2]$ é idêntico à união de dois eventos disjuntos

$$[X = 1 \text{ ou } X = 2] = [X = 1] \cup [X = 2]$$

Figure 6.4: $\mathbb{F}(x) = \mathbb{P}(X \leq x)$ para $x < 3$.Figure 6.5: $\mathbb{F}(x) = \mathbb{P}(X \leq x)$.

Eles são disjuntos pois, pela definição de uma função matemática, não podemos ter um elemento $\omega \in \Omega$ tal que $X(\omega) = 1$ e, ao mesmo tempo, $X(\omega) = 2$. Assim, temos

$$\mathbb{F}(2) = \mathbb{P}(X \leq 2) = \mathbb{P}([X = 1] \cup [X = 2]) = \mathbb{P}(X = 1) + \mathbb{P}(X = 2) = p(1) + p(2) = 0.1 + 0.4 = 0.5$$

Veja que a altura do salto é igual a $p(2)$, a probabilidade $p(2) = \mathbb{P}(X = 2)$. Para qualquer x entre 2 e 3, tal como $x = 2.72$, temos

$$\mathbb{F}(x) = \mathbb{P}(X \leq x) = \mathbb{P}(X \leq 2) + \mathbb{P}(2 < X \leq x) = \mathbb{P}(X = 1) + \mathbb{P}(X = 2) + 0 = p(1) + p(2) = 0.5$$

Continuando desta forma, vemos que $\mathbb{F}(x)$ vai dar saltos em $x = 3$ e $x = 4$. A altura do salto em $x = k$ é igual à probabilidade $p(k) = \mathbb{P}(X = k)$. Quando escolhermos um valor x maior que todos os pontos possíveis de X teremos $\mathbb{F}(x) = 1$. Por exemplo, se $x = 4.5$, claramente teremos

$$\mathbb{F}(4.5) = \mathbb{P}(X \leq 4.5) = 1$$

pois, com certeza, teremos $X \leq 4.5$ já que o maior valor possível de X é 4. O gráfico completo de $\mathbb{F}(x)$ é mostrado a seguir.

■

Caso geral de $\mathbb{F}(x)$

Suponha que temos uma v.a. aleatória discreta X com valores possíveis x_i e probabilidades associadas $p(x_i) = \mathbb{P}(X = x_i)$ dadas por

Valores possíveis	x_1	x_2	x_3	...
Probab assoc	$p(x_1)$	$p(x_2)$	$p(x_3)$...

Como a função distribuição acumulada de probabilidade é definida como:

$$\mathbb{F}(x) = \mathbb{P}(X \leq x) = \sum_{x_i \leq x} p(x_i)$$

então $\mathbb{F}(x)$ é o valor acumulado (a soma) das probabilidades $p(x_i)$ dos pontos possíveis x_i que são menores ou iguais a x .

6.8 Valor Esperado $\mathbb{E}(X)$

Definition 6.8.1 — Esperança matemática $\mathbb{E}(X)$ de uma v.a. X . O valor esperado $\mathbb{E}(X)$ de uma v.a. discreta X é a soma dos seus valores possíveis ponderados pelas suas respectivas probabilidades. Suponha que temos uma v.a. aleatória discreta X com valores possíveis x_i e probabilidades associadas $p(x_i) = \mathbb{P}(X = x_i)$ dadas por

Valores possíveis	x_1	x_2	x_3	...
Probab assoc	$p(x_1)$	$p(x_2)$	$p(x_3)$...

Então, por definição, temos

$$\mathbb{E}(X) = \sum_i x_i p(x_i)$$

O valor esperado $\mathbb{E}(X)$ também é chamado de *esperança matemática* da v.a. X .

A esperança $\mathbb{E}(X)$ é um valor teórico, matemático, associado com a distribuição de probabilidade da v.a X . Não é necessário nenhum dado estatístico para calcular $\mathbb{E}(X)$. Bastam as duas listas, a de valores possíveis e a de probabilidades associadas.

■ **Example 6.9** A distribuição de probabilidade de uma v.a. X é especificada pelas duas listas abaixo:

Valores possíveis k	1	2	3	4
Probab assoc $p(k)$	0.1	0.4	0.2	0.3

O valor esperado de X é igual a

$$\mathbb{E}(X) = \sum_{k=1}^4 k\mathbb{P}(X = k) = 1 \times 0.1 + 2 \times 0.4 + 3 \times 0.2 + 4 \times 0.3 = 2.7$$

■

Observe pelo exemplo acima que o valor mais provável da v.a. X é 2, co $\mathbb{P}(X = 2) = 0.4$. Portanto, $\mathbb{E}(X) = 2.7$ não é o valor provável. $\mathbb{E}(X)$ também não corresponde a nenhum dos valores possíveis da v.a. X .

6.9 Interpretando $\mathbb{E}(X)$

Qual o significado empírico deste número $\mathbb{E}(X)$? Como interpretá-lo na prática já que ele não corresponde ao valor mais provável e nem precisa ser um dos valores possíveis da v.a.? Suponha

uma v.a. discreta X com valores possíveis x_i e probabilidades associadas $p(x_i) = \mathbb{P}(X = x_i)$. Temos uma enorme amostra de N instâncias independentes de X . Nesta amostra, x_i apareceu N_i vezes. Podemos estimar as probabilidades pela frequência relativa da ocorrência de x_i na amostra:

$$p(x_i) = \mathbb{P}(X = x_i) \approx \frac{N_i}{N}$$

Assim,

$$\mathbb{E}(X) = \sum_i x_i p(x_i) \approx \sum_i x_i \frac{N_i}{N}$$

Como x_i apareceu N_i vezes na amostra, isto é o mesmo que somar todos os N valores da amostra e dividir por N :

$$\mathbb{E}(X) \approx \sum_i \frac{x_i N_i}{N} = \frac{x_1 + \dots + x_{N_1} + x_2 + \dots + x_{N_2} + \dots}{N}$$

onde x_1 aparece N_1 vezes, x_2 aparece N_2 vezes, etc. Isto é, se a amostra é grande, devemos ter o número teórico $\mathbb{E}(X)$ aproximadamente igual à *média aritmética* dos N elementos da amostra.

Vamos reforçar: $\mathbb{E}(X)$ é um número real, uma constante, associada com as duas listas, a de valores possíveis e a de probabilidades associadas, que constituem uma v.a. $\mathbb{E}(X)$ não é, ela mesma, uma v.a. pois não tem duas listas, é apenas um número. $\mathbb{E}(X)$ é um *resumo teórico* da distribuição de X , ou um resumo das duas listas. É aproximadamente igual à média aritmética dos valores de uma grande amostra de instâncias de X .

Talvez você se pergunte: e se a série de valores não convergir? No capítulo ??, veremos que, nas situações usuais de análise de dados, a probabilidade disso acontecer é zero. Este resultado é chamado de Lei Forte dos Grandes números.

6.10 Principais Distribuições Discretas

Existem infinitas distribuições de probabilidade. Dado um conjunto de valores possíveis, qualquer atribuição de números não-negativos que somem 1 constituem uma distribuição de probabilidade. Entretanto, algumas poucas distribuições recebem nomes especiais. Estas distribuições aparecem com frequência na análise de dados e são matematicamente tratáveis. Podemos pensar no analista de dados abordando um problema prático com um saco de distribuições de probabilidade bem conhecidas. Ele gostaria de não precisar inventar uma nova distribuição mas sim, de usar uma das aquelas que já estão no seu embornal. Vamos ver algumas das mais populares agora. Elas são as seguintes:

- Bernoulli $Ber(\theta)$;
- Binomial $Bin(n, \theta)$;
- Multinomial $M(n; \theta_1, \dots, \theta_k)$;
- Geométrica $Geo(\theta)$;
- Zipf e Pareto.

6.11 Bernoulli

É a distribuição discreta mais simples possível: dois resultados possíveis apenas. $X(\omega)$ só assume dois valores possíveis: 0 ou 1. Costumamos dizer que o valor corresponde a um sucesso e o valor 0 a um fracasso. Estes são apenas nomes para as duas categorias, sem maior significado. Por exemplo, podemos dizer a morte de um indivíduo num período corresponde a um sucesso.

Temos $X(\omega) \in \{0, 1\}$ para todo $\omega \in \Omega$. Definimos duas probabilidades:

$$p(1) = \mathbb{P}(X = 1) = \mathbb{P}(\omega \in \Omega : X(\omega) = 1)$$

$$p(0) = \mathbb{P}(X = 0) = \mathbb{P}(\omega \in \Omega : X(\omega) = 0)$$

Temos $p(0) + p(1) = 1$ o que implica que $p(1) = 1 - p(0)$. É comum escrever $p(1) = \theta$ e $p(0) = 1 - \theta$. Outra notação comum é $p(1) = p$ e $p(0) = q$.

Tipicamente, temos $0 < \theta < 1$. Se $\theta = 0$, a chance de um sucesso é zero e só vamos observar fracassos. Do mesmo modo, o caso $\theta = 1$ é um caso extremo, em que o sucesso ocorre com certeza absoluta.

Se $p(1) = \theta$ e $p(0) = 1 - \theta$, temos

$$\mathbb{E}(X) = 1 \times \theta + 0 \times (1 - \theta) = \theta$$

Observe que $\mathbb{E}(X) = \theta$ não é igual a nenhum valor possível de X , que são apenas 0 ou 1.

Se tivermos uma grande amostra de instâncias de X , cada uma delas igual a 0 ou 1, devemos ter o valor de $\mathbb{E}(X) = \theta$ aproximadamente igual a média aritmética dos valores 0 ou 1 observados. Mas uma média aritmética de valores 0 ou 1 é apenas a proporção de 1's na amostra. Isto é, como obviamente esperado, devemos ter

$$\mathbb{E}(X) \approx \hat{\theta} = \frac{1}{N} \sum_i x_i.$$

Em resumo, temos a definição

Definition 6.11.1 — Distribuição de Bernoulli. A v.a. X possui distribuição Bernoulli com parâmetro $\theta \in [0, 1]$ se $X(\omega) \in \{0, 1\}$ com $\mathbb{P}(X = 1) = \theta$ e $\mathbb{P}(X = 0) = 1 - \theta$.

Notation 6.3. Escrevemos que $X \sim Ber(\theta)$ para significar que a v.a. X possui distribuição Bernoulli com parâmetro θ .

6.12 Binomial

A distribuição Binomial corresponde ao número de sucessos em n repetições *independentes* de um experimento binário (de Bernoulli). A probabilidade de sucesso é constante e igual a $\theta \in [0, 1]$ em todas as repetições. Como a v.a. X conta o número total de sucessos em n repetições, a lista de valores possíveis é formada por $\{0, 1, 2, \dots, n\}$. A lista de probabilidades associadas é dada por $\{(1 - \theta)^n, n\theta(1 - \theta), \dots, \theta^n\}$. A fórmula geral dessas probabilidades é a seguinte:

$$\mathbb{P}(X = k) = \frac{n!}{k!(n-k)!} \theta^k (1 - \theta)^{n-k}.$$

Vamos explicar como chegar nesta fórmula após apresentar a distribuição multinomial, na próxima seção.

Temos $\mathbb{E}(X) = n\theta$. Este resultado é intuitivo. Se temos a probabilidade de sucesso numa repetição igual a, por exemplo, $\theta = 0.20$ devemos esperar que 20% das repetições sejam sucesso. Isto é, devemos esperar que a proporção de sucessos seja igual a 0.20. Assim, o número de sucessos em n repetições que se espera observar é igual a $n \times 0.20$. A prova formal exige que usemos alguns

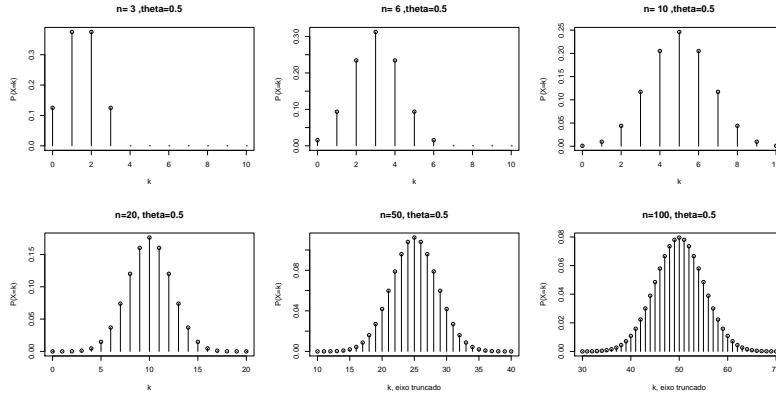


Figure 6.6: $\mathbb{P}(X = k)$ com $\theta = 1/2$ e diferentes valores para n para $X \sim \text{Bin}(n, \theta)$.

truques algébricos.

$$\begin{aligned}
 \mathbb{E}(X) &= \sum_{k=0}^n k \binom{n}{k} \theta^k (1-\theta)^{n-k} \\
 &= n\theta \sum_{k=0}^n k \frac{(n-1)!}{(n-k)!k!} \theta^{k-1} (1-\theta)^{(n-1)-(k-1)} \\
 &= n\theta \sum_{k=1}^n \frac{(n-1)!}{((n-1)-(k-1))!(k-1)!} \theta^{k-1} (1-\theta)^{(n-1)-(k-1)} \\
 &= n\theta \sum_{k=1}^n \binom{n-1}{k-1} \theta^{k-1} (1-\theta)^{(n-1)-(k-1)} \\
 &= n\theta \sum_{\ell=0}^{n-1} \binom{n-1}{\ell} \theta^\ell (1-\theta)^{(n-1)-\ell} \quad \text{com } \ell = k-1 \\
 &= n\theta \sum_{\ell=0}^m \binom{m}{\ell} \theta^\ell (1-\theta)^{m-\ell} \quad \text{com } m = n-1 \\
 &= n\theta(\theta + (1-\theta))^m \\
 &= n\theta
 \end{aligned}$$

Definition 6.12.1 — Distribuição Binomial. A v.a. X possui distribuição Binomial com parâmetros n e $\theta \in [0, 1]$ se $X(\omega) \in \{0, 1, \dots, n\}$ com

$$\mathbb{P}(X = k) = \frac{n!}{k!(n-k)!} \theta^k (1-\theta)^{n-k}$$

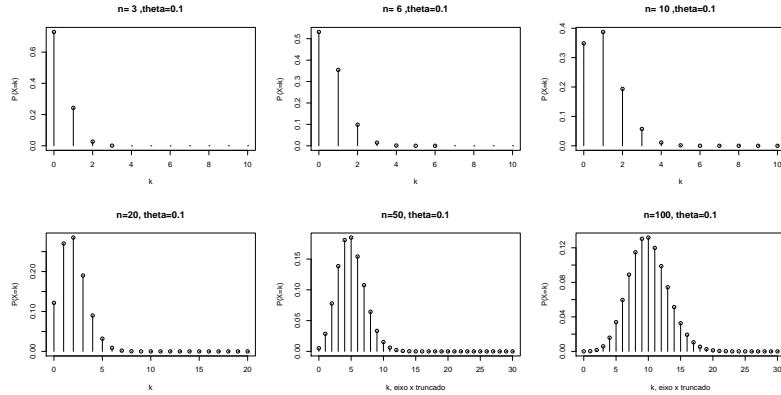
para $k = 0, 1, \dots, n$.

Notation 6.4. Escrevemos que $X \sim \text{Bin}(n, \theta)$ para significar que a v.a. X possui distribuição Binomial com parâmetros n e θ .

A forma da distribuição binomial depende de θ e de n . A Figura 6.6 mostra a função de probabilidade $\mathbb{P}(X = k)$ com $\theta = 1/2$ e diferentes valores para n .

A Figura 6.12 $\theta = 0.1$ com diferentes valores para n .

■ **Example 6.10 — Testes de soros ou vacinas.** Este exemplo foi retirado de [10]. Uma doença atinge o gado de certa região com uma incidência de 25%. Procura-se testar a eficácia de uma vacina recentemente descoberta. Injetamos a vacina em n animais sadios. Como avaliar o resultado?



Imaginamos um experimento binomial. Seja Y o número sadios após algum tempo depois da aplicação da vacina e após estarem expostos da maneira usual á doença. Para cada animal, teremos um sucesso caso ele permaneça sadio. Se a vacina for completamente inócuia teremos um sucesso em cada animal com probab $\theta = 0.75$. Ainda no caso de uma vacina inócuia, assumindo que os resultados de diferentes animais são independentes, teremos $Y \sim \text{Bin}(n, \theta)$.

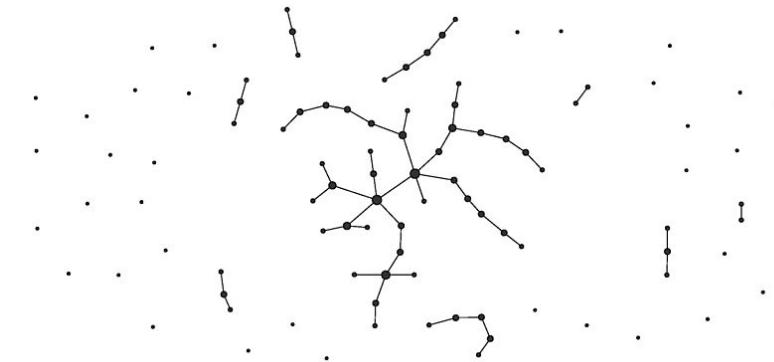
A probabilidade de que k dos n animais estejam sadios é $\mathbb{P}(Y = k) = n!/(k!(n-k)!)\theta^k(1-\theta)^{n-k}$. Se tivermos usado $n = 10$ animais, a chance de que todos estejam sadios é igual a $\mathbb{P}(Y = 10) = 0.75^{10} = 0.056$. Se tivermos usado $n = 12$ animais, a probabilidade de todos estarem sadios vale $\mathbb{P}(Y = 12) = 0.75^{12} = 0.032$, um valor menor que o anterior. Assim, se num total de 10 ou 12 animais, nenhum é contaminado, teremos uma forte indicação de que o soro teve algum efeito. A razão é que, se a vacina fosse inócuia (e portanto $\theta = 0.75$), dificilmente observaríamos sadios todos os $n = 10$ ou $n = 12$ animais. Embora esse resultado não se constitua em prova conclusiva.

Vimos que, com $n = 10$, tivemos $\mathbb{P}(Y = 10) = 0.056$. Sem a vacina, a probabilidade de que dentre 17 animais, *no máximo* um deles fique infectado é igual a $\mathbb{P}(Y \leq 1) = \mathbb{P}(Y = 0) + \mathbb{P}(Y = 1) = 0.75^{17} + 17 \times 0.75^{16} \times 0.25 = 0.0501$. Portanto, a evidência a favor da vacina é *mais forte* quando há 1 contaminado em 17 do que quando há 0 em 10! Para $n = 23$, temos $\mathbb{P}(Y \leq 2) = 0.0492$. Assim, 2 ou menos infectados em 23 é, outra vez, uma evidência mais forte em favor da vacina, do que 1 em 17 ou 0 em 10.

■ **Example 6.11 — Binomial em redes sociais.** A distribuição binomial aparece com destaque no modelo de Erdős-Rényi para grafos sociais. Cada ator numa rede social é um vértice num grafo. Arestras conectam os amigos. Temos n vértices formando $n(n-1)/2$ pares de possíveis arestras não-direcionadas. Para cada par de vértices, jogue uma “moeda” com probabilidade de sucesso igual a θ . Se der cara, conecte-os por uma arestra. A moeda de probabilidade θ é lançada independentemente para cada um dos $n(n-1)/2$ pares de vértices.

Fixe um vértice qualquer e seja Y o número de arestas incidentes. Y conta o número de amigos conectados ao vértice em questão. Então $Y \sim \text{Bin}(n-1, \theta)$. Veja que $\mathbb{E}(Y) = (n-1)\theta \approx n\theta$ se n for grande. A Figura 6.11 mostra um grafo gerado pelo modelo binomial de Erdős and Rényi com $\theta = 0.01$ e $n = 100$. Observe que esperamos ao redor de $n\theta = 100 \times 0.01 = 1$ amigo conectado em cada vértice. Alguns vértices não possuem nenhuma aresta, outros possuem duas ou três.

Nosso curso não é sobre redes sociais e portanto não vamos explorar este modelo. Apenas como curiosidade, vamos comentar sobre alguns dos resultados probabilísticos provados por [9] sobre grafos aleatórios gerados por este modelo supondo que $n \rightarrow \infty$. Considere $n\theta \approx \mathbb{E}(Y)$, o número esperado de vizinhos de um vértice qualquer. O tipo de grafo aleatório que vai ser gerado depende do valor esperado do número de amigos de um vértice. É claro que quanto maior o valor de $\mathbb{E}(Y)$, mais denso será o grafo resultante. É difícil falar algo geral quando o número n de vértices é



pequeno. Entretanto, quando n começa a crescer, emerge uma estabilidade probabilística. Dado um grafo não-direcionado, podemos olhar para o seu maior componente conectado, também chamado de componente gigante. Este é o maior subgrafo extraído do grafo original tal que qualquer par de vértices possui pelo menos um caminhos conectando-os. Quando n cresce, este subgrafo pode dominar o grafo. Temos

- Se $n\theta > 1$ e se n cresce então o grafo terá um componente gigante da ordem de n (da mesma ordem de grandeza que o grafo completo) e o segundo maior componente será de ordem $\leq O(\log(n))$ (ou seja, muito menor que o grafo original).
- Além disso, o grafo gerado quase certamente não terá um componente conectado maior que $O(\log(n))$.
- Se $n\theta > (1 + \varepsilon)\log(n)$ (um pouco maior que $\log(n)$), então o grafo quase certamente será completamente conectado.
- Por outro lado, se $n\theta < (1 - \varepsilon)\log(n)$ então o grafo quase certamente terá vértices isolados
- Etc, etc, etc... vários bonitos e não-óbvios no artigo [9].

Como saber se um dado grafo foi gerado pelo modelo de Erdős e Rényi? Uma maneira óbvia é comparar a distribuição do número de vizinhos realmente observada no grafo real com a distribuição derivada do modelo de Erdős-Rényi. Contamos a proporção de vértices isolados, a proporção de vértices com 1 amigo, a proporção de vértices com 2 amigos, etc. Em seguida, comparamos estas frequências com a probabilidade $\mathbb{P}(Y = k)$ de que um vértice possua um número Y de amigos igual a k . Se forem muito diferentes as frequências relativas e as probabilidades teóricas, teremos razões para desconfiar do modelo de Erdős e Rényi como um modelo gerador para o grafo observado. Caso contrário, teremos alguma evidência de que o modelo pode ser o gerador do grafo. Como medir a distância entre as frequências relativas e as probabilidades teóricas, entre o que observamos e o que esperamos sob o modelo? Temos uma resposta genérica para isto: usando o teste qui-quadrado, a ser visto na seção ??.

6.13 Distribuição Multinomial

A distribuição multinomial é uma generalização da distribuição binomial. A binomial conta o número de sucessos em n repetições de um experimento *binário*. Em cada repetição temos duas categorias para classificar o resultado: sucesso ou fracasso. Quando tivermos mais de duas categorias em cada repetição, teremos a distribuição multinomial. Na distribuição multinomial nós também repetimos um experimento independentemente n vezes. Entretanto, em cada experimento, existem k possibilidades e não apenas duas, como na binomial. O resultado do experimento é a contagem de quantas vezes cada uma das k possibilidades apareceu nas n repetições.

O exemplo canônico da distribuição multinomial é o lançamento de um dado. Imagine que um dado é lançado n vezes. Em cada repetição ocorre uma “categoria”: 1, 2, 3, 4, 5 ou 6. As

probabilidades de cada categoria são: $\theta_1, \theta_2, \dots, \theta_6$. Se o dado for perfeitamente balanceado, as seis probabilidades θ_i serão todas iguais a $1/6$. Se o dado for desbalanceado, elas não serão todas iguais. Elas deverão ser números entre 0 e 1 e devemos ter $\theta_1 + \dots + \theta_6 = 1$.

Qualquer que seja o dado, balanceado ou não, ao fim dos n lançamentos teremos as contagens:

$$\begin{aligned} N_1 &= \text{no. de lançamentos na cat. 1} \\ N_2 &= \text{no. de lançamentos na cat. 2} \\ \vdots &= \vdots \\ N_6 &= \text{no. de lançamentos na cat. 6} \end{aligned}$$

O resultado é um vetor aleatório multinomial com 6 posições contando o número de ocorrência de cada categoria.

Notation 6.5 (Distribuição Multinomial).

$$(N_1, N_2, \dots, N_6) \sim \mathcal{M}(n; \theta_1, \dots, \theta_6)$$

A binomial pode ser vista como um caso simples da multinomial. Seja $X \sim \text{Bin}(n, \theta)$, onde X é o número de sucessos em n repetições de um experimento binário. De forma bastante redundante, poderíamos registrar o fenômeno aleatório na forma de um vetor com o número de sucessos e o número de fracassos: $(X, n - X)$. Este vetor é uma multinomial com duas categorias. Na nossa notação, teríamos $(X, n - X) \sim \mathcal{M}(n; \theta, 1 - \theta)$.

Voltando ao caso do dado desbalanceado lançado n vezes, temos:

$$\mathbf{N} = (N_1, N_2, \dots, N_6) \sim \mathcal{M}(n; \theta_1, \dots, \theta_6)$$

Qual o suporte deste vetor aleatório \mathbf{N} ? Para qualquer sequência de lançamentos, o resultado será um vetor (N_1, \dots, N_6) de inteiros ≥ 0 com $n_1 + \dots + n_6 = n$. Assim, o número de valores possíveis para \mathbf{N} será um número finito. Embora finito, este número será bem grande a menos que n seja muito pequeno.

Quais as probabilidades associadas aos elementos do suporte? Vamos calcular um caso particular antes de dar a fórmula geral. Usando $n = 8$ lançamentos do dado, vamos calcular a probabilidade

$$\mathbb{P}(\mathbf{N} = (2, 0, 2, 1, 0, 3))$$

Isto é, queremos a chance de rolar o dado 8 vezes e terminar tendo a face 1 aparecendo duas vezes, a face 2 nenhuma vez, a face 3 aparecendo duas vezes, a face 4 uma vez, a face 5 zero vezes e a face 6 três vezes. Existem várias sequências ω de 8 lançamentos que levam ao resultado acima.

Por exemplo, se os 8 lançamentos sucessivos forem $\omega = (3, 1, 6, 6, 1, 4, 6, 3)$ teremos

$$\mathbf{N}(\omega) = (N_1(\omega), \dots, N_6(\omega)) = (2, 0, 2, 1, 0, 3)$$

Esta não é a única sequência produzindo estas contagens mas vamos nos concentrar nela por enquanto. Qual é a probabilidade $\mathbb{P}(\omega)$ desta sequência de 8 lançamentos? Como os lançamentos são independentes teremos:

$$\begin{aligned} \mathbb{P}(\omega = (3, 1, 6, 6, 1, 4, 6, 3)) &= \mathbb{P}(\text{sair 3 no 1o. E sair 1 no 2o. E ... sair 3 no 8o.}) \\ &= \mathbb{P}(\text{sair 3 no 1o.}) \mathbb{P}(\text{sair 1 no 2o.}) \dots \mathbb{P}(\text{sair 3 no 8o.}) \\ &= \theta_3 \theta_1 \theta_6 \theta_6 \theta_1 \theta_4 \theta_6 \theta_3 \\ &= \theta_1^2 \theta_2^0 \theta_3^2 \theta_4^1 \theta_5^0 \theta_6^3 \end{aligned}$$

Generalizando, se a sequência ω de n lançamentos tiver

- n_1 aparições da face 1
- n_2 aparições da face 2
- \vdots
- n_6 aparições da face 6

teremos

$$\mathbb{P}(\omega) = \theta_1^{n_1} \theta_2^{n_2} \theta_3^{n_3} \theta_4^{n_4} \theta_5^{n_5} \theta_6^{n_6}$$

Voltando aos $n = 8$ lançamentos do dado, vamos calcular a probabilidade $\mathbb{P}(\mathbf{N} = (2, 0, 2, 1, 0, 3))$. Seja A o evento formado por todos os ω (sequências de $n = 8$ lançamentos) tais que existem 2 1's, 0 2's, 2 3's, 0 4's, e 3 6's. Da mesma forma que calculamos antes, todo ω neste evento A terá a mesma probabilidade dada por

$$\mathbb{P}(\omega) = \theta_1^2 \theta_2^0 \theta_3^2 \theta_4^1 \theta_5^0 \theta_6^3$$

Assim,

$$\mathbb{P}(\mathbf{N} = (2, 0, 2, 1, 0, 3)) = \mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\omega) = C \times \theta_1^2 \theta_2^0 \theta_3^2 \theta_4^1 \theta_5^0 \theta_6^3$$

onde C é o número de sequências de tamanho 8 onde colocamos um elemento de $\{1, 2, \dots, 6\}$ em cada posição e em que temos exatamente 2 1's, 0 2's, ..., 3 6's.

Este número de possibilidades é igual a

$$\binom{8}{2, 0, 2, 1, 0, 3} = \frac{8!}{2!0!2!1!0!3!} = 1680$$

Este é o número de permutações distintas do vetor $\omega = (3, 1, 6, 6, 1, 4, 6, 3)$.

Definition 6.13.1 — Distribuição multinomial. Um vetor $\mathbf{N} = (N_1, N_2, \dots, N_6)$ possui distribuição multinomial com parâmetros n e $(\theta_1, \dots, \theta_k)$ com $\theta_i \geq 0$ e $\sum_i \theta_i = 1$ se o conjunto de valores possíveis são os inteiros $n_i \geq 0$ com $n_1 + \dots + n_k = n$ e probabilidades associadas dadas por

$$\mathbb{P}(\mathbf{N} = (n_1, n_2, \dots, n_k)) = \frac{n!}{n_1! n_2! \dots n_k!} \theta_1^{n_1} \theta_2^{n_2} \dots \theta_k^{n_k}$$

Notação: $\mathbf{N} = (N_1, N_2, \dots, N_k) \sim \mathcal{M}(n; \theta_1, \dots, \theta_k)$.

■ **Example 6.12** Suponha que temos uma amostra de $n = 22343$ indivíduos escolhidos independentemente da população brasileira e classificados em $k = 6$ categorias de religião. Cada indivíduo da amostra é como se fosse um lançamento de um dado desbalanceado com seis “faces” representadas pelas categorias de religião: 1: Católica, 2: Protestante, 3: Sem Religião, 4: Espírita, 5: Outras Religiões Cristãs, 6: Outras. É comum assumir as contagens aleatórias do número de pessoas em cada categoria seguem uma distribuição multinomial com probabilidades $(\theta_1, \dots, \theta_6)$ para as categorias. Estas probabilidades são conhecidas (aproximadamente) a partir de grandes pesquisas conduzidas pelo IBGE:

$$(\theta_1, \dots, \theta_6) = (0.75, 0.15, 0.07, 0.01, 0.01, 0.01).$$

Assim, dizemos que o vetor de contagens é multinomial:

$$\mathbf{N} = (N_1, N_2, \dots, N_6) \sim \mathcal{M}(22343; (0.75, 0.15, 0.07, 0.01, 0.01, 0.01))$$

A tabela abaixo mostra o resultado das contagens a parti de uma amostra de 22343 indivíduos.

Categorias i	Católica	Protestante	Sem Relig	Espírita	Outras Crist.	Outras
θ_i	0.75	0.15	0.07	0.01	0.01	0.01
N_i	16692	3398	1568	241	221	223

■ **Example 6.13** Suponha que uma amostra de $n = 538$ indivíduos escolhidos independentemente dentre pacientes com linfoma de Hodgkins (um tipo de câncer do sistema linfático) são classificados em 12 categorias de acordo com sua resposta a um certo tratamento e seu tipo histológico. As contagens estão na tabela abaixo. Temos $4 \times 3 = 12$ categorias no total e cada indivíduo é como o resultado do lançamento de um dado desbalanceado de 12 faces.

Tipo Histológico	Resposta			Total
	Positiva	Parcial	Sem Resposta	
LP	74	18	12	104
NS	68	16	12	96
MC	154	54	58	266
LD	18	10	44	72
Total	314	98	126	538

As contagens aleatórias do número de indivíduos em cada categoria seguem uma distribuição multinomial

$$\mathbf{N} = (N_1, N_2, \dots, N_{12}) \sim \mathcal{M}(538; (\theta_1, \dots, \theta_{12}))$$

Como a amostra é grande, podemos obter estimativas, valores aproximados, para as probabilidades θ_i a partir da proporção de elementos da amostra que caíram na categoria i . É o princípio frequentista de estimar a probabilidade pela proporção de vezes que o dado rolado mostrou a “face” i . Por exemplo, $\theta_7 = \mathbb{P}(\text{MC e Parcial}) \approx 54/538 \approx 0.1$.

6.14 Distribuição de Poisson

Esta distribuição recebeu o seu nome (Poisson, pronunciando-se como *puasson*) por causa do matemático francês Siméon-Denis Poisson, que viveu entre 1781 e 1840. Ele estudou vários problemas envolvendo probabilidades e usou esta distribuição em vários deles. Um grande número de situações em que ela aparece envolve a contagem do número de certas ocorrências num certo intervalo de tempo sem um limite claro para o número máximo que poderia ser obtido. Exemplos usuais seriam:

- número de colisões no tráfego de BH durante o ano.
- número de automóveis entrando na UFMG entre 7 e 8 da manhã
- número de consultas médicas que um cliente de um plano de saúde faz durante o ano

Como é uma v.a. discreta, precisamos apenas listas os valores possíveis e as probabilidades associadas.

Definition 6.14.1 — Distribuição de Poisson. Uma v.a. X possui distribuição de Poisson se o conjunto suporte é o conjunto dos números naturais $\mathbb{N} = \{0, 1, 2, \dots\}$. A probabilidade de que $X = k$ depende de uma constante positiva λ e é igual a

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

onde $k = 0, 1, 2, \dots$

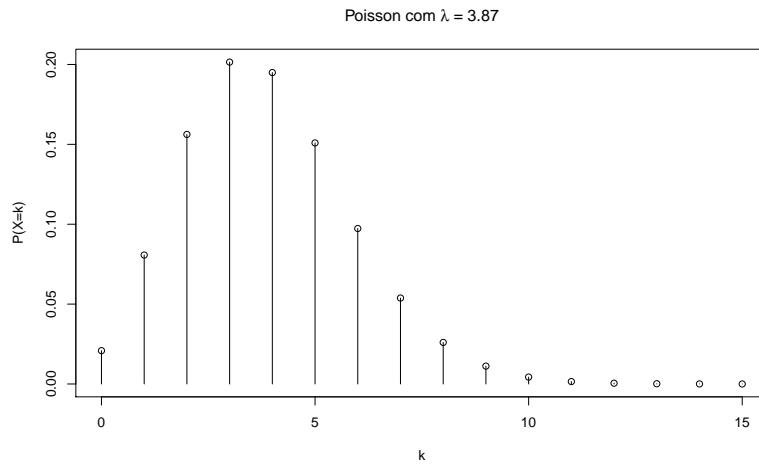


Figure 6.7: Função de probabilidade Poisson com $\lambda = 3.87$.

Na distribuição de Poisson temos esta constante $\lambda > 0$ que influencia no cálculo das probabilidades. Os valores possíveis são $0, 1, 2, \dots$ e as probabilidades associadas são dadas por:

- $\mathbb{P}(Y = 0) = \frac{\lambda^0}{0!} e^{-\lambda} = e^{-\lambda}$
- $\mathbb{P}(Y = 1) = \frac{\lambda^1}{1!} e^{-\lambda} = \lambda e^{-\lambda}$
- $\mathbb{P}(Y = 2) = \frac{\lambda^2}{2!} e^{-\lambda}$
- $\mathbb{P}(Y = 3) = \frac{\lambda^3}{3!} e^{-\lambda}$
- $\mathbb{P}(Y = 4) = \frac{\lambda^4}{4!} e^{-\lambda}$
- Etc.

Para ilustrarmos com um caso particular, vamos supor que $\lambda = 3.87$. Então

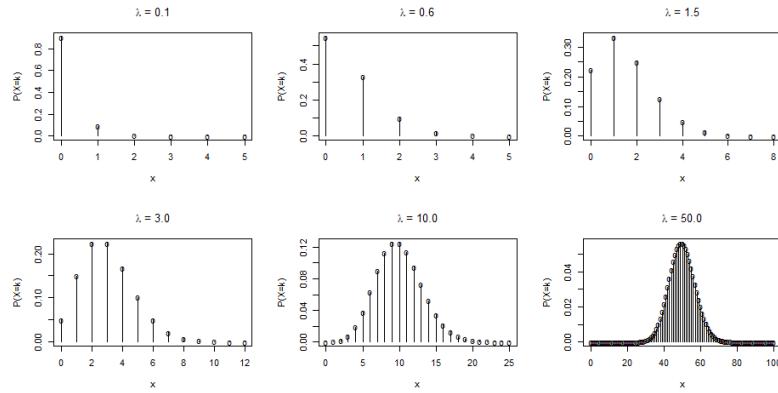
- $\mathbb{P}(Y = 0) = e^{-3.87} = 0.021$
- $\mathbb{P}(Y = 1) = 3.87 \times e^{-3.87} = 0.081$
- $\mathbb{P}(Y = 2) = 3.87^2 / 2! \times e^{-3.87} = 0.156$
- $\mathbb{P}(Y = 3) = 3.87^3 / 3! e^{-3.87} = 0.201$
- Etc.

A função de probabilidade de uma distribuição de Poisson é ilustrada na Figura 6.7. Ela mostra a função de probabilidade usando $\lambda = 3.87$.

A Figura 6.8 mostra várias funções de probabilidade Poisson variando o valor de λ . Usamos $\lambda = 0.1, 0.6$, e 1.5 nos gráficos da linha superior, indo da esquerda para a direita. Na linha de baixo, usamos $\lambda = 3, 10, 50$. Note como as probabilidades ficam concentradas nos menores inteiros quando λ é pequeno. À medida que λ cresce, as probabilidades de inteiros maiores aumentam e, ao mesmo tempo, os inteiros próximos de zero ficam com probabilidades desprezíveis. Na verdade, você deve ter observado que os inteiros que possuem probabilidades mais altas em cada gráfico são aqueles em torno do valor de λ .

De fato, isto não é uma coincidência. O valor esperado $\mathbb{E}(Y)$ de uma v.a. X com distribuição de Poisson com parâmetro λ é $\mathbb{E}(Y) = \lambda$. A prova usa a expansão de Taylor em torno do zero da função exponencial como uma série de potência:

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \dots$$

Figure 6.8: Funções de probabilidade Poisson variando o valor de λ .

Temos

$$\begin{aligned}
 \mathbb{E}(Y) &= \sum_{k=0}^{\infty} k \mathbb{P}(Y=k) = 0 \times \mathbb{P}(Y=0) + 1 \times \mathbb{P}(Y=1) + 2 \times \mathbb{P}(Y=2) + \dots \\
 &= 1 \times \lambda e^{-\lambda} + 2 \times \frac{\lambda^2}{2!} e^{-\lambda} + 3 \times \frac{\lambda^3}{3!} e^{-\lambda} \dots \\
 &= \lambda e^{-\lambda} \left(1 + \frac{2\lambda}{2!} + \frac{3\lambda^2}{3!} + \frac{4\lambda^3}{4!} + \frac{5\lambda^4}{5!} + \dots \right) \\
 &= \lambda e^{-\lambda} \left(1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \frac{\lambda^4}{4!} + \dots \right) \\
 &= \lambda e^{-\lambda} (e^\lambda) \\
 &= \lambda
 \end{aligned}$$

■ **Example 6.14 — Mortes por coices de cavalos.** A distribuição de Poisson costuma ser chamada de distribuição (da contagem de) eventos raros. Em 1898, Ladislaus von Botkiewicz foi um dos primeiros a usar a distribuição de Poisson contando a ocorrência de um evento bastante raro em seu livro *Das Gesetz der kleinen Zahlen*, que pode ser traduzido como *A lei dos pequenos números*. Ele mostrou que acontecimentos relativamente raros para um dado indivíduo, quando observado em uma grande população, apresentam regularidades que são bem aproximadas pela distribuição de Poisson. Tornaram-se clássicos os dados apresentados por ele no seu livro com o número de homens mortos por coices de cavalo em certas corporações do exército prussiano durante vinte anos (1875-1894).

Em cada um dos 20 anos, ele anotou o número de mortos em cada uma das 10 corporações. Assim, temos 200 contagens vindas de 10 corporações vezes 20 anos. Em geral, não ocorria nenhuma morte: 109 das 200 contagens foram iguais a zero. Em 65 corporações-ano tivemos apenas 1 morte. A situação completa encontra-se na segunda coluna na tabela abaixo. Ela mostra o número de corporações-ano com zero mortes, 1 morte, etc.

k mortos no ano	Frequência observada	$\mathbb{P}(Y=k)$	Frequência esperada
0	109	0.5434	108.7
1	65	0.3314	66.3
2	22	0.1011	20.2
3	3	0.0206	4.1
4	1	0.0031	0.7
Total	200	0.9995	200

A terceira coluna apresenta a fórmula de probabilidade $\mathbb{P}(Y = k)$ da distribuição de Poisson. Para isto, precisamos primeiro de um valor para o parâmetro λ . Se as 200 contagens são realizações de uma v.a. que segue a distribuição de Poisson, a média aritmética dessas 200 contagens deveria ser aproximadamente igual a $\mathbb{E}(Y) = \lambda$. Esta média aritmética é igual a $(0 \times 109 + 1 \times 65 + 2 \times 22 + 3 \times 3 + 4 \times 1)/200 = 122/200 = 0.61$. Usando $\lambda = 0.61$ calculamos $\mathbb{P}(Y = k) = 0.61^k e^{0.61}/k!$ para $k = 0, 1, 2, 3, 4$ na terceira coluna da tabela. A quarta coluna apresenta quantas corporações deveríamos esperar com zero mortes no ano, com uma morte no ano, com duas mortes no ano, etc. Por exemplo, se existe uma probabilidade igual a 0.5434 de que uma corporação-ano tenha zero mortes, ao final de 200 corporações-ano deveríamos ter aproximadamente $200 \times 0.5434 = 108.7$ com zero mortes. De modo análogo, deveríamos ter aproximadamente $200 \times 0.3314 = 66.3$ corporações-ano com uma morte. A proximidade entre as frequências observadas e esperadas mostra que estes dados poderiam estar sendo gerados por uma distribuição de Poisson e portanto ela é uma distribuição adequada para modelar estes dados. ■

■ **Example 6.15 — Errata.** Um exemplo de uso da distribuição de Poisson que costuma ser citado é o número de erros de digitação ou tipográficos numa página de documento. O professor Kim Border, do Caltech, em suas notas de aula não publicadas, conta o caso de seu colega Phil Hoffman. Este colega escreveu o livro *How Did Europe Conquer the World?* e recebeu da editora, antes de sua publicação as provas para verificação final. Em $n = 261$ páginas havia um número total de 43 erros. Havia 222 páginas com zero erros, um erro em 35 páginas e 2 erros em 4 páginas. Em média, havia $43/261 = 0.165$ erro por página e este número serve como estimativa para o parâmetro λ . A tabela com as contagens observadas e as esperadas sob o modelo de Poisson estão na tabela abaixo. A proximidade dos valores é bastante impressionante.

	0	1	2	≥ 3
obs	222	35	4	0
esp	221.4	36.5	3.0	0.17

Note que a última célula precisa do valor da probabilidade de que $\mathbb{P}(Y \geq 3)$. Como a distribuição de Poisson tem infinitos valores você terá de calcular uma série infinita: $\mathbb{P}(Y \geq 3) = \mathbb{P}(Y = 3) + \mathbb{P}(Y = 4) + \dots$. Entretanto, podemos obter este valor por subtração após obtermos as primeiras probabilidades pois

$$\mathbb{P}(Y \geq 3) = 1 - \mathbb{P}(Y = 0) - \mathbb{P}(Y = 1) - \mathbb{P}(Y = 2).$$

■

6.14.1 A gênese de uma Poisson

Como esta distribuição de Poisson aparece? Existe um exemplo clássico mostrando que as contagens de emissões de partículas radioativas por uma massa atômica segue uma distribuição de Poisson (Figura 6.9). Um contador Geiger-Müller registra o número de partículas atingindo uma placa num intervalo de 7.5 segundos. Os valores possíveis para a contagem desse número de partículas é $\mathbb{N} = \{0, 1, 2, \dots\}$. As contagens das partículas é aleatória e quaisquer duas contagens em dois períodos de tempo iguais provavelmente não serão a mesma. Existe uma boa dose de variabilidade nestas contagens, mesmo que o tempo de observação seja o mesmo.

Para obter as probabilidades, vamos adotar dois caminhos: uma mais teórico, outro mais empírico. Um modelo teórico para a emissão de partículas por uma massa radioativa foi proposto por físicos no século passado e consiste em três hipóteses que eram bem embasadas na observação prática do fenômeno radioativo.

- **Hipótese 1:** A probabilidade da chegada de k partículas num intervalo de tempo $(t, t + \Delta)$ depende apenas do comprimento Δ do intervalo e não do momento t de seu início. Isto é, tome dois intervalos de tempo (em segundos) de igual duração tais como, por exemplo,



Figure 6.9: Esquerda: Material radioativo. Direita: Um membro da equipe que cuida da segurança de um reator nuclear.

$(1, 1 + \Delta)$ (um segundo após o início do experimento) e $(7200, 7200 + \Delta)$ (duas horas após seu início). A probabilidade de observar, digamos, 5 partículas no primeiro intervalo é a mesma que no segundo intervalo. Não existe um “desgate” da massa atômica (pelo menos durante experimentos de duração não excessiva). Horas depois do início do experimento tudo se passa como se estivéssemos no início dele. A probabilidade de emitir k partículas num intervalo Δ de tempo não depende de quando começamos o intervalo. A probabilidade depende de Δ : intervalos de tempo longos terão contagens mais maiores. Entretanto, a probabilidade de contar k partículas depende apenas desse comprimento Δ e não do momento de início do período de observação.

- **Hipótese 2:** Os números de partículas em intervalos de tempo disjuntos são v.a.’s independentes. Suponha que, num intervalo de tempo de Δ segundos observemos em média 5 partículas. Se num dado intervalo $(t, t + \Delta)$ observarmos bem mais que a média de 5 partículas (digamos, com 10 partículas), então no próximo intervalo $(t + \Delta, t + 2\Delta)$ de duração Δ não haverá nenhuma tendência para corrigir o excesso do primeiro intervalo, nem nenhum estímulo para continuar emitindo mais partículas que a média de 5.
- **Hipótese 3:** As partículas chegam sozinhas, elas não simultaneamente.

Pode-se provar matematicamente (ver [17, pag. 22]) que um sistema estocástico com estas três propriedades ou hipóteses deverá ter necessariamente a distribuição de probabilidade de Poisson para as contagens num intervalo de tempo: $\mathbb{P}(Y = k) = \frac{\lambda^k}{k!} e^{-\lambda}$ para $k = 0, 1, 2, \dots$ e onde λ é uma constante positiva associada com a massa radioativa e representa o valor esperado de emissões no intervalo de tempo.

Esta dedução matemática é condizente com a realidade? [21] “repetiram” o experimento um grande número de vezes. Eles contaram o número de partículas emitidas em 2608 intervalos de tempo consecutivos de 7.5 segundos cada um. Sejam $y_1 = 4, y_2 = 3, y_3 = 0, \dots, y_{2608} = 4$ as contagens de partículas emitidas em cada intervalo. Vamos assumir que eles são os valores instanciados das v.a.’s i.i.d $Y_1, Y_2, \dots, Y_{2608}$, todas com distribuição Poisson(λ). Se este modelo Poisson para a emissão de partículas estiver correto o que podemos esperar ver nas contagens observadas? Vamos comparar os valores teóricos $\mathbb{P}(Y = k) = \frac{\lambda^k}{k!} e^{-\lambda}$ com a frequência observada de intervalos com contagens iguais a k .

Primeiro, vamos calcular $\mathbb{P}(\text{emitir } k \text{ partículas em 7.5 segundos})$ usando o modelo de Poisson. Calculamos a média aritmética das observações como uma aproximação para λ . Assim, usamos $\hat{\lambda} = (y_1 + y_2 + \dots + y_{2608})/2608 = 3.87$. Podemos agora calcular $\mathbb{P}(Y = k) = \frac{3.87^k}{k!} e^{-3.87}$ para os diferentes valores de k . Estas probabilidades estão na segunda coluna da tabela abaixo.

k	$\mathbb{P}(Y = k)$	Frequência empírica
0	0.02086	57/2608 = 0.02186
1	0.08072	203/2608 = 0.07784
2	0.15619	0.14686
3	0.20149	0.20130
4	0.19495	0.20399
5	0.15089	0.15644
6	0.09732	0.10968
7	0.05381	0.05329
8	0.02603	0.01725
9	0.01119	0.01035

A terceira coluna mostra a proporção dos 2608 intervalos em que obtivemos k partículas. Veja que não usamos nenhum modelo de probabilidade aqui, apenas os dados observados. Se o modelo estiver correto, estes dois valores devem ser parecidos para todo k . De fato, a proximidade dos valores é muito grande. Isto mostra que o modelo de Poisson pode muito bem ser o mecanismo que gera os dados observados. O modelo ajusta-se muito bem aos dados empíricos.

Existem duas situações em que a distribuição de Poisson aparece. Quando contamos número de ocorrências raras sem um limite claro para o número máximo tais como:

- número de colisões no tráfego de BH entre as 18 e 19 horas num dia de semana.
- número de automóveis entrando na UFMG entre 7 e 8 da manhã.
- número de consultas médicas que um cliente de um plano de saúde faz durante o ano.
- número de falhas por dia registradas num sistema de gerência de redes.
- número de infecções por vírus num servidor de arquivos em um data center por dia.
- número de visitas numa página Web por minuto.
- número de ligações num call center por minuto.

Se as três hipóteses acerca do decaimento radioativo também forem aproximadamente válidas para estes eventos aleatórios listados acima, podemos esperar ver a distribuição de Poisson aparecendo. Por exemplo, no primeiro caso acima, se a chance de observar 3 colisões em BH num pequeno intervalo Δ de tempo se mantiver constante entre 18 e 19 horas, se as colisões ocorrem de forma independente e não ocorrem de forma simulatânea, podemos esperar uma contagem de Poisson como resultado. Observe que em situações do mundo real, devido ao aspecto de sazonalidade, a probabilidade de ocorrer k eventos pode permanecer constante apenas dentro de intervalos de tempo limitados.

Uma outra situação em que a distribuição de Poisson aparece naturalmente é como uma aproximação para a distribuição de uma v.a. X seguindo a distribuição binomial $\text{Bin}(n, \theta)$ quando n é grande e θ é pequeno. Como $\mathbb{E}(X) = n\theta$, se uma distribuição de Poisson for uma boa aproximação, devemos ter $\lambda \approx n\theta$. Exemplos para este caso:

- número de mortos por câncer de esôfago durante o ano em BH. Cada indivíduo de BH lança uma moeda no início do ano para determinar se ele vai falecer de câncer de esôfago durante aquele ano ou não. Temos um grande número n de lançamentos da moeda e uma probabilidade de “sucesso” θ bem pequena.
- número de apólices de seguro de automóveis com 2 ou mais sinistros durante um certo ano. A carteira de apólices na seguradora possui centenas de milhares de clientes e este é o n . A probabilidade θ de que um segurado tenha dois ou mais sinistros durante o ano é pequena.

■ **Example 6.16 — Bombas em Londres.** No final da Segunda Guerra Mundial, os alemães desenvolveram os primeiros mísseis balísticos guiados de longo alcance, as bombas V1 e V2 (Figura 6.10). Eles eram lançados do continente europeu através do Canal da Mancha sobre a Inglaterra e, em especial, sobre Londres. Depois de certo tempo, certos bairros de Londres estavam

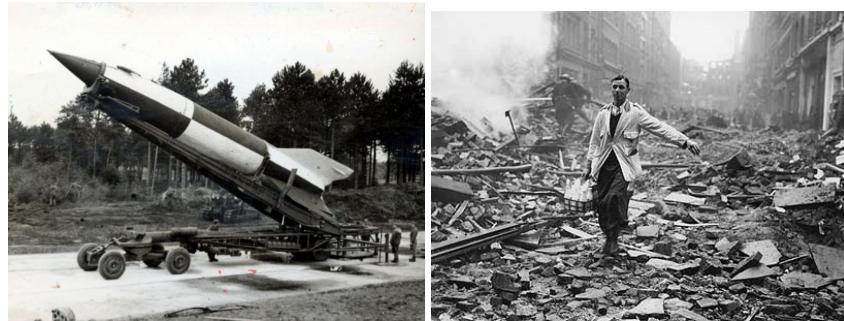


Figure 6.10: Esquerda: Bomba V2 pronta para ser lançada sobre Londres. Direita: A vida em Londres seguia durante a guerra

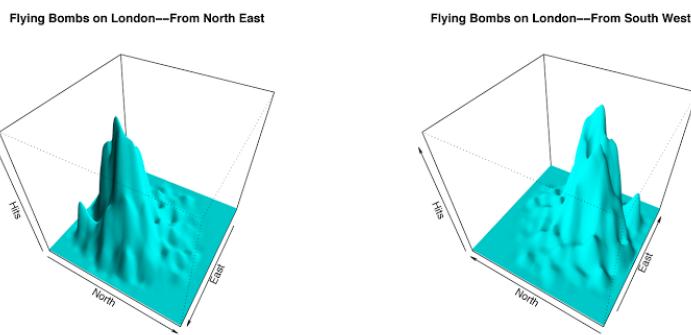


Figure 6.11: Superfície representando a densidade de bombas por km^2 em Londres a partir de duas perspectivas. Extrido blog <http://madvis.blogspot.com.br/2010/09/flying-bombs-on-london-summer-of-1944.html>.

sendo duramente atingidos, enquanto outros não. Parecia que os alemães tinham capacidades de bombardeio excepcionalmente precisas. Se não havia razão aparente para que os alemães evitassem atingir alguns quarteirões, levantou-se a suspeita de que seus espiões viviam ali. Este era um segredo militar importante da época: saber a precisão dessas bombas. Elas estavam caindo desordenadamente sobre a cidade ou estavam atingindo seus alvos pretendidos? Será que os alemães realmente haviam conseguido fazer uma bomba auto-guiada com precisão?

Em 1946, R.D.Clarke, um atuário britânico publicou uma nota descrevendo o trabalho que ele havia feito durante o período de guerra para responder a esta questão ([5]). Esta análise ficou famosa e aparece em todo livro de probabilidade. Charles Franklin, professor da University of Wisconsin, reconstruiu os dados deste artigo antigo a partir dos mapas originais dos locais das bombas guardados nos arquivos britânicos em Kew. Ele publicou uma visão tridimensional da densidade de bombas em seu blog <http://madvis.blogspot.com.br/2010/09/flying-bombs-on-london-summer-of-1944.html> e reproduzidas na Figura 6.11. Ela mostra a superfície de densidade de bombas caídas de Junho a agosto de 1944. Vemos as superfícies a partir de dois ângulos, do Nordeste (esquerda) e do Sudoeste (direita). Quanto mais alta a densidade, maior o número de impactos por km^2 .

Do ponto de vista da grande Londres, a maior densidade do ataque está bem definida, com os impactos concentrados numa certa região. Assim, obviamente, certa precisão havia. Mas havia precisão suficiente para distinguir alvos dentro dessa região de maior concentração? A análise de Clarke focou na área de maior densidade. Dentro dessa área, sua análise não encontrou nenhuma



Figure 6.12: Esquerda: Sir Churchill visitando os locais atingidos por bombas em 10 de setembro de 1940 em Londres durante a Segunda Guerra. Extraído de http://www.bbc.co.uk/history/events/germany_bombs_london. Direita: Desenho esquemático ilustrando a situação em que temos a região sul de Londres dividida em $N = 576$ quadradinhos com área igual a 0.25km^2 . Os pontos representam os locais onde bombas caíram

evidência de aglomeração que não pudesse ser explicada pela variação de uma v.a. Poisson. O raciocínio baseia-se numa grade fina com $N = 576$ quadradinhos, cada um com 0.25km^2 , na região sul de Londres, a mais atingida (ver Figura 6.12). Se os alemães não possuíssem mira, o número de bombas num desses pequenos quadrados seria um valor vindo de uma v.a. com distribuição Poisson(λ), o mesmo λ onde quer que estivesse o quadradinho. Por quê?

O raciocínio é o seguinte. Fixe um pequeno quadrado no mapa da Figura 6.12. Seja X o número de bombas no quadrado. Tivemos um grande número B de bombas sendo lançadas sobre o mapa. Existe uma pequena probabilidade θ de atingir um quadrado específico. A contagem das bombas num quadradinho específico é como contar o número de “sucessos” em B lançamentos de uma moeda com probabilidade de sucesso θ pequena. Usamos a aproximação da binomial pela Poisson com $\lambda = n\theta$. A questão crucial é que, se os alemães não têm mira, a probabilidade θ é a mesma para todo pequeno quadrado. Temos $N = 576$ quadradinhos com contagens Y_1, \dots, Y_N , todas sendo v.a.’s Poisson com o mesmo parâmetro (λ). Este modelo teórico ajusta-se aos dados? Se sim, isto seria uma evidência a favor da hipótese de não haver mira.

Como antes, calculamos as probabilidades $\mathbb{P}(Y = 0)$, $\mathbb{P}(Y = 1)$, $\mathbb{P}(Y = 2)$, etc, usando o modelo de Poisson. A seguir, obtemos a proporção dos quadradinhos em que tivemos contagens $Y = 0$, $Y = 1$, $Y = 2$, etc, usando apenas os dados empíricos. A seguir, compararmos as probabilidades teóricas de Poisson com as frequências baseadas apenas nos dados. Se forem similares, os dados são compatíveis com o modelo.

O número total de bombas em Londres foi 537 e existem $N = 576$ quadradinhos. Assim, o número médio de bombas por quadradinho é $\hat{\lambda} = 537/576 = 0.9323$. Seja Y_i o número de bombas no quadradinho i . Supomos que Y_1, \dots, Y_n são i.i.d. com distribuição Poisson(λ) com $\lambda = 0.9323$. Na tabela abaixo, k é o número de bombas num quadradinho, N_k é o número de quadradinhos que foram atingidos por k bombas, $N_k/576$ é a proporção de quadradinhos atingidos por k bombas e $\mathbb{P}(Y = k) = 0.9323^k/k!e^{-0.9323}$ é a probabilidade de uma v.a. Poisson($\lambda = 0.9323$) ser igual a k .

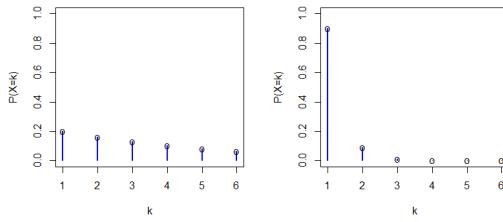


Figure 6.13: Função de probabilidade $\mathbb{P}(Y = k)$ de uma distribuição geométrica com $\theta = 0.2$ (esquerda) e $\theta = 0.9$ (direita).

É impressionante a proximidade da frequência empírica e as probabilidades teóricas.

k	N_k	$N_k/576$	$\mathbb{P}(Y = k)$
0	229	0.398	0.394
1	211	0.366	0.367
2	93	0.161	0.171
3	35	0.061	0.053
4	7	0.012	0.012
≥ 5	1	0.002	0.003
Total	576	1	1

6.15 Geométrica

Esta distribuição aparece a partir do onipresente experimento de lançar sucessivamente uma moeda com probabilidade de sucesso θ . O interesse é conhecer quantos lançamentos são necessários até que o primeiro sucesso seja observado. É claro que este tempo de espera pelo primeiro sucesso depende de θ . Se θ for um valor próximo de 1 devemos obter o primeiro sucesso logo na primeiro ou segundo lançamento, dificilmente necessitando de muitos lançamentos para parar. Por outro lado, se $\theta \approx 0$, não devemos esperar poucos lançamentos. Vamos estudar mais precisamente como se dá este tempo de espera.

Seja Y é o número de lançamentos em uma sequência de ensaios independentes de Bernoulli até que o primeiro sucesso seja observado. Em cada ensaio a probabilidade de sucesso é θ . O evento $Y = 1$ significa que o primeiro ensaio foi um sucesso S . Temos $\mathbb{P}(Y = 0) = \mathbb{P}(S) = \theta$. O evento $Y = 2$ significa que o primeiro ensaio foi um fracasso F e o segundo foi um sucesso S . Como os lançamentos da moeda são independentes, a probabilidade de observar FS é o produto das probabilidades dos resultados de cada lançamento individual: $\mathbb{P}(Y = 2) = \mathbb{P}(FS) = (1 - \theta)\theta$. O caso $Y = 3$ funciona de forma semelhante: $\mathbb{P}(Y = 3) = \mathbb{P}(FFS) = (1 - \theta)^2\theta$. E o caso geral sai fácil agora: $\mathbb{P}(Y = k) = (1 - \theta)^{k-1}\theta$, para $k = 1, 2, \dots$.

A Figura 6.13 mostra a função de probabilidade $\mathbb{P}(Y = k)$ de uma distribuição geométrica com $\theta = 0.2$ (esquerda) e $\theta = 0.9$ (direita). O eixo vertical é o mesmo nos dois gráficos. Veja como as probabilidades $\mathbb{P}(Y = k)$ são concentradas fundamentalmente nos valores $k = 1$ e $k = 2$ quando $\theta = 0.9$. Podemos calcular a chance de precisar de jogar a moeda 3 ou mais vezes para obter o primeiro sucesso neste segundo caso: $\mathbb{P}(Y \geq 3) = 1 - \mathbb{P}(Y = 1) - \mathbb{P}(Y = 2) = 1 - 0.9 - 0.9 \times 0.1 = 0.01$. Assim, apenas 1% das vezes que fizermos este experimento com uma moeda com $\theta = 0.9$ podemos esperar ter de lançar a moeda 3 vezes ou mais. Já no outro caso, com $\theta = 0.2$, temos $\mathbb{P}(Y \geq 3) = 1 - 0.2 - 0.2 \times 0.8 = 0.64$. Portanto, neste caso, a chance de ter de esperar 3 ou mais lançamentos é maior que a de parar antes de 3.

Quando Y é geométrica com parâmetro de sucesso igual a θ , podemos mostrar que $\mathbb{E}(Y) = 1/\theta$. Assim, se $\theta = 0.1$ esperamos ter de lançar a moeda $\mathbb{E}(Y) = 1/0.1 = 10$ vezes antes de observar o primeiro sucesso. Se a moeda é honesta e $\theta = 0.5$ então $\mathbb{E}(Y) = 1/0.5 = 2$ enquanto $\mathbb{E}(Y) = 1/0.01 = 100$ se $\theta = 0.01$.

Vamos agora calcular $\mathbb{E}(Y)$ usando truques simples de séries infinitas. Lembre-se da série geométrica: se $x \in (0, 1)$ então

$$1 + x + x^2 + x^3 + \dots = \frac{1}{1-x}.$$

Temos

$$\begin{aligned}\mathbb{E}(Y) &= \sum_{k=1}^{\infty} k\mathbb{P}(Y=k) = 1 \times \mathbb{P}(Y=1) + 2 \times \mathbb{P}(Y=2) + \dots \\ &= 1 \times \theta + 2 \times \theta \times (1-\theta) + 3 \times \theta \times (1-\theta)^2 + \dots \\ &= \theta(1 + 2 \times (1-\theta) + 3 \times (1-\theta)^2 + \dots) \\ &= \theta(0 + 1 + 2 \times (1-\theta) + 3 \times (1-\theta)^2 + \dots) \\ &= \theta \left(\frac{d}{d\theta}(-1) - \frac{d}{d\theta}(1-\theta) - \frac{d}{d\theta}(1-\theta)^2 - \frac{d}{d\theta}(1-\theta)^3 + \dots \right) \\ &= \theta \frac{d}{d\theta} (-1 \times (1+(1-\theta)) + (1-\theta)^2 + (1-\theta)^3 + \dots) \\ &= \theta \frac{d}{d\theta} \left(-\frac{1}{1-(1-\theta)} \right) \\ &= \theta \frac{d}{d\theta} \left(-\frac{1}{\theta} \right) \\ &= \frac{1}{\theta}\end{aligned}$$

Definition 6.15.1 — Distribuição geométrica. A v.a. Y possui distribuição geométrica se o seu suporte for o conjunto $\{1, 2, \dots\}$ com função de probabilidade dada por $\mathbb{P}(Y=k) = (1-\theta)^{k-1}\theta$, para $k = 1, 2, \dots$

6.16 Distribuição de Pareto ou Zipf

A distribuição de Pareto ganhou muita importância em tempos mais recentes quando vários fenômenos exibiram o comportamento típico que encontramos nestas distribuições, chamado de *comportamento de cauda pesada*. No caso discreto, ela costuma ser chamada de distribuição de Zipf.

Definition 6.16.1 — Distribuição de Zipf (Pareto discreta). Seja X uma v.a. com suporte igual ao conjunto $\{1, 2, 3, \dots, N\}$. O valor do número máximo N pode ser finito ou infinito. Ela é chamada *distribuição de Zipf* ou *Pareto discreta* se as probabilidades forem da seguinte forma: $\mathbb{P}(X=k) = \frac{C}{k^{1+\alpha}}$ com $\alpha > 0$ onde $C > 0$ é uma constante tal que as probabilidades somam 1.

Quando $\alpha = 1$, temos $\mathbb{P}(X=k) = C/k^2$. Se $\alpha = 2.5$, temos $\mathbb{P}(X=k) = C^*/k^{3.5}$. Usamos C^* neste caso apenas para enfatizar que a constante para o caso $\alpha = 2.5$ é diferente daquela para o caso $\alpha = 1$. Podemos ter $0 < \alpha < 1$. Por exemplo, se $\alpha = 0.5$ então $\mathbb{P}(X=k) = C^{**}/k^{1.5}$. O que *realmente* importa é o seguinte: a probabilidade $\mathbb{P}(Y=k)$ decresce de acordo com uma potência de k . Por isto, ela é chamada de uma distribuição de lei de potência (*power law*, em inglês). Ela não cai com uma rapidez exponencial como é o caso de uma Poisson e uma geométrica. Voltamos a comparar estas distribuições no final desta seção.

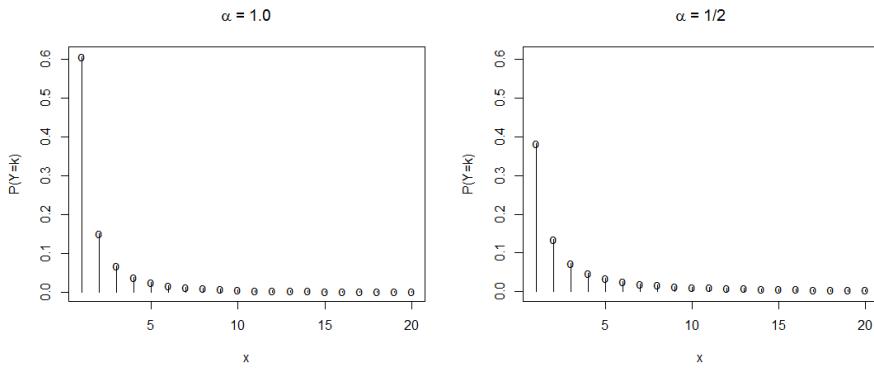


Figure 6.14: Probabilidades $\mathbb{P}(Y = k) = c/k^{1+\alpha}$ de uma distribuição discreta de Pareto (ou de Zipf) com $\alpha = 1$ (esquerda) e $\alpha = 1/2$ (direita). A escala do eixo vertical é a mesma nos dois casos.

A constante C na fórmula acima da distribuição deve garantir que as probabilidades somem 1. Assim, se fixarmos um valor para α , o valor de C fica determinado pois devemos ter

$$\begin{aligned} 1 &= \mathbb{P}(X = 1) + \mathbb{P}(X = 2) + \mathbb{P}(X = 3) + \dots \\ &= C \left(\frac{1}{1^{1+\alpha}} + \frac{1}{2^{1+\alpha}} + \frac{1}{3^{1+\alpha}} + \dots \right) \\ &= C \sum_{k=1}^N \frac{1}{k^{1+\alpha}} \end{aligned}$$

o que implica que

$$C = \frac{1}{\sum_{k=1}^N 1/k^{1+\alpha}}.$$

Quando $N = \infty$, esta constante está associada com a função zeta de Riemann definida como

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}.$$

para $s > 1$. Esta função foi estudada extensamente em matemática. Usando os valores calculados para esta função zeta de Riemann, quando $N = \infty$ e $\alpha = 0.5$, devemos ter uma constante $C = 1/\zeta(1.5) \approx 1/2.612 = 0.383$, enquanto $C = 1/\zeta(2.0) \approx 1/1.645 = 0.608$ quando $\alpha = 1.0$.

A Figura 6.14 mostra a função de probabilidade $\mathbb{P}(Y = k)$ de uma distribuição de Pareto com $N = \infty$ e $\alpha = 1$ (esquerda) e $\alpha = 0.5$ (direita). O eixo vertical é o mesmo nos dois gráficos. O valor da probabilidade $\mathbb{P}(Y = 1)$ é bastante diferente nos dois casos, sendo maior que 0.6 quando $\alpha = 1$ mas aproximadamente 0.4 quando $\alpha = 0.5$. Exceto por este valor bem diferente, é muito difícil identificar visualmente diferenças marcantes entre os dois gráficos. E no entanto elas existem. Sabemos que a soma das probabilidades é igual a 1 nos dois casos. Então, a diminuição de 0.2 no valor de $\mathbb{P}(Y = 1)$ ao passar do gráfico da esquerda para o da direita precisa ter vazado para as demais probabilidades que ganharam estes 0.2 adicionais. Mas, no gráfico da direita, quase não vemos as alturas de $\mathbb{P}(Y = k)$ maiores que as do lado esquerdo quando $k \geq 2$. Mas essas diferenças estão lá.

Para se ter uma ideia da diferença entre os dois gráficos e começar a entender o comportamento de cauda pesada, vamos calcular $\mathbb{P}(Y \leq k)$ em cada caso, com $\alpha = 1$ e com $\alpha = 0.5$. A Figura 6.15 mostra o gráfico da função de distribuição de probabilidade acumulada $F(k) = \mathbb{P}(Y \leq k)$ para diferentes valores de k com $\alpha = 1$ (esquerda) e $\alpha = 1/2$ (direita). Agora sim, vemos uma

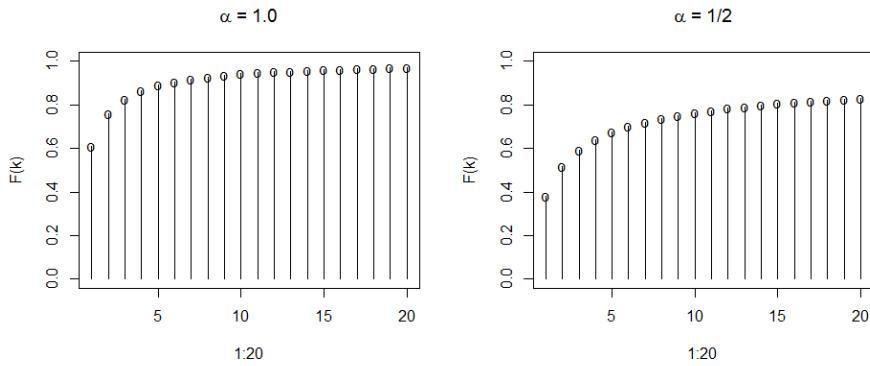


Figure 6.15: Probabilidades $\mathbb{F}(k) = \mathbb{P}(Y \leq k)$ de uma distribuição discreta de Pareto (ou de Zipf) com $\alpha = 1$ (esquerda) e $\alpha = 1/2$ (direita). A escala do eixo vertical é a mesma nos dois casos.

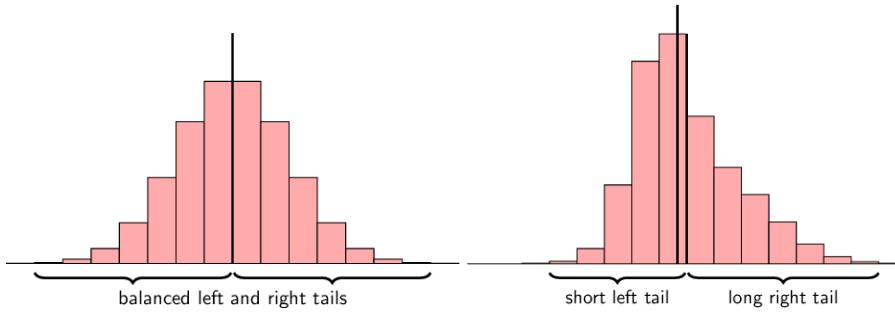


Figure 6.16: Distribuições simétrica (esquerda) e assimétrica (direita).

diferença enorme das duas distribuições, diferença que não era óbvia na Figura 6.14. A distribuição acumulada $\mathbb{F}(k)$ com $\alpha = 1$ chega mais próxima do máximo 1 mais rapidamente que a distribuição acumulada com $\alpha = 1/2$. Por exemplo, com $\alpha = 1$, temos $\mathbb{F}(15) = \mathbb{P}(Y \leq 15) = 0.96$ enquanto que, com $\alpha = 1/2$, temos apenas $\mathbb{F}(15) = 0.81$.

Entretanto, o aspecto mais notável da Figura 6.15 é exibir claramente o comportamento de cauda pesada (heavy tail). Para entender este conceito, vamos classificar as distribuições que podem ser simétricas, como no caso da distribuição no lado esquerdo da Figura 6.16, ou assimétricas, como no lado direito da Figura 6.16. Uma distribuição simétrica é aquela em que os lados esquerdo e direito da distribuição estão aproximadamente equilibrados em torno de um ponto central. Pode ser provado que, no caso simétrico, este ponto central coincide com o valor esperado $\mathbb{E}(X)$. As caudas da distribuição são as partes à esquerda e à direita, distantes do centro. A cauda é a parte onde os valores de $\mathbb{P}(X = k)$ tornam-se menores. Para uma distribuição simétrica, as caudas à esquerda e à direita são igualmente equilibradas, significando que as probabilidades $\mathbb{P}(X = k)$ têm aproximadamente o mesmo decaimento nas duas direções. No caso assimétrico, a cauda da distribuição num dos lados é mais longa do que no outro lado. No caso da Figura 6.16, o lado direito espalha-se mais que o lado esquerdo e dizemos que temos assimetria à direita.

Nos gráficos da Figura 6.14, vemos que as distribuições de Pareto são inclinadas e assimétricas à direita. Isto é, são aglomeradas à esquerda e com uma “cauda” estendendo-se para a direita. No caso da distribuição de Pareto, esta cauda à direita concentra muita probabilidade mesmo que isto não transpareça ao olhar a Figura 6.14. Olhando os gráficos da Figura 6.15, podemos perceber como a distribuição de Pareto com $\alpha = 1/2$ cai tão lentamente que mesmo valores muito grandes

de k ainda possuem probabilidade não desprezível. Na Figura 6.14 parece não haver massa de probabilidade $\mathbb{P}(X = k)$ com relevância prática para $k \geq 10$ nos dois casos ($\alpha = 1$ e $\alpha = 1/2$). Entretanto, nos dois casos, ainda existe uma probabilidade não desprezível de que vejamos valores de X maiores que 10. Temos $\mathbb{P}(X > 10) = 0.06$ no caso $\alpha = 1$ e $\mathbb{P}(X > 10) = 0.24$ no caso $\alpha = 1/2$. Isto é, se você gerar uma v.a. Pareto com $\alpha = 1/2$ no computador, a chance de observar um valor que 10 é de 25%.

Bem, talvez este efeito acabe logo. Por exemplo, será que podemos ter uma Pareto com $\alpha = 1$ ou com $\alpha = 1/2$ gerando números que dificilmente passem de 20. Assim, a Pareto geraria valores mas eles não se espalhariam para muito longe da região onde a maioria deles estaria concentrada. Na verdade, isto não acontece. Temos $\mathbb{P}(X > 20) = 0.03$ para $\alpha = 1$ e $\mathbb{P}(X > 10) = 0.16$ para $\alpha = 1/2$. Indo bem mais longe, especialmente no caso $\alpha = 1/2$, leva a resultados surpreendentes. Veja a tabela abaixo para alguns valores de $\mathbb{P}(X > k)$. Note como a probabilidade $\mathbb{P}(X > k)$ cai muito lentamente. Se simularmos uma Pareto, teremos a maioria dos seus valores sendo pequenos mas vão aparecer com bastante facilidade valores *ordens de grandeza* maiores que a maioria dos dados. Este é o fenômeno da cauda pesada. Nas distribuições binomial, Poisson ou geométrica, números muito maiores que a maioria são muito improváveis, quase impossíveis.

k	101	501	1001	5001	50000
$\alpha = 1$	0.0060	0.0013	0.0006	0.0002	0.00005
$\alpha = 1/2$	0.0759	0.0340	0.0241	0.0107	0.0033

6.16.1 Exemplos de distribuições de Pareto

A distribuição de Pareto é muito comum em estudos da web. Uns poucos sites possuem milhões de páginas mas centenas de milhões de sites possuem apenas umas poucas páginas. Poucos sites contêm milhões de links, enquanto a imensa maioria deles não possui mais que uma dezena de links. Centenas de milhões de usuários visitam uns poucos sites dando pouca atenção a bilhões de outros. Não somente na web. A renda do trabalho dos indivíduos e o patrimônio das famílias mostram um comportamento de Pareto. Poucos possuem fortunas, enquanto a maioria possui relativamente pouco. Distribuições de tamanhos, tais como o tamanho de empresas e o tamanho das cidades no mundo também mostram um padrão de Pareto.

Numa operadora de planos de saúde, a maioria dos clientes ocasiona um custo anual irrisório mas uns poucos indivíduos são responsáveis por uma quantidade desproporcional do custo total. Em [3], encontramos o resultado típico da concentração dos gastos de saúde. Ordenando os indivíduos pelo que custaram às seguradoras nos EUA e tomando o grupo de 1% dos que mais tiveram gastos, verifica-se que este pequeno grupo consumiu 27% do gasto total anual. O grupo dos top 5% consumiram 55%, enquanto os top 10% consumiram 69% do total. Outro exemplo é em relação à frequência das palavras numa língua qualquer. Algumas palavras são usadas com grande frequência mas maioria possui uma frequência bastante pequena.

Todos estes são casos de extremo desequilíbrio, em que a maioria dos valores são relativamente pequenos mas existe uma cauda pesada persistentemente gerando valores ordens de grandeza acima da maioria. São todos eles candidatos a serem modelados pela distribuição de Pareto, seja em sua versão discreta como vimos aqui, seja na versão contínua, no próximo capítulo.

Na seção 6.16.3 vamos discutir como verificar se os dados observados seguem uma distribuição é de Poisson.

6.16.2 Distribuição de Zipf

A distribuição de Zipf é um caso particular da distribuição de Pareto, quando $\alpha = 0$, e ela foi popularizada em estudos de linguagem pelo americano George Kingsley Zipf (1902–1950). Zipf achou uma regularidade estatística em textos escritos em diferentes linguagens. Considerando

palavra	posto (rank)	frequência
de	1	79607
a	2	48238
ser	27	4033
amor	802	174
chuva	2087	70
probabilidade	8901	12
iterativo	14343	6
algoritmo	21531	3

Table 6.1: Posto (ou rank) de algumas palavras e frequência de sua aparição por milhão de palavras em textos de português brasileiro

uma grande coleção de textos em um idioma qualquer, ele notou que algumas palavras aparecem pouco, são raramente usadas. Outras aparecem com grande frequência. Por exemplo, na tabela abaixo, extraída de www.linguatec.pt, temos algumas das palavras mais frequentemente usadas no português brasileiro, bem como algumas palavras menos usadas. A tabela mostra a posição (o posto ou *rank*) de algumas palavras na segunda coluna. Por exemplo, a preposição *de* é a mais usada no português e por isto ela tem o posto 1. O artigo *a* é a segunda palavra mais usada e portanto tem o posto 2. A palavra *ser* tem o posto 27, *amor* ocupa a posição 802, e assim por diante. A terceira coluna mostra a frequência da aparição dessas palavras por milhão de palavras. Assim, a palavra *de* aparece numa média de 79607 vezes em cada grupo de 1 milhão de palavras num texto. A palavra *chuva* ocupa a posição 2087 e só aparece com frequência de 70 vezes em cada milhão de palavras.

Imagine o seguinte experimento aleatório: primeiro, após processar uma grande quantidade de textos, ordene as palavras do idioma de acordo com o seu posto. A seguir, escolha uma palavra completamente ao acaso de um texto. Seja Y o posto (ou rank) dessa palavra escolhida ao acaso. Se $Y = 1$, significa que a palavra escolhida ao acaso é a palavra mais frequente. Se $Y = 17$, a palavra escolhida é a 17^a mais frequente. Zipf verificou que existia uma grande regularidade na chance de escolher a palavra de rank k . Ele descobriu que $\mathbb{P}(Y = k)$ é aproximadamente proporcional a $1/k$. Assim, de forma aproximada, temos

$$\begin{aligned}\mathbb{P}(Y = 1) &\propto 1 \\ \mathbb{P}(Y = 2) &\propto 1/2 \\ \mathbb{P}(Y = 3) &\propto 1/3, \text{ e etc.}\end{aligned}$$

Se a distribuição de Zipf for um bom modelo para os dados de um idioma devemos ter

$$\mathbb{P}(Y = k) \approx \frac{C}{k} = \frac{C}{k^{1+\alpha}}.$$

Esta é uma distribuição de Pareto com $\alpha = 0$. Na prática, costuma-se encontrar $\alpha \approx 0$.

6.16.3 Como verificar se a distribuição é Pareto?

Por causa da extrema assimetria presente na distribuição de Pareto, não é apropriado usar diretamente o gráfico da função de probabilidade como aqueles da Figura 6.14. Voltando aos dados da tabela 6.16.2, sabemos pela visão frequentista de probabilidade, sabemos que $n_k = 10^6 \mathbb{P}(Y = k)$ é aproximadamente igual à frequência (por milhão) da palavra de posto k . Isto é, a terceira coluna da tabela com a frequência empírica n_k deveria ser aproximadamente igual a 10^6 vezes a probabilidade

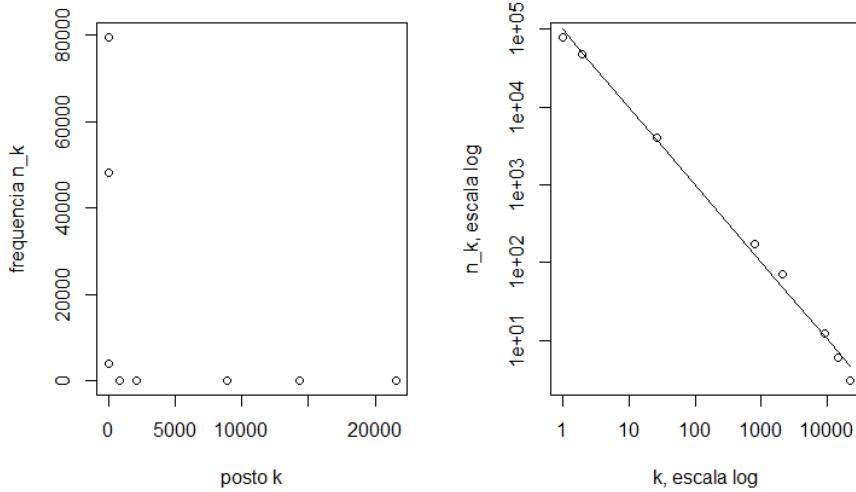


Figure 6.17: Gráfico do posto k versus a freqüência n_k para algumas palavras do português (esquerda). O gráfico da direita mostra os mesmos dados num gráfico log-log (isto é, os pontos são $(\log(k), \log(n_k))$). A reta foi obtida pela técnica de mínimos quadrados (ver capítulo ??) e é igual a $\log(n_k) = 11.51 - 0.999 \log(k)$.

$\mathbb{P}(Y = k)$. Se o modelo Zipf-Pareto for adequado, teremos

$$n_k \approx 10^6 \mathbb{P}(Y = k) = 10^6 \frac{C}{k^{1+\alpha}}.$$

Ao tomarmos log dos dois lados teremos:

$$\log(n_k) \approx \log(10^6) + \log(\mathbb{P}(Y = k)) = \underbrace{(\log(10^6) + \log(C))}_a - \underbrace{(1 + \alpha) \log(k)}_b = a - b \log(k).$$

Assim, se Zipf-Pareto é adequado, um plot de $\log n_k$ versus $\log(k)$ deveria ser aproximadamente uma linhareta com intercepto $a = 6\log(10) + \log(C)$ e inclinação $-b = -(1 + \alpha)$.

Vamos checar se isto ocorre olhando os dados das palavras do português brasileiro que estão da tabela 6.16.2. O resultado está na Figura 6.17.

Como identificar se um modelo Pareto é um bom ajuste para dados sobre os valores discretos $\{1, 2, \dots\}$? Como vimos, uma maneira é contar o número de vezes n_k que o valor k aparece numa amostra. A seguir, fazemos um gráfico de $\log(k)$ (no eixo x) versus o log da freqüência $\log(n_k)$ (no eixo y). Se o modelo Pareto for adequado, devemos observar aproximadamente uma linhareta neste plot.

Embora esta seja uma técnica simples e útil, temos uma maneira mais efetiva de checar se a distribuição de Pareto é um bom modelo para os dados. Usamos a distribuição acumulada $F(x) = \mathbb{P}(Y \leq x)$ para isto. Veremos esta técnica mais efetiva quando estudarmos a Pareto no caso contínuo no próximo capítulo.

6.17 Comparação entre as distribuições

Poisson \times geométrica \times Pareto: qual a diferença mais relevante entre elas? Todas são distribuições sobre os inteiros positivos. A principal diferença está no comportamento *na cauda*:

- Poisson tem cauda curta, valores com probabilidades significativas estão concentrados em uma faixa estreita torno de sua esperança $\mathbb{E}(Y) = \lambda$.
- Pareto gera facilmente valores muito grandes, ordens de grandeza maiores que $\mathbb{E}(Y)$.
- Geométrica é um caso intermediário.

Comparando as três:

- Poisson:

$$\frac{\mathbb{P}(Y = k+1)}{\mathbb{P}(Y = k)} = \frac{e^{-\lambda} \lambda^{k+1}/(k+1)!}{e^{-\lambda} \lambda^k/k!} = \frac{\lambda}{k+1} \rightarrow 0$$

se $k \rightarrow \infty$. Isto é $\mathbb{P}(Y = k+1) << \mathbb{P}(Y = k)$ se k é grande.

- Geométrica:

$$\frac{\mathbb{P}(Y = k+1)}{\mathbb{P}(Y = k)} = \frac{(1-\theta)^{k+1}\theta}{(1-\theta)^k\theta} = 1 - \theta < 1,$$

constante em k . Isto é, $\mathbb{P}(Y = k+1) = (1-\theta)\mathbb{P}(Y = k)$, uma queda geométrica ou exponencial.

- Pareto:

$$\frac{\mathbb{P}(Y = k+1)}{\mathbb{P}(Y = k)} = \left(\frac{k}{k+1}\right)^\theta \rightarrow 1$$

Isto é $\mathbb{P}(Y = k+1) \approx \mathbb{P}(Y = k)$ se k é grande, uma queda muito lenta.



7. Variáveis Aleatórias Contínuas

7.1 Introdução

Temos dois tipos principais de variáveis aleatórias: discretas, vistas no capítulo anterior, e contínuas, que veremos neste capítulo. A especificação informal de uma v.a. contínua continua sendo como no caso de uma v.a. discreta, a combinação de duas listas. Informalmente, temos

Definition 7.1.1 — V.A. contínua. Uma variável aleatória é chamada *contínua* quando ela for especificada com:

- um ou mais intervalos da reta real que compõe o conjunto dos valores possíveis.
- uma função densidade de probabilidade $f(x)$ definida neste intervalo.

A única restrição é que a densidade $f(x)$ deve ser maior ou igual a zero para todo x e sua integral sobre o intervalo de valores possíveis deve ser igual a 1.

Considere o gráfico da função $f(x)$ na figura 7.1. Ela tem valores maiores que zero apenas entre -5 e 15. Além disso, a integral de $f(x)$ no intervalo $(-5, 15)$ (isto é, a área total sob a curva, entre $(-5, 15)$) vale 1. Dessa forma, a função $f(x)$ pode representar a densidade de probabilidade de uma v.a. contínua X que pode assumir qualquer valor no intervalo real $(-5, 15)$.

No caso contínuo, probabilidades estão associadas com áreas sob a função densidade:

$$\mathbb{P}(X \in (a, b)) = \int_a^b f(x)dx$$

A Figura 7.2 mostra como as probabilidades são calculadas no caso contínuo. As regras usuais de probabilidades são válidas, claro. Por exemplo, o evento X pertencer a dois intervalos disjuntos, tal como $[X \in (1, 2) \text{ ou } (4, 5)]$, é a união de dois eventos *disjuntos*, $[X \in (1, 2)] \cup [X \in (4, 5)]$. Eles são disjuntos pois $X(\omega)$ não pode estar ao mesmo tempo em $(1, 2)$ e em $(4, 5)$. Assim,

$$\mathbb{P}(X \in (1, 2) \text{ ou } (4, 5)) = \mathbb{P}(X \in (1, 2)) + \mathbb{P}(X \in (4, 5)).$$

Olhando o gráfico de $f(x)$ sabemos quais as regiões da reta real com probabilidade mais alta: são aquelas regiões que possuem maior área abaixo da curva $f(x)$. A ideia intuitiva é que a função $f(x)$ mostra como a probabilidade total (igual a 1) foi distribuída no eixo real indicando quais

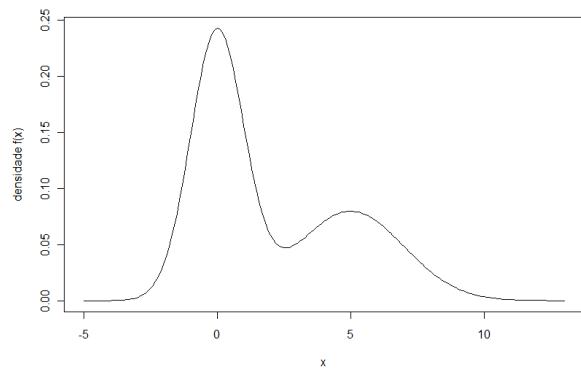


Figure 7.1: Exemplo de um função densidade de probabilidade $f(x)$. Ela é maior ou igual a zero e sua área abaixo da curva é igual a 1.

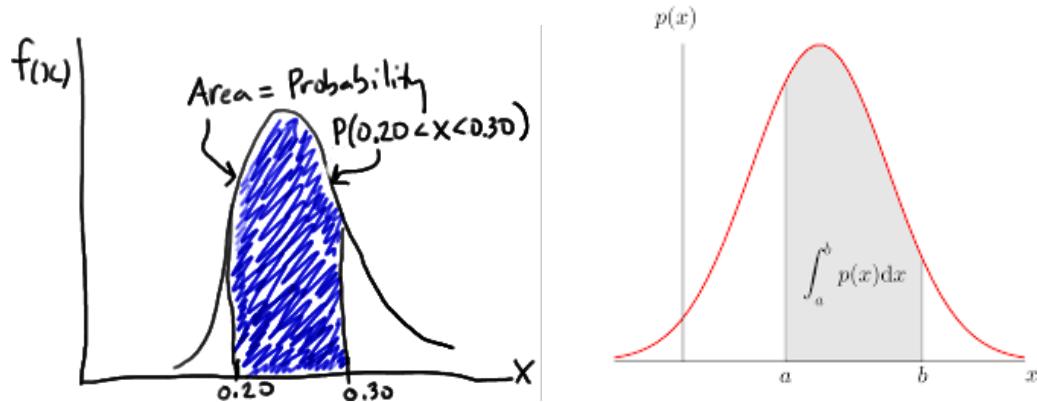


Figure 7.2: Probabilidades são áreas sob a função densidade $f(x)$.

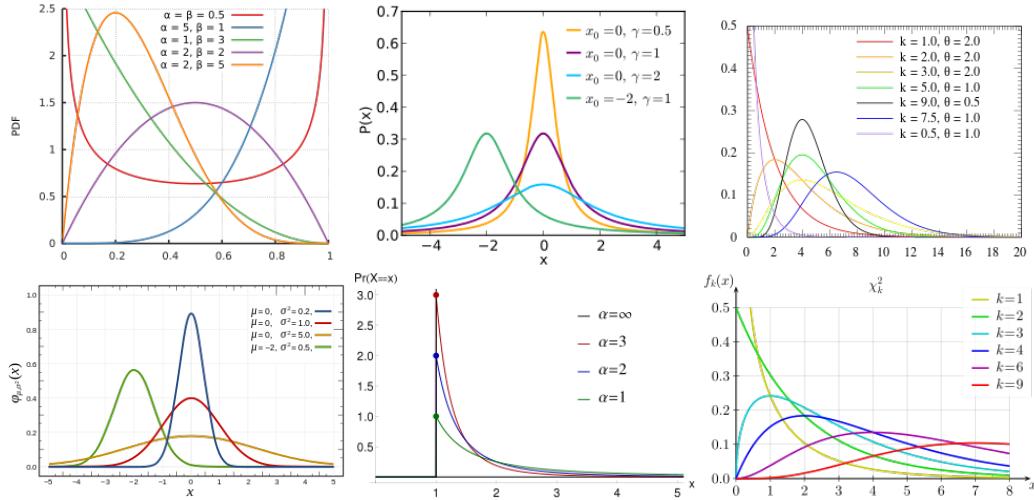


Figure 7.3: Exemplos de algumas das principais densidades de probabilidade de v.a.'s contínuas: beta, Cauchy, gamma (linha superior) e normal, Pareto e qui-quadrado (linha inferior). Cada curva é uma densidade obtida variando os parâmetros da distribuição.

regiões tem mais chance de produzir um resultado para X (as regiões com $f(x)$ mais altos) e quais as regiões com pequena probabilidade de gerar um valor de X (regiões onde $f(x) \approx 0$).

Voltando à densidade da Figura 7.1, considere os quatro intervalos $(-5, -2.5)$, $(-2.5, 0)$, $(5, 7.5)$ e $(7.5, 10)$, todos de igual comprimento. Qual tem a maior probabilidade? Isto é, comparando as probabilidade de que X venha de cada um desses intervalos, qual delas é a maior de todas?

- $\mathbb{P}(X \in (-5, -2.5))$
- $\mathbb{P}(X \in (-1.0, 1.5))$
- $\mathbb{P}(X \in (5, 7.5))$
- $\mathbb{P}(X \in (7.5, 10))$

Devemos olhar para a área debaixo de $f(x)$ em cada um desses intervalos. Neste caso, claramente $\mathbb{P}(X \in (-1.0, 1.5))$ é a maior probabilidade dentre estas quatro enquanto $\mathbb{P}(X \in (-5, -2.5))$ é a menor de todas.

A Figura 7.3 mostra exemplos de algumas das principais densidades de probabilidade de v.a.'s contínuas. Indo da esquerda para a direita, vemos exemplos de densidades das distribuições beta, Cauchy, gamma (linha superior) e normal, Pareto e qui-quadrado (linha inferior). Cada curva é uma densidade obtida variando os parâmetros da distribuição.

Mesmo quando a variável não for contínua, uma aproximação com uma distribuição contínua pode ser útil. Imagine uma amostra de 5000 lotes que constituem uma fazenda e onde se cultiva somente soja. Seja y_i a colheita do lote i . É muito pouco prático e um tanto sem sentido trabalharmos com uma distribuição discreta para uma situação como essa. Mesmo que estejamos interessados apenas nos 5000 lotes (um número finito de valores), é mais útil assumirmos que as colheitas dos lotes são os resultados de 5000 realizações de uma certa variável aleatória *contínua* que possui uma forma simples e já conhecida.

Qual seria a densidade desta v.a. Y ? Para saber isto, fazemos um histograma padronizado (com área total = 1) como na Figura 7.4. Quebre o eixo horizontal em pequenos intervalos de comprimento Δ . Em cada pequeno intervalo i , conte o número n_i de elementos em sua amostra que caíram no intervalo. Levante uma barra cuja altura seja igual a esta contagem. Este é o histograma não padronizado e que está no gráfico à esquerda na Figura 7.4. O histograma padronizado tem área total = 1. Para isto, basta dividir cada barra pelo valor constante $n\Delta$ produzindo barras com

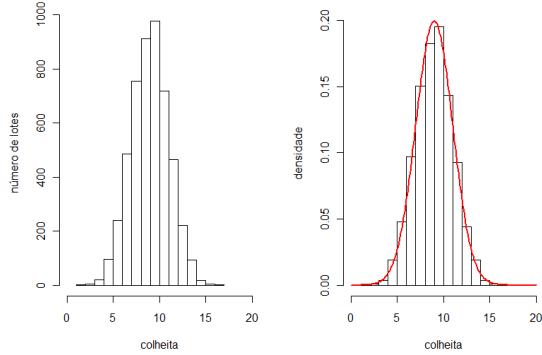


Figure 7.4: Histograma simples e padronizado.

altura $= n_i / (n\Delta)$. O resultado está no gráfico à direita na Figura 7.4. No histograma padronizado, sobreponha uma densidade candidata. O histograma se parece com uma certa densidade de uma distribuição chamada gaussiana (ou normal, $N(9, 4)$) cuja densidade é a curva em vermelho. Isto significa que a distribuição real será *aproximada* por esta distribuição normal. Veremos mais tarde como escolher uma distribuição candidata e testar se ela se ajusta bem aos dados.

De fato, esta é uma maneira prática, rápida e eficiente de achar uma densidade candidata. Após conhecer as principais distribuições, teremos uma certa coleção de possibilidades para propor uma densidade candidata ao olharmos para o histograma e enxergarmos de forma aproximada a forma das densidades que conhecemos. É claro que o histograma pode ter uma forma muito estranha, que não se pareça com as formas das distribuições básicas. Nestes casos podemos apelar para misturas dessas distribuições básicas (ver capítulo ??).

Mas vamos primeiro ver como enxergar as distribuições básicas nos histogramas. A Figura 7.5 mostra o histograma padronizado de amostras de tamanho 1000 compostas por dados simulados no computador. Como eles foram simulados, temos certeza sobre qual foi a distribuição geradora em cada caso. Começando da linha superior e indo da esquerda para a direita, estas distribuições foram a exponencial, log-normal, uniforme e beta, respectivamente. As linhas vermelhas são as respectivas densidades de probabilidade. Podemos ver que, de fato, os histogramas possuem o mesmo formato que as densidades. Ou seja, olhar para o histograma deveria sugerir a forma da densidade de probabilidade.

Na prática, não temos as linhas vermelhas da Figura 7.5 pois não conhecemos a densidade de probabilidade que gerou os dados que estamos analisando. Saber que olhar para o histograma dá uma boa indicação de qual é esta densidade desconhecida é uma maneira de aprender diretamente dos dados o mecanismo oculto que está produzindo os dados observados.

Qual a justificativa para esta proximidade entre o histograma e a densidade de probabilidade desconhecida? Seja $f^*(x)$ a densidade verdadeira que gerou os dados. Em geral, $f^*(x)$ é desconhecida. Seja $f(x)$ a densidade de uma distribuição retirada do nosso limitado catálogo de distribuições conhecidas, que possuem nomes próprios e que são bem estudadas. Se o histograma da amostra é bem aproximado por $f(x)$ então acreditamos que $f(x) \approx f^*(x)$. Por quê?

Para responder, seja $(y_0 - \delta/2, y_0 + \delta/2)$ um pequeno intervalo do histograma centrado em y_0 e de pequeno comprimento δ . Podemos aproximar a área debaixo da curva pelo retângulo com base δ centrada em y_0 e com altura $f^*(y_0)$:

$$P(Y \in (y_0 - \delta/2, y_0 + \delta/2)) = \int_{y_0 - \delta/2}^{y_0 + \delta/2} f^*(y) dy \approx f^*(y_0)\delta.$$

Vamos usar agora a ideia de estimar uma probabilidade pela proporção de vezes que o evento

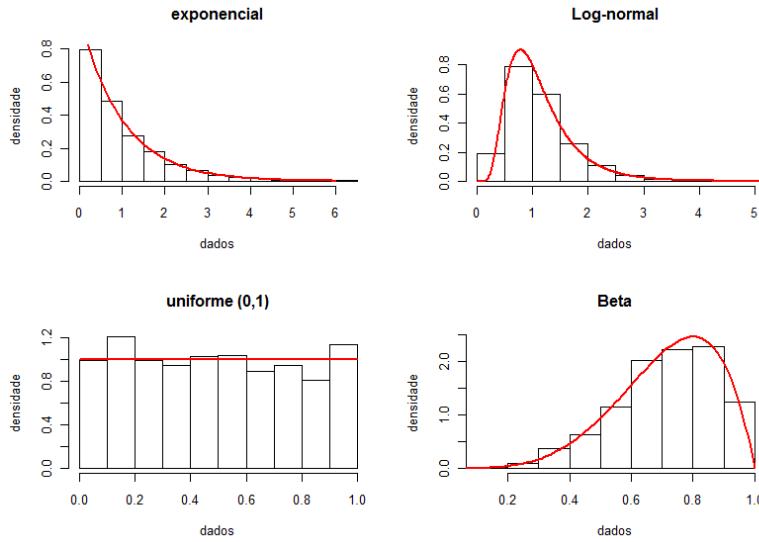


Figure 7.5: Histogramas padronizados de amostras obidas de distribuições cujas densidades são mostradas com a linhas vermelha.

acontece num grande número de repetições. A mesma probabilidade $P(Y \in (y_0 - \delta/2, y_0 + \delta/2))$ também pode ser aproximada pela fração de elementos da amostra que caíram no intervalo $(y_0 - \delta/2, y_0 + \delta/2)$:

$$P(Y \in (y_0 - \delta/2, y_0 + \delta/2)) \approx \frac{\#\{Y_i \in (y_0 - \delta/2, y_0 + \delta/2)\}}{n}$$

Igualando as duas aproximações para a probabilidade $P(Y \in (y_0 - \delta/2, y_0 + \delta/2))$ e dividindo por δ dos dois lados, temos

$$\frac{\#\{Y_i \in (y_0 - \delta/2, y_0 + \delta/2)\}}{n\delta} \approx f^*(y_0)$$

O lado esquerdo é a altura do histograma no ponto y_0 . O lado direito é a altura da curva densidade no mesmo ponto y_0 . Assim, a altura do histograma no ponto central y_0 de um dos intervalos é aproximadamente igual à densidade *desconhecida* $f^*(y_0)$. Isto é, olhar o histograma é essencialmente olhar a densidade desconhecida.

7.2 $\mathbb{F}(X)$ no caso contínuo

Já definimos a função acumulada de probabilidade $\mathbb{F}(x) = \mathbb{P}(X \leq x)$. No caso contínuo, a função $\mathbb{F}(x)$ associa a cada valor x da reta real toda a área de sob a curva densidade no intervalo $(-\infty, x)$. Gaste um tempo estudando a Figura 7.6. Do lado esquerdo ela mostra as duas curvas, $f(x)$ e $\mathbb{F}(x)$, em dois plots separados. No de cima, temos a densidade $f(x)$. No de baixo, a distribuição acumulada $\mathbb{F}(x)$. Para qualquer ponto arbitrário x da reta, obtemos toda a área sob a densidade $f(x)$ até o ponto x . Esta é a área sombreada no plot superior e ela é igual a $\mathbb{P}(X \leq x) = \mathbb{F}(x)$. Como a área total é 1, temos um valor entre 0 e 1. Este valor $\mathbb{F}(x)$ é colocado no gráfico de baixo como a *altura* da curva \mathbb{F} no ponto x . Note como a altura da curva $\mathbb{F}(x)$ varia a medida que deslizamos o valor de x a partir da extrema esquerda. $\mathbb{F}(x)$ começa valendo quase zero, refletindo o fato de não haver quase nenhuma área à esquerda de um x na extrema esquerda. A medida que avançamos para a direita com x , acumulamos área no gráfico de cima e a altura de $\mathbb{F}(x)$ sobre até chegar ao valor de

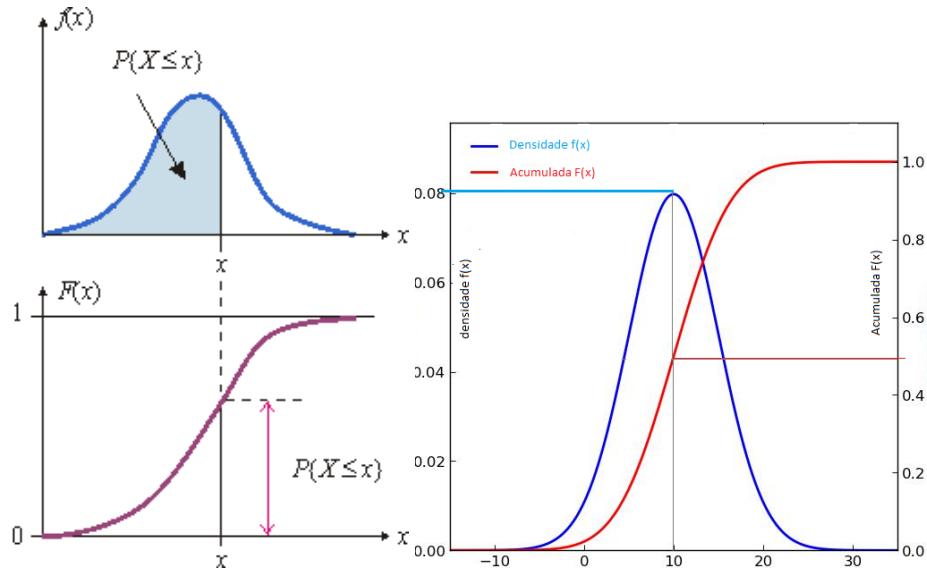


Figure 7.6: Densidade de probabilidade $f(x)$ e função acumulada $\mathbb{F}(x)$.

1 (ou assintotar em 1), que é a área total. O gráfico à direita na Figura 7.6 mostra as duas curvas, $f(x)$ e $\mathbb{F}(x)$, num mesmo plot. Os valores de $f(x)$ são lidos no eixo vertical à esquerda, enquanto os valores de $\mathbb{F}(x)$ são lidos no eixo vertical à direita.

Como áreas sob a curva são integrais, temos como passar da densidade $f(x)$ para a função acumulada $\mathbb{F}(x)$:

$$\mathbb{F}(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x f(t) dt.$$

Como consequência, podemos também reverter este resultado, passando da função acumulada $\mathbb{F}(x)$ para a densidade $f(x)$, um resultado cuja prova é imediata a partir do Teorema Fundamental do Cálculo:

Theorem 7.2.1 Se X é v.a. contínua e $\mathbb{F}(x)$ for diferenciável em x então $f(x) = \frac{d\mathbb{F}(x)}{dx}$.

■ **Example 7.1 — Densidade e Acumulada.** Suponha que X seja uma v.a. contínua com suporte $\mathcal{S} = (0, \infty)$ e densidade $f(x) = 3e^{-3x}$, para $x > 0$. Para $x \leq 0$, e portanto fora do suporte, temos $f(x) = 0$. Ver o plot superior na Figura 7.7.

Começando com o caso mais simples, aquele em que $x \leq 0$, pela definição temos

$$\mathbb{F}(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x f(t) dt = \int_{-\infty}^x 0 dt = 0.$$

Portanto, para um $x > 0$ arbitrário no suporte, temos

$$\mathbb{F}(x) = \mathbb{P}(X \leq x) = \int_{\infty}^x f(t) dt = 0 + \int_{\infty}^x 3e^{-3t} dt = 1 - e^{-3x}.$$

Para $x > 0$, se tomarmos a derivada de $\mathbb{F}(x) = 1 - e^{-3x}$ encontramos $\mathbb{F}'(x) = 3e^{-3x}$, que é exatamente $f(x)$. Para $x < 0$, a derivada de $\mathbb{F}(x) \equiv 0$ é $\mathbb{F}'(x) = 0$, que é novamente o valor da densidade para $x < 0$. ■

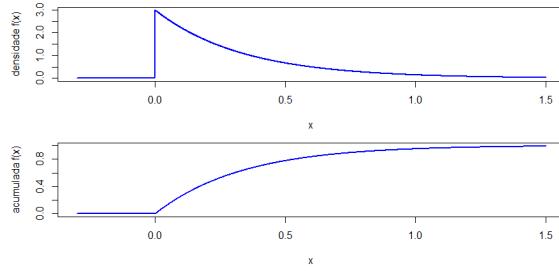


Figure 7.7: Densidade $f(x)$ e distribuição acumulada $F(x)$ correspondente.

7.3 $\mathbb{E}(X)$ no caso contínuo

Já estudamos a definição e o significado empírico do valor esperado de uma v.a. X no caso discreto. Suponha que X é uma v.a. discreta com valores possíveis são $\{x_1, x_2, \dots\}$. Observe que o número de valores possíveis pode ser infinito. Então

$$\mathbb{E}(X) = \sum_{x_i} x_i \mathbb{P}(X = x_i).$$

Suponha que temos uma grande amostra aleatória $\{X_1, X_2, \dots, X_n\}$ de tamanho n dessa v.a. A conexão semântica entre a definição e a amostra com n valores neste caso discreto é que valor teórico $\mathbb{E}(X)$ deve ser aproximadamente igual à média aritmética $\bar{X} = (X_1 + \dots + X_n)/n$ dos n elementos da amostra.

Definition 7.3.1 — $\mathbb{E}(X)$ no caso contínuo. Seja X uma v.a. contínua com conjunto suporte \mathcal{S} na reta real e densidade de probabilidade $f(x)$. O valor esperado de X é igual a $\mathbb{E}(X) = \int_{\mathcal{S}} x f(x) dx$.

Ocasionalmente, para algumas distribuições esta integral pode não existir ou não estar definida. Nestes casos, a v.a. não possui valor esperado. Para as distribuições mais comuns na prática da análise de dados, este valor esperado existe sem problema algum.

O caso contínuo é a versão discreta levada ao limite. Podemos raciocinar intuitivamente exatamente como no caso discreto. Particione todo o eixo da reta real em pequenos intervalos (ou bins) de comprimento Δ e centrados em $\dots, x_{-2}, x_{-1}, x_0, x_1, x_2, \dots$. Em cada pequeno bin $(x_i - \Delta/2, x_i + \Delta/2)$, sabemos que a integral de qualquer função $h(x)$ pode ser aproximada pelo retângulo de base Δ e altura $h(x_i)$:

$$\int_{\text{bin}_i} h(x) dx = \int_{x_i - \Delta/2}^{x_i + \Delta/2} h(x) dx \approx h(x_i) \Delta.$$

Na verdade, a integral de Riemann é *definida* tomando a soma sobre todos os bins e levando ao limite com o comprimento de cada bin indo a zero e o número de bins indo a infinito:

$$\int_{\mathbb{R}} h(x) dx = \lim_{\Delta \rightarrow 0} \sum_i h(x_i) \Delta.$$

A Figura 7.8 mostra uma curva vermelha representando a função $h(x)$. A área sob a curva é a integral $\int h(x) dx$ e ela é aproximada pela soma dos retângulos. Cada retângulo possui uma base de comprimento Δ e altura $h(x_i^*)$, igual ao valor da função $h(x)$ no ponto central x_i^* da base do retângulo.

No caso da esperança de uma v.a. contínua com densidade $f(x)$, se fizermos a função genérica $h(x)$ ser igual a $xf(x)$ (isto é, $h(x) = xf(x)$) então a aproximação acima fica

$$\int_{\text{bin}_i} x f(x) dx = \int_{x_i - \Delta/2}^{x_i + \Delta/2} x f(x) dx \approx x_i f(x_i) \Delta$$

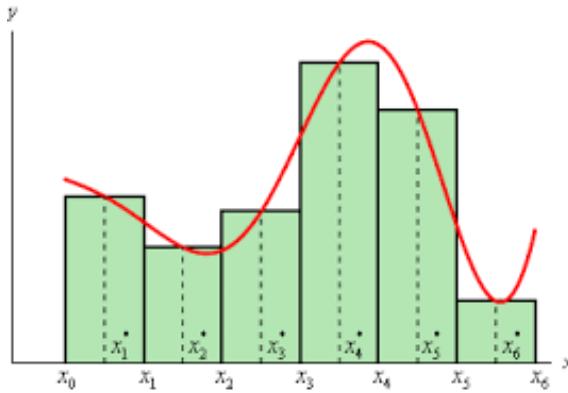


Figure 7.8: Integral $\int h(x)dx$ onde $h(x)$ é a curva vermelha e ela é aproximada pela soma dos retângulos de base Δ e altura igual ao valor de $h(x)$ no ponto central x_i^* da base do retângulo.

Mas temos também uma aproximação de probabilidades como áreas sob a curva densidade de modo que

$$\mathbb{P}(X \in (x_i - \Delta/2, x_i + \Delta/2)) = \int_{x_i - \Delta/2}^{x_i + \Delta/2} f(x)dx \approx f(x_i)\Delta$$

Assim, podemos obter a seguinte aproximação para $\mathbb{E}(X)$ no caso contínuo:

$$\begin{aligned}\mathbb{E}(X) &= \int_{-\infty}^{\infty} yf(y)dy \\ &= \dots + \int_{x_{-1}-\Delta/2}^{x_{-1}+\Delta/2} xf(x)dx + \int_{x_0-\Delta/2}^{x_0+\Delta/2} xf(x)dx + \int_{x_1-\Delta/2}^{x_1+\Delta/2} xf(x)dx + \int_{x_2-\Delta/2}^{x_2+\Delta/2} xf(x)dx + \dots \\ &\approx \dots + x_{-1}f(x_{-1})\Delta + x_0f(x_0)\Delta + x_1f(x_1)\Delta + x_2f(x_2)\Delta + \dots \\ &\approx \dots + x_{-1}\mathbb{P}(X \in \text{bin}_{-1}) + x_0\mathbb{P}(X \in \text{bin}_0) + x_1\mathbb{P}(X \in \text{bin}_1) + x_2\mathbb{P}(X \in \text{bin}_2) + \dots\end{aligned}$$

Esta última expressão é igual à esperança de uma v.a. discreta que assume os possíveis valores $\dots, x_{-2}, x_{-1}, x_0, x_1, x_2, \dots$ com probabilidades $\mathbb{P}(X \in \text{bin}_i)$. Isto é, a definição da esperança no caso contínuo como sendo $\int xf(x)dx$ é apenas a expressão do caso discreto levada ao limite contínuo.

7.4 Distribuição Uniforme

A partir desta seção, vamos construir nosso pequeno catálogo de distribuições contínuas. Vamos apresentar algumas das mais importantes distribuições contínuas com algumas ilustrações muito simples sobre seu uso na prática da análise de dados. Começamos com a mais simples de todas, a distribuição Uniforme.

Definition 7.4.1 — Distribuição Uniforme. A distribuição uniforme sobre um intervalo (a, b) , com $a < b$, é definida pela densidade

$$f(x) = \frac{1}{(b-a)}$$

se $x \in (a, b)$. Ver Figura 7.9. No caso de querermos estender a definição de $f(x)$ para todo $x \in \mathbb{R}$

podemos definir a densidade de forma que ela seja igual a zero fora do intervalo (a, b) :

$$f(x) = \begin{cases} \frac{1}{(b-a)}, & \text{if } x \in [a, b] \\ 0, & \text{caso contrário.} \end{cases}$$

■ **Notation 7.1.** Se X possui distribuição uniforme no intervalo (a, b) escrevemos $X \sim U(a, b)$.

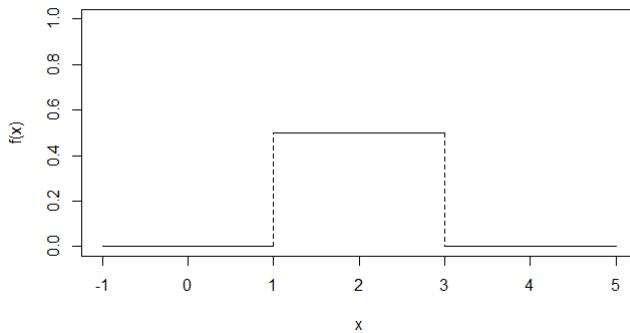


Figure 7.9: Densidade da distribuição Uniforme(1,2)

Variando o comprimento do intervalo, $f(x)$ muda. Por exemplo, se $X \sim U(3, 10)$ então $f(x) = 1/7$ para $x \in (3, 10)$.

A distribuição uniforme mais famosa é $X \sim U(0, 1)$, definida no intervalo $(0, 1)$. Neste caso, $f(x) = 1$ para $x \in (0, 1)$. A probabilidade de que $X \sim U(0, 1)$ caia num intervalo (a, b) contido em $(0, 1)$ é a área sob a densidade uniforme:

$$\mathbb{P}(X \in (a, b)) = \int_a^b f(x) dx = \int_a^b 1 dx = b - a$$

Assim, no caso da $U(0, 1)$ a probabilidade de um intervalo (a, b) contido em $(0, 1)$ é o seu comprimento. Por exemplo, $\mathbb{P}(X \in (1/2, 1)) = 1/2$ e $\mathbb{P}(X \in (0.75, 0.78)) = 0.03$.

