Statistical Foundations for Data Scientists

a view toward applications

volume I

Renato Assunção

Copyright © 2024 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 2020

Contents

Preface 11

I Part One

1 Introduction 15
1.1 The calculus of probabilities 15
1.2 Statistics: data, data, data ... 19
1.2.1 Using statistics to distinguish between models 21
1.3 Probability and Statistics: summary 23
1.4 Credit risk: data and probabilistic model 24
1.4.1 The unreasonable effectiveness of data 25
1.5 Probability models for statistical data analysis 25

2 Data and their statistical description 29
2.1 Dados Estatísticos 29
2.2 Tipologia de variáveis 30
2.3 R, uma linguagem para análise de dados 31
2.4 Dados tabulares em R 31
2.5 Vetores e resumos numéricos 33
2.6 Visualizando dados numéricos 34
2.6.1 Histograma 34
2.6.2 Ramo-e-folhas 40
2.6.3 Boxplot 41
2.6.4 Scatterplot 47

2.6.5 Scatterplot 3-dim . . . 51
2.7 Vetores ou colunas para dados categóricos 53
2.8 Objetos em R 55
2.8.1 Escalares . . . 55
2.8.2 Vetores . . . 56
2.8.3 Matrizes . . . 56
2.8.4 Dataframes . . . 58
2.8.5 Listas . . . 59
2.8.6 Funções . . . 60
2.8.7 Vetorizar sempre que possível . . . 61
2.8.8 Comando `apply` . . . 62
2.9 Material inicial sobre R 63
2.9.1 Material mais avançado em R . . . 63
2.9.2 Cursos online gratuitos . . . 63

3 Basic Probability . . . 65
3.1 Introduction 65
3.2 Probability Space 65
3.3 The probability function $\mathbb{P}$ 68
3.3.1 The Three Kolmogorov Axioms . . . 69
3.4 How to establish a function $\mathbb{P}$? 70
3.4.1 Ω is a finite set . . . 71
3.4.2 Frequentist view . . . 74
3.4.3 $\mathbb{P}$ when Ω is countably infinite . . . 77
3.4.4 $\mathbb{P}$ when Ω is uncountable . . . 78
3.4.5 Função densidade de probabilidade . . . 80
3.5 Consequences of Kolmogorov's Axioms 82
3.6 Reviewing the probability space - Optional reading 83
3.6.1 The sample space Ω . . . 83
3.6.2 Reviewing σ-álgebra $\mathscr{A}$ . . . 88
3.6.3 Reviewing the probability function $\mathbb{P}$ . . . 91

4 Conditional Probability . . . 93
4.1 Conditional Probability 93
4.1.1 What is a conditional probability . . . 93
4.1.2 Conditional Probability and Data Science . . . 93
4.1.3 Defining Conditional Probability . . . 94
4.1.4 Data science and conditional probability, again . . . 95
4.1.5 Intuition for the definition . . . 96
4.2 Venn Diagrams 97
4.2.1 Conditional Probability in Venn Diagrams . . . 97
4.2.2 $\mathbb{P}(B|A)$ and $\mathbb{P}(B)$ . . . 98
4.3 Independence of events 98
4.3.1 How does independence arise? . . . 98
4.3.2 Independence in the Venn diagram . . . 99

4.4 Bayes Rule 100
4.4.1 Bayes and diagnostic test . . . 101
4.4.2 Sensitivity and Specificity . . . 101
4.4.3 Total probability rule . . . 104
4.4.4 Extension of the Bayes Rule . . . 105
4.5 Conditional as a new probability measure 109
4.6 Mutual Independence 110
4.7 Paradoxes with conditional probability 111
4.8 Precision and recall 112

5 Introduction to Classification . . . 115
5.1 Introduction 115
5.2 Classification trees 117
5.3 Classification trees in R 121
5.4 Some examples of classification trees 122
5.4.1 Predicting myocardial infarctions . . . 122
5.4.2 Predictive models for outcome after severe head injury . . . 122
5.4.3 Autism Distinguished from Controls Using Classification Tree Analysis . . . 123
5.5 Naive Bayes 124
5.5.1 Os dados: rótulos . . . 124
5.5.2 Os dados: os atributos dos e-mails . . . 124

6 Discrete Random Variables . . . 127
6.1 Random variables: formalism 127
6.2 Random Variables and Data Tables 129
6.3 Types of random variables 129
6.4 Discrete Random Variables 130
6.4.1 Discrete R.V.: examples . . . 131
6.4.2 The σ-algebra and the probability function . . . 132
6.4.3 Cumulative distribution function . . . 134
6.4.4 Expected value $\mathbb{E}(X)$ . . . 137
6.4.5 Interpreting $\mathbb{E}(X)$ . . . 137
6.5 Main Discrete Distributions 138
6.5.1 Bernoulli . . . 138
6.5.2 Binomial . . . 139
6.5.3 Multinomial . . . 142
6.5.4 Poisson . . . 145
6.5.5 Geometric . . . 153
6.5.6 Pareto ou Zipf . . . 154
6.6 Text classification and the multinomial distribution 159
6.7 Comparison between distributions 165

7 Continuous Random Variables . . . 167
7.1 Introduction 167
7.2 Density is approximated by the histogram 169

7.3 $\mathbb{F}(X)$ in the continuous case 172
7.4 $\mathbb{E}(X)$ in the continuous case 174
7.5 Uniform Distribution 175
7.6 Beta Distribution 177
7.7 Exponential Distribution 182
7.8 Distribuição normal ou gaussiana 185
7.9 Distribuição gama 191
7.10 Distribuição Weibull 194
7.11 Distribuição de Pareto 198
7.11.1 Simulando uma Pareto . . . 200
7.11.2 Ajustando e visualizando uma Pareto . . . 201

8 Independence and Transformation . . . 203
8.1 Independence of r.v.'s 203
8.2 Como saber quando as v.a.'s são independentes? 205
8.3 Transformação de uma v.a. 207
8.4 Distribuição de $Y = h(X)$ 209
8.5 Esperança de $Y = h(X)$ 212
8.6 Probabilidades e variáveis aleatórias indicadoras 213
8.7 Multiple random variables - OPTIONAL READING 214
8.8 Transformações não-explícitas algebricamente 218
8.8.1 Distribuição dos autovalores de matriz aleatória . . . 219
8.8.2 Distribuições em alta dimensão . . . 219

9 Variance and Inequalities . . . 221
9.1 Variabilidade e desvio-padrão 221
9.2 Desigualdade de Tchebyshev 228
9.2.1 Opcional: A otimizalidade de Tchebyshev . . . 229
9.2.2 Força e fraqueza de Tchebyshev . . . 230
9.3 Outras desigualdades 230

10 Fitting Distributions to Data . . . 231
10.1 Teste qui-quadrado 231
10.2 A estatística Qui-quadrado 234
10.2.1 A distribuição de χ^2 . . . 235
10.3 Como usar este resultado de Pearson? O p-valor 236
10.4 Ajustando os graus de liberdade 238
10.5 Teste quando a v.a. é contínua 239
10.6 A função acumulada empírica 241
10.7 Distância de Kolmogorov 243
10.8 Convergência de D_n 245
10.9 Resumo da ópera 246

10.10 Teste de Kolmogorov com parâmetros estimados da amostra 246
10.11 Kolmogorov versus Qui-quadrado 249
10.12 Teste de Kolmogorov-Smirnov 249
10.13 Teste de Anderson-Darling 249
10.14 Exemplos da literatura 250
10.15 Duas colunas 250
10.16 Testes de ajustes de distribuição na prática de análise de dados 250
10.17 Prova do teste de Kolmogorovn e Kolmogorov-Smirnov 250
10.18 Como provar o resultado de Pearson? Um esboço 250

11 Monte Carlo Simulation 253
11.1 O que é uma simulação Monte Carlo 253
11.2 Geradores de números aleatórios $U(0,1)$ 254
11.2.1 Gerador congruencial misto 254
11.3 Simulação de v.a. Bernoulli 257
11.4 Simulação de v.a. Binomial 263
11.5 Simulação de v.a. Multinomial 263
11.6 Simulação de v.a.'s discretas arbitrárias 264
11.7 Gerando Poisson 265
11.8 Método da transformada inversa 267
11.9 Gerando v.a. com distribuição Gomperz 269
11.10 Gerando v.a. com distribuição de Pareto 271
11.11 Summary of the Inverse Transform Method 273
11.12 Gerando v.a. gaussiana ou normal 274
11.13 Monte Carlo para estimar integrais 275
11.13.1 Integrais com limites genéricos 276
11.14 Método da aceitação-rejeição 277
11.14.1 Dois teoremas 281
11.14.2 Sobre o impacto de M 281
11.15 Importance Sampling 281
11.15.1 The intuition 281
11.15.2 Interest quantities as expectations 282
11.15.3 Interest quantities as expectations 283
11.16 Self-Normalized Importance Sampling 287
11.17 História do método Monte Carlo 289
11.18 Aplicação em seguros: valor presente atuarial 291
11.19 Simulando um fundo de pensão 293
11.20 Processo de Poisson: sinistros no tempo 298
11.20.1 Outra abordagem 299
11.20.2 Processo de Poisson não-homogêneo 299
11.21 Provas dos teoremas: opcional 301

12 Random Vectors 305
12.1 Introdução 305
12.2 Conjunta discreta 308
12.3 Marginal discreta 308
12.4 Independência de duas v.a.'s 309
12.5 Marginal discreta com várias v.a.'s 311
12.6 Simulação de $\mathbf{X}$ discreto 314
12.7 Um outro arranjo no caso bi-dimensional 315
12.8 Um longo exemplo: Mobilidade social no Brasil em 1988 316
12.9 Condicional discreta 317
12.10 De volta à mobilidade social 321
12.11 Distribuição condicional de X 322
12.12 Exemplos de distribuições condicionais discretas 323
12.13 Esperança condicional discreta 329
12.14 Variância condicional discreta 330
12.15 Distribuição conjunta contínua 330
12.16 Definição formal de densidade 332
12.16.1 Jointly distributed random variables 335
12.16.2 Geometric probability 337
12.17 Marginal contínua 338
12.18 Condicional contínua 339
12.19 Esperança condicional 342
12.20 Variância condicional 342
12.21 Simulação de um vetor contínuo 343

13 Multivariate Gaussian Distribution 347
13.1 Normal bivariada: introdução 347
13.1.1 A distribuição condicional $(Y_2|Y_1 = y)$ 348
13.1.2 A intuição para os momentos condicionais 350
13.1.3 A densidade conjunta bivariada de $\mathbf{Y} = (Y_1, Y_2)$ 351
13.2 O desvio padronizado 353
13.3 O índice ρ de correlação de Pearson 354
13.4 Propriedades de ρ 355
13.5 Matriz de correlação 355
13.6 Propriedades de ρ 359
13.7 Estimando ρ 360
13.8 Distância Estatística de Mahalanobis 360
13.9 Autovetor e autovalor de Σ 365
13.9.1 Formas quadráticas 365
13.9.2 Matrizes positivas definidas 366
13.9.3 Autovetores e autovalores 367
13.9.4 Teorema Espectral 369

13.10 Densidade da normal multivariada **369**

14 Principal Component Analysis 371
14.1 Introdução **371**
14.2 Teoria **371**
14.3 Reconhecimento de faces com componentes principais **371**

15 Factor Analysis 383
15.0.1 Representação algébrica da análise fatorial 384
15.0.2 Resumo até aqui... 386
15.0.3 Suposições do modelo fatorial 386
15.0.4 Interpretando as cargas dos fatores 389
15.0.5 Métodos de estimação 389
15.0.6 Resumo prático 390
15.0.7 Exemplos de Johnson & Wichern 390
15.0.8 Identificabilidade do modelo 390
15.0.9 Procedimento VARIMAX 396

16 Classification 397
16.0.1 De Mahalanobis para razão de densidades 401
16.0.2 O caso geral para classificação 403
16.0.3 Expected cost of misclassification (ECM) 405
16.0.4 EMC: Expected misclassification cost 407
16.0.5 Classificação ótima com duas gaussianas 411

17 Linear Discriminant Analysis 415

18 Iterated and Conditional Expectation 419
18.1 Introduction **419**
18.2 Conditional expectation as a random variable **421**
18.3 Continuous Case **423**
18.4 The expectation of the conditional expectation **424**
18.5 Law of iterated expectation **428**
18.6 Law of iterated variance **430**
18.7 Proof of the law of iterated variance **432**

19 Limit Theorems 435
19.1 Introduction **435**
19.2 The limit of a sequence of real numbers **435**
19.3 Convergence of sequences of random variables **437**
19.4 The Strong Law of Large Numbers **440**
19.4.1 Limit of Sums of Functions of Random Variables 445
19.5 Almost surely convergence **446**
19.6 Convergence in probability **448**
19.7 Weak Law of Large Numbers **449**

19.8 Comparison Between Weak and Strong Laws **451**
19.9 Convergence in distribution **451**
19.10 Moment Generating Function **451**
19.11 Central Limit Theorem - CLT **451**
19.12 Proof of the CLT **451**
19.12.1 Moment Generating Function 451
19.12.2 The proof 451
19.13 Applications of the CLT **451**
19.14 (The Law of the Iterated Logarithm **451**

bibliography **453**
Books **453**
Articles **454**

Index **457**

Preface

Nas últimas duas décadas, houve uma importante mudança na teoria e na prática ...

Falar e agradecer sobre os alunos e colegas que colaboraram:

- Karen: material de classificação.
- Ramon Lopes: leu versões preliminares com grande atenção e ofereceu várias sugestões de melhoria.
- Raphaella Carvalho: material de análise fatorial

12.14 Variância condicional discreta
12.15 Distribuição conjunta contínua
12.16 Definição formal de densidade
12.17 Marginal contínua
12.18 Condicional contínua
12.19 Esperança condicional
12.20 Variância condicional
12.21 Simulação de um vetor contínuo

Part One

13 Multivariate Gaussian Distribution . . . 347
13.1 Normal bivariada: introdução
13.2 O desvio padronizado
13.3 O índice ρ de correlação de Pearson
13.4 Propriedades de ρ
13.5 Matriz de correlação
13.6 Propriedades de ρ
13.7 Estimando ρ
13.8 Distância Estatística de Mahalanobis
13.9 Autovetor e autovalor de Σ
13.10 Densidade da normal multivariada

14 Principal Component Analysis 371
14.1 Introdução
14.2 Teoria
14.3 Reconhecimento de faces com componentes principais

15 Factor Analysis 383

16 Classification 397

17 Linear Discriminant Analysis 415

18 Iterated and Conditional Expectation 419
18.1 Introduction
18.2 Conditional expectation as a random variable
18.3 Continuous Case
18.4 The expectation of the conditional expectation
18.5 Law of iterated expectation
18.6 Law of iterated variance
18.7 Proof of the law of iterated variance

19 Limit Theorems 435
19.1 Introduction
19.2 The limit of a sequence of real numbers
19.3 Convergence of sequences of random variables
19.4 The Strong Law of Large Numbers
19.5 Almost surely convergence
19.6 Convergence in probability
19.7 Weak Law of Large Numbers
19.8 Comparison Between Weak and Strong Laws
19.9 Convergence in distribution
19.10 Moment Generating Function
19.11 Central Limit Theorem - CLT
19.12 Proof of the CLT
19.13 Applications of the CLT
19.14 (The Law of the Iterated Logarithm

bibliography 453
Books
Articles

Index 457

1. Introduction

1.1 The calculus of probabilities

Let's start by understanding the difference between probability and statistics. Probability is a branch of pure mathematics. It allows you to make mathematical calculations about random phenomena. No need for statistical data collected in the real world. Probability is a theoretical activity, a mathematical theory that does not require empirical data.

We can summarize the functioning of this theoretical activity of the probabilist as follows. First, establish a probabilistic model that describes the phenomenon under study. Next, mathematically calculate the probability of the events of interest.

■ **Example 1.1 — Probability: a long sequence of heads.** A coin is repeatedly flipped up and the result, Heads (K) or Tails (C), is noted. What is the chance of seeing a streak of at least 8 successive heads at some point over 100 tosses of the coin? Is this such a rare event that we should be surprised if it happens? Or, on the contrary, the chance of having 8 heads in a row is great so that if we observe 8 successive heads at some point during the 100 tosses, this would be considered a novelty event.

In calculating odds, we don't need to flip the even coin even once to get the chance that we have an unbroken streak of 8 or more heads over 200 tosses. This calculation is done mathematically, based on simple rules for manipulating probabilities of elementary events.

When the number of flips of a balanced coin is small, this mathematical calculation is obtained from listing all possible configurations. For example, suppose we want to know the chance of observing a sequence of at least three successive heads over 4 tosses of a balanced coin. The list of all possible outcomes for the four throws is made up of 16 sequences:

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

There are 5 sequences with three or four successive heads: $KKKK, KKKC, KKCK, KCKK, CKKK$. As we will learn in chapter ???, the calculus of probabilities leads us to conclude that the probability

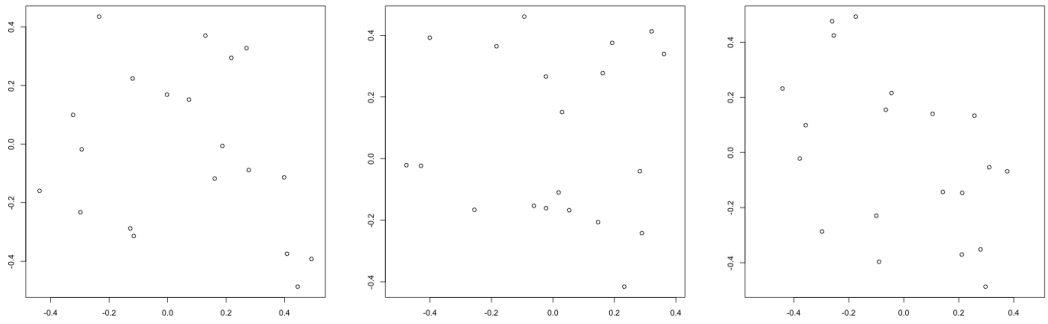


Figure 1.1: Three independent experiments of tossing n points completely at random in a square

of seeing 3 or more successive heads in a sequence of 4 tosses is equal to $5/16 \approx 0.31$, or 31% chance, a considerable value .

Unfortunately this same simple calculation based on enumeration cannot be done in this simple way when the number of coin flips starts to grow. With 100 tosses, the number of possible configurations is equal to 2^{100} and finding in this set those configurations with 8 or more heads in succession is very difficult. ??? Será???

There are special techniques for doing this calculation. Will we see in the chapter ?? how it is done but now it only matters to say that it is done with paper and pencil (or a computer), exactly as a mathematical operation of addition, multiplication and division is carried out. It is a mental calculation, without the need to carry out the physical experiment of flipping the coin and collecting the results of successive flips.

What is the use of this calculation with coins? We can use it to check whether the *hot hand* hypothesis in sports is plausible. The *hot hand* represents the belief held by fans and sports professionals that during a game some players or the entire team get hot, momentarily achieving a higher level of concentration, dexterity and skill to the point of being able to score several points in sequence in a game. For example, suppose a basketball or volleyball team plays against another and both have compatible skills. This means that, at each new point, the chance that it will come to the A team is equal to 1/2. Thus, the sequence of points in a game indexed by the label of the team that took that point would be similar to the result of tossing a fair coin successively. In fact, we want to check if this similarity is valid or if, on the contrary, the "coin" has a probability of success that fluctuates around 1/2. This fluctuation throughout the game would be such that during certain extended periods it is much greater than 1/2 (when the team enters the *hot hand* phase) and periods when it is below 1/2. If this is true, there should be prolonged periods following points where we see many successive "heads", far more than what we would expect if the coin engine is really at work.

However, several studies have shown that this is just an illusion, that successive dot patterns that seem exceptional can happen with reasonable probability if the calculation is done correctly. To fully understand this, we need to talk about statistical data collection, which is the subject of the **??** section.

■

■ **Example 1.2 — Probability: spatial model 1.** Imagine that n points are played completely at random on a square of area 1 centered on the origin $(0,0)$. See three independent realizations of this experiment in Figure 1.2. We want to know the probability that there is no point in a radius r around the origin $(0,0)$. Let's denote this probability by $\mathbb{P}_1(r)$.

All probability is a number between 0 and 1. Unlikely events have a probability close to zero,

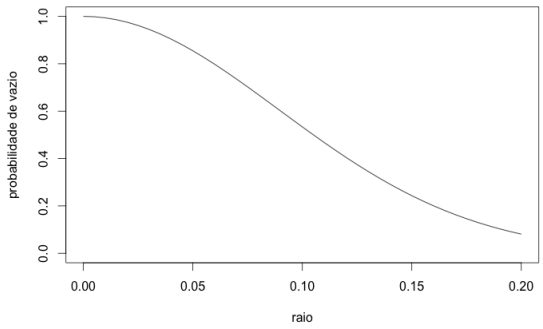


Figure 1.2: Graph $f(r) = \exp(-n\pi r^2)$, which is approximately $\mathbb{P}_1(r)$, the probability that all n random points in a unit square are at least a distance r from the origin $(0,0)$.

while very likely events have a probability close to 1. The probability $\mathbb{P}_1(r)$ depends on some aspects of the problem. For example, it is natural to expect $\mathbb{P}_1(r)$ to depend on the number n of points played at random in the square of area 1. If n is very large, it will be difficult for even one of the many points played do not end up falling inside the r circle. On the other hand, if n is very small, it is relatively easy for the disk of radius r to end up empty.

Another aspect that impacts the value of probability $\mathbb{P}_1(r)$ is the radius r. Let's keep the number n of points fixed at some value. If r is quite small (ie $r \approx 0$), it will be difficult for random points to fall inside the circle. Thus, the probability $\mathbb{P}_1(r)$ that there is no point in a radius r around the origin $(0,0)$ must be close to 1, a high probability. When r increases, the probability $\mathbb{P}_1(r)$ must decrease to zero. Through rigorous mathematical calculations, it can be shown that $\mathbb{P}_1(r)$ is approximately equal to $\exp(-n\pi r^2)$. Figure 1.2 shows the graph of this function with $n = 20$ points.

■

The probability $\mathbb{P}_1(r) \approx \exp(-n\pi r^2)$ is obtained *without statistical data*, nor computer simulation. it is a purely mathematical calculation. Figure 1.3 with the three configurations of random points is only illustrative. It was not used in the calculation of $\mathbb{P}_1(r)$. For this random point model, several other probabilities can be calculated. For example, what is the probability that there are at least 2 points in a certain region of area α? It is approximately equal to

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

It is not necessary to collect any statistical data to make this calculation.

■ **Example 1.3 — Probability: spatial model 2.** Another probabilistic model for generating points in the square leads to quite different results in calculating probabilities. For example, suppose that only 5 parent points are played completely at random on the square of area 1 centered on the origin $(0,0)$. Next, each parent point generates 4 child points so that we have 20 child points at the end. The children scatter randomly around the parents up to a maximum distance of 0.1 Consider the spatial pattern of points composed only by the children. See three independent realizations of this new experiment in Figure 1.3.

Figure 1.4 contrasts three random realizations of the first spatial model (top row) with three other realizations of the second spatial model (bottom row). There are no obvious differences between these plots.

In the case of model 2, we can also calculate the probability $\mathbb{P}_2$ of the same previous event, that there is no point in a radius r around the origin $(0,0)$. As in the 1 model, we have $\mathbb{P}_2 \approx 1$ and decreasing to zero as the radius r increases. But this probability decays quite differently in the two

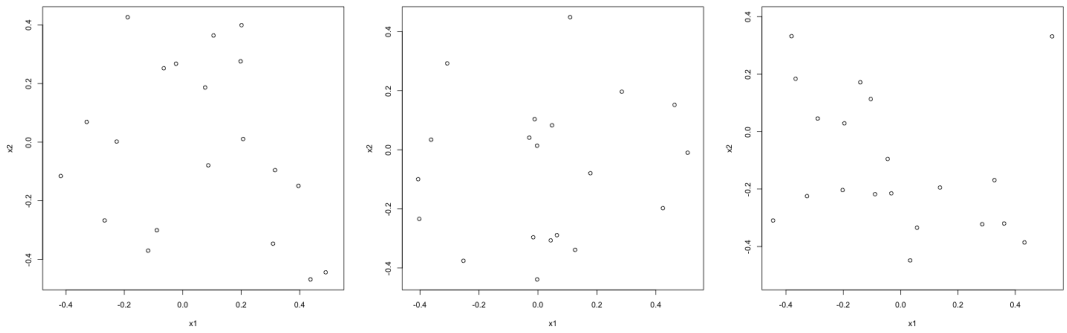


Figure 1.3: Three independent experiments of the second model of tossing n points at random in a square.

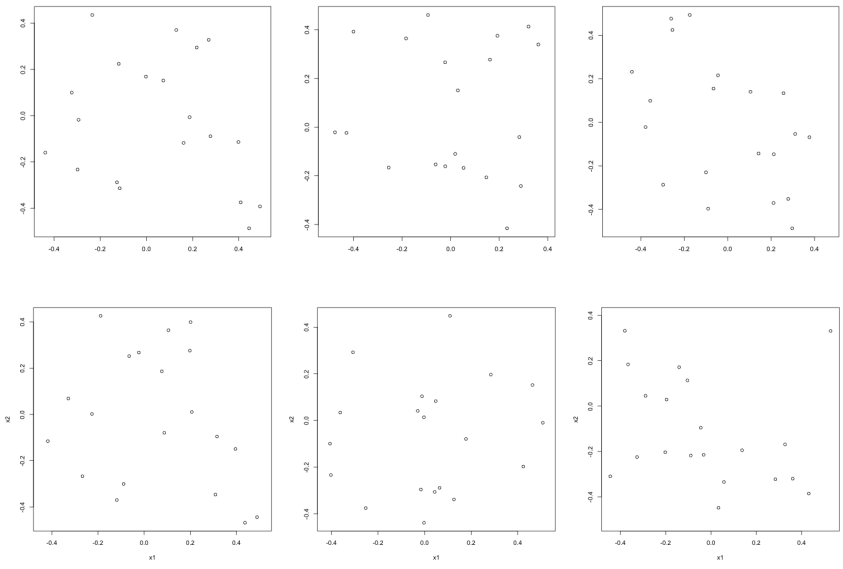


Figure 1.4: Contrasting realizations of the first spatial model (top row) with the second spatial model (bottom row).

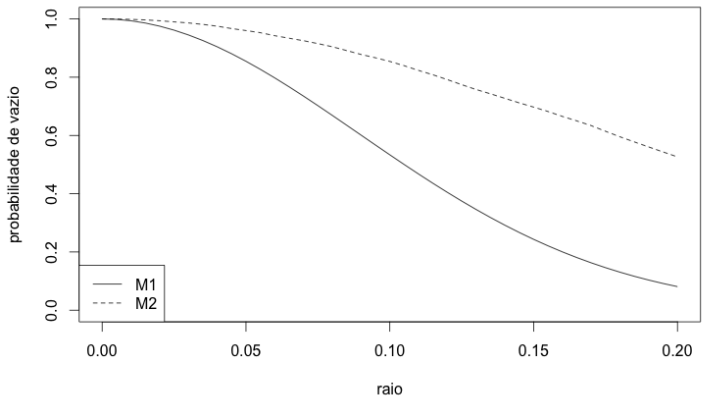


Figure 1.5: Probability that there is no point in a radius r around the origin $(0,0)$ by model 1 (M1) and model 2 (M2).

models, as can be seen in Figure 1.3. Model 2 has a much slower decay of its probability than model 1. With a radius $r = 0.15$, the probability that there is no point inside the disk is approximately 0.20 in the case of model 1 but approximately equal to 0.70 in the case of model 1. model 2 case.

To explain this big difference, note that the random child points of model 2 are strongly influenced by the positions of the few (only five) parent points. If these five parent points fall far from the center $(0,0)$, there is a high chance that there are no child points within the disk around the origin. In model 1, all the n points are played independently of each other, so there is a small chance that they all move away from the center at the same time.

■

1.2 Statistics: data, data, data ...

"Data!data!data!" he cried impatiently. "I can't make bricks without clay."

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

In contrast to probability, in statistics we are always dealing with experimentally obtained data. Statistics is a branch of applied mathematics. It needs statistical data, collected through sampling processes. What do we do with this data? We tried to infer which was the probabilistic model that generated the observed data. Thus, in probability we establish a probabilistic model and we wonder about the probabilities of various events. In statistics, nature or the world generates random data and the objective is to make inferences about the model that generated this data. This inference uses probability calculation as a tool to help identify the model that generates the statistical data. After identifying the probabilistic model that apparently generated the observed data, we wish to use this model to calculate certain probabilities. Interest is usually focused on calculating the chance of events that are possible but not yet observed. Another type of event of interest for which we would like to obtain an accurate estimate are those events that are infrequent but can carry substantial risk.

■ **Example 1.4 — Back to the long sequences.** In a basketball game, the sum of the number of points each team has is on the order of 150. If we imagine that a basket from team A represents *heads* and a basket from team B represents *tails* , we can conceive of the sequence of points in the game as the result of the successive tosses of a coin. Thus, the probability of *heads* is the probability that a basket occurred during the game belongs to team A. As team A can be better than team B

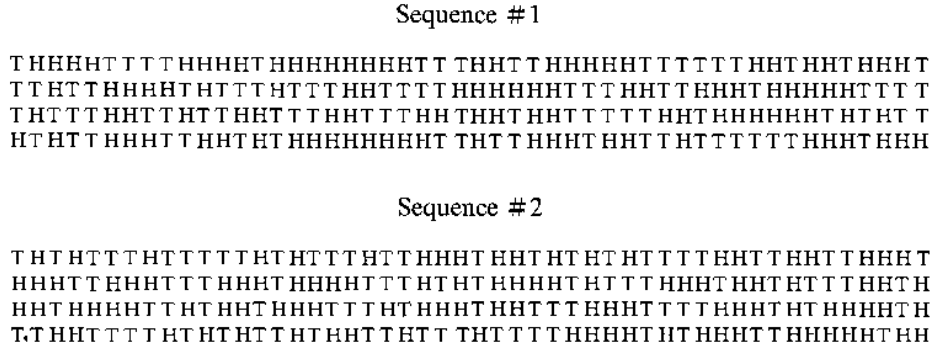


Figure 1.6: Two sequences of 200 tosses of a coin. One of them has a heads probability that changes over the toss. Which one?

we must imagine an unbalanced coin, in which the probability of *heads* can be greater than the probability of *tails*. For example, if team *A* makes 100 out of 150 baskets in a game, we can assume that the probability of *heads* is twice as high as the probability of tails.

The unbalanced currency model for representing the sequence of baskets has several consequences. If this model is a good representation, then we must conclude that the points appear as a result of a coin flip. This means that since the coin has no memory of previous tosses, the outcome of the next basket does not depend on what happened before. Thus, the probability of the next basket is the same from the beginning, it does not change as a result of the previous results. More than the influence of previous results, the probability remains static no matter what. That is, there is no mechanism making the probability of *heads* vary over the turns.

If the *hot hand* hypothesis is true then this coin model is not valid for representing the results of a game. That is, with *hot hand* operating the occurrence of baskets from team *A* are not like *heads* results in successive flips of an unbalanced coin. The *hot hand* hypothesis leads us to think of another model to represent the game, a model in which the probability of *heads* fluctuates during the game, not remaining constant. It could fluctuate as a result of past game results or due to factors outside the game. In any case, it would not remain constant, destroying the proximity to the coin model.

Which one is the correct one? It's not as easy to guess as you might initially think. The two sequences shown in Figure **??** represent the result of 200 tosses of a well-balanced coin, with the chance of *heads* (or heads, *H*) being equal to *tails* (or tails, *T*). However, while one of them actually represents the balanced coin, the other was generated by a mechanism whereby the probability of the next toss changes as a function of the number of heads in previous tosses. Try to quickly guess which of these two is the "false" sequence.

This example is inspired by Schilling (ref ??). He explains that Révész, a probabilist, used to divide his students into two blocks. Each student in block 1 should flip a coin 200 times and record the sequence obtained. The second block should try to generate, from its own head, a sequence that resembles, as best as possible, what would be obtained by flipping a coin. The sequences of the two groups were scrambled and a key with the identification of the group was kept. He would then look at the generated sequences to see if they conform to what would typically be observed with balanced coin flips. He could almost always correctly identify which group had generated each of the sequences.

The "false" sequence, the one that was not obtained with a balanced coin being tossed in succession, is the second. How could this be identified? We'll see this later in this course.

■

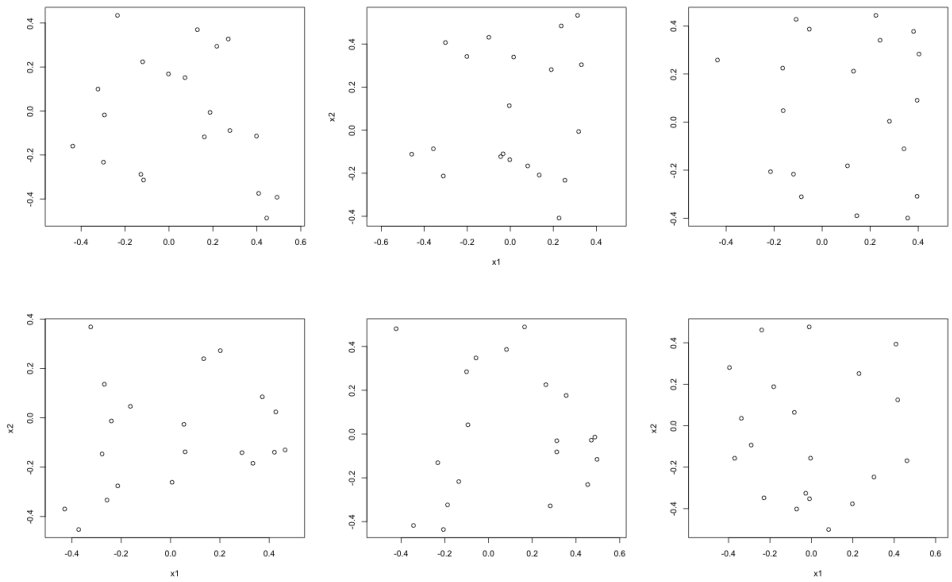


Figure 1.7: Six random realizations, some obtained with the first spatial model and some with the second spatial model. Can you identify which is which?

■ **Example 1.5 — Back to the spatial models.** In Figure 1.5, we have six random realizations of points played in the unit square. Some of them were obtained with the first spatial model described above while the others were obtained with the second spatial model. Which of the two models generated each of the six plots in this figure?

This is not a simple task as the differences between the point patterns generated by the two models are not very large. As each figure was generated by me, I know which model was used in each case. Figure 1.5 reveals the model behind each of the plots in Figure 1.5.

■

1.2.1 Using statistics to distinguish between models

The purpose of this section is to illustrate the difference between mathematical probability calculation and statistical data analysis. Let's use a statistical test to discriminate between the two models behind each plot in Figure 1.5.

For each random point, I found the distance to its nearest neighbor point. Fixing a radius r, I counted the proportion of points in a plot that had a distance smaller than r. For example, for $r = 0.10$, I got the proportion G of observed points that had their nearest neighbor within a distance of less than 0.10. Considering several different radii r in each plot, I obtained the proportion of points that had a distance smaller than r.

Then, by calculating probabilities, without using statistical data, with pure mathematics, I got limits (m, M) such that, if the data actually comes from model 1, the value of the proportion G should be between m and M with very high probability. If out of bounds, model 2 should be correct.

Figure 1.9 shows how the statistical test is performed. On the horizontal axis we have the radius r. The vertical axis represents the probability values of the distance to the nearest neighbor being less than r. The two dashed lines were obtained with probabilistic, purely mathematical calculations. We won't bother to describe these calculations right now. The dashed lines represent limits for the empirical curve $G(r)$, represented by the solid line. This empirical curve was obtained by calculating the proportion of G of observed points that had their nearest neighbor at a distance smaller than r. I made this calculation using some values $r_1 < \ldots < r_k$ of r obtaining the proportions

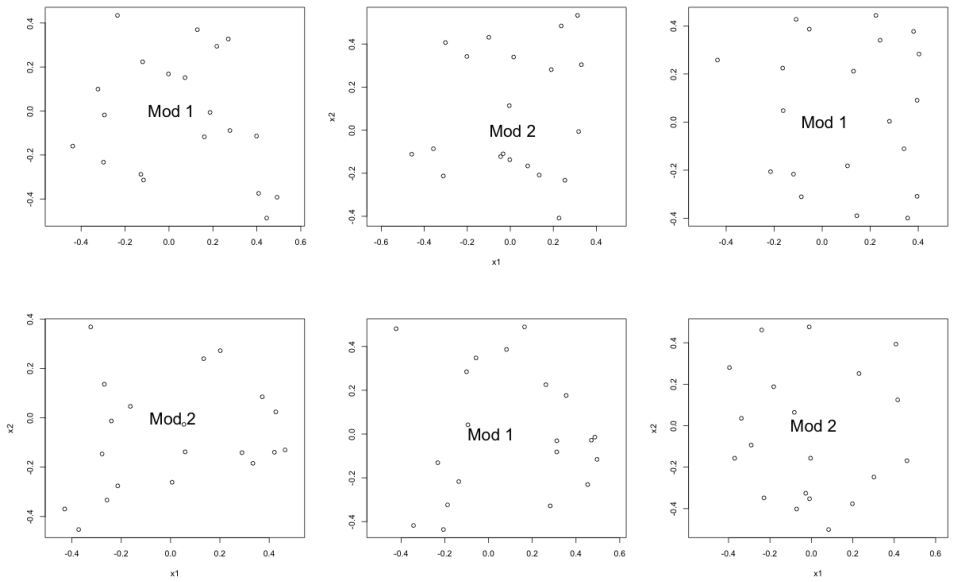


Figure 1.8: Identifying the spatial model behind each plot of Figure 1.5.

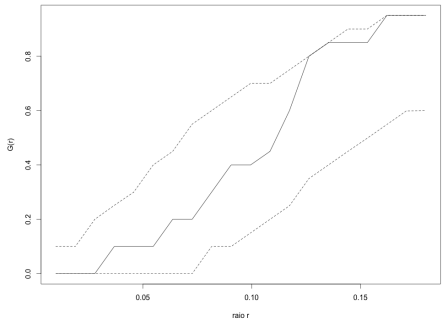


Figure 1.9: Empirical curve $G(r)$ (solid line) and its limits under model 1 (dashed lines) versus radius r.

$G(r_1) < \ldots < G(r_k)$. The empirical curve $G(r)$ is obtained by connecting the points $(r_j, G(r_j))$ with line segments.

In Figure 1.9 we see the empirical curve $G(r)$ within the limits determined by the dashed lines. The statistical test then recommends inferring that Model 1 was used to generate the corresponding spatial data. The reasoning is that, if model 1 is correct, the empirical curve $G(r)$ should be between the dashed lines with high probability. On the other hand, if model 2 is the model that is generating the data, the empirical curve $G(r)$ would tend to leave the dashed limits. The test can lead to errors: model 1 can be correct and still the empirical curve $G(r)$ goes out of its limits. The opposite can also occur: model 2 is correct but the empirical curve $G(r)$ is within the dashed limits. However, the dashed lines are calculated so that this does not happen too often. That is, the dashed lines are obtained in such a way that, with very high probability, the empirical curve $G(r)$ falls within its limits if model 1 is correct.

Let's emphasize once more: the dashed lines were obtained assuming that model 1 is the correct one and this was done by calculating probabilities, *no data*. The continuous curve $G(r)$ represents an empirical calculation, based on experimental data. The value of $G(r)$ is a proportion calculated

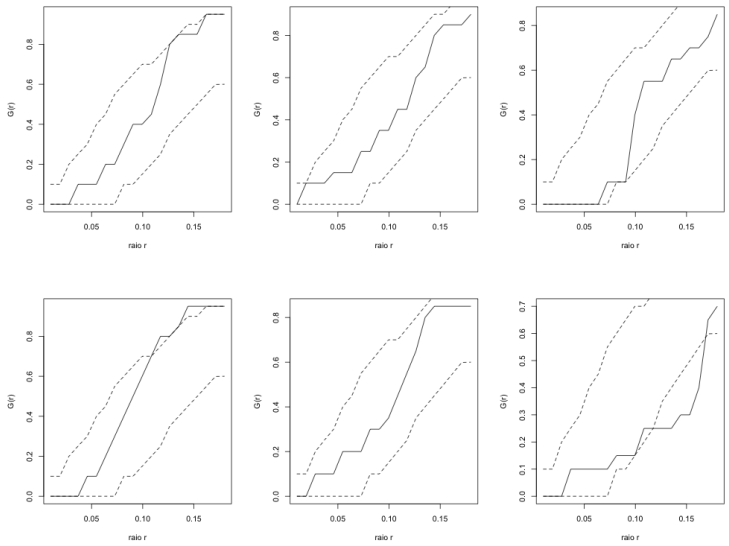


Figure 1.10: The result of applying a statistical test to each of the plots in Figure 1.5.

from the statistical data, the proportion of points whose distance to the nearest neighbor point is less than r. If the continuous curve $G(r)$ calculated with the data falls within the dashed bounds obtained by model 1 theory, then we bet that model 1 generated the observed spatial data. If it goes out of bounds, we bet on model 2 to be the data generator.

Figure 1.2.1 shows the result of applying this statistical test to each of the plots in Figure 1.5. Remember that the identity of the probabilistic model that actually generated these spatial patterns was revealed in Figure 1.5. Considering the decision recommended by the statistical test in Figure 1.2.1, we see that we would only make a wrong decision in the plot $(1,2)$, whose data are generated by model 2.

1.3 Probability and Statistics: summary

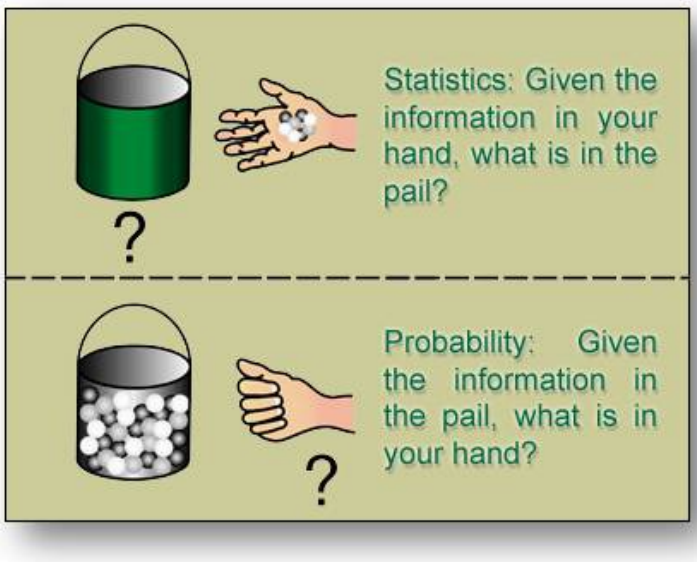
Probability: From a mathematical model of a phenomenon that generates non-deterministic results, the calculation of probabilities allows the mathematical deduction of the probability of the occurrence of several events. You don't need any statistics for this.

Statistics: with data collected in the real world, we build a table of numbers. We want to find out which probabilistic model generated these data. Once the model has been identified, we want to use it to calculate the probability of the occurrence of possible events that have not yet been observed or that are not very frequent.

The image in Figure 1.3 shows the differences between statistics and probability. This image is taken from `http://herdingcats.typepad.com/my_weblog/`.

Although different, statistics and probabilities feed each other. One could not exist healthily without the other. The real world problems for which we want to calculate probabilities inspire the development of probabilistic theories and concepts, as well as the obtaining of mathematical theorems about these probabilities, often contradicting our intuition. On the other hand, the ability to create models and make probabilistic calculations increasingly complex, leads us to collect data with sophisticated structures and then be able to infer which of these models are operating in reality. In the next section we present a well-established example of using data to establish a sophisticated prediction model for granting financial credit.

Figure 1.11: Difference between statistics and probability.



1.4 Credit risk: data and probabilistic model

Customers apply for credit or borrow from financial agents. These agents want to know, for each customer, whether he will pay back the loan made on time. A *credit risk* model assesses the probability of this occurring given that the customer has certain attributes. If the probability is low, he is a potential risk and credit should be denied. If the probability is high, credit should be granted.

In fact, this probability is highly customized and must depend on several aspects related to the customer and the business environment. For example, the probability of paying back the loan on time should depend on the average balance on the customer's account relative to the loan amount. If the customer wants a loan that represents 25% of what he usually has in his account balance or if he wants a loan that is 10 times his average balance, the risk of default appears to be higher in the second case.

Many other factors must affect the calculation of this probability. How many years has the borrower been a customer of the institution? What is your past history in terms of loans and your payments? Is the economic environment one of growth and therefore favorable to new investments or is it an environment of recession?

We need a probability model to make these calculations. There are many possible models currently being used by financial institutions. Some are better than others because they can better predict what customers will do in the future.

What data is needed to identify such a probability model? A statistical sample of those who took out a loan is sought among the bank's recent customers. For each of these customers, a binary response Y is noted:

- $Y = 1$ if the customer paid back in due time.
- $Y = 0$ otherwise.

In addition, we have a set of attributes that can influence the behavior of these customers. For each customer in the sample, note the following characteristics that could potentially affect the likelihood of repaying the borrowed loan on time:

- 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

Do we really need a probabilistic model? In the days of big data, doesn't data answer everything? After all, we can do straightforward and simple calculations from the data directly.

For example, what is the probability that a customer over 60 years old and an average balance greater than 5 thousand reais does not pay the credit? Separate the sub-sample of customers over 60 years old and balance greater than 5 thousand. If this sub-sample is not too small... (say, greater than 1000 individuals)... Among the individuals in this sub-sample, obtain the proportion of those who did not pay the credit. This proportion is approximately the probability of non-payment. Very simple, just count in the database.

It's not always that simple. The customer has many attributes, not just age and average balance. For each customer, we have more than 15 attributes. If each attribute has only two possible values, we have $2^{15} = 32768$ customer settings. In each of these possible configurations, we want the probability of non-payment. We need at least 100 individuals in each configuration to estimate the probability. This gives 32768000, or over 32 million individuals in the database.

There is simply no base with recent customers of this size for this problem. Suppose there is no individual in the database with age x, balance y, etc. Or maybe there are only 3 individuals with these attributes. How to properly estimate the probability of non-payment of a new customer with these attributes?

Another situation where things are not so straightforward for a statistician is when the event of interest is relatively rare. Consider, for example, the financial losses associated with typhoons in Taiwan. What is the probability of a typhoon causing more than 4 million losses in the next 10 years? Since there is no typhoon that so far has caused a loss greater than 4 million, should we estimate this probability to be zero?

■ **Example 1.6 — Another example.** Data $T_1, T_2, \ldots, T_n$: the survival time of n patients undergoing a new medical treatment. We want to estimate the expected time $\mathbb{E}(T)$ of survival after treatment. Simple: take the arithmetic mean of the observed n times.

Assume that the experiment needs to provide an estimate one year after the start of the study. One year after the study, 50% of the patients died (and therefore the value of T for these individuals is known). But 50% have not died yet and T is not known for these other individuals. The mean of the known values will tend to underestimate the expected survival value. How to do in this case?

■

1.5 Probability models for statistical data analysis

We need a *conceptual statistical model*.

Definition 1.5.1 — Conceptual Statistical Model. A hypothetical probability distribution describing how the observed data could have been generated.

Modelling is the design of a mathematical framework capable of generating the data. The data that interest us are not deterministic. So this mathematical model is usually a probabilistic or stochastic model. Let's list some of the desired properties of a good statistical model.

Figure 1.12: Financial losses in rice farms caused by typhoons in Taiwan.

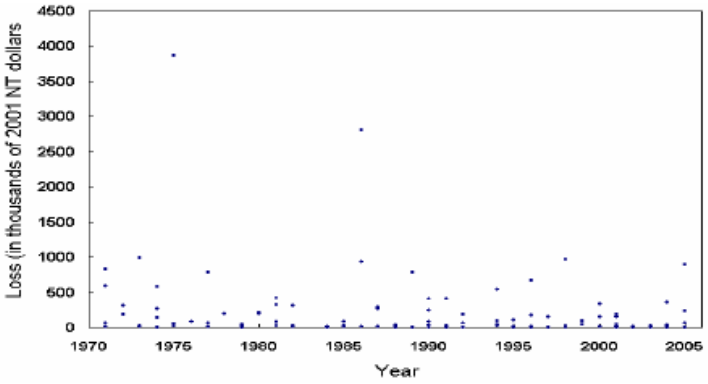


Figure 1. Scatter plot of Taiwan typhoon rice loss

The probabilistic model must be able to simulate data with statistical characteristics similar to those observed in reality. For example, it must be able to predict events more or less well that actually occur in reality. The model proposes a plausible mechanism, which corresponds in some sense to what actually happens in reality. A plausible mechanism might suggest interventions or actions that alter reality in some desired way (preventing disease and fraud, for example). Finally, the model must be easily manipulated mathematically and conceptually. We need to do probability calculations with the model. If it's too complex, we won't be able to do that.

Properties often conflict

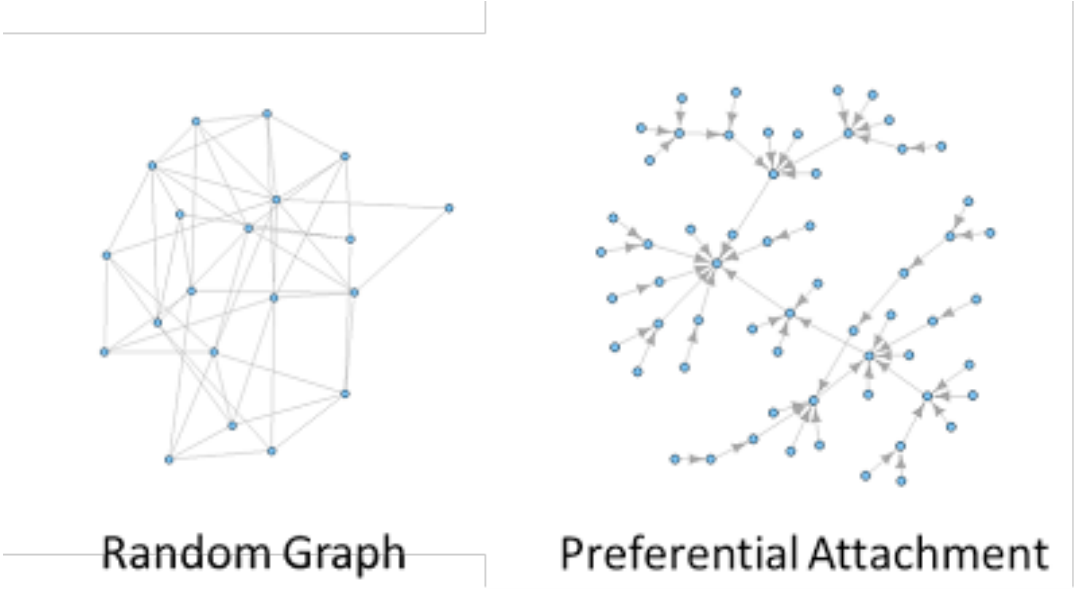
It is often not possible to have all three properties simultaneously. For example, a model to generate data that is very realistic may have to become very complicated. This means that it will likely be difficult to analyze mathematically. Therefore, it may be reasonable to consider models that reproduce only some of the characteristics of the underlying data. We want to reproduce in the model the main features that we are most interested in at the moment. The modeling process is often difficult, requires experience, and is often both a science and an art.

Models for what?

Why are we interested in building mathematical models for our observed data? A good model gives some meaning to our data and helps to roughly understand the mechanism by which the data is created. Often, the model is just a caricature of the real situation. A caricature is a drawing of a real-life character that emphasizes and exaggerates some of the person's physical or behavioural characteristics in a humorous way. In general, by far the caricature is a faithful portrait of the individual. However, she represents him in such a way that, when we see the caricature, we immediately recognize who it is. Models are like caricatures: they capture the essentials of representing a situation so that model properties can then be applied to the real situation.

■ **Example 1.7 — Models for complex networks.** One active area of data analysis is interested in complex networks represented by real-world networks such as social networks, computer networks, biological networks, and brain networks. They are all represented by nodes or vertices connected by edges. A recurring aspect of real-word networks is that most of their vertices have a small number

Figure 1.13: The two models of social networks: example of achievements. Pólya-Erdös model (left) and Barabási-Albert model (right).



of incident edges. However, a few vertices have many edges (they are the hubs of the network). Let $\mathbb{P}(K)$ be the probability that a vertex has k edges. Empirically, when analysing their statistical aspects, we almost always find that $\mathbb{P}(K) \approx c/k^{\gamma}$ where c and γ are constants. For example, we may find that $\mathbb{P}(K) \approx 2/k^3$. This is called a power-law probability distribution (inverse power of k). How can this type of probability distribution for the number of connections appear in practice?

The Pólya-Erdös model

Suppose each pair of vertices tosses a coin in the air. If heads, a link is established between them. If it comes tails, they don't bond. By mere chance, some vertices will have a greater number of links than others.

However, this model is not capable of generating the power-law characteristic of reality. The number of links has little variation around the average, never generating the dominant hubs that we see in real cases. The nodes will all have approximately the same number of edges. This is not a good model for the complex networks of reality.

Preferential attachment model

The Barabási-Albert *preferential-attachment* social network model is a much better alternative than the Pólya-Erdös model. Start with a few vertices randomly linked together by the previous model. Produce new vertices sequentially A new vertex connects to an existing node with a probability proportional to the number of edges the old node already has.

Figure 1.13 shows two examples of social networks generated from the two models, the Pólya-Erdös model (left) and the Barabási-Albert model (right).

The Barabási-Albert preferential attachment model is not a perfect model for real complex networks. But it induces a distribution in the degrees of the vertices of complex networks that has a power-law form, with a heavy tail. We have in hand, then, a hypothetical mechanism that produces a very visible and characteristic aspect of complex networks. We have a caricature of the actual generating process of complex networks.

■

The intention of the data analyst is to formulate a simple but non-trivial mathematical structure that represents the essential and most relevant aspects of the random phenomenon of interest.

Similar to a caricature, a good probabilistic model is not a faithful and perfect portrait of a real situation, but a sketch that reproduces and even amplifies or exaggerates its most striking features in order to make it easily recognizable.

Another use of a good model is to make predictions. A good credit risk rating model will do this. Based on various characteristics (or *features*, in English) of a user, we can predict whether or not he will pay a possible loan on the agreed date. This is done with historical data: we have a huge collection of individuals who borrowed and what the result was ($Y = 1$, paid; $Y = 0$, did not pay). For each individual, we also have its characteristics collected as a vector $\mathbf{x}$. Some of the characteristics are: gender, age, time as an account holder, average balance, etc.

With these statistical data, we find a model for $\mathbb{P}(Y = 1|\mathbf{x})$. That is, a model for the probability of paying given that it has the characteristics $\mathbf{x}$. This model is used to predict the behavior of future borrowers. A customer with the characteristics $\mathbf{x}$ arrives and asks for a loan. Calculate $\mathbb{P}(Y = 1|\mathbf{x})$ using the model. If the probability is low, do not grant the loan.

Decision making based on predictions appears all the time. Should we grant the loan? Offer a discount to a customer if they are likely to buy a very expensive item. Cutting the connection to a network if there is a chance that certain activities on the network are hacking. Build a new weather station at location (x, y) if this position minimizes the uncertainty of forecasts for the region as a whole from the existing network plus the new station.

We may be experiencing the beginnings of a period in which decision-making will be radically transformed. In an insightful paper, [24] describes this moment starting from a commentary of the economist Tim Bresnahan in the book *Predicition Machines*: "computers do arithmetic and nothing more. The advent and commercialization of computers made arithmetic cheap. When arithmetic became cheap, not only did we use more of it for traditional applications of arithmetic, but we also used the newly cheap arithmetic for applications that were not traditionally associated with arithmetic." [24] then complements this thought: "By reducing the general cost of something important (in this example, computation), technology changed the world. Similarly, developments in AI are transformative insofar as they are causing a dramatic drop in the price of something that is foundational to decision-making: useful prediction. This is true despite the reality that the technologies are a long way from being done... Because it so efficiently draws upon the infrastructure of past revolutions–from electricity, to the computer revolution, to the rise of the intensely networked era that began in the 1970s, known as the Internet revolution–a whole world of AI applications is emerging around us. The net effect is a massive price drop for useful prediction."

These words still reverberate with myself. Just think about how cheap became some tasks: Waze predicting right now the best route to commute; Spotify predicting what I will like to listen during the week (and getting it right!); Twitter predicting the news that will interest me most; GoogleAds showing me some products that I really need, etc. We can imagine that soon enough we will be able to cheaply make very personalized predictions about health consequences of taking specific medicine or having micro-robots monitoring vital signals and predicting emerging problems before they become unmanageable. Although "Prediction is very difficult, especially if it's about the future" (Niels Bohr), there is one that is easy: statistical thinking and techniques will be crucial elements in this brave new world. I hope you obtain a good knowledge of them in the rest of this book.

2. Data and their statistical description

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.

Essas 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 predições, 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.

De uma maneira mais formal (mais ainda vagamente descrita), nós vamos procurar criar uma função matemática g que tem como parâmetro de entrada as carecterísticas da mensagem representadas por um vetor $\mathbf{x}$ onde $\mathbf{x} = (\texttt{num_ char}, \texttt{line_ breaks}, \ldots, \texttt{\%obs})$ e cuja saída seja a variável $Y = \texttt{spam}$.

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. `ratioti` 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, subtraimos 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 *nminais*, 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

Ela 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"
```

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 / Height2 *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.

```
[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/Ng/ssc2003/BPMain.htm`. O dataframe `pressao` contem 500 pacientes, cada um deles numa linha da tabela. Estas são os casos ou instâncias ou observaccões 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

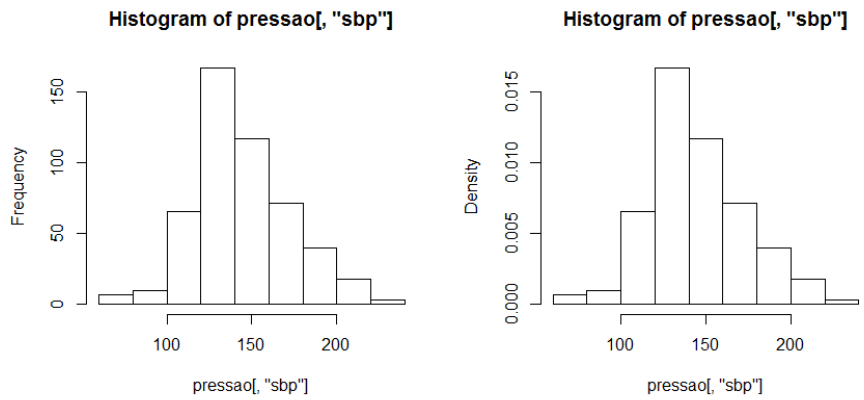


Figure 2.1: Histograma dos 500 valores da variável `sbp`, pressão sistólica, vindos da tabela 2.2.

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 egião 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 retâgulo 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 obte-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.

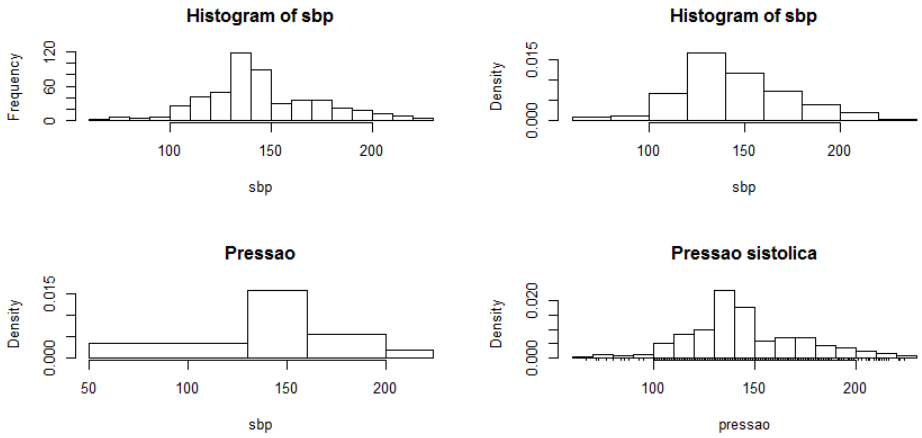


Figure 2.2: Plots mostrando algumas opções ao usar o comando `hist`.

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

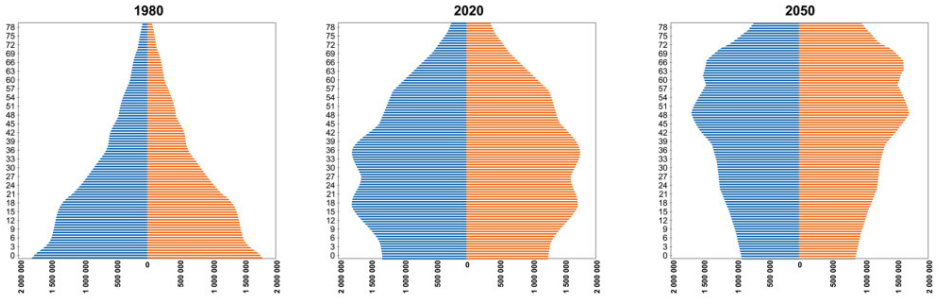


Figure 2.3: Pirâmides etárias do Brasil em 1980 e projeções em 2020 e 2050. Fonte: IBGE.

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 90^o. 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 compreende-los melhor. 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")
```

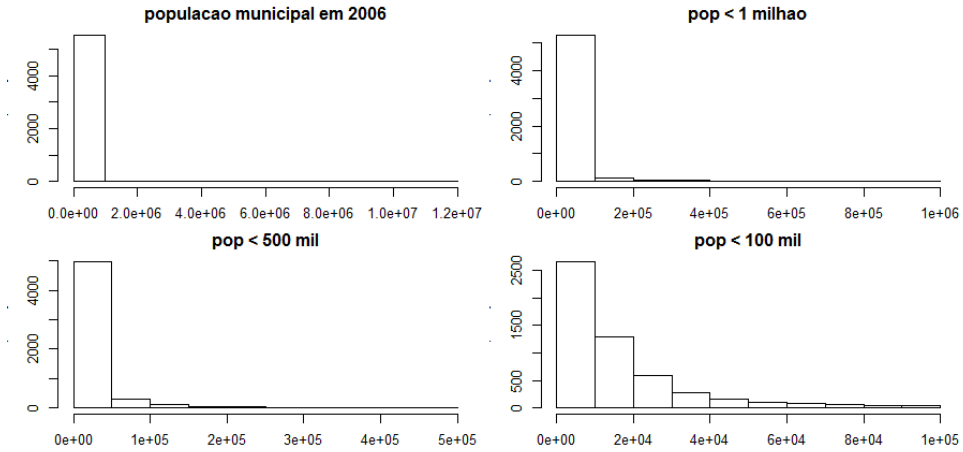


Figure 2.4: População dos 5564 municípios brasileiros em 2006. Fonte: IBGE.

```
> 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 menos 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 possui menos de 40 mil habitantes. De fato, a proporção exata é dada por

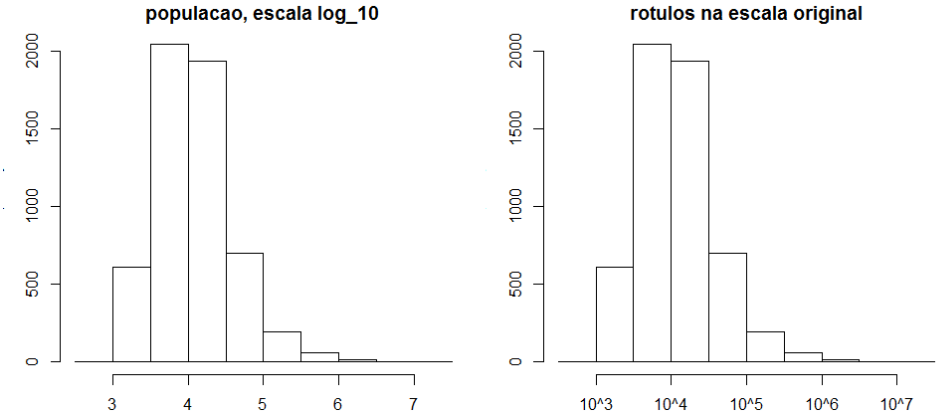


Figure 2.5: População na escala logarítimica (base 10) dos 5564 municípios brasileiros em 2006. Fonte: IBGE.

```
> 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 logarítimica. 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* $3, 4, \ldots, 7$ (e apenas esses rótulos) foram substituídos pelos rótulos $10^3, 10^4, \ldots, 10^7$ para ajudar a entender melhor o que as barras representam na escala original de tamanho populacional.

O que significa olhar um gráfico na escala logarítimica? 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-populaçção 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 logarítimica? 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, $A \to B$. Para ter um impacto similar a este primeiro, talvz 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     Internacional  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 $31, 31, 34, 36, \ldots, 39$. 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

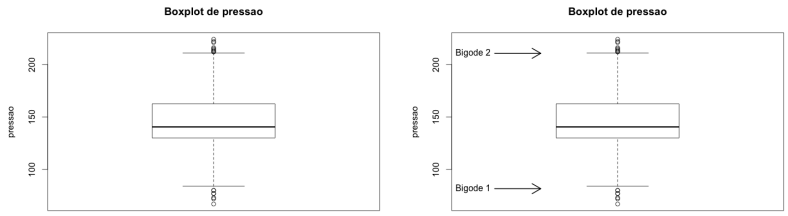


Figure 2.6: Boxplot usando os dados da variável `sbp`, pressão sistólica, vindos da tabela 2.2.

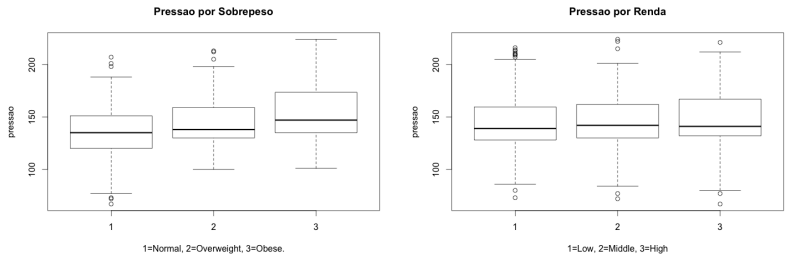


Figure 2.7: Boxplot de pressão sistólica versus a categoria de peso do indivíduo (esquerda) e versus o seu nível de renda (direita).

bigode. Caso o máximo dos dados seja menor que o limiar do bigode superior, o bigode vai apenas até o próprio máaximo 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 visualizaąão 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 dos indivíduos obsesos 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 etre 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 deslocamente 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:

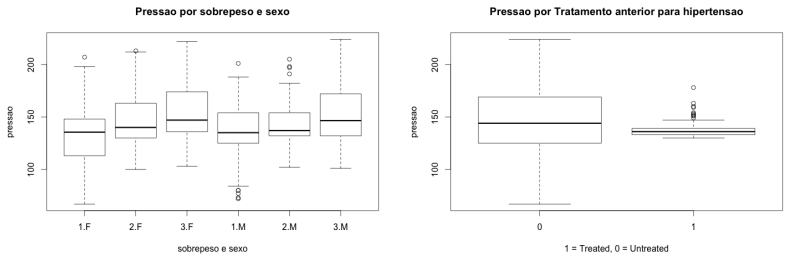


Figure 2.8: Esquerda: Pressão sistólica para as 6 categorias criadas cruzando as variáveis categóricas sobrepeso e sexo. Direita: Pressão sistólica para dois grupos de pacientes, aqueles que receberam um tratamento anterior para hipertensão e aqueles não tratados.

```
> par(mfrow=c(1,2))
> boxplot(sbp ~ overwt, ylab="pressao", xlab="1 = Normal, 2 = Overweight, 3 = Obese.")
> title("Pressao por Sobrepeso")
> 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 sobrepeso e sexo. Este gráfico foi criado com os comandos seguintes:

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

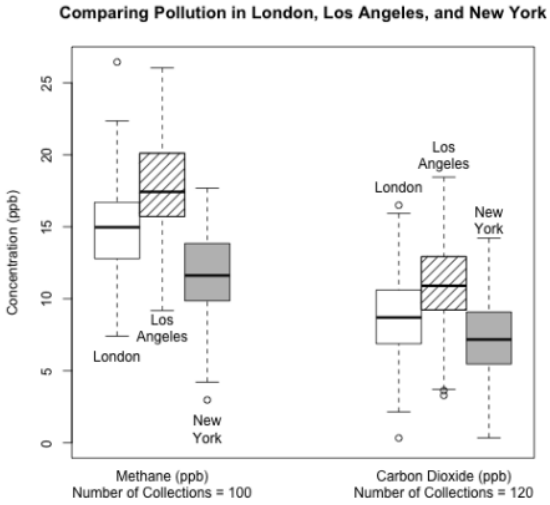


Figure 2.9: Níveis de poluição de dois gases ligados ao efeito estufa, CO2 e CH4, em três grandes cidades: Londres, Nova York e Los Angeles. Fonte: `http://bit.ly/2np7ikU`.

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 estálocalizada), 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)
```

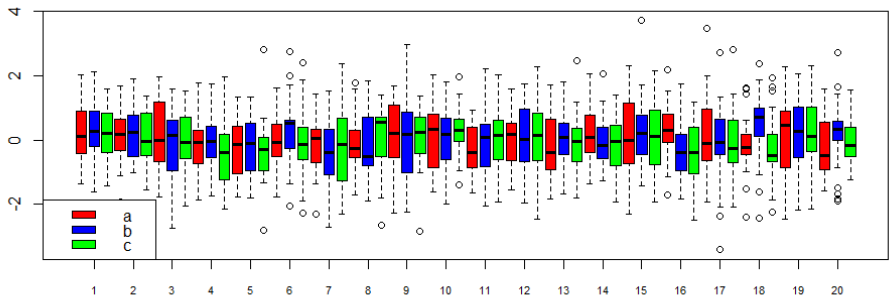


Figure 2.10: Box-plots para cada uma de muitas categorias. Cada grupo de 3 boxplots representa são mostrados em grupos de 3. Fonte: `http://bit.ly/2nGyg3G`

```
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. O valores dos indivíduos tratados quase não variam em trono 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 menos 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 FiveThirtyEigth, `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 $Q1$ e $Q3$) 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.

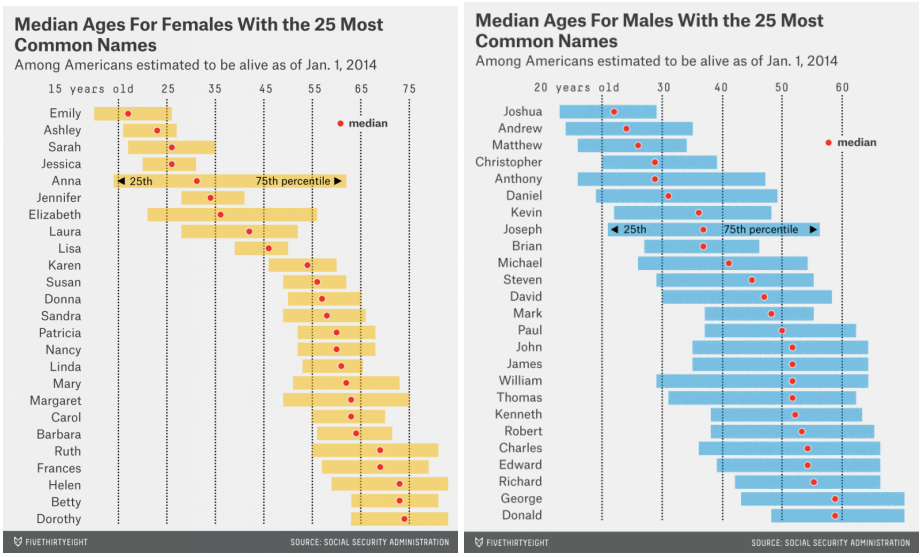


Figure 2.11: Boxplot das idades em 2014 das mulheres (esquerda) e homens (direita) que possuíam um dos 25 nomes femininos ou masculinos mais populares nos EUA. Fonte: site *FiveThirtyEigth*, `http://53eig.ht/2mHEmAE`.

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 vivasem 2014 tenham menos de 14 anos (nascidas a partir de 2000). Já o nome Joseph, elativamente 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.

■

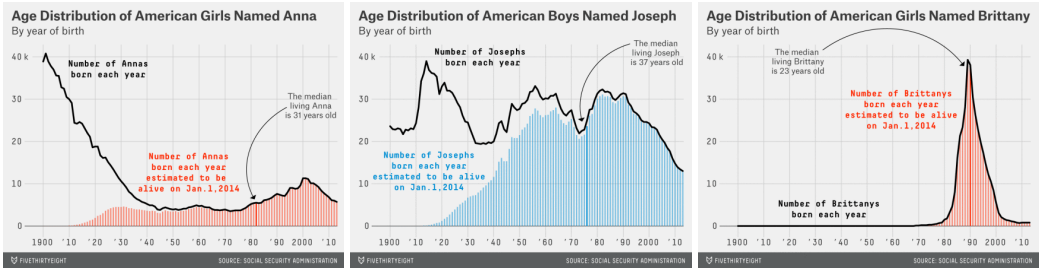


Figure 2.12: Annas, Josephs e Brittanys ao longo do tempo e em 2014. Fonte: site *FiveThirtyEigth*, `http://53eig.ht/2mHEmAE`.

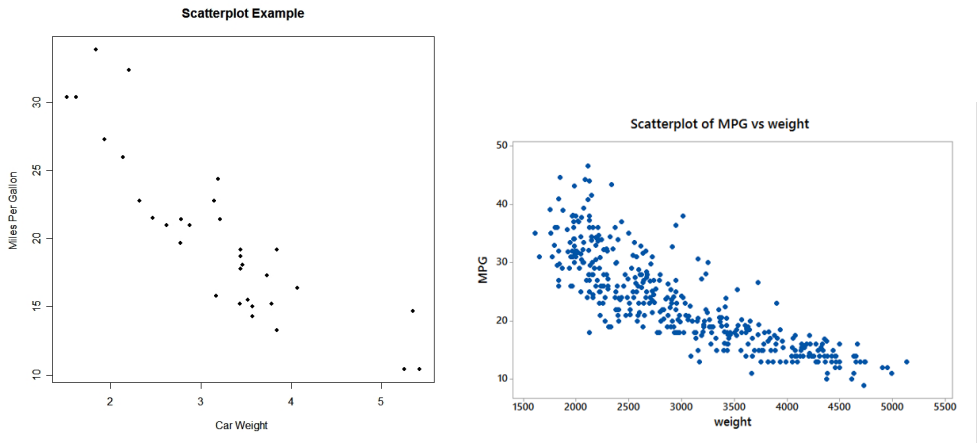


Figure 2.13: Gráfico de dispersão. Cada ponto é um modelo de automóvel. O eixo horizontal mostra seu peso, o eixo vertical mostra o seu desempenho em termos de milhas percorridas por galão de gasolina consumido. O gráfico da direita possui mais modelos de carros que o da esquerda e permite visualizar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

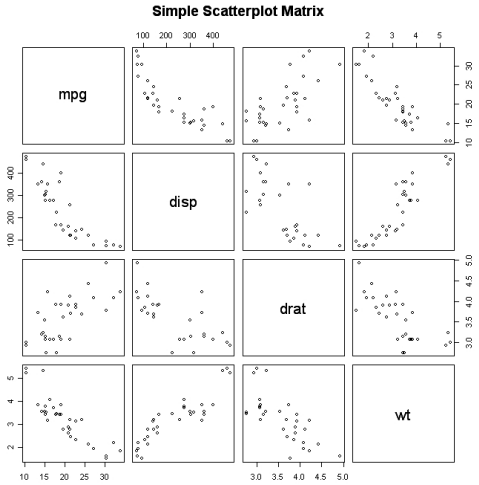


Figure 2.15: Matriz de scatterplotsde 4 variáveis de um dataframe usando o comando `pairs()`.

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

Na posiç ão (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

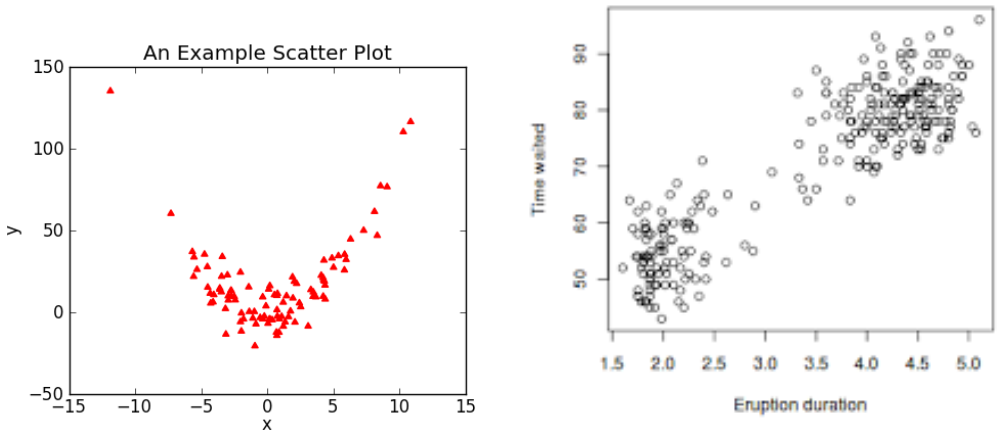


Figure 2.16: Scatterplots com relações mais complexas entre as variáveis.

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 nenhum 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 erupçẽoes 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

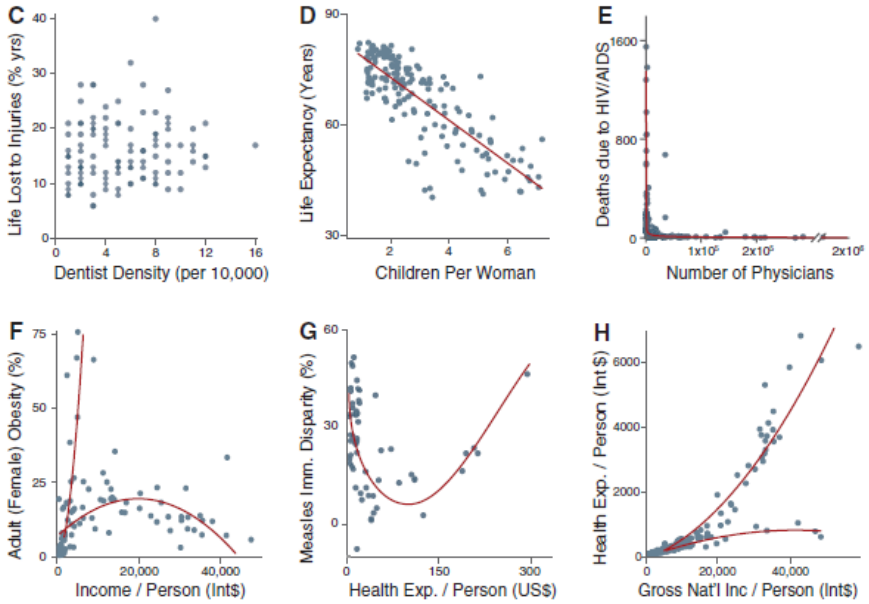


Figure 2.17: Scatterplots onde cada ponto é um país e as variáveis são indicadores sociais ou de saúde coletados pela Organização Mundial de Saúde (OMS). Fonte: [28].

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 [28]. 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 némero 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 relaąão 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

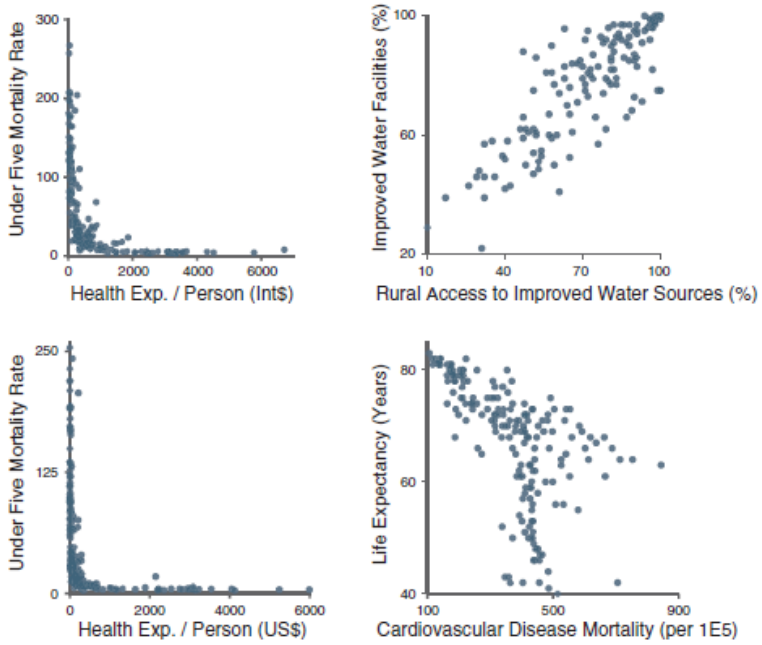


Figure 2.18: Mais scatterplots com os dados de indicadores sociais de países coletados pela Organização Mundial de Saúde (OMS). Fonte: [28].

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 [28], 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 relaąão entre x e y. A Figura 2.19, também extraída de [28], 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 nével 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,

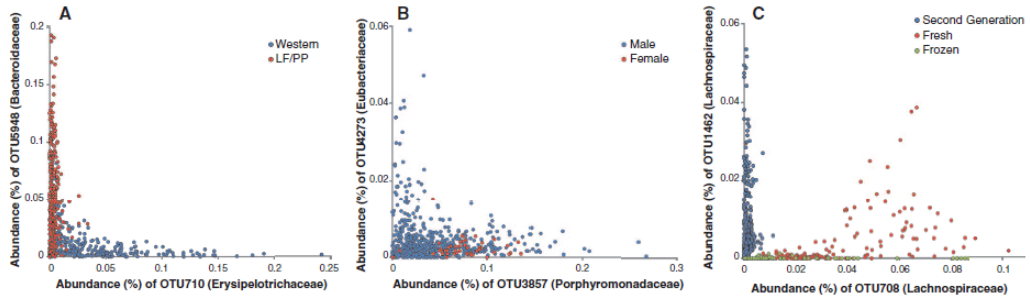
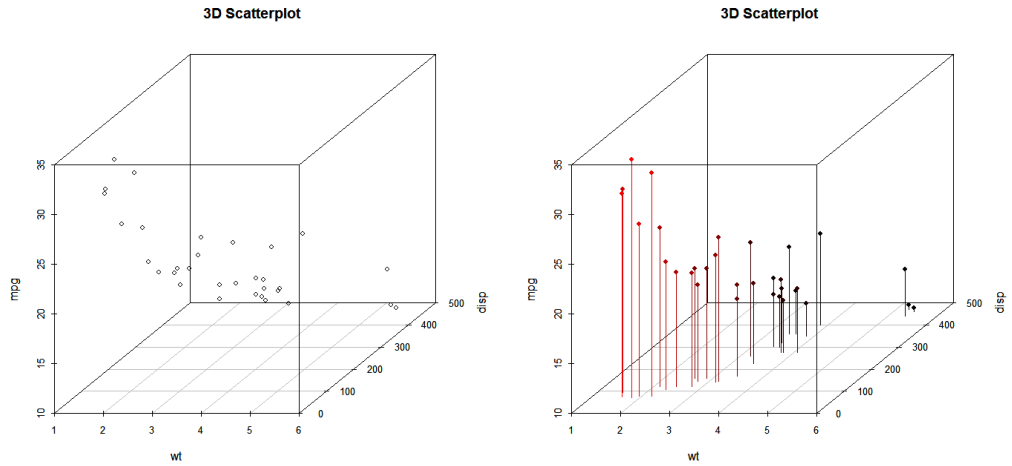


Figure 2.19: Abundância de diferentes espécies de bactérias comuns no intestino humano. Fonte: [28].



eu não acho estes gráficos muito úteis para análise.

```
# 3D Scatterplot, codigo 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 apreciar a utilidade desta ferramenta, use o código abaixo no R para experimentar, após instalar a library `rgl`.

```
# Spinning 3d Scatterplot, codigo 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)
```

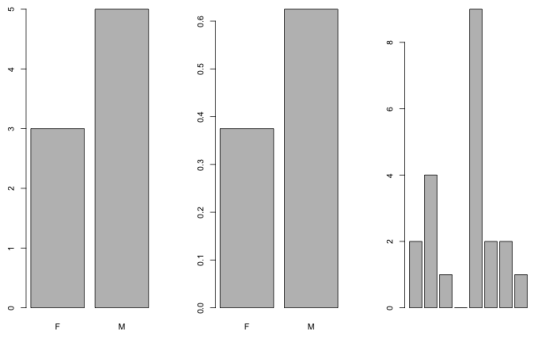


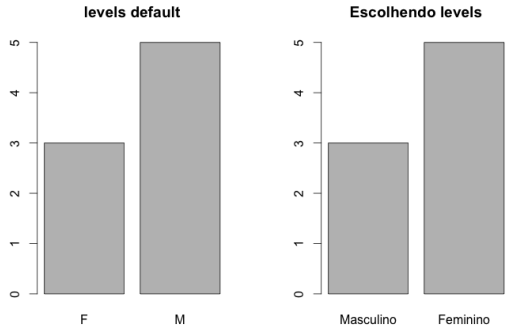
Figure 2.20: Barplot.

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 dist
> 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")
```

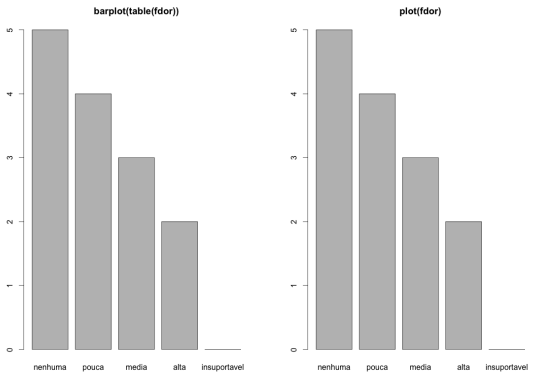
A questão não é apenas ser um vetor de caracteres. Vetores com números podem representar categorias. Por exemplo, poderíamos ter um vetor com os números 0 e 1 onde 0 representaria um caso "Masculino" e 1 representaria um caso "Feminino". R tem uma classe de objetos para trabalhar com variáveis categóricas: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

Nãao é só a questão de ser caracter para que um vetor seja uma variável categórica. Núumeros 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 "fact
> levels(fdor) = c("nenhuma", "pouca", "media", "alta", "insuportavel")  # expressando os
```




```
> 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: $0 \to 1$, $1 \to 2$, $3 \to 3$, $5 \to 4$, e $1000 \to 5$. 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
```

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 Vetorizar 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-stre`
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
```

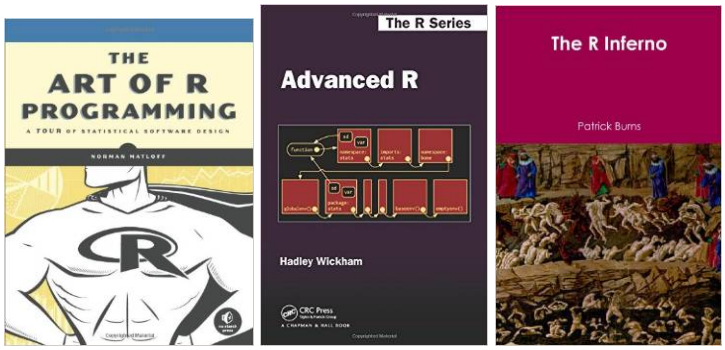
Ver capítulos 3 e 4 do livro *The R Inferno*, free pdf em `http://www.burns-stat.com/documents/books/the-r-inferno/` Abstract: If you are using R and you think you're in hell, this is a map for you. Even if it doesn't help you with your problem, it might amuse you (and hence distract you from your sorrow).

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

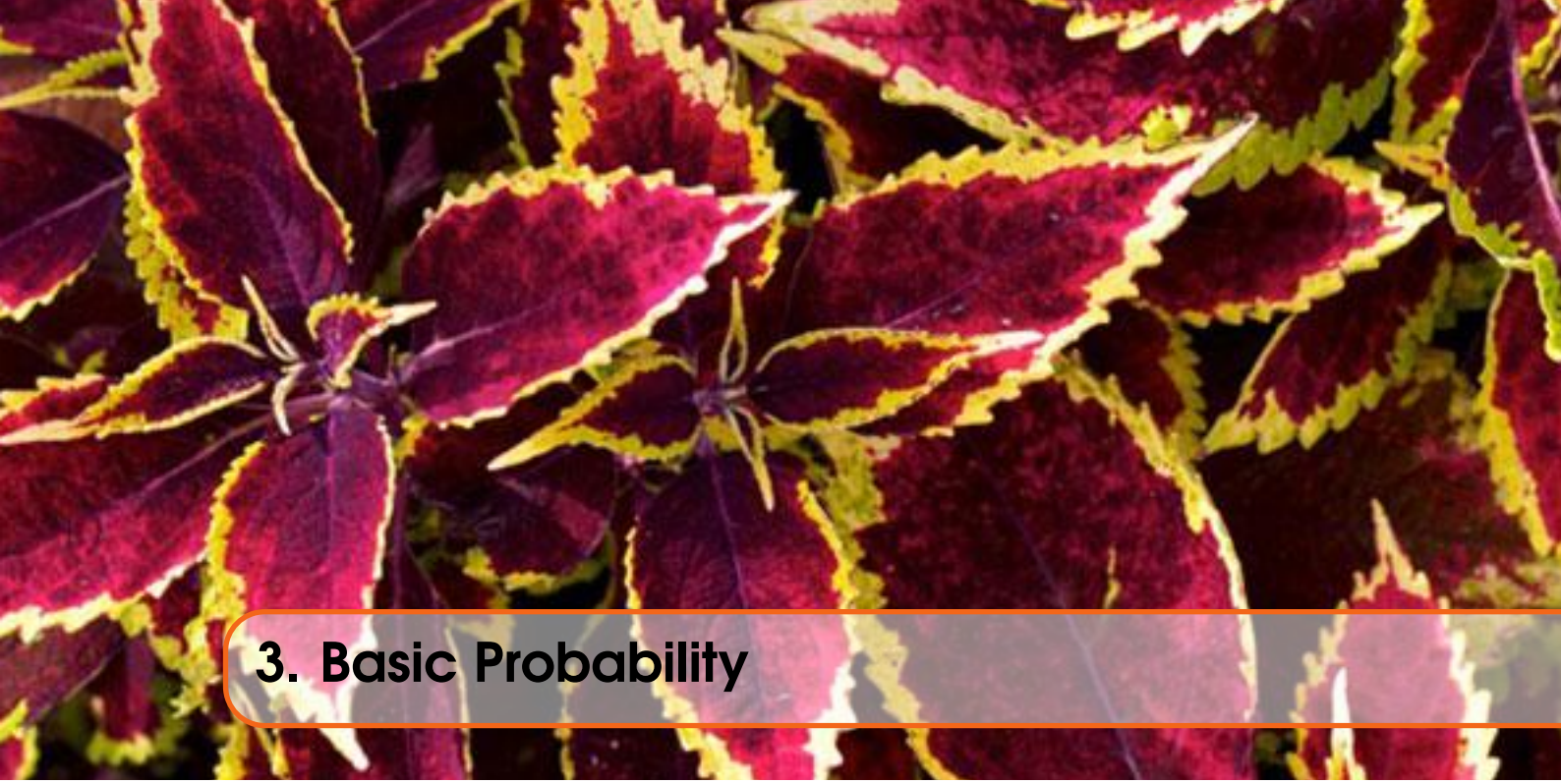
- Comece lendo na wikipedia: `http://en.wikipedia.org/wiki/R_(programming_language)`
- Download R para Linux, Mac, Windows em `http://cran.r-project.org/`
- Rstudio, free IDE para R: `http://www.rstudio.com/`
- Muitos tutoriais disponíveis no CRAN e na web. Escolha o seu sabor...
- R-tutorial: excelente - `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/UC4OWDcPB1peiBXDfCSZ3h-w/feed`



3. Basic Probability

3.1 Introduction

In this chapter, we will introduce probability spaces in two ways. We begin with a condensed version, distilling the entire discussion to its essential elements. This initial portion of the chapter provides sufficient foundation to carry you through your studies until the book's conclusion. It will cover several key details, mathematical nuances, and additional examples that elucidate and justify the necessity of certain advanced concepts (such as sigma-algebras) that may arise in more in-depth studies of probability. The second part, starting from the section labeled 3.6, is optional and can be omitted during the first reading. Revisiting it later will not detract from the comprehension of the remaining content of the book. However, if you possess a curiosity for mathematics and seek to understand the underlying rationale behind certain concepts, it is certainly worth exploring.

3.2 Probability Space

We are going to deal with non-deterministic, probabilistic, random phenomena. The mathematical model for **any** probabilistic phenomenon is the probability space $(\Omega, \mathscr{A}, \mathbb{P})$, consisting of three elements that satisfy Kolmogorov's three axioms: the sample space Ω, the sigma-algebra of events $\mathscr{A}$, and the probability function.

The sample space Ω

The first of these is a set Ω containing **all** possible outcomes of the random phenomenon. Each possible result is represented by a single element $\omega \in \Omega$.

■ **Example 3.1** If we look at the result of tossing a coin in the air we can represent Ω as the set $\{C, \tilde{C}\}$ representing heads and tails, respectively. We can use which two symbols for these results. For example, we can adopt the representation $\Omega = \{0,\ 1\}$ for heads and tails, respectively.

If we roll a 6-sided die, we can take $\Omega = \{1,2,3,4,5,6\}$.

If we roll two dice in a row, we get $\Omega = \{(i,j) : i \in D, \text{ e } j \in D\}$ where $D = \{1,2,3,4,5,6\}$. That is, Ω is the set of all pairs formed by the integers between 1 and 6.

If we look at the number of followers of a user on a social network, the possible results are the natural numbers: $\Omega = \{0,1,2,\ldots\}$. In this example, we could try to limit Ω by putting a maximum limit on the number of followers but this ends up bringing many practical and theoretical complications that end up not paying off (see the section **??** for more discussion). ■

■ **Example 3.2** Sometimes Ω involves intervals of the line or regions of $\mathbb{R}^n$. If we look at the time until the first comment appears on a YouTube video, a natural candidate would be the positive real ray: $\Omega = (0,\infty)$. In a video game, the random â θ with which a user throws an object (such as an angry bird in Figure 3.1. In this case, $\Omega = [0,2\pi)$. Another simple example is the observation of weight (in kg) and height (in m) of an adult individual. we can take $\Omega = (0,\infty)\times(0,\infty)$. This set seems excessive because it has points with very large or very small values, which evidently cannot be the result of observing heights or weights of individuals of the human species. However, this is not a problem: Ω must contain all possible results but does not have to contain *only* possible results. see more details in the 3.6 section).

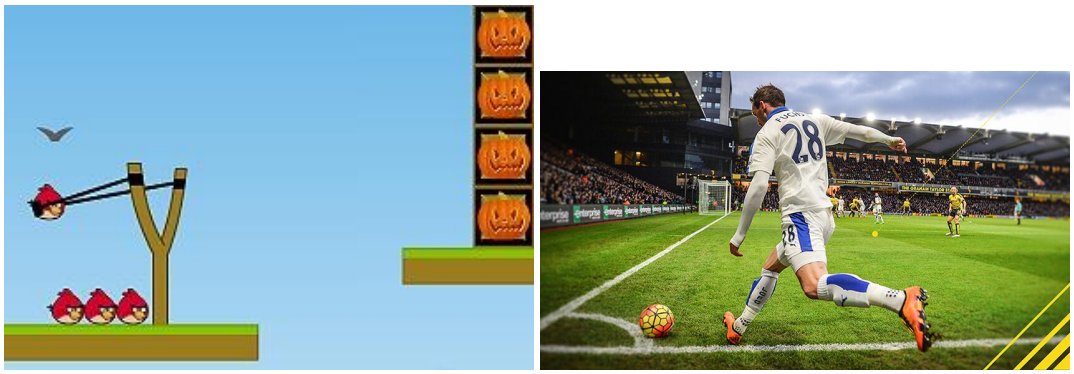


Figure 3.1: Random angle θ when launching an object in a digital game.

■

The $\sigma-$algebra $\mathscr{A}$

After setting Ω, the second item is highly technical. It starts very simply. What is the probability that the outcome of rolling a die is an even face? This is equivalent to asking what is the probability that the result $\omega \in \Omega$ is an element of the subset $B = \{2,4,6\} \subset \Omega$. Choosing a real number at random in the range $[0,1]$, what is the probability that it is at one end of the range? More specifically, what is the probability that the chosen random number ω falls out of the subset $B = [0,0.1]\cup[0.9,1]$.

Let B be a subset of Ω. If the random result ω belongs to B, we say that the subset B has occurred. Otherwise, the B subset did not occur. If the die is rolled and comes up 5, we say that the subset $\{2,4,6\}$ did not occur but we can say that the subset $\{4,5,6\}$ did. In fact, leaving the 5 face, there are several subsets for which we can say that they occurred. For example, we can say that the subsets $\{1,5,6\}$, or the subset $\{2,4,5,6\}$, or $\{5,6\}$ occurred, and even the subset $\{5\}$ formed only by the result $\omega =$ face 5. It is important to extend this notion to the maximum limit: we say that the set $\Omega = \{1,2,3,4,5,6\}$ occurred if the 5 came out. Of course, because it contains all possible outcomes, the set Ω will always occur, no matter what the random result.

Likewise, there are several subsets that do not occur when the 5 face comes out. For example, the subsets $\{1,2,3\}$ and $\{1\}$ do not occur. Extending this idea as far as possible, we say that the empty subset $\emptyset$ does not occur if the face 5 comes out. The empty set $\emptyset$ is a subset of Ω (of any Ω) and since it has no elements, we will always have $\omega \notin$ *emptyset* and it will never occur.

If B is a subset of Ω, we use the notation $\mathbb{P}(B)$ to represent the probability that the random result ω belongs to this subset. The second item in the probability space is $\sigma-$algebra $\mathscr{A}$, the

collection of subsets of $B \subset \Omega$ for which we can calculate probabilities. Are there subsets for which we will not be able to calculate probabilities? The answer is: it depends on Ω.

When Ω is a finite set, the $\sigma-$algebra $\mathscr{A}$ is easy: it is the set formed by all subsets of Ω. When Ω is an infinite but countable set, such as the set of natural numbers $\mathbb{N}$ or integers $\mathbb{Z}$, a $\sigma-$algebra $\mathscr{A}$ will also be the collection of all subsets of Ω.

■ **Example 3.3** For the coin example, with C representing and $\tilde{C}$ representing tails, the possibilities for the events $B \subset \{C, \tilde{C}\}$ are limited. The event B can be the empty set $\emptyset$, the set formed by the result heads (that is, $B = \{C\}$), the set $\{\tilde{C}\}$ and the set itself $\Omega = \{C, \tilde{C}\}$. So,

$$\mathscr{A} = \{\emptyset, \{C\}, \{\tilde{C}\}, \Omega\}$$

■

■ **Example 3.4** For $\Omega = \{1,2,3,4,5,6\}$ in the case of rolling the dice, every subset of Ω is in $\mathscr{A}$. For example, we want to calculate the probability of the event where the die results in an even number. This means calculating the probability one of the results ω contained in the set (or event) occurs $B = \{2,4,6\} \subset \Omega$. The $\mathscr{A}$ collection is larger in this example, formed by the $2^6 = 64$ subsets of Ω:

$$\mathscr{A} = \{\emptyset, \{1\}, \ldots, \{6\}, \{1,2\}, \ldots, \{5,6\}, \{1,2,3\}, \ldots, \{2,3,4,5,6\}, \Omega\}$$

■

■ **Example 3.5** Tossing a coin twice has the sample space $\Omega = \{(C,C), (C,\tilde{C}), (\tilde{C},C), (\tilde{C},\tilde{C})\}$. We want to calculate probabilities of events such as $B =$ *at least one head comes up*, represented by the subset $B = \{(C,C), (C,\tilde{C}), (\tilde{C},C)\}$. A $\sigma-$algebra $\mathscr{A}$ has all $2^4 = 16$ subsets of Ω:

$$\mathscr{A} = \{\emptyset, \{(C,C)\}, \ldots, \{(\tilde{C},\tilde{C})\}, \{(C,C), (C,\tilde{C})\}, \ldots, \{(\tilde{C},C), (\tilde{C},\tilde{C})\}, \{(C,C), (C,\tilde{C}), (\tilde{C},C)\}, \ldots, \Omega\}$$

■

The difficulty arises when Ω is not a countable set, such as the range $[0,1]$ or plane $\mathbb{R}^2$. In this case there are subsets $B \subset \Omega$ to which we cannot consistently assign a probability $\mathbb{P}(B)$.

These special subsets, for which we cannot assign probabilities, are called *non-measurable*. The other subsets $B \subset \Omega$, for which we can calculate their probabilities $\mathbb{P}(B)$ are called *events* and belong to $\sigma-$algebra $\mathscr{A}$. This is the highly technical part of defining the second element of the probability space, the $\sigma-$algebra $\mathscr{A}$. A rigorous treatment of this topic is given in the books on Measure Theory of Sets and the Lebesgue Integral.