A distribuição uniforme $U(0, 1)$ é crucial nos métodos de simulação Monte Carlo. Programas de computador chamados geradores de números aleatórios fornecem uma sucessão de valores que são semelhantes em seu comportamento estatístico a observações sucessivas e independentes obtidas de uma distribuição $U(0, 1)$. Por exemplo, se gerarmos uma grande número de valores $U(0, 1)$ usando a função `runif` em R obtemos o quadro abaixo. Ele apresenta a frequência f_i dentre 100 mil números aleatórios gerados pelo R que pertencem ao intervalo $I_i = (0.i, 0.(i+1)]$ $i = 0, \dots, 9$. Veja como a proporção de números em cada intervalo é aproximadamente 0.1, como deveria ser se os dados seguirem uma $U(0, 1)$. O código usado está abaixo.

```
> x = runif(100000)
> intervals = cut(x, breaks = seq(0, 1, by=0.1))
> table(intervals)
intervals
(0,0.1] (0.1,0.2] (0.2,0.3] (0.3,0.4] (0.4,0.5] (0.5,0.6]
9881     10153     9982     10134     10104     9872
(0.6,0.7] (0.7,0.8] (0.8,0.9] (0.9,1]
9983     9951     9967     9973
```

A esperança de uma v.a. $X \sim U(a, b)$ com distribuição uniforme em (a, b) é facilmente obtida:

$$\mathbb{E}(X) = \int_a^b x \frac{1}{b-a} dx = \frac{1}{b-a} \int_a^b x dx = \frac{1}{b-a} \left(\frac{b^2}{2} - \frac{a^2}{2} \right) = \frac{(b-a)(b+a)}{2(b-a)} = \frac{a+b}{2}$$

Isto é, a esperança de uma v.a. $X \sim U(a, b)$ é o ponto central do intervalo (a, b) .

■ **Example 7.2** Se $X \sim U(0, 1)$ então $\mathbb{E}(X) = 1/2$. Se $X \sim U(90, 100)$ então $\mathbb{E}(X) = 95$. ■

A função de distribuição acumulada de probabilidade é também muito fácil de ser obtida:

$$F(x) = \mathbb{P}(X \leq x) = \begin{cases} 0, & \text{se } x < a \\ x/(b-a), & \text{se } a \leq x \leq b \\ 1, & \text{se } x > b \end{cases}$$

■ **Example 7.3** A distribuição uniforme pode servir para aproximar uma distribuição contínua num pequeno intervalo. Por exemplo, atuários costumam aproximar a distribuição idade ao morrer de um indivíduo usando este truque. Suponha que é sabido de alguma forma que um indivíduo completou seu aniversário de k anos mas faleceu antes de atingir a idade $k+1$. Se X é a idade exata (contínua) ao morrer desse indivíduo, podemos aproximar $X \sim U(x, x+1)$. Isto é, ele pode morrer em qualquer momento ao longo do ano com igual chance. Nenhum dia ou mês ao longo do ano de morte teria mais probabilidade de acontecer. Esta é uma aproximação apenas mas costumam possibilitar vários cálculos que, de outro modo, seriam impossíveis. Ver [4]. ■

■ **Example 7.4 — Medindo papel.** Este exemplo é do livro clássico de [24]. Quinhentos espécimes de um novo tipo de papel sob teste foram recobertos com um polímero líquido específico e em seguida foram enxutos e distribuídos de acordo com seus pesos em cinco categorias. A espessura dos papéis da categoria central foi medida para determinar se eles estavam uniformemente distribuídos. A espessura da folha era calculada tomando-se uma média aritmética de cinco pontos, um no centro e nas 4 quinas. Os dados da tabela abaixo representam a espessura dos 116 espécimes nesta classe central. A mensuração está em unidades codificadas, representando desvios para baixo ou para cima em relação a uma norma.

x	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
f_i	13	14	8	14	7	5	18	6	10	9	10	12

A Figura 7.10 é um histograma dos 116 valores da tabela acima junto com o gráfico da função constante $f(x) = 1/12$ que seria a densidade de probabilidade de uma distribuição uniforme. O valor esperado da altura de uma barra do histograma seria a linha em vermelho caso a distribuição de X fosse uniforme. Para testar se a hipótese de que a distribuição seja uniforme precisamos de técnicas como o teste qui-quadrado, a ser ensinado no Capítulo ???. Adiantando este tópico, podemos dizer que, apesar da aparente discrepância entre os dados observados e a distribuição uniforme, os dados são perfeitamente compatíveis com uma uniforme. Isto é, não existe evidência nos dados de que a distribuição uniforme não tenha gerados estes dados. As diferenças entre as barras e a linha teórica são perfeitamente naturais e plausíveis para uma amostra de tamanho 116. O código para o barplot está abaixo.

```
counts = c(13, 14, 8, 14, 7, 5, 8, 6, 10, 9, 10, 12)
freqrel = counts/sum(counts)
barplot(freqrel, names.arg=as.character(-5:6))
abline(h=1/12, lwd=2, col="red")
1-pchisq(sum((counts - 116/12)^2 / 116/12), 11) # p-valor do teste qui-quadrado
```

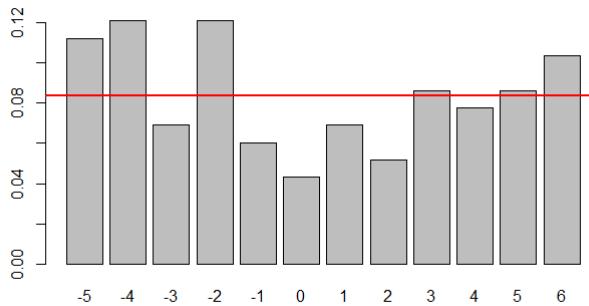


Figure 7.10: Barplot das proporções de papéis que caíram em diferentes intervalos de mesmo comprimento de acordo com sua espessura média.

7.5 Distribuição Beta

Existem vários fenômenos cujas variáveis de interesse tem seus valores limitados acima e abaixo por números conhecidos a e b . Um exemplo típico é constituído por dados que aparecem sob a forma de proporção:

- em cada empresa industrial brasileira é registrada a proporção de gastos em salários na produção total durante um ano. Podemos também registrar, por exemplo, a proporção de gastos em energia na produção total no mesmo ano.
- a razão entre o comprimento do fêmur e o comprimento total da perna de um indivíduo.

Veja que estas proporções assumem valores no intervalo contínuo $(0, 1)$. Elas diferem de outras proporções baseadas em contagens tais como a proporção de sucessos em 10 lançamentos de uma moeda. Neste caso, a proporção assume um dos valores $\{0, 0.1, 0.2, \dots, 0.9, 1.0\}$ e assim é uma variável discreta. Compare este caso discreto com a situação contínua nos exemplos acima.

Uma classe de distribuições que é rica o suficiente para fornecer modelos para a maioria das variáveis aleatórias que têm valores limitados é a classe de distribuições beta. Ela possui seus valores concentrados no intervalo $(0, 1)$.

Definition 7.5.1 — Distribuição Beta. A variável aleatória X tem distribuição beta com parâmetros α e β onde $\alpha > 0$ e $\beta > 0$ se Y tem densidade dada por

$$f(x) = Cx^{\alpha-1}(1-x)^{\beta-1}$$

para $x \in (0, 1)$. A constante C é tal que a densidade integra 1 no intervalo $(0, 1)$.

■ Notation 7.2. Se X possui distribuição beta em $(0, 1)$ com parâmetros $\alpha > 0$ e $\beta > 0$ escrevemos $X \sim \text{Beta}(\alpha, \beta)$.

Por exemplo, com $\alpha = 3$ e $\beta = 5$ temos $f(x) = Cx^2(1-x)^4$ (ver gráfico à esquerda na Figura 7.11). Já com $\alpha = 8$ e $\beta = 2$, temos $f(x) = C^*x^7(1-x)$ (gráfico à direita na Figura 7.11). Escrevemos a constante como C^* neste último exemplo apenas para enfatizar que a constante do primeiro exemplo é diferente da constante deste segundo exemplo. Pela densidade da Beta(3, 5), a região que gera um valor com maior probabilidade é bastante ampla, indo de 0.1 a 0.7. Já a densidade da Beta(8, 2) está bem concentrada numa faixa bem mais estreita e deslocada para o extremo superior do intervalo, entre 0.6 e 1.0.

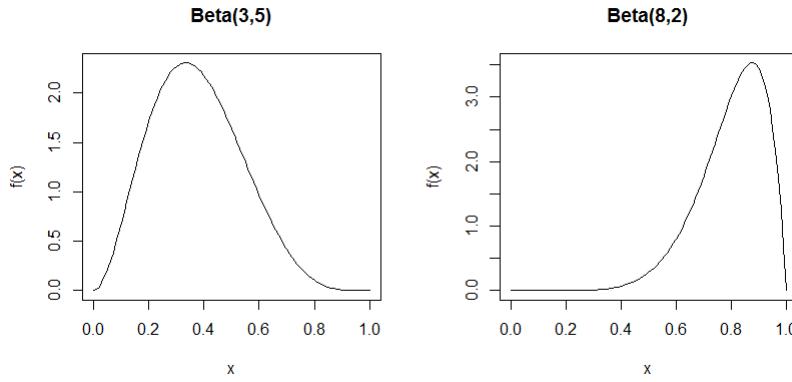


Figure 7.11: Densidade da distribuição Beta(3,5) e Beta(8,2).

Analisando a expressão da densidade da distribuição beta, vemos que $f(x) = Cx^{\alpha-1}(1-x)^{\beta-1}$ é o resultado do produto de dois monômios, o primeiro sendo $x^{\alpha-1}$ e o segundo sendo $(1-x)^{\beta-1}$. Por exemplo, com $\alpha = 3$ e $\beta = 5$ temos $f(x) = Cx^2(1-x)^4$ sendo obtido com o produto de x^2 por $(1-x)^4$. Note que o primeiro fator, x^2 , cresce com x , enquanto que o segundo fator, $(1-x)^4$, decresce com x . Como os dois fatores atingem zero num dos extremos, o seu produto é zero em $x = 0$ e em $x = 1$. O produto dos fatores vai subir e depois descer ao longo do intervalo $(0, 1)$. O seu ponto de máximo e a forma dessa subida e descida são controlados pelos parâmetros α e β . Este raciocínio é válido desde que $\alpha > 1$ e $\beta > 1$. Os outros casos, com valores de $\alpha \leq 1$ ou $\beta \leq 1$, serão comentados daqui a pouco.

A constante C da densidade de probabilidade é obtida de forma que a densidade tenha área total abaixo de $f(x)$ igual a 1:

$$\begin{aligned} 1 &= \int_0^1 Cx^{\alpha-1}(1-x)^{\beta-1} dx \\ &= C \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx \end{aligned}$$

onde concluímos que

$$C = \frac{1}{\int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx}$$

Pode-se mostrar que esta integral no denominador por ser expressa em termos de uma função matemática conhecida como função gama e denotada por $\Gamma(z)$:

$$C = \frac{1}{\int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \quad (7.1)$$

A função gama $\Gamma(z)$ é uma generalização do fatorial de número inteiro positivo. Isto é, quando z é um inteiro positivo temos $\Gamma(z) = (z-1)!$. Embora não seja relevante neste livro, a definição da função gama é a seguinte:

$$\Gamma(z) = \int_0^\infty y^{z-1} e^{-y} dy.$$

A função $\Gamma(z)$ é contínua e portanto pode ser imaginada como um tipo de interpolação entre os fatoriais dos inteiros. A outra propriedade fundamental da função gama é que $\Gamma(z+1) = z\Gamma(z)$ para todo $z > 0$. Ela está implementada em R e pode ser chamada com o comando `gamma`:

```
> gamma(c(4, 4.72, 4.73, 5, 5.25, 6, 7))
[1] 6.00000 15.88223 16.11313 24.00000 35.21161 120.00000 [7] 720.00000
```

Em geral, a integral no denominador de (7.1) deve ser obtida numericamente a não ser que α e β sejam inteiros positivos. Neste caso particular, temos

$$\int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx = \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!}.$$

A distribuição beta inclui a distribuição uniforme pois, se $\alpha = 1$ e $\beta = 1$, teremos $f(x) = Cx^{1-1}(1-x)^{1-1} = C$ para $x \in (0, 1)$. Como a integral em $(0, 1)$ deve ser 1, temos $C = 1$. Assim, $f(x) = 1$ para $x \in (0, 1)$, que é a distribuição uniforme.

Quando $0 < \alpha < 1$ ou $0 < \beta < 1$ temos uma densidade com um formato bem diferente. Neste caso, a densidade vai para infinito em pelo menos um dos dois extremos, 0 ou 1. Por exemplo, se $\alpha = 1/2$ e $\beta = 4$ teremos $f(x) = C(1-x)^3/\sqrt{x}$. Quando $x \rightarrow 0$ teremos $f(x) \rightarrow \infty$. Curiosamente, apesar disso, a área total sob a função $g(x) = (1-x)^3/\sqrt{x}$ é finita e portanto a constante C pode ser calculada e uma densidade $f(x)$ realmente existe com as escolhas de $\alpha = 1/2$ e $\beta = 4$.

A Figura 7.12 mostra a grande diversidade de formas tomada pelo gráfico da densidade $f(x)$ da distribuição beta em função dos valores dos parâmetros $\alpha > 0$ e $\beta > 0$. As curvas no gráfico (1, 1) são exemplos com $\alpha > 1$ e $\beta > 1$. Eles têm um único máximo bem definido na posição $x = \alpha/(\alpha + \beta)$. Existe uma simetria nos parâmetros α e β . As curvas de, por exemplo, uma Beta(15, 5) e uma Beta(5, 15) são uma reflexão da outra ao redor do ponto central $x = 1/2$. As curvas no gráfico (1, 2) são exemplos com seu ponto de máximo fixado na posição $\alpha/(\alpha + \beta) = 0.75$ mas com os dois parâmetros crescendo. O efeito de fazer α e β crescerem mantendo a razão $\alpha/(\alpha + \beta)$ fixa é tornar a densidade cada vez mais concentrada em torno de seu ponto de máximo. As curvas no gráfico (1, 2) são exemplos com $\alpha = \beta$. Neste caso, as densidades são simétricas em torno de seu máximo em $x = 1/2$. Finalmente, as curvas no gráfico (2, 2) são exemplos com um dos parâmetros menor que 1 e o caso $\alpha = \beta = 1$, que equivale à distribuição uniforme. No caso em que $\alpha = \beta = 1/2$, a curva assintota em direção ao infinito nos dois extremos do intervalo $(0, 1)$ e tem uma forma mais ou menos uniforme na maior parte do miolo do intervalo. Esta distribuição é muito importante em vários modelos bayesianos modernos tal como o modelo *Latent Dirichlet Allocation* para tratamento de textos. A distribuição beta com apenas um dos parâmetros menor que 1, como a Beta(0.5, 5) neste gráfico, possui uma assintóta para o infinito em $x = 0$ e diminui para zero quando x cresce para 1. O caso $\beta < 1$ e $\alpha > 1$ obedece à simetria do espelho observada antes. O código R para esta figura está abaixo.

```
opar <- par()      # make a copy of current settings
par(mfrow=c(2,2), mar=c(1,1,1,1))

x = seq(0,1,by=0.001)
plot(x, dbeta(x,15,5), type="l", ylab="f(x)", axes=F); box()
lines(x, dbeta(x, 10, 10), col="red"); lines(x, dbeta(x, 5, 15), col="blue")
legend("right", c("b(15,5)","b(10,10)","b(5,15)"), lty=1,
       col=c("black", "red", "blue"))

plot(x, dbeta(x,150,50), type="l", ylab="f(x)", axes=F); box()
lines(x, dbeta(x, 30, 10), col="red");
lines(x, dbeta(x, 15, 5), col="blue");
lines(x, dbeta(x, 6, 2), col="green");
legend("left", c("b(150,50)","b(30,10)","b(15,5)","b(6,2)"), lty=1,
       col=c("black", "red", "blue", "green"))

plot(x, dbeta(x,100,100), type="l", ylab="f(x)", axes=F); box()
lines(x, dbeta(x, 50, 50), col="red");
```

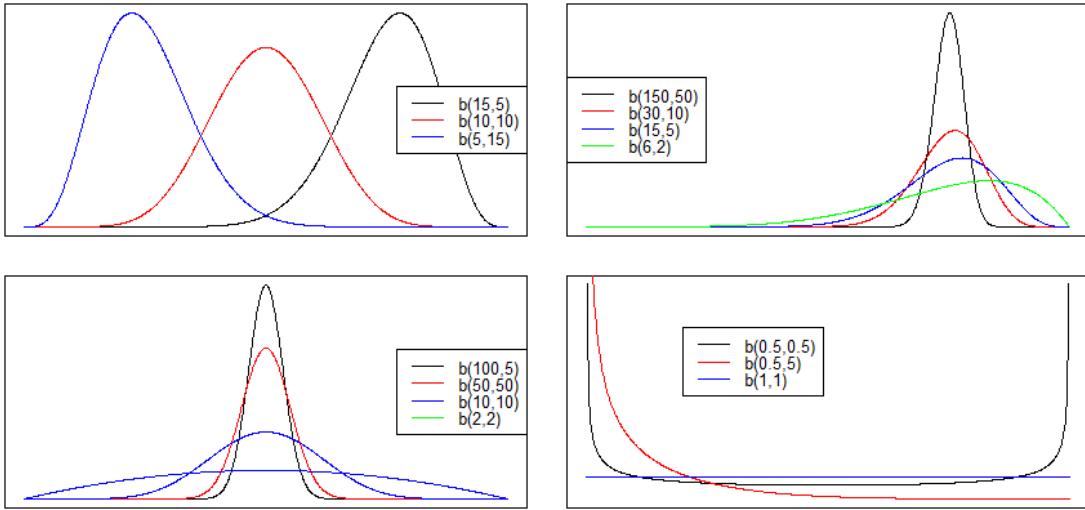


Figure 7.12: Gráficos da função densidade $\text{Beta}(\alpha, \beta)$ variando os parâmetros α e β . Direita, acima: exemplos com $\alpha > 1$ e $\beta > 1$; Esquerda, acima: efeito de fazer α e β crescerem mantendo a razão $\alpha/(\alpha + \beta)$ fixa; Direita, embaixo: $\alpha = \beta > 1$; Esquerda, embaixo: $\alpha < 1$.

```

lines(x, dbeta(x, 10, 10), col="blue");
lines(x, dbeta(x, 2, 2), col="blue");
legend("right", c("b(100,5)", "b(50,50)", "b(10,10)", "b(2,2)"), lty=1,
       col=c("black", "red", "blue", "green"))

plot(x, dbeta(x,0.5,0.5), type="l", ylab="f(x)", ylim=c(0,10), axes=F); box()
lines(x, dbeta(x, 0.5, 5), col="red");
lines(x, dbeta(x, 1, 1), col="blue");
legend(0.2,8, c("b(0.5,0.5)", "b(0.5,5)", "b(1,1)"), lty=1,
       col=c("black", "red", "blue", "green"))

par(opar) # restore original graphical parameters

```

■ Example 7.5 — Preços de Ações. Vamos representar por X a variável que mede a proporção da variação do preço médio diários de ações quando estes preços caem. Isto é, se o preço médio de uma ação num dia é p_1 e o preço desta ação no dia seguinte é $p_2 < p_1$ então $X = (p_1 - p_2)/p_1$. Se o preço sobe de um dia para o outro a v.a. X não é registrada para aquela ação. Um total de 2314 ações com preços que caíram de um dia para o outro foram observadas. Os dados contínuos foram agrupados em bins e estão resumidos na forma de distribuição de frequência f_i na tabela abaixo. A Figura 7.13 apresenta um histograma desses dados junto com a densidade de uma distribuição beta com parâmetros $\alpha = 1.04$ e $\beta = 10.63$. Usamos o método de momentos para escolher estes valores para os parâmetros e este método será assunto do capítulo ???. Do ponto de vista puramente visual, parece que a distribuição beta fornece um modelo razoável para a variável X pois a densidade da beta ajustou-se razoavelmente bem ao histograma.

y	0.02	0.06	0.10	0.14	0.18	0.22	0.26	0.30	0.34	0.38	0.42	0.46	0.50	0.54
f_i	780	567	373	227	147	84	49	43	17	12	9	3	2	1

O código em R para produzir esta figura está abaixo. Existe um truque para colocar um gráfico de linha ou pontos sobre um barplot em R. Salvando o valor de retorno do comando `barplot` temos `df.bar`, um objeto matriz com uma única coluna contendo os valores que são usados pelo `barplot` no eixo x . Ajustando sua linha com esta escala, tudo dá certo no final. Veja o código.

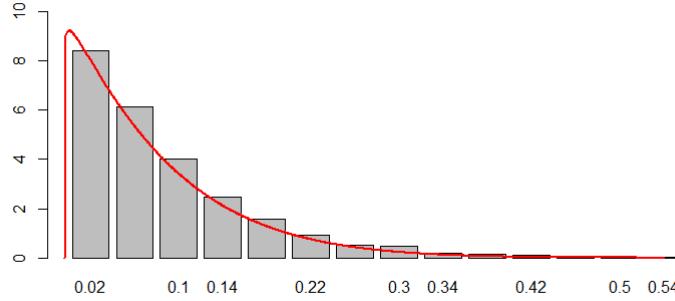


Figure 7.13: Histograma padronizado com os dados da proporção de queda do valor diário de 2314 ações que viram seu preço cair de um dia para o outro. A curva em vermelho é a densidade de uma Beta(1.04, 10.63).

```
counts = c(780,567,373,227,147,84,49,43,17,12,9,3,2,1)
freqrel = counts/(sum(counts)*0.04)
df.bar = barplot(freqrel, names.arg=as.character(seq(0.02, 0.54, by=0.04)), ylim=c(0,10))
lines(seq(0,max(df.bar),len=1000), dbeta(seq(0, 0.60, len=1000), 1.04, 10.63), lwd=2, col="red")
```

A esperança de uma v.a. $X \sim \text{beta}(\alpha, \beta)$ é obtida fazendo a integral:

$$\begin{aligned}\mathbb{E}(X) &= \int_0^1 x f(x) dx = C \int_0^1 x x^{\alpha-1} (1-x)^{\beta-1} dx \\ &= C \int_0^1 x^\alpha (1-x)^{\beta-1} dx\end{aligned}$$

Afirmamos anteriormente que a constante C de uma Beta(α, β) pode ser escrita como $C = \Gamma(\alpha + \beta)/(\Gamma(\alpha)\Gamma(\beta))$ onde $\Gamma(z)$ é a função gama. A função gama cria uma interpolação entre os fatoriais pois $\Gamma(n) = (n-1)!$ quando n é um número inteiro positivo. A sua definição formal é

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx \quad (7.2)$$

Pode-se mostrar que usando esta função possui uma propriedade recursiva: $\Gamma(z+1) = z \Gamma(z)$.

Note agora que uma Beta($\alpha + 1, \beta$) teria densidade $C^* x^{\alpha+1-1} (1-x)^{\beta-1}$ que corresponde ao núcleo dentro da última integral no desenvolvimento acima. Assim, esta integral tem de ser igual a $1/C^*$ e portanto

$$\mathbb{E}(X) = \frac{C}{C^*} = \frac{\Gamma(\alpha+\beta)/(\Gamma(\alpha)\Gamma(\beta))}{\Gamma(\alpha+1+\beta)/(\Gamma(\alpha+1)\Gamma(\beta))} = \frac{\alpha}{\alpha+\beta}$$

Em conclusão, no caso de uma distribuição Beta(α, β), a esperança é $\alpha/(\alpha + \beta)$.

A distribuição acumulada de probabilidade não possui uma fórmula fechada, uma expressão analítica. Ela é bem estudada numericamente e pode ser usada em cálculos de probabilidade sem dificuldade.

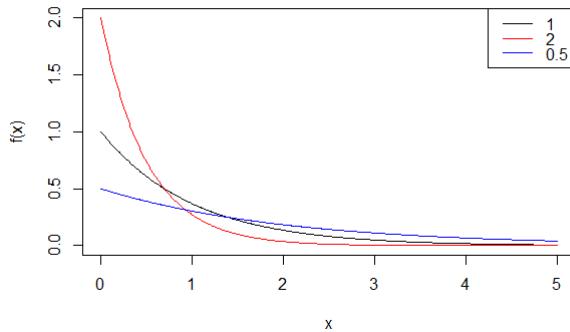


Figure 7.14: Densidade da distribuição exponencial com λ igual a 1, 2 e 0.5.

7.6 Distribuição exponencial

A distribuição exponencial costuma ser um bom modelo para o tempo *entre* eventos aleatórios que ocorrem continuamente no tempo e a uma taxa λ constante. Ela foi usada de forma pioneira por Erlang, um matemático dinamarquês, em 1909 mostrando que o tempo de espera entre chamadas telefônicas num servidor tinha distribuição exponencial com um mesmo parâmetro λ e que elas eram v.a.'s independentes. Veremos a definição de v.a.'s independentes no capítulo 8 mas, por ora, basta saber que, se as últimas ligações chegaram muito rapidamente, sem precisar esperar muito, as próximas ligações não são afetadas, continuam seguindo a mesma distribuição.

Outros exemplos de tempo de espera entre ocorrências aleatórias mas que possuem uma taxa mais ou menos constante são os seguintes:

- O tempo de espera até o próximo decaimento radioativo de uma massa atômica (tempo entre contagens de um contador Geiger)
- distância entre mutações numa cadeia de DNA
- tempo entre comunicações em redes sociais (ver [2]).
- tempo entre acessos numa página web

Na prática, a suposição de uma taxa λ constante de ocorrência de eventos é irreal. Por exemplo, a taxa de chegada de chamadas telefônicas muda ao longo do dia. Entretanto, se considerarmos apenas uma faixa de tempo, tal como entre 14 e 16 horas de um dia útil, ela pode ser considerada aproximadamente constante e a distribuição exponencial pode ser um bom modelo para o tempo entre chamadas.

Definition 7.6.1 — Distribuição exponencial. Uma v.a. X com suporte $(0, \infty)$ é chamada de exponencial se sua densidade de probabilidade é dada por $f(x) = \lambda e^{-\lambda x}$ onde $\lambda > 0$ é o parâmetro da distribuição.

■ **Notation 7.3.** Se a v.a. X segue a distribuição exponencial com parâmetro λ escrevemos $X \sim \exp(\lambda)$.

A Figura 7.14 mostra três exemplos da densidade exponencial. Ela não apresenta muita variabilidade. É sempre apenas um decaimento exponencial a partir da origem. O ritmo do decaimento é ditado pelo parâmetro λ . Quanto maior o valor de λ , mais depressa o decaimento e portanto mais concentrado em torno de zero é o valor de X .

Refletindo este controle do parâmetro λ nos valores de X , a esperança de X pode ser obtida

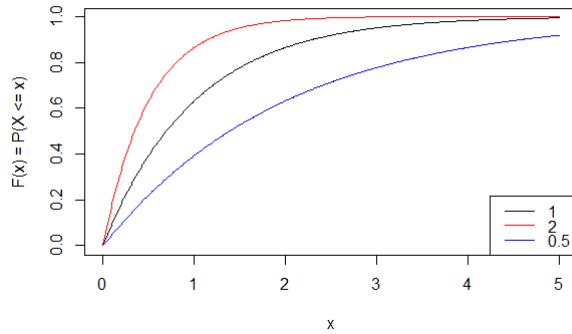


Figure 7.15: Densidade da distribuição exponencial com λ igual a 1, 2 e 0.5.

fazendo-se integração por partes com $u = \lambda x$ e $dv = e^{-\lambda x}$:

$$\mathbb{E}(X) = \int_0^\infty x \lambda e^{-\lambda x} dx = -\frac{x\lambda e^{-\lambda x}}{\lambda} \Big|_0^\infty + \int_0^\infty e^{-\lambda x} dx = -(0 - 0) + \left(-\frac{e^{-\lambda x}}{\lambda} \Big|_0^\infty \right) = \frac{1}{\lambda}.$$

A distribuição acumulada de probabilidades $\mathbb{F}(x) = \mathbb{P}(X \leq x)$ é facilmente obtida. Como o suporte da distribuição exponencial é o semi-eixo $(0, \infty)$ teremos $\mathbb{F}(x) = \mathbb{P}(X \leq x) = 0$ para todo $x < 0$. Para $x \geq 0$, temos:

$$\mathbb{F}(x) = \mathbb{P}(X \leq x) = \int_0^x \lambda e^{-\lambda x} dx = 1 - e^{-\lambda x}.$$

O gráfico desta função acumulada com diferentes valores para λ está na Figura 7.15.

■ **Example 7.6 — Impulsos elétricos.** Psicólogos e biofísicos tem interesse no tempo X entre impulsos elétricos sucessivos na medula espinhal de vários mamíferos. W. J. Megill fez várias medições do tempo decorrido entre os impulsos na medula de um gato. Foram registrados 391 intervalos e o histograma destes dados, na Figura 7.16, sugere a adoção de um modelo exponencial.

■ **Example 7.7 — Ar condicionado em Boeings.** [20] registrou os tempos entre falhas sucesivas dos sistemas de ar condicionado de uma frota de 13 jatos Boeing 730. Estes tempos entre falhas durações de tempo estão listadas na tabela abaixo. Assim, o avião de número 7907 teve uma falha de seu sistema de ar condicionado após 194 horas de serviço, uma terceira falha 41 horas de serviço após a segunda falha, e assim por diante. Mais ou menos após 2000 horas de serviço, cada avião receberia uma inspeção geral. Se um intervalo de tempo entre duas falhas do sistema de ar condicionado incluisse a inspeção geral, o comprimento deste intervalo de tempo não era registrado (e portanto não está na tabela abaixo) pois sua magnitude pode ter sido afetada por possíveis reparos feitos durante a inspeção geral.

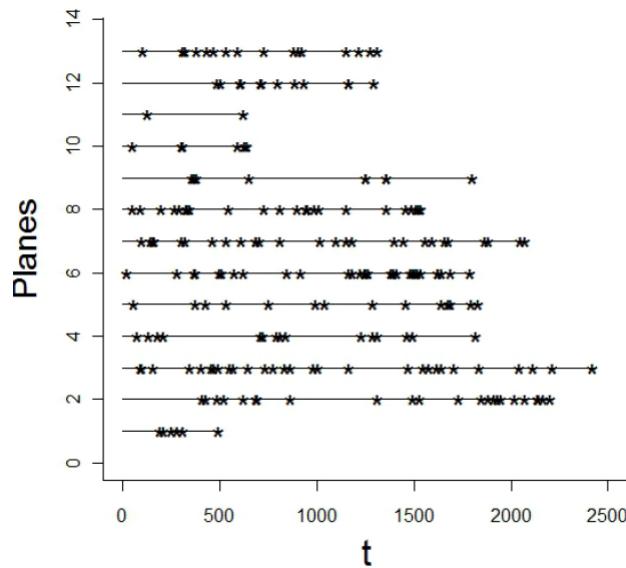


Figure 7.17: Visualização dos tempos de reparos de ar condicionados em Boeings. Cadalinha é um avião. Cada estrela marca o momento de um reparo.

Um histograma destes 213 dados sugere que o modelo exponencial fornece um bom ajuste aos dados. A Figura 7.18 mostra o histograma padronizado dos dados e uma densidade exponencial usando $\lambda = 0.0107$, igual ao inverso da média aritmética. A razão para esta escolha de λ é que $\mathbb{E}(X) = 1/\lambda$ no caso exponencial, e $\mathbb{E}(X)$ é aproximadamente igual à média aritmética dos dados. Assim, devemos ter $\lambda \approx 1/\bar{x}$, onde \bar{x} é a média aritmética dos dados.

7.7 Distribuição normal ou gaussiana

A distribuição normal ou gaussiana é a mais famosa distribuição de probabilidade. O requerimento mínimo para adotarmos o modelo normal para um conjunto de dados contínuos é que seu histograma seja aproximadamente simétrico em torno do ponto central, que também deve ser o ponto de máximo. Histogramas razoavelmente simétricos não são muito comuns. Eles aparecem quando lidamos com medidas biométricas, principalmente com medidas antropométricas, como aquelas da Figura 7.19. Ela mostra a distribuição de frequência das alturas (em polegadas) de homens adultos nascidos nas Ilhas Britânicas, segundo dados publicados por uma comissão da British Association em 1883. Estas distribuições de frequência são do tipo simétrico em que a distribuição normal ou gaussiana se ajusta bem.

Definition 7.7.1 — Distribuição normal ou gaussiana. Uma v.a. X com suporte na reta real $\mathbb{R} = (-\infty, \infty)$ possui distribuição normal ou gaussiana com parâmetros $\mu \in \mathbb{R}$ e $\sigma^2 > 0$ se sua densidade de probabilidade for igual a

$$f(x) = C \exp\left(-\frac{1}{2} \left(\frac{x-\mu}{\sigma}\right)^2\right)$$

■ **Notation 7.4.** Se a v.a. X segue a distribuição normal ou gaussiana com parâmetros μ e σ^2 escrevemos $X \sim \mathcal{N}(\mu, \sigma^2)$.

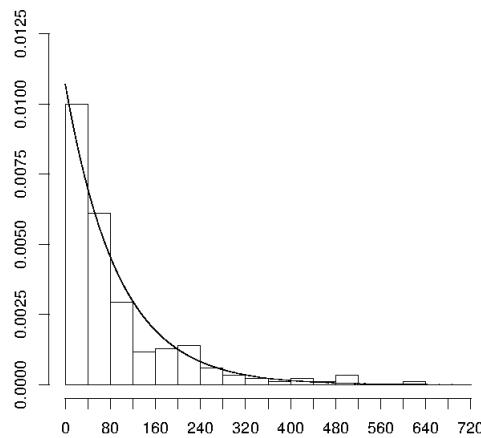


Figure 7.18: Histograma padronizado e densidade exponencial usando λ igual ao inverso da média aritmética dos tempos na amostra.

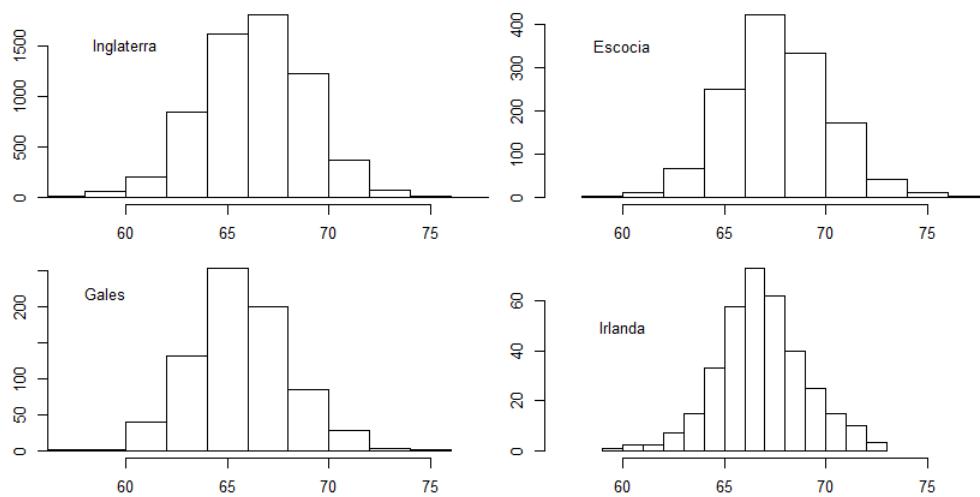


Figure 7.19: Histogramas de alturas (em polegadas) de amostras de indivíduos adultos d sexo masculino da Grã-Bretanha em 1883.

■ **Example 7.8 — Medição biométrica.** A Figura 7.20 mostra o histograma padronizado de medições do diâmetro transversal da cabeça de 1000 estudantes de Cambridge. As medidas foram tomadas ao décimo de polegada mais próximo. A curva em vermelho é a densidade de uma v.a. normal com μ igual à média aritmética e σ igual ao desvio-padrão amostral (ver capítulo 9).

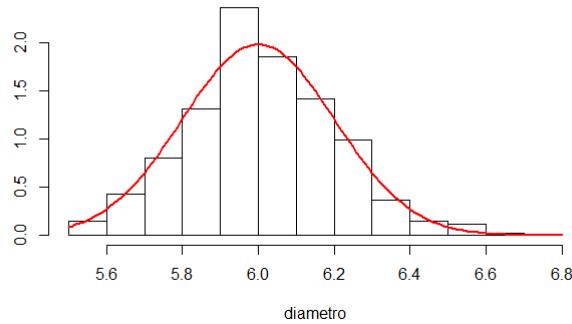


Figure 7.20: Histogramas dos diâmetros transversais da cabeça de 1000 estudantes da Universidade de Cambridge

■ **Example 7.9 — Diâmetros Biacromial e Biiliac.** Este exemplo usa dados de 21 medidas de dimensão corporal, bem como idade, peso, altura e sexo em 507 indivíduos saudáveis e que fazem exercícios físicos regularmente várias horas por semana. São 247 homens e 260 mulheres concentrados entre os 20 e 30 anos. Os dados apareceram em [15]. A Figura ?? mostra três locais de algumas das medições feitas pelo autores. A Figura ?? mostra os histogramas para homens (linha superior) e mulheres (linha inferior) com os diâmetros Biacromial, Biiliac e Bitrochanteric, respectivamente, da esquerda para a direita. O eixo horizontal é o mesmo para homens e mulheres para permitir a comparação entre eles. Em cada histograma foi ajustada uma densidade normal. Visualmente o ajuste parece ser bem adequado.

Quando os dados são simétricos e quando queremos comparar várias distribuições, o boxplot é uma excelente ferramenta. Compare a dificuldade em contrastar os histogramas maculinos e

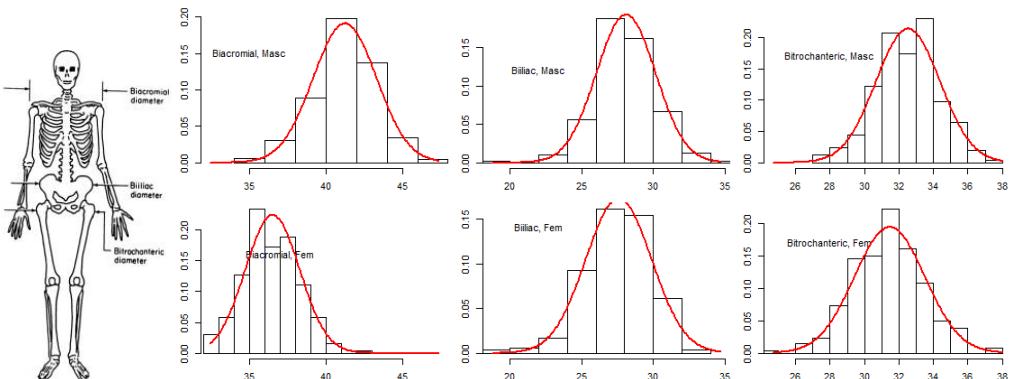


Figure 7.21: Histogramas para homens (linha superior) e mulheres (linha inferior) com os diâmetros Biacromial, Biiliac e Bitrochanteric, respectivamente, da esquerda para a direita.

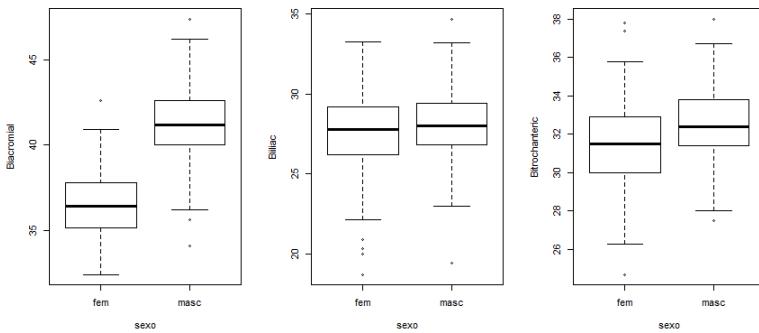


Figure 7.22: Histogramas para homens (linha superior) e mulheres (linha inferior) com os diâmetros Biacromial, Biliac e Bitrochanteric, respectivamente, da esquerda para a direita.

femininos na Figura ?? com os boxplots da Figura 7.22. Esta segunda visualização torna muito mais fácil a tarefa do analista de dados.

O código para estes boxplots segue abaixo.

```
mat = matrix(scan("body.dat.txt"), ncol=25, byrow=T)
sx=mat[,25]
par(mfrow=c(1,3), mar=c(5, 4, 4, 2) + 0.1)
boxplot(mat[,1] ~ sx, xlab="sexo", ylab="Biacromial", names=c("fem","masc"))
boxplot(mat[,2] ~ sx, xlab="sexo", ylab="Biliac", names=c("fem","masc"))
boxplot(mat[,3] ~ sx, xlab="sexo", ylab="Bitrochanteric", names=c("fem","masc"))
```

Apesar desses exemplos em que a distribuição normal $\mathcal{N}(\mu, \sigma^2)$ serve como modelo para os dados, na maioria das vezes os dados vão apresentar aspectos, tais como assimetria, que vão tornar o modelo gaussiano inapropriado. A importância da distribuição normal não está na sua presença como modelo para dados diretamente mensuráveis mas sim na sua aparição quando somamos várias variáveis aleatórias. O Teorema Central do Limite, como o nome está dizendo, é um teorema central para a probabilidade e estatística e é assunto do capítulo 16. Em linhas gerais, ele prova que v.a.'s somadas em grande quantidade converge para uma distribuição normal $\mathcal{N}(\mu, \sigma^2)$. A imensa importância deste fato vai ficar mais clara ao longo deste livro.

A Figura 7.23 mostra o efeito de variar μ e σ no caso de uma densidade normal. O efeito de variar μ (deixando σ fixo) é o de deslocar rigidamente a curva densidade que tem seu ponto de máximo no ponto $x = \mu$. O gráfico da esquerda mostra a densidade de três normais: $\mathcal{N}(-2, 1)$, $\mathcal{N}(0, 1)$, $\mathcal{N}(3, 1)$. O gráfico da direita mostra três densidades com o mesmo valor $\mu = 0$ e com $\sigma = 0.5, 4, 1$. O parâmetro σ controla a dispersão dos pontos em torno de μ . Quanto maior o valor de σ , mais achatada e espalhada em volta de μ é a densidade. Reduzindo σ faz a densidade ficar mais concentrada em torno de μ . Como a área total tem de ser sempre igual a 1, quando a densidade fica mais concentrada, a altura do ponto de máximo se eleva.

```
x = seq(-6, 6, by=0.05)
par(mfrow=c(1,2))
plot(x, dnorm(x, 0, 1), ylab="f(x)", lwd=2, type="l")
lines(x, dnorm(x, -2, 1), col="red", lwd=2)
lines(x, dnorm(x, 3, 1), col="blue", lwd=2)
```

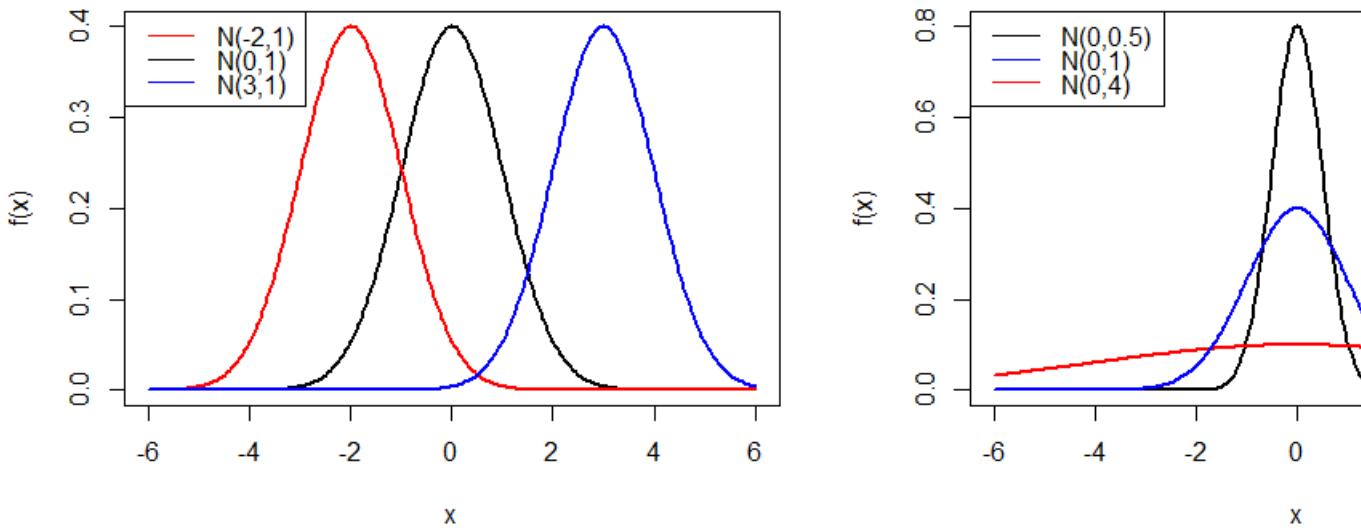


Figure 7.23: Ilustração do efeito de variar μ e σ no caso de uma densidade da distribuição normal $\mathcal{N}(\mu, \sigma^2)$.

```
legend("topleft",c("N(-2,1)", "N(0,1)", "N(3,1)"), lty=1, col=c("red", "black", "blue"))

plot(x, dnorm(x, 0, 0.5), ylab="f(x)", lwd=2, type="l")
lines(x, dnorm(x, 0, 4), col="red", lwd=2)
lines(x, dnorm(x, 0, 1), col="blue", lwd=2)
legend("topleft",c("N(0,0.5)", "N(0,1)", "N(0,4)"), lty=1, col=c( "black","blue" , "red"))
```

Definition 7.7.2 — Regra de 2σ . No caso de uma distribuição normal $\mathcal{N}(\mu, \sigma^2)$, existe uma regra simples e muito útil. Qualquer que sejam os valores dos parâmetros, a área (e portanto, a probabilidade) que fica entre $\mu - \sigma$ e $\mu + \sigma$ é aproximadamente igual a 0.68 (ver Figura 7.24, lado esquerdo). A área que fica localizada a dois σ de μ (istoé, a área entre $\mu - 2\sigma$ e $\mu + 2\sigma$) é aproximadamente igual a 0.95. A uma distância de 3σ de μ fica uma área (e probabilidade) de 0.997, aproximadamente. Assim, é de 5% a chance de uma valor selecionado de uma $\mathcal{N}(\mu, \sigma^2)$ se afastar por mais de 2σ de μ . A chance de se fastar mais de 3σ é bastante pequena.

A constante de integração na densidade $f(x)$ é o valor que faz a área total debaixo da curva ser igual a 1. Este valor é conhecido como uma fórmula fechada: $C = 1/(\sqrt{2\pi\sigma^2})$. Está além dos objetivos deste texto demonstrar este fato. O leitor interessado deve consultar [17]. A esperança de uma uma distribuição normal $\mathcal{N}(\mu, \sigma^2)$ pode ser obtida explicitamente explorando a propriedade da simetria da densidade em torno de μ . Pode-se mostrar que $\mathbb{E}(X) = \mu$. Finalmente, a função distribuição acumulada $\mathbb{F}(x)$ de uma v.a. $\mathcal{N}(\mu, \sigma^2)$ não tem uma forma funcional simples e tem de ser obtida numericamente. O lado direito da Figura 7.24 mostra a densidade $f(x)$ de uma $\mathcal{N}(0, 1)$ na parte de cima e a função distribuição acumulada $\mathbb{F}(x)$ na parte de baixo. Note que os eixos horizontais são os mesmos para facilitar a comparação entre elas.

We usually write $\phi(x)$ for the pdf and $\Phi(x)$ for the cdf of the standard normal.

This is a rather important probability distribution. This is partly due to the central limit theorem,

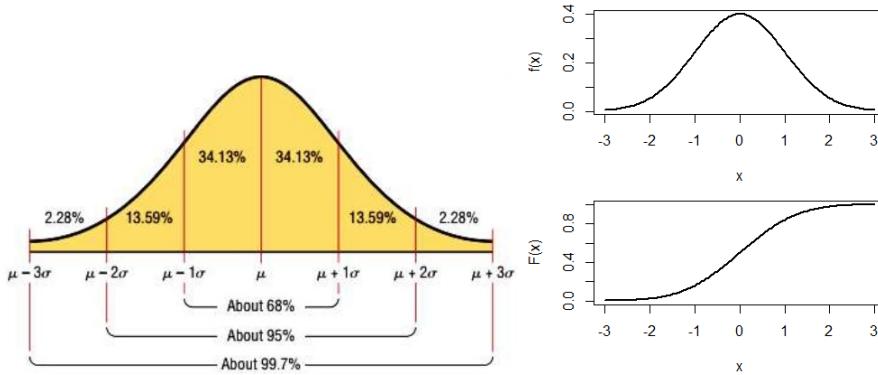


Figure 7.24: Ilustração da regra de 2σ e de 3σ no caso de uma distribuição normal $\mathcal{N}(\mu, \sigma^2)$.

which says that if we have a large number of iid random variables, then the distribution of their averages are approximately normal. Many distributions in physics and other sciences are also approximately or exactly normal.

We first have to show that this makes sense, i.e.

Proposition 7.7.1

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx = 1.$$

Proof: Substitute $z = \frac{(x-\mu)}{\sigma}$. Then

$$I = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz.$$

Then

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \\ &= \int_0^{\infty} \int_0^{2\pi} \frac{1}{2\pi} e^{-r^2/2} r dr d\theta \\ &= 1. \end{aligned}$$

We also have

Proposition 7.7.2 $\mathbb{E}[X] = \mu$.

Proof:

$$\begin{aligned} \mathbb{E}[X] &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} xe^{-(x-\mu)^2/2\sigma^2} dx \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x-\mu)e^{-(x-\mu)^2/2\sigma^2} dx + \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \mu e^{-(x-\mu)^2/2\sigma^2} dx. \end{aligned}$$

The first term is antisymmetric about μ and gives 0. The second is just μ times the integral we did above. So we get μ .

Also, by symmetry, the mode and median of a normal distribution are also both μ .

Proposition 7.7.3 $\mathbb{V}(X) = \sigma^2$.

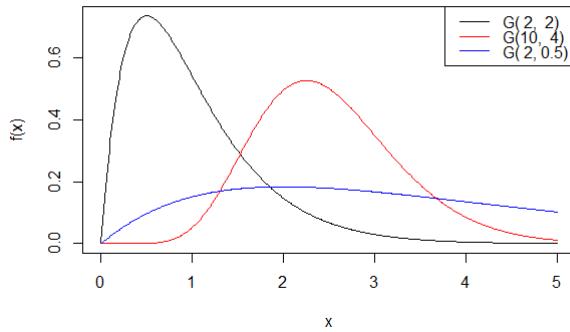


Figure 7.25: Exemplos de Gama.

Proof: We have $\mathbb{V}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$. Substitute $Z = \frac{X-\mu}{\sigma}$. Then $\mathbb{E}[Z] = 0$, $\mathbb{E}[Z^2] = \frac{1}{\sigma^2} \mathbb{E}[X^2]$. Then

$$\begin{aligned}\mathbb{V}(Z) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2} dz \\ &= \left[-\frac{1}{\sqrt{2\pi}} z e^{-z^2/2} \right]_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} dz \\ &= 0 + 1 \\ &= 1\end{aligned}$$

So $\mathbb{V}X = \sigma^2$.

7.8 Distribuição gama

A distribuição gama tem sido usada para modelar v.a.'s positivas com certa assimetria. A Figura 7.25 mostra alguns exemplos da densidade de uma distribuição gama. As densidades começam iguais a zero perto da origem, crescem num ritmo que depende de parâmetros da distribuição e, após atingir um pico, descrescem em direção à zero. O suporte da distribuição é o semi-eixo positivo $(0, \infty)$.

Definition 7.8.1 — Distribuição gama. Uma v.a. X com $(0, \infty)$ como suporte tem distribuição gama com parâmetros $\alpha > 0$ e $\beta > 0$ se a sua densidade é dada por

$$f(x) = Cx^{\alpha-1}e^{-\beta x}$$

para $x > 0$. A constante C é obtida para garantir que a área sob a densidade é igual a 1.

■ **Notation 7.5.** Se a v.a. X segue a distribuição gama com parâmetros α e β escrevemos $X \sim \Gamma(\alpha, \beta)$.

Na Figura 7.25 temos as densidades das distribuições gama com parâmetros α e β variados: $\Gamma(2, 2)$, $\Gamma(10, 4)$ e $\Gamma(2, 1/2)$. Considere uma dessas densidades com $\alpha > 1$ e $\beta > 1$. Por exemplo, a densidade $f(x) = Cx^9e^{-4x}$ de uma $\Gamma(10, 4)$. Quando $x \approx 0$ (e maior que zero), teremos $x^9 \approx 0$ e $e^{-4x} \approx 1$ levando a uma densidade $f(x) \approx 0$. Se tomarmos x indo para ∞ teremos $x^9 \rightarrow \infty$ e $e^{-4x} \rightarrow 0$. O produto desses dois termos ficará próximo de zero pois o decrescimento exponencial domina qualquer crescimento polinomial. Assim, $f(x) \approx 0$ tanto se $x \approx 0$ quanto se $x \rightarrow \infty$. Para x



Figure 7.26: Foto de besouro da espécie *Tribolium castaneum* e amostra de milho infestado por eles cercado de milhos sadios.

no meio desses dois extremos a densidade terá um valor moderado com um único ponto de máximo bem definido. A forma exata da densidade será ditada pelos dois valores α e β .

A distribuição gama aparece naturalmente quando trabalhamos com distribuições exponenciais. Suponha que o tempo de espera entre dois eventos siga uma distribuição exponencial com parâmetro λ . Assuma também que os tempos sucessivos sejam independentes. Então o tempo de espera por k eventos sucessivos segue uma distribuição $\Gamma(k, \lambda)$. Isto é, se T_1, T_2, \dots, T_k são v.a.'s $\exp(\lambda)$ e independentes então $X = T_1 + T_2 + \dots + T_k \sim \Gamma(k, \lambda)$. Este caso especial da gama, com α igual a um inteiro positivo, aparece com tanta frequência em aplicações que acabou ganhando o nome de distribuição de Erlang, em homenagem a Agner Krarup Erlang, um matemático dinamarquês que viveu entre 1878 e 1929 e estudou as propriedades probabilísticas do tráfego telefônico, uma indústria nascente na época.

Outra ligação entre a distribuição exponencial e a distribuição gama é que a distribuição exponencial é um caso particular da distribuição gama. Se fizermos $\alpha = 1$ o termo polinomial da densidade de uma $\Gamma(1, \beta)$ desaparece e ficamos simplesmente com uma $\exp(\beta)$.

Vamos apresentar outro exemplo de que a função gama aparece naturalmente através da manipulação de outras v.a.'s. Considere um vetor de dimensão n em que cada entrada do vetor é uma variável aleatória gaussiana $\mathcal{N}(0, 1)$, com $\mu = 0$ e $\sigma^2 = \sigma = 1$. Quando as entradas são valores aleatórios independentes uns dos outros, o comprimento ao quadrado do vetor também será aleatório e terá uma distribuição $\Gamma(n/2, 2\sigma^2)$. Este fato é fundamental na análise de variância e em modelos de regressão, como veremos no capítulo ??.

■ **Example 7.10 — Populações de Besouro de Farinha.** Os besouros da espécie *Tribolium castaneum*, conhecidos como besoura de farinha, são pragas que atacam produtos armazenados tendo preferência por cereais moídos, como farelo, rações, farinhas e fubá. Estes insetos são responsáveis pela perda total em armazéns de estocagem. A Figura 7.26 mostra um espécime desses besouros e um milho danificado por eles cercado por milhos sadios.

[6] estudou matematicamente e empiricamente o crescimento de populações desses insetos sob diversas condições. Eles apresentaram justificativas ecológicas para afirmar que o número de indivíduos numa população, após certo tempo, era um valor incerto mas que seguia uma distribuição gama. Eles encontraram fortes evidências em dados de laboratório de que isto era verdade. Repetindo o experimento de deixar crescer uma população a partir de certo número inicial de indivíduos eles verificaram que, ao final de certo período, um histograma do tamanho das diferentes populações ajustava-se muito bem a uma distribuição gama. Ver Figura 7.27. Outros estudos têm mostrado que a distribuição gama costuma ser um bom ajuste para a abundância de espécies por razões teóricas ([8]) ou empíricas. Por exemplo, [22] mostraram que a distribuição gama foi a melhor distribuição para ajustar dados de tamanho de comunidades de 128 dentre 136 diferentes

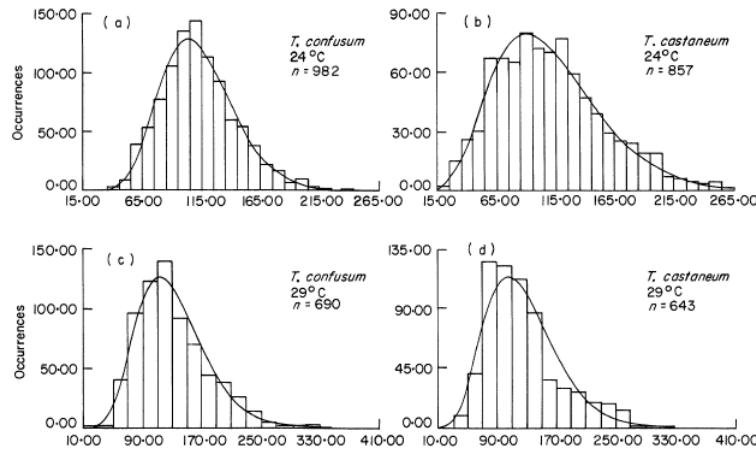


Figure 7.27: Histogramas dos tamanhos de n populações de besouros crescidas sob diferentes condições e ajuste da distribuição gama em cada caso. Parâmetros da gama foram obtidos por máxima verossimilhança, assunto do capítulo ??.

espécies de invertebrados marítimos. ■

A constante de integração na densidade $f(x) = Cx^{\alpha-1}e^{-\beta x}$, para $x > 0$, é o valor que faz a área total debaixo da curva ser igual a 1. Fazendo a substituição de variáveis $\beta x = y$ e $dx = dy$ temos:

$$\begin{aligned} 1 &= \int_0^\infty C x^{\alpha-1} e^{-\beta x} dx \\ &= C \frac{1}{\beta^\alpha} \int_0^\infty y^{\alpha-1} e^{-y} dy \\ &= C \frac{\Gamma(\alpha)}{\beta^\alpha} \end{aligned}$$

onde usamos a definição da função gama em 7.2. Portanto, $C = \beta^\alpha / \Gamma(\alpha)$. Quando $\alpha = k$ for um inteiro positivo $\Gamma(k) = (k-1)!$ e portanto a constante é conhecida exatamente. Caso contrário, temos uma aproximação numérica pois não existe fórmula fechada para $\Gamma(\alpha)$ quando α não é um inteiro.

A esperança de uma v.a. $X \sim \Gamma(\alpha, \beta)$ pode ser obtida de forma explícita:

$$\mathbb{E}(X) = \int_0^\infty x C x^{\alpha-1} e^{-\beta x} dx = C \int_0^\infty x^{\alpha+1-1} e^{-\beta x} dx = C/C^*$$

onde C^* é a constante de integração de uma $\Gamma(\alpha+1, \beta)$. Como já sabemos obter esta constante, e como $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$, temos

$$\mathbb{E}(X) = \frac{\beta^\alpha / \Gamma(\alpha)}{\beta^{\alpha+1} / \Gamma(\alpha+1)} = \frac{\alpha}{\beta}$$

A função distribuição acumulada $\mathbb{F}(x)$ de uma v.a. $X \sim \Gamma(\alpha, \beta)$ não tem uma expressão analítica simples exceto em certos casos excepcionais. Portanto, o valor

$$\mathbb{F}(x) = \int_0^x C t^{\alpha-1} e^{-\beta t} dt$$

tem de ser obtido numericamente.

7.9 Distribuição Weibull

A distribuição de Weibull pode aparecer de maneira natural quando consideramos uma distribuição de probabilidade para o tempo aleatório de espera até que uma falha aconteça. Este é um assunto fundamental nos estudos de confiabilidade (*reliability*, em inglês) de sistema e máquinas. Ela aparece também no estudo de tempos de sobrevivência de humanos e outros seres vivos. Suponha que T seja o tempo de vida aleatório de um componente. Tomando um pequeno intervalo Δt , os estudos de confiabilidade desejam calcular $\mathbb{P}(t < T < t + \Delta t | T > t)$. Isto é, queremos a probabilidade do componente falhar durante o pequeno intervalo de tempo $[t, t + \Delta t]$ dado que ele sobreviveu até o tempo t .

Considere, por exemplo, o significado dessa probabilidade em dois momentos. Um deles é após o “nascimento” do espécime ou logo após o componente ser posto em funcionamento. Qual a chance dele sobreviver por, digamos, $\Delta t = 1$ minuto dia dado que ele está novo em folha, tendo funcionado por uma hora? Estamos considerando o momento $t = 60$ próximo do “nascimento” e querendo saber $\mathbb{P}(t < T < t + \Delta t | T > t) = \mathbb{P}(60 < T < 61 | T > 60)$.

Agora queremos saber a chance de sobreviver o próximo 1 minuto dado que o componente já funcionou por 5 anos seguidos. Quanto deveria ser $\mathbb{P}(2628000 < T < 2628001 | T \geq 2628000)$? Esta probabilidade deveria igual, maior ou menor que a anterior? Temos três alternativas básicas. A primeira, a mais comum, é aquela que o material envelhece, deteriorando-se com o tempo, sofrendo desgaste e aumentando sua fragilidade. Neste caso, a chance dele falhar nos próximos minutos quando ele está novo em folha seria bem menor do que a chance dele falhar nos mesmos próximos minutos dado que ele está velho. Pense na vida humana típica e compare a probabilidade de falecer dentro de um ano dado que você está vivo com 15 anos com a probabilidade de falecer em um ano dado que está vivo com 94 anos. Claramente, a segunda probabilidade é bem maior que a primeira.

A segunda alternativa é que o material nunca envelhece, nunca sofre desgaste. Isto significa que, após anos de funcionamento, ele continua tão bom quanto estava quando novo em folha. Isto não significa que o componente seja eterno e nunca falhe. Significa que a chance de falhar no próximo Δt intervalo de tempo não muda com a idade do material. Esta situação é chamada *falta de memória*. É como se o componente não registrasse a passagem do tempo, não guardasse memória (ou qualquer outro sinal) de que já funcionou por algum tempo.

A terceira alternativa pode parecer menos natural: o material melhora com o passar do tempo de modo que a probabilidade de falhar no próximo Δt intervalo de tempo *diminui* com a idade do material. Embora isto pareça pouco prático, pense na mortalidade infantil em humanos. Os primeiros meses de vida são muito mais arriscados do que depois que a criança atinge um ou dois anos de vida.

Deseja-se estudar $\mathbb{P}(t < T < t + \Delta t | T > t)$ para Δt bem pequeno, numa abordagem similar a de equações diferenciais. Aprendemos como um sistema funciona num intervalo Δt pequeno e, a seguir, usando matemática, conseguimos projetar até um tempo mais longuíquo.

Naturalmente, temos $\lim_{\Delta t \rightarrow \infty} \mathbb{P}(t < T < t + \Delta t | T > t) = 0$. Pense que, dado que um espécime está vivo agora, a chance dele não sobreviver nos próximos segundos é bem pequena e que a chance de não sobreviver nos próximos milisegundos é menor ainda. Uma maneira de se fazer este estudo sem ficar preso neste resultado óbvio é padronizar a probabilidade $\mathbb{P}(t < T < t + \Delta t | T > t)$ calculando-a *por unidade de tempo* e levando-a ao limite quando Δt vai a zero.

Definition 7.9.1 — Taxa de Falha Instantânea. A taxa de falha instantânea de uma v.a. T é definida como

$$\lambda(t) = \lim_{\Delta t \rightarrow \infty} \frac{\mathbb{P}(t < T < t + \Delta t | T > t)}{\Delta t} \quad (7.3)$$

Assim, se tivermos a função $\lambda(t)$ e se Δt for pequeno, teremos

$$P(t < t < t + \Delta t \mid T > t) \approx \lambda(t)\Delta t.$$

A questão então é: qual deve ser o tipo da função $\lambda(t)$? Na primeira possibilidade, em que o material envelhece, sofrendo desgaste e aumentando sua fragilidade, temos $\lambda(t)$ crescente com t . A opção mais simples para este crescimento é criar um crescimento linear ou parabólico: $\lambda(t) = b t$ ou $\lambda(t) = b t^2$ onde b é uma constante positiva. Podemos deixar de forma geral um crescimento polinomial $\lambda(t) = b t^{\alpha-1}$ com $\alpha > 1$.

A situação de um distribuição sem memória tem $\lambda(t) = b$ para todo t . Isto equivale a $\lambda(t) = b t^0 = b t^{1-1}$. Ou seja, equivale a manter a definição polinomial $\lambda(t) = b t^{\alpha-1}$ incluindo agora o caso $\alpha = 0$.

A situação de “quanto mais velho, melhor” pode ser representada pela mesma fórmula polinomial $\lambda(t) = b t^{\alpha-1}$ mas tomando $0 < \alpha < 1$. Por exemplo, se $\alpha = 1/2$ então $\lambda(t) = b t^{\alpha-1} = b/\sqrt{t}$, uma taxa de falhas decrescente com t .

Theorem 7.9.1 — Weibull e a taxa de falha. Assuma que uma v.a. T possui uma taxa de falha da falha da forma $\lambda(t) = b t^{\alpha-1}$ onde $b > 0$ e $\alpha > 0$. Então a sua densidade de probabilidade tem de ser a seguinte:

$$f(t) = Ct^{\alpha-1}e^{-(\frac{t}{\beta})^\alpha} \quad (7.4)$$

para $t > 0$. Esta distribuição é chamada de Weibull com parâmetros α e β .

A constante $\beta > 0$ em (7.4) está associada com a constante b da taxa de falha. C é uma constante de integração para que a densidade $f(t)$ integre 1 em $(0, \infty)$. Pode-se mostrar que $C = \alpha/\beta^\alpha$.

Weibull e a taxa de falha. Here is my proof: omitida. ■

■ **Notation 7.6.** Se a v.a. X segue a distribuição Weibull com parâmetros α e β escrevemos $X \sim \mathcal{W}(\alpha, \beta)$.

Como você imaginar, a forma da densidade de uma Weibull $\mathcal{W}(\alpha, \beta)$ depende dos parâmetros α e β . A Figura 7.28 mostra o gráfico da função densidade de uma Weibull $\mathcal{W}(\alpha, \beta)$ com diferentes valores para os parâmetros. No gráfico da esquerda temos $\alpha = 2$ e $\beta = 0.5, 1, 2$. No gráfico do centro temos $\alpha = 10$ e $\beta = 0.5, 1, 2$. No gráfico da direita tomamos $\alpha = 1/2$ e variamos $\beta = 0.5, 1, 2$. O parâmetro α é chamado de *shape parameter*: ele muda a forma da curva. O parâmetro β é chamado de *scale parameter*: este parâmetro apenas faz uma mudança de escala no eixo horizontal.

■ **Example 7.11 — Weibull na velocidade do vento.** A velocidade do vento muda constantemente de acordo com a hora do dia e a época do ano. Mesmo fixando uma estação e uma hora no dia, a velocidade está sempre mudando. Um método de apresentar dados de velocidade do vento é produzir um histograma do número de horas a cada ano que a velocidade do vento está dentro de uma determinada faixa. A Figura 7.29, à esquerda, mostra o histograma padronizado, em que os dados são normalizados dividindo-se pelo número total de horas e tendo área total 1. A distribuição de Weibull costuma ser usada para modelar a velocidade do vento. No norte da Europa, o valor de α fica em torno de 2. O gráfico da direita mostra o ajuste de uma Weibull a outros dados de velocidade do vento. ■

A esperança de uma v.a. $X \sim \mathcal{W}(\alpha, \beta)$ envolve a função gama $\Gamma(Z)$: $\mathbb{E}(X) = \beta\Gamma((\alpha+1)/\alpha)$. A função distribuição acumulada $F(x)$ tem uma forma funcional simples. Para $x > 0$ temos

$$F(x) = \int_0^x C x^{\alpha-1} e^{-(x/\beta)^\alpha} dx = 1 - e^{-(x/\beta)^\alpha}$$

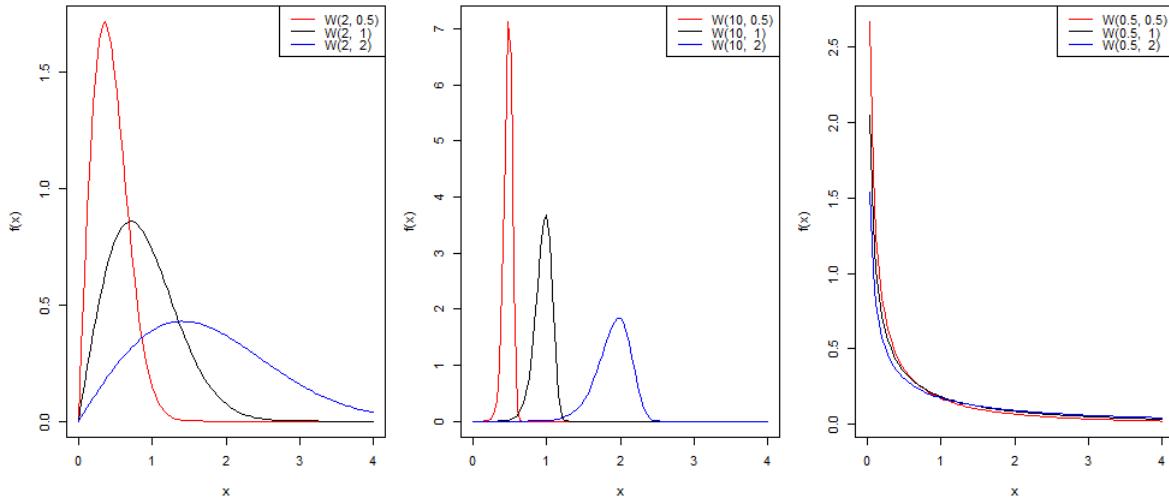


Figure 7.28: Densidade da distribuição Weibull $\mathcal{W}(\alpha, \beta)$. Esquerda: $\alpha = 2$ e $\beta = 0.5, 1, 2$. Centro: $\alpha = 10$ e $\beta = 0.5, 1, 2$. Direita: $\alpha = 0.5$ e $\beta = 0.5, 1, 2$.

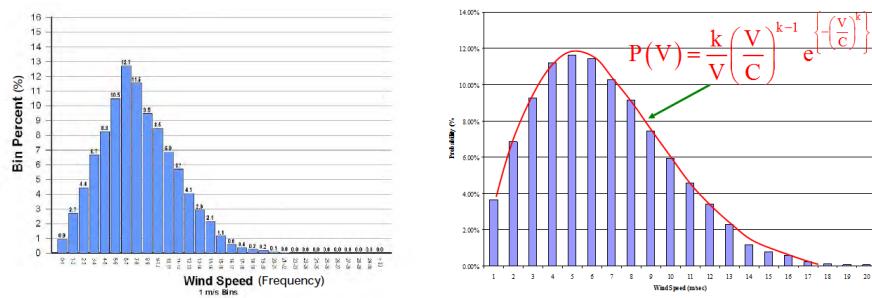


Figure 7.29: Histograma de dados de velocidade do vento (esquerda). Outro histograma com o ajuste de uma Weibull (direita).

Heterogeneous Networks

- Multiple object types and/or multiple link types

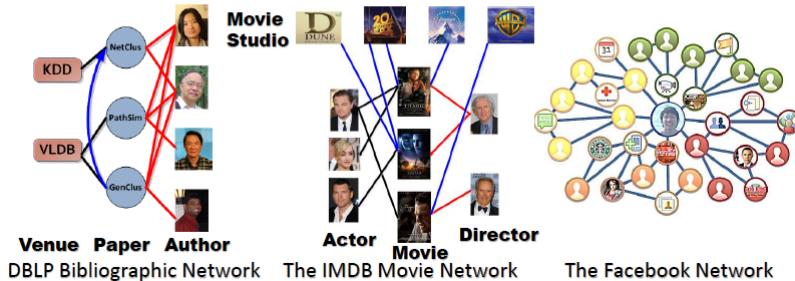


Figure 7.30: Redes heterogêneas. Extraído de [23].

■ **Example 7.12 — Weibull em redes sociais.** [23] usaram a distribuição de Weibull num sofisticado modelo para predizer quando um link iria ocorrer em redes heterogêneas. Redes homogêneas são aquelas em que todos os nós e arestas-relacionamentos são de um mesmo tipo. Por exemplo, uma rede de autores de artigos (os vértices) com links entre aqueles que foram co-autores em algum artigo. Ou uma rede onde filmes formam os vértices e uma aresta entre dois filmes é criada se existir um ator trabalhando em ambos os filmes. Redes heterogêneas são aquelas em que objetos e relacionamentos são de vários tipos. A Figura 7.30 mostra alguns exemplos de redes heterogêneas. Na área de saúde, podemos ter uma rede em que os nós são classificados como médicos, pacientes, hospitais, doenças e tratamentos. Existem arestas entre vértices do mesmo tipo e vértices entre nós de diferentes tipos. Um médico conecta-se com alguns de seus colegas por terem a mesma especialidade, pacientes são conectados aos seus médicos, e assim por diante. Num repositório de códigos, temos como vértices os projetos, os desenvolvedores, as linguagens de programação. Num site de e-commerce, temos os vendedores, os clientes, os produtos e as revisões. A distribuição de Weibull foi usada para predizer quando uma aresta entre dois vértices seria formada. O modelo usado está dentro da classe dos modelos lineares generalizados, a ser estudado no capítulo 22.

■

7.10 Distribuição de Pareto

Estudamos a distribuição de Pareto no caso discreto. No caso contínuo, a distribuição de Pareto surgiu em análises econômicas, especialmente para representar distribuições de renda e valores de perdas em certas classes de seguros. Como antes, teremos a densidade de probabilidade $f(x)$ decrescendo com x na forma polinomial. O fenômeno da cauda pesada também ocorre no caso contínuo. A grande maioria dos dados fica numa faixa estreita de variação mas uma certa proporção não desprezível tem valores ordens de grandeza maior que o valor esperado.

Definition 7.10.1 — Distribuição de Pareto. Uma v.a. X possui distribuição de Pareto com parâmetros $x_o > 0$ e $\alpha > 0$ se ela possui como suporte o conjunto $\mathcal{S} = (x_o, \infty)$ e densidade de probabilidade da forma

$$f(x) = C \frac{1}{x^{\alpha+1}}$$

para $x > x_o$.

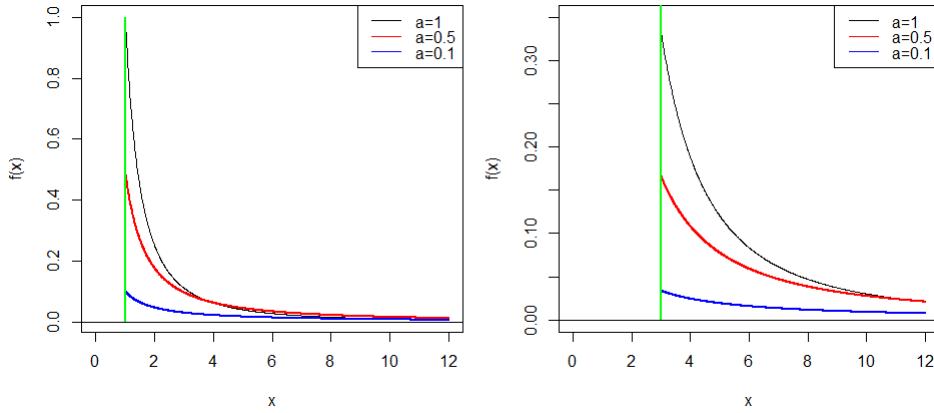


Figure 7.31: Exemplos de densidades de Pareto. Casos a esquerda têm $x_o = 1$ e $\alpha = 1, 0.5, 0.1$. Casos a direita têm $x_o = 3$ e os mesmos valores para α .

A constante C é o valor que faz com que a área abaixo da curva no intervalo (x_o, ∞) seja igual a 1. Pode-se mostrar sem dificuldade que esta constante é $C = \alpha x_o^\alpha$:

$$\begin{aligned} 1 &= \int_{x_o}^{\infty} C \frac{1}{x^{\alpha+1}} dx = C \left(\frac{x^{1-\alpha-1}}{-\alpha} \Big|_{x_o}^{\infty} \right) \\ &= \frac{C}{\alpha} \left(-\frac{1}{\infty^\alpha} - \frac{-1}{x_o^\alpha} \right) \\ &= \frac{C}{\alpha x_o^\alpha} \end{aligned}$$

o que implica em $C = \alpha x_o^\alpha$.

Note que a densidade $f(x)$ é maior que zero apenas para $x > x_o$ e, por sua vez, $x_o > 0$. Por exemplo, se X é a renda de indivíduo escolhido ao acaso de uma população, olhamos apenas aqueles casos em que a renda fica acima de certa quantidade mínima x_o . Os dados relativos à renda costumam ser obtidos através do imposto de renda e pessoas de baixa renda não pagam imposto. Outro exemplo, são os valores pagos por uma seguradora quando sinistros ocorrem com seus segurados. A seguradora só toma conhecimento dos valores acima de um valor mínimo x_o determinando pela franquia do seguro.

A Figura 7.31 mostra exemplos da densidade Pareto. O gráfico à esquerda têm $x_o = 1$ e $\alpha = 1, 0.5, 0.1$. O gráfico à direita têm $x_o = 3$ e os mesmos valores para α que o gráfico anterior.

Notation 7.7. Se a v.a. X segue a distribuição Pareto com parâmetros $x_o > 0$ e $\alpha > 0$ escrevemos $X \sim \text{Pareto}(\alpha, x_o)$.

Example 7.13 Um exemplo interessante e antigo é a distribuição da renda anual de 2476 proprietários de terra na Inglaterra, em 1715. Esta distribuição pode ser vista a Figura 7.32 e corresponde à forma de uma densidade de Pareto. O gráfico foi extraído do livro clássico do estatístico inglês G. U. Yule (ver [24]).

Outro exemplo é a duração temporal de incêndios florestais no Norte e no Sul da África a partir de dados de monitoramento por satélite. O gráficos da direita mostram os dados das duas regiões com o ajuste de uma distribuição de Pareto.

A distribuição de Pareto é muito usada por seguradoras e reseguradoras para modelar as perdas financeiras que elas tem com as apólices. Quais os valores típicos de α na prática de seguros e

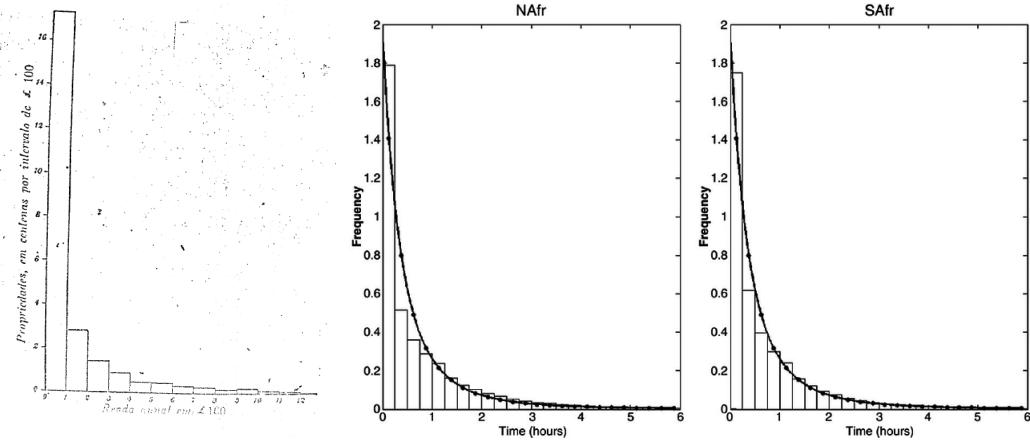


Figure 7.32: Número de propriedades de terra na Inglaterra em 1717. Ajuste de distribuição de Pareto a dados da duração temporal dos incêndios florestais no Sul e no Norte da África.

resseguros? A Swiss Re, a maior companhia européia de resseguros, fez um estudo. Nos casos de perdas associadas com incêndios, $\alpha \in (1, 2.5)$. Esta faixa pode ser mais detalhada: para incêndios em instalações industriais de maior porte, temos $\alpha \approx 1.2$. Para incêndios ocorrendo em pequenos negócios e serviços temos $\alpha \in (1.8, 2.5)$. No caso de perdas associadas com catástrofes naturais: $\alpha \approx 0.8$ para o caso de perdas decorrentes de terremotos; $\alpha \approx 1.3$ para furacões, tornados e vendavais.

A esperança de uma v.a. X com distribuição de Pareto(α, x_o) pode ser obtida de forma explícita pois integral de polinômios é fácil de se realizar. Quando $\alpha > 1$ podemos fazer as seguintes operações tomando cuidado com os vários sinais negativos na integral:

$$\mathbb{E}(X) = \int_0^\infty x C \frac{1}{x^{1+\alpha}} dx = \underbrace{\alpha x_o^\alpha}_C \left(\frac{1}{(1-\alpha)} x^{\alpha-1} \Big|_{x_o}^\infty \right) = \frac{\alpha}{\alpha-1} x_o$$

usando que $C = \alpha x_o^\alpha$.

A esperança $\mathbb{E}(X) = x_o \alpha / (\alpha - 1)$ só vale se $\alpha > 1$. Se $\alpha < 1$ a fórmula acima daria um valor negativo, o que nem faz sentido. O que acontece quando $0 < \alpha < 1$? Por exemplo, vamos olhar o caso $\alpha = 1/2$. Neste caso, a esperança-integral fica

$$\mathbb{E}(X) = \int_0^\infty x C \frac{1}{x^{1+\alpha}} dx = C \int_0^\infty \frac{1}{\sqrt{x}} dx = \infty$$

Como é conhecido de cursos de cálculo, esta integral diverge. Embora a curva $xf(x) = C/\sqrt{x}$ descreça com o aumento de x , ela o faz tão lentamente que a área abaixo da curva $xf(x) = C/\sqrt{x}$ cresce sem limites e a integral é ilimitada (ou infinita). Este é um caso matematicamente curioso que tem implicações práticas. Por exemplo, quando tivermos uma grande amostra e tiramos a sua média aritmética \bar{x} deveríamos ter $\bar{x} \approx \mathbb{E}(X)$. Mas o que podemos esperar quando $\alpha < 1$ e portanto $\mathbb{E}(X) = \infty$? Veja a simulação mais abaixo para um dica sobre o que acontece.

A função distribuição acumulada $\mathbb{F}(x)$ de uma v.a. Pareto(α, x_o) é facilmente obtida. Para $x > x_o$ temos

$$\mathbb{F}(x) = \int_{x_o}^x C \frac{1}{t^{1+\alpha}} dt = \underbrace{\alpha x_o^\alpha}_C \left(\frac{1}{-\alpha} t^\alpha \Big|_{x_o}^x \right) = 1 - \left(\frac{x_o}{x} \right)^\alpha \quad (7.5)$$

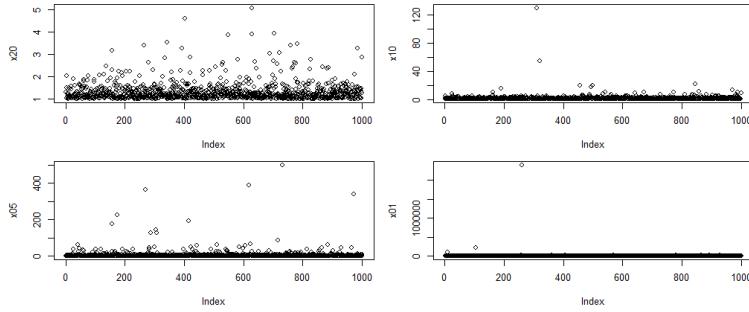


Figure 7.33: Amostras de 1000 valores Pareto com $x_o = 1$ e α igual a 4 (canto superior esquerdo), 2 (canto superior direito), 1 (canto inferior esquerdo) e 0.5 (canto inferior direito).

7.10.1 Simulando uma Pareto

No capítulo ?? veremos algumas das técnicas básicas para gerar amostra de uma v.a. A geração de v.a. com distribuição de Pareto com parâmetros $x_o > 0$ e $\alpha > 0$ é muito simples. Nós usamos o método da transformada inversa com a função 7.5 do seguinte modo: gere $U \sim U(0, 1)$ e então transforme este valor aleatório obtendo $X \sim x_o / (1 - U)^{-1/\alpha}$. Este valor aleatório X é uma v.a. com $\text{Pareto}(x_o, \alpha)$. Em R, basta usar

```
xo * (1-runif(n))^{(-1/alpha)}
```

para gerar n valores simulados. Os seguintes comandos foram usados para gerar mil valores desta distribuição:

```
par(mfrow=c(2, 2), mar=c(4, 4, 1, 1))
xo = 1; alpha = 4; x20 = xo * (1-runif(1000))^{(-1/alpha)}; plot(x20)
xo = 1; alpha = 2; x10 = xo * (1-runif(1000))^{(-1/alpha)}; plot(x10)
xo = 1; alpha = 1; x05 = xo * (1-runif(1000))^{(-1/alpha)}; plot(x05)
xo = 1; alpha = 0.5; x01 = xo * (1-runif(1000))^{(-1/alpha)}; plot(x01)
```

O resultado desses comandos está na Figura 7.33. Os valores da Pareto são lidos no eixo vertical. O eixo horizontal apenas indexa a ordem em que os 1000 valores foram gerados. O valor esperado $\mathbb{E}(X)$ das duas primeiras Pareto é igual a $4/3$ e 2. Este valor esperado não é um mau resumo do que acontece quando $\alpha = 4$. Com $\alpha = 2$ tivemos 7 valores próximo ou acima de 20, dez vezes maiores que $\mathbb{E}(X)$ portanto. Tivemos valores da ordem de 100, ou 50 vezes maiores que o seu valor esperado. Quando $\alpha \leq 1$, temos $\mathbb{E}(X) = \infty$ e esta presença de valores muito diferentes dos demais torna-se extrema. Com $\alpha = 1$, temos 80% dos valores menores que 5, mas 1% deles acima de 100 e 6 pontos acima de 20. Com $\alpha = 1/2$, a maioria dos valores se espalham numa faixa mais larga: 91% deles estão abaixo de 100. Entretanto, 3% são pelo menos 10 vezes maiores que o limite de 100, alcançando 1000 ou mais. Não para aí: 1% deles são maiores que 10 mil e 3 deles chegam a valores superiores a 100 mil. Compare com a faixa $(0, 100)$ onde encontram-se 91% deles.

7.10.2 Ajustando e visualizando uma Pareto

A presença de uma porção considerável de valores ordens de grandeza maiores que a maioria torna pouco útil o uso de histogramas como os da Figura 7.32. Tipicamente, histogramas de distribuições de Pareto, especialmente se o parâmetro α for menor que 2, serão similares àqueles da Figura 7.34 onde, mesmo após truncar brutalmente o eixo horizontal, não conseguimos visualizar adequadamente se os dados seguem uma Pareto. Na linha superior vemos os dados gerados de uma $\text{Pareto}(x_o = 1, \alpha = 2)$. Mostramos apenas os dados que são menores que 20, 10 e 5, sucessivamente. É difícil julgar se este decaimento é polinomial ou exponencial. Além disso, não

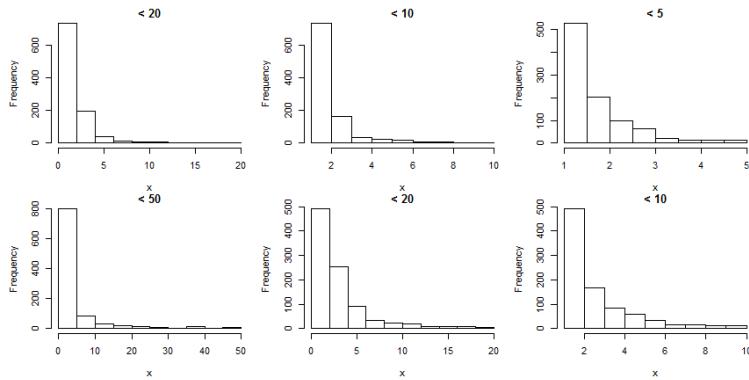


Figure 7.34: Amostras de 1000 valores Pareto com $x_o = 1$ e α igual a 4 (canto superior esquerdo), 2 (canto superior direito), 1 (canto inferior esquerdo) e 0.5 (canto inferior direito).

estamos olhando todos os dados mas apenas aqueles que não ultrapassaram um limiar, e portanto temos apenas uma informação parcial. Na linha inferior, com $\text{Pareto}(x_o = 1, \alpha = 1)$, esta situação se repete.

Uma maneira mais eficiente de visualizar a possível adequação do modelo Pareto é usando a função $\bar{F}(x)$. Trataremos disso no próximo capítulo.



8. Independência e Transformações de v.a.'s



9. Variância e Desigualdades

9.1 Variabilidade e desvio-padrão

Suponha que você vai gerar no computador valores aleatórios vindos de uma distribuição de probabilidade. Dizemos que simulamos no computador o experimento aleatório de gerar valores de uma distribuição de probabilidade. Como resumir grosseiramente esta longa lista de números antes mesmo de começar a gerá-los? O valor teórico em torno do qual eles vão variar é a esperança $\mathbb{E}(Y)$. Às vezes, teremos $Y > \mathbb{E}(Y)$, e às vezes, teremos $Y < \mathbb{E}(Y)$. Podemos esperar os valores gerados de oscilando Y em torno de $\mathbb{E}(Y)$. Mas até onde pode ir esta oscilação? Podemos ter situações em que os valores de Y oscilam muito pouco em torno de $\mathbb{E}(Y)$ e situações em que podem oscilar muito. No primeiro caso, $\mathbb{E}(Y)$ dará uma boa ideia dos valores aleatórios Y , todos muito próximos de $\mathbb{E}(Y)$. No segundo caso, $\mathbb{E}(Y)$ vai dar uma menos precisa dos valores Y já que eles podem se afastar muito de $\mathbb{E}(Y)$.

Para medir este grau de variabilidade de uma v.. em torno de seu valor esperado $\mathbb{E}(Y)$ usamos o *desvio-padrão*. Como o nome está dizendo, o desvio-padrão é o padrão para se medir desvios em relação ao valor esperado $\mathbb{E}(Y)$. O desvio-padrão é a régua, o metro que usamos para saber se uma v.a. oscila muito ou pouco em torno de seu valor esperado $\mathbb{E}(Y)$. Como no caso do valor esperado, o desvio-padrão é um valor teórico, deduzido a partir da distribuição de probabilidade (isto é, das duas listas) de uma v.a. Não é necessário nenhum dado estatístico para obter o desvio-padrão.

Vamos relembrar a definição de valor esperado $\mathbb{E}(Y)$ de uma v.a. no caso discreto,

$$\mathbb{E}(X) = \sum_{x_i} x_i \mathbb{P}(X = x_i)$$

e no caso contínuo,

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx$$

Antes de definir formalmente o desvio-padrão de uma v.a., vamos entender o conceito que queremos quantificar. Vamos chamar de desvio-padrão, e abreviar por DP, esta medida de variabilidade de uma v.a. em torno de seu valor esperado $\mathbb{E}(X)$. Se ele for definido de alguma forma razoável, quem deveria ter um maior DP, X e Y na Figura 9.1?

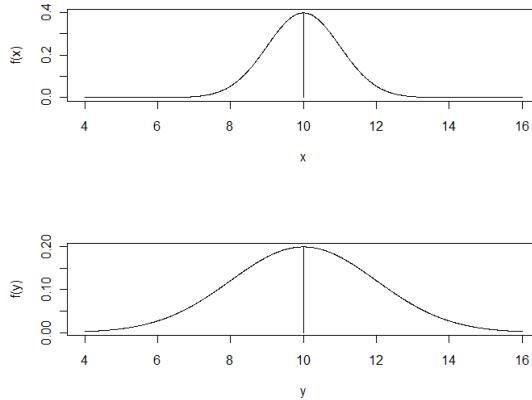


Figure 9.1: Densidades de X e Y com mesmo valor esperado: $\mathbb{E}(X) = \mathbb{E}(Y) = 10$.

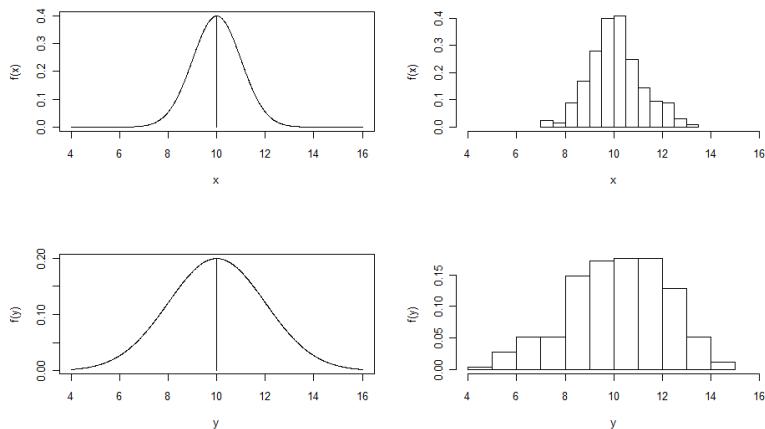


Figure 9.2: Histogramas de amostras e densidades de X e Y com mesmo valor esperado: $\mathbb{E}(X) = \mathbb{E}(Y) = 10$. Qual das amostras varia mais em torno do seu valor esperado?

Note que $\mathbb{E}(X) = \mathbb{E}(Y)$, ambos iguais a 10 e que a escala d eixo horizontal é a mesma para os dois gráficos, permitindo sua comparação visual. O DP é uma medida de variabilidade em torno do valor esperado. Os valores de X e Y vindos das densidades na Figura 9.1 vão se afastar or mais, ora menos em torno de seus valores esperados (que são iguais aqui). Qual das duas distribuições vai tender a se afastar mais de seu valor esperado?

Olhando para as densidades vemos que $f(y)$ espalha-se mais em torno de seu valor esperado $\mathbb{E}(Y) = 10$. De fato, a área abaixo de 8 ou acima de 12 é muito maior no caso da densidade $f(y)$ que nos caso da densidade $f(x)$. Isto quer dizer que valores distantes de 10 são gerados mais facilmente sob a densidade $f(y)$ do que sob a densidade $f(x)$. A Figura 9.2 mostra as densidades anteriores com histogramas de amostras geradas dessas mesmas densidades ao seu lado. Verificamos que amostras de Y tendem a se afastar mais de seu valor esperado $\mathbb{E}(Y)$ do que amostras de X . Podemos portanto esperar que, ao definirmos o desvio-padrão, devemos encontrar esta medida maior no caso Y do que no caso X .

Na Figura 9.1 colocamos as duas variáveis com o mesmo valor esperado $\mathbb{E}(X) = \mathbb{E}(Y) = 10$. Entretanto, isto não é necessário. Podemos medir a variabilidade de cada variável em torno de seu respectivo valor esperado, mesmo que $\mathbb{E}(X) \neq \mathbb{E}(Y)$. Por exemplo, a Figura 9.3 mostra duas

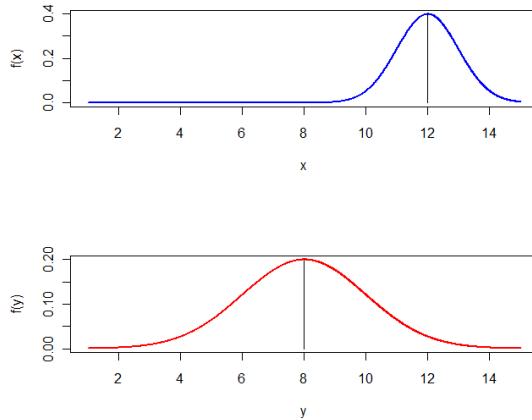


Figure 9.3: Densidades de X e Y com diferentes valores esperados: $\mathbb{E}(X) \neq \mathbb{E}(Y)$. Qual das amostras varia mais em torno do seu valor esperado?

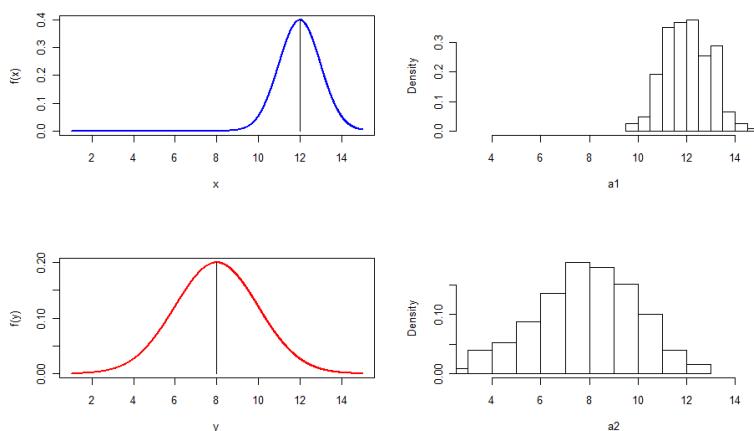


Figure 9.4: Histogramas de amostras e densidades de X e Y com diferentes valores esperados: $\mathbb{E}(X) \neq \mathbb{E}(Y)$. Qual das amostras varia mais em torno do seu valor esperado?

densidades, das v.a.'s X e Y , com diferentes valores esperados: $\mathbb{E}(X) = 12$ e $\mathbb{E}(Y) = 8$. Antes de dar a definição formal, queremos saber intuitivamente qual delas possui maior DP.

Novamente, olhando para as áreas sob as duas curvas densidade, vemos que tipicamente X fica entre 10 e 14, e portanto afastando-se tipicamente por menos que 2 unidades de seu valor esperado $\mathbb{E}(X) = 12$. Ao olharmos para a densidade de Y , vemos que afastamentos por mais de 2 unidades de seu valor esperado $\mathbb{E}(Y) = 8$ tem uma probabilidade substancial. Realmente, a área abaixo de 6 ou acima de 10 é uma fração considerável da área total (igual a 1). A Figura 9.4 mostra histogramas de amostras simuladas a partir das densidades da Figura 9.3 e vemos que valores de Y tendem a se afastar mais de seu valor esperado que valores da v.a. X .

Na Figuras anteriores estivemos usando densidades simétricas mas isto também não é necessário na definição do DP. A Figura 9.5 mostra as densidades de probabilidade $f(x)$ e $f(y)$ das variáveis aleatórias X e Y . Elas são assimétricas mas possuem o mesmo valor esperado: $\mathbb{E}(X) = \mathbb{E}(Y) = 1^1$. Este valor esperado é marcado pela linha vertical.

¹Estamos usando duas densidades gama aqui, com $\alpha = \beta$.

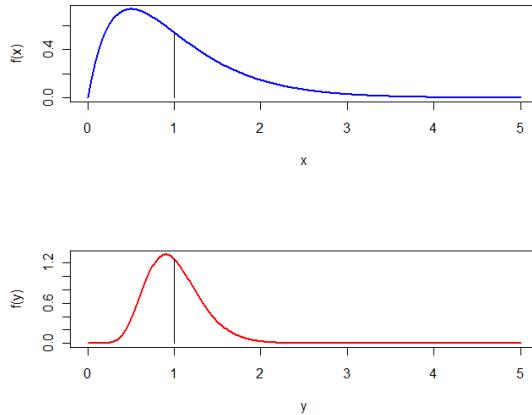


Figure 9.5: Densidades assimétricas de X e Y mas mesmo valor esperado: $\mathbb{E}(X) = \mathbb{E}(Y) = 1$.

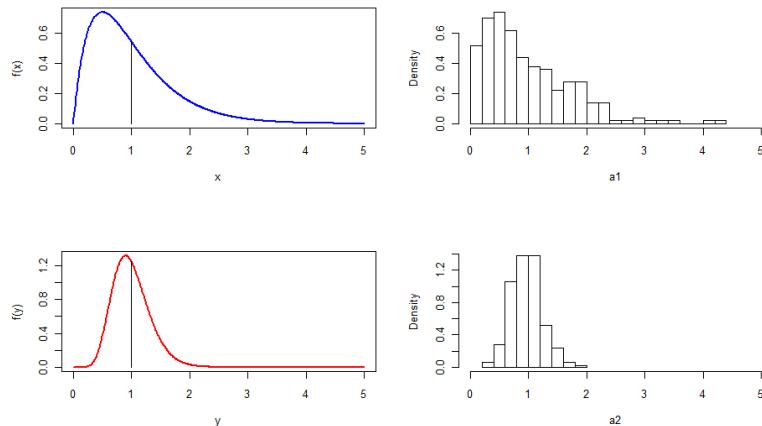


Figure 9.6: Histogramas de amostras e densidades assimétricas de X e Y com mesmo valor esperado: $\mathbb{E}(X) = \mathbb{E}(Y) = 1$. Qual das amostras varia mais em torno do seu valor esperado?

Novamente considerando que áreas sob a curva num intervalo indicam probabilidade de ocorrer um valor naquele intervalo, vemos que a distribuição de Y tem sua área total mais concentrada em torno de seu valor esperado. Isto é, X deve gerar mais facilmente valores que se afastam mais de seu valor esperado. Com as amostras de cada distribuição na Figura 9.6, vemos que esta intuição se confirma.

Na Figura 9.7 mostramos um caso em que X e Y possuem densidades assimétricas e têm $1 = \mathbb{E}(X) \neq \mathbb{E}(Y) = 3$. Considerando áreas sob as curvas, espera-se que Y tenha maior variação em torno de seu $\mathbb{E}(Y) = 3$ do que X em torno de seu respectivo valor esperado $\mathbb{E}(X) = 1$. As amostras no lado direito confirmam esta intuição.

A distribuições não precisam ser contínuas. A Figura 9.8 mostra as funções de probabilidade $\mathbb{P}(Y = y)$ e $\mathbb{P}(X = x)$ de duas variáveis discretas X e Y . Elas possuem diferentes valores esperados. Usamos duas v.a.'s de Poisson aqui, $X \sim \text{Poisson}(1.2)$ e $Y \sim \text{Poisson}(3.3)$. São relativamente maiores as barras de probabilidades alocadas a valores de y mais afastados de $\mathbb{E}(Y) = 3.3$ do que as barras de probabilidade alocadas x . Estas últimas tendem a estar bastante concentradas em torno de $\mathbb{E}(X) = 1.1$, indicando que valores da v.a. X tendem a se afastar pouco de seu valor

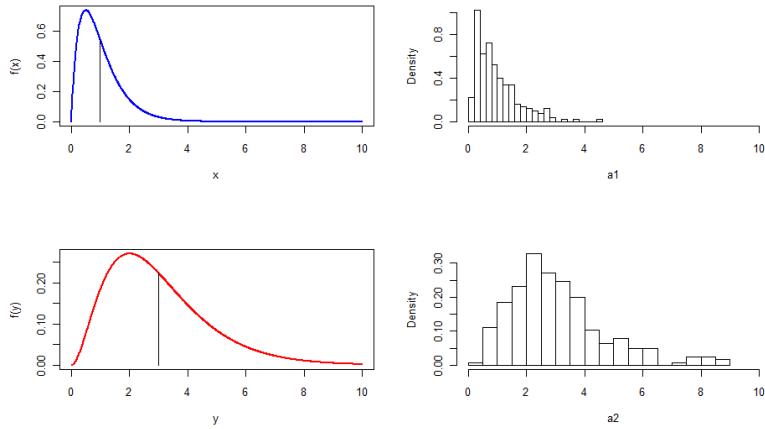


Figure 9.7: Histogramas e densidades assimétricas de X e Y com $1 = \mathbb{E}(X) \neq \mathbb{E}(Y) = 3$. Qual das amostras varia mais em torno do seu valor esperado?

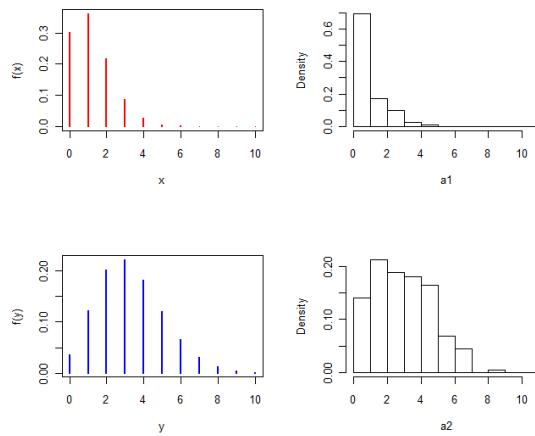


Figure 9.8: Histogramas e funções de probabilidade de duas Poisson com $1.2 = \mathbb{E}(X) \neq \mathbb{E}(Y) = 3.3$. Qual das amostras varia mais em torno do seu valor médio?

esparedo. É intuitivo que a variabilidade de Y em torno de $\mathbb{E}(Y)$ deve ser maior que a variabilidade de X em torno de $\mathbb{E}(X)$. Isto é confirmado com as amostras visualizadas ao lado das funccões de probabilidade $\mathbb{P}(Y = y)$ e $\mathbb{P}(X = x)$.

Você deve ter agora uma boa ideia do que queremos medir. Como então *definir* o DP de uma v.a.? Falta definir matematicamente esta noção intuitiva. Queremos medir o grau de variação da v.a. Y em torno de seu valor esperado $\mu = \mathbb{E}(Y)$. Podemos olhar para o *desvio* $Y - \mu$. As vezes, o desvio $Y - \mu$ é positivo, as vezes ele é negativo. Queremos ter uma ideia do tamanho do desvio e não de seu sinal. Vamos olhar então para o desvio absoluto $|Y - \mu|$.

Um ponto fundamental a ser observado é que $|Y - \mu|$ é uma variável aleatória. Vamos ter certeza de que este ponto está claro tendo uma visão empírica do desvio $|Y - \mu|$. Suponha que Y seja uma v.a. qualquer (discreta ou contínua) com $\mathbb{E}(Y) = \mu$. Simule Y várias vezes por Monte Carlo. Os valores aleatórios gerados sucessivamente vão variar em torno de μ . As vezes, eles serão apenas um pouco maiores ou menores que μ . As vezes, serão muito maiores ou muito menores que μ . Queremos ter uma ideia do tamanho do desvio $|Y - \mu|$. Mas como fazer isto se $|Y - \mu|$ é

aleatório?

A resposta é encontrada se pensarmos em termos abstratos. Como caracterizamos uma v.a. Y ? Fazemos isto fornecendo a sua densidade de probabilidade $f(y)$ (caso contínuo) ou sua função de probabilidade $\mathbb{P}(Y = y)$ (caso discreto). Isto é, fornecemos duas “listas”, a de valores possíveis (o suporte) e a de probabilidades associadas. Mas isto é muita coisa para fornecer para o desvio aleatório $|Y - \mu|$. Queremos uma forma mais econômica, mais simples, de resumir toda a distribuição do desvio aleatório $|Y - \mu|$. Será que não existe uma forma de ter apenas um único número resumindo toda a distribuição, mesmo que de forma grosseira.

Esta pergunta retórica tem uma resposta simples: sim, podemos usar o valor esperado do desvio absoluto de Y em torno de seu valor esperado $\mu = \mathbb{E}(Y)$. Isto é, podemos usar $E(|Y - \mu|)$ para representar de forma geral o tamanho do desvio. $E(|Y - \mu|)$ é o valor *esperado* do desvio de Y em torno de seu valor esperado μ .

Isto parece resolver nossa busca. Se $E(|Y - \mu|)$ for muito grande, então o desvio de Y tende a ser grande. Se $E(|Y - \mu|)$ for próximo de zero, então tipicamente Y varia pouco em torno de μ .

Entretanto, existe uma dificuldade com esta medida de variação. Cálculos matemáticos mais avançados com valor absoluto são muito difíceis. Em particular, a função $f(x) = |x|$ possui mínimo em $x = 0$, um ponto em que $f(x)$ não possui derivada (esboce o gráfico de $f(x)$ para ver isto). Assim, o mínimo de $f(x) = |x|$ não pode ser obtido derivando-se $f(x)$ e igualando a derivada a zero. Isto tem consequências de longo alcance em otimização. Em reusno, teremos problemas mais a frente se insistirmos em usar $E(|Y - \mu|)$ como definição da medida de variabilidade de uma v.a. A saída para este problema é calcularmos a variância $\sigma^2 = E(|Y - \mu|^2)$, que é mais fácil, e a seguir “corrigir” este cálculo tomando a sua raiz quadrada (o desvio-padrão).

Definition 9.1.1 — Variância e Desvio-padrão DP. Dada uma v.a. Y com valor esperado μ definimos a sua variância $\sigma^2 = E(|Y - \mu|^2)$ e o seu desvio padrão DP ou $\sigma = \sqrt{\sigma^2} = \sqrt{E(|Y - \mu|^2)}$.

■ **Notation 9.1 — Variância..** Escrevemos a variância $\sigma^2 = E(|Y - \mu|^2)$ como $\mathbb{V}(Y)$. Vamos escrever $DP(Y)$ para seu desvio-padrão.

O desvio-padrão $\sigma = \sqrt{E(|Y - \mu|^2)}$ usualmente é diferente da medida mais intuitiva $E(|Y - \mu|)$ mas eles costumam não ser muito diferentes. Assim, a interpretação do DP σ como sendo o tamanho esperado do desvio é aproximadamente correta.

Nos dois exemplos a seguir, vamos mostrar como calcular $\sigma^2 = E(|Y - \mu|^2)$ no caso discreto e no caso contínuo. Para compreender todo o cálculo, você precisa aprender a distribuição de transformações de v.a.’s, assunto do capítulo ???. Por enquanto, apenas aceite que os cálculos apresentados são válidos.

■ **Example 9.1 — Variância e DP, caso discreto.** Seja Y uma v.a. discreta com apenas 4 valores possíveis e probabilidades associadas:

y	1	2	3	4
$\mathbb{P}(Y = y)$	0.50	0.40	0.07	0.03

Temos

$$\mathbb{E}(Y) = \sum_{y=1}^4 y\mathbb{P}(Y = y) = 1 \times 0.50 + 2 \times 0.40 + 3 \times 0.07 + 4 \times 0.03 = 1.63$$

e portanto, usando $\mu = 1.63$, temos a variável aleatória do desvio $|Y - \mu| = |Y - 1.63|$ com suas duas listas, a de valores possíveis (o suporte) e a de probabilidades associadas. A lista de valores

possíveis é igual a $\mathcal{S} = \{|1 - 1.63|, |2 - 1.63|, |3 - 1.63|, |4 - 1.63|\} = \{|-0.63|, 0.37, 1.37, 2.37\}$. Vamos denotar por $d \in \mathcal{S}$ um elemento genérico do suporte do desvio aleatório $|Y - 1.63|$. As probabilidades associadas são imediatas pois, por exemplo, $\mathbb{P}(|Y - 1.63| = 1.37) = \mathbb{P}(Y = 3) = 0.07$. Portanto, a distribuição do desvio aleatório $|Y - 1.63|$ é dada por

d	0.63	0.37	1.37	2.37
$\mathbb{P}(Y - 1.63 = d)$	0.50	0.40	0.07	0.03

Finalmente, podemos calcular a variância de Y como o produto dos valores possíveis do desvio $|Y - 1.63|$ pelas suas probabilidades associadas:

$$\begin{aligned}\mathbb{V}(Y) &= E(|Y - \mu|^2) = E(|Y - 1.63|^2) \\ &= 0.63^2 \times 0.50 + 0.37^2 \times 0.40 + 1.37^2 \times 0.07 + 2.37^2 \times 0.03 = 0.55\end{aligned}$$

O desvio-padrão é igual a $DP(Y) = \sigma = \sqrt{0.55} = 0.74$. Assim, o desvio de Y em relação a seu esperado 1.63 é, em média, igual a 0.74. ■

■ **Example 9.2 — Variância e DP, caso contínuo.** Seja Y uma v.a. contínua com suporte $\mathcal{S} = (0, \infty)$ e densidade $f(y) = 3 \exp(-3y)$. Temos

$$\mathbb{E}(Y) = \int_0^\infty y f(y) dy = \int_0^\infty y 3e^{-3y} dy = \frac{1}{3}$$

O desvio quadrático é a variável aleatória $|Y - 1/3|^2$, que é contínua. Para obter a variância, precisamos calcular sua esperança $\mathbb{E}(|Y - 1/3|^2)$. Para isto, precisamos do seu suporte e densidade, assuntos que aprenderemos na parte de distribuição de transformações de v.a.'s, no capítulo ???. Entretanto, adiantando este assunto, podemos afirmar que a esperança $\mathbb{E}(|Y - 1/3|^2)$ pode ser obtida simplesmente multiplicando-se cada valor possível de $|Y - 1/3|^2$ pela densidade de $f(y)$:

$$\begin{aligned}\mathbb{V}(Y) &= \mathbb{E}(|Y - \mu|^2) = \mathbb{E}(|Y - 1/3|^2) \\ &= \int_0^\infty |y - 1/3|^2 f(y) dy \\ &= \int_0^\infty |y - 1/3|^2 3e^{-3y} dy \\ &= 1/3^2 \quad \text{após manipulações de cálculo.}\end{aligned}$$

Este exemplo é um caso particular da v.a. exponencial com densidade $f(y) = \lambda \exp(-\lambda y)$ onde λ é uma constante positiva (ver seção ??). Usamos acima o caso particular $\lambda = 3$. No caso geral, temos $\mathbb{E}(Y) = 1/\lambda$ e $\mathbb{V}(Y) = 1/\lambda^2$. Assim, no caso exponencial, $DP = \mathbb{E}(Y) = 1/\lambda$. ■

Os dois exemplos mostram como calcular a variância e o desvio-padrão na prática. Vamos apresentar estes resultados sob a forma de um teorema. A prova do teorema é uma consequência imediata dos resultados sobre transformações de v.a.'s na seção ???. Assim, vamos omitir a demonstração neste momento.

Theorem 9.1.1 — Cálculo de $\mathbb{V}(Y)$. Dada uma v.a. Y com suporte \mathcal{S} e valor esperado μ , a

variância $\mathbb{V}(Y) = \sigma^2 = E(|Y - \mu|^2)$ pode ser obtida da seguinte maneira:

$$\text{Caso discreto: } \sum_{y_i \in \mathcal{S}} (y_i - \mu)^2 \mathbb{P}(Y = y_i). \quad (9.1)$$

$$\text{Caso contínuo: } \int_{y \in \mathcal{S}} (y - \mu)^2 f(y) dy \quad (9.2)$$

9.2 Desigualdade de Tchebyshev

Como o nome está dizendo, o desvio-padrão é um padrão para medir desvios de uma v.a. Y (em torno do seu valor esperado). O DP é uma métrica universal, serve para qualquer v.a., discreta ou contínua. A desigualdade de Tchebyshev justifica esta universalidade do desvio-padrão. Ela diz que o desvio-padrão dá uma boa ideia do afastamento máximo que se pode esperar de uma v.a.

Para entender a desigualdade de Tchebyshev, considere o seguinte problema. Seja Y uma v.a. Y com valor esperado $\mathbb{E}(Y) = \mu$ e desvio-padrão σ . Se o desvio-padrão é uma métrica para medir desvios, e se σ é aproximadamente o valor esperado do desvio absoluto $|Y - \mu|$, não deveríamos observar um desvio $|Y - \mu|$ muito grande em termos de desvios-padrão. Por exemplo, poderíamos imaginar que deveria ser pequena a chance de observar um desvio $|Y - \mu|$ maior que 10 desvios-padrão. Isto é, a probabilidade de ocorrer o evento $|(Y - \mu)| > 10\sigma$ deveria ser pequena. Isto é realmente verdade? E quanto mudarmos o multiplicador para 100 ou para 3? Quanto é a probabilidade de vermos um desvio $|Y - \mu|$ maior que 2σ ? É possível dar uma resposta universal, que valha para toda e qualquer variável aleatória. A resposta surpreendente é sim e ela está no teorema de Tchebyshev (as vezes, escreve-se Chebyshev).

Theorem 9.2.1 — Desigualdade de Tchebyshev. Seja Y uma v.a. com $\mathbb{E}(Y) = \mu$ e desvio-padrão σ . Então

$$\mathbb{P}(|Y - \mu| > k\sigma) \leq 1/k^2.$$

Por exemplo, se tomarmos $k = 2$ então, para *qualquer* v.a., temos $\mathbb{P}(|Y - \mu| > 2\sigma) \leq 1/4$. A chance de que Y se desvie de seu valor esperado por mais que 2 desvios-padrão é menor que 0.25. Esta probabilidade pode ser bem menor que 0.25 no caso de certas distribuições mas o que podemos garantir é que, com certeza, ela nunca vai ultrapassar 0.25, qualquer que seja a distribuição de Y .

Para $k = 4$, a probabilidade se reduz a 0.06: $\mathbb{P}(|Y - \mu| > 4\sigma) \leq 1/16 = 0.06$. Assim, a chance de vermos um desvio maior do que 4 desvios-padrão é apenas 6% e isto vale para toda e qualquer v.a. O DP serve como uma métrica universal de desvios estatísticos: desviar-se por mais de 4 DPs de seu valor esperado μ pode ser considerado um evento um tanto raro.

Observe que a probabilidade decai com $1/k^2$. Nos primeiros inteiros temos uma queda rápida mas depois temos uma queda lenta:

k	2	4	6	10	20
$100\% \times \mathbb{P}$	25%	6%	3%	1%	0.3%

Vejamos agora a prova da desigualdade de Tchebyshev. Vamos considerar apenas o caso contínuo. O caso discreto é similar e é deixado como exercício. Seja $f(y)$ a densidade da v.a. Y com $\mathbb{E}(Y) = \mu$ e desvio-padrão σ . Queremos calcular $\mathbb{P}(|Y - \mu| > k\sigma)$. Como Y é uma v.a. contínua, esta probabilidade é a área sob a densidade $f(y)$ na região da reta que corresponde ao evento $|(Y - \mu)| > k\sigma$. Mas este evento ocorre significa que o valor $Y(\omega) = y$ da v.a. foi tal que y foi maior que $\mu + k\sigma$ ou foi menor que $\mu - k\sigma$. Isto é, o evento $|(Y - \mu)| > k\sigma$ é a união dos eventos

$[Y < \mu - k\sigma]$ e $[Y > \mu + k\sigma]$. Estes dois eventos são disjuntos que não existe que resultado ω tal que, ao mesmo tempo, tenhamos $Y(\omega) < \mu - k\sigma$ e $Y(\omega) > \mu + k\sigma$. Assim,

$$\mathbb{P}(|Y - \mu| > k\sigma) = \mathbb{P}([Y < \mu - k\sigma] \cup [Y > \mu + k\sigma]) \quad (9.3)$$

$$= \mathbb{P}([Y < \mu - k\sigma]) + \mathbb{P}([Y > \mu + k\sigma]) \quad (9.4)$$

$$= \int_{-\infty}^{\mu - k\sigma} f(y) dy + \int_{\mu + k\sigma}^{\infty} f(y) dy \quad (9.5)$$

Para $y \in (\mu + k\sigma, \infty)$ temos $1 < (y - \mu)/(k\sigma)$ ou ainda $1 = 1^2 < (y - \mu)^2/(k^2\sigma^2)$. Assim, podemos limitar a segunda integral acima por

$$\int_{\mu + k\sigma}^{\infty} 1 \times f(y) dy \leq \int_{\mu + k\sigma}^{\infty} \frac{(y - \mu)^2}{k^2\sigma^2} \times f(y) dy$$

De maneira análoga, podemos também limitar a primeira integral:

$$\int_{-\infty}^{\mu - k\sigma} 1 \times f(y) dy \leq \int_{-\infty}^{\mu - k\sigma} \frac{(y - \mu)^2}{k^2\sigma^2} \times f(y) dy$$

Assim, usando estes dois limites superiores para as duas integrais, temos

$$\mathbb{P}(|Y - \mu| > k\sigma) \leq \int_{-\infty}^{\mu - k\sigma} \frac{(y - \mu)^2}{k^2\sigma^2} \times f(y) dy + \int_{\mu + k\sigma}^{\infty} \frac{(y - \mu)^2}{k^2\sigma^2} \times f(y) dy$$

Como $(y - \mu)^2/(k^2\sigma^2) \geq 0$ para todo y na reta real, teremos

$$\int_{\mu - k\sigma}^{\mu + k\sigma} \frac{(y - \mu)^2}{k^2\sigma^2} \times f(y) dy \geq 0$$

e portanto

$$\begin{aligned} \mathbb{P}(|Y - \mu| > k\sigma) &\leq \int_{-\infty}^{\mu - k\sigma} \frac{(y - \mu)^2}{k^2\sigma^2} \times f(y) dy + \int_{\mu + k\sigma}^{\infty} \frac{(y - \mu)^2}{k^2\sigma^2} \times f(y) dy + \int_{\mu - k\sigma}^{\mu + k\sigma} \frac{(y - \mu)^2}{k^2\sigma^2} \times f(y) dy \\ &= \int_{-\infty}^{\infty} \frac{(y - \mu)^2}{k^2\sigma^2} \times f(y) dy \\ &= \frac{1}{k^2\sigma^2} \int_{-\infty}^{\infty} (y - \mu)^2 \times f(y) dy \\ &= \frac{1}{k^2\sigma^2} \sigma^2 = \frac{1}{k^2} \end{aligned}$$

Existem demonstrações mais curtas que esta mostrada acima mas elas usam outra desigualdade, a de Markov, que teria de ser demonstrada antes.

9.2.1 Opcional: A otimizalidade de Tchebyshev

Esta seção pode ser omitida sem prejuízo do restante do livro. A desigualdade de Tchebyshev é ótima, a melhor possível. Não conseguimos melhorar esta desigualdade. O sentido disso é seguinte. Suponha que exista uma função $g(k) \leq 1/k^2$ tal que a nova desigualdade

$$\mathbb{P}(|Y - \mu| > k\sigma) \leq g(k) \leq 1/k^2$$

valha para *toda* v.a. Y . Vamos mostrar que ela teremos de ter $g(k) = 1/k^2$, que é o limite da desigualdade de Tchebyshev.

Para isto, como a desigualdade tem de ser universal, valendo para toda v.a., vamos considerar uma v.a. particular. Fixe um inteiro positivo k qualquer. Seja Y a v.a. discreta com

$$Y = \begin{cases} -1, & \text{com probab } \frac{1}{2k^2} \\ 0, & \text{com probab } 1 - \frac{1}{k^2} \\ 1, & \text{com probab } \frac{1}{2k^2} \end{cases}$$

Ela tem $\mathbb{E}(Y) = 0$ e $DY = \sigma^2 = 1/k$. Então, para esta v.a., o evento $|Y - \mu| \geq k\sigma$ significa $|Y - 0| = |Y| \geq k/k = 1$. Como Y é no máximo igual a 1, então $|Y| \geq 1$ é equivalente a $Y = -1$ ou $Y = 1$. Sabemos que $\mathbb{P}(Y = -1) = 1/(2k^2)$ e que $\mathbb{P}(Y = 1) = 1/(2k^2)$. Assim,

$$\mathbb{P}(|Y - \mu| \geq k\sigma) = \mathbb{P}(|Y| \geq 1) = \mathbb{P}(Y = -1) + \mathbb{P}(Y = 1) = \frac{1}{2k^2} + \frac{1}{2k^2} = \frac{1}{k^2}.$$

Mas este é exatamente o limite dado pela desigualdade de Tchebyshev. Isto é, para esta v.a. a desigualdade de Tchebyshev transforma-se numa igualdade.

Como o limite $g(k) \leq 1/k^2$, que deveria ser melhor que aquele fornecido pela desigualdade de Tchebyshev, tem de valer para toda e qualquer v.a., ele teria de valer também para esta v.a. Y . Mas então vimos que $g(k)$ teria de igual a $1/k^2$.

Em resumo, valendo para *toda* v.a., não é possível obter uma cota mais apertada (menor) que $1/k^2$, que é aquela fornecida pela desigualdade de Tchebyshev.

9.2.2 Força e fraqueza de Tchebyshev

A força da desigualdade de Tchebyshev é a sua generalidade: ela vale para toda e qualquer v.a. A fraqueza da desigualdade de Tchebyshev é, bem, a sua generalidade. Para ser válida para toda e qualquer v.a., a desigualdade acaba não sendo muito “apertada”. Isto é, *quando consideramos apenas uma distribuição específica*, podemos obter cotas muito melhores que $1/k^2$ para a chance de ter um desvio grande.

Por exemplo, se $Y \sim N(\mu, \sigma^2)$ e $k = 2$, então sabemos que

$$\mathbb{P}(|Y - \mu| \geq 2\sigma) \approx 1/20 = 0.05$$

enquanto a desigualdade de Tchebyshev garante apenas que

$$\mathbb{P}(|Y - \mu| \geq 2\sigma) \leq 1/4$$

9.3 Outras desigualdades

A completar no futuro. Desigualdade de Hoeffding e de Mill.



10. Ajuste de distribuição

Neste capítulo, vamos aprender a verificar se uma amostra aleatória da v.a.'s i.i.d. X_1, X_2, \dots, X_n segue uma certa distribuição de probabilidade. Vamos aprender o teste de Kolmogorov e o teste qui-quadrado. O teste de Kolmogorov é aplicado quando o modelo é de uma v.a. contínua. O teste qui-quadrado pode ser aplicado a distribuições contínuas ou discretas mas exige uma categorização arbitrária.

10.1 Teste qui-quadrado

Em linhas gerais, o teste qui-quadrado funciona assim. Primeiro, particionamos a faixa de variação de uma v.a. Y em categorias. Por exemplo, podemos criar as categorias $[Y < -2]$, $[-2 \leq Y < 0]$, $[0 \leq Y < 2]$ e $[Y \geq 2]$. A seguir, contamos quantos elementos da amostra caem em cada categoria. Comparamos as contagens observadas com as que teoricamente deveriam cair na categoria de acordo com a distribuição de probabilidade sendo testada. Se a discrepância entre o observado e o esperado sob o modelo for grande, a distribuição de probabilidade sob teste é rejeitada. Se a discrepância for pequena, a distribuição é aceita como um modelo compatível com os dados. O teste possui duas vantagens principais: ele pode ser usado com distribuições contínuas ou discretas; e ele sabe como lidar com quantidades estimadas a partir dos dados (sobre isto, ver seção ??).

O teste qui-quadrado assume que os dados de uma amostra Y_1, Y_2, \dots, Y_n são instâncias i.i.d. que seguem uma certa distribuição de probabilidade (ou modelo teórico). O modelo teórico pode ser qualquer distribuição de probabilidade, contínua ou discreta. Vamos denotar esta distribuição teórica pela sua função de distribuição acumulada $\mathbb{F}(y)$. A vantagem dessa notação é que $\mathbb{F}(y)$ existe tanto para distribuições contínuas quanto discretas.

Nós vamos explicar o teste numa situação pouco prática, uma em que o modelo teórico é especificado completamente. Assim, inicialmente vamos assumir que $\mathbb{F}(y)$ poderia ser uma gaussiana $N(10, 4)$, e não uma $N(\mu, \sigma^2)$ com os parâmetros μ e σ^2 desconhecidos. Ter os parâmetros completamente especificados, conhecidos, é uma situação na prática. Ela pode acontecer quando, por exemplo, temos uma especificação técnica que estabelece que certos produtos sendo fabricados devem ter comprimento médio $\mu = 10$ e um desvio-padrão tolerável de $\sigma = \sqrt{4}$. Neste caso, vamos testar se a amostra conforma-se com estas especificações técnicas. Outra situação pode

ser o tempo de sobrevida com um novo medicamento. Suponha que, a partir de muitos dados acumulados nos últimos anos, sabe-se que o medicamento usual produz uma sobrevida aleatória que segue com uma distribuição exponencial $\exp(\lambda)$ e que o tempo esperado de sobrevida é conhecido (a partir dos muitos dados acumulados) e é igual a $1/\lambda = 24$ meses. Assim, a distribuição do tempo de sobrevida com o medicamento usual é uma $\exp(\lambda = 1/24)$. Com os dados da sobrevida de uns poucos pacientes tratados com o novo medicamento, queremos verificar se os tempos de vida continuam seguindo a mesma distribuição $\exp(\lambda = 1/24)$ ou se elas mostram que o novo medicamento mudou esta distribuição (e para melhor, deseja-se).

Dessa forma, nesta parte inicial do capítulo, vamos supor que a distribuição de interesse é completamente especificada, ela não tem parâmetros desconhecidos. Ela poderia ser $N(10, 4)$, mas não uma $N(\mu, \sigma^2)$; poderia ser uma $\text{Bin}(20, 0.1)$, e não uma $\text{Bin}(20, \theta)$ com a probabilidade θ desconhecida; poderia ser uma $\text{Poisson}(5)$, mas não uma $\text{Poisson}(\lambda)$ com λ desconhecido. Na seção ?? vamos discutir como atacar o problema mais geral em que a forma da distribuição é especificada mas não seus parâmetros, que são tratados como desconhecidos.

A pergunta de interesse do teste qui-quadrado é: a amostra Y_1, Y_2, \dots, Y_n é composta de v.a.'s i.i.d. seguindo o modelo teórico $\mathbb{F}(y)$? O passo 1 do teste qui-quadrado é particionar o conjunto de valores possíveis de Y em N categorias (ou intervalos). Por exemplo:

- Se o modelo teórico é uma $\text{Bin}(20, 0.1)$, podemos criar 5 categorias de valores possíveis: $Y = 0, Y = 1, Y = 2, Y = 3$ e $Y \geq 4$.
- Se o modelo é uma $\text{Poisson}(5)$, podemos criar 12 categorias: $Y = 0, Y = 1, \dots, Y = 10$ e $Y \geq 11$.
- Se o modelo é uma $\exp(10)$, podemos criar 5 categorias-intervalos: $[0, 0.05), [0.05, 0.1), [0.1, 0.2), [0.2, 0.4), [0.4, \infty)$
- Se o modelo é uma $N(0, 1)$, podemos criar 4 categorias-intervalos: $(-\infty, -2), [-2, -1), [-1, 0), [0, 1), [1, 2), e (2, \infty)$.

Em princípio, estes intervalos-categorias $[a, b)$ são arbitrários mas, na prática, nós os escolhemos de forma que não tenham nem probabilidades $\mathbb{P}(Y_i \in [a, b))$ muito altas, nem muito baixas.

O passo 2 do teste qui-quadrado é calcular o número de elementos da amostra Y_1, Y_2, \dots, Y_n que caem em cada intervalo-categoria. Vamos denotar por N_k o número de Y_i 's que caem no intervalo k . N_k é chamada de *frequência observada* na amostra.

Calcule também E_k , o número *esperado* de observações que deveriam cair no intervalo k . Isto é, calcule $E_k = n \times P(Y \in \text{Intervalo } k)$. Antes de justificar esta fórmula, vamos ver uns exemplos.

Suponha que o modelo teórico é uma $\text{Bin}(20, 0.1)$, que temos amostra de tamanho $n = 53$ e que a categoria é $Y = 0$. Então o número esperado é $E = 53 * \mathbb{P}(Y = 0) = 53 * (1 - 0.1)^{20} = 6.44$. Se observamos 53 repetições de uma $\text{Bin}(20, 0.1)$ esperamos que 6.44 delas sejam iguais a zero.

Outro exemplo: o modelo teórico é uma $\text{Poisson}(2)$. Temos amostra de tamanho $n = 97$. A categoria é $Y \geq 4$. Então o número esperado nesta categoria é

$$E = 97 \times \mathbb{P}(Y \geq 4) = 97 \times (1 - \mathbb{P}(Y \leq 3)) = 97 \times \left(1 - \sum_{j=0}^3 \frac{2^j \exp(-2)}{j!}\right) = 13.86$$

Se observamos 97 repetições independentes de uma $\text{Poisson}(2)$, esperamos que 13.86 delas sejam maiores ou iguais a 4.

Mais um exemplo, agora com uma distribuição contínua. O modelo teórico é uma $\exp(10)$. Temos uma amostra de tamanho $n = 147$. O intervalo-categoria é $X \in [0.2, 0.4)$. Então o número esperado de observações neste intervalo é

$$E = 147 \times \int_{0.2}^{0.4} 10 \exp(-10x) dx = 17.20$$

Repete-se o cálculo nos demais intervalos.

A justificativa para esta forma de obter os números esperados E_k no intervalo-categoría $[a_k, b_k]$ é simples. Para cada elemento i da amostra de tamanho n , defina uma variável aleatória indicadora I_i (um ensaio de Bernoulli) tal que $I_i = 1$ (um “sucesso”) se $Y_i \in [a_k, b_k]$, e $I_i = 0$ se $Y_i \notin [a_k, b_k]$. Então $N_k = \sum_i I_i$ é o número de “sucessos” dentre estes n ensaios de Bernoulli. Como os Y_i são independentes, as indicadoras I_i são ensaios de Bernoulli independentes. Além disso, a probabilidade de sucesso em cada um deles permanece constante e igual a $p_k = \mathbb{P}(I_i = 1) = \mathbb{P}(Y_i \in [a_k, b_k])$. Assim, N_k segue uma distribuição binomial $\text{Bin}(n, p_k)$ e portanto, $\mathbb{E}(N_k) = np_k$.

O passo 3 do teste qui-quadrado é comparar frequências observadas N_k e as frequências esperadas E_k . E_k é o valor esperado da contagem N_k caso o modelo teórico seja verdadeiro. A ideia intuitiva é que, caso E_k e N_k sejam muito diferentes, teremos uma evidência de que o modelo teórico não é próximo da realidade. Caso E_k e N_k sejam parecidos, teremos uma evidência de que o modelo é capaz de produzir valores parecidos com os observados.

Se E_k e N_k forem parecidos, isto quer dizer que os dados observados REALMENTE sigam o modelo teórico? Não. Existem pelo menos três razões para este não:

1. Suponha que temos uma única amostra e dois (ou mais) modelos diferentes: os valores teóricos dos dois modelos podem estar bem próximos dos valores observados e não termos nenhum deles claramente melhor (mais próximo) que o outro.
2. Este aspecto do modelo (as contagens nos intervalos) é próximo da realidade. Outros aspectos do modelo, quando comparados com a realidade, podem mostrar que o modelo não é adequado. Por exemplo, uma análise de resíduos de um modelo (um assunto futuro neste livro) pode mostrar alguns problemas que não são aparentes na comparação entre E_k e N_k .
3. Finalmente, ninguém acredita que a realidade siga fielmente uma fórmula matemática perfeita. Precisamos apenas que a fórmula seja uma boa aproximação para a realidade.

Ainda considerando o passo 3, como então comparar as frequências observadas N_k e as frequências esperadas E_k ? Podemos ter uma boa aproximação numa categoria-intervalo mas uma péssima aproximação em outra categoria-intervalo. Assim, precisamos de um resumo, uma idéia global de como é a aproximação em geral, considerando todas as categorias. A medida-resumo é uma espécie de “média” das diferenças $|N_k - E_k|$. Note a presença do valor absoluto $|N_k - E_k|$ ao invés das diferenças $N_k - E_k$. Se a medida-resumo for pequena, então $N_k \approx E_k$ e adotamos o modelo teórico. Se a medida-resumo for grande, vamos precisar adotar outro modelo teórico para os dados.

■ Example 10.1 — Bombas em Londres. Considere o exemplo das bombas em Londres visto no capítulo ???. Temos 576 quadrados com a contagem em cada um deles. O modelo para estas 576 contagens é uma $\text{Poisson}(\lambda)$ com $\lambda = 0.9323$. Este valor de λ foi obtido a partir dos dados, como explicamos no capítulo ??.

Particione o conjunto de valores possíveis em intervalos: $Y = 0$, $Y = 1, \dots, Y = 5$, e $Y \geq 6$. Calcule N_k , E_k e a diferença $N_k - E_k$ para cada intervalo.

k	0	1	2	3	4	5 e acima
N_k	229	211	93	35	7	1
E_k	226.74	211.39	98.54	30.62	7.14	1.5
$N_k - E_k$	2.26	-0.39	-5.54	4.38	-0.14	-0.50

Nesta tabela, temos $E_k = 576 \times \mathbb{P}(Y = k) = 576 \frac{0.9323^k}{k!} e^{-0.9323}$ para $k = 0, \dots, 4$. Para a última categoria, calculamos $\mathbb{P}(Y \geq 5) = 1 - \mathbb{P}(Y \leq 4) = 1 - \sum_{j=0}^4 \mathbb{P}(Y = j)$. A medida-resumo sugerida antes é a média das diferenças (em valor absoluto):

$$\frac{1}{6} \sum_{k=0}^5 |N_k - E_k|$$

Entretanto, como argumentamos a seguir, esta não é uma boa idéia de como resumir a discrepância. Vamos ver porque. ■

Imagine um problema em que temos apenas três categorias com as seguintes diferenças $|N_k - E_k|$: 11.5, 10.6 e 0.9. Estas diferenças são grandes ou pequenas? Bem, depende... Depende do quê? Do valor esperado nessas categorias. Considere duas possíveis situações com apenas estas três categorias. Vamos diferenciar a segunda situação usando um asterisco nas variáveis:

k	0	1	2
N_k	20	1	6
E_k	8.5	11.6	6.9
$ N_k - E_k $	11.5	10.6	0.9
N_k^*	1020	1001	1006
E_k^*	1008.5	1011.6	1006.9
$ N_k^* - E_k^* $	11.5	10.6	0.9

As diferenças são *idênticas* mas, relativamente ao que esperamos contar em cada categoria, as diferenças são muito menores na segunda situação. Quando esperamos contar 11.6 numa categoria e observamos apenas 1, erramos por 10.6 e este erro parece grande. Mas quando esperamos 1011.6 e observamos 1001 o erro parece pequeno mesmo que a diferença absoluta seja a mesma de antes. Parece razoável considerarmos as diferenças $|N_k - E_k|$ maiores (em algum sentido) do que as diferenças $|N_k^* - E_k^*|$.

Assim, uma medida-resumo de comparação mais apropriada seja então a média das diferenças relativas ao esperado em cada categoria. Isto é, com N categorias ao todo, um candidato a medida-resumo seria:

$$\frac{1}{N} \sum_k \frac{|N_k - E_k|}{E_k}$$

Karl Pearson (1857-1936) estudou esta medida e achou que, embora intuitiva e simples, ela não era matematicamente manejável. A razão é que o comportamento dessa média-resumo dependia de aspectos específicos do problema sendo analisado. Ele dependia do tamanho da amostra, da distribuição particular sob estudo (binomial, Poisson, exponencial, etc). Num toque de gênio, ele propôs uma medida-resumo diferente.

10.2 A estatística Qui-quadrado

A medida-resumo de Pearson calcule N_k , E_k e a diferença $N_k - E_k$ para cada intervalo-categoria. Ao invés de calcular

$$\frac{1}{N} \sum_k \frac{|N_k - E_k|}{E_k},$$

ele calcula a estatística qui-quadrado de Pearson:

$$\chi^2 = \sum_k \frac{(N_k - E_k)^2}{E_k}$$

A estatística usa a letra grega χ (pronuncia-se “qui”), e não a letra “X”.

No caso das bombas em Londres, temos

$$\chi^2 = \frac{(2.26)^2}{226.74} + \frac{(-0.39)^2}{211.39} + \dots + \frac{(-0.50)^2}{1.5} = 1.13$$

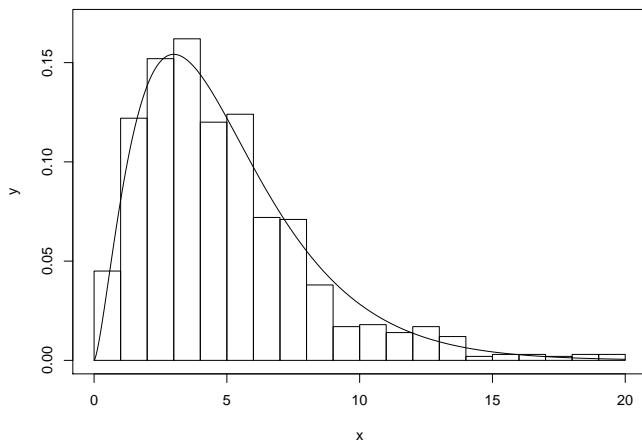


Figure 10.1: Histograma de 1000 simulações e densidade de Qui-quadrado com 5 g.l. superimposta

Como saber se a discrepância entre N_k e E_k refletida em χ^2 é grande ou pequena? A resposta precisou do gênio de Pearson e, ao mesmo tempo, ela justifica por que usamos esta medida-resumo particular e pouco intuitiva.

10.2.1 A distribuição de χ^2

Considerando o nosso problema das bombas em Londres como ilustração, vamos entender o que Pearson se perguntou. Suponha que o modelo teórico Poisson(λ) seja realmente verdadeiro. Suponha que as contagens das bombas nos quadrados realmente sejam independentes e sigam uma distribuição Poisson(λ) com $\lambda = 0.9323$. Mesmo neste caso, χ^2 nunca será exatamente igual a zero. Dependendo da amostra, ele pode ser pequenino e próximo de zero ou pode ser um pouco maior. Não deve ser um valor muito muito grande pois o modelo é verdadeiro e portanto N_k deveria estar próximo de E_k . Mas ele não será exatamente zero. A pergunta que Pearson se fez é qual é a variação natural de χ^2 quando o modelo teórico é verdadeiro? Vamos responder isto com um experimento no R.

Execute o seguinte algoritmo em R:

- Crie um vetor E de dimensão 6 com as contagens esperadas de $X = 0, X = 1, \dots, X = 4$, e $X \geq 5$ em 576 Poisson(0.9323). Isto é, $E = c(226.74, 211.39, 98.54, 30.62, 7.14, 1.5)$.
- Crie um vetor Qui com 1000 posições.
- for(i in 1:1000) faça:
 - Gere X_1, \dots, X_{576} iid Poisson($\lambda = 0.9323$)
 - Conte o número N_k de X_i 's iguais a $0, 1, \dots, 4, \geq 5$
 - Faça $Qui[i] \leftarrow X^2 = \sum_k (N_k - E_k)^2 / E_k$
- Faça um histograma dos 1000 valores gerados do vetor Qui.

O resultado deste experimento está na Figura 10.1. O histograma mostra a variabilidade que se pode esperar da estatística χ^2 quando o modelo é verdadeiro. Isto é, a estatística χ^2 é uma variável aleatória com uma lista de valores possíveis (o eixo positivo $(0, \infty)$) e probabilidades associadas. O histograma dá uma ideia de quais regiões do eixo $(0, \infty)$ são mais prováveis equais são menos prováveis.

O que Karl Pearson descobriu matematicamente é que a distribuição de χ^2 era (aproximadamente) a mesma qualquer que fosse a distribuição do modelo teórico (Poisson, normal, gama, ou qualquer outro modelo para os dados). A distribuição de χ^2 é uma distribuição universal. Não

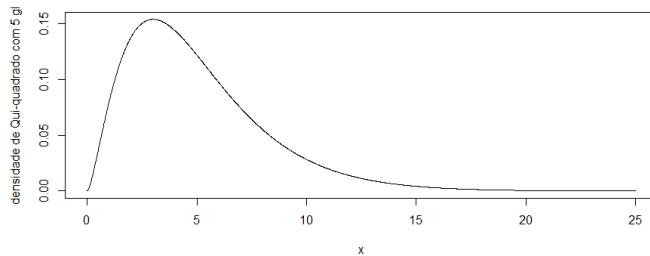


Figure 10.2: Densidade de uma distribuição qui-quadrado com $v = 5$ g.l.

importa o problema, a forma de medir a discrepância via χ^2 comporta-se estatisticamente do mesmo modo. Se isto é assim, poderemos saber quando uma discrepância é excessiva ou quando ela é pequena em todo e qualquer problema. E a forma de saber se a discrepância χ^2 é grande ou pequena é a mesma, sempre. Karl Pearson conseguiu uma fita métrica que mede desvios dos dados observados em relação a qualquer modelo teórico. É um resultado fantástico: qualquer que seja a distribuição $F(x)$ do modelo teórico, se ele realmente estiver gerando os dados, então a distribuição da estatística χ^2 é uma só: uma distribuição chamada qui-quadrado. Não precisamos fazer nenhuma simulação Monte Carlo para encontrar quais os valores razoáveis para χ^2 quando o modelo teórico for correto.

Vamos colocar um pouco mais de rigor e menos entusiasmo aqui. A distribuição de χ^2 não é *exatamente* igual a uma distribuição qui-quadrado (com a densidade que acabei de mostrar no gráfico da Figura 10.1). Ela é *aproximadamente* igual a uma qui-quadrado quando o tamanho da amostra é grande. O que é uma amostra grande? A resposta pode depender do problema. No caso Poisson, com $\lambda \approx 1$, basta ter $n > 200$ para obtermos uma boa aproximação. Com λ 's maiores, $n = 100$ já é suficiente. O fato é que com amostras não muito grandes já podemos usar a aproximação qui-quadrado.

Outro ponto que precisa de esclarecimento. A distribuição qui-quadrado não é uma só. Ela possui um parâmetro chamado de *número de graus de liberdade*. Este parâmetro é igual ao número de categorias-intervalos menos 1 e menos p , onde p é o número de parâmetros do modelo teórico $F(x)$ que precisaram ser estimados. Nesta parte inicial do texto supomos que todos os parâmetros do modelo teórico $F(x)$ são conhecidos. Portanto, o número de graus de liberdade é apenas o número de categorias-intervalos em χ^2 menos 1. No caso das bombas de Londres, por exemplo, se $\lambda = 0.9323$ for conhecido, o número de graus de liberdade é 6, o número de categorias, menos 1. Então, o número de graus de liberdade é $6 - 1 = 5$.

10.3 Como usar este resultado de Pearson? O p-valor

O valor de χ^2 é aleatório, não pode ser previsto de forma determinística. Ele varia de amostra para amostra, mesmo que o modelo probabilístico que gera os dados siga sendo o mesmo. Entretanto, quando o modelo teórico $F(x)$ é verdadeiro (de fato, está gerando os dados observados), então χ^2 varia aproximadamente como uma distribuição qui-quadrado com v graus de liberdade onde v é ao número de categorias menos 1, caso não existam parâmetros desconhecidos em $F(x)$. No caso das bombas em Londres, temos $v = 6 - 1 = 5$ graus de liberdade (abreviado como g.l. daqui por diante).

Quais são os valores típicos de uma qui-quadrado com $v = 5$ g.l.? E quais são os valores não-típicos, os valores que dificilmente viriam de uma qui-quadrado com 5 g.l.? A Figura 10.2 mostra a densidade de probabilidade de uma distribuição qui-quadrado com $v = 5$ g.l.

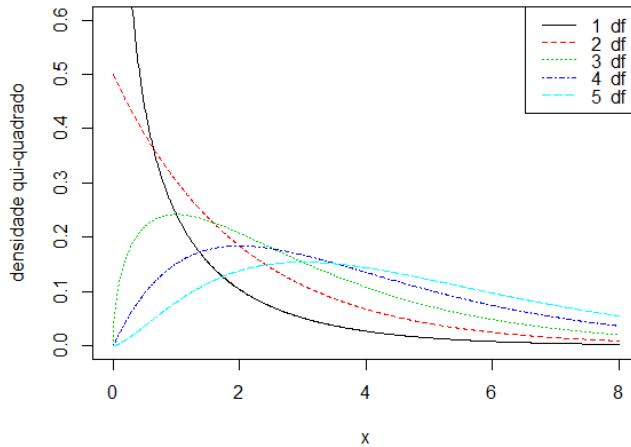


Figure 10.3: Densidades da distribuição qui-quadrado com $k = 1, 2, 3, 4, 5$

Na verdade, a distribuição qui-quadrado com v graus de liberdade é um caso particular da distribuição gama (ver seção ?? no capítulo ??). Uma qui-quadrado com v graus de liberdade é o mesmo que uma Gama($v/2, 1/2$), que possui densidade de probabilidade dada por

$$f(x) = \frac{1}{2^{v/2}\Gamma(k/2)} x^{v/2-1} e^{-x/2} = (\text{cte.}) x^{v/2-1} e^{-x/2}$$

Assim, a densidade da qui-quadrado é o produto de $x^{v/2-1}$, uma potência crescente de x , por um decrescimento exponencialmente rápido em x , $e^{-x/2}$. Mesmo com um número v de graus de liberdade grande, o decrescimento exponencial eventualmente domina o produto de modo que $x^{v/2-1} e^{-x/2} \rightarrow 0$ quando $x \rightarrow \infty$. A Figura 10.3 mostra as densidades $f(x)$ com diferentes valores para os graus de liberdade v .

Mas, como usar o teste qui-quadrado? Vamos voltar ao caso das bombas de Londres. Os valores típicos de uma χ^2 com $v = 5$ graus de liberdade estão na Figura 10.2. Eles são aqueles entre 0 e 10. Os valores entre 10 e 15 são mais raros, tendo uma probabilidade $\mathbb{P}(\chi^2 \in (10, 15)) \approx 0.064$, obtida com o comando `pchisq(15, 5) - pchisq(10, 5)`. Os valores acima de 15 são possíveis mas bastante improváveis. Eles ocorrem com probabilidade $\mathbb{P}(\chi^2 > 15) \approx 0.01$, obtida com o comando `1-pchisq(15, 5)`.

Calcule o valor realizado de χ^2 usando os dados da amostra. Este valor realizado é um número real positivo. Por exemplo, no caso das bombas em Londres, tivemos $\chi^2 = 1.13$ com $v = 5$ g.l. O valor 1.13 é um valor típico de uma qui-quadrado com 5 g.l.? Se for, a discrepância medida pela estatística χ^2 é pequena. Se for atípico e grande, a discrepância dificilmente poderia ser produzida se o modelo teórico for o verdadeiro gerador dos dados. A Figura 10.4 mostra a densidade de uma qui-quadrado com 5 g.l. com o valor observado 1.13 da estatística χ^2 .

É óbvio que 1.13 é um valor típico de uma qui-quadrado com 5 g.l. Ele está bem no meio da faixa de variação razoável dos valores dessa distribuição. Isto é sinal de que as diferenças entre as contagens observadas na amostra e as contagens esperadas pelo modelo são aquelas que se espera quando o modelo é o verdadeiro. Uma forma de expressar quão discrepante é o valor observado de χ^2 é calcular a probabilidade de observar uma v.a. qui-quadrado com 5 g.l. maior ou igual a 1.13. Esta probabilidade é chamada de *p-valor* e, no caso das bombas em Londres, ela é igual a 0.95, obtido com o comando `1-pchisq(1.13, 5)`.

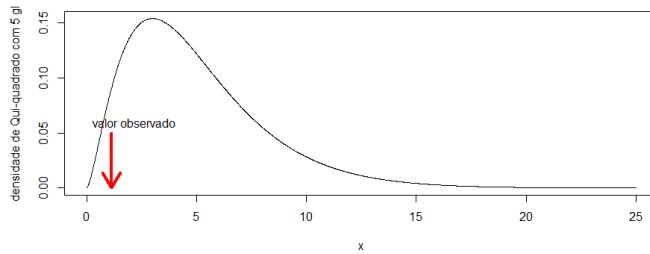


Figure 10.4: 1.13 é o valor observado de χ^2 no caso das bombas em Londres. Gráfico da densidade de uma qui-quadrado com 5 g.l.

O p-valor é a área da densidade da qui-quadrado com 5 g.l. que está acima do valor 1.13 observado com a amostra. Um p-valor próximo de zero é sinal de que o modelo não se ajusta bem aos dados. Não foi este o caso aqui.

10.4 Ajustando os graus de liberdade

Estivemos até agora supondo que o modelo teórico é completamente especificado, que não existem parâmetros desconhecidos. Com o que aprendemos até o momento, podemos verificar se uma $N(10, 4)$ ajusta-se aos dados, mas não podemos checar se uma $N(\mu, \sigma^2)$ ajusta-se aos dados, procurando no processo também estimar os parâmetros μ e σ^2 . A razão é que a distribuição da estatística qui-quadrado é afetada quando estimamos parâmetros para obter os números esperados E_k nas categorias-intervalos. Não poderemos mostrar isto neste livro. Este não é um fato óbvio, nem facilmente perceptível. De fato, o próprio Karl Pearson ignorava este problema. Ele acreditava que a distribuição da estatística qui-quadrado não era afetada ao usarmos parâmetros estimados. Ronald Fisher, outro gigante da estatística, ainda muito jovem e começando sua carreira, corrigiu este erro do velho Pearson e isto teve consequências negativas para sua carreira pois Karl Pearson não aceitou de bom grado esta correção. Assim são os aspectos humanos na ciência.

Felizmente, na maioria dos casos, a correção é muito simples. O número correto de graus de liberdade v é o número de categorias menos 1 e menos o número de parâmetros estimados com os dados. Vamos explicar com um exemplo clássico e curioso.

Ladislaus Josephovich von Bortkiewicz (1868-1931) foi um economista, estatístico e matemático que trabalhou em Berlim na época do Império Prussiano. Em 1898 ele publicou um livro, *Das Gesetz der kleinen Zahlen*, que significa A Lei dos Pequenos Números. Nele, apresentou vários estudos usando a distribuição de Poisson. Um deles ficou famoso. Ele obteve o número de soldados mortos por coices de cavalo em certas corporações do exército prussiano durante vinte anos (1875-1894). Em cada um dos 20 anos, ele anotou o número de mortos em cada uma de 10 corporações. Temos $20 \times 10 = 200$ contagens N_k mostradas na segunda coluna da tabela abaixo. Na primeira coluna, temos o número k de mortos na corporação-ano.

k	N_k	E_k
0	109	108.7
1	65	66.3
2	22	20.2
3	3	4.1
≥ 4	1	0.7
Total	200	200

A terceira coluna mostra o número esperado E_k se estas 200 contagens forem instâncias



Figure 10.5: Os dados coletados por von Bortkiewicz incluíram o número de mortes por coices de cavalos em corporações do exército prussiano no século XIX. Essas mortes seguem uma distribuição de Poisson.

i.i.d. de uma mesma v.a. de Poisson. Isto é, para $k = 0, 1, 2, 3$, temos $E_k = 200\mathbb{P}(X = k)$ onde $X \sim \text{Poisson}(\lambda)$. Para fazer este cálculo, precisamos do valor do parâmetro λ já que $\mathbb{P}(X = k) = \lambda^k/k! \exp(-\lambda)$. Como $\mathbb{E}(X) = \lambda$ no caso de uma Poisson, usamos os próprios dados para encontrar um valor para λ . A média aritmética das contagens será um valor próximo de $\mathbb{E}(X)$ qualquer que seja o modelo. Assim, uma estimativa para λ é $\hat{\lambda}$ dado por

$$\hat{\lambda} = \frac{1}{200} (0 \times 109 + 1 \times 65 + 2 \times 22 + 3 \times 3 + 1 \times 4) = 0.61.$$

Por exemplo, para a primeira categoria, temos $E_0 = 200 \times \mathbb{P}(X = 0) = 200\exp(-0.61) = 108.6702$. A última categoria teve uma corporação-ano com exatamente 4 mortes. O seu valor esperado é obtido fazendo-se

$$E_4 = 200 \times \mathbb{P}(X \geq 4) = 200 \times \mathbb{P}(X \leq 3) = 200 \times (\mathbb{P}(X = 0) + \mathbb{P}(X = 3))$$

Os valores de N_k e E_k são muito próximos e seria uma surpresa se o modelo de Poisson fosse rejeitado pelo teste qui-quadrado. A estatística qui-quadrado é igual a

$$\chi^2 = \frac{(109 - 108.7)^2}{108.7} + \frac{(65 - 66.3)^2}{66.3} + \frac{(22 - 20.2)^2}{20.2} + \frac{(3 - 4.1)^2}{4.1} + \frac{(1 - 0.7)^2}{0.7} = 0.61$$

Um parâmetro desconhecido teve ser estimado. Portanto, se o modelo Poisson é adequado, a estatística qui-quadrado seguiria uma distribuição qui-quadrado com $v = 5 - 1 - 1 = 3$ graus de liberdade. A área acima do valor $\chi^2 = 0.61$ numa densidade qui-quadrado com 3 graus de liberdade é obtida no R com `1 - pchisq(0.61, 3)` resultando em 0.894. Ver Figura 10.6.

10.5 Teste quando a v.a. é contínua

Quando a distribuição é contínua, além de um teste qui-quadrado podemos olhar os histogramas e as densidades de modelos contínuos para conferir o ajuste. Na Figura 10.7, mostramos os histogramas de amostras de tamanho $n = 1000$ geradas de 4 distribuições, com o histograma padronizado e a densidade correspondente sobreposta.

Olhar o histograma pode não ser suficiente. Na Figura 10.8 um histograma de uma amostra de uma v.a. contínua com uma densidade candidata sobreposta. Os dados são de tempos de vida de componentes eletrônicos. A curva contínua é a densidade de probabilidade $f(x) = 0.024 \exp(-0.024x)$ para $x > 0$ de uma distribuição exponencial com parâmetro $\lambda = 0.024$. Não parece óbvio que a densidade ajusta-se perfeitamente ao histograma. Até onde podemos tolerar um desajuste?

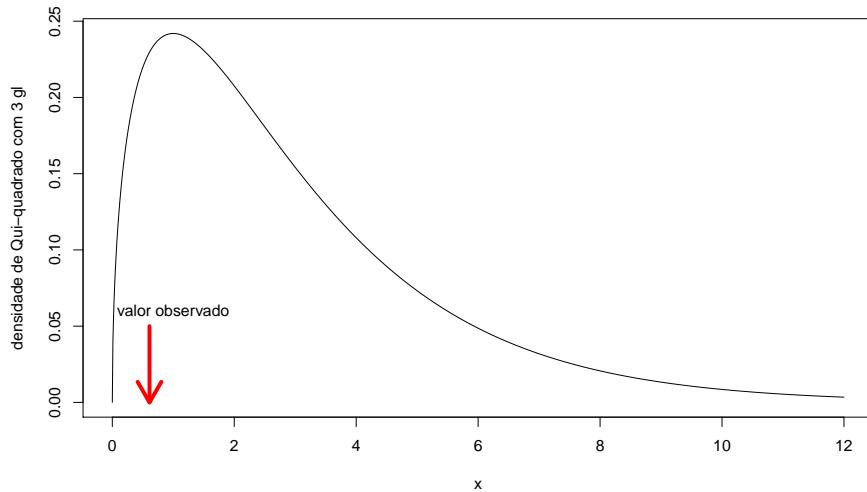


Figure 10.6: Valor observado $\chi^2 = 0.61$ e densidade de qui-quadrado com 3 g.l.

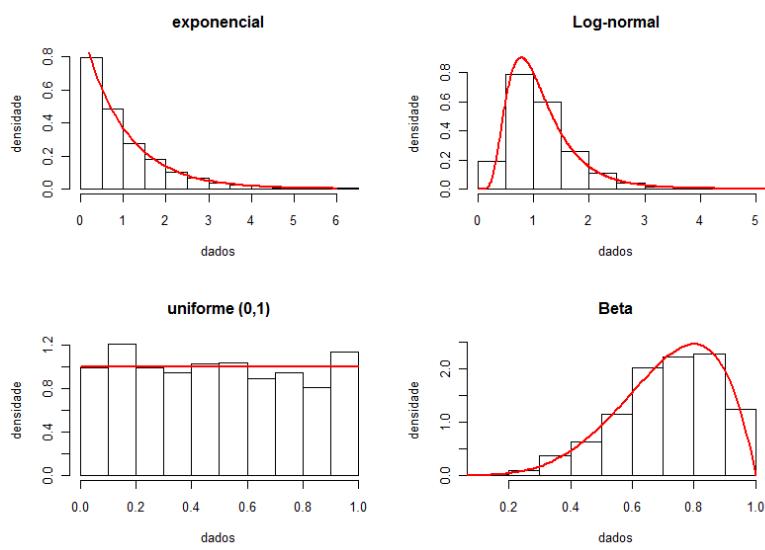


Figure 10.7: Histogramas de amostras das distribuições $\exp(1)$, log-normal $(0,0.5)$, uniforme $U(0,1)$ e beta $(5,2)$ com as densidades correspondentes.

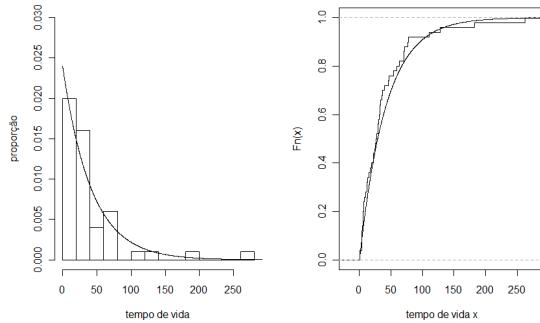


Figure 10.8: Esquerda: Gráfico de histograma de dados de tempos de vida de componentes eletrônicos e densidade de uma exponencial com parâmetro $\lambda = 0.024$. Direita: Função distribuição acumulada empírica e teórica de uma exponencial

O gráfico à direita é o da função de distribuição acumulada de probabilidade, menos intuitiva mas muito útil. Neste gráfico temos a versão empírica e teórica desta função. A função em forma de escada é a função distribuição acumulada empírica $\hat{F}_n(x)$ dos dados de tempos de vida e a curva contínua e suave é a função distribuição acumulada $F(x) = 1 - \exp(-0.024x)$ para $x > 0$ de uma exponencial com parâmetro λ igual a 0.024. Neste gráfico, estas duas funções são muito próximas. A ideia básica do teste de Kolmogorov é comparar esta duas funções acumuladas.

10.6 A função acumulada empírica

Definition 10.6.1 — A função acumulada empírica. Seja x_1, x_2, \dots, x_n um conjunto de números reais. A função distribuição acumulada empírica $\hat{F}_n(x)$ é uma função $\hat{F}_n : \mathbb{R} \rightarrow [0, 1]$ tal que, para qualquer $x \in \mathbb{R}$ temos

$$\hat{F}_n(x) = \frac{\#\{x_i \leq x\}}{n} = \text{Proporção dos } x_i \text{ que são } \leq x$$

A função $\hat{F}_n(x)$ é definida para todo x na reta real, e não apenas para os x iguais aos n valores x_i da amostra. O símbolo do chapéu em $\hat{F}_n(x)$ é para enfatizar que esta função é baseada nos dados (e daí, o adjetivo empírica). A Figura 10.9 mostra novamente a função acumulada empírica $\hat{F}_n(x)$ para os dados dos tempos de vida de equipamentos eletrônicos. Ela foi obtida com os seguintes comandos R:

```
Fn <- ecdf(dados)
plot(Fn, verticals= T, do.p=F, main="", xlab="tempo de vida x")
```

Suponha que X seja uma v.a. contínua. Adotamos um modelo para X , tal como uma exponencial com parâmetro $\lambda = 0.024$. Este é um modelo candidato, que queremos verificar se ajusta-se bem aos dados. Calculamos a função acumulada teórica $F(x)$. Não precisa dos dados para isto, este é um cálculo matemático-probabilístico. A seguir, com base nos dados da amostra, *e somente nela*, sem uso do modelo teórico, construímos a função distribuição acumulada empírica $\hat{F}_n(x)$. Se tivermos $\hat{F}_n(x) \approx F(x)$ para todo x na reta, como na Figura 10.10, nós concluímos que o modelo adotado ajusta-se bem aos dados. Como saber se $\hat{F}_n(y) \approx F(y)$? Veremos a resposta na próxima seção.

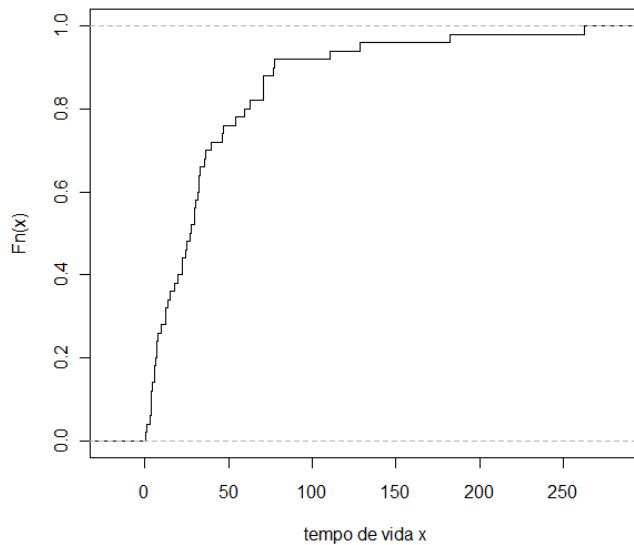


Figure 10.9: Função distribuição acumulada empírica $\hat{F}_n(x)$ dos dados de tempos de vida.

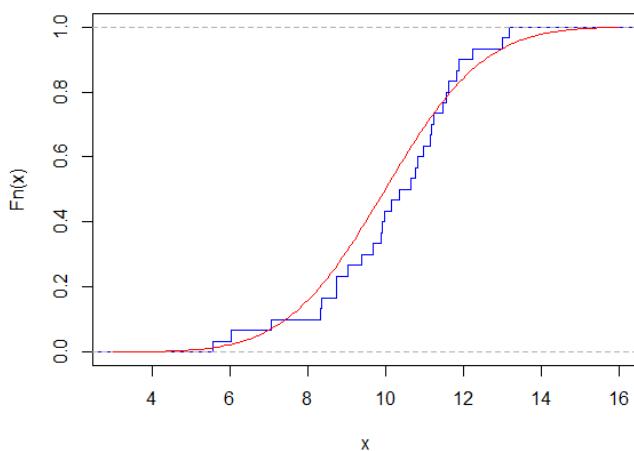


Figure 10.10: Empírica $\hat{F}_n(x)$ e a teórica $F(x)$.

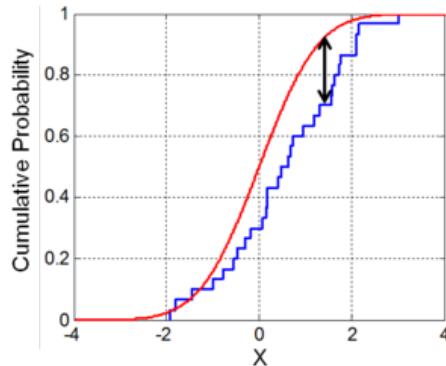


Figure 10.11: Empírica $\hat{F}_n(x)$ e a teórica $F(x)$ com a distância de Kolmogorov D_n .

10.7 Distância de Kolmogorov

Para cada ponto x na reta, olhe a distância vertical $|\hat{F}_n(x) - F(x)|$ entre as duas curvas $\hat{F}_n(x)$ e $F(x)$. Varra o eixo horizontal x procurando a maior distância entre as curvas. Vamos definir esta maior distância por D_n .

Definition 10.7.1 — Distância de Kolmogorov. Considere $D_n = \max_x |\hat{F}_n(x) - F(x)|$, a maior distância vertical entre as duas curvas $\hat{F}_n(x)$ e $F(x)$.

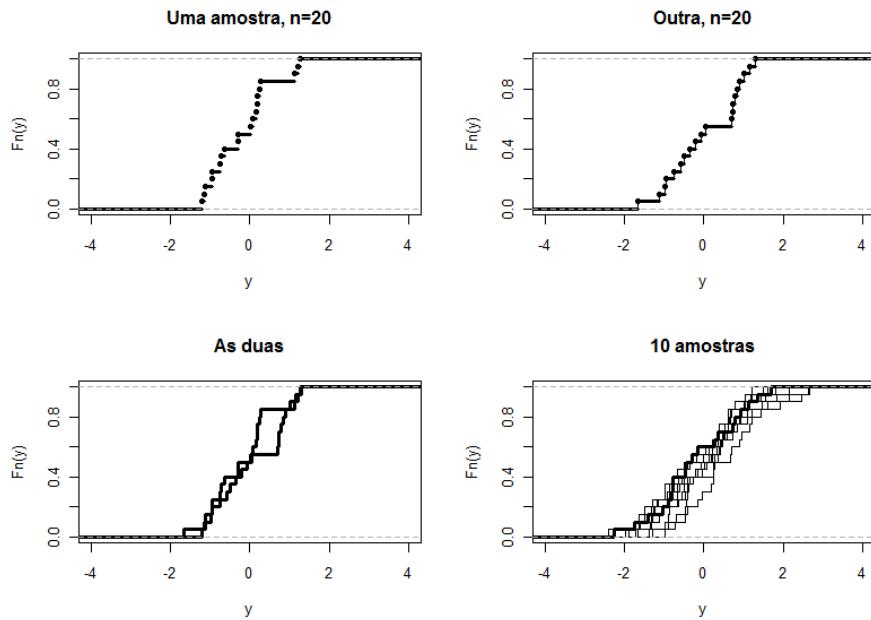
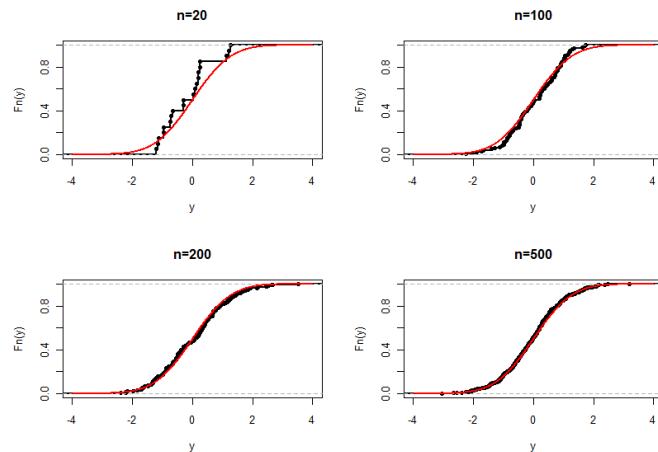
A Figura 10.11 mostra o ponto x em que as curvas $\hat{F}_n(x)$ e $F(x)$ estão separadas pela maior distância vertical ao longo do eixo horizontal. Se $D_n \approx 0$, então o modelo adotado ajusta-se bem aos dados. Como saber se $D_n \approx 0$? O grande matemático russo Andrei Kolmogorov (1903-1987) estudou o comportamento da estatística D_n .

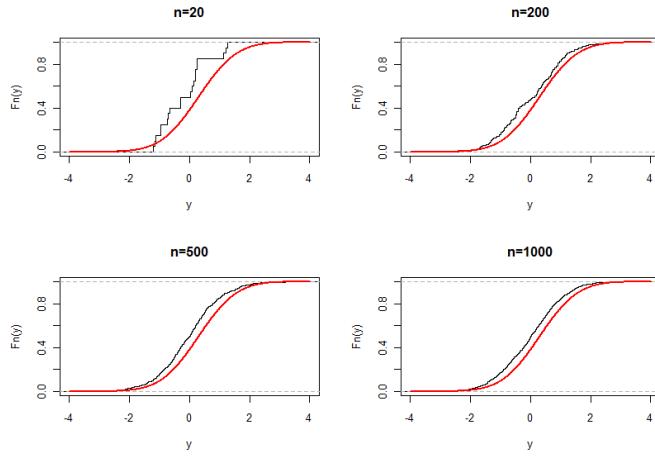
A primeira coisa a se observar é que $\hat{F}_n(x)$ é uma função aleatória. A Figura 10.12 mostra a função $\hat{F}_n(x)$ obtida com uma amostra de tamanho $n = 20$ de uma gaussiana $N(0, 1)$. A seguir, vemos outra função $\hat{F}_n(x)$, construída com uma segunda amostra do mesmo modelo $N(0, 1)$. O terceiro gráfico mostra claramente como estas duas funções empíricas são diferentes. O quarto gráfico dá uma ideia da variabilidade de $\hat{F}_n(x)$ a partir de 10 amostras distintas, todas de tamanho $n = 20$ de uma $N(0, 1)$. O código usado para gerar a Figura 10.12 foi o seguinte:

```
set.seed(3); x1 <- rnorm(20); x2 <- rnorm(20)
par(mfrow=c(2, 2))
plot(ecdf(x1), xlim=c(-4, 4), do.p=T, verticals=F, lwd=3, main="", xlab="y", ylab="Fn(y)")
plot(ecdf(x2), xlim=c(-4, 4), do.p=T, verticals=F, lwd=3, main="", xlab="y", ylab="Fn(y)")
lines(ecdf(x2), verticals=T, lty=2)
plot(ecdf(x1), xlim=c(-4, 4), do.p=F, verticals=T, lwd=3, main="", xlab="y", ylab="Fn(y)")
lines(ecdf(x2), lwd=3, do.p=F, verticals=T)
plot(ecdf(rnorm(20)), xlim=c(-4, 4), do.p=F, verticals=T, lwd=3, main="", xlab="y", ylab="Fn(y)")
for(i in 2:9) lines(ecdf(rnorm(20)), do.p=F, verticals=T)
```

Apesar de aleatório, podemos afirmar algumas coisas sobre $\hat{F}_n(x)$. Suponha que $F(x)$ é o verdadeiro modelo gerador dos dados. Na Figura 10.13 usei o modelo $N(0, 1)$. Pode-se mostrar matematicamente que, apesar de $\hat{F}_n(x)$ (e portanto, D_n também) flutuar com a amostra, temos D_n convergindo para zero quando n cresce: $D_n \rightarrow 0$ se $n \rightarrow \infty$, qualquer que seja modelo contínuo $F(x)$.

O que acontece com D_n quando n cresce se estivermos usando o modelo teórico *incorrecto*? Suponha que $F(x)$ não seja o modelo que gera os dados da amostra. Na Figura 10.14, eu uso o modelo $F(x) \sim N(0, 1)$ mas, na verdade, os dados são gerados de uma $N(0.3, 1)$. Então, pode-se

Figure 10.12: Mostrando o caráter aleatório de $\hat{F}_n(x)$.Figure 10.13: $D_n \rightarrow 0$ se o modelo é correto

Figure 10.14: $D_n \rightarrow 0$ se o modelo é correto

mostrar que D_n converge para um valor maior que zero.

Resumindo:

- Suponha que $\mathbb{F}(x)$ é o modelo verdadeiro. Então $D_n \rightarrow 0$ se $n \rightarrow \infty$.
- Se $\mathbb{F}(x)$ não é o modelo verdadeiro, $D_n \rightarrow a > 0$.

Mas continuamos com o problema de como decidir na prática: quão próximo de zero D_n tem de ser para aceitarmos o modelo teórico $\mathbb{F}(x)$? $D_n = 0.01$ é pequeno? Com certeza, a resposta depende de n já que $D_n \rightarrow 0$ se $n \rightarrow \infty$. A resposta depende do modelo teórico $\mathbb{F}(x)$ considerado? Por exemplo, o comportamento de D_n quando $\mathbb{F}(x)$ for uma gaussiana é diferente do comportamento quando $\mathbb{F}(x)$ for uma exponencial?

Vimos que $D_n \rightarrow 0$ se $n \rightarrow \infty$. Com que rapidez ele decresce em direção a 0? Kolmogorov mostrou que:

- $nD_n \rightarrow \infty$ (degenera).
- $\log(n)D_n \rightarrow 0$ (degenera).
- $\sqrt{n}D_n \not\rightarrow 0$ e também $\not\rightarrow \infty$.
- $\sqrt{n}D_n$ fica (aleatoriamente) estabilizado. Qualquer outra potência diferente de $-1/2$ leva a resultados degenerados.
- $n^{0.5+\epsilon}D_n \rightarrow \infty$.
- $n^{0.5-\epsilon}D_n \rightarrow 0$.

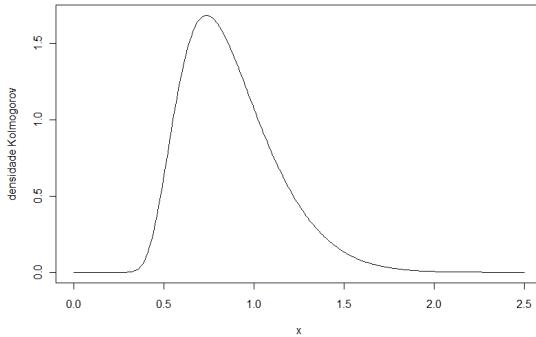
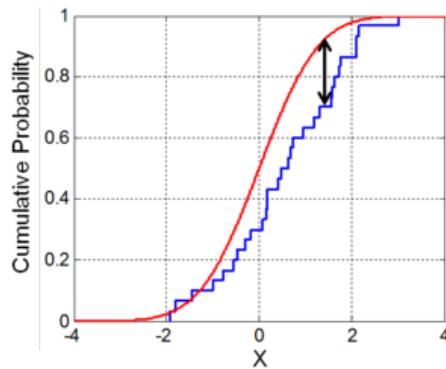
10.8 Convergência de D_n

Mas e daí? Como saber se D_n é pequeno? Suponha que o modelo $\mathbb{F}(x)$ usado na distância seja realmente o modelo verdadeiro. Kolmogorov mostrou que

$$\sqrt{n}D_n \rightarrow K$$

onde K é uma v.a. cuja distribuição de probabilidade *não depende* de $\mathbb{F}(x)$. Isto é, $\sqrt{n}D_n$ é aleatório mas sua distribuição é a mesma em todos os problemas! Assim, sabemos como $\sqrt{n}D_n$ pode variar se o modelo for verdadeiro, qualquer que seja este modelo verdadeiro. Isto significa que temos uma métrica *universal* para medir a distância entre a acumulada empírica $\hat{\mathbb{F}}_n(x)$ e a distribuição verdadeira $\mathbb{F}(x)$, *qualquer que seja* esta distribuição verdadeira.

Que distribuição universal é esta? A v.a. K segue a distribuição de uma ponte browniana, assunto muito técnico para nosso livro. Apenas como curiosidade, a densidade de K é dada por $f(x) = 8x \sum_{k=1}^{\infty} (-1)^{k+1} k^2 e^{-2k^2 x^2}$ para $x > 0$. O gráfico da Figura 10.15 mostra esta densidade

Figure 10.15: Densidade de $K \approx \sqrt{n}D_n$ Figure 10.16: Empírica $\hat{F}_n(y)$ e a teórica $F(y)$.

$f(x)$. Se calcularmos D_n usando o verdadeiro modelo $\mathbb{F}(x)$ que gerou os dados então $\sqrt{n}D_n$ deve estar entre 0.4 e 1.5 com alta probabilidade. Se não usarmos o modelo verdadeiro, sabemos que $\sqrt{n}D_n \rightarrow \infty$.

Nunca teremos $\sqrt{n}D_n$ exatamente igual a zero. Se $\sqrt{n}D_n > 1.8$ teremos uma forte evidência de que o modelo $\mathbb{F}(x)$ escolhido não é o modelo gerador dos dados. Um ponto de corte menos extremo: se $\mathbb{F}(x)$ é o modelo que gerou os dados, então a probabilidade de $\sqrt{n}D_n > 1.36$ é apenas 5%.

10.9 Resumo da ópera

Temos dados de uma amostra: x_1, x_2, \dots, x_n . Eles foram gerados i.i.d. com a distribuição $\mathbb{F}(x)$? Aqui, vamos igualar distribuição, hipótese e modelo. Como decidir? Calcule a distribuição acumulada empírica $\hat{F}_n(x)$. Calcule a distância de Kolmogorov $D_n = \max_y |\hat{F}_n(x) - \mathbb{F}(x)|$ (ver Figura 10.16) Se $\sqrt{n}D_n > 1.36$, rejeite $\mathbb{F}(x)$ como modelo gerador dos dados da amostra. Se $\sqrt{n}D_n \leq 1.36$, siga em frente com o modelo $\mathbb{F}(x)$. Ele é compatível com os dados. Na prática, nunca saberemos se $\mathbb{F}(x)$ é o modelo que gerou os dados. Sabemos apenas que o modelo proposto é compatível com o que observamos nos dados.

10.10 Kolmogorov versus Qui-quadrado

Dados X_1, X_2, \dots, X_n formam uma amostra i.i.d. de uma distribuição-modelo $\mathbb{F}(x)$? Temos duas opções para fazer um teste: Kolmogorov e Qui-quadrado. Para o teste de Kolmogorov, o modelo

$\mathbb{F}(x)$ tem de ser contínuo. A teoria não vale se for a distribuição for discreta (binomial, Poisson, etc). Além disso, o teste de Kolmogorov só é válido se não precisarmos estimar parâmetros de $F(y)$. Por exemplo, se X_1, X_2, \dots, X_n seguem uma $N(\mu, \sigma^2)$ mas não sabemos o valor de μ e σ^2 , a bela teoria de Kolmogorov não é válida. Se μ e σ^2 forem especificados de antemão, antes de olhar os dados, OK, é válido. Se eles não são especificados de antemão mas, ao contrário, precisam ser estimados a partir dos dados observados, então a distribuição de $\sqrt{n}D_n$ não é conhecida e não podemos usar Kolmogorov a não ser informalmente.

O teste qui-quadrado de Pearson pode ser aplicado com qualquer modelo, contínuo ou discreto. Ele consegue incorporar o efeito de estimar parâmetros de $\mathbb{F}(x)$, se isto for necessário. A sua implementação é muito fácil. Entretanto, precisamos especificar os intervalos ou classes onde as contagens vão ser feitas. Qual o efeito desta escolha? Em princípio, quanto mais classes, melhor. Mas usar muitas classes pode levar a categorias com probabilidades próximas de zero e então a aproximação da distribuição χ^2 não funciona bem. Devemos escolher classes de forma que o número esperado em cada uma delas seja, de preferência, pelo menos 5. Classes com contagens esperadas menores que 1 devem ser evitadas.



11. Simulação Monte Carlo

11.1 O que é uma simulação Monte Carlo

O verbo *simular* quer dizer fazer aparecer como real uma coisa que não é, fingir. Em engenharia e ciência dos dados, a *simulação* é a imitação do comportamento ou das características de um sistema probabilístico utilizando um gerador de números aleatórios num computador. Chamamos este processo de *simulação Monte Carlo*.

A simulação Monte Carlo é usada em situações onde cálculos matemáticos exatos são impossíveis ou muito difíceis de serem feitos. Outra situação onde ela também é usada é quando existem soluções exatas mas não para o problema de interesse, e sim para uma versão tão simplificada do problema real que coloca-se em dúvida a qualidade das respostas oferecidas pelo método matemático.

Estes números aleatórios gerados no computador possuem uma distribuição de probabilidade de interesse. Pode ser a distribuição normal (gaussiana), de Poisson, de Pareto (power law) ou outra qualquer. Os números aleatórios gerados servem para estudar propriedades complexas de algoritmos ou aspectos do problema que não podem ser deduzidos analiticamente, por meio de fórmulas matemáticas.

Tudo começa com a distribuição uniforme. Existe uma base para gerar números aleatórios de qualquer distribuição. Praticamente todos os métodos conhecidos começam gerando uma variável aleatória U com distribuição $U(0, 1)$, uniforme no intervalo $(0, 1)$. Isto é, U é um número escolhido ao acaso em $(0, 1)$ com densidade uniforme. A probabilidade de selecionar X num intervalo (a, b) é o seu comprimento: $b - a$. A seguir, esse métodos transformam U de forma a obter uma variável com a distribuição de interesse. Assim, todas as variáveis são obtidas a partir da distribuição $U(0, 1)$.

De fato, os números aleatórios gerados no computador não são realmente aleatórios, mas sim determinísticos. Muito trabalho de pesquisa já foi feito para criar bons geradores de números aleatórios. São procedimentos que geram uma sequência de valores U_1, U_2, \dots . Para todos os efeitos práticos, eles podem ser considerados i.i.d. com distribuição uniforme em $(0, 1)$. Além disso, por causa da representação finita nos computadores, não conseguimos de fato gerar números reais com precisão infinita.

Não veremos em detalhes os geradores de números com distribuição uniforme no intervalo $(0, 1)$. Este é um assunto bastante técnico e de pouco uso na prática da análise de dados. Vamos dar apenas um ligeira ideia de como eles funcionam. Vamos ver um dos algoritmos mais simples existentes.

11.2 Geradores de números aleatórios $U(0, 1)$

Os geradores de números aleatórios $U(0, 1)$ dependem da operação de divisão inteira. Suponha que r é o resto da divisão inteira de n por p onde r, n , e p são inteiros positivos. Por exemplo,

- $21 = 7 \times 3 + 0$ e a divisão inteira de 21 por 7 deixa resto 0.
- $22 = 7 \times 3 + 1$ e o resto é 1.
- $27 = 7 \times 3 + 6$ e o resto é 6.
- $4 = 7 \times 0 + 4$ e o resto é 4.
- Finalmente, $0 = 7 \times 0 + 0$ e o resto é 0.

Os restos possíveis da divisão inteira por 7 são $0, 1, \dots, 6$.

De forma geral, n pode ser escrito de forma única como $n = kp + r$ onde k é inteiro e $r = 0, \dots, p - 1$. O resto é o valor r sendo r igual a um dos valores $0, \dots, p - 1$. Usa-se a notação $n \equiv r \pmod{p}$ e dizemos que n é congruente com o resto r módulo p .

11.2.1 Gerador congruencial misto

Começamos com um valor inicial inteiro positivo x_0 arbitrário, chamado de *semente* (*seed*, em inglês). Recursivamente, calcule x_1, x_2, \dots por meio da fórmula:

$$ax_{i-1} + b \equiv x_i \pmod{p}$$

onde a, b , e p são inteiros positivos. x_i é um dos inteiros $0, 1, \dots, p - 1$. Por exemplo, considere o gerador dado por

$$32749x_{i-1} + 3 \equiv x_i \pmod{32777}$$

Iniciando-se com a semente $x_0 = 100$, obtenha $32749 \times 100 + 3 = 3274903$. A seguir, obtemos o resto da divisão inteira por $p = 32777$. Temos $3274903 = 99 \times 32777 + 29980$. Assim, $x_1 = 29980$. O segundo valor x_2 é obtido de forma análoga. Temos

$$32749 \times 29980 + 3 = 981815023 = 29954 \times 32777 + 12765$$

Assim, $x_2 = 12765$.

Estes valores x_1, x_2, \dots são os restos da divisão inteira módulo 32777. Portanto, todos eles são inteiros no conjunto $\{0, 1, 2, \dots, 32775, 32776\}$. Para obtermos números aleatórios no intervalo $(0, 1)$ fazemos a divisão desses restos por 32777. Assim, o primeiro número aleatório entre 0 e 1 é

$$u_1 = x_1/p = 29980/32777 = 0.9146658.$$

O segundo é

$$u_2 = x_2/p = 12765/32777 = 0.3894499,$$

e assim por diante produzindo $u_1 = 0.91466577$, $u_2 = 0.38944992$, $u_3 = 0.09549379$, $u_4 = 0.32626537$, $u_5 = 0.86466120$, $u_6 = 0.78957806$, $u_7 = 0.89190591, \dots$

A sequência

$$u_1 = x_1/p, u_2 = x_2/p, \dots$$

é uma aproximação para uma sequência de valores de variáveis *independentes* e com distribuição uniforme em $(0, 1)$. A qualidade desta aproximação é atestada pela incapacidade de vários testes estatísticos em detectar padrões não aleatórios nas sequências geradas por bons geradores.

Estes números não possuem realmente uma distribuição $U(0, 1)$. Para começar, eles não são contínuos. O gerador do nosso exemplo pode gerar, no máximo, 32776 restos x_i distintos: os inteiros $0, 1, \dots, 32776$. Assim, existem apenas 32776 números u_i no intervalo $(0, 1)$ que podem ser gerados por este procedimento:

$$0/32777, 1/32777, 2/32777, \dots, 32776/32777$$

Quanto maior o valor de p , maior o número de valores u_i distintos possíveis. Mas mesmo com um p bastante grande, existe apenas um número finito de valores possíveis.

Em segundo lugar, os números não realmente aleatórios, mas sim pseudo-aleatórios. Os valores u_1, u_2, \dots resultam de uma função matemática aplicada de forma recursiva. Usando o mesmo gerador e a mesma semente x_0 , vamos obter sempre os mesmos números u_i .

Além disso, por causa do número finito de possibilidades, a sequência de números pseudo-aleatórios começa a se repetir depois de um tempo. Por exemplo, se $a = 3$, $b = 0$, $m = 30$ e $x_0 = 1$, teremos a sequência $\{3, 9, 27, 21, 3, 9, 27, 21, 3, 9, 27, 21, 3, 9, 27, 21, 3, \dots\}$. Com probabilidade 1, depois de certo tempo, obtém-se um valor x_i igual a algum valor x_{i-k} já obtido anteriormente. A partir daí, teremos a sequência repetindo-se com $x_{i+j} = x_{i-k+j}$. O número de passos k até obter-se uma repetição numa sequência é chamado de *período* do gerador.

Uma importante biblioteca de subrotinas científicas, chamada NAG, utiliza um gerador congruencial com $a = 13^{13}$, $b = 0$ e $p = 2^{59}$, que possui um período igual a $2^{57} \approx 1.4410^{17}$. Bons geradores tem períodos tão grandes que podem ser ignorados na prática.

A semente x_0 costuma ser determinada pelo relógio interno do computador. Pode também ser pré-especificada pelo usuário. Isto garante que se repita a mesma sequência de números aleatórios. De qualquer forma, é um número arbitrário para iniciar o processo.

Este gerador que aprendemos aqui é um exemplo muito simples mas que possui as principais características de todos os geradores, inclusive aqueles bem mais sofisticados e melhores por possuírem períodos mais longos e maior granularidade. Vamos assumir no resto deste capítulo que temos algum gerador de números (pseudo)-aleatórios reais no intervalo $(0, 1)$. Isto é, sabemos gerar $U \sim U(0, 1)$. Vamos ignorar as sutilezas desses geradores e supor que U escolhe um número real completamente ao acaso no intervalo $(0, 1)$. Se (a, b) é um intervalo contido em $(0, 1)$, então $\mathbb{P}(U \in (a, b)) = (b - a)$, o comprimento do intervalo. O comando `runif(1)` em R gera um valor $U(0, 1)$. `runif(n)` gera n valores $U(0, 1)$ independentes.

11.3 Simulação de v.a.'s binomiais

Vamos começar com o caso mais simples, uma Bernoulli. Como podemos gerar a v.a. binária X onde

$$X \sim \text{Bernoulli}(p) : \quad \begin{cases} P(X = 1) = p \\ P(X = 0) = 1 - p \end{cases}$$

A ideia é muito simples. Selecione U ao acaso no intervalo $(0, 1)$. Se $U < p$, diga que $X = 1$ ocorreu. Se $U > p$, diga que $X = 0$ ocorreu.

Qual a probabilidade de gerarmos $X = 1$? Temos

$$\mathbb{P}(X = 1) = \mathbb{P}(U \in (0, p)) = p - 0 = p,$$

exatamente a probabilidade desejada. Por exemplo, para gerar uma $\text{Bernoulli}(0.35)$, podemos usar o código abaixo:

```
p = 0.35
U = runif(1)
if(U <= p) X = 1
else X = 0
```

Isto fica mais simples ainda em R: `X = runif(1) <= p`. Se quisermos gerar 215 valores i.i.d., usamos `X = runif(215) <= p`

E como gerar uma variável binomial com parâmetros n e p ? Para gerar $X \sim \text{Bin}(n, p)$, basta repetir o algoritmo Bernoulli n vezes independentemente. Supondo que $n = 100$ e $p = 0.17$, por exemplo, temos

```
n <- 100; p <- 0.17; X <- 0
for(i in 1:n){
  if(runif(1) < p) X <- X + 1
}
```

Embora correto, este procedimento não é o melhor pois precisamos fazer um número maior de operações que um outro procedimento que, além de mais eficiente, é mais genérico servindo para várias outras distribuições discretas. Antes de ver este procedimento mais geral, vamos falar um pouco mais da geração de binomiais no ambiente R.

Em R, vetorizando fica muito mais simples gerar $X \sim \text{Bin}(n, p)$: `X = sum(runif(n) <= p)`. Na verdade, o R já possui um gerador de binomial $\text{Bin}(m, \theta)$. Usando o Help do R, vemos que `rbinom(n, size, prob)` gera n valores, cada um deles vindo de uma $\text{Bin}(\text{size}, \text{prob})$. WARNING: No HELP do R, o argumento n refere-se a quantos valores binomiais $\text{Bin}(\text{size}, \text{prob})$ queremos gerar. Não confundir com a notação usual em que escrevemos $\text{Bin}(n, \theta)$. Por exemplo, para gerar $n = 10$ valores independentes de uma $\text{Bin}(100, 0.17)$ (isto é, `size=100` e `prob= theta= 0.17`), digitamos:

```
> rbinom(10, 100, 0.17)
[1] 14 20 20 14 8 14 12 13 17 14
```

A função `dbinom(x, size, prob)` calcula a probabilidade $P(X = x)$ quando X é uma v.a. binomial $\text{Bin}(\text{size}, \text{prob})$. Por exemplo, se $X \sim \text{Bin}(100, 0.17)$ então $\mathbb{P}(X = 13)$ é

```
> dbinom(13, 100, 0.17)
[1] 0.06419966
```

Podemos pedir vários valores de uma única vez:

```
> dbinom(13:17, 100, 0.17)
[1] 0.06419966 0.08171369 0.09595615 0.10441012 0.10566807
```

11.4 Simulação de v.a.'s discretas arbitrárias

Vamos ver um procedimento geral, que serve para qualquer distribuição discreta, mesmo para aquelas com infinitos valores, como a Poisson, Geométrica e Pareto.

Suponha que X é uma variável aleatória discreta com suporte $\{x_1, x_2, \dots\}$. Temos $\mathbb{P}(X = x_i) = p_i > 0$ para $i = 0, 1, \dots$ e com $\sum_i p_i = 1$. Por exemplo, poderíamos ter X com distribuição de Poisson com parâmetro $\lambda = 1.61$. Assim, $x_i = i$ e $p_i = (1.61)^i / i! \exp(-1.61) = 0.1998876 (1.61)^i / i!$ com $i = 0, 1, 2, \dots$.

A distribuição de X é dada por:

x_i	$P(x = x_i) = p_i$
x_1	p_1
x_2	p_2
x_3	p_3
\vdots	\vdots
Total	$\sum_i p_i = 1$

Acumulamos as probabilidades obtendo $F(x_k) = P(X \leq x_k) = \sum_{i=1}^k p_i$. Por exemplo,

$$\begin{aligned} F(x_1) &= p_1 \\ F(x_2) &= p_1 + p_2 \\ F(x_3) &= p_1 + p_2 + p_3 \text{ Etc.} \end{aligned}$$

O algoritmo é muito simples:

- Se $0 < U < F(x_1) = p_1$, faça $X = x_1$
- Se $p_1 \leq U < p_1 + p_2$, faça $X = x_2$
- Se $p_1 + p_2 \leq U < p_1 + p_2 + p_3$, faça $X = x_3$
- Etc.

De maneira mais formal, fazemos $X = g(U)$ onde g é a função matemática definida da seguinte forma:

$$X = g(U) = \begin{cases} x_0, & \text{se } U < p_0 \\ x_1, & \text{se } p_0 \leq U < p_0 + p_1 \\ x_2, & \text{se } p_0 + p_1 \leq U < p_0 + p_1 + p_2 \\ \dots & \dots \\ x_i, & \text{se } \sum_{k=1}^{i-1} p_k \leq U < \sum_{k=0}^i p_k \\ \dots & \dots \end{cases}$$

De forma mais resumida, podemos escrever que, se $F(x_k) = P(X \leq x_k) = \sum_{i=1}^k p_i$ então $X = g(U) = x_j$ se $F(x_{j-1}) \leq U < F(x_j)$.

■ **Example 11.1 — Caso simples.** Um exemplo simples de uso desta técnica é o seguinte: suponha que desejamos gerar um valor de uma variável aleatória X com a seguinte distribuição de probabilidade discreta:

$$X = \begin{cases} -1, & \text{com probabilidade } p_0 = 0.25 \\ 2, & \text{com probabilidade } p_1 = 0.35 \\ 7, & \text{com probabilidade } p_2 = 0.17 \\ 12, & \text{com probabilidade } p_3 = 0.23 \end{cases}$$

Basta então gerar um valor $U \sim U(0, 1)$ e decidir sobre o valor de X a partir do intervalo em que U cair:

$$g(U) = X = \begin{cases} -1, & \text{se } U < 0.25 \\ 2, & \text{se } 0.25 \leq U < 0.60 \\ 7, & \text{se } 0.60 \leq U < 0.77 \\ 12, & \text{se } 0.77 \leq U < 1.00 \end{cases}$$

Por exemplo, se o valor simulado de U for igual a 0.4897 então o valor simulado de X será 2 pois $0.25 \leq 0.4897 < 0.60$. ■

11.5 Gerando Poisson

Para o caso de $X \sim \text{Poisson}(1.61)$ teríamos:

$$X = g(U) = \begin{cases} 0 & \text{se } U < 0.1998876 \\ 1 & \text{se } 0.1998876 \leq U < 0.5217067 \\ 2 & \text{se } 0.5217067 \leq U < 0.7807710 \\ \dots & \dots \\ i & \text{se } 0.1998876 \sum_{k=1}^{i-1} (1.61)^k / k! \leq U < 0.1998876 \sum_{k=0}^i (1.61)^k / i! \\ \dots & \dots \end{cases}$$

Neste exemplo da Poisson, existe uma dificuldade: é impossível listar os infinitos possíveis valores de X e só então verificar onde o valor de X caiu. Uma maneira mais apropriada é trabalhar sequencialmente: verifique se U cai no primeiro intervalo. Se sim, atribua o valor 0 à X e pare o procedimento. Senão, calcule o intervalo seguinte e verifique se U cai neste novo intervalo. Se sim, atribua o valor 1 à X e pare o procedimento. Senão, calcule o intervalo seguinte e etc.

Para facilitar o cálculo podemos ainda usar uma relação de recorrência entre as probabilidades sucessivas de uma Poisson com parâmetro λ :

$$p_{i+1} = \frac{\lambda}{i+1} p_i$$

O código em R para este procedimento com $\lambda = 1.61$ seria:

```
lambda <- 1.61
x <- -1
i <- 0; p <- exp(-lambda); F <- p
while(x == -1){
  if(runif(1) < F) x <- i
  else{
    p <- lambda*p/(i+1)
    F <- F + p
    i <- i+1
  }
}
```

■ **Example 11.2 — Gerando Poisson numa seguradora.** Suponha que você é um atuário trabalhando numa pequena companhia de seguros. Sua tarefa é simular a perda agregada que a companhia pode experimentar no próximo ano em um tipo bem particular de apólice. Uma das etapas exige a simulação do número de sinistros mensais que, com base na sua experiência passada e na de outros atuários, você decide assumir que é Poisson com valor esperado $\lambda = 1.7$. Você usa um gerador de números aleatórios i.i.d $U(0,1)$ para produzir a seguinte seqüência: 0.670, 0.960, 0.232, 0.224, 0.390, 0.494. Obtenha os valores correspondentes da distribuição de Poisson(1.7).

Como os valores acumulados $P(X \leq k)$ para $k = 0, 1, 2, 3, 4$ de uma Poisson(1.7) são iguais a 0.183, 0.493, 0.757, 0.907, 0.970, os valores simulados da Poisson são iguais a 2, 4, 1, 1, 1, 2, respectivamente. ■

■ **Example 11.3 — Recursão em Binomial.** Mostre que, para o caso de $X \sim \text{Bin}(n, \theta)$, temos

$$p_{i+1} = \frac{p(n-i)}{(1-p)(i+1)} p_i$$

Escreva um código em R para gerar variáveis binomiais com um procedimento similar ao da Poisson.

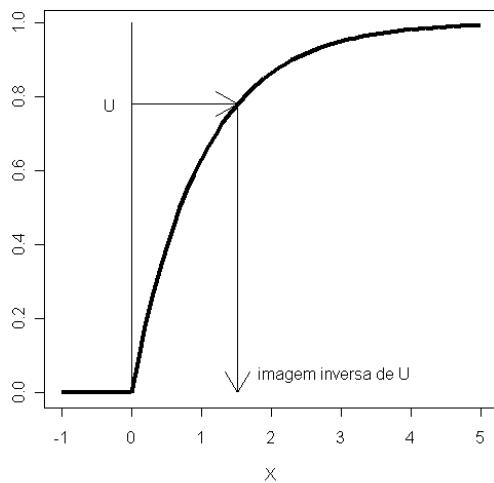


Figure 11.1: Gráfico da função distribuição acumulada $F(x)$ de certa variável aleatória. U é um número aleatório com distribuição $U(0, 1)$. O gráfico mostra sua imagem inversa $F^{-1}(U)$ através da distribuição acumulada de X . Esta imagem é aleatória e possui distribuição igual a de X .

```

x <- -1
i <- 0; c <- p/(1-p); pr <- (1-p)^n; F <- pr
while(x == -1){
  if(runif(1) < F) x <- i
  else{
    pr <- ((c*(n-i))/(i+1))*pr
    F <- F + pr
    i <- i+1
  }
}
%

```

■

11.6 Método da transformada inversa

Suponha que certa variável aleatória contínua X possua função distribuição acumulada dada por $F(x)$. Por exemplo, se $X \sim \exp(3)$ então $F(x) = 1 - \exp(-3x)$ para $x \geq 0$ e $F(x) = 0$ se $x < 0$. Um método muito poderoso para gerar X é gerar uma variável uniforme $U \sim U(0, 1)$ e transformá-la usando a função matemática $Y = F^{-1}(U)$. A variável aleatória Y possui a mesma distribuição que X . Isto é, a função distribuição acumulada de Y é exatamente $F(x)$.

A Figura 11.1 mostra de forma gráfica como este método trabalha. Primeiro, gere um número $U \sim U(0, 1)$ e coloque-o no eixo vertical. A seguir, obtenha a imagem inversa deste valor aleatório U por meio da função matemática F_X^{-1} . Esta imagem inversa é uma variável aleatória pois é função do valor aleatório U . Além disso, esta imagem inversa possui distribuição de probabilidade igual à distribuição desejada. Isto é, a distribuição acumulada da variável aleatória $F_X^{-1}(U)$ é F_X^{-1} .

Por exemplo, se $X \sim \exp(3)$ então $F(x) = 1 - \exp(-3x)$ para $x \geq 0$. Calcule a inversa de F . Basta igualar a expressão de $F(x)$ a um valor u e inverter isolando x como função de u . Se

$$u = 1 - \exp(-3x)$$

então

$$1 - u = \exp(-3x).$$

Tomando log dos dois lados, temos

$$\log(1 - u) = -3x$$

e portanto

$$x = -\frac{1}{3} \log(1 - u).$$

Assim, $\mathbb{F}^{-1}(u) = -1/3 \log(1 - u)$.

Agora, gere uma variável uniforme $U \sim U(0, 1)$. A seguir, transforme este número aleatório usando $Y = \mathbb{F}^{-1}(U) = -1/3 \log(1 - U)$.

Esta v.a. Y possui a mesma distribuição que X . Isto é, $Y \sim \exp(3)$. A função distribuição acumulada de Y no ponto x é exatamente $\mathbb{F}(x) = 1 - \exp(-3x)$. De fato, se $x > 0$, nós temos

$$\begin{aligned} \mathbb{P}(Y \leq x) &= \mathbb{P}(-1/3 \log(1 - U) \leq x) \\ &= \mathbb{P}(\log(1 - U) \geq -3x) \\ &= \mathbb{P}(1 - U \geq e^{-3x}) \quad \text{tomando exp dos dois lados} \\ &= \mathbb{P}(U \leq 1 - e^{-3x}) \\ &= \mathbb{P}(U \in (0, 1 - e^{-3x})) \\ &= 1 - e^{-3x} \quad \text{pois } U \sim U(0, 1) \end{aligned}$$

Repetindo o argumento acima, pode-se gerar uma exponencial com qualquer parâmetro. Se $Y \sim \exp(\lambda)$ usando a transformação $Y = -1/\lambda \log(1 - U)$ pode ser usada. Note que, como U e $1 - U$ possuem a mesma distribuição uniforme $U(0, 1)$ (você pode mostrar isto?), então $Y = -1/\lambda \log(U)$ também é exponencial com parâmetro λ .

A mesma prova dada acima pode ser usada para mostrar o resultado de forma geral, para qualquer v.a. contínua. Seja $U \sim U(0, 1)$. Defina a v.a. $Y = \mathbb{F}^{-1}(U)$. Como uma função de distribuição acumulada é não decrescente, se $a \leq b$, então $\mathbb{F}(a) \leq \mathbb{F}(b)$. Além disso, $\mathbb{P}(U \leq a) = a$ se $a \in [0, 1]$. Assim,

$$\begin{aligned} \mathbb{F}_Y(x) &= \mathbb{P}(Y \leq x) \\ &= \mathbb{P}(\mathbb{F}_X^{-1}(U) \leq x) \\ &= \mathbb{P}(\mathbb{F}_X(\mathbb{F}_X^{-1}(U)) \leq \mathbb{F}_X(x)) \\ &= \mathbb{P}(U \leq \mathbb{F}_X(x)) \\ &= \mathbb{F}_X(x) \end{aligned}$$

11.7 Gerando v.a. com distribuição Gomperz

Uma distribuição muito importante para o mercado de seguros é a distribuição de Gompertz. Ela modela muito bem o tempo de vida a partir dos 22 anos. Seu suporte é o intervalo $(0, \infty)$ e, para x neste intervalo, temos a densidade de probabilidade dada por

$$f(x) = Bc^x e^{-B(c^x - 1)/\log(c)}$$

onde $c > 1$ e $B > 0$. O parâmetro c usualmente possui um valor em torno de 1.09. Um valor típico para B é 1.02×10^{-4} . A função de distribuição acumulada $F(x)$ é

$$\mathbb{F}(x) = 1 - \exp\left(-\frac{B}{\log(c)}(c^x - 1)\right)$$

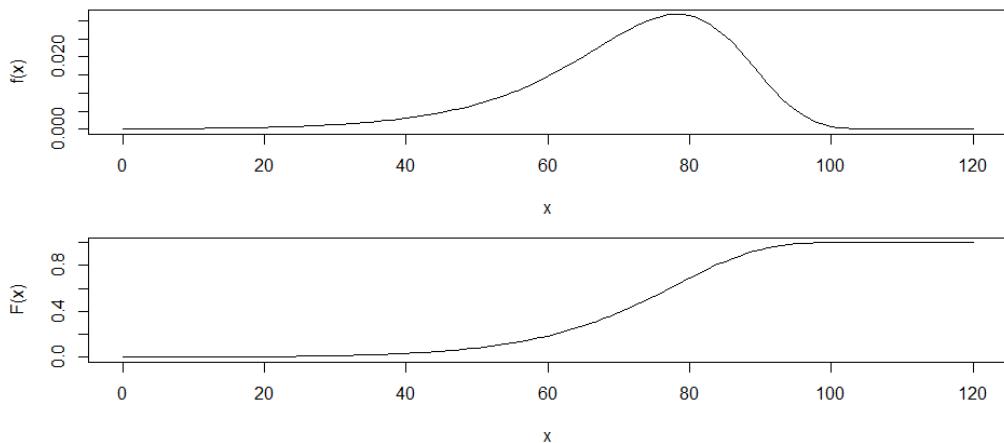


Figure 11.2: Densidade $f(x)$ (acima) e distribuição acumulada $\mathbb{F}(x)$ de uma v.a. Gomperz com parâmetros $c = 1.09$ e $B = 0.000102$.

onde $c > 1$ e $B > 0$. A Figura 11.2 mostra a função densidade $f(x)$ (acima) e a distribuição acumulada $\mathbb{F}(x)$ de uma v.a. Gomperz com parâmetros $c = 1.09$ e $B = 0.000102$.

```
ce <- 1.09; B <- 0.000102; k <- B/log(ce)
eixox <- seq(0,120,by=1)
dens <- B * ce^eixox * exp(-k * (ce^eixox - 1))
Fx = 1 - exp( - (B/log(ce)) * (ce^eixox -1) )
par(mfrow=c(2,1), mar=c(4,4,1,1))
plot(eixox, dens, type="l", xlab="x", ylab="f(x)")
plot(eixox, Fx, type="l", xlab="x", ylab="F(x)")
```

A transformada inversa de uma Gomperz é facilmente obtida:

$$\mathbb{F}^{-1}(u) = \log(1 - \log(c) \log(1 - u)/B) / \log(c)$$

Com isto obtemos a amostra (ver Figura 11.3). O código em R para obter uma amostra é o seguinte:

```
# Amostra de 10 mil valores iid de Gompertz
## fixa as constantes
ce <- 1.09; B <- 0.000102; k <- B/log(ce)
u <- runif(10000) ## gera valores iid U(0,1)
## Gompertz pelo metodo da transformada inversa
x <- 1/log(ce) * (log( 1- log(1-u)/k))
# fazendo histograma e densidade
hist(x, prob=T)
eixox <- seq(0,120,by=1)
dens <- B * ce^eixox * exp(-k * (ce^eixox - 1))
lines(x,y)
```

11.8 Gerando v.a. com distribuição de Pareto

Uma distribuição muito usada para valores extremos de perdas em seguros é a distribuição de Pareto. Ela possui dois parâmetros. O primeiro, $x_0 > 0$, é o valor mais baixo que uma perda pode ter.

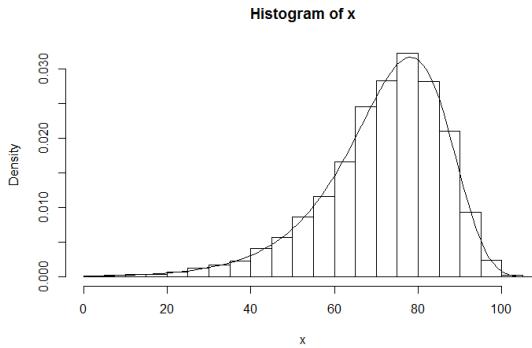


Figure 11.3: Histograma dos valores gerados de uma Gomperz com parâmetros $c = 1.09$ e $B = 0.000102$ e densidade de probabilidade.

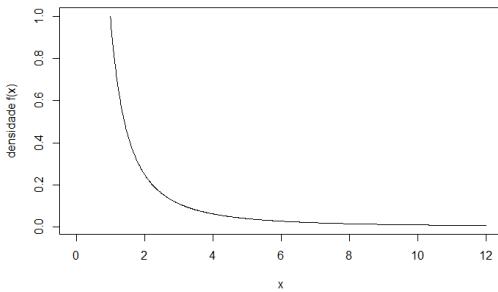


Figure 11.4: Densidade da Pareto com $x_0 = 1$ e $\alpha = 1$

Você pode pensar em x_0 como um valor de franquia ou como um valor *stop-loss* de uma seguradora. Uma seguradora vai cobrir toda a perda acima do valor x_0 e esta seguradora só toma conhecimento de sinistros com valores acima de x_0 .

O segundo parâmetro, $\alpha > 0$, controla o peso da cauda superior da distribuição em relação aos valores mais baixos e próximos de x_0 . Quanto menor α , maior a chance de observarmos valores extremos numa perda que segue a distribuição de probabilidade de Pareto.

A densidade de probabilidade de uma variável aleatória X que possui distribuição de Pareto com parâmetros (x_0, α) é dada por

$$f(x) = \begin{cases} 0, & \text{se } x \leq x_0 \\ \frac{\alpha}{x_0} \left(\frac{x_0}{x}\right)^{\alpha+1}, & \text{se } x > x_0 \end{cases}$$

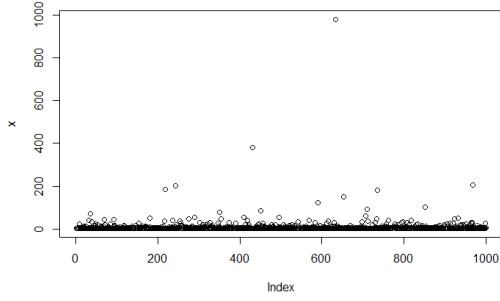
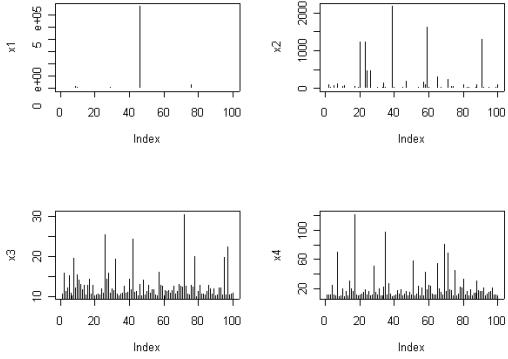
e pode ser visualizada na Figura 11.4.

As propriedades da distribuição de Pareto dependem essencialmente do valor de α . Por exemplo, se $\alpha < 1$, então

$$\mathbb{E}(X) \int_{x_0}^{\infty} f_X(x) dx = \infty$$

e portanto o valor esperado não existe neste caso. Se $\alpha > 1$ o valor esperado sempre existe mas se $\alpha < 2$ é a variância de X que não existe (é infinita).

Quais os valores típicos de α na prática de seguros e resseguros? A Swiss Re, a maior companhia européia de resseguros, fez um estudo. Nos casos de perdas associadas com incêndios,

Figure 11.5: Amostra de 1000 valores i.i.d. de uma Pareto com $x_0 = 1$ e $\alpha = 1$ Figure 11.6: Amostras de 100 valores Pareto com (x_0, α) igual a $(1.3, 0.25)$ (canto superior esquerdo), $(1.3, 0.5)$ (canto superior direito), $(10, 5)$ (canto inferior esquerdo) e $(10, 2)$ (canto inferior direito).

$\alpha \in (1, 2.5)$. Esta faixa pode ser mais detalhada: para incêndios em instalações industriais de maior porte, temos $\alpha \approx 1.2$. Para incêndios ocorrendo em pequenos negócios e serviços temos $\alpha \in (1.8, 2.5)$. No caso de perdas associadas com catástrofes naturais: $\alpha \approx 0.8$ para o caso de perdas decorrentes de terremotos; $\alpha \approx 1.3$ para furacões, tornados e vendavais.

A função distribuição acumulada para uma variável aleatória de Pareto é dada por

$$\mathbb{F}(x) = \mathbb{P}(X \leq x) = \begin{cases} 0, & \text{se } x \leq x_0 \\ \int_0^x \frac{\alpha}{x_0} \left(\frac{x_0}{y}\right)^{\alpha+1} dy = 1 - \left(\frac{x_0}{x}\right)^\alpha, & \text{se } x > x_0 \end{cases}$$

Então, para gerar uma v.a. Pareto, use

$$X \sim F_X^{-1}(U) = x_0 / (1 - U)^{-1/\alpha}.$$

Em R, basta digitar `x0/(1-runif(1000))^(1/a)` e o resultado está na Figura 11.5.

O efeito do parâmetro α na geração de valores extremos fica mais claro na Figura 11.6. Ela mostra gráficos de linha de quatro amostras de tamanho 100 cada uma de uma distribuição de Pareto. Os pontos dos gráficos possuem coordenadas (i, x_i) onde x_i é o i -ésimo dado de uma amostra da Pareto, $i = 1, \dots, 100$. As linhas conectam os pontos $(i, 0)$ até (i, x_i) . Os gráficos da linha superior possuem parâmetros (x_0, α) iguais a $(1.3, 0.25)$ (esquerda) e $(1.3, 0.5)$ (direita) enquanto os gráficos da linha inferior possuem $(10, 5)$ (esquerda) e $(10, 2)$ (direita).

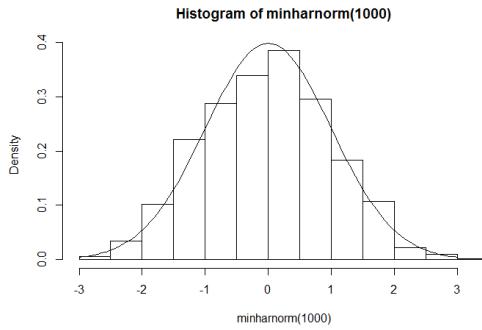


Figure 11.7: Histograma padronizado de 1000 valores $N(0,1)$ gerados com função `minhanorm` com a densidade $f(x)$ sobreposta.

11.9 Gerando v.a. gaussiana ou normal

A distribuição normal é muito especial pois ela é a aproximação para a soma de variáveis independentes (Teorema Central do Limite). Como a distribuição acumulada $\mathbb{F}(x)$ de uma normal não possui uma fórmula fechada, o uso da técnica de transformação $F^{-1}(U)$ de variáveis uniformes não pode ser usado.

Box e Muller propuseram um método muito simples para este caso especial da distribuição gaussiana. É possível mostrar que, se $\theta \sim U(0, 2\pi)$ e $V \sim \exp(0.5)$, então $X = \sqrt{V} \cos(\theta) \sim N(0, 1)$. Como você sabe gerar uniformes e exponenciais, você pode usar este resultado para gerar normais padronizadas. Em R, basta criar a função abaixo:

```
minhanorm = function(n) sqrt(rexp(n, 0.5)) * cos(runif(n, 0, 2 * pi))
```

Gerando uma amostra (com o resultado na Figura 11.7):

```
set.seed(123)
minhanorm = function(n){ sqrt(rexp(n, 0.5)) * cos(runif(n, 0, 2 * pi)) }
hist(minhanorm(1000), prob=T)
plot(dnorm, -3,3, add=T)
```

Como gerar gaussianas $N(\mu, \sigma^2)$, centrada em μ e com desvio-padrão σ em torno de μ . Usamos uma propriedade da distribuição gaussiana: se $Z \sim N(0, 1)$ então $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$. Nós sabemos gerar $Z \sim N(0, 1)$ usando o algoritmo de Box-Muller. Se quisermos, por exemplo, $X \sim N(10, 4)$, basta gerar Z e em seguida tomar $X = 10 + \sqrt{4}Z$. Em R:

```
minhanorm = function(n){ sqrt(rexp(n, 0.5)) * cos(runif(n, 0, 2 * pi)) }
x = 10 + sqrt{4} * minhanorm(100)
```

É claro que, sendo a gaussiana uma distribuição tão importante, R já possui um gerador de gaussianas: `rnorm(100, mean=0, sd=1)`.

11.10 Monte Carlo para estimar integrais

Queremos calcular uma integral

$$\theta = \int_0^1 g(x) dx$$

Podemos ver a integral θ como a esperança de uma v.a.: se $U \sim U(0, 1)$ então $\theta = E[g(U)]$. Usamos agora que, se U_1, U_2, \dots, U_n são i.i.d. $U(0, 1)$ então as v.a.'s $Y_1 = g(U_1), Y_2 = g(U_2), \dots, Y_n = g(U_n)$

também são i.i.d. com esperança θ . Pela Lei dos Grandes Números (ver capítulo 16), se $n \rightarrow \infty$,

$$\frac{1}{n} \sum_{i=1}^n Y_i = \frac{1}{n} \sum_{i=1}^n g(U_i) \rightarrow E[g(U)] = \theta$$

Assim, se n é grande, θ é aproximadamente a média aritmética dos valores simulados $g(u_i)$.

■ **Example 11.4 — Calculando uma integral simples.** Vamos estimar a integral

$$\theta = \int_0^1 x^2 dx = \frac{1}{3}$$

usando Monte Carlo. Uma amostra i.i.d. de 1000 variáveis aleatórias $U(0, 1)$ é gerada:

$$u_1 = 0.4886415, u_2 = 0.1605763, u_3 = 0.8683941, \dots, u_{1000} = 0.3357509$$

Calculamos então

$$\begin{aligned}\hat{\theta} &= (u_1^2 + u_2^2 + \dots + u_{1000}^2) / 1000 \\ &= ((0.4886415)^2 + (0.1605763)^2 + \dots + (0.3357509)^2) / 1000 \\ &= 0.33406 \approx \theta\end{aligned}$$

Se fizermos uma nova geração das U_i , com uma semente diferente, vamos produzir $\hat{\theta}$ ligeiramente diferente. Outros 1000 valores da uniforme produzem $\hat{\theta} = 0.3246794$. Aumentando o tamanho da amostra a variação de uma simulação para outra diminui: a escolha do tamanho da amostra precisa de desigualdades em probabilidade (logo mais). ■

■ **Example 11.5 — Integrais gaussianas.** Se $X \sim N(0, 1)$ então

$$\mathbb{P}(X \in (0, 1)) = \int_0^1 \frac{\exp(-x^2/2)}{\sqrt{2\pi}} dx = \theta$$

Não existe fórmula para esta integral, que deve ser obtida numericamente. Usando as funções nativas em R, obtemos o melhor que a análise numérica pode fornecer: `pnorm(1) - pnorm(0)` que retorna $0.8413447 - 0.5 = 0.3413447$

Vamos obter este valor (aproximadamente) por meio de simulação Monte Carlo. Gere 1000000 valores i.i.d. de uma $U(0, 1)$ e calcule $(y_1 + y_2 + \dots + y_{1000}) / 1000$ onde $y_i = (2\pi)^{-0.5} \exp(-u_i^2/2)$.

Por exemplo, se $u_i = 0.4886$ então $y_i = (2\pi)^{-0.5} \exp(-0.4886^2/2) = 0.3541$.

Em R: `mean((2*pi)^(-0.5) * exp(-runif(100000)^2/2))`

Quatro simulações sucessivas (e independentes) com 1000000 valores: 0.3414839, 0.3413451, 0.3411779, 0.3412634. Com uma amostra de tamanho 100 mil, estamos tendo alguma variação na quarta decimal. Comparando com $\theta = 0.3413447$, os erros de estimação são pequenos. ■

11.10.1 Integrais com limites genéricos

Nem sempre a integral de interesse terá os limites 0 e 1. No caso geral,

$$\theta = \int_a^b g(x) dx$$

Como fazer neste caso? Simples. Gere $U_i \sim U(0, 1)$ e a seguir transforme para uma $U(a, b)$ com $X_i = a + (b - a)U_i$. Agora, calcule a média aritmética dos valores $g(X_i)$ e multiplique o resultado por $b - a$. A razão é a seguinte:

$$\theta = \int_a^b g(x) dx = (b - a) \int_a^b g(x) \frac{1}{b - a} dx = (b - a) \mathbb{E}(g(X))$$

onde $X \sim U(a, b)$ e portanto tem densidade $f(x) = 1/(b - a)$.

■ **Example 11.6 — Integral com limites genéricos.** Calcule o valor aproximado de

$$\theta = \int_3^9 \log(2 + |\sin(x)|) e^{-x/20} dx$$

Uma amostra U_1, U_2, \dots i.i.d. de 100000 $U(0, 1)$ é gerada e calcula-se $V_i = 3 + 6U_i$. A seguir, obtemos

$$W_i = g(V_i) = \log(2 + |\sin(V_i)|) e^{-V_i/20}$$

e estimamos a integral com

$$(9 - 3)\bar{W} = 6 \frac{1}{100000} (W_1 + \dots + W_{100000})$$

Em código R:

```
v = 3 + 6*runif(100000, 0, 1) # na verdade, podemos usar v = runif(100000, 3, 9)
w = log(2 + abs(sin(v))) * exp(-v/20)
mean(w) * 6
```

Três simulações deram: 4.309863, 4.308165, 4.30991 e 4.310968. Neste exemplo, não sabemos o verdadeiro valor θ da integral mas as simulações dão aproximadamente o mesmo valor. Isto é um sinal de que, ao usar qualquer um deles como estimativa, a integral deve estar sendo estimada com pequeno erro. ■

11.11 Método da rejeição

Queremos gerar amostra de densidade $f(x)$. Não conseguimos obter $\mathbb{F}(x)$ analiticamente e o método da transformada inversa não pode ser usado. Uma alternativa: usar o *método de aceitação-rejeição*. A ideia básica deste método é a seguinte: gere de *outra distribuição* que seja fácil. A seguir, retemos alguns dos valores gerados e descartamos os outros. Isto é feito de tal maneira que a amostra que resta é gerada exatamente da densidade $f(x)$.

A essência dessa ideia está na Figura 11.8. Suponha que sabemos gerar com facilidade da densidade $g(x)$ (linha tracejada). Uma amostra gerada a partir de $g(x)$ produz o histograma abaixo. Mas queremos amostra de $f(x)$. Eliminamos de forma seletiva alguns dos valores gerados. Se o processo seletivo for feito de maneira adequada, terminamos com uma amostra que, no fim dos dois processos (geração e aceitação-rejeição), é gerada de $f(x)$. O resultado é aquele na Figura 11.9.

Como funciona exatamente este método? Fixe uma densidade-alvo $f(x)$. Quais $g(x)$ podemos escolher? Em princípio, qualquer uma desde que o suporte de $g(x)$ seja igual ou maior que aquele de $f(x)$. Isto é, se $f(x) > 0$ então $g(x) > 0$. $g(x)$ pode gerar valores impossíveis sob $f(x)$. Mas não podemos permitir que valores possíveis sob $f(x)$ sejam impossíveis sob $g(x)$. Isto é bem razoável: se inicialmente, usando $g(x)$, gerarmos valores impossíveis sob $f(x)$, podemos rejeitá-los no segundo passo do algoritmo. Mas se nunca gerarmos valores de certas regiões possíveis sob $f(x)$, nossa amostra final não será uma amostra de $f(x)$.

Em seguida, devemos encontrar uma constante $M > 1$ tal que $f(x) \leq Mg(x)$ para todo x . Isto é, multiplicamos a densidade $g(x)$ de onde sabemos amostrar por uma constante $M > 1$ implicando em empurrar o gráfico de $g(x)$ para cima até que ele cubra a densidade $f(x)$. Por exemplo, se $M = 2$, compararmos o valor de $f(x)$ com $2g(x)$, duas vezes a altura da densidade g no ponto x . Devemos ter sempre $f(x) \leq Mg(x)$.

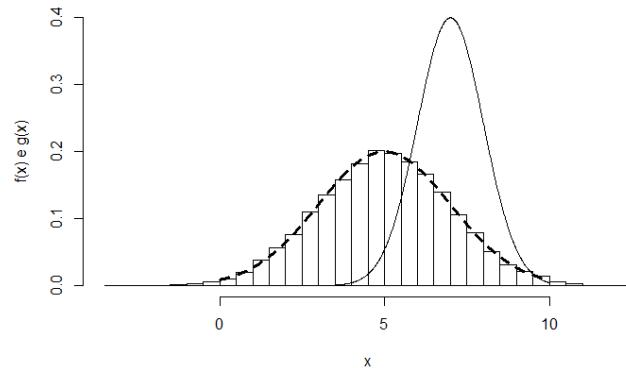


Figure 11.8: Linha contínua: densidade $f(x)$ de onde queremos amostrar. Linha tracejada: densidade $g(x)$ de onde sabemos amostrar. Histograma de amostra de 20000 elementos de $f(x)$.

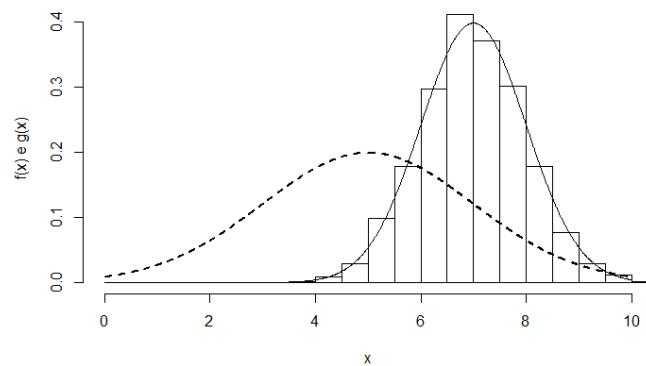


Figure 11.9: Linha contínua: densidade $f(x)$ de onde queremos amostrar. Linha tracejada: densidade $g(x)$ de onde sabemos amostrar. histograma de amostra de 20000 elementos de $f(x)$. Histograma dos 3696 elementos da amostra anterior que restaram após rejeitar seletivamente 16304 dos elementos gerados.

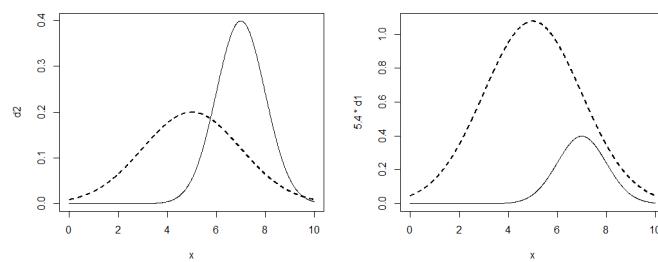


Figure 11.10: $f(x)$ e $Mg(x)$.

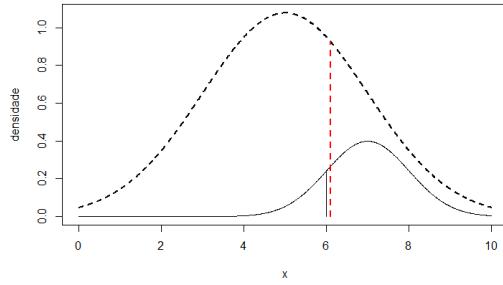


Figure 11.11: Em $x = 6.0$, temos as alturas $f(6.0)$ (linha contínua) e a altura $5.4g(6.0)$ (tracejada).

Na Figura 11.10, a curva com a linha contínua é a densidade $f(x)$ de onde queremos amostrar. A curva com a linha tracejada é a densidade $g(x)$ de onde sabemos amostrar. A direita temos o gráfico de $f(x)$ e de $5.4g(x)$. Observe que temos $f(x) \leq 5.4g(x)$ para todo x

Agora, temos $f(x)$ e $Mg(x)$ tal que $f(x) < Mg(x)$. Veja a Figura 11.11. No ponto $x = 6.0$ temos a altura $f(6.0)$ (contínua) e a altura $5.4g(6.0)$ (tracejada). Para todo x , definimos a razão entre estas alturas

$$r(x) = \frac{f(x)}{Mg(x)} \leq 1 \quad \text{para todo } x.$$

Sejam x_1, x_2, \dots os elementos da amostra de $g(x)$. Quais destes valores vamos reter e quais vamos rejeitar? Calcule $r(x_1), r(x_2), \dots$ Se $r(x_i) \approx 0$, tipicamente vamos rejeitar x_i . Se $r(x_i) \approx 1$, tipicamente vamos reter x_i .

Para cada elemento x_i gerado por $g(x)$, jogamos uma moeda com probabilidade de cara igual a $r(x_i)$. Se sair cara, retemos x_i como um elemento vindo de $f(x)$. Se sair coroa, eliminamos x_i da amostra final. Se começarmos com n elementos retirados de $g(x)$, o tamanho final da amostra é aleatório e geralmente menor que n devido à rejeição de vários elementos.

Y é um valor inicialmente gerado a partir de $g(x)$ e X é um dos valores finalmente aceitos no final do processo.

Algorithm 1 Método da Rejeição.

```

1:  $I \leftarrow \text{True}$ 
2: while  $I$  do
3:   Gere  $Y \sim g(y)$ 
4:   Gere  $U \sim \mathcal{U}(0, 1)$ 
5:   if  $U \leq r(Y) = f(Y)/Mg(Y)$  then
6:      $X \leftarrow Y$ 
7:      $I = \text{False}$ 
8:   end if
9: end while

```

■ **Example 11.7 — Gerando de uma gama.** Queremos gerar $X \sim \text{Gamma}(3, 3)$ com densidade:

$$f(x) = \begin{cases} 0 & , \text{ se } x \leq 0 \\ \frac{27}{2}x^2e^{-3x} & , \text{ se } x \geq 0 \end{cases} \quad (11.1)$$

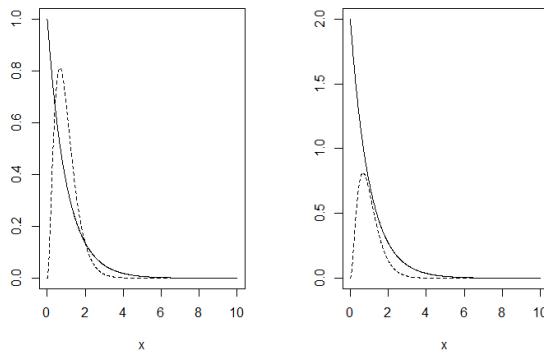


Figure 11.12: Esquerda: Densidade-alvo $f(x)$ (linha tracejada) e densidade $g(x)$ de onde sabemos gerar (linha contínua). Direita: Densidade $f(x)$ e a função $2g(x)$.