The good news is that this difficult topic has very little practical impact. The reason is that all events of practical utility, even highly complex ones, are measurable. Non-measurable events are rare sets of purely theoretical interest. They do not get in the way of understanding and mastering the main statistical data analysis techniques and can be ignored at first. Let's ignore these mathematical niceties and assume that $\mathscr{A}$ contains all the subsets that can be formed from Ω.

■ **Example 3.6** For $\Omega = [0, 2\pi)$, the collection of events B that constitute $\mathscr{A}$ will contain subsets such as subranges $(a,b) \subset [0,2\pi)$ (think $(\pi/6, \pi/4)$, for example). Also, $\mathscr{A}$ will contain events formed by any pair of subintervals such as $B = (\pi/12, \pi/8) \cup (\pi/6, 1) \subset [0,2\pi) = \Omega$. Events formed by unions and intersections of 3, 4, or any other number of intervals will be part of $\mathscr{A}$. Is not possible to enumerate a list containing all possible events in this example. However, any event of practical interest will be part of $\mathscr{A}$. ■

In summary, in the practice of data analysis, we can assume that we can obtain the probability associated with each and every subset $B \subset \Omega$. This **is not** strictly true in cases where the set Ω is an interval of the line or plane or in other situations (technically, when Ω is an uncountable set).

However, for all practical purposes, we can ignore this complication and assume that $\mathscr{A}$ contains any and all subsets of Ω that you can think of. I find this subject fascinating as it is linked to themes involving the difficulties of thinking rigorously with infinite sets. In the section 3.6, we try to show in an informal way what are the difficulties involved and why Measure Theory is necessary. These sections are optional on a first reading of this book.

Summary of notation

- The sample space Ω is an set of all possible results of a random experiment.
- A result ω is an element of a sample space
- An event A is an set (a subset of the sample space Ω).
- An event may have no results (the empty set $\emptyset$, which has zero probability of occurring, as we will see).
- An event can contain a single result ω (single event or atomic event)
- An event can be the entire sample space Ω

3.3 The probability function $\mathbb{P}$

The third element of the probability space is the probability function itself. For each event $A \subset \Omega$, we must have a numeric value for $\mathbb{P}(A)$ and this number must be in the closed range $[0,1]$. The function $\mathbb{P}$ has an event (subset) B as a parameter.

■ **Example 3.7** For the example of the die, if it is well balanced, it will be natural to establish that the faces have equal probabilities of occurrence. That is, that $\mathbb{P}(\{1\}) = 1/6$ as well as $\mathbb{P}(\{k\}) = 1/6$ for every face $k = 1,\ldots,6$. How is the assignment of probabilities to events B composed of more than one face? For example, what is the probability of an even face? That is, what is the probability of an outcome ω that is in the event subset $\{2,4,6\}$. For any event B in this case of well-balanced data, let's define $\mathbb{P}(B) = m/6$ where m is the number of elements in B. So $\mathbb{P}(\{2,4,6\}) = 3/6 = 1/2$. Later we will see a more general way of specifying the $\mathbb{P}$ function. ■

The meaning of $\mathbb{P}(B)$ is the probability of an outcome that belongs to event B. Only one result $\omega \in \Omega$ occurs, only one of the elements ω of the set Ω. But we want to assign probabilities to subsets of Ω. For example, in the video game example, we want to know the probability that an object is thrown at an angle θ between $\pi/6$ and $\pi/4$ degrees. That is, we want the probability that the random angle θ that will occur at the launch is in the range $B = (\pi/6, \pi/4)$. In the language of probability we say that the event (subset) B occurs if the random realization generates an element of the event set B. Thus, the B subset event occurs if one of its elements occurs.

We need to define $\mathbb{P}(A)$ for every $A \subset \Omega$ of *consistently*. What properties should this assignment have so that we don't get inconsistent or contradictory results, even when Ω is a complex set like the examples in the 3.6 section? What are the minimum requirements that this probability assignment must satisfy, no matter how complicated Ω and $\mathscr{A}$ are? For example, we don't want to deduce that, from a certain probability assignment, we end up finding probabilities that are negative or greater than 1. Or we get $\mathbb{P}(A) < \mathbb{P}(B)$ despite knowing that A contains all the results that belong to B and therefore we should have $\mathbb{P}(A) \geq \mathbb{P}(B)$.

It is somewhat surprising that it takes very little for a function $\mathbb{P}$ to be a valid probability assignment, which will not yield inconsistent results. In 1933, the Russian mathematician Kolmogorov established a set of three conditions that this function $\mathbb{P}(B)$ must satisfy in order to be a valid probability function. That is, satisfying these conditions (or axioms), we will not have inconsistencies or logical contradictions throughout all probability calculations, no matter how complex the situation under analysis is. To be valid, just $\mathbb{P}$ is any function as defined below.

3.3.1 The Three Kolmogorov Axioms

Definition 3.3.1 — Probability function $\mathbb{P}$. Given a sample space Ω and a σ-event algebra $\mathscr{A}$, a probability function $\mathbb{P}$ is any function such that

$$\begin{aligned} \mathbb{P} : \mathscr{A} &\longrightarrow [0,1] \\ A &\longrightarrow \mathbb{P}(A) \end{aligned}$$

and that obeys the three *Kolmogorov axioms*:
Axiom 1 $\mathbb{P}(A) \geq 0 \ \forall \ A \in \mathscr{A}$
Axiom 2 $\mathbb{P}(\Omega) = 1$
Axiom 3 $\mathbb{P}(A_1 \cup A_2 \cup A_3 \cup \ldots) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) + \ldots$
if events $A_1, A_2, \ldots$ are disjoint (ie mutually exclusive).

The first two axioms are just setting a scale for probability. That is, the probability $\mathbb{P}(A)$ of any event A must be a number between 0 and 1. Since Ω contains all possible outcomes of the experiment, the event Ω occurs with certainty and has maximum probability. Axiom 2 says that this maximum probability is equal to 1. The choices established by the first two axioms are arbitrary. For example, the second axiom could have chosen $\mathbb{P}(\Omega) = 100$ instead of 1.

What is really important is the third axiom:

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

if the A_i's are disjoint. In mathematical jargon, we say that the probability function is an σ-additive function by satisfying this property.

To understand a little better what this property means, let's consider a particular case of Axiom 3. Let's take any two events A and B and make $A_1 = A$, $A_2 = B$, and $A_n = \emptyset$, the empty set, for $n \geq 3$. Let's also assume that $A \cap B = \emptyset$. So, in this particular case,

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

and Axiom 3 allows us to conclude that if $A \cap B = \emptyset$,

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

To proceed, let's show that $\mathbb{P}(\emptyset) = 0$. How $\Omega = \Omega \cup \emptyset \cup \emptyset \cup \emptyset \cup \ldots$ we use $(P2)$ and $(P3)$ to write

$$\begin{aligned} 1 &= \mathbb{P}(\Omega) \\ &= \mathbb{P}(\Omega \cup \emptyset \cup \emptyset \cup \emptyset \cup \ldots) \\ &= \mathbb{P}(\Omega) + \mathbb{P}(\emptyset) + \mathbb{P}(\emptyset) + \ldots \\ &= 1 + \mathbb{P}(\emptyset) + \mathbb{P}(\emptyset) + \ldots \end{aligned}$$

Cutting the 1 from both sides of the equality, we get 0 equals the sum of $\mathbb{P}(\emptyset)$ infinite times. So $\mathbb{P}(\emptyset)$ must be 0.

Returning to our main problem, we replace $\mathbb{P}(\emptyset)$ for 0 in (3.1), concluding so that if $A \cap B = \emptyset$, we have

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) \tag{3.2}$$

(see Figure 3.2).

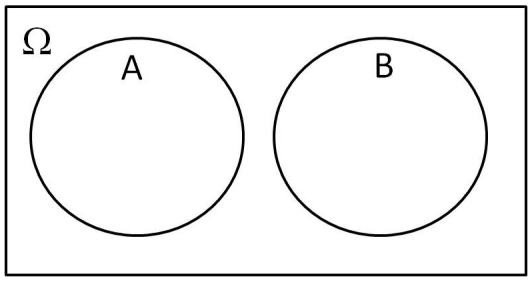


Figure 3.2: Venn diagram with $A \cap B = \varnothing$.

The axioms define the function $\mathbb{P}$ as being an additive function over disjoint sets. Intuitively, the probability of an outcome ω occurring in the union $A \cup B$ of the events A and B should be greater than the probability $\mathbb{P}(A)$ of the outcome ω occurs only in the A event. After all, in $A \cup B$ there are all possible outcomes that were in A plus the outcomes of B. The probability $\mathbb{P}(A \cup B)$ should also be greater than $\mathbb{P}(B)$. But why should it be the sum of individual probabilities? Couldn't another definition be reasonable? For example, maybe a formula such as $\mathbb{P}(A \cup B) = \left(\sqrt{\mathbb{P}(A)} + \sqrt{\mathbb{P}(B)}\right)^2$ couldn't perhaps be used? Is the decision to specify the probability of the disjoint union as the sum of the individual probabilities completely arbitrary?

Not really. The formula that makes practical sense is the one used in Kolmogorov's axioms. To see this, consider a simple situation where the Ω space is partitioned into two events. disjoint, each with probability 1/2. For example, in the case of throwing a balanced die we can write $\Omega = A \cup B$ with $A = \{1,2,3\}$ and $B = \{4,5,6\}$. Intuitively, each of them has probability equal to $1/2$, $\mathbb{P}(A) = 1/2$ and $\mathbb{P}(B) = 1/2$, because the die is well balanced . The union of A and B is the set Ω itself and therefore $\mathbb{P}(A \cup B)$ must equal $\mathbb{P}(\Omega) = 1$, the sum of the individual probabilities $\mathbb{P}(A)$ and $\mathbb{P}(B)$. The formula $\mathbb{P}(A \cup B) = \left(\sqrt{\mathbb{P}(A)} + \sqrt{\mathbb{P}(B)}\right)^2$ would lead to the erroneous conclusion that $1 = 2$ because

$$1 = \mathbb{P}(\Omega) = \mathbb{P}(A \cup B) = \left(\frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}\right)^2 = 2\,,$$

a contradiction. Similar reasoning with other partitions of Ω intuitively justify axiom 3. Thus, to make practical sense a probability function must be additive over disjoint events.

Any function $\mathbb{P}(B)$ that satisfies Kolmogorov's three axioms, no matter what Ω and $\mathscr{A}$ are, will be a valid probability function. This is not to say that each and every function is good or of practical use. The rule above just says that certain functions $\mathbb{P}$ will be mathematically valid as probability assignments. It is enough that $\mathbb{P}$ satisfies Kolmogorov's three axioms for $\mathbb{P}$ to be a valid probability assignment.

3.4 How to establish a function $\mathbb{P}$?

OK, we have seen that any function $\mathbb{P}$ that satisfies Kolmogorov's axioms is valid as a probability assignment function. But how do we choose one of these valid functions in a practical case? We used a combination of mathematical convenience (ease of handling) with a good approximation to reality. There is a trade-off between these two aspects. If we focus only on the use of mathematically very simple models, we will end up with models that are very far from the reality of the phenomenon, that do not represent it well. If we insist on incorporating all the aspects that can affect a phenomenon, we will have a probabilistic model that is unfeasible from a mathematical and computational point of view.

To define the probability function $\mathbb{P}$ we must consider three cases:

- Ω is finite: $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$
- Ω is infinite enumerable: $\Omega = \{\omega_1, \omega_2, \ldots\}$
- Ω is non-enumerable, such as $\Omega = (0,1)$ or $\Omega = \mathbb{R}^2$.

As we can expect, the third case has a few more complications than the other two.

3.4.1 Ω is a finite set

Let's start with the case where Ω is finite. Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$ where the ω_i are separate individual results. They are also often called *atomic events*.

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

You can assign values $\mathbb{P}(\omega_i) \geq 0$ completely arbitrarily as long as they satisfy the constraint that their sum equals 1:

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

Let $A = \{\omega_{i_1}\omega_{i_2}, \ldots, \omega_{i_n}\} \subset \Omega$ any event. What is the probability $\mathbb{P}(A)$? The probabilities for the events $A \subseteq \Omega$ are derived from assigning probabilities $\mathbb{P}(\omega)$ to atomic events using Axiom 3 of the definition **??**. The event A is the union of the events composed of the individual results, $A = \{\omega_{i_1}\} \cup \ldots \cup \{\omega_{i_n}\}$. Thus, we use Axiom 3 to write the probability of A as the sum of the individual events:

$$\mathbb{P}(A) = \mathbb{P}\left(\bigcup_j \{\omega_{i_j}\}\right) = \sum_{j=1}^{n} \mathbb{P}\left(\{\omega_{i_j}\}\right) = \sum_{\omega_i \in A} \mathbb{P}(\omega_i) = \mathbb{P}(\omega_{i_1}) + \ldots + \mathbb{P}(\omega_{i_n})$$

That is, the probability $\mathbb{P}(A)$ of any event A when Ω is finite is the sum of the probabilities of the atomic outcomes that compose it:

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\omega)\,. \tag{3.3}$$

To repeat: assign non-negative probabilities and adding 1 to the atomic events $\omega \in \Omega$. For any $A \subset \Omega$ event:

- identify which elements ω_i belong to A;
- add your probabilities $\mathbb{P}(\omega_i)$:

$$\mathbb{P}(A) = \mathbb{P}\left(\{\omega_{i_1}, \ldots \omega_{i_n}\}\right) = \sum_{\omega_i \in A} mathbbP(\omega_i)\,.$$

Equiprobable outcomes

A large number of problems and studies consider situations in which the N outcomes of Ω are equally likely, called equiprobable outcomes. In this case, we will have $\mathbb{P}(\omega) = 1/N$ for all $\omega \in \Omega$. Also, calculating $\mathbb{P}(A)$ becomes a counting problem since $\mathbb{P}(A) = |A|/|\Omega = |A|/N$ where $|A|$ is the number of elements in the set A. In these problems, everything boils down to combinatorial analysis, which can get quite complicated.

■ **Example 3.8** For example, flipping a coin 3 times generates a sample space with 8 elements, each with probability $1/8$. What is the probability of getting heads on the 2nd toss? As $A = \{CCC, CC\tilde{C}, \tilde{C}CC, \tilde{C}C\tilde{C}\}$, we have $\mathbb{P}(A) = 4/8$. ■

■ **Example 3.9 — Taken from William Feller's book, volume I.** Each of the 50 US states has two senators. A committee of 50 members chosen at random from among the senators is formed. What is the probability that a certain state (say, Oregon) is represented on the committee? What is the probability that everyone is represented?

The sample space is constituted by the different commissions of 50 members that can be formed. How many? the binomial number

$$\binom{N}{k} = \frac{N!}{k!(N-k)!}$$

calculates the number of ways to form groups (sets) of k elements chosen from the possible N. So there are $\binom{100}{50}$ possible commissions on Ω. How many committees are there where at least one of your two Oregon senators is present? For each of the possible commissions, there are 0 Oregon representatives present or more than zero representatives (1 or 2) present. Let's count the number of committees that have zero Oregon representatives. In this case, the 50-member committee must be formed with 98 remaining senators. We have $\binom{98}{50}$ possible commissions this way and therefore the number of commissions with at least one rep is equal to $\binom{100}{50} - \binom{98}{50}$. Therefore, the desired probability is equal to

$$\frac{\binom{100}{50} - \binom{98}{50}}{\binom{100}{50}} = 1 - \frac{98!50!}{100!48!} = \approx 0.753$$

For the second probability of interest, we need to count the number of commissions of 50 individuals that can be formed in which each of the 50 states is represented. Imagine lining up the 50 commission positions according to the name of each of the 50 states. Each position must be filled with two senators from that state. So we have 2^{50} commissions of this type and the desired probability is

$$\frac{2^{50}}{\binom{100}{50}} \approx 4.13\ 10^{-14}$$

The above calculation was done with the help of Stirling's approximation, a mathematical formula to approximate the value of very large factorial numbers. ■

■ **Example 3.10 — Taken from the website `cut-the-knot.org`.** This example has been slightly adapted from the excellent material on the website `https://www.cut-the-knot.org/Probability/Probabilities.shtml`. Roll two well-balanced dice and record the result of both sides. What is the probability that the sum of the two faces is 11? And what about 12? A *wrong* way of thinking about this problem is to imagine that these events have equal probabilities. This erroneous reasoning argues that the sum being 11 means that one die is 5 and the other is 6. On the other hand, the sum being 12 means that one die is 6 and the other is also 6. Thus, each case has the same number of outcomes and they should have equal probabilities.

To identify the error of this reasoning, we will explain the sample space. Imagine that the two dice are labeled Dice 1 and Dice 2. Each die has 6 faces and there are 36 possible outcomes considering both dice simultaneously. All 36 outcomes are equally likely. Denote by S_k the event that the sum of the two dice is k. We want $\mathbb{P}(S_{12})$ and $\mathbb{P}(S_{11})$.

The table below shows in its margins all six possible outcomes of each of the two dice. In the inner cells of the table, we have the sum of the faces of the two dice. Since the atomic results are equiprobable, to get the probability of any event S_k, just count how many table cells have the sum equal to k and divide by 36. it is easy to see that the sum 12 occurs only in a single result, $(6,6)$ and therefore $\mathbb{P}(S_{12}) = 1/36$, while the sum 11 occurs with the results (5.6) and (6.5) and therefore $\mathbb{P}(S_{11}) = 2/36$, twice as large.

Die # 1	Die # 2					
	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

■

Not equiprobable outcomes

Many complicated combinatorics arise in situations of equiprobable outcomes, but in most practical situations, the atomic outcomes have non-identical probabilities. This means that the probability of events A is no longer reduced simply by counting how many atomic results ω are in the event A. In these cases, after identifying the results $\omega \in A$, we use the general formulation $\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\omega)$ in (3.3). The probabilities $\mathbb{P}(\omega)$ of each atomic result ω are usually obtained in an approximated way from data analysis as in the examples below.

■ **Example 3.11 — Data mining in a micromarket.** Suppose there are only three products in a micromarket: A, B, and C. Ω is made up of the possible 8 product baskets:

- 0 (or no product),
- only A, only B, only C,
- only AB products together, only AC together, only BC together,
- the 3 products ABC together.

Let us represent $\Omega = \{0, A, B, C, AB, AC, BC, ABC\}$. A statistical analysis of the purchase pattern of several customers was carried out. It was observed, for example, that approximately 17% of customers left with basket A and 3% left with basket AB. This allowed to obtain approximately the probabilities the possibilities of each $\omega \in \Omega$. For example, $\mathbb{P}(A) \approx 0.17$ and $\mathbb{P}(AB) \approx 0.03$. Thus, we can assign probabilities to the atomic elements of Ω using these approximate probabilities:

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

See that they are greater than or equal to zero and add up to 1.

Here are some composite $E \subseteq \Omega$ events and their probabilities:

- E means to take product A in the basket, or $E = \{A, AB, AC, ABC\}$ and therefore $\mathbb{P}(E) = 0.17 + 0.03 + 0.18 + 0.12 = 0.49$.
- $E =$ take product A but not product C. That is, $E = \{A, AB\}$ and $\mathbb{P}(E) = 0.17 + 0.03 = 0.20$.
- $E =$ an empty basket, or $E = \{0\}$ and therefore $\mathbb{P}(E) = \mathbb{P}(0) = 0.02$.
- $E =$ an empty basket or with at least one product. So $E = \Omega$ and therefore $\mathbb{P}(E) = \mathbb{P}(\Omega) = 1$.
- $E =$ basket with 4 different products. Oops, this event doesn't exist. So $E = \varnothing$, the empty set, with $\mathbb{P}(\varnothing) = 1 - \mathbb{P}(\varnothing^c) = 1 - \mathbb{P}(\Omega) = 1 - 1 = 0$.

■

■ **Example 3.12 — Wind Direction and Speed.** At airports, planes land and take off in different directions of the runway depending on the wind direction and speed communicated by the control tower or, in smaller airports, by the windsock (Figure 3.3).

At a certain airport, the wind direction and speed are measured and classified into ranges of values. An analysis of data collected over the last 10 years allows us to calculate the statistical frequency of occurrence of wind direction and speed (in miles per hour) during the month of July according to the table below:

Figure 3.3: Control of wind direction and speed in airports.

Direction	Speed (mph)						Total
	0-7	8-12	13-18	19-24	25-38	39+	
N	0	1/2	2	2	1	1/2	6.0
NE	0	1	4	4	2	1/2	11.5
E	1/2	2	6	4	2	1	15.5
SE	1/2	2	6	4	2	1	15.5
S	1/2	1	4	6	4	2	17.5
SW	1/2	1	4	6	4	4	19.5
W	0	1	2	4	2	1/2	9.5
NW	0	1/2	1	2	1	1/2	5.0
Total	2	9	29	32	18	10	100

In this problem, we will take the sample space as the set of ordered pairs $\omega = (d,v) \in \Omega$ representing the category of direction and wind speed at the same instant of time chosen completely at random during the month of July at this airport. The possible outcomes are represented by the $8 \times 6 = 48$ combinations making up the cell IDs in the table above. The probability that an atomic outcome $\omega = (d,v)$ falls in a given cell will be approximated by the frequencies shown in the table (divided by 100). For example, the probability that we observe an outcome $\omega = (d,v) = (SE, 8-12)$ is $\mathbb{P}(\omega = (SE, 8-12)) = 2/100$. Clearly, the 48 atomic outcomes are not equiprobable.

If the interest is to know the probability that, at a certain moment in the month of July, we observe a wind in the direction SW or S. This corresponds to the sum of the probabilities of the pairs (v,d) with $v = SW$ or $v = W$ and is equal to $(9.5 + 19.5)/100 = 0.29$. The probability of observing a wind in the direction SW or W and, at the same time, with speed in the range $19-24$ or $25-38$ is equal to $(6+4+4+2)/100 = 0.16$.

■

3.4.2 Frequentist view

If the phenomenon under analysis can be repeated:

- indefinitely,
- under the same conditions
- independently (without one repetition affecting the next),

so $\mathbb{P}(\omega) \approx \dfrac{m}{N}$ where m is the number of times the result ω occurred in your N repetitions. We can take $\mathbb{P}(\omega) = \dfrac{m}{N}$, ignoring the built-in sampling approximation. This is called the *frequentist* view of probability.

And $\mathbb{P}(A)$ for $A = \{\omega_{i_1}, \ldots, \omega_{i_n}\}$? We have two possibilities. Take $\mathbb{P}(\omega_{i_j}) = \dfrac{m_{i_j}}{N}$ for each $\omega_{i_j} \in A$ and add these probabilities:

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

The other option is to simply check how many times the event A occurred in the N independent repetitions and take

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

where m is the number of times the event A occurred in the N iterations.

The second option produces the same result as the first option.

■ **Example 3.13 — Frequentist in the micromarket.** In the example, we have three products: A, B, and C. A statistical analysis allowed to obtain approximately the probabilities:

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

How were they obtained? A large number of N of customers were observed: these are the repetitions. The number of times m in which the basket was BC was counted. Finally we had $\mathbb{P}(BC) = m/N = 0.21$ or 21% of customers. ■

This is a long optional remark on a first reading. Let's exercise a critical eye on the mini-market example. By the frequentist argument, for $\mathbb{P}(BC) \approx m/N = 0.21$, we should have repetitions under the same conditions. Perhaps this is unreasonable. Some clients are old, others are young. Some buy in winter and others in summer. The conditions under which the repetitions are taking place do not appear to be identical. If the conditions are not identical, the probabilities may not remain constant. For example, the chance of product A being in the customer's basket is high if the customer is young or a purchase in the summer but the probability is low otherwise.

Another assumption is that the repetitions are independent. We will formalize this concept of probabilistic independence in a moment, but basically it means that the outcome of one repetition does not affect the probabilities of any other. This too can be questioned. Some customers may influence others via phone or comments. Another reason is that, if the customers are not all different, the purchases of the same customer can be very similar. To think of an extreme situation, imagine that only a single customer who always buys the same basket has been observed. Estimating the odds based on this single customer's data is not a good idea.

Another assumption is that repetitions can be done indefinitely. Suppose we are interested in $\Omega = \{TGG, \widetilde{TGG}\}$ where TGG means the chance of a third world major in the next 5 years and $\widetilde{TGG}$ its non-occurrence . It does not seem reasonable to want to establish probabilities by invoking frequencies in prolonged repetitions under the same conditions for this type of event.

A so-called *Bayesian* approach assumes that probabilities are subjective and can be manipulated with the rules of probability calculus. We will not see this approach in this book. We will see throughout the course that there are several ways to adapt the basic version of the frequentist approach to more realistic situations, with repetitions not having to be in the same conditions and also depending on each other.

■ **Example 3.14 — Dados de Weldon.** Walter Raphael Weldon (1860-1906) was an English biologist fascinated by the theory of evolution, as were many brilliant scientists in the first half of the 20th century (see Figure fig:WeldonEDados). Darwin published *The Origin of Species* in 1859

with immediate success and impact. For the following decades, biologists sought to accumulate evidence of the evolution of species, showing that biological diversity is the result of a process of descent with variability (children are not identical to parents) and with a force (natural selection) acting so that some of the individuals are more likely to leave offspring. He was one of the most important biologists defending the theory of evolution and together with Francis Galton and Karl Pearson he founded the journal *Biometrika*, one of the most important in statistics to date. Weldon writes in one of his articles that "... 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 ". The rediscovery of Mendel's work in 1900 sparked a debate involving Weldon about how it would be possible to combine the theory of evolution and the genetic theory. This junction was made in the 1930s *Sir* Ronald Fisher, the greatest statistician who ever lived and whom we will meet several times in this book. Having sufficient variability is crucial to creating evolutionary flexibility in the face of environmental changes that create the forces of natural selection. Evolutionary biology and modern statistics came together, forming an inextricable pair.

In an 1894 letter to Francis Galton, Weldon writes that he rolled a set of 12 dice 26306 times. His motivation was " to judge whether the differences between a series of group frequencies and a theoretical law, taken as a whole, were not more than might be attributed to the chance fluctuation of random sampling ".

Figure 3.4: Walter Raphael Weldon (1860-1906), evolutionary biologist and pioneer in the application of statistics to biometric problems. He wrote in a letter about the experiment of throwing 12 dice 26306 times and this experiment was analyzed by Karl Pearson in 1900 in the paper that presented the chi-square test.

With each roll of the 12 dice, he recorded the number of dice that showed either a face 5 or a face 6. Calling a success the appearance of a 5 or 6 on one of the 12 dice, on each roll of the 12 dice he could have zero successes. (none of the 12 dice with a 5 or 6) to 12 successes (all 12 dice showing a 5 or a 6). The results obtained are in the table below represented as N_k where $k = 0, \ldots, 12$. The f_k line shows the relative frequency of the event *get k successes* (multiplied by 10000, to avoid too many zeros after the decimal point). From the frequentist view, the probability of getting k successes when rolling the 12 dice should be approximately equal to these f_k values.

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	

In 1900, Karl Pearson[27] publishes a fundamental article in the history of statistics, presenting the chi-square test, capable of measuring, in a certain sense, the distance between empirical data and predictions based on probabilistic models. This test will be studied in chapter 10, where we

will return to this experiment by Weldon. In that article, Karl Pearson he analyzes the Weldon counts looking to see if they are compatible with the hypothesis or model that the data are well balanced, that all 6 faces of the dice are equally likely. If this is true, the probability that we get k successes when rolling 12 dice is what you see in the p_k row of the table above (counts would follow a binomial $\text{Bin}(12, 1/3)$ model, see chapter ???). These probabilities are also multiplied by 10 thousand.

For most people the two lines of numbers, f_k and p_k, are pretty close together. However, Pearson showed that if the data were well balanced, it would be highly unlikely, almost impossible, for us to have differences between f_k and p_k of the size we see in the table above. The probability of having differences in size from the ones we see in the table (which look small) is 0.000016 (this is the p-value of the chi-square test, see chapter 10). That is, the data is not well balanced. The last row of the table shows the differences $f_k - p_k$. They are negative for small k and positive for large k. This means that there are "too many" events with too many data showing 5 or 6. Experiment data tend to show more faces 5 or 6 than expected under the well-balanced data model. Pearson suggested that this was due to the data construction process. Casino dice are carefully constructed. Cheap wooden dice, the kind Weldon used, are unbalanced. By digging the wood to make the indentations and form the faces of the dice we create an imbalance. Face 6, the most excavated, is on one side of the die and face 1, the least excavated, is on the opposite side. Thus, face 6 is lighter than its opposite face (face 1) and tends to come out more often. In conclusion, using relative frequency as an approximation to the real probabilities that govern the throwing of cheap dice, Weldon's experiment showed that they were not well balanced. ∎

3.4.3 $\mathbb{P}$ when Ω is countably infinite

Suppose $\Omega = \{\omega_1, \omega_2, \ldots\}$ is a countably infinite set. This set has infinite elements and they can be put into a orderly sequence indexed by the natural numbers $1, 2, 3, \ldots$. Basically, there is a first element, a second element, etc. Typical examples of countably infinite sets are the natural numbers $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$ or the integers $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$. And additional example is the set $\mathbb{Z}^2$ of discrete points (i, j) in the plane forming a regular infinite grid with $i \in \mathbb{Z}$ and $j \in \mathbb{Z}$. A set that has infinite elements but which is *not* countably infinite is the real interval $(0, 1)$. In this case, mathematicians have proved that there is no way to enumerate all the elements in this interval. It seems that we are creating a big fuss about an irrelevant difference but, indeed, there is some difficulties in the $(0, 1)$ case that will require us to consider radically different ways to assign probability. But this is the subject of the next section.

The countably infinite case is identical to the finite case. Just assign $\mathbb{P}(\omega_i) \geq 0$ to the atomic elements of Ω such that the probabilities add up to 1. To get $\mathbb{P}(A)$ for some composite event A , sum the values $\mathbb{P}(\omega)$ of all elements $\omega \in A$.

Suppose $\Omega = \{\omega_1, \omega_2, \ldots\}$. This case is identical to the finite case. Just assign $\mathbb{P}(\omega_i) \geq 0$ to the atomic elements of Ω such that the probabilities add up to 1. *Any* assignment in this conditions will be a valid probability function for Ω. To get $\mathbb{P}(A)$ for some composite event A, add the values $\mathbb{P}(\omega)$ of all elements $\omega \in A$.

For example, a fair coin is tossed repeatedly until we see the first tails $\tilde{c}$. The sample space is infinite and composed of the atomic elements representing the observed sequence

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

A valid probability assignment is as follows:

$$\mathbb{P}(\omega_i) = \begin{cases} \mathbb{P}(\tilde{c}) = 1/2 \\ \mathbb{P}(c\tilde{c}) = (1/2)(1/2) = (1/2)^2 \\ \mathbb{P}(cc\tilde{c}) = (1/2)(1/2)(1/2) = (1/2)^3 \\ \vdots \end{cases}$$

It is valid because $\mathbb{P}(\omega_i) \geq 0$ for all $\omega_i \in \Omega$ and the sum of all of them is 1:

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

Consider the event B where the coin is tossed less than 5 times. Then

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

Let A be the composite event where the coin is tossed an even number of times:

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

We have

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

3.4.4 $\mathbb{P}$ when Ω is uncountable

The non-enumerable sets appear here again. Countably infinite sets are small infinites. Uncountable infinite sets are big infinites. There are several difficulties in rigorously dealing with them. We will give just one example for these sets.

Select a completely random real number in the interval [0,1]. We have $\Omega = [0,1]$. How to assign probabilities to events $A \subseteq [0,1]$? Let's try the same procedure as in the case where Ω is finite or enumerable. That is, assign a value $\mathbb{P}(\omega)$ to each real number $\omega \in [0,1]$ and set

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

This is not going to work. No point is favored as we choose a point "completely at random". Therefore, we must set $\mathbb{P}(\omega) = \xi > 0$ for all $\omega \in \Omega$. That is, no point has a better chance of being chosen than another, the probability of selecting ξ is the same for all possible ξ values.

What then is $\mathbb{P}(A)$? Suppose $A = \{1/2, 1/4, 1/8, 1/16, \ldots\}$. By Axiom 3 of 3.3.1 we have

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

since $\xi > 0$. So something is wrong because, to be valid, a probability must be a number between 0 and 1.

The mistake is to assume a positive probability, that $\mathbb{P}(\omega) = \xi > 0$. The correct and surprising is that every point in $[0,1]$ has probability 0! We will have to accept that $\mathbb{P}(\omega) = 0$ for every point $\omega \in [0,1]$. But if every number in $[0,1]$ has probability *zero*, how can we get $\mathbb{P}([0,1/2]) = 1/2$?

This paradox always appears when we represent reality with numbers on the real line. For example, in physics, the representation of reality with real numbers generates similar paradoxes. Suppose the interval $[0,1]$ represents a segment of wire with a mass of 1 gram. Assume that the wire has its mass perfectly and evenly distributed on the wire. We say that it has a constant mass density.

What is the mass of a point $x \in [0,1]$? Suppose the point has a mass $\xi > 0$. Since the density is constant, all points must have the same mass $\xi > 0$. Since there are infinitely many points in the segment $[0,1]$, the total mass should be $\xi \times \infty = \infty$ and not 1 gram.

The model that represents the wire by a line segment is incorrect from a physical point of view, it is just an idealized approximation of reality. The wire has atomic units that have mass. Its representation as a continuous line leads to paradoxes. The mathematical solution to making the representation ***useful*** is to assume that:

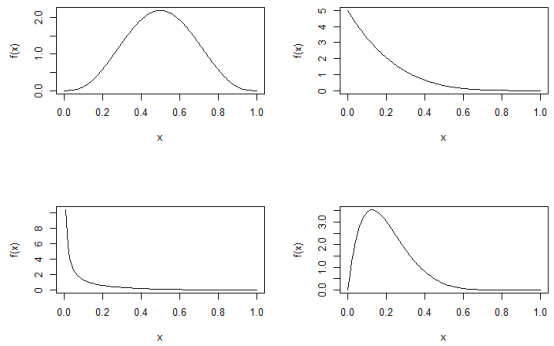


Figure 3.5: Four different probability mass density functions $f(x)$.

- Every point on the wire, seen in isolation, has zero mass.
- The mass associated with a segment $[a,b]$ is directly proportional to its length.
- Since the total mass of $[0,1]$ is 1 gram, the mass of $[a,b] \in [0,1]$ is $b-a$.
- For example, $[0.1/2]$ has a mass of 1/2 gram, $[1/2,3/4]$ has a mass of 1/4 gram, etc.
- Note that the point x is also the interval $[x,x]$, which has mass 0 because it has length 0.

A slightly more complicated way is to use a mass density function. This function will be very useful when the dough is not evenly distributed. The mass density function is a function $f(x)$ defined for each x in the segment $[0,1]$. This function is such that the mass in the segment $[a,b]$ is its integral between a and b:

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

If we take $f(x) = 1$ for all $x \in [0,1]$ we have

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

This is the density function $f(x)$ for the yarn with mass uniformly distributed in $[0,1]$.

The idea of a more general function than the constant function is to spread the total mass of the object over the segment $[0,1]$ via the function $f(x)$. The dough may not be evenly distributed. For example, suppose that the material of the linear segment $[0,1]$ is an alloy composed of the amalgamation of two elements, copper and zinc. In certain regions of the $[0.1]$ segment, there is more copper than zinc. In other regions, zinc predominates. The wire density will vary depending on the proportions of zinc and copper at the location. It may be more concentrated in some regions of the wire than in others. This is immediately reflected in the density function $f(x)$. In regions where the mass is more concentrated, $f(x)$ will be higher. Figure 3.5 shows examples of different mass densities $f(x)$ for $x \in [0,1]$. The mass of any subset is obtained by integration into the subset.

In probability, with uncountable Ω sets such as $\Omega \subseteq \mathbb{R}^n$, we adopt the same procedure as described above. The total probability mass of Ω is 1 because $\mathbb{P}(\Omega) = 1$. Spread over Ω this total mass (1) of probability using $f(x)$ for $x \in \Omega$. The probability mass of any event $A \subset \Omega$ is obtained by integration:

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

■ **Example 3.15** Think about the experiment of choosing a completely random point in $[0.1]$. Take $f(x) = 1$ for $x \in [0,1]$. Let's consider the event $A = [a,b]$, an interval. The experiment chooses a

single point in $[0,1]$, not intervals. If the event A occurs, it means that the chosen point belongs to the range $A = [a,b]$.

$$\mathbb{P}(\text{Interval[a,b]}) = \text{Length of interval [a,b]} \tag{3.4}$$

Every single point in the range $[0,1]$ has zero probability. Let A be a subset of $\Omega = [0,1]$ such as the union of two or more disjoint intervals: $A = [0,1/4] \cup [3/4,1]$. Then

$$\mathbb{P}(A) = \int_A f(x)dx = \int_A 1\, dx = 1 \times (1/4 + 1/4) = 1/2$$

■

R In $\mathscr{A}$, every event can be approximated with a finite or infinite number of countable operations $\cup$, $\cap$, and ^{C} on intervals of the line. We know how to calculate interval probabilities. As every event is obtained from set operations over intervals, the probability $\mathbb{P}(A)$ can be established for every event $A \in \mathscr{A}$. This is essentially the result of the Caratheodory Extension Theorem. We start with probabilities over basic sets (actually as intervals on the line) and this theorem assures us that, by extension, the probability is determined for all other events.

3.4.5 Função densidade de probabilidade

Random variables, the subject of the next chapter, make $\Omega \subset \mathbb{R}^n$ in practice. And this case is very easy because then the density $f(\omega)$ can be any function

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

such that:

- $f(\omega) \geq 0$, (so we never get a negative probability).
- and $\int_\Omega f(\omega)\, d\omega = 1$

■ **Example 3.16 — Tempo de espera.** We are interested in the waiting time (in hours) for the first comment after a YouTube video has been posted to a certain channel. What is the sample space? The waiting time will be a value on the positive semi-axis of the real line. As there is no well-defined upper bound, let's take $\Omega = (0,\infty)$ as the sample space. The events will be subsets of this semi-axis. And the density $f(x)$? There are several alternatives that we can adopt depending on the channel or the type of video posted by the channel. For example, $f(x)$ can vary if the channel has millions of followers or just a few; if it is appealing to generate views; if the video is about a topic of little interest, etc. The graph in Figure 3.6 shows some possibilities for $f(x)$ using all vertical axes and horizontal axes on the same scale for ease of comparison. The total area under the curves is equal to 1 on all four of these graphs.

The two top line graphs show exponential decay. The graph on the left has a more rapid decay. accelerated, $f(x) = 0.5e^{-0.5x}$, while the graph on the right has $f(x) = 0.2e^{-0.2x}$. In both cases, the integral of the function $f(x)$ on the semi-axis $(0,\infty)$ is equal to 1. The two density functions on the top row assign probabilities $\mathbb{P}(B)$ very different from the event $B = (0,1]$ representing the waiting time taking less than an hour. This probability $\mathbb{P}((0,1])$ is the area under each of the two curves in the range $(2,4]$:

$$\mathbb{P}((0,1]) = \int_0^1 f(x)dx = \begin{cases} \int_0^1 0.5e^{-0.5x}dx \approx 0.39 & \text{on the left} \\ \int_0^1 0.2e^{-0.2x}dx \approx 0.18 & \text{on the right} \end{cases}$$

In these two graphs, the density $f(x)$ systematically decays from the origin (time 0). In both cases, the probability of the waiting time falling in the range $(k,k+1]$ is equal to $\mathbb{P}([k,k+1]) =$

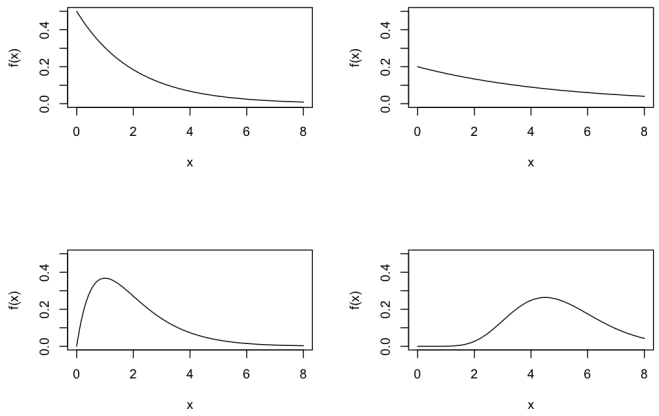


Figure 3.6: Four possible $f(x)$ probability density models for waiting time (in hours) for the first comment after posting a video on a You Tube channel. Visually, see how different the probabilities are $\mathbb{P}(A) = \mathbb{P}((0,1])$.

$\int_k^{k+1} f(x)dx$ where $k = 0,1,2,\ldots$. Visually, this probability is the area under the curve $f(x)$ in the range $(k,k+1]$. What happens with $\mathbb{P}((k,k+1])$ à as k increases on the top line graphs? Clearly, the area decreases with increasing k. For example, for the graph on the left, we have $\mathbb{P}((0,1]) = 0.39$, $\mathbb{P}((1,2]) = 0.24$, $\mathbb{P}((2,3]) = 0.14$, $\mathbb{P}((3,4]) = 0.09$. For the graph on the right, these probabilities again decay with k but this time, more slowly: $\mathbb{P}((0,1]) = 0.18$, $\mathbb{P}((1,2]) = 0.15$, $\mathbb{P}((2,3]) = 0.12$, $\mathbb{P}((3,4]) = 0.10$.

The bottom line graphs in Figure 3.6 show another possible situation. The curve on the left has $f(x) = xe^{-x}$ while the one on the right has $f(x) = x^9 exp(-2x)$. In them, unlike the top row graphs, the probability of the first comment appearing immediately after your post (immediately $x = 0$) is small. For the two top line graphs, the probability $\mathbb{P}((0,5/60))$ of the first comment appearing before 5 minutes after posting is 0.04 (left) and 0.02 (right). As for the two bottom line graphs we have these much smaller probabilities, equal to 0.003 (left) and 3.9 10^{-15} (right). The left curve is quite concentrated in waiting times of less than 2 hours. In sharp contrast, on the most ŕight curve, the probability of a comment appearing before $x = 2$ hours is only 0.008. ■

■ **Example 3.17 — Darts in a circular target.** Darts are thrown at a circular target of radius 1. A player has an ability that makes the chance of hitting a region A near the center greater than if that same region is near the edge. This ability is represented by the density

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

for x,y on the unit disk and c is a constant to ensure that $\int_\Omega f(x,y)dxdy = 1$. It can be shown that then $c = 14\pi/12$.

Let $r = r(x,y) = \sqrt{x^2+y^2}$ be the distance from (x,y) to the origin. We can rewrite the previous density as follows:

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

This makes it easier to visualize density with a heat map or contour lines. For any region A within the disk, we have

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

Individuals with different abilities will have their density different. Density should express which regions are most likely to be reached. What would a heatmap of the density $f(x,y)$ of a "blind" player look like? What about an extremely skilled player? What about a player who has a bias to the right? That is, a player who tends to throw the dart displaced to the right of the target. ∎

R [Remarks] We should have $f(\omega) \geq 0$, a lower bound. But we can have $f(\omega) > 1$: there is no upper bound. The fundamental constraint is that the integral over all Ω must be 1. The value $f(\omega)$ in each $\omega \in \Omega$ is not required to be less than 1. To obtain the probability of an event $A \subset \Omega$ just integrate $f(x)$ over the region A:

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

So a probability $\mathbb{P}(A)$ is area (or volume) under a density curve (or surface).

3.5 Consequences of Kolmogorov's Axioms

Like all good mathematical theory, all the rest of probability theory, all its surprising results, are all rigorously deduced from these three axioms of Kolmogorov described in section 3.3.1. For example, if an event A_2 is greater than another event A_1 (meaning that A_1 is contained in the subset A_2), we would expect $\mathbb{P}(A_2)$ to be greater than $\mathbb{P}(A_1)$. We don't need to declare this as an additional property of a probability function. It is deduced as a logical consequence of the three axioms. This and four other properties can be derived immediately if $\mathbb{P}$ satisfies Kolmogorov's three axioms.

Proposition 3.5.1 — Properties of $\mathbb{P}$.

(P1) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$. The probability of the complementary event A^c is 1 minus the probability of the event A.

(P2) $0 \leq \mathbb{P}(A) \leq 1$ for every event $A \in \mathscr{A}$. The interval $[0,1]$ is the scale on which probabilities are measured.

(P3) If $A_1 \subseteq A_2 \Longrightarrow \mathbb{P}(A_1) \leq \mathbb{P}(A_2)$. If one event contains another, its probability must be greater (or equal, at least).

(P4) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$. This is a very important property, very useful in practice.

(P5) $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$. This is an immediate consequence of the above property. See proof below.

(P6) $\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_i\right) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_i)$. This is a useful property for demonstrating many other more advanced probability results.

Let us just assume that $\mathbb{P}$ satisfies all three of Kolmogorov's axioms, nothing more. We start by proving the **P1** property. We have $\mathbb{P}(\Omega) = 1$ and $\Omega = A \cup A^c$. As $A \cap A^c = \varnothing$, we can use (**??**). So,

$$1 = \mathbb{P}(\Omega) = \mathbb{P}\left(A \cup A^c\right) = \mathbb{P}\left(A\right) + \mathbb{P}\left(A^c\right) \quad \Rightarrow \quad \mathbb{P}(A^C) = 1 - \mathbb{P}(A)\,.$$

To prove the property **P2**, see that Axiom 1 of the definition 3.3.1 already states that $0 \leq \mathbb{P}(A)$ for any $A \subseteq \Omega$. By the same Axiom 1 of the definition 3.3.1, we have $0 \leq \mathbb{P}\left(A^c\right)$ since A^c is also an event. To verify that $\mathbb{P}(A) \leq 1$, just use the proof of **P1**, where we conclude that

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

since $\mathbb{P}\left(A^c\right) \geq 0$. Thus, $0 \leq \mathbb{P}(A)$ and $\mathbb{P}(A) \leq 1$. That is, $0 \leq \mathbb{P}(A) \leq 1$.

Now let us prove **P3**. Remembering a set theory definition and notation, the set $B - A$ is formed by the elements that belong to B and do not belong to A. Thus, $B - A = B \cap A^c$. Returning to à

proposition **P3** now, if $A_1 \subseteq A_2 \implies \mathbb{P}(A_1) \leq \mathbb{P}(A_2)$. As $A_1 \subseteq A_2$, we can write $A_2 = A_1 \cup (A_2 - A_1)$. Since $A_1 \cap (A_2 - A_1) = \varnothing$, by axiom 3 we have

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

since, by axiom 1, $\mathbb{P}(A_2 - A_1)$ must be greater than or equal to zero.

Jumping to **P4**: The idea of this proof is to decompose $A \cup B$ into two disjoint events and then apply Axiom 3. We have $A \cup B = A \cup (B - A)$ and $A \cap (B - A) = \emptyset$. So, using Axiom 3, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B - A)$. Consider now the decomposition of the set BA into two disjoint subsets: $B = (B - A) \cup (A \cap B)$ and therefore $\mathbb{P}(B) = \mathbb{P}(B - A) + \mathbb{P}(A \cap B)$. Moving the last term to the left side, we find $\mathbb{P}(B - A) = \mathbb{P}(B) - \mathbb{P}(A \cap B)$. Substituting this expression in what we found earlier for $\mathbb{P}(A \cup B)$ yields the desired result:

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

This last property is the general case of $\mathbb{P}(A \cup B)$, when $A \cap B$ can be different from the empty set.

The **P5** property is very simple. Since $\mathbb{P}(A \cap B) \geq 0$ by Axiom 1, we use the **P4** property that we just proved to conclude that:

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \underbrace{\mathbb{P}(A \cap B)}_{\geq 0} \leq \mathbb{P}(A) + \mathbb{P}(B)$$

We are not going to prove the property **P6**, which is more interesting as a tool for demonstrating other results.

With this, we finish the minimal part necessary to continue the studies from the next chapter of this book. The next sections in this chapter are optional. They can be ignored on first reading by someone who is in a hurry or less mathematically curious. However, I believe that these next sections should not be completely ignored. I have tried to justify the existence and necessity of several subtle or more advanced concepts that you will surely find if you delve into the study of machine learning, statistics or (of course) probability. More than that, these optional sections constitute teaching material that, in general, is not present in other elementary textbooks that I know of. In them, these most advanced concepts are defined without much concern to justify their need. As I said, you may proceed by completely ignoring the next sections. They are like the blue or red pill dilemma that Morpheus presented to Neo in *The Matrix*: "You take the blue pill...the story ends, you wake up in your bed and believe whatever you want to believe. You take the red pill...you stay in Wonderland, and I show you how deep the rabbit hole goes." Neo chose the red pill as I hope you will eventually do.

3.6 Reviewing the probability space - Optional reading

We will look at each of the three elements of probability spaces in more detail below. There will be some repetition with respect to the themes already exposed in the previous sections.

3.6.1 The sample space Ω

Ω is a set representing **all** possible outcomes of the phenomenon. In artificial intelligence, we talk about all possible "states of the world" rather than possible outcomes. Each possible outcome must be completely specified and unique in Ω. There cannot be two elements in Ω representing the same possible result. See Figure 3.8.

The "world" represented in Ω is limited. It concerns the world you observe or study at the moment. To every state in the world corresponds one, and only one, element $\omega \in \Omega$.

A curious point is that Ω can have more elements than world states. That is, it may have elements that represent impossible outcomes. We will see the practical use of this in a moment. First, let's see some examples of Ω.

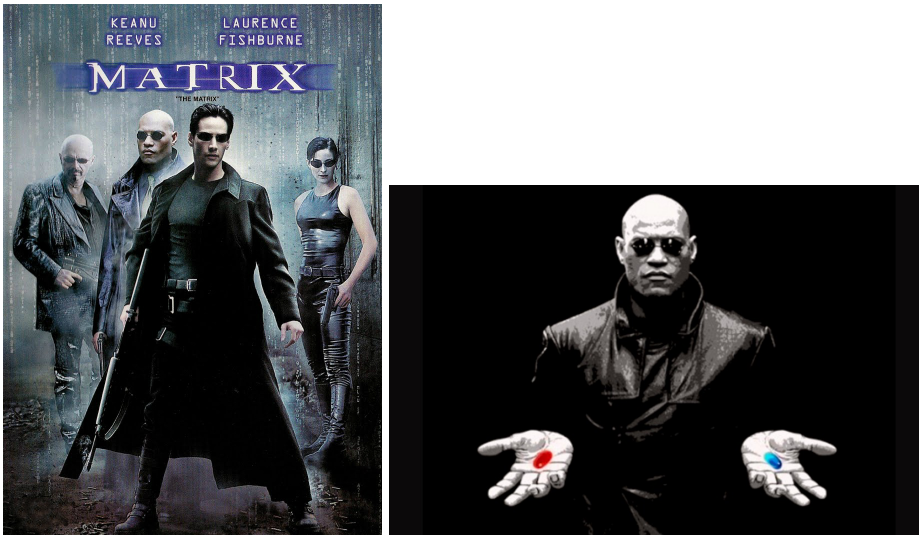


Figure 3.7: Red or Blue Pill? What's gonna be?