Sabemos gerar $W \sim \exp(1)$ pois basta tomar $W = -\log(1-U)$ onde $U \sim \mathcal{U}(0,1)$. A densidade de W é:

$$g(x) = \begin{cases} 0, & \text{se } x < 0 \\ e^{-x}, & \text{se } x \geq 0 \end{cases} \quad (11.2)$$

O suporte das duas distribuições é o mesmo, o semi-eixo real positivo. Ver Figura 11.12. Então:

$$0 \leq \frac{f(x)}{g(x)} = \frac{\frac{27}{2}x^2e^{-3x}}{e^{-x}} = \frac{27}{2}x^2e^{-2x} \quad (11.3)$$

Derivando e igualando a zero temos o ponto de máximo em $x = 1$. Como $\frac{f(1)}{g(1)} = \frac{27}{2}1^2e^{-2} = 1.827 < 2$, temos $f(x) < 2g(x)$ para todo x . Assim, tomamos $M = 2$.

```
set.seed(123); M = 2; nsim = 10000
x = rexp(nsim, 1)
razao = dgamma(x, 3, 3)/(M * dexp(x, 1))
aceita = rbinom(10000, 1, razao)
amostra = x[aceita == 1]
par(mfrow=c(2,1))
xx = seq(0, 4, by=0.1); yy = dgamma(xx, 3, 3)
hist(x, prob=T, breaks=50, xlim=c(0, 8),
     main="f(x) e amostra de g(x)")
lines(xx, yy)
hist(amostra, breaks=20, prob=T, xlim=c(0,8),
     main="f(x) e amostra de f(x)")
lines(xx, yy)
```

O resultado pode ser visto na Figura 11.13. Um script R mais simples que o anterior é o seguinte:

```
set.seed(123)
M = 2; nsim = 10000
x = rexp(nsim, 1)
amostra = x[ runif(nsim)<dgamma(x,3,3)/(M*dexp(x, 1)) ]
```

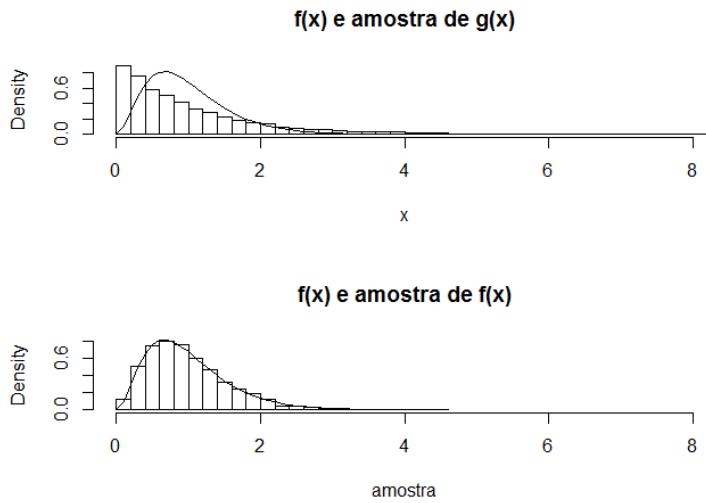


Figure 11.13: Amostra de 10 mil valores de uma $g(x) = \exp(1)$; rejeitando aproximadamente 5000 valores terminamos com amostra de $f(x)$ = Gama(3, 3).

Pseudo-code:

```

1:  $I \leftarrow true$ 
2: while  $I$  do
3:   Selecione  $U \sim \mathcal{U}(0, 1)$ 
4:   Selecione  $U^* \sim \mathcal{U}(0, 1)$ 
5:   Calcule  $\omega = -\log(1 - U)$ 
6:   if  $U^* \leq \frac{f(\omega)}{2g(\omega)} = (27/4)\omega^2 \exp(-2\omega)$  then
7:      $x \leftarrow \omega$ 
8:      $I = False$ 
9:   end if
10: end while

```

11.11.1 Dois teoremas

Temos dois teoremas para este método.

Theorem 11.11.1 — Aceitação-Rejeição gera valores de $f(x)$. A variável aleatória X gerada pelo método de aceitação-rejeição possui densidade $f(x)$.

Prova: Leitura opcional no final deste capítulo.

Theorem 11.11.2 — Impacto de M . O número de iterações necessários até que um valor seja aceito possui distribuição geométrica com valor esperado M .

Prova: Leitura opcional no final deste capítulo.

11.11.2 Sobre o impacto de M

O método de aceitação-rejeição funciona com qualquer M tal que $f(x) \leq Mg(x)$. Suponha que M_1 é muito maior que M_2 , ambos satisfazendo a condição. Se rodarmos o método em paralelo com

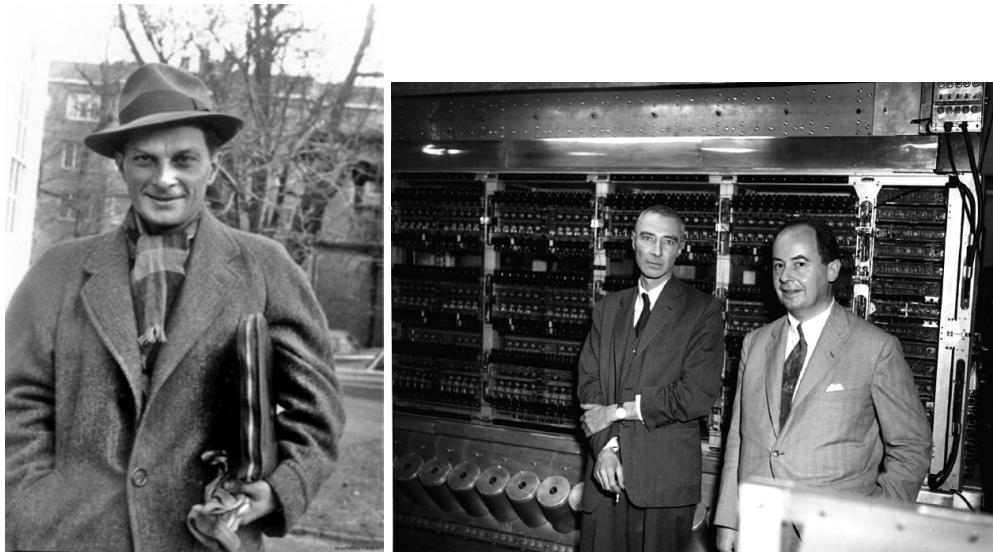


Figure 11.14: Esquerda: Stanislaw Ulam na Polônia. Direita: John von Neumann (à direita) e Robert Oppenheimer em frente a um pequeno pedaço do ENIAC, um dos primeiros computadores do mundo.

os dois valores de M , aquele com o maior valor rejeitaria mais frequentemente que o método com o M menor. Pelo teorema, devemos selecionar, em média, M valores até que aceitemos um deles. Quanto menor M , menos rejeição.

Não é difícil provar que M deve ser maior ou igual a 1. O máximo de eficiência é obtido quando $M = 1$. Mas neste caso, como a área total debaixo de $f(x)$ e $g(x)$ é igual a 1, devemos ter $f(x) = g(x)$. Isto é, a densidade de onde geramos é idêntica à densidade-alvo $f(x)$ e todos os valores são aceitos. É claro que esta não é a situação em que estamos interessados em usar o método de aceitação-rejeição.

Se selecionarmos $g(x)$ muito diferente de $f(x)$, especialmente se tivermos $g(x) \approx 0$ numa região em que $f(x)$ não é desprezível, é possível que tenhamos de usar um valor de M muito grande para satisfazer $f(x) \leq Mg(x)$ para todo x . Esta será uma situação em que o método de aceitação-rejeição será pouco eficiente pois muitas amostras devem ser propostas (em média, M) para que uma delas seja eventualmente aceita).

11.12 História do método Monte Carlo

Os dois métodos principais que vimos neste capítulo, o da transformada inversa e o da rejeição, foram criados praticamente ao mesmo tempo por Stanislaw Ulam (1909-1984, transformada inversa) e John von Neumann (1903-1957, rejeição, e pronuncia-se Nóimánn). A Figura 11.14 mostra estes dois cientistas. Eles haviam trabalhado no projeto Manhattan em Los Alamos, responsável pelo desenvolvimento das armas atômicas nos EUA durante a Segunda Guerra Mundial. Os dois eram matemáticos emigrados, fugindo do terror nazista, e que acharam abrigo nos EUA. Stanislaw Ulam era polonês e John von Neumann, que dispensa apresentações, era húngaro.

O método Monte Carlo tem até uma certidão de nascimento. Após o fim da guerra, Stan Ulam deixa Los Alamos e vai para a California trabalhar como professor. Com dores de cabeça insuportáveis e suspeitando de um câncer e ele se submete a uma cirurgia no cérebro (nas condições daquela época). Os cirurgiões abrem-no e ficam aliviados ao verem que era apenas uma inflamação. Fecham o seu crânio e o põem em repouso. Neste período de convalescência ele pensa no método Monte Carlo em geral como uma maneira de resolver integrais muito complicadas que apareciam

no seu trabalho com a física atômica da época. Ele escreve para seu amigo von Neumann, que responde como uma carta de duas páginas. Ele resume a ideia de Stan Ulam e, em seguida, em 4 linhas, descreve sua ideia do método de rejeição. Esta carta sobreviveu e está na Figura 11.15 e 11.16. Note que, nas últimas linhas da carta, von Neumann menciona que pode ser útil manter seu método em mente, especialmente depois que ENIAC estiver disponível. Nós mantemos este método na nossa mente até hoje, muito tempo depois que o ENIAC desapareceu.

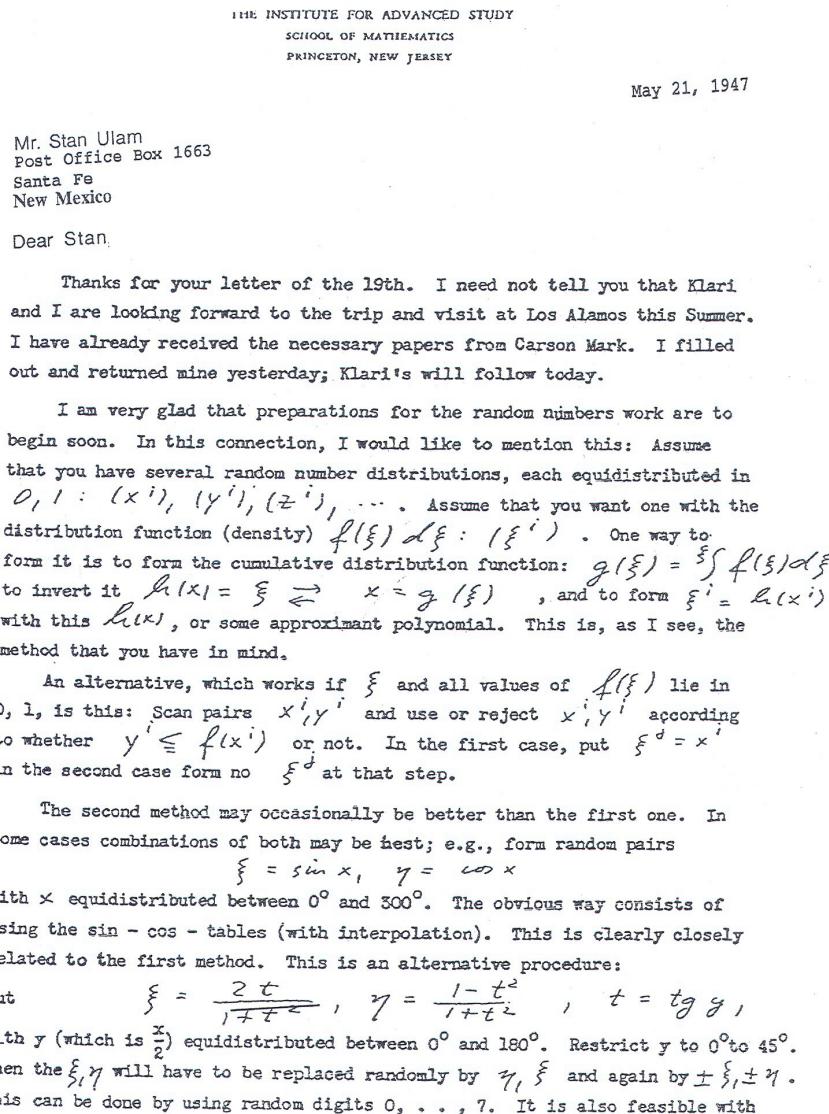


Figure 11.15: Primeira página da carta de von Neumann a Stan Ulam.

.igits 0, . . . , 9:
 random

0	Replace ξ, η by ξ, η
1	" $- \xi, \eta$
2	" $\xi, -\eta$
3	" $-\xi, -\eta$
4	" η, ξ
5	" $\eta, -\xi$
6	" $-\eta, \xi$
7	" $-\eta, -\xi$
8	Reject this digit
9	" " "

Now $t = \tan y$, $0^\circ \leq y \leq 45^\circ$, lies between 0 and 1, and its distribution function is $\frac{dt}{1+t^2}$. Hence one may pick pairs of numbers t, s both (independently) equidistributed between 0 and 1, and then

use t } for $(1+t^2)s \leq 1$
 reject t, s and }
 form no t at } for $(1+t^2)s > 1$
 this step }

Of course, the first pair requires a divider, but the method may still be worth keeping in mind, especially when the ENIAC is available.

* * *

With best regards from house to house.

Yours, as ever,


John von Neumann

JvN:MW

Figure 11.16: Segunda página da carta de von Neumann a Stan Ulam.

11.13 Aplicação em seguros: valor presente atuarial

Esta seção apresenta algumas aplicações de simulação Monte Carlo em problemas de seguros de vida. Ela não se preocupa em explicar os conceitos de atuária ou matemática financeira e pode ser de leitura difícil para quem não conhece os conceitos básicos desses assuntos. Esta seção pode ser omitida sem prejuízo do entendimento do restante do livro.

Suponha que o tempo adicional de vida de um indivíduo (x) possua distribuição $T \sim \exp(\lambda)$ e que desejamos calcular o valor presente atuarial (abreviado como VPA) de um seguro de vida que paga 10 unidades monetárias (abreviado como u.m.) no momento de morte com taxa de juros instantânea δ . Este VPA é denotado por θ e é igual ao valor esperado do valor presente do pagamento do benefício. Ele é dado por

$$\theta = e[g(T)] = E[10e^{-\delta T}] = 10 \int_0^{\infty} e^{-\delta t} \lambda e^{-\lambda t} dt = 10\lambda / (\delta + \lambda)$$

Neste caso, conhecemos uma fórmula para θ e não é necessário estimar por simulação. No entanto, apenas para ilustrar o uso do método, vamos estimar por Monte Carlo o valor de θ .

Uma forma de estimar θ é gerar uma amostra da distribuição da variável de interesse (T , neste caso), e tomar a média da função $g(T)$: Gere um grande número de valores t_1, t_2, \dots, t_n i.i.d. com distribuição $\exp(\lambda)$. A seguir tome a média aritmética dos valores w_1, w_2, \dots, w_n onde $w_i = g(t_i) = 10\exp(-\delta t_i)$. Isto é,

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n (10\exp(-\delta t_i))$$

Exemplo Suponha que $\delta = 0.05$ e que $\lambda = 1/40$. Gerando 1000 valores de uma $\exp(1/40)$ encontramos $\hat{\theta} = 3.279681$. Outras duas simulações adicionais fornecem as estimativas 3.562084 e 3.261478. Estas estimativas estão variando a partir da segunda casa decimal. Isto não é satisfatório, mostra que o erro de estimação pode ser maior que 10% do valor a ser estimado. Podemos aumentar o tamanho da amostra para obter uma estimativa que varie menos. Gerando estimativas com 100 mil valores exponenciais fornecem estimativas que variam a partir da terceira casa decimal. Isto também não é muito satisfatório mas poderá servir se não é necessária muita precisão. Existem métodos mais eficientes que reduzem esta variabilidade tais como o método de amostragem por importâncias. Não veremos estes métodos neste livro. \square

Exercício Suponha que o tempo de vida de uma população feminina segue uma distribuição de Makeham com parâmetros $A = 0.0005$, $B = 0.000075858$, $c = 1.09144$. Uma mulher com idade $x = 43$ anos fica viúva e passa a receber uma anuidade contínua temporária de 25 anos que paga à taxa de 23 u.m. por ano. Seja T o tempo de vida adicional desta mulher e δ a taxa de juros instantânea anual com fator de desconto $v = \exp(-\delta)$. O valor presente atuarial do benefício é

$$\theta = 23E\left(\frac{1-v^T}{\delta}I_{[0,25]}(T)\right) = 23 \int_0^{25} \frac{1-v^t}{\delta} f_T(t) dt$$

onde $I_{[0,25]}(t) = 1$ se $t \in [0, 40]$ e é igual a zero, caso contrário. Estime o valor de θ usando uma amostra de tamanho 30 mil de X com distribuição de Makeham condicionada a $X > 43$.

Solução Gere um grande número n de tempos de vida adicional T a partir de uma Makeham e a seguir use a aproximação

$$\theta \approx \hat{\theta} = \frac{23}{n} \sum_{i=1}^n \left(\frac{1-v^{t_i}}{\delta} \right) I_{[0,25]}(t_i)$$

Exercício Suponha que X tem uma distribuição de Gompertz com parâmetros $B = 0.000072$, $c = 1.087$. Calcule

$$\theta = P(65 < X \leq 85) = E(I_{(65,85]}(T))$$

de forma aproximada.

Solução Gere um grande número n de tempos de vida a partir do nascimento X e a seguir use a aproximação $\theta \approx \hat{\theta} = k/n$ onde k é o número de vezes em que X caiu no intervalo $(65, 85]$ dentre os valores simulados.

Exercício O tempo de vida de uma população segue uma distribuição de Makeham com parâmetros $A = 0.0005$, $B = 0.000075858$, $c = 1.09144$. Um indivíduo com idade $x = 31$ anos faz um seguro de vida temporário que paga um benefício $b(t)$ a seu filho que acabou de nascer se ele falecer t unidades de tempo após a assinatura do contrato. Para evitar riscos de anti-seleção, o contrato estabelece que $b(t) = 0$ se $t < 2$ anos. Para $2 \leq t < 18$, temos $b(t) = 30$ e, se $18 \leq t < 25$, $b(t) = 15 + 15\exp(-0.2(t - 18))$. A seguradora deseja calcular o valor presente atuarial (VPA) desta apólice dado por

$$\theta = \int_0^{25} b(t) v^t f_T(t) dt$$

onde $v = 0.951$. Esboce a função $b(t)$ para você verificar que tipo de benefício está sendo pago. A seguir, use simulação Monte Carlo para calcular o VPA. \square

Exercício No exercício anterior, imagine que a anuidade é variável pagando à taxa de $b(t)$ no instante t (no caso anterior $b(t) = 23$ para todo t). o valor da anuidade do benefício é variável e igual a $b(t) = 15(1 + \cos(t\pi/50))$. Esboce o gráfico de $b(t)$ para $t \in (0, 25)$. Estime o valor presente atuarial desta anuidade:

$$\theta = 15 \int_0^{25} (1 + \cos(t\pi/50)) \frac{1-v^t}{\delta} f_T(t) dt$$

O segundo momento (e o terceiro, quarto, etc.) de variáveis aleatórias contínuas também são integrais que podem ser estimadas por meio de simulação Monte Carlo. Variâncias também podem ser calculadas por Monte Carlo.

Seja X uma variável aleatória contínua com densidade $f_X(x)$ e segundo momento

$$m_2 = E(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x) dx$$

Suponha que X_1, X_2, \dots, X_n seja uma amostra aleatória de variáveis aleatórias i.i.d. com distribuição X . Como $X_1^2, X_2^2, \dots, X_n^2$ é uma amostra de v.a.'s i.i.d., pela Lei dos Grandes Números,

$$\frac{1}{n} (X_1^2 + X_2^2 + \dots + X_n^2) \rightarrow m_2 = E(X^2)$$

quando $n \rightarrow \infty$.

Assim, se n é grande podemos esperar a média dos valores X_i^2 próxima do valor desconhecido m_2 . Assim, uma estimativa para m_2 pode ser $\widehat{m}_2 = (X_1^2 + X_2^2 + \dots + X_n^2)/n$ onde X_1, X_2, \dots, X_n é uma grande amostra da variável X .

A variância de X é $\sigma^2 = \text{Var}(X) = E(X^2) - (E(X))^2 = m_2 - m_1^2$ onde $m_1 = E(X)$. Assim, uma estimativa para σ^2 é

$$\widehat{\sigma}^2 = (X_1^2 + X_2^2 + \dots + X_n^2)/n - ((X_1 + X_2 + \dots + X_n)/n)^2$$

Exercício Suponha que a taxa de juros anual é $\delta = 0.05$ e que $v = \exp(-\delta)$. Gere um grande número de tempos de vida adicionais T para indivíduos que possuem atualmente $x = 35$ anos e cuja idade ao morrer a partir do nascimento possui distribuição de Gompertz com parâmetros $B = 0.0000785$ e $c = 1.0892$. A seguir, calcule valores aproximados para a *esperança* (valor presente atuarial) e a *variância* das seguintes quantidades aleatórias:

- $Y = 10v^T I_{[5,25]}(T)$, valor presente de seguro de vida temporário de 20 anos, diferido de 5 anos.
- $Y = (1 + 3T)v^T$, valor presente de seguro de vida inteira com benefício variável (crescente linearmente com o tempo).
-

$$Y = \begin{cases} 30v^T, & \text{se } T < 10 \\ 30(1 - (T - 10)/30)v^T, & \text{se } 10 \leq T < 20 \\ 20v^T, & \text{se } T \geq 20 \end{cases}$$

11.14 Simulando um fundo de pensão

Esta seção simula um fundo de pensão. Ela pode ser omitida sem prejuízo do entendimento do restante do livro.

Suponha que o tempo total de vida ou idade ao morrer X de uma indivíduo escolhido ao acaso de uma população possui uma distribuição de Gomperz com parâmetros dados por $B = 1.02 \times 10^{-4}$

e $c = 1.0855$. Um grupo de 100 indivíduos desta população, todos com $x = 40$ anos de idade em $t = 0$, estão constituídos num fundo que paga 10 u.m. a um beneficiário no momento em que cada um dos 100 indivíduos falece. No instante $t = 0$, o fundo possui 175 u.m. que vai aplicar a uma taxa de juros de $\delta = 0.06$ ao ano. Como não haverá nenhum aporte adicional de capital a não ser aquele obtido através da aplicação financeira do capital existente hoje, deseja-se saber se existem recursos suficientes para pagar todos os benefícios.

A questão deve ser mais bem especificada. Se todos os indivíduos morrerem logo após o instante $t = 0$, o fundo precisaria de um pouco menos que $10 \times 100 = 1000$ u.m. para honrar seus compromissos, o que é bem menos que seu capital. Assim, existe uma chance de que o fundo não possa cumprir com suas obrigações. Como todos os 100 indivíduos morrerem logo após os seus 40 anos seria um evento raro, esta chance de insolvência seria pequena. Assim, existe uma chance de que o fundo fique insolvente mas é possível que esta chance seja pequena. Deste modo, a pergunta mais apropriada seria: qual a probabilidade de que o fundo eventualmente fique insolvente?

Para responder a isto, precisamos apenas obter o valor presente de todas as obrigações futuras do fundo que é igual a

$$S = \sum_{i=1}^{100} Z_i = \sum_{i=1}^{100} 10 \exp(-0.06 T_i)$$

Se $S > 175$, o fundo não terá como honrar seus compromissos. Se $S \leq 175$, todos os benefícios serão pagos. Só a história futura do fundo poderá dizer qual dos eventos vai realmente ocorrer, se $S > 175$ ou se $S \leq 175$.

Podemos usar a teoria de probabilidades para calcular aproximadamente estas probabilidades. Pelo Teorema Central do Limite, a soma das 100 variáveis aleatórias i.i.d. Z_1, \dots, Z_{100} tem uma distribuição aproximadamente normal e

$$P(S \leq 175) = P\left(\frac{S - 100\mu}{10\sigma} < \frac{175 - 100\mu}{10\sigma}\right) \approx P\left(N(0, 1) < \frac{175 - 100\mu}{10\sigma}\right)$$

onde $\mu = E(Z_i)$ e $\sigma^2 = Var(Z_i)$. A última probabilidade pode ser obtida consultando-se uma tabela de uma normal padrão (ou usando um programa estatístico qualquer) desde que os valores de μ e σ estejam disponíveis.

Para μ temos

$$\mu = \int_0^\infty 10 \exp(-0.06t) f_T(t) dt$$

onde $f_T(t)$ é a densidade de $40 + T$, a idade ao morrer de uma variável Gompertz condicionada a ter o valor maior que 40. Esta integral não é simples de ser calculada mas é muito fácil de ser estimada por simulação Monte Carlo.

Para isto, gere um grande número (digamos, 10 mil) de variáveis T produzindo os valores t_1, \dots, t_{10000} e calcule a média

$$\hat{\mu} = \frac{1}{10000} \sum_{k=1}^{10000} 10 \exp(-0.06t_k)$$

Numa simulação em meu computador obtive $\hat{\mu} = 1.638743$. Com estes mesmos 10 mil números, calcula-se uma estimativa do desvio padrão:

$$\hat{\sigma} = \frac{1}{10000} \sum_{k=1}^{10000} (10 \exp(-0.06t_k)) - (\hat{\mu})^2 = 2.509787$$

Com estes dois valores, podemos então estimar

$$P(S \leq 175) \approx P\left(N(0,1) < \frac{175 - 163.8743}{25.09787}\right) = 0.758747$$

Assim, a probabilidade de insolvência é $1 - 0.758747 = 0.241253$, um valor extremamente elevado.

Vamos chamar o método acima de método 1. Ele depende da aproximação normal dada pelo teorema central do limite. Em outros problemas de atuária, nem sempre será possível usar esta aproximação e assim seria útil ter um método que não dependa do uso do teorema.

Num método 2, vamos usar simulações para gerar várias possíveis histórias do fundo, todas igualmente prováveis, e verificar então qual a chance de insolvência $P(S > 175)$. Para isto, basta gerar várias vezes (digamos 10 mil vezes) um grupo de 100 tempos de vida adicionais e verificar em cada um deles se o evento $S > 175$ ocorreu. A proporção de vezes em que este evento ocorrer nas 10 mil repetições será uma estimativa da probabilidade desejada.

Exercício Estude o seguinte código R e verifique que ele executa o procedimento delineado acima.

```
nrep <- 0; conta <- 0
while(nrep < 10001){
  nrep <- nrep + 1
  T <- (1/log(ce)) * log( 1-log( (1-runif(100))*(1-Fa) )/k ) - a
  Z <- 10*exp(-0.06*T)
  if(sum(Z) > 175) conta <- conta + 1
}
conta/10000      # obtive 0.2217 com minhas 10 mil repeticoes
```

□

Exercício Qual o teorema que justifica o procedimento acima ?

Solução A lei dos grandes números. Para cada uma das 10 mil repetições , defina a variável aleatória W_k que é binária e vale $W_k = 1$ se o evento $S > 175$ ocorre na k -ésima repetição e $W_k = 0$ caso contrário. Assim, $E(W_k) = P(W_k = 1) = P(S > 175)$. Pela lei dos grandes números, a média aritmética das 10 mil variáveis W_k deve ser um valor aleatório próximo da constante $P(S > 175)$:

$$\bar{W} = \frac{1}{10000} \sum_{k=1}^{10000} W_k \approx P(S > 175)$$

□

Esta técnica pode ser usada para estudar várias questões adicionais. Por exemplo, como esta probabilidade de insolvência varia em função da taxa de juros e do valor inicial do fundo em $t = 0$. Ou como ela varia em função dos parâmetros da distribuição de mortalidade Gompertz? Antes de passar a estas outras questões, vamos explorar um pouco mais o R para mostrar como pode ser a evolução histórica de um fundo. Isto vai servir para ilustrar a alta dose de variabilidade, incerteza ou risco que existe nas questões atuariais.

Vamos considerar a evolução temporal do capital depositado no fundo a medida que o tempo passa. Seja $C(t)$ o valor do depositado no fundo no instante t . Este valor depende do número e dos momentos t_1, t_2, \dots em que benefícios foram pagos antes de t e ele pode ser calculado da seguinte maneira:

$$C(t) = \begin{cases} 175 e^{0.06t}, & \text{se } t < t_1 \\ (175 e^{0.06t_1} - 10) e^{0.06(t-t_1)}, & \text{se } t_1 \leq t < t_2 \\ ((175 e^{0.06t_1} - 10) e^{0.06(t_2-t_1)} - 10) e^{0.06(t-t_2)}, & \text{se } t_2 \leq t < t_3 \\ \dots \end{cases}$$

Fazendo as multiplicações em cada segmento de tempo e simplificando, obtemos

$$C(t) = \begin{cases} 175e^{0.06t}, & \text{se } t < t_1 \\ (175 - 10e^{0.06t_1})e^{0.06t}, & \text{se } t_1 \leq t < t_2 \\ (175 - 10e^{0.06t_1} - 10e^{0.06t_2})e^{0.06t}, & \text{se } t_2 \leq t < t_3 \\ \dots & \end{cases}$$

Podemos criar um vetor (chamado cap2) com os valores de $C(t)$ numa grade bem fina de valores t (chamada te) com os seguintes comandos no R:

```
ce <- 1.086; B <- 0.000102; k <- B/log(ce)
a <- 40 ; Fa <- 1 - exp(-k*(ce^a - 1)); capinicial <- 175
# Gerando 100 tempos de vida adicionais e ordenando-os
T <- sort((1/log(ce)) * log(1-log( (1-runif(100))*(1-Fa) )/k ) - a)
tmax <- max(T) # maximo de T foi 56.77 na minha simulacao

te <- seq(0,tmax+1, length=1001) # vetor com eixo "continuo" de tempo
## O PROXIMO COMANDO E' SUTIL, E' O MAIS CRUCIAL
## pos tera' as posicoes no eixo te "continuo" de tempo onde ocorrem
## as mortes
pos <- trunc(T/T[100] * 1001)
## se primeira morte for muito proxima de zero, podemos ter pos[1] zero
pos[pos == 0] <- 1

cap <- rep(0,1001) # vetor para receber os valores do capital no tempo te
cap[1:pos[1]] <- capinicial
for(i in 1:(length(pos)-1)){
  if(pos[i] < pos[i+1]) cap[(pos[i]+1):pos[i+1]] <- cap[pos[i]] - 10 * exp(-0.06 * T[i])
  else cap[pos[i]+1] <- cap[pos[i]] - 10 * exp(-0.06 * T[i])
}
cap2 <- cap * exp(0.06 * te)
```

Queremos agora fazer um gráfico desta evolução temporal de $C(t)$. Em cada instante t , o eixo vertical mostra o valor do capital $C(t)$ depositado no fundo no momento. Os comandos abaixo mostram como obter o gráfico no lado esquerdo da Figura 11.17:

```
## Mostra-se o desenvolvimento do fundo apenas ate o momento em que ele
## fica eventualmente insolvente (enquanto cap2 > 0; no instante seguinte,
## cap2 < 0, o fundo nao possui recursos para pagar o beneficio de 10 u.m.)
aux <- cap2 > 0
plot(te[aux], cap2[aux], type="l",
      xlab="t", ylab="C(t)", xlim=range(te), ylim=c(-50, max(cap2)))
abline(h=0)

## Um grafico mostrando apenas os pagamentos dos 10 primeiros beneficios
plot(te[1:pos[10]], cap2[1:pos[10]], type="l",
      xlab="t", ylab="C(t)", ylim=c(150, max(cap2[1:pos[10]])))
```

O lado direito da Figura 11.17 mostra o desenvolvimento da do gráfico do lado esquerdo apenas até as primeiras 10 mortes. Nesse novo gráfico podemos visualizar melhor os decréscimos constantes de 10 u.m. a cada morte. Em cada instante t_i em que ocorre um falecimento, o fundo

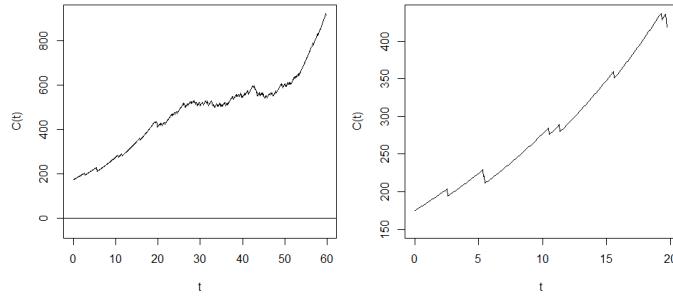


Figure 11.17: Esquerda: Gráfico de $C(t)$ versus t . Direita: Gráfico de $C(t)$ versus t até o pagamentos dos 10 primeiros benefícios. Ver texto para detalhes.

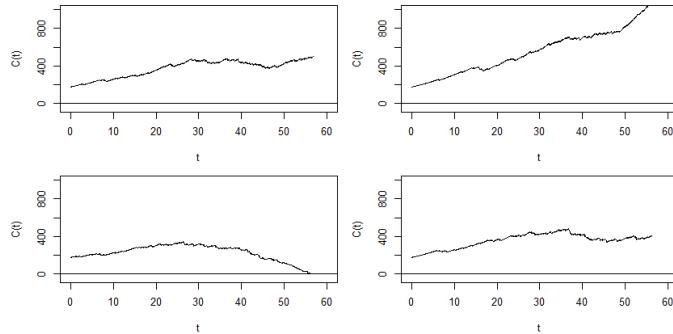


Figure 11.18: Gráficos com 4 desenvolvimentos independentes do fundo, todos gerados nas mesmas condições . O gráfico da segunda linha à esquerda mostra uma situação em que o fundo fica insolvente por volta de $t = 57$.

diminui de 10 u.m. e, a partir do novo patamar alcançado, continua a crescer exponencialmente entre mortes à taxa $\delta = 0.06$.

A Figura 11.18 mostra o desenvolvimento de 4 simulações independentes do fundo. Observe que o terceiro gráfico possui a linha interrompida em $t = 57$ aproximadamente. Neste momento, mais um benefício deveria ser pago mas o fundo não possuía recursos para saldar o compromisso. Ele ficou insolvente. Nos outros gráficos, o fundo não teve problemas para pagar todos os benefícios e ainda terminou com um saldo positivo.

Vamos agora mostrar os comandos para gerar este processo 50 vezes e mostrar a evolução temporal de todos as 50 possíveis realizações do fundo num mesmo gráfico. Para isto, use os comandos abaixo. O resultado está na Figura .

```
# grafico sem exibir nada (opcao type="n"), s\'{o} para criar os eixos
# ele vai receber as linhas a serem geradas
par(mfrow=c(1,1))
plot(c(0,80), c(-50,1000), type="n" , xlab="t", ylab="C(t)"); abline(h=0)
n <- 50
for(i in 1:n){
  T <- sort((1/log(ce)) * log( 1-log( (1-runif(100))*(1-Fa) )/k ) - a)
  tmax <- max(T); te <- seq(0,tmax+1, length=1001)
  pos <- trunc(T/T[100] * 1001); pos[pos == 0] <- 1

  cap <- rep(0,1001); cap[1:pos[1]] <- capinicial
  for(i in 1:(length(pos)-1)){
```

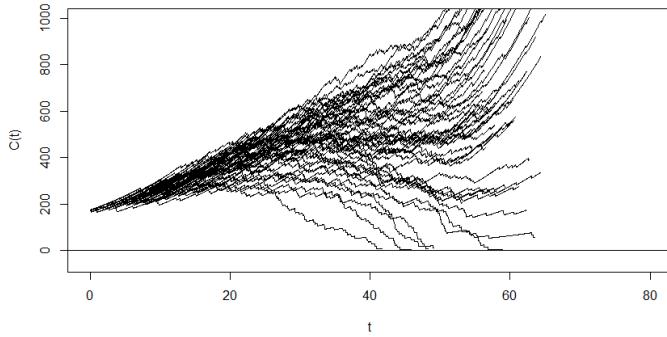


Figure 11.19: Gráfico com 50 desenvolvimentos independentes do fundo $C(t)$, todos gerados nas mesmas condições . Aqueles que ficaram insolventes tiveram suas linhas interrompidas imediatamente antes de tornarem-se negativos.

```

    if(pos[i] < pos[i+1]) cap[(pos[i]+1):pos[i+1]] <- cap[pos[i]] - 10 * exp(-0.06 * T[i])
      else cap[pos[i]+1] <- cap[pos[i]] - 10 * exp(-0.06 * T[i])
  }
  cap2 <- cap * exp(0.06 * te)
  aux <- cap2 > 0
  lines(te[aux], cap2[aux])
}

```

11.15 Processo de Poisson: sinistros no tempo

Sinistros aparecem ao longo do tempo de acordo com um processo de Poisson com taxa constante λ . Isto é, o número médio de sinistros em qualquer intervalo de tempo (a, b) é uma variável de Poisson com média $(b - a)\lambda$. Em particular, se o intervalo possui comprimento $b - a = 1$ então o número médio de sinistros é λ . Observe que não importa onde está o intervalo, seja ele por exemplo $(0, 3)$, $(1, 4)$ ou $(10001, 100004)$, a distribuição é a mesma, uma Poisson com média 3λ . A notação $N(I)$ é usada para a variável aleatória que conta o número de sinistros no intervalo I .

A outra propriedade importante de um processo de homogêneo é que as contagens em intervalos *disjuntos* são variáveis aleatórias independentes. Assim, a contagem $N((0, 1))$ de sinistros no intervalo $(0, 1)$ é independente da contagem $N((1, 2))$ no intervalo $(1, 2)$, mesmo estando um intervalo ao lado do outro. Se $\lambda = 5$, por exemplo, e se observarmos uma contagem $N((0, 1)) = 15$, bem maior que seu valor esperado de $\lambda = 5$, não poderemos prever se a contagem $N((1, 2))$ no intervalo $(1, 2)$ estará acima ou abaixo de sua média.

Outra propriedade que é provada no curso de processos estocásticos é que os tempos entre ocorrências de um processo de Poisson com taxa λ são i.i.d. com distribuição $\exp(\lambda)$. Isto é, seja $T_1 = X_1$ o tempo de espera até a ocorrência do primeiro sinistro, $T_2 = X_1 + X_2$ o tempo de espera até o segundo sinistro, etc, de modo que X_i é o tempo entre o $(i-1)$ -ésimo e o i -ésimo sinistros. Por convenção, $T_1 = X_1$ é o tempo entre a origem $t = 0$ e o primeiro evento. Então X_1, X_2, X_3, \dots são i.i.d. $\exp(\lambda)$.

Sabemos como gerar os X_i 's pelo método da transformação inversa: $X_i = -1/\lambda \log(U_i)$ onde U_1, U_2, \dots são i.i.d. com distribuição $U(0, 1)$. Para gerar eventos de um processo de Poisson com taxa $\lambda = 2$ no intervalo $[0, 12]$ usamos o algoritmo abaixo, resultado na Figura 11.20.

```

lambda <- 2; tf <- 12; t <- 0; i <- 0
s <- numeric(); n <- numeric()

```

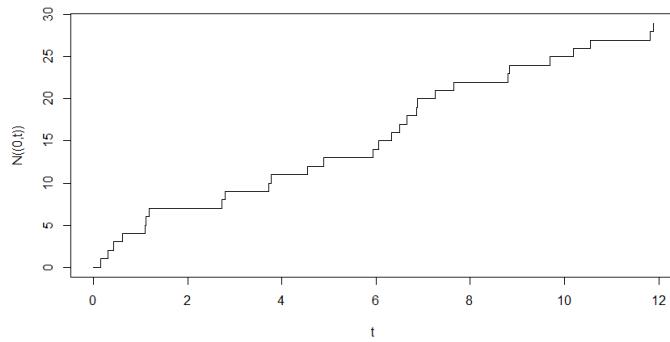


Figure 11.20: Gráfico com uma realização de um processo de Poisson de taxa $\lambda = 2$ em $(0, 12)$. Para cada tempo t , o eixo vertical mostra o número total $N([0, t])$ de eventos que ocorreram até o instante t .

```

while(t <= tf){
  s <- c(s,t); n <- c(n,i); i <- i+1
  t <- t - (1/lambda)*log(runif(1))
}
plot(s, n, type="s", xlab="t", ylab="N((0,t))")

```

Exercício Considerando a realização do processo de Poisson de taxa $\lambda = 2$ da Figura 11.20 responda: a realização gerou um número de eventos maior, menor ou igual que o número esperado no intervalo $[0, 12]$? Quantos eventos ocorreram no intervalo $(4, 6)$ e quantos eram esperados? \square .

11.15.1 Outra abordagem

O algoritmo acima gera os eventos sequencialmente. Uma outra abordagem gera todos os eventos de uma única vez usando uma propriedade do processo de Poisson. Sejam $T_1 = X_1$, $T_2 = X_1 + X_2$, $T_3 = X_1 + X_2 + X_3$, etc. Dado que existem $N([0, t_F]) = n$ eventos no intervalo $[0, t_F]$, os tempos desses n eventos distribuem-se entre 0 e t_F como n variáveis aleatórias i.i.d. com distribuição $U(0, t_f)$. Assim, os tempos *ordenados* T_1, T_2, \dots, T_n são as estatísticas de ordem (os valores ordenados) de n variáveis i.i.d. $U(0, t_f)$.

Para gerar os eventos basta então usar o seguinte algoritmo que explora esta propriedade:

```

tf <- 12; lambda <- 2
ntf <- rpois(1, lambda = 12*2)
tempos <- c(0, sort(tf * runif(ntf)))
plot(tempos, 0:ntf, type="s", xlab="t", ylab="N((0,t))")

```

11.15.2 Processo de Poisson não-homogêneo

Em geral, a taxa de ocorrência de eventos não é constante no tempo. Ela varia suavemente no tempo e é representada pela função $\lambda(t)$. A interpretação desta função é que, num pequeno intervalo de tempo $[t, t + dt]$, o número *esperado* de eventos é dado por $E(N([t, t + dt])) \approx \lambda(t) dt$. Assim, o número esperado de eventos *por unidade de tempo* em torno de t é dado por $E(N([t, t + dt]))/dt \approx \lambda(t)$. Por exemplo, se $\lambda(3.3) = 5$ para $t = 3.3$, então o número médio de eventos num pequeno intervalo $[3.3, 3.3 + 0.1]$ é dado aproximadamente por $E(N([3.3, 3.4])) \approx 50.1 = 0.5$, meio evento no intervalo de comprimento 0.1. Por outro lado o número esperado *por unidade de tempo* na vizinhança de $t = 3.3$ é aproximadamente $E(N([3.3, 3.4]))/0.1 \approx \lambda(3.3) = 5$.

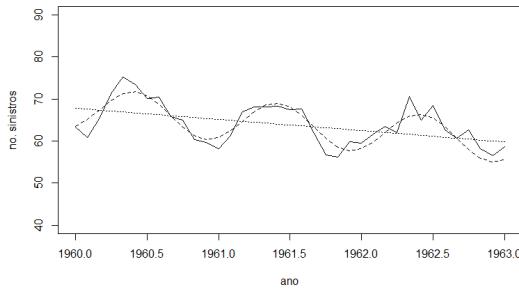


Figure 11.21: Gráfico com dados mensais de número de acidentes com motocicletas ao longo do tempo. Ver texto para detalhes.

Alguns riscos são claramente sazonais, ocorrendo mais intensamente em certas épocas do ano tais como no verão (ou no inverno) ou no início do ano (período de férias). Além disso, é comum a existência de tendências de crescimento histórico que perduram por longos períodos. Finalmente, existem também ciclos, movimentos oscilatórios mas que não são observados com regularidade tais como os movimentos sazonais. Modelos para este tipo de dados são estudados em séries temporais.

A Figura 11.21 é adaptada de [BeardBook1984] e mostra dados de acidentes de trânsito com motocicletas entre os anos de 1960 a 1962 em Londres. O eixo vertical mostra o número de acidentes enquanto o eixo horizontal é o eixo do tempo. Dados mensais são conectados numa linha ziguezageada mostrando a flutuação das contagens mensais ao longo do período. Uma linha reta mostra a tendência histórica de decrescimento neste período de três anos e a curva senoidal mostra o movimento sazonal dos sinistros.

Essa figura mostra que não é razoável esperar que os sinistros ocorram a uma taxa constante no tempo. Assim, suponha que os sinistros ocorram como um processo de Poisson *não-homogêneo* com taxa

$$\lambda(t) = 68 - 0.22t + 5\cos(\pi(1+t/6)) \quad (11.4)$$

onde t é medido em meses com a convenção de que $t = 0$ é o dia 01/01/1960.

Um processo de Poisson não-homogêneo com taxa $\lambda(t)$ é definido como o único processo pontual em que as contagens em intervalos disjuntos de tempo são independentes e em que o número aleatório de eventos num intervalo $[a, b]$ é uma variável de Poisson com valor esperado igual a $\int_a^b \lambda(t)dt$.

Um método prático para gerar um processo de Poisson não-homogêneo num intervalo de tempo $[0, t_f]$ é o de afinamento ou emagrecimento (*thining*, em inglês). Suponha que $\lambda(t) < k$ para todo $t \in [0, t_F]$. Por exemplo, em (11.4), $\lambda(t) < 72$ para todo $t \in [0, 37]$.

A seguir, gere um processo de Poisson *homogêneo* com taxa k em $[0, t_F]$ obtendo os tempos $0 < t_1 < t_2 < \dots < t_n < t_F$. Para cada evento gerado:

- $p_i = \lambda(t_i)/k \in (0, 1)$
- retenha t_i com probabilidade p_i e apague-o com probabilidade $1 - p_i$.

Os eventos retidos no final formam um processo de Poisson não-homogêneo com intensidade $\lambda(t)$ (interessados podem ver a demonstração em livros de processos estocásticos).

Assim, a geração pode ser feita por meio dos seguintes comandos em R, com o resultado mostrado na Figura 11.22:

```
t <- 0; i <- 0; tf <- 3; s <- 0; k <- 72
lambdat <- function(t){
  68 - 0.22*t + 5*cos(pi * (t + 1))
}
```

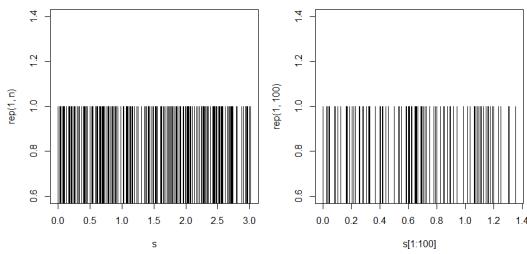


Figure 11.22: Processo pontual de Poisson com posterior afinamento.

```

while(t < tf){
  t <- t - (1/k) * log(runif(1))
  if(runif(1) <= lambdat(t)/k) i <- i+1; s <- c(s,t)
}

n <- length(s)
par(mfrow=c(1,2))
plot(s, rep(1,n), type="h")
plot(s[1:100], rep(1,100), type="h")

```

Exercício Suponha que $\lambda(t) = 3 + 5 \exp(\cos((t+1)/2))$ no intervalo $t \in (0, 36)$. Escreva linhas de código R para gerar um processo de Poisson não-homogêneo com esta intensidade. Faça um gráfico para visualizar a intensidade e os eventos gerados.

11.16 Provas dos teoremas: opcional

Este seção apresenta uma demonstração de que o algoritmo do método de aceitação-rejeição de fato simula variáveis aleatórias com a distribuição desejada. Esta demonstração depende da regra da probabilidade total, assunto coberto no capítulo de distribuições multivariadas.

Vamos denotar por Y um valor qualquer inicialmente gerado a partir de $g(x)$ e por X um dos valores finalmente aceitos no final do processo. O algoritmo de aceitação-rejeição é o seguinte:

Algorithm 2 Método da Rejeição.

- 1: $I \leftarrow \text{True}$
- 2: **while** I **do**
- 3: Gere $Y \sim g(y)$
- 4: Gere $U \sim \mathcal{U}(0, 1)$
- 5: **if** $U \leq r(Y) = f(Y)/Mg(Y)$ **then**
- 6: $X \leftarrow Y$
- 7: $I = \text{False}$
- 8: **end if**
- 9: **end while**

Assuma que a distribuição acumulada associada com $f(x)$ e $g(x)$ é igual a $F(x)$ e $G(x)$, respectivamente. Isto é, $f(x) = F'(x)$ e $g(x) = G'(x)$.

Vamos usar a regra da probabilidade total: para qualquer evento B e qualquer variável aleatória contínua Y com densidade $g(y)$, podemos escrever

$$\mathbb{P}(B) = \int \mathbb{P}(B|Y = y)g(y)dy.$$

Theorem 11.16.1 — Aceitação-Rejeição gera valores de $f(x)$. A variável aleatória X gerada pelo método de aceitação-rejeição possui densidade $f(x)$. Além disso, o número de iterações necessários até que um valor seja aceito possui distribuição geométrica com valor esperado M .

Proof. Vamos mostrar que $\mathbb{P}(X \leq x) = F(x)$. Ou seja, a variável aleatória gerada X possui a distribuição acumulada $F(x)$ e portanto, possui a densidade $f(x)$ associada a $F(x)$. Vamos inicialmente calcular

$$\begin{aligned} \mathbb{P}(Y \leq t \mid Y \text{ gerado é aceito}) &= \mathbb{P}\left(Y \leq t \mid U \leq \frac{f(Y)}{Mg(Y)}\right) \\ &= \frac{\mathbb{P}\left(Y \leq t \text{ and } U \leq \frac{f(Y)}{Mg(Y)}\right)}{\mathbb{P}\left(U \leq \frac{f(Y)}{Mg(Y)}\right)} \end{aligned} \quad (11.5)$$

Denote por B o evento $B = [Y \leq t \text{ and } U \leq \frac{f(Y)}{Mg(Y)}]$ do numerador. Aplicando a regra da probabilidade total a este evento, temos

$$\mathbb{P}\left(Y \leq t \text{ and } U \leq \frac{f(Y)}{Mg(Y)}\right) = \int_{-\infty}^{\infty} \mathbb{P}\left(Y \leq t \text{ and } U \leq \frac{f(Y)}{Mg(Y)} \mid Y = z\right) g(z) dz$$

Por causa da condição $Y = z$, temos que

$$\mathbb{P}(Y \leq t \mid Y = z) = \begin{cases} 0, & \text{se } t < z \\ 1, & \text{se } t \geq z \end{cases}$$

Temos também que

$$\mathbb{P}\left(Y \leq t \text{ and } U \leq \frac{f(Y)}{Mg(Y)} \mid Y = z\right) = \begin{cases} 0, & \text{se } t < z \\ \mathbb{P}\left(U \leq \frac{f(z)}{Mg(z)}\right) = \frac{f(z)}{Mg(z)}, & \text{se } t \geq z \end{cases}$$

Assim, o numerador de 11.5 é igual a

$$\begin{aligned} \int_{-\infty}^{\infty} \mathbb{P}\left(Y \leq t \text{ and } U \leq \frac{f(Y)}{Mg(Y)} \mid Y = z\right) g(z) dz &= \int_{-\infty}^t \frac{f(z)}{Mg(z)} g(z) dz \\ &= M^{-1} \int_{-\infty}^t f(z) dz \\ &= M^{-1} F(t) \end{aligned}$$

O denominador de 11.5 é calculado de modo semelhante:

$$\begin{aligned} \mathbb{P}\left(U \leq \frac{f(Y)}{Mg(Y)}\right) &= \int_{-\infty}^{\infty} \mathbb{P}\left(U \leq \frac{f(Y)}{Mg(Y)} \mid Y = z\right) g(z) dz \\ &= \int_{-\infty}^{\infty} \frac{f(z)}{Mg(z)} g(z) dz \\ &= \frac{1}{M} \int_{-\infty}^{\infty} f(z) dz \\ &= \frac{1}{M} \quad \text{pois } f(x) \text{ é uma densidade e integra 1} \end{aligned}$$

Portanto, retornando a 11.5, podemos escrever

$$\mathbb{P}(Y \leq t \mid Y \text{ gerado é aceito}) = \frac{F(t)/M}{1/M} = F(t)$$

O valor aleatório X da saída é o valor aleatório Y gerador por g dado que este Y foi aceito. Assim, o resultado acima está mostrando que X é um valor aleatório com distribuição acumulada $F(x)$ e portanto, com densidade $f(x)$. Em resumo, o valor X que termina sendo aceito no método possui densidade $f(x)$. ■ ■

O teorema abaixo resume o processo de rejeição até que um valor seja aceito e o papel de M neste processo.

Theorem 11.16.2 — Impacto de M. O número de iterações necessárias até que um valor seja aceito possui distribuição geométrica com valor esperado M .

Proof. O processo de aceitação-rejeição pode ser pensado da seguinte forma. Considere uma sequência i.i.d. de pares de variáveis (U_i, Y_i) com $i = 1, 2, \dots$. A variável Y_i é gerada através de $g(x)$ e U_i possui distribuição $U(0, 1)$, independente de Y_i . Seja I_i uma variável indicadora valendo 0 se Y_i for rejeitada e valendo 1 caso contrário. O número de simulações necessárias até que um valor seja aceito é o número de simulações necessárias para a aparição do primeiro valor 1 para I_i . Calculamos anteriormente a probabilidade de que $I_i = 1$:

$$\mathbb{P}(I_i = 1) = \mathbb{P}\left(U_i \leq \frac{f(Y_i)}{Mg(Y_i)}\right) = \frac{1}{M}$$

Este valor é constante, não depende de i .

Assim, queremos obter a distribuição do número de simulações necessárias para a aparição do primeiro sucesso (o valor 1) quando temos variáveis binárias I_1, I_2, \dots i.i.d. com probabilidade de sucesso constante e igual a $1/M$. Por definição, esta é a distribuição geométrica com valor esperado M . ■ ■

Assim, se adotarmos certo valor M , vamos selecionar, em média, M valores para cada valor aceito.



12. Vetores Aleatórios

12.1 Introdução

Neste capítulo vamos lidar um vetor de v.a.'s, e não com uma única v.a. Teremos $\mathbf{X} = (X_1, \dots, X_k)$, um vetor aleatório de dimensão k . Cada uma das entradas X_i do vetor \mathbf{X} é uma variável aleatória medida no mesmo resultado ω do experimento estocástico. A importância vital de se lidar com vetores aleatórios é que uma v.a. (uma das entradas no vetor) vai dar alguma informação sobre o valor de outra v.a. (outra entrada do vetor).

O arcabouço matemático é o seguinte. Temos um espaço amostral Ω com uma medida de probabilidade sobre subconjuntos \mathcal{A} de Ω . Ω é “complexo”: cada resultado do experimento aleatório pode ter muitas características de interesse: X_1, X_2, X_3, \dots Coletamos estas várias medidas num vetor aleatório $\mathbf{X} = (X_1, \dots, X_n)$. As variáveis X_1, X_2, X_3, \dots são medições feitas no *mesmo resultado* $\omega \in \Omega$ do experimento. Ver Figura 12.1.

■ **Example 12.1 — Imagem como vetor.** Seja $\Omega = \{ \text{imagens } n \times m \}$. Selecione ao acaso uma das imagens tal como aquela na Figura 12.2. Características de interesse: intensidade de cinza em cada um dos pixels da imagem. Assim, $\mathbf{X} = (X_{11}, X_{12}, \dots, X_{nm})$ Veja que todas as medições são sobre um mesmo resultado do experimento: a imagem selecionada. ■

■ **Example 12.2 — Vértices e vetores.** Considere uma rede social vista como um grafo, como na Figura 12.2. Os vértices são os usuários e as arestas direcionadas são as relações de seguidor-seguido. O experimento consiste em selecionar um vértice ao acaso. Ω é a coleção de vértices e arestas do grafo com suas muitas características associadas. Suponha que existam k características de interesse do vértice selecionado:

- X_1 = idade do nó (intrínseco ao nó)
- X_2 = número de outlinks (relacional)
- X_3 = número de inlinks (relacional)

Assim, $\mathbf{X} = (X_1, X_2, X_3)$. O objetivo é descobrir qual é a relação probabilística entre o número de outlinks do nó com a idade do nó. ■

■ **Example 12.3 — problema de classificação supervisionada.** Ω é a coleção de itens classificados em dois (ou mais) grupos. Por exemplo:

O modelo teórico $X = (X_1, X_2, \dots, X_n)$

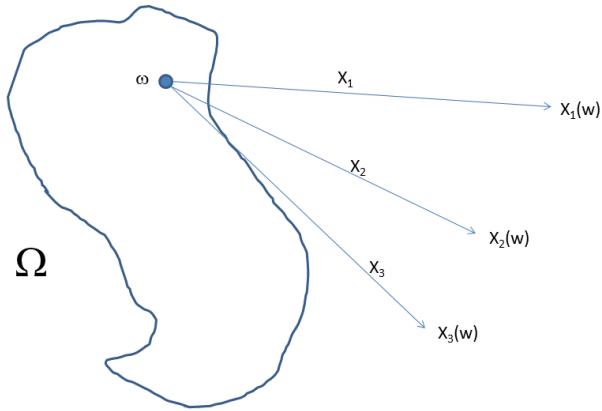


Figure 12.1: O arcabouço teórico para vetores aleatórios.



Figure 12.2: Esquerda: Imagens como vetores. Direita: Características de vértices em redes sociais como vetores.



Figure 12.3: Esquerda: E-mails e spams. Direita: Predizendo o preço de imóveis.

- Coleção de e-mails: spam versus não-spam;
- Coleção de tomadores de empréstimos num banco: pagam versus não pagam de volta o empréstimo dentro do prazo;
- Crâneos humanos em escavação arqueológica: masculinos versus femininos.

A coleção pode nem existir ainda de forma completa. Por exemplo, para a detecção de spams (Figura 12.3), nosso interesse reside nos e-mails já enviados mas principalmente nos que ainda serão enviados no futuro.

Em cada item da coleção, medimos dois tipos de variáveis aleatórias. No caso de spam *versus* não-spam, temos Y , uma v.a. binária: 1 se o e-mail é um spam, e 0 se não-spam. No caso do risco de crédito, temos o tomador de empréstimo como inadimplente ou não. Para os crâneos, os dois sexos, masculino ou feminino.

Além dessa variável Y binária representando a classe do item, temos outras v.a.'s representando atributos adicionais do mesmo item. Por exemplo, no caso do spam, imagine que temos um conjunto de $k = 3$ atributos medido em cada e-mail:

- X_1 : número de vezes em que aparece a palavra “sale”;
- X_2 : número de vezes em que aparece a palavra “offer”;
- X_3 : número de vezes em que aparece a palavra “Viagra”.

O vetor aleatório final combina a variável binária e os atributos: $\mathbf{X} = (Y, X_1, X_2, X_3)$ O objetivo é predizer o valor de Y a partir dos atributos. Qual o valor da probabilidade condicional $\mathbb{P}(Y = \text{spam} | X_1 = 3, X_2 = 1, X_3 = 3)$? Como esta probabilidade muda quando alteramos alguns dos atributos X ? Não esperamos que $\mathbb{P}(Y = \text{spam} | X_1 = 0, X_2 = 0, X_3 = 0)$ seja igual a $\mathbb{P}(Y = \text{spam} | X_1 = 5, X_2 = 3, X_3 = 3)$. Mas como ocorre esta mudança, quais são os valores envolvidos? ■

■ **Example 12.4 — Regressão e o preço de imóveis.** Alguns apartamentos custam 200 mil reais, outros curtam 10 vezes mais (Figura 12.3). O que faz com que os preços Y de apartamentos variem tanto? Os corretores de imóveis dizem que existem três aspectos fundamentais determinando o preço de um imóvel. Em primeiro lugar, sua localização. Em segundo lugar, sua localização, e em terceiro também. Depois vêm os demais aspectos tais como área, idade do imóvel, etc.

Para um imóvel escolhido ao acaso numa região, sejam X_1 sua localização, X_2 a idade, X_3 a área, X_4 o número de quartos, e X_5 uma indicadora de que o prédio possui piscina. O vetor aleatório é $\mathbf{X} = (Y, X_1, X_2, X_3, X_4, X_5)$. O interesse principal é conhecer a distribuição da variável aleatória Y o preço do imóvel *condicionado* no valor das demais variáveis. Por exemplo, deseja-se saber a distribuição da v.a.

$$(Y | X_1 = \text{Sion}, X_2 = 10 \text{ anos}, X_3 = 200m^2, X_4 = 4, X_5 = \text{não})$$

Quais os valores típicos de Y dado que os valores do vetor \mathbf{X} estão fixados nestes valores? Qual a chance de $Y > 500$ mil quando $X_1 = \text{Sion}, X_2 = 10 \text{ anos}, X_3 = 200m^2, X_4 = 4, X_5 = \text{não}$? E como esta distribuição de Y muda quando alteramos alguns dos atributos em \mathbf{X} ? ■

A principal ideia de trabalhar com vetores aleatórios é explorar interrelações entre as variáveis componentes do vetor. Se $\mathbf{X} = (X_1, X_2, \dots, X_k)$, podemos analisar cada v.a. separadamente das demais e ajustar um modelo a cada uma delas, seja uma binomial, uma Poisson, exponencial, normal, etc. Isto é chamado de análise *marginal*. É o que viemos fazendo até agora. O mais interessante é quando analisamos as variáveis *conjuntamente*. A análise conjunta procura explorar a existência de relações probabilísticas entre as variáveis.

12.2 Conjunta discreta

Se $\mathbf{X} = (X_1, X_2, \dots, X_k)$ é composto apenas de v.a.'s discretas, a distribuição conjunta das v.a.'s é muito simples. Como no caso de apenas uma v.a. discreta, precisamos apenas especificar uma lista de valores possíveis para o vetor \mathbf{X} e a lista de probabilidades associadas. Se X_i tem m_i valores possíveis, a lista de valores possíveis do vetor discreto $\mathbf{X} = (X_1, X_2, \dots, X_k)$ terá $m_1 \times m_2 \times \dots \times m_k$ possibilidades. Basta agora atribuir uma probabilidade ≥ 0 a cada um deles de forma que somem 1. Todas as probabilidades de interesse são obtidas a partir desta lista de probabilidades básicas.

■ **Example 12.5 — Exemplo muito simples.** Seja Ω o conjunto de pacientes em visita ao otorinolaringologista com problemas na garganta (faringoamigdalite aguda). Incluímos em Ω os pacientes do futuro. Em geral, estes pacientes têm a suspeita da presença de infecção pela bactéria *estreptococcus*. Existem dois tipos de testes em cada paciente:

- Teste padrão-ouro, cultura em placa agar-sangue: resultado positivo ou negativo.
- Teste rápido, barato com resultados positivo ou negativo MAS com menor qualidade.

Vamos discutir a validação de um teste diagnóstico. Considere o vetor $\mathbf{X} = (TO, TR)$ onde TO significa um teste padrão-ouro e TR , um teste rápido. A v.a. TO possui dois valores: 0 ou 1. A v.a. TR também possui dois valores: 0 ou 1. O vetor \mathbf{X} possui 4 resultados possíveis

Teste-ouro	Teste-rápido	Probabilidade
0	0	?
0	1	?
1	0	?
1	1	?

As probabilidades $\mathbb{P}(TO = x_1, TR = x_2)$ dos 4 resultados possíveis devem ser maiores que (ou iguais a) zero. Por exemplo, uma atribuição válida de probabilidades é a que está na tabela seguinte. Esta tabela fornece a distribuição conjunta do vetor $\mathbf{X} = (TO, TR)$.

Teste-ouro	Teste-rápido	Probabilidade
0	0	0.40
0	1	0.19
1	0	0.03
1	1	0.38
Total		1

■

12.3 Marginal discreta

■ **Definition 12.3.1 — Distribuição Marginal.** Se $\mathbf{X} = (X_1, X_2, \dots, X_k)$ é composto apenas de v.a.'s discretas, a *distribuição marginal* de uma v.a. X_i é a distribuição dessa única v.a., dada por $\mathbb{P}(X_i = x)$, ignorando os valores das demais v.a.'s.

Como obter $\mathbb{P}(X_i = x)$ a partir da distribuição conjunta do vetor $\mathbf{X} = (X_1, X_2, \dots, X_k)$? A distribuição marginal de uma v.a. X_i de um vetor discreto é a soma das probabilidades conjuntas sobre todos os valores das outras variáveis.

Definição 12.3.2 — Distribuição Marginal - 2. Se $\mathbf{X} = (X_1, X_2, \dots, X_k)$ é composto apenas de v.a.'s discretas, a *distribuição marginal* de uma v.a. X_i é dada por

$$\mathbb{P}(X_i = x) = \sum_S \mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_i = x, \dots, X_k = x_k)$$

onde a soma é sobre todos os valores possíveis de todas as v.a.'s exceto X_i , que tem seu valor fixo em x , representados no conjunto S .

■ **Example 12.6 — De volta ao exemplo muito simples.** A distribuição conjunta do vetor $\mathbf{X} = (TO, TR)$ foi obtida em 12.5. Vamos obter a distribuição marginal da v.a. TO . Para isto, precisamos de $\mathbb{P}(TO = 0)$ e de $\mathbb{P}(TO = 1)$

$$\begin{aligned}\mathbb{P}(TO = 0) &= \mathbb{P}(TO = 0 \wedge TR = 0) + \mathbb{P}(TO = 0 \wedge TR = 1) \\ &= 0.40 + 0.19 = 0.59 \\ \mathbb{P}(TO = 1) &= \mathbb{P}(TO = 1 \wedge TR = 0) + \mathbb{P}(TO = 1 \wedge TR = 1) \\ &= 0.03 + 0.38 = 0.41\end{aligned}$$

Para obter a distribuição marginal da v.a. TR , precisamos de $\mathbb{P}(TR = 0)$ e de $\mathbb{P}(TR = 1)$:

$$\begin{aligned}\mathbb{P}(TR = 0) &= \mathbb{P}(TR = 0 \wedge TO = 0) + \mathbb{P}(TR = 0 \wedge TO = 1) \\ &= 0.40 + 0.03 = 0.43 \\ \mathbb{P}(TR = 1) &= \mathbb{P}(TR = 1 \wedge TO = 0) + \mathbb{P}(TR = 1 \wedge TO = 1) \\ &= 0.19 + 0.38 = 0.57\end{aligned}$$

Na prática, como $\mathbb{P}(TR = 1) = 1 - \mathbb{P}(TR = 0)$, basta obtermos uma delas, a outra sendo obtida por subtração. ■

12.4 Independência de duas v.a.'s

Duas v.a.'s discretas X e Y são *independentes* se os eventos $[X = x]$ e $[Y = y]$ são independentes para qualquer combinação de x e y . Isto é,

Definição 12.4.1 — Independência no caso discreto. As v.a.'s discretas X e Y são independentes se

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \mathbb{P}(Y = y) \quad (12.1)$$

para todo par (x, y) .

Theorem 12.4.1 — Definição equivalente de independência. Se X e Y são independentes se, e somente se,

$$\mathbb{P}(X = x | Y = y) = \mathbb{P}(X = x) \quad (12.2)$$

para todo par (x, y) .

Este teorema mostra que podemos definir a independência de v.a.'s discretas de uma maneira (como em 12.1) ou de outra (como em 12.2). Uma implica a outra. Para provar o resultado, vamos

começar assumindo que (12.1) é válido para todo par (x, y) . Então, pela definição de probabilidade condicional, podemos concluir que

$$\mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{\mathbb{P}(X = x) \mathbb{P}(Y = y)}{\mathbb{P}(Y = y)} = \mathbb{P}(X = x)$$

e portanto (12.2) é válida se (12.1) é válida.

Por outro lado, se (12.2) é válida para todo par x e y então, usando novamente a definição de probabilidade condicional, concluímos que

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x|Y = y) \mathbb{P}(Y = y) = \mathbb{P}(X = x) \mathbb{P}(Y = y)$$

e (12.2) sendo válida implica que (12.1) é válida também. Em suma, podemos definir a independência de v.a.'s discretas como (12.2) ou (12.1).

■ **Example 12.7 — TO e TR não são independentes.** Apresentamos anteriormente no exemplo 12.5 a distribuição conjunta do vetor $\mathbf{X} = (TO, TR)$. Vamos verificando dois casos:

$$\mathbb{P}(TO = 1, TR = 1) = 0.38 \neq 0.23 = (0.03 + 0.38) \times (0.19 + 0.38) = \mathbb{P}(TO = 1) \mathbb{P}(TR = 1)$$

$$\mathbb{P}(TO = 0, TR = 0) = 0.40 \neq 0.25 = (0.40 + 0.19) \times (0.40 + 0.03) = \mathbb{P}(TO = 0) \mathbb{P}(TR = 0)$$

Se TO e TR fossem independentes, $TO = 1$ ocorreria junto com $TR = 1$ apenas 23% das vezes mas eles ocorrem juntos 38% de acordo com a tabela. $TO = 0$ e $TR = 0$ ocorreriam juntos 25% das vezes se independentes mas a tabela fornece 40%. Os dois testes tendem a concordar muito mais frequentemente do que se fossem independentes. Isto é esperado, claro. Os dois testes servem para diagnosticar a mesma doença. Mesmo o teste rápido TR sendo pior que o teste-ouro TO , os dois testes devem ter uma concordância além do mero acaso. ■

■ **Example 12.8 — Defeitos em produtos.** Imagine que aparelhos eletrônicos saindo da linha de produção podem ter dois tipos de defeitos: defeitos graves (que inviabilizam o seu uso) e defeitos menores (de acabamento, que permitem seu uso normal). Sejam as v.a.'s X e Y que indicam a presença ou não desses defeitos num produto ω . Se a distribuição conjunta de X e Y é dada na tabela abaixo, mostre que elas são independentes:

X	Y	$\mathbb{P}(X = x, Y = y)$
0	0	0.075
0	1	0.225
1	0	0.175
1	1	0.525
Total		1

Primeiro, obtemos as marginais:

$$\mathbb{P}(X = 0) = \mathbb{P}(X = 0 \wedge Y = 0) + \mathbb{P}(X = 0 \wedge Y = 1) = 0.075 + 0.225 = 0.300$$

$$\mathbb{P}(Y = 0) = \mathbb{P}(Y = 0 \wedge X = 0) + \mathbb{P}(Y = 0 \wedge X = 1) = 0.075 + 0.175 = 0.250$$

com $\mathbb{P}(X = 1) = 1 - \mathbb{P}(X = 0) = 0.700$ e $\mathbb{P}(Y = 1) = 1 - \mathbb{P}(Y = 0) = 0.750$. Nos quatro casos possíveis para X e Y , temos a probabilidade conjunta como o produto das marginais.

$$\mathbb{P}(X = 0, Y = 0) = 0.075 = 0.30 \times 0.250 = \mathbb{P}(X = 0) \mathbb{P}(Y = 0)$$

$$\mathbb{P}(X = 0, Y = 1) = 0.225 = 0.30 \times 0.750 = \mathbb{P}(X = 0) \mathbb{P}(Y = 1)$$

$$\mathbb{P}(X = 1, Y = 0) = 0.175 = 0.70 \times 0.250 = \mathbb{P}(X = 1) \mathbb{P}(Y = 0)$$

$$\mathbb{P}(X = 1, Y = 1) = 0.525 = 0.70 \times 0.750 = \mathbb{P}(X = 1) \mathbb{P}(Y = 1)$$

Precisamos verificar, como fizemos aqui, a igualdade da conjunta como produto das marginais para todos os valores de X e Y .

■

12.5 Marginal discreta com várias v.a.'s

A definição de independência estende-se para mais de duas variáveis discretas. Por exemplo, suponha que $\mathbf{X} = (X_1, X_2, X_3, X_4)$ onde

- X_1 = diagnóstico de uma doença, presente (1) ou ausente (0)
- X_2 = sexo, masculino (1) ou feminino (0)
- X_3 = idade, classificada em três categorias: criança (1), adulto jovem (2), idoso (3)
- X_4 = fumante (1) ou não-fumante (0)

Existem $2 \times 2 \times 3 \times 2 = 24$ valores possíveis para \mathbf{X} . Precisamos alocar probabilidades aos 24 valores possíveis, todas não-negativas e somando 1. Esta alocação constitui a distribuição conjunta das variáveis no vetor \mathbf{X} . Por exemplo, imagine que tenhamos o seguinte:

X_1	X_2	X_3	X_4	$100\% \times \mathbb{P}$
0	0	1	0	4.6
0	0	1	1	6.7
0	0	2	0	4.1
0	0	2	1	5.3
0	0	3	0	5.2
0	0	3	1	6.6
0	1	1	0	1.6
0	1	1	1	1.8
0	1	2	0	4.9
0	1	2	1	0.2
0	1	3	0	3.1
0	1	3	1	4.4
1	0	1	0	1.1
1	0	1	1	6.8
1	0	2	0	0.5
1	0	2	1	4.0
1	0	3	0	4.0
1	0	3	1	2.9
1	1	1	0	3.6
1	1	1	1	3.7
1	1	2	0	6.6
1	1	2	1	6.8
1	1	3	0	6.6
1	1	3	1	4.9
Total				100.0

A distribuição marginal de X_1 é obtida como antes: para cada valor possível de X_1 , somamos sobre todas as combinações das demais variáveis. Por exemplo, para obter $\mathbb{P}(X_1 = 0)$ somamos todas as probabilidades em azul, que são aquelas em que $X_1 = 0$:

X_1	X_2	X_3	X_4	$100\% \times \mathbb{P}$
0	0	1	0	4.6
0	0	1	1	6.7
0	0	2	0	4.1
0	0	2	1	5.3
0	0	3	0	5.2
0	0	3	1	6.6
0	1	1	0	1.6
0	1	1	1	1.8
0	1	2	0	4.9
0	1	2	1	0.2
0	1	3	0	3.1
0	1	3	1	4.4
1	0	1	0	1.1
1	0	1	1	6.8
1	0	2	0	0.5
1	0	2	1	4.0
1	0	3	0	4.0
1	0	3	1	2.9
1	1	1	0	3.6
1	1	1	1	3.7
1	1	2	0	6.6
1	1	2	1	6.8
1	1	3	0	6.6
1	1	3	1	4.9

$\mathbb{P}(X_1 = 0) = \sum_{i,j,k} \mathbb{P}(X_1 = 0, X_2 = i, X_3 = j, X_4 = k)$

$$\begin{aligned} &= (4.6 + 5.7 + 4.1 + 5.3 + 5.2 + 6.6 + \\ &\quad + 1.6 + 1.8 + 4.9 + 0.2 + 3.1 + 4.4) / 100 \\ &= 0.475 \end{aligned}$$

Em seguida, obtemos $\mathbb{P}(X_1 = 1)$ por subtração já que X_1 só pode ser 0 ou 1:

$$\mathbb{P}(X_1 = 1) = 1 - \mathbb{P}(X_1 = 0) = 1 - 0.475 = 0.525$$

Vamos saltar X_2 e obter a distribuição marginal de X_3 . Para obter $\mathbb{P}(X_3 = 1)$ somamos as probabilidades das linhas em azul, que são aquelas em que $X_3 = 1$:

X_1	X_2	X_3	X_4	$100\% \times \mathbb{P}$
0	0	1	0	4.6
0	0	1	1	6.7
0	0	2	0	4.1
0	0	2	1	5.3
0	0	3	0	5.2
0	0	3	1	6.6
0	1	1	0	1.6
0	1	1	1	1.8
0	1	2	0	4.9
0	1	2	1	0.2
0	1	3	0	3.1
0	1	3	1	4.4
1	0	1	0	1.1
1	0	1	1	6.8
1	0	2	0	0.5
1	0	2	1	4.0
1	0	3	0	4.0
1	0	3	1	2.9
1	1	1	0	3.6
1	1	1	1	3.7
1	1	2	0	6.6
1	1	2	1	6.8
1	1	3	0	6.6
1	1	3	1	4.9

$\mathbb{P}(X_3 = 1) = \sum_{i,j,k} \mathbb{P}(X_1 = i, X_2 = j, X_3 = 1, X_4 = k)$

$$\begin{aligned} &= (4.6 + 6.7 + 1.6 + 1.8 \\ &\quad + 1.1 + 6.8 + 3.6 + 3.7) / 100 \\ &= 0.299 \end{aligned}$$

Já temos $\mathbb{P}(X_3 = 1) = 0.299$. Obtemos em seguida $\mathbb{P}(X_3 = 2)$ somando as probabilidades das linhas em azul, que são aquelas em que $X_3 = 2$:

X_1	X_2	X_3	X_4	$100\% \times \mathbb{P}$
0	0	1	0	4.6
0	0	1	1	6.7
0	0	2	0	4.1
0	0	2	1	5.3
0	0	3	0	5.2
0	0	3	1	6.6
0	1	1	0	1.6
0	1	1	1	1.8
0	1	2	0	4.9
0	1	2	1	0.2
0	1	3	0	3.1
0	1	3	1	4.4
1	0	1	0	1.1
1	0	1	1	6.8
1	0	2	0	0.5
1	0	2	1	4.0
1	0	3	0	4.0
1	0	3	1	2.9
1	1	1	0	3.6
1	1	1	1	3.7
1	1	2	0	6.6
1	1	2	1	6.8
1	1	3	0	6.6
1	1	3	1	4.9

Tendo calculado $\mathbb{P}(X_3 = 1) = 0.299$ e $\mathbb{P}(X_3 = 2) = 0.324$, podemos obter $\mathbb{P}(X_3 = 3)$ por subtração:

$$\mathbb{P}(X_3 = 3) = 1 - \mathbb{P}(X_3 = 1) - \mathbb{P}(X_3 = 2) = 0.377.$$

Você deve ter percebido como funciona a marginalização, em geral. Seja $\mathbf{X} = (X_1, X_2, \dots, X_k)$ um vetor de v.a.'s discretas. Suponha que X_i tenha n_i valores possíveis. Queremos $\mathbb{P}(X_1 = x)$ onde x é um dos seus n_1 valores possíveis. Para cada valor de x , a probabilidade $\mathbb{P}(X_1 = x)$ é uma soma de $n_2 \times n_3 \times \dots \times n_k$ termos da tabela de distribuição conjunta. Se todas as v.a.'s são binárias temos 2^{k-1} parcelas para cada valor de x . Se quisermos $\mathbb{P}(X_1 = x)$ para todos os n_1 valores x possíveis para X_1 , precisamos fazer o cálculo anterior n_1 vezes. Na verdade, $n_1 - 1$ vezes pois o último é obtido por subtração:

$$\mathbb{P}(X_1 = x_{n_1}) = 1 - \mathbb{P}(X_1 = x_1) - \dots - \mathbb{P}(X_1 = x_{n_1-1})$$

Às vezes, falar em distribuição marginal ou conjunta pode soar ambíguo. No exemplo que estamos considerando com $\mathbf{X} = (X_1, X_2, X_3, X_4)$, podemos estar interessados na distribuição marginal (ou conjunta?) de X_1 e X_2 , após somarmos sobre as possibilidades para X_3 e X_4 . Sem nos prender em terminologia, o que queremos é a distribuição de probabilidade das v.a.'s X_1 e X_2 e isto é facilmente obtido. Por exemplo, para obter $\mathbb{P}(X_1 = 0, X_2 = 1)$ somamos sobre todas as linhas azuis da tabela abaixo, que são aquelas em que temos $X_1 = 0$ e $X_2 = 1$:

X_1	X_2	X_3	X_4	$100\% \times \mathbb{P}$
0	0	1	0	4.6
0	0	1	1	6.7
0	0	2	0	4.1
0	0	2	1	5.3
0	0	3	0	5.2
0	0	3	1	6.6
0	1	1	0	1.6
0	1	1	1	1.8
0	1	2	0	4.9
0	1	2	1	0.2
0	1	3	0	3.1
0	1	3	1	4.4
1	0	1	0	1.1
1	0	1	1	6.8
1	0	2	0	0.5
1	0	2	1	4.0
1	0	3	0	4.0
1	0	3	1	2.9
1	1	1	0	3.6
1	1	1	1	3.7
1	1	2	0	6.6
1	1	2	1	6.8
1	1	3	0	6.6
1	1	3	1	4.9

Definition 12.5.1 — Independência de v.a's discretas. Seja $\mathbf{X} = (X_1, \dots, X_k)$ um vetor aleatório composto de v.a.'s discretas. Elas são *independentes* se

$$\mathbb{P}(X_1 = x_1, \dots, X_k = x_k) = \mathbb{P}(X_1 = x_1) \dots \mathbb{P}(X_k = x_k)$$

para qualquer configuração de valores possíveis (x_1, \dots, x_k) .

Theorem 12.5.1 — Independência: outra maneira. Se o vetor \mathbf{X} é composto de v.a.'s independentes então

$$\mathbb{P}(X_1 = x_1 | X_2 = x_2, \dots, X_k = x_k) = \mathbb{P}(X_1 = x_1)$$

para qualquer configuração de valores possíveis (x_1, \dots, x_k) .

Este resultado é válido se X_1 trocar de posição com qualquer outra v.a.

12.6 Simulação de \mathbf{X} discreto

Suponha que queiramos simular, via Monte Carlo, um vetor de v.a.'s discretas com distribuição conjunta dada pela tabela abaixo. Como fazer? Não podemos gerar as v.a.'s separadamente já que elas tipicamente não são independentes. O método é simples e, fundamentalmente, equivale ao método da transformada inversa. Podemos usar o mesmo procedimento aprendido para uma v.a. discreta. Simule $U \sim U(0, 1)$ e veja em que segmento U caiu na coluna de soma acumulada. Este segmento determina o vetor \mathbf{X} gerado. Por exemplo, se $U = 0.3215$ então $\mathbf{X} = (0, 0, 3, 1)$ é selecionado. A geração não é feita separadamente para cada v.a. do vetor com base na sua distribuição marginal, a menos que as v.a.'s sejam independentes.

X_1	X_2	X_3	X_4	$100\% \times \mathbb{P}$	Soma Acum.
0	0	1	0	4.6	4.6
0	0	1	1	6.7	11.3
0	0	2	0	4.1	15.4
0	0	2	1	5.3	20.7
0	0	3	0	5.2	25.9
0	0	3	1	6.6	32.5
0	1	1	0	1.6	34.1
0	1	1	1	1.8	35.9
0	1	2	0	4.9	40.8
0	1	2	1	0.2	41.0
0	1	3	0	3.1	44.1
0	1	3	1	4.4	48.5
1	0	1	0	1.1	49.6
1	0	1	1	6.8	56.4
1	0	2	0	0.5	56.9
1	0	2	1	4.0	60.9
1	0	3	0	4.0	64.9
1	0	3	1	2.9	67.8
1	1	1	0	3.6	71.4
1	1	1	1	3.7	75.1
1	1	2	0	6.6	81.7
1	1	2	1	6.8	88.5
1	1	3	0	6.6	95.1
1	1	3	1	4.9	100.0

12.7 Um outro arranjo no caso bi-dimensional

Às vezes, temos um arranjo bi-dimensional. No caso de termos apenas duas v.a.'s discretas, é comum apresentar a distribuição conjunta de probabilidade como um array de duas entradas. Vamos voltar ao exemplo 12.5, aos dois testes de diagnósticos: teste-outo e teste-rápido. Temos a probabilidade conjunta abaixo:

	$TR = 0$	$TR = 1$
$T0 = 0$	0.40	0.19
$T0 = 1$	0.03	0.38

Colocamos os valores possíveis de X_1 nas linhas. Colocamos os valores de X_2 nas colunas. Na posição (i, j) do array colocamos a probabilidade $\mathbb{P}(X_1 = x_i, X_2 = x_j)$. E nas marginais (ou margens) da tabela, temos as distribuições marginais da variável coluna e da variável-linha. A marginal de $T0$ é obtida somando as colunas:

$TR = 0$	$TR = 1$	$Total$
$T0 = 0$	0.40	$\mathbb{P}(T0 = 0) =$ 0.40 + 0.19 = 0.51
$T0 = 1$	0.03	$\mathbb{P}(T0 = 1) =$ 0.03 + 0.38 = 0.49

Isto explica o nome distribuição *marginal* para a distribuição de uma única variável: elas ficam nas margens da tabela.

Somando as linhas encontramos a marginal de TR :

	$TR = 0$	$TR = 1$	$Total$
$T0 = 0$	0.40	0.19	0.51
$T0 = 1$	0.03	0.38	0.49
$Total$	$\mathbb{P}(TR = 0) =$ $0.40 + 0.03 = 0.43$	$\mathbb{P}(TR = 1) =$ $0.19 + 0.38 = 0.57$	1.00

A soma dos valores na marginal-linha ou na marginal-coluna é o total das probabilidades: 1.

12.8 Um longo exemplo: Mobilidade social no Brasil em 1988

Selecione um adulto brasileiro ω ao acaso em 1988. Para cada ω , vamos definir duas v.a.'s:

- $SF(\omega)$: o status sócio-econômico da sua ocupação (6 valores): 1,2, ..., 6. As ocupações estão categorizadas de acordo com características de renda e educação: Baixo inferior, Baixo superior, Médio inferior, Médio, Médio superior, Alto.
- $SP(\omega)$: status social da ocupação de seu pai quando o pai tinha 45 anos (6 valores): 1,2, ..., 6. Usa-se as mesmas categorias que n caso do filho.

Nesta categorização, executivos e juízes de tribunais superiores estavam na categoria *Alto*. Já os trabalhadores braçais, em ocupações que exigiam nenhuma instrução, estavam na categoria *Baixo inferior*.

O arcabouço teórico para a mobilidade social é o seguinte. Selecione um indivíduo ω ao acaso em 1988. Para cada indivíduo ω selecionado, meça o vetor $\mathbf{X}(\omega) = (SF(\omega), SP(\omega))$. Existem 36 valores possíveis para o vetor aleatório \mathbf{X} e as probabilidades associadas são

$$\theta_{ij} = \mathbb{P}(\text{pai ter status } i \wedge \text{filho ter status } j) = \mathbb{P}(SP = i, SF = j).$$

Não esperamos que as v.a.'s SP e SF sejam v.a.'s independentes. Existe uma grande inérvia na sociedade: filhos de pais de status baixo tendem a continuar com status baixo e filhos de pais de status alto geralmente possuem status alto. Como quantificar esta inérvia? Como comparar diferentes sociedades quanto ao seu grau de mobilidade social? Este é um assunto fascinante e estudado por vários autores [16].

Vamos estimar as probabilidades θ_{ij} usando os dados de uma pesquisa do IBGE. A Pesquisa Nacional por Amostra de Domicílios, em 1988, entrevistou uma amostra de 42137 homens chefes de família entre 20 e 64 anos e a partir dessa amostra, usando proporções, a tabela 12.8 foi criada. Nela, os valores aproximados de θ_{ij} estão multiplicados por 100.

SP : status do pai	SF: status do indivíduo em 1988.					
	Baixo Inf.	Baixo Sup.	Médio Inf.	Médio	Médio Sup.	Alto
BI	21.7	12.8	13.2	4.6	2.1	1.0
BS	0.7	4.2	3.6	2.5	2.5	1.3
MI	0.6	3.7	7.1	2.7	2.7	1.5
M	0.6	1.9	2.0	2.2	1.2	0.9
MS	0.3	0.6	0.6	0.7	0.7	0.5
A	0.1	0.3	0.3	0.6	0.6	0.9

Algumas das questões de interesse associadas com esta tabela são as seguintes: Como mudou a distribuição do status entre duas gerações? Existe uma maior proporção de pessoas empregadas no status alto na geração mais recente? Filhos de pais com status muito baixo passam com facilidade para um status mais alto? A estrutura de ocupação mudou drasticamente na década

de 70 no Brasil devido ao milagre econômico nos anos dos governos militares. Houve uma expansão acelerada da indústria e do setor de serviços neste período. O Brasil deixou de ser uma sociedade agrária e foram abertos novos postos de trabalho qualificados, requerendo mais qualificação profissional. Engenheiros ainda na faculdade eram recrutados com altos salários. No setor de serviços administrativos isto também ocorreu. Como resultado, houve a necessidade de recrutar pessoas vindas de pais com status mais baixos para preencher estas novas vagas de empregos nos estratos superiores. Quanto da mobilidade social pode ser explicada por esta expansão ou deslocamento temporal da estrutura de emprego?

Vamos começar obtendo as distribuições marginais de SP e SF , a estrutura de status das ocupações para a geração dos pais e dos filhos. No caso desse arranjo bi-dimensional, basta somar as probabilidades ao longo das linhas e das colunas para encontrar os valores nas margens da tabela:

SP : status do pai	SF : status do indivíduo em 1988 .						TOTAL
	Baixo Inf.	Baixo Sup.	Médio Inf.	Médio	Médio Sup.	Alto	
BI	21.7	12.8	13.2	4.6	2.1	1.0	55.4
BS	0.7	4.2	3.6	2.5	2.5	1.3	14.8
MI	0.6	3.7	7.1	2.7	2.7	1.5	18.3
M	0.6	1.9	2.0	2.2	1.2	0.9	8.8
MS	0.3	0.6	0.6	0.7	0.7	0.5	3.4
A	0.1	0.3	0.3	0.6	0.6	0.9	2.8
TOTAL	24.0	23.5	26.8	13.3	9.8	6.1	100%

Focando apenas nos números que estão nas marginais, vemos que, na geração dos pais, 55% das ocupações estavam no estrato *Baixo inferior* e isto foi reduzido a apenas 24% das ocupações na geração dos filhos. Nos dois níveis de status mais elevados, a porcentagem passa de 6% para 16% entre as duas gerações. Há um deslocamento de ocupações em direção aos status mais elevados.

Vamos fazer alguns cálculos com a distribuição conjunta de (SP, SF) . Seja A o evento “pai pobre, filho rico”: o indivíduo tem status pelo menos *Médio superior* e seu pai tem status menor ou igual a *Baixo superior*. Este evento corresponde ao par $(SP(\omega), SF(\omega))$ cair em uma de 4 células da tabela: (1,5), (1,6), (2,5) e (2,6). Temos

$$\mathbb{P}(A) = \mathbb{P}(SF \geq 5 \wedge SP \leq 2) = \frac{2.1 + 1.0 + 2.5 + 1.3}{100} = 0.069$$

ou 6.9%.

Seja B o evento reverso, “pai rico, filho pobre”: o indivíduo tem status menor ou igual a *Baixo superior* e seu pai tem status pelo menos *Médio superior*. Temos

$$\mathbb{P}(B) = \mathbb{P}(SP \geq 5 \wedge SF \leq 2) = \frac{0.3 + 0.6 + 0.1 + 0.3}{100} = \frac{1.3}{100} = 0.013$$

ou 1.3%. Era mais fácil que um filho de pobre se tornasse muito rico que um filho de rico ficasse muito pobre.

12.9 Condicional discreta

Vamos ver como obter a distribuição de uma v.a. condicionada nos valores de outras no vetor no caso discreto¹. Vamos voltar ao exemplo em que temos $\mathbf{X} = (X_1, X_2, X_3, X_4)$ um vetor aleatório composto de v.a.’s discretas com a distribuição conjunta dada na tabela abaixo:

¹ A maneira de apresentar a distribuição condicional nesta seção é igual a de Daphne Koller em seu curso *Probabilistic Graphical Model* na plataforma Coursera.

X_1	X_2	X_3	X_4	$100\% \times \mathbb{P}$
0	0	1	0	4.6
0	0	1	1	6.7
0	0	2	0	4.1
0	0	2	1	5.3
0	0	3	0	5.2
0	0	3	1	6.6
0	1	1	0	1.6
0	1	1	1	1.8
0	1	2	0	4.9
0	1	2	1	0.2
0	1	3	0	3.1
0	1	3	1	4.4
1	0	1	0	1.1
1	0	1	1	6.8
1	0	2	0	0.5
1	0	2	1	4.0
1	0	3	0	4.0
1	0	3	1	2.9
1	1	1	0	3.6
1	1	1	1	3.7
1	1	2	0	6.6
1	1	2	1	6.8
1	1	3	0	6.6
1	1	3	1	4.9
Total			100.0	

Aprendemos a obter a distribuição *marginal* de uma ou mais v.a.'s: some sobre os valores das demais v.a.'s. Queremos agora a distribuição de algumas v.a. *condicionada* nos valores de uma ou mais das outras v.a.'s. Por exemplo, queremos a distribuição do vetor (X_1, X_2, X_4) dado que $X_3 = 2$:

$$\mathbb{P}(X_1 = i, X_2 = j, X_3 = k | X_3 = 2)$$

para diferentes valores de i, j, k . A partir da tabela da distribuição conjunta, nós simplesmente eliminamos as linhas em que $X_3 \neq 2$. A razão é simples: fomos informados que $X_3 = 2$ e portanto os demais casos não importam mais, estamos restritos ao mundo em que X_3 está fixado em 2 e apenas as outras v.a.'s estão liberadas podem assumir valores. Em suma, queremos

$$\mathbb{P}(X_1 = i, X_2 = j, X_3 = k | X_3 = 2)$$

Como $X_3 = 2$, podemos eliminar de consideração todos os outros resultados em que $X_3 \neq 2$. Este é novo conjunto de valores possíveis para o vetor \mathbf{X} , apenas aqueles em que X_3 possui o valor 2. Dentro deste novo “mundo”, as probabilidades devem somar 1. Basta normalizarmos: divida os valores originais das probabilidades pela soma dos seus termos.

X_1	X_2	X_3	X_4	$100\% \times \mathbb{P}$
0	0	2	0	4.1
	0	2	1	5.3
0	1	2	0	4.9
	1	2	1	0.2
1	0	2	0	0.5
	0	2	1	4.0
1	1	2	0	6.6
	1	2	1	6.8

Renormalize a tabela resultante para que suas probabilidades somem 1. Isto é, dividimos cada probabilidade que restou pela sua soma de forma que os valores agora vão somar 1. A tabela resultante é a distribuição condicional de (X_1, X_2, X_4) dado que $X_3 = 2$. A distribuição de qualquer conjunto de v.a.'s condicionado nos valores das demais é obtido do mesmo modo.

X_1	X_2	X_4	$100\% \times \mathbb{P}(\dots X_3 = 2)$
0	0	0	100% ($4.1/32.4$) = 12.7
0	0	1	100% ($5.3/32.4$) = 16.4
0	1	0	100% ($4.9/32.4$) = 15.1
0	1	1	100% ($0.2/32.4$) = 0.6
1	0	0	100% ($0.5/32.4$) = 1.5
1	0	1	100% ($4.0/32.4$) = 12.3
1	1	0	100% ($6.6/32.4$) = 20.4
1	1	1	100% ($6.8/32.4$) = 21.0
Total			100%

Vamos obter agora a distribuição de $(X_1, X_4 | X_2 = 0, X_3 = 2)$. Elimine todas as linhas da tabela original com a distribuição conjunta em que $X_2 \neq 0$ OU que $X_3 \neq 2$. A seguir, renormalize as linhas restantes e simplifique a tabela.

X_1	X_2	X_3	X_4	$100\% \times \mathbb{P}$
0	0	2	0	4.1
0	0	2	1	5.3
1	0	2	0	0.5
1	0	2	1	4.0

Renormalizando as linhas restantes e simplificando a tabela, temos:

X_1	X_4	$100\% \times \mathbb{P}(X_1 = i, X_4 = j X_2 = 0, X_3 = 2)$
0	0	4.1/13.9 = 29.5
0	1	5.3/13.9 = 38.1
1	0	0.5/13.9 = 3.6
1	1	4.0/13.9 = 28.8
Total		100

Podemos obter uma visão um pouco mais algébrica da distribuição condicional. Queremos a distribuição de $(X_1, X_4 | X_2 = 0, X_3 = 2)$. Isto é, queremos as probabilidades $\mathbb{P}(X_1 = i, X_4 = j | X_2 = 0, X_3 = 2)$ para toda combinação de i, j . Pela definição de probabilidade condicional:

$$\mathbb{P}(X_1 = i, X_4 = j | X_2 = 0, X_3 = 2) = \frac{\mathbb{P}(X_1 = i, X_4 = j, X_2 = 0, X_3 = 2)}{\mathbb{P}(X_2 = 0, X_3 = 2)}$$

O numerador são os elementos que restam na tabela das probabilidades originais, da distribuição conjunta, após eliminarmos as linhas em que não temos $[X_2 = 0, X_3 = 2]$. O denominador é o fator de normalização já que

$$\mathbb{P}(X_2 = 0, X_3 = 2) = \sum_{k,l} \mathbb{P}(X_1 = k, X_4 = l, X_2 = 0, X_3 = 2)$$

Assim, esta visão gráfica de eliminar as linhas da tabelas etc corresponde a esta operação algébrica.

12.10 De volta à mobilidade social

Vamos obter a distribuição condicional do status SP do pai dado que o status SF do filho é Alto. Dado que o filho está na elite, de onde ele veio? Pela definição de probabilidade condicional,

$$\mathbb{P}(SP = i|SF = 6) = \frac{\mathbb{P}(SP = i, SF = 6)}{\mathbb{P}(SF = 6)} = \text{cte } \mathbb{P}(SP = i, SF = 6)$$

O numerador da fração acima é dos elementos na coluna 6 da tabela da distribuição conjunta, a coluna em que $SF = 6$. Assim, para termos $\mathbb{P}(SP = i|SF = 6)$ basta tomar os os números da coluna 6, $\mathbb{P}(SP = i, SF = 6)$, e normalizá-los para que somem 1 (isto é, dividir os por sua soma $\mathbb{P}(SF = 6) = \sum_i \mathbb{P}(SP = i, SF = 6)$): para que somem 1:

ALTO	
1.0	.8 0.16
1.3	.8 0.21
1.5	.8 0.25
0.9	.8 0.15
0.5	.8 0.08
0.9	.8 0.15
$\Sigma = 6.1$	$\Sigma = 1$

Esses valores finais são os valores de $\mathbb{P}(Y_1 = i|Y_2 = 6)$ para os diferentes valores de i .

Note que $\mathbb{P}(SP = 1|SF = 6) = 0.20$, isto é, 20% da elite veio dos estratos mais baixos da sociedade daquela época. Vamos ver na outra direção agora. Vamos olhar para $\mathbb{P}(SF = j|SP = 1)$: dado que o pai era lavrador manual ou similar, aonde foram parar seus filhos? Estas probabilidades são proporcionais aos elementos da linha 1 da tabela da distribuição conjunta:

$$\mathbb{P}(SF = j|SP = 1) = \frac{\mathbb{P}(SF = j, SP = 1)}{\mathbb{P}(SP = 1)} \propto \mathbb{P}(SF = j|SP = 1)$$

Basta normalizar os números da linha 1 da tabela de probabilidade conjunta para obter estas probabilidades condicionais:

B.I.	21.7	12.8	13.2	4.6	2.1	1.0
$\mathbb{P}(SF = j SP = 1)$	j=1	j=2	j=3	j=4	j=5	j=6
	0.39	0.23	0.24	0.08	0.04	0.02

Assim, $\mathbb{P}(SF = 6|SP = 1) = 0.02$, mas $\mathbb{P}(SP = 1|SF = 6) = 0.20$, uma ordem de grandeza de diferença. Como explicar esta disparidade? A enorme massa de 55% de pais de baixo status enviou apenas 2% de seus filhos para a elite. Mas 2% de 55% formam 1% da população total. A elite da geração dos filhos forma 5% da população total. Estes 5% da população total dividem-se em 1% vindos de pais de status baixo e os outros 4% vindos de pais com status maior. Assim, estes 1% *dentre os 5% da elite de hoje* formam os 20% da elite que veio de baixo na pirâmide social.

12.11 Distribuição condicional de X

Vamos ver agora a definição mais formal de distribuição condicional. Seja $\mathbf{X} = (X_1, X_2, \dots, X_k)$ um vetor aleatório. Queremos a distribuição de probabilidade da v.a. X_1 dados os valores das demais. Por exemplo, queremos a distribuição de X_1 quando $X_2 = 0, \dots, X_k = 2$.

$$(X_1 | X_2 = 0, \dots, X_k = 2) \sim ??$$

O que é a distribuição de uma v.a. discreta? Duas coisas...

- Uma lista $\{a_1, \dots, a_m\}$ dos valores possíveis de X_1 quando $X_2 = 0, \dots, X_k = 2$
- Uma lista com as probabilidades associadas quando $X_2 = 0, \dots, X_k = 2$:

$$\mathbb{P}(X_1 = a_i | X_2 = 0, \dots, X_k = 2)$$

Ao mudar os valores condicionados de X_2, \dots, X_k esta distribuição também muda. Por exemplo, as probabilidades de $\mathbb{P}(X_1 = a_i | X_2 = 1, \dots, X_k = 0)$ usualmente são diferentes das de $\mathbb{P}(X_1 = a_i | X_2 = 0, \dots, X_k = 2)$. A distribuição é função dos valores em que estamos condicionando as demais variáveis X_2, \dots, X_k .

Vamos nos fixar em obter $\mathbb{P}(X_1 = a_i | X_2 = 0, \dots, X_k = 2)$. Para estes valores fixos $X_2 = 0, \dots, X_k = 2$ das variáveis condicionantes, a distribuição é encontrada pela fórmula de probabilidade condicional:

$$\mathbb{P}(X_1 = a_i | X_2 = 0, \dots, X_k = 2) = \frac{\mathbb{P}(X_1 = a_i, X_2 = 0, \dots, X_k = 2)}{\mathbb{P}(X_2 = 0, \dots, X_k = 2)}$$

O denominador não depende de a_i e portanto não varia com o valor de a_i . Isto é, se $a_i \neq a_j$, temos

$$\frac{\mathbb{P}(X_1 = a_i | X_2 = 0, \dots, X_k = 2)}{\mathbb{P}(X_1 = a_j | X_2 = 0, \dots, X_k = 2)} = \frac{\frac{\mathbb{P}(X_1 = a_i, X_2 = 0, \dots, X_k = 2)}{\mathbb{P}(X_2 = 0, \dots, X_k = 2)}}{\frac{\mathbb{P}(X_1 = a_j, X_2 = 0, \dots, X_k = 2)}{\mathbb{P}(X_2 = 0, \dots, X_k = 2)}} = \frac{\mathbb{P}(X_1 = a_i, X_2 = 0, \dots, X_k = 2)}{\mathbb{P}(X_1 = a_j, X_2 = 0, \dots, X_k = 2)}$$

Podemos enxergar a distribuição condicional de uma v.a. diretamente da tabela original de probabilidade conjunta. X_3 possui 3 valores possíveis: 1, 2, 3. Comparando a chance (condicional) de $X_3 = 1$ versus $X_3 = 2$

$$\frac{\mathbb{P}(X_3 = 1 | X_1 = 0, X_2 = 1, X_4 = 0)}{\mathbb{P}(X_3 = 2 | X_1 = 0, X_2 = 1, X_4 = 0)} =$$

$$= \frac{\mathbb{P}(X_3 = 1, X_1 = 0, X_2 = 1, X_4 = 0)}{\mathbb{P}(X_3 = 2, X_1 = 0, X_2 = 1, X_4 = 0)} =$$

$$= \frac{1.6}{4.9} = 0.33$$

Se \mathbf{x}_1 e \mathbf{x}_2 são dois dos vetores-valores possíveis para o vetor \mathbf{X} e se $\mathbb{P}(\mathbf{X} = \mathbf{x}_1)$ for duas vezes maior que $\mathbb{P}(\mathbf{X} = \mathbf{x}_2)$ então esta razão ainda será respeitada entre as condicionais correspondentes.

X_1	X_2	X_3	X_4	$100\% \times \mathbb{P}$
0	0	1	0	4.6
0	0	1	1	6.7
0	0	2	0	4.1
0	0	2	1	5.3
0	0	3	0	5.2
0	0	3	1	6.6
0	1	1	0	1.6
0	1	1	1	1.8
0	1	2	0	4.9
0	1	2	1	0.2
0	1	3	0	3.1
0	1	3	1	4.4
1	0	1	0	1.1
1	0	1	1	6.8
1	0	2	0	0.5
1	0	2	1	4.0
1	0	3	0	4.0
1	0	3	1	2.9
1	1	1	0	3.6
1	1	1	1	3.7
1	1	2	0	6.6
1	1	2	1	6.8
1	1	3	0	6.6
1	1	3	1	4.9
Total				100.0

12.12 Exemplos de distribuições condicionais discretas

Nesta seção, vamos apresentar alguns exemplos ilustrando o conceito de distribuição conjunta, marginal e condicional de variáveis discretas.

■ **Example 12.9 — Sementes de maçãs.** A maioria das maçãs possuem cinco câamaras (ou carpelos) contendo suas sementes. Cada carpelo contém até duas sementes e portanto cada maçã tipicamente tem um máximo de 10 sementes. O número de sementes viáveis, de boa qualidade, depende da variedade de maçã e do vigor e saúde da planta. Árvores mais saudáveis produzem frutos melhores com mais e maiores sementes.

A tabela 12.4 mostra a distribuição do número de maçãs por quantidade de sementes viáveis em três diferentes variedades de maçãs [crandall1917seed]. A partir desses dados, podemos ajustar uma distribuição binomial para a v.a. X que conta o número de sementes de uma maçã. Dividindo o número médio de sementes em cada variedade por 10 teremos a chance de cada uma das 10 possíveis semeentes de uma maçã tornar-se uma semente viável. Este valor está representado na última coluna, rotulada como θ na tabela abaixo.

	# of Seeds											θ
	0	1	2	3	4	5	6	7	8	9	10	
Apple	0	1	2	3	4	5	6	7	8	9	10	
Ben Davis	9	18	27	54	67	61	72	55	36	11	4	0.495
Collins	12	22	40	26	26	12	8	4	0	0	0	0.280
Grimes	0	0	6	22	50	47	49	39	21	11	7	0.562

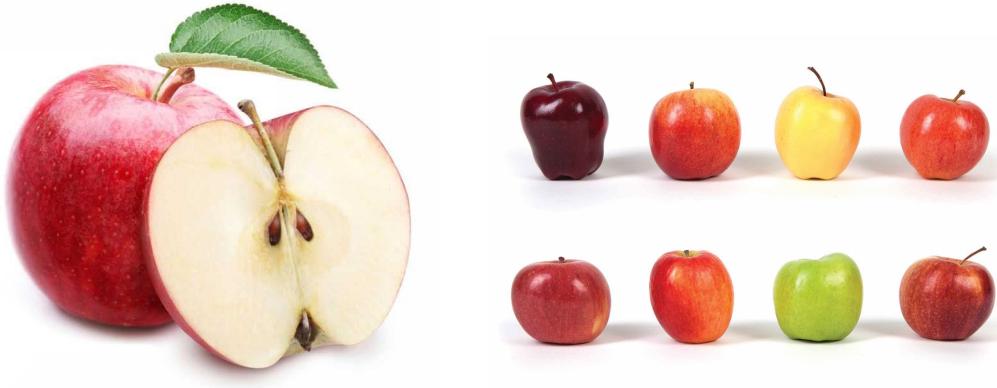


Figure 12.4: O número de sementes numa maçã depende da sua variedade.

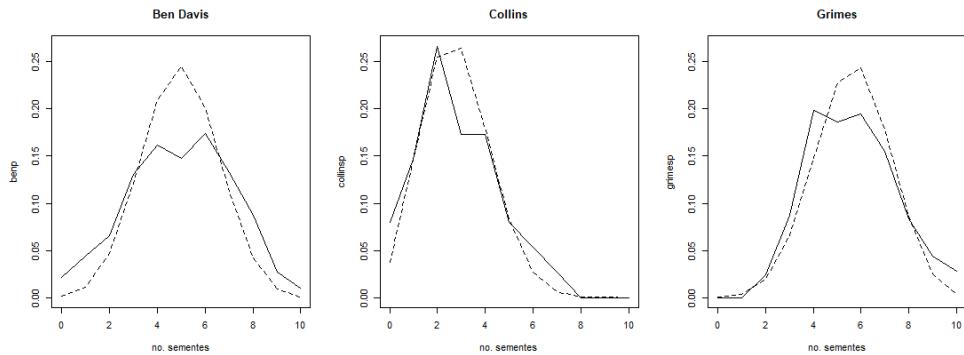


Figure 12.5: Frequênciia relativa (linha sólida) e teórica (linha tracejada, derivada de um modelo binomial) para o número de sementes de três variedades de maçãs.

Seja C a classe de maçã considerada, com $C \in \{Ben, Col, Gri\}$. Estamos sugerindo que

$$\begin{aligned} (X | C = Ben) &\sim \text{Bin}(10, 0.495) \\ (X | C = Col) &\sim \text{Bin}(10, 0.280) \\ (X | C = Gri) &\sim \text{Bin}(10, 0.562) \end{aligned}$$

A Figura 12.5 mostra um ajuste binomial aos dados da tabela acima. O eixo horizontal mostra os possíveis valores do número x de sementes. O eixo vertical mostra as probabilidades $\mathbb{P}(X = x)$. A linha sólida é a estimativa empírica simples, a proporção de maçãs com x sementes em cada variedade. A linha tracejada são as probabilidades derivadas do modelo binomial. Enquanto o ajuste para as variedades *Collins* e *Grimes* parecem razoáveis, a variedade *Ben Davis* parece não seguir a distribuição binomial. Comparado com o esperado (a linha tracejada), esta variedade possui mais maçãs nos dois extremos, com poucas e com muitas sementes. Entretanto, como a comparação visual entre as duas curvas mostra bastante similaridade entre elas, vamos seguir com este modelo binomial nas três variedades.

```
ben = c(9, 18, 27, 54, 67, 61, 72, 55, 36, 11, 4)
collins = c(12, 22, 40, 26, 26, 12, 8, 4, 0, 0, 0)
grimes = c(0, 0, 6, 22, 50, 47, 49, 39, 21, 11, 7)
```

```
benp = ben/sum(ben); collinsp = collins/sum(collins); grimesp = grimes/sum(grimes)
```



Figure 12.6: ...

```

mb = sum((0:10)*benp)/10; mc = sum((0:10)*collinsp)/10; mg = sum((0:10)*grimesp)/10

par(mfrow=c(1,3)); aux = range(benp, collinsp, grimesp)
plot(0:10, benp, type="l", ylim=aux, main="Ben Davis", xlab="no. sementes")
lines(0:10, dbinom(0:10, 10, mb), lty=2)
plot(0:10, collinsp, type="l", ylim=aux, main="Collins", xlab="no. sementes")
lines(0:10, dbinom(0:10, 10, mc), lty=2)
plot(0:10, grimesp, type="l", ylim=aux, main="Grimes", xlab="no. sementes")
lines(0:10, dbinom(0:10, 10, mg), lty=2)

```

■ **Example 12.10 — Apartamentos em Belo Horizonte.** numero de quartos versus vaga de garagem, por idade do apto.

■ **Example 12.11 — Infecção conjunta.** Este exemplo veio das notas de aula do professor Jonathan Jordan, da unIversidade de Sheffield, na Inglaterra. Duas pessoas moram em uma casa e, na semana 1, ambas estão sob risco de pegar um resfriado. Suponha que cada pessoa tenha uma probabilidade de 0.1 de pegar um resfriado na semana 1, independentemente um do outro. Se ninguém pegar um resfriado, as probabilidades de pegar um resfriado não mudam na semana 2. Se exatamente uma pessoa pegar um resfriado na semana 1, o outro ocupante da casa que estava infectado passa a ter uma probabilidade de 0.2 de pegar um resfriado na semana 2. Suponha que ninguém vai pegar um resfriado duas vezes. Por exemplo, se ambos pegarem um resfriado na semana 1, ninguém pode pegar um resfriado na semana 2. Crie uma tabela com a distribuição conjunta de W_1 e W_2 o número de pessoas refriadadas nas semanas 1 e 2, respectivamente.

Começamos encontrando a distribuição marginal de W_1 . Esta variável conta o número de infectados (sucessos) na primeira semana. Temos dois indivíduos, ambos com probabilidade de infecção 0.1 e independentes. Portanto, $W_1 \sim \text{Bin}(2, 0.1)$ e assim encontramos $\mathbb{P}(W_1 = k) = 2/(k!(2-k)!)\cdot 0.1^k\cdot 0.9^{2-k}$. A coluna marginal da tabela 12.1 mostra estas probabilidades marginais.

Para encontrar as células internas da distribuição conjunta, vamos obter uma linha por vez. Começando da última linha, queremos

$$\mathbb{P}(W_1 = 2, W_2 = k) = \mathbb{P}(W_2 = k|W_1 = 2)\mathbb{P}(W_1 = 2) = \mathbb{P}(W_2 = k|W_1 = 2)(0.1)^2$$

Pela descrição do problema, se os dois tiverem se infectado na primeira semana, ninguém mais vai se infectar na segunda semana e portanto $\mathbb{P}(W_2 = k|W_1 = 2) = 0$ se $k = 1$ ou $k = 2$ e $\mathbb{P}(W_2 = 0|W_1 = 2) = 1$. Assim, a terceira linha está completa.

		Week 2			
Week 1		0	1	2	Total
0		$(0.9)^4$	$2(0.1)(0.9)^3$	$((0.1)(0.9))^2$	$(0.9)^2 = 0.81$
1		$2(0.8)(0.1)(0.9)$	$(0.2)2(0.1)(0.9)$	0	$2(0.1)(0.9) = 0.18$
2		$(0.1)^2$	0	0	$(0.1)^2 = 0.01$
Total		0.8101	0.1818	0.0081	

Table 12.1: Distribuição conjunta do número de infectados na primeira e segunda semanas.

Passando para a segunda linha, se um deles estiver infectado, não poderemos os dois infectados na segunda semana e portanto a célula $\mathbb{P}(W_1 = 1, W_2 = 2) = 0$. O indivíduo infectado na primeira semana estará sadio na segunda semana e portanto W_2 depende simplesmente do indivíduo não-infectado na primeira semana. Ele fica infectado com probabilidade 0.2 e portanto

$$\mathbb{P}(W_1 = 1, W_2 = 1) = \mathbb{P}(W_2 = 1|W_1 = 1)\mathbb{P}(W_1 = 1) = (0.2) \times 2(0.1)(0.9)$$

e

$$\mathbb{P}(W_1 = 1, W_2 = 0) = \mathbb{P}(W_2 = 0|W_1 = 1)\mathbb{P}(W_1 = 1) = (0.8) \times 2(0.1)(0.9)$$

Finalmente, para a primeira linha, se ninguém se infectou na primeira semana, a segunda semana segue a mesma distribuição binomial de W_1 . Isto é, $(W_2|W_1 = 0) \sim \text{Bin}(2, 0.1)$ e portanto

$$\mathbb{P}(W_1 = 0, W_2 = k) = \mathbb{P}(W_2 = k|W_1 = 0)\mathbb{P}(W_1 = 0) = \binom{2}{k} 0.1^k 0.9^{(2-k)} \times 0.9^2$$

■

Existem duas maneiras de obter a distribuição condicional. A primeira delas é derivando a distribuição condicional a partir da distribuição conjunta do vetor aleatório. A segunda maneira é quando, ao invés de fornecermos a distribuição conjunta para que a condicional seja deduzida, nós fornecemos diretamente a distribuição conjunta sem nunca nos preocuparmos em fornecer a conjunta. Muitos modelos de análise de dados, tais como os modelos de regressão, são dessa forma, baseados apenas na especificação da distribuição condicionada. Eles são chamados de *modelos discriminativos* e serão estudados nos capítulos ??, ?? e ?. O próximo exemplo ilustra como este tipo de modelagem estatística funciona.

■ **Example 12.12 — Besouros e morte.** [Bliss1935] é um paper clássico que introduziu a técnica de regressão logística, assunto do capítulo ?. Num experimento para desenvolver novos produtos para controle de pragas agrícolas, besouros adultos foram expostos por cinco horas a um concentrado de dissulfureto de carbono gasoso (CS_2) em concentrações crescentes. Espera-se que a chance de morte por exposição aumente com a dose aplicada. A proporção de besouros mortos em cada um dos cinco níveis do concentrado está na tabela a seguir:

Dose	# Expostos	# Mortos	Proporção	Modelo ($Y D = d$)
49.1	59	6	0.102	$\text{Bin}(59, p_{49.1})$
53.0	60	13	0.217	$\text{Bin}(60, p_{53.0})$
56.9	62	18	0.290	$\text{Bin}(62, p_{56.9})$
60.8	56	28	0.500	$\text{Bin}(56, p_{60.8})$
64.8	63	52	0.825	$\text{Bin}(63, p_{64.8})$
68.7	59	53	0.898	$\text{Bin}(59, p_{68.7})$
72.6	62	61	0.984	$\text{Bin}(62, p_{72.6})$
76.5	60	60	1.000	$\text{Bin}(60, p_{76.5})$

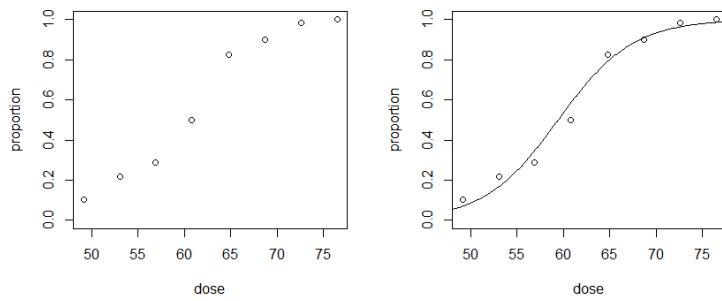


Figure 12.7: Esquerda: Proporção de besouros mortos versus dose aplicada. Direita: Curva obtida por modelo de regressão logística fornecendo uma estimativa da probabilidade de morte para qualquer dose d num intervalo contínuo.

A medida que a dose aumenta, a proporção de besouros que morrem aumenta também. No nível mais baixo de concentração, 49.1, temos apenas 10% dos besouros morrendo mas mais de 95% morrem se a dose é 72.6 ou maior. Suponha que a dose é uma v.a. D a ser escolhida pelo produtor e dependente de condições climáticas (vento, umidade, etc.). Seja Y a variável que conta o número de “sucessos” (besouros mortos) após a exposição prolongada ao CS_2 . A partir da tabela podemos obter um modelo para a distribuição condicional de Y condicionada na dose aplicada. Precisamos especificar a probabilidade de morte de cada besouro. Ela depende da dose D . Por isto vamos escrever p_d para esta probabilidade quando a dose aplicada for $D = d$. Supondo que os besouros morrem independentemente uns dos outros (fixada a dose), temos um modelo binomial para as contagens em cada linha da tabela acima. Este modelo está na última coluna da tabela. Isto é,

$$(Y|D = d) \sim \text{Bin}(N_d, p_d)$$

onde N_d é o número de insetos expostos à dose d e p_d é a probabilidade de morte à dose d . Como é a probabilidade p_d varia como função da dose d ? Usando a ideia de frequência relativa em do evento morte em cada dose aplicada, podemos obter um valor aproximado para p_d simplesmente olhando para a proporção de besouros mortos em cada dose. A Figura 12.7 mostra a proporção de besouros mortos para cada dose d . Neste gráfico, mostramos também uma curva em forma de S derivada pelo modelo de regressão logística (a ser estudado no capítulo ??).

```

dose=c(49.1, 53.0, 56.9, 60.8, 64.8, 68.7, 72.6, 76.5)
number=c(59,60,62,56,63,59,62,60)
killed=c(6,13,18,28,52,53,61,60)
proportion = killed/number
aux = glm(cbind(killed,number-killed) ~ dose, family=binomial(link=logit))$coef
dd = seq(47, 79, by=0.1)
pr = 1/(1+exp(-(aux[1]+ aux[2]*dd)))

par(mfrow=c(1,2))
plot(dose,proportion,ylim=c(0,1))
plot(dose,proportion,ylim=c(0,1))
lines(dd, pr)

```



Figure 12.8: Esquerda: Proporção de besouros mortos versus dose aplicada. Direita: Curva obtida por modelo de regressão logística fornecendo uma estimativa da probabilidade de morte para qualquer dose d num intervalo contínuo.

■ **Example 12.13 — Caranguejos-ferradura.** Este exemplo usa dados do livro *Foundations of linear and generalized linear models* [1]. Os caranguejos-ferradura (*horseshoe-crab*, em inglês) são animais que praticamente não mudaram durante as últimas centenas de milhões de anos, já existindo na forma atual 200 milhões de anos antes dos dinossauros aparecerem, o que é um fato surpreendente do ponto de vista evolutivo. Eles possuem uma forma que lembra a ferradura de um cavalo e isto deu origem ao seu nome (Figura 12.8). Eles se reproduzem nas praias formando imensos ninhos com centenas de milhares deles ao longo da costa. Os machos chegam primeiro e aguardam as fêmeas para reprodução. Quando as fêmeas vêm para a praia, elas liberam feromônios que atraem os machos. Os machos são menores que as fêmeas e eles interceptam aquelas que passam por perto deles, agarrando-as às suas costas suas garras dianteiras especializadas (Figura 12.8). A fêmea cava de 4 a 5 buracos na areia e deposita milhares de ovos em cada um deles. O macho procura fertilizar estes ovos. Tipicamente, de 2 a 6 machos, chamados machos satélites, não conseguem agarrar-se às fêmeas mas ficam a sua volta conseguindo muitas vezes ser bem sucedidos em fertilizar os ovos.

Este estudo contou o número de machos satélites para cada uma de 173 fêmeas e procurou determinar os fatores que afetam este número. Como isto foi feito é um mistério para mim dada a confusão orgiástica que parece reinar na praia (Figura 12.8). As características do caranguejo-fêmea que poderiam afetar este número de satélites incluem a sua cor, a condição da espinha, o peso e a largura da carapaça.

Suponha que um modelo inicial para estes dados seja adotar uma distribuição de probabilidade Poisson para o número de satélites Y de uma fêmea dado intervalo em que cai a largura da sua carapaça. Seja X uma variável discreta valendo 1, 2, 3 ou 4 dependendo da carapaça da fêmea cair no intervalo $(20, 24]$, $(24, 26]$, $(26, 28]$ ou $(28, 35]$, respectivamente. Assim, temos o vetor (Y, X) medido no mesmo caranguejo-ferradura fêmea. Nossa modelo inicial é que $(Y|X = x) \sim \text{Poisson}(\lambda_x)$ onde o valor esperado λ_x vai mudar com o valor assumido pela v.a. X .

A Figura 12.9 mostra um boxplot das contagens Y para cada categoria de carapaça. O valor mediano de Y cresce com X . Talvez um modelo inicial pudesse ser, por exemplo, $(Y|X = x) \sim \text{Poisson}(\lambda_x)$ com $\lambda_x = 0.3 + x$. Este é apenas um exemplo ilustrativo, sem querer dizer que a análise mais apropriada para estes dados seja este modelo. Várias modificações podem ser feitas para melhorar o modelo. Por exemplo, podemos usar as demais características da fêmea propondo um modelo em que distribuição Poisson de Y dependerá também da cor, espinha e peso da fêmea. Podemos também dispensar a categorização da variável largura, usando-a da forma contínua como ela foi coletada originalmente. Veja o capítulo ?? para estas modificações.

```
crabs = read.table("http://www.stat.ufl.edu/~aa/glm/data/Crabs.dat", header=T)
aux = cut(crabs$width, c(20, 24, 26, 28, 35))
levels(aux)
```

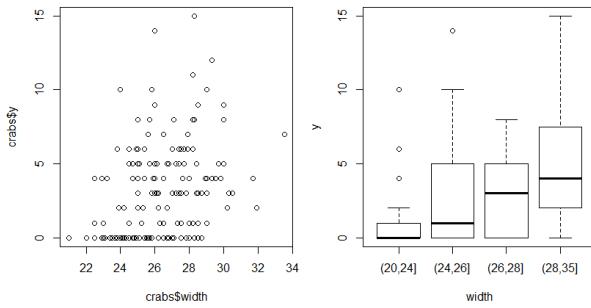


Figure 12.9: Esquerda: Gráfico do número Y de machos satélites versus a largura da carapaça da fêmea. Direita: Boxplots de Y versus a categoria X de carapaça.

```
# [1] "(20,24]" "(24,26]" "(26,28]" "(28,35]"
par(mfrow=c(1,2), mar=c(4,4,1,1))
plot(crabs$width, crabs$y)
boxplot(crabs$y ~ aux, ylab="y", xlab="width")
```

■

12.13 Esperança condicional discreta

Considere o vetor $Y = (Y_1, Y_2, \dots, Y_n)$. Mostramos como obter a distribuição condicional de Y_1 dados os valores de Y_2, \dots, Y_p . Se temos uma distribuição de probabilidade (condicional), temos as duas listas: valores possíveis e probabilidades associadas. Todas as coisas que fizemos com uma v.a. usual, nós podemos fazer também com a distribuição condicional. Por exemplo, podemos calcular o valor esperado de Y_1 dados (ou fixados) os valores de Y_2, \dots, Y_p . É simplesmente a definição usual de esperança de v.a.'s discretas mas agora usando a distribuição condicional: a soma dos valores possíveis vezes as probabilidades condicionais associadas.

Definition 12.13.1 — Esperança condicional discreta. Seja $Y = (Y_1, Y_2, \dots, Y_n)$ um vetor aleatório de variáveis discretas. O valor esperado de Y_1 condicionado nos valores $Y_2 = a_2, \dots, Y_p = a_p$ das demais v.a.'s é definido como

$$\mathbb{E}(Y_1 | Y_2 = a_2, \dots, Y_p = a_p) = \sum_y y \mathbb{P}(Y_1 = y | Y_2 = a_2, \dots, Y_p = a_p).$$

O valor esperado de qualquer das outras v.a.'s condicionado nas demais é definido de forma análoga.

Assim, a esperança condicional é simplesmente a média ponderada dos valores possíveis de Y_1 mas usando a distribuição condicional $\mathbb{P}(Y_1 = y | Y_2 = a_2, \dots, Y_p = a_p)$ de Y_1 como peso, ao invés de usar a distribuição marginal $\mathbb{P}(Y_1 = y)$ de Y_1 .

■ **Example 12.14 — Caranguejos-ferradura, de novo.** No exemplo 12.13, nós modelamos os dados dizendo que o número Y de machos-satélites em volta de uma fêmea com carapaça de largura $X = x$ seguia uma distribuição $(Y|X = x) \sim \text{Poisson}(\lambda_x)$ com $\lambda_x = 0.3 + x$. Assim, como a esperança de uma v.a. Poisson é o seu parâmetro, temos $\mathbb{E}(Y|X = x) = \lambda_x = 0.3 + x$. Isto é, $\mathbb{E}(Y|X = x)$ é uma função linear da categoria x de tamanho de carapaça. ■

12.14 Variância condicional discreta

Relembre: Se $\mu = \mathbb{E}(Y)$ então

$$\mathbb{V}(Y) = \mathbb{E}(Y - \mu)^2 = \sum_y (y - \mu)^2 \mathbb{P}(Y = y)$$

Podemos calcular a variabilidade de Y_1 em torno de sua esperança $\mathbb{E}(Y_1|Y_2 = a_2, \dots, Y_p = a_p)$ condicionada em valores das outras v.a.s (Y_2, \dots, Y_p):

$$\mathbb{V}(Y_1|Y_2 = a_2, \dots, Y_p = a_p) = \sum_y (y - m)^2 \mathbb{P}(Y_1 = y|Y_2 = a_2, \dots, Y_p = a_p)$$

onde $m = \mathbb{E}(Y_1|Y_2 = a_2, \dots, Y_p = a_p)$ é a esperança condicional. Pode-se mostrar que

$$\mathbb{V}(Y_1|Y_2 = a_2, \dots, Y_p = a_p) = \mathbb{E}(Y_1^2|Y_2 = a_2, \dots, Y_p = a_p) - m^2$$

■ **Example 12.15 — Caranguejos-ferradura, mais uma vez.** No exemplo 12.13, o número Y de machos-satélites em volta de uma fêmea com carapaça de largura $X = x$ seguia uma distribuição $(Y|X = x) \sim \text{Poisson}(\lambda_x)$ com $\lambda_x = 0.3 + x$. No caso de uma v.a. Poisson temos a sua esperança igual à sua variância, e ambas iguais ao parâmetro. Isto é, $\mathbb{V}(Y|X = x)\lambda_x = 0.3 + x$. ■

12.15 Distribuição conjunta contínua

Para explicar a distribuição conjunta no caso em que todas as v.a.'s são contínuas, vamos considerar o caso bivariado inicialmente. Seja (Y_1, Y_2) um vetor aleatório bivariado de v.a.'s contínuas. Assim, Y_1 é uma v.a. contínua e Y_2 também é uma v.a. contínua: ambas possuem densidades marginais $f_1(y)$ e $f_2(y)$. Mas ao invés de analisarmos as v.a.'s isoladamente, queremos estudar o modo como elas interagem. Existe uma versão bivariada da densidade. Vamos ver o seu significado empírico olhando para histogramas tri-dimensionais sem nos preocuparmos por enquanto em definir formalmente a função densidade.

Relembre a relação entre o histograma feito com uma amostra de uma v.a. Y e a densidade subjacente. O histograma “imita” a densidade $f(x)$. A probabilidade é igual a área debaixo da curva densidade. No caso bivariado, suponha que tenhamos uma amostra de tamanho n do vetor aleatório bivariado (Y_1, Y_2) :

$$(y_{11}, y_{12}), (y_{21}, y_{22}), (x_{31}, y_{32}), \dots, (y_{n1}, y_{n2})$$

A amostra é composta por n vetores (y_1, y_2) selecionados no plano de acordo com uma função densidade $f(y_1, y_2)$. Um histograma tri-dimensional tem aproximadamente a mesma forma que a superfície contínua $f(y_1, y_2)$ de modo que ao ver o histograma 3-dim estamos praticamente vendo a densidade $f(y_1, y_2)$. Para fazer o histograma tri-dimensional, crie uma grade regular sobre o plano e conte número de vetores (Y_1, Y_2) que caem em cada célula. A seguir, levante uma pilastra de altura proporcional a esta contagem. Regiões com mais pontos terão pilastras mais altas. Podemos dividir as alturas das pilastras por uma constante para que o volume total das pilastras seja igual a 1.

Outro exemplo ilustrativo segue na Figura 12.11.

Na Figura 12.12 temos uma amostra de 250 dados de (Y_1, Y_2) com o histograma 3d, a densidade $f(y_1, y_2)$ e suas curvas de nível.

Uma distribuição bi-dimensional mais complexa pode ser vista na Figura 12.13. Ela mostra dados do dataset quakes. Ele fornece informações sobre 1000 terremotos com magnitude maior que 4.0 na escala Richter em torno da ilha Fiji na Oceania a partir de 1964. No gráfico à esquerda, temos a longitude e latitude do epicentro desses 1000 eventos. Podemos ver a posição do epicentro

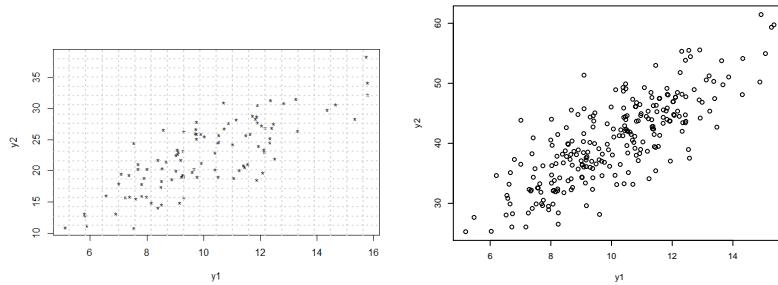


Figure 12.10: Esquerda: Amostra de 100 instâncias do vetor aleatório (Y_1, Y_2) e grade regular sobreposta. Direita: Histograma tri-dimensional baseado em amostra de vetor (X, Y) .

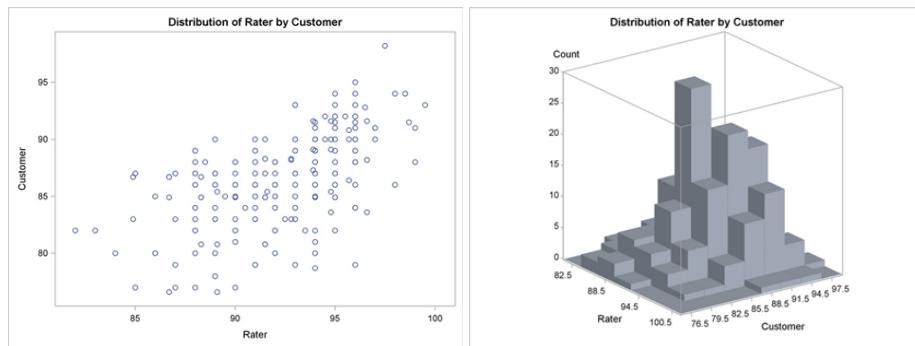


Figure 12.11: Amostra de n pontos do vetor aleatório (X, Y) e histograma tri-dimensional baseado nesta amostra.

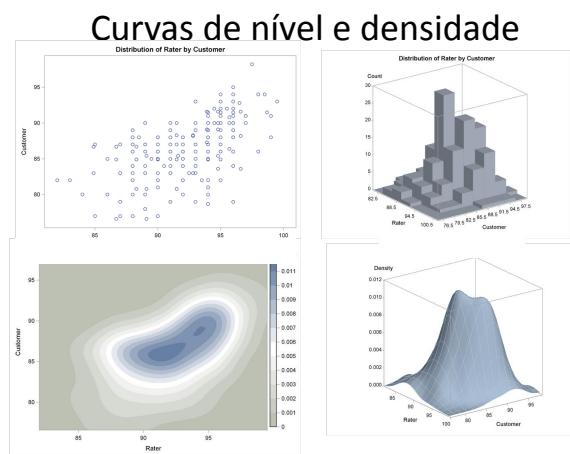


Figure 12.12: Amostra de 250 dados de (Y_1, Y_2) com histograma 3d, densidade $f(y_1, y_2)$ e suas curvas de nível

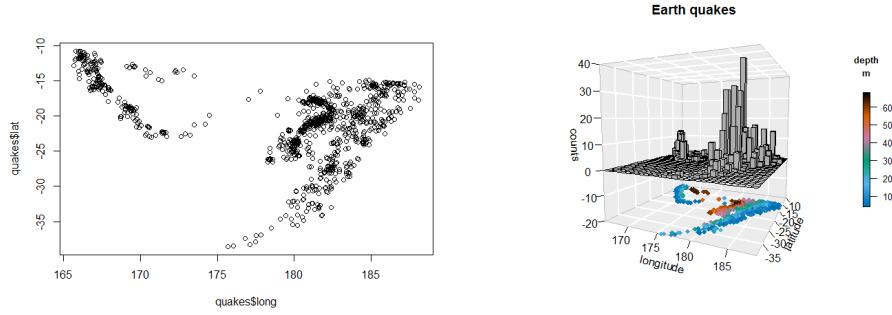


Figure 12.13: Esquerda: longitude e latitude do epicentro de 1000 terremotos. Direita: Histograma construído com `hist3D(x = xmid, y = ymid, z = xy)`.

como um vetor aleatório (X, Y) com certa densidade de probabilidade $f(x, y)$. Para visualizar mentalmente esta densidade de probabilidade, veja o histograma do lado direito da Figura 12.13.

Em cada estrela, medem-se duas v.a.'s continuas: $Y_1 = \log(\text{intensidade da luz})$ e $Y_2 = \log(\text{temperatura à superfície})$. O vetor aleatório $Y = (Y_1, Y_2)$ possui uma densidade de probabilidade $f(y_1, y_2)$ representada na Figura 12.14. Através dessa superfície $f(y_1, y_2)$, podemos responder: quais as combinações de Y_1 e Y_2 que são mais prováveis? Quais as regiões do espaço das medições em $Y = (Y_1, Y_2)$ onde existe chance razoável de se observar uma estrela?

Old Faithful é o nome de um geiser localizado no Parque Nacional Yellowstone, no estado de Wyoming, nos Estados Unidos (Figura 12.15). Ele ganhou este nome pela regularidade com que emana seus gases. O tempo de espera pela próxima erupção pode demorar aproximadamente 50 minutos ou 90 minutos. Isto depende da duração da última erupção. Uma erupção mais prolongada leva a um tempo maior de espera pela próxima. Por exemplo, uma erupção de 2 minutos leva a uma espera de aproximadamente 50 minutos enquanto que um erupção de 4.5 minutos resulta numa espera de 90 minutos.

A Figura 12.16 mostra dados do vetor (X, Y) com a duracção X em minutos de uma erupção e o tempo Y de espera pela próxima erupção. Do lado direito a densidade de probabilidade $f(x, y)$ desse vetor.

A Figura 12.17 mostra uma visão tri-dimensional da densidade $f(x, y)$ e dos pontos aleatórios vindos dessa densidade.

Amostra e densidade

Embora bonitos, gráficos tri-dimensionais não são muito úteis. Nós nunca conseguimos ver o que fica atrás dos picos e, por efeito de perspectiva, é difícil avaliar as alturas da superfícies $f(x, y)$ em diferentes posições do plano. Por isto, o melhor é usar as curvas de nível ou uma imagem com um mapa de calor para visualizar a superfície $f(x, y)$ no plano, como na Figura 12.18.

12.16 Definição formal de densidade

Seja $Y = (Y_1, Y_2)$ um vetor bivariado de v.a.'s contínuas. Uma função densidade de probabilidade é qualquer função tal que:

- $f(y_1, y_2) \geq 0$
-

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(y_1, y_2) dy_1 dy_2 = 1$$

No caso uni-dimensional, probabilidades são áreas debaixo da curva-densidade $f(x)$. No caso bi-dimensional, probabilidades são volumes debaixo da superfície-densidade $f(x, y)$. A probab do

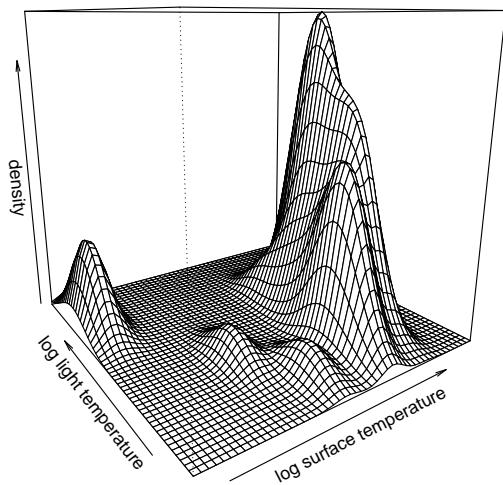


Figure 12.14: Densidade do vetor com Y_1 = Intensidade da luz e Y_2 = Temperatura de estrelas.



Figure 12.15: Old Faithful geyser.

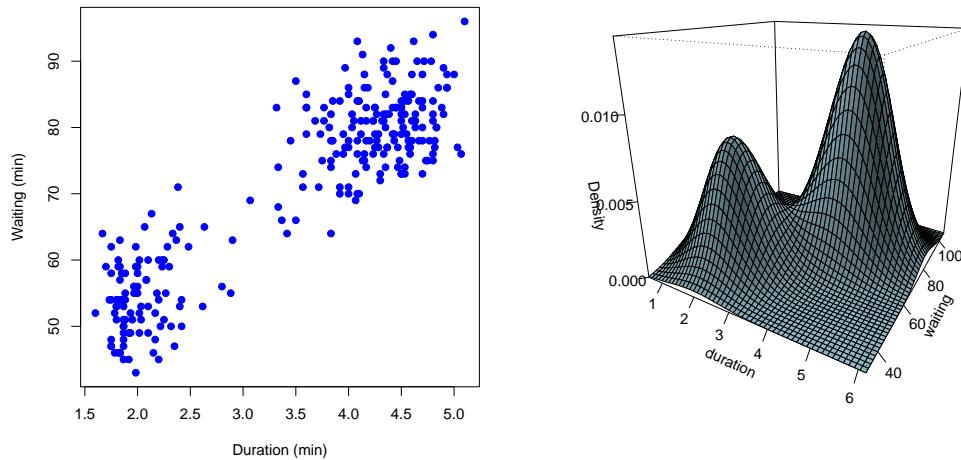


Figure 12.16: Old Faithful geyser: waiting time and eruption duration

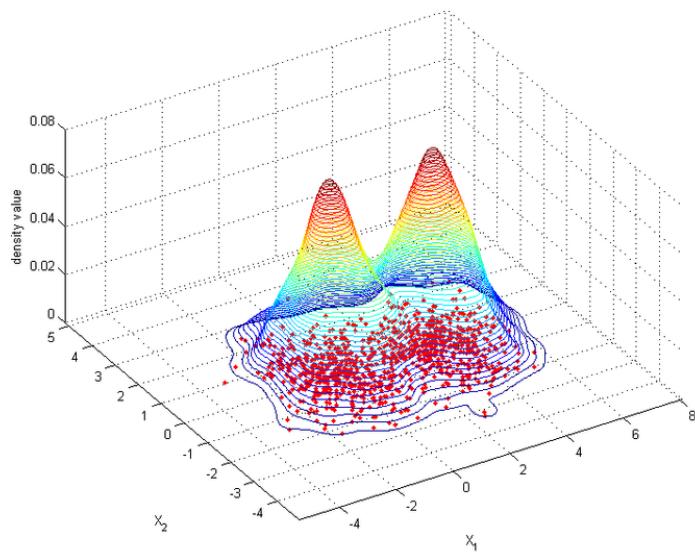


Figure 12.17: Old Faithful geyser: tempo de espera e duração de erupção.

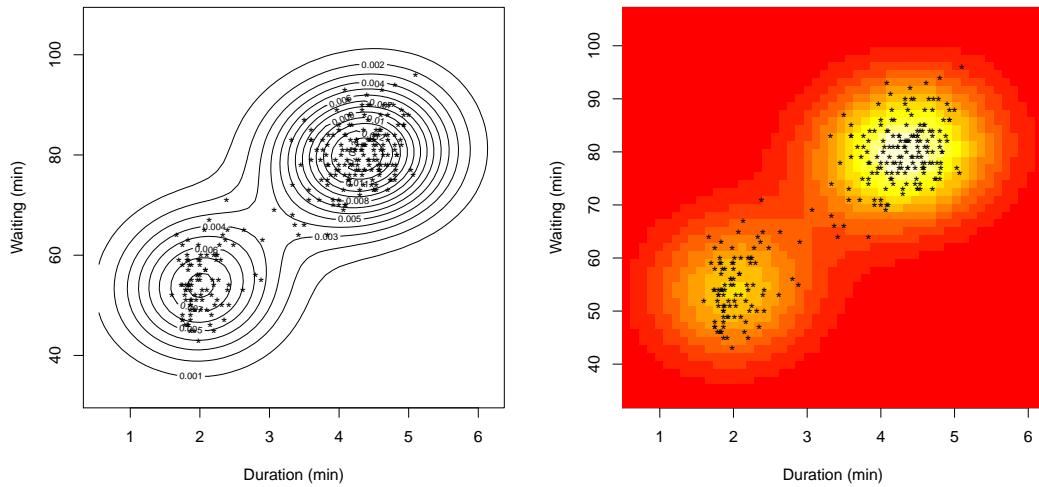
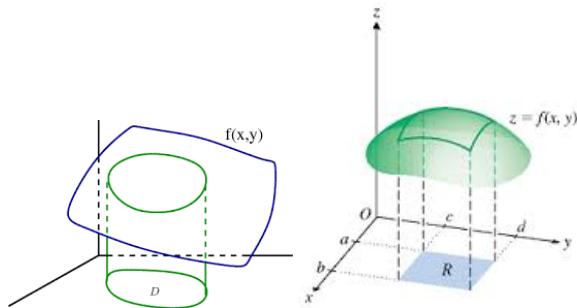


Figure 12.18: Old Faithful geyser: waiting time and eruption duration

Figure 12.19: Probabilidade de (X, Y) cair em D é igual ao volume sob a superfície.

vetor (X, Y) cair numa região D do plano é

$$\mathbb{P}((X, Y) \in D) = \int \int_D f(x, y) dx dy$$

A Figura 12.19 ilustra a situação.

No caso geral de um vetor aleatório k -dimensional $\mathbf{Y} = (Y_1, \dots, Y_k)$ com k v.a.'s contínuas, a densidade de probabilidade é qualquer função tal que:

- $f(\mathbf{y}) = f(y_1, \dots, y_k) \geq 0$ para todo ponto $\mathbf{y} \in \mathbb{R}^k$
-

$$1 = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(y_1, \dots, y_k) dy_1 \dots dy_k$$

A probabilidade do vetor \mathbf{Y} cair numa região D de \mathbb{R}^k é dada por

$$\mathbb{P}((Y_1, \dots, Y_k) \in D) = \int \dots \int_D f(y_1, \dots, y_k) dy_1 \dots dy_k$$

12.16.1 Jointly distributed random variables

Definition 12.16.1 — Joint distribution. Two random variables X, Y have *joint distribution* $F : \mathbb{R}^2 \mapsto [0, 1]$ defined by

$$F(x, y) = \mathbb{P}(X \leq x, Y \leq y).$$

The *marginal distribution* of X is

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(X \leq x, Y < \infty) = F(x, \infty) = \lim_{y \rightarrow \infty} F(x, y)$$

Definition 12.16.2 — Jointly distributed random variables. We say X_1, \dots, X_n are *jointly distributed continuous random variables* and have *joint pdf* f if for any set $A \subseteq \mathbb{R}^n$

$$\mathbb{P}((X_1, \dots, X_n) \in A) = \int_{(x_1, \dots, x_n) \in A} f(x_1, \dots, x_n) \, dx_1 \cdots dx_n.$$

where

$$f(x_1, \dots, x_n) \geq 0$$

and

$$\int_{\mathbb{R}^n} f(x_1, \dots, x_n) \, dx_1 \cdots dx_n = 1.$$

In the case where $n = 2$,

$$F(x, y) = \mathbb{P}(X \leq x, Y \leq y) = \int_{-\infty}^x \int_{-\infty}^y f(x, y) \, dy \, dx.$$

If F is differentiable, then

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y).$$

Theorem 12.16.1 If X and Y are jointly continuous random variables, then they are individually continuous random variables.

Proof: We prove this by showing that X has a density function.

We know that

$$\begin{aligned} \mathbb{P}(X \in A) &= \mathbb{P}(X \in A, Y \in (-\infty, +\infty)) \\ &= \int_{x \in A} \int_{-\infty}^{\infty} f(x, y) \, dy \, dx \\ &= \int_{x \in A} f_X(x) \, dx \end{aligned}$$

So

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) \, dy$$

is the (marginal) pdf of X .

Definition 12.16.3 — Independent continuous random variables. Continuous random variables X_1, \dots, X_n are independent if

$$\mathbb{P}(X_1 \in A_1, X_2 \in A_2, \dots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1)\mathbb{P}(X_2 \in A_2) \cdots \mathbb{P}(X_n \in A_n)$$

for all $A_i \subseteq \Omega_{X_i}$.

If we let F_{X_i} and f_{X_i} be the cdf, pdf of X_i , then

$$F(x_1, \dots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$$

and

$$f(x_1, \dots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$

are each individually equivalent to the definition above.

To show that two (or more) random variables are independent, we only have to factorize the joint pdf into factors that each only involve one variable.

If (X_1, X_2) takes a random value from $[0, 1] \times [0, 1]$, then $f(x_1, x_2) = 1$. Then we can see that $f(x_1, x_2) = 1 \cdot 1 = f(x_1) \cdot f(x_2)$. So X_1 and X_2 are independent.

On the other hand, if (Y_1, Y_2) takes a random value from $[0, 1] \times [0, 1]$ with the restriction that $Y_1 \leq Y_2$, then they are not independent, since $f(x_1, x_2) = 2I[Y_1 \leq Y_2]$, which cannot be split into two parts.

Proposition 12.16.2 For independent continuous random variables X_i ,

1. $\mathbb{E}[\prod X_i] = \prod \mathbb{E}[X_i]$
2. $\mathbb{V}(\sum X_i) = \sum \mathbb{V}(X_i)$

12.16.2 Geometric probability

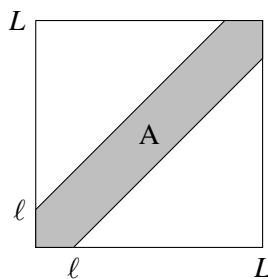
Often, when doing probability problems that involve geometry, we can visualize the outcomes with the aid of a picture.

■ **Example 12.16** Two points X and Y are chosen independently on a line segment of length L . What is the probability that $|X - Y| \leq \ell$? By “at random”, we mean

$$f(x, y) = \frac{1}{L^2},$$

since each of X and Y have pdf $1/L$.

We can visualize this on a graph:



Here the two axes are the values of X and Y , and A is the permitted region. The total area of the white part is simply the area of a square with length $L - \ell$. So the area of A is $L^2 - (L - \ell)^2 = 2L\ell - \ell^2$. So the desired probability is

$$\int_A f(x, y) \, dx \, dy = \frac{2L\ell - \ell^2}{L^2}.$$

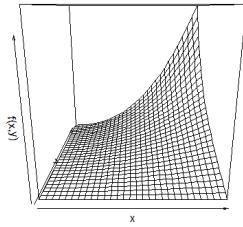


Figure 12.20: Gráfico de $f(x,y) = 60/13 (x^2y + x^3y^4)$ no suporte $[0, 1] \times [0, 1]$.

■

12.17 Marginal contínua

Distribuição marginal

No caso discreto, a distribuição marginal de uma v.a. é obtida somando-se sobre todos os valores das demais variáveis. No caso contínuo, substituímos a soma por uma integral. No caso bi-dimensional (X, Y) , a densidade de probabilidade da v.a. contínua X é obtida integrando sobre os valores de Y . Para diferenciar as densidades, vamos escrever $f_X(x)$ para a densidade marginal de X no ponto x e $f_{XY}(x,y)$ para o valor da densidade conjunta de (X, Y) no ponto (x,y) . Por exemplo, $f_X(0)$ e $f_X(1.2)$ são os valores da densidade marginal de X nos pontos $x = 0$ e $x = 1.2$. $f_{XY}(0.2, 1.5)$ é o valor da densidade conjunta no ponto $(x,y) = (0.2, 1.5)$. Para um ponto genérico x

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y) dy$$

Exercício básico

Vetor contínuo (X, Y) com suporte em $[0, 1] \times [0, 1]$ (isto é, densidade é zero fora desta região). Densidade: $f(x,y) = k(x^2y + x^3y^4)$ para $(x,y) \in [0, 1]^2$. Encontrar a constante de normalização k :

$$\begin{aligned} 1 &= \int \int_{[0,1]^2} k(x^2y + x^3y^4) dx dy = k \int_{[0,1]} \left(\frac{x^2}{2} + \frac{x^3}{5} \right) dx \\ &= k \left(\frac{1}{6} + \frac{1}{20} \right) = \frac{13k}{60} \end{aligned}$$

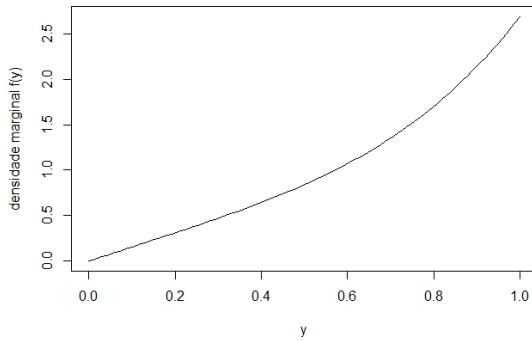
e portanto $k = 60/13$

Exercício básico

Encontrar a marginal $f_Y(y)$ para $y \in [0, 1]$:

$$f_Y(y) = \int_{[0,1]} \frac{60}{13} (x^2y + x^3y^4) dx = \frac{5}{13} (4y + 3y^4)$$

Veja que, avaliada no ponto $y = 0.1$, temos $f_Y(0.1) = 5/13 (4(0.1) + 30 \cdot 0.1^4) = 0.165$ enquanto que, no ponto $y = 0.9$, temos $f_Y(0.9) = 5/13 (4(0.9) + 30 \cdot 0.9^4) = 2.319$.

Figure 12.21: Gráfico de $f_Y(y) = 5/13(4y + 3y^4)$ no suporte $[0, 1]$.

12.18 Condicional contínua

Distribuição Condicional

No caso bi-dimensional (X, Y) , a densidade de probabilidade de X CONDICIONADA ao evento $Y = y$ é dada por

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)}$$

Por exemplo, $f_{X|Y}(x|y = 0.2)$ é a densidade da v.a. X condicionada ao evento $Y = 0.2$ e avaliada num ponto x genérico:

$$f_{X|Y}(x|y = 0.2) = \frac{f_{XY}(x, 0.2)}{f_Y(0.2)}$$

Observe que esta é uma densidade da v.a. X (variando em x) e que o denominador não depende de x . O valor $f_Y(0.2)$ é o mesmo para qualquer valor x . $f_{X|Y}(x = 0.3|y = 0.2)$ é esta densidade condicional de X avaliada no ponto $x = 0.3$:

$$f_{X|Y}(x = 0.3|y = 0.2) = \frac{f_{XY}(0.3, 0.2)}{f_Y(0.2)}$$

Exercício básico

Densidade: $f(x, y) = 60/13 (x^2 y + x^3 y^4)$ para $(x, y) \in [0, 1]^2$. Marginal $f_Y(y) = 5/13 (4y + 3y^4)$ para $y \in [0, 1]$: Densidade de X condicionada ao evento $Y = 0.2$:

$$f_{X|Y}(x|y = 0.2) = \frac{f_{XY}(x, 0.2)}{f_Y(0.2)} = \frac{60/13 (0.2 x^2 + 0.2^4 x^3)}{5/13 (4 \cdot 0.2 + 3 \cdot 0.2^4)} = \frac{12}{0.8048} (0.2 x^2 + 0.0016 x^3)$$

Exercício básico

Comparando duas densidades condicionais de X : condicionada ao evento $Y = 0.20$ e ao evento $Y = 0.95$.

$$\begin{aligned} f_{X|Y}(x|y = 0.95) &= \frac{f_{XY}(x, 0.95)}{f_Y(0.95)} \\ &= \frac{60/13 (0.95 x^2 + 0.95^4 x^3)}{5/13 (4 \cdot 0.95 + 3 \cdot 0.95^4)} = \frac{60}{31.217} (0.95 x^2 + 0.95^4 x^3) \end{aligned}$$

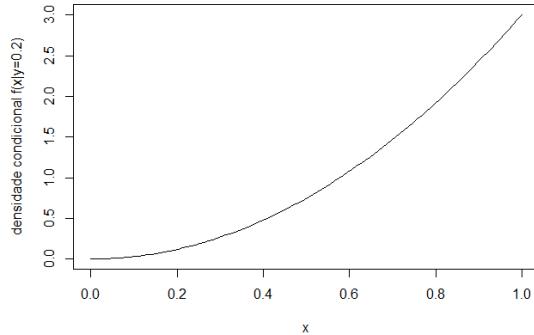


Figure 12.22: Gráfico de $f_{X|Y}(x|y = 0.2) = 12/0.8048 (0.2x^2 + 0.0016x^3)$ no suporte $[0, 1]$.

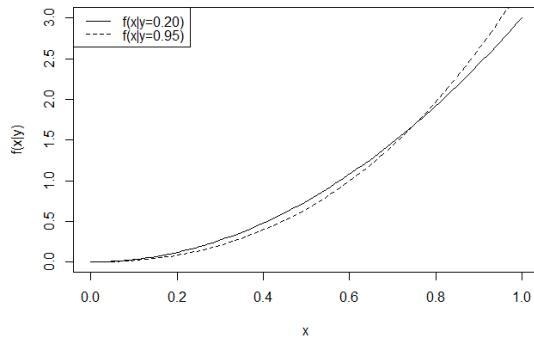


Figure 12.23: $f_{X|Y}(x|y = 0.20)$ e $f_{X|Y}(x|y = 0.95)$.

Não são muito diferentes neste exemplo particular.

Mais um exemplo - gaussiana

Densidade para (X, Y) é

$$f_{XY}(x, y) = \frac{1}{2\pi\sqrt{0.51}} \exp\left(-\frac{x^2 + y^2 - 1.4xy}{1.02}\right)$$

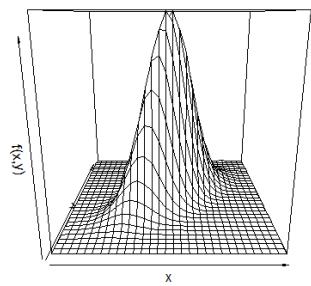
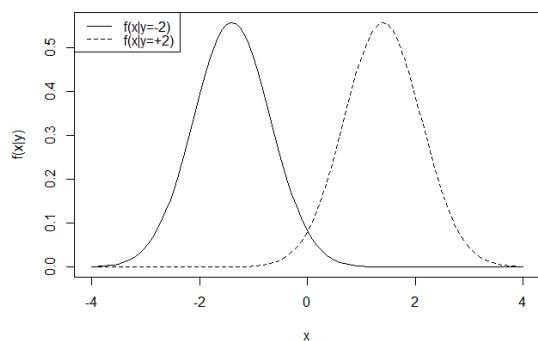
com suporte em \mathbb{R}^2 . Esta é a densidade de uma gaussiana bivariada onde a correlação é igual a $\rho = 0.7$ e as marginais são $X \sim N(0, 1)$ e $Y \sim N(0, 1)$. Marginal $f_Y(y) = 1/\sqrt{2\pi} \exp(-y^2/2)$ para $y \in \mathbb{R}$.

Densidade $(X|Y = -2)$:

$$\begin{aligned} f_{X|Y}(x|y = -2) &= \frac{f_{XY}(x, -2)}{f_Y(-2)} = \frac{\frac{1}{2\pi\sqrt{0.51}} \exp\left(-\frac{x^2 + (-2)^2 - 1.4x(-2)}{1.02}\right)}{1/\sqrt{2\pi} \exp(-(-2)^2/2)} \\ &= 1/\sqrt{1.02\pi} \exp\left(-\frac{(x + 1.4)^2}{1.02}\right) \end{aligned}$$

De forma similar, obtemos $f_{X|Y}(x|y = +2)$. Gráficos abaixo.

Vendo a condicional na conjunta

Figure 12.24: Densidade gaussiana bivariada $f_{XY}(x,y)$.Figure 12.25: Gráfico de $f_{X|Y}(x|y = -2)$ e $f_{X|Y}(x|y = +2)$.

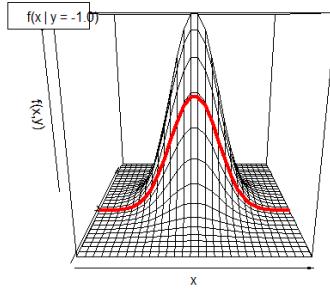


Figure 12.26: Gráfico de $f_{XY}(x, -1.0)$, que é proporcional a $f_{X|Y}(x|y = -1.0)$.

Olhar a superfície da densidade $f(x, y)$ mostra imediatamente a forma (shape) da densidade condicional. Por exemplo,

$$f_{X|Y}(x|y = 0.2) = \frac{f_{XY}(x, 0.2)}{f_Y(0.2)} \propto f_{XY}(x, 0.2)$$

pois o denominador é uma constante *COM RESPEITO A x*. Assim, se quisermos saber como $f_{X|Y}(x|y = 0.2)$ varia como função de x , basta olharmos na superfície $f(x, y)$ a curva obtida se fixarmos $y = 2$.

Vendo a condicional na conjunta

$f(x|y = 1.0)$ tem a mesma forma(shape) que a curva em vermelho, que é $f_{XY}(x, -1.0)$, os valores da densidade conjunta com $y = -1.0$ fixo. A densidade condicional é esta curva multiplicada por uma constante positiva.

12.19 Esperança condicional

Considere o vetor $Y = (Y_1, Y_2, \dots, Y_p)$. Calculamos a distribuição condicional de Y_1 dados os valores de Y_2, \dots, Y_p . Podemos calcular o valor esperado de Y_1 dados (ou fixados) os valores de Y_2, \dots, Y_p . É simplesmente como na definição usual de esperança de v.a.'s discretas:

$$\mathbb{E}(Y_1|Y_2 = a_2, \dots, Y_p = a_p) = \int y f_{Y_1|Y_2 \dots Y_p}(y|y_2 = a_2, \dots, y_p = a_p) dy$$

Média ponderada dos valores possíveis de Y_1 MAS USANDO a densidade condicional de Y_1 como peso, ao invés de usar a distribuição marginal de Y_1 .

12.20 Variância condicional

Relembre: Se $\mu = \mathbb{E}(Y)$ então

$$\mathbb{V}(Y) = \mathbb{E}(Y - \mu)^2 = \int (y - \mu)^2 f_Y(y) dy$$

Podemos calcular a variabilidade de Y_1 em torno de sua esperança CONDICIONADA nos valores das outras v.a.s (Y_2, \dots, Y_p):

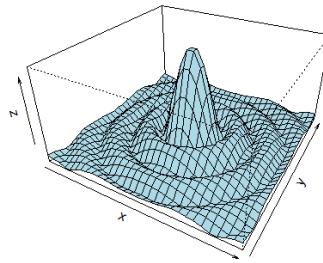


Figure 12.27: Gráfico de $f_{XY}(x,y)$, de onde queremos simular.

$$\mathbb{V}(Y_1|Y_2 = a_2, \dots, Y_p = a_p) = \int (y - m)^2 f_{Y_1|Y_2 \dots Y_p}(y|y_2 = a_2, \dots, y_p = a_p) dy$$

onde $m = \mathbb{E}(Y_1|Y_2 = a_2, \dots, Y_p = a_p)$ é a esperança condicional.
Pode-se mostrar que

$$\mathbb{V}(Y_1|Y_2 = a_2, \dots, Y_p = a_p) = \mathbb{E}(Y_1^2|Y_2 = a_2, \dots, Y_p = a_p) - m^2$$

12.21 Simulação de um vetor contínuo

Simulando um vetor contínuo

Queremos simular uma amostra do vetor aleatório bivariado (X, Y) com densidade $f(x, y)$. Existem vários métodos (ver disciplina PGM - Probabilistic Graphical Models) Um método simples é o de aceitação-rejeição. Obtenha uma densidade $g(x, y)$ de onde você saiba simular. Encontre M tal que $f(x, y) \leq M g(x, y)$ para todo ponto (x, y) .

```
while(contador < nsim){
    gere (x,y) de g(x,y)
    jogue moeda com P(cara) = f(x,y)/(M*g(x,y))
    se cara:
        aceite (x,y)
        contador = contador + 1
}
```

Exemplo: Simulando um vetor contínuo

Queremos simular 100 pontos aleatórios (x, y) seguindo a densidade $f_{XY}(x, y)$ com suporte em $[-10, 10]^2$ e dada por

$$f_{XY}(x, y) = \frac{|\sin(r(x, y))|}{44 r(x, y)}$$

onde $r(x, y) = \sqrt{x^2 + y^2}$ é a distância de (x, y) à origem. O máximo de $f_{XY}(x, y)$ ocorre em $(x, y) = (0, 0)$ e é igual a $1/44 \approx 0.0228$.

Exemplo: Simulando um vetor contínuo

Vamos simular (X, Y) em $[-10, 10]^2$ com uma distribuição uniforme. Isto é, a densidade é igual a $g(x, y) = 1/20^2$ em $[-10, 10]^2$ e igual a zero fora dessa região. Gerar desta $g(x, y)$ é muito fácil pois X e Y são independentes e cada uma delas segue uma uniforme em $[-10, 10]$. Assim, gere a coordenada $X \sim U(-10, 10)$ e independentemente a coordenada $Y \sim U(-10, 10)$.

```
x = runif(1000, -10, 10)
y = runif(1000, -10, 10)
```

A seguir, retenha ou descarte estes valores com probabilidade $f(x,y)/(Mg(x,y))$. Quem é M ?

Exemplo: Simulando um vetor contínuo

Temos $g(x,y) = 1/400$ para todo (x,y) na região. Queremos $1 > f(x,y)/(Mg(x,y)) = 400f(x,y)/M$. Como o máximo de $f(x,y)$ ocorre na origem e é igual a $1/44$, podemos ter certeza que

$$\frac{f(x,y)}{Mg(x,y)} = \frac{400f(x,y)}{M} \leq \frac{400f(0,0)}{M} = \frac{400}{44M} < 1$$

se tomarmos $M > 400/44 = 9.090909$. Vamos tomar $M = 10$. Assim, basta reter os pontos (x,y) tais que a sua “moeda” resulte em cara onde

$$\mathbb{P}(\text{cara}) = \frac{400f_{XY}(x,y)}{10} = 40f_{XY}(x,y)$$

Exemplo: Simulando um vetor contínuo

```
n = 1000; contador = 0
amostra = matrix(0, ncol=2, nrow=n)
while(contador < n)
{
  x = runif(1, -10, 10)
  y = runif(1, -10, 10)
  r = sqrt(x^2+y^2)
  fxy = abs(sin(r))/(44*r)
  prob = 40 * fxy
  if(runif(1) < prob){
    contador = contador + 1
    amostra[contador, ] = c(x,y)
  }
}
plot(amostra, asp=1)
```

Amostra gerada de $f(x,y)$

Amostra gerada de $f(x,y)$

Amostra gerada de $f(x,y)$

```
x <- seq(-10, 10, length= 30)
y <- x
f <- function(x, y) {
  r <- sqrt(x^2+y^2);
  abs(sin(r))/(44*r)
}
z <- outer(x, y, f)
image(x,y,log(z), asp=1)
points(amostra)
```

?? FUNCOES DE MAIS DE UMA V.A. ?? MINIMUM E MAXIMUM ??? PROPRIEDADES DE ESPERANCA E VARIANCIA

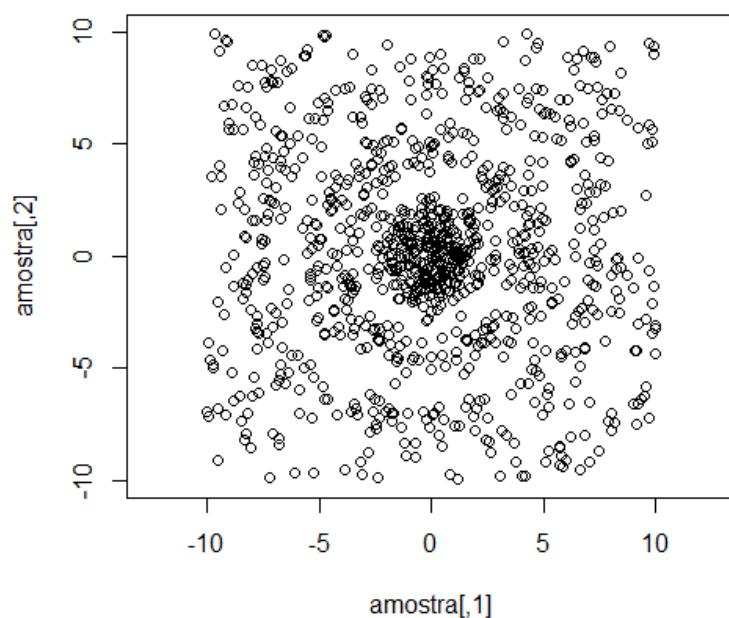


Figure 12.28: Amostra gerada de $f(x,y)$ por aceitação-rejeição

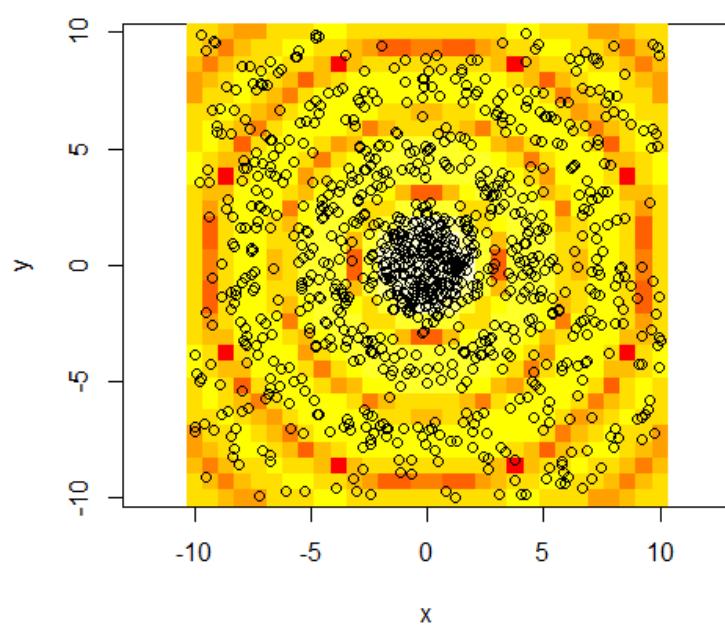


Figure 12.29: Amostra de $f(x,y)$ e imagem heatmap da densidade.



13. Distribuição Normal Multivariada

13.1 Normal bivariada: introdução

A distribuição gaussiana multivariada é extremamente importante para a análise de dados por causa do Teorema Central do Limite. Vamos começar estudando o caso bi-dimensional, em que temos um vetor aleatório $Y = (Y_1, Y_2)$. Cada uma das v.a's separadamente segue uma gaussiana com sua própria esperança μ_j e variância σ_j^2 . Isto é, $Y_1 \sim N(\mu_1, \sigma_1^2)$ e $Y_2 \sim N(\mu_2, \sigma_2^2)$. No caso de uma gaussiana bivariada, as amostras do vetor bivariado Y formam nuvens de pontos no plano em forma de elipses centradas em (μ_1, μ_2) . Além disso, elas não são (em geral) independentes: A distribuição de Y_2 muda se soubermos o valor d.v.a. Y_1 . Um único parâmetro $\rho \in [-1, 1]$ controla como ocorre esta dependência. Este parâmetro ρ é chamado de índice de correlação ou associação linear entre Y_1 e Y_2 .

A Figura 13.1 mostra o gráfico da densidade de probabilidade $f(y_1, y_2)$ de uma distribuição gaussiana com $Y_1 \sim N(\mu_1 = 0, \sigma_1^2 = 1)$, $Y_2 \sim N(\mu_2 = 0, \sigma_2^2 = 1)$ e com correlação linear $\rho = 0$. Ela mostra também uma amostra aleatória de $n = 100$ instâncias do vetor aleatório $Y = (Y_1, Y_2)$ extraídas desta distribuição. Isto é, vemos a amostra composta pelos $n = 100$ pares de valores $Y_i = (Y_{i1}, Y_{i2})$ com $i = 1, \dots, 100$.

```
#densidade normal bivariada padrao
x = seq(-5, 5, length= 40); y <- x
f = function(x,y) { dnorm(x)*dnorm(y) }
z = outer(x, y, f)
par(mfrow=c(1,2), mar=c(5,5,1,1))
persp(x, y, z, theta = 30, phi = 30, expand = 0.5, col = "lightblue")
plot(rnorm(200), rnorm(200), xlab="y_1", ylab="y_2")
```

A Figura 13.2 mostra o gráfico da densidade de probabilidade $f(y_1, y_2)$ de uma distribuição gaussiana com Y_1 e Y_2 com as mesmas distribuições marginais da Figura 13.1 mas com correlação linear $\rho = 0.7$. N ldo direito, temos uma amostra aleatória de $n = 100$ exemplos de $Y = (Y_1, Y_2)$. O código abaixo mostra como gerar a figura, a superfície da densidade de probabilidade e a amostra.

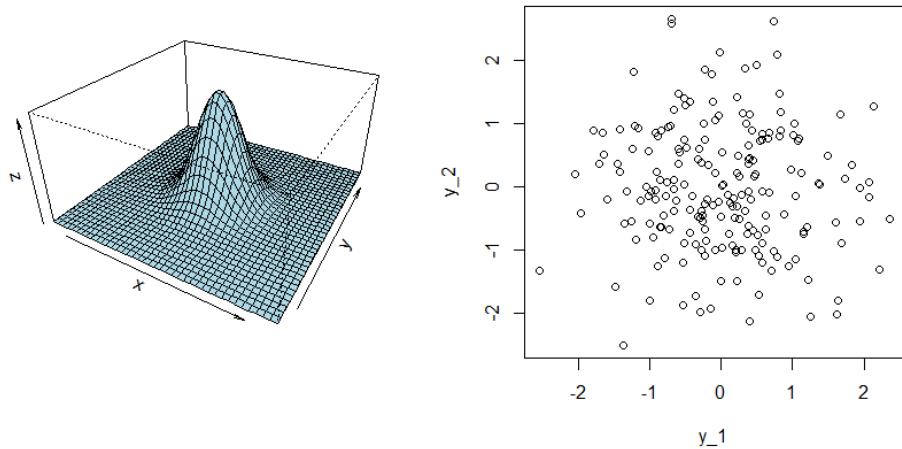


Figure 13.1: Esquerda: Densidade $f(y_1, y_2)$ de uma distribuição normal bivariada com $Y_1 \sim N(\mu_1 = 0, \sigma_1^2 = 1)$ e $Y_2 \sim N(\mu_2 = 0, \sigma_2^2 = 1)$ e com correlação $\rho = 0$. Direita: amostra com $n = 100$ instâncias do vetor Y .

Usamos a biblioteca *MASS* para gerar a amostra. Os detalhes do código serão explicados ao longo deste capítulo.

```
# densidade normal bivariada (0,1)^2 mas com rho=0.7
f <- function(x,y, rho=0.7){exp(-(x^2 - 2*rho*x*y + y^2)/(2*(1-rho^2)))/(2*pi*sqrt(1-rho^2))
z <- outer(x, y, f)
par(mfrow=c(1,2), mar=c(5,5,1,1))
persp(x, y, z, theta = 30, phi = 30, expand = 0.5, col = "lightblue")
library(MASS); set.seed(3)
plot(mvrnorm(n = 200, c(10,50),
matrix(c(2.5^2, 0.7*2.5*15, 0.7*2.5*15, 15^2), ncol=2)), xlab="y1", ylab="y2")
```

A partir das margens do gráfico com a nuvem de pontos desta última amostra, podemos reconhecer facilmente os dois primeiros momentos marginais. Para a v.a. Y_1 , temos $\mathbb{E}(Y_1) = \mu_1 \approx 10$ e $\sigma_1 \approx 2.5$. Para a v.a. Y_2 , temos $\mathbb{E}(Y_2) = \mu_2 \approx 50$ e $\sqrt{\mathbb{V}(Y_2)} = \sigma_2 \approx 15$.

Os valores de Y_1 e Y_2 medidos num mesmo elemento amostral ω não são independentes. O valor da v.a. Y_1 dá informação sobre o valor da v.a. Y_2 . Como assim? Vamos ser mais específicos. Qual a distribuição de Y_2 dado que $Y_1 = 14$? O que podemos dizer do valor esperado de Y_2 dado que $Y_1 = 14$? Este valor esperado continua igual à esperança marginal $\mu_2 = 50$? A Figura 13.3 mostra uma linha vertical na posição $Y_1 = 14$. Dizer que condicionamos a v.a. $Y_1 = 14$ significa dizer que o vetor aleatório Y é da forma $Y = (Y_1, Y_2) = (14, Y_2)$. A primeira coordenada já está fixada e apenas na segunda coordenada ainda existe incerteza sobre seu valor. A partir da amostra, vemos que fixando $Y_1 = 14$, não é mais razoável esperar que Y_2 oscile em torno de $\mu_2 \approx 50$. O valor esperado condicional, $\mathbb{E}(Y_2|Y_1 = 14)$, deve ser maior que $\mu_2 = 50$.

Qual a sua estimativa para $\mathbb{E}(Y_2|Y_1 = 14)$ no olhômetro? Suponha que um ponto aleatório será escolhido da distribuição condicional de Y_2 dado que $Y_1 = 14$. O ponto aleatório Y estará na linha vertical $(14, y_2)$. Os pontos (y_1, y_2) da amostra que possuem $y_1 \approx 14$ indicam o que deve ser o comportamento probabilístico da v.a. Y_2 dado que $Y_1 = 14$. A partir desses pontos com $Y_1 \approx 14$, vemos que $\mathbb{E}(Y_2|Y_1 = 14) \approx 70$. Assim, $\mathbb{E}(Y_2|Y_1 = 14)$ é muito maior que $50 = \mathbb{E}(Y_2) = \mu_2$, a esperança marginal da v.a. Y_2 . A esperança condicional $\mathbb{E}(Y_2|Y_1 = 14)$ é bem maior que a esperança marginal $\mathbb{E}(Y_2)$.

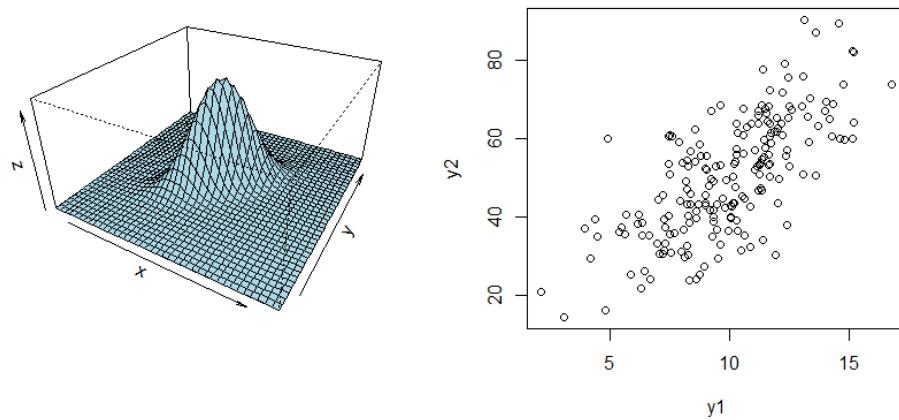


Figure 13.2: Esquerda: Densidade $f(y_1, y_2)$ de uma distribuição normal bivariada com $Y_1 \sim N(\mu_1 = 0, \sigma_1^2 = 1)$ e $Y_2 \sim N(\mu_2 = 0, \sigma_2^2 = 1)$ e com correlação $\rho = 0.7$. Direita: amostra com $n = 100$ instâncias do vetor Y .

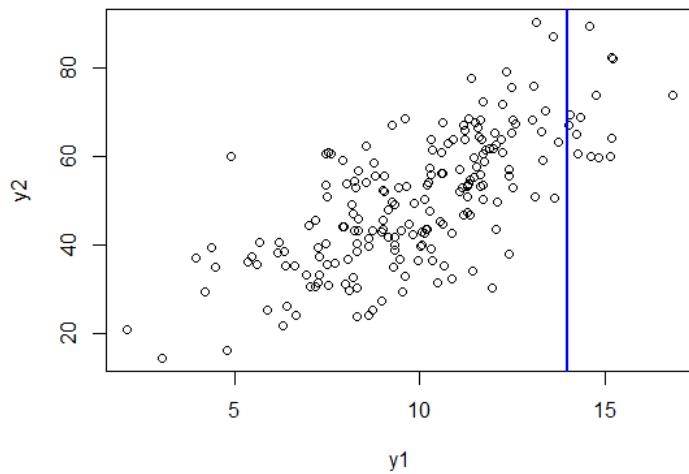


Figure 13.3: Amostra de normal bivariada. Se soubermos que o valor da v.a. Y_1 é igual a 14, o vetor aleatório Y é da forma $Y = (Y_1, Y_2) = (14, Y_2)$.

Se $\mathbb{E}(Y_2|Y_1 = 14) \approx 70$, quanto é o desvio-padrão $\sqrt{\mathbb{V}(Y_2|Y_1 = 14)}$ da distribuição de Y_2 condicionada em $Y_1 = 14$? Olhando os pontos (y_1, y_2) que possuem $y_1 \approx 14$, qual o tamanho médio dos desvios de Y_2 em torno de sua esperança condicional $\mathbb{E}(Y_2|Y_1 = 14) \approx 70$? Grosseiramente, esses pontos estão no intervalo de $[50, 80]$. Eu chutaria (ou estimaria) que $\sqrt{\mathbb{V}(Y_2|Y_1 = 14)} \approx (80 - 30)/4 = 7.5$. Veja que $7.5 < 15 = \sqrt{\mathbb{V}(Y_2)}$, que é o desvio-padrão marginal de Y_2 .

Estes são os dois primeiros momentos condicionais da v.a. $(Y_2|Y_1 = 14)$, a esperança e variância condicionais. Eles são apenas resumos da distribuição de probabilidade da v.a. $(Y_2|Y_1 = 14)$. Qual é a distribuição de probabilidade de $(Y_2|Y_1 = 14)$? Uma normal? Uma gama? Uma uniforme? Pode-se mostrar que, se o vetor $Y = (Y_1, Y_2)$ segue uma normal bivariada, então $(Y_2|Y_1 = 14)$ é uma v.a. com distribuição normal. Isto é, $(Y_2|Y_1 = 14) \approx N(70, 7.5^2)$

E que tal $Y_2|Y_1 = y$ com y genérico? Conseguimos obter uma fórmula geral para expressar qual é esta distribuição genérica. Ela depende do coeficiente de correlação ρ que neste exemplo vale $\rho = 0.7$. Temos

$$(Y_2|Y_1 = y) \sim N(\mu_{Y_2|Y_1=y}, \sigma_{Y_2|Y_1=y}^2)$$

com

$$\mu_{Y_2|Y_1=y} = \mu_2 + \frac{\rho\sigma_2}{\sigma_1}(y - \mu_1)$$

e

$$\sigma_{Y_2|Y_1=y} = \sigma_2 \sqrt{1 - \rho^2}.$$

Por exemplo, com $y_2 = 14$ temos

$$\mu_{Y_2|Y_1=14} = \mu_2 + \frac{\rho\sigma_2}{\sigma_1}(14 - \mu_1) = 50 + \frac{0.7 * 15}{2.5}(14 - 10) = 66.8$$

e

$$\sigma_{Y_2|Y_1=y} = \sigma_2 \sqrt{1 - \rho^2} = 15 \sqrt{1 - 0.7^2} = 10.71$$

e portanto

$$(Y_2|Y_1 = 14) \sim N(69.2, 10.71^2)$$

Como sabemos essa fórmula? Fazendo o cálculo matemático da densidade condicional:

$$f_{Y_2|Y_1}(y_2|y_1 = a) = \frac{f_Y(a, y_2)}{f_{Y_1}(a)}$$

a partir da densidade conjunta da normal bivariada. Até agora não mostramos a expressão da densidade conjunta $f(y_1, y_2)$ de uma gaussiana bivariada. Mostramos apenas gráficos dessa densidade. Para apresentar esta densidade conjunta, vamos começar definindo a matriz 2×2 de covariância Σ dada por

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

onde ρ é um parâmetro com valor no intervalo $[-1, 1]$, e onde σ_1 e σ_2 são os desvios padrões de cada marginal. Esta matriz é simétrica já que o elemento $(1, 2)$ é igual ao elemento $(2, 1)$. Esta matriz ser simétrica terá consequências importantes até o final o capítulo.

Seja o vetor-coluna 2×1 das esperanças marginais:

$$\mu = (\mu_1, \mu_2)' = (\mathbb{E}(Y_1), \mathbb{E}(Y_2))'$$

A fórmula geral da densidade de uma normal bivariada é igual a

$$f_Y(\mathbf{y}) = \text{cte} \times \exp\left(-\frac{1}{2} d^2(\mathbf{y}, \mu)\right)$$

onde $d^2(\mathbf{y}, \mu)$ é uma medida de distância entre o ponto $\mathbf{y} = (y_1, y_2)'$ e o vetor esperado $\mu = (\mu_1, \mu_2)'$. Quanto mais distante $\mathbf{y} = (y_1, y_2)'$ estiver do vetor esperado $\mu = (\mu_1, \mu_2)'$, menor o valor da densidade. Assim, a densidade é maior em pontos $\mathbf{y} = (y_1, y_2)'$ que estejam próximos de $\mu = (\mu_1, \mu_2)'$ e decai exponencialmente à medida que $\mathbf{y} = (y_1, y_2)'$ se afasta de $\mu = (\mu_1, \mu_2)'$.

Esta medida de distância é muito importante e ela *não* é a distância euclidiana. Ela é chamada de distância de Mahalanobis e é dada por

$$d^2(\mathbf{y}, \mu) = (\mathbf{y} - \mu)' \Sigma^{-1} (\mathbf{y} - \mu) \quad (13.1)$$

Vamos estudá-la na seção 13.7.

A fórmula da densidade da normal bivariada generaliza-se para um vetor de dimensão p . Um vetor normal multivariado tem uma densidade conjunta que é proporcional à uma exponencial que decai com uma medida de distância entre o vetor $\mathbf{y} = (y_1, \dots, y_p)'$ e o vetor esperado $\mu = (\mu_1, \dots, \mu_p)'$:

$$f_Y(\mathbf{y}) = \text{cte} \times \exp\left(-\frac{1}{2} d^2(\mathbf{y}, \mu)\right)$$

onde $d^2(\mathbf{y}, \mu)$ é a distância de Mahalanobis dada em (13.1). Antes de estudarmos a distância de Mahalanobis, vamos analisar a matriz de covariância Σ .

13.2 O índice ρ de correlação de Pearson

Podemos *resumir* a distribuição de probabilidade de uma v.a. Y com os resumos numéricos e teóricos representados pela esperança $\mathbb{E}(Y)$ e o desvio-padrão $DP_Y = \sqrt{\mathbb{V}(Y)}$. Os resumos $\mathbb{E}(Y)$ e $DP_Y = \sqrt{\mathbb{V}(Y)}$ não dependem de dados estatísticos. São resultados de cálculos matemáticos e resumem a distribuição teórica de uma v.a. $\mathbb{E}(Y)$ é um valor em torno do qual os dados tendem a oscilar e $DP_Y = \sqrt{\mathbb{V}(Y)}$ o valor típico do afastamento dos dados em relação a $\mathbb{E}(Y)$.

Vamos agora passar a olhar os dados estatísticos. Suponha que temos uma amostra aleatória de Y . Isto é, v.a.'s Y_1, Y_2, \dots, Y_n i.i.d. com a mesma distribuição que Y . Estes n números ficam numa das colunas de nossa tabela de dados. Para ter uma idéia de TODA A distribuição de probabilidade de Y podemos fazer um histograma dos dados. A forma do histograma segue aproximadamente a densidade $f(y)$.

A contraparte empírica dos resumos: podemos estimar os resumos *teóricos* $\mathbb{E}(Y)$ e $\sigma = DP_Y = \sqrt{\mathbb{V}(Y)}$ a partir dos dados. Pela Lei dos Grandes Números (ver capítulo 16), se o tamanho n da amostra é grande, temos a média aritmética $\bar{Y} = (Y_1 + \dots + Y_n) \approx \mathbb{E}(Y)$ e $DP_{\text{amostral}} S = \sqrt{\sum_i (Y_i - \bar{Y})^2 / n} \approx \sigma$. Às vezes, define-se o DP amostral S usando $n - 1$ no denominador. A diferença é mínima a não ser que n seja muito pequeno. Note que $\bar{Y} \neq \mathbb{E}(Y)$ e $S \neq \sigma$. As estatísticas \bar{Y} e S dependem dos dados e variam de amostra para amostra, mesmo que o mecanismo gerador das amostras não mude.

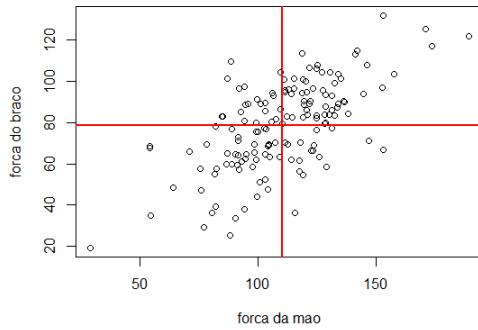


Figure 13.4: Relação entre força de preensão (do aperto de mão) e força do braço para 147 pessoas que trabalham em empregos fisicamente extenuantes.

Definition 13.2.1 — Desvio padronizado. O desvio da v.a. Y em relação a seu valor esperado $\mu = \mathbb{E}(Y)$ é a v.a. $Y - \mu$. O desvio padronizado é definido como $Z = (Y - \mu)/\sigma$.

Assim, o desvio padronizado é medido relativamente ao desvio-padrão σ da v.a. Y . Um desvio padronizado $Z = 2$ significa um afastamento da v.a. Y de 2 DPs em relação a μ . Pela desigualdade de Tchebyshev, vimos que, qualquer que seja a distribuição de Y , temos que o evento $Z > 4$ é muito raro (tem baixa probabilidade de ocorrer).

Como medir a associação entre duas variáveis Y_1 e Y_2 medidas num mesmo elemento ω ? Estas variáveis poderiam ser qualquer par de colunas da nossa tabela de dados. Seja $Z_1 = (Y_1 - \mu_1)/\sigma_1$ o desvio padronizado de Y_1 e $Z_2 = (Y_2 - \mu_2)/\sigma_2$ o desvio padronizado de Y_2 . Quando Z_1 é grande existe alguma tendência de também termos Z_2 grande? Se sim, diremos que Y_1 e Y_2 possuem um grau de associação ou correlação. Como formalizar este conceito?

Vamos começar com a versão empírica da associação. A Figura 13.4 mostra os dados de uma amostra de 147 pessoas (os itens) trabalhando em ocupações fisicamente demandantes. Em cada indivíduo, medimos o par de variáveis (Y_1, Y_2) onde Y_1 é a força do aperto de mão (ou *grip strength*) e Y_2 é a força do braço (ou *arm strength*). As linhas vertical e horizontal em vermelho indicam aproximadamente os valores de $\mathbb{E}(Y_1) = \mu_1$ e $\mathbb{E}(Y_2) = \mu_2$. A maioria dos pontos está nos quadrantes 1 e 3. Quando $Z_1 > 0$, em geral, temos também $Z_2 > 0$. E quando $Z_1 < 0$, costumamos ter $Z_2 < 0$.

Existem várias formas intuitivas de medir a associação entre Y_1 e Y_2 . Uma forma *não intuitiva* mas que tem excelentes propriedades teóricas é o índice de correlação de Pearson. Considere o produto dos desvios padronizados:

$$Z_1 Z_2 = \frac{Y_1 - \mu_1}{\sigma_1} \times \frac{Y_2 - \mu_2}{\sigma_2}$$

Se desvios grandes e positivos de Y_1 tendem a ocorrer com desvios grandes e positivos de Y_2 , seu produto será maior ainda. Ao mesmo tempo, se os desvios grandes e negativos de Y_1 tendem a ocorrer com desvios grandes e positivos de Y_2 , seu produto será maior ainda. A Figura 13.5 mostra o comportamento do produto dos desvios padronizados. Tipicamente, em média, o produto dos desvios padronizados $Z_1 Z_2$ é positivo (esquerda), próximo de zero (centro) e negativo (direita).

Vamos olhar um pouco mais a natureza de $Z_1 Z_2$. (Y_1, Y_2) é um vetor aleatório, com os valores de duas v.a.'s são medidas no mesmo item. Considere

$$Z_1 Z_2 = \frac{Y_1 - \mu_1}{\sigma_1} \times \frac{Y_2 - \mu_2}{\sigma_2}$$

Sabemos que é μ_1 é uma constante, um valor numérico fixo e teórico obtido a partir da distribuição de Y_1 . O mesmo vale para μ_2 , σ_1 e σ_2 , todas constantes. E o produto $Z_1 Z_2$? Este produto é uma

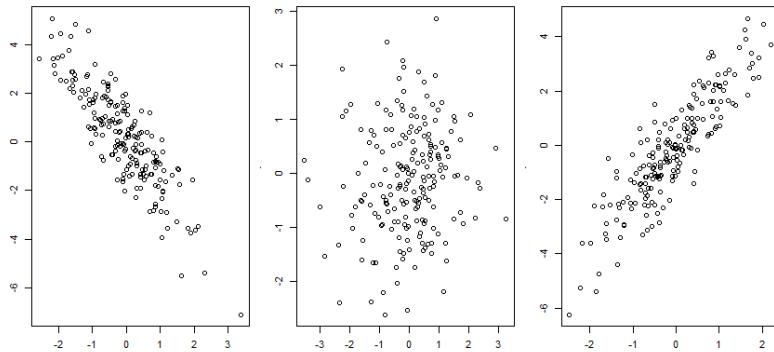


Figure 13.5: Tipicamente, em média, o produto dos desvios padronizados Z_1Z_2 é positivo (esquerda), próximo de zero (centro) e negativo (direita).

v.a. Como tal, possui lista de valores possíveis e lista de probabilidades associadas. Ao invés de obtermos esta duas listas, uma tarefa complicada na maioria dos casos, vamos nos contentar em obter apenas um resumo teórico da distribuição da v.a. Z_1Z_2 . Como resumir esta v.a. num único número? Já sabemos fazer isto com qualquer v.a.: tomamos o seu valor esperado. Isto é, vamos calcular

$$\rho = \text{Corr}(Y_1, Y_2) = \mathbb{E}(Z_1Z_2) = \mathbb{E}\left(\frac{Y_1 - \mu_1}{\sigma_1} \times \frac{Y_2 - \mu_2}{\sigma_2}\right)$$

Este resumo é o índice de correlação de Pearson.

13.3 Propriedades de ρ

- O índice ρ de correlação de Pearson está sempre entre -1 e 1. Esta é uma das razões para usar ρ como medida de associação entre Y_1 e Y_2 : ficamos com uma escala fixa em qualquer problema variando entre -1 e 1 sempre.
- Além disso, pela definição, a correlação não depende de uma ordem das variáveis:

$$\text{Corr}(Y_1, Y_2) = \mathbb{E}(Z_1Z_2) = \text{Corr}(Y_2, Y_1)$$

- Também temos que $\text{Corr}(Y, Y) = 1$: a correlação de uma v.a. consigo mesma é 1.
- Se Y_1 é uma v.a. independente da v.a. Y_2 então $\rho = 0$. Neste caso, uma amostra de valores do vetor (Y_1, Y_2) formará um gráfico de dispersão com forma indistinta, uma nuvem sem inclinação.
- Se $\rho \approx 1$ ou se $\rho \approx -1$ então Y_2 é aproximadamente uma função linear perfeita de Y_1 . Isto é, uma amostra de valores do vetor (Y_1, Y_2) formará uma gráfico de dispersão na forma aproximada de uma linha reta.

A Figura 13.6 mostra como a associação entre as variáveis muda quando ρ muda de valor.

13.4 Matriz de correlação

Correlação é uma medida de associação entre duas v.a.'s. E quando tivermos p v.a.'s simultaneamente, todas medidas no mesmo item? Suponha que tenhamos um vetor (Y_1, Y_2, \dots, Y_p) de v.a.'s. Podemos fazer uma matriz $p \times p$ de correlação. Na posição (i, j) teremos

$$\rho_{ij} = \text{Corr}(Y_i, Y_j) = \mathbb{E}\left(\frac{Y_i - \mu_i}{\sigma_i} \times \frac{Y_j - \mu_j}{\sigma_j}\right)$$

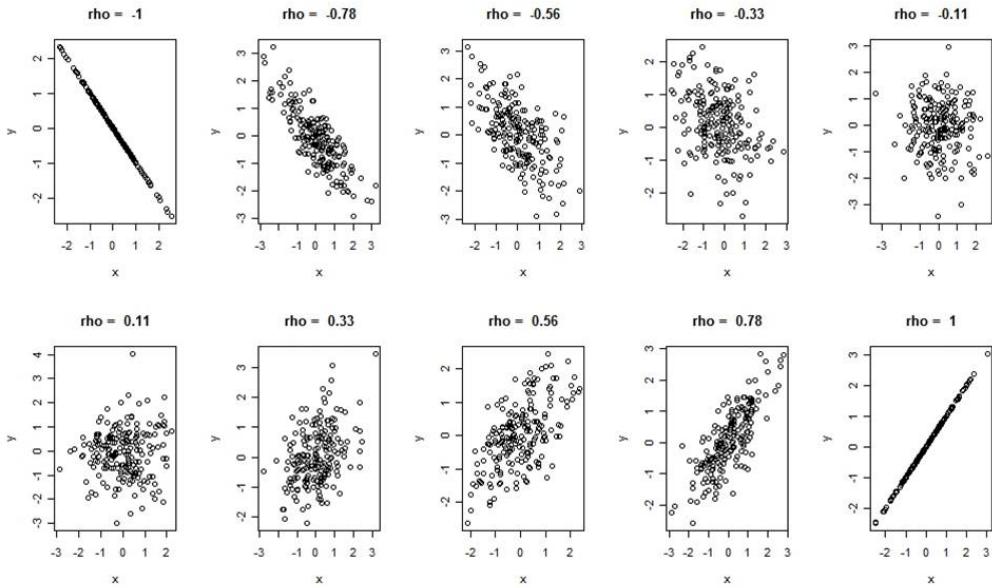


Figure 13.6: Mostrando como a associação entre as variáveis muda quando ρ muda de valor.

Como $\text{Corr}(Y_i, Y_j) = \text{Corr}(Y_j, Y_i)$, a matriz é simétrica. Como $\text{Corr}(Y_i, Y_i) = 1$, a diagonal principal é toda de 1's.

Por exemplo, vamos considerar um vetor aleatório $\mathbf{Y} = (Y_1, Y_2, \dots, Y_9)$ com 9 v.a.'s. As 9 variáveis aleatórias são escores obtidos em 9 testes de habilidade cognitiva, todos aplicados num mesmo indivíduo. As v.a.'s são as seguintes:

- 3 v.a.'s medindo habilidade verbal: Word Meaning, Sentence Completion, and Odd words;
- 3 v.a.'s medindo habilidade quantitativa: Mixed Arithmetic, Remainders, and Missing numbers;
- 3 v.a.'s medindo habilidade espacial: Gloves, Boots, and Hatchets.

Como poderia ser a matriz de correlação 9×9 entre estas v.a.'s? A Figura ?? mostra a matriz de correlação entre os pares formados a partir dessas 9 v.a.'s medidas num mesmo indivíduo em um teste de personalidade.

No lado esquerdo da Figura 13.8, temos a matriz de scatterplots de pares formados com essas

Variable	1	2	3	4	5	6	7	8	9
WrdMean	1								
SntComp	0.75	1							
OddWrds	0.78	0.72	1						
MxdArit	0.44	0.52	0.47	1					
Remndrs	0.45	0.53	0.48	0.82	1				
MissNum	0.51	0.58	0.54	0.82	0.74	1			
Gloves	0.21	0.23	0.28	0.33	0.37	0.35	1		
Boots	0.30	0.32	0.37	0.33	0.36	0.38	0.45	1	
Hatchts	0.31	0.30	0.37	0.31	0.36	0.38	0.52	0.67	1

WrdMean, word meaning; SntComp, sentence completion; OddWrds, odd words;
MxdArit, mixed arithmetic; Remndrs, remainders; MissNum, missing numbers,
Hatchts, hatchets.

Figure 13.7: Matriz de correlação entre pares formados a partir de 9 medidas feitas num mesmo indivíduo em um teste de personalidade. Matriz de scatterplots desses 9 variáveis.

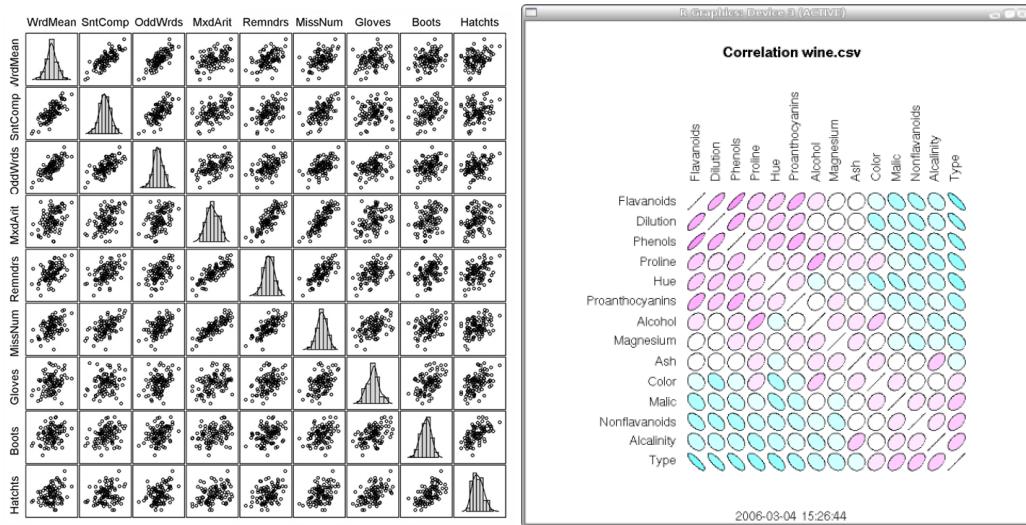


Figure 13.8: Esquerda: Matriz de scatterplots entre os pares formados a partir das 9 medidas de um teste de personalidade. Direita: Matriz estilizada dos scatterplots. A partir dos dados amostrais mostra-se o formato da nuvem de pontos de uma amostra de vinhos com 14 variáveis medidas em cada um dos vinhos. Gráfico feito com o pacote `rattle`.

9 variáveis onde a matriz de correlação pode ser visualizada. No lado direito, temos uma matriz estilizada dos scatterplots. Usa-se uma amostra de vinhos. Em cada desses vinhos, 14 variáveis são medidas. Este é um gráfico feito com o pacote `rattle`. Correlações positivas são representadas por elipses rosas e negativas pr elipses azuis. Quanto mais alongada a elipse, maior a correlação (em valor absoluto).

Outras visualizações são possíveis. No lado esquerdo da Figura 13.9, temos mais uma visualização da matriz de correlação usando o pacote `rattle`. No lado direito, temos uma visualização com o pacote `qgraph`. Ela é útil quando temos um número muito grande de variáveis inviabilizando o uso das matrizes de scatterplots. Neste gráfico, as v.a.'s são vértices e correlações são arestas. Correlações próximas de zero não são mostradas. Correlações positivas são mostradas em verde e correlações negativas são mostradas em vermelho.

Nem sempre as associações entre as variáveis são simples. Os gráficos podem ter formas bem diferentes temos mostrado até agora, sempre como uma nuvem de pontos em forma de elipse. Na Figura 13.10, os scatterplots mostram nuvens de pontos altamente concentradas nos canto inferior esquerdo sem uma clara associação entre as variáveis. Vimos anteriormente outros exemplos em que o coeficiente de correlação linear não captura bem a relação entre duas variáveis, se é que esta associação existe. Veja os gráficos das Figuras 2.16, 2.17, 2.18 e 2.19. Para medir associações mais complexas como estas precisamos usar outras medidas tais como a informação mútua e o coeficiente de informação maximal. Tratamos dessas outras medidas no capítulo ??.

Um tipo complexo de associação que pode ser decomposto em tipos mais simples é aquele em que modelamos os dados como um distribuição de mistura. A Figura 13.11 ilustra esta situação usando o dataset `iris`, uma coleção de 4 medições em três espécies de flor (*Iris setosa*, *Iris virginica* e *Iris versicolor*). Em cada flor individual foram medidas a o comprimento e a largura (em centímetros) das pétalas e também o comprimento e a largura das sépalas (uma espécie de pétala menor, mais rígida e localizada na base da flor). Este dataset é famoso pois, em 1936, Sir Ronald A. Fisher [11] desenvolveu um modelo discriminante linear (ver capítulo 15) para distinguir as espécies umas das outras. A Figura 13.11 mostra um matriz de scatterplot das 4 medições. A relação parece complicada por causa da presença de nuvens claramente distintas em cada gráfico.

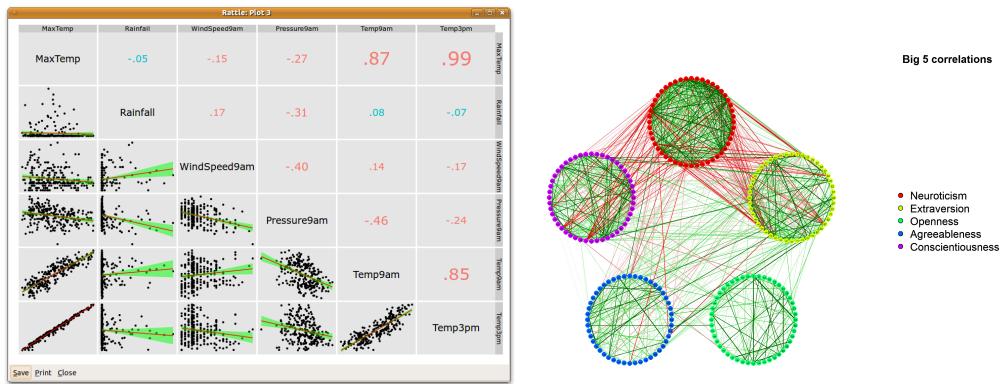


Figure 13.9: Esquerda: Mais uma visualização da matriz de correlação com R + rattle. Direita: Uma visualização com qgraph. V.a.'s são vértices e correlações são arestas. Uma aresta verde significa uma correlação positiva e enquanto vermelha significa uma correlação negativa. As arestas mais grossas e saturadas tem $|\rho|$ grande.

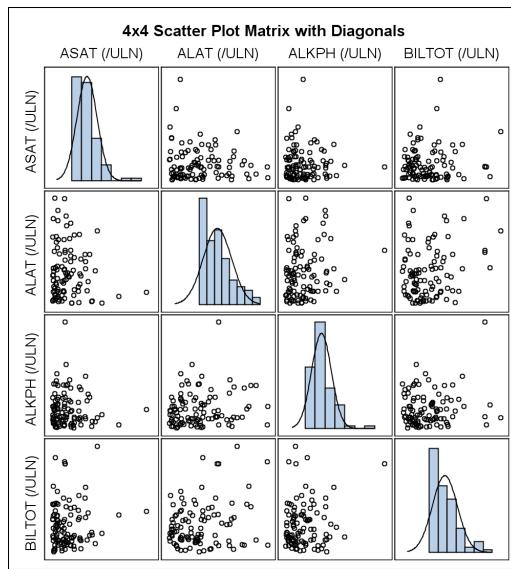


Figure 13.10: Scatterplot matrix of 4 lab variables to test liver functioning commonly used in clinical research

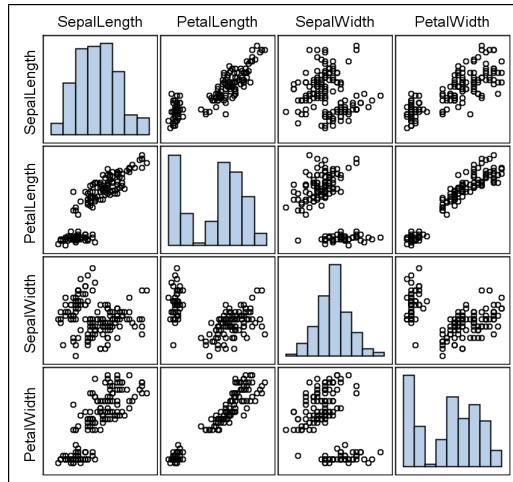


Figure 13.11: Scatterplot matrix de 4 variáveis medidas numa flor: comprimento de pétala, largura de pétala, comprimento de sétala, largura de sétala. Três espécies distintas misturadas. Relação entre as variáveis é diferente, ela depende da espécie.

Isto ocorre porque, em cada gráfico, temos a mistura de duas populações (ou espécies) de flores, cada espécie associada com uma nuvem. Se fixarmos o olhar apenas numa única espécie, caímos na situação tradicional que temos estudado até aqui. Assim, sob certas condições, uma situação que parece mais complexa pode se transformar na situação tradicional de nuvens em formas de elipses se considerarmos que o gráfico mostra uma mistura de diferentes grupos ou populações.

Neste exemplo, a mistura foi óbvia pois sabemos que existem três espécies distintas de flores nos dados e as nuvens estão claramente separadas. A dificuldade é quando esta informação adicional sobre a existência de diferentes populações não está disponível e quando as nuvens não são claramente separadas como na Figura 13.11. Nestas situações mais difíceis, temos de inferir sobre a existências dessas diferentes populações. Uma técnica para isto é o algoritmo EM, a ser estudado no capítulo 24. Lá, nós voltaremos a tratar os modelos de mistura e suas complicações.

13.5 Propriedades de ρ

Se $\rho = -1$ ou $\rho = +1$, podemos predizer o valor de Y_2 como função linear de Y_1 , sem erro, de forma perfeita. Isto é, se $\rho = \pm 1$, temos $Y_2 = \alpha + \beta Y_1$ onde α e β são duas constantes. Se $\rho = 0$ pode acontecer que Y_1 seja fortemente relacionada a Y_2 de uma forma não-linear. São casos raros na prática e não vamos nos ater a eles.

O parâmetro ρ é invariante por mudança linear de escala. Isto significa que a correlação entre Y_1 e Y_2 não muda se trocarmos Y_2 por $Y_2^* = a + bY_2$ onde a e b são constantes com $b > 0$. Por exemplo, suponha que Y_1 é o estoque de café num certo mês e Y_2 é o preço do café em reais no mesmo mês. Seja $\rho = \text{Corr}(Y_1, Y_2)$. Suponha que outra variável seja usada: o preço Y_3 do café, mas agora medido em dólares. Se a taxa de câmbio é fixa e igual a 2.3 teremos $Y_3 = 2.3Y_2$. Então,

$$\text{Corr}(Y_1, Y_3) = \text{Corr}(Y_1, 2.3Y_2) = \text{Corr}(Y_1, Y_2)$$

Do mesmo modo, se medirmos a temperatura em graus centígrados (Y_2) ou em graus Farenheit ($Y_3 = 32 + 1.8Y_2$), a correlação de temperatura com uma outra variável Y_1 é a mesma:

$$\text{Corr}(Y_1, Y_3) = \text{Corr}(Y_1, 32 + 1.8Y_2) = \text{Corr}(Y_1, Y_2)$$

13.6 Estimando ρ

ρ é um resumo teórico da distribuição conjunta de duas v.a.'s. Ele não depende de dados para ser obtido, é uma conta matemática. Relembre a definição:

$$\rho = \text{Corr}(Y_1, Y_2) = \mathbb{E} \left(\frac{Y_1 - \mu_1}{\sigma_1} \times \frac{Y_2 - \mu_2}{\sigma_2} \right)$$

Precisamos de $\mu_1 = \mathbb{E}(Y_1)$, $\sigma_1^2 = \mathbb{V}(Y_1)$, etc. Em seguida, precisamos calcular (usando teoria de probabilidade) o valor esperado do produto dos desvios. Para várias distribuições, esta conta matemática é inviável (não-analítica). No entanto, com dados, podemos estimar ρ facilmente.

Como $\frac{1}{n}(Y_1 + \dots + Y_n) = \bar{Y} \approx \mathbb{E}(Y)$ e como $S = \sqrt{\sum_i (Y_i - \bar{Y})^2 / n} \approx \sigma$ podemos aproximar

$$\rho = \mathbb{E} \left(\frac{Y_1 - \mu_1}{\sigma_1} \times \frac{Y_2 - \mu_2}{\sigma_2} \right) \approx \mathbb{E} \left(\frac{Y_1 - \bar{Y}_1}{S_1} \times \frac{Y_2 - \bar{Y}_2}{S_2} \right)$$

onde \bar{Y}_1 é a média aritmética dos n valores da variável 1, etc. Isto é, \bar{Y}_1 é média aritmética da coluna associada com a variável 1 na tabela de dados. Mas ainda precisaríamos calcular uma esperança matemática de uma função de várias v.a.'s, o que é inviável na maioria dos casos.

Solução: calcule o desvio observado em cada um dos n valores das duas variáveis. Para a variável 1 com os n valores y_{11}, \dots, y_{n1} da coluna 1 da tabela, calcule uma nova coluna d comprimento n formada por

$$z_{i1} = \frac{y_{i1} - \bar{y}_1}{s_1}$$

Faça o mesmo para a coluna 2, criando uma outra coluna de desvios padronizados empíricos:

$$z_{i2} = \frac{y_{i2} - \bar{y}_2}{s_2}$$

A seguir, multiplique as duas colunas de desvios padronizados e tire a sua média aritmética calculando

$$r = \frac{1}{n} \sum_{i=1}^n z_{i1} z_{i2} = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_{i1} - \bar{y}_1}{s_1} \right) \left(\frac{y_{i2} - \bar{y}_2}{s_2} \right)$$

Pela Lei dos Grandes Números (de novo), teremos $r = \frac{1}{n} \sum_{i=1}^n z_{i1} z_{i2} \approx \rho$ se n for grande.

13.7 Distância Estatística de Mahalanobis

A pressão sistólica mede a força do sangue nas artérias, à medida que o coração contrai para impulsionar o sangue através do corpo. Se ela foralta, ela pode levar à doenças do coração, angina e doenças vasculares nas pernas. Uma pressao sistólica saudável situa-se entre 120 e 140 mm Hg. Uma pressão sistólica maior que 140 mm Hg não é saudável A pressao diastólica é similar e deve ficar em torno de 80. Acima de 100, ela não é saudavel.

Uma amostra de 250 indivíduos (as instâncias, casos ou exemplos) foi selecionada e a pressão sistólica e diastólica foi medida em cada um deles. A Figura 13.12 mostra um gráfico com estes dois atributos para os 250 indivíduos. Vamos ver estes 250 pontos como realizações ou instanciações do vetor aleatório $\mathbf{Y} = (Y_1, Y_2)$. O vetor μ tem os valores esperados de cada variável de \mathbf{Y} :

$$\mathbb{E}(\mathbf{Y}) = \mathbb{E}(Y_1, Y_2) = (\mathbb{E}(Y_1), \mathbb{E}(Y_2)) = (\mu_1, \mu_2) = \mu .$$

As linhas vermelhas, vertical e horizontal, mostram as posições de $\mu_1 = 120$ e $\mu_2 = 80$.

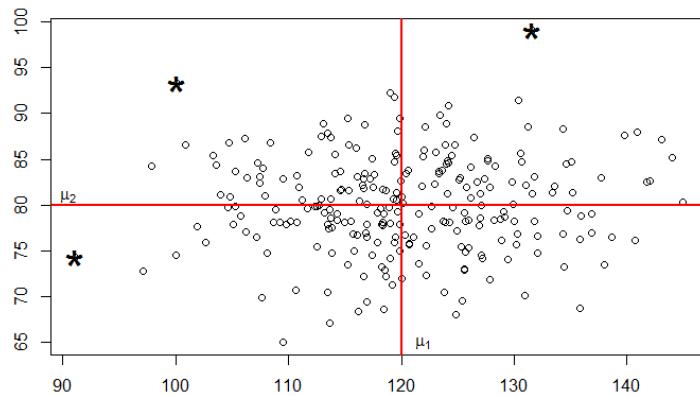


Figure 13.12: Amostra de $\mathbf{y}_i = (y_{i1}, y_{i2})$ com $i = 1, 2, \dots, 250$.

O centro $\mu = (\mu_1, \mu_2)$ é o perfil esperado ou típico para o vetor aleatório \mathbf{Y} . Quem está longe do perfil típico μ ? Quais as instâncias $\mathbf{y}_i = (y_{i1}, y_{i2})$ que são anômalas? Podemos usar uma medida baseada na distância euclidiana entre um ponto $\mathbf{y}_i = (y_{i1}, y_{i2})$ e o vetor esperado $\mu = (\mu_1, \mu_2) = (120, 80)$. Esta distância é dada por $d(\mathbf{y}, \mu) = \sqrt{(y_1 - \mu_1)^2 + (y_2 - \mu_2)^2} = \sqrt{(y_1 - 120)^2 + (y_2 - 80)^2}$. Fixe um círculo com centro em μ e com raio r . Todos os pontos neste círculo estão igualmente distantes do perfil médio, uma distância igual ao raio r . Isto é, pontos à igual distância de μ são aqueles localizados num círculo com centro em μ . Neste gráfico da Figura 13.12, est é uma forma razoável de medir o grau de afastamento de um ponto \mathbf{y} do perfil esperado μ . Por este critério, os três pontos destacados como estrelas na figura estão aproximadamente à mesma distância de μ e todos eles são razoavelmente anômalos. Eles são anomalias estatísticas porque não existem outros indivíduos com valores \mathbf{y} similares aos seus. Estatisticamente, eles estão igualmente distantes do perfil médio.

Mas, e se o segundo atributo (pressão diastólica) for como na Figura 13.13? O perfil esperado $\mu = (\mu_1, \mu_2)$ continua o mesmo de antes. Apenas o desvio-padrão do segundo atributo mudou, ficando bem mais reduzido agora. Nesta nova situação, quem está distante do centro? Quem é anômalo? Não parece mais razoável considerar todos os quatro pontos estrelados, em vermelho e localizados no círculo como igualmente anômalos ou igualmente distantes de μ . Enquanto os dois pontos vermelhos localizados perto do eixo vertical, (122, 96.9) e (118, 63.1), parecem estar estatisticamente bem distantes de μ , os outros dois pontos perto do eixo horizontal, (136.9, 81.8) e (103.2, 77.4), parecem pontos moderadamente razoáveis. Estes dois últimos pontos possuem a pressão sistólica um tanto distante de $\mu_1 = 120$ mas não extrema demais, e a pressão diastólica perfeitamente razoável. Pela distância euclidiana estes 4 pontos estão todos a uma igual distância de μ . A distância euclidiana não é mais uma medida de distância estatística razoável. Qualé a medida de distância que estamos usando implicitamente, sem nem mesmo perceber? Não é a distância euclidiana.

Como fazer alguns dos pontos vermelhos mais distantes que os outros? A ideia principal é que afastar-se do centro por poucas unidades na direção norte-sul nos leva para fora da nuvem de pontos e passamos a ter uma anomalia. Precisamos andar mais unidades na direção leste-oeste para sair fora da nuvem de pontos do que na direção norte-sul. Então x unidades na direção leste-oeste valem o mesmo que x/k na direção norte-sul, onde k é alguma constante maior que 1. Como achar este k ? Como equalizar as distâncias nas duas direções? Resposta: medindo distâncias em unidades de desvios-padrão.

Temos um desvio-padrão σ para cada eixo, um σ para cada atributo. O DP σ mede quanto, em média, um atributo aleatório desvia-se de seu valor esperado. Por exemplo, um DP $\sigma = 10$

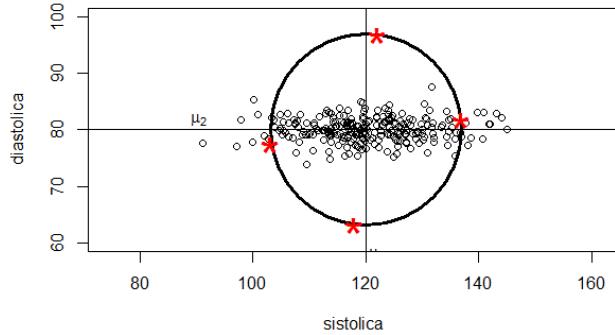


Figure 13.13: Amostra de $\mathbf{y}_i = (y_{i1}, y_{i2})$ com $i = 1, 2, \dots, 250$. A distância euclidiana continua sendo uma medida de distância estatística razoável?

significa que, em geral, observações desviam-se de 10 unidades em torno de seu valor esperado. Às vezes mais de 10 unidades, às vezes, menos de 10 unidades. Em média, temos um afastamento de 10 unidades. Est é significado prático do DP σ .

Sendo assim, qual o desvio padrão σ de cada variável na Figura 13.13? Temos o centro $\mathbb{E}(\mathbf{Y}) = \mu = (\mu_1, \mu_2) = (120, 80)$. Visualmente, não é difícil aceitar que $DP_1 = \sigma_1 = 10$ e $DP_2 = \sigma_2 = 2$. Voltando aos pontos estrelados em vermelhos na Figura 13.13, vemos que o ponto $(136.9, 81.8)$ afastou-se do centro μ praticamente apenas ao longo do eixo horizontal e este afastamento foi de aproximadamente $136.9 - 120 \approx 17$ unidades ou $1.7 \times \sigma_1$. Já o ponto $(122, 96.9)$ afastou-se do centro apenas ao longo do eixo vertical e este afastamento foi de aproximadamente $96.9 - 80 \approx 17$ unidades ou $8.5 \times \sigma_2$. Portanto, o segundo ponto está muito mais distante do centro μ em termos de DPs do que o primeiro ponto.

Como generalizar este raciocínio para pontos que afastam-se do centro não somente ao longo de um dos eixos? A ideia é medir distâncias em termos de desvios-padrão. Como $(\mu_1, \mu_2) = (120, 80)$ e $(\sigma_1, \sigma_2) = (10, 2)$, afastar-se $x\sigma_1$ unidades ao longo do eixo 1 é equivalente a afastar-se $x\sigma_2$ ao longo do eixo 2. Por exemplo, 20 unidades ao longo do eixo 1 (ou $2 \times \sigma_1$) é estatisticamente equivalente a 4 unidades ao (ou $2\sigma_2$) longo do eixo 2.

Vamos medir o desvio em cada eixo em unidades de seu desvio-padrão e calcular a distância com estes desvios padronizados. O desvio padronizado ao longo do eixo 1 é $z_1 = \frac{y_1 - \mu_1}{\sigma_1} = \frac{y_1 - 120}{10}$. O desvio padronizado ao longo do eixo 2 é $z_2 = \frac{y_2 - \mu_2}{\sigma_2} = \frac{y_2 - 80}{2}$. A distância é então a *distância euclidiana desde que os desvios sejam medidos em unidades de desvio-padrão*. Isto é, vamos fazer

$$\begin{aligned} d(\mathbf{y}, \mu) &= \sqrt{z_1^2 + z_2^2} \\ &= \sqrt{\left(\frac{y_1 - 120}{10}\right)^2 + \left(\frac{y_2 - 80}{2}\right)^2} \\ &= \sqrt{\left(\frac{y_1 - \mu_1}{\sigma_1}\right)^2 + \left(\frac{y_2 - \mu_2}{\sigma_2}\right)^2} \end{aligned}$$

Nesta nova métrica, quais os pontos (y_1, y_2) que estão a uma mesma distância do centro (μ_1, μ_2) ? Tome uma distância fixa (por exemplo, 1). Eles formam uma elipse centrada em (μ_1, μ_2)

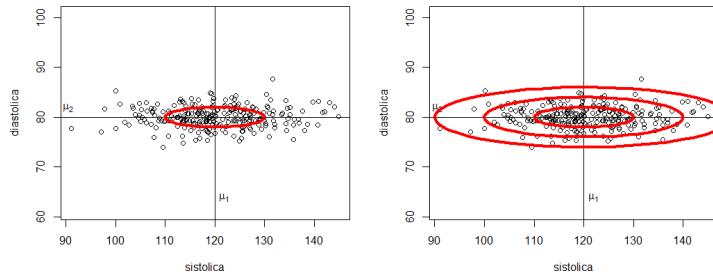


Figure 13.14: Esquerda: Lugar geométrico dos pontos a igual distância $c = 1$ do centro: uma elipse. Estes são pontos (y_1, y_2) que estão a uma distância c igual a 1 do centro (μ_1, μ_2) . Direita: variando $c = 1, 2, 3$ obtemos elipses concêntricas. Isto é, os pontos de cada elipse satisfazem $d(\mathbf{y}, \mu) = c$ para diferentes c 's.

e com eixos paralelos aos eixos coordenados. Isto é, os pontos $\mathbf{y} = (y_1, y_2)$ que satisfazem à equação

$$d(\mathbf{y}, \mu) = \sqrt{\left(\frac{y_1 - 120}{10}\right)^2 + \left(\frac{y_2 - 80}{2}\right)^2} = 1$$

formam uma elipse. Esta é a equação de uma elipse e o gráfico desses pontos está na Figura 13.14.

Os pontos que estão a uma distância $c > 0$ genérica do centro $\mu = (\mu_1, \mu_2)$ são aqueles que satisfazem a equação

$$d(\mathbf{y}, \mu) = \sqrt{\left(\frac{y_1 - 120}{10}\right)^2 + \left(\frac{y_2 - 80}{2}\right)^2} = c$$

e eles formam uma elipse. Os eixos desta elipse são paralelos aos eixos coordenados e têm comprimentos iguais a $c\sigma_1$ e $c\sigma_2$. O eixo maior da elipse está na direção da variável com maior DP. Quantas vezes maior é o eixo maior da elipse em relação ao seu eixo menor? Se σ_1 é o maior DP, então

$$\frac{\text{eixo maior}}{\text{eixo menor}} = \frac{c\sigma_1}{c\sigma_2} = \frac{\sigma_1}{\sigma_2}$$

Assim, se σ_1 é x vezes maior que σ_2 então o eixo maior da elipse associada com a distância c será também x vezes maior que o eixo menor. Esta razão é constante, não depende da distância c : variando c , teremos elipses concêntricas. Isto é ilustrado no lado direito da Figura 13.14.

Em matemática, preferimos trabalhar com a distância ao quadrado. A razão, não óbvia, é o Teorema de Pitágoras: soma dos quadrados dos catetos é igual ao quadrado da hipotenusa. A generalização deste teorema para espaços vetoriais de dimensão \mathbb{R}^n leva naturalmente a trabalhar com distâncias ao quadrado. Você verá um uso desta versão generalizada do teorema de Pitágoras no capítulo 18. Além de jogar fora a raiz quadrada, vamos escrever a fórmula de distância de uma maneira que parece mais complicada. Afinal, se podemos complicar, por quê simplificar?

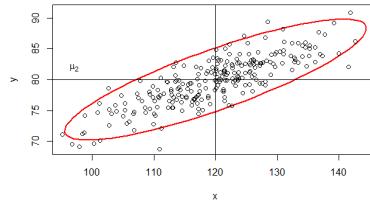


Figure 13.15: Amostra de vetor bivariado $\mathbf{Y} = (Y_1, Y_2)$ em que as variáveis aleatórias possuem correlação positiva.

$$\begin{aligned}
 d^2(\mathbf{y}, \boldsymbol{\mu}) &= \left(\frac{y_1 - \mu_1}{\sigma_1} \right)^2 + \left(\frac{y_2 - \mu_2}{\sigma_2} \right)^2 \\
 &= (y_1 - \mu_1, y_2 - \mu_2) \begin{bmatrix} 1/\sigma_1^2 & 0 \\ 0 & 1/\sigma_2^2 \end{bmatrix} \begin{pmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \end{pmatrix} \\
 &= (y_1 - \mu_1, y_2 - \mu_2) \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}^{-1} \begin{pmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \end{pmatrix} \\
 &= \begin{pmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \end{pmatrix}' \Sigma^{-1} \begin{pmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \end{pmatrix} \\
 &= (\mathbf{y} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{y} - \boldsymbol{\mu})
 \end{aligned}$$

Se $\mathbf{x} \in \mathbb{R}^n$ e \mathbf{A} é uma matriz simétrica $n \times n$, a expressão $\mathbf{x}' \mathbf{A} \mathbf{x}$ é chamada de *forma quadrática*. Como vimos acima, se $c > 0$ e

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

então $d^2(\mathbf{y}, \boldsymbol{\mu}) = (\mathbf{y} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{y} - \boldsymbol{\mu}) = c$ é a equação de uma elipse centrada no vetor $\boldsymbol{\mu} = (\mu_1, \mu_2)$. Quando, como acima, a matriz Σ é diagonal com elementos positivos (com as variâncias σ_i 's), então esta elipse tem eixos paralelos aos eixos e o tamanho de cada eixo é proporcional ao σ_i da variável associada.

Numa situação mais realista, as v.a.'s são associadas, não são independentes. Dizemos que elas são correlacionadas e isto significa que existe certa redundância de informação nas duas variáveis. O valor de uma variável numa certa instância ω dá informação sobre o valor da outra variável na mesma instância ω . Pode-se predizer (com algum erro) uma variável em função da outra.

O gráfico da Figura 13.15 mostra um caso típico onde a correlação entre as variáveis é positiva: quando uma variável está acima de sua média, a outra tende a estar também acima de sua média. Pelo mesmo raciocínio intuitivo que fizemos antes, os pontos na elipse da Figura 13.15 tendem a estar a igual distância do perfil esperado $\mathbb{E}(\mathbf{Y}) = \boldsymbol{\mu} = (\mu_1, \mu_2)$. Pontos estatisticamente equidistantes de $\boldsymbol{\mu}$ não estão mais numa elipse com eixos paralelos aos eixos do sistema de coordenadas. A elipse está inclinada seguindo a associação entre as variáveis.

A medida de distância que produz estas elipses de pontos equidistantes do centro é a mesma forma quadrática anterior:

$$d^2(\mathbf{y}, \boldsymbol{\mu}) = (\mathbf{y} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{y} - \boldsymbol{\mu})$$

Ela é a mesma expressão matricial de distância que usamos antes, mas a matriz Σ não é mais diagonal. Quem é Σ ? A matriz Σ é uma matriz 2×2 simétrica chamada de *matriz de covariância*

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

onde $\rho = \text{Corr}(Y_1, Y_2)$ é o índice de correlação de Pearson entre Y_1 e Y_2 . Temos sempre $-1 \leq \rho \leq 1$. Os elementos fora da diagonal, $\rho\sigma_1\sigma_2$, são chamados de covariância entre Y_1 e Y_2 . Costumamos escrever $\text{Cov}(Y_1, Y_2) = \rho\sigma_1\sigma_2 = \sigma_{12}$.

Esta matriz Σ determina a forma da elipse na Figura 13.15. Os eixos desta elipse estão na direção dos *autovetores* da matriz Σ . O tamanho de cada eixo é proporcional à raiz do *autovalor* correspondente. Vamos rever os conceitos de autovetor e autovalor na próxima seção. Eles são a base de algumas técnicas importantes de análise de dados tais como a análise de componentes principais (capítulo 14) e a análise discriminante linear (capítulo 15). Depois desta revisão, vamos retornar à distância estatística e à distribuição normal multivariada.

13.8 Autovetor e autovalor de Σ

Não vamos fazer uma discussão geral sobre autovetores e autovalores. Vamos focar apenas nos resultados relevantes para o que precisamos. Assim, como a matriz de covariância é simétrica, vamos considerar apenas os resultados para matrizes simétricas cujas entradas são números reais. Além disso, matrizes de covariância tipicamente são positivas definidas, um conceito que vamos definir abaixo. Assim, vamos revisar autovetores e autovalores apenas para matrizes simétricas e positivas definidas.

13.8.1 Formas quadráticas

Começamos definindo forma quadrática. Seja $\mathbf{v} = (v_1, \dots, v_p)'$ um vetor em \mathbb{R}^p . Um vetor será sempre um vetor-coluna neste livro. Seja Σ uma matriz $p \times p$.

Definition 13.8.1 — Forma quadrática. A forma quadrática associada com a matriz Σ é a expressão

$$\mathbf{v}' \Sigma \mathbf{v} = \sum_{ij} \Sigma_{ij} v_i v_j$$

Por exemplo, se $\mathbf{v} = (v_1, v_2)$ e Σ for uma matriz 2×2 , teremos

$$(v_1, v_2) \Sigma \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \Sigma_{11}v_1^2 + \Sigma_{12}v_1v_2 + \Sigma_{21}v_2v_1 + \Sigma_{22}v_2^2$$

A forma quadrática envolve as combinações lineares dos produtos de pares de variáveis (produto de duas variáveis distintas ou produto de uma variável por ela mesma).

Alguns exemplos específicos de forma quadrática com matrizes 2×2 :

$$(v_1, v_2) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = v_1^2 + v_2^2$$

$$(v_1, v_2) \begin{bmatrix} 9 & 0 \\ 0 & 4 \end{bmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = 9v_1^2 + 4v_2^2$$

$$(v_1, v_2) \begin{bmatrix} 9 & 3 \\ 3 & 4 \end{bmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = 9v_1^2 + 4v_2^2 + 3v_1v_2 + 3v_2v_1 = 9v_1^2 + 4v_2^2 + 6v_1v_2$$

Um caso tri-dimensional

$$(v_1, v_2, v_3) \begin{bmatrix} 9 & 3 & -2 \\ 3 & 10 & -6 \\ -2 & -6 & 6 \end{bmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = 9v_1^2 + 10v_2^2 + 6v_3^2 + 6v_1v_2 - 4v_3v_1 - 12v_3v_2$$

A matriz Σ de uma forma quadrática pode sempre ser considerada uma matriz simétrica. A razão é que, se Σ não for simétrica, podemos encontrar outra matriz Σ^* simétrica tal que, para todo vetor \mathbf{v} temos

$$\mathbf{v}' \Sigma \mathbf{v} = \mathbf{v}' \Sigma^* \mathbf{v}$$

Por exemplo, no caso bi-dimensional,

$$\begin{aligned} \mathbf{v}' \Sigma \mathbf{v} &= (v_1, v_2) \begin{bmatrix} 9 & 2 \\ 4 & 4 \end{bmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \\ &= 9v_1^2 + 4v_2^2 + 2v_1v_2 + 4v_2v_1 \\ &= 9v_1^2 + 4v_2^2 + 6v_1v_2 \\ &= (v_1, v_2) \begin{bmatrix} 9 & 3 \\ 3 & 4 \end{bmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \end{aligned}$$

Vamos deixar o caso geral como exercício. De agora em diante, a matriz Σ nas formas quadráticas é sempre uma matriz simétrica (como são as matrizes de covariância).

13.8.2 Matrizes positivas definidas

Queremos que uma medida de distância mais geral que a euclidiana. Estamos usando a distância estatística

$$d^2(\mathbf{y}, \mu) = (\mathbf{y} - \mu)' \Sigma^{-1} (\mathbf{y} - \mu).$$

Se usarmos a matriz $b\mathbf{s}\mathbf{A}$ no lugar de Σ^{-1} , a distância estatística entre um vetor \mathbf{y} e a origem (ou o vetor nulo) $b\mathbf{s}\mathbf{0}$ é uma forma quadrática:

$$d^2(\mathbf{y}, \mathbf{0}) = (\mathbf{y} - \mathbf{0})' \mathbf{A} (\mathbf{y} - \mathbf{0}) = \mathbf{y}' \mathbf{A} \mathbf{y}.$$

Para que esta medida de distância seja útil, queremos garantir que, para todo vetor \mathbf{y} que não seja o vetor nulo tenhamos a forma quadrática sempre maior que zero:

$$d^2(\mathbf{y}, \mathbf{0}) = \mathbf{y}' \mathbf{A} \mathbf{y} = \sum_{ij} \mathbf{A}_{ij} y_i y_j > 0$$

Uma matriz \mathbf{A} que atende a esta condição é chamada de matriz definida positiva.

Definition 13.8.2 — Matriz positiva definida. Uma matriz $p \times p$ real e simétrica \mathbf{A} é chamada *positiva definida* se sua correspondente forma quadrática for maior que zero para todo vetor $\mathbf{v} \neq \mathbf{0}$. Isto é, não sendo \mathbf{v} o vetor nulo, então

$$0 < \mathbf{v}' \mathbf{A} \mathbf{v} = [v_1 \ \dots \ v_n] \mathbf{A} \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = \sum_{i,j} \mathbf{A}_{i,j} v_i v_j$$

A matriz é chamada de *semi-positiva definida* se $\mathbf{v}' \mathbf{A} \mathbf{v} \geq 0$ para vetor $\mathbf{v} \neq \mathbf{0}$.

Assim, pedir que $\mathbf{v}' \mathbf{A} \mathbf{v} > 0$ para todo $\mathbf{v} \neq \mathbf{0}$ é o mesmo que pedir que todo ponto \mathbf{v} diferente da origem tenha uma distância *positiva* em relação à origem. Não faria sentido termos distâncias negativas. Também não queremos ter um vetor não nulo com uma distância zero até a origem. Assim, gostaríamos que a matriz Σ^{-1} na distância estatística atendesse a esta condição de ser uma matriz positiva definida.