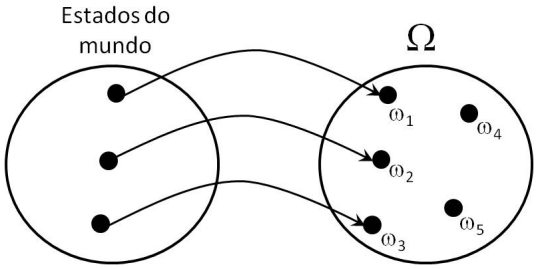


Figure 3.8: The sample space Ω

Examples of Ω

■ **Example 3.18 — The canonical example: tossing a coin once..** A coin is tossed. In this case, we can take

- Ω = { heads, tails }
- ou Ω = {c,c̃}
- ou Ω = {0, 1}
- ou Ω = {T, F}

Either of these options is valid, two symbols to represent the two possible outcomes of the probabilistic experiment of flipping a coin. ■

■ **Example 3.19 — More coins.** There are three successive tosses of a coin. Then

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

Ω has 8 elements. The world of this second example is wider than that of the first observer-example. In this new world we can calculate the probability of the second coin toss being *heads*. In the world of the first observer we cannot calculate the probabilities referring to the second or third tosses of the coin as they do not belong to the Ω of that world. ■

■ **Example 3.20 — Web Links.** The number of out-links from a randomly chosen web page is observed.

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

Ω is the infinite set of natural numbers. This is strange because at any given time there is a maximum number of web pages. Does it make sense to let pages have any number of links, an infinite number of possibilities? Makes sense? We want to let a page be able to have any number of links. A maximum number of links must exist but is unknown. Also, the maximum number changes over time.

We could put a maximum number that clearly must exceed the maximum (say, a billion links) but interestingly this makes the mathematical manipulation much more difficult and less productive than if we assume that $\Omega = \mathbb{N}$. The reason is that we have several probability models when $\Omega = \mathbb{N}$ but not when $\Omega = \{0,1,2,3,\ldots,10^9\}$. These models with $\Omega = \mathbb{N}$ are well studied and we know how to extract a lot of useful information from them.

Thus, we use $\mathbb{N}$ for the mathematical convenience of working with probability distributions defined over $\mathbb{N}$. Let's assign a probability strictly greater than zero to each of the infinite elements of Ω. We "correct" the excess of elements in Ω by assigning probabilities very close to zero to elements of $\mathbb{N}$ that are too large, absurdly large numbers to represent in-links. A power-law probability distribution will be one way to do this.

Anticipating what we will see later, we could do:

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

for $k = 0,1,2,\ldots$ where $C = 6/\pi^2$ (the constant was obtained by Euler (Basel problem)). In this case, we will have

- $\mathbb{P}(\omega = 100) \approx 6.110^{-5}$
- $\mathbb{P}(\omega = 10000) \approx 6.110^{-9}$
- and even smaller probabilities for numbers larger than these.

■

■ **Example 3.21 — In-links and Out-links.** The experiment consists of observing the number of in-links and out-links of n web pages. We can then define

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

■

■ **Example 3.22 — Itens in a supermarket.** A supermarket has 350000 products. The number of items of each product purchased by a customer is recorded. Then

$$\Omega = \left\{(n_1,n_2,...,n_{350000}) \in \mathbb{N}^{350000}\right\}$$

■

■ **Example 3.23 — Space of graphs.** We have a finite set V with n vertices. The interest lies in the non-directional relationships between the vertices. Who connects with whom? We can define

$$\Omega = \{\text{Undirected graphs in V}\},$$

a set with 2^p elements, where $p = n(n-1)/2$, the number of unordered pairs of vertices. ■

■ **Example 3.24 — Space of images.** An image with 512×512 pixels, each pixel has a shade of gray. The grayscale of each pixel is encoded with an integer between 0 and 255. 8 bits, $2^8 = 256$ possible shades: 0 is black and 255 is white.

Ω is the set of all M matrices of dimension 512×512 and with $M(i,j) \in \{0,1,\ldots,255\}$. Thus, Ω is a finite set but with a huge number of elements. In fact, the cardinality of the set Ω is equal to 256^{512^2}. Two elements of Ω are two grayscale 512×512 images.

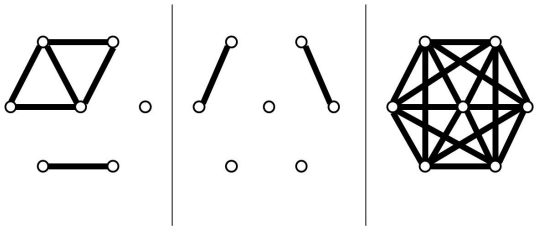


Figure 3.9: Three world states of the graph set, three elements of Ω.

Two examples are in Figure 3.24. The image on the left is a "structured" image, having several objects. The image on the right is one where each pixel is a random number between 0 and 255, completely unstructured. How to assign probabilities in this set? We will learn later to assign probabilities so that some images are more likely than others.

■

■ **Example 3.25 — Another canonical example: choosing a real number at random.** Imagine the experiment of choosing a real number $x \in [0.1]$ at random. This experiment cannot actually be done on a computer where numbers have a finite binary (and decimal) representation. However, we can conceptually imagine this experiment. In this case, $\Omega = [0,1]$. We will have more to say about this example when we want to better understand the assignment of probabilities in continuous or non-enumerable spaces. ■

■ **Example 3.26 — Height.** Select an adult inhabitant of Belo Horizonte (Brazilian city) at random and measure its height in meters. Is $\Omega = (1.30, 3.0)$ a good choice? Who knows (1.0,4.0)? Or $\Omega = (0,\infty)$? Or $\Omega = (-\infty,\infty) = \mathbb{R}$

For this example, in practice, $\Omega = (0,\infty)$ or $\Omega = \mathbb{R}$ are the preferred choices although obviously there is no negative height or height greater than, say, 5 meters. As in the case of the number of links on a web page, we are creating a set Ω that, of course, has all possible results but also has a number of impossible results. We will correct this "excess" through the third element of the 3-tuple of the probability space, the probability function $\mathbb{P}$. Whichever you choose, Ω is a non-enumerable infinite set. ■

■ **Example 3.27 — Tweets, a more complex case.** Collect all tweets issued in Belo Horizonte starting at t=0. Every tweet is logged. Our "World" is interested in calculating the probability that:

- a tweet talk about music,
- the wait time between two tweets about music is stable in time,
- the number of characters used in the tweet tends to be higher than average for music Tweets.

Ω can be represented by the set of all ladder functions of the type shown in Figure 3.10. The times $t_1 < t_2 < t_3 < \ldots$ are the times when the tweets arrive. A gray rung of the ladder function represents a tweet about music. One black step, one on non-music. The step width is proportional to the number of characters.

The graph in Figure 3.10 above is a possible result, an element $\omega \in \Omega$. Ω is formed by the infinite functions of this type varying the t_i's, the colors and widths of the steps. It is a

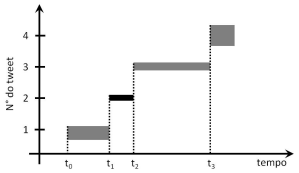


Figure 3.10: An element ω from the sample space Ω in the examples of tweets.

non-enumerable infinite set. How to assign probabilities in this set? ■

■ **Example 3.28 — Brownian Motion.** This example is important to motivate the necessary abstraction and mathematical complication of probability spaces. We won't see this example for the rest of the course but it is a typical example of a stochastic process, a crucial subject in higher probability. This example is of historical importance. Einstein published a fundamental paper in 1905 explaining Brownian motion as an effect of atomic motion and won a Nobel Prize for it a few years later.

It is somewhat complicated to present the precise definition of a Brownian motion at this point but it is not difficult to understand intuitively. Let (X_t, Y_t) be the position of a particle at the instant of time t. It starts from the origin so $(X_0, Y_0) = (0,0)$. The particle's motion depends only on where it is at any given time. Assuming it is in (X_t, Y_t) at time t, at time $t+\Delta$ it has the position $(X_{t+\Delta}, Y_{t+\Delta})$ where $X_{t+\Delta} = X_t + \mathcal{N}_x(0,\Delta)$ and $Y_{t+\Delta} = Y_t + \mathcal{N}_y(0,\Delta)$. The increment $\mathcal{N}_x(0,\Delta)$ has a Gaussian or normal distribution with zero mean and variance equal to Δ, and similarly for $\mathcal{N}_y(0,\Delta)$. Will we see in the chapter ?? the definition of the Gaussian distribution, but at this point you can imagine that the increment on each coordinate axis has an equal chance of going forward or backward and has an average length approximately equal to $0.80\sqrt{\Delta}$. The Brownian motion is the result of taking $\Delta \to 0$, which generates a curve in the plane similar to the one on the left in Figure 3.11. This curve is random. Repeating the experiment under the same conditions generates different curves such as the four shown on the left side of Figure 3.11.

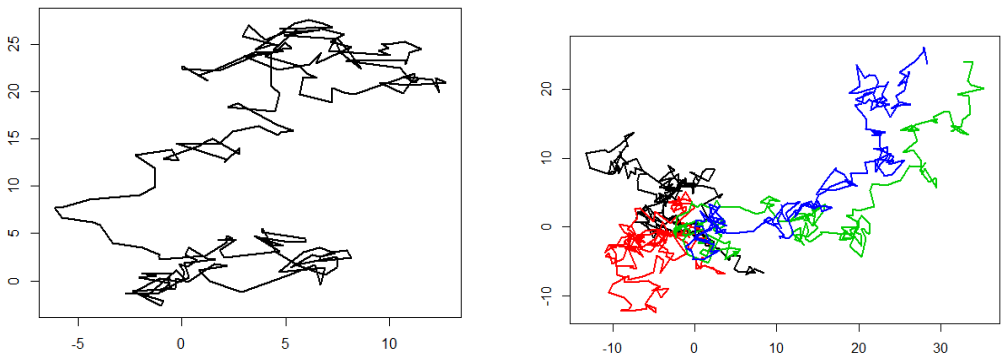


Figure 3.11: Left: Erratic motion of a pollen grain on the water surface observed every 1 second. Right: Four realizations of Brownian motion.

Ω is the set of all plane curves that may appear as a result of the experiment, four of which appear in Figure 3.11. Ω is a set with infinite curves. These curves have very curious mathematical properties. For example, they are continuous but have no derivative at any point, and furthermore, the length of the curve at any time interval is infinite, even if the time interval is quite small.

These and many other non-intuitive properties are rigorously derived with advanced probability tooling. The question that we are not going to answer but that justifies the 3-tuple paraphernalia of the probability space is: how to assign probabilities to this sample space Ω composed of these functions? ∎

∎ **Example 3.29 —** $\{0,1\}^\infty$ **.** Let us look at one last non-trivial example that will also not be studied in the rest of the course. A coin is flipped independently *indefinitely*. That is, the number of tosses is infinite. Let's represent by 0 and 1 the results of a coin toss. As there is no limit to the number of coin tosses, the sample space will have elements of the form

$$\omega = (a_1, a_2, a_3, \ldots)$$

where each a_i will be equal to 0 or 1. That is, the element $\omega \in \Omega$ will be a vector of infinite length where each entry is 0 or 1. The sample space Ω is composed of all infinite elements ω in this way. Fun fact: $\Omega = [0,1]$ because the expansion of a real number between 0 and 1 in base 2 is in the form ω above. ∎

3.6.2 Reviewing σ-álgebra $\mathscr{A}$

O segundo elemento do espaço de probabilidade $(\Omega, \mathscr{A}, \mathbb{P})$ é a σ-álgebra $\mathscr{A}$. 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

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

For example, in our simple example of rolling a balanced die we have $\Omega = \{1,2,3,4,5,6\}$. In addition to wanting to calculate probabilities such as $\mathbb{P}(\text{occurring face } 4)$, we will also want to calculate probabilities such as $\mathbb{P}(\text{occurring face greater than } 3)$ or $\mathbb{P}(\text{occurred even face})$. These probabilities are equivalent to calculating the probability that the result ω of the experiment belongs to the subset $\{4,5,6\}$ in the first case and to the subset $\{2,4,6\}$ in the second case. So we want to calculate probabilities of a specific outcome $\omega \in \Omega$ but also probabilities that the outcome comes from a subset $A \subseteq \Omega$.

Definition 3.6.1 A σ-algebra $\mathscr{A}$ is the set of subsets $A \subseteq \Omega$ for which we can calculate $\mathbb{P}(A)$. The subsets $A \in \mathscr{A}$ are called *events*. Subsets $A \subseteq \Omega$ of the form $A = \{\omega\}$, composed of a single element of Ω, are called *atomic events*.

Ideally, we want to calculate $\mathbb{P}(A)$ for each and every subset $A \subset \Omega$. Unfortunately, *in some cases*, we cannot calculate $\mathbb{P}(A)$ for each and every subset $A \subseteq \Omega$. The σ-algebra $\mathscr{A}$ is simply the class of subsets of Ω for which we can calculate $\mathbb{P}(A)$.

The σ-algebra $\mathscr{A}$ is very simple if Ω is

- a finite set of elements (such as $\Omega = \{0,1,2,3\}$)
- or an infinite but enumerable set (such as $\Omega = \mathbb{N} = \{0,1,2,3,\ldots\}$)

In these two cases $\mathscr{A} = 2^\Omega$, equal to the set of parts of Ω, the set of all subsets of Ω. If Ω is finite or infinitely enumerable, $\mathscr{A}$ contains all subsets of Ω. In this case we don't have to worry: we can calculate $\mathbb{P}(A)$ for each and every subset $A \subset \Omega$. Every subset of Ω is an event for which we will have a probability.

The σ-algebra $\mathscr{A}$ is a more complicated concept if Ω is an uncountable set such as the interval $[0,1]$ or the real line. In this case, *not* will be able to calculate $\mathbb{P}(A)$ for each and every subset $A \subset \Omega$. The σ-algebra $\mathscr{A}$ will not contain all subsets of Ω. Not every subset of Ω will be an event.

Why is it not possible to have a probability $\mathbb{P}(A)$ for each and every subset $A \subset \Omega$ in these cases? In measure (or size) theory of sets, it is proved that there is no *mathematically consistent* way to calculate the size (or measure) of all subsets of an uncountable set. We can measure (or size) soooo weird sets but we can't measure all subsets.

The theoretical consequence of this is that σ-algebra $\mathscr{A}$ does not contain all subsets of Ω if it is a non-enumerable set (such as the range $[0,1]$ or the real line). However, in the practice of data analysis, we can ignore this and continue working as if the σ-algebra $\mathscr{A}$ contained all subsets. The reason for this is that all the subsets A that we can conceive of in data analysis, even if very complicated, belong to $\mathscr{A}$. But if so, what doesn't belong to $\mathscr{A}$? The sets not in $\mathscr{A}$ are so weird they can't be displayed, we don't have a formula to get them.

For example, suppose $\Omega = [0,1]$ or $\Omega = \mathbb{R}$. It can be shown that in the smallest σ-algebra of some practical relevance (called σ-Borel's algebra) every event can be written with a finite number or an infinite number of enumerables of operations of $\cup$, $\cap$, and ^{c} of intervals of the line. It is very difficult to think in a set that is of practical interest and that can not be obtained by the use of these set operations.

Strange sets but with measure

As we said in the 3.6.2 section, when Ω is an uncountable set, such as $\Omega = [0,1]$ or $\Omega = (0,\infty)$, the σ-algebra $\mathscr{A}$ does not contain all subsets of Ω.

To understand this a little better, consider the case where we choose a real number at random in the range $[0,1]$. So $\Omega = [0,1]$. Let's say the measure (or size) of $\Omega = [0.1]$ is 1. Notation: $\mu([0,1]) = 1$.

The measure $\mu(I)$ of a subinterval $I \subset [0,1]$ will be equal to its length. For example, $\mu((0,1/2)) = 1/2$ and $\mu((1/2,3/4)) = 1/4$.

What if we have a set formed by joining k disjoint subintervals? As they do not overlap, the measure of the union will be the sum of the measures of the component intervals. For example, $\mu((0,1/2) \cup (3/4,1)) = \mu((0,1/2)) + \mu((3/4,1)) = 3/4$.

What if $A = \{1/2\}$, the subset formed only by the point 1/2? To be *consistent* with assigning the measure of an interval equal to its length, we will have to make the measure (or size) equal to zero: $\mu(\{1/2\}) = 0$.

What if we have multiple points in $[0,1]$? For example, $A = \{1/4,\ 2/4,\ 3/4\}$? To be consistent with the notion of a measure equal to a length, we'll need to take $\mu(\{1/4,\ 2/4,\ 3/4\}) = 0$.

What if we have infinite points in $[0,1]$? The answer depends on the type of infinity. Without intending to formally demonstrate the following statements, let's just give an idea of the need to consider things as complicated as a σ-algebra.

Suppose A is an infinite but enumerable set. For example, suppose that $A = \{1, \frac{1}{2},\ \frac{1}{3},\ \frac{1}{4},\ \frac{1}{5},\ \ldots, \frac{1}{n}, \ldots\}$. In this case, it can be shown that, to be consistent with what we have assumed so far, we will have to take its measure or size as $\mu(A) = 0$.

In general, if A is an infinite but enumerable set we should have $\mathbb{P}(A) = 0$. For example, consider the subset A formed by *all rational numbers* of $[0,1]$. That is, all fractions in $[0,1]$. This subset has measure (or size) 0, even though A is a dense set at $[0,1]$. This is surprising because, for any two points $a < b$ of $[0,1]$, however small the distance between a and b, there are infinitely many rational points between them. Still, even though the rationals are dense, despite being on any minimal line segment in infinite quantity, if we want a consistent measure of size, we have to say that the measure of all the rationals in $[0,1]$ is zero . On the other hand, the set of irrationals in $[0,1]$ has measure (or size) equal to 1, the same measure as the total set $[0,1]$.

■ **Example 3.30 — The Cantor set K .** Mathematicians thought of sets much stranger than rational ones, like the Cantor set K, for example. This set is represented in Figure 3.12 and is constructed as follows.

- Delete the middle third of the range $[0,1]$.
- Next, delete the middle third of each of the two sub-ranges.
- Iterate ad infinitum (so this is not an algorithm).
- The "final" result is the Cantor set.

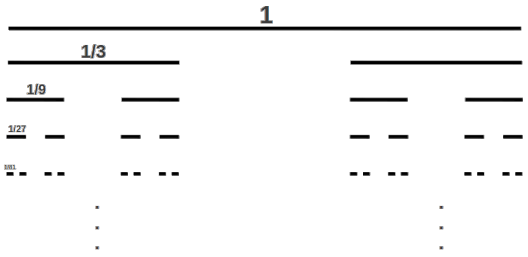


Figure 3.12: The Cantor set K.

The Cantor set K has remarkable properties (see `https://en.wikipedia.org/wiki/Cantor_Set`). Is there anything left at the end? Yes, that's a lot: K has an uncountable number of points. K is equivalent to the set of numbers that only use the digits 0 and 2 in their base 3 representation.

For every point $x \in K$, we have a sequence of distinct points $x_n \to x$. Thus, K has no isolated points. Around every point of K there are infinitely many other points of K, no matter how small the neighborhood. And yet K is totally disconnected (contains no gaps)!! Well, it can be rigorously proved that the measure (or size) of this set K is equal to zero.

■

Non-measurable sets

Subsets of non-enumerable sets can be very strange. So strange that some cannot be measured. It proves that there is no *consistent* way to extend the concept of measure (or size) to *all* subsets of $[0,1]$. The consequence is that not every subset of $[0.1]$ can have a size associated with it.

Who are these strange sets, so strange that we can't associate a measurement with their size? They are called non-measurable sets. See `https://en.wikipedia.org/wiki/Non-measurable_set` for more details. No one has ever "seen" one of these non-measurable sets. There is no constructive way to generate these non-measurable sets. What is proved is that *there are* unmeasurable sets but we cannot *build* (and display) one of them.

All known examples of non-measurable sets use the Axiom of Choice (see `https://pt.wikipedia.org/wiki/Axiom_of_choice`) and so we cannot show one of them *explicitly*. An example of a non-measurable set is the Vitali set (see `http://en.wikipedia.org/wiki/Vitali_set`)

The Borel σ-algebra

The most popular σ-algebras are Borel's. If $\Omega = \mathbb{R}$ or $\Omega = (0,1)$, take all ranges and generate *all* possible sets using $\cup$, $\cap$ and $mathrm^C$ in *enumerable* number. $\mathscr{A}$ is closed for $\cup$, $\cap$ and ^{C} in enumerable number: if $A_n \in \mathscr{A}$ then $\bigcup_{n=1}^{\infty} A_n \in \mathscr{A}$. If $\Omega = \mathbb{R}^n$, start with the "cubes" in $\mathbb{R}^n$ and generate all and generate all possible sets using $\cup$, $\cap$ and ^{C} in enumerable number.

Borel's σ-algebra $\mathscr{A}$ is large enough to contain all cases of practical interest. If you're able to "think' a set, it's probably in Borel's σ-algebra.

Non-measurable sets are so strange that they can be ignored in the practice of data analysis. The measurable sets are so many and so diverse that they include all the subsets of Ω that can appear in engineering and applied mathematics problems. So let's ignore this complication that σ-algebra $\mathscr{A}$ does not include all subsets of Ω. Every subset you can think of, believe me, will be a measurable set and will be part of the σ-algebra $mathcalA$.

In summary...

- The theory of σ-algebras and measurable sets is important in the rigorous study of the fundamentals of probability.

- In the study of stochastic processes, where Ω is quite complicated, these mathematical questions are important.
- For example, in the case of Brownian motion, Ω is an infinite set of continuous curves of the plane. In the case of coin flipping *indefinitely*, $\Omega = \{0,1\}^{\infty}$.
- σ-algebra also appears when defining conditional distributions rigorously.
- In our discipline and in the practice of data analysis and probabilistic machine learning, this subject is not very relevant.
- It is an important topic for the rigorous study of *theory* of statistics and machine learning. Especially nowadays, when many tools are being developed to search approximations in function spaces, this subject is of interest to many theorists. Regression trees, for example, is a problem where Ω is quite complicated.
- We'll ignore the complications of σ-algebra for the rest of this book. Let's assume that every subset that we are able to conceive of will be in σ-algebra $\mathscr{A}$.

3.6.3 Reviewing the probability function $\mathbb{P}$

When Ω is a complex set

Defining probability densities for Ω sample spaces that are complicated can be quite difficult. Worse: it may be impossible because we do not know how to explain a density in several cases. Again, the infinite sets come to haunt us. It is always the difficulty of mathematically dealing with "excessive" infinity. There are practical situations that require working with these Ω sets and we have to solve this somehow.

For example, in the case of Brownian motion, where we observe the erratic movement of a pollen grain on the water surface observed every 1 second (see Figure **??**). Ω is the set of all erratic brownian motion curves. How to define events (subsets of Ω) here? We want to calculate, for example, the probability that the particle's trajectory does not self-intersect in the first 10 minutes. This event corresponds to a huge set of Ω curves. What is the probability of its occurrence? How to consistently assign probabilities to all events?

In a second example, consider flipping a fair coin indefinitely and having $\Omega = \{0.1\}^{\infty}$. How to set probabilities consistently for all events in this case? Events must take into account the infinite launches. Let f_n be the ratio of 1's in the first n tosses. Track f_n along a sequence $\omega \in \Omega$ such as $(0,1,0,0,0,1,0,\ldots)$. What happens to f_n when n grows? Based on our experience, we expect to see $f_n \to 1/2$. But does this happen for sure? With probability 1? Or is there any chance, however minimal, that f_n will not converge to $1/2$?

Or maybe f_n doesn't converge to any value, oscillating in the $(0,1)$ range without permanently stabilizing around any value. After all, we can think of many (infinite!) $\omega \in \Omega$ sequences such that $f_n \nrightarrow 1/2$. For example, $\omega = (0,1,0,1,1,1,1,1,1,1,\ldots)$. Or $\omega = (1,0.1,0.0,1,0,0,0,1,0,0,0,0,1,\ldots)$. What is the probability that one of these infinite sequences with $f_n \nrightarrow 1/2$ occurs? The answer is: the probability is equal to zero (it is that of the theorem Strong Law of Large Numbers, see chapter ??). In an infinite sequence of tosses of a balanced coin, the probability that f_n does not converge to 1/2 is zero. But there are infinitely many such sequences in Ω, which do not converge. And yet they have zero probability of occurrence.

And if the coin has a very small probability of getting heads, say $\theta \approx 0$. What is the probability that f_n does not converge to this θ very close to zero? is zero again. This is fun. Let's see another event. Take the length of the longest streak of unbroken 1's in the infinite series of flips of the fair coin. What is the probability that this length is at least 2000? Curious? The probability is equal to 1, it will happen for sure over the infinite sequence of coin flips. This conclusion is a direct result of the Borel-Cantelli Theorem, subject of Chapter ??.

An even more interesting case. Suppose Ω is the set formed by all continuous functions. This Ω could be the result of the experiment of observing the continuous temperature curve for 24 hours

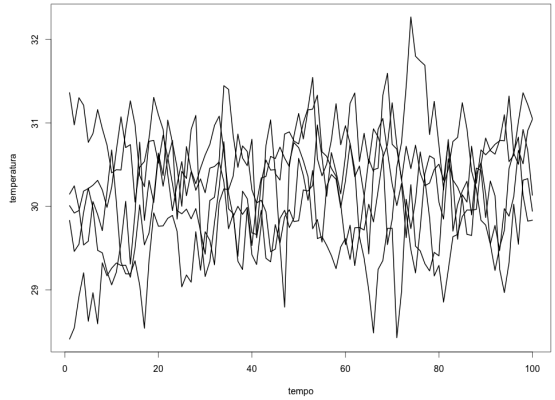


Figure 3.13: Ω is formed by the infinite set of continuous functions. The figure shows a small number of copies of this Ω set.

at a certain location. Figure 3.13 shows some results of this experiment. That is, the element $\omega \in \Omega$ is one of the infinite possible curves. Events are subsets of curves of this Ω set.

How to consistently assign probabilities to all possible events? For example, if A and B are two events (two subsets of curves) such that $A \subset B$ then we must have $\mathbb{P}(A) \leq \mathbb{P}(B)$. What could a probability density be in this Ω set of continuous curves? How to integrate in this set? We need a more complex notion of integral than the Riemann integral, a notion of measure or size of sets. This is the subject of advanced courses in probability and stochastic processes.

Solution in sight

For the rest of the book, we will avoid all the complications seen in these optional sections. In the simplest practice of data analysis we do not work directly with Ω. We have reduced the stochastic phenomenon to a few numerical characteristics with which we describe the random experiment. These features are called *random variables*, subject of the chapters 6 and 7. In practice, this will mean that, in the "worst case", we will have Ω equivalent to subsets of $\mathbb{R}^n$, for which we can define probability densities.

4. Conditional Probability

4.1 Conditional Probability

4.1.1 What is a conditional probability

Let B be an event in Ω and $\mathbb{P}(B)$ its probability of occurrence. Without being able to see the result of the experiment directly, we are only told that another event A has occurred. Does this change the probability of B occurring? For example, two well-balanced dice are rolled in a row. You bet on the occurrence of B: the first die will result in a 6. If you know that the sum of the two dice was less than 8 (event A) and you could revise your bet, you would place more chips on the occurrence of B? Or less chips?

With the information that a certain event A has occurred, we want to recalculate the chances of other events $B_1, B_2, \ldots$ We call this the probability of B conditioned to the occurrence of the event A, or the probability of B given that A occurred, or, even shorter, probability of B given A.

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

4.1.2 Conditional Probability and Data Science

The vast majority of data science techniques are algorithms for doing conditional probability calculations. A patient is diagnosed with stomach cancer. B represents the event that the patient will have at least 1 more year to live from that moment. Suppose, based on several similar cases, we know that approximately 70% of patients survive for more than a year from the diagnosis of this cancer. We are using the frequentist idea. Among all the patients observed in a similar situation in the recent past, 70% of them lived more than one year from the diagnosis. Therefore, assuming that $\mathbb{P}(B) = 0.70$ is reasonable.

Let A be the event that this stomach cancer patient has a biopsy of a sample of stomach tissue with the result that the tumor in that sample is not malignant. The A event has occurred. We imagine that now $\mathbb{P}(B|A)$ is greater than the previous value, $\mathbb{P}(B) = 0.70$. How to recalculate the probability of the occurrence of B conditional on this additional information that A occurred? If we have a large number of patients initially diagnosed and with a subsequent biopsy indicating benign, we count the proportion of those who survive more than one year within this subgroup of individuals.

This will be a good approximation for $\mathbb{P}(B|A)$.

The problem is much more complicated if the event A represents the following joint information:

- biopsy specimen indicates that it is not a malignant tumor,
- the patient is 45 years old,
- he is male,
- has always lived in Santa Catarina,
- is a smoker
- and always eats salami and smoked sausages.

There will not be a very large sample of patients in these same conditions. Perhaps none or only two or three people have been observed with all these characteristics. This prevents us from using the simple frequency occurring in these very few cases to approximate $\mathbb{P}(B|A)$. Data science tools calculate these probabilities using various tricks. They seek to extract as much information from the data as possible. We'll look at some of these tools in chapters throughout this book.

In general, given the characteristics represented by the event A, what is the chance of another event B occurring? Given that the robot's sensors say that A has occurred (for example, the ambient temperature is 42 degrees Fahrenheit and the brightness is low), what is the chance that it is in the B region? Given that the user has purchased the A set of items on their visits in the last 3 months, what is the probability that they will purchase the B item on their next purchase? What is the B item that maximizes this probability? Given certain behavior of a stock in the financial market over the last 3 years, what is the probability that B occurs where B represents the stock rising 10% or more within 30 days? Given certain A characteristics of an email, how likely is it to be spam?

Conditional probability is extremely important in theory but it is even more important in the practice of data analysis. It can be difficult to calculate and is the source of almost all paradoxes in probability calculus, even in very simple problems (see section 4.7 for some classical paradoxes).

4.1.3 Defining Conditional Probability

First question: how to go from $\mathbb{P}(B)$ to $\mathbb{P}(B|A)$? What is the relationship between $\mathbb{P}(B)$ and $\mathbb{P}(B|A)$? Can we have $\mathbb{P}(B) = \mathbb{P}(B|A)$? We will see that, in some cases, yes. *In those cases where* $\mathbb{P}(B) = \mathbb{P}(B|A)$ the occurrence of A does not change the chances of the occurrence of B. However, most of the time we will have $\mathbb{P}(B) \neq \mathbb{P}(B|A)$. We want to identify those cases where this probability changes and know how this change occurs: we have $\mathbb{P}(B) < \mathbb{P}(B|A)$ or, on the contrary, $\mathbb{P}(B) > \mathbb{P}(B|A)$. More than that, we want a formula that allows us to accurately calculate $\mathbb{P}(B|A)$ in any probability space.

Some obvious cases

Some cases are easy to calculate because they represent extreme situations. For example, rolling a well-balanced die and noting the face: $\Omega = \{1,2,\ldots,6\}$. Let $B = \{4,5,6\}$ with $\mathbb{P}(B) = 3/6$. Let's consider an event $A \subset B$. For example, $A = \{5,6\}$. Intuitively, what should $\mathbb{P}(B|A)$ be? Knowing that a face came out of the event $\{5,6\}$, what is the probability that the face that came out belongs to the set $\{4,5,6\}$? Now, if we know that the result ω of rolling the dice satisfies $\omega \in \{5,6\} = A$, then of course this element ω also satisfies $\omega \in \{4,5,6\} = B$ because $A \subset B$. Knowing that a result $\omega \in A \subset B$ occurred, we automatically infer that the result $\omega \in B$ and therefore B also occurred. Intuitively, we should define the conditional probability $\mathbb{P}(B|A)$ so that $\mathbb{P}(B|A) = 1$ in this example. Note that in this case $\mathbb{P}(B|A) = 1 > \mathbb{P}(B) = 3/6$. Anyway, the obvious case is that if $A \subset B$, then we must have $\mathbb{P}(B|A) = 1$.

Another obvious case is when $A \cap B = \emptyset$. Intuitively, what should $\mathbb{P}(B|A)$ be? If the event that occurred is in A, it cannot be in B because A and B are disjoint. Thus, when we know that an event $\omega \in A$ has occurred, we automatically infer that B has not occurred. So we should have $\mathbb{P}(B|A) = 0 \leq \mathbb{P}(B)$.

■ **Example 4.1** Example of balanced die with $\Omega = \{1,2,\ldots,6\}$. Let B be the *EVEN FACE* event.

That is, $B = \{2,4,6\}$ and $\mathbb{P}(B) = \frac{1}{2}$. Let $A = \{5\}$. of course $A \cap B = \emptyset$. Intuitively, if face 5 occurred, what is the chance of an even face? This chance is zero. Or would you bet on the occurrence of B in this case? ∎

Thus, two intuitively obvious cases are:

- If $A \subset B$ then $\mathbb{P}(B|A) = 1$.
- If $A \cap B = \emptyset$ then $\mathbb{P}(B|A) = 0$.

And the general case? Let A and B be two events with $\mathbb{P}(A) > 0$ and with $A \cap B \neq \emptyset$. How to calculate $\mathbb{P}(B|A)$? The definition is presented below.

Definition 4.1.1 — Conditional Probability. Let A and B be two events with $\mathbb{P}(A) > 0$. So, by definition,

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

So, to calculate the probability that B occurred *given that* A occurred:

- Find the probability $\mathbb{P}(A \cap B)$ that A and B both occurred
- Increase this probability by multiplying it by $1/\mathbb{P}(A) > 1$.

∎ **Example 4.2** Consider flipping a fair coin 5 times in a row with C representing heads and $\tilde{C}$ representing tails:

$$\Omega = \{CCCCC, CCCC\tilde{C}, \ldots, \tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}\}$$
$$\hookrightarrow 32 \text{ elementos}$$

We have $\mathbb{P}(\omega) = 1/32$ because all outcomes are equally likely. Let $B = \{\omega \in \Omega\,;\, 1^o \text{ element is } C\}$. In B we have 16 elements out of the 32 in Ω. Since they are equally likely, we conclude that $\mathbb{P}(B) = 16/32 = 1/2$.

The following information is provided: $A = \{$ There was only one tail in the 5 flips $\}$. That is, most (4) of the 5 flips are heads (C), only one position contains a tail. The universe of interest is now restricted to a subset $A \subset \Omega$ with many (4) heads. So, intuitively, the chance of having one of these many faces in the top position should be greater than in the total universe. That is, we anticipate that $\mathbb{P}(B|A)$ must be greater than $\mathbb{P}(B) = 1/2$. In fact, computing $\mathbb{P}(B|A)$ by the definition, we find:

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

So the probability has changed a lot when we know that A has occurred:

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

Knowing that only one tail has occurred in five tosses makes it highly likely that the 1^a position is heads. ∎

4.1.4 Data science and conditional probability, again

The last example illustrates one of the main goals of data science. When we have a complex system, involving many factors, we get some information at a low cost. This information is represented by the A event. We use this low-cost information to recalculate the probabilities of events B that we don't know have occurred: $\mathbb{P}(B|A)$. With these recalculated probabilities we can make decisions. We don't know if B occurred for several possible reasons:

- because B is in the future,
- because looking at B directly is very expensive,
- because it is impossible to observe B directly (for example, to know the real motivation behind a certain action of a person),

- or it is unethical to know B (eg, forcing an individual to smoke for 20 years to know if he will develop lung cancer).

4.1.5 Intuition for the definition

We saw the definition $\mathbb{P}(B|A) = \mathbb{P}(A \cap B)/\mathbb{P}(A)$. Why use this formula? Why this definition and not another such as $\mathbb{P}(A \cup B)/\mathbb{P}(A)$ or $\mathbb{P}(A \cap B)/\mathbb{P}(B)$? The answer is: to be consistent with experience, with empirical knowledge.

To see this, let's find $\mathbb{P}(B|A)$ in two different ways in a simple case to simulate on the computer. One way is to count the event B among those cases where A occurs. This is the natural way of estimating probabilities: by the relative frequency of the occurrence of events in a sequence of repetitions of the random experiment. The second way will be by the definition $\mathbb{P}(B|A) = \mathbb{P}(A \cap B)/\mathbb{P}(A)$. We will see that the two coincide and therefore that the definition (**??**) has to be adopted if we are to be consistent with our everyday experience.

Roll a well-balanced die twice. So $\Omega = \{(1,1),(1,2),\ldots,(6,6)\}$ and $\mathbb{P}(\omega) = \frac{1}{36}$ for all $\omega \in \Omega$. Let $B = [1^o$ given is a 6$]$. So $B = \{(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}$ and $\mathbb{P}(B) = 6/36 = 1/6$.

Let $A = [$ Sum of faces is greater than 8 $]$. We have

$$A = \{(3.6),(4.5),(4.6),(5.4),(5.5),(5.6),(6.3),(6,4),(6.5),(6.6)\}$$

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

How much is $\mathbb{P}(B|A)$? Should we expect it to be greater or less than $\mathbb{P}(B)$? The sum of the faces ranges from 2 to 12. Being greater than 8 means that we had a large value and that we can expect the two faces to be at least moderately large. Of course, if one of them is 1 or 2, we cannot have the sum of the faces greater than 8.

In fact, using the formula,

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

Let's now calculate $\mathbb{P}(B|A)$ by simulating the data on a computer. We will not use the formula **??**) but only relative frequencies and we will see that we will get the same result as if we use (**??**). Simulate the double tosses a large number of N times (for example, N = 100K). We will present the results in Table **??**.

Repetition	1	2	3	4	5	6	7	8	9	10	...
Die # 1	2	5	5	2	6	4	2	1	6	6	...
Di2 # 2	1	5	1	3	1	5	3	6	4	3	...
B occurred?	n	n	n	n	y	n	n	n	y	y	...
A occurred?	n	y	n	n	n	y	n	n	y	y	...

Table 4.1: Successive rolls of two dice.

Let's just consider the times when A occurred. In my simulation I got 13886 times. *Out of these 13886 occurrences*, check how many times the event B occurred. I got 5623 times. it is natural to expect $\mathbb{P}(B|A) \approx 5623/13886 = 0.405$. Why? Considering only the 13886 times that A occurred, we verified the proportion of times that B occurred. This is the way to empirically estimate, with data only, the value of $\mathbb{P}(B|A)$: by the relative frequency of occurrence of the event B *given that the event* A occurred .

Let us now verify that this empirical calculation leads to the formula (**??**) considering the numerator and denominator of the definition $\mathbb{P}(B|A) = \mathbb{P}(A \cap B)/\mathbb{P}(A)$. We know that

$$\mathbb{P}(A) \approx \frac{\text{number of times } A \text{ occurred}}{N} \qquad \Rightarrow \mathbb{P}(A) \approx \frac{13886}{N} \qquad \Rightarrow 13886 \approx N\,\mathbb{P}(A).$$

Likewise, by interpreting probability as frequency in long repetitions,

$$\mathbb{P}(A \cap B) \approx \frac{\text{number of times } A \text{ and } B \text{ occur}}{N}$$

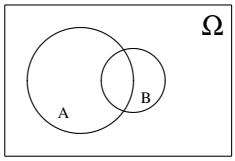
But A and B occur 5623 times in N. Remember we separated the 13886 cases where A occurred and then counted the cases where B occurred among these 13886 cases). So $\mathbb{P}(A \cap B) \approx 5623/N \Rightarrow N\,\mathbb{P}(A \cap B) \approx 5623$. Thus,

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

Our conclusion is that: if we want to keep intact our idea that the probability of an event is approximately equal to its relative frequency in a long series of independent repetitions, then the definition of conditional probability $\mathbb{P}(B|A)$ *must be* $\mathbb{P}(A \cap B)/\mathbb{P}(A)$. No other definition will yield results consistent with the experiments we've done.

4.2 Venn Diagrams

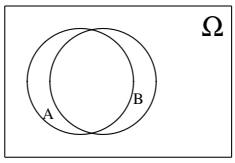
It is common to represent events using Venn set diagrams.



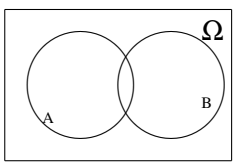
Ω is the largest enclosing rectangle. Events are figures with sizes proportional to their probability. So, what is the approximate value of $\mathbb{P}(A)$? $\mathbb{P}(A) \approx 0.90$? Or is it more reasonable to assume $\mathbb{P}(A) \approx 1/4$? Or $\mathbb{P}(A) \approx 1/8$? Or $\mathbb{P}(A) \approx 0.01$? With these same four answer choices, what is the approximate value of $\mathbb{P}(B)$? Answer in the footnote[1].

4.2.1 Conditional Probability in Venn Diagrams

How to see the definition $\mathbb{P}(B|A) = \mathbb{P}(A \cap B)/\mathbb{P}(A)$ in the diagram below? $\mathbb{P}(B|A)$ is the size of $A \cap B$ *relative to the size of* A. Answer: what is the approximate value of $\mathbb{P}(B|A)$: $0.85, 1/3, 1/8$, or 0.05? We have $\mathbb{P}(B|A)$ much larger than $\mathbb{P}(B) \approx 1/3$. So the correct answer would be $\mathbb{P}(B|A) \approx 0.85$.

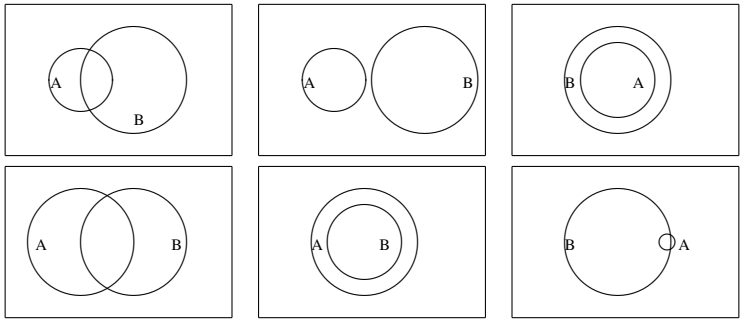


Now consider the following situation from the next diagram, shown below. What is the approximate value of $\mathbb{P}(B|A)$: $0.85, 1/3, 1/8$, or 0.05? We have $\mathbb{P}(B|A)$ much smaller than $\mathbb{P}(B) \approx 1/3$ and maybe the best option is $\mathbb{P}(B|A) \approx 0.05$.



[1] Answer: $\mathbb{P}(A) \approx 1/4$ and $\mathbb{P}(B) \approx 1/8$

In all cases below we have $\mathbb{P}(B) \approx 1/5$. Get $\mathbb{P}(B|A)$ approximately in each case.

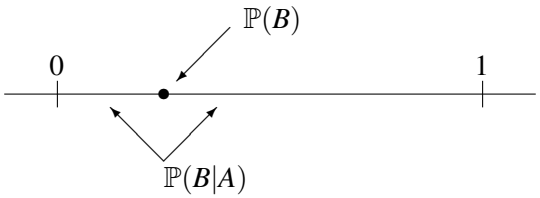


Considering the top row and starting from the left, the first diagram has $\mathbb{P}(B|A) \approx 0.40$. The second has $\mathbb{P}(B|A) = 0$ while the third has $\mathbb{P}(B|A) = 1.0$. Moving to the bottom row, from the left, the first diagram has $\mathbb{P}(B|A) \approx 0.40$, as in the first one in the top row. In the second we have $\mathbb{P}(B|A) \approx 0.85$ and in the third we have $\mathbb{P}(B|A) \approx 0.9$.

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

If $A \subset B$ then $\mathbb{P}(B|A) = 1$. The information that A occurred makes *certain* the occurrence of a result $\omega \in B$. If $A \cap B = \emptyset$ then $\mathbb{P}(B|A) = 0$. The information that A has occurred makes *impossible* the occurrence of any $\omega \in B$ These are *extreme situations*: knowing that A has occurred leads to an uncertain knowledge of the occurrence of B.

However, for the most part, knowing that A has occurred will not eliminate uncertainty about whether B has occurred. We will have $0 < \mathbb{P}(B|A) < 1$.



We can have $\mathbb{P}(B|A) > \mathbb{P}(B)$ or $\mathbb{P}(B|A) < \mathbb{P}(B)$.

4.3 Independence of events

There is another important case. It appears when the information that A has occurred has no bearing on the uncertainty about the occurrence of B. That is, there are cases where $\mathbb{P}(B|A) = \mathbb{P}(B)$

Definition 4.3.1 We say that A and B are *independent events* if $\mathbb{P}(B|A) = \mathbb{P}(B)$.

This definition of independence is equivalent to this one: A and B are independent events if and only if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$. To verify this claim, let's start by recalling that, by definition, $\mathbb{P}(B|A) = \mathbb{P}(A \cap B)/\mathbb{P}(A)$. If $\mathbb{P}(B|A) = \mathbb{P}(B)$, replacing $\mathbb{P}(B|A)$ with its defining expression, we get $\mathbb{P}(A \cap B)/\mathbb{P}(A) = \mathbb{P}(B)$ and therefore $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$. To find the reverse result, assuming $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$ is true, we substitute $\mathbb{P}(A \cap B)$ in the expression $\mathbb{P}(B|A) = \mathbb{P}(A \cap B)/\mathbb{P}(A)$ finding $\mathbb{P}(B|A) = \{\cancel{\mathbb{P}(A)}\mathbb{P}(B)\}/\cancel{\mathbb{P}(A)} = \mathbb{P}(B)$. Thus, saying that the events A and B are independent is the same as saying $\mathbb{P}(B|A) = \mathbb{P}(B)$ or, equivalently, saying that $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$.

4.3.1 How does independence arise?

Event independence can arise in two different ways. It can arise because we *assume* that events are independent. For example, thinking about the physical mechanism involved, we assume

that successive tosses of a coin are independent: the coin has no memory of what happened. Thus $\mathbb{P}($ Heads on 2nd. $|$ Heads on the 1st $) = mathbbP($ Heads on the 2nd $) = 1/2$. We do not mathematically deduce that the events are independent. We assume they are independent thinking about the mechanism involved in coin flipping.

The other way in which event independence can arise is when *we verify* mathematically that $\mathbb{P}(A \cap B) = \mathbb{P}(A)\ \mathbb{P}(B)$ or that $\mathbb{P}(B|A) = \mathbb{P}(B)$. Sometimes we cannot easily intuit that A and B are independent. In these cases, we calculate $\mathbb{P}(B|A)$ and $\mathbb{P}(B)$ and, voilá: if the probabilities are equal, the events are independent.

■ **Example 4.3 — Independence of events.** Let's start with a simple example of independence considering the repetitions of certain experiments such as rolling a die twice and noting the result. Events related to the first roll only must be independent of events related to the second roll of the die only. This is intuitive from our experience with this type of situation. The probabilities must remain the same: rolling a die does not physically change it to the point of affecting the probabilities of the 6 faces. Also, the die has no memory of what rolled before so that one roll does not affect the next.

It is this understanding of the physical situation of the problem that makes us *suppose* that the results are equally likely: $\mathbb{P}(\omega) = 1/36$ for $\omega = (r_1, r_2) \in$ *Omega* with $r_1, r_2 = 1, \ldots, 6$. In fact, let's assume that the occurrence of a 5 on the first roll (event $r_1 = 5$) increases the chance of seeing another 5 on the second roll (event $r_2 = 5$). In this case, we should have $\mathbb{P}(r_2 = 5|r_1 = 5) > \mathbb{P}(r_2 = 5)$. However, assuming that the results (r_1, r_2) are all equally likely, we find that $\mathbb{P}(r_2 = 5|r_1 = 5) = \mathbb{P}(r_2 = 5 \cap r_1 = 5)/\mathbb{P}(r_1 = 5) = (1/36)/(6/36) = 1/6$, which is equal to $\mathbb{P}(r_2 = 5) = 6/36 = 1/6$.

Thus, by assuming that the outcomes (r_1, r_2) are all equally likely, we are implicitly imposing that the events of the first roll are independent of the second roll. We can always mathematically check the validity of the condition $\mathbb{P}(A \cap B) = \mathbb{P}(A)\ \mathbb{P}(B)$ when A refers to an associated event to the first roll and B refers to the second roll only. If A is the event that the first roll is even then

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

and $\mathbb{P}(A) = 18/36 = 1/2$. Now let B be the event that the second roll is divisible by 3. So, $B = \{(1,3), \ldots, (6,3), (1,6), \ldots, (6,6)\}$. and $\mathbb{P}(B) = 12/36 = 1/3$.

We have $A \cap B = \{(2,3), (4,3), (6,3), (2,6), (4,6), (6,6)\}$ and, as expected, A and B are independent because

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

■

■ **Example 4.4 — Less obvious example.** A well-balanced die is rolled only once. Let $A = \{2,4,6\}$ and $B = \{1,2,3,4\}$. We have $\mathbb{P}(A) = 1/2$ and $\mathbb{P}(B) = 2/3$. Also, $A \cap B = \{2,4\}$ and $\mathbb{P}(A \cap B) = 1/3$. Like

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

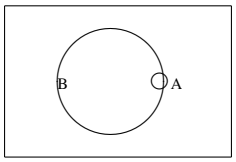
we conclude that the events A and B are independent. Thus, knowing that an even face came out does not make it more likely (or less likely) that one of the first 4 faces will occur. this conclusion it is not obvious at first sight and ends up requiring the mathematical verification of the independence condition. ■

4.3.2 Independence in the Venn diagram

If A and B are disjoint events on a Venn diagram, they are not independent. The reason is that if they are disjoint then $A \cap B = \emptyset$ and therefore $\mathbb{P}(A \cap B) = 0$. If $\mathbb{P}(A)$ and $\mathbb{P}(B)$ are > 0 then $0 = \mathbb{P}(A \cap B) \neq \mathbb{P}(A)\ \mathbb{P}(B)$. Therefore, they cannot be independent events.

If $A \subset B$ (with $\mathbb{P}(A) \neq 0$ and $\mathbb{P}(B) \neq 1$) then $\mathbb{P}(A \cap B) = \mathbb{P}(A) \neq \mathbb{P}(A)\ \mathbb{P}(B)$ and therefore A and B are not independent.

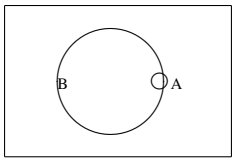
Except in these two cases, it is difficult to visually verify that A and B are independent on a Venn diagram. We would have to be able to see that the size of $A \cap B$ relative to Ω is equal to the product of the proportions of the sizes of A and B. However, if $\mathbb{P}(B|A)$ is very different from $\mathbb{P}(B)$ we can safely say that A and B are not independent. For example, without doing any math we can say that A and B are not independent in this case:



Visually it is obvious that $\mathbb{P}(B) \approx 1/3$ but that $\mathbb{P}(B|A) \approx 1$. Therefore, the occurrence of A increases the chances of the occurrence of B. Explanation: A is a rare event as $\mathbb{P}(A) \approx 0$. However, most of A is in B. If the rare event A occurs, it is highly likely that it is one of the $\omega \in A \cap B$.

4.4 Bayes Rule

The probabilities $\mathbb{P}(A|B)$ and $\mathbb{P}(B|A)$ can be completely different. For example, see the Venn diagram below:



We have $\mathbb{P}(B|A) \approx 1$ but $\mathbb{P}(A|B) \approx 1/25$. A more extreme case in the case of a well-balanced die roll: $A =$ comes up even, $B =$ comes up 2. In this case, $1 = \mathbb{P}(A|B) \neq \mathbb{P}(B|A) = 1/3$. With a mocking tone, consider a situation that shows how the two conditional probabilities can be completely different: $\mathbb{P}($ be Dracula $|$ not sleep at night $) \approx 0$ but $\mathbb{P}($ not sleep at night $|$ be Dracula $) \approx 1$.

Less silly examples involve considering conditional probabilities such as, for example, $\mathbb{P}($ is the murderer $|$ was seen In general, this conditional probability is smaller than the inverse probability $\mathbb{P}($ be seen at the scene $|$ is the murderer $)$. Or a problem we'll see later when considering the results of a clinical diagnostic test to verify if someone has some disease or carry a virus: let $A =$ [is infected] and $B =$ [positive test]. Then, we typically will have $\mathbb{P}(A|B)$ with a value very different from $\mathbb{P}(B|A)$.

There is a simple mathematical relationship between $\mathbb{P}(A|B)$ and $\mathbb{P}(B|A)$. We have

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

Equating the two expressions found for $\mathbb{P}(A \cap B)$ we have

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

Definition 4.4.1 — Bayes Rule - Simplified Version.

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

The main use of Bayes' rule is when we have one of the conditional probabilities, say $\mathbb{P}(A|B)$, and we want to calculate the inverse: $\mathbb{P}(B|A)$. In this case, we simply multiply $\mathbb{P}(A|B)$ by the ratio $\mathbb{P}(B)/\mathbb{P}(A)$. We will see in the 4.4.1 section that one of the two conditional probabilities is easily estimated from statistical data while the other ends up being obtained using Bayes' rule. For this we also need the non-conditional probabilities $\mathbb{P}(A)$ and $\mathbb{P}(B)$. In fact, we will now see that there is still one small addition we need to make to the formula presented in **??** because the denominator is not always readily available. The full version of Bayes' rule will be presented after the next example using disease diagnostic tests.

4.4.1 Bayes and diagnostic test

Should the Brazilian population be screened with an HIV test? It is believed that 0.1% of Brazilians are HIV positive, which means that approximately 200K out of the 200 million inhabitants are infected with the HIV virus. If we select a Brazilian at random, we will denote by V the event indicating that *the selected person is infected with HIV*. It has probability $\mathbb{P}(V) \approx 0.001$. Let's denote by $\overline{V}$ the complementary event, that the selected individual is not infected by HIV, with probability $\mathbb{P}(\bar{V}) = 1 - \mathbb{P}(V) = 0.999$.

A diagnostic test for the presence of the virus applied to this individual has two possible results: positive for the detected presence of the virus and represented by T, or negative, and represented by $\overline{T}$. No test is perfect. Because of this, we can define the confusion table considering the individual's actual situation and the result indicated by the diagnostic test:

Virus?	Test Result	
	T	$\overline{T}$
V	ok	error
$\overline{V}$	error	ok

A patient receives the test result and it is positive (the event T has occurred). The question is: does it actually have the virus (the V event occurred) or did an error occur? The main medical problem is calculating $\mathbb{P}(V|T)$. How to get this? By Bayes' rule we have $\mathbb{P}(T|V)$, as explained below.

4.4.2 Sensitivity and Specificity

The diagnostic test is applied to two large groups of individuals:

- a group composed of individuals known to have the HIV virus.
- another group in which the components are known not to carry the HIV virus.

The probability of any event A occurring in the first group will be a conditional probability, $\mathbb{P}(A|V)$, given that the individual is V. For example, the proportion of individuals in the first group who test positive will be an estimate of $\mathbb{P}(T|V)$. In the second group, the proportion of individuals for which A occurs will serve as an estimate for the conditional probability $\mathbb{P}(A|\overline{V})$, given that the individual is $\overline{V}$.

Definition 4.4.2 — Sensitivity and Specificity. Based on the frequency of T responders in each group, the laboratory producing the diagnostic test estimates the following amounts:

- $\boxed{Sensitivity}$: $\mathbb{P}(T|V) \approx 0.99$.
- $\boxed{Specificity}$: $\mathbb{P}(\overline{T}|\overline{V}) \approx 0.95$.

These are sensitivity and specificity estimates only, because they are typically based on samples of a few hundred individuals in each group. If another subset of subjects were tested, the sensitivity and specificity estimates would be slightly different.

The higher these two probabilities, the better. Ideally, we would like both to be equal to 1. In practice, diagnostic tests make errors and their sensitivity and specificity are less than 1. Some tests have very high sensitivity and specificity (such as Western blot) but are expensive, time-consuming, and require specialized personnel. Other tests have these lower probabilities but are faster, cheaper and more automatic.

Another relevant practical aspect is the selection of individuals who are part of the two groups, V and $\overline{V}$. It is important that the subjects included in the study include the full spectrum of characteristics of future patients. If important subgroups of patients are not adequately represented, sensitivity and specificity estimates can be biased. For example, if a study has only very healthy individuals in the $\overline{V}$ group, the estimates of $\mathbb{P}(\overline{T}|\overline{V})$ may be higher because the test may work better among these individuals than among those who, while being $\overline{V}$, have other co-morbidities that may confound diagnostic testing. Likewise, if only individuals in advanced stages of the disease are part of the V, omitting the intermediate and generally more difficult to diagnose cases, the estimates of $\mathbb{P}(T|V)$ may be overly optimistic. In these cases, the estimates obtained do not adequately generalize to the rest of the future cases.

The reason for the names of the two probabilities is as follows:

- Is the test sensitive to the presence of the virus? That is, if the virus is present, does the test produce the event T?
- Is the test specific for the HIV virus? That is, if the patient has anything other than the virus, the test should not be positive.

Suppose $\mathbb{P}(T|V) = 0.99$ (sensitivity) and $\mathbb{P}(\overline{T}|\overline{V}) = 0.95$ (specificity). Both odds are relatively high, which is good. But we will see that this is not enough for a mass test.

Definition 4.4.3 Complementary probabilities are associated with misdiagnosis and doctors use two terms for them:

- *False positive* (FP): T for a patient who is $\overline{V}$
- *False Negative* (FN): $\overline{T}$ for a patient who is V

FP and FN probabilities are derived directly from sensitivity and specificity:

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

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

From the frequencies in the confusion table, we can estimate these false positive probabilities. $\mathbb{P}(FP)$ and false negative $\mathbb{P}(FN)$:

Virus?	Test Result		Total
	T	$\overline{T}$	
V	sens $\mathbb{P}(T\|V)$	1 - sens $\mathbb{P}(FN) = \mathbb{P}(\overline{T}\|V)$	1.0
$\overline{V}$	1-esp $\mathbb{P}(FP) = \mathbb{P}(T\|\overline{V})$	esp $\mathbb{P}(\overline{T}\|\overline{V})$	1.0

But we don't just want $\mathbb{P}(FP) = \mathbb{P}(T|\overline{V})$ and $\mathbb{P}(FN) = \mathbb{P}(\overline{T}|V)$. More important than getting these two probabilities is calculating the *inverse conditional probabilities*. The doctor has the T result of a patient's examination in hand. Given that he has this result T, what is the probability that the patient has the virus? That is, what is the value of $\mathbb{P}(V|T)$? Likewise, we want to know the

Figure 4.1: Left: Event T and Ω partitioned as $\Omega = V \cup \overline{V}$. Right: Decomposing the event $T = [T \cap V] \cup [T \cap \overline{V}]$.

value of the other inverted conditional probability, the probability $\mathbb{P}(\overline{V}|\overline{T})$. Armed with an estimate of $\mathbb{P}(V)$, we use Bayes' rule to obtain these inverse probabilities.

We have $\mathbb{P}(V) = 0.001$, a rough estimate provided by the Ministry of Health. This is the estimate of the prevalence of the virus in the general population. If we don't have much confidence in this estimator value of $\mathbb{P}(V)$, we can calculate the probabilities with several plausible scenarios. giving different values for $\mathbb{P}(V)$, we can check how the probabilities of interest ($\mathbb{P}(V|T)$ and $\mathbb{P}(\overline{V}|\overline{T})$) change. Perhaps they will not change much, or even if they do, the conclusions we will draw about appropriate health policies may remain the same.

By Bayes' rule, we have:

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

To get $\mathbb{P}(T)$, we are going to use a very useful trick based on set intersection. The left side of Figure 4.1 shows the event T and the decomposition of Ω obtained by the partition $\Omega = V \cup \overline{V}$ with $V \cap \overline{V} = \emptyset$. On the right side of Figure 4.1, the event T and the partition $\Omega = V \cup \overline{V}$ are mixed. Using elementary set operations, we write the event T as a union of two disjoint sets:

$$\begin{aligned} T &= T \cap [V \cup \overline{V}] && (4.2)\\ &= [T \cap V] \cup [T \cap \overline{V}] && (4.3) \end{aligned}$$

Now we can calculate $\mathbb{P}(T)$ easily:

$$\begin{aligned} \mathbb{P}(T) &= \mathbb{P}([T \cap V] \cup [T \cap \overline{V}]) \\ &= \mathbb{P}(T \cap V) + \mathbb{P}(T \cap \overline{V}) \quad \text{because they are disjoint events} \\ &= \mathbb{P}(T|V) \cdot \mathbb{P}(V) + \mathbb{P}(T|\overline{V}) \cdot \mathbb{P}(\overline{V}) \text{using prob definition. conditional} \\ &= 0.99 \cdot 0.001 + 0.05 \cdot 0.999 = 0.05094 \end{aligned}$$

Finishing the calculation of Bayes' rule, we have:

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

So if we have the test produce a positive result, (T), the chance of the patient being $\overline{V}$, or not having the virus, will be very high. Only 2% of individuals with a positive test (T) actually have the virus. Ditto, calculating the other inverse probability:

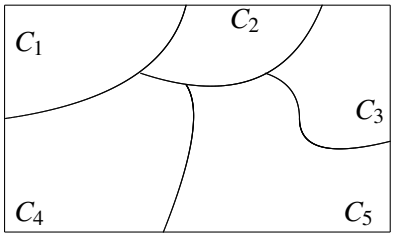
$$\begin{aligned}\mathbb{P}(\overline{V}|\overline{T}) &= \frac{\mathbb{P}(\overline{T}|\overline{V})\mathbb{P}(\overline{V})}{\mathbb{P}(\overline{T})} \\ &= \frac{0.95\times(1-0.001)}{1-0.05094} = 0.9999895\end{aligned}$$

If the test is negative, the individual is virtually certain to be $\overline{V}$.

These calculations show why we don't do a mass screening of the Brazilian population. As $\mathbb{P}(V|\overline{T}) = 1-\mathbb{P}(\overline{V}|\overline{T}) \approx 0$, if the test does not detect the virus, the chance of the individual being really infected is very low. OK, this is a desired result. The problem is that $\mathbb{P}(\overline{V}|T) \approx 1$. That is, almost everyone detected by the positive test will not be infected. How many people would test positive (falsely or correctly) for the test? That is, how many would have the result T? Approximately 200 million $\times\,\mathbb{P}(T) \approx 10$ million, a huge number of people. Of these, 98% (or 9.8 million) do not have HIV: the vast majority of a huge number of people. The logistical difficulties of ensuring that everyone gets tested, the political discussions about the individual's right not to be subject to this imposition of the state, the enormous cost involved, all this leads to another strategy: to actively search among people from groups of risk (which would have much higher $\mathbb{P}(V)$).

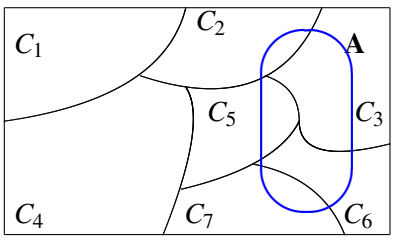
4.4.3 Total probability rule

In Bayes' rule, we derive a very useful formula, called the total probability formula. Let's see the general case. The sample space Ω is partitioned into the events $C_1, C_2, \ldots, C_k$.



$$\Omega = C_1 \cup C_2 \cup \ldots \cup C_k$$
e
$$C_i \cap C_j = \emptyset \quad se\ i \neq j$$

With these C_j's, we can break an arbitrary event A into disjoint pieces and write A as the union of disjoint subsets:



$$\begin{aligned}A &= A \cap \Omega \\ &= A \cap (C_1 \cup C_2 \cup \ldots \cup C_k) \\ &= (A \cap C_1) \cup (A \cap C_2) \cup \ldots \cup (A \cap C_k)\end{aligned}$$

We have then

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

4.4.4 Extension of the Bayes Rule

Considering the sample space Ω partitioned into the events $C_1, \ldots, C_k$, we can express the denominator $\mathbb{P}(A)$ of Bayes' rule using the conditional probabilities $mathbbP(A|C_j)$:

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

This is the general form of Bayes' rule which we have now prominently stated:

Definition 4.4.4 — Bayes Rule - Full Version.

$$\mathbb{P}(C_i|A) = \frac{\mathbb{P}(A|C_i)\mathbb{P}(C_i)}{\mathbb{P}(A|C_1)\mathbb{P}(C_1)+\ldots+\mathbb{P}(A|C_k)\mathbb{P}(C_k)}\,.$$

■ **Example 4.5 — Bayes Rule 01.** Daily, a website produces articles divided into three distinct topics: Politics (P), Sports (E), and Culture (C). This is the partition of the sample space composed of site articles. The topic P is responsible for 50% of the articles produced while the topic E produces another 40% and C produces the rest. Thus, $\mathbb{P}(P) = 0.50$, $\mathbb{P}(E) = 0.40$ and $\mathbb{P}(C) = 0.10$. The word *framework* appears in 50% of cultural texts, in 30% of political texts and in only 5% of sports texts.

How were all these numbers obtained? A large collection of articles was separated from the website and they were separated into three groups based on their classified *manual* in each of the three topics. Thus, a stack of articles P, a stack of articles E and a stack of articles C were formed. The size of each of these stacks (the frequency of the 3 topics in the complete collection) gave the estimates $\mathbb{P}(P) = 0.50$, $\mathbb{P}(E) = 0.40$ and $\mathbb{P}(C) = 0.10$. Next, the proportion of articles in which the word *framework* appeared at least once in each of the three groups was measured. This provided the estimates below:

- $\mathbb{P}(\text{framework}|C) = 0.50$,
- $\mathbb{P}(\text{framework}|P) = 0.30$,
- $\mathbb{P}(\text{framework}|E) = 0.05$,

Note that the probabilities above do not add up to 1. They refer to different universes. The complement of the probability $\mathbb{P}(\text{framework}|E) = 0.05$ is the probability $1 - \mathbb{P}(\text{framework}|E) = \mathbb{P}(\text{not appear } \textit{framework} \text{ in the article}|E) = 0.95$.

An automatic text classifier receives an article as input and checks the words present. Let's denote by A the event in which the word *framework* is present in a given article. How likely is it to be from the Culture topic? By the Bayes rule,

$$\begin{aligned}\mathbb{P}(C|A) &= \frac{\mathbb{P}(A|C)\mathbb{P}(C)}{\mathbb{P}(A|C)\mathbb{P}(C)+\mathbb{P}(A|P)\mathbb{P}(P)+\mathbb{P}(A|E)\mathbb{P}(E)} \\ &= \frac{0.50\cdot 0.10}{0.50\cdot 0.10+0.30\cdot 0.50+0.05\cdot 0.40} \\ &= 0.23\end{aligned}$$

■

■ **Example 4.6 — Bayes Rule 02.** Let's expand a little on the example **??** keeping the same probabilities of the three topics: $\mathbb{P}(P) = 0.50$, $\mathbb{P}(E) = 0.40$ and $\mathbb{P}(C) = 0.10$. Instead of a single word, real topic identification models use the entire dictionary. Let's stay in an "intermediate" position using just 2 distinct words: $w_1 = \textit{capacity}$ and $w_2 = \textit{pretentious}$. The single A event

from the previous example now needs to be extended. The possibilities now must contemplate the occurrence or not of the two words simultaneously. Let's use w or $\overline{w}$ to denote the presence or absence of the word w in a text. We will then have four possible events considering w_1 and w_2:

- $A_{11} = [w_1, w_2]$ (both present)
- $A_{10} = [w_1, \overline{w_2}]$ (only w_1 is present)
- $A_{01} = [\overline{w_1}, w_2]$ (only w_2 is present)
- $A_{01} = [\overline{w_1}, \overline{w_2}]$ (both missing)

We also need the probabilities $\mathbb{P}(A_{ij}|P)$, $\mathbb{P}(A_{ij}|E)$, $\mathbb{P}(A_{ij}|C)$. They are in the table below:

	P	E	C
A_{11}	0.05	0.01	0.15
A_{10}	0.15	0.05	0.25
A_{01}	0.10	0.04	0.25
A_{00}	0.70	0.09	0.35
Total	1	1	1

In this table, the sum along each column is 1 but the sum along each row will not be 1 (unless by chance). A new text appears and the A_{01} event occurs. How likely is this new text to be cultural? Direct answer by the Bayes rule:

$$\begin{aligned}
\mathbb{P}(C|A_{01}) &= \frac{\mathbb{P}(A_{01}|C)\mathbb{P}(C)}{\mathbb{P}(A_{01}|C)\mathbb{P}(C)+\mathbb{P}(A_{01}|P)\mathbb{P}(P)+\mathbb{P}(A_{01}|E)\mathbb{P}(E)} \\
&= \frac{0.25\cdot 0.10}{0.25\cdot 0.10+0.10\cdot 0.50+0.04\cdot 0.40} \\
&= 0.27
\end{aligned}$$

The probability of the text being of culture went from the probability *a priori* $\mathbb{P}(C) = 0.10$ to $\mathbb{P}C|A_{01}) = 0.27$ after observing partial information over the text, after observing the event A_{01} corresponding to two words. We say that we *update* the probability *a priori* after receiving certain information.

The largest practical difficulty appears when trying to expand this example to a large collection of words. Within the collection of texts manually labelled as *Culture*, we need to estimate the chance of the two words occurring together or not. Imagine that, instead of two words, we use 1000 words from the dictionary. The table we need now has 2^{1000} lines (we will have 2^{1000} events representing the occurrence or not of $w_1, w_2, \ldots, w_{1000}$. it is not feasible to estimate the probabilities in this table astronomical in size. The Naive Bayes method seeks to solve this difficulty by assuming a certain type of independence from the events involved. See section ??(to be completed).

■

■ **Example 4.7 — Bayes Rule 03.** Components are shipped in batches of 10 units. Assume that 95% of the lots have none of their 10 defective items, 4% have 1 defective item, and 1% have two or more defectives. Two items are selected at random from each lot and they are tested. Let A_0, A_1 and A_2 be the events in which 0, 1, and 2 of the tested items are defective. Knowing that one of these events happened, we can recalculate the probabilities of: L_0, the lot has 0 defectives; L_1, the batch has 1 defective; L_2, the batch has 2 or more defectives.

Let's initially assume that A_0 occurs. We want $\mathbb{P}(L_0|A_0)$, $\mathbb{P}(L_1|A_0)$ and $\mathbb{P}(L_2|A_0)$. When using Bayes' rule, we're going to need the inverse probabilities. To obtain them, let's recall a result of discrete mathematics: in how many different ways can we choose a subset of k elements from the possible n? This number is the binomial coefficient:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

This formula is valid even in extreme cases, with $k = 0$ and $k = n$. So, in a batch of 10 pieces, the number of ways to select two of your items is $10!/(2!8!) = 45$. As the items are chosen at random, the sample space is composed of these 45 pairs of elements, all with the same probability 1/45.

- $\mathbb{P}(A_0|L_0) = 1$ because, if the entire lot has no defects, then the sample will certainly not contain defects.
- $\mathbb{P}(A_0|L_1)$: if the batch has 1 defective and 9 non-defective, the number of pairs of elements where none is defective is equal to the number of ways to select 2 elements from the 9 that are in good condition: $9!/(2!7!) = 36$. Therefore, since any pair has probability $1/45$, we have $\mathbb{P}(A_0|L_1) = 36/45$.
- $\mathbb{P}(A_0|L_2)$: this calculation is more complicated as we need to know exactly how many defective items exist on the lot. One approximation is to assume that 3 or more defectives in a batch is an event so rare that it can be ignored. So, assuming there are exactly two bad items in the 10-component lot, there are $8!/(2!6!) = 28$ ways to pick two of your good items. So, $\mathbb{P}(A_0|L_2) = 28/45$.

Now using Bayes' rule:

$$\begin{aligned}\mathbb{P}(L_0|A_0) &= \frac{\mathbb{P}(A_0|L_0)\cdot\mathbb{P}(L_0)}{\mathbb{P}(A_0|L_0)\cdot\mathbb{P}(L_0)+\mathbb{P}(A_0|L_1)\cdot\mathbb{P}(L_1)+\mathbb{P}(A_0|L_2)\cdot\mathbb{P}(L_2)}\\ &= \frac{1\cdot 0.95}{1\cdot 0.95+0.8\cdot 0.04+0.62\cdot 0.01} = 0.961\end{aligned}$$

$$\begin{aligned}\mathbb{P}(L_1|A_0) &= \frac{\mathbb{P}(A_0|L_1)\cdot\mathbb{P}(L_1)}{\mathbb{P}(A_0|L_0)\cdot\mathbb{P}(L_0)+\mathbb{P}(A_0|L_1)\cdot\mathbb{P}(L_1)+\mathbb{P}(A_0|L_2)\cdot\mathbb{P}(L_2)}\\ &= \frac{0.8\cdot 0.04}{1\cdot 0.95+0.8\cdot 0.04+0.62\cdot 0.01} = 0.032\end{aligned}$$

$$\begin{aligned}\mathbb{P}(L_2|A_0) &= \frac{\mathbb{P}(A_0|L_2)\cdot\mathbb{P}(L_2)}{\mathbb{P}(A_0|L_0)\cdot\mathbb{P}(L_0)+\mathbb{P}(A_0|L_1)\cdot\mathbb{P}(L_1)+\mathbb{P}(A_0|L_2)\cdot\mathbb{P}(L_2)}\\ &= \frac{0.62\cdot 0.01}{1\cdot 0.95+0.8\cdot 0.04+0.62\cdot 0.01} = 0.006\end{aligned}$$

The denominator was the same in the three calculations above.

Let us now obtain the probability assuming that the observed event was A_1, a defective one in the 2-item sample. We have:

- $\mathbb{P}(A_1|L_0) = 0$: since the batch has no defects, it is impossible to obtain a sample with defects.
- $\mathbb{P}(A_1|L_1)$: if the batch has 1 defective and 9 non-defective, let's count the number of pairs of elements where exactly one is defective. One of the sample elements is the only one defective in the lot. There are 9 good items left from which we must choose one to compose the rest of the sample. So there are only 9 possibilities and $\mathbb{P}(A_1|L_1) = 9/45 = 0.20$.
- $\mathbb{P}(A_1|L_2)$: As before, assuming there are exactly two defective items in the batch of 10 components, there are $8!/(1!7!) = 8$ ways to choose one of your own 8 good items. For each of these choices, we have 2 defective items from which to select one. So we have $8\cdot 2 = 16$ possible pairs and therefore $\mathbb{P}(A_1|L_2) = 16/45 = 0.36$

Using Bayes' rule again:

$$\begin{aligned}\mathbb{P}(L_0|A_1) &= \frac{\mathbb{P}(A_1|L_0)\cdot\mathbb{P}(L_0)}{\mathbb{P}(A_1|L_0)\cdot\mathbb{P}(L_0)+\mathbb{P}(A_1|L_1)\cdot\mathbb{P}(L_1)+\mathbb{P}(A_1|L_2)\cdot\mathbb{P}(L_2)}\\ &= \frac{0\cdot 0.95}{0\cdot 0.95+0.20\cdot 0.04+0.36\cdot 0.01} = 0\end{aligned}$$

$$\begin{aligned}\mathbb{P}(L_1|A_1) &= \frac{\mathbb{P}(A_1|L_1)\cdot\mathbb{P}(L_1)}{\mathbb{P}(A_1|L_0)\cdot\mathbb{P}(L_0)+\mathbb{P}(A_1|L_1)\cdot\mathbb{P}(L_1)+\mathbb{P}(A_1|L_2)\cdot\mathbb{P}(L_2)}\\ &= \frac{0.20\cdot 0.04}{0\cdot 0.95+0.20\cdot 0.04+0.36\cdot 0.01}=0.690\end{aligned}$$

$$\begin{aligned}\mathbb{P}(L_2|A_1) &= \frac{\mathbb{P}(A_0|L_2)\cdot\mathbb{P}(L_2)}{\mathbb{P}(A_1|L_0)\cdot\mathbb{P}(L_0)+\mathbb{P}(A_1|L_1)\cdot\mathbb{P}(L_1)+\mathbb{P}(A_1|L_2)\cdot\mathbb{P}(L_2)}\\ &= \frac{0.62\cdot 0.01}{0\cdot 0.95+0.20\cdot 0.04+0.36\cdot 0.01}=0.310\end{aligned}$$

The denominator was the same in the three calculations above. Let's leave it as an exercise to consider the case where A_2 occurs.

■

■ **Example 4.8 — Bayes Rule 04.** In an urn, there are 6 balls of unknown colors. Three balls are drawn at random without replacement and are all black. Find the probability that there are no black balls left in the urn.

Let's define the event A representing the withdrawal of 3 black balls from the urn. Let C_i be the event that there are i black balls in the urn with $i = 0,1,\ldots,6$. We want to calculate $\mathbb{P}(C_3|A)$. A much easier probability is the inverse conditional. Let's start with:

$$\mathbb{P}(A|C_0)=\frac{\mathbb{P}(A\cap C_0)}{\mathbb{P}(C_0)}=\frac{0}{\mathbb{P}(C_0)}=0$$

because if we know that there are no black balls in the urn (C_0 is given) then the probability is zero of getting a sample with 3 black balls from the same urn. By the same argument, we have $\mathbb{P}(A|C_1)=\mathbb{P}(A|C_2)=0$.

Consider now $\mathbb{P}(A|C_3)$. There are $6!/(3!3!)=20$ ways to select 3 of the 6 balls in the urn, all equally likely. If there are 3 black balls in the 6-ball urn, in how many ways can we select a sample of 3 elements from the urn, all three being black balls. There is only one way, which is the sample formed by the exact 3 existing black balls. So $\mathbb{P}(A|C_3)=1/20$.

For $\mathbb{P}(A|C_4)$, in how many ways can we select 3 black balls from the urn that contains 4 of them? There are $4!/(3!1!)=4$ ways to do this and so $\mathbb{P}(A|C_4)=4/20=0.20$. For $\mathbb{P}(A|C_5)$, we have $5!/(3!2!)=10$ ways to select the samples with 3 black balls and therefore $\mathbb{P}(A|C_5)=10/20=0.50$. Finally, in the last case, given C_6, all the balls in the urn are black and we are guaranteed to have a sample with 3 black balls. Therefore, $\mathbb{P}(A|C_6)=1$.

To calculate $\mathbb{P}(C_3|A)$, we use Bayes' rule:

$$\mathbb{P}(C_3|A)=\frac{\mathbb{P}(A|C_3)*\mathbb{P}(C_3)}{\sum_{j=0}^{6}\mathbb{P}(A|C_j)*\mathbb{P}(C_j)}$$

$$=\frac{\frac{1}{20}*\mathbb{P}(C_3)}{0+0+0+\frac{1}{20}\mathbb{P}(C_3)+\frac{1}{5}\mathbb{P}(C_4)+\frac{1}{2}\mathbb{P}(C_5)+1\mathbb{P}(C_6)}=??$$

We need to establish the value of $\mathbb{P}(C_j)$, the probabilities that there are j black balls in the urn. This depends on the mechanism that puts the balls in the urn, something that was not explained in the problem. Let's show some possibilities for $\mathbb{P}(C_j)$.

A first scenario is that any number of black balls between 0 and 6 has the same probability. So $\mathbb{P}(C_j)=\frac{1}{7}$ for all j. Just substitute these values now into the above formula for $\mathbb{P}(C_3|A)$ finding 0.029, a very small probability.

As a second scenario, suppose that the balls are chosen preferably of a single color. So the values of $\mathbb{P}(C_j)$ for $j=0$ and $j=6$ would be the largest, with a minimum value for $j=3$. For example, $\mathbb{P}(C_j)=\frac{1}{35}(1+(j-3))^2$. This means that $\mathbb{P}(C_j)$ is equal to $10/35, 5/35, 2/35, 1/35, 2/35, 5/35, 10/35$ for $j=0,\ldots,6$, respectively. Substituting into the above formula for $\mathbb{P}(C_3|A)$ we find 0.0038, an even smaller probability than the previous one.

Another scenario option is as follows: There are 10 different colors and the color of each of the 6 balls in the urn is chosen at random among the 10 possible colors. The chance of placing a black ball in the urn is $1/10$. It can be shown, using the binomial distribution presented in the 6 chapter, that the chance of putting j black bags in the 6-ball urn is

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

This means that $\mathbb{P}(C_j)$ is equal to $0.5314, 0.3543, 0.0984, 0.0146, 0.0012, 10^{-5}, 10^{-6}$ for $j=0$, $ldots$, 6, respectively. Substituting into the above formula for $\mathbb{P}(C_3|A)$ we find 0.729, a substantially higher probability than the previous ones. So the answer depends on additional knowledge about how the balls appear in the urn.

■

4.5 Conditional as a new probability measure

Let $(\Omega,\mathscr{A},\mathbb{P})$ be any probability space. Let's fix an event $A\subseteq\Omega$ with $\mathbb{P}(A)>0$. We know that for any event $B\subseteq\Omega$, we calculate $\mathbb{P}(B|A)$ through the definition $\mathbb{P}(B|A)=\mathbb{P}(A\cap B)/\mathbb{P}(A)$. Instead of thinking about this conditional probability just for a pair of events A and B, we can fix A and calculate the conditional probability given A for a series of events B_1, B_2, $B_3,\ldots$. In fact, we want to fix A and recalculate the probability (conditioned on the occurrence of A) of the event B for each and every $B\subseteq\Omega$. That is, we want to assign a new probability measure $\mathbb{P}_A$ to the events of σ-algebra $\mathscr{A}$. Otherwise, we want to create a new probability space $(\Omega,\mathscr{A},\mathbb{P}_A)$ where Ω and $\mathscr{A}$ are the same as in original probability space but the probability function changes from $\mathbb{P}$ to $\mathbb{P}_A$. Now, for each event B in σ-algebra $\mathscr{A}$, we calculate its probability as

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

See that we are defining a probability function $\mathbb{P}_A(\,\cdot\,)$ with an argument in $\mathscr{A}$:

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

For this function to be a valid probability assignment to events $B\subseteq\mathscr{A}$, it must obey Kolmogorov's three axioms:

Axiom 1 $\mathbb{P}_A(B)\geq 0\ \ \forall\ \ B\in\mathscr{A}$

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

Axiom 3 $\mathbb{P}_A(B_1\cup B_2\cup B_3\cup\ldots)=\mathbb{P}_A(B_1)+\mathbb{P}_A(B_2)+\mathbb{P}_A(B_3)+\ldots$ if events $B_1,B_2,\ldots$ are disjoint (ie, mutually exclusive).

This actually happens because:

- $\mathbb{P}_A(B)=\mathbb{P}(A\cap B)/\mathbb{P}(A)\geq 0$ because $\mathbb{P}(A\cap B)\geq 0$ and $\mathbb{P}(A)>0$.
- $\mathbb{P}_A(\Omega)=\mathbb{P}(\Omega\cap A)/\mathbb{P}(A)=\mathbb{P}(A)/\mathbb{P}(A)=1$
- As for the last property, if $B_1,B_2,\ldots$ are disjoint, then the events $[B_1\cap A],[B_2\cap A],\ldots$ are

also disjoint. So,

$$\begin{aligned}
\mathbb{P}_A(B_1 \cup B_2 \cup B_3 \cup \ldots) &= \frac{\mathbb{P}([B_1 \cup B_2 \cup B_3 \cup \ldots] \cap A)}{\mathbb{P}(A)} \\
&= \frac{\mathbb{P}((B_1 \cap A) \cup (B_2 \cap A) \cup \ldots)}{\mathbb{P}(A)} \\
&= \frac{\mathbb{P}(B_1 \cap A) + \mathbb{P}(B_2 \cap A) + \ldots}{\mathbb{P}(A)} \\
&= \frac{\mathbb{P}(B_1 \cap A)}{\mathbb{P}(A)} + \frac{\mathbb{P}(B_2 \cap A)}{\mathbb{P}(A)} + \ldots \\
&= \mathbb{P}(B_1|A) + \mathbb{P}(B_2|A) + \ldots \\
&= \mathbb{P}_A(B_1) + \mathbb{P}_A(B_2) + \ldots
\end{aligned}$$

Thus, the probability function $\mathbb{P}_A(\,\cdot\,) = \mathbb{P}(\,\cdot\,|A)$ reassigns probabilities to the events of σ-algebra $\mathscr{A}$ creating a new probability space. This means that *all* properties valid for any probability function are also valid for the particular probability function $\mathbb{P}_A$. For example, the properties seen in **??** are valid for $\mathbb{P}_A(\,\cdot\,) = \mathbb{P}(\,\cdot\,|A)$. We have

(P1) $\mathbb{P}(B^c|A) = 1 - \mathbb{P}(B|A)$.
(P2) $0 \leq \mathbb{P}(B|A) \leq 1$ for every event $B \in \mathscr{A}$.
(P3) if $B_1 \subseteq B_2 \Longrightarrow \mathbb{P}(B_1|A) \leq \mathbb{P}(B_2|A)$
(P4) $\mathbb{P}(\bigcup_{n=1}^{\infty} B_i|A) \leq \sum_{n=1}^{\infty} \mathbb{P}(B_i|A)$
(P5) $\mathbb{P}(B \cup C|A) = \mathbb{P}(B|A) + \mathbb{P}(C|A) - \mathbb{P}(B\ capC|A)$

No further demonstration of these facts is required. These properties are valid and follow immediately from the fact that $\mathbb{P}_A$ is a probability measure over Ω that satisfies Kolmogorov's three axioms.

4.6 Mutual Independence

We speak of the independence of two events A and B. They are independent events if

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

or, equivalently,

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

And when we have multiple events $E_1, E_2, \ldots, E_n$? When are these events independent of each other? How to move from a pair of events to a set of events? Unfortunately, it is not enough to look at all pairs of events and verify that the previous definition is valid for all of them. Events $E_1, E_2, \ldots, E_n$ are independent events if every combination of events satisfies the product rule:

$$\mathbb{P}(E_{i_1} \cap E_{i_2} \cap \ldots \cap E_{i_m}) = \mathbb{P}(E_{i_1}) \ldots\ mathbbP(E_{i_m})$$

for every selection of indices $i_1, i_2, \ldots, i_m$ and for every m between 2 and n. These events are called *mutually independent*.

We can deduce that if A, B, and C are independent then C is also independent of $A \cap B$, of $A \cap B^c$, of $A \cup B$, of B^c etc. (see list of exercises).

If the events are mutually independent then any pair of events is independent. A curious result is that the inverse is not true. We can have pairwise independent events that are not mutually independent. For example, we can have independent A and B, independent A and C, and independent B and C but dependent A, B, C. A practical use of this distinction appears in a technique for sharing passwords in cryptography (see [12], section 8.9.2).

4.7 Paradoxes with conditional probability

Sometimes calculations involving conditional probabilities lead to curious and unintuitive results. Conditional probability is a constant source of paradoxical results. We'll show you some in this section.

■ **Example 4.9 — Not knowing the color of the withdrawn ball.** An urn contains two pink balls and two brown balls (Figure 4.2). One of them is chosen completely at random. Suppose this first ball drawn is pink. It *not* is put back in the urn. Set the R_1 event to $R_1 = [\text{1a. pink ball}]$. A second ball is drawn by choosing one of the remaining three completely at random. Let R_2 be the event $R_2 = [\text{2a. pink ball}]$. What is the probability that R_2 occurs given that R_1 has occurred? Since R_1 has occurred, one blue and two brown balls remain in the urn. So $\mathbb{P}(R_2|R_1) = 1/3$. If the first ball drawn is brown, an event represented by M_1, we have $\mathbb{P}(R_2|M_1) = 2/3$.

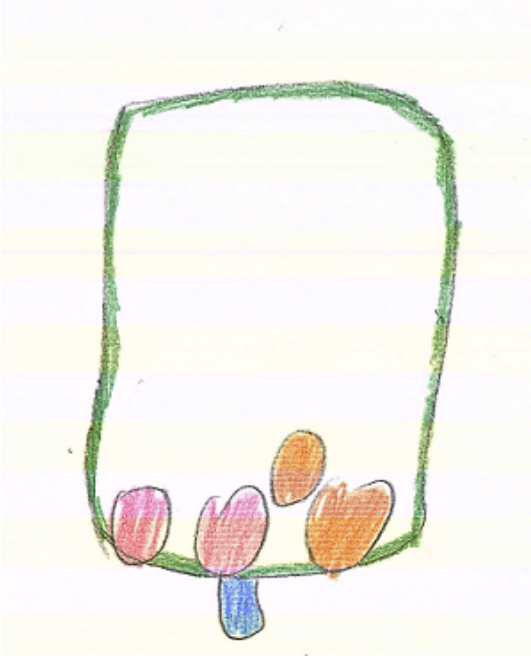


Figure 4.2: Urn with two pink balls and two brown balls, designed by Sofia. A ball is drawn at random and is not replaced in the urn. If its color is not recorded, what is the probability that a second ball drawn is pink?

The curious thing comes now: suppose a ball is drawn but we don't know what color it is. It was definitely a pink or brown ball, we just don't know which one. Let's denote by B the event that a ball of unknown color was drawn. Note that $B = R_1 \cup M_1$ and that R_1 and M_1 are disjoint: $R_1 \cap M_1 = \emptyset$. Let's calculate $\mathbb{P}(R_2|B) = \mathbb{P}(R_2|R_1 \cup M_1)$ and check that this probability is equal to 1/2, the same probability $\mathbb{P}(R_1)$ to draw a pink ball when the urn is full. That is, drawing a ball of unknown color does not change the probability of R_2 occurring, but if the color of the drawn ball is revealed, then the probability changes greatly.

As $B = R_1 \cup M_1$ with $R_1 \cap M_1 = \emptyset$. We then have $R_2 \cap B = R_2 \cap [R_1 \cup M_1] = [R_2 \cap R_1] \cup [R_2 \cap M_1]$, the union of two disjoint sets. So,

$$\begin{aligned}
\mathbb{P}(R_2 \cap B) &= \mathbb{P}(R_2 \cap R_1) + \mathbb{P}(R_2 \cap M_1) \\
&= \mathbb{P}(R_2|R_1)\mathbb{P}(R_1) + \mathbb{P}(R_2|M_1)\mathbb{P}(M_1) \\
&= \frac{1}{3}\frac{1}{2} + \frac{2}{3}\frac{1}{2} = \frac{1}{2}
\end{aligned}$$

■

■ **Example 4.10 — Paradox of the three prisoners.** Three prisoners in a medieval dungeon, A, B and C, were sentenced to death. Feeling blessed by the birth of his heir, the feudal lord decides that one of them will be chosen at random by lot and released. A learns of the decision and tells his jailer: " I already know that at least one of the other two prisoners, B or C, will be executed. Could you then tell me the name of one of them that will be executed? As we are incommunicado, this information will be of no use. "Convinced by the argument, the jailer says that C will be executed. Prisoner A reasons that his chance of being released was 1/3, but now that he knows it's just him and B left, his chance has gone up to 1/2. Or not?

Initially, there are three possible outcomes, all with equal probabilities:

- A will be released, with probability $1/3$.
- B will be released, with probability $1/3$.
- C will be released, with probability $1/3$.

With the information to be provided by the jailer, there are four possibilities with different probabilities:

- $[A \text{ lib}, B \text{ exec}]$: A will be released and A is informed that B will be executed, with probability $1/6$.
- $[A \text{ lib}, C \text{ exec}]$: A will be released and A is informed that C will be executed, with probability $1/6$.
- $[B \text{ lib}, C \text{ exec}]$: B will be released and A is informed that C will be executed, with probability $1/3$.
- $[C \text{ lib}, B \text{ exec}]$: C will be released and A is informed that B will be executed, with probability $1/3$.

The probability that the jailer reports that C will be executed is equal to

$$\mathbb{P}(C \text{ exec}) = \mathbb{P}([A \text{ lib}, C \text{ exec}] \cup [B \text{ lib}, C \text{ exec}]) = \mathbb{P}(A \text{ lib}, C \text{ exec}) + (B \text{ lib}, C \text{ exec}) = \frac{1}{6} + \frac{1}{3} = \frac{1}{2}$$

Given that the jailer informs that C will be executed, the conditional probability of interest is

$$\mathbb{P}(A \text{ lib} \mid C \text{ exec}) = \frac{\mathbb{P}([A \text{ lib}} \cap C \text{ exec }])}{\mathbb{P}(C \text{ exec})} = \frac{1/6}{1/2} = \frac{1}{3}.$$

Thus, the probability that A will be released does not change with the information provided. ■

■ **Example 4.11 — Monty Hall.** To complete - see `https://pt.wikipedia.org/wiki/Problema_de_Monty_Hall` ■

■ **Example 4.12 — Paradox of boy-girl.** To complete - see `https://en.wikipedia.org/wiki/Boy_or_Girl_paradox` ■

■ **Example 4.13 — Simpson's paradox.** To complete - see `https://stats.stackexchange.com/questions/21896/basic-simpsons-paradox?rq=1`

■

■ **Example 4.14 — Difference between disjunction and independence.** Philosophy of cause and effect and conditioning in the future. See Ben-Naim.

■

■ **Example 4.15 — Conditioning direction.** Increased information can lead to conflict and decrease support from one event to another. See Ben-Naim, page 135.

■

4.8 Precision and recall

???

Precision, recall, accuracy, false positive, false negative, ROC. Common metrics in classification, recommendation and information retrieval problems. Recall

5. Introduction to Classification

5.1 Introduction

Supervised classification is a data analysis task that seeks to learn from statistical data how to use the attributes of objects to separate them into two or more distinct sets called classes. Consider a dataset statistics in the usual tabular format where the attributes (columns) are divided into two types. One of the attributes is a nominal variable with the label identifying the real class of the object-row of the table. The other variables are attributes that we want to use to predict the label or class.

Schematically, imagine that we have the following data table with k attributes and n cases (see Tables in (5.1)). The variables (or columns) $X_1, \ldots, X_k$ represent the attributes of the objects. The variable (or column) Y contains the two labels or classes (0 and 1) into which these objects are divided. The attributes of the row i of the table are represented by the vector k-dimensional

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$$

while its corresponding label (or class) is represented by Y_i. The purpose of supervised classification is to learn a function h that takes as input the attributes of an individual and provides as a response a good approximation to the probability of the case receive a label. In the case of only two labels, 0 or 1, we want a function h that gives the probability that the label Y is equal to 1 conditioned on the values of the attributes. That is, given that the vector of attributes of a certain individual is $\mathbf{x}$, approximately obtain the probability that its label or class is 1:

$$\mathbb{P}(Y = 1 | \mathbf{X} = \mathbf{x}) \approx h(\mathbf{x}) \,.$$

For example, for the case of row $i = 3$ of Table **??**, we want the function to provide an estimate or an approximation for

$$\mathbb{P}(Y_3 = 1 | \mathbf{X} = \mathbf{x}_3) = \mathbb{P}\left(Y_3 = 1 | \mathbf{X} = (5.09, 3,\ ldots, 27.91)\right) \approx h(5.09, 3, \ldots, 27.91) = h(\mathbf{x}_3)$$

$$\left[\begin{array}{ccccc} X_1 & X_2 & \dots & X_k & Y \\ \hline x_{11} & x_{12} & \dots & x_{1k} & Y_1 \\ x_{21} & x_{22} & \dots & x_{2k} & Y_2 \\ x_{31} & x_{32} & \dots & x_{3k} & Y_3 \\ x_{41} & x_{42} & \dots & x_{4k} & Y_4 \\ x_{51} & x_{52} & \dots & x_{5k} & Y_5 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} & Y_n \end{array}\right] = \left[\begin{array}{ccccc} X_1 & X_2 & \dots & X_k & Y \\ \hline 1.37 & 3 & \dots & -24.97 & 0 \\ 2.75 & 2 & \dots & 39.55 & 1 \\ 5.09 & 3 & \dots & 27.91 & 0 \\ 3.11 & 3 & \dots & 2.36 & 1 \\ 7.36 & 1 & \dots & 12.99 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 2.22 & 2 & \dots & 65.96 & 0 \end{array}\right] \tag{5.1}$$

In row 3 of the table, the value Y_3 actually observed for this case was label or class 0 (that is, we had the occurrence of the event $Y_3 = 0$) but the function h is giving the probability that Y_3 is equal to 1. However, as there are only two classes in this example, if we want $\mathbb{P}(Y_3 = 0|\mathbf{X} = \mathbf{x}_3)$ just make a subtraction:

$$\mathbb{P}(Y_3 = 0|\mathbf{X} = \mathbf{x}_3) = 1 - \mathbb{P}(Y_3 = 1|\mathbf{X} = \mathbf{x}_3) \approx 1 - h(\mathbf{x}_3)$$

■ **Example 5.1** When a financial institution makes a loan to an individual or a company, it needs to assess the *credit risk* associated with that loan. Credit risk is the chance or probability that the institution will incur a financial loss resulting from the borrower's failure to repay the loan or to meet all contractual obligations (paying too late, for example). When deciding whether or not the requested loan should be granted to that particular customer, the institution estimates the probability that the customer will not honor the terms of the contract in the future. How does she do this? Using statistical data from the past and from other customers. We'll see algorithms for this later.

Having in hand the estimate of the probability of non-payment of that loan in the future by that customer (an estimate of its credit risk, in the jargon of the area), the institution makes its decisions. If the probability is too high, the institution refuses to grant the loan to that customer. If it is moderate, the institution may grant the loan but charging higher interest than usual. If the probability is low, it grants the loan at a low interest rate.

At the time of applying for credit, it is impossible to know exactly what the future result will be. The decision is based on an approximation to the probability that the customer is a high risk and this approximation to the probability is learned or estimated from the statistical data.

Approximations for credit risks (or nonpayment probabilities) are calculated from an algorithm that provides the output $h(\mathbf{x})$ for an individual who has certain attributes. Among the most common attributes used by financial institutions, we can mention: $X_1 =$ the average bank balance in recent months, $X_2 =$ the loan amount relative to this average balance, $X_=$ age of the customer, $X_5 =$ how long they have been a customer, $X_6 =$ their gender (M or F), etc.

Given a profile $bsX = (x_1, x_2, \dots, x_6) = \mathbf{x}$, what is the approximate probability that its label is 1 (non-payment)? That is, we want $\mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) \approx h(\mathbf{x})$. In the next sections, we will see two algorithms to obtain this approximation from historical data. ■

The supervised classification task is one of the most studied in machine learning. Other examples where Y represents the class and $\mathbf{X}$ the attribute set are as follows:

- Classify certain insects (the objects) into one of three subspecies (the three classes of Y) using as attributes $\mathbf{x}$ a set of 12 measurements of morphological (shape) characteristics.
- Classify women as carriers or non-carriers of a genetic disorder (the two classes of Y) using as attributes a vector of measurements in blood proteins and family history.
- Rate the quality of a new cell phone battery as good or bad (the classes in Y) based on some preliminary measurements.
- Classify email messages as spam or non-spam based on characteristics of their header and content.

This supervised classification problem will be studied in detail in chapter 17. Here, we will only superficially present some algorithms. The purpose of the current chapter is to present an interesting example of using the ideas of conditional probability.

5.2 Classification trees

Classification trees is a statistical technique that uses both statistical and algorithmic methods. It does not require knowledge of probability from the user. Trees classify objects by selecting from a large number of variables those that are most important for predicting outcomes. The analysis is based on a binary recursive segmentation. To understand this algorithm, let's start with a very simple example. We want to get an approximation $h(\mathbf{x})$ to the credit risk $\mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x})$ using only two attributes, age The individual's X_1 and their average income of X_2 in the last 6 months. We collected data from 100 customers who recently applied for credit from the institution. The first 10 rows of the table with the collected data are as follows:

```
> cbind(income,age, credit)[1:10,]
      income age credit
 [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
```

Figure 5.1 shows, on its left side, the data of all customers through a scatter plot of the two attributes (age and income). Next, we repeat this graph but now identify each individual as being a bad payer (circle) or a good payer (star). The individuals of these two different labels are in quite separate regions. We can try to partition or segment the variable space (the plan, in this example) so that only one payer class is within each region. A poor attempt to create regions that separate the classes can be seen in the third graph of Figure 5.1 where rectangular tiles are created by the red lines. This attempt is bad because the created regions have both bad and good payers. A much better attempt can be seen in the last graph of Figure 5.1. Note that the separation is now excellent, with each of the four regions containing individuals of only one of the two classes, either the good payers or the bad payers.

The segmentation of the variable space in the last graph of Figure 5.1 can be represented by a binary tree such as the one on the left side of Figure 5.2. Initially, we decide whether the individual has income less than 13K or not. If so, the case goes to the left branch. If negative, it goes to the right branch. Within the left branch (that is, for cases where income is less than 13k), we make a new decision. If the age is less than 53 years old, go to the left branch. Otherwise, go to the right branch. For cases where the income is greater than 13k, the targeting is different. Instead of looking if the age is less than 53 years old, we check if the age is less than 38 years old. If yes, go left. Otherwise, go right. In the end, we are left with four terminal leaves on the tree. In each of the sheets we have cases that all had a single class indicated in the graph of Figure 5.2. On the one hand, every segmentation of the plane with tiles in the form of rectangles can be expressed by a binary tree like this. On the other hand, every binary tree created with the branches dividing each branch based on a single attribute above or below a certain threshold generates a segmentation of

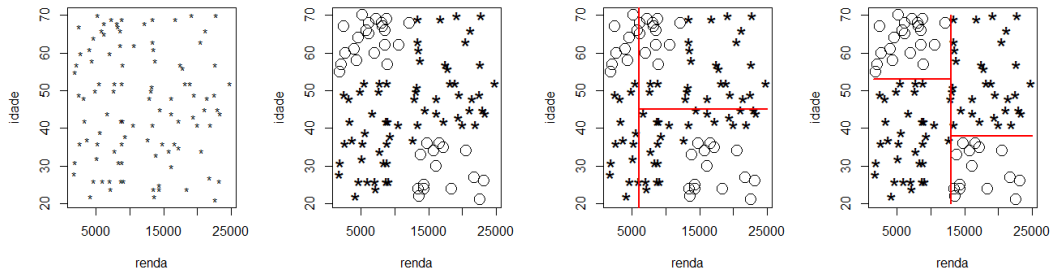


Figure 5.1: Left: Income and age data for borrowers. on the right, the same graph identifying the bad ones (circle) and the good payers (star). Next, a bad segmentation and a good segmentation in terms of class separation based on income and age attributes.