Exemplos de matriz positiva definida:

$$(y_1, y_2) \begin{bmatrix} 9 & 0 \\ 0 & 4 \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = 9y_1^2 + 4y_2^2 > 0$$

e

$$(y_1, y_2) \begin{bmatrix} 9 & -3 \\ -3 & 4 \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = 9y_1^2 + 4y_2^2 + 3y_1y_2 + 3y_2y_1 = 9y_1^2 + 4y_2^2 - 6y_1y_2 > 0$$

Vamos ver agora alguns exemplos de matrizes que *não são* positivas definidas.

$$(y_1, y_2) \begin{bmatrix} 9 & 0 \\ 0 & -4 \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = 9y_1^2 - 4y_2^2$$

não é positiva definida pois é menor que zero se, por exemplo, tomarmos $(y_1, y_2) = (0, 1)$. Neste caso, teremos o resultado igual a $9 \times 0 - 4 \times 1^2 = -4$. Um outro exemplo:

$$(y_1, y_2) \begin{bmatrix} 1 & -2 \\ -2 & 1 \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = y_1^2 + y_2^2 - 4y_1y_2,$$

que é menor que zero se tomarmos $(y_1, y_2) = (1, 1)$. Neste caso, teremos $1^2 + 1^2 - 4 = -2$.

Não é óbvio como identificar se uma matriz é positiva definida, especialmente se sua dimensão p for alta. Ter todas as suas entradas positivas, por exemplo, não é um critério válido. Para confirmar isto, considere o caso abaixo

$$\mathbf{v}' \mathbf{A} \mathbf{v} = [1 \ -2] \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = -3$$

Como verificar, em geral, se uma matriz \mathbf{A} de dimensão $p \times p$ e simétrica é positiva definida? É claro que não podemos checar todos os infinitos \mathbf{v} . Temos um resultado que ajuda nesta tarefa, a ser visto na próxima seção.

Theorem 13.8.1 Seja Σ uma matriz de números reais quadrada $p \times p$ simétrica e positiva definida. Então existe a matriz inversa Σ^{-1} e ela também é simétrica e positiva definida.

Não veremos a prova deste teorema.

13.8.3 Autovetores e autovalores

Definition 13.8.3 — Autovetor e autovalor. Seja Σ uma matriz de números reais quadrada $p \times p$ simétrica. Um autovetor de Σ é um vetor $\mathbf{v} \in \mathbb{R}^p$ não-nulo tal que

$$\Sigma \cdot \mathbf{v} = \lambda \mathbf{v}$$

onde λ é uma constante real. A constante λ é chamada de autovalor associado ao autovetor \mathbf{v} .

Se \mathbf{v} é autovetor de Σ então qualquer múltiplo $c\mathbf{v}$ também é um autovetor se $c \neq 0$ pois, pelas propriedades usuais de multiplicação matricial,

$$\Sigma(c\mathbf{v}) = c(\Sigma \mathbf{v}) = c(\lambda \mathbf{v}) = \lambda(c\mathbf{v}).$$

Em geral, vamos querer falar da direção determinada por um autovetor. Assim, se \mathbf{v} é autovetor, vamos preferir trabalhar com \mathbf{v}/c onde $c = \|\mathbf{v}\| = \sqrt{\mathbf{v}' \mathbf{v}}$ é o comprimento (ou norma) euclidiano do vetor \mathbf{v} dado por $\sqrt{\mathbf{v}' \mathbf{v}} = \sqrt{v_1^2 + v_2^2 + \dots + v_p^2}$. Assim, vamos assumir daqui por diante que um autovetor terá comprimento unitário.

Em geral, transformar um vetor \mathbf{v} aplicando-lhe uma matriz Σ e gerando um novo vetor $\Sigma\mathbf{v}$ é uma operação cujo resultado final é difícil de ser antecipado. Não é fácil antecipar o que será o vetor $\Sigma\mathbf{v}$ a não ser que façamos todas as contas matriciais envolvidas. O autovetor é uma direção muito especial em \mathbb{R}^n . É uma direção na qual a operação $\Sigma\mathbf{v}$ é facilmente antecipada. Na direção \mathbf{v} , o efeito da matriz Σ é simplesmente espichar o vetor \mathbf{v} se $\lambda > 1$, encolher o vetor \mathbf{v} (se $0 < \lambda < 1$). Veremos que o caso $\lambda < 0$ não ocorre se Σ é simétrica e positiva definida, as únicas matrizes que nos interessam.

No caso geral, mesmo quando uma matriz Σ possui números reais em todas as suas entradas, é possível que um autovetor e o seu autovalor correspondente envolvam números complexos. Este não é o caso quando Σ for uma matriz simétrica.

Theorem 13.8.2 — Autovetores de Σ . Seja Σ uma matriz quadrada $p \times p$ simétrica. Então os autovalores de Σ são números reais.

Não veremos a prova deste teorema. O importante é podemos daqui por diante considerar os autovalores λ bem como seus autovetores como compostos apenas por números reais.

Theorem 13.8.3 — Autovetores de Σ e Σ^{-1} . Seja Σ uma matriz quadrada $p \times p$ simétrica e positiva definida com inversa Σ^{-1} . Então temos os seguintes resultados:

- Os autovalores de Σ são todos números reais maiores que zero.
- Os autovetores de Σ e Σ^{-1} são os mesmos.
- Se λ é autovalor de Σ , então $1/\lambda$ é autovalor de Σ^{-1} .

Prova: Suponha que $\mathbf{v} \neq \mathbf{0}$ é um autovetor de Σ com autovalor λ . Vamos provar que $\lambda > 0$. Como \mathbf{v} é autovetor, por definição, ele não é o vetor nulo e $\Sigma \mathbf{v} = \lambda \mathbf{v}$. Pré-multiplicando dos dois lados por \mathbf{v}' temos

$$\mathbf{v}' \Sigma \mathbf{v} = \lambda \mathbf{v}' \mathbf{v} = \lambda \|\mathbf{v}\|^2 \quad (13.2)$$

onde $\|\mathbf{v}\|^2$ é o comprimento (ou norma) do vetor \mathbf{v} . Como Σ é positiva definida, temos $\mathbf{v}' \Sigma \mathbf{v} > 0$ e portanto, por (13.2), $\lambda \|\mathbf{v}\|^2 > 0$. Como $\mathbf{v} \neq \mathbf{0}$ e a norma $\|\mathbf{v}\|^2$ de um vetor não-nulo é maior que zero, então temos de ter $\lambda > 0$.

Os autovetores de Σ e Σ^{-1} são os mesmos porque, se \mathbf{v} é autovetor de Σ , pré-multiplicando dos dois lados de $\Sigma \mathbf{v} = \lambda \mathbf{v}$ por Σ^{-1} , temos

$$\Sigma^{-1} \Sigma \mathbf{v} = \Sigma^{-1} (\lambda \mathbf{v})$$

ou seja

$$\mathbf{v} = \lambda \Sigma^{-1} \mathbf{v}$$

ou ainda, como $\lambda > 0$,

$$\frac{1}{\lambda} \mathbf{v} = \Sigma^{-1} \mathbf{v}$$

Assim, \mathbf{v} também é autovetor de Σ^{-1} com autovalor $1/\lambda$.

Na seção anterior não respondemos como verificar se uma matriz é positiva definida. Uma maneira um tanto computacionalmente intensiva de verificar isto é usar o teorema abaixo.

Theorem 13.8.4 Seja Σ uma matriz de dimensão $p \times p$ e simétrica. Ela é positiva definida se, e somente se, todos os seus autovalores forem positivos.

Não veremos a prova deste teorema. Sem querer descer a detalhes mais técnicos e concentrando apenas nas matrizes de covariância de vetores aleatórios \mathbf{Y} , podemos afirmar que a matriz de covariância Σ e sua inversa Σ^{-1} são, ambas, semi-positivas definidas e, na maioria das vezes, serão positivas-definidas.

13.8.4 Teorema Espectral

Seja Σ uma matriz $p \times p$ simétrica e positiva definida. Existem p autovalores associados com Σ . Estes p autovalores são números reais pois Σ é simétrica. Estes autovalores são positivos pois Σ é positiva definida. A cada autovalor corresponde um autovetor ou direção em \mathbb{R}^p . O que podemos falar desses autovetores? Os p autovetores são ortogonais entre si. Como existem p deles, tomando-os com comprimento 1, eles formam uma base ortonormal do espaço vetorial \mathbb{R}^p .

Colocando-os como p colunas de uma matriz P , teremos $\mathbf{P}'\mathbf{P} = \mathbf{I}$ pois eles são ortonormais. Seja \mathbf{D} uma matriz diagonal $p \times p$ com os autovalores (na mesma ordem que as colunas de \mathbf{P}). O teorema espectral afirma que $\Sigma = \mathbf{P} \mathbf{D} \mathbf{P}'$

O que isto significa: Σ age simplesmente como uma matriz diagonal \mathbf{D} (que é fácil de ser entendida) se trabalharmos no sistema de coordenadas dos autovetores (que são as colunas de \mathbf{P}). Dizemos que Σ é diagonalizável.

Essencialmente, no sistema de coordenadas dos autovetores, a matriz Σ funciona como uma matriz diagonal. \mathbf{x} no novo sistema de coordenadas dos autovetores é $\mathbf{x}^* = \mathbf{P}\mathbf{x}$. Se \mathbf{x}^* é o conjunto de coordenadas no sistema de autovetores, para voltar ao sistema original simplesmente multiplique pela inversa de P que é ... P' . Lembre-se que $\mathbf{P}'\mathbf{P} = \mathbf{I}$.

Resumindo:

- Pontos na ELIPSE tendem a estar a igual distância do perfil esperado $\mu = (\mu_1, \mu_2)$.
- A maneira correta de medir distância ao perfil esperado $\mu = (\mu_1, \mu_2)$ é pela forma quadrática

$$d^2(\mathbf{y}, \mu) = (\mathbf{y} - \mu)' \Sigma^{-1} (\mathbf{y} - \mu)$$

- A elipse é determinada pelos autovetores e autovalores de Σ^{-1} , a inversa da matriz de covariância das v.a.'s envolvidas.
- Os autovetores de Σ^{-1} e de Σ são os mesmos.
- Os autovalores de Σ^{-1} são os inversos $1/\lambda$ dos autovalores λ de Σ .

13.9 Densidade da normal multivariada

Finalmente podemos apresentar a densidade da distribuição normal multivariada. Seja $\mathbf{Y} = (Y_1, \dots, Y_p)$ um vetor aleatório de v.a.'s contínuas. Seja $\mu = (\mu_1, \dots, \mu_p)$ o vetor esperado $\mathbb{E}(\mathbf{Y}) = (\mathbb{E}(Y_1), \dots, \mathbb{E}(Y_p))$. Seja Σ a matriz $p \times p$ de covariância do vetor \mathbf{Y} .

Definition 13.9.1 — Densidade da normal multivariada. O vetor aleatório \mathbf{Y} de dimensão p segue uma distribuição normal (ou gaussiana) multivariada se sua densidade conjunta for da forma

$$f_{\mathbf{Y}}(\mathbf{y}) = \text{cte} \times \exp \left(-\frac{1}{2} d^2(\mathbf{y}, \mu) \right)$$

onde

$$d^2(\mathbf{y}, \mu) = (\mathbf{y} - \mu)' \Sigma^{-1} (\mathbf{y} - \mu)$$

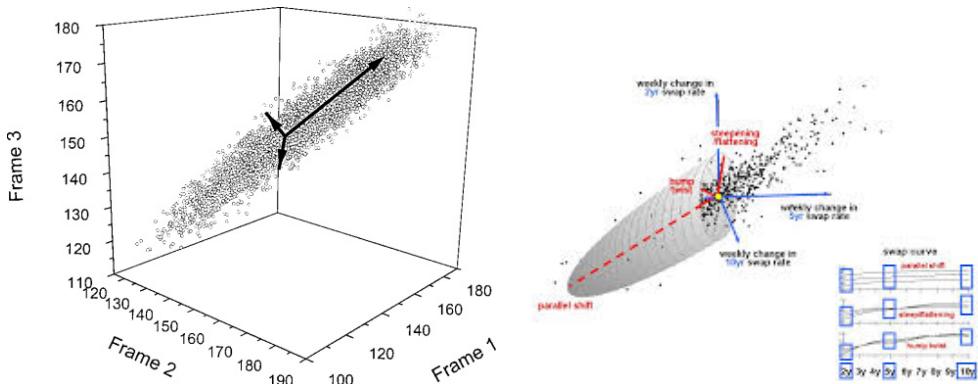


Figure 13.16: Nuvem de pontos de uma normal tri-dimensional $\mathbf{Y} \sim N_3(\mu, \Sigma)$.

$\| \cdot \|$ é a distância estatística (de Mahalanobis) entre \mathbf{y} e μ . Notação: $\mathbf{Y} \sim N_k(\mu, \Sigma)$

A densidade decresce com d^2 . As superfícies de nível da densidade são elipsóides concêntricos centrados em μ . Os eixos do elipsóide estão na direção dos autovetores de Σ e com comprimentos proporcionais à raiz do autovalor correspondente. A Figura fig:nuvemnormal3dim mostra a nuvem de uma normal tri-dimensional.

Caso 3-dim



14. Análise de Componente Principal e Fatorial

14.1 Introdução

Teste, teste,teste



15. Classificação: Análise Discriminante Linear

Classificação

- Outros nomes:
 - Análise discriminante de Fisher;
 - Classificação supervisionada.
- Vamos começar com a situação mais simples: duas classes
- Indivíduos são amostrados e medimos um conjunto de p variáveis em cada um deles.

$$X = (X_1, X_2, \dots, X_p)$$

- Com base nas medições em X , queremos inferir se $X \in \text{pop1}$ ou $X \in \text{pop2}$.
- Para construir uma regra de classificação de novos intens, usamos uma amostra com as classes conhecidas (rotuladas):

Item	Classe ou População	Variáveis $X_1 \ X_2 \ \dots \ X_p$
1	π_1	$X_{1,1} \ X_{1,2} \ X_{1,3} \ \dots \ X_{1,p}$
2	π_1	$X_{2,1} \ X_{2,2} \ X_{2,3} \ \dots \ X_{2,p}$
\vdots	\vdots	\vdots
m_1	π_1	$X_{m_1,1} \ X_{m_1,2} \ X_{m_1,3} \ \dots \ X_{m_1,p}$
1	π_2	$X_{m_1+1,1} \ X_{m_1+1,2} \ X_{m_1+1,3} \ \dots \ X_{m_1+1,p}$
2	π_2	$X_{m_1+2,1} \ X_{m_1+2,2} \ X_{m_1+2,3} \ \dots \ X_{m_1+2,p}$
\vdots	\vdots	\vdots
m_2	π_2	$X_{m_1+m_2,1} \ X_{m_1+m_2,2} \ X_{m_1+m_2,3} \ \dots \ X_{m_1+m_2,p}$
Novo Item	?	$X_1^* \ X_2^* \ X_3^* \ \dots \ X_p^*$

- ? → queremos inferir a classe
- $X_1^* \ X_2^* \ X_3^* \ \dots \ X_p^*$ → conhecido, observado.

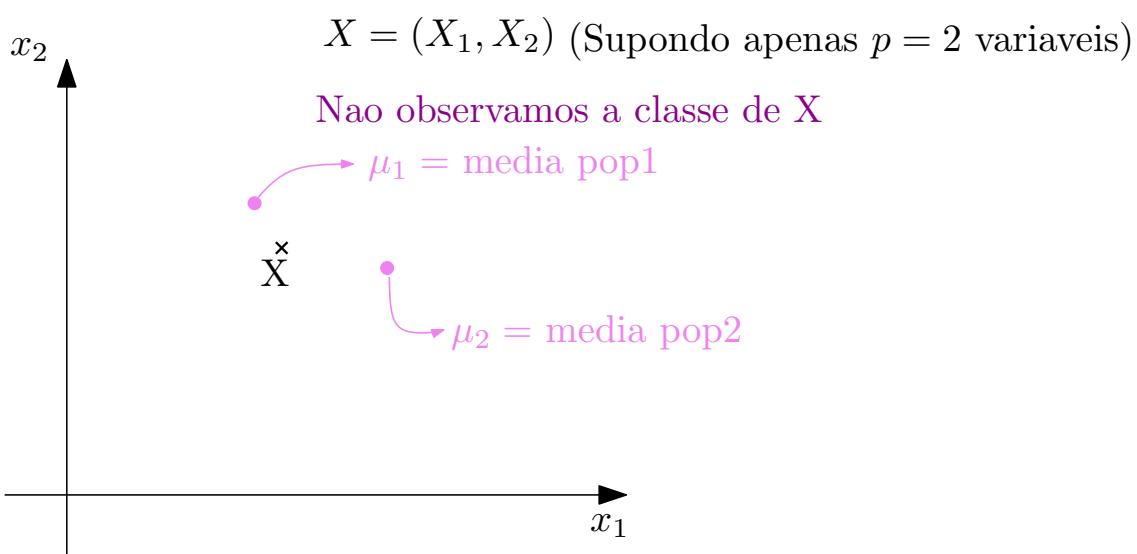
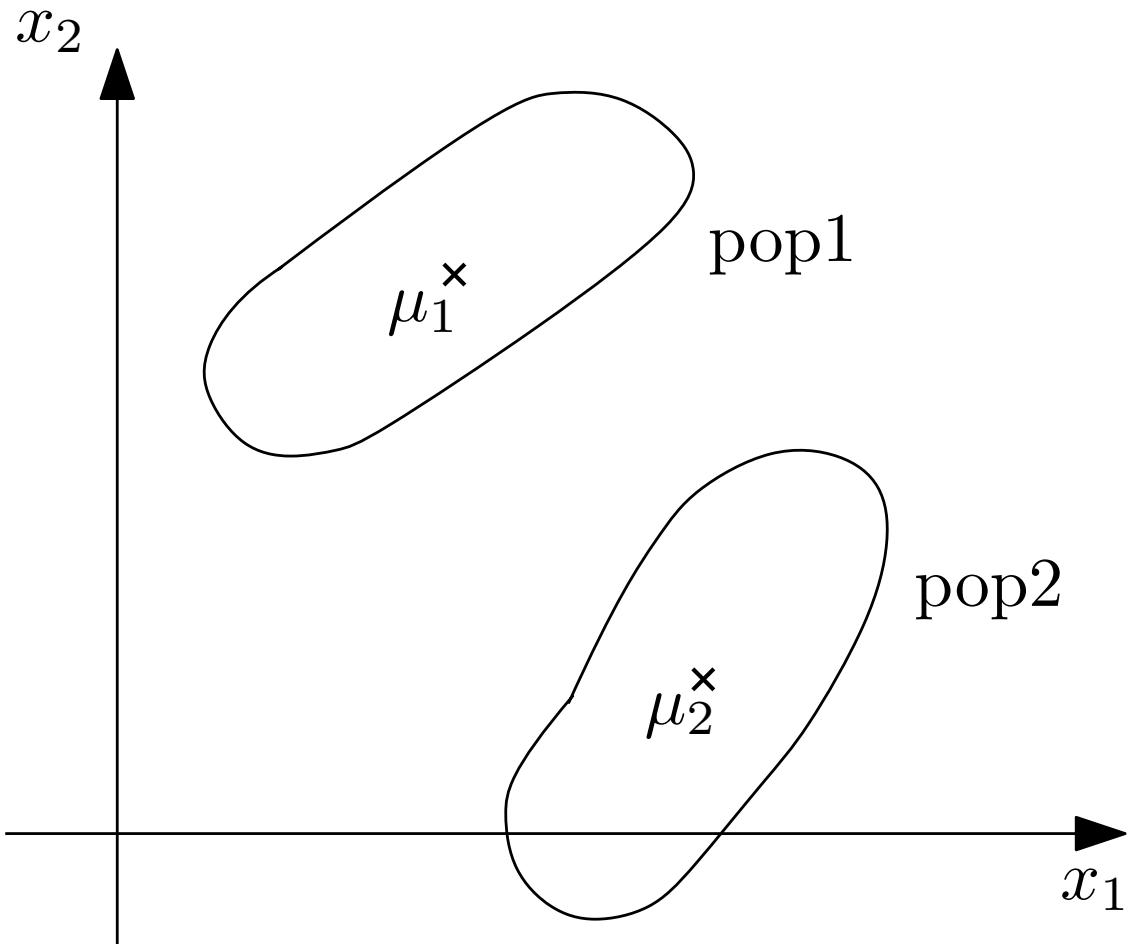
Exemplos

Exemplos

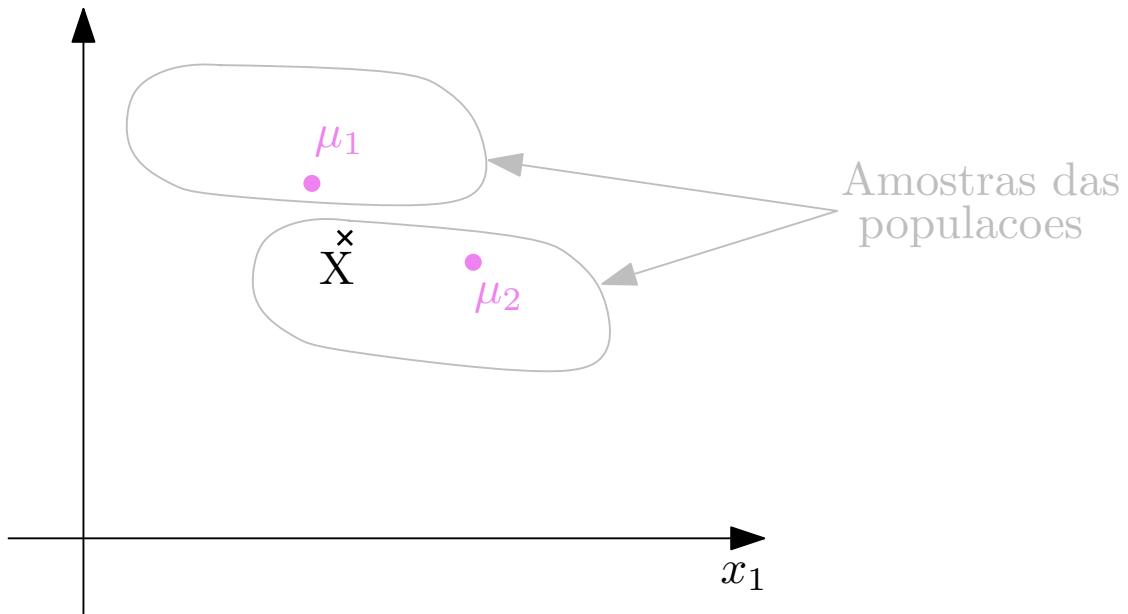
Por que precisamos predizer a classe de um item novo?

Populações π_1 e π_2	Variáveis $X_1 \dots X_p$
Risco de Crédito: Empresas tomadoras de crédito em um banco $\pi_1 \rightarrow$ créditos bons $\pi_2 \rightarrow$ créditos ruins	- % do empréstimo frente ao faturamento anual da empresa - tempo como cliente - nº de empréstimos anteriores pagos a tempo - saldo mensal
Crânios em um sítio arqueológico $\pi_1 \rightarrow$ homens $\pi_2 \rightarrow$ mulheres	- Circunferência - Largura - Altura
Pessoas com úlcera e normais	- Medidas de grau de ansiedade - Grau de perfeccionismo - Grau de sentimento de culpa - Grau de dependência
Textos de James Madison ou Alexander Hamilton	- Frequências de palavras distintas e comprimento das sentenças
Populações π_1 e π_2	Variáveis $X_1 \dots X_p$
Duas espécies de flor	- Comprimento da pétala - Largura da pétala - Comprimento da sépala - Largura da sépala
Usuários que clicam e não clicam em um anúncio	- Posição do anúncio na página - Tamanho do anúncio - Tem imagem? - Número de palavras
Alunos que evadem e que completam um curso online (curso a distância, curso noturno, curso de formação de professores)	- Nota do exame de entrada no curso - Medidas de motivação a partir de questionário na entrada - Renda familiar - Idade

- Classe pode ser conhecida apenas no futuro. Ex.: Risco de crédito: No momento em que o crédito é solicitado, não sabemos se o crédito do Indivíduo é bom ou ruim.
- Informação sobre a classe não é conhecida com certeza. Ex.: Crânios arqueológicos danificados.
- Obter a classe implica em destruir o item. Ex.: Queremos classificar um paciente chegando ao pronto socorro com lesão na cabeça como UTI ou não-UTI, com base em algumas medidas rápidas.
- Cada uma das populações possui uma distribuição conjunta para as p variáveis:
- $X = (X_1, X_2, \dots, X_p)$
- População 1
 $(X | \in pop1) \sim N_p(\mu_1, \Sigma_1)_{px1 \ pxp}$
- População 2
 $(X | \in pop2) \sim N_p(\mu_2, \Sigma_2)_{px1 \ pxp}$
- Com base na amostra rotulada (classe conhecida), podemos obter estimativas da distribuição da população 1 e da população 2.
- Novo item
- Olhar a distância do novo item a μ_1 e $\mu_2 \Rightarrow$ parece razoável alocar X à população 1, pois a distância a μ_1 é menor.
- Distância Euclidiana de X à μ_1 é menor que sua distância a μ_2 .
- No entanto, X parece pertencer à população 2!
- Precisamos levar em conta as correlações.
- Precisamos olhar a distância estatística ou a distância de Mahalanobis do novo item X a cada



Item	Classe ou População	Variáveis $X_1 X_2 \dots X_p$
1	π_1	$X_{1,1} X_{1,2} X_{1,3} \dots X_{1,p}$
2	π_1	$X_{2,1} X_{2,2} X_{2,3} \dots X_{2,p}$
\vdots	\vdots	\vdots
m_1	π_1	$X_{m_1,1} X_{m_1,2} X_{m_1,3} \dots X_{m_1,p}$
Médias		$\bar{x}_{11} \bar{x}_{12} \bar{x}_{13} \dots \rightarrow \bar{x}_1$ vetor = $\widehat{\mu}_1$
1	π_2	$X_{m_1+1,1} X_{m_1+1,2} X_{m_1+1,3} \dots X_{m_1+1,p}$
2	π_2	$X_{m_1+2,1} X_{m_1+2,2} X_{m_1+2,3} \dots X_{m_1+2,p}$
\vdots	\vdots	\vdots
m_2	π_2	$X_{m_1+m_2,1} X_{m_1+m_2,2} X_{m_1+m_2,3} \dots X_{m_1+m_2,p}$
Médias		$\bar{x}_{21} \bar{x}_{22} \bar{x}_{23} \dots \rightarrow \bar{x}_2$ vetor = $\widehat{\mu}_2$



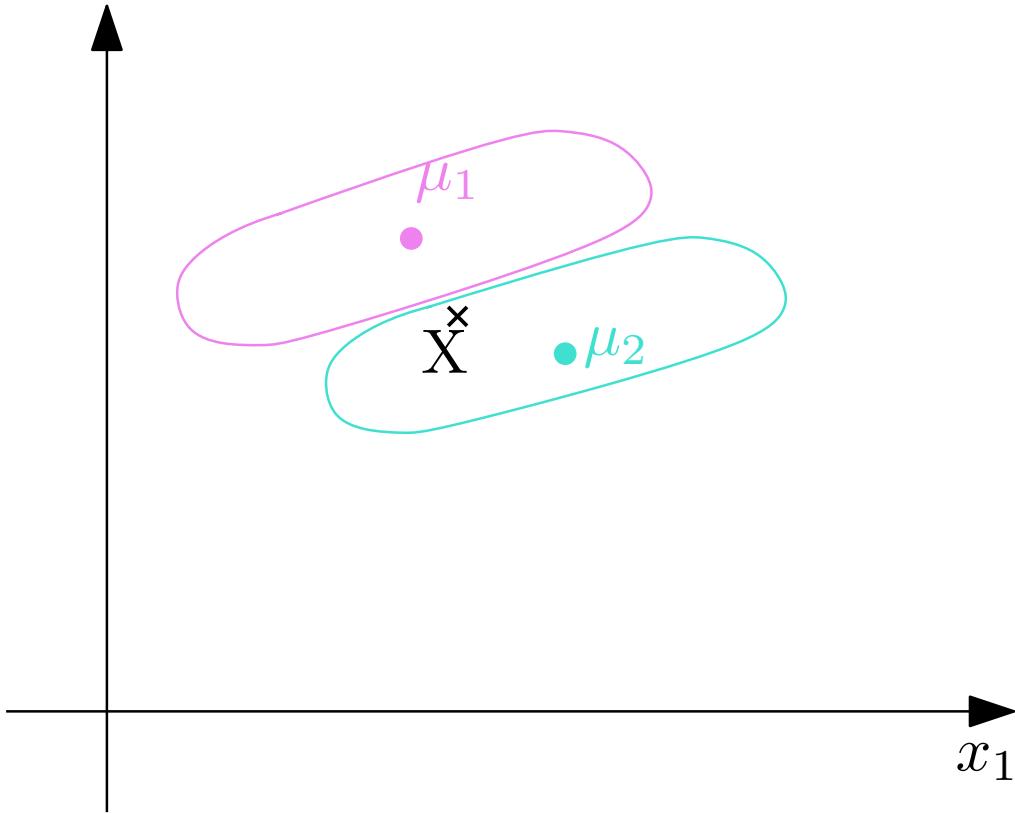
um dos centros μ_1 e μ_2 .

Mahalanobis

- $\mathbb{E}(X) = \mu$ = vetor com os valores esperados de cada uma das p variáveis
- $\mathbb{V}(X) = \Sigma$ = matriz de variâncias e covariâncias do vetor X
- $d_1^2 = d^2(X, \mu_1)$
 $= (X, \mu_1)^t \Sigma_1^{-1} (X, \mu_1)$
- $d_2^2 = d^2(X, \mu_2)$
 $= (X, \mu_2)^t \Sigma_2^{-1} (X, \mu_2)$

Regra de Classificação (Inicial)

$$d^2(X, \mu) = \boxed{(X - \mu)^t} \cdot \boxed{\Sigma^{-1}} \cdot \boxed{\frac{X - \mu}{\|X - \mu\|}}$$



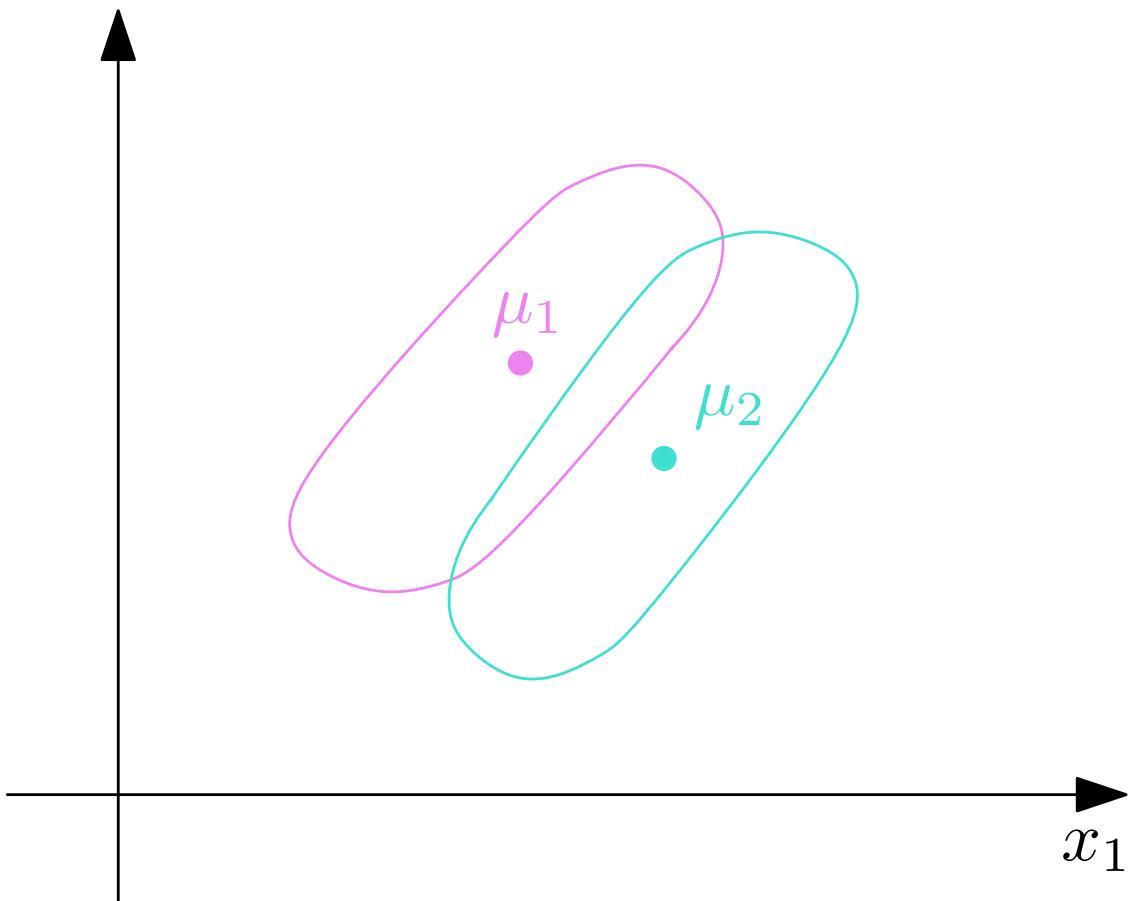
- Aloque X à população com menor $d^2(X, \mu)$, onde d é a distância de Mahalanobis.
- Isto é,
 - Se $d^2(X, \mu_1) < d^2(X, \mu_2) \Rightarrow$ aloque X à pop1;
 - Caso contrário, aloque X à pop2.
- Espaço \mathbb{R}^p é dividido em duas regiões:
 - $R_1 = \{x \in \mathbb{R}^p \mid d^2(X, \mu_1) < d^2(X, \mu_2)\}$
 - $R_2 = \mathbb{R}^p - R_1 =$ pontos que serão alocados à pop2.
- Quais são essas duas regiões?
- A seguir uma visão intuitiva.
- Resultado mais rigoroso vem a seguir.
- Obtenha o perfil médio das duas populações.
- Assuma que $\Sigma_1 = \Sigma_2$
 \Rightarrow autovetores de Σ_1 e Σ_2 são os mesmos.
 \Leftarrow as duas regiões.
- Outra maneira de ver a regra de classificação:

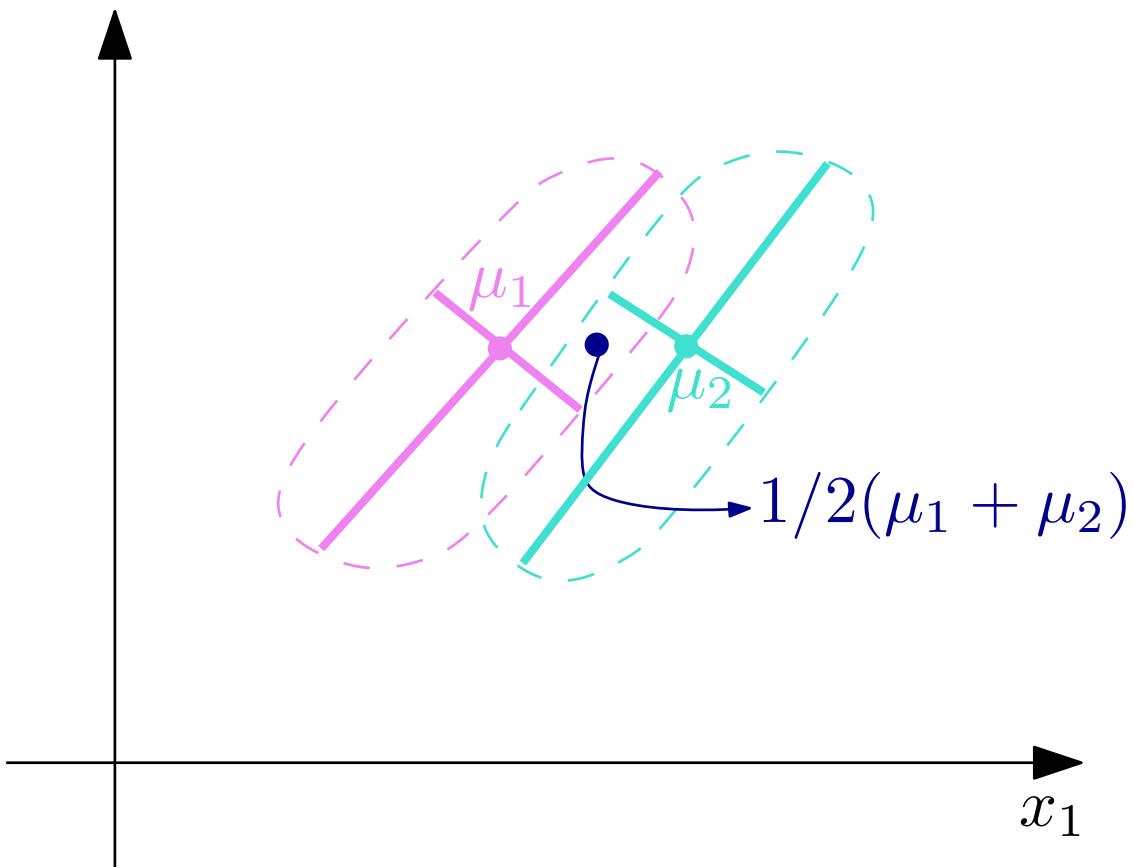
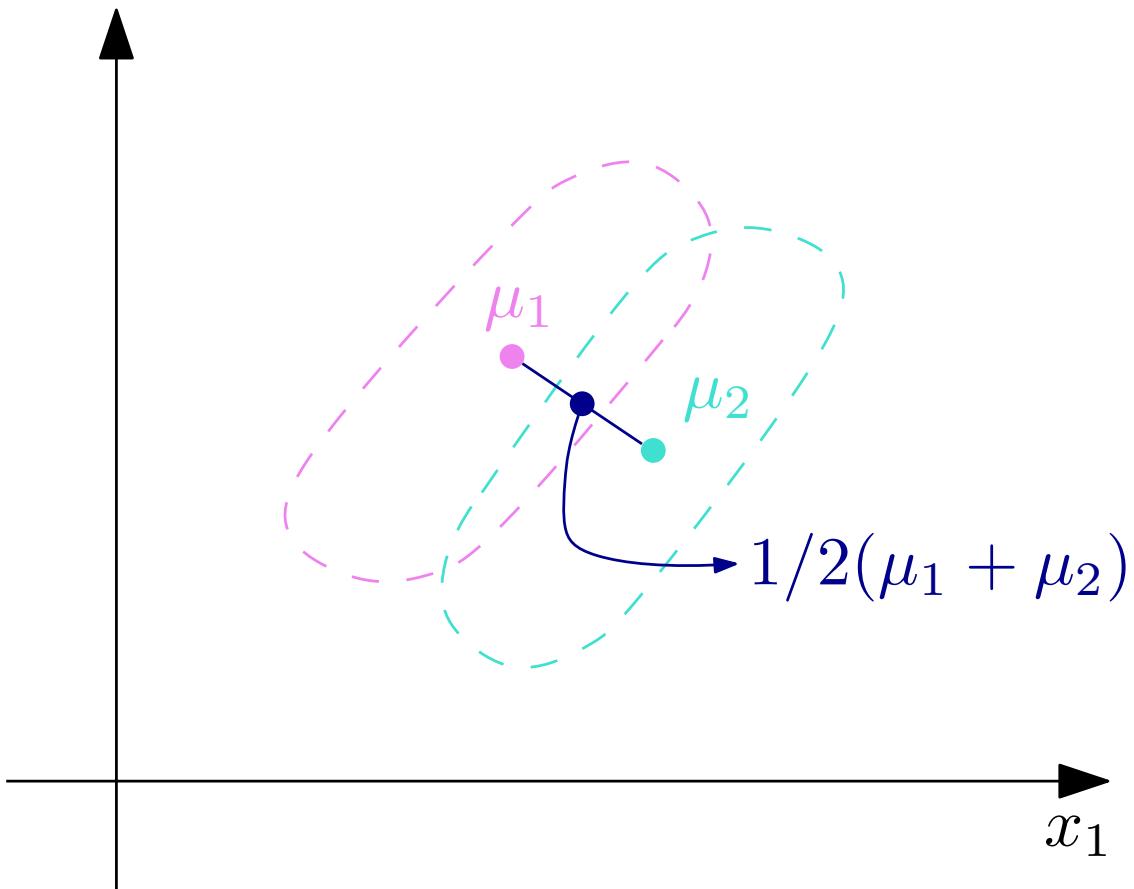
$$\begin{cases} \pi_1 = \text{pop1} \\ \pi_2 = \text{pop2} \end{cases}$$

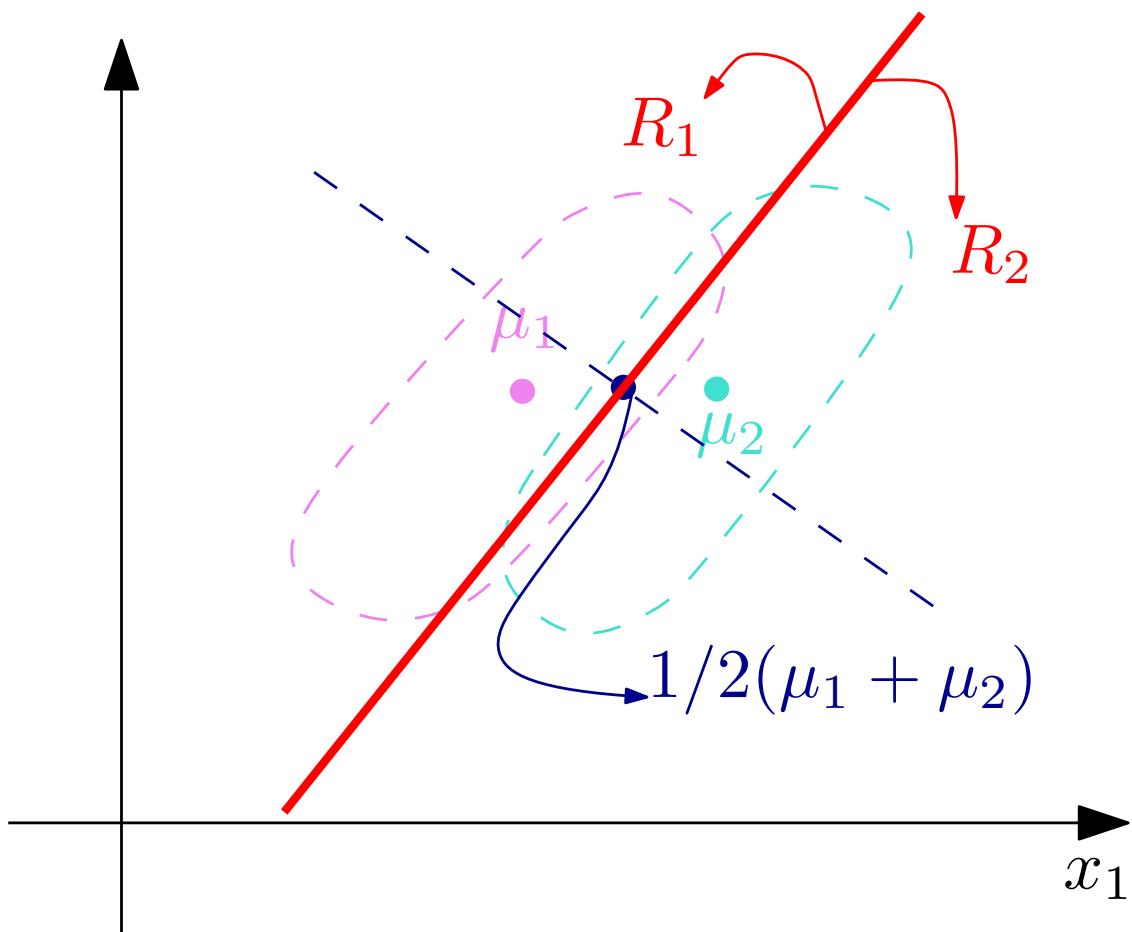
$$X \sim \begin{cases} N(\mu_1, \Sigma), \text{ se } \in \pi_1 \\ N(\mu_2, \Sigma), \text{ se } \in \pi_2 \end{cases} \Rightarrow \text{Assumindo } \Sigma_1 = \Sigma_2, \text{ por enquanto.}$$

- $f_1(x) =$ densidade do vetor X se $\in \pi_1$
- $f_2(x) =$ densidade do vetor X se $\in \pi_2$
- Tomando a razão das densidades no mesmo ponto X:

$$\frac{f_1(x)}{f_2(x)} = \frac{\varphi^e \exp\left(-\frac{1}{2}(X-\mu_1)^t \Sigma^{-1} (X-\mu_1)\right)}{\varphi^e \exp\left(-\frac{1}{2}(X-\mu_2)^t \Sigma^{-1} (X-\mu_2)\right)} = \exp\left(-1/2(d^2(X - \mu_1) - d^2(X - \mu_2))\right)$$







$$f_1(x) = \left[\frac{1}{(2\pi)^{p/2} |\Sigma^{1/2}|} \right] \cdot \exp\left(-\frac{1}{2} (X - \mu_1)^t \Sigma^{-1} (X - \mu_1)\right)$$

↓ Constante em X ↓ dist. de Mahalanobis

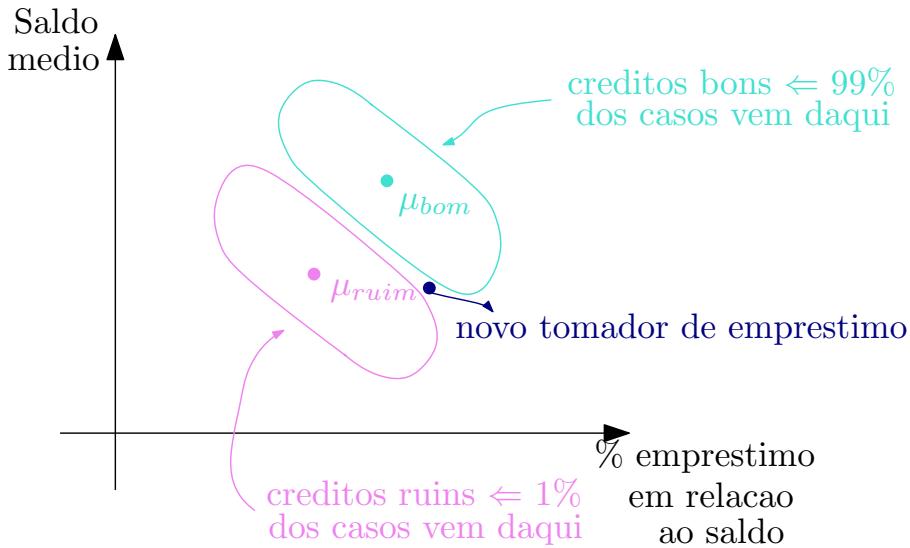
$$\frac{f_1(x)}{f_2(x)} > 1 \Leftrightarrow d^2(X - \mu_1) < d^2(X - \mu_2)$$

↓
condicao para alocar a pop1

- Veja que:
- Podemos definir a região de classificação à pop1 usando a razão de densidades.
- Este segundo modo de ver o problema é muito útil para a situação mais geral.
- Uma situação mais geral:
 - Custo de classificação errada pode variar
 - Uma das populações é muito mais comum do que a outra
 - A distribuição pode não ser gaussiana
- Exemplo de risco de crédito:
 - Cliente solicita empréstimo no banco
 - Queremos saber, no momento do empréstimo, se ele é um bom risco (pagará no prazo) ou um mau risco.
 - Nos baseamos em várias características (features) medidas no momento do empréstimo:
 - * idade, sexo, tempo como cliente, saldo médio,
 - * % do empréstimo em relação ao saldo,
 - * já pegou empréstimo antes?
- Custo de classificação em uma matriz:

		classificado em π_1	classificado em π_2
População Verdadeira	π_1 Bom crédito	custo = 0	$c(2 \in \pi_1)$
	π_2 Mau crédito	$c(1 \in \pi_2)$	custo = 0

- $c(2| \in \pi_1)$ = custo de classificar como mau crédito um bom pagador;
= custo de perder um bom cliente;
= perder o pequeno ganho a ser obtido por juros do empréstimo.
- $c(1| \in \pi_2)$ = custo de classificar como bom crédito um mau pagador;
= custo de perder todo \$\$ emprestado;
= perder todo o valor emprestado
- Em geral, nesse problema $c(1| \in \pi_2) >> c(2| \in \pi_1)$
- Isso tem impacto numa regra de classificação pois, se quisermos minimizar o custo esperado de uma decisão ruim, devemos levar em conta custos muito diferentes.
- Como fazer isso?
- O segundo ponto que queremos considerar é o tamanho desbalanceado das duas populações
- Maus pagadores são muito mais raros do que bons pagadores.
- E daí?
- Suponha que os custos de classificação incorreta sejam iguais: $c(1| \in \pi_2) = c(2| \in \pi_1)$
- Aonde você classificaria o novo item se $(d^2(X - \mu_1) = d^2(X - \mu_2))$???
- Se $(d^2(X - \mu_1) = d^2(X - \mu_2))$, estamos dizendo que não existe evidência nos dados X para saber se $\in \pi_1$ ou $\in \pi_2$
- Os dados X estão igualmente distantes das duas populações
- Resta a informação a priori que diz que, com alta probabilidade (0.99) um novo caso vem de π_1 .



- Então, se os custos são os mesmos, deveríamos alocar em π_1 .
- Como misturar custos e probabilidade a priori nos casos mais extremos?
- Um terceiro ponto a ser considerado:
- A distribuição dos dados pode não ser gaussiana.
- No caso gaussiano, como $f(x) = c^{te} \cdot \exp\left(\frac{-1}{2}(X - \mu_1)^t \Sigma^{-1}(X - \mu_1)\right)$, comparar distâncias de Mahalanobis é equivalente a comparar duas densidades de probabilidades (estritamente, apenas se $\Sigma_1 = \Sigma_2$).
- Vamos considerar o caso de $f(x)$ arbitrária geral.
Expected cost of minimization (ECM)
 - As duas densidades
 - R_1 e R_2 definidas por alguma regra de classificação (não necessariamente boa).

Veja que:

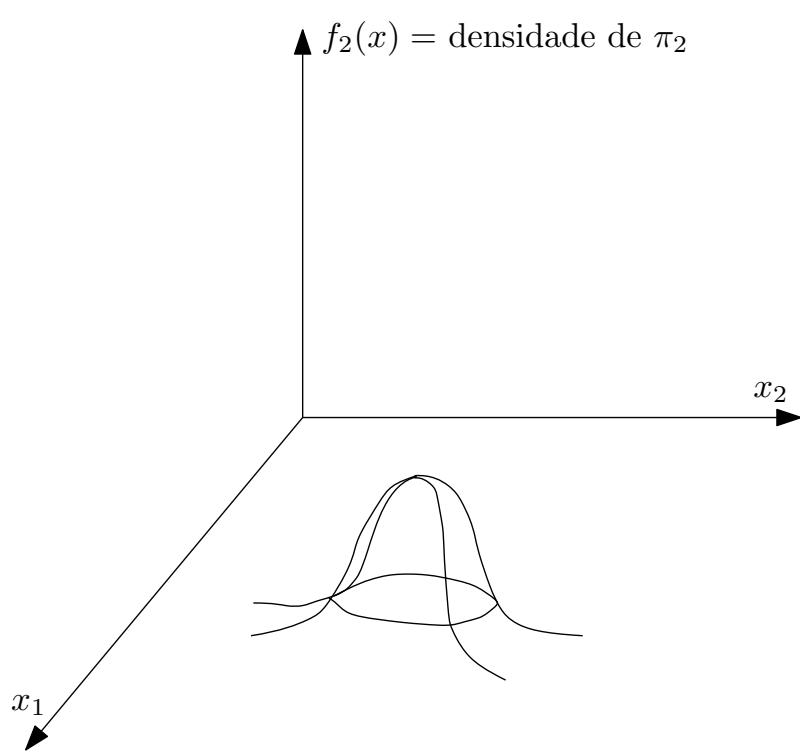
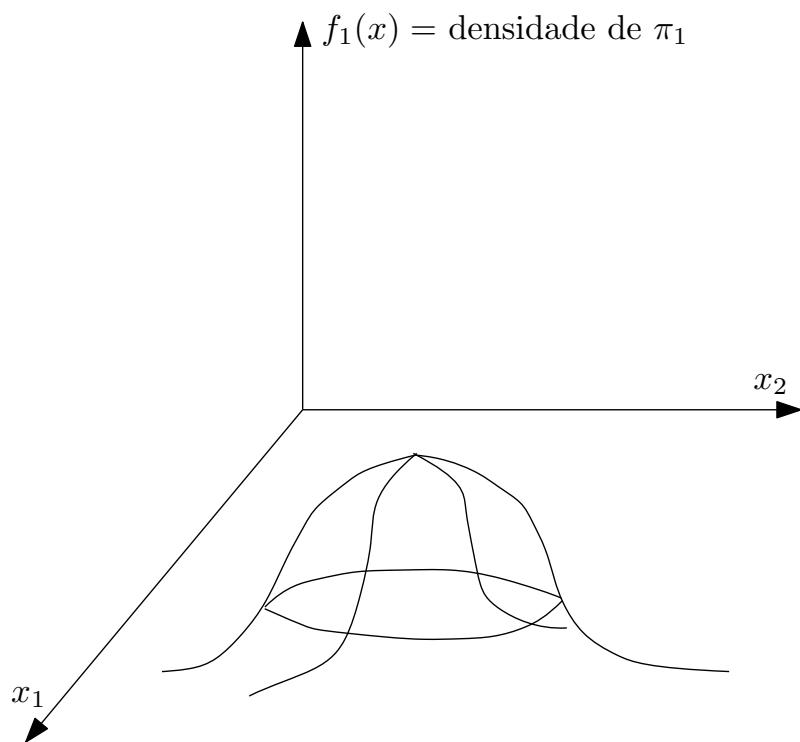
- (a) estabelecer uma partição de \mathbb{R}^p , $R_1 = \mathbb{R}^p - R_2$ implica em definir uma regra de classificação (aloque X a π_1 , se $X \in R_1$)
- (b) uma regra de classificação implica em uma partição do \mathbb{R}^p :

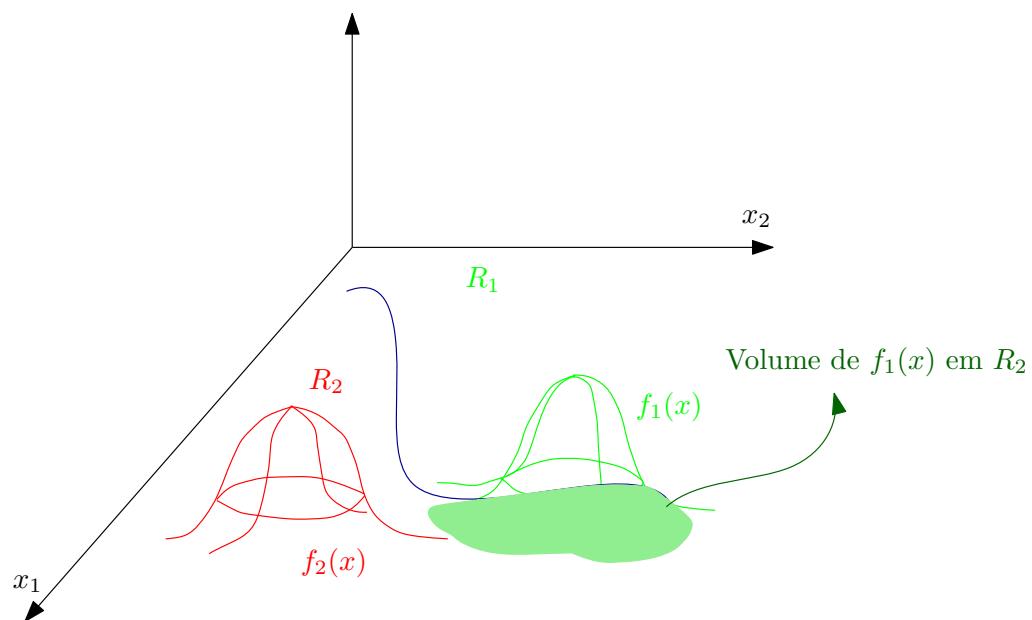
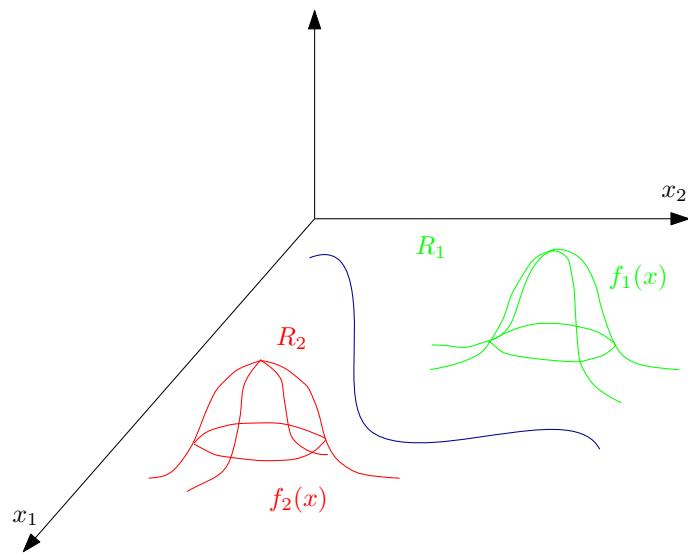
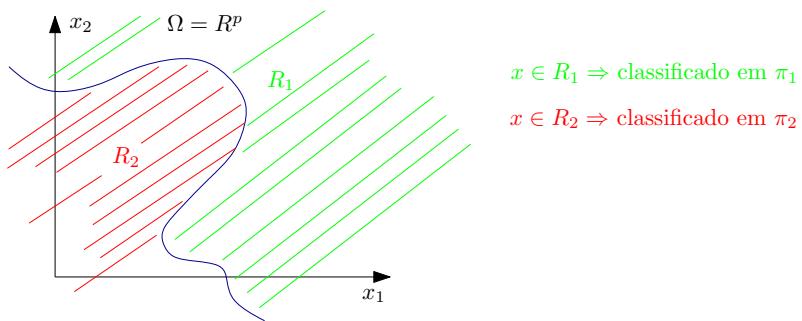
$$\begin{aligned} R_1 &= \{X \in \mathbb{R}^p \mid \text{se } x \in R_1, \text{ a regra aloca } X \text{ a } \pi_1\} \\ R_2 &= \mathbb{R}^p - R_1 \end{aligned}$$

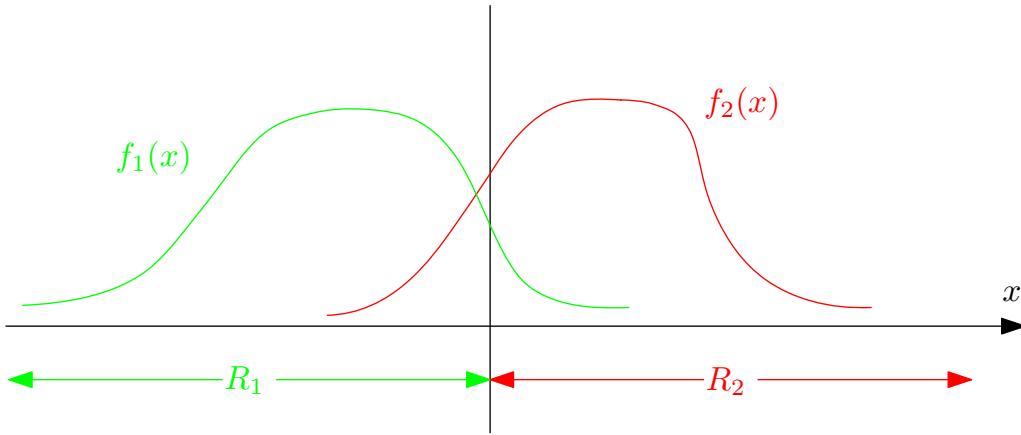
Assim, uma regra de classificação é equivalente a uma partição.

- Probabilidade condicional de classificar um objeto em π_2 quando, de fato, ele está em π_1 é:
 $\mathbb{P}(\text{Class. em } \pi_2 \mid \in \pi_1) = \mathbb{P}(X \in R_2 \mid \in \pi_1) = \int_{R_2} f_1(x) dx$
- Similarmente,
 $\mathbb{P}(\text{Class. em } \pi_1 \mid \in \pi_2) = \mathbb{P}(X \in R_1 \mid \in \pi_2) = \int_{R_1} f_2(x) dx$
- Vamos ver o caso em que $p = 1$ (uma única variável)
- As densidades $f_1(x)$ e $f_2(x)$, e R_1 e R_2 :
- A probabilidade de classificação incorreta:
- Veja que quando procuramos diminuir $\mathbb{P}(\text{Class. em } \pi_1 \mid \in \pi_2)$ estamos aumentando $\mathbb{P}(\text{Class. em } \pi_2 \mid \in \pi_1)$.
- Existe um trade-off entre essas probabilidades
- Como escolher uma boa partição de R_1 e R_2 do espaço \mathbb{R}^p ?
- Como os dois erros possuem custos diferentes, vamos minimizar o custo médio de má classificação.
- Temos também a probabilidade a priori de que os objetos venham da pop1 ou da pop2:

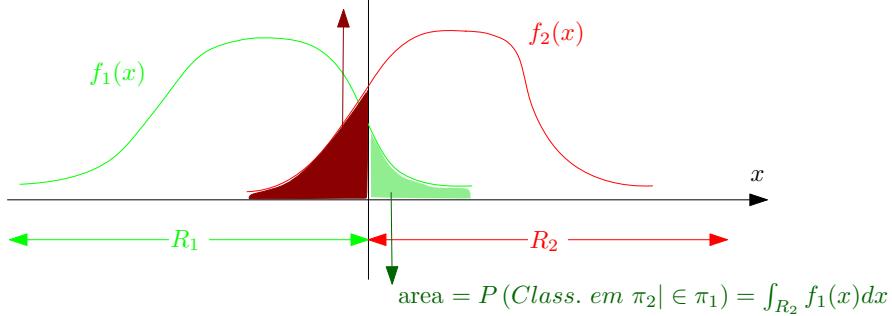
$$\begin{cases} p_1 = \mathbb{P}(\in \pi_1) \\ p_2 = \mathbb{P}(\in \pi_2) = 1 - \mathbb{P}(\in \pi_1) = 1 - p_1 \end{cases}$$
- Quadro geral:





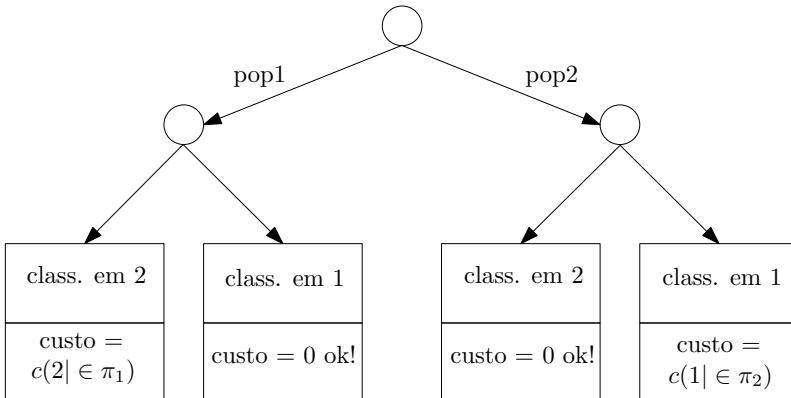


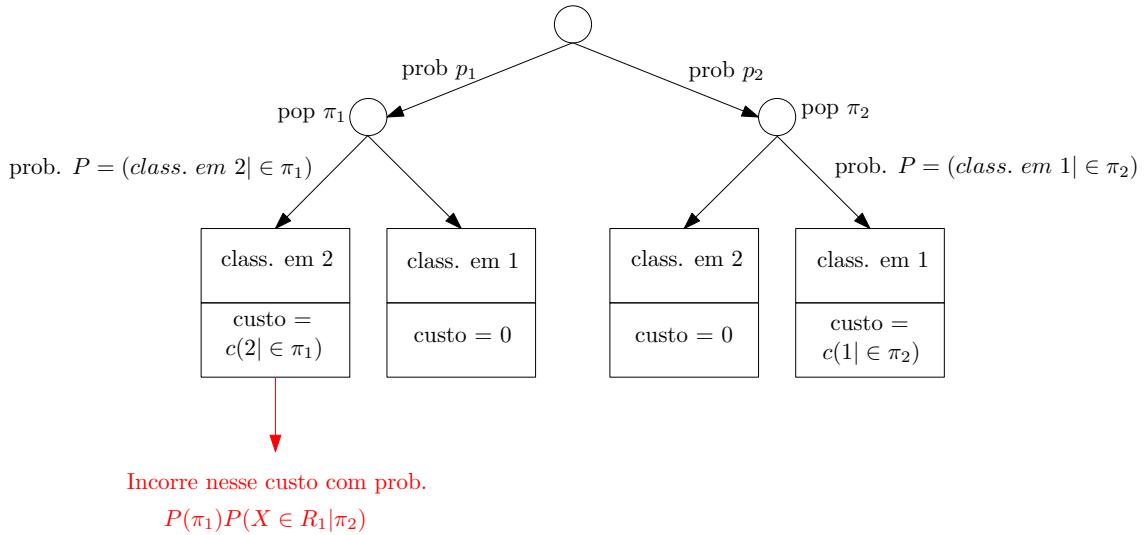
$$\text{area} = P(\text{Class. em } \pi_1 | \in \pi_2) = \int_{R_1}^{R_2} f_2(x) dx$$



$$\text{area} = P(\text{Class. em } \pi_2 | \in \pi_1) = \int_{R_2}^{R_1} f_1(x) dx$$

- Podemos ter $c(2| \in \pi_1) \neq c(1| \in \pi_2)$
- Às vezes ocorre o custo mais elevado
- Queremos uma regra que, em geral (ou, em média) leve a um custo pequeno \Rightarrow queremos um custo médio (ou esperado) pequeno.
- Custo esperado (ou custo médio):
 - $EMC = c(2| \in \pi_1)\mathbb{P}(X \in R_2 | \in \pi_1)\mathbb{P}(\pi_1) + c(1| \in \pi_2)\mathbb{P}(X \in R_1 | \in \pi_2)\mathbb{P}(\pi_2)$
 - $EMC \rightarrow$ custo esperado de má classificação.
- Queremos achar as regiões R_1 e R_2 que minimizam o ECM.
- Solução:
- Prova: Queremos R_1 e R_2 que minimizam ECM:
 - como $R_1 \cup R_2 = \mathbb{R}^p$ então $1 = \int_{\mathbb{R}^p} f_1(x) dx = \int_{R_1} f_1(x) dx + \int_{R_2} f_1(x) dx$.
 - Portanto:
 - Conseguimos escrever:





$$R_1 = \left\{ x \text{ tais que } \frac{f_1(x)}{f_2(x)} \geq \frac{c(1| \in \pi_2)}{c(2| \in \pi_1)} \cdot \frac{p_2}{p_1} \right.$$

↓ razao de densidades

↓ razao dos custos

↓ razao das probabilidades a priori

$$R_2 = R^p - R_1$$

$$EMC = c(2| \in \pi_1) P(X \in R_2| \in \pi_1) P(\pi_1) + c(1| \in \pi_2) P(X \in R_1| \in \pi_2) P(\pi_2)$$

$$\int_{R_2} f_1(x) dx$$

$$\int_{R_1} f_2(x) dx$$

$$EMC = c(2| \in \pi_1) \boxed{1 - \int_{R_1} f_1(x) dx} P(\pi_1) + c(1| \in \pi_2) \boxed{1 - \int_{R_1} f_2(x) dx} P(\pi_2)$$

$$= c(2| \in \pi_1) P(\pi_1) + \int_{R_1} (-c(2| \in \pi_1) P(\pi_1) f_1(x) + c(1| \in \pi_2) P(\pi_2) f_2(x)) dx$$

todos ≥ 0

$$ECM = c(2| \in \pi_1) P(\pi_1) + \int_{R_1} (c(1| \in \pi_2) P(\pi_2) f_2(x) - c(2| \in \pi_1) P(\pi_1) f_1(x)) dx$$

(*)

$$(*) = c(1| \in \pi_2)P(\pi_2)f_2(x) - c(2| \in \pi_1)P(\pi_1)f_1(x) < 0$$

$$\Leftrightarrow c(1| \in \pi_2)P(\pi_2)f_2(x) < c(2| \in \pi_1)P(\pi_1)f_1(x)$$

$$\Leftrightarrow \frac{c(1|\in\pi_2)P(\pi_2)}{c(2|\in\pi_1)P(\pi_1)} < \frac{f_1(x)}{f_2(x)}$$

$R_1 = \{x \text{ tais que } isto \text{ eh valido}\}$

R_2 = complementar de R_1

$$f_i(x) = \left[\frac{1}{(2\pi)^{p/2} |\Sigma^{1/2}|} \right] \cdot \exp\left(\frac{-1}{2} (X - \mu_1)^t \Sigma^{-1} (X - \mu_1)\right)$$

↓

↓

A mesma constante nas duas populações dist. de Mahalanobis

- Queremos escolher R_1 de forma que ECM seja mínimo.
 - O 1º termo não envolve R_1 .
 - O 2º termo: Escolher R_1 é escolher a região em que (*) será integrada.
 - Escolhemos $R_1 = \{x \text{ tais que } (*) < 0\} \rightarrow$ Isso minimiza ECM.
 - Mas:

Classificação ótima com duas gaussianas

- Caso 1: $\Sigma_1 = \Sigma_2 = \Sigma$
 - Portanto:
 - $R_1 \rightarrow$ Aloque x a π_1 se:
 - Exercício: Abrindo as expressões acima e manipulando encontramos:
 - Aloque x a π_1 se:
 - Caso contrário aloque a π_2 .
 - Caso 2: $\Sigma_1 \neq \Sigma_2$
 - A solução geral é válida:

■ **Example 15.1** Sensores baratos foram usados para identificar a espécie de inseto que passava por uma ambiente a partir da captura do som da batida das asas [chen2014flying]. A Figura 15.1 mostra na parte superior o histograma das frequências sonoras (em Hertz, Hz) das batidas das asas de 3000 insetos, sendo 1000 de cada uma de três espécies: *Cx. stigmatosoma*, *Aedes aegypti*, *Cx. tarsalis*. Na parte inferior, as densidades gaussianas que ajustam-se razoavelmente aos histogramas acima. Seja X a frequência em Hertz da batida das asas de um inseto aleatório e seja C a sua espécie, com três valores possíveis. Assim, as densidades gaussianas na parte inferior representam a densidade condicional ($X|C = c$) para $c = 1, 2, 3$.

$$\frac{f_1(x)}{f_2(x)} = \frac{\varphi^{te} \exp\left(\frac{-1}{2}(X - \mu_1)^t \Sigma^{-1} (X - \mu_1)\right)}{\varphi^{te} \exp\left(\frac{-1}{2}(X - \mu_2)^t \Sigma^{-1} (X - \mu_2)\right)} = \exp\left(\frac{-1}{2}(X - \mu_1)^t \Sigma^{-1} (X - \mu_1) + \frac{-1}{2}(X - \mu_2)^t \Sigma^{-1} (X - \mu_2)\right)$$

$$R_1 = \{x \text{ tais que } \exp\left(\frac{-1}{2}(X - \mu_1)^t \Sigma^{-1} (X - \mu_1) + \frac{-1}{2}(X - \mu_2)^t \Sigma^{-1} (X - \mu_2)\right) \geq \frac{c(1| \in \pi_2)}{c(2| \in \pi_1)} \frac{p_1}{p_2}$$

Tome log's

$$-(X - \mu_1)^t \Sigma^{-1} (X - \mu_1) + (X - \mu_2)^t \Sigma^{-1} (X - \mu_2) \geq 2 \log \left(\frac{c(1| \in \pi_2)}{c(2| \in \pi_1)} \frac{p_1}{p_2} \right)$$

Se o interesse fosse usar apenas uma única variável, X , para inferir a espécie C de um dado inseto, a tarefa seria simples se estivéssemos considerando apenas *Cx. tarsalis* versus uma das outras duas. Como as duas gaussianas são bem separadas no eixo x , quando uma das densidades tem algum valor significativo, a outra é praticamente zero. Já usar X para classificar *Cx. stigmatosoma* versus *Aedes aegypti* é uma tarefa bem mais complicada e sujeita a muitos erros de classificação.

Suponha que os custos de má-classificação sejam iguais e que as probabilidades a priori de qualquer duas espécies também sejam iguais. Intuitivamente, qual seria a região R_1 de classificação para a espécie $C = 1$ (*Cx. stigmatosoma*) versus $C = 3$ (*Cx. tarsalis*)? E para classificar $C = 1$ versus $C = 2$ (*Aedes aegypti*)? Sabemos que, para decidir entre $C = 1$ e $C = 2$, devemos alocar o inseto a $C = 1$ se $x \in R = \{x \text{ tais que } f_1(x) > f_2(x)\}$. Assim, basta comprara as gaussianas e veirficar qual delas tem maior altura num dado valor de x . A espécie desta densidade mais elevada é a regra de alocação ótima.

A Figura 15.2 apresenta as mesmas três espécies e as estimativas das densidades de probabilidade da v.a. T , o tempo em que um inseto é observado, em função de sua espécie. Cada curva tem uma área toda igual a 1. Assim, *Cx. stigmatosoma* é um inseto de comportamento predominantemente noturno. Já *Cx. tarsalis* prefere voar de manhã mas também durante a noite. Como seria a região R para diferenciar estas duas espécies? Seja R a região para classificar na espécie *Cx. tarsalis*.

■

15.0.1 Bayes' Theorem for random vectors is analogous to Bayes' Theorem for events.

Now Bayes' Theorem suppose X and Y are random vectors with a joint density $f(x, y)$. Substituting

$f(x, y) = f_{Y|X}(y|x)f(x)$ into (??), we have

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f_{Y|X}(x, y)}{f_Y(y)} \\ &= \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)}. \end{aligned} \tag{15.1}$$

This is a form of Bayes' Theorem.

Bayes' Theorem for Random Vectors If X and Y are continuous random vectors and $f_Y(y) > 0$ we have

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{\int f_{Y|X}(y|x)f_X(x)dx}.$$

If X and Y are discrete random vectors and $f_Y(y) > 0$ we have

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{\sum_x f_{Y|X}(y|x)f_X(x)}.$$

$$(\mu_1 - \mu_2)^t \Sigma^{-1} x \geq \ln \left(\frac{c(1| \in \pi_2)}{c(2| \in \pi_1)} \frac{p_1}{p_2} \right) + \frac{1}{2} (\mu_1 - \mu_2)^t \Sigma^{-1} (\mu_1 + \mu_2)$$

$$f_1(x) = \left[\frac{1}{(2\pi)^{p/2} |\Sigma_1^{1/2}|} \right] \cdot \exp\left(\frac{-1}{2} (X - \mu_1)^t \Sigma_1^{-1} (X - \mu_1)\right)$$

$$f_2(x) = \left[\frac{1}{(2\pi)^{p/2} |\Sigma_2^{1/2}|} \right] \cdot \exp\left(\frac{-1}{2} (X - \mu_1)^t \Sigma_2^{-1} (X - \mu_1)\right)$$

$\frac{f_1(x)}{f_2(x)} = \dots =$ a constante nao se cancela e
a expressao matricial mais
complicada

$$R_1 = \left\{ x; \frac{f_1(x)}{f_2(x)} > \frac{c(1| \in \pi_2)}{c(2| \in \pi_1)} \frac{p_1}{p_2} \right\}$$

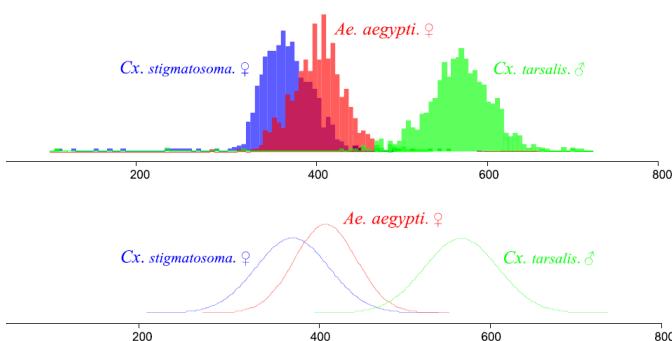


Figure 15.1: bla

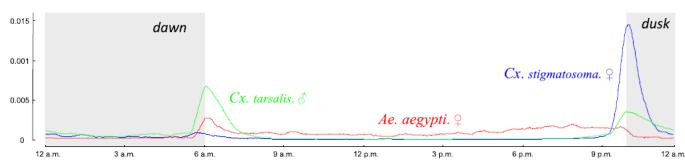


Figure 15.2: bla

Proof: These results follow by using the definition of marginal pdf in the denominator of (15.1). \square

The resemblance of this result to Bayes' Theorem for events may be seen by comparing the formula (??), identifying X with A and Y with B . The theorem also holds, as a special case, if X and Y are random variables.

15.0.2 Bayes classifiers are optimal.

Suppose X is a random variable (or random vector) that may follow one of two possible distributions having pdf $f(x)$ or $g(x)$. If x is observed, which distribution did it come from? This is the problem of *classification*. Typically, there is a random sample X_1, \dots, X_n and the problem is to classify (to one of the two distributions) each of the many observations. A *decision rule* or *classification rule* is a mapping that assigns to each possible x a classification (that is, a distribution). What is the best classification rule? A classification error is made if either $X \sim f(x)$ and the observation $X = x$ is classified as coming from $g(x)$ or $X \sim g(x)$ and the observation $X = x$ is classified as coming from $f(x)$.

Theorem Suppose X is drawn from a distribution having pdf $f(x)$, where $f(x) > 0$ for all x , with probability π and from a distribution having pdf $g(x)$, where $g(x) > 0$ for all x , with probability $1 - \pi$. Then the probability of committing a classification error is minimized if $X = x$ is classified as arising from $f(x)$ whenever $\pi f(x) > (1 - \pi)g(x)$, and is classified as arising from $g(x)$ when $(1 - \pi)g(x) \geq \pi f(x)$.

Before proving the theorem let us interpret it. Let C_1 refer to the case $X \sim f(x)$ and C_2 to $X \sim g(x)$, where we use the letter C to stand for “class,” so that the problem is to classify x as falling either in class C_1 or class C_2 . We take $P(C_1) = \pi$ and $P(C_2) = 1 - \pi$. The *Bayes classifier* assigns to each x the class having the maximal posterior probability, $P(C_1|X = x)$ versus $P(C_2|X = x)$, given by

$$P(C_1|X = x) = \frac{f(x)\pi}{f(x)\pi + g(x)(1 - \pi)}$$

and

$$P(C_2|X = x) = \frac{g(x)(1 - \pi)}{f(x)\pi + g(x)(1 - \pi)}.$$

The theorem says that *the Bayes classifier minimizes the probability of misclassification*.

Details:

Proof details:

We consider the case in which the two distributions are discrete and, for simplicity, we assume $\pi = \frac{1}{2}$. Let $R = \{x : f(x) \leq g(x)\}$. We want to show that the classification rule assigning $x \rightarrow g(x)$ whenever $x \in R$ has a smaller probability of error than the classification rule $x \rightarrow g(x)$ whenever $x \in A$ for any set A that is different than R . To do this we decompose R and its complement R^c as $R = (R \cap A) \cup (R \cap A^c)$ and $R^c = (R^c \cap A) \cup (R^c \cap A^c)$. We have

$$\sum_{x \in R} f(x) = \sum_{x \in R \cap A} f(x) + \sum_{x \in R \cap A^c} f(x) \tag{15.2}$$

and

$$\sum_{x \in R^c} g(x) = \sum_{x \in R^c \cap A} g(x) + \sum_{x \in R^c \cap A^c} g(x). \tag{15.3}$$

By the definition of R we have, for every $x \in R$, $f(x) \leq g(x)$ and, in particular, for every $x \in R \cap A^c$, $f(x) \leq g(x)$. Therefore, from (15.2) we have

$$\sum_{x \in R} f(x) \leq \sum_{x \in R \cap A} f(x) + \sum_{x \in R \cap A^c} g(x). \tag{15.4}$$

Similarly, from (15.3) we have

$$\sum_{x \in R^c} g(x) < \sum_{x \in R^c \cap A} f(x) + \sum_{x \in R^c \cap A^c} g(x). \quad (15.5)$$

Strict inequality holds in (15.5) because A is distinct from R ; if $A = R$ then $R^c \cap A = \emptyset$ and the first sums in both (15.3) and (15.5) become zero. Combining (15.4) and (15.5) we get

$$\sum_{x \in R} f(x) + \sum_{x \in R^c} g(x) < \sum_{x \in R \cap A} f(x) + \sum_{x \in R \cap A^c} g(x) + \sum_{x \in R^c \cap A} f(x) + \sum_{x \in R^c \cap A^c} g(x)$$

and the right-hand side reduces to $\sum_{x \in A} f(x) + \sum_{x \in A^c} g(x)$. In other words, we have

$$\sum_{x \in R} f(x) + \sum_{x \in R^c} g(x) < \sum_{x \in A} f(x) + \sum_{x \in A^c} g(x). \quad (15.6)$$

The left-hand side of (15.6) is the probability of an error using the rule $x \rightarrow g(x)$ whenever $x \in R$ while the right-hand side of (15.6) is the probability of an error using the rule $x \rightarrow g(x)$ whenever $x \in A$. Therefore the rule $x \rightarrow g(x)$ whenever $x \in R$ has the smallest probability of classification error.

The case for general π is essentially the same, and the continuous case replaces sums with integrals.

□

Corollary Suppose that with equal probabilities X is drawn either from a distribution having pdf $f(x)$, where $f(x) > 0$ for all x , or from a distribution having pdf $g(x)$, where $g(x) > 0$ for all x . Then the probability of committing a classification error is minimized if $X = x$ is classified to the distribution having the higher likelihood.

The theorem extends immediately to finitely many alternative classes (distributions). We state it in the language of Bayes classifiers.

Theorem Suppose X is drawn from a distribution having pdf $f_i(x)$, where $f_i(x) > 0$ for all x , with probability π_i , for $i = 1, \dots, m$, where $\pi_1 + \dots + \pi_m = 1$, and let C_i be the class $X \sim f_i(x)$. Then the probability of committing a classification error is minimized if $X = x$ is classified as arising from the distribution having pdf $f_k(x)$ for which H_k has the maximum posterior probability

$$P(C_k | X = x) = \frac{f_k(x)\pi_k}{\sum_{i=1}^m f_i(x)\pi_i} \quad (15.7)$$

among all the classes C_i .

Corollary Suppose n observations X_1, \dots, X_n are drawn, independently, from a distribution having pdf $f_i(x)$, where $f_i(x) > 0$ for all x , with probability π_i , for $i = 1, \dots, m$, where $\pi_1 + \dots + \pi_m = 1$, and let C_i be the class $X \sim f_i(x)$. Then the expected number of misclassifications is minimized if each $X_j = x_j$ is classified as arising from the distribution having pdf $f_k(x_j)$ for which C_k has the maximum posterior probability

$$P(C_k | X_j = x_j) = \frac{f_k(x_j)\pi_k}{\sum_{i=1}^m f_i(x_j)\pi_i}$$

among all the classes C_i .

Proof: Let $Y_i = 1$ if X_i is misclassified, and 0 otherwise. The theorem says that $P(Y_i = 1) = P(Y_1 = 1)$ is minimized by the Bayes classifier, which maximizes (15.7). The expected number of misclassifications is then $E(\sum_i Y_i)$ and we have

$$\begin{aligned} E(\sum_i Y_i) &= \sum_i E(Y_i) \\ &= \sum_i P(Y_i = 1) \\ &= nP(Y_1 = 1). \end{aligned}$$

$$\left\{ \begin{array}{l} Y' s \text{ de pop1} \\ Y' s \text{ de pop2} \end{array} \right. \rightarrow \text{o mais separado possivel}$$

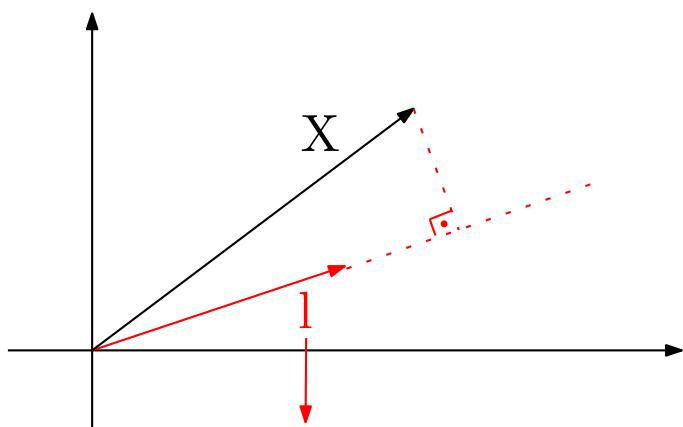
Therefore, the expected number of misclassifications is minimized by the Bayes classifier. \square

Example ?? (continued from page ??) We described previously the use of Bayes' theorem in decoding saccade direction from the activity of neurons in the supplementary eye field. This may be considered an application of Bayesian classification. Previously we took the events A_1, A_2, A_3 , and A_4 to be the saccade directions up, right, down, left. To put this in the notation of the corollary above, we may write $C_i : A_i$, for $i = 1, 2, 3, 4$. The observations X_i are then random vectors of length 55 representing spike counts among 55 neurons. The unpublished work, previously cited, by Kass and Ventura, took the neural spike counts to be independent (they were, in fact, recorded separately) and Poisson distributed. Initial data (usually called *training data*) were used to estimate the 55 Poisson parameters. This provided the pdfs $f_k(x)$ that appear in the corollary above. The cited prediction accuracy of 95% from Bayesian classification ("Bayesian decoding") Bayesian decoding was achieved on separate data (*test data*). \square

The fundamental result given in the theorem extends to the case in which different penalties result from the various incorrect classifications. This more general situation is treated by *decision theory*. Suppose $d(x)$ is a mapping that assigns to each x a class (a distribution). Such a mapping is called a *decision rule*. Let us denote the possible values of any such rule by a (for *action*), so that a may equal any of the integers $1, 2, \dots, m$. The penalties associated with various classifications, or decisions, may be specified by a *loss function* $L(a, i)$, where each $L(a, i)$ is the non-negative number representing the penalty for deciding to classify x as arising from $f_a(x)$ when actually it arose from $f_i(x)$. We then may consider the expected loss $E(L(d(X), i))$, i.e., the average behavior of the decision rule, which is also known as the *risk* of the decision rule. The decision rule with the smallest risk is called the *optimal decision rule*. Assuming that the distribution with pdf $f_i(x)$ has probability π_i , for $i = 1, \dots, m$, this optimal rule turns out to be the *Bayes rule*, Bayes rule which is found by minimizing the expected loss computed from the posterior distribution. The theorem above then becomes the special case in which $L(a, i) = 0$ if $a = i$ and $L(a, i) = 1$ otherwise, for then the risk is simply the probability of misclassification.

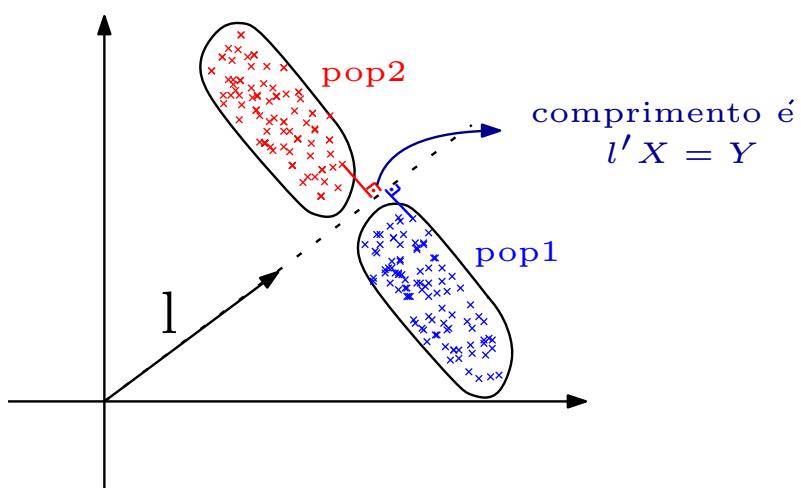
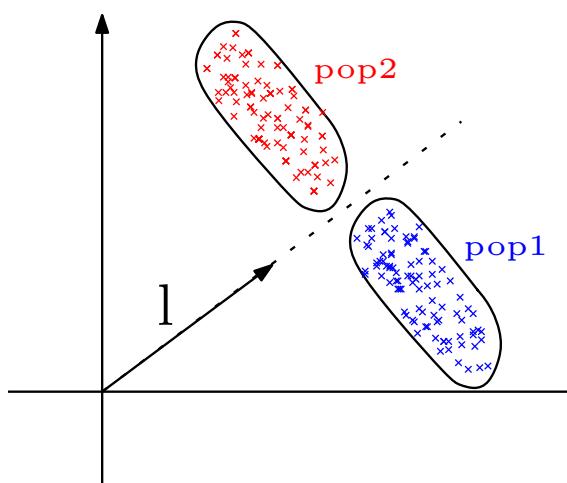
Função Discriminante de Fisher

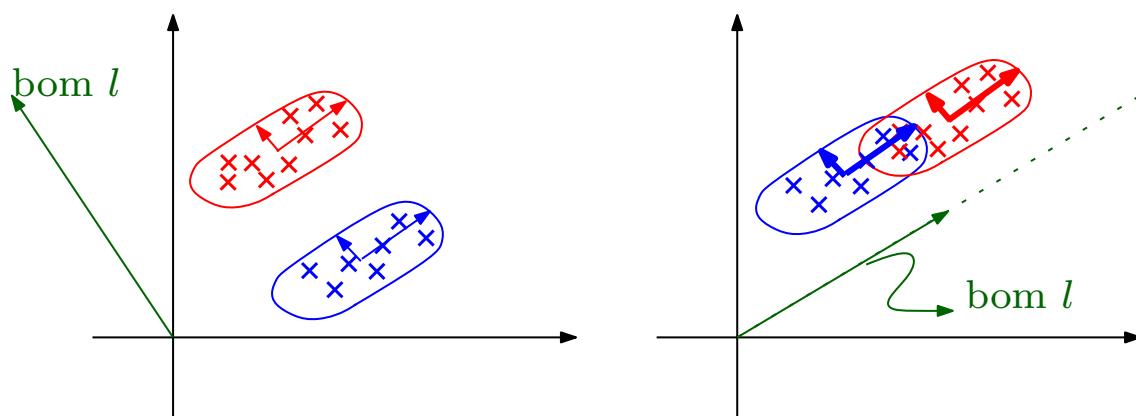
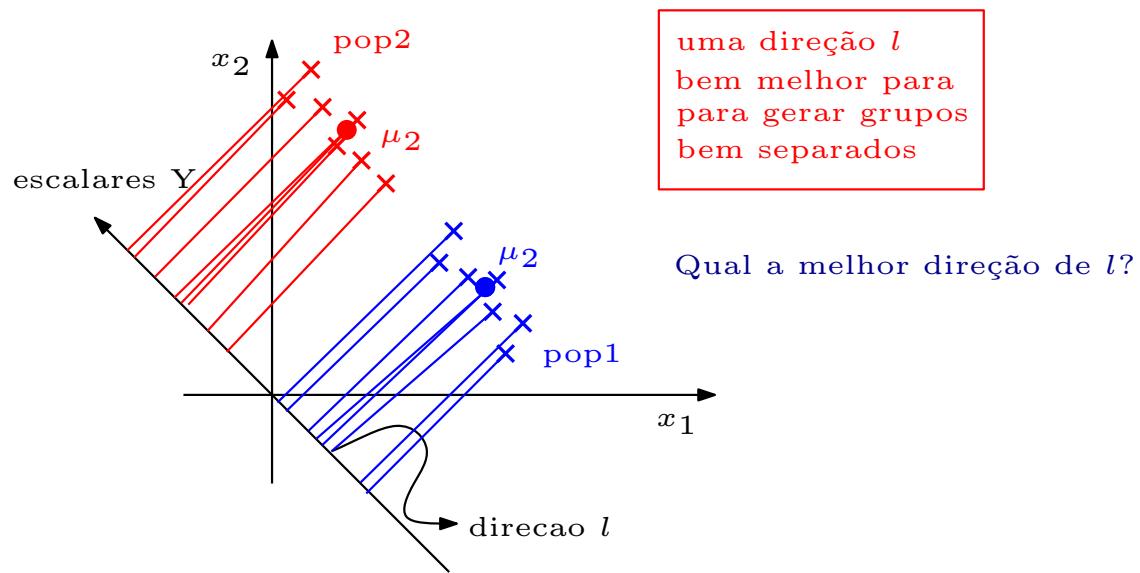
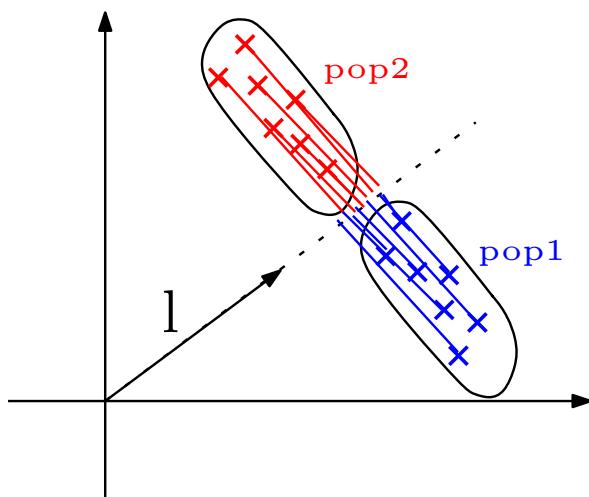
- *Linear Discriminant Analysis (LDA)*
- $X(X_1, X_2, \dots, X_p)$
- Fisher: Vamos criar um índice univariado (escalar) calculando $Y = l'X$, onde $l' = (l_1, l_2, \dots, l_p)$ é um vetor de constantes de forma que os:
- Projeção ortogonal de X em l :
- Suponha $\|l\|^2 = 1$
- Estamos buscando direção l em que $l'X = Y$ dos dois grupos sejam maximalmente separados
- l acima é uma má escolha!
- Projeção $Y = l'X$ de dois vetores X : um de pop1, outro de pop2.
- Projeção de todos os vetores gerando os escalares $Y_{11} Y_{12} \dots Y_{1m_1}$ (pop1) $Y_{21} Y_{22} \dots Y_{2m_2}$ (pop2).
- Pop1 e pop2 não ficam separadas.
- Melhor direção não tem associação simples com os autovetores de Σ
- A esquerda, bom $l \approx 2^\circ$ (menor) autovetor.
- A direita, bom $l \approx 1^\circ$ (maior) autovetor.
- Na maioria das vezes, não é nenhum dos autovetores de Σ

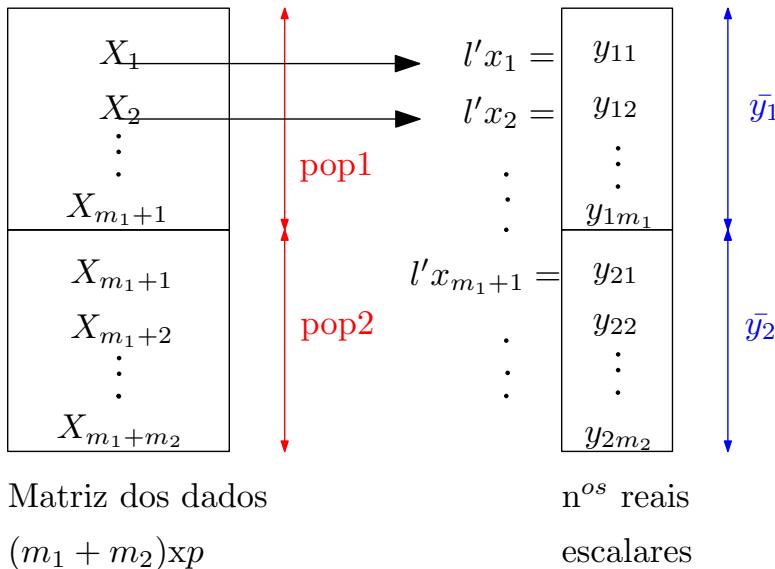


$$\frac{X' l}{\|l\|^2} \cdot l = (X' l) \cdot l, \text{ se } \|l\|^2 = 1$$

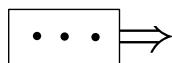
↓
“tamanho das projeções”







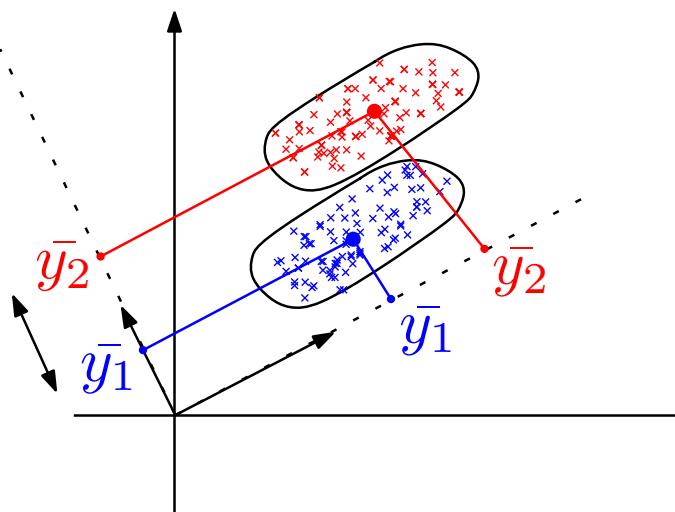
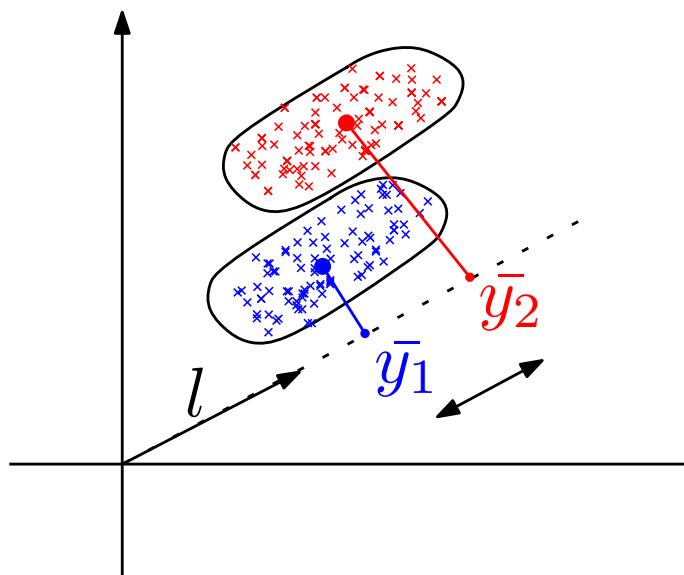
Separação entre os grupos das duas populações $= \|y\bar{1} - y\bar{2}\|$



- Para encontrar a solução Fisher raciocinou assim inicialmente:
 - Projete cada item de dado l gerando o escalar $Y = l'X$
 - Calcule a média de: $\text{pop1} = y\bar{1}$, $\text{pop2} = y\bar{2}$
- Procure a direção l em que $\|y\bar{1} - y\bar{2}\|$ seja máxima.
- Mas isso tem um problema:
 - $\|y\bar{1} - y\bar{2}\|$ é grande mas as projeções não estão bem separadas.
 - Outra direção em que $\|y\bar{1} - y\bar{2}\|$ é menor do que a primeira, mas que separa melhor as duas populações.

Solução de Fisher

- Separação $= \frac{\|y\bar{1} - y\bar{2}\|}{S_y}$
 - $S_y = \text{DP}$ das observações no eixo da projeção $S_y^2 = \frac{1}{m_1+m_2-2} \left(\sum_{j=1}^{m_1} (y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{m_2} (y_{2j} - \bar{y}_2)^2 \right)$
- Teorema: A combinação linear $Y = l'X$ que maximiza a separação $\frac{\|y\bar{1} - y\bar{2}\|}{S_y}$ é dada por $l' = (\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1}$.
- $S_{pooled} = \frac{m_1}{m} S_{pop1} + \frac{m_2}{m} S_{pop2}$;
 - $S_{pop1} \rightarrow$ Matriz de variância e covariância da pop1.
- Prova: Johnson & Wichern, Applied Multivariate Statistical Analysis.
- LDA de Fisher assume que as matrizes de covariância Σ_1 e Σ_2 sejam iguais.
- Assim, Fisher LDA fornece uma direção l em que, se projetarmos os dados, teremos o máximo de separação entre as populações.
- Podemos usar LDA para classificação, mas...
- Este procedimento resulta na mesma regra de classificação vista antes no caso gaussiano.





16. Teoremas Limite

16.1 Introdução

Bla

16.2 Convergence

Let X_1, X_2, \dots be a sequence of random variables, and let X be another random variable with distribution P . Let F_n be the cdf of X_n and let F be the cdf of X .

1. X_n converges *almost surely* to X , $X_n \xrightarrow{q.s.} X$, if for every $\varepsilon > 0$,

$$P\left(\lim_{n \rightarrow \infty} |X_n - X| < \varepsilon\right) = 1. \quad (16.1)$$

2. X_n converges *in probability* to X , $X_n \xrightarrow{P} X$, if for every $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} P(|X_n - X| < \varepsilon) = 1. \quad (16.2)$$

3. X_n converges *in L_p* to X , $X_n \xrightarrow{L_p} X$, if

$$\lim_{n \rightarrow \infty} \int |X_n - X|^p dP = \lim_{n \rightarrow \infty} \mathbb{E}[|X_n - X|^p] = 0. \quad (16.3)$$

4. X_n converges *in distribution* to X , $X_n \rightsquigarrow X$, if

$$\lim_{n \rightarrow \infty} F_n(t) = F(t) \quad (16.4)$$

for all t for which F is continuous.

■ **Example 16.1** This example shows that convergence in probability does not imply almost sure convergence. Let $S = [0, 1]$. Let P be uniform on $[0, 1]$. We draw $S \sim P$. Let $X(s) = s$ and let

$$X_1 = s + I_{[0,1]}(s)$$

$$X_4 = s + I_{[0,1/3]}(s)$$

$$X_2 = s + I_{[0,1/2]}(s)$$

$$X_5 = s + I_{[1/3,2/3]}(s)$$

$$X_3 = s + I_{[1/2,1]}(s)$$

$$X_6 = s + I_{[2/3,1]}(s)$$

etc. Then $X_n \xrightarrow{P} X$ since $P(|X_n - X| > \varepsilon)$ is equal to the probability of an interval of s values whose length is going to zero. However, for every s , $X_n(s)$ alternates between the values s and $s + 1$ infinitely often, so this convergence does not occur almost surely. ■

Theorem 16.2.1 The following relationships hold:

- (a) $X_n \xrightarrow{L_p} X$ implies that $X_n \xrightarrow{P} X$.
- (b) $X_n \xrightarrow{P} X$ implies that $X_n \rightsquigarrow X$.
- (c) If $X_n \rightsquigarrow X$ and if $P(X = c) = 1$ for some real number c , then $X_n \xrightarrow{P} X$.
- (d) $X_n \xrightarrow{q.c.} X$ implies that $X_n \xrightarrow{P} X$.

Theorem 16.2.2 Let X_n, X, Y_n, Y be random variables. Let g be a continuous function. Let c be a constant.

- (a) If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then $X_n + Y_n \xrightarrow{P} X + Y$.
- (b) If $X_n \xrightarrow{L_p} X$ and $Y_n \xrightarrow{L_p} Y$, then $X_n + Y_n \xrightarrow{L_p} X + Y$.
- (c) If $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow c$, then $X_n + Y_n \rightsquigarrow X + c$.
- (d) If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then $X_n Y_n \xrightarrow{P} XY$.
- (e) If $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow c$, then $X_n Y_n \rightsquigarrow cX$.
- (f) If $X_n \xrightarrow{\text{converges somehow}} X$, then $g(X_n) \xrightarrow{\text{converges the same}} g(X)$.
 - Parts (c) and (e) are known as *Slutsky's theorem*.
 - Part (f) is known as *The continuous mapping theorem*.

16.3 Lei dos Grandes Números

We are now in a position to prove our first fundamental theorem of probability. We have seen that an intuitive way to view the probability of a certain outcome is as the frequency with which that outcome occurs in the long run, when the experiment is repeated a large number of times. We have also defined probability mathematically as a value of a distribution function for the random variable representing the experiment. The Law of Large Numbers, which is a theorem proved about the mathematical model of probability, shows that this model is consistent with the frequency interpretation of probability. This theorem is sometimes called the law of averages. To find out what would happen if this law were not true, see the article by Robert M. Coates.²

Relembre a desigualdade de Tchebyshev.

Theorem 16.3.1 — Tchebyshev Inequality. Let X be a random variable with expected value $\mu = \mathbb{E}(X)$ and variance $\mathbb{V}(X) = \sigma^2$. Let let $\varepsilon > 0$ be any positive real number. Then

$$P(|X - \mu| \geq \varepsilon) \leq \frac{\mathbb{V}(X)}{\varepsilon^2}.$$

Note that X in the above theorem can be *any* random variable, and ε any positive number.

Estamos agora na posição em que podemos provar o primeiro teorema. Através dele, veremos de maneira formal, rigorosa, que a nossa intuição está correta: se X_1, X_2, \dots são v.a.'s i.i.d. com valor esperado μ então a média aritmética dessas variáveis aleatórias converge para μ . Em que sentido se dá esta convergência? Lembre-se que a média aritmética é uma v.a. Com que rapidez a convergência ocorre?

²R. M. Coates, "The Law," The World of Mathematics, ed. James R. Newman (New York: Simon and Schuster, 1956).

Theorem 16.3.2 — Law of Large Numbers. Let X_1, X_2, \dots, X_n be an independent trials process, with finite expected value $\mu = E(X_j)$ and finite variance $\sigma^2 = V(X_j)$. Let $S_n = X_1 + X_2 + \dots + X_n$. Then for any $\varepsilon > 0$,

$$P\left(\left|\frac{S_n}{n} - \mu\right| \geq \varepsilon\right) \rightarrow 0$$

as $n \rightarrow \infty$. Equivalently,

$$P\left(\left|\frac{S_n}{n} - \mu\right| < \varepsilon\right) \rightarrow 1$$

as $n \rightarrow \infty$.

Prova: Since X_1, X_2, \dots, X_n are independent and have the same distributions, we can apply Theorem ???. We obtain

$$V(S_n) = n\sigma^2 ,$$

and

$$V\left(\frac{S_n}{n}\right) = \frac{\sigma^2}{n} .$$

Also we know that

$$E\left(\frac{S_n}{n}\right) = \mu .$$

By Chebyshev's Inequality, for any $\varepsilon > 0$,

$$P\left(\left|\frac{S_n}{n} - \mu\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2} .$$

Thus, for fixed ε ,

$$P\left(\left|\frac{S_n}{n} - \mu\right| \geq \varepsilon\right) \rightarrow 0$$

as $n \rightarrow \infty$, or equivalently,

$$P\left(\left|\frac{S_n}{n} - \mu\right| < \varepsilon\right) \rightarrow 1$$

as $n \rightarrow \infty$.

Note that this theorem is not necessarily true if σ^2 is infinite. Caso da Pareto.

Law of Averages

Note that S_n/n is an average of the individual outcomes, and one often calls the Law of Large Numbers the “law of averages.” It is a striking fact that we can start with a random experiment about which little can be predicted and, by taking averages, obtain an experiment in which the outcome can be predicted with a high degree of certainty. The Law of Large Numbers, as we have stated it, is often called the “Weak Law of Large Numbers” to distinguish it from the “Strong Law of Large Numbers” described in Exercise ??.

Consider the important special case of Bernoulli trials with probability p for success. Let $X_j = 1$ if the j th outcome is a success and 0 if it is a failure. Then $S_n = X_1 + X_2 + \dots + X_n$ is the number of successes in n trials and $\mu = E(X_1) = p$. The Law of Large Numbers states that for any $\varepsilon > 0$

$$P\left(\left|\frac{S_n}{n} - p\right| < \varepsilon\right) \rightarrow 1$$

as $n \rightarrow \infty$. The above statement says that, in a large number of repetitions of a Bernoulli experiment, we can expect the proportion of times the event will occur to be near p . This shows that our mathematical model of probability agrees with our frequency interpretation of probability.

Coin Tossing

Let us consider the special case of tossing a coin n times with S_n the number of heads that turn up. Then the random variable S_n/n represents the fraction of times heads turns up and will have values between 0 and 1. The Law of Large Numbers predicts that the outcomes for this random variable will, for large n , be near $1/2$.

In Figure ??, we have plotted the distribution for this example for increasing values of n . We have marked the outcomes between .45 and .55 by dots at the top of the spikes. We see that as n increases the distribution gets more and more concentrated around .5 and a larger and larger percentage of the total area is contained within the interval (.45, .55), as predicted by the Law of Large Numbers.

Die Rolling

■ **Example 16.2** Consider n rolls of a die. Let X_j be the outcome of the j th roll. Then $S_n = X_1 + X_2 + \dots + X_n$ is the sum of the first n rolls. This is an independent trials process with $E(X_j) = 7/2$. Thus, by the Law of Large Numbers, for any $\varepsilon > 0$

$$P\left(\left|\frac{S_n}{n} - \frac{7}{2}\right| \geq \varepsilon\right) \rightarrow 0$$

as $n \rightarrow \infty$. An equivalent way to state this is that, for any $\varepsilon > 0$,

$$P\left(\left|\frac{S_n}{n} - \frac{7}{2}\right| < \varepsilon\right) \rightarrow 1$$

as $n \rightarrow \infty$. ■

Numerical Comparisons

It should be emphasized that, although Chebyshev's Inequality proves the Law of Large Numbers, it is actually a very crude inequality for the probabilities involved. However, its strength lies in the fact that it is true for any random variable at all, and it allows us to prove a very powerful theorem.

In the following example, we compare the estimates given by Chebyshev's Inequality with the actual values.

■ **Example 16.3** Let X_1, X_2, \dots, X_n be a Bernoulli trials process with probability .3 for success and .7 for failure. Let $X_j = 1$ if the j th outcome is a success and 0 otherwise. Then, $E(X_j) = .3$ and $V(X_j) = (.3)(.7) = .21$. If

$$A_n = \frac{S_n}{n} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

is the average of the X_i , then $E(A_n) = .3$ and $V(A_n) = V(S_n)/n^2 = .21/n$. Chebyshev's Inequality states that if, for example, $\varepsilon = .1$,

$$P(|A_n - .3| \geq .1) \leq \frac{.21}{n(.1)^2} = \frac{21}{n} .$$

Thus, if $n = 100$,

$$P(|A_{100} - .3| \geq .1) \leq .21 ,$$

or if $n = 1000$,

$$P(|A_{1000} - .3| \geq .1) \leq .021 .$$

These can be rewritten as

$$P(.2 < A_{100} < .4) \geq .79 ,$$

$$P(.2 < A_{1000} < .4) \geq .979 .$$

These values should be compared with the actual values, which are (to six decimal places)

$$\begin{aligned} P(.2 < A_{100} < .4) &\approx .962549 \\ P(.2 < A_{1000} < .4) &\approx 1. \end{aligned}$$

The program **Law** can be used to carry out the above calculations in a systematic way. ■

Uniform Case

■ **Example 16.4** Suppose we choose at random n numbers from the interval $[0, 1]$ with uniform distribution. Then if X_i describes the i th choice, we have

$$\begin{aligned} \mu &= E(X_i) = \int_0^1 x dx = \frac{1}{2}, \\ \sigma^2 &= V(X_i) = \int_0^1 x^2 dx - \mu^2 \\ &= \frac{1}{3} - \frac{1}{4} = \frac{1}{12}. \end{aligned}$$

Hence,

$$\begin{aligned} E\left(\frac{S_n}{n}\right) &= \frac{1}{2}, \\ V\left(\frac{S_n}{n}\right) &= \frac{1}{12n}, \end{aligned}$$

and for any $\varepsilon > 0$,

$$P\left(\left|\frac{S_n}{n} - \frac{1}{2}\right| \geq \varepsilon\right) \leq \frac{1}{12n\varepsilon^2}.$$

This says that if we choose n numbers at random from $[0, 1]$, then the chances are better than $1 - 1/(12n\varepsilon^2)$ that the difference $|S_n/n - 1/2|$ is less than ε . Note that ε plays the role of the amount of error we are willing to tolerate: If we choose $\varepsilon = 0.1$, say, then the chances that $|S_n/n - 1/2|$ is less than 0.1 are better than $1 - 100/(12n)$. For $n = 100$, this is about .92, but if $n = 1000$, this is better than .99 and if $n = 10,000$, this is better than .999. 5.0trueinPSfig8-2Illustration of Law of Large Numbers — uniform case.fig 8.2

We can illustrate what the Law of Large Numbers says for this example graphically. The density for $A_n = S_n/n$ is determined by

$$f_{A_n}(x) = nf_{S_n}(nx).$$

We have seen in Section ??, that we can compute the density $f_{S_n}(x)$ for the sum of n uniform random variables. In Figure ?? we have used this to plot the density for A_n for various values of n . We have shaded in the area for which A_n would lie between .45 and .55. We see that as we increase n , we obtain more and more of the total area inside the shaded region. The Law of Large Numbers tells us that we can obtain as much of the total area as we please inside the shaded region by choosing n large enough (see also Figure ??). ■

Normal Case

■ **Example 16.5** Suppose we choose n real numbers at random, using a normal distribution with mean 0 and variance 1. Then

$$\begin{aligned} \mu &= E(X_i) = 0, \\ \sigma^2 &= V(X_i) = 1. \end{aligned}$$

n	$P(S_n/n \geq .1)$	Chebyshev
100	.31731	1.00000
200	.15730	.50000
300	.08326	.33333
400	.04550	.25000
500	.02535	.20000
600	.01431	.16667
700	.00815	.14286
800	.00468	.12500
900	.00270	.11111
1000	.00157	.10000

Table 16.1: Chebyshev estimates.

Hence,

$$\begin{aligned} E\left(\frac{S_n}{n}\right) &= 0, \\ V\left(\frac{S_n}{n}\right) &= \frac{1}{n}, \end{aligned}$$

and, for any $\varepsilon > 0$,

$$P\left(\left|\frac{S_n}{n} - 0\right| \geq \varepsilon\right) \leq \frac{1}{n\varepsilon^2}.$$

In this case it is possible to compare the Chebyshev estimate for $P(|S_n/n - \mu| \geq \varepsilon)$ in the Law of Large Numbers with exact values, since we know the density function for S_n/n exactly (see Example ??). The comparison is shown in Table 16.1, for $\varepsilon = .1$. The data in this table was produced by the program **LawContinuous**. We see here that the Chebyshev estimates are in general not very accurate. ■

Monte Carlo Method

Here is a somewhat more interesting example.

■ **Example 16.6** Let $g(x)$ be a continuous function defined for $x \in [0, 1]$ with values in $[0, 1]$. In Section ??, we showed how to estimate the area of the region under the graph of $g(x)$ by the Monte Carlo method, that is, by choosing a large number of random values for x and y with uniform distribution and seeing what fraction of the points $P(x, y)$ fell inside the region under the graph (see Example ??).

Here is a better way to estimate the same area (see Figure ??). Let us choose a large number of independent values X_n at random from $[0, 1]$ with uniform density, set $Y_n = g(X_n)$, and find the average value of the Y_n . Then this average is our estimate for the area. To see this, note that if the density function for X_n is uniform,

$$\begin{aligned} \mu &= E(Y_n) = \int_0^1 g(x)f(x)dx \\ &= \int_0^1 g(x)dx \\ &= \text{average value of } g(x), \end{aligned}$$

while the variance is

$$\sigma^2 = E((Y_n - \mu)^2) = \int_0^1 (g(x) - \mu)^2 dx < 1,$$

since for all x in $[0, 1]$, $g(x)$ is in $[0, 1]$, hence μ is in $[0, 1]$, and so $|g(x) - \mu| \leq 1$. Now let $A_n = (1/n)(Y_1 + Y_2 + \dots + Y_n)$. Then by Chebyshev's Inequality, we have

$$P(|A_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} < \frac{1}{n\varepsilon^2}.$$

3trueinPSfig8-3Area problem.fig 8.3

This says that to get within ε of the true value for $\mu = \int_0^1 g(x) dx$ with probability at least p , we should choose n so that $1/n\varepsilon^2 \leq 1-p$ (i.e., so that $n \geq 1/\varepsilon^2(1-p)$). Note that this method tells us how large to take n to get a desired accuracy. ■

The Law of Large Numbers requires that the variance σ^2 of the original underlying density be finite: $\sigma^2 < \infty$. In cases where this fails to hold, the Law of Large Numbers may fail, too. An example follows.