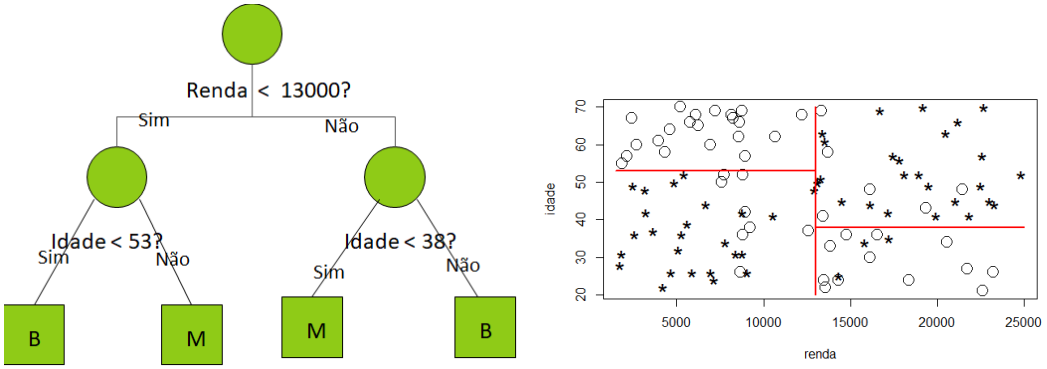


Figure 5.2: Left: Tree representation of the segmentation in the last graph of Figure 5.1. A slightly more realistic scatterplot: no four-region segmentation can separate complete good and bad payers.

the space partitioned into rectangles with sides parallel to the coordinate axes. From now on, we will alternatively use the representation of a given segmentation as a binary tree or as rectangular tiles in the plane.

The segmentation in the last graph of Figure 5.1 represented in this tree is perfect but it rarely happens in practice. In general, a segmentation with straight lines like the ones we have done so far will not be able to completely separate the data into groups composed of only a single class. Consider, for example, the example in the second graph of Figure 5.2. In this case, no segmentation into four regions can separate the good and bad payers. We can create a larger number of regions that have less label variation, or we can just be satisfied with the best possible segmentation into four regions. Ideally, we want perfect separation: in each segment, we have only one class, good or bad. In practice, we look for the percentage of good inside the terminal nodes (or regions of the segmentation) close to 100% or 0% in each segment. We calculate a measure of impurity on each segment created: the further away from 100% or 0%, the more impure the node or region created.

How to get good segmentation? There is a simple iterative algorithm. Let's start by getting the first separation.

- Traverse the horizontal axis (income) stopping at each possible value and doing a segmentation:
  - Calculate the impurity of the two resulting segments.
  - Choose that income value that produces the least impurity in the segments.
  - This is the best possible separation considering an income-based cut.

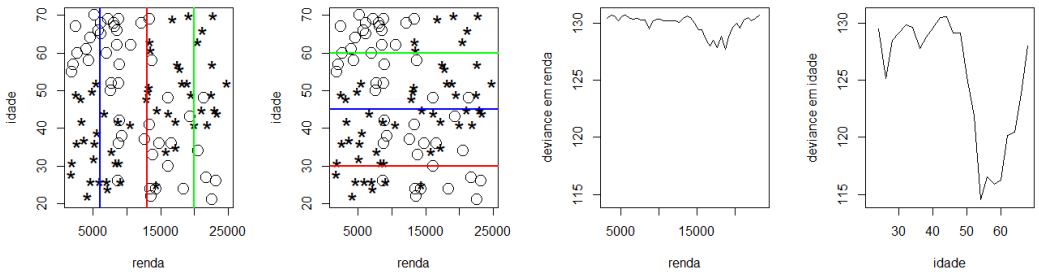


Figure 5.3: Left:

  - Don't do this segmentation yet; wait for the result of the next step described below.
- Returning to the original graph, traverse the vertical axis (age) stopping at every possible value and doing a segmentation:
  - Calculate the impurity of the two resulting segments.
  - Choose that age value that produces the least impurity in the segments.
  - This is the best possible separation considering an age-based cut.

One of the two, either the best age-based targeting or the best income-based targeting, should be better. Choose the best one and finally do the first segmentation.

In Figure 5.3 , we show in the first graph three partition alternatives (vertical colored lines) based on three cut-off points of the income variable (6K, 13K, 20K). In the second graph, we have three partition alternatives (horizontal colored lines) based on three age variable cut-off points (30, 45, 60). Considering these 6 partition alternatives, the one that leads to the lowest degree of impurity in the resulting nodes (in the resulting regions) is chosen. In this case, the degree of impurity of the vertical lines 6K, 13K and 20K is 130.6, 130.3, 129.9, while the degree of impurity of the horizontal lines 30, 45 and 60 is 129.1, 130.0 and 116.2. Thus, among these 6 options, the best one is the horizontal line at age equal to 60 years.

The degree of impurity generated by a segmentation can be calculated in several ways and here we use the value of *deviance*. The deviance formula is not relevant now, but for the record, it is a sum over all the leaves on the tree. In this first step, it is just the sum over the two regions generated by the vertical line or the two regions generated by the horizontal line. Let $i = 1.2$ be an íindex for the generated region and $j = 0.1$ an index for the class. Let n_{ij} be the number of objects of class j in region i and $n_{i+} = n_{i0} + n_{i1}$. As $p_{ij} = n_{ij}/n_{i+}$, we then have the deviance given by

$$D = -2\sum_i \sum_j n_{ij} \log(p_{ij})$$

This measure is associated with the famous entropy measure of a probability distribution. If value $n_{ij} = 0$ we have $p_{ij} = n_{ij}/n_{i+} = 0$ and $n_{ij}\log(p_{ij})$ is an expression of undefined value. Using the fact that the limit $x\log(x) \to 0$ when $x \to 0$, we define that $n_{ij}\log(p_{ij}) = 0$ if $n_{ij} = 0$.

Having a defined degree of impurity measure, like the measure D above, we don't need to limit ourselves to just the six alternatives presented in the first two graphs of Figure 5.3 (the three segmentations vertical and the three horizontal segmentations). Let's sweep each coordinate axis, first income and then age, calculating the impurity value at each possible value of a regular grid imposed on the axis. The last two graphs in Figure 5.3 show the value of the impurity measure considering each possible value of the attributes (income or age). As the ideal is the smallest possible impurity, the best segmentation is a horizontal line separating the objects below or above the age of 55 years approximately.

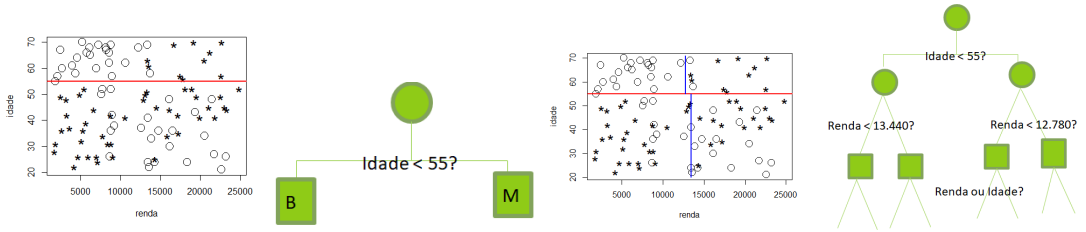


Figure 5.4: First and second segmentation with the corresponding trees.

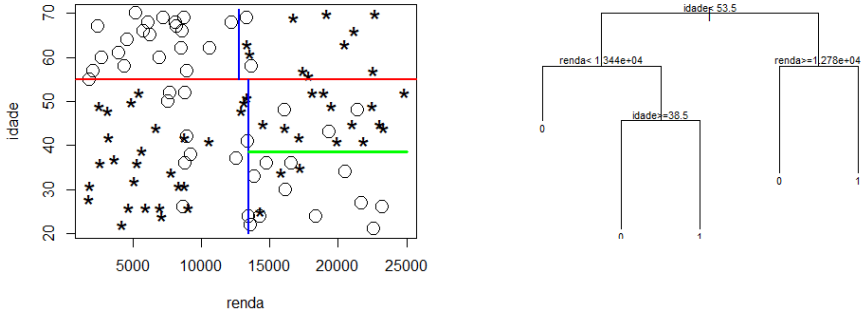


Figure 5.5: Final partition and corresponding tree.

If we do this first segmentation, separating the cases in which Age is less than 55 years old from those in which the Age is greater than or equal to 55 years, we will have the dot plot on the left side of Figure 5.4.

There are now two segments: customers under the age of 55 and those aged 55 or over. The idea of the algorithm is to iterate within each segment created in this first step. Thus, *in the segment aged less than 55 years*, we traverse each of the two coordinate axes, the horizontal axis (income) and the vertical axis *between 20 and 55 years old*, finding on each axis the point that segments vertically or horizontally with the least possible impurity the rectangle of individuals aged less than 55 years. In this case, the best possible segmentation is to use a vertical line cutting this first group into two sub-groups: those with income less than 13,400 and those with income greater than or equal to 13,440.

We repeated this procedure within the second group, that of individuals over 55 years of age. Checking which cut along each of the two attributes generates the least impurity, we found that it is best to partition along the income also breaking between those with income less than 12,780 and those with income greater than or equal to 12,780. The result is the third graph of Figure 5.4.

At this point, the data is segmented into four groups. Now, we simply repeat the search procedure within each of these four segments. In some of the subgroups created, it is not worth segmenting further because they are already reasonably pure (with the proportion of good payers close to 0% or 100%). In others, it may be worth targeting more. In this example, we just do an additional segmentation, along the age attribute, in one of the four segments already created and the final result is in the graph of Figure 5.5. The corresponding tree is also in this figure. The criterion used to stop segmentation in a given group is the result of a statistical test that will not be discussed at this time.

It's easy to see that if we have more than two attributes, the algorithm works the same way. Simply walk through the range one attribute at a time choosing that cutoff point and attribute that best separates the two object classes (minimizes impurity). Then iterate through the algorithm

Branch 1	Branch 2
married	single, divorced, widowed
single	married, divorced, widowed
divorced	single, married, widowed
widowed	single, married, divorced
married, single	divorced, widowed
married, divorced	single, widowed
married, widowed	divorced, single

Table 5.1: Table with possible segmentations based on the categorical attribute *marital status*.

within each segment created. Proceed within each segment that is being created until a statistical test decides that it is no longer worthwhile to continue segmenting.

Sometimes the attribute is not a numeric variable such as income or age. It can be categorical, such as the marital status, with the values *married, single, divorced, widowed* or religion, with the values *catholic, evangelical, atheist, other*. What to do in the case of a categorical attribute like these? If we have m categories in the attribute, then there are $2^{m-1}-1$ possible segments. For example, with four marital states, there are $2^{4-1}-1=7$ possible segments.

For each possible categorical segmentation, calculate the impurity reduction when doing the segmentation. Choose that categorical targeting that produces the greatest impurity reduction. Compare with the impurity reduction that the other attributes (numerical or categorical) provide. Choose the variable and segmentation that produce the maximum impurity reduction.

Proceed by segmenting iteratively. When to stop? When there is no more statistical evidence that further segmenting produces a significant decrease in tree impurity. The statistical test is performed at each possible new segmentation. final tree is the best segmentation you can do with the available data. There are other ways to choose the final tree. For example, we can continue segmenting until there is complete purity in each leaf, even if the leaf contains only a single individual. Then we can go back pruning the tree by deleting the branches that are not really needed. A more in-depth discussion of this algorithm and these pruning rules will be done in the **??** chapter.

So far, the target variable is categorical with two categories: good and bad payers. This tree-based sorting method also works if:

- The target variable has more than two classes;
- The target variable is numeric, rather than classes (number of days overdue, for example). In this case, we just need a proper definition of what the impurity is at the nodes of the tree. it is common for the impurity to be measured through the standard deviation within the nodes.

5.3 Classification trees in R

In R, we recommend using the `rpart` package. The main command to create the tree is:

```
rpart(formula, data= ) where
```

where `formula` has the following format: `outcome predictor1 + predictor2 + predictor3 + etc` and `data =` specifies the data frame where the variables are. If we use the formula `outcome .`, with a dot in place of a list summing the variable names, we will be telling the `rpart` command to use all the variables available in the dataset except the outcome variable with the classes of individuals. The dataframe can be ignored if the variables exist as separate objects.

```
# Classification Tree with rpart
library(rpart)
fit <- rpart(credit ~ income + age)
```

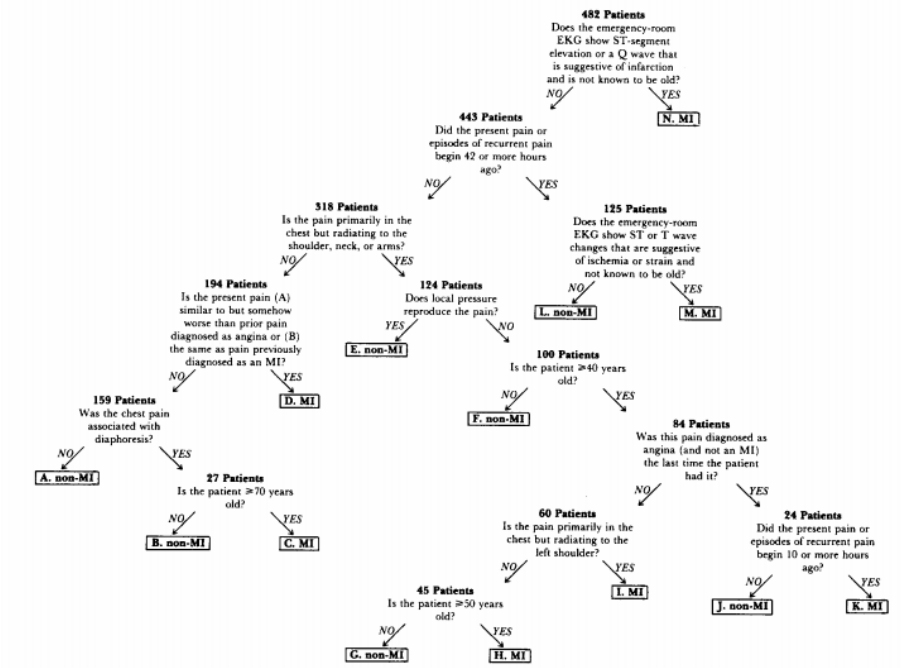


Figure 5.6: Classification tree to diagnose patients suffering from myocardial infarction among those seeking the emergency room with acute chest pain. Figure extracted from[19]

```
plot(fit)
text(fit, cex=0.7)
summary(fit) # information about the tree splits
```

COMPLETE WITH MORE SUBSTANTIAL EXAMPLE

5.4 Some examples of classification trees

This section illustrates some real examples of classification trees. *Rewrite completely. Currently, it only has copy-paste of abstracts of medical articles.*

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, [19] 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.

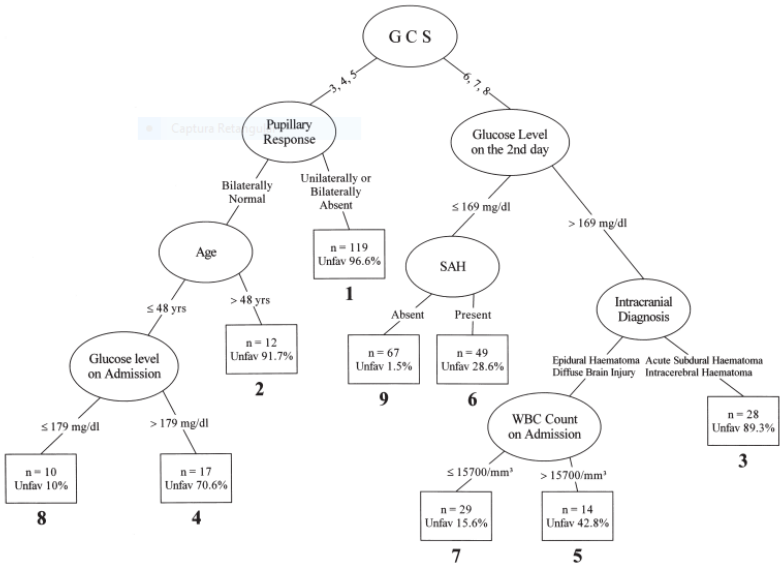


Figure 5.7: Prediction tree based on 345 patients with severe head injury. Ovals denote intermediate subgroups subject to further splitting; squares denote terminal prognostic subgroups. The numbers below the squares represent the prognostic rank of each subgroup based on the proportion of unfavorable outcomes. GCS, Glasgow Coma Scale; SAH, subarachnoid hemorrhage; WBC, white blood cells. Figura extraída de [30].

The goal of this study [30] was to develop a simple, easy to use a model 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 [26], 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.

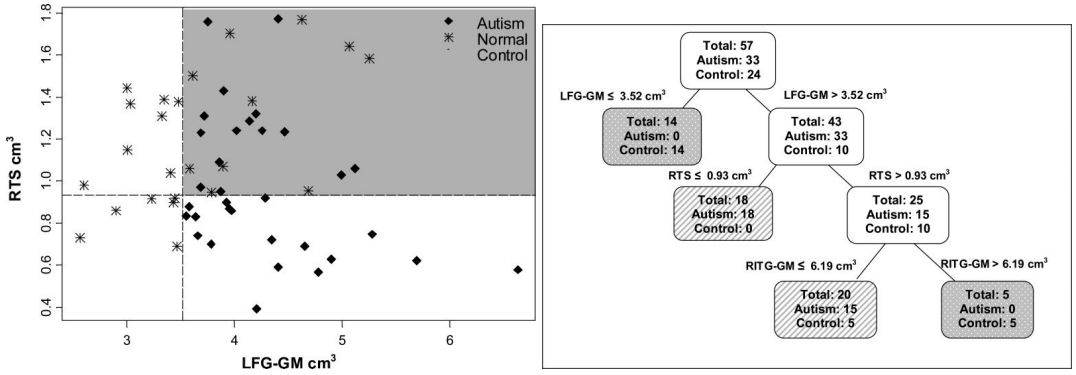


Figure 5.8: **Left**: The two dimensional classification tree shows how the actual observations are separated based on the first two splits of the regression tree. The left fusiform gyrus gray matter (LFG-GM) is on the x-axis. This represents the first split formed by the tree and the 14 red stars on 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.

5.5 Naive Bayes

Naïve Bayes é uma técnica para classificar objetos. Por exemplo, ela serve para classificar automaticamente os e-mails como spam ou não-spam. Spams são e-mails não solicitados contendo propaganda, pronografia, ou com esquemas mal-intencionados (tais como pirâmides, ou você foi sorteado e basta enviar uma taxa). Geralmente, estes spams são geralmente enviados para um imenso número de pessoas. O objetivo é filtrar automaticamente estes e-mails de spam antes que eles cheguem à caixa de entrada de e-mail do usuário. Naive Bayes é uma técnica que aprende como construir um algoritmo para isto a partir de dados estatísticos.

5.5.1 Os dados: rótulos

Como classificar automaticamente com Naïve Bayes? Nós coletamos uma grande amostra de e-mails. Estes e-mails são cuidadosamente rotulados num processo manual (e portanto trabalhoso, demorado e caro). Cada e-mail ganha um desses dois rótulos ou duas classes: *spam* ou *não-spam*. Os emails então ficam particionados em dois sub-grupos: n spams e m não-spams.

5.5.2 Os dados: os atributos dos e-mails

Os atributos são características dos e-mails que podem indicar de alguma forma a classe (ou rótulo) do e-mail (spam ou não-spam). O algoritmo Naïve Bayes usa estes atributos para criar um escore, uma nota, que indica o grau de suspeita de que o e-mail seja um spam. Esta nota é criada automaticamente e portanto não deve envolver nenhum tratamento humano prévio do e-mail. São características extraídas automaticamente dos e-mails a medida que chegam aos servidores.

A ideia fundamental do algoritmo Naïve Bayes é usar os dados estatísticos da amostra para aprender a classificar como *spam* ou *não-spam* os futuros e-mails que chegarão ao servidor. Como queremos classificar bem estes futuros e-mails, nós procuramos usar os rótulos manualmente contruídos para avaliar a qualidade das possíveis regras de classificação e escolher a que parece ser a melhor delas.

spam	num_char	line_breaks	format	number	ratioti	%obs
no	21,705	551	html	small	2.45	52.30
no	7,011	183	html	big	0.00	12.30
yes	631	28	text	none	0.00	0.00
⋮	⋮	⋮	⋮	⋮	⋮	
no	15,829	242	html	small	34.71	0.03

variable	**description**
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
inherit	The number of times the word "inherit" (or "inheritance" and similar) appeared.
viagra	The number of times the word "viagra" appeared in the email.

Como combinar estas atributos para criar a nota ou escore?
dataset email de
https://rdrr.io/cran/openintro/man/email.html

6. Discrete Random Variables

6.1 Random variables: formalism

Probability is a math subject. Establish a probability space $(\Omega, \mathscr{A}, \mathbb{P})$ and we do mathematical probability calculations. Statistics, data mining and machine learning are subjects that deal with data. In the simplest and most usual case, we have a table full of numbers (or labels for categories such as *Male* and *Female*). In this table, rows are items, columns are attributes measured on items. How to link these two subjects? The binding is provided by the concept of *random variable*.

The probability space $(\Omega, \mathscr{A}, \mathbb{P})$ is the mathematical basis of probability. The probability space must assign probability to every event (or subset) $A \subset \Omega$. If Ω is too complicated, we may only be interested in *a few* specific aspects of the random experiment. *Random variables* (r.v.) is the tool to reduce the probability space $(\Omega, \mathscr{A}, \mathbb{P})$ to the minimum necessary in practice. Random variables are numerical characteristics of the random phenomenon with associated probabilities.

Definition 6.1.1 — Random Variables. Formally, a random variable is a (measurable) mathematical function X from Ω to $\mathbb{R}$. That is,

$$\begin{array}{rcl} X : \Omega & \to & \mathbb{R} \\ \omega & \to & X(\omega) \end{array}$$

Thus, X is a function that associates the real value $X(\omega)$ with the result ω of the experiment. Each result ω has a value of $X(\omega)$. Of course, different ω can have the same $X(\omega)$ value. We commented on the restriction that a random variable be a function measurable in the remark 6.1. From now on, we will abbreviate the expression *random variable* by r.v.

■ **Example 6.1** Let's go back to this basic, recurring example: the repeated flip n times of a dishonest coin that is likely to come up heads equal to $\theta \in (0,1)$. The sample space is composed of all n-tuples with C or $\tilde{C}$ in each position. That is, $\Omega = \{\omega = (s_1, s_2, \ldots, s_n)\ ;\ s_i = C \text{ or } \tilde{C} \text{ for } i = 1, \ldots, n\}$. Let's define r.v.'s that will associate a real number with each $\omega \in \Omega$. Then consider the

following r.v.'s:

$$\begin{array}{rcl} X\colon \Omega & \to & \mathbb{R} \\ \omega & \to & X(\omega) = \text{ no. of heads in the } n \text{ releases} \end{array}$$

or the proportion of heads in the n coin flips, which means that we will have a value in the set $\{0/n, 1/n, \ldots, (n-1)/n, 1\}$:

$$\begin{array}{rcl} Y\colon \Omega & \to & \mathbb{R} \\ \omega & \to & Y(\omega) = \text{proportion of heads in } n \text{ flips} \end{array}$$

The next random variable records the order of toss associated with the last appearance of a head in the sequence (if there are no heads, we assign the value 0):

$$\begin{array}{rcl} Z\colon \Omega & \to & \mathbb{R} \\ \omega & \to & Z(\omega) = \max\{\{i \text{ such that } s_i = C \text{ for } i = 1, \ldots, n\} \cup \{0\}\} \end{array}$$

■

Unlike traditional notation that uses f, g, h to denote math functions, we use uppercase letters from the end of the alphabet, such as X, Y, Z or W to denote the random variables. This is a tradition from which we cannot escape as this is the notation used all over the world. At first this is confusing and, for some, irritating. Then we get used to it.

■ **Example 6.2** Consider the example of Figure 3.13, where the sample space Ω is the set formed by all continuous functions in the period of $[0,24]$ hours. That is, the element $\omega \in \Omega$ is a continuous function f with domain $[0,24]$. Events are subsets of curves of this set Ω with infinite curves f.

We can define a series of random variables of potential interest in this set. To make the situation of this example more explicit, let us now denote the element *omega* by f, the continuous function that is the result of the experiment of observing the temperature for 24 hours. We then have, for example, the maximum temperature throughout the day.

$$\begin{array}{rcl} X\colon \Omega & \to & \mathbb{R} \\ f & \to & X(f) = \max\limits_{x \in [0.24]} f(x) \end{array}$$

the temperature at noon

$$\begin{array}{rcl} Y\colon \Omega & \to & \mathbb{R} \\ f & \to & Y(f) = f(12) \end{array}$$

the average temperature throughout the day

$$\begin{array}{rcl} Z\colon \Omega & \to & \mathbb{R} \\ f & \to & Z(f) = \dfrac{1}{24} \displaystyle\int_0^{24} f(x)\, dx \end{array}$$

■

R A random variable is almost any math function X which goes from Ω to $\mathbb{R}$. The only restriction is that the function must be *measurable*, a very technical condition. In practice, this condition of measurability of the function can be ignored and we can think of a random variable as any function that has some relevance for data analysis. Every "practical" function is measurable. Every function involving a finite number or enumerable infinity of operations involving logs, exponentials, polynomials, trigonometric functions, ladder functions, they are all measurable. it's hard to think of a practically useful function that can't be written that way.

6.2 Random Variables and Data Tables

Remember the statistical data table. In the rows, we have the items, individuals, cases, instances or examples (such as different cancer patients in a hospital or different customers in a bank). In the columns, we have characteristics or attributes of the items. For example, we might have columns representing gender, age and cancer stage, or average current account balance, time as an account holder. Informally, *random variables* (r.v.). are the *mathematical or probabilistic* representations of these attribute columns in the statistical data table.

How is the connection between the data table and the probabilistic model? The probability space is formed by the trio $(\Omega, \mathscr{A}, \mathbb{P})$. An example of a data table is in table 6.2. We say that Ω is the set of all emails already received and to be received. Note that Ω is a set of undefined and probably infinite size. The data array contains only a *sample* of elements from Ω. Each *row* in the table corresponds to a distinct element of Ω. Each column represents different characteristics or measurements about the emails. In general, we assume that the different emails in the sample (the different rows in the table) represent events that are independent of each other. That is, an email with certain characteristics does not influence the characteristics of other emails. In the same row of the table, the entries measure different characteristics of the *same* ω (same e-mail). Thus, the elements *within the same table row* are not usually independent. On the contrary, because multiple measurements try to capture the same thing (the spam/non-spam nature of the email), they tend to be associated or correlated. We will see the precise definition of correlation later, but suffice it to say that when the number of characters is low we have an indication that the variable `spam` must have the value `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 — Informal view of a r.v.. Instead of its formal definition, every r.v. can be thought of simply as a combination of two components:

- a set of possible values on the real line;
- probabilities associated with these possible values.

Possible values are associated with the specific meaning of random variables. The probabilities come from a model $(\Omega, \mathscr{A}, \mathbb{P})$ which often *does not* need to be explicitly presented. This makes life a lot easier.

6.3 Types of random variables

We have three basic types of *statistical data* in statistical data tables:

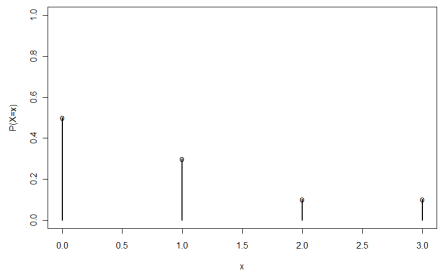


Figure 6.1: Function $p(x_i)$ where x_i is one of the possible values of X. Also called probability mass function. X has possible values $\{0,1,2,3\}$ with probabilities $\{0.5,0.3,0.1,0.1\}$, respectively.

- categorical or non-numerical data, which can be nominal (such as gender or religion) or ordinal (for example, the answer to a question such as "Do you trust members of Congress very, little, or not at all?")
- discrete numeric data: number of children, number of requests in the last two hours.
- continuous numeric data: current account balance, temperature, inflation index.

These data are represented by two types of random variables:

- **Discrete r.v.:** For categorical or discrete numeric data.
- **Continuous r.v.:** For continuous numeric data.

6.4 Discrete Random Variables

Definition 6.4.1 — Discrete R.V.. Discrete random variables serve to model data table columns that have discrete numeric or categorical values. We can think of an r.v. discrete as being composed of two enumerable lists.

- A list of possible values for the r.v.: $\{x_1, x_2, \ldots\}$.
- A list with the probability associated with each of these values: $\{p(x_1), p(x_2), \ldots\}$.

The two lists together define what we call the *probability distribution* of the r.v. X.

Definition 6.4.2 In general, the list of possible values in the above definition only has the values x where $p(x)$ is strictly greater than zero. That is, only enter values where $p(x) > 0$. The set of points x such that $p(x) > 0$ is called the *support set* of the distribution.

We can represent the two lists in a table:

Possible values	x_1	x_2	x_3	$\ldots$
Probab assoc	$p(x_1)$	$p(x_2)$	$p(x_3)$	$\ldots$

The list of probabilities must have values ≥ 0 and they must add up to 1. Together these two tables define a function $x_i \to p(x_i)$ from the set of possible values x_i for the probabilities $p(x_i)$. This function is called *mass probability function* of r.v. X.

The best way to visualize a discrete random variable is with a graph of the associated probabilities. Figure **??** shows the two lists of the informal definition 6.1 in the case of an r.v. discrete X. The ordered list of possible values is $\{0,1,2,3\}$ and the associated probabilities are in the ordered list $\{0.5,0.3,0.1,0.1\}$, respectively. The horizontal axis contains the possible values $\{0,1,2,3\}$ and the vertical axis shows the probabilities $p(x_i)$ where x_i is one of the possible values of X.

6.4.1 Discrete R.V.: examples

■ **Example 6.3 — Binary r.v..** A column of the data table indicates the sex of an individual ω chosen from a population. Arbitrarily, we assign the value 0 to MASC and 1 to FEM. That is, $X(\omega) = 0$ if ω is male and $X(\omega) = 1$ if ω is female. For each individual ω we only look at its sex, represented by $X(\omega) \in \{0.1\}$. To finish the specification of this r.v. discrete, we need to specify the associated probability list. Say, $p(0) = 0.35$ and $p(1) = 1 - 0.35 = 0.65$. ■

■ **Example 6.4 — Gas pump monitoring.** At a gas station, the use of its 4 vehicle fuel pumps is monitored every 5 minutes during peak hours. Every 5 minutes, the number of pumps in use is recorded. The items or instances are the different instants of time. The data are discrete numeric and, at each instant, can be 0, 1, 2, 3 or 4. Let ω be one of the instants of time. $X(\omega)$ is the number of pumps in use. You must also specify the probabilities of each possible value for X. For example, the table **??** gives a possible specification for the probabilities.

Possible values	0	1	2	3	4
Associated Probab	$p(0) = 0.32$	$p(1) = 0.42$	$p(2) = 0.21$	$p(3) = 0.04$	$p(4) = 0.01$

■

■ **Example 6.5 — Number of followers on a social network.** In a social network, choose n users-vertices at random and count the number of incident edges of each one of them (user's followers). The items or instances are the different users. The data is discrete numeric and can be $0, 1, 2, 3, \ldots$ without a natural upper bound. Let ω be one of the users and $X(\omega)$ be your number of followers. $X(\omega) \in \{0, 1, 2, 3, \ldots\} = \mathbb{N}$. Specifying the probabilities (without explaining where we got this from):

Possible values k	0	1	2	3	...	223	...
Associated probab $p(k)$	0.001	0.002	0.002	0.04	...	0.002	...

The (infinite) list of probabilities must have values ≥ 0 and they must add up to 1. That is, $1 = \sum_{k=0}^{\infty} p(k)$. ■

■ **Example 6.6 — Response in sampling survey.** A sample of individuals (the instances) is asked what their religion is: Catholic, Protestant, non-religious, other Christian religions, Spiritist, others. There are six possible categories for each answer, clearly non-numerical and unordered. Let's represent this column of data with a random variable X. Since X is a function of Ω to $\mathbb{R}$, we will arbitrarily assign a number to each category of the response.

Let $X(\omega)$ be a random variable that, for each individual ω in the population, assigns a number as follows:

$$X(\omega) = \begin{cases} 1, & \text{if } \omega \text{ is Catholic} \\ 2, & \text{if } \omega \text{ is Protestant} \\ 3, & \text{if } \omega \text{ has no religion} \\ 4, & \text{if } \omega \text{ is from other Christian religions} \\ 5, & \text{if } \omega \text{ is spiritist} \\ 6, & \text{if } \omega \text{ is of some other religion} \end{cases}$$

The association between the categories and the corresponding numbers is completely arbitrary. Any

other association would be valid. For example, we could have defined:

$$X(\omega) = \begin{cases} -2, & \text{if } \omega \text{ is Catholic} \\ -1, & \text{if } \omega \text{ is Protestant} \\ 0, & \text{if } \omega \text{ has no religion} \\ 1, & \text{if } \omega \text{ is from other Christian religions} \\ 5, & \text{if } \omega \text{ is spiritist} \\ 999, & \text{if } \omega \text{ is of some other religion} \end{cases}$$

In practice, with these non-numeric attributes, the values of the random variable will only be used as a (numeric) label for the category.

Let's go back to the previous specification, where $X(\omega) \in \{1,\ldots,6\}$. To complete the specification of the r.v., we also need to state the probabilities associated with each category of religion (or every possible value of the r.v.). For example, using IBGE data, in the 1980s, when choosing an individual at random from the Brazilian population, we have the following probabilities:

Possible values 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

■

6.4.2 The σ-algebra and the probability function

Assigning probability to each possible value of an r.v. X is a consequence of the probabilities defined in the probability space $(\Omega, \mathscr{A}, \mathbb{P})$. For example, flip a coin 6 times, with $C =$ heads and $\tilde{C} =$ tails. $\Omega = \{CCCCCC, \tilde{C}CCCCC, C\tilde{C}CCCC, \ldots, \tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}\}$. Ω has 36 elements and $\mathbb{P}(\omega) = 1/36$. If we are not interested in the order in which the results appear, but only in the total number of heads, we can only focus on a reduced version of the probability space. We define $X(\omega)$ to be the number of C's in ω. Formally, we have

$$\begin{aligned} X : \Omega &\to \mathbb{R} \\ \omega &\to X(\omega) = \text{ number of } C\text{'s in } \omega \end{aligned}$$

So $X(\omega) \in \{0,1,\ldots,6\} \subset \mathbb{R}$. These are the possible values of the r.v. X.

Each of these possible values has a probability that is induced by the original probability space $(\Omega, \mathscr{A}, \mathbb{P})$. A proposition about the value of an r.v. X in $\mathbb{R}$ determines an event A in Ω. For example, the proposition $[X = 6]$ is equivalent to the event that the 6 tosses had 6 heads. The proposition $[X \geq 5]$ is equivalent to the event that the 6 tosses had at least 5 heads. Some examples of propositions about the value of r.v. X and the equivalent events are as follows:

$$[X = 6] = \{\omega \in \Omega \,:\, X(\omega) = 6\} = \{\omega = (CCCCCC)\}$$

Remember that the ":" symbol should be read as "such that". That is, the set $\{\omega \in \Omega \,:\, X(\omega) = 6\}$ must be read as $\{\omega \in \Omega$ such that $X(\omega) = 6\}$. Other examples follow:

$$[X = 5] = \{\omega \in \Omega \,:\, X(\omega) = 5\} = \{(\tilde{C}CCCCC), (C\tilde{C}CCCC), \ldots, (CCCCC\tilde{C})\}$$

$$[X = 1] = \{\omega \in \Omega \,:\, X(\omega) = 1\} = \{(C\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}), (\tilde{C}C\tilde{C}\tilde{C}\tilde{C}\tilde{C}), \ldots, (\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}C)\}$$

$$[X = 0] = \{\omega \in \Omega \,:\, X(\omega) = 0\} = \{\omega = (\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C})\}$$

More examples and notation:

$$[X \geq 5] = \{\omega \in \Omega \,:\, X(\omega) \geq 5\} = \{(CCCCCC), (\tilde{C}CCCCC), (C\tilde{C}CCCC), \ldots, (CCCCC\tilde{C})\}$$

or else

$$[X \leq 1] = \{\omega \in \Omega \,:\, X(\omega) \leq 1\} = \{(\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}), (C\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}), (\tilde{C}C\tilde{C}\tilde{C}\tilde{C}\tilde{C}), \ldots, (\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}C)\}$$

Being subsets, events can be handled with the usual operations of union, intersection and complement. In fact, compound propositions about the value of X are equivalent to an event resulting from operations on simpler events. For example:

$$\begin{aligned}[X \leq 5 \text{ and } X > 4] &= \{\omega \in \Omega \,:\, X(\omega) \leq 5\} \cap \{\omega \in \Omega \,:\, X(\omega) > 4\} \\ &= \{(\tilde{C}CCCCC), (C\tilde{C}CCCC), \ldots, (CCCCC\tilde{C})\}\end{aligned}$$

Notation 6.1. *For an r.v. X and a real number x, the notation $[X \leq x]$ means the event $\{\omega \in \Omega \,:\, X(\omega) \leq x\}$. The condition that X must be a measurable function in the definition 6.1.1 is the guarantee that this event actually belongs to σ-algebra $\mathscr{A}$ for all $x \in \mathbb{R}$.*

Let a and b be two arbitrary real numbers and consider the interval $(a,b]$. The notation $[a < X \leq b]$ or $[X \in (a,b]]$ is equivalent to the event $\{\omega \in \Omega;\, X(\omega) \in (a,b]\}$. Other ranges such as (a,b), $[a,b)$, $[a,b]$ or (a,b) have analogous notations, note that $[X \leq x] = [X \in [x,\infty)]$.

Since $[X \in (a,b]]$ is an event, and therefore a subset of Ω, we can handle many such events with set operations. As in the coin toss example, we might be interested on the probability that X does not belong to an interval $(a,b]$. We have

$$[X \notin (a,b]] = \{\omega \in \Omega \,:\, X(\omega) \notin (a,b]\} = \{\omega \in \Omega \,:\, X(\omega) \in (a,b]\}^c = [X \in (a,b]]^c \,.$$

■ **Example 6.7** The random variable K counts the number of years completed by an individual in a certain population at the time of his death. So the possible values for K are in the set $\{0,1,2,\ldots,95,96,\ldots\}$ without an upper bound. If we are interested in finding the probability that the individual lives at least 60 years but dies before reaching age 64, then we are interested in the event

$$[X \geq 60 \cap X < 64] = [X = 60] \cup [X = 61] \cup [X = 62] \cup [X = 63]$$

■

Interest may be focused on obtaining the probability that an r.v. X is in the range $(1,2)$ or is greater than 3. That is, we are interested in calculating the probability of an event A given by

$$\begin{aligned}[X \in (1,2) \text{ or } X > 3] &= \{\omega \in \Omega;\, X(\omega) \in (1,2) \text{ or } X(\omega) > 3\} \\ &= \{\omega \in \Omega;\, X(\omega) \in (1,2)\} \cup \{\omega \in \Omega;\, X(\omega) > 3\} \\ &= [X \in (1,2)] \cup [X(\omega) > 3]\end{aligned}$$

Note that the events $[X \in (1,2)]$ and $[X(\omega) > 3]$ are disjoint because we cannot have a value ω such that, at the same time, $X(\omega) \in (1,2)$ and $X(\omega) > 3$.

Notation 6.2. *In general, we are interested in calculating probabilities of events defined by values of r.v.'s. Let $B \subseteq \mathbb{R}$ be a subset of the real line. We denote*

$$\mathbb{P}(X \in B) = \mathbb{P}(\{\omega \in \Omega \,:\, X(\omega) \in B\})$$

For example, for the event $[X \leq 2]$ we have

$$\mathbb{P}([X \leq 2]) = \mathbb{P}(\{\omega \in \Omega \,:\, X(\omega) \leq 2\}) = \mathbb{P}(\{\omega \in \Omega \,:\, X(\omega) \in [2,\infty)\}) \,.$$

Or $\mathbb{P}([X = 2]) = \mathbb{P}(\{\omega \in \Omega \,:\, X(\omega) = 2\})$. In general, we just write $\mathbb{P}(X \leq 2)$ and $\mathbb{P}(X = 2)$ instead of the more heavily loaded notation $\mathbb{P}([X \leq 2])$ and $\mathbb{P}([X = 2])$.

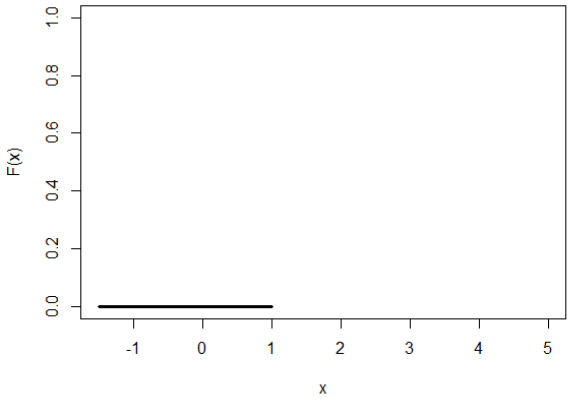


Figure 6.2: $\mathbb{F}(x) = \mathbb{P}(X \leq x)$ for $x < 1$.

6.4.3 Cumulative distribution function

Definition 6.4.3 The *accumulated probability distribution function* of r.v. X is the function *mathematics* $\mathbb{F}(x)$ defined for all $x \in \mathbb{R}$ and given by

$$\begin{array}{rcl} \mathbb{F}\colon \mathbb{R} & \to & [0,1] \\ x & \to & \mathbb{F}(x) = \mathbb{P}(X \leq x) \end{array}$$

This function is simple and does not have *any* additional information beyond what is contained in the probability list of the informal definition 6.4.1. However, despite not bringing additional information to these two lists, it is very important both in the theory and in the practice of data analysis appearing in statistical tests (such as the Kolmogorov test) and as a tool to obtain proofs of certain theorems. Therefore, we will study it carefully. Let's start by calculating $\mathbb{F}(x)$ in a particular case.

■ **Example 6.8 — Calculation of $\mathbb{F}(x)$ for a discrete r.v..** Suppose we have an r.v. discrete random X with possible values $\{0,1,2,3\}$ and associated probabilities $p(k) = \mathbb{P}(X = k)$ given by

Possible values k	1	2	3	4
Associated Probab $p(k)$	0.1	0.4	0.2	0.3

Let's calculate $\mathbb{F}(x) = \mathbb{P}(X \leq x)$ for some of the values of x. Consider, for example, $x = -1$. How much is $\mathbb{F}(-1)$, the value of the function $\mathbb{F}$ evaluated at the point -1. By the definition of $\mathbb{F}$ we should get

$$\mathbb{F}(-1) = \mathbb{P}(X \leq -1) = \mathbb{P}(\{\omega \; in \Omega \; : \; X(\omega) \leq -1\}) \; .$$

This probability is zero because, for all ω, we have $X(\omega) \in \{1,2,3,4\}$ and therefore $X(\omega)$ is always a value greater than -1. That is, there is no $\omega \in \Omega$ such that $X(\omega) \leq -1$. So, $\{\omega \in \Omega \; : \; X(\omega) \leq -1\} = \emptyset$ and therefore $\mathbb{F}(-1) = \mathbb{P}(X \leq -1) = \mathbb{P}(\emptyset) = 0$. By the same argument, for any $x < 1$, we get $\mathbb{F}(x) = \mathbb{P}(X \leq x) = 0$.

Exactly at the point $x = 1$, the function $\mathbb{F}(x)$ makes a jump. In fact, the event $[X \leq 1]$ is identical to the event $[X = 1]$ since there is no ω such that $X(\omega) < 1$:

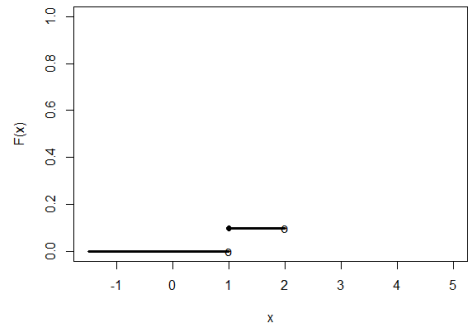


Figure 6.3: $\mathbb{F}(x) = \mathbb{P}(X \leq x)$ for $x < 2$.

$$\begin{aligned}
\{\omega \in \Omega \,:\, X(\omega) \leq 1\} &= \{\omega \in \Omega \,:\, X(\omega) < 1\} \cup \{\omega \in \Omega \,:\, X(\omega) = 1\} \\
&= \emptyset \cup \{\omega \in \Omega \,:\, X(\omega) = 1\} \\
&= \{\omega \in \Omega \,:\, X(\omega) = 1\} \\
&= [X = 1]
\end{aligned}$$

In this way, we have

$$\mathbb{F}(1) = \mathbb{P}(X \leq 1) = \mathbb{P}(X < 1) + \mathbb{P}(X = 1) = 0 + p(1) = 0.1$$

The $\mathbb{F}(x)$ function jumps from 0 to $x < 1$ to 0.1 at the point $x = 1$. For $x = 1.5$ we have

$$\begin{aligned}
\mathbb{F}(1.5) &= \mathbb{P}(X \leq 1.5) && (6.1)\\
&= \mathbb{P}([X < 1] \cup [X = 1] \cup [1 < X \leq 1.5]) && (6.2)\\
&= \mathbb{P}(\emptyset \cup [X = 1] \cup \emptyset) && (6.3)\\
&= \mathbb{P}(X = 1) = 0.1 && (6.4)
\end{aligned}$$

By the same argument, for any x such that $1 < x < 2$ we have $\mathbb{F}(x) = \mathbb{P}(X = 1) = 0.1$. Exactly at the point $x = 2$, the function $\mathbb{F}(x)$ makes one more jump. The event $[X \leq 2]$ is identical to the union of two disjoint events

$$[X = 1 \text{ or } X = 2] = [X = 1] \cup [X = 2]$$

They are disjoint because, by the definition of a mathematical function, we cannot have an element $\omega \in \Omega$ such that $X(\omega) = 1$ and, at the same time, $X(\omega) = 2$. So we have

$$\mathbb{F}(2) = \mathbb{P}(X \leq 2) = \mathbb{P}([X = 1] \cup [X = 2]) = \mathbb{P}(X = 1) + \mathbb{P}(X = 2) = p(1) + p(2) = 0.1 + 0.4 = 0.5$$

See that the jump height is equal to $p(2)$, the probability $p(2) = \mathbb{P}(X = 2)$. For any x between 2 and 3, such as $x = 2.72$, we have

$$\mathbb{F}(x) = \mathbb{P}(X \leq x) = \mathbb{P}(X \leq 2) + \mathbb{P}(2 < X \leq x) = \mathit{mathbbP}(X = 1) + \mathbb{P}(X = 2) + 0 = p(1) + p(2) = 0.5$$

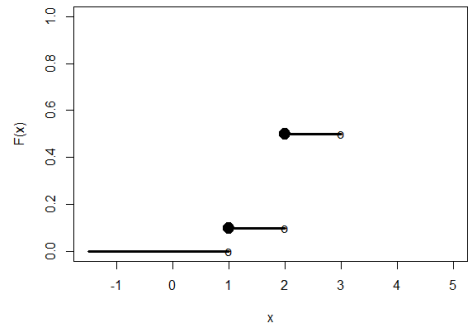


Figure 6.4: $\mathbb{F}(x) = \mathbb{P}(X \leq x)$ for $x < 3$.

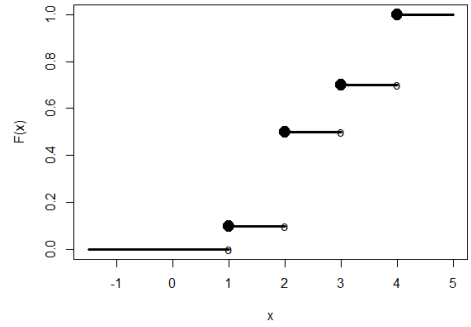


Figure 6.5: $\mathbb{F}(x) = \mathbb{P}(X \leq x)$.

Continuing in this way, we see that $\mathbb{F}(x)$ will jump in $x = 3$ and e $x = 4$. The jump height at $x = k$ is equal to the probability $p(k) = \mathbb{P}(X = k)$. When we choose a value x greater than all possible points of X we will have $\mathbb{F}(x) = 1$. For example, if $x = 4.5$, we clearly have

$$\mathbb{F}(4.5) = \mathbb{P}(X \leq 4.5) = 1$$

because, of course, we will have $X \leq 4.5$ since the largest possible value of X is 4. The full graph of $\mathbb{F}(x)$ is shown below.

■

General case of $\mathbb{F}(x)$

Suppose we have an r.v. discrete random X with possible values x_i and associated probabilities $p(x_i) = \mathbb{P}(X = x_i)$ given by

Possible values	x_1	x_2	x_3	...
Associated probab	$p(x_1)$	$p(x_2)$	$p(x_3)$	...

As the cumulative probability distribution function is defined as:

$$\mathbb{F}(x) = \mathbb{P}(X \leq x) = \sum_{x_i \leq x} p(x_i)$$

then $\mathbb{F}(x)$ is the cumulative value (the sum) of the probabilities $p(x_i)$ of the possible points x_i that are less than or equal to x.

6.4.4 Expected value $\mathbb{E}(X)$

Definition 6.4.4 — Mathematical expectation $\mathbb{E}(X)$ of a r.v. X. The *expected value* $\mathbb{E}(X)$ of a discrete r.v. X is the sum of their possible values weighted by their respective probabilities. Suppose we have a discrete r.v. X with possible values x_i and associated probabilities $p(x_i) = \mathbb{P}(X = x_i)$ given by

Possible values	x_1	x_2	x_3	...
Associated probab	$p(x_1)$	$p(x_2)$	$p(x_3)$	...

So, by definition, we have

$$\mathbb{E}(X) = \sum_i x_i p(x_i)$$

The expected value $\mathbb{E}(X)$ is also called the *mathematical hope* of r.v. X.

The expectation $\mathbb{E}(X)$ is a theoretical, mathematical value associated with the probability distribution of the v.a X. No statistical data is needed to calculate $\mathbb{E}(X)$. The two lists, the one of possible values and the one of associated probabilities, are enough.