Cauchy Case

■ **Example 16.7** Suppose we choose n numbers from $(-\infty, +\infty)$ with a Cauchy density with parameter $a = 1$. We know that for the Cauchy density the expected value and variance are undefined (see Example ??). In this case, the density function for

$$A_n = \frac{S_n}{n}$$

is given by (see Example ??)

$$f_{A_n}(x) = \frac{1}{\pi(1+x^2)},$$

that is, the density function for A_n is the same for all n . In this case, as n increases, the density function does not change at all, and the Law of Large Numbers does not hold. ■

Historical Remarks

The Law of Large Numbers was first proved by the Swiss mathematician James Bernoulli in the fourth part of his work *Ars Conjectandi* published posthumously in 1713.³ As often happens with a first proof, Bernoulli's proof was much more difficult than the proof we have presented using Chebyshev's inequality. Chebyshev developed his inequality to prove a general form of the Law of Large Numbers (see Exercise ??). The inequality itself appeared much earlier in a work by Bienaym  , and in discussing its history Maistrov remarks that it was referred to as the Bienaym  -Chebyshev Inequality for a long time.⁴

In *Ars Conjectandi* Bernoulli provides his reader with a long discussion of the meaning of his theorem with lots of examples. In modern notation he has an event that occurs with probability p but he does not know p . He wants to estimate p by the fraction \bar{p} of the times the event occurs when the experiment is repeated a number of times. He discusses in detail the problem of estimating, by this method, the proportion of white balls in an urn that contains an unknown number of white and black balls. He would do this by drawing a sequence of balls from the urn, replacing the ball drawn after each draw, and estimating the unknown proportion of white balls in the urn by the proportion of the balls drawn that are white. He shows that, by choosing n large enough he can obtain any desired accuracy and reliability for the estimate. He also provides a lively discussion of the applicability of his theorem to estimating the probability of dying of a particular disease, of different kinds of weather occurring, and so forth.

³J. Bernoulli, *The Art of Conjecturing IV*, trans. Bing Sung, Technical Report No. 2, Dept. of Statistics, Harvard Univ., 1966

⁴L. E. Maistrov, *Probability Theory: A Historical Approach*, trans. and ed. Samuel Kotz, (New York: Academic Press, 1974), p. 202

In speaking of the number of trials necessary for making a judgement, Bernoulli observes that the “man on the street” believes the “law of averages.”

Further, it cannot escape anyone that for judging in this way about any event at all, it is not enough to use one or two trials, but rather a great number of trials is required. And sometimes the stupidest man—by some instinct of nature per se and by no previous instruction (this is truly amazing)—knows for sure that the more observations of this sort that are taken, the less the danger will be of straying from the mark.⁵

But he goes on to say that he must contemplate another possibility.

Something further must be contemplated here which perhaps no one has thought about till now. It certainly remains to be inquired whether after the number of observations has been increased, the probability is increased of attaining the true ratio between the number of cases in which some event can happen and in which it cannot happen, so that this probability finally exceeds any given degree of certainty; or whether the problem has, so to speak, its own asymptote—that is, whether some degree of certainty is given which one can never exceed.⁶

Bernoulli recognized the importance of this theorem, writing:

Therefore, this is the problem which I now set forth and make known after I have already pondered over it for twenty years. Both its novelty and its very great usefulness, coupled with its just as great difficulty, can exceed in weight and value all the remaining chapters of this thesis.⁷

Bernoulli concludes his long proof with the remark:

Whence, finally, this one thing seems to follow: that if observations of all events were to be continued throughout all eternity, (and hence the ultimate probability would tend toward perfect certainty), everything in the world would be perceived to happen in fixed ratios and according to a constant law of alternation, so that even in the most accidental and fortuitous occurrences we would be bound to recognize, as it were, a certain necessity and, so to speak, a certain fate.

I do now know whether Plato wished to aim at this in his doctrine of the universal return of things, according to which he predicted that all things will return to their original state after countless ages have past.⁸

16.3.1 Uma forma mais geral

A seguir, uma forma mais geral do teorema, em que as v.a.’s não precisam ter a mesma distribuição:

Theorem 16.3.3 — Weak law of large numbers. Let X_1, X_2, \dots be independent random variables, each with finite mean and variance. Define $S_n = X_1 + \dots + X_n$. Then $\frac{1}{n}(S_n - \mathbb{E}[S_n]) \xrightarrow{P} 0$.

16.3.2 A versão forte

⁵Bernoulli, op. cit., p. 38.

⁶ibid., p. 39.

⁷ibid., p. 42.

⁸ibid., pp. 65–66.

Theorem 16.3.4 — Strong law of large numbers. Let X_1, X_2, \dots be iid random variables with common mean m . Define $S_n = X_1 + \dots + X_n$. Then $S_n/n \xrightarrow{q.c.} m$.

The laws of large numbers tell us that the probability mass of an average of random variables “piles up” near its expectation. In just a minute, we will see even more: how fast this piling occurs. But first we should talk about the distribution of the average.

16.4 Moment generating functions

If X is a continuous random variable, then the analogue of the probability generating function is the moment generating function: [Moment generating function] The *moment generating function* of a random variable X is

$$m(\theta) = \mathbb{E}[e^{\theta X}].$$

For those θ in which $m(\theta)$ is finite, we have

$$m(\theta) = \int_{-\infty}^{\infty} e^{\theta x} f(x) \, dx.$$

We can prove results similar to that we had for probability generating functions.

We will assume the following without proof:

Theorem 16.4.1 The mgf determines the distribution of X provided $m(\theta)$ is finite for all θ in some interval containing the origin.

Definition 16.4.1 — Moment. The r th *moment* of X is $\mathbb{E}[X^r]$.

Theorem 16.4.2 The r th moment X is the coefficient of $\frac{\theta^r}{r!}$ in the power series expansion of $m(\theta)$, and is

$$\mathbb{E}[X^r] = \left. \frac{\theta^r}{r!} m(\theta) \right|_{\theta=0} = m^{(r)}(0).$$

Proof: We have

$$e^{\theta X} = 1 + \theta X + \frac{\theta^2}{2!} X^2 + \dots$$

So

$$m(\theta) = \mathbb{E}[e^{\theta X}] = 1 + \theta \mathbb{E}[X] + \frac{\theta^2}{2!} \mathbb{E}[X^2] + \dots$$

■ **Example 16.8** Let $X \sim \mathcal{E}(\lambda)$. Then its mgf is

$$\mathbb{E}[e^{\theta X}] = \int_0^{\infty} e^{\theta x} \lambda e^{-\lambda x} \, dx = \lambda \int_0^{\infty} e^{-(\lambda - \theta)x} \, dx = \frac{\lambda}{\lambda - \theta},$$

where $0 < \theta < \lambda$. So

$$\mathbb{E}[X] = m'(\theta) = \left. \frac{\lambda}{(\lambda - \theta)^2} \right|_{\theta=0} = \frac{1}{\lambda}.$$

Also,

$$\mathbb{E}[X^2] = m''(0) = \frac{2\lambda}{(\lambda - \theta)^3} \Big|_{\theta=0} = \frac{2}{\lambda^2}.$$

So

$$\mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

■

Theorem 16.4.3 If X and Y are independent random variables with moment generating functions $m_X(\theta), m_Y(\theta)$, then $X + Y$ has mgf $m_{X+Y}(\theta) = m_X(\theta)m_Y(\theta)$.

Proof:

$$\mathbb{E}[e^{\theta(X+Y)}] = \mathbb{E}[e^{\theta X} e^{\theta Y}] = \mathbb{E}[e^{\theta X}] \mathbb{E}[e^{\theta Y}] = m_X(\theta)m_Y(\theta).$$

16.4.1 Caso Normal

Proposition 16.4.4 The moment generating function of $N(\mu, \sigma^2)$ is

$$\mathbb{E}[e^{\theta X}] = \exp\left(\theta\mu + \frac{1}{2}\theta^2\sigma^2\right).$$

Proof:

$$\mathbb{E}[e^{\theta X}] = \int_{-\infty}^{\infty} e^{\theta x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\sigma^2(x-\mu)^2} dx.$$

Substitute $z = \frac{x-\mu}{\sigma}$. Then

$$\begin{aligned} \mathbb{E}[e^{\theta X}] &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\theta(\mu+\sigma z)} e^{-\frac{1}{2}z^2} dz \\ &= e^{\theta\mu + \frac{1}{2}\theta^2\sigma^2} \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-\theta\sigma)^2} dz}_{\text{pdf of } N(\sigma\theta, 1)} \\ &= e^{\theta\mu + \frac{1}{2}\theta^2\sigma^2}. \end{aligned}$$

Theorem 16.4.5 Suppose X, Y are independent random variables with $X \sim N(\mu_1, \sigma_1^2)$, and $Y \sim (\mu_2, \sigma_2^2)$. Then

1. $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.
2. $aX \sim N(a\mu_1, a^2\sigma_1^2)$.

Proof:

1.

$$\begin{aligned} \mathbb{E}[e^{\theta(X+Y)}] &= \mathbb{E}[e^{\theta X}] \cdot \mathbb{E}[e^{\theta Y}] \\ &= e^{\mu_1\theta + \frac{1}{2}\sigma_1^2\theta^2} \cdot e^{\mu_2\theta + \frac{1}{2}\sigma_2^2\theta^2} \\ &= e^{(\mu_1+\mu_2)\theta + \frac{1}{2}(\sigma_1^2+\sigma_2^2)\theta^2} \end{aligned}$$

which is the mgf of $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

2.

$$\begin{aligned}\mathbb{E}[e^{\theta(aX)}] &= \mathbb{E}[e^{(\theta a)X}] \\ &= e^{\mu(a\theta) + \frac{1}{2}\sigma^2(a\theta)^2} \\ &= e^{(a\mu)\theta + \frac{1}{2}(a^2\sigma^2)\theta^2}\end{aligned}$$

Finally, suppose $X \sim N(0, 1)$. Write $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ for its pdf. It would be very difficult to find a closed form for its cumulative distribution function, but we can find an upper bound for it:

$$\begin{aligned}\mathbb{P}(X \geq x) &= \int_x^\infty \phi(t) dt \\ &\leq \int_x^\infty \left(1 + \frac{1}{t^2}\right) \phi(t) dt \\ &= \frac{1}{x} \phi(x)\end{aligned}$$

To see the last step works, simply differentiate the result and see that you get $(1 + \frac{1}{x^2})\phi(x)$. So

$$\mathbb{P}(X \geq x) \leq \frac{1}{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.$$

Then

$$\log \mathbb{P}(X \geq x) \sim -\frac{1}{2}x^2.$$

16.5 Teorema Central do Limite

Theorem 16.5.1 — Central limit theorem. Let X_1, X_2, \dots be iid with mean μ and variance $\sigma^2 < \infty$. Let $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$. Then,

$$Z_n := \frac{\bar{X}_n - \mu}{\sqrt{\mathbb{V}[\bar{X}_n]}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \rightsquigarrow Z, \quad (16.5)$$

where $Z \sim N(0, 1)$.

Suppose X_1, \dots, X_n are iid random variables with mean μ and variance σ^2 . Let $S_n = X_1 + \dots + X_n$. Then we have previously shown that

$$\mathbb{V}(S_n/\sqrt{n}) = \mathbb{V}\left(\frac{S_n - n\mu}{\sqrt{n}}\right) = \sigma^2.$$

Theorem 16.5.2 — Central limit theorem. Let X_1, X_2, \dots be iid random variables with $\mathbb{E}[X_i] = \mu$, $\mathbb{V}(X_i) = \sigma^2 < \infty$. Define

$$S_n = X_1 + \dots + X_n.$$

Then for all finite intervals (a, b) ,

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b\right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt.$$

Note that the final term is the pdf of a standard normal. We say

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \rightarrow_D N(0, 1).$$

To show this, we will use the continuity theorem without proof:

Theorem 16.5.3 — Continuity theorem. If the random variables X_1, X_2, \dots have mgf's $m_1(\theta), m_2(\theta), \dots$ and $m_n(\theta) \rightarrow m(\theta)$ as $n \rightarrow \infty$ for all θ , then $X_n \rightarrow_D$ the random variable with mgf $m(\theta)$.

We now provide a sketch-proof of the central limit theorem: **Proof:** wlog, assume $\mu = 0, \sigma^2 = 1$ (otherwise replace X_i with $\frac{X_i - \mu}{\sigma}$).

Then

$$\begin{aligned} m_{X_i}(\theta) &= \mathbb{E}[e^{\theta X_i}] = 1 + \theta \mathbb{E}[X_i] + \frac{\theta^2}{2!} \mathbb{E}[X_i^2] + \dots \\ &= 1 + \frac{1}{2}\theta^2 + \frac{1}{3!}\theta^3 \mathbb{E}[X_i^3] + \dots \end{aligned}$$

Now consider S_n/\sqrt{n} . Then

$$\begin{aligned} \mathbb{E}[e^{\theta S_n/\sqrt{n}}] &= \mathbb{E}[e^{\theta(X_1 + \dots + X_n)/\sqrt{n}}] \\ &= \mathbb{E}[e^{\theta X_1/\sqrt{n}}] \dots \mathbb{E}[e^{\theta X_n/\sqrt{n}}] \\ &= \left(\mathbb{E}[e^{\theta X_1/\sqrt{n}}] \right)^n \\ &= \left(1 + \frac{1}{2}\theta^2 \frac{1}{n} + \frac{1}{3!}\theta^3 \mathbb{E}[X^3] \frac{1}{n^{3/2}} + \dots \right)^n \\ &\rightarrow e^{\frac{1}{2}\theta^2} \end{aligned}$$

as $n \rightarrow \infty$ since $(1 + a/n)^n \rightarrow e^a$. And this is the mgf of the standard normal. So the result follows from the continuity theorem.

Note that this is not a very formal proof, since we have to require $\mathbb{E}[X^3]$ to be finite. Also, sometimes the moment generating function is not defined. But this will work for many “nice” distributions we will ever meet.

The proper proof uses the characteristic function

$$\chi_X(\theta) = E[e^{i\theta X}].$$

An important application is to use the normal distribution to approximate a large binomial.

Let $X_i \sim B(1, p)$. Then $S_n \sim B(n, p)$. So $\mathbb{E}[S_n] = np$ and $\mathbb{V}(S_n) = p(1-p)$. So

$$\frac{S_n - np}{\sqrt{np(1-p)}} \rightarrow_D N(0, 1).$$

■ **Example 16.9** Suppose two planes fly a route. Each of n passengers chooses a plane at random. The number of people choosing plane 1 is $S \sim B(n, \frac{1}{2})$. Suppose each has s seats. What is

$$F(s) = \mathbb{P}(S > s),$$

i.e. the probability that plane 1 is over-booked? We have

$$F(s) = \mathbb{P}(S > s) = \mathbb{P}\left(\frac{S - n/2}{\sqrt{n \cdot \frac{1}{2} \cdot \frac{1}{2}}} > \frac{s - n/2}{\sqrt{n}/2}\right).$$

Since

$$\frac{S - np}{\sqrt{n}/2} \sim N(0, 1),$$

we have

$$F(s) \approx 1 - \Phi\left(\frac{s - n/2}{\sqrt{n}/2}\right).$$

For example, if $n = 1000$ and $s = 537$, then $\frac{S_n - n/2}{\sqrt{n}/2} \approx 2.34$, $\Phi(2.34) \approx 0.99$, and $F(s) \approx 0.01$. So with only 74 seats as buffer between the two planes, the probability of overbooking is just $1/100$. ■

■ **Example 16.10** An unknown proportion p of the electorate will vote Labour. It is desired to find p without an error not exceeding 0.005. How large should the sample be?

We estimate by

$$p' = \frac{S_n}{n},$$

where $X_i \sim B(1, p)$. Then

$$\begin{aligned} \mathbb{P}(|p' - p| \leq 0.005) &= \mathbb{P}(|S_n - np| \leq 0.005n) \\ &= \mathbb{P}\left(\underbrace{\frac{|S_n - np|}{\sqrt{np(1-p)}}}_{\approx N(0,1)} \leq \frac{0.005n}{\sqrt{np(1-p)}}\right) \end{aligned}$$

We want $|p' - p| \leq 0.005$ with probability ≥ 0.95 . Then we want

$$\frac{0.005n}{\sqrt{np(1-p)}} \geq \Phi^{-1}(0.975) = 1.96.$$

(we use 0.975 instead of 0.95 since we are doing a two-tailed test) Since the maximum possible value of $p(1-p)$ is $1/4$, we have

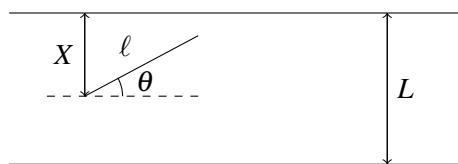
$$n \geq 38416.$$

In practice, we don't have that many samples. Instead, we go by

$$\mathbb{P}(|p' - p| \leq 0.03) \geq 0.95.$$

This just requires $n \geq 1068$. ■

■ **Example 16.11 — Estimating π with Buffon's needle.** Recall that if we randomly toss a needle of length ℓ to a floor marked with parallel lines a distance L apart, the probability that the needle hits the line is $p = \frac{2\ell}{\pi L}$.



Suppose we toss the pin n times, and it hits the line N times. Then

$$N \approx N(np, np(1-p))$$

by the Central limit theorem. Write p' for the actual proportion observed. Then

$$\begin{aligned}\hat{\pi} &= \frac{2\ell}{(N/n)L} \\ &= \frac{\pi 2\ell / (\pi L)}{p'} \\ &= \frac{\pi p}{p + (p' - p)} \\ &= \pi \left(1 - \frac{p' - p}{p} + \dots \right)\end{aligned}$$

Hence

$$\hat{\pi} - \pi \approx \frac{p - p'}{p}.$$

We know

$$p' \sim N\left(p, \frac{p(1-p)}{n}\right).$$

So we can find

$$\hat{\pi} - \pi \sim N\left(0, \frac{\pi^2 p(1-p)}{np^2}\right) = N\left(0, \frac{\pi^2(1-p)}{np}\right)$$

We want a small variance, and that occurs when p is the largest. Since $p = 2\ell/\pi L$, this is maximized with $\ell = L$. In this case,

$$p = \frac{2}{\pi},$$

and

$$\hat{\pi} - \pi \approx N\left(0, \frac{(\pi-2)\pi^2}{2n}\right).$$

If we want to estimate π to 3 decimal places, then we need

$$\mathbb{P}(|\hat{\pi} - \pi| \leq 0.001) \geq 0.95.$$

This is true if and only if

$$0.001 \sqrt{\frac{2n}{(\pi-2)(\pi^2)}} \geq \Phi^{-1}(0.975) = 1.96$$

So $n \geq 2.16 \times 10^7$. So we can obtain π to 3 decimal places just by throwing a stick 20 million times! Isn't that exciting? ■

16.5.1 Historical remarks

The first version of the central limit theorem was proved by DeMoivre around 1733 for the special case when the X_i are binomial random variables with $p = 1/2 = q$. This proof was subsequently extended by Laplace to the case of arbitrary $p \neq q$. Laplace also discovered the more general form of the Central Limit Theorem presented here. His proof however was not completely rigorous, and in fact, cannot be made completely rigorous. A truly rigorous proof of the Central Limit Theorem was first presented by the Russian mathematician Liapunov in 1901-1902. As a result, the Central Limit Theorem (or a slightly stronger version of the CLT) is occasionally referred to as Liapunov's theorem. A theorem with weaker hypotheses but with equally strong conclusion is Lindeberg's Theorem of 1922. It says that the sequence of random variables need not be identically distributed, but instead need only have zero means, and the individual variances are small compared to their sum.

16.5.2 Accuracy of the Approximation by the Central Limit Theorem

The statement of the Central Limit Theorem does not say how good the approximation is. In general the approximation given by the Central Limit Theorem applied to a sequence of Bernoulli random trials or equivalently to a binomial random variable is acceptable when $np(1 - p) > 18$. The normal approximation to a binomial deteriorates as the interval (a, b) over which the probability is computed moves away from the binomial's mean value np .

The Berry-Esséen Theorem gives an explicit bound: For independent, identically distributed random variables X_i with $\mu = \mathbb{E}[X_i] = 0$, $\sigma^2 = \mathbb{E}[X_i^2]$, and $\rho = \mathbb{E}[|X^3|]$, then

$$\left| \mathbb{P}[S_n / (\sigma\sqrt{n}) < a] - \int_{-\infty}^a \frac{1}{\sqrt{2\pi}} e^{u^2/2} du \right| \leq \frac{33}{4} \frac{\rho}{\sigma^3} \frac{1}{\sqrt{n}}.$$

Illustration 1

We expect the normal distribution to arise whenever the outcome of a situation, results from numerous small additive effects, with no single or small group of effects dominant. Here is an illustration of that principle.

This illustration is adapted from *Dicing with Death: Chance, Health, and Risk* by Stephen Senn, Cambridge University Press, Cambridge, 2003.

Consider the following data from an American study called the National Longitudinal Survey of Youth (NLSY). This study originally obtained a sample of over 12,000 respondents aged 14-21 years in 1979. By 1994, the respondents were aged 29-36 years and had 15,000 children among them. Of the respondents 2,444 had exactly two children. In these 2,444 families, the distribution of children was boy-boy: 582; girl-girl 530, boy-girl 666, and girl-boy 666. It appears that the distribution of girl-girl family sequences is low compared to the other combinations, our intuition tells us that all combinations are equally likely and should appear in roughly equal proportions. We will assess this intuition with the Central Limit Theorem.

Consider a sequence of 2,444 trials with each of the two-child families. Let $X_i = 1$ (success) if the two-child family is girl-girl, and $X_i = 0$ (failure) if the two-child family is otherwise. We are interested in the probability distribution of

$$S_{2444} = \sum_{i=1}^{2444} X_i.$$

In particular, we are interested in the probability $\mathbb{P}[S_{2444} \leq 530]$, that is, what is the probability of seeing as few as 530 girl-girl families or even fewer in a sample of 2444 families? We can use the Central Limit Theorem to estimate this probability.

We are assuming the family “success” variables X_i are independent, and identically distributed, a reasonable but arguable assumption. Nevertheless, without this assumption, we cannot justify the use of the Central Limit Theorem, so we adopt the assumption. Then $\mu = \mathbb{E}[X_i] = (1/4) \cdot 1 + (3/4) \cdot 0 = 1/4$ and $\mathbb{V}[X_i] = (1/4)(3/4) = 3/16$ so $\sigma = \sqrt{3}/4$. Hence

$$\begin{aligned}\mathbb{P}[S_{2444} \leq 530] &= \mathbb{P}\left[\frac{S_{2444} - 2444 \cdot (1/4)}{(\sqrt{3}/4 \cdot \sqrt{2444})} \leq \frac{530 - 2444 \cdot (1/4)}{(\sqrt{3}/4 \cdot \sqrt{2444})}\right] \\ &\approx \mathbb{P}[Z \leq -3.7838] \\ &\approx 0.0000772\end{aligned}$$

Therefore, we are justified in thinking that under our assumptions, the proportion of girl-girl families is low. It is highly unlikely that under our assumptions such a proportion would have occurred. We then begin to suspect our assumptions, one of which was the implicit assumption that the appearance of girls was equally likely as boys, leading to equal proportions of the four types of families. In fact, there is ample evidence that the birth of boys is more likely than the birth of girls.

Illustration 2

We expect the normal distribution to arise whenever the outcome of a situation, results from numerous small additive effects, with no single or small group of effects dominant. Here is another illustration of that principle.

The following is adapted from *An Introduction to Probability Theory and Its Applications, Volume I*, second edition, William Feller, J. Wiley and Sons, 1957, Chapter VII.3(e), page 175.

The Central Limit Theorem can be used to assess risk. Two large banks compete for customers to take out loans. The banks have comparable offerings. Assume that each bank has a certain amount of funds available for loans to customers. Any customers seeking a loan beyond the available funds will cost the bank, either as a lost opportunity cost, or because the bank itself has to borrow to secure the funds to loan to the customer. If too few customers take out loans, then that also costs the bank, since now the bank has unused funds.

We create a simple mathematical model of this situation. We suppose that the loans are all of equal size and for definiteness each bank has funds available for a certain number (to be determined) of these loans. Then suppose n customers select a bank independently and at random. Let $X_i = 1$ if customer i selects bank H with probability $1/2$ and $X_i = 0$ if customers select bank T, also with probability $1/2$. Then $S_n = \sum_{i=1}^n X_i$ is the number of loans from bank H to customers. Now there is some positive probability that more customers will turn up than can be accommodated. We can approximate this probability with the Central Limit Theorem:

$$\begin{aligned}\mathbb{P}[S_n > s] &= \mathbb{P}[(S_n - n/2)/((1/2)\sqrt{n}) > (s - n/2)/((1/2)\sqrt{n})] \\ &\approx \mathbb{P}[Z > (s - n/2)/((1/2)\sqrt{n})] \\ &= \mathbb{P}[Z > (2s - n)/\sqrt{n}]\end{aligned}$$

Now if n is large enough that this probability is less than (say) 0.01, then the number of loans will be sufficient in 99 of 100 cases. Looking up the value in a normal probability table,

$$\frac{2s - n}{\sqrt{n}} > 2.33$$

so if $n = 1000$, then $s = 537$ will suffice. If both banks assume the same risk of sellout at 0.01, then each will have 537 for a total of 1074 loans, of which 74 will be unused. In the same way, if the bank is willing to assume a risk of 0.20, i.e. having enough loans in 80 of 100 cases, then they would need funds for 514 loans, and if they want to have sufficient seats in 999 out of 1000 cases, they should have 549 loans available.

Now the possibilities for generalization and extension are apparent. A first generalization would be to allow the loan amounts to be random with some distribution. Still we could apply the Central Limit Theorem to approximate the demand on available funds. Second, the cost of either unused funds or lost business could be multiplied by the chance of occurring. The total of the products would be an expected cost, which could then be minimized.

16.6 A concentration inequality: Hoeffding's inequality

We know that sample means of iid random variables will, for large sample sizes, “concentrate” around the population mean. A concentration inequality gives a bound on the probability that the sample mean is outside a neighborhood of the population mean. Often, concentration inequalities are the key to proving limit theorems and even some finite-sample results in statistics and machine learning.

Here we prove a famous but relatively simple concentration inequality for sums of independent bounded random variables. By “bounded random variables” we mean X_i such that $\mathbb{P}(a_i \leq X_i \leq b_i) = 1$. For one thing, boundedness implies existence of moment generating functions. We start with a simple result for one bounded random variable with mean zero; the proof uses some properties of convex functions.

Proposition 16.6.1 Let X be a random variable with mean zero, bounded within the interval $[a, b]$. Then the moment generating function $M_X(t) = \mathbb{E}(e^{tX})$ satisfies

$$M_X(t) \leq e^{t^2(b-a)^2/8}.$$

Proof:

Write $X = Wa + (1 - W)b$, where $W = (X - a)/(b - a)$. The function $z \mapsto e^{tz}$ is convex, so we get

$$e^{tX} \leq We^{ta} + (1 - W)e^{tb}.$$

Taking expectation, using the fact that $\mathbb{E}(X) = 0$, gives

$$M_X(t) \leq -\frac{a}{b-a}e^{ta} + \frac{b}{b-a}e^{tb}.$$

The right-hand side can be rewritten as $e^{h(\zeta)}$, where

$$\zeta = t(b-a) > 0, \quad h(z) = -cz + \log(1 - c + ce^z), \quad c = -a/(b-a) \in (0, 1).$$

Obviously, $h(0) = 0$; similarly, $h'(z) = -c + ce^z/(1 - c + ce^z)$, so $h'(0) = 0$. Also,

$$h''(z) = \frac{c(1-c)e^z}{(1-c+ce^z)^2}, \quad h'''(z) = \frac{c(1-c)e^z(1-c-ce^z)}{(1-c+ce^z)^3}.$$

It is easy to verify that $h'''(z) = 0$ iff $z = \log(\frac{1-c}{c})$. Plugging this z value in to h'' gives $1/4$, and this is the global maximum. Therefore, $h''(z) \leq 1/4$ for all $z > 0$. Now, for some $z_0 \in (0, \zeta)$, there is a second-order Taylor approximation of $h(\zeta)$ around 0:

$$h(\zeta) = h(0) + h'(0)\zeta + h''(z_0)\frac{\zeta^2}{2} \leq \frac{\zeta^2}{8} = \frac{t^2(b-a)^2}{8}.$$

Plug this bound in to get $M_X(t) \leq e^{h(\zeta)} \leq e^{t^2(b-a)^2/8}$.

Proposition 16.6.2 — Chernoff. For any random variable X , $\mathbb{P}(X > x) \leq \inf_{t>0} e^{-tx}\mathbb{E}(e^{tX})$.

Proof: See Exercise ??

Now we are ready for the main result, Hoeffding's inequality. The proof combines the results in the two previous lemmas.

Theorem 16.6.3 — Hoeffding's inequality. Let Y_1, Y_2, \dots be independent random variables, with $\mathbb{P}(a \leq Y_i \leq b) = 1$ and mean μ . Then

$$\mathbb{P}(|\bar{Y}_n - \mu| >) \leq 2e^{-2n^2/(b-a)^2}.$$

Proof: We can take $\mu = 0$, without loss of generality, by working with $X_i = Y_i - \mu$. Of course, X_i is still bounded, and the length of the bounding interval is still $b - a$. Write

$$\mathbb{P}(|\bar{X}_n| >) = \mathbb{P}(\bar{X}_n >) + \mathbb{P}(-\bar{X}_n >).$$

Start with the first term on the right-hand side. Using Lemma 16.6.2,

$$\mathbb{P}(\bar{X}_n >) = \mathbb{P}(X_1 + \dots + X_n > n) \leq \inf_{t>0} e^{-tn} M_X(t)^n,$$

where $M_X(t)$ is the moment generating function of X_1 . By Lemma 16.6.1, we have

$$\mathbb{P}(\bar{X}_n >) \leq \inf_{t>0} e^{-tn} e^{nt^2(b-a)^2/8}.$$

The minimizer, over $t > 0$, of the right-hand side is $t = 4/(b-a)^2$, so we get

$$\mathbb{P}(\bar{X}_n >) \leq e^{-2n^2/(b-a)^2}.$$

To complete the proof, apply the same argument to $\mathbb{P}(-\bar{X}_n >)$, obtain the same bound as above, then sum the two bounds together.

There are lots of other kinds of concentration inequalities, most are more general than Hoeffding's inequality above. Exercise ?? walks you through a concentration inequality for normal random variables and a corresponding strong law. Modern work on concentration inequalities deals with more advanced kinds of random quantities, e.g., random functions or stochastic processes.



17. Esperança Condisional e Aproximação

17.1 Introdução

Vimos na seção 12.13 e na seção 12.14 as definições de esperança e variância condicional no caso discreto. O caso contínuo foi tratado nas seções 12.19 e 12.20.

17.1.1 The conditional expectation $\mathbb{E}(Y|X = x)$ is called the regression of Y on X .

The conditional expectationconditional expectation of $Y|X$ is

$$\mathbb{E}(Y|X = x) = \int y f_{Y|X}(y|x) dy$$

whereregression the integral is taken over the range of y .

Illustration: Spike count pairs (continued) For the joint distribution of spike counts let us compute $\mathbb{E}(X|Y = 0)$. We previously found $f_{X|Y}(0|0) = .60$, $f_{X|Y}(1|0) = .30$, $f_{X|Y}(2|0) = .10$. Then

$$\mathbb{E}(X|Y = 0) = 0(.6) + 1(.3) + 2(.1) = .5.$$

□

Note that $\mathbb{E}(Y|X = x)$ is a function of x , so we might write $M(x) = \mathbb{E}(Y|X = x)$ and thus $M(X) = \mathbb{E}(Y|X)$ is a random variable. An important result concerning $M(X)$ is often called the law of total expectation.

Theorem: Law of total expectation. Suppose X and Y arelaw of total expectation random variables and Y has finite expectation. Then we have

$$\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(Y).$$

Proof: From the definition we compute

$$\begin{aligned}\mathbb{E}(\mathbb{E}(Y|X=x)) &= \int \left(\int y f_{Y|X}(y|x) dy \right) f_X(x) dx \\ &= \int \int y f_{Y|X}(y|x) f_X(x) dx dy \\ &= \int \int y f_{(X,Y)}(x,y) dx dy \\ &= \int y f_Y(y) dy = \mathbb{E}(Y).\square\end{aligned}$$

There are also the closely-related law of total probability and law of total variance.

Theorem: Law of total probability. Suppose law of total probability X and Y are random variables. Then we have

$$\mathbb{E}(P(Y \leq y|X)) = F_Y(y).$$

Proof: The proof follows a series of steps similar to those in the proof of the law of total expectation. \square

We may also define the conditional variance of $Y|X$

$$\mathbb{V}(Y|X=x) = \int (y - \mathbb{E}(Y|X=x))^2 f_{Y|X}(y|x) dy$$

and then get the following, which has important applications.

Theorem 17.1.1 — Law of total variance. Suppose X and Y are random variables law of total variance and Y has finite variance. Then we have

$$\mathbb{V}(Y) = \mathbb{V}(\mathbb{E}(Y|X)) + \mathbb{E}(\mathbb{V}(Y|X)).$$

Proof: The proof is similar to that of the law of total expectation. \square

■ **Example 17.1** Suponha que uma seguradora possua uma linha de seguros tais como seguros de automóveis. No mercado de seguros, a palavra *sinistro* refere-se a qualquer evento em que o bem segurado (um automóvel, por exemplo) sofre um acidente ou prejuízo material. Seja N o total aleatório de sinistros que ela vai registrar num certo período. Um segurado pode ter mais de um sinistro no período. Estamos simplesmente somando o total de sinistros que aconteceu com todos os segurados.

Cada um dos N sinistros ocasiona uma perda monetária aleatória X . Quando $N = n$, estas perdas são representadas por X_1, X_2, \dots, X_n e $S = \sum_{i=1}^n X_i$. Entretanto, o total de sinistros é aleatório e portanto a perda total no período é igual a

$$S = \sum_{i=1}^N X_i = X_1 + X_2 + \dots + X_N$$

Esta soma possui duas fontes de variabilidade. Na primeira delas, os termos X_1, X_2, \dots são aleatórios e ocasionam variabilidade em S . Na segunda fonte, o número de termos que são somados também é aleatório.

A distribuição de X é bem conhecida a partir das perdas individuais. Um histograma das milhares de perdas que a seguradora teve de indenizar em períodos passados fornece uma boa base empírica para estimar a distribuição de X e, em particular, $\mathbb{E}(X) = \mu_X$ e $\mathbb{V}(X) = \sigma_X^2$.

Também a partir do passado, a seguradora sabe que, com o total de clientes que ela possui na carteira, ela pode esperar N em torno de certo valor $\mathbb{E}(N)$ e com certa variância $\mathbb{V}(N)$. Por

exemplo, talvez a seguradora já tenha estudado o número de perdas e descoberto que ela segue uma distribuição de Poisson com média proporcional ao número de clientes. Em resumo, de alguma forma a seguradora tem uma boa estimativa para $\mathbb{E}(N)$ e $\mathbb{V}(N)$. Vamos também assumir que N e as perdas individuais X_1, X_2, \dots são v.a.'s independentes. Se por acaso no período acontecer um número N de sinistros maior que o valor esperado $\mathbb{E}(N)$ isto não vai afetar a distribuição dos X_i 's individuais. As perdas individuais não tenderão a ser maiores que sua média, por exemplo.

A seguradora tem interesse em conhecer o comportamento probabilístico de $S = (X_1 + \dots + X_N)$. Em particular, queremos $\mathbb{E}(S)$ e $\mathbb{V}(S)$, a esperança e variância de S . A seguradora quer saber como as características das perdas individuais ($\mathbb{E}(X) = \mu_X$ e $\mathbb{V}(X) = \sigma_X^2$) e do número de perdas ($\mathbb{E}(N)$ e $\mathbb{V}(N)$) se combinam para gerar $\mathbb{E}(S)$ e $\mathbb{V}(S)$.

Para apreciar como a presença das duas fontes de variabilidade complicam as coisas, vamos pensar inicialmente em $\mathbb{E}(S) = \mathbb{E}(X_1 + \dots + X_N)$. Como a esperança soma de v.a.'s é a soma das esperanças das v.a.'s, um primeiro impulso é pensar que esta regra pode ser aplicada aqui e escrever:

$$\mathbb{E}(S) = \mathbb{E}(X_1 + \dots + X_N) = \mathbb{E}(X_1) + \dots + \mathbb{E}(X_N)$$

Entretanto, esta propriedade não é válida aqui. Ela vale quando somamos um número fixo de termos, não um número aleatório. Note, por exemplo, como não sabemos quantos termos estamos somando no lado direito da expressão acima.

Entretanto, usando a esperança iterada, é fácil obter $\mathbb{E}(S)$. Suponha que $N = n$. Neste caso, $\mathbb{E}(S|N = n) = \mathbb{E}(X_1 + \dots + X_n|N = n)$. Agora o número de termos na soma está fixado (em n) e podemos usar a propriedade de esperança de somas de v.a.'s: $\mathbb{E}(X_1 + \dots + X_n|N = n) = \mathbb{E}(X_1|N = n) + \dots + \mathbb{E}(X_n|N = n)$. Como os X_i 's são independentes de N e como $\mathbb{E}(X_i|N = n) = \mu_X$ para todo i , temos $\mathbb{E}(S|N = n) = n\mu_X$. Portanto, para um N aleatório, não especificado, teremos $\mathbb{E}(S|N) = N\mu_X$. Não deixe passar batido que esta última expressão é uma v.a. Isto é, $\mathbb{E}(S|N)$ é uma v.a. O termo N aparecendo nesta expressão é uma v.a. e é por isto que podemos tomar a esperança da v.a. $\mathbb{E}(S|N)$:

$$\mathbb{E}[\mathbb{E}(S|N)] = \mathbb{E}[N\mu_X] = \mu_X \mathbb{E}[N].$$

Um tanto desapontados, concluímos que

$$\mathbb{E}(S) = \mathbb{E}[\mathbb{E}(S|N)] = \mu_X \mathbb{E}[N] = \mathbb{E}[X] \mathbb{E}[N].$$

O desapontamento é porque provavelmente isto é o você que chutaria de cara: o valor esperado de uma soma aleatória de termos X_i i.i.d. é a esperança μ_X dos termos individuais vezes o esperado $\mathbb{E}[N]$ do número de termos. So, no big deal, then. Muito cuidado, muita consideração, muito não pode isto ou aquilo para descobrir no final que o mais intuitivo e direto era o correto desde o início.

Entretanto, este raciocínio simples e direto não funciona em geral e, por exemplo, vai se quebrar no próximo passo, quando vamos calcular $\mathbb{V}(S)$. Como a variância de uma soma de v.a.'s independentes é a soma das variâncias das v.a.'s, de forma similar ao da esperança podemos esperar que a variância da soma aleatória $\mathbb{V}(S)$ seja igual a $\mathbb{E}(N) \sigma_X^2$. Veremos agora que isto é incorreto. A variância correta é um valor maior que este.

Vamos usar a fórmula da decomposição da variância condicionando em N . Já obtivemos $\mathbb{E}(S|N) = N\mu_X$. Vamos agora obter $\mathbb{V}(S|N)$. Como antes, escolhendo um valor arbitrário *mas fixo*, temos

$$\mathbb{V}(S|N = n) = \mathbb{V}(X_1 + \dots + X_n|N = n) = \mathbb{V}(X_1|N = n) + \dots + \mathbb{V}(X_n|N = n).$$

pois os X_i 's são independentes de N . Temos $\mathbb{V}(X_i|N = n) = \sigma_X^2$ para todo i . Assim, $\mathbb{V}(S|N = n) = n\sigma_X^2$ e portanto, tomando N aleatório, $\mathbb{V}(S|N) = N\sigma_X^2$.

Usando agora a decomposição de variância:

$$\mathbb{V}(S) = \mathbb{V}(\mathbb{E}(S|N)) + \mathbb{E}(\mathbb{V}(S|N)) \quad (17.1)$$

$$= \mathbb{V}(\mu_X N) + \mathbb{E}(\sigma_X^2 N) \quad (17.2)$$

$$= \mu_X^2 \mathbb{V}(N) + \sigma_X^2 \mathbb{E}(N) \quad (17.3)$$

Assim, em relação ao chute incorreto que comentamos antes, $\mathbb{V}(S)$ possui o termo $\mu_X^2 \mathbb{V}(N)$ adicionalmente ao termo já esperado $\sigma_X^2 \mathbb{E}(N)$. Este termo adicional é causado pela natureza aleatória do número N de termos na soma S . Se tivermos um valor $N = n$ fixo, teremos $\mathbb{V}(X_1 + \dots + X_n) = n\sigma_X^2$. Se N for aleatório, não basta substituir n por $\mathbb{E}(N)$ pois $\mathbb{V}(X_1 + \dots + X_N) \neq \mathbb{E}(N)\sigma_X^2$. A variância neste caso é a soma deste termo à direita mais $\mu_X^2 \mathbb{V}(N)$.

Na verdade, existe uma situação em que $\mathbb{V}(X_1 + \dots + X_N) = \mathbb{E}(N)\sigma_X^2$. Isto ocorre quando $\mu_X = 0$ pois então o termo adicional $\mu_X^2 \mathbb{V}(N)$ é zero.

FAZER GRAFICO PARA INTERPRETAR OS DOIS FATORES. Grafico de $S \times N$ com $E(N)$ marcado no eixo x . Variância de S aumentando com N . O primeiro deles é $\mathbb{E}(N)\sigma_X^2$ e representa quanto podemos esperar de variância? O segundo termo é $\mu_X^2 \mathbb{V}(N)$ e representa a variabilidade do que podemos esperar da soma?

■ **Example 17.2** In the spike count pairs illustration, we computed the conditional expectation $\mathbb{E}(X|Y = y)$ for a single value of y . We could evaluate it for each possible value of y . When we consider $\mathbb{E}(X|Y = y)$ as a function of y , this function is called the *regression* of X on Y . Similarly, the function $\mathbb{E}(Y|X = x)$ is called the regression of Y on X . To understand this terminology, and the interpretation of the conditional expectation, consider the case in which (X, Y) is bivariate normal.

figure=../figs.dir/galton-contour.ps,angle=-90,width=2.5in figure=../figs.dir/galton.ps,angle=-90,width=2.5in

Figure 17.1: Conditional expectation for bivariate normal data mimicking Pearson and Lee’s data on heights of fathers and sons. Left panel shows contours of the bivariate normal distribution based on the means, standard deviations, and correlation in Pearson and Lee’s data. The dashed vertical lines indicate the averaging process used in computing the conditional expectation when $X = 64$ or $X = 72$ inches: we average y using the probability $f_Y|X(y|x)$, which is the probability, roughly, in between the dashed vertical lines, integrating across y . In the right panel we generated a sample of 1,078 points (the sample size in Pearson and Lee’s data set) from the bivariate normal distribution pictured in the left panel. We then, again, illustrate the averaging process: when we average the values of y within the dashed vertical lines we obtain the two values indicated by the red x . These fall very close to the least-squares regression line (the solid line).

■ **Example 17.3 — Regression of son’s height on father’s height.** A famous regression data set, from Pearson and Lee (1903) (Pearson, K. and Lee, A. (1903) On the laws of inheritance in man, *Biometrika*, 2: 357–462.), has been used frequently as an example of regression. (See Freedman, Pisani, and Purves (2007).) (Freedman, D., Pisani, R., and Purves, R. (2007) *Statistics*, Fourth Edition, W.W. Norton.) Figure 17.1 displays both a bivariate normal pdf and a set of data generated from the bivariate normal pdf—the latter are similar to the data obtained by Pearson and Lee (who did not report the data, but only summaries of them). The left panel of Figure 17.1 shows the theoretical regression line. The right panel shows the regression based on the data, fitted by the method of least-squares, was discussed briefly in Chapter ?? and will be discussed more extensively in Chapter ???. In a large sample like this one, the least-squares regression line (right panel) is close

to the theoretical regression line (left panel). The purpose of showing both is to help clarify the averaging process represented by the conditional expectation $\mathbb{E}(Y|X = x)$.

The terminology “regression” is illustrated in Figure 17.1 by the slope of the regression line being less than that of the dashed line. Here, $\sigma_Y = \sigma_X$, because the variation in sons’ heights and fathers’ heights was about the same, while $(\mu_X, \mu_Y) = (68, 69)$, so that the average height of the sons was about an inch more than the average height among their fathers. The dashed line has slope $\sigma_Y/\sigma_X = 1$ and it goes through the point (μ_X, μ_Y) . Thus, the points falling on the dashed line in the left panel, for example, would be those for which a theoretical son’s height was exactly 1 inch more than his theoretical father. Similarly, in the plot on the left, any data points falling on the dashed line would correspond to a real son-father pair for which the son was an inch taller than the father. However, if we look at $\mathbb{E}(Y|X = 72)$ we see that among these taller fathers, their son’s height tends, on average, to be less than the 1 inch more than the father’s predicted by the dashed line. In other words, if a father is 3 inches taller than average, his son will likely be *less than* 3 inches taller than average. This is the tendency for the son’s height to “regress toward the mean.” regress toward the mean We understand the phenomenon as follows. First, the father is tall partly for genetic reasons and partly due to environmental factors which pushed him to be taller. If we represent the effect due to the environmental factors as a random variable U , and assume its distribution follows a bell-shaped curve centered at 0, then for any positive u we have $P(U < u) > 1/2$. Thus, if u represents the effect due to environmental factors that the father received and U the effect that the son receives, the son’s environmental effect will tend to be smaller than the father’s whenever the father’s effect is above average. For a tall father, while the son will inherit the father’s genetic component, his positive push toward being tall from the environmental factors will tend to be somewhat smaller than his father’s had been. This is *regression toward the mean*. The same tendency, now in the reverse direction, is apparent when the father’s height is $X = 64$. Regression to the mean is a ubiquitous phenomenon found whenever two variables vary together. \square ■

17.2 Optimal Prediction of $(Y|X = x)$

In general, the regression $\mathbb{E}(Y|X = x)$ could be a nonlinear function of x but in Figure 17.1 it is a straight line. This is not an accident: if (X, Y) is bivariate normal, the regression of Y on X is linear with slope $\rho \cdot \sigma_Y/\sigma_X$. Specifically,

$$\mathbb{E}(Y|X = x) = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X} (x - \mu_X). \quad (17.4)$$

We say that Y has a regression on X with regression coefficient $\beta = \rho \frac{\sigma_Y}{\sigma_X}$. This means that when $X = x$, the *average* value of Y is given by (17.4). We should emphasize, again, that we are talking about random variables, which are *theoretical* quantities, as opposed to observed data. In data-analytic contexts the word “regression” almost always refers to least-squares regression, illustrated in the right panel of Figure 17.1.

Compare Equation (17.4) to Equations (??) and (??). From (??) and (??) we have that the best linear predictor of Y based on X is $f(X)$ where

$$f(x) = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X} (x - \mu_X). \quad (17.5)$$

In general, we may call this the *linear regression* of Y on X . In the case of bivariate normality, the regression of Y on X is equal to the *linear regression* of Y on X , i.e., the regression is linear. We derived (17.5) as the best linear predictor of Y based on X by minimizing mean squared error. More generally, let us write the regression function as $M(x) = \mathbb{E}(Y|X = x)$. It is easy to show that $M(x)$ is the best predictor of Y in the sense of minimizing mean squared error.

Theorem 17.2.1 — Prediction. The function prediction $f(x)$ that minimizes $\mathbb{E}((Y - f(X))^2)$ is the conditional expectation $f(x) = M(x) = \mathbb{E}(Y|X = x)$.

Proof: *Proof Details:* Note that $\mathbb{E}(Y - M(X)) = \mathbb{E}(Y) - \mathbb{E}(\mathbb{E}(Y|X))$ and by the law of total expectation (page 355) this is zero. Now write $Y - f(X) = (Y - M(X)) + (M(X) - f(X))$ and expand $\mathbb{E}((Y - f(X))^2)$ to get

$$\mathbb{E}((Y - f(X))^2) = \mathbb{E}((Y - M(X))^2) + 0 + \mathbb{E}((M(X) - f(X))^2).$$

The right-hand term $\mathbb{E}((M(X) - f(X))^2)$ is always non-negative and it is zero when $f(x)$ is chosen to equal $M(x)$. Therefore the whole expression is minimized when $f(x) = M(x)$. \square

17.2.1 Inspection paradox

Cena de *Quatro Casamentos e Um Funeral* (ou *Four Weddings and One Funeral*, no original): Charles and Carrie conversam após uma noite de amor e Charles pergunta quantos parceiros Carrie teve antes dele: *Well, come on. Tell me. I've seen the dress. We have no secrets.* Bem, ela começa a enumerar. Quando a conta chega a 12, Charles decide interromper sabendo apenas que foi mais que a Princesa Diana e esperando que tenha sido menos que Madonna.

Este é um fenômeno conhecido como o paradoxo da amizade [feld1991yourfriend]. Seja N o número de parceiros sexuais de um indivíduo escolhido ao acaso. Então $\mathbb{E}(N)$ é o número médio de parceiros sexuais nesta população. Para este indivíduo, considere o número de parceiros que cada um dos seus X parceiros já teve: M_1, M_2, \dots, M_N . Tome a média aritmética desses X números: $\bar{M} = (M_1 + \dots + M_N)/N$. Esperamos encontrar $\bar{M} \approx \mathbb{E}(N)$ o que encontramos de fato, na maioria das vezes, é que \bar{M} é significativamente maior que $\mathbb{E}(N)$. Ou seja, provavelmente seus parceiros tiveram, em média, mais parceiros que você.

Isto não está restrito à vida sexual. Numa rede social, pegue um indivíduo ao acaso. Escolhemos Raquel. Verifique quantos amigos Raquel possui. A seguir, calcule o número médio de amigos dos amigos de Raquel. Usualmente, eles terão uma rede de amizade maior que a de Raquel.

Como isto é possível? Isso pode ser explicado como uma forma de amostragem viciada (bias sampling). Os amigos de Raquel não formam uma amostra aleatória de indivíduos da rede social. Na verdade, os indivíduos que possuem muitos amigos têm mais probabilidade de pertencer ao grupo de amigos de um indivíduo escolhido ao acaso. Para explicar isto com clareza, vamos pegar um exemplo extremo, pouco realista, mas que vai mostrar o que está acontecendo.

Considere uma rede social formada por $n + 1$ indivíduos rotulados $i = 1, 2, \dots, n + 1$. O último indivíduo, digamos Paulo, é muito popular, sendo amigo de todos os demais n membros da rede. Os primeiros n indivíduos tem poucos amigos: apenas o super-popular Paulo e mais um dentre os n primeiros. Escolhemos um dos $n + 1$ indivíduos ao acaso (com probabilidade igual a $1/(n + 1)$). Seja N o número de amigos do indivíduo escolhido. Então N é uma v.a. discreta com apenas dois valores possíveis:

$$N = \begin{cases} 2 & \text{com probab } n/(n+1) \\ n & \text{com probab } 1/(n+1) \end{cases}.$$

Isto é, $N = n$ se Paulo for o escolhido, e $N = 2$ se qualquer um dos outros n indivíduos for escolhido. O valor esperado é aproximadamente 3 se n for grande:

$$\mathbb{E}(N) = 2\frac{n}{n+1} + n\frac{1}{n+1} = 3\left(1 - \frac{1}{n+1}\right) \approx 3.$$

Se n não é pequeno, a imensa maioria das vezes em que selecionarmos alguém, veremos $N = 2$. Por exemplo, se $n = 100$ então $N = 2$ será observado com probabilidade aproximadamente 0.99.

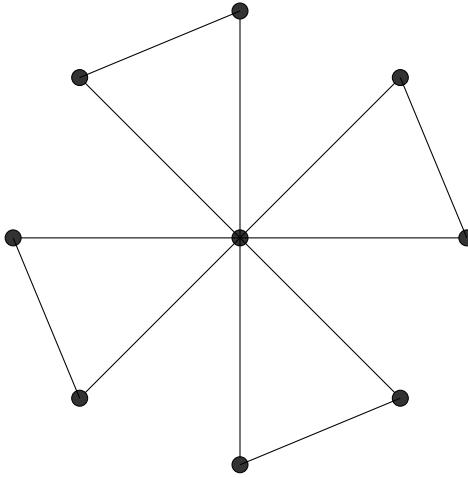


Figure 17.2: Rede social com um super-usuário no centro, amigo de todos os demais.

Considere agora o grupo de amigos dessa pessoa escolhida ao acaso. Se qualquer um dos primeiros n indivíduos for o escolhido, Paulo fará parte desse grupo. Assim, a chance de Paulo entrar no grupo de amigos do indivíduo escolhida ao acaso é igual a $100/101 \approx 0.99$. Por outro lado, supondo que cada indivíduo só possui um único amigo além de Paulo e que este outro indivíduo só possui 2 amigos, temos uma probabilidade igual a $1/(n+1) \approx 0.01$ de qualquer outro indivíduo diferente de Paulo fazer parte do grupo de amigos do indivíduo selecionado ao acaso. Em suma, os indivíduos possuem probabilidades muito distintas de pertencer ao grupo de amigos da pessoa sorteada ao acaso. Paulo tem probabilidade 0.99 e cada um dos n demais indivíduos têm probabilidade 0.01.

Vamos obter a distribuição de probabilidade de \bar{M} , o número médio de amigos da pessoas sorteada. Temos

$$\bar{M} = \begin{cases} \frac{1}{2}(2+n) = \frac{n}{2} + 1 & \text{com probab } n/(n+1) \\ \frac{1}{n}(2n) = 2 & \text{com probab } 1/(n+1) \end{cases}.$$

Dessa forma, na maior parte das vezes em que selecionarmos alguém, teremos $\bar{M} = n/2 + 1 >> 3 \approx \mathbb{E}(N)$. Isto só não vai ocorrer quando Paulo for o indivíduo selecionado, o que acontece com probabilidade $1/(n+1)$.

17.2.2 Inspection paradox

Suppose that n families have children attending a school. Family i has X_i children at the school, where X_1, \dots, X_n are iid random variables, with $\mathbb{P}(X_i = k) = p_k$. Suppose that the average family size is μ .

Now pick a child at random. What is the probability distribution of his family size? Let J be the index of the family from which she comes (which is a random variable). Then

$$\mathbb{P}(X_J = k | J = j) = \frac{\mathbb{P}(J = j, X_j = k)}{\mathbb{P}(J = j)}.$$

The denominator is $1/n$. The numerator is more complex. This would require the j th family to have k members, which happens with probability p_k ; and that we picked a member from the j th family, which happens with probability $\mathbb{E}\left[\frac{k}{k + \sum_{i \neq j} X_i}\right]$. So

$$\mathbb{P}(X_J = k | J = j) = \mathbb{E}\left[\frac{nkp_k}{k + \sum_{i \neq j} X_i}\right].$$



Figure 17.3: Inspection paradox. Escolha uma criança ao caso numa escola. Ela tem maior chance de vir de uma família grande e portanto o valor esperado do tamanho de sua família é maior que a família média.

Note that this is independent of j . So

$$\mathbb{P}(X_J = k) = \mathbb{E} \left[\frac{nk p_k}{k + \sum_{i \neq j} X_i} \right].$$

Also, $\mathbb{P}(X_1 = k) = p_k$. So

$$\frac{\mathbb{P}(X_J = k)}{\mathbb{P}(X_1 = k)} = \mathbb{E} \left[\frac{nk}{k + \sum_{i \neq j} X_i} \right].$$

This is increasing in k , and greater than 1 for $k > \mu$. So the average value of the family size of the child we picked is greater than the average family size. It can be shown that X_J is stochastically greater than X_1 .

This means that if we pick the children randomly, the sample mean of the family size will be greater than the actual mean. This is since for the larger a family is, the more likely it is for us to pick a child from the family.

Part Two

18	Régressão Linear	365
18.1	Introdução	
19	Régressão Logística	367
19.1	Introdução	
20	Regularização e Fatores Latentes	369
20.1	Introdução	
21	Estimador de Máxima Verossimilhança 371	
21.1	Motivação	
21.2	Estimando o tempo médio de sobrevida	
21.3	Quais valores do parâmetro são verossímeis?	
21.4	Resumo informal da máxima verossimilhança	
21.5	Por quê a máxima verossimilhança?	
21.6	Modelos paramétricos	
21.7	Estimador de máxima verossimilhança	
21.8	Obtendo o EMV	
21.9	EMV: soluções analíticas	
22	GLM	391
22.1	Introdução	
23	Teoria de Estimação	393
23.1	Outros métodos de estimação	
23.2	Estimadores são variáveis aleatórias	
23.3	Comparando dois estimadores	
23.4	Estimação Pontual	
24	Algoritmo EM	403
24.1	Misturas de distribuições: introdução	
24.2	Misturas de distribuições: formalismo	
24.3	Estimando uma distribuição de mistura	
24.4	Dados e rótulos	
24.5	O algoritmo EM	
24.6	Exemplos de uso do algoritmo EM	
24.7	Convergência do algoritmo EM	
25	Testes de Hipótese	419
25.1	Introdução	
26	Seleção de Modelos	421
26.1	Introdução	
	bibliography	427
	Books	
	Articles	



18. Regressão Linear

18.1 Introdução

Neste capitulo, apresentar apenas o MODELO de regressao linear com regressors tipo x^2 e categoricas. Dizer que UM metodo de estimacao dos parametros é o metodo de minimos quadrados: projecao linear. Coincide com o EMV, a ser explicado nos capitulos 21 e 23. Por ora, apenas saiba que este metodo está embutido na funcao glm (ou lm) do R. Como usar lm. Como interpretar os parametros.



19. Regressão Logística

19.1 Introdução

Neste capítulo, apresentar apenas o MODELO de regressao logistica e dizer que UM metodo de estimacao dos parametros é o EMV. Ele será explicado nos capitulos 21 e 23. Por ora, apenas saiba que est metodo está embutido na funcao `glm` do R. Como usar `glm`. Como interpretar os parametros.



20. Regularização e Fatores Latentes

20.1 Introdução

Teste, teste,teste



21. Estimador de Máxima Verossimilhança

21.1 Motivação

O método de máxima verossimilhança foi criado por Sir Ronald Fisher, o maior estatístico que já existiu (Figura ??). Ele foi uma espécie de Isaac Newton da estatística, responsável pelos principais conceitos e resultados da inferência estatística. Estes conceitos e resultados constituem a maior parte da estrutura da pesquisa científica em estatística e de suas aplicações até os dias de hoje. Como Newton, ele parece ter tido uma epifania que organizou o trabalho das décadas seguintes, completando 100 anos aproximadamente nos dias de hoje. A Figura ?? mostra como era Sir Ronald Fisher.

Existem muitos *sites* sobre Fisher que, além de estatístico, foi também um dos maiores geneticistas que já existiu. Ao lado de Sewal Wright e Haldane, ele foi responsável por conseguir conciliar a teoria da evolução de Darwin com a genética de Mendel. Um desses excelentes *sites* é

<http://www.economics.soton.ac.uk/staff/aldrich/fisherguide/rafreader.htm>

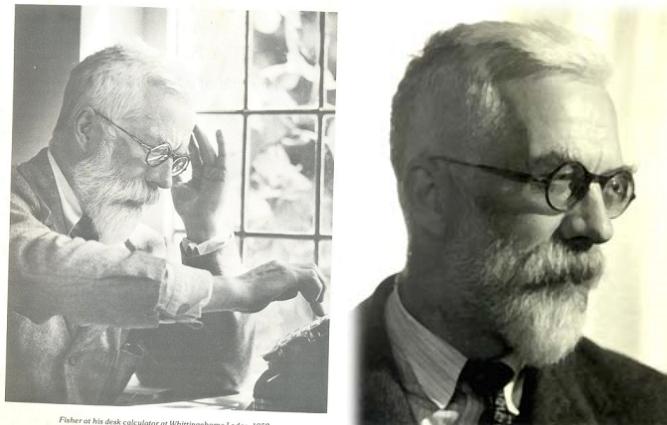


Figure 21.1: Sir Ronald A. Fisher.

Figure 21.2: Os artigos fundamentais de Sir Ronald Fisher, de 1921 (esquerda) e de 1925 (direita).

Suas idéias principais em inferência foram publicadas em dois artigos publicados em 1922 e 1925, intitulados *On the mathematical foundations of theoretical statistics*[12] e *Theory of statistical estimation* [13] (Figura 21.2). Alguns dos principais conceitos (verossimilhança, suficiência e eficiência, por exemplo) e resultados que serão estudados no curso apareceram neste último artigo espetacular, publicado quando ele tinha 35 anos de idade.

Uma das idéias mais fundamentais de Fisher é a do estimador de máxima verossimilhança. Para explicar o que é este estimador, vamos apresentá-lo de maneira bastante informal no contexto de um exemplo.

21.2 Estimando o tempo médio de sobrevida

Glioblastoma multiforme IV é o mais agressivo tipo de câncer do cérebro. Este é um câncer devastador e o tempo de vida após o diagnóstico é curto. Suponha que, usando o tratamento cirúrgico e terapêutico padrão nestes casos, o tempo médio de sobrevida seja de 12 meses. Uma inovação médica parece promissora mas é muito mais cara. Como não existe a certeza de que o novo tratamento seja realmente melhor que o anterior, existem também restrições éticas quanto a sua adoção indiscriminada. Tanto as seguradoras de saúde quanto os pacientes e médicos envolvidos precisam tomar uma decisão mais bem informada sobre a adoção do novo tratamento em substituição ao antigo.

Suponha que X_1, \dots, X_n sejam os tempos de vida em meses de n indivíduos após o novo tratamento cirúrgico. Suponha também que elas sejam variáveis aleatórias i.i.d. com distribuição contínua. O interesse é em fazer inferência sobre o valor esperado de X_i . Este é chamado de tempo médio de sobrevida após o diagnóstico. Isto é, queremos fazer inferência sobre $E(X_i) = \mu$. Se μ for maior que 12 meses, o novo procedimento deveria ser considerado atentamente. Para decidir se μ é maior que 12, vamos estimar μ a partir dos dados da amostra usando a sua média aritmética: $\bar{X} = (X_1 + \dots + X_n)/n$. Se a amostra é grande $\bar{X} \approx \mathbb{E}(X_i) = \mu$. Isto é garantido pelo teorema da Lei dos Grandes Números (LGN).

Qual a natureza de \bar{X} ? É uma constante? É uma função matemática? É uma v.a.? X_1, X_2, \dots, X_n são v.a.'s iid com $\mathbb{E}(X_i) = \mu$ e $\mathbb{V}(X_i) = \sigma^2$. Portanto, sua combinação na estatística $\bar{X} = (X_1 + \dots + X_n)/n$, é também uma v.a. Em cada amostra particular, ela fica instanciada num número específico. Alguns números são mais prováveis que outros. Qual é o valor esperado da v.a. \bar{X} ? Temos $\mathbb{E}(\bar{X}) = \mu$, o valor esperado populacional. Assim, \bar{X} oscila em torno do mesmo valor esperado que as observações individuais. Às vezes, um pouco acima de μ , à vezes, um pouco

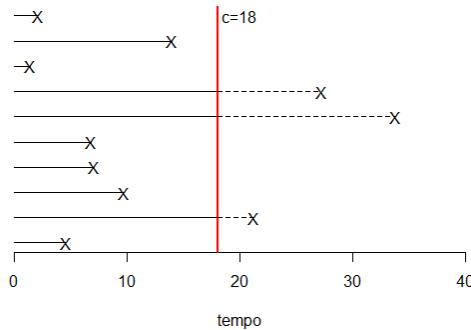


Figure 21.3: Uma amostra particular de $n = 10$ tempos de vida x_i . Três valores são maiores que 18 e portanto não são observados até o fim. Sabe-se apenas que, nestes casos, $x_i \geq 18$.

abaixo. Quanto de variação temos? Qual a variância da v.a. \bar{X} ? Temos $\mathbb{V}(\bar{X}) = \sigma^2/n$, a variância populacional reduzida pelo divisor n , o tamanho da amostra. Assim, \bar{X} varia muito menos em torno de μ que as observações originais. Mais ainda, pela LGN, sabemos que $\bar{X} \rightarrow \mu$.

Assim, checar o valor de \bar{X} dá uma boa base para uma tomada de decisão acerca do valor de μ , principalmente se a amostra é grande. Entretanto, para obter \bar{X} é necessário esperar que todos os indivíduos da amostra faleçam para conhecermos todos os valores X_1, \dots, X_n e isto pode demorar um longo tempo. Se o novo tratamento não for melhor nem pior que o tratamento padrão, podemos ter, por exemplo, $\mathbb{P}(X_i > 36 \text{ meses}) = 0.10$. Numa amostra de 100 indivíduos, 3 anos após o início dos estudos teríamos aproximadamente 10 pacientes ainda vivos.

Nem sempre é possível esperar tanto tempo. Os diversos interessados na decisão (pacientes, familiares, seguradoras e médicos) precisam tomar decisões, mesmo que sujeitas a revisões posteriores. As decisões precisam ser bem informadas mas não podem esperar tanto tempo pela coleta dos dados. É sempre possível rever decisões errôneas mas, num dado momento, alguma decisão deve ser tomada. E de preferência, ela deve ser uma decisão baseada nas evidências disponíveis no momento.

Uma solução muito comum nos experimentos bioestatísticos é usar a *amostra censurada*. Observamos os pacientes até um tempo limite. Digamos, 18 meses. Para aqueles que viverem mais que o tempo limite, simplesmente anotamos que este evento ocorreu. Isto é, a amostra é composta das variáveis aleatórias Y_1, \dots, Y_n onde Y_i é igual ao tempo de vida X_i , se $X_i < 18$. Caso ocorra o evento $X_i > 18$, então anota-se $Y_i = 18$ como um símbolo de que o evento $[X_i > 18]$ ocorreu. Em notação matemática:

$$Y_i = \min\{X_i, 18\} = \begin{cases} X_i, & \text{se } X_i < 18 \\ 18, & \text{se } X_i \geq 18 \end{cases}$$

A Figura 21.3 mostra os tempos de vida X_i de uma amostra com $n = 10$ indivíduos. Cada indivíduo é representado por uma linha horizontal que inicia-se no tempo 0 e que prolonga-se até o falecimento do indivíduo. O ponto de censura é de 18 meses. Três indivíduos têm seus tempos de vida censurados. Nestes três casos sabe-se apenas que $X_i > 18$ mas não se conhece seu valor exato. Para representar este fato, nós anotamos o valor de Y_i como sendo igual a 18. A Tabela 21.1 apresenta os valores dos tempos de sobrevida x_i de todos os 10 indivíduos, bem como os tempos censurados y_i .

Nosso problema continua sendo estimar o tempo esperado de sobrevida mas agora usando estes dados censurados. Nenhuma resposta vêm à mente de imediato. A opção anterior, quando os dados

i	1	2	3	4	5	6	7	8	9	10
x_i	4.6	21.2	9.7	7.1	6.8	33.8	27.2	1.4	14.0	2.1
y_i	4.6	18	9.7	7.1	6.8	18	18	1.4	14.0	2.1

Table 21.1: Dados de tempos de sobrevida x_i de uma amostra de 10 indivíduos e os dados censurados y_i que seriam realmente registrados. O tempo de censura é 18 meses.

não eram censurados, era simplesmente tomar a média aritmética das variáveis aleatórias X_i . O que nós temos agora são as variáveis Y_i que nunca superam o tempo 18. Os maiores tempos de sobrevida (os três valores acima de 18 na Tabela 21.1) são substituídos pelo tempo de censura (18 meses, no exemplo). Portanto, a média aritmética dos tempos censurados y_i será menor que a média aritmética dos tempos não-censurados x_i . Ela tende a subestimar o verdadeiro tempo esperado de sobrevida e por isto ela não é um estimador razoável para μ .

Outra opção poderia ser então ignorar os tempos que foram de fato censurados e tomar a média aritmética apenas dos tempos em que se chegou a observar o falecimento do indivíduo. Esta também não é uma boa idéia. Ela daria um estimador até pior que a média de todos os Y_i pois teríamos apenas os tempos de sobrevida de quem faleceu muito rapidamente.

Fisher fez um raciocínio muito engenhoso que produz um candidato a estimador que parece razoável neste problema. Não só neste problema mas em quase qualquer outro modelo estatístico o mesmo raciocínio pode ser aplicado. Mais surpreendente ainda, este raciocínio gera estimadores imbatíveis num certo sentido. Nenhum estimador pode ser melhor do que aquele que vamos obter aplicando o método criado por Fisher. A bem da verdade, esta afirmação é um tanto exagerada e precisa ser melhor qualificada. Vamos explicar as condições exatas em que a afirmação é válida no Capítulo 23. Entretanto, por enquanto, podemos assumir que, de forma aproximada, os estimadores obtidos através do método de Fisher são os melhores possíveis na maioria das situações práticas e usuais da análise de dados e daí sua popularidade. Isto é mesmo notável: num certo sentido, nada pode ser melhor que o estimador baseado nas ideias de Fisher! O fato é que este raciocínio que vamos expor agora mudou completamente a história da estatística em 1925. Ele transformou o que era um conjunto de ideias e técnicas desconectadas numa ciência, a ciência dos dados.

21.3 Quais valores do parâmetro são verossímeis?

Para aplicar o método de Fisher, precisamos de um modelo de probabilidade para os dados observados. Vamos assumir que os tempos de vida são v.a.'s i.i.d. Na prática, a distribuição de X_i depende da idade, sexo, estágio do tumor no diagnóstico, entre outras variáveis. O comum é que um modelo de regressão GLM (capítulo 22) seja usado em problemas como este. Entretanto, para não complicar desnecessariamente e focar apenas no essencial, vamos assumir que os pacientes são idênticos com relação a todas estas características que poderiam afetar a distribuição de X_i . Isto é, vamos supor que eles tenham a mesma idade, mesmo sexo, mesmo estágio de tumor no diagnóstico, etc. Então as v.a.'s X_i são i.i.d.

Para explicar o método, assuma que $X_i \sim \exp(\lambda)$. O método da máxima verossimilhança requer a escolha de alguma classe de distribuições para os dados. Esta classe depende de parâmetros desconhecidos que serão estimados pelo método. Poderíamos escolher outros modelos comuns na análise de dados de tempo de espera tais como a Weibull (seção ??). Entretanto, vamos ficar um modelo muito simples, a distribuição exponencial com um único parâmetro λ , para focarmos no essencial. Veremos o método de Fisher para este modelo particular mas ele funciona do mesmo modo para qualquer modelo paramétrico que você assuma. Assim, queremos estimar $\mu = \mathbb{E}(X_i)$. O valor de μ está associado com o parâmetro λ da distribuição exponencial pois $\mu = 1/\lambda$. Então, estimar μ é o mesmo que estimar $\lambda = 1/\mu$.

Considere os 10 dados y_1, \dots, y_{10} registrados na amostra censurada e exibidos no gráfico da Figura 21.3:

$$4.6, 18^c, 9.7, 7.1, 6.8, 18^c, 18^c, 1.4, 14.0, 2.1$$

Os três tempos de vida censurados são denotados por 18^c para diferenciá-los dos demais. Fisher percebeu que alguns valores de λ não eram compatíveis com os dados observados. Por exemplo, não parece plausível que $\lambda = 100$ pois neste caso o tempo esperado de sobrevida seria apenas $E(X_i) = 1/100 = 0.01$, um valor evidentemente muito menor que os dados observados, mesmo que estes estejam distorcidos pela censura. Do mesmo modo, os dados observados não dão suporte à afirmação de que $\lambda = 0.01$ pois, neste caso, a esperança de X_i seria $\mu = 1/0.01 = 100$, um valor muito acima da maioria dos tempos de sobrevida observados na amostra.

Estes valores extremos para λ são facilmente descartados. Para tentar ser mais preciso sobre os valores de λ que parecem razoáveis tendo em vista a amostra em mãos, Fisher procurou pensar qual é o princípio lógico que nós usamos para descartar os valores extremos discutidos acima.

Ele imaginou o seguinte: fixado algum valor para o parâmetro desconhecido λ , é possível calcular a chance de observar uma amostra tal como aquela realmente registrada. Por exemplo, o primeiro elemento da amostra foi igual a $y_1 = 4.6$. Supondo que λ é um valor fixo e arbitrário para o parâmetro desconhecido, vamos calcular a probabilidade de Y_1 ser um valor aproximadamente igual ao valor 4.6 realmente observado. Vamos considerar um pequeno intervalo $(4.6 - \Delta/2, 4.6 + \Delta/2)$ de comprimento Δ e centrado em 4.6, e que denotaremos por $(4.6 \pm \Delta/2)$. Como 4.6 está longe da região de censura, temos $Y_i = X_i$ e a probabilidade é igual a

$$\mathbb{P}(Y_1 \in (4.6 \pm \Delta/2)) = \mathbb{P}(X_1 \in (4.6 \pm \Delta/2)) \approx \lambda \exp(-4.6\lambda)\Delta,$$

onde aproximamos a probabilidade pela área do retângulo de base Δ e altura igual a $f_\lambda(4.6)$, a densidade da distribuição exponencial no ponto 4.6. De maneira análoga, calculamos a probabilidade para todos os outros elementos da amostra em que o valor registrado foi de fato o tempo de sobrevida.

Passemos agora aos três elementos da amostra que tiveram o valor registrado $y_i = 18^c$. Nós só registramos $y_i = 18^c$ se, e somente se, o valor correspondente x_i tiver sido maior que 18, pois o tempo de sobrevida foi superior ao tempo de censura de 18 meses. Neste caso, a probabilidade de registrarmos $y_i = 18$ é dada por

$$\mathbb{P}(Y_i = 18^c) = \mathbb{P}(X_i \geq 18) = \exp(-18\lambda)$$

Como as variáveis Y_1, \dots, Y_{10} são i.i.d., a probabilidade conjunta de extraírmos uma amostra tal qual a que realmente obtivemos é dada por

$$\mathbb{P}(Y_1 \in (4.6 \pm \Delta/2), Y_2 = 18^c, \dots, Y_{10} \in (2.1 \pm \Delta/2))$$

que é igual ao produto das probabilidades marginais:

$$\mathbb{P}(Y_1 \in (4.6 \pm \Delta/2)) \mathbb{P}(Y_2 = 18^c) \dots \mathbb{P}(Y_{10} \in (2.1 \pm \Delta/2))$$

e esta por sua vez é igual a

$$\begin{aligned} \lambda \exp(-4.6\lambda)\Delta \exp(-18\lambda) \dots \lambda \exp(-2.1\lambda)\Delta &= \lambda^7 \exp(-\lambda(4.6 + 18 + \dots + 2.1))\Delta^7 \\ &= \lambda^7 \exp(-99.7\lambda)\Delta^7 \end{aligned}$$

Note que o expoente da exponencial multiplica λ pela soma 99.7 dos 10 valores registrados de y_i , somando tanto os 3 valores censurados quanto os 7 outros não censurados.

Considerando os dados da amostra como números fixos, a expressão acima é uma função de λ e Δ . O valor de Δ é completamente arbitrário e é escolhido pelo usuário sem relação com o verdadeiro valor do parâmetro ou com os valores dos dados na amostra. Já que sua escolha é subjetiva e arbitrária, vamos denotar a probabilidade anterior por $L(\lambda)\Delta^7$ enfatizando que apenas o primeiro fator é uma função de λ :

$$\begin{aligned}\mathbb{P}(Y_1 \approx 4.6, Y_2 = 18^c, \dots, Y_{10} \approx 2.1) &\approx \lambda^7 \exp(-\lambda(4.6 + 18 + \dots + 2.1))\Delta^7 \\ &= \lambda^7 \exp(-99.7\lambda)\Delta^7 \\ &\equiv L(\lambda)\Delta^7\end{aligned}$$

Isto é,

$$L(\lambda) = \lambda^7 \exp(-99.7\lambda). \quad (21.1)$$

Note que se variarmos λ , o valor de Δ não se altera. Assim, com respeito a λ , o valor de Δ é uma constante.

Para diferentes valores de λ teremos valores diferentes da probabilidade aproximada $L(\lambda)\Delta^7$ de obter uma amostra tal como a que realmente obtivemos. Isto fica claro na Figura 21.4. Nela, temos um gráfico de $L(\lambda)$ versus λ para valores de λ entre 0 e 0.2. O que podemos ver é que, para valores tais como $\lambda > 0.15$, a probabilidade de obter a amostra é praticamente zero.

Observe também *não estamos* dizendo que a probabilidade de λ ser maior que 0.15 é praticamente zero. É uma diferença sutil. Estamos dizendo que: caso λ seja maior que 0.15, os dados que acabamos de receber como amostra constituem um fenômeno raríssimo. Escrevendo de maneira bem informal, estamos calculando

$$\mathbb{P}(\text{termos os dados observados} | \lambda = 0.15) \approx 0,$$

e não

$$\mathbb{P}(\lambda = 0.15 | \text{temos os dados observados}) = ??,$$

que não sabemos calcular.

A verossimilhança $L(\lambda)$ não fornece a probabilidade de λ ser o valor verdadeiro do parâmetro. O parâmetro λ é desconhecido mas não é aleatório. Ele é fixo e desconhecido, pode ser um valor qualquer positivo mas não existem probabilidades associadas com esses diversos valores. O raciocínio correto é o seguinte: caso λ seja algum valor maior que 0.15, a amostra que realmente temos em mãos é um evento com probabilidade muito baixa de acontecer. Isto é, se λ é algum valor maior que 0.15 um evento muito raro (a amostra em mãos) acabou de ocorrer.

O mesmo se dá com valores de $\lambda < 0.01$. Estes são também valores de λ para os quais a amostra que temos em mãos tem uma probabilidade muito baixa de ocorrência.

Fisher dizia que estes valores tão extremos para λ não são *verossímeis*. A palavra verossímil é formada pelos radicais *vero* (de verdadeiro, real, autêntico) e *símil* (de semelhante, similar). De acordo com os dicionários, algo é verossímil se parece verdadeiro, se não repugna à verdade, se é semelhante à verdade, se é coerente o suficiente para se passar por verdade. Portanto, ao dizer que algo é verossímil, não dizemos que é verdadeiro mas que parece verdadeiro pois está de acordo com todas as evidências disponíveis. A notação $\ell(\theta)$ é devido à palavra (*likelihood*), que significa verossimilhança em inglês.

Fica no ar a questão: devem existir outros valores de λ que, caso sejam os verdadeiros valores do parâmetro, a ocorrência dos dados observados não seja um evento tão improvável.

Compare dois valores de λ quanto a sua verossimilhança, quanto a sua suposta veracidade, levando em conta os dados que foram observados. Por exemplo, vamos comparar os valores de

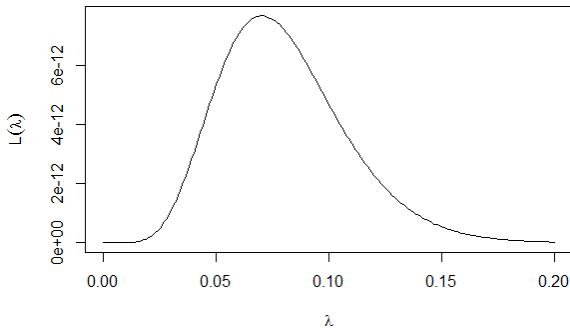


Figure 21.4: Gráfico da função $L(\lambda) = \lambda^7 \exp(-99.7\lambda)$ versus λ .

$\lambda = 0.06$ e $\lambda = 0.15$ para ver se um deles é mais verossímil que o outro. Para isto, vamos calcular a razão entre $L(0.06)$ e $L(0.15)$:

$$\frac{L(0.06)}{L(0.15)} = \frac{0.06^7 \exp(-99.7 * 0.06)}{0.15^7 \exp(-99.7 * 0.15)} = 12.92$$

O que isto quer dizer? Quando $\lambda = 0.06$, a probabilidade de obter uma amostra como a que realmente obtivemos é quase 13 vezes maior que a mesma probabilidade quando $\lambda = 0.15$. Neste sentido, o valor $\lambda = 0.06$ é mais verossímil que o valor $\lambda = 0.15$. Ambos podem ser considerados como valores possíveis para λ mas os dados da amostra podem ocorrer com probabilidade muito maior quando $\lambda = 0.06$ do que quando $\lambda = 0.15$. Se temos que inferir sobre o verdadeiro valor de λ com base nesta amostra, porquê alguém iria preferir $\lambda = 0.15$ a $\lambda = 0.06$?

É neste sentido que Fisher dizia que valores de λ maiores que 0.15 ou menores que 0.01 eram inverossímeis. A probabilidade de obter uma nova amostra similar a que temos em mãos e que foi realmente observada é muito maior quando $\lambda \in (0.05, 0.10)$ do que se λ estiver fora desse intervalo.

A idéia então é acompanhar os valores $L(\lambda)$ à medida que λ varre o espaço paramétrico. Quanto maior o valor de $L(\lambda)$, mais verossímil o valor correspondente de λ . O valor de λ que leva ao valor máximo de $L(\lambda)$ é chamado de estimativa de *máxima verossimilhança*.

Pela Figura 21.4, vemos que o valor de λ que vai maximizar $L(\lambda)$ está por volta de 0.075. Podemos obter analiticamente este máximo. Basta derivar $L(\lambda)$ com respeito a λ e igualar a zero:

$$\frac{dL(\lambda)}{d\lambda} = \frac{d}{d\lambda} (\lambda^7 \exp(-99.7\lambda)) = 7\lambda^6 e^{-99.7\lambda} - 99.7\lambda^7 e^{-99.7\lambda} = 0$$

o que implica em

$$7 - 99.7\lambda = 0,$$

cuja solução é $\hat{\lambda} = 7/99.7$. Portanto, uma estimativa para o tempo médio de sobrevida é

$$\frac{1}{\hat{\lambda}} = \frac{99.7}{7} = \frac{10}{7} \frac{99.7}{10} = \frac{10}{7} \bar{Y}.$$

Isto é, a estimativa de máxima verossimilhança do tempo médio pode ser pensada em dois passos: primeiro calcule a média aritmética \bar{Y} de todos os valores da amostra, censurados e não-censurados,

obtendo $99.7/10$. Esta estimativa tende a subestimar o valor verdadeiro e deveríamos aumentá-la. Este é o objetivo do fator $10/7 > 1$ que, multiplicando a média anterior, vai aumentar a estimativa mais para perto do valor verdadeiro.

No caso geral, usando k para representar o número de observações censuradas e $\sum_i Y_i$ para denotar a soma aleatória de todos os n valores registrados (k censurados e $n - k$ não censurados), teremos o seguinte estimador de máxima verossimilhança para o tempo médio de sobrevida:

$$\frac{1}{\hat{\lambda}} = \frac{n}{n-k} \frac{\sum_i Y_i}{n} \quad (21.2)$$

Se não houver nenhuma observação censurada, o estimador é a media aritmética simples $\sum_i Y_i/n$ de todas as observações registradas. Se houver alguma observação censurada, a fração $n/(n-k)$ será maior que 1. Como neste caso $\sum_i Y_i/n$ tende a subestimar o verdadeiro tempo médio de sobrevida, o efeito da fração $n/(n-k)$ em (21.2) é dilatar a subestimativa, possivelmente trazendo-a mais para perto do verdadeiro valor que queremos estimar. A fração $n/(n-k)$ aumenta com o número de observações censuradas. Se todas as observações forem censuradas (isto é, se $k = n$), não existe o estimador de máxima verossimilhança.

21.4 Resumo informal da máxima verossimilhança

O método de máxima verossimilhança pode ser aplicado em praticamente toda situação de inferência em que os dados aleatórios sigam um modelo estatístico paramétrico, representado por uma família de distribuições de probabilidade para os dados y indexadas por um vetor de dimensão finita θ . O método pode ser resumido de maneira informal da seguinte maneira:

- Suponha que y_1, \dots, y_n sejam os dados da amostra.
- Usando o modelo estatístico, calcule o valor aproximado da probabilidade de observar os dados da amostra e obtenha a *função de verossimilhança* $L(\theta)$ onde apenas θ pode variar e as v.a.s estão com os seus valores fixados com os dados realmente observados na amostra.
- Obtenha o valor $\hat{\theta}$ que maximiza $L(\theta)$. Este valor é a estimativa de máxima verossimilhança.

Em resumo, o método de máxima verossimilhança encontra o valor $\hat{\theta}$ de θ que é o mais verossímil tendo em vista os dados à mão. O valor $\hat{\theta}$ é aquele em que, aproximadamente, a probabilidade de observar os dados realmente observados é máxima.

21.5 Por quê a máxima verossimilhança?

O método de máxima verossimilhança parece uma idéia engenhosa. Fisher chamou $L(\theta)$ de função de verossimilhança e propôs que seu ponto de máximo $\hat{\theta}$ fosse usado como estimativa de θ . As razões para isto são as seguintes:

- generalidade: o método usado para obter a estimativa $\hat{\theta}$ é muito geral e é útil quando a intuição não conseguir sugerir bons estimadores para θ . Ele pode ser aplicado sempre que houver um modelo estatístico paramétrico. Como veremos a seguir, a função de verossimilhança $L(\theta)$ se reduz à função de densidade conjunta (se os dados forem contínuos) ou à função de probabilidade conjunta (se os dados forem discretos). Assim, é fácil obter $L(\theta)$ e basta maximizá-la em θ .
- Fisher provou um resultado teórico: se a amostra cresce então a estimativa $\hat{\theta}$ converge para o verdadeiro valor θ do parâmetro.
- outro resultado teórico de Fisher: se a amostra cresce então a estimativa $\hat{\theta}$ é aproximadamente não-viciada para θ . Isto é, $\hat{\theta}$ não estará sistematicamente subestimando ou superestimando θ .
- mais um resultado teórico de Fisher, e este é sensacional: qualquer estimador não-viciado ou aproximadamente não-viciado terá um erro de estimativa médio maior que o estimador de

máxima verossimilhança. Assim, ele terá o menor erro possível para o médio de estimação. E isto é válido em praticamente *qualquer* modelo estatístico.

- Finalmente, o estimador de máxima verossimilhança possui distribuição aproximadamente gaussiana, não importa quão complicada seja a sua fórmula. Este é um fato fundamental para podermos construir intervalos de confiança e para realizar teste de hipóteses, como veremos nos capítulos 23 e 25.

Os resultados teóricos foram apresentados acima de maneira informal, não rigorosa. Vamos definir precisamente os conceitos envolvidos (ser viciado, erro de estimação, etc) e veremos as condições em que os resultados são válidos no capítulo 23. No entanto, de forma intuitiva e aproximada, o que importa no momento é que vale a pena considerar com carinho e afeto os estimadores de máxima verossimilhança. Numa quantidade enorme de aplicações importantes nada pode ser melhor que eles. Podemos até ter estimadores tão bons quanto o estimador de máxima verossimilhança, mas não melhores. Isto não será válido sempre mas é válido tão frequentemente que tornou o método de máxima verossimilhança imensamente popular.

21.6 Modelos paramétricos

Definition 21.6.1 — Modelo paramétrico. Seja $\mathbf{Y} = (Y_1, \dots, Y_n)$ um vetor aleatório. Um modelo estatístico paramétrico para este vetor é uma família de distribuições de probabilidade $\mathcal{P}_\theta = \{f(\mathbf{y}; \theta)\}$ indexadas por um vetor de dimensão finita θ pertencente a um conjunto Θ , chamado de espaço paramétrico.

■ **Example 21.1 — Modelo Poisson i.i.d..** Seja $\mathbf{Y} = (Y_1, \dots, Y_n)$ um vetor aleatório composto de v.a.'s i.i.d. com distribuição Poisson(λ). Assim, $\mathcal{P}_\theta = \{f(\mathbf{y}; \theta)\}$ com distribuição

$$f(\mathbf{y}; \theta) = \mathbb{P}(Y_1 = y_1, \dots, Y_n = y_n) = \prod_{i=1}^n \frac{\lambda^{y_i}}{y_i!} e^{-\lambda}$$

indexada pelo parâmetro unidimensional $\theta = \lambda$. O espaço paramétrico Θ é a semireta $(0, \infty)$. ■

■ **Example 21.2 — Modelo Gaussiano i.i.d..** Seja $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$ um vetor aleatório composto de v.a.'s i.i.d. com distribuição $N(\mu, \sigma^2)$. Vamos definir $\theta = (\mu, \sigma^2)$, um vetor bi-dimensional com o espaço paramétrico $\Theta = \mathbb{R} \times (0, \infty)$. Então $\mathcal{P}_\theta = \{f(\mathbf{y}; \theta)\}$ com

$$\begin{aligned} f(\mathbf{y}; \theta) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - \mu)^2}{2\sigma^2}\right\} \\ &= \left[\frac{1}{\sqrt{2\pi\sigma^2}}\right]^n \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2\right\} \end{aligned}$$

■

■ **Example 21.3 — Modelo de regressão linear.** Seja $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$ um vetor aleatório composto de v.a.'s independentes mas não identicamente distribuídas. Temos $Y_i \sim N(\mu_i, \sigma^2)$. O valor esperado μ_i depende de atributos da i -ésima observação: é uma soma ponderada desses atributos. Mais especificamente, com os atributos reunidos num vetor $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ip})$ de dimensão $p+1$, temos $\mu_i = \mathbf{x}'_i \beta = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$. Vamos definir

$$\theta = (\beta, \sigma^2) = (\beta_0, \beta_1, \dots, \beta_p, \sigma^2),$$

um vetor $(p+2)$ -dimensional, com o espaço paramétrico $\Theta = \mathbb{R}^{p+1} \times (0, \infty)$. Assim, o modelo

$\mathcal{P}_\theta = \{f(\mathbf{y}; \theta)\}$ é definido por

$$\begin{aligned} f(\mathbf{y}; \theta) &= \prod_{i=1}^n N(\mathbf{x}'_i \beta, \sigma^2) \\ &= f(\mathbf{y} | \mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip})}{\sigma}\right)^2\right) \end{aligned}$$

■

■ **Example 21.4 — Modelo de regressão logística.** Seja $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$ um vetor aleatório composto de v.a.'s independentes mas não identicamente distribuídas. As v.a.'s Y_i são ensaios de Bernoulli, binárias: $Y_i \sim \text{Bernoulli}(p_i)$ onde $p_i = \mathbb{P}(Y_i = 1)$. A probabilidade de sucesso p_i *não* é a mesma para todas as observações. Algumas têm mais chance de ser sucesso do que outras. A probabilidade p_i varia de observação para observação em função de atributos, regressores ou variáveis independentes, medidos em cada exemplo: $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ip})$, um vetor de dimensão $p+1$

No modelo de regressão logística, assumimos uma forma funcional específica para modelar esta dependência de p_i em função dos atributos: a função logística. Pegue um preditor linear do sucesso da i -ésima observação: $\eta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$. Transforme este preditor com a função logística para cair no intervalo $(0, 1)$, que é a faixa de variação de probabilidades.

$$p_i = \frac{1}{1 + e^{-\eta_i}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip})}} = \frac{1}{1 + e^{-\mathbf{x}'_i \theta}}$$

onde $\theta = (\beta_0, \beta_1, \dots, \beta_p)$ é o vetor $(p+1)$ -dimensional de pesos ou parâmetros desconhecidos, com o espaço paramétrico $\Theta = \mathbb{R}^{p+1}$. Assim, as v.a.'s binárias Y_1, Y_2, \dots, Y_n são independentes e, para cada exemplo, temos o modelo

$$Y_i \sim \text{Bernoulli}(p_i)$$

O modelo $\mathcal{P}_\theta = \{f(\mathbf{y}; \theta)\}$ é especificado por

$$\begin{aligned} f(\mathbf{y}; \theta) &= \prod_{i=1}^n (\mathbb{P}(Y_i = 1)^{y_i} \mathbb{P}(Y_i = 0)^{1-y_i}) \\ &= \prod_{i=1}^n \left(\left(\frac{1}{1 + e^{-\mathbf{x}'_i \theta}} \right)^{y_i} \left(1 - \frac{1}{1 + e^{-\mathbf{x}'_i \theta}} \right)^{1-y_i} \right) \end{aligned}$$

Neste exemplo, o vetor θ que indexa as distribuições coincide com o vetor β .

■

21.7 Estimador de máxima verossimilhança

Suponha que o vetor $\mathbf{Y} = (Y_1, \dots, Y_n)$ seja composto por variáveis aleatórias. Vamos usar a mesma notação tanto para a função de probabilidade conjunta $\mathbb{P}(\mathbf{Y} = \mathbf{y}) = f(\mathbf{y})$ de v.a.'s discretas quanto para a densidade conjunta $f(\mathbf{y})$ de v.a.'s contínuas. Para evitar ficar distinguindo as duas quando o procedimento for o mesmo para os dois casos, discreto e contínuo, vamos chamá-la simplesmente de densidade conjunta. Como as distribuições que nos interessam no caso do método de máxima verossimilhança pertencem a famílias paramétricas indexada por $\theta \in \Theta$, vamos denotar isto explicitamente escrevendo a densidade conjunta como $f(\mathbf{y}; \theta)$.

■ **Definition 21.7.1 — Função de Verossimilhança.** Considere a densidade conjunta $f(\mathbf{y}, \theta)$ do vetor $\mathbf{Y} = (Y_1, \dots, Y_n)$ como uma função de θ para \mathbf{y} fixo. Nós chamamos esta função de *função de verossimilhança* do parâmetro θ e vamos denotá-la por $L(\theta)$.

Se o vetor aleatório \mathbf{Y} é composto por variáveis aleatórias discretas então, a função $L(\theta) = f(\mathbf{y}, \theta)$ é a probabilidade de observar \mathbf{Y} igual ao valor \mathbf{y} *realmente* observado na amostra quando θ é o verdadeiro valor do parâmetro. O valor de $L(\theta)$ é uma medida de quão verossímil é o valor de θ , verossímil no sentido de ser capaz de gerar produzir o vetor observado \mathbf{x} observado. Se θ_1 e θ_2 são dois valores distintos de θ e se $L(\theta_1)$ é k vezes maior que $L(\theta_2)$ então dizemos que θ_1 é k vezes mais verossímil que θ_2 .

No caso de v.a.'s conínuas, por causa dos problemas com conjuntos não-enumeráveis, qualquer instânciação \mathbf{y} tem probabilidade zero. Entretanto, podemos usar intuitivamente o mesmo conceito que no caso discreto. A probabilidade do vetor aleatório \mathbf{Y} ser aproximadamente o vetor \mathbf{y} realmente observado é

$$\mathbb{P}(\mathbf{Y} = \mathbf{y} \pm \Delta) = \mathbb{P}(Y_1 \in (y_1 - \Delta, y_1 + \Delta), \dots, Y_n \in (y_n - \Delta, y_n + \Delta)) = f(\mathbf{y}, \theta)(2\Delta)^n = L(\theta)(2\Delta)^n$$

O fator $(2\Delta)^n$ não envolve o parâmetro θ .

Definition 21.7.2 — Função de Log-Verossimilhança. Chamamos $\ell(\theta) = \log(f(\mathbf{y}, \theta))$ de função de log-verossimilhança do parâmetro θ .

■ **Example 21.5 — Modelo Poisson i.i.d..** Neste modelo, temos $\theta = \lambda$ e a função de verossimilhança

$$L(\lambda) = \prod_{i=1}^n \frac{\lambda^{y_i}}{y_i!} e^{-\lambda} = \frac{\lambda^{\sum_i y_i}}{\prod_i y_i!} e^{-n\lambda}$$

e a função de log-verossimilhança

$$\ell(\lambda) = (\sum_i y_i) \log(\lambda) - n\lambda - \sum_i \log(y_i!)$$

Por exemplo, se $n = 3$ e $y_1 = 1, y_2 = 0$ e $y_3 = 1$ então

$$L(\lambda) = \frac{\lambda^{1+0+1}}{1!0!1!} e^{-3\lambda}$$

e

$$\ell(\lambda) = 2\log(\lambda) - 3\lambda - (\log(1!) + \log(0!) + \log(1!))$$

■ **Example 21.6 — Modelo Gaussiano i.i.d..** Neste modelo, temos a função de verossimilhança

$$L(\theta) = L(\mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi \sigma^2}} \exp\left(-\frac{1}{2} \left(\frac{y_i - \mu}{\sigma}\right)^2\right) \quad (21.3)$$

$$= \left(\sqrt{2\pi \sigma^2}\right)^{-n/2} \exp\left(-\frac{1}{2} \sum_{i=1}^n \left(\frac{y_i - \mu}{\sigma}\right)^2\right) \quad (21.4)$$

$$= \left(\sqrt{2\pi \sigma^2}\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2\right) \quad (21.5)$$

$$(21.6)$$

e a função de log-verossimilhança

$$\ell(\theta) = \ell(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi \sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2$$

Por exemplo, se $n = 4$ e $y_1 = 1.11$, $y_2 = 0.32$ e $y_3 = -1.77$ e $y_4 = 0.91$, com $\sum_i y_i = 1.11 + \dots + 0.91 = 0.57$, então

$$L(\theta) = L(\mu, \sigma^2) = \left(\sqrt{2\pi \sigma^2} \right)^{-2} \exp \left(-\frac{1}{2\sigma^2} ((1.11 - \mu)^2 + (0.32 - \mu)^2 + (-1.77 - \mu)^2 + (0.91 - \mu)^2) \right)$$

e

$$\ell(\theta) = \ell(\mu, \sigma^2) = -2 \log (2\pi \sigma^2) - \frac{1}{2\sigma^2} ((1.11 - \mu)^2 + (0.32 - \mu)^2 + (-1.77 - \mu)^2 + (0.91 - \mu)^2)$$

■

O vetor \mathbf{y} aparece na expressão da verossimilhança $L(\theta)$ mas ele é considerado fixo, como nos exemplos acima. Lembre-se que \mathbf{y} significa o conjunto de valores realmente obtidos em um experimento, os valores realizados do vetor aleatório \mathbf{Y} .

■ **Example 21.7 — Modelo de regressão linear.** Completar

■

■ **Example 21.8 — Modelo de regressão logística.** Completar

■

Definition 21.7.3 — Estimador de máxima verossimilhança. O método de estimação de máxima verossimilhança consiste em encontrar o valor $\hat{\theta}$ dentro do espaço paramétrico Θ que é o mais verossímil em termos de gerar os dados \mathbf{y} da amostra. Isto é, se observamos $\mathbf{Y} = \mathbf{y}$, nós procuramos $\hat{\theta}$ que satisfaça

$$L(\hat{\theta}) = f(\mathbf{y}, \hat{\theta}) = \max_{\theta \in \Theta} L(\theta) = \max_{\theta \in \Theta} f(\mathbf{y}, \theta)$$

■ **Example 21.9 — Modelo Poisson i.i.d..** Suponha que $\mathbf{y} = (y_1, y_2, y_3) = (1, 0, 1)$ no modelo Poisson i.i.d. com $n = 3$. Então pode ser verificado (ver a próxima seção) que tomando $\hat{\lambda} = (y_1 + y_2 + y_3)/3 = 2/3 \approx 0.667$ maximizamos $L(\lambda)$:

$$L(2/3) = \frac{(2/3)^{1+0+1}}{1!0!1!} e^{-32/3} \geq \frac{\lambda^{1+0+1}}{1!0!1!} e^{-3\lambda} = L(\lambda)$$

■

■ **Example 21.10 — Modelo Gaussiano i.i.d..** Neste modelo, temos $\theta = (\mu, \sigma^2)$. Se tivermos apenas $n = 4$ observações com $y_1 = 1.11$, $y_2 = 0.32$ e $y_3 = -1.77$ e $y_4 = 0.91$, então pode ser verificado (ver a próxima seção) que tomando $\hat{\mu}$ como a média aritmética,

$$\hat{\mu} = \bar{y} = \sum_i y_i / 4 = (1.11 + \dots + 0.91) / 4 = 0.1425$$

e $\hat{\sigma}^2$ igual à variância amostral

$$\hat{\sigma}^2 = S^2 = \sum_i (y_i - \bar{y})^2 / 4 = \sum_i (y_i - 0.1425)^2 / 4 \approx 1.3036$$

maximizamos $L(\theta)$:

$$L(\theta) = L(\mu, \sigma^2) \leq L(\bar{y}, S^2) = L(\hat{\mu}, \hat{\sigma}^2) = \left(\sqrt{2\pi 1.3036} \right)^{-2} \exp \left(-\frac{1}{21.3036} ((1.11 - 0.1425)^2 + (0.32 - 0.1425)^2 + (-1.77 - 0.1425)^2 + (0.91 - 0.1425)^2) \right)$$

■

O estimador de máxima verossimilhança $\hat{\theta}$ depende dos dados \mathbf{y} da amostra. No caso Poisson, por exemplo, $\hat{\lambda} = \bar{y} = (y_1 + \dots + y_n)/n$. No caso gaussiano, $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2) = (\bar{y}, S^2)$ onde S^2 é função dos dados \mathbf{y} . Assim, para enfatizar este fato, vamos escrever o estimador de máxima verossimilhança como $\hat{\theta}(\mathbf{y})$.

21.8 Obtendo o EMV

A principal razão para definirmos a função de log-verossimilhança $\ell(\theta)$ é facilitar a obtenção do EMV $\hat{\theta}$. Na maioria dos problemas, a densidade conjunta é o produto de várias funções de θ . A função log transforma produtos em somas e maximizar uma soma de funções é mais simples que maximizar um produto de funções. Como o log é uma função crescente, tomar o log muda a função mas não seu ponto de máximo. Se θ_1 e θ_2 são dois pontos do espaço paramétrico tais que $L(\theta_1) > L(\theta_2) > 0$ então $\log(L(\theta_1)) > \log(L(\theta_2))$ pois a função log é crescente (se $a > b > 0$ então $\log(a) > \log(b)$).

Isto significa que, se existir um ponto de máximo $\hat{\theta}$ tal que $L(\hat{\theta}) \geq L(\theta)$ para todo $\theta \in \Theta$ então

$$\ell(\hat{\theta}) = \log(L(\hat{\theta})) \geq \log(L(\theta)) = \ell(\theta)$$

Assim, procurar o ponto de máximo de $L(\theta)$ pode ser reduzida a procurar o ponto de máximo de $\ell(\theta)$. A Figura 21.5 ilustra a situação num caso em que θ é uni-dimensional.

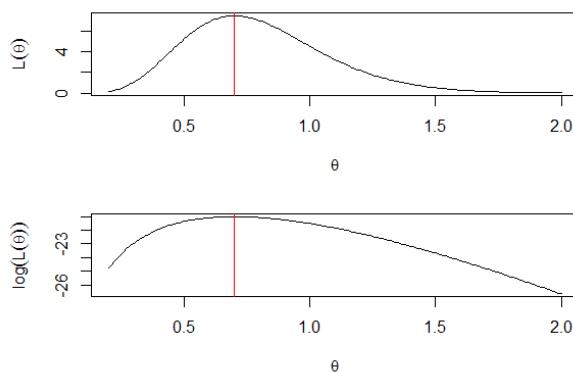


Figure 21.5: Gráfico da função $L(\theta)$ e $\ell(\theta) = \log(L(\theta))$. O ponto de máximo $\hat{\theta}$ é o mesmo nas duas funções embora os valores de $L(\theta)$ e $\ell(\theta) = \log(L(\theta))$ sejam completamente diferentes.

Então problema é, como encontrar o ponto de máximo $\hat{\theta}$ de $\ell(\theta)$? Pontos de máximo e mínimo de funções bem comportadas (com primeira e segunda derivadas contínuas, por exemplo) podem ser usualmente encontrados buscando os pontos críticos da função.

Seja $\theta = (\theta_1, \dots, \theta_k)$ o parâmetro. Queremos $\theta \in \Theta$ que maximize a log-verossimilhança $\ell(\theta)$:

$$\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k) = \arg \max_{\theta} \ell(\theta)$$

Temos uma amostra $= (Y_1, \dots, Y_n)$ composta de variáveis aleatórias discretas ou contínuas com log-verossimilhança $\ell(\theta) = \log f(|)$. Então

Usualmente, o EMV $\hat{\theta}$ é obtido resolvendo-se *simultaneamente* as k equações:

$$\left\{ \begin{array}{l} \frac{\partial \ell(\theta)}{\partial \theta_1} = 0 \\ \frac{\partial \ell(\theta)}{\partial \theta_2} = 0 \\ \vdots \\ \frac{\partial \ell(\theta)}{\partial \theta_k} = 0 \end{array} \right.$$

Representamos o vetor gradiente como o vetor-coluna

$$\nabla \ell(\theta) = \frac{\partial \ell(\theta)}{\partial \theta} = \left(\frac{\partial \ell(\theta)}{\partial \theta_1}, \frac{\partial \ell(\theta)}{\partial \theta_2}, \dots, \frac{\partial \ell(\theta)}{\partial \theta_k} \right)^t$$

Assim, o sistema de equações pode ser escrito de forma vetorial:

$$\nabla \ell(\theta) = \frac{\partial \ell(\theta)}{\partial \theta} = (0, 0, \dots, 0)^t \quad (21.7)$$

Uma solução desse sistema é um *ponto crítico*. Estes sistema

Quando um ponto crítico é um ponto de máximo de $\ell(\theta)$? Para responder a isto, nós olhamos para a matriz de segunda derivada de $\ell(\theta)$ avaliada no ponto $\hat{\theta}$:

$$D^2 \ell(\theta) \Big|_{\theta=\hat{\theta}} = \begin{bmatrix} \frac{\partial^2 \ell(\theta)}{\partial \theta_1^2} & \frac{\partial^2 \ell(\theta)}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \ell(\theta)}{\partial \theta_1 \partial \theta_k} \\ \frac{\partial^2 \ell(\theta)}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \ell(\theta)}{\partial \theta_2^2} & \cdots & \frac{\partial^2 \ell(\theta)}{\partial \theta_2 \partial \theta_k} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 \ell(\theta)}{\partial \theta_k \partial \theta_1} & \frac{\partial^2 \ell(\theta)}{\partial \theta_k \partial \theta_2} & \cdots & \frac{\partial^2 \ell(\theta)}{\partial \theta_k^2} \end{bmatrix} \Big|_{\theta=\hat{\theta}}$$

Verificamos se ela é definida negativa. Isto é, se $\theta^t D^2 \ell(\theta) \theta < 0$ para todo vetor $\theta \neq \mathbf{0}$.

Em alguns exemplos simples, a equação de verossimilhança (21.7) pode ser resolvida algebricamente resultando numa fórmula para $\hat{\theta}(x)$. Veremos alguns desses exemplos na próxima seção (ver seção 21.9). Entretanto, na maioria das situações práticas e mais realistas, é necessário resolver a equação de verossimilhança numericamente usando algum algoritmo de busca de raízes ou de maximização. Veremos alguns exemplos de maximização numérica, incluindo a regressão logística e modelos GLM, na seção ??.

Outros métodos que não dependem do gradiente $\nabla \ell(\theta)$ são necessários se o máximo $\hat{\theta}$ ocorre na fronteira do espaço paramétrico ou quando um dos parâmetros é restrito a um conjunto discreto de valores. Aparecem situações nas quais $\hat{\theta}$ não pode ser obtido através da equação de máxima verossimilhança. Por exemplo, o máximo global da função $\ell(\theta)$ pode ocorrer na fronteira de Θ e este máximo pode não ser raiz de $\ell'(\theta) = 0$ (ver exemplo mais a frente). Similarmente, se Θ é restrito a um conjunto discreto de valores tais como os inteiros então a equação de verossimilhança não se aplica pois neste caso não poderemos derivar $\ell(\theta)$ com relação a θ . Nestes casos, procedimentos especiais devem ser aplicados caso a caso.

21.9 EMV: soluções analíticas

Nesta seção vamos ver em detalhes alguns casos em que soluções exatas, analíticas, sob a forma de fórmulas matemáticas podem ser obtidas. Embora estes sejam casos muito simplificados, eles servem para mortrar algumas propriedades do EMV.

21.9.1 Verossimilhança Bernoulli

Um experimento de Bernoulli é realizado independentemente 10 vezes e observa-se a sequência de resultados. Seja θ a probabilidade de sucesso em um experimento. O verdadeiro valor do parâmetro θ é desconhecido. Sabemos apenas que o espaço paramétrico Θ é o intervalo $[0, 1]$.

Suponha que foi observada o evento-sequência

$$E = (\tilde{C}, C, \tilde{C}, \tilde{C}, C, \tilde{C}, C, \tilde{C}, C, \tilde{C})$$

onde C representa sucesso e \tilde{C} representa fracasso. Podemos representar o experimento através de variáveis aleatórias i.i.d. X_1, \dots, X_{10} com distribuição de Bernoulli com parâmetro θ . A sequência observada é representada pelos dados $\mathbf{x} = (0, 1, 0, 0, 1, 0, 1, 0, 1, 0)$ onde 1 indica C e 0 indica \tilde{C} . A probabilidade da ocorrência de \mathbf{x} , que é também a função de verossimilhança de θ , é dada por :

$$L(\theta) = \mathbb{P}(\mathbf{X} = \mathbf{x}) = \mathbb{P}(\mathbf{X} = (0, 1, 0, 0, 1, 0, 1, 0, 1, 0)) = \theta^4(1 - \theta)^6. \quad (21.8)$$

Esta probabilidade depende do valor desconhecido de θ . Antes do experimento sabíamos apenas que $\theta \in [0, 1]$. Agora, com o experimento realizado, vemos que alguns valores de θ são muito mais verossímeis que outros. O gráfico da Figura 21.6 mostra a função de verossimilhança $L(\theta)$ versus θ .

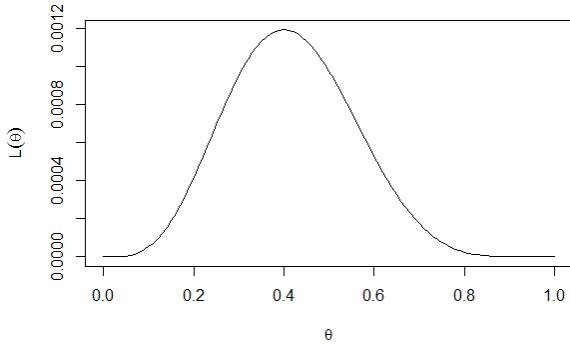


Figure 21.6: Gráfico da função de verossimilhança $L(\theta) = \theta^4(1 - \theta)^6$ versus θ .

Vê-se do gráfico que o valor de θ que maximiza a verossimilhança $L(\theta)$ é $\hat{\theta} = 4/10$, a frequência de sucessos nos dez experimentos. Para verificar isto, vamos obter a função log-verossimilhança:

$$\ell(\theta) = \log L(\theta) = 4\log(\theta) + 6\log(1 - \theta)$$

e a equação de verossimilhança é

$$\ell'(\theta) = \frac{4}{\theta} - \frac{6}{1 - \theta} = 0$$

o que implica na solução $\hat{\theta} = 0.4$. A partir do gráfico, já sabemos que esta solução é um máximo global.

Não precisamos esperar a realização do experimento para saber quem é a estimativa de máxima verossimilhança. Podemos escrever a estimativa de máxima verossimilhança em função dos valores genéricos de x_1, \dots, x_n que vão aparecer como resultado da amostragem. Seja $X_i = 1$ ou $X_i = 0$ caso o i -ésimo lançamento da moeda seja cara ou coroa, respectivamente. Seja $\mathbf{x} = (x_1, \dots, x_{10})$ uma realização do experimento. Então \mathbf{x} é uma lista de 1's e 0's e $k = \sum_{i=1}^{10} x_i$ é igual ao número de caras que ocorreram nesta particular realização do experimento. A probabilidade de ocorrência do evento representado por \mathbf{x} é igual a

$$\mathbb{P}(\mathbf{x}; \theta) = \theta^k(1 - \theta)^{10 - k} = L(\theta)$$

O valor de θ que maximiza a verossimilhança $L(\theta)$ é encontrado facilmente:

$$\frac{d}{d\theta} \log \mathbb{P}(\mathbf{x}; \theta) = \frac{d}{d\theta} (k \log(\theta) + (10 - k) \log(1 - \theta)) = \frac{k}{\theta} - \frac{10 - k}{1 - \theta} = 0$$

o que produz $\hat{\theta} = k/10$. Este é um ponto de máximo pois $L(\theta) = 0$ se $\theta = 0$ ou se $\theta = 1$ e obviamente $L(\theta) > 0$ se $\theta > 0$ e $0 < k < 10$. Portanto, se $0 < k < 10$, tem de existir um ponto de máximo no interior do intervalo. Nos casos extremos em que $k = 0$ (e portanto $L(\theta) = (1 - \theta)^{10}$) ou em que $k = 10$ (e portanto $L(\theta) = \theta^{10}$), a solução $\hat{\theta} = k/10$ ainda é válida.

Assim, em todos os casos, a estimativa de máxima verossimilhança de θ quando o vetor de observações é \mathbf{x} é dado por

$$\hat{\theta} = \frac{k}{10} = \frac{\sum_{i=1}^{10} x_i}{10}$$

, a média aritmética do vetor de 0's e 1's.

Note que $\hat{\theta}$ obtido acima é uma função do vetor \mathbf{x} . Assim, podemos escrever a estimativa de máxima verossimilhança de θ como $\hat{\theta}(\mathbf{x})$ onde \mathbf{x} é o vetor de observações obtido no experimento.

Este exemplo descreve o método de máxima verossimilhança numa situação simples e, de certo modo, frustrante. Ao final de todo o arrazoado, o estimador de máxima verossimilhança é simplesmente o bom e velho \bar{x} , um estimador que todo mundo já teria proposto sem precisar de um conceito elaborado.

Uma situação muito mais interessante foi apresentada na seção inicial desse capítulo que lidou com o caso de observações de tempo de sobrevivência com dados censurados, quando a característica de interesse não possui um estimador óbvio. No entanto, mesmo no caso em que o estimador de máxima verossimilhança coincide com um estimador intuitivo e simples, será reconfortante saber que este estimador simples é também um estimador ótimo. Esta conclusão virá das propriedades do estimador de máxima verossimilhança que serão estudadas no capítulo 23.

Um comentário importante é sobre o caráter da função de verossimilhança. A função $L(\theta)$ é derivada a partir de uma função de probabilidade conjunta e, no caso de variáveis aleatórias discretas, representa uma probabilidade. Entretanto, ela não é a probabilidade de θ ser o verdadeiro valor do parâmetro. Ela fornece uma medida de quão verossímil é cada valor possível do parâmetro θ tendo em vista os dados que foram obtidos.

Considerando o exemplo anterior dos sucesso e fracassos de uma moeda, fica mais claro que $L(\theta)$ não é uma probabilidade nem uma função densidade de probabilidade. A partir de (21.8), sabemos que $L(\theta) = \theta^4(1-\theta)^6$ para $\theta \in (0, 1)$. A integral de $L(\theta)$ no intervalo $(0, 1)$ não é igual a 1. Ela é igual a

$$\int_0^1 L(\theta) d\theta = \int_0^1 \theta^4(1-\theta)^6 d\theta = \frac{6!4!}{11!},$$

resultado que pode ser obtido facilmente a partir da conhecida constante de integração na expressão da densidade de uma distribuição beta com parâmetros $\alpha = 5$ e $\beta = 7$.

21.9.2 Verossimilhança Binomial

Suponha que desejamos estimar θ , a fração de pessoas que tem tuberculose em uma grande população homogênea. Para isto, nós selecionamos aleatoriamente n indivíduos para um teste e encontramos x deles com a doença. Como a população é grande e homogênea, assumimos que os n indivíduos testados são independentes e cada um deles tem probabilidade θ de ter tuberculose. A verossimilhança de θ é então

$$L(\theta) = \mathbb{P}(X = x, \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

onde $0 \leq \theta \leq 1$. O espaço paramétrico é $\Theta = [0, 1]$. Maximizar $L(\theta)$ em θ é o mesmo que maximizar $\theta^x(1-\theta)^{n-x}$ e isto foi feito no exemplo anterior. Portanto, a estimativa de máxima verossimilhança de θ é dada por $\hat{\theta}(x) = x/n$.

21.9.3 Verossimilhança Poisson

Para o monitoramento de risco ambiental, alguns testes de laboratório são realizados em amostras de água de um rio para determinar se a água é está dentro de limites toleráveis para a saúde humana. De particular interesse é a concentração de certos tipos de bactérias na água. O número destas bactérias é determinado em n amostras de volume unitário da água do rio, dando n contagens observadas x_1, x_2, \dots, x_n . O problema é estimar λ , o número médio de bactérias por unidade de volume no rio.

Nós assumimos que as bactérias estão dispersas na água do rio de modo que as localizações das bactérias são pontos aleatórios no espaço seguindo um processo de Poisson. Então a probabilidade de contar x_i bactérias em uma amostra de volume unitário é dada por uma distribuição de Poisson:

$$p(x_i, \lambda) = \frac{\lambda^{x_i}}{x_i!} \exp(-\lambda)$$

onde $x_i = 0, 1, \dots$ e $0 < \lambda < \infty$. Como volumes disjuntos são independentes, a função de verossimilhança de λ para as contagens observadas x_1, x_2, \dots, x_n é dada por

$$L(\lambda) = p(\mathbf{x}, \lambda) = \prod_{i=1}^n p(x_i, \lambda) = \prod_{i=1}^n \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \frac{\lambda^{\sum x_i} e^{-n\lambda}}{x_1! \dots x_n!}$$

A função log-verossimilhança e suas derivadas são as seguintes:

$$\ell(\lambda) = -\log(x_1!, \dots, x_n!) + \left(\sum_{i=1}^{10} x_i \right) \log \lambda - n\lambda$$

$$\frac{d\ell(\lambda)}{d\lambda} = \frac{1}{\lambda} \sum_{i=1}^{10} x_i - n;$$

$$\frac{d^2\ell(\lambda)}{d\lambda^2} = -\frac{1}{\lambda^2} \sum_{i=1}^{10} x_i;$$

Se $\sum_i x_i > 0$, a equação de máxima verossimilhança $\ell'(\theta) = 0$ tem uma única solução $\hat{\lambda}(\mathbf{x}) = \sum_i x_i/n = \bar{x}$. A segunda derivada é negativa neste ponto, indicando que nós temos um máximo relativo. Como $L(0) = 0$ e $L(\lambda) \rightarrow 0$ quando $\lambda \rightarrow \infty$ então $\hat{\lambda}$ é o máximo global.

Se $\sum_i x_i = 0$, temos $L(\lambda) = e^{-n\lambda}$ e $\ell'(\theta) = 0$ não tem solução. Neste caso, o máximo ocorre na fronteira do espaço paramétrico: $\hat{\lambda} = 0 = \sum_i x_i/n$.

Assim, em qualquer caso, nós temos $\hat{\lambda}(\mathbf{x}) = \bar{x}$ sendo a estimativa de máxima verossimilhança de λ .

21.9.4 Verossimilhança Poisson truncada

Usualmente não é possível contar o número Poisson de bactérias em uma amostra de água do rio. Pode-se somente determinar se as bactérias estão ou não presentes na amostra. São incubados e testados n tubos de testes, cada um deles contendo um volume V de água do rio. Um teste negativo mostra que não há bactérias presentes enquanto um teste positivo mostra que existe pelo menos uma bactéria presente no tubo. Se y tubos dentre os n testados dão resultados negativos, qual é a estimativa de máxima verossimilhança de λ ?

O resultado do experimento é $\mathbf{x} = (x_1, \dots, x_n)$ onde $x_i = 1$ ou 0 conforme o teste seja negativo ou positivo, respectivamente. Então $y = \sum_i x_i$.

Supusemos que a contagem do número de bactérias segue uma distribuição de Poisson com λ bactérias, em média, por unidade de volume. Portanto, a probabilidade de que existam k bactérias em um volume V de água do rio segue uma Poisson com parâmetro λV :

$$p(k) = (\lambda V)^k \frac{e^{-\lambda V}}{k!} \quad x = 0, 1, 2, \dots$$

A probabilidade de uma reação negativa (sem bactérias) é $\pi = p(0) = \exp(-\lambda V)$ e a probabilidade de uma reação positiva (no mínimo uma bactéria) é $1 - \pi = 1 - \exp(-\lambda V)$.

Assumindo que volumes disjuntos são independentes, os n tubos de teste constituem ensaios aleatórios independentes. A probabilidade de observar y reações negativas dentre os n tubos testados é igual a

$$L(\lambda) = p(y; \lambda) = \binom{n}{y} \pi^y (1 - \pi)^{n-y} = \binom{n}{y} e^{-y\lambda V} (1 - e^{-\lambda V})^{n-y} \quad (21.9)$$

onde $0 \leq \lambda < \infty$.

Podemos tomar o logaritmo de $L(\lambda)$ encontrando

$$l(\lambda) = \text{cte} - y\lambda V + (n - y) \log(1 - e^{-\lambda V}) \quad (21.10)$$

onde cte é uma constante em termos de λ . Esta função pode ser derivada e igualada a zero produzindo $\hat{\lambda} = -\log(y/n)/V$.

Alternativamente, por (21.9), vemos que a função de verossimilhança é um múltiplo da função $\pi^y (1 - \pi)^{n-y}$. Vimos no exemplo anterior que esta função atinge seu valor máximo quando $\pi = y/n$. Se $y > 0$, o valor correspondente de λ é obtido resolvendo a equação $\pi = e^{-V\lambda}$, que produz $\lambda = -\log(\pi)/V$. Assim, nós obtemos $\hat{\lambda} = -\log(y/n)/V = (\log(n) - \log(y))/V$, se $y > 0$. Observe que, se $y = n$, então $\hat{\lambda} = 0$, no limite do espaço paramétrico.

Caso $y = 0$, a estimativa de máxima verossimilhança não existe. De fato, neste caso, deveríamos ter $0/n = 0 = e^{-V\hat{\lambda}}$, o que implicaria em $\hat{\lambda} = \infty$.

Como ilustração numérica, suponha que $n = 40$ tubos de teste, cada um contendo $V = 10$ ml de água do rio estão sob teste. Se $y = 28$ dão testes negativos e $n - y = 12$ dão testes positivos, então

$$\hat{\lambda} = \frac{\log 40 - \log 28}{10} = 0,0357$$

A concentração de bactérias coliformes por EMV é estimada por 0,0357.

O gráfico da esquerda na Figura 21.7 mostra a função de verossimilhança $L(\lambda)$ dada em (21.9) para λ entre 0.00 e 0.20. O gráfico da direita mostra a função de log-verossimilhança $l(\lambda)$ em (??).

Figure 21.7: Gráfico da função $L(\lambda)$ (a esquerda) e de $l(\lambda)$ (a direita)

Observe que quanto maior a concentração de bactérias no rio, maior a probabilidade de que todos os n tubos dêem resultados positivos. Isto acontece porque a probabilidade de um tubo ser negativo (contagem igual a zero) é dada por $\exp(-\lambda V)$, uma função que decresce com λ .

Isto é relevante porque os dois casos extremos, $y = 0$ e $y = n$, são indesejáveis. O primeiro porque produz $\hat{\lambda} = 0$, uma estimativa baixa demais pois, se $\lambda = 0$, teríamos uma água absolutamente

pura. O segundo caso extremo produz $\hat{\lambda} = \infty$, uma estimativa absurdamente grande. A implicação prática é que devemos escolher um volume V que não seja nem tão pequeno a ponto de gerar apenas tubos negativos (e $y = n$), nem V tão grande que todos os tubos sejam positivos (com $y = 0$). Isto é, se queremos estimar bem o número médio de bactérias por unidade de volume devemos ter tantos tubos positivos quanto tubos negativos na amostra. Isto traz a preocupação prática de escolher um volume V adequado para não terminar com estimativas muito ruins. Usualmente, escolhem-se vários volumes, indo dos volumes maiores, onde sempre há a presença de bactérias, aos volumes menores, onde elas quase nunca estão presentes.



22. GLM

22.1 Introdução

Teste, teste,teste



23. Teoria de Estimação

23.1 Outros métodos de estimação

Aprendemos um método para estimar o vetor de parâmetros θ de um modelo estatístico para os dados $\mathbf{Y} = (Y_1, \dots, Y_n)$: , o método de máxima verossimilhança, que produz o estimador de máxima verossimilhança (MLE). Este método é intuitivo e pode ser usado sempre que pudermos escrever a verossimilhança $L(\theta)$. Mas o MLE não é o único método para estimar parâmetros. Existem vários outros métodos, todos tão intuitivos quanto o MLE. Nem sempre eles coincidem nas suas estimativas de θ . Se eles não coincidem, como escolher um deles? Vamos definir propriedades desejáveis que um bom método deveria satisfazer. Em seguida, vamos verificar que o MLE satisfaz estas propriedades desejáveis. Antes de definir estas propriedades, vamos ver exemplos de diferentes métodos de estimação. Eles são úteis para, por exemplo, fornecer valores iniciais para obter o MLE pelo procedimento de Newton.

23.1.1 Método de Momentos

Já usamos este método informalmente várias vezes antes. O caso inicial mais simples é aquele em que $\mathbf{Y} = (Y_1, \dots, Y_n)$ é formado por variáveis i.i.d. com uma densidade que depende de um único parâmetro. Por exemplo, \mathbf{Y} poderia ser composto por v.a.'s Poisson(θ) ou poderiam ser $N(\theta, 1)$. Como $\mathbb{E}(Y_i) = \theta$, sugerimos usar a média aritmética $\bar{Y} = (Y_1 + \dots + Y_n)/n$ como estimador de $\mathbb{E}(Y_i) = \theta$.

Em geral, o momento teórico $\mathbb{E}(Y_i)$ será uma função matemática $g(\theta)$ do parâmetro θ . Nos casos acima, tivemos $g(\theta) = \theta$ mas podemos ter uma função $g(\theta)$ diferente da função identidade.

■ **Example 23.1 — Pareto e o método de momentos.** Se $\mathbf{Y} = (Y_1, \dots, Y_n)$ é formado por variáveis i.i.d. com uma densidade de Pareto dependente apenas do parâmetro α e dada por

$$f(y) = \begin{cases} \alpha/y^{\alpha+1} & \text{se } y \geq 1 \\ 0, & \text{se } y < 1 \end{cases}$$

Temos $\mathbb{E}(Y_i) = g(\alpha) = \alpha/(\alpha - 1)$, quando $\alpha > 1$. Se $\alpha < 1$, o método de momentos não pode ser aplicado não existe a esperança de Y_i pois a integral $\int xf(x)dx$ não converge. Igualamos

$\mathbb{E}(Y_i) = g(\alpha) = \alpha/(\alpha - 1)$ com a média aritmética \bar{Y} dos dados e invertemos de modo a isolar $\alpha = g^{-1}(\bar{Y})$, obtendo assim o estimador de momentos de α :

$$\alpha/(\alpha - 1) = \bar{Y} \Rightarrow \hat{\alpha}_{MM} = \frac{\bar{Y}}{\bar{Y} - 1}$$

No caso deste exemplo, o MLE é uma expressão muito diferente: $\hat{\alpha}_{MLE} = \frac{n}{\sum_i \log(Y_i)}$, o inverso da média aritmética dos logaritmos dos dados. Qual dos dois estimadores deveria ser escolhido? E por que não fazer combinações deles, tais como uma média ponderada $\hat{\alpha}_{comb} = 0.7 \times \hat{\alpha}_{MLE} + 0.3 \hat{\alpha}_{MM}$ ou $\hat{\alpha}_{comb} = 0.5 \times \hat{\alpha}_{MLE} + 0.5 \hat{\alpha}_{MM}$? ■

Muitas vezes, a distribuição dos dados depende de mais de um parâmetro tais como $N(\mu, \sigma^2)$ ou uma Gama(α, β) onde $\theta = (\mu, \sigma^2)$ e $\theta = (\alpha, \beta)$, respectivamente. A ideia do método de momentos é obter os dois primeiros momentos teóricos, $\mathbb{E}(Y_i)$ e $\mathbb{E}(Y_i^2)$, que serão ambos funções do parâmetro θ . Igualamos estes dois momentos teóricos aos dois primeiros momentos amostrais e resolvemos o sistema de duas equações com duas incógnitas.

■ **Example 23.2 — Gama e o método de momentos.** Se $\mathbf{Y} = (Y_1, \dots, Y_n)$ é formado por variáveis i.i.d. com uma densidade Gama dependente do parâmetro $\theta = (\alpha, \beta)$ e dada por

$$f(y) = \begin{cases} \text{cte. } y^{\alpha-1} e^{-\beta y} & \text{se } y > 0 \\ 0, & \text{caso contrário} \end{cases}$$

A Figura 23.1 mostra um histograma de amostra de $n = 200$ valores retirados de uma distribuição gama com parâmetros α e β . O problema prático de estimação é: tendo escolhido o modelo gama para os dados, usar apenas os dados da amostra para inferir os valores desconhecidos de α e β . Precisamos dos dois primeiros momentos teóricos: $\mathbb{E}(Y_i) = \alpha/\beta$ e $\mathbb{E}(Y_i^2) = \alpha(\alpha + 1)/\beta^2$. Veja que cada um desses momentos teóricos é uma função do vetor de parâmetros $\theta = (\alpha, \beta)$.

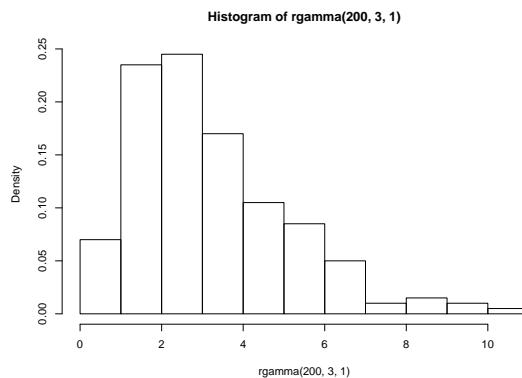


Figure 23.1: Amostra de $n = 200$ valores de Gama($\alpha = 3, \beta = 1$)

Precisamos também dos dois primeiros momentos amostrais, inteiramente baseados nos dados: $m_1 = \sum_i Y_i/n = \bar{Y}$ e $m_2 = \sum_i Y_i^2/n = \bar{Y}^2$. Igualamos os momentos correspondentes, teóricos e amostrais: $\mathbb{E}(Y) = m_1$ e $\mathbb{E}(Y^2) = m_2$. Isto é:

$$\begin{aligned} \frac{\alpha}{\beta} &= m_1 \\ \frac{\alpha(\alpha + 1)}{\beta^2} &= m_2 \end{aligned}$$

Em seguida, resolvemos este sistema de equações não lineares para encontrar a solução $\hat{\alpha}$ e $\hat{\beta}$. Com simples manipulação algébrica, encontramos a solução: $\hat{\alpha} = m_1^2/(m_2 - m_1^2)$ e $\hat{\beta} = m_1/(m_2 - m_1^2)$. Estes são os estimadores de momentos de $\theta = (\alpha, \beta)$.

O que é o MLE neste caso de v.a.'s iid gama? O MLE não dá o mesmo resultado que o método de momentos. A log-verossimilhança é:

$$\ell(\alpha, \beta) = n\beta \log(\alpha) - n \log(\Gamma(\alpha)) + (\alpha - 1) \sum_i \log(y_i) - \beta \sum_i y_i$$

Derivando em relação a α e β , temos:

$$\begin{aligned}\frac{\partial \ell}{\partial \beta} &= \frac{n\alpha}{\beta} - \sum_i y_i \\ \frac{\partial \ell}{\partial \alpha} &= n \log(\beta) - n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_i \log(y_i)\end{aligned}$$

Igualando a zero a derivada em β produz $\hat{\beta} = \hat{\alpha}/\bar{y}$. A seguir, igualando a zero a derivada em α produz

$$n \log(\beta) - n \Gamma'(\alpha)/\Gamma(\alpha) + \sum_i \log(y_i) = 0$$

Substituindo $\beta = \alpha/\bar{y}$ teremos

$$n \log(\alpha) - n \log(\bar{y}) - n \Gamma'(\alpha)/\Gamma(\alpha) + \sum_i \log(y_i) = 0$$

Esta é uma equação com uma única incógnita, α . Precisamos encontrar a solução numericamente. A maioria das linguagens possuem implementada a função $\psi(\alpha) = \Gamma'(\alpha)/\Gamma(\alpha)$, chamada de função digama. Pode-se usar o estimador de momentos como valores iniciais num procedimento de Newton. O ponto importante para nós é que o MLE $\hat{\theta}$ difere do estimador de momentos neste problema. Qual dos dois estimadores deveria ser usado?

Passando para o caso mais geral, suponha que $\mathbf{Y} = (Y_1, \dots, Y_n)$ é formado por variáveis i.i.d. cuja distribuição depende do vetor de parâmetros $\theta \in \mathbb{R}^k$. A ideia do método de momentos é igualar os k primeiros momentos teóricos, que serão funções de θ , aos k momentos amostrais:

Momento Teórico	Momento Amostral
$\mathbb{E}(Y) = g_1(\theta)$	$m_1 = \bar{Y} = \sum_{i=1}^n Y_i/n$
$\mathbb{E}(Y^2) = g_2(\theta)$	$m_2 = \sum_i Y_i^2/n$
\vdots	\vdots
$\mathbb{E}(Y^k) = g_k(\theta)$	$m_k = \sum_i Y_i^k/n$

Resolve-se o sistema de equações não lineares produzindo o estimador de momentos $\hat{\theta}$.

A justificativa do método de momentos é que, pela lei dos grandes números (ver capítulo ??), $m_1 = \sum_i Y_i/n$ converge para $\mathbb{E}(Y)$ quando $n \rightarrow \infty$. Assim, se o tamanho n da amostra é grande, podemos esperar $m_1 \approx \mathbb{E}(Y)$. Pela mesma lei dos grandes números, $m_k = \sum_i Y_i^k/n$ converge para $\mathbb{E}(Y^k)$. Assim, $m_2 \approx \mathbb{E}(Y^2)$, $m_3 \approx \mathbb{E}(Y^3)$, etc. Trocamos a aproximação por uma igualdade para obter a estimativa de momentos.

23.1.2 Um método sem nome

Um terceiro método é baseado na estatística de Kolmogorov. Até onde ue saiba, este método está sendo inventado aqui, neste livro, e tem caráter puramente didático. O objetivo é apenas ilustrar que outros métodos intuitivamente razoáveis para estimar parâmetros são facilmente derivados.

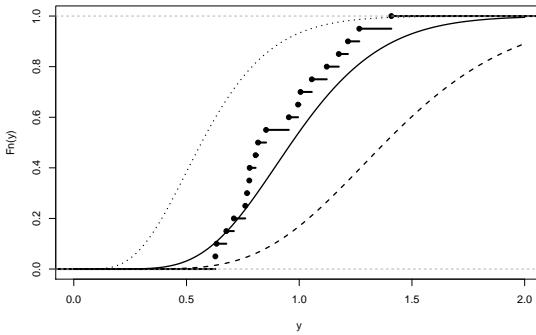


Figure 23.2: Função distribuição acumulada empírica $\mathbb{F}_n(y)$ e três funções de distribuições acumuladas: Gama(10,10) (linha sólida), Gama(10, 7) (linha tracejada) e Gama(6,10) (linha pontilhada).

Considere o modelo de v.a.'s i.i.d. Gama(α, β). Construa a função acumulada empírica com os dados observados. Para cada α e β fixos, podemos calcular a distribuição acumulada teórica de uma Gama(α, β). A ideia então é obter os parâmetros α e β que tornam as duas funções acumuladas, empírica e teórica, o mais próximas possível.

A Figura 23.2 mostra como este método funciona. Nela, vemos um gráfico com a função distribuição acumulada empírica $\mathbb{F}_n(y)$ de uma amostra com $n = 20$ dados (a função escada). Esta função é calculada usando apenas os dados, sem nenhuma referência ou uso de algum modelo teórico. O gráfico mostra também três funções de distribuição acumulada teórica de um modelo gama com diferentes parâmetros α e β : Gama(10,10) (linha sólida), Gama(10, 7) (linha tracejada) e Gama(6,10) (linha pontilhada). É mais ou menos claro que, dentre as três curvas apresentadas, aquela associada com a distribuição Gama(10,10) é a mais próxima da função acumulada empírica. É a melhor possível?

Vamos usar como critério para estimar $\theta = (\alpha, \beta)$ a minimização da distância de Kolmogorov. Para cada valor possível para o parâmetro $\theta = (\alpha, \beta)$, obtenha a função distribuição acumulada teórica $F(y; \theta)$ de uma Gama com parâmetros α e β . A seguir, calcule a distância de Kolmogorov entre esta acumulada teórica $F(y; \theta)$ e a distribuição acumulada empírica $\mathbb{F}_n(y)$:

$$D_n(\mathbf{y}, \theta) = \max_y |\mathbb{F}_n(y) - F(y; \theta)|$$

A distância $D_n(\mathbf{y}, \theta)$ é função dos dados \mathbf{y} e do valor escolhido para θ . Com os dados \mathbf{y} fixos, ela é função apenas de θ . Obtenha agora o valor de $\theta = (\alpha, \beta)$ que minimiza a distância $D_n(\mathbf{y}, \theta)$:

$$\hat{\theta} = (\hat{\alpha}, \hat{\beta}) = \arg \min_{\theta} D_n(\mathbf{y}, \theta)$$

Pois bem, este método é intuitivamente razoável também, não? E no entanto, ele vai dar um resultado diferente do método de momentos e do MLE. Como escolher entre um deles? Ou devemos combiná-los de algum modo tal como, por exemplo, através de uma média ponderada desses três estimadores?

23.1.3 Mais um método sem nome

Lembre-se do modelo de regressão logística com k covariáveis e vetor de coeficientes $\theta = (\beta_0, \beta_1, \dots, \beta_k)$. O i -ésimo indivíduo tem suas características ou covariáveis agrupadas no vetor $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ik})'$. O preditor linear da probabilidade de sucesso do i -ésimo indivíduo é a combinação linear $\eta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} = \mathbf{x}'_i \theta$ das suas características. Este preditor linear

é o que governa a probabilidade de sucesso, dada pelo modelo logístico

$$\begin{aligned}\mathbb{P}(Y_i = 1) &= \frac{1}{1 + e^{-\eta_i}} \\ &= \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}))} \\ &= p_i(\theta)\end{aligned}$$

Já vimos como obter o MLE de $\theta = (\beta_0, \beta_1, \dots, \beta_k)$. Um outro método, simples e intuitivo, é o seguinte. Chute um valor qualquer para $\theta = (\beta_0, \beta_1, \dots, \beta_k)$ e obtenha as probabilidades $p_i(\theta)$ para cada indivíduo. Compare com os dados e ache o valor de θ que mais se ajuste aos dados. Por exemplo, ache o valor de θ que minimize a soma das diferenças (em valor absoluto) entre a resposta binária y_i a probabilidade predita $p_i(\theta)$:

$$\hat{\theta} = \min_{\theta} D(\mathbf{y}, \theta)$$

onde

$$D(\mathbf{y}, \theta) = \sum_i |y_i - p_i(\theta)|.$$

A função $D(\mathbf{y}, \theta)$ é uma medida da discrepância entre os dados \mathbf{y} e o modelo parametrizado por θ .

Para exemplificar este estimador, considere o caso em que temos uma única covariável, como no caso do teste de desenvolvimento infantil de Denver visto no capítulo 19. Na Figura 23.3, nos dois gráficos, vemos os dados (x_i, y_i) de $n = 173$ crianças onde x_i é a idade da i -ésima criança em meses e y_i é a variável binária indicando o sucesso ou fracasso na dessa criança na realização do teste. A curva logística depende do vetor paramétrico $\theta = (\beta_0, \beta_1)$ e é da forma $p(\theta) = 1/(1 + \exp(-\beta_0 - \beta_1 x))$. A Figura 23.3 mostra duas possíveis curvas logísticas: com $\theta = (-6.3, -0.35)$ (à esquerda) e $\theta = (-5.85, -0.39)$ (à direita).

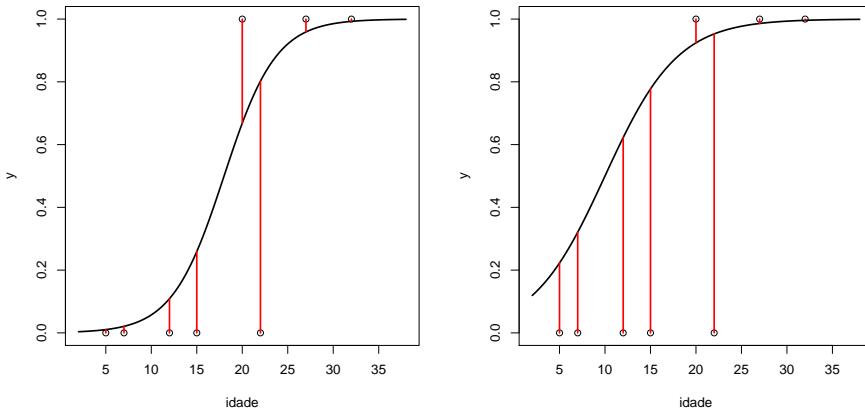


Figure 23.3: Duas possíveis curvas com os valores de $|y_i - p_i(\theta)|$.

A medida de discrepancia $D(\mathbf{y}, \theta) = \sum_i |y_i - p_i(\theta)|$ é a soma das barras verticais conectando a resposta y_i de cada indivíduo com a sua probabilidade de sucesso $p_i(\theta)$. Estas barras serão pequenas quando $y_i = 1$ e $p_i(\theta) \approx 1$ ou então quando $y_i = 0$ e $p_i(\theta) \approx 0$. Nos casos em que existe grande discrepancia entre os dados e a predição do modelo com um valor especificado para o parâmetro θ teremos muitos indivíduos com $y_i = 0$ e $p_i(\theta) \approx 1$ ou então com $y_i = 1$ e $p_i(\theta) \approx 0$. O

estimador do parâmetro θ é encontrado obtendo o valor $\hat{\theta}$ que minimiza a medida de discrepância $D(\mathbf{y}, \theta)$.

Podemos inventar muitos critérios razoáveis neste problema. Por exemplo:

- Mudando ligeiramente a medida de discrepancia para considerar as diferenças ao quadrado: encontre $\hat{\theta}$ que minimize $D(\mathbf{y}, \theta) = \sum_i (y_i - p_i(\theta))^2$.
- Dando mais peso aos pontos que possuem probabilidade próxima de 1/2, procuramos $\hat{\theta}$ que minimize $\sum_i (p_i(\theta)(1 - p_i(\theta))) |y_i - p_i(\theta)|$. Veja que $p(\theta)(1 - p(\theta)) \approx 0$ se $p(\theta) \approx 0$ ou ≈ 1 . Assim, damos mais peso aos pontos no “centro”, aqueles em que $p(\theta) \approx 1/2$.

Podemos misturar estimadores obtidos por métodos diferentes. Suponha que $\hat{\theta}_{MLE}$ seja o estimador MLE na regressão logística. Seja $\hat{\theta}_D$ o estimador que minimiza $D(\theta) = \sum_i |y_i - p_i(\theta)|$. Defina então seu estimador preferido como: $\hat{\theta} = (\hat{\theta}_{MLE} + \hat{\theta}_D)/2$, a média aritmética de $\hat{\theta}_{MLE}$ e $\hat{\theta}_D$. Podemos preferir usar outra média ponderada qualquer tal como: $\hat{\theta} = 0.75 \hat{\theta}_{MLE} + 0.25 \hat{\theta}_D$.

A lição com estes exemplos é que podemos pensar em *infinitos* estimadores, todos aparentemente razoáveis. Como escolher entre eles? Veremos uma teoria que nos ajuda neste sentido a seguir. Mais importante ainda: ela diz que, num certo sentido que vamos tornar preciso, o MLE é o melhor que podemos fazer. Um outro método qualquer pode até ser tão bom quanto o MLE mas não melhor que ele. Esta afirmação é demasiado geral e precisa ser melhor qualificada mas ela explica a imensa popularidade do método de máxima verossimilhança na análise de dados.

23.2 Estimadores são variáveis aleatórias

Para escolher bons estimadores, precisamos de uma teoria que nos guie. Nesta teoria, será *fundamental* ver os estimadores como variáveis ou vetores aleatórios. O que é uma variável ou vetor aleatório? Já aprendemos a entoar o mantra: são duas listas, uma com os valores possíveis, e outra com as probabilidades associadas.

Associada com esta necessidade de diferenciar estimadores como variáveis aleatórias, é importante entender a diferença conceitual entre parâmetros, estimadores e estimativas. Vamos diferenciar μ e \bar{Y} . Suponha que Y_1, Y_2, \dots, Y_n sejam i.i.d. $N(\mu, 1)$. Queremos estimar $\mu = \mathbb{E}(Y_i)$. Usamos $\bar{Y} = (Y_1 + \dots + Y_n)/n$

Qual a diferença entre μ e \bar{Y} ? Gerei no R cinco v.a.’s $N(0, 1)$. Assim, $\mu = 0$. O resultado foi: -0.962 -0.293 0.259 -1.152 0.196. Tivemos a estimativa $\bar{y} = -0.390$ para o parâmetro μ . Fazendo uma nova simulação, tivemos 0.030 0.085 1.117 -1.219 1.267 e nova estimativa: $\bar{y} = 0.256$.

Note que μ não muda de valor quando uma nova amostra é retirada. Temos sempre $\mu = 0$ aqui. Entretanto, \bar{y} muda de valor de amostra para amostra. Isto indica que μ e \bar{y} não podem ser as mesmas coisas. De fato, μ é um número real. Ele é desconhecido em geral mas não é uma variável aleatória. O parâmetro μ não é composto por duas listas, uma de valores possíveis e e outra com probabilidades associadas.

Para diferenciar entre estimador e estimativa, vamos realizar o experimento de retirar duas amostras de tamanho 5 de $N(0, 1)$. Você vai retirar a segunda amostra com nova semente. Antes de fazer o experimento, responda: qual das duas amostras, a 1a. ou a 2a., será a melhor para estimar μ ? Isto é, qual das duas amostras vai produzir uma média aritmética \bar{Y} mais próxima de μ ? Mesmo sabendo o verdadeiro valor de μ , é impossível responder a isto *antes* de retirarmos as amostras.

A razão é que o estimador \bar{Y} é uma *variável aleatória*. Sendo assim, \bar{Y} é simplesmente a coleção de duas listas: uma com os valores possíveis, e outra com as probabilidades associadas. Quando obtemos uma amostra específica, tal como -0.962 -0.293 0.259 -1.152 0.196, calculamos $\bar{y} = (y_1 + \dots + y_5)/5 = (-0.962 + \dots + 0.196)/5 = -0.390$. Este número \bar{y} é uma *estimativa*: é a instância específica do estimador \bar{Y} , que se materializa numa amostra particular. A estimativa \bar{y} é apenas um número, enquanto \bar{Y} é uma variável aleatória (e portanto, duas listas) Em termos

de notação, a única diferença entre o estimador \bar{Y} e a estimativa \bar{y} é o uso de letras maiúscula e minúsculas.

Esta diferença entre estimativa e estimador é similar à diferença entre uma função $f(x) = x^2$, que é uma *regra de associação entre dois conjuntos*, e o valor da função num ponto específico, tal como $x = 2$ e que portanto é igual a $f(2) = 2^2 = 4$. O que causa confusão é que muitas vezes queremos nos referir a um ponto x específico mas arbitrário, e neste caso escrevemos $f(x) = x^2$, como na definição da função.

No caso de estimadores como \bar{Y} , vamos querer falar do seu valor \bar{y} numa amostra específica mas arbitrária. Assim, o número real obtido com uma amostra específica mas arbitrária é deixado nesta notação geral \bar{y} que confunde-se um pouco com a notação \bar{Y} adotada para o estimador.

23.3 Comparando dois estimadores

Suponha que Y_1, Y_2, \dots, Y_5 sejam i.i.d. $N(\mu, 1)$. Queremos estimar μ . Podemos usar a variável aleatória $\bar{Y} = (Y_1 + \dots + Y_5)/5$. Algumas vezes (em algumas amostras) teremos a estimativa produzindo um erro de estimação $|\bar{y} - \mu| \approx 0$ mas algumas vezes podemos ter $|\bar{y} - \mu| >> 0$.

Alternativamente, poderíamos decidir usar outro estimador, a variável aleatória mediana amostral M . Este estimador é obtido assim:

- ordene a amostra: $Y_{(1)} = \min\{Y_1, \dots, Y_5\}$,
- $Y_{(2)}$ é o segundo menor valor da amostra, etc.
- até obtermos $Y_{(5)} = \max\{Y_1, \dots, Y_5\}$
- Então tome $M = Y_{(3)}$, o elemento central na lista ordenada, como um estimador de μ

Quem é melhor para estimar μ : a variável aleatória M ou a variável aleatória \bar{Y} ? Vamos tentar responder a isto simulando num computador. Com as duas amostras de uma $N(0, 1)$ que vimos antes, $-0.962 \ -0.293 \ 0.259 \ -1.152 \ 0.196$ e $0.030 \ 0.085 \ 1.117 \ -1.219 \ 1.267$, nós temos:

- Primeira amostra: $\bar{y} = -0.390$ e $m = -0.292$
- Segunda amostra: $\bar{y} = 0.256$ e $m = 0.085$

Note que neste caso simulado, diferentemente do que acontece na análise de dados reais, nós sabemos de antemão que $\mu = 0$. Sendo assim, nem faz sentido prático querer estimar μ se sabemos seu valor verdadeiro, com precisão absoluta. Entretanto, estamos apenas procurando verificar com a simulação qual dos dois estimadores, \bar{Y} ou M , é o melhor estimador.

Nas duas amostras acima, a v.a. M esteve mais próxima de $\mu = 0$ que \bar{Y} . Isto talvez seja um indicativo de que M tem um erro de estimação sempre menor que \bar{Y} . Isto é *falso*: numa terceira amostra temos os dados $-0.745 \ -1.131 \ -0.716 \ 0.253 \ 0.152$ com $\bar{y} = -0.437$ e $m = -0.716$. Neste caso, tivemos \bar{Y} mais próximo de μ que M . O fato é que, às vezes, teremos $|\bar{y} - \mu| < |m - \mu|$ mas às vezes teremos o contrário. O que acontece *em geral*? Qual o comportamento *estatístico* das v.a.'s M e \bar{Y} ?

Vamos fazer um estudo de simulação um pouco mais extenso para estudar o comportamento estatístico dos estimadores M e \bar{Y} . Vamos simular 1000 amostras de tamanho 5, calcular M e \bar{Y} em cada delas e verificar o erro que cada estimador cometeu em cada uma das 1000 amostras. O código abaixo mostra o que foi feito. Primeiro 5000 valores i.i.d. de uma $N(0, 1)$ foram organizados numa matriz `mat` de dimensão 1000×5 . E seguida, calculamos a média aritmética \bar{Y} de cada linha obtendo o vetor `media` de tamanho 1000. Fizemos o mesmo calculando a mediana M e obtendo o vetor `med` de tamanho 1000.

```
mat = matrix(rnorm(5*1000), ncol=5)
media = apply(mat, 1, mean) # media de cada linha
med = apply(mat, 1, median) # mediana de cada linha
aux = range(c(med, media)) # min e max das estimativas
```

```
par(mfrow=c(1, 2))
plot(media, med, asp=1, ylab="mediana") # media x mediana
abline(0,1)
plot(abs(media), abs(med), asp=1) # |media| x |mediana|
sum(abs(media - 0) > abs(med - 0))
[1] 427
```

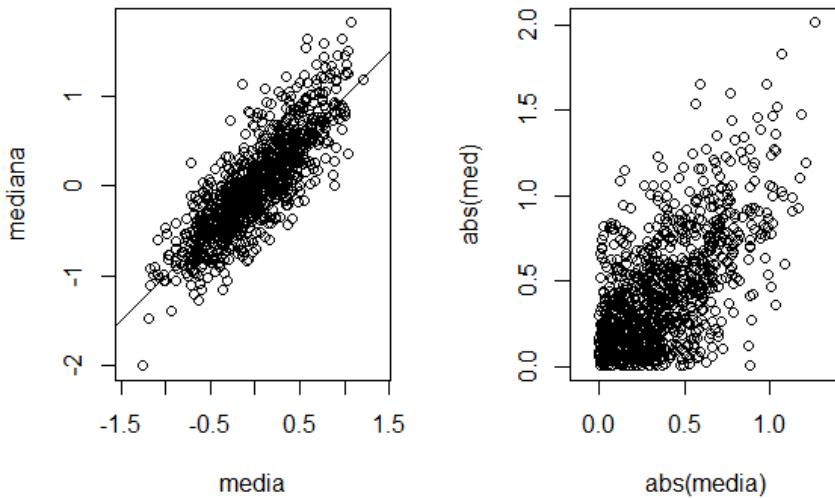


Figure 23.4: Esquerda: Gráfico de dispersão dos 1000 valores de M e \bar{Y} , estimadores de $\mu = 0$ em amostras de 5 observações independentes de uma $N(0, 1)$. Direita: Gráfico de dispersão dos erros de estimação dos dois estimadores (em valor absoluto) em cada amostra: $|M - \mu| = |M|$ e $|\bar{Y} - \mu| = |\bar{Y}|$.

A Figura 23.4 mostra um gráfico de dispersão dos 1000 valores de M e \bar{Y} . Vemos que eles são positivamente correlacionados. Quando \bar{y} sobreestima $\mu = 0$ (isto é, quando $\bar{y} > 0$), temos a tendência de também observar m acima de $\mu = 0$. Isto é razoável. Quando $M > 0$, temos uma amostra em que 3 ou mais de seus 5 valores estão acima de 0. Neste caso, podemos esperar que a média aritmética \bar{y} desses mesmos 5 valores \bar{y} tenha uma boa chance de também ser positiva. No gráfico à direita mostramos os erros de estimação dos dois estimadores (em valor absoluto) em cada amostra: $|m - \mu| = |m|$ versus $|\bar{y} - \mu| = |\bar{y}|$. Embora seja mais difícil de visualizar, parece que $|\bar{y}|$ tende a estar mais próximo de zero que o correspondente valor $|m|$. No último comando do script acima, verificamos que isto é verdade. Calculamos o número de amostras dentre as 1000 em que tivemos o erro de estimação \bar{Y} , dado por $|\bar{Y} - \mu| = |\bar{Y} - 0|$, maior que o erro de estimação de M , dado por $|M - \mu| = |M - 0|$. Obtivemos 427 em 1000. Assim, usando a ideia de frequência relativa, podemos estimar $\mathbb{P}(|\bar{Y} - \mu| > |M - \mu|) \approx 0.427$.

Podemos concluir que o estimador \bar{Y} é melhor que o estimador M sempre? Para todo tamanho de amostra n , e não apenas com $n = 5$ como neste exemplo? Para todo valor de μ , e não apenas para $\mu = 0$? Para todo valor do desvio-padrão σ , e não apenas para $\sigma = 1$? O tamanho de amostras foi adequado? Mil simulações parece um número pequeno para alcançar decisões muito firmes. Note, por exemplo, que a estimativa da probabilidade de ter um erro de estimação menor usando \bar{Y} foi 0.427, muito próximo de 0.5, quando não haveria diferença entre eles. Como concluir de forma

geral e definitiva sobre a qualidade de um estimador frente a outro? Vamos dar mais um passo nesta direção na próxima seção.

23.4 Estimação Pontual

Temos uma amostra aleatória $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$ de v.a.'s. A distribuição conjunta de \mathbf{Y} pertence a uma família (ou modelo) paramétrico com densidade conjunta $f(\mathbf{y}, \theta) = f(y_1, y_2, \dots, y_n; \theta)$. O parâmetro θ é um vetor de dimensão k pertencente a um conjunto $\Theta \subseteq \mathbb{R}^k$ chamado *espaço paramétrico*. Deseja-se inferir sobre θ . Conhecendo-se θ podemos (em princípio, pelo menos) calcular a probabilidade de qualquer evento ocorrer.

Definition 23.4.1 — Estatística. Uma estatística é uma função matemática $g(\mathbf{Y}) = g(Y_1, \dots, Y_n)$ que tenha como argumento \mathbf{Y} e que tome valores em \mathbb{R}^h . Uma estatística não pode envolver os parâmetros desconhecidos θ .

Definition 23.4.2 — Estimador pontual. Um estimador pontual de θ ou, de forma mais geral, de uma função $q(\theta)$ será qualquer estatística $g(\mathbf{Y}) = g(Y_1, \dots, Y_n)$. A única diferença entre uma estatística e um estimador é que ao definir um estimador precisamos declarar o quê ele está estimando (declarar $q(\theta)$).

O motivo para lidarmos com a notação mais geral $q(\theta)$ é que, às vezes, podemos estar interessados apenas num aspecto da distribuição dos dados, sem muito interesse pela distribuição como um todo. Por exemplo, se $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$ é uma amostra de v.a.'s i.i.d. com distribuição gama com parâmetros $\theta = (\alpha, \beta)$, podemos estar interessado apenas na sua esperança $\mu = q(\alpha, \beta) = \alpha/\beta$. Assim, podemos tentar estimar diretamente μ sem necessariamente precisar estimar ambos, α e β . Entretanto, vamos nos concentrar neste texto no problema de estimar θ , o parâmetro que indexa a família de densidades $f(\mathbf{y}, \theta)$ associada com os dados da amostra.

■ **Example 23.3** Seja uma amostra $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$ de v.a.'s i.i.d. gaussianas $N(\mu, \sigma^2)$. Isto é, $\mathbb{E}(Y_i) = \mu$ e $\text{Var}(Y_i) = \mathbb{E}(Y_i - \mu)^2 = \sigma^2$. Seja $\bar{Y} = \frac{1}{n} \sum_i Y_i$, a média aritmética das v.a.'s. A média aritmética \bar{Y} é usualmente tomada como um estimador de μ . A variância amostral $S^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2$ é usualmente um estimador para σ^2 . ■

■ **Example 23.4** Seja uma amostra $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$ de v.a.'s i.i.d. com distribuição Poisson(θ). No caso da distribuição de Poisson, temos não apenas $\mathbb{E}(Y_i) = \theta$ mas também $\text{Var}(Y_i) = \theta$. Seja $\bar{Y} = \frac{1}{n} \sum_i Y_i$, a média aritmética das v.a.'s. Como $\bar{Y} \approx \mathbb{E}(Y_i)$ quando o tamanho amostral n é grande, usamos \bar{Y} como estimador de θ no caso da Poisson.

O caso curioso aqui é que \bar{Y} é também um estimador da variância das v.a.'s já que $\text{Var}(Y_i)$ é também igual a θ . Mais curioso ainda, como vimos no capítulo 16, pela Lei dos Grandes Números, a variância amostral $S^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2$ converge para a variância populacional se tivermos v.a.'s i.i.d. Assim, $S^2 \approx \text{Var}(Y_i) = \theta$ e portanto S^2 serve como estimador tanto para a média populacional quanto para a variância populacional no caso de dados Poisson. Na verdade, esperamos que $\bar{Y} \approx S^2$ se n é grande no caso de uma Poisson. Esta ideia de que a razão S^2/\bar{Y} deveria ser aproximadamente igual a 1 é explorada em alguns métodos para testar se os dados seguem de fato uma distribuição de Poisson. ■

■ **Example 23.5** Seja uma amostra $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$ de v.a.'s i.i.d. sem que especifiquemos uma classe de densidades ou modelo seguido pelos dados. Sejam $\mathbb{E}(Y_i) = \mu$ e $\text{Var}(Y_i) = \sigma^2$. Então $\bar{Y} = \frac{1}{n} \sum_i Y_i$ é um estimador intuitivamente razoável para μ pois, pela Lei dos Grandes Números, esperamos $\bar{Y} \approx \mu$. Pela mesma razão, a variância amostral $S^2 \approx \sigma^2$ e portanto S^2 é um estimador intuitivamente razoável para σ^2 . ■

■ **Example 23.6 — Estimador pontual como vetor.** Considere uma amostra $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$ de v.a.'s i.i.d. gaussianas $N(\mu, \sigma^2)$. Seja que $\theta = (\mu, \sigma^2)$. É comum usarmos o vetor bi-dimensional

$$\mathbf{T} = g(\mathbf{Y}) = (\bar{Y}, S^2)$$

para fazer inferência sobre θ . θ_y é uma estatística bi-dimensional já que cada entrada de θ_y é uma função dos dados \mathbf{Y} . A primeira entrada do vetor θ_y é a média aritmética $\sum_i Y_i/n$ e ela é usada para inferir sobre o valor desconhecido de μ . A segunda entrada é uma medida empírica da variação dos dados em torno de \bar{Y}_n e ela é usada para inferir sobre o valor de $\sigma^2 = \mathbb{E}(Y - \mu)^2$.

Quando algo é um estimador?

Definição de estimador permite que QUALQUER estatística $g(\mathbf{Y})$ seja estimador de θ . Mas então podemos usar \bar{Y} ou $W = \max\{Y_1, \dots, Y_n\}$ como estimadores de $\sigma^2 = \mathbb{V}(Y_i)$? Podemos mas não devemos. Vamos ver que \bar{Y} e W tem propriedades muito ruins como estimadores de σ^2 . Podemos facilmente encontrar estimadores de σ^2 , tais como $S^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2$, que são muito melhores que \bar{Y} ou $W = \max\{Y_1, \dots, Y_n\}$.

Outros exemplos de estimadores pontuais

$\theta_y = g(\mathbf{Y}) = (\bar{Y}, (Y_{(n)} - Y_{(1)})/2)$, a média e metade da variação total (range) da amostra
 $\theta_y = g(\mathbf{Y}) = \hat{F}_n(3.2) = \#\{Y_i; Y_i \leq 3.2\}/n$ onde $\#A$ é o número de elementos (ou cardinalidade) do conjunto A . Isto é, $\hat{F}_n(3.2)$ é a proporção de elementos da amostra que são menores ou iguais a 3.2. Poderíamos substituir o ponto $x = 3.2$ por qualquer outro no exemplo acima.

Estimadores não-intuitivos

Alguns estimadores possuem fórmulas matemáticas complicadas e não-intuitivas. Por exemplo, para variáveis aleatórias Y_i positivas (isto é, com $P(Y_i > 0) = 1$) podemos definir a estatística $\theta_y = g(\mathbf{Y}) =$



24. Algoritmo EM

24.1 Misturas de distribuições: introdução

Os modelos básicos de distribuições que dispomos são flexíveis mas não dão conta de tudo que ocorre. Será raro que um conjunto de instâncias seja muito bem modelado por uma das poucas distribuições que aprendemos até agora. Temos duas alternativas:

- Aumentar o nosso “dicionário de distribuições” criando uma looooooonga lista de distribuições para ajustar aos dados reais.
- Misturar os tipos básicos já definidos criando tipos compostos que ampliam a classe de distribuições disponíveis para análise.

Uma forma de misturar distribuições é construir um modelo de regressão: Cada i -ésimo indivíduo tem uma distribuição Y_i (como a gaussiana, por exemplo) que é modulada pelas suas variáveis independentes ou *features* \mathbf{x}_i . Essas variáveis independentes controlam o valor esperado $\mathbb{E}(Y_i) = \mathbf{x}'_i \beta$. Assim, a coleção de v.a.’s Y_1, \dots, Y_n é uma mistura de diferentes gaussianas, cada uma tendo um valor esperado próprio.

Vai ser comum não termos covariáveis para usarmos como *features* mas ainda assim termos claramente dados vindos de 2 ou mais distribuições. Também será comum nos modelos de fatores latentes que, mesmo usando as *features* para controlar as distribuições, ainda tenhamos que usar mais misturas de distribuições.

Um exemplo clássico em aprendizagem de máquina é o de dados de erupções do geyser Faithful do Parque Yellowstone nos EUA, mostrados na Figura 24.1. Cada ponto representa uma erupção. No eixo horizontal, temos a duração de cada erupção. No eixo vertical, temos o intervalo entre a erupção em questão e a erupção seguinte. Parece que existem duas nuvens elípticas de pontos, indicando que estes dados podem vir de duas normais bivariadas misturadas.

Para entender e dominar as ferramentas para lidar com misturas de distribuições, vamos retornar para o caso de v.a.’s unidimensionais gaussianas. Vamos exemplificar o que é a mistura de três gaussianas. A Figura 24.2 mostra a densidade de probabilidade das gaussianas $N(0, 1)$, $N(3, 0.5^2)$ e $N(10, 1)$ e a densidade de sua mistura nas respectivas proporções 0.6, 0.1 e 0.3. Note como podemos visualizar as densidades originais através dos picos remanescentes na densidade da mistura. Note também que a densidade misturada com a menor proporção, a gaussiana $N(3, 0.5^2)$ aparece com o

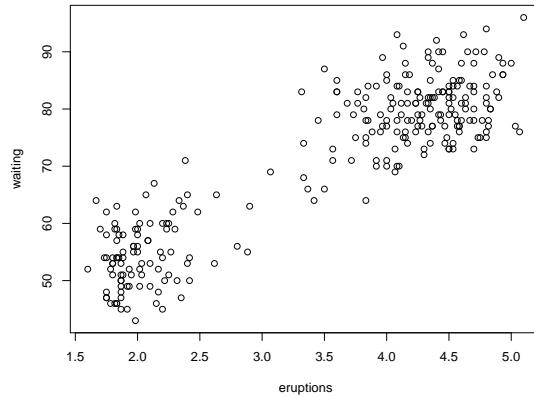


Figure 24.1: Dados de $n = 272$ erupções da geyser Faithful do Parque Yellowstone nos EUA. No eixo horizontal, a duração de cada erupção. No eixo vertical, temos o intervalo entre a erupção em questão e a erupção seguinte. Parece que existem duas normais bivariadas misturadas.

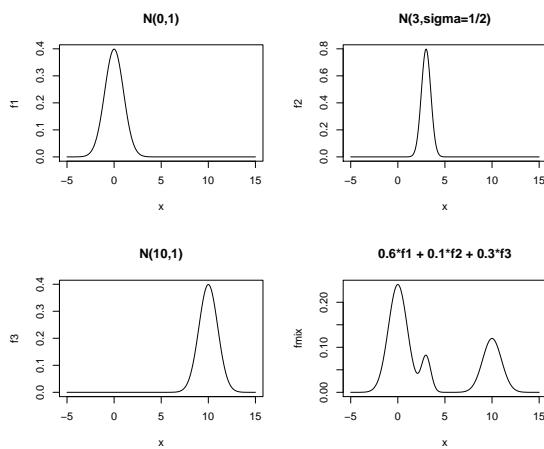


Figure 24.2: Mistura de 3 normais $N(0,1)$, $N(3,0.5^2)$ e $N(10,1)$ nas proporções 0.6, 0.1 e 0.3, respectivamente.

menor pico na densidade da mistura. Antes de cermos como esta densidade da mistura é obtida, vamos mostrar na Figura 24.3 uma amostra de v.a.'s i.i.d vindas dessa distribuição de mistura. Na mesma Figura, mostramos a densidade da mistura sobreposta ao histograma padronizado. Desse modo, quando a msitura fo relativamente simples como nesse caso, poderemos visualmente reconhecer a presença das distribuições componentes.

O caso de uma mistura de v.a.'s discretas é similar. Xiao et al (1999) estudam o período de internação em hospitais para modelagem de custos. Um tipo de financiamento de custos de saúde bastante usado mundo afora é baseado no DRG ?? Explicar ?? Considerando apenas as internações devido a partos...?? É bastante conhecido que partos por cesáriana levam tipicamente a tempos mais longos de internação que partos naturais.

Hospital-maternidade na Austrália. The total sample size is 5648. The data, based on separation dates, are available from July 1992 to September 1996. Patients' socio-economic characteristics (age, gender, Aboriginality, marital status, country of birth, occupation, employment and insurance status), health provision factors (mode of separation, accommodation status, admission type, source

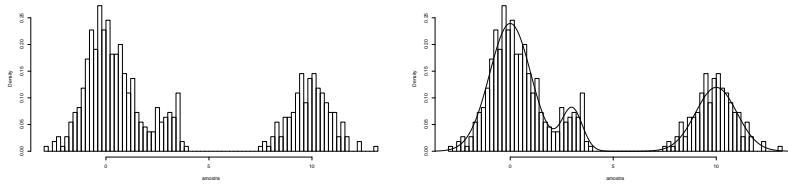


Figure 24.3: Amostra de $n = 550$ dados vindos da densidade mistura de 3 normais. A mesma amostra com a densidade da mistura sobreposta.

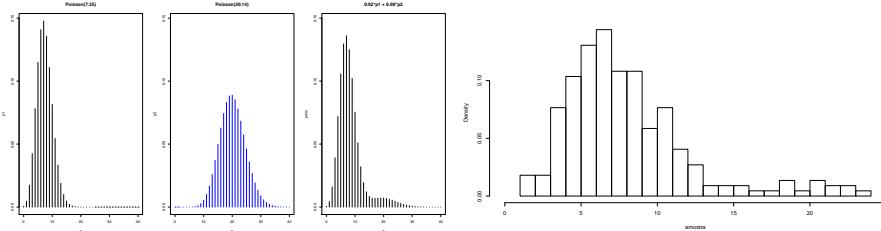


Figure 24.4: Esquerda: Mistura com 92% vindo de uma Poisson($\lambda = 7.35$) e os outros 8% vindo de uma Poisson($\lambda = 20.1$). Amostra de $n = 222$ casos de parto cesáreo com complicações graves. Dados adaptados de Xiao et. al. (1999). Mistura: 92% vêm de uma Poisson($\lambda = 7.35$) e os outros 8% vêm de Poisson($\lambda = 20.1$).

of referral, distance from the hospital and treating medical officer) and other relevant factors (number of diagnoses, number of procedures and number of inpatient theatre attendance) were reviewed and selected from the main database. These were considered as potential factors influencing LOS based on our previous study. Only significant factors were included in the final models.

A Figura 24.4

24.2 Misturas de distribuições: formalismo

Vamos considerar inicialmente o caso contínuo. Estamos olhando um atributo Y . Suponha que temos três sub-populações: 1, 2 e 3. Represente as medições nas diferentes sub-populações como v.a.'s Y_1 , Y_2 , e Y_3 . As sub-populações são diferentes e isto implica que as v.a.'s têm densidades diferentes. As densidades são: $f_1(y)$, $f_2(y)$ e $f_3(y)$ e as respectivas distribuições acumuladas são $F_1(y)$, $F_2(y)$ e $F_3(y)$. Assim, $F'_1(y) = f_1(y)$, $F'_2(y) = f_2(y)$ e $F'_3(y) = f_3(y)$. Para fixar as ideias, considere mentalmente o caso da Figura 24.2 em que a população 1 seguia uma $N(0, 1)$, a densidade 2 era $N(3, 1/2^2)$ e a população 3 era uma $N(10, 1)$.

A variável realmente observada é representada por Y . Qual a distribuição de probabilidade da v.a. Y ? Se o indivíduo vier da população 1, Y terá a mesma distribuição que a v.a. Y_1 . Se vier da população 2, $Y \sim Y_2$, e se vier da população 3, $Y \sim Y_3$. O indivíduo da população mistura vem aleatoriamente de *uma* das três populações. Ele vem das 3 populações com as seguintes probabilidades:

- vem da população 1 com probabilidade θ_1 ;
- vem da população 2 com probabilidade θ_2 ;
- vem da população 3 com probabilidade θ_3

com $\theta_1 + \theta_2 + \theta_3 = 1$ pois só existem estas três alternativas de onde Y é extraída.

Assim, a medição Y tem a seguinte estrutura aleatória: Y tem a mesma distribuição que Y_k com probab θ_k ou, de forma mais compacta:

- $Y \sim Y_1$ com probabilidade θ_1

- $Y \sim Y_2$ com probabilidade θ_2
- $Y \sim Y_3$ com probabilidade θ_3

Qual a densidade de Y ? Usamos a fórmula da probabilidade total para calcular a função de distribuição acumulada $\mathbb{F}(y) = \mathbb{P}(Y \leq y)$. Vamos condicionar no resultado de qual população Y foi amostrada e a seguir somamos (de forma ponderada) sobre as três possíveis populações. Temos

$$\mathbb{F}(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(Y \leq y \text{ e vem de alguma pop})$$

Temos a igualdade de eventos

$$\begin{aligned} [Y \leq y] &= [Y \leq y \cap \text{vem de pop 1}] \cup [Y \leq y \cap \text{vem de pop 2}] \\ &\quad \cup [Y \leq y \cap \text{vem de pop 3}] \end{aligned}$$

Como os eventos são disjuntos, a probab da sua união é a soma das probabilidades:

$$\begin{aligned} \mathbb{F}(y) &= \mathbb{P}(Y \leq y \cap \text{vem de pop 1}) + \mathbb{P}(Y \leq y \cap \text{vem de pop 2}) + \mathbb{P}(Y \leq y \cap \text{vem de pop 3}) \\ &= \mathbb{P}(Y \leq y | \text{pop 1})\mathbb{P}(\text{pop 1}) + \mathbb{P}(Y \leq y | \text{pop 2})\mathbb{P}(\text{pop 2}) + \mathbb{P}(Y \leq y | \text{pop 3})\mathbb{P}(\text{pop 3}) \\ &= \mathbb{P}_1(Y \leq y)\theta_1 + \mathbb{P}_2(Y \leq y)\theta_2 + \mathbb{P}_3(Y \leq y)\theta_3 \\ &= \mathbb{F}_1(y)\theta_1 + \mathbb{F}_2(y)\theta_2 + \mathbb{F}_3(y)\theta_3 \end{aligned}$$

Assim, descobrimos que a função de distribuição acumulada $\mathbb{F}(y)$ é uma média ponderada das dist acumuladas $\mathbb{F}_i(y)$ das componentes da mistura. Outra maneira de dizer isto é: a distribuição acumulada da mistura é a mistura das distribuições acumuladas.

A distribuição acumulada não é muito intuitiva. A densidade é mais interpretável por sua ligação com os histogramas dos dados observados. Se temos a distribuição acumulada, podemos obter a densidade de Y derivando $\mathbb{F}(y)$:

$$\begin{aligned} f(y) &= \mathbb{F}'(y) = \mathbb{F}'_1(y)\theta_1 + \mathbb{F}'_2(y)\theta_2 + \mathbb{F}'_3(y)\theta_3 \\ &= f_1(y)\theta_1 + f_2(y)\theta_2 + f_3(y)\theta_3 \end{aligned}$$

Assim, a densidade da mistura Y é a mistura das densidades das componentes Y_1 , Y_2 e Y_3 .

Reveja a Figura 24.2 para enxergar esta fórmula. O código R para obtê-la é o seguinte:

```
x <- seq(-5, 15, by=0.01)
f1 <- dnorm(x) # densidade N(0,1) nos pontos de x
f2 <- dnorm(x, 3, 1/2) # densidade de N(mu=3, sigma=1/2)
f3 <- dnorm(x, 10, 1)
fmix <- 0.6*f1 + 0.1*f2 + 0.3*f3
par(mfrow=c(2,2))
plot(x, f1, type="l"); title("N(0,1)")
plot(x, f2, type="l"); title("N(3,sigma=1/2)")
plot(x, f3, type="l"); title("N(10,1)")
plot(x, fmix, type="l"); title("0.6*f1 + 0.1*f2 + 0.3*f3")
```

Se quisermos gerar uma amostra de tamanho $n = 550$ da mistura de três normais, usamos um algoritmo muito simples que reproduz o conceito de mistura:

```
for(i in 1:550){
  Seleccione a pop k = 1, 2 ou 3 com probabs p1, p2, p3
  Y = um valor da normal da pop k
}
```

O script R correspondente é o seguinte:

```

## gerando amostra da mistura (n=550)
## 3 subpops normais, probabs = c(0.6, 0.1, 0.3)
## numero de cada subpop
num <- rmultinom(n=1, size=550, prob=c(0.6, 0.1, 0.3))
num # gerou (321, 56, 173)
amostra <- c(rnorm(num[1]), rnorm(num[2], 3, 1/2), rnorm(num[3], 10, 1))

```

No caso de v.a.'s discretas, os resultados são os mesmos do caso contínuo. Suponha que Y seja uma mistura de três v.a.'s discretas: Y_1, Y_2, Y_3 (por exemplo, 3 Poissons) As 3 v.a.'s têm distribuições acumuladas $\mathbb{F}_k(y)$ e função de probabilidade $p_k(y) = \mathbb{P}(Y_k = y)$ para $k = 1, 2, 3$. Então, a distribuição acumulada da mistura Y é dada por

$$\mathbb{F}(y) = \mathbb{P}(Y \leq y) = \mathbb{F}_1(y)\theta_1 + \mathbb{F}_2(y)\theta_2 + \mathbb{F}_3(y)\theta_3,$$

idêntico ao caso contínuo. A função de massa de probabilidade é dada por

$$\begin{aligned} p(y) &= \mathbb{P}(Y = y) \\ &= \mathbb{F}(y) - \mathbb{F}(y-1) \\ &= p_1(y)\theta_1 + p_2(y)\theta_2 + p_3(y)\theta_3 \end{aligned}$$

24.2.1 Misturas de normais multivariadas

Voltamos aos dados de erupção do geyser Faithful mostrados na Figura 24.1. Aparentemente temos duas normais bivariadas misturadas nestes dados. Olhando os dados, podemos chutar grosseiramente os valores dos parâmetros de cada componente. Considerando o componente 1, no canto inferior esquerdo do gráfico, podemos chutar que o vetor de valores esperados é $\mu_1 = (\mu_{11}, \mu_{12}) = (2.1, 52)$ e a matriz de covariância é

$$\Sigma_1 = \begin{bmatrix} \sigma_{11}^2 & \rho_1\sigma_{11}\sigma_{12} \\ \rho_1\sigma_{11}\sigma_{12} & \sigma_{12}^2 \end{bmatrix} = \begin{bmatrix} (0.25)^2 & 0.3\sigma_{11}\sigma_{12} \\ 0.3\sigma_{11}\sigma_{12} & 4^2 \end{bmatrix}$$

Para o componente 2, no canto superior direito do gráfico, temos $\mu_2 = (\mu_{21}, \mu_{22}) = (4.5, 80)$ e a matriz de covariância:

$$\Sigma_2 = \begin{bmatrix} \sigma_{21}^2 & \rho_2\sigma_{21}\sigma_{22} \\ \rho_2\sigma_{21}\sigma_{22} & \sigma_{22}^2 \end{bmatrix} = \begin{bmatrix} (0.35)^2 & 0.7\sigma_{21}\sigma_{22} \\ 0.7\sigma_{21}\sigma_{22} & 5^2 \end{bmatrix}$$

A proporção do componente 1 pode ser estimada grosseiramente em 35% ou $\theta_1 = 0.35$

Cmo no caso univariado, a densidade conjunta do vetor bivariado $\mathbf{Y} = (Y_1, Y_2)$ é uma mistura de duas densidades gaussianas bivariadas:

$$f(\mathbf{y}) = f(y_1, y_2) = \theta_1 f_1(y_1, y_2) + \theta_2 f_2(y_1, y_2)$$

onde $\theta_1 + \theta_2 = 1$ com $\theta_1 \geq 0$ e $\theta_2 \geq 0$ e com $f_1(y_1, y_2)$ sendo a densidade do componente 1 (uma normal bivariada) e $f_2(y_1, y_2)$ sendo a densidade do componente 2 (também uma normal bivariada).

A Figura 24.2.1 mostra a densidade da mistura e a amostra que seria gerada por ela.

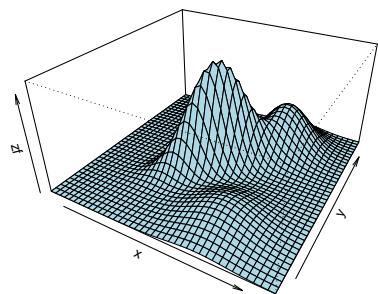
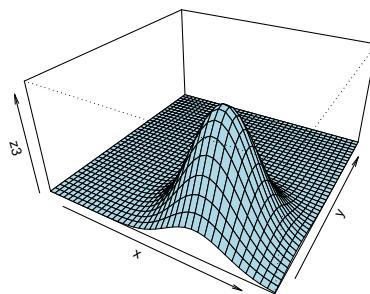
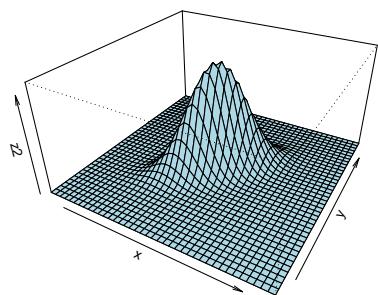
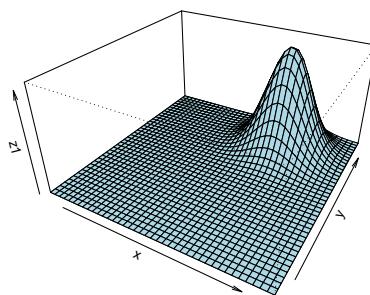
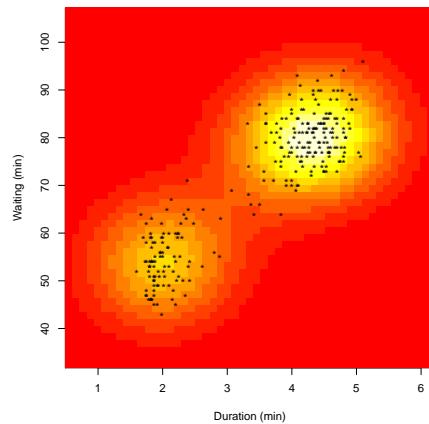
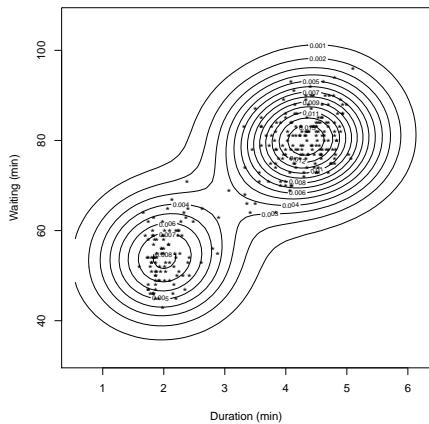
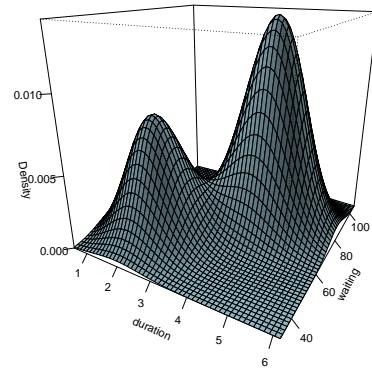
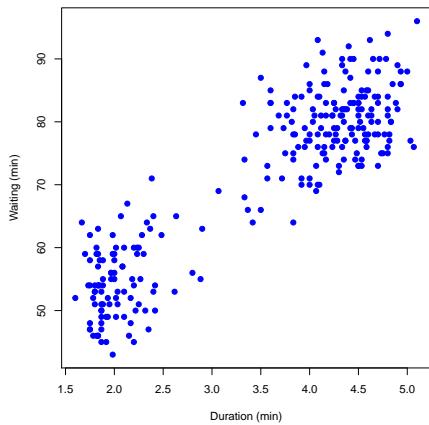
Uma outra visualização da densidade da mistura está na Figura 24.2.1.

A Figura 24.2.1 mostra a densidade da mistura de 3 normais bivariadas. O código R para esta figura está abaixo.

```

par(mfrow=c(2,2), mar=c(0,0,0,0))
x <- seq(-5, 5, length= 40); y <- x
f <- function(x,y) { dnorm(x,2,1)*dnorm(y,3,1) }
z1 <- outer(x, y, f)

```



```

persp(x, y, z1, theta = 30, phi = 30, expand = 0.5, col = "lightblue")

# densidade normal bivariada com rho=0.7
f <- function(x,y, rho=0.7){exp(-(x^2 - 2*rho*x*y + y^2)/(2*(1-rho^2)))/(2*pi*sqrt(1-rho^2))
z2 <- outer(x, y, f)
persp(x, y, z2, theta = 30, phi = 30, expand = 0.5, col = "lightblue")

f <- function(x,y) { dnorm(x,2,1.2)*dnorm(y,-3,1.2) }
z3 <- outer(x, y, f)
persp(x, y, z3, theta = 30, phi = 30, expand = 0.5, col = "lightblue")

zf <- 0.3*z1 + 0.5*z2 + 0.2*z3
persp(x, y, zf, theta = 30, phi = 30, expand = 0.5, col = "lightblue")

```

24.3 Estimando uma distribuição de mistura

O algoritmo para gerar dados de uma mistura é simples:

- Input: Número de grupos k
- Input: Densidade de cada grupo: $f_1(\mathbf{y}), f_2(\mathbf{y}), \dots, f_k(\mathbf{y})$
- Input: proporções de cada grupo: $\theta_1, \theta_2, \dots, \theta_k$
- Gerar amostra de mistura $\theta_1 f_1(\mathbf{y}) + \dots + \theta_k f_k(\mathbf{y})$ de k componentes: fácil.
- Passo 1: Escolha uma das k componentes ao acaso com probabilidades $\theta_1, \dots, \theta_k$.
- Passo 2: Selecione Y da distribuição $f_i(\mathbf{y})$ da componente i selecionada no passo anterior.

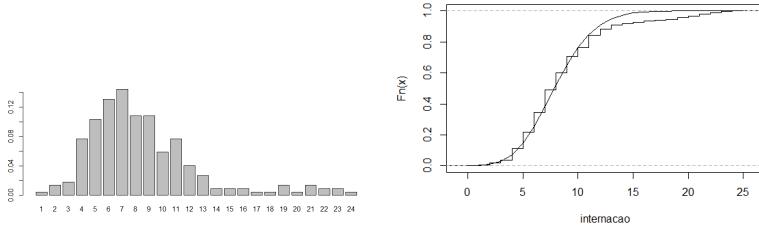
Isto é, dado o mecanismo (o modelo) aleatório de uma mistura, podemos gerar dados sintéticos. Mas o problema *realmente* relevante é o contrário. Como ajustar um modelo de mistura a dados estatísticos? Isto é, recebemos os dados e queremos inferir qual o modelo que foi usado para gerá-los. Esta tarefa não é tão simples. Primeiro, precisamos saber quantos componentes estão presentes. Vamos discutir isto mais tarde. Neste ponto, vamos supor que o número de componentes é conhecido e é igual K . A outra peça importante é o tipo de distribuição presente na mistura: gaussiana, gama, Poisson, Weibull? Vamos também supor que os possíveis tipos vêm de uma mesma família de distribuições e que esta família é conhecida. Voltaremos a discutir estes dois pontos mais a frente. Assim, se (e este é um grande se)

- soubermos o número K de componentes na mistura (digamos, $K = 3$)
 - soubermos a classe da distribuição de probabilidade de cada componente (digamos, normal).
- então poderemos usar o algoritmo EM para ajustar o modelo de mistura. A seleção de k é feita via técnicas de escolha de modelos: ajustamos vários modelos com diferentes k e escolhemos o “melhor”. Veremos seleção de modelos mais tarde neste curso.

24.4 Dados e rótulos

Nem sempre é fácil obter o MLE devido a dificuldades na otimização da função de verossimilhança. Um problema difícil é quando temos variáveis latentes ou ocultas (hidden or latent states). Por exemplo, nos problemas de mistura que aparecem em diversas áreas como análise de imagens, de textos, etc. Uma das maneiras de estimar os parâmetros num modelo de análise fatorial é através do algoritmo EM. Vamos começar o estudo do algoritmo EM em problemas simples de misturas.

Suponha que estamos analisando o número de dias de internação de 222 mulheres após um parto cesáreo com complicações. O resultado está no lado esquerdo da Figura 24.4. A cauda estende-se por uma faixa muito longa para vir de uma única Poisson. De fato, se supusermos que uma única Poisson(λ) gerou estes dados, vamos usar a média amostral igual a $\bar{y} = 8.51$ como estimador de λ . Podemos fazer um gráfico da distribuição acumulada empírica $\mathbb{F}_n(y)$ junto com o



da distribuição acumulada teórica de uma Poisson($\lambda = 8.51$). O resultado está no lado direito da Figura 24.4. Pela curva teórica deveríamos ter quase toda a amostra abaixo de $y = 15$. Entretanto, a curva empírica mostra que ainda existem vários dados acima desse valor.

```
mean(amostra)
Fn <- ecdf(amostra)
plot(Fn, verticals= T, do.p=F, main="", xlab="internacao")
x <- 0:25
y <- ppois(x, mean(amostra))
lines(x,y)
```

Como a distribuição é discreta, o teste de Kolmogorov não é válido. Usamos então o teste qui-quadrado. Para escolher as classes, devemos ter pelo menos 5 observações em cada classe para que a distribuição assintótica do teste qui-quadrado seja uma boa aproximação.

```
> table(amostra)
amostra
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
 1  3  4 17 23 29 32 24 24 13 17  9  6  2  2  2  1  1  3  1  3  2  2  1
```

Vamos agrupar as contagens iniciais criando uma classe $[Y \leq 3]$, as classes $[Y = 4], \dots, [Y = 13]$, e as classes $[14 \leq Y \leq 19]$ e $[Y \geq 20]$. Assim, terminamos com $m = 13$ classes indexadas por $j = 1, \dots, 13$ e com as contagens C_j . As probabilidades p_j de uma Poisson($\lambda = 8.51$) em cada uma dessas categorias é facilmente obtida em R e assim as contagens esperadas $E_j = 222p_j$. Com isto a estatística do teste qui-quadrado

$$X^2 = \sum_j \frac{(C_j - E_j)^2}{E_j} \quad (24.1)$$

é calculada junto com seu p-valor.

```
> x = table(amostra)
> esp = 222*c(ppois(3,8.51), dpois(4:13, 8.51), ppois(19,8.51)-ppois(13, 8.51), 1 - ppois(19, 8.51))
> obs = c(sum(x[1:3]), x[4:13], sum(x[14:19]), sum(x[20:24]))
> xisq = sum( (obs - esp)^2/esp )
> xisq
[1] 674.7081
> 1-pchisq(xisq, 13-2)
[1] 0
> round((obs - esp)/sqrt(esp),2)
0.53  2.31  1.56  1.11  0.62 -1.18 -0.90 -2.33 -0.46 -1.22 -0.95 -0.11 25.59
```

O p-valor é aproximadamente igual a zero e portanto, uma (única) distribuição de probabilidade não é adequada para modelar estes dados. O último comando acima mostra os desvios ($C_j -$

$E_j)/\sqrt{E_j}$ usados (ao quadrado) na estatística qui-quadrado X^2 em (24.1). Observe que, comparado com o esperado sob uma única Poisson, os desvios qui-quadrado mostram que existe um excesso substancial na última categoria, com as contagens mais altas.

Com esta motivação, vamos partir para uma mistura de duas Poissons, uma delas servindo para capturar a pequena parcela de mulheres que ficam um tempo relativamente bastante longo após o parto cesáreo com complicações. Vamos assumir que uma proporção α dos dados vem de uma Poisson com parâmetro λ_a . A proporção $1 - \alpha$ restante vem de uma Poisson com parâmetro λ_b . Queremos inferir de maneira automática sobre $\theta = (\lambda_a, \lambda_b, \alpha)$. Como fazer isto?

A partir do gráfico na Figura 24.4, podemos conjecturar que temos uma Poisson($\lambda_a \approx 2$) e uma Poisson($\lambda_b \approx 10$). Queremos o melhor estimador possível, o MLE. Seria muito fácil obter esse MLE se soubéssemos a qual grupo cada observação pertence. Neste caso, bastaria ajustar uma Poisson separadamente a cada um dos dois grupos de dados. O MLE do parâmetro λ de uma Poisson simples é a média aritmética dos dados. Infelizmente não sabemos isto: observamos apenas os dados numéricos, e não sua classe.

Mas como seria no caso em que conhecêssemos os rótulos dos grupos? Se soubéssemos, o vetor de dados com a informação completa, da contagem e do rótulo do grupo, pode ser representado por

$$(\mathbf{y}, \mathbf{z}) = (y_1, \dots, y_{222}, z_1, \dots, z_{222})$$

onde y_i é a contagem da mulher i e z_i é o rótulo do seu grupo, com λ_a ou com λ_b . Temos $z_i = 0$ se a i -ésima mulher for do grupo que se interna menos e portanto a y_i contagem vem de uma Poisson com parâmetro λ_a . Se $z_i = 1$ então a i -ésima mulher é do grupo 2 e $y_i \sim \text{Poisson}(\lambda_b)$. Os dados realmente observados são apenas as contagens $\mathbf{y} = (y_1, \dots, y_{222})$. As variáveis não-observadas em $\mathbf{z} = (z_1, \dots, z_{222})$ são chamadas de variáveis latentes ou ocultas (hidden, latent). O vetor de parâmetros é $\theta = (\lambda_a, \lambda_b, \alpha)$.

24.4.1 A verossimilhança completa

O vetor (y_i, z_i) é composto por duas v.a.'s discretas com distribuição conjunta dada por

$$\begin{aligned}\mathbb{P}(y_i = y, z_i = 0) &= \mathbb{P}(y_i = y | z_i = 0)\mathbb{P}(z_i = 0) = \frac{\lambda_a^y}{y!} e^{-\lambda_a} \cdot (1 - \alpha) \\ \mathbb{P}(y_i = y, z_i = 1) &= \mathbb{P}(y_i = y | z_i = 1)\mathbb{P}(z_i = 1) = \frac{\lambda_b^y}{y!} e^{-\lambda_b} \cdot \alpha\end{aligned}$$

para $y \in \mathbb{N}$. Podemos escrever estas duas expressões com uma única linha. Para $z = 0$ ou $z = 1$, temos

$$\mathbb{P}(y_i = y, z_i = z) = \left[\frac{\lambda_a^y e^{-\lambda_a}}{y!} (1 - \alpha) \right]^{1-z} \left[\frac{\lambda_b^y e^{-\lambda_b}}{y!} \alpha \right]^z \quad (24.2)$$

Estamos supondo que as contagens de diferentes mulheres são v.a.'s independentes. Então, a verossimilhança de $\theta = (\lambda_a, \lambda_b, \alpha)$ baseada nos dados completos é

$$L^c(\theta | \mathbf{y}, \mathbf{z}) = \prod_{i=1}^{222} \left(\frac{\lambda_a^{y_i} e^{-\lambda_a}}{y_i!} (1 - \alpha) \right)^{1-z_i} \left(\frac{\lambda_b^{y_i} e^{-\lambda_b}}{y_i!} \alpha \right)^{z_i}$$

Tomando logaritmo, temos a log-verossimilhança completa

$$\ell^c(\theta | \mathbf{y}, \mathbf{z}) = \sum_{i=1}^{222} (1 - z_i) \log \left(\frac{\lambda_a^{y_i} e^{-\lambda_a}}{y_i!} (1 - \alpha) \right) + z_i \log \left(\frac{\lambda_b^{y_i} e^{-\lambda_b}}{y_i!} \alpha \right)$$

O MLE de $\theta = (\lambda_a, \lambda_b, \alpha)$ no caso da informação completa (\mathbf{y}, \mathbf{z}) estar disponível é muito simples (fica como exercício):

$$\hat{\alpha} = \frac{1}{222} \sum_{i=1}^{222} z_i = \text{proporção das pacientes com } z_i = 1 \quad (24.3)$$

$$\hat{\lambda}_a = \frac{\sum_{i=1}^{222} y_i(1-z_i)}{\sum_{i=1}^{222} (1-z_i)} = \text{média das pacientes com } z_i = 0 \quad (24.4)$$

$$\hat{\lambda}_b = \frac{\sum_{i=1}^{222} y_i z_i}{\sum_{i=1}^{222} z_i} = \text{média das pacientes com } z_i = 1 \quad (24.5)$$

(24.6)

Se pelo menos tivéssemos o vetor completo (\mathbf{y}, \mathbf{z}) Mas o que temos é apenas o vetor \mathbf{y} das contagens. Precisamos da verossimilhança de $\alpha, \lambda_a, \lambda_b$ usando *apenas* \mathbf{y} .

24.4.2 A verossimilhança incompleta

Vamos obter a verossimilhança marginal dos dados observados \mathbf{y} . Como as mulheres são independentes, basta encontrar a distribuição da contagem (y_i) do i -ésimo bloco.

$$\begin{aligned} \mathbb{P}(Y_i = y) &= \mathbb{P}(Y_i = y, Z_i = 0) + \mathbb{P}(Y_i = y, Z_i = 1) \\ &= \alpha \frac{\lambda_a^y e^{-\lambda_a}}{y!} + (1-\alpha) \frac{\lambda_b^y e^{-\lambda_b}}{y!} \end{aligned}$$

Com isto, obtemos a verossimilhança baseada apenas nos dados realmente observados

$$L(\theta|\mathbf{y}) = \prod_{i=1}^{222} \mathbb{P}(Y_i = y_i) = \prod_{i=1}^{222} \left(\frac{\alpha \lambda_a^{y_i} e^{-\lambda_a}}{y_i!} + \frac{(1-\alpha) \lambda_b^{y_i} e^{-\lambda_b}}{y_i!} \right)$$

Esta função já não é tão simples de ser maximizada (na verdade, neste toy example, ela é muito simples). O algoritmo EM vem em nosso socorro, especialmente em problemas mais complicados.

24.5 O algoritmo EM

A primeira coisa a se fazer é obter a distribuição $\mathbb{P}(\mathbf{z}|\mathbf{y}, \theta)$ dos dados faltantes \mathbf{Z} condicionados nos valores \mathbf{y} observados. Temos

$$\mathbb{P}(\mathbf{z}|\mathbf{y}, \theta) = \prod_{i=1}^{222} \mathbb{P}(z_i|y_i, \theta) = \prod_{i=1}^{222} \frac{\mathbb{P}(y_i, z_i|\theta)}{\mathbb{P}(y_i|\theta)}$$

E agora, o principal truque do algoritmo EM: como não sabemos quem é \mathbf{z} , vamos deixá-lo aleatório e tomar o seu valor esperado. Este é o passo 1 do algoritmo EM no problema de mistura: obter a log-verossimilhança ℓ^c baseada nos dados completos mas deixando os dados faltantes como variáveis aleatórias e, a seguir, calcular o seu valor *esperado* $\mathbb{E}(\ell^c)$.

24.5.1 A distribuição de $\mathbf{Z}|\mathbf{Y} = \mathbf{y}$

Mais precisamente, calculamos a log-verossimilhança $\ell^c = \log L^c$ de θ baseada nos dados completos:

$$\begin{aligned}
\ell^c(\theta|\mathbf{y}, \mathbf{z}) &= \log L^c(\theta|\mathbf{y}, \mathbf{z}) \\
&= \log \left[\prod_{i=1}^{222} \left(\frac{\lambda_a^{y_i} e^{-\lambda_a}}{y_i!} (1-\alpha) \right)^{1-z_i} \left(\frac{\lambda_b^{y_i} e^{-\lambda_b}}{y_i!} \alpha \right)^{z_i} \right] \\
&= \sum_{i=1}^{222} \left[(1-z_i) \log \left(\frac{\lambda_a^{y_i} e^{-\lambda_a}}{y_i!} (1-\alpha) \right) + z_i \log \left(\frac{\lambda_b^{y_i} e^{-\lambda_b}}{y_i!} \alpha \right) \right]
\end{aligned}$$

A seguir, substituímos os valores desconhecidos z_i pelas variáveis aleatórias Z_i fazendo com que $\ell^c(\theta|\mathbf{y}, \mathbf{Z})$ seja a variável aleatória:

$$\ell^c(\theta|\mathbf{y}, \mathbf{Z}) = \sum_{i=1}^{222} \left[(1-Z_i) \log \left(\frac{\lambda_a^{y_i} e^{-\lambda_a}}{y_i!} (1-\alpha) \right) + Z_i \log \left(\frac{\lambda_b^{y_i} e^{-\lambda_b}}{y_i!} \alpha \right) \right]$$

Precisamos agora calcular o valor esperado de $\ell^c(\theta|\mathbf{y}, \mathbf{Z})$. Observe que, em $\ell^c(\theta|\mathbf{y}, \mathbf{Z})$, estamos deixando \mathbf{y} fixado em seus valores observados na amostra. A única coisa aleatória em $\ell^c(\theta|\mathbf{y}, \mathbf{Z})$ é o vetor \mathbf{Z} . Então, ao calcular a esperança de $\ell^c(\theta|\mathbf{y}, \mathbf{Z})$ precisamos lembrar que calculamos uma esperança condicionada a $\mathbf{Y} = \mathbf{y}$. Assim,

$$\mathbb{E}[\ell^c(\theta|\mathbf{y}, \mathbf{Z})|\mathbf{Y} = \mathbf{y}] = \sum_{i=1}^{222} \left[\log \left(\frac{\lambda_a^{y_i} e^{-\lambda_a}}{y_i!} (1-\alpha) \right) \mathbb{E}(1-Z_i|\mathbf{Y} = \mathbf{y}) + \log \left(\frac{\lambda_b^{y_i} e^{-\lambda_b}}{y_i!} \alpha \right) \mathbb{E}(Z_i|\mathbf{Y} = \mathbf{y}) \right] \quad (24.7)$$

Existe outra sutileza no nosso caminho. O valor de θ na verossimilhança $\ell^c(\theta|\mathbf{y}, \mathbf{Z})$ é um valor θ arbitrário pertencente ao espaço paramétrico Θ_0 . Queremos maximizar ℓ^c com respeito a este parâmetro. Mas, ao calcular $\mathbb{E}(Z_i|\mathbf{Y} = \mathbf{y})$ em (24.7), precisamos usar *algum valor* para o parâmetro θ . Por isto, vamos usar um valor inicial $\theta^{(0)}$ para este parâmetro no cálculo desta esperança. Para deixar tudo bastante explícito, vamos usar uma notação um pouco mais carregada reescrevendo (24.7) como:

$$\mathbb{E}[\ell^c(\theta|\mathbf{y}, \mathbf{Z})|\mathbf{Y} = \mathbf{y}, \theta^{(0)}] = \sum_{i=1}^{222} \mathbb{E}\left(1-Z_i|\mathbf{Y} = \mathbf{y}, \theta^{(0)}\right) \log \left(\frac{\lambda_a^{y_i} e^{-\lambda_a}}{y_i!} (1-\alpha) \right) + \mathbb{E}\left(Z_i|\mathbf{Y} = \mathbf{y}, \theta^{(0)}\right) \log \left(\frac{\lambda_b^{y_i} e^{-\lambda_b}}{y_i!} \alpha \right)$$

Assim, queremos calcular

$$\mathbb{E}[\ell^c(\theta|\mathbf{y}, \mathbf{Z})|\mathbf{Y} = \mathbf{y}, \theta^{(0)}] = \sum_{i=1}^{180} \mathbb{E}\left(1-Z_i|\mathbf{Y} = \mathbf{y}, \theta^{(0)}\right) \log \left(\frac{\lambda_a^{y_i} e^{-\lambda_a}}{y_i!} (1-\alpha) \right) + \mathbb{E}\left(Z_i|\mathbf{Y} = \mathbf{y}, \theta^{(0)}\right) \log \left(\frac{\lambda_b^{y_i} e^{-\lambda_b}}{y_i!} \alpha \right)$$

O vetor \mathbf{Y} está fixado no seu valor observado \mathbf{y} . A esperança de Z_i usa um *valor inicial e fixo* $\theta^{(0)}$ para o parâmetro desconhecido. θ é o valor genérico do parâmetro. \mathbf{Z} é o vetor aleatório que torna a função $\ell^c(\theta|\mathbf{y}, \mathbf{Z})$ uma variável aleatória. Vamos denotar $\theta = (\lambda_a, \lambda_b, \alpha)$ e $\theta^{(0)} = (\lambda_a^{(0)}, \lambda_b^{(0)}, \alpha^{(0)})$.

24.5.2 Escolhendo $\theta^{(0)}$

O valor inicial $\theta^{(0)} = (\lambda_a^{(0)}, \lambda_b^{(0)}, \alpha^{(0)})$ pode ser obtido fazendo uma inspeção grosseira dos dados. Por exemplo, considerando o gráfico de barras para as 222 contagens na Figura 24.4, podemos chutar $\theta^{(0)} = (\lambda_a^{(0)}, \lambda_b^{(0)}, \alpha^{(0)}) = (7, 17, 0.10)$.

24.5.3 $\mathbb{E}(Z_i|\mathbf{Y} = \mathbf{y}, \theta^{(0)})$

Para calcular $\mathbb{E}(Z_i|\mathbf{Y} = \mathbf{y}, \theta^{(0)})$ lembramos que Z_i depende apenas de Y_i e que Z_i é uma variável aleatória binária. Portanto,

$$\begin{aligned}
\mathbb{E}(Z_i | \mathbf{Y} = \mathbf{y}, \theta^{(0)}) &= \mathbb{P}(Z_i = 1 | Y_i = y_i, \theta^{(0)}) \\
&= \frac{\mathbb{P}(Z_i = 1, Y_i = y_i | \theta^{(0)})}{\mathbb{P}(Z_i = 1, Y_i = y_i | \theta^{(0)}) + \mathbb{P}(Z_i = 0, Y_i = y_i | \theta^{(0)})} \\
&= \frac{\left(\frac{\lambda_b^{(0)y_i}}{y_i!} e^{-\lambda_b^{(0)}} \right) \cdot \alpha_0}{\frac{\lambda_b^{(0)y_i}}{y_i!} e^{-\lambda_b^{(0)}} \cdot \alpha_0 + \frac{\lambda_a^{(0)y_i}}{y_i!} e^{-\lambda_a^{(0)}} \cdot (1 - \alpha_0)} \\
&= \frac{\lambda_b^{(0)y_i} e^{-\lambda_b^{(0)}} \alpha_0}{\lambda_b^{(0)y_i} e^{-\lambda_b^{(0)}} \alpha_0 + \lambda_a^{(0)y_i} e^{-\lambda_a^{(0)}} (1 - \alpha_0)}
\end{aligned}$$

onde $\theta^{(0)} = (\lambda_a^{(0)}, \lambda_b^{(0)}, \alpha_0)$.

Por exemplo, considerando $\theta^{(0)} = (\lambda_a^{(0)}, \lambda_b^{(0)}, \alpha^{(0)}) = (7, 17, 0.10)$, temos

$$\begin{aligned}
\mathbb{E}(Z_i | \mathbf{Y} = \mathbf{y}, \theta^{(0)}) &= \mathbb{P}(Z_i = 1 | Y_i = y_i, \theta^{(0)}) \\
&= \frac{\lambda_b^{(0)y_i} e^{-\lambda_b^{(0)}} \alpha_0}{\lambda_b^{(0)y_i} e^{-\lambda_b^{(0)}} \alpha_0 + \lambda_a^{(0)y_i} e^{-\lambda_a^{(0)}} (1 - \alpha_0)} \\
&= \frac{17^{y_i} e^{-17} * 0.10}{17^{y_i} e^{-17} * 0.10 + 7^{y_i} e^{-7} * 0.90}
\end{aligned}$$

Note que o parâmetro desconhecido θ não está presente nesta expressão. Ele foi substituído por um valor inicial fixo $\theta^{(0)} = (7, 17, 0.10)$ um tanto arbitrário. O valor $\mathbb{E}(Z_i | \mathbf{Y} = \mathbf{y}, \theta^{(0)})$ é computável, é simplesmente um número real.

Tendo o valor de $\mathbb{E}(Z_i | \mathbf{Y} = \mathbf{y}, \theta^{(0)})$ podemos então calcular $\mathbb{E}[l^c(\theta | \mathbf{y}, \mathbf{Z}) | \mathbf{Y} = \mathbf{y}, \theta^{(0)}]$. Este último valor pode ser calculado em pontos arbitrários θ se $\theta^{(0)}$ é fixado. Vamos definir:

$$Q(\theta | \theta^{(0)}, \mathbf{y}) = \mathbb{E}[l^c(\theta | \mathbf{y}, \mathbf{Z}) | \mathbf{Y} = \mathbf{y}, \theta^{(0)}] \quad (24.8)$$

Esta expressão é crucial no algoritmo EM. Lembre-se: $\theta^{(0)}$ é um chute inicial e fixo para o parâmetro, θ é um valor arbitrário para o parâmetro e os Z 's são os rótulos dos grupos das observações. $\theta^{(0)}$ será atualizado ao longo das iterações, como explicaremos em breve.

Os dados \mathbf{y} estarão fixos ao longo das iterações do algoritmo EM. É importante perceber que $Q(\theta | \theta^{(0)}, \mathbf{y})$ é função de dois valores para o parâmetro, um valor genérico e arbitrário, não-especificado, θ e o valor inicial especificado e fixo $\theta^{(0)}$.

24.5.4 Os passos do algoritmo EM

O primeiro passo é chamado *E-step*: trata-se de obter a esperança da log-verossimilhança, a expressão $Q(\theta | \theta^{(0)}, \mathbf{y})$ onde $\theta^{(0)}$ é um valor inicial usado para calcular $E(Z | Y = y)$ e θ é um valor arbitrário θ . O segundo passo do algoritmo EM é chamado *M-step*. Lembrando que $\theta^{(0)}$ é um valor inicial fixado pelo usuário, no passo *M* encontramos o valor de θ que maximiza $Q(\theta | \theta^{(0)}, \mathbf{y})$. Isto é, encontramos o valor θ_1 do argumento θ que maximiza $Q(\theta | \theta^{(0)}, \mathbf{y})$ mantendo $\theta^{(0)}$ fixo num valor conhecido:

$$\theta^{(1)} = \arg_{\theta \in \Theta} \max Q(\theta | \theta^{(0)}, \mathbf{y})$$

No caso de mistura de Poissons, esta maximização é muito simples. Para simplificar, vamos escrever $\mathbb{E}(Z_i | \mathbf{Y} = \mathbf{y}, \theta^{(0)})$ como $\mathbb{E}(Z_i)$. Então

$$\hat{\alpha}^{(1)} = \frac{1}{222} \sum_{i=1}^{222} \mathbb{E}(Z_i) \quad (24.9)$$

$$\hat{\lambda}_a^{(1)} = \frac{\sum_{i=1}^{222} y_i (1 - \mathbb{E}(Z_i))}{\sum_{i=1}^{222} (1 - \mathbb{E}(Z_i))} \quad (24.10)$$

$$\hat{\lambda}_b^{(1)} = \frac{\sum_{i=1}^{222} y_i \mathbb{E}(Z_i)}{\sum_{i=1}^{222} \mathbb{E}(Z_i)} \quad (24.11)$$

Estas expressões são quase idênticas ao MLE no caso de dados completos. Compare as expressões acima com aquelas de (24.3)-(24.5). Veja que os z_i , conhecidos no caso completo, são substituídos pelo valor corrente de sua estimativa, $\mathbb{E}(Z_i)$, no algoritmo EM. O valor de $\mathbb{E}(Z_i)$ é atualizado tão logo as novas estimativas $\theta^{(1)}$ para o parâmetro são obtidas via (24.3)-(24.5).

24.5.5 Resumo do algoritmo EM

Começamos com um valor de $\theta^{(0)}$ inicial para o parâmetro θ . Calculamos $Q(\theta | \theta^{(0)}, \mathbf{y})$ como uma função de θ (com $\theta^{(0)}$ fixo). A seguir, maximizamos $Q(\theta | \theta^{(0)}, \mathbf{y})$ com respeito a θ obtendo $\theta^{(1)}$. O processo é iterado:

- calculamos $Q(\theta | \theta^{(j)}, \mathbf{y})$ (passo E)
- A seguir, maximizamos em θ para obter $\theta^{(j+1)}$ (passo M)

Este processo iterativo converge para o EMV de θ . A principal desvantagem do algoritmo EM é que esta convergência pode ser lenta. O que muda de problema para problema é a expressão de $Q(\theta | \theta^{(j)}, \mathbf{y})$.

Uma grande vantagem adicional do algoritmo EM é que terminamos também com uma estimativa de $\mathbb{E}(Z_i) = \mathbb{P}(Z_i = k)$, a probabilidade de cada observação pertencer ao grupo k .

24.6 Exemplos de uso do algoritmo EM

Terminar EM para o caso Poisson no R Caso normal multivariado: ver wikipedia.

24.7 Convergência do algoritmo EM

Mas por quê o algoritmo EM funciona? Existe uma prova de que o EM converge para um máximo local (ou global) da log-verossimilhança, como veremos neta seção.

24.7.1 Definições preliminares

Definição 24.7.1 — Função convexa. Uma função $g(x)$ é uma *função convexa* se a curva está sempre abaixo da secante. Outra definição equivalente: é convexa se a reta tangente em cada ponto está abaixo da curva. Ou então se a derivada $g'(x)$ é crescente. Ou ainda se a derivada segunda $g''(x)$ é positiva (ou melhor, não-negativa).

O exemplo clássico de função convexa é a parábola $g(x) = x^2$. Cheque cada uma das definições acima para verificar que esta função, de fato, é convexa. Outros exemplos clássicos é $g(x) = e^x$.

Para quê tantas caracterizações de função convexa? A razão é que, ao generalizar a definição para funções de várias variáveis, algumas das caracterizações podem ser verificadas mais facilmente.

Definição 24.7.2 — Função convexa multivariada. Uma função $g(\mathbf{x}) : A \subset \mathbb{R}^n \rightarrow \mathbb{R}$ é convexa se a matriz de derivadas parciais de segunda ordem $D^2 g$ é uma matriz definida positiva. Isto é, se $\mathbf{x}^t D^2 g \mathbf{x} > 0$ para todo ponto \mathbf{x} .

■ **Example 24.1 — Parabolóide é convexo.** Generalizando o caso da parábola, a função $g(\mathbf{x}) = \beta_1x_1^2 + \beta_2x_2^2 + \dots + \beta_nx_n^2$ com $\beta_j > 0$ é convexa pois

$$D^2g = \begin{bmatrix} \frac{\partial^2 g}{\partial x_1^2} & \frac{\partial^2 g}{\partial x_1 \partial x_2} & \dots & \frac{\partial^2 g}{\partial x_1 \partial x_n} \\ \frac{\partial^2 g}{\partial x_2 \partial x_1} & \frac{\partial^2 g}{\partial x_2^2} & \dots & \frac{\partial^2 g}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \dots & \vdots \\ \frac{\partial^2 g}{\partial x_n \partial x_1} & \frac{\partial^2 g}{\partial x_n \partial x_2} & \dots & \frac{\partial^2 g}{\partial x_n^2} \end{bmatrix} = 2 \begin{bmatrix} \beta_1 & 0 & \dots & 0 \\ 0 & \beta_2^2 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & \beta_n \end{bmatrix}$$

Agora, como D^2g é uma matriz diagonal é fácil verificar que, para todo ponto $\mathbf{x} \in \mathbb{R}^n$ e diferente de zero, temos

$$\mathbf{x}^T D^2g \mathbf{x} = \sum_i \beta_i x_i^2.$$

Se $\mathbf{x} \neq \mathbf{0}$ então pelo menos uma de suas coordenadas deve ser estritamente maior que zero e portanto, como $x_i^2 \geq 0$ e $\beta_i > 0$, temos $\mathbf{x}^T D^2g \mathbf{x} > 0$. ■

■ **Example 24.2 — Exponencial multivariada é função convexa.** Generalizando o caso da exponencial, a função $g(\mathbf{x}) = \exp(\beta_1x_1 + \beta_2x_2 + \dots + \beta_nx_n) = \exp(\mathbf{x}'\boldsymbol{\beta})$ com $\beta_j > 0$ é convexa pois

$$\nabla g = g(\mathbf{x}) \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

e portanto

$$D^2g = g(\mathbf{x}) \begin{bmatrix} \beta_1^2 & \beta_1\beta_2 & \dots & \beta_1\beta_n \\ \beta_2\beta_1 & \beta_2^2 & \dots & \beta_2\beta_n \\ \vdots & \vdots & \dots & \vdots \\ \beta_n\beta_1 & \beta_n\beta_2 & \dots & \beta_n^2 \end{bmatrix}$$

Função g é côncava se $-g$ é convexa. No caso côncavo, desigualdade é invertida. ■

24.7.2 Desigualdade de Jensen

A desigualdade de Jensen é uma desigualdade fundamental em probabilidade. Usualmente, ela aparece nos livros de probabilidade depois da definição de esperança de uma v.a. Entretanto, como seu principal uso neste livro só aparece agora, deixamos para apresentá-la mais tarde, junto com seu primeiro uso no texto.

Seja X uma v.a. qualquer com $E(X) = \mu$. Seja $g(x)$ uma função convexa. Crie uma nova v.a. $Y = g(X)$. Então a esperança dessa nova v.a., $\mathbb{E}(Y) = \mathbb{E}(g(X)) \geq g(\mu) = g(\mathbb{E}(X))$ Exemplo: $\mathbb{E}(g(X)) = \mathbb{E}(X^2) \geq g(\mathbb{E}(X)) = [\mathbb{E}(X)]^2 = \mu^2$

Função LOG é côncava: $E(\log(X)) \leq \log[E(X)]$

Notação

Seja (\mathbf{y}, \mathbf{z}) o vetor de dados completos com densidade $f(\mathbf{y}, \mathbf{z}|\theta)$. Vamos também denotar $\ell^c(\theta|\mathbf{y}, \mathbf{z}) = \log f(\mathbf{y}, \mathbf{z}|\theta)$. Seja $f(\mathbf{y}|\theta) = \int_{\mathbf{z}} f(\mathbf{y}, \mathbf{z}|\theta) d\mathbf{z}$ a densidade marginal de \mathbf{Y} . Esta é também a log-verossimilhança de θ baseada apenas nos dados observados \mathbf{y} . Isto é, $\ell(\theta|\mathbf{y}) = \log f(\mathbf{y}|\theta)$. Seja

$$k(\mathbf{z}|\mathbf{y}, \theta) = \frac{f(\mathbf{y}, \mathbf{z}|\theta)}{f(\mathbf{y}|\theta)}$$

a densidade condicional de \mathbf{Z} dados as observações \mathbf{y} . Vamos usar a letra k para denotar esta densidade condicional e assim evitar mais usos da letra f para densidades.

Escolha d número de componentes.

Como uma rotina numérica para otimização, o algoritmo EM é geralmente estável e confiável. Uma propriedade atraente é que a verossimilhança aumenta a cada iteração sucessiva do algoritmo, resultado do Teorema ???. Assim, a menos que as iterações sucessivas empurrem $\theta^{(n)}$ para infinito, o algoritmo EM vai convergir para algum valor. Mas a convergência para o valor desejado, o MLE, não é garantida: se a função de verossimilhança tem vários modas (ou pontos de máximo local), o algoritmo pode convergir para um desses máximos locais, e não para o máximo global. As condições suficientes para a convergência para o máximo global são dadas por Wu (1983). Apesar do algoritmo EM convergir em geral, a convergência pode ser lenta. Esta é um dos principais pontos fracos desse notável algoritmo.



25. Testes de Hipótese

25.1 Introdução

Teste, teste,teste



26. Seleção de Modelos

26.1 Introdução

26.1.1 The dependence of two random vectors may be quantified by mutual information.

It often happens that the deviation of one distribution from another must be evaluated. Consider two continuous pdfs $f(x)$ and $g(x)$, both being positive on (A, B) . The *Kullback-Leibler (KL) discrepancy* is the quantity

$$D_{KL}(f, g) = \mathbb{E}_f \left(\log \frac{f(X)}{g(X)} \right)$$

where the subscript on the expectation \mathbb{E}_f signifies that the random variable X has pdf $f(x)$. In other words, we have

$$D_{KL}(f, g) = \int_A^B f(x) \log \frac{f(x)}{g(x)} dx.$$

The KL discrepancy may also be defined, analogously, for discrete distributions. Note that $D_{KL}(f, g)$ may also be written in the difference form

$$D_{KL}(f, g) = \mathbb{E}_f (\log f(X)) - \mathbb{E}_f (\log g(X)). \quad (26.1)$$

In fact, the KL discrepancy is essentially unique among all discrepancies $D(f, g)$ that satisfy

- (i) $D(f, g) = \mathbb{E}_f(\varphi(f(X))) - \mathbb{E}_f(\varphi(g(X)))$ for some differentiable function φ , and
- (ii) $D(f, g)$ is minimized over g by $g = f$.

■ **Example 26.1** *Details:* When there are finitely many outcomes (so that sums replace integrals in the definition of $D_{KL}(f, g)$) it may be shown that the form of φ must be logarithmic, i.e., φ must satisfy $\varphi(f(x)) = a + b \log f(x)$ for some a, b , with $b > 0$. See Konishi and Kitagawa (2008, Section 3.1). (Konishi, S. and Kitagawa, G. (2008) *Information Criteria and Statistical Modeling*, Springer.) □ ■

In addition to having the special difference-of-averages property in (26.1), the KL discrepancy takes a simple and intuitive form when applied to normal distributions.

Illustration: Two normal distributions Suppose $f(x)$ and $g(x)$ are the $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$ pdfs. Then, from the formula for the normal pdf we have

$$\log \frac{f(x)}{g(x)} = -\frac{(x - \mu_1)^2 - (x - \mu_2)^2}{2\sigma^2} = \frac{2x(\mu_1 - \mu_2) - (\mu_1^2 - \mu_2^2)}{2\sigma^2}$$

and substituting X for x and taking the expectation (using $\mathbb{E}_X(X) = \mu_1$), we get

$$D_{KL}(f, g) = \frac{2\mu_1^2 - 2\mu_1\mu_2 - \mu_1^2 + \mu_2^2}{2\sigma^2} = \left(\frac{\mu_1 - \mu_2}{\sigma} \right)^2.$$

That is, $D_{KL}(f, g)$ is simply the squared standardized difference between the means. This is a highly intuitive notion of how far apart these two normal distributions are. \square

■ **Example 26.2 Auditory-dependent vocal recovery in zebra finches** Auditory-dependent vocal recovery in zebra finches Song learning among zebra finches has been heavily studied. When microlesions are made in the HVC region of an adult finch brain, songs become destabilized but the bird will recover its song within about 1 week. Thompson *et al.* (2007) ablated the output nucleus (LMAN) of the anterior forebrain pathway of zebra finches in order to investigate its role in song recovery. (Thompson, J.A., Wu, W., Bertram, R., and Johnson, F. (2007) Auditory-dependent vocal recovery in adult male zebra finches is facilitated by lesion of a forebrain pathway that includes basal ganglia, *J. Neurosci.*, 27: 12308–12320.) They recorded songs before and after the surgery. The multiple bouts of songs, across 24 hours, were represented as individual notes having a particular spectral composition and duration. The distribution of these notes post-surgery was then compared to the distribution pre-surgery. In one of their analyses, for instance, the authors compared the distribution of pitch and duration before and after surgery. Their method of comparison was to compute the KL discrepancy. Thompson *et al.* found that deafening following song disruption produced a large KL discrepancy whereas LMAN ablation did not. This indicated that the anterior forebrain pathway is not the neural locus of the learning mechanism that uses auditory feedback to guide song recovery. ■

The Kullback-Leibler discrepancy may be used to evaluate the association of two random vectors X and Y . We define the *mutual information* of X and Y as

$$I(X, Y) = D_{KL}(f_{(X,Y)}, f_X f_Y) = \mathbb{E}_{(X,Y)} \log \frac{f_{(X,Y)}(X, Y)}{f_X(X)f_Y(Y)}.$$

In other words, the mutual information between X and Y is the Kullback-Leibler discrepancy between their joint distribution and the distribution they would have if they were independent. In this sense, the mutual information measures how far a joint distribution is from independence.

Illustration: Bivariate normal If X and Y are bivariate normal with correlation ρ some calculation following application of the definition of mutual information gives

$$I(X, Y) = -\frac{1}{2} \log(1 - \rho^2). \quad (26.2)$$

Thus, when X and Y are independent, $I(X, Y) = 0$ and as they become highly correlated (or negatively correlated) $I(X, Y)$ increases indefinitely. \square

Theorem For random variables X and Y that are either discrete or jointly continuous having a positive joint pdf, mutual information satisfies (i) $I(X, Y) = I(Y, X)$, (ii) $I(X, Y) \geq 0$, (iii) $I(X, Y) = 0$ if and only if X and Y are independent, and (iv) for any one-to-one continuous transformations $f(x)$ and $g(y)$, $I(X, Y) = I(f(X), g(Y))$.

Proof: Omitted. See, e.g, Cover and Thomas (1991). (Cover, T.M. and Thomas, J.Y. (1991) *Elements of Information Theory*, New York: Wiley.) \square

Property (iv) makes mutual information quite different from correlation. mutual infomative versus correlation We noted that correlation is a measure of *linear* association and, as we saw in the illustration on page ??, it is possible to have $\text{Cor}(X, X^2) = 0$. In contrast, by property (iv), we may consider mutual information to be a measure of more general forms of association, and for the continuous illustration on page ?? we would have $I(X, X^2) = \infty$.

The use here of the word “information” is important. For emphasis we say, in somewhat imprecise terms, what we think is meant by this word.

Roughly speaking, information about a random variable Y is associated with the random variable X if the uncertainty in Y is larger than the uncertainty in $Y|X$.

For example, we might interpret “uncertainty” in terms of variance. If the regression of Y on X is linear, as in (17.4) (which it is if (X, Y) is bivariate normal), we have

$$\sigma_{Y|X}^2 = (1 - \rho^2)\sigma_Y^2. \quad (26.3)$$

In this case, information about Y is associated with X whenever $|\rho| > 0$ so that $1 - \rho^2 < 1$. A slight modification of (26.3) will help us connect it more strongly with mutual information. First, if we redefine “uncertainty” to be standard deviation rather than variance, (26.3) becomes

$$\sigma_{Y|X} = \sqrt{1 - \rho^2}\sigma_Y. \quad (26.4)$$

Like Equation (26.3), Equation (26.4) describes a multiplicative (proportional) decrease in uncertainty in Y associated with X . An alternative is to redefine “uncertainty,” and rewrite (26.4) in an *additive* form, so that the uncertainty in $Y|X$ is obtained by *subtracting* an appropriate quantity from the uncertainty in Y . To obtain an additive form we define “uncertainty” as the log standard deviation. Assuming $|\rho| < 1$, $\log \sqrt{1 - \rho^2}$ is negative and, using $\log \sqrt{1 - \rho^2} = \frac{1}{2} \log(1 - \rho^2)$, we get

$$\log \sigma_{Y|X} = \log \sigma_Y - \left(-\frac{1}{2} \log(1 - \rho^2) \right). \quad (26.5)$$

In words, Equation (26.5) says that $-\frac{1}{2} \log(1 - \rho^2)$ is the amount of information associated with X in reducing the uncertainty in Y to that of $Y|X$. If (X, Y) is bivariate normal then, according to (26.2), this amount of information associated with X is the mutual information.

Formula (26.5) may be generalized by quantifying “uncertainty” in terms of *entropy*, entropy which leads to a popular interpretation of mutual information.

Details: We say that the *entropy* of a discrete random variable X is

$$H(X) = - \sum_x f_X(x) \log f_X(x) \quad (26.6)$$

We may also call this the entropy of the distribution of X . In the continuous case the sum is replaced by an integral (though there it is defined only up to a multiplicative constant, and is often called *differential entropy*). The entropy of a distribution was formalized analogously to Gibbs entropy in statistical mechanics by Claude Shannon, Claude in his development of communication theory. As in statistical mechanics, the entropy may be considered a measure of disorder in a distribution. For example, the distribution over a set of values $\{x_1, x_2, \dots, x_m\}$ having maximal entropy is the uniform distribution (giving equal probability $\frac{1}{m}$ to each value) and, roughly speaking, as a distribution becomes concentrated near a point its entropy decreases.

For ease of interpretation the base of the logarithm is often taken to be 2 so that, in the discrete case,

$$H(X) = - \sum_x f_X(x) \log_2 f_X(x).$$

Suppose there are finitely many possible values of X , say x_1, \dots, x_m , and someone picks one of these values with probabilities given by $f(x_i)$, then we try to guess which value has been picked by asking “yes” or “no” questions (e.g., “Is it greater than x_3 ?”). In this case the entropy (using \log_2 , as above) may be interpreted as the minimum average number of yes/no questions that must be asked in order to determine the number, the average being taken over replications of the game.

Entropy may be used to characterize many important probability distributions. The distribution on the set of integers $0, 1, 2, \dots, n$ that maximizes entropy subject to having mean μ is the binomial. The distribution on the set of all non-negative integers that maximizes entropy subject to having mean μ is the Poisson. In the continuous case, the distribution on the interval $(0, 1)$ having maximal entropy is the uniform distribution. The distribution on the positive real line that maximizes entropy subject to having mean μ is the exponential. The distribution on the positive real line that maximizes entropy subject to having mean μ and variance σ^2 is the gamma. The distribution on the whole real line that maximizes entropy subject to having mean μ and variance σ^2 is the normal.

Now, if Y is another discrete random variable then the entropy in the conditional distribution of $Y|X = x$ may be written

$$H(Y|X = x) = - \sum_y f_{Y|X}(y|x) \log f_{Y|X}(y|x)$$

and if we average this quantity over X , by taking its expectation with respect to $f_X(x)$, we get what is called the *conditional entropy* of Y given X :

$$H(Y|X) = \sum_x \left(- \sum_y f_{Y|X}(y|x) \log f_{Y|X}(y|x) \right) f_X(x).$$

Algebraic manipulation then shows that the mutual information may be written

$$I(X, Y) = H(Y) - H(Y|X).$$

This says that the mutual information is the average amount (over X) by which the entropy of Y decreases given the additional information $X = x$. In the discrete case, working directly from the definition we find that entropy is always non-negative and, furthermore, $H(Y|Y) = 0$. The expression for the mutual information, above, therefore also shows that in the discrete case $I(Y, Y) = H(Y)$. (In the continuous case we get $I(Y, Y) = \infty$.) For an extensive discussion of entropy, mutual information, and communication theory see Cover and Thomas (1991) or MacKay (2003). (Mackay, D. (2003) *Information Theory, Inference, and Learning Algorithms*, Cambridge.)

Mutual information has been used extensively to quantify the information about a stochastic stimulus (Y) associated with a neural response (X). In that context the notation is often changed by setting $Y = S$ for “stimulus” and $X = R$ for neural “response,” and the idea is to determine the amount of information about the stimulus that is associated with the neural response.

■ Example 26.3 Temporal coding in inferotemporal cortex Temporal coding in inferotemporal cortex In an influential paper, Optican and Richmond (1987) reported responses of single neurons in inferotemporal (IT) cortexinferotemporal cortex of monkeys while the subjects were shown various checkerboard-style grating patterns as visual stimuli. (Optican, L.M. and Richmond, B.J. (1987) Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. III. Information theoretic analysis. *J. Neurophysiol.*, 57: 162–178.) Optican and Richmond computed the mutual information between the 64 randomly-chosen stimuli (the random variable

Y here taking 64 equally-likely values) and the neural response (X), represented by firing rates across multiple time bins. They compared this with the mutual information between the stimuli and a single firing rate across a large time interval and concluded that there was considerably more mutual information in the time-varying signal. Put differently, more information about the stimulus was carried by the time-varying signal than by the overall spike count. \square ■

In both Example 26.1 and Example 26.2 the calculations were based on pdfs that were *estimated* from the data. We discuss probability *density estimation* in Chapter ??.



Bibliography

Books

- [Agr15] Alan Agresti. *Foundations of linear and generalized linear models*. John Wiley & Sons, 2015 (cited on page 270).
- [Bow+97] Newton L Bowers et al. *Actuarial Mathematics, 2nd Edition*. The Society of Actuaries, Schaumburg, Illinois, 1997 (cited on page 156).
- [Dav12] Ernest Davis. *Linear algebra and probability for computer science applications*. CRC Press, 2012 (cited on page 99).
- [Fel68] William Feller. *An Introduction to Probability Theory and its Applications*. Wiley, 1968 (cited on page 126).
- [Gia12] Eduardo Giannetti. *O valor do amanhã*. Editora Companhia das Letras, 2012.
- [Jam96] Barry R James. *Probabilidade: Um Curso em Nível Intermediário, 2ª Edição*. IMPA, Coleção Euclides, 1996 (cited on pages 135, 169).
- [Ric81] Frank Proschan Richard E. Barlow. *Mathematical Theory of Reliability*. SIAM, 1981 (cited on page 163).
- [RCE10] Ernest Rutherford, James Chadwick, and Charles Drummond Ellis. *Radiations from radioactive substances*. Cambridge University Press, 2010 (cited on page 135).
- [YK58] G. Udny Yule and Maurice Kendall. *Introduction to the Theory of Statistics*. Griffin, London, 1958 (cited on pages 156, 178).

Articles

- [Ash+01] Arlene S Ash et al. “Finding future high-cost cases: comparing prior cost versus diagnosis-based methods.” In: *Health services research* 36.6 Pt 2 (2001), page 194 (cited on page 143).
- [Cla46] R D Clarke. “An Application of the Poisson Distribution”. In: *Journal of the Institute of Actuaries* 72 (1946), page 481 (cited on page 137).

- [CD81] RF Costantino and RA Desharnais. “Gamma distributions of adult numbers for *Triboium* populations in the region of their steady states”. In: *The Journal of Animal Ecology* (1981), pages 667–681 (cited on page 172).
- [DP84] B Dennis and GP Patil. “The gamma distribution and weighted multimodal gamma distributions as models of population abundance”. In: *Mathematical Biosciences* 68.2 (1984), pages 187–212 (cited on page 172).
- [ER60] P Erdős and A Rényi. “On the evolution of random graphs”. In: *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* 5 (1960) (cited on pages 127, 128).
- [Fis36] Ronald A Fisher. “The use of multiple measurements in taxonomic problems”. In: *Annals of eugenics* 7.2 (1936), pages 179–188 (cited on page 297).
- [Fis22] Ronald Aylmer Fisher. “On the mathematical foundations of theoretical statistics”. In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 222 (1922), pages 309–368 (cited on page 372).
- [Fis25] Ronald Aylmer Fisher. “Theory of statistical estimation”. In: 22.05 (1925), pages 700–725 (cited on page 372).
- [Hei+03] Grete Heinz et al. “Exploring relationships in body dimensions”. In: *Journal of Statistics Education* 11.2 (2003) (cited on page 167).
- [Hou83] Michael Hout. “Mobility tables.” In: (1983) (cited on page 258).
- [Pea00] Karl Pearson. “X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50.302 (1900), pages 157–175 (cited on page 75).
- [Res+11] David N Reshef et al. “Detecting novel associations in large data sets”. In: *Science* 334.6062 (2011), pages 1518–1524 (cited on pages 48–50).
- [SG85] Günter H Schmidt and David J Garbutt. “Species abundance data from fouling communities conform to the gamma distribution”. In: *Marine Ecology Progress Series* (1985), pages 287–290 (cited on page 172).