■ **Example 6.9** The probability distribution of an r.v. X is specified by the two lists below:

Possible values k	1	2	3	4
Associated probab $p(k)$	0.1	0.4	0.2	0.3

The expected value of X is equal to

$$\mathbb{E}(X) = \sum_{k=1}^{4} k\mathbb{P}(X = k) = 1 \times 0.1 + 2 \times 0.4 + 3 \times 0.2 + 4\, times 0.3 = 2.7$$

■

Note from the example above that the most likely value of r.v. X is 2, with $\mathbb{P}(X = 2) = 0.4$. So $\mathbb{E}(X) = 2.7$ is not the most probable value. $\mathbb{E}(X)$ also does not match any of the possible values of r.v. X.

6.4.5 Interpreting $\mathbb{E}(X)$

What is the empirical meaning of this number $\mathbb{E}(X)$? How to interpret it in practice since it does not correspond to the most probable value and does not even need to be one of the possible values of the r.v.? Suppose an r.v. discrete X with possible values x_i and associated probabilities $p(x_i) = \mathbb{P}(X = x_i)$. We have a huge sample of N independent instances of X. In this sample, x_i appeared N_i times. We can estimate the probabilities by the relative frequency of occurrence of x_i in the sample:

$$p(x_i) = \mathbb{P}(X = x_i) \approx \frac{N_i}{N}$$

So,

$$\mathbb{E}(X) = \sum_i x_i p(x_i) \approx \sum_i x_i \frac{N_i}{N}$$

Since x_i appeared N_i times in the sample, this is the same as adding up all the N values in the sample and dividing by N:

$$\mathbb{E}(X) \approx \sum_i \frac{x_i N_i}{N} = \frac{x_1 + \ldots x_1 + x_2 + \ldots x_2 + \ldots}{N}$$

where X_1 appears N_1 times, x_2 appears N_2 times, etc. That is, if the sample is large, we should have the theoretical number $\mathbb{E}(X)$ approximately equal to the *arithmetic mean* of the N elements in the sample.

Let's reinforce: $\mathbb{E}(X)$ is a real number, a constant, associated with the two lists, the one of possible values and the one of associated probabilities, which constitute an r.v. $\mathbb{E}(X)$ is not itself an r.v. because it doesn't have two lists, it's just a number. $\mathbb{E}(X)$ is a *theoretical summary* of the distribution of X, or a summary of the two lists. It is approximately equal to the arithmetic mean of the values of a large sample of instances of X.

You may wonder: what if the series of values do not converge? In chapter ??, we will see that, in the usual situations of data analysis, the probability of this happening is zero. This result is called the Strong Law of Large Numbers.

6.5 Main Discrete Distributions

There are infinite probability distributions. Given a set of possible values, any assignment of non-negative numbers that add up to 1 constitutes a probability distribution. However, a few assignments are given special names. These distributions appear frequently in data analysis and are mathematically tractable. We can think of the data analyst tackling a practical problem with a bag of well-known probability distributions. He would like not to invent a new distribution, but rather to use one of those already in his bag. Let's see some of the most popular ones now. They are as follows:

- Bernoulli $\text{Ber}(\theta)$;
- Binomial $\text{Bin}(n, \theta)$;
- Multinomial $\mathscr{M}(n; \theta_1, \ldots, \theta_k)$;
- Geometric $\text{Geo}(\theta)$;
- Zipf and Pareto.

6.5.1 Bernoulli

It is the simplest possible discrete distribution: only two possible outcomes. $X(\omega)$ only takes on two possible values: 0 or 1. We usually say that the value corresponds to a success and the value 0 to a failure. These are just names for the two categories, with no further meaning. For example, we can say the death of an individual in a period corresponds to a success.

We have $X(\omega) \in \{0.1\}$ for all $\omega \in \Omega$. We define two probabilities:

$$p(1) = \mathbb{P}(X = 1) = \mathbb{P}(\omega \in \Omega : X(\omega) = 1)$$

$$p(0) = \mathbb{P}(X = 0) = \mathbb{P}(\omega \in \Omega : X(\omega) = 0)$$

We have $p(0) + p(1) = 1$ which implies that $p(1) = 1 - p(0)$. It is common to write $p(1) = \theta$ and $p(0) = 1 - \theta$. Another common notation is $p(1) = p$ and $p(0) = q$.

Typically, we have $0 < \theta < 1$. If $\theta = 0$, the chance of a success is zero and we will only observe failures. Likewise, the case $\theta = 1$ is an extreme case, where success occurs with absolute certainty.

If $p(1) = \theta$ and $p(0) = 1 - \theta$, we have

$$\mathbb{E}(X) = 1 \times \theta + 0 \times (1 - \theta) = \theta$$

Note that $\mathbb{E}(X) = \theta$ is not equal to any possible value of X, which is just 0 or 1.

If we have a large sample of instances of X, each of which is equal to 0 or 1, we should have the value of $\mathbb{E}(X) = \theta$ approximately equal to the arithmetic mean of the observed values 0 or 1 . But an arithmetic mean of values 0 or 1 is just the proportion of ons in the sample. That is, as obviously expected, we must have

$$\mathbb{E}(X) \approx \hat{\theta} = \frac{1}{N}\sum_i x_i\,.$$

In summary, we have the definition

Definition 6.5.1 — Bernoulli distribution. The r.v. X has a Bernoulli distribution with parameter $\theta \in [0,1]$ if $X(\omega) \in \{0,1\}$ with $\mathbb{P}(X=1) = \ theta$ and $\mathbb{P}(X=0) = 1-\theta$.

Notation 6.3. *We write $X \sim Ber(\theta)$ to mean that the r.v. X has Bernoulli distribution with parameter θ.*

6.5.2 Binomial

The Binomial distribution corresponds to the number of successes in n *independent* repetitions of a binary (Bernoulli) experiment. The probability of success is constant and equal to $\theta \in [0.1]$ in all repetitions. As the r.v. X counts the total number of successes in n repetitions, the list of possible values is formed by $\{0,1,2,\ldots,n\}$. The list of associated probabilities is given by $\{(1-\theta)^n, n\theta(1-\theta), \ldots, \theta^n\}$. The general formula for these probabilities is as follows:

$$\mathbb{P}(X=k) = \frac{n!}{k!(n-k)!}\theta^k(1-\theta)^{n-k}\,.$$

We will explain how to arrive at this formula after presenting the multinomial distribution in the next section.

We have $\mathbb{E}(X) = n\theta$. This result is intuitive. If we have a probability of success on a repetition equal to, for example, $\theta = 0.20$ we should expect 20% of the repetitions to be successful. That is, we should expect the ratio of successes to be equal to 0.20. Thus, the number of successes in n repetitions expected to be observed is equal to $n \times 0.20$. The formal proof requires us to use some algebraic tricks.

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{k=0}^{n} k\binom{n}{k}\theta^k(1-\theta)^{n-k} \\
&= n\theta\sum_{k=0}^{n} k\frac{(n-1)!}{(n-k)!k!}\theta^{k-1}(1-\theta)^{(n-1)-(k-1)} \\
&= n\theta\sum_{k=1}^{n} \frac{(n-1)!}{((n-1)-(k-1))!(k-1)!}\theta^{k-1}(1-\theta)^{(n-1)-(k-1)} \\
&= n\theta\sum_{k=1}^{n} \binom{n-1}{k-1}\theta^{k-1}(1-\theta)^{(n-1)-(k-1)} \\
&= n\theta\sum_{\ell=0}^{n-1} \binom{n-1}{\ell}\theta^{\ell}(1-\theta)^{(n-1)-\ell} \qquad \text{com } \ell = k-1 \\
&= n\theta\sum_{\ell=0}^{m} \binom{m}{\ell}\theta^{\ell}(1-\theta)^{m-\ell} \qquad \text{com } m = n-1 \\
&= n\theta(\theta+(1-\theta))^m \\
&= n\theta
\end{aligned}$$

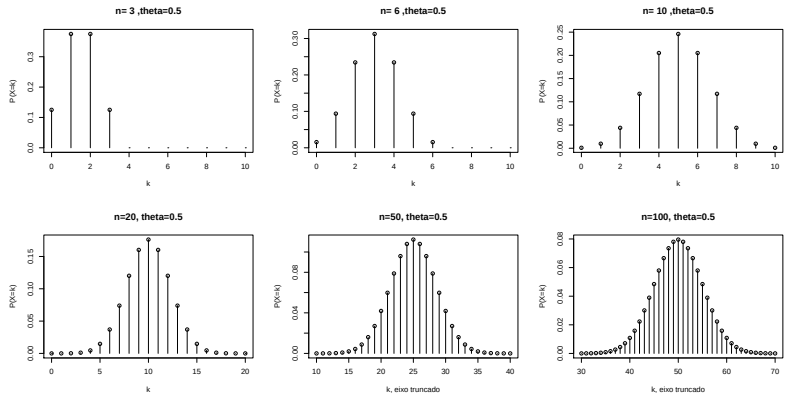
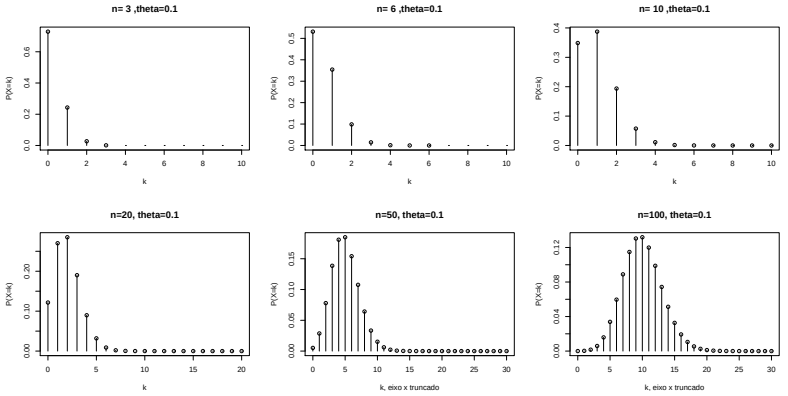


Figure 6.6: $\mathbb{P}(X=k)$ with $\theta = 1/2$ and different values for n for $X \sim \text{Bin}(n,\theta)$.



Definition 6.5.2 — Binomial Distribution. The r.v. X has Binomial distribution with parameters n and $\theta \in [0,1]$ if $X(\omega) \in \{0,1,\ldots,n\}$ with

$$\mathbb{P}(X=k) = \frac{n!}{k!(n-k)!}\theta^k(1-\theta)^{n-k}$$

for $k = 0,1,\ldots,n$.

Notation 6.4. *We write $X \sim Bin(n,\theta)$ to signify that the r.v. X has Binomial distribution with parameters n and θ.*

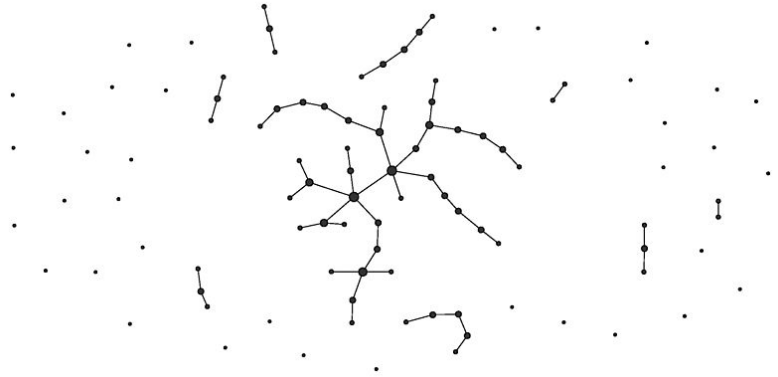
The shape of the binomial distribution depends on θ and n. Figure 6.6 shows the probability function $\mathbb{P}(X=k)$ with $\theta = 1/2$ and different values for n.

Figure 6.5.2 $\theta = 0.1$ with different values for n.

■ **Example 6.10 — Serum or vaccine tests.** This example is taken from [16]. A disease affects cattle in a certain region with an incidence of 25%. It seeks to test the effectiveness of a recently discovered vaccine. We inject the vaccine into n healthy animals. How to evaluate the result?

We imagine a binomial experiment. Let Y be the number healthy after some time after vaccination and after being exposed in the usual way to the disease. For each animal, we will have a success if it remains healthy. If the vaccine is completely innocuous we will have a success in each animal with probability $\theta = 0.75$. Even in the case of an innocuous vaccine, assuming that the results of different animals are independent, we will have $Y \sim \text{Bin}(n,\theta)$.

The probability that k of the n animals is healthy is $\mathbb{P}(Y=k) = n!/(k!(n-k)!)0.75^k 0.25^{n-k}$. If we have used $n = 10$ animals, the chance that they are all healthy is equal to $\mathbb{P}(Y=10) =$



$0.75^{10} = 0.056$. If we have used $n = 12$ animals, the probability that they are all healthy is $\mathbb{P}(Y = 12) = 0.75^{12} = 0.032$, a lower value than the previous one. So, if in a total of 10 or 12 animals, none is contaminated, we will have a strong indication that the serum had some effect. The reason is that, if the vaccine were innocuous (and therefore $\theta = 0.75$), we would hardly observe healthy all the $n = 10$ or $n = 12$ animals. However, this result does not constitute conclusive evidence.

We saw that with $n = 10$, we had $\mathbb{P}(Y = 10) = 0.056$. Without the vaccine, the probability that out of 17 animals, *at most one* of them becomes infected is equal to $\mathbb{P}(Y \leq 1) = \mathbb{P}(Y = 0) + \ mathbbP(Y = 1) = 0.75^{17} + 17 \times 0.75^{16} \times 0.25 = 0.0501$. Therefore, the evidence in favor of the vaccine is *stronger* when there are 1 infected in 17 than when there are 0 in 10! For n = 23, we have $\mathbb{P}(Y \leq 2) = 0.0492$. So 2 or fewer infected in 23 is, again, stronger evidence in favor of the vaccine than 1 in 17 or 0 in 10.

■

■ **Example 6.11 — Binomial in social networks.** The binomial distribution appears prominently in the Erdös-Rényi model for social graphs. Each actor in a social network is a vertex in a graph. Edges connect friends. We have n vertices forming $n(n-1))/2$ pairs of possible undirected edges. For each *pair* of vertices, toss a "coin" with a success probability equal to θ. If heads, connect them by an edge. The probability coin θ is flipped independently for each of the $n(n-1)/2$ vertex pairs.

Fix any vertex and let Y be the number of incident edges. Y counts the number of friends connected to the vertex in question. So $Y \sim \text{Bin}(n-1, \theta)$. See that $\mathbb{E}(Y) = (n-1)\theta \approx n\theta$ if n is large. Figure 6.11 shows a graph generated by the binomial model of Erdós and Rényi with $\theta = 0.01$ and $n = 100$. Note that we expect around $n\theta = 100 \times 0.01 = 1$ friend connected at each vertex. Some vertices have no edges, others have two or three.

Our course is not about social networks and therefore we will not explore this model. Just as a curiosity, we will comment on some of the probabilistic results proved by [14] on random graphs generated by this model assuming that $n \to \infty$. Consider $n\theta \approx \mathbb{E}(Y)$, the expected number of neighbors of any vertex. The type of random graph that will be generated depends on the expected value of the number of friends of a vertex. it is clear that the larger the value of $\mathbb{E}(Y)$, the denser the resulting graph. it is difficult to say something general when the number n of vertices is small. However, when n starts to grow, a probabilistic stability emerges. Given an undirected graph, we can look at its largest connected component, also called the giant component. This is the largest subgraph extracted from the original graph such that any pair of vertices has at least one path connecting them. When n grows, this subgraph can dominate the graph. We have

- If $n\theta > 1$ and if n grows then the graph will have a giant component of the order of n (of the same order of magnitude as the complete graph) and the second largest component will be of order $\ leqO(\log(n))$ (that is, much smaller than the original graph).

- Also, the generated graph will almost certainly not have a connected component greater than $O(\log(n))$.
- If $n\theta > (1+\varepsilon)\log(n)$ (slightly larger than $\log(n)$), then the graph will almost certainly be fully connected.
- On the other hand, if $n\theta < (1-\varepsilon)\log(n)$ then the graph will almost certainly have isolated vertices
- Etc, etc, etc... several nice and non-obvious ones in the article [14].

How to know if a given graph was generated by the model of Erdös and Rényi? An obvious way is to compare the distribution of the number of neighbors actually observed in the real graph with the distribution derived from the model of Erdös -Renyi. We count the proportion of isolated vertices, the proportion of vertices with 1 friend, the proportion of vertices with 2 friends, etc. Then, we compare these frequencies with the probability $\mathbb{P}(Y = k)$ that a vertex has a number Y of friends equal to k. If the relative frequencies and the theoretical probabilities are very different, we will have reason to distrust the Erdös and Rényi model as a generating model for the observed graph. Otherwise, we will have some evidence that the model can be the generator How to measure the distance between relative frequencies and theoretical probabilities, between what we observe and what we expect under the model? We have a generic answer for this: using the chi-square test, to be seen in section ??.

■

6.5.3 Multinomial

The multinomial distribution is a generalization of the binomial distribution. The binomial counts the number of successes in n repetitions of a *binary* experiment. In each repetition we have two categories to classify the result: success or failure. When we have more than two categories in each repetition, we have the multinomial distribution. In the multinomial distribution we also independently repeated an experiment n times. However, in each experiment, there are k possibilities and not just two, as in the binomial. The result of the experiment is the count of how many times each of the k possibilities appeared in the n repetitions.

The canonical example of the multinomial distribution is the roll of a die. Imagine that a die is rolled n times. In each repetition there is a "category": $1,2,3,4,5$ or 6. The probabilities for each category are: $\theta_1, \theta_2, \ldots, \theta_6$. If the die is perfectly balanced, the six probabilities θ_i will all equal $1/6$. If the die is unbalanced, they will not all be the same. They must be numbers between 0 and 1 and we must have $\theta_1 + \ldots + \theta_6 = 1$. Qualquer que seja o dado, balanceado ou não, ao fim dos n lançamentos teremos as contagens:

$$\begin{aligned} N_1 &= \text{no. of releases in cat. 1} \\ N_2 &= \text{no. of releases in cat. two} \\ \vdots &= \vdots \\ N_6 &= \text{no. of releases in cat. 6} \end{aligned}$$

The result is a multinomial random vector with 6 positions counting the random number of occurrences of each category.

Notation 6.5 (Multinomial Distribution)**.**

$$(N_1, N_2, \ldots, N_6) \sim \mathscr{M}(n; \theta_1, \ldots, \theta_6)$$

The binomial distribution is a simple case of the multinomial distribution. Let $X \sim \text{Bin}(n, \theta)$, where X is the number of successes in n repetitions of a binary experiment. Quite redundantly, we

could record the random phenomenon in the form of a vector with the number of successes and the number of failures: $(X, n-X)$. This vector is a multinomial with two categories. In our notation, we would have $(X, n-X) \sim \mathscr{M}(n; \theta, 1-\theta)$.

Returning to the case of the unbalanced die thrown n times, we have:

$$\mathbf{N} = (N_1, N_2, \ldots, N_6) \sim \mathscr{M}(n; \theta_1, \ldots, \theta_6)$$

What is the support of this random vector $\mathbf{N}$? For any sequence of tosses, the result will be a vector $(N_1, \ldots, N_6)$ of integers ≥ 0 with $n_1 + \ldots n_6 = n$. So the number of possible values for $\mathbf{N}$ will be a number finite. Although finite, this number will be quite large unless n is very small.

What are the probabilities associated with the support elements? Let's calculate a particular case before giving the general formula. Using $n = 8$ rolls of the die, let's calculate the probability

$$\mathbb{P}(\mathbf{N} = (2,0,2,1,0,3))$$

That is, we want the chance to roll the die 8 times and end up with face 1 appearing twice, face 2 not appearing, face 3 appearing twice, face 4 once, face 5 zero times and face 6 3 times. There are several 8 toss ω sequences that lead to the above result.

For example, if the 8 successive tosses are $\omega = (3,1,6,6,1,4,6,3)$ we will have

$$\mathbf{N}(\omega) = (N_1(\omega), \ldots, N_6(\omega)) = (2,0,2,1,0,3)$$

This isn't the only sequence producing these counts but we'll focus on it for now. What is the probability $\mathbb{P}(\omega)$ of this sequence of 8 tosses? As the releases are independent we will have:

$$\begin{aligned}
\mathbb{P}(\omega = (3,1,6,6,1,4,6,3)) &= \mathbb{P}(\text{exit 3 on 1st. And come out 1 on the 2nd. And ... come out 3 on the 8th.}) \\
&= \mathbb{P}(\text{exit 3 on 1st.})\,\mathbb{P}(\text{exit 1 on 2nd.})\ldots\mathbb{P}(\text{ exit 3 on 8th. }) \\
&= \theta_3\ \theta_1\ \theta_6\ \theta_6\ \theta_1\ \theta_4\ \theta_6\ \theta_3 \\
&= \theta_1^2\ \theta_2^0\ \theta_3^2\ \theta_4^1\ \theta_5^0\ \theta_6^3
\end{aligned}$$

Generally speaking, if the sequence ω of n releases has

- n_1 face appearances 1
- n_2 face appearances 2
- $\vdots$
- n_6 face appearances 6

we will have

$$\mathbb{P}(\omega) = \theta_1^{n_1}\ \theta_2^{n_2}\ \theta_3^{n_3}\ \theta_4^{n_4}\ \theta_5^{n_5}\ \theta_6^{n_6}$$

Returning to the $n = 8$ rolls of the die, let's calculate the probability $\mathbb{P}(\mathbf{N} = (2,0,2,1,0,3))$. Let A be the event formed by all ω (sequences of $n = 8$ tosses) such that there are 2 1's, 0 2's, 2 3's, 0 4's, and 3 6's. As we calculated before, every ω in this event A will have the same probability given by

$$\mathbb{P}(\omega) = \theta_1^2\ \theta_2^0\ \theta_3^2\ \theta_4^1\ \theta_5^0\ \theta_6^3$$

So,

$$\mathbb{P}(\mathbf{N} = (2,0,2,1,0,3)) = \mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\omega) = C \times \theta_1^2\ \theta_2^0\ \theta_3^2\ \theta_4^1\ \theta_5^0\ \theta_6^3$$

where C is the number of sequences of length 8 where we place an element of $\{1,2,\ldots,6\}$ in each position and where we have exactly 2 1's, 0 2's, ... , 36's.

This number of possibilities is equal to

$$\binom{8}{2,0,2,1,0,3} = \frac{8!}{2!0!2!1!0!3!} = 1680$$

This is the number of distinct permutations of the vector $\omega = (3,1,6,6,1,4,6,3)$.

Definition 6.5.3 — Multinomial distribution. A vector $\mathbf{N} = (N_1, N_2, \ldots, N_6)$ has a multinomial distribution with parameters n and $(\theta_1, \ldots, \theta_k)$ with $\theta_i \geq 0$ and $\sum_i \theta_i = 1$ if the set of possible values are the integers $n_i \geq 0$ with $n_1 + \ldots + n_k = n$ and associated probabilities given by

$$\mathbb{P}(\mathbf{N} = (n_1, n_2, \ldots, n_k)) = \frac{n!}{n_1! n_2! \ldots n_k!} \theta_1^{n_1} \ \theta_2^{n_2} \ldots \theta_k^{n_k}$$

Notation: $\mathbf{N} = (N_1, N_2, \ldots, N_k) \sim \mathscr{M}(n; \theta_1, \ldots, \theta_k)$.

■ **Example 6.12** Suppose we have a sample of $n = 22343$ individuals chosen independently from the Brazilian population and classified into $k = 6$ religious categories. Each individual in the sample is like throwing an unbalanced die with six "faces" represented by the categories of religion: 1: Catholic, 2: Protestant, 3: No Religion, 4: Spiritualist, 5: Other Christian Religions, 6: Other. it is common to assume the random counts of the number of people in each category follow a multinomial distribution with probabilities $(\theta_1, \ldots, \theta_6)$ for the categories. These probabilities are known (approximately) from large surveys conducted by the IBGE:

$$(\theta_1, \ldots, \theta_6) = (0.75, 0.15, 0.07, 0.01, 0.01, 0.01) .$$

Thus, we say that the vector of counts is multinomial:

$$\mathbf{N} = (N_1, N_2, \ldots, N_6) \sim \mathscr{M}(22343; (0.75, 0.15, 0.07, 0.01, 0.01, 0.01))$$

The table below shows the results of the counts from a sample of 22343 individuals.

Categories i	Catholic	Protestant	Non-Religious	Spiritism	Other Crist.	Others
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** Suppose that a sample of $n = 538$ individuals chosen independently from patients with Hodgkins lymphoma (a type of cancer of the lymphatic system) are classified into 12 categories according to their response to a certain treatment and their histological type. The counts are in the table below. We have $4 \times 3 = 12$ categories in total and each individual is like the result of rolling an unbalanced 12-sided die.

	Response			
Histological Type	Positive	Partial	No Response	Total
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

Random counts of the number of individuals in each category follow a multinomial distribution

$$\mathbf{N} = (N_1, N_2, \ldots, N_{12}) \sim \mathscr{M}(538; (\theta_1, \ldots, \theta_{12})$$

As the sample is large, we can obtain estimates, approximate values, for the probabilities θ_i from the proportion of elements in the sample that fell into the i category. is the frequentist principle of estimating the probability by the proportion of times the die rolled showed the "face" i. For example, $\theta_7 = \mathbb{P}(\text{ MC and Partial }) \approx 54/538 \approx 0.1$.

■

6.5.4 Poisson

This distribution got its name (Poisson, pronounced as *pwa-ssawn*) after the French mathematician Siméon-Denis Poisson, who lived between 1781 and 1840. He studied various problems involving probabilities and used this distribution in several of them. A large number of situations in which it appears involve counting of the number of certain occurrences in a certain time interval without a clear limit to the maximum number that could be obtained. Usual examples would be:

- number of collisions in BH traffic during the year.
- number of cars entering UFMG between 7 and 8 am
- number of medical appointments that a customer of a health plan makes during the year

How is an r.v. discrete, we just need to list the possible values and the associated probabilities.

Definition 6.5.4 — Poisson distribution. an r.v. X has a Poisson distribution if the support set is the set of natural numbers $\mathbb{N} = \{0,1,2,\ldots\}$. The probability that $X = k$ depends on a positive constant λ and is equal to

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

where $k = 0,1,2,\ldots$

In the Poisson distribution we have this constant $\lambda > 0$ which influences the calculation of probabilities. The possible values are $0,1,2,\ldots$ and the associated probabilities are given by:

- $\mathbb{P}(Y = 0) = \frac{\lambda^0}{0!} e^{-\lambda} = e^{-\lambda}$
- $\mathbb{P}(Y = 1) = \frac{\lambda^1}{1!} e^{-\lambda} = \lambda\ e^{-\lambda}$
- $\mathbb{P}(Y = 2) = \frac{\lambda^2}{2!} e^{-\lambda}$
- $\mathbb{P}(Y = 3) = \frac{\lambda^3}{3!} e^{-\lambda}$
- $\mathbb{P}(Y = 4) = \frac{\lambda^4}{4!} e^{-\lambda}$
- Etc.

To illustrate with a particular case, let's assume that $\lambda = 3.87$. Then

- $\mathbb{P}(Y = 0) = e^{-3.87} = 0.021$
- $\mathbb{P}(Y = 1) = 3.87 \times\ e^{-3.87} = 0.081$
- $\mathbb{P}(Y = 2) = 3.87^2/2! \times\ e^{-3.87} = 0.156$
- $\mathbb{P}(Y = 3) = 3.87^3/3!\ e^{-3.87} = 0.201$
- Etc.

The probability function of a Poisson distribution is illustrated in Figure 6.7. It shows the probability function using $\lambda = 3.87$.

Figure 6.8 shows several Poisson probability functions varying the value of λ. We use $\lambda = 0.1, 0.6$, and 1.5 in the top line charts, going from left to right. On the bottom line, we use $\lambda = 3, 10, 50$. Notice how the probabilities are concentrated on the smallest integers when λ is small. À As λ grows, the probabilities of larger integers increase and, at the same time, the integers close to zero are left with negligible probabilities. In fact, you may have noticed that the integers that have the highest probabilities on each graph are those around the value of λ.

In fact, this is not a coincidence. The expected value $\mathbb{E}(Y)$ of an r.v. X with Poisson distribution with parameter λ is $\mathbb{E}(Y) = \lambda$. The proof uses the Taylor expansion around zero of the exponential function as a power series:

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \ldots$$

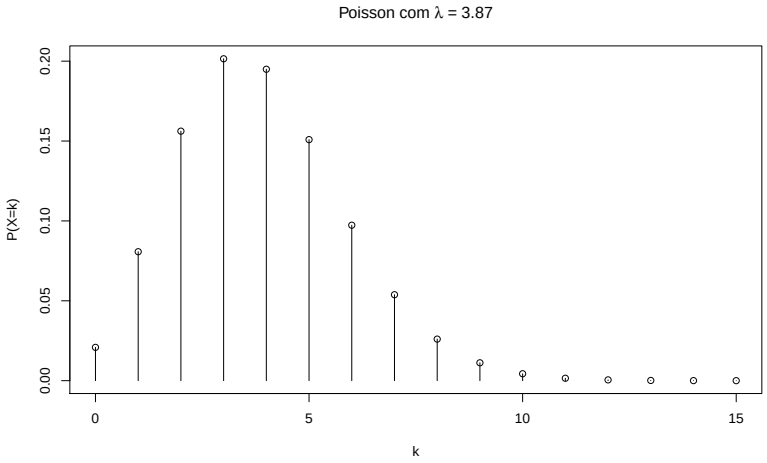


Figure 6.7: Poisson probability function with $\lambda = 3.87$.

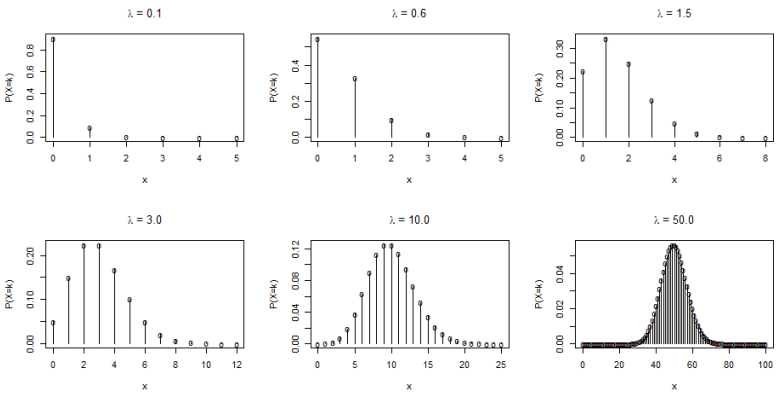


Figure 6.8: Poisson probability functions varying the value of λ.

We have

$$\begin{aligned}
\mathbb{E}(Y) &= \sum_{k=0}^{\infty} k\mathbb{P}(Y=k) = 0\times\mathbb{P}(Y=0)+1\times\mathbb{P}(Y=1)+2\times\mathbb{P}(Y=2)+\ldots \\
&= 1\times\lambda\, e^{-\lambda}+2\times\frac{\lambda^2}{2!}\,e^{-\lambda}+3\times\frac{\lambda^3}{3!}\,e^{-\lambda}\ldots \\
&= \lambda e^{-\lambda}\left(1+\frac{2\lambda}{2!}+\frac{3\lambda^2}{3!}+\frac{4\lambda^3}{4!}+\frac{5\lambda^4}{5!}+\ldots\right) \\
&= \lambda e^{-\lambda}\left(1+\frac{\lambda}{1!}+\frac{\lambda^2}{2!}+\frac{\lambda^3}{3!}+\frac{\lambda^4}{4!}+\ldots\right) \\
&= \lambda e^{-\lambda}\left(e^{\lambda}\right) \\
&= \lambda
\end{aligned}$$

■ **Example 6.14 — Deaths from horse kicks.** The Poisson distribution is often called the rare event (count) distribution. In 1898, Ladislaus von Botkiewicz was one of the first to use the Poisson distribution by counting the occurrence of a very rare event in his book *Das Gesetz der kleinen Zahlen*, which can be translated as *The law of small numbers* , He showed that relatively rare events for a given individual, when observed in a large population, have regularities that are well approximated by the Poisson distribution. The data presented by him in his book on the number of men killed by horse kicks in certain Prussian army corps during twenty years (1875-1894) became classics.

In each of the 20 years, he recorded the death toll in each of the 10 corporations. So we have 200 counts coming from 10 corporations times 20 years. In general, no deaths occurred: 109 of the 200 counts were zero. In 65 corporation-years we had only 1 death. The complete situation can be found in the second column in the table below. It shows the number of corporation-years with zero deaths, 1 death, etc.

k dead in the year	Observed frequency	$\mathbb{P}(Y=k)$	Expected frequency
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

The third column presents the probability formula $\mathbb{P}(Y=k)$ from the Poisson distribution. For this, we first need a value for the λ parameter. If the 200 counts are achievements of an r.v. which follows the Poisson distribution, the arithmetic mean of these 200 counts should be approximately equal to $\mathbb{E}(Y)=\lambda$. This arithmetic mean is equal to $(0\times 109+1\times 65+2\times 22+3\times 3+4\times 1)/200 = 122/200 = 0.61$. Using $\lambda = 0.61$ we calculate $\mathbb{P}(Y=k) = 0.61^k e^{0.61}/k!$ for $k=0,1,2,3,4$ in the third column of the table . The fourth column shows how many corporations we should expect with zero deaths in the year, with one death in the year, with two deaths in the year, etc. For example, if there is a probability equal to 0.5434 that a corporation-year has zero deaths, at the end of 200 corporation-years we should have approximately $200\times 0.5434 = 108.7$ with zero deaths. Similarly, we should have approximately $200\times 0.3314 = 66.3$ corporation-years with one death. The proximity between the observed and expected frequencies shows that these data could be generated by a Poisson distribution and therefore it is an adequate distribution to model these data.

■

Figure 6.9: Left: Radioactive material. Right: A staff member who looks after the safety of a nuclear reactor.

■ **Example 6.15 — Errata.** An example of the use of the Poisson distribution that is often cited is the number of typos or typographical errors on a document page. Professor Kim Border of Caltech, in his unpublished lecture notes, recounts the case of his colleague Phil Hoffman. This colleague wrote the book *How Did Europe Conquer the World?* and received proofs for final verification from the publisher before publication. In $n = 261$ pages there were a total number of 43 errors. There were 222 pages with zero errors, one error in 35 pages, and 2 errors in 4 pages. On average, there was $43/261 = 0.165$ error per page and this number serves as an estimate for the λ parameter. The table with the observed and expected counts under the Poisson model is in the table below. The proximity of values is quite impressive.

	0	1	2	≥ 3
obs	222	35	4	0
esp	221.4	36.5	3.0	0.17

Note that the last cell needs the probability value that $\mathbb{P}(Y \geq 3)$. As the Poisson distribution has infinitely many values you will have to calculate an infinite series: $\mathbb{P}(Y \geq 3) = \mathbb{P}(Y = 3) + \mathbb{P}(Y = 4) + \ldots$. However, we can obtain this value by subtraction after obtaining the first probabilities because

$$\mathbb{P}(Y \geq 3) = 1 - \mathbb{P}(Y = 0) - \mathbb{P}(Y = 1) - \mathbb{P}(Y = 2)\,.$$

■

The genesis of a Poisson

How does this Poisson distribution appear? There is a classic example showing that counts of radioactive particle emissions by an atomic mass follow a Poisson distribution (Figure **??**). A Geiger-Müeller counter records the number of particles hitting a plate in an interval of 7.5 seconds. Possible values for counting this number of particles are $\mathbb{N} = \{0, 1, 2, \ldots\}$ Particle counts are random and any two counts in two equal time periods are unlikely to be the same. There is a good deal of variability in these counts, even if the observation time is the same.

To obtain the probabilities, we will adopt two paths: one more theoretical, the other more empirical. A theoretical model for the emission of particles by a radioactive mass was proposed by physicists in the last century and consists of three hypotheses that were well grounded in the practical observation of the radioactive phenomenon.

- **Hypothesis 1:** The probability of the arrival of k particles in a time interval $(t, t + \Delta)$ depends only on the length Δ of the interval and not on the moment t of its start. That is, take two time intervals (in seconds) of equal duration such as, for example, $(1, 1 + \Delta)$ (one second after the

start of the experiment) and $(7200, 7200+\Delta)$ (two hours after its start). The probability of observing, say, 5 particles in the first range is the same as in the second range. There is no "wear out" of atomic mass (at least during experiments of non-excessive duration). Hours after the start of the experiment, everything happens as if we were at the beginning of it. The probability of emitting k particles in an intervalΔ of time does not depend on when we started the interval. Probability depends on Δ: long time intervals will have higher counts. However, the probability of counting k particles depends only on this length Δ and not on the momentum start of the observation period.

- **Hypothesis 2:** The numbers of particles in disjoint time intervals are independent r.v.'s. Suppose that, in a time interval of Δ seconds, we observe on average 5 particles. If in a given interval $(t, t+\Delta)$ we observe much more than the average of 5 particles (say, with 10 particles), then in the next interval $(t+\Delta, t+2\Delta)$ of duration Δ there will be no tendency to correct the excess of the first interval, nor any stimulus to continue emitting more particles than the mean of 5.
- **Hypothesis 3:** The particles arrive alone, not simultaneously.

It can be mathematically proved (see [22, page 22]) that a stochastic system with these three properties or hypotheses must necessarily have the Poisson probability distribution for the counts in a time interval: $\mathbb{P}(Y=k) = \frac{\lambda^k}{k!}e^{-\lambda}$ for $k = 0, 1, 2, \ldots$ and where λ is a positive constant associated with the radioactive mass and represents the expected value of emissions in the time interval.

Is this mathematical deduction consistent with reality? [31] "repeat" the experiment a large number of times. They counted the number of particles emitted in 2608 consecutive time intervals of 7.5 seconds each. Let $y_1 = 4, y_2 = 3, y_3 = 0, \ldots, y_{2608} = 4$ be the particle counts emitted in each interval. Let's assume they are the instantiated values of the r.v.'s i.i.d $Y_1, Y_2, \ldots, Y_{2608}$, all with Poisson($\lambda$) distribution. If this Poisson model for particle emission is correct what can we expect to see in the observed counts? Let's compare the theoretical values $\mathbb{P}(Y=k) = \frac{\lambda^k}{k!}e^{-\lambda}$ with the observed frequency of intervals with counts equal to k.

First, let's calculate $\mathbb{P}$(emit k particles in 7.5 seconds) using the Poisson model. We calculate the arithmetic mean of the observations as an approximation to λ. So we use $\hat{\lambda} = (y_1 + y_2 + \ldots y_{2608})/2608 = 3.87$. We can now calculate $\mathbb{P}(Y=k) = \frac{3.87^k}{k!}e^{-3.87}$ for the different values of k. These probabilities are in the second column of the table below.

k	$\mathbb{P}(Y=k)$	Empirical frequency
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

The third column shows the proportion of the 2608 intervals in which we obtained k particles. Note that we don't use any probability models here, just the observed data. If the model is correct, these two values should be similar for all k. In fact, the proximity of values is very great. This shows that the Poisson model may well be the mechanism that generates the observed data. The model fits the empirical data very well.

There are two situations in which the Poisson distribution appears. When counting number of rare occurrences without a clear limit to the maximum number such as:

- number of collisions in BH traffic between 6 pm and 7 pm on a weekday.
- number of cars entering UFMG between 7 and 8 am.
- number of medical appointments that a customer of a health plan makes during the year.
- number of failures per day registered in a network management system.
- Number of virus infections on a file server in a data center per day.
- number of visits to a web page per minute.
- number of calls to a call center per minute.

If the three hypotheses about radioactive decay are also approximately valid for these random events listed above, we can expect to see the Poisson distribution appearing. For example, in the first case above, if the chance of observing 3 collisions in BH in a small interval Δ of time remains constant between 18 and 19 hours, if the collisions occur independently and do not occur simultaneously, we can expect a Poisson count as a result. Note that in real-world situations, due to the seasonality aspect, the probability of k events occurring can only remain constant within limited time intervals.

Another situation in which the Poisson distribution appears naturally is as an approximation to the distribution of an r.v. X following the binomial distribution $\text{Bin}(n,\theta)$ when n is large and θ is small. Since $\mathbb{E}(X) = n\theta$, if a Poisson distribution is a good approximation, we should have $\lambda \approx n\theta$. Examples for this case:

- number of deaths from esophageal cancer during the year in BH. Each individual in BH flips a coin at the beginning of the year to determine whether he will die of esophageal cancer during that year or not. We have a large number n of coin flips and a very small probability of "success" θ.
- number of car insurance policies with 2 or more claims during a given year. The insurance policy portfolio has hundreds of thousands of customers and this is the n. The probability θ that an insured has two or more claims during the year is small.

■ **Example 6.16 — Bombs in London.** At the end of the Second World War, the Germans developed the first long-range guided ballistic missiles, the bombs $V1$ and $V2$ (Figure 6.10). They were dropped from the European continent across the English Channel into England and, in particular, London. After a while, certain boroughs of London were being hit hard, while others were not.

At that time, it was usually necessary to drop large numbers of bombs blindly and randomly to ensure the destruction of a target. The suspicion began to arise that the Germans had achieved some innovation in this respect. Some blocks were of military interest and were hit hard by the bombs, leading to suspicions that the Germans had more accurate bombing capabilities than the Allies. Some blocks were of military interest but were relatively spared. For these cases, the suspicion arose that their spies lived there. This was an important military secret at the time: knowing the accuracy of these bombs. Were they falling haphazardly over the city or were they hitting their intended targets? Had the Germans really managed to make a precision-guided bomb?

In 1946, R.D.Clarke, a British actuary, published a note describing the work he had done during the wartime to answer this question ([9]). This analysis became famous and appears in every probability book. Charles Franklin, a professor at the University of Wisconsin, reconstructed the data in this old article from the original maps of the bomb sites held in the British archives at Kew. He published a three-dimensional view of bomb density on his blog. `http://madvis.blogspot.com.br/2010/09/flying-bombs-on-london-summer-of-1944.html` and reproduced in Figure 6.11. It shows the density surface of dropped bombs from June to August 1944. We view the surfaces from two angles, northeast (left) and southwest (right). The higher the density, the greater the number of impacts per km^2.

Figure 6.10: Left: $V2$ bomb ready to be dropped on London. Right: Life in London continued during the war

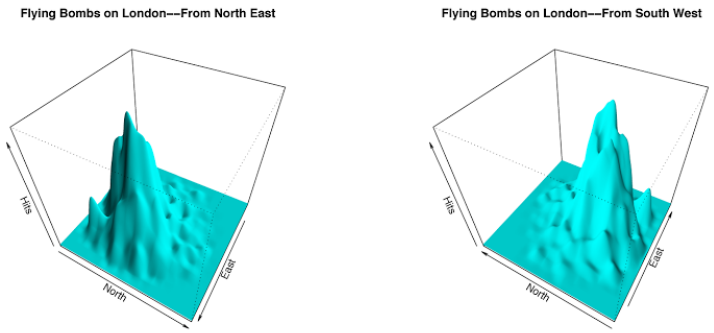


Figure 6.11: Surface representing the bomb density per km^2 in London from two perspectives. Taken from blog `http://madvis.blogspot.com.br/2010/09/flying-bombs-on-london-summer-of-1944.html`.

Figure 6.12: Left: Sir Churchill visiting bomb-hit sites on 10 September 1940 in London during World War II. Taken from `http://www.bbc.co.uk/history/events/germany_bombs_london`. Right: Schematic drawing illustrating the situation in which we have South London divided into $N = 576$ squares with an area equal to 0.25km^2. The dots represent the places where bombs fell.

From the perspective of Greater London, the highest density of attack is well defined, with impacts concentrated in a certain region. So, obviously, there was some precision. But was there enough precision to distinguish targets within this region of highest concentration? Clarke's analysis focused on the area of greatest density. Within that area, his analysis found no evidence of agglomeration that could not be explained by the variation of an r.v. Poisson. The reasoning is based on a fine grid with $N = 576$ squares, each with $0.25km^2$, in the region south of London, the most affected (see Figure 6.12. If germans didn't have crosshairs, the number of bombs in one of those little squares would be a value coming from an r.v. with Poisson(λ) distribution, the *same* λ wherever the little square was. Why?

The reasoning is as follows. Fix a small square on the map of Figure **??**. Let X be the number of bombs in the square. We had a large number of B bombs being dropped over the map. There is a small probability θ of hitting a specific square. Counting the bombs in a specific square is like counting the number of "successes" in B flips of a coin with a small success probability θ. We use Poisson's binomial approximation with $\lambda = n\theta$. The crucial point is that, *if the Germans don't have crosshairs*, the probability θ is the *same* for every small square. We have $N = 576$ squares with counts $Y_1, \ldots, Y_N$, all being r.v.'s Poisson with the same parameter (λ). Does this theoretical model fit the data? If so, this would be evidence in favor of the no crosshair hypothesis.

As before, we calculate the probabilities $\mathbb{P}(Y = 0)$, $\mathbb{P}(Y = 1)$, $\mathbb{P}(Y = 2)$, etc, using the model of Poisson. Next, we obtain the proportion of the squares where we had counts $Y = 0$, $Y = 1$, $Y = 2$, etc., using only the empirical data. Next, we compare the theoretical Poisson probabilities with the frequencies based on the data alone. If they are similar, the data is compatible with the model.

The total number of bombs in London was 537 and there are $N = 576$ squares. So the average number of bombs per square is $\hat{\lambda} = 537/576 = 0.9323$. Let Y_i be the number of bombs in the box i. We assume that $Y_1, \ldots, Y_n$ are i.i.d. with Poisson(λ) distribution with $\lambda = 0.9323$. In the table below, k is the number of bombs in a square, N_k is the number of squares that were hit by k bombs, $N_k/576$ is the proportion of squares hit by k bombs, and $\mathbb{P}(Y = k) = 0.9323^k/k!e^{-0.9323}$ is

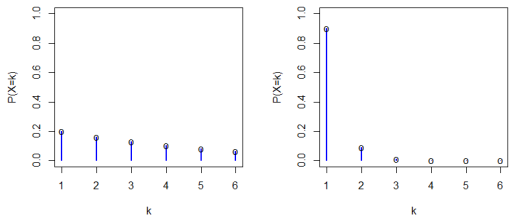


Figure 6.13: Probability function $\mathbb{P}(Y=k)$ of a geometric distribution with $\theta=0.2$ (left) and $\theta=0.9$ (right).

the probability of an r.v. Poisson$(\lambda=0.9323)$ is equal to k. The proximity between the empirical frequency and the theoretical probabilities is impressive.

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

■

6.5.5 Geometric

This distribution appears from the ubiquitous experiment of successively flipping a coin with a success probability θ. The interest is to know how many launches are needed until the first success is observed. of course this wait time for the first success depends on θ. If θ is a value close to 1 we should get the first success right on the first or second launch, hardly needing many launches to stop. On the other hand, if $\theta \approx 0$, we shouldn't expect too few releases. Let's study more precisely how this waiting time takes place.

Let Y be the number of throws in a sequence of independent Bernoulli trials until the first success is observed. In each trial the probability of success is θ. The event $Y=1$ means that the first trial was a success S. We have $\mathbb{P}(Y=0)=\mathbb{P}(S)=\theta$. The event $Y=2$ means that the first trial was a failure F and the second was a success S. Since coin flips are independent, the probability of observing FS is the product of the probabilities of the outcomes of each individual toss: $\mathbb{P}(Y=2)=\mathbb{P}(FS)=(1-\theta)\theta$. The case $Y=3$ works similarly: $\mathbb{P}(Y=3)=\mathbb{P}(FFS)=(1-\theta)^2\theta$. And the general case comes out easy now: $\mathbb{P}(Y=k)=(1-\theta)^{k-1}\theta$, for $k=1,2,\ldots$.

Figure 6.13 shows the probability function $\mathbb{P}(Y=k)$ of a geometric distribution with $\theta=0.2$ (left) and $\theta=0.9$ (right) . The vertical axis is the same in both graphs. See how the probabilities $\mathbb{P}(Y=k)$ are fundamentally concentrated on the values $k=1$ and $k=2$ when $\theta=0.9$. We can calculate the chance of needing to flip the coin 3 or more times to get the first success in this second case: $\mathbb{P}(Y\geq 3)=1-\mathbb{P}(Y=1)-\ mathbbP(Y=2)=1-0.9-0.9\times 0.1=0.01$. So only 1% of the times we do this experiment with a coin with $\theta=0.9$ can we expect to have to flip the coin 3 times or more. In the other case, with $\theta=0.2$, we have $\mathbb{P}(Y\geq 3)=1-0.2-0.2\times 0.8=0.64$. So, in this case, the chance of having to wait 3 or more rolls is greater than that of stopping before 3.

When Y is geometric with success parameter equal to θ, we can show that $\mathbb{E}(Y)=1/\theta$. So if $\theta=0.1$ we expect to have to flip the coin $\mathbb{E}(Y)=1/0.1=10$ times before seeing the first success. If the coin is fair and $\theta=0.5$ then $\mathbb{E}(Y)=1/0.5=2$ while $\mathbb{E}(Y)=1/0.01=100$ if $\theta=0.01$.

Let's now calculate $\mathbb{E}(Y)$ using simple infinite series tricks. Remember the geometric series: if $x \in (0,1)$ then

$$1+x+x^2+x^3+\ldots = \frac{1}{1-x}.$$

We have

$$\begin{aligned}
\mathbb{E}(Y) &= \sum_{k=1}^{\infty} k\mathbb{P}(Y=k) = 1\times\mathbb{P}(Y=1)+2\times\mathbb{P}(Y=2)+\ldots \\
&= 1\times\theta+2\times\theta\times(1-\theta)+3\times\theta\times(1-\theta)^2+\ldots \\
&= \theta\left(1+2\times(1-\theta)+3\times(1-\theta)^2+\ldots\right. \\
&= \theta\left(0+1+2\times(1-\theta)+3\times(1-\theta)^2+\ldots\right. \\
&= \theta\left(\frac{d}{d\theta}(-1)-\frac{d}{d\theta}(1-\theta)-\frac{d}{d\theta}(1-\theta)^2-\frac{d}{d\theta}(1-\theta)^3+\ldots\right. \\
&= \theta\,\frac{d}{d\theta}\left(-1\times\left(1+(1-\theta)+(1-\theta)^2+(1-\theta)^3+\ldots\right.\right. \\
&= \theta\,\frac{d}{d\theta}\left(-\frac{1}{1-(1-\theta)}\right) \\
&= \theta\,\frac{d}{d\theta}\left(-\frac{1}{\theta}\right) \\
&= \frac{1}{\theta}
\end{aligned}$$

Definition 6.5.5 — Geometric distribution. The r.v. Y has a geometric distribution if its support is the set $\{1,2,\ldots\}$ with probability function given by $\mathbb{P}(Y=k)=(1-\theta)^{k-1}\theta$, for $k=1,2,\ldots$.

6.5.6 Pareto ou Zipf

The Pareto distribution gained a lot of importance in more recent times when several phenomena exhibited the typical behavior that we find in these distributions, called *heavy tail behavior*. In the discrete case, it is often called the Zipf distribution.

Definition 6.5.6 — Zipf distribution (Discrete Pareto). Let X be an r.v. with support equal to the set $\{1,2,3,\ldots,N\}$. The value of the maximum number N can be finite or infinite. It is called *Zipf distribution* or *Discrete Pareto* if the probabilities are as follows: $\mathbb{P}(X=k)=\frac{C}{k^{1+\alpha}}$ with $\alpha>0$ where $C>0$ is a constant such that the probabilities add up to 1.

When $\alpha=1$, we have $\mathbb{P}(X=k)=C/k^2$. If $\alpha=2.5$, we have $\mathbb{P}(X=k)=C^*/k^{3.5}$. We use C^* in this case just to emphasize that the constant for the case $\alpha=2.5$ is different from that for the case $\alpha=1$. We can have $0<\alpha<1$. For example if $\alpha=0.5$ then $\mathbb{P}(X=k)=C^{**}/k^{1.5}$. What *really* matters is this: the probability $\mathbb{P}(Y=k)$ decreases with a power of k. For this reason, it is called a power law distribution. It does not fall exponentially as is the case with a Poisson and a geometric distributions. We return to comparing these distributions at the end of this section.

The constant C in the formula above the distribution should guarantee that the probabilities add up to 1. Thus, if we set a value for α, the value of C is determined because we must have

$$\begin{aligned}
1 &= \mathbb{P}(X=1)+\mathbb{P}(X=2)+\mathbb{P}(X=3)+\ldots \\
&= C\left(\frac{1}{1^{1+\alpha}}+\frac{1}{2^{1+\alpha}}+\frac{1}{3^{1+\alpha}}+\ldots\right) \\
&= C\sum_{k=1}^{N}\frac{1}{k^{1+\alpha}}
\end{aligned}$$

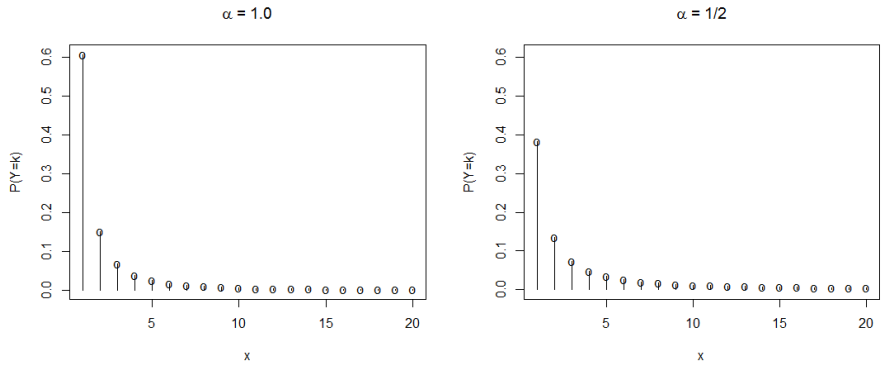


Figure 6.14: Probability $\mathbb{P}(Y = k) = c/k^{1+\alpha}$ of a discrete Pareto (or Zipf) distribution with $\alpha = 1$ (left) and $\alpha = 1/2$ (right). The vertical axis scale is the same in both cases.

which implies that

$$C = \frac{1}{\sum_{k=1}^{N} 1/k^{1+\alpha}}\,.$$

When $N = \infty$, this constant is associated with the Riemman zeta function defined as

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}\,.$$

for $s > 1$. This function has been studied extensively in mathematics. Using the values calculated for this Riemman zeta function, when $N = \infty$ and $\alpha = 0.5$, we should have a constant $C = 1/\zeta(1.5) \approx 1/2,612 = 0.383$, while $C = 1/\zeta(2.0) \approx 1/1,645 = 0.608$ when $\alpha = 1.0$.

Figure 6.14 shows the probability function $\mathbb{P}(Y = k)$ of a Pareto distribution with $N = \infty$ and $\alpha = 1$ (left) and $\alpha = 0.5$ (right). The vertical axis is the same in both graphs. The probability value $\mathbb{P}(Y = 1)$ is quite different in the two cases, being greater than 0.6 when $\alpha = 1$ but approximately 0.4 when $\alpha = 0.5$. Except for this very different value, it is very difficult to visually identify marked differences between the two graphs. And yet they exist. We know that the sum of the probabilities is equal to 1 in both cases. So, the 0.2 decrease in the value of $\mathbb{P}(Y = 1)$ when going from the left to the right graph must have leaked to the other odds that won this additional 0.2. But in the graph on the right, we hardly see the heights of $\mathbb{P}(Y = k)$ greater than those on the left when $k \geq 2$. But those differences are there.

To get an idea of the difference between the two graphs and start to understand the heavy tailed behavior, let's calculate $\mathbb{P}(Y \leq k)$ in each case, with $\alpha = 1$ and with $\alpha = 0.5$. Figure 6.15 shows the graph of the cumulative probability distribution function $\mathbb{F}(k) = \mathbb{P}(Y \leq k)$ for different values of k with $\alpha = 1$ (left) and $\alpha = 1/2$ (right). Now, we see a huge difference between the two distributions, a difference that was not obvious in Figure 6.14. The cumulative distribution $\mathbb{F}(k)$ with $\alpha = 1$ gets closer to the maximum 1 faster than the cumulative distribution with $\alpha = 1/2$. For example, with $\alpha = 1$, we have $\mathbb{F}(15) = \mathbb{P}(Y \leq 15) = 0.96$ while, with $\alpha = 1/2$, we have just $\mathbb{F}(15) = 0.81$.

However, the most notable aspect of Figure 6.15 is that it clearly displays the heavy tail behavior. To understand this concept, we will classify distributions that can be symmetrical, as in the case of the distribution on the left side of Figure 6.16, or asymmetric, as in the right side of Figure 6.16. A symmetrical distribution is one in which the left and right sides of the distribution are approximately balanced around a central point. It can be proved that, in the symmetric case, this central point coincides with the expected value $\mathbb{E}(X)$. The tails of the distribution are the left and right parts away from the center. The tail is the part where the values of $\mathbb{P}(X = k)$ get smaller. For

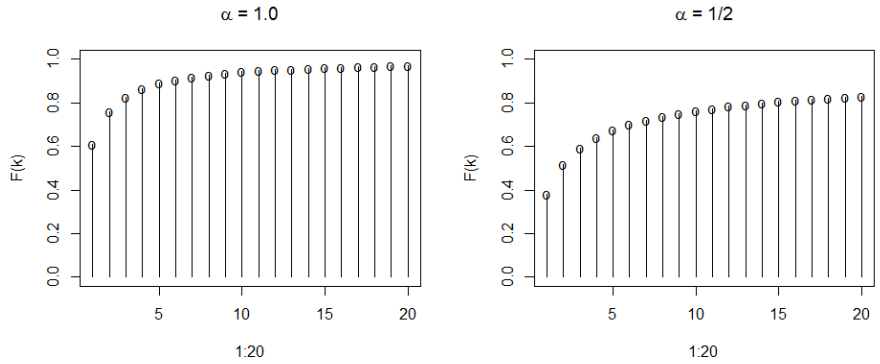


Figure 6.15: Probabilities $\mathbb{F}(k) = \mathbb{P}(Y \leq k)$ of a discrete Pareto (or Zipf) distribution with $\alpha = 1$ (left) and $\alpha = 1/2$ (right). The vertical axis scale is the same in both cases.

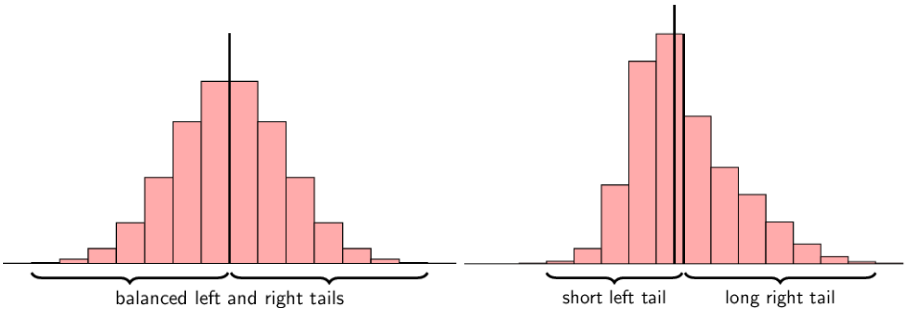


Figure 6.16: Symmetric (left) and asymmetric (right) distributions.

a symmetric distribution, the left and right tails are equally balanced, meaning that the probabilities $\mathbb{P}(X = k)$ have approximately the same decay in both directions. In the asymmetric case, the tail of the distribution on one side is longer than on the other side. In the case of Figure 6.16, the right side spreads out more than the left side and we say that we have asymmetry on the right.

In the graphs of Figure **??**, we see that Pareto distributions are skewed and skewed to the right. That is, they are clustered on the left and with a "tail" extending to the right. In the case of the Pareto distribution, this tail on the right concentrates a lot of probability, even though this does not appear when looking at Figure **??**. Looking at the graphs in Figure 6.15, we can see how the distribution with $\alpha = 1/2$ falls so slowly that even very large values of k still have a non-negligible probability. In Figure 6.14 there seems to be no probability mass $\mathbb{P}(X = k)$ with practical relevance to $k \geq 10$ in both cases ($\alpha = 1$ and $alpha = 1/2$). However, in both cases, there is still a non-negligible probability that we will see values of X greater than 10. We have $\mathbb{P}(X > 10) = 0.06$ in the case $\alpha = 1$ and $\mathbb{P}(X > 10) = 0.24$ in the case $\alpha = 1/2$. That is, if you generate an r.v. Pareto with $\alpha = 1/2$ on the computer, the chance of observing a value that 10 is 25%.

Well, maybe this effect will wear off soon. For example, can we have a Pareto with $\alpha = 1$ or with $\alpha = 1/2$ generating numbers that are unlikely to exceed 20. Thus, Pareto would generate values but they would not spread very far from the region where most of them would be concentrated. In fact, this does not happen. We have $\mathbb{P}(X > 20) = 0.03$ for $\alpha = 1$ and $\mathbb{P}(X > 10) = 0.16$ for $\alpha = 1/2$. Going much further, especially in the case $\alpha = 1/2$, leads to surprising results. See the table below for some values of $\mathbb{P}(X > k)$. Note how the probability $\mathbb{P}(X > k)$ drops very slowly. If we simulate a Pareto distribution, most of the data will have small values. However, numbers

orders of magnitude larger than most data will show up quite easily. This is the phenomenon of the heavy tail. In binomial, Poisson, or geometric distributions, numbers much larger than most are very unlikely, almost impossible.

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

Examples of Pareto distribution

The Pareto distribution is very common in web studies. A few sites have millions of pages but hundreds of millions of sites have just a few pages. Few sites contain millions of links, while the vast majority of them do not have more than a dozen links. Hundreds of millions of users visit a few sites paying little attention to billions of others. Not just on the web. Individuals' labor income and families' wealth show Pareto behavior. Few have fortunes, while most have relatively little. Size distributions such as the size of companies and the size of cities in the world also show a Pareto pattern.

At a health care provider, most customers incur a paltry annual cost, but a few individuals are responsible for a disproportionate amount of the total cost. In [4], we find the typical result of the concentration of health expenditures. Ranking individuals by what they cost insurers in the US and taking the group of 1% of those who had the most expenses, it appears that this small group consumed 27% of the total annual expenditure. The top 5% group consumed 55%, while the top 10% consumed 69% of the total. Another example is in relation to the frequency of words in any language. Some words are used very frequently but most are used very little.

These are all cases of extreme imbalance, where most values are relatively small but there is a heavy tail persistently generating values orders of magnitude above the majority. They are all candidates to be modeled by the Pareto distribution, either in its discrete version, as we have seen here, or in the continuous version, in the next chapter.

In the 6.5.6 section we will discuss how to check whether the observed data follows a Poisson distribution.

Distribuição de Zipf

The Zipf distribution is a particular case of the Pareto distribution, when $\alpha = 0$, and it was popularized in language studies by the American George Kingsley Zipf (1902–1950). zipf found one statistical regularity in texts written in different languages. Considering a large collection of texts in any language, he noticed that some words appear little, are rarely used. Others appear with great frequency. For example, in the table below, extracted from `www.linguatec.pt`, we have some of the most frequently used words in Brazilian Portuguese, as well as some less frequently used words. The table shows the position or *rank* of some words in the second column. For example, the preposition *de* (of) is the most used in Portuguese and therefore has rank 1. The article *a* (the) is the second most used word and therefore has rank 2. The word *ser* (to be) has rank 27, *amor* (love) ranks 802, and so on. The third column shows the frequency of appearance of these words per million words. Thus, the word *de* appears an average of 79607 times in each group of 1 million words in a text. The word *chuva* (rain) occupies position 2087 and appears only 70 times in every million words.

Imagine the following random experiment: first, after processing a large amount of text, order the language words according to their rank. Next, choose a word completely at random from a text. Let Y be the rank (or rank) of this randomly chosen word. If $Y = 1$, it means that the word chosen at random is the most frequent word. If $Y = 17$, the chosen word is the 17^a most frequent. Zipf found that there was a great deal of regularity in the chance of choosing the rank word k. He found that $\mathbb{P}(Y = k)$ is approximately proportional to $1/k$. So, approximately, we have

word	rank	frequency
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: Rank of some words and frequency of their appearance per million words in Brazilian Portuguese texts

$$\begin{aligned}\mathbb{P}(Y=1) &\propto 1\\ \mathbb{P}(Y=2) &\propto 1/2\\ \mathbb{P}(Y=3) &\propto 1/3, \text{ e etc.}\end{aligned}$$

If the Zipf distribution is a good model for language data we should have

$$\mathbb{P}(Y=k) \approx \frac{C}{k} = \frac{C}{k^{1+\alpha}}\,.$$

This is a Pareto distribution with $\alpha = 0$. In practice, you usually find $\alpha \approx 0$.

How to check if the distribution is Pareto?

Because of the extreme asymmetry present in the Pareto distribution, it is not appropriate to directly use the graph of the probability function like those in Figure 6.14. Going back to the data from the table 6.5.6, we know from the frequentist view of probability, we know that $n_k = 10^6\mathbb{P}(Y=k)$ is approximately equal to the frequency (per million) of the word of rank k. That is, the third column of the table with the empirical frequency n_k should be approximately equal to 10^6 times the probability $\mathbb{P}(Y=k)$. If the Zipf-Pareto model is adequate, we have

$$n_k \approx 10^6\ \mathbb{P}(Y=k) = 10^6\ \frac{C}{k^{1+\alpha}}\,.$$

By taking logs from both sides we will have:

$$\log(n_k) \approx \log(10^6) + \log(\mathbb{P}(Y=k)) = \underbrace{\big(\log(10^6) + \log(C)\big)}_{a} - \underbrace{(1+\alpha)}_{b}\log(k) = a - b\log(k)\,.$$

Thus, if Zipf-Pareto is adequate, a plot of $\log n_k$ versus $\log(k)$ should be approximately a straight line with intercept $a = 6\log(10) + \log(C)$ and inclination $-b = -(1+\alpha)$.

Let's check if this happens by looking at the Brazilian Portuguese word data in the 6.5.6 table. The result is shown in Figure 6.17.

How to identify if a Pareto model is a good fit for data on the discrete values $\{12,\ldots\}$? As we have seen, one way is to count the number of times n_k that the value k appears in a sample. Next, we plot $\log(k)$ (on the x axis) versus the log frequency $\log(n_k)$ (on the y axis). If the Pareto model is suitable, we should observe approximately a straight line in this plot.

While this is a simple and useful technique, we have a more effective way of checking whether the Pareto distribution is a good model for the data. We use the cumulative distribution $\mathbb{F}(x) = \mathbb{P}(Y \leq k)$ for this. We will see this technique more effective when we study Pareto in the continuous case in the next chapter.

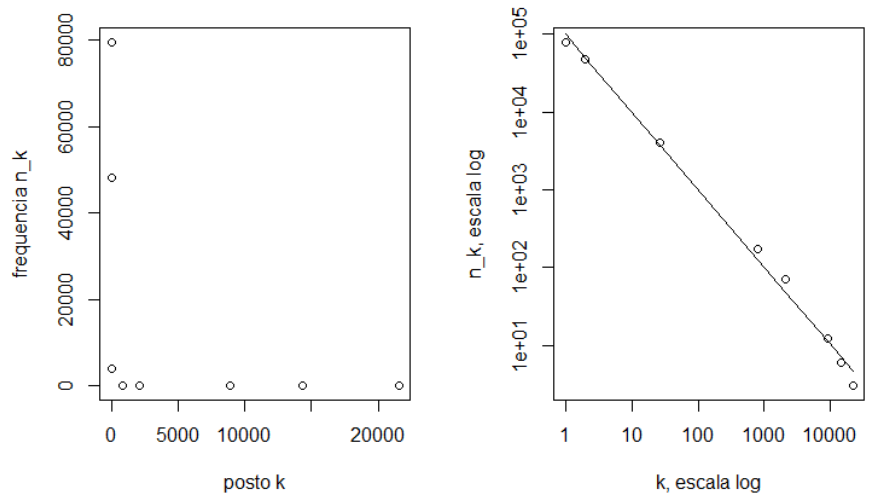


Figure 6.17: Graph of rank k versus frequency n_k for some Portuguese words (left). The graph on the right shows the same data in a log-log graph (ie the points are $(\log(k), \log(n_k))$. The line was obtained by the least squares technique (see chapter ??) and is equal to $\log(n_k) = 11.51 - 0.999\log(k)$).

6.6 Text classification and the multinomial distribution

This section presents another example of the multinomial distribution in action in the problem of text classification. It is a basic building block model for more sophisticated models aimed at the same task. We will describe the main aspects of a basic model, the *b*ag of words, by glossing over the details and practical problems that need to considered in a data analysis. However, the main probabilistic aspects will be covered and explained.

Imagine a large collection of texts (called a *corpus*), such as newspaper articles, where each document is classified under one of 3 topics: *sports*, *politics* or *others*. This collection is a sample of texts and each of them has been manually classified into one of 3 possible categories. This required a large amount of man-hours of work. The objective now is to use this sample of texts labeled by their categories to create an automatic classification rule. Other texts not considered, such as texts to be written in the future or other texts from the database must be classified in one of these 3 categories of no human intervention. We will extract information about the three categories from the collection of labeled texts and use this information to create a classification rule. Let's do this using the multinomial distribution to model the texts.

Texts as a string of words. Consider one of the topics. For example, *sport*. Let's think of a *generative model* of a *sport* text. The text will be generated as a sequence of words chosen at random from an ordered list of words. Ignoring some practical details for now, let's call this ordered list of D distinct words in the Portuguese language *dictionary*. The choice of words that will compose the text is made independently of each other. We associate a probability θ_i with the vocabulary word i.

A die with D faces We can imagine a "dice" with D faces, each face representing one of the distinct words in the vocabulary. The chance of choosing the word i is the probability θ_i of the "dice" producing the face i.

face-word	1	2	...	D	Soma
probabilities	θ_1	θ_2	...	θ_D	1

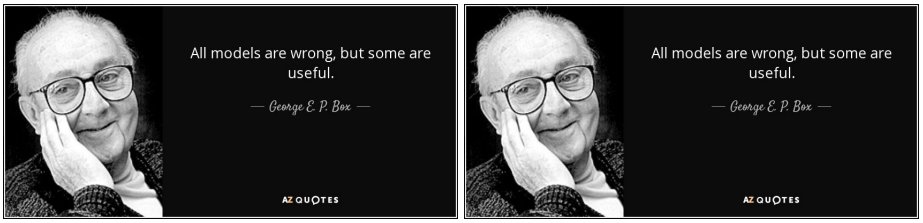


Figure 6.18: George Box and some of his aphorisms.

The text is generated by rolling the "dice" successively and independently. Thus, a sports text could be generated from this model producing, for example, the following result: *gol, Neymar, game, goal, full, goal, etc.*

Model is not realistic. Obviously, the chance of this generative model generating a text that is minimally similar to real text is very small. The sequence of words to be generated will hardly have the syntactic structure of Portuguese and will not even have a semantic sense. It is surprising that it is useful yet so simple and so blatantly false as a model for generating texts from reality. This is one of the best examples of the beautiful aphorism of the statistician George Box (Figure 6.18): "All models are wrong, but some are useful". There is even a page on wikipedia entirely to this aphorism: `https://en.wikipedia.org/wiki/All_models_are_wrong`. What he means is that all statistical models describe the world in a very simplified and therefore false way. However, despite this, they generate useful predictions and serve to capture essential aspects of the problem. All models drastically reduce the complexities of the world so that they are mathematically tractable. As put by [36], "these constraints come in the form of assumptions that seek to capture the probabilistic essence of the process. The objective is to formulate a simple, but not trivial, mathematical structure representing the essential and most relevant aspects of the phenomenon. Similar to a caricature, a good probabilistic model is not a faithful and perfect picture of an individual, but a sketch that reproduces and even amplifies or exaggerates the most salient features in order to make it easily recognizable."

The list of probabilities will vary from topic to topic. The list of D distinct words (the dictionary) is the same for all topics. However, each topic will assign probabilities differently. In the *sport* topic, the words *goal, player, net* will have higher probabilities θ_i than under the *politics* and *other* topics.

Getting the probabilities of each topic The probabilities of the words for each topic are obtained from the simple frequencies calculated in the manually labeled collection. Take all documents in the collection that have been classified as *sport*. Put all the words used in the texts in a bag of words (model *bag of words*). If the word *goal* appears 523 times throughout the texts, 523 copies of the word *goal* will be placed in the word bag. Count how many times each of the D words in the dictionary appears in the word bag and divide by the total number of words in the bag to get proportions that add up to 1. For example, if the word *goal* appears 1.5% of the time in the *sport* bag so $\theta_{goal} = 0.015$. This is repeated for each topic creating a different bag of words and therefore different θ_i probabilities.

Avoiding $\theta_i = 0$ Generally, many dictionary words will have $\theta_i = 0$ at the end of this probability estimation process. The reason is that, for example, the word *ineffable* may not have appeared even once in the *sport* collection. Words of very infrequent use in a topic will have probabilities equal to zero associated with them. This would indicate that this word could never appear in the future in a *sport* text. A collection of *sport* texts, even if very large, will have zero occurrences of many words that, although unlikely in a *sport* text, are not impossible. We want to prevent this future impossibility from appearing in our model.

A simple solution is to place a copy of each of the separate D words from the dictionary in the

word bag, in addition to the words themselves from the *sport* text collection. Suppose there are D distinct words and the text collection has a total of M words. The word i appears m_i times in the collection, where m_i can be zero. We add a copy of the word i to the bag so that it has m_i+1 occurrences of the word i. Even if $m_i = 0$, we will have at least one occurrence of the word i. In addition, the word bag has a total of M words (from the texts) plus D words, one copy for each word in the dictionary. So there is a total of $M+D$ words in the bag. Instead of estimating θ_i by the fraction m_i/M, use the estimator $\hat{\theta}_i = (m_i+1)/(M+D)$, the relative frequency of the word i in this bag enriched with copies. This estimator is called *Laplace estimator*. Note that if $m_i = 0$, we have $\hat{\theta}_i = 1/(M+D)$, a very small value but strictly greater than zero.

Completing the dictionary. Even with a very extensive dictionary, new words can appear that were not listed. New slang, foreign words, modern abbreviations, names of people or places, all are words that may not have been part of the initial dictionary. There are several solutions for this but a very simple solution to implement here is to add a pseudo-word *UNK* (from "unknown") to the dictionary that represents all these situations. So every time, when scrolling through the texts labeled as sport, we find any new word that was not in the dictionary, a copy of the pseudo-word *UNK* in the word bag. In practice, *UNK* becomes a (pseudo)-word in the dictionary representing all unlisted words. There are better solutions but this one will serve our purposes.

The probabilities of each topic

Suppose the probabilities of the distinct D words in the dictionary were estimated using Laplace's estimator on each of the three collections of texts, from *sport*, from *politics*, and *other*. The result is in a table:

word	1	2	...	D	Sum
θ_{1i}, esporte	θ_{11}	θ_{12}	...	θ_{1D}	1
θ_{2i}, política	θ_{21}	θ_{22}	...	θ_{2D}	1
θ_{3i}, outros	θ_{31}	θ_{32}	...	θ_{3D}	1

A new text appears and we want to automatically sort it into one of three categories. We use the multinomial distribution for this.

Sorting new text

The new text has M words in all, some repeated several times throughout the text:

$$\text{New text } = (x_1, x_2, x_3, \ldots, x_M)$$

where x_j is the word j of the new text. Using the *bag of words* model, what is the probability that this new text was written using the θ_{1i} probabilities of the *sport* topic? Let N_i be the random number of times the dictionary word i will appear in the new text. If the text is from sports and the *bag of words* model is valid, we have a multinomial for these counts:

$$\mathbf{N} = (N_1, N_2, \ldots, N_D) \sim \mathscr{M}(M; (\theta_1, \ldots, \theta_D))$$

The probability of the text Suppose the counts actually observed in the new text were the integers n_1, n_2, etc. We now calculate the probability of observing this new text *given the topic is sport*:

$$\mathbb{P}\left(\mathbf{N} = (n_1, n_2, \ldots, n_D) | \text{ sport } \right) = \frac{n!}{n_1! n_2! \ldots n_D!} \theta_{11}^{n_1} \ \theta_{12}^{n_2} \ldots \theta_{1D}^{n_D}$$

We do the same calculation for the other two topics:

$$\mathbb{P}\left(\mathbf{N} = (n_1, n_2, \ldots, n_D) | \text{ policy } \right) = \frac{n!}{n_1! n_2! \ldots n_D!} \theta_{21}^{n_1} \ \theta_{22}^{n_2} \ldots \theta_{2D}^{n_D}$$

and

$$\mathbb{P}\left(\mathbf{N} = (n_1, n_2, \ldots, n_D) | \text{ others } \right) = \frac{n!}{n_1! n_2! \ldots n_D!} \theta_{31}^{n_1} \ \theta_{32}^{n_2} \ldots \theta_{3D}^{n_D}$$

Note that the multinomial constant involving the factorials is the same in the three cases above.

Avoid computing the constant We can set one of the topic categories as a reference and compare the probabilities against this base category. For example, suppose we fix the category *sport* and calculate two ratios. The first of them:

$$\begin{aligned} r_{p.e} &= \frac{\mathbb{P}(\mathbf{N}=(n_1,n_2,\ldots,n_D)|\text{ policy })}{\mathbb{P}(\mathbf{N}=(n_1,n_2,\ldots,n_D)|\text{ sport })} \\ &= \left(\frac{\theta_{21}}{\theta_{11}}\right)^{n_1}\cdots\left(\frac{\theta_{2D}}{\theta_{1D}}\right)^{n_D} = \prod_{i=1}^{D}\left(\frac{\theta_{2i}}{\theta_{1i}}\right)^{n_i} \end{aligned}$$

The constant disappeared. If the word i does not appear in the new text (and therefore $n_i = 0$), then the factor $(\theta_{2i}/\theta_{1i})^{n_i}$ is equal to 1 and not needs to be calculated. Since most dictionary words will probably not appear in the new text, the calculations are greatly simplified.

Using the reasons

We also calculate the second ratio:

$$r_{o.e} = \frac{\mathbb{P}\Big(\mathbf{N}=(n_1,n_2,\ldots,n_D)|\text{ other }\Big)}{\mathbb{P}(\mathbf{N}=(n_1,n_2,\ldots,n_D)|\text{ sport })}$$

Suppose $r_{p.e}$ is greater than 1. For example $r_{p.e} = 4.3$. This means that the chance of having these counts n_1, n_2, etc. (ie having this new text) when the topic is *politics* is 4.3 times greater than the same chance when the topic is *sports*:

$$\mathbb{P}(\mathbf{N}=(n_1,n_2,\ldots,n_D)|\text{ policy }) = 4.3\ \ \mathbb{P}(\mathbf{N}=(n_1,n_2,\ldots,n_D)|\text{ sport })$$

If $r_{p.e}$ is less than 1, the reasoning is the opposite. For example, $r_{p.e} = 0.1$, the probability of the text under the topic *politics* is 10 times less than the same probability under the topic *sport*.

Making decisions

A decision rule then might be as follows:

- If $\max\{r_{p.e}, r_{o.e}\} < 1$, assign the new text to the reference category *sport*.
- If $\max\{r_{p.e}, r_{o.e}\} > 1$, assign the new text to the numerator-category which leads to the maximum of reasons.
- If $\max\{r_{p.e}, r_{o.e}\} = 1$, there is not enough evidence in the new text to allocate to one of the categories. One can randomly choose one of the categories that make up a ratio that is equal to 1.

This decision rule works well in many cases but in some situations it can (and should) be improved. The question is related to the difference between $\mathbb{P}(A|B)$ and $\mathbb{P}(B|A)$.

What probability do we want?

Our decision rule is based on comparing probabilities as

$$\mathbb{P}(\mathbf{N}=(n_1,n_2,\ldots,n_D)|\text{ sport })$$

But what we would really like to know is the value of the inverse probability:

$$\mathbb{P}(\text{ sport }|\mathbf{N}=(n_1,n_2,\ldots,n_D))$$

The first probability calculates the chance of having the new text GIVEN that it was written in the *sport* category. The second probability calculates the chance that the text is in the *sport* category given that it has the observed word configuration. In general, these probabilities are not equal and, in fact, can be quite different.

Calculating the inverse probability

We can use Bayes' rule to invert the probabilities of interest. For example,

$$\mathbb{P}\Big(\text{ sport }\Big|\mathbf{N}=(n_1,n_2,\ldots,n_D)\Big)=\frac{\mathbb{P}\Big(\mathbf{N}=(n_1,n_2,\ldots,n_D)\Big|\text{ sport }\Big)\,\mathbb{P}(\text{ sport })}{\mathbb{P}(\mathbf{N}=(n_1,n_2,\ldots,n_D))}$$

and

$$\mathbb{P}\Big(\text{ policy }\Big|\mathbf{N}=(n_1,n_2,\ldots,n_D)\Big)=\frac{\mathbb{P}(\mathbf{N}=(n_1,n_2,\ldots,n_D)|\text{ policy })\,\mathbb{P}(\text{ policy })}{\mathbb{P}(\mathbf{N}=(n_1,n_2,\ldots,n_D))}$$

Note that the right-hand denominators of the two expressions are identical and will disappear if we take the odds ratios:

$$\begin{aligned}\frac{\mathbb{P}(\text{ policy }|\mathbf{N}=(n_1,n_2,\ldots,n_D))}{\mathbb{P}(\text{ sport }|\mathbf{N}=(n_1,n_2,\ldots,n_D))} &= \frac{\mathbb{P}(\mathbf{N}=(n_1,n_2,\ldots,n_D)|\text{ policy })\,\mathbb{P}(\text{ policy })}{\mathbb{P}(\mathbf{N}=(n_1,n_2,\ldots,n_D)|\text{ sport })\,\mathbb{P}(\text{ sport })}\\ &= r_{p.e}\,\frac{\mathbb{P}(\text{ politics })}{\mathbb{P}(\text{ sport })}\end{aligned}$$

Prior decision We conclude that the decision should be based on the previously calculated r_e ratio BUT corrected by the product of the ratio $\mathbb{P}(\text{ policy })/\mathbb{P}(\text{ sport })$. This ratio calculates how often a text on *politics* appears in relation to the frequency of a text on *sports*. For example, suppose *sport* texts are 100 times more frequent than *politics* texts so that $\mathbb{P}(\text{ policy })/\mathbb{P}(\text{ sport }) = \frac{1}{100}$.
Updating a priori A priori, without looking at the new text, we know that there are many more articles on *sport* than *politics*. A reasonable decision rule, which decides *a priori*, without even looking at the new text, is to allocate any text that appears to the category *sport*. With the new text in hand, looking at the configuration of the words, we can change our decision rule *a priori* by allocating the new text to *policy*. But *sport* is so frequent that we will only do this if the evidence in favor of *politics* in the new text is very strong. For example, if $r_{p.e} = 1.1$, there is some but not much evidence in favor of *policy*. After all, this means that the chance of having the new text wordset when the topic is *politics* is only 10% greater than the same chance when the topic is *sport*.

With *sport* being 100 times more frequent than *politics* in general and with the new text having $r_{p.e} = 1.1$ we get

$$\frac{\mathbb{P}(\text{ policy }|\mathbf{N}=(n_1,n_2,\ldots,n_D))}{\mathbb{P}(\text{ sport }|\mathbf{N}=(n_1,n_2,\ldots,n_D))} = r_{p.e}\,\frac{\mathbb{P}(\text{ politics })}{\mathbb{P}(\text{ sport })} = \frac{1.1}{100} = 0.011$$

That is, the chance of being in *politics* remains approx. 100 times less than the chance of being *politics* even though $r_{p.e} > 1$. We continue to assign the text to *sport*.
A numerical problem Let $(\theta_1,\ldots,\theta_D)$ be a vector where θ_i is the probability that the dictionary word i will be used in a text on a certain topic (sports, say). The odds add up to 1. A specific text is parsed and you get the counts $n_1,\ldots,n_D$ so that n_i is the number of times the dictionary word i appeared in this text. Given that the text is really from sports, the probability that it has these counts is given by the multinomial model:

$$\mathbb{P}\left(\mathbf{N}=(n_1,n_2,\ldots,n_D)|\text{ sport }\right)=\frac{n!}{n_1!n_2!\ldots n_D!}\theta_1^{n_1}\ \theta_2^{n_2}\ldots\theta_D^{n_D}$$

We have already seen that the constant does not need to be evaluated and therefore your problem is to get the numerical value of the expression

$$\theta_1^{n_1}\ \theta_2^{n_2}\ldots\theta_D^{n_D}=\prod_{i=1}^{D}\theta_i^{n_i}$$

The D number of words in the dictionary is too large. Most θ_i probabilities are numbers close to zero. The product of many of them raised to the power n_i quickly leads to the limit of numerical precision of the machine. We end up with *underflow* and the product is transformed to 0. An illustrative example:

```
p1 = runif(1000) # one thousand random numbers between 0 and 1
# standardizing so p1 have probabilities summing to 1
p1 = p1/sum(p1)
# generating 1000 random counts between 1 and 100
contagens = sample(1:100, 1000, replace=T)
# calculating  theta^n
aux = (p1)^contagens
# obtaining their product: returns zero
prod(aux)
[1] 0
```

Solution: take logs The trick to doing this calculation is to use logarithms. On the log scale, products are transformed into sums and, therefore, tend to be much more numerically stable.

$$\begin{aligned}
\log \mathbb{P}(\text{text}) &= \log\left(\mathbb{P}\left(\mathbf{N}=(n_1,n_2,\ldots,n_D)|\text{ sport }\right)\right) \\
&= \text{cte.}+\log(\theta_1^{n_1})+\log(\theta_2^{n_2})+\ldots+(\theta_D^{n_D}) \\
&= \text{cte.}+n_1\log(\theta_1)+n_2\log(\theta_2)+\ldots+n_D(\theta_D) \\
&= \text{cte.}+\sum_{i=1}^{D} n_i\log(\theta_i)
\end{aligned}$$

In practice, the constant can be ignored (and does not need to be calculated) as it will be the same across all topics (sports, politics, etc). In R:

```
lp1 = log(p1)
aux = contagens*lp1
sum(aux)
# being more concise, in a single R line code
sum(contagens * log(p1))
```

In general, we want to calculate the probability of observing a certain text given that it is from politics *divided* by the probability of observing that same text given that it is from sports. This ratio is equal to $r_{p.e}$:

$$\begin{aligned}
r_{p.e} &= \frac{\mathbb{P}(\mathbf{N}=(n_1,n_2,\ldots,n_D)|\text{ policy })}{\mathbb{P}(\mathbf{N}=(n_1,n_2,\ldots,n_D)|\text{ sport })} \\
&= \left(\frac{\theta_{21}}{\theta_{11}}\right)^{n_1}\cdots\left(\frac{\theta_{2D}}{\theta_{1D}}\right)^{n_D} \\
&= \prod_{i=1}^{D}\left(\frac{\theta_{2i}}{\theta_{1i}}\right)^{n_i}
\end{aligned}$$

The same trick of taking logarithms applies here. The constant has already disappeared and using logarithms we have

$$\log(r_{p.e}) = \sum_{i=1}^{D}\log\left(\frac{\theta_{2i}}{\theta_{1i}}\right)^{n_i} = sum_{i=1}^{D} n_i\left(\log(\theta_{2i})-\log(\theta_{1i})\right)$$

If `p1` and `p2` are the probability vectors of the two topics in R, just write

```
sum(counts * (log(p1)-log(p2)))
```

This value is on the log scale. Thus, $r_{p.e} > 1$ implies $\log(r_{p.e}) > 0$ while $r_{p.e} < 1$ implies $\log(r_{p.e}) < 0$.

6.7 Comparison between distributions

Poisson $\times$ geometric $\times$ Pareto: what is the most relevant difference between them? All are distributions over positive integers. The main difference is in the behavior *on the tail*:

- Poisson has a short tail, values with significant probabilities are concentrated in a narrow band around their expectation $\mathbb{E}(Y) = \lambda$.
- Pareto easily generates very large values, orders of magnitude larger than $\mathbb{E}(Y)$.
- Geometric is an intermediate case.

Comparing the three:

- Poisson:

$$\frac{\mathbb{P}(Y = k+1)}{\mathbb{P}(Y = k)} = \frac{e^{-\lambda}\lambda^{k+1}/(k+1)!}{e^{-\lambda}\lambda^{k}/k!} = \frac{\lambda}{k+1} \to 0$$

  if $k \to \infty$. This is $\mathbb{P}(Y = k+1) <<< \mathbb{P}(Y = k)$ if k is large.
- Geometric:

$$\frac{\mathbb{P}(Y = k+1)}{\mathbb{P}(Y = k)} = \frac{(1-\theta)^{k+1}\theta}{(1-\theta)^{k}\theta} = 1-\theta < 1\,,$$

  constant in k. That is, $\mathbb{P}(Y = k+1) = (1-\theta)\mathbb{P}(Y = k)$, a geometric or exponential drop.
- Pareto:

$$\frac{\mathbb{P}(Y = k+1)}{\mathbb{P}(Y = k)} = \left(\frac{k}{k+1}\right)^{\theta} \to 1$$

  This is $\mathbb{P}(Y = k+1) \approx \mathbb{P}(Y = k)$ if k is large, a very slow drop.

7. Continuous Random Variables

7.1 Introduction

We have two main types of random variables: discrete, seen in the previous chapter, and continuous, which we will see in this chapter. The informal specification of a continuous r.v., as seen in the case of a discrete r.v., is the combination of two lists. Informally, we have

Definition 7.1.1 — R.V. to be continued. A random variable is called *continuous* when it is specified with:

- one or more intervals of the real line that compose the set of possible values.
- a probability density function $f(x)$ defined on this interval.

The only constraint is that the density $f(x)$ must be greater than or equal to zero for all x and its integral over the range of possible values must be equal to 1. The set of points x where $f(x) > 0$ é called the *support set* of the density $f(x)$.

Consider the graph of the function $f(x)$ in the figure 7.1. It has $f(x) > 0$ between -5 and 15 and so, the support of this $f(x)$ is $(-5, 15)$. Outside this range, $f(x) = 0$. Furthermore, the integral of $f(x)$ in the interval $(-5, 15)$ is 1. That is, the total area under the curve $f(x)$ between $(-5, 15))$ is equal to 1. In this way, the function $f(x)$ can represent the probability density of a continuous r.v. X that can take any real value in the real range $(-5, 15)$.

In the continuous case, probabilities are associated with areas under the density function. Consider the event $B = [\{\omega \in \Omega : X(\omega) \in (a,b)\}]$. We want to obtain the probability $\mathbb{P}(B)$ of the event B occurrence. In general, we shorten the event B representation by writing

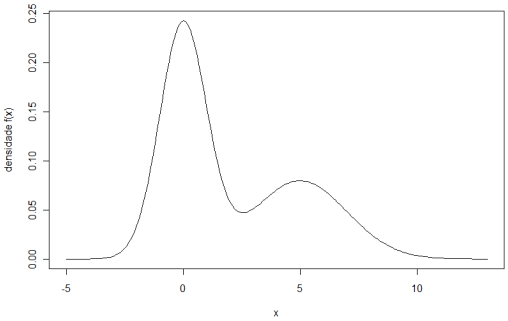


Figure 7.1: Example of a probability density function $f(x)$. It is greater than or equal to zero and its area under the curve is equal to 1.

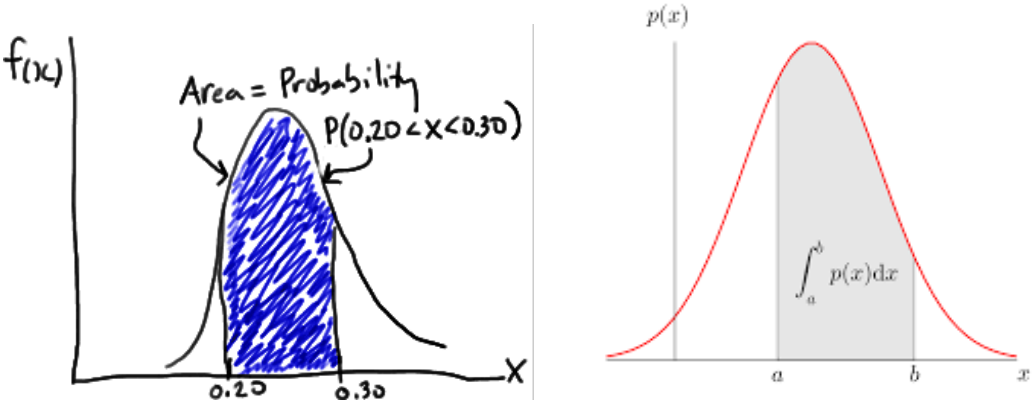


Figure 7.2: Probabilities are areas under the density function $f(x)$.

only $B = [X \in (a,b)]$ and we get

$$\mathbb{P}(X \in (a,b)) = \int_a^b f(x)dx$$

Figure **??** shows how probabilities are calculated in the continuous case. The usual rules of probabilities are valid, of course. For example, the event that X belongs to two disjoint intervals, such as $[X \in (1,2) \text{ or } (4,5)]$, is the union of the two *disjoint* events, $B_1 \cup B_2$ where $B_1 = [X \in (1,2)] = [\{\omega \in \Omega : X(\omega) \in (1,2)\}]$ e $B_2 = [X \in (4,5)] = [\{\omega \in \Omega : X(\omega) \in (4,5)\}]$. The events B_1 and B_2 are disjoint since $\omega \in B_1 \cap B_2$ means that $X(\omega)$ belongs at the same time to the intervals $(1,2)$ and (4.5). So,

$$\mathbb{P}(X \in (1,2) \text{ or } (4,5)) = \mathbb{P}(X \in (1,2)) + \mathbb{P}(X \in (4,5))\,.$$

Looking at the graph of $f(x)$, we know which regions of the real line have the highest probability: they are those regions that have greater area under the curve $f(x)$. The intuitive idea is that the function $f(x)$ shows how the total probability of occurrence of an event (which is equal to 1) was distributed on the real axis indicating which regions are most likely to produce a result for X (the regions with the highest $f(x)$) and which regions have a small probability of generating a value of X (regions where $f(x) \approx 0$).

Returning to the density in Figure 7.1, consider the four intervals $(-5,-2.5)$, $(-2.5,0)$, $(5,7.5)$ and $(7.5,10)$, all of equal length. Which is more likely to occur? That is, comparing the probabilities that X comes from each of these intervals, which one is has the largest probability?

- $\mathbb{P}(X \in (-5,-2.5))$
- $\mathbb{P}(X \in (-1.0,1.5))$
- $\mathbb{P}(X \in (5,7.5))$
- $\mathbb{P}(X \in (7.5,10))$

We must look at the area under $f(x)$ in each of these intervals. In this case, clearly $\mathbb{P}(X \in (-1.0,1.5))$ is the highest probability among these four while $\mathbb{P}(X \in (-5,-2.5))$ is the smallest.

Figure **??** shows examples of some of the main probability densities of continuous r.v.'s. Going from left to right, we see examples of densities of beta, Cauchy, gamma (top row) and normal, Pareto and chi-square (bottom row) distributions. Each curve é is a density obtained by varying the parameters of these distributions.

These distributions will be studied in more detail in the final sections of this chapter. For now, just observe the diversity of shapes that can be obtained with these basic distributions. However, also notice how none of them seem to resemble the density shown in Figure **??** where the density has two humps. None of the densities shown in Figure **??** has this property of having two maximum

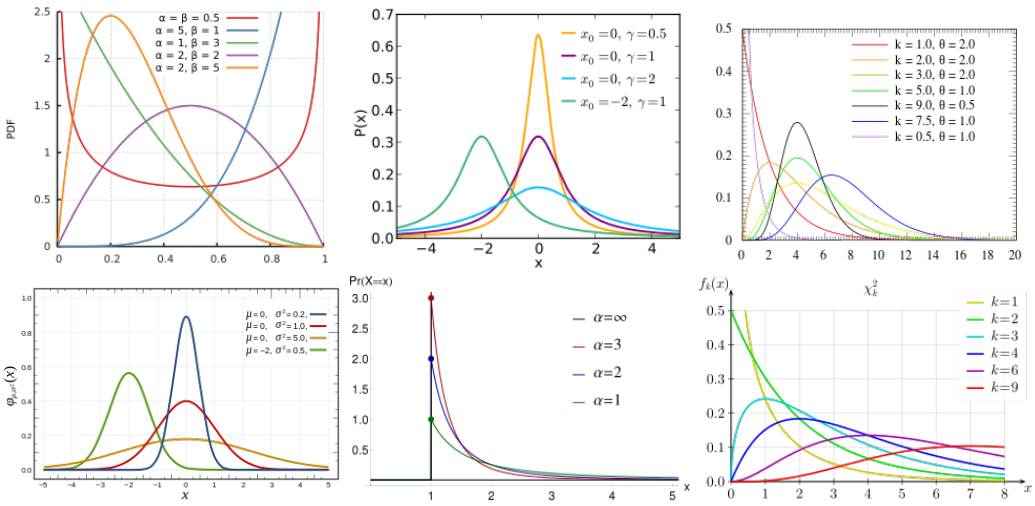


Figure 7.3: Examples of some of the main probability densities of continuous r.v.'s: beta, Cauchy, gamma (top row) and normal, Pareto, and chi-square (bottom row). Each curve is the density function obtained by varying the parameters of the distributions.

points. Thus, the collection of basic distributions is rich but not enough to cover all situations that appear in practice. However, in a way, they are enough. As we will see in chapter ??, we can generate virtually any desirable density by mixing the basic distributions. For example, the density of Figure **??** can be very well approximated by making a mixture of two Gaussian densities. Hence, the study of these basic distributions is important as its properties will be transfered for these mixed distributions later on.

7.2 Density is approximated by the histogram

Even when the variable is not continuous, an approximation with a continuous distribution might be useful. Imagine a sample of 5000 lots that make up a farm and where only soy is grown. Let y_i be the harvest from lot i. It is impractical and somewhat meaningless to work with a discrete distribution in a situation like that. Even if we are only interested in the 5000 lots (a finite number of values), it is more useful to assume that the lots' harvests are the results of 5000 realizations of a certain *continuous* random variable that has a simple and already known shape.

What would be the density of this v.a. Y? To know this, we make a standardized histogram (with total area equal to 1) as in Figure 7.4. Break the horizontal axis into small intervals of length Δ. In each small range i, count the number n_i of elements in your sample that fell in the range. Raise a bar whose height is equal to this count. This is the non-standard histogram, which is in the left hand side in Figure 7.4. The standardized histogram has a total area equal to 1. To do this, just divide each bar by the constant value $n\Delta$ producing bars with height $= n_i/(n\Delta)$. The result is in the central graph in Figure 7.4. On the standardized histogram, overlay a candidate density. The histogram looks like a certain density of a so-called Gaussian (or normal, $N(9,4)$) distribution whose density function is the curve in red. This means that the actual distribution will be *approximated* by this normal distribution. We will see later how to choose a candidate distribution and to test if it fits the data well.

However, it is important to keep in mind that the basic distributions that we know and that are used to model the observed data do not cover the entire possible spectrum of histograms we observe with empirical data. Most of the time, a distribution chosen to model the data, however well-fitted to the data it appears, is almost certainly not the true and unknown distribution that generated the

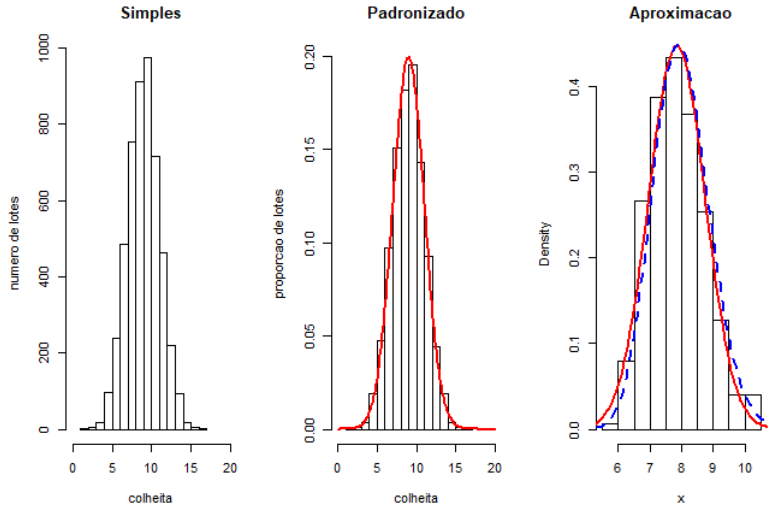


Figure 7.4: Simple and standardized histogram. Histogram with the actual density $f^*(x)$ in blue dashed line and an approximate density using the Gaussian or normal distribution (red line).

data. For example, in the leftmost graph in Figure ??, we see a histogram with two probability densities. The one in the blue dashed line is the true probability density $f^*(x)$ associated with this data. I know this as a fact because I generated the data in the computer using techniques to be seen in chapter ?? and using a density $f^*(x)$ that I chose it myself. The curve in red is $f(x)$, a normal or Gaussian density fitted to the histogram data. Although $f^(x) \neq f(x)$, the fit is is very good, with the blue and red curves being very similar ($f^*(x) \approx f(x)$ for all x). In practice, the unknown density $f^*(x)$ in blue is approximated by one of the known densities (the curve $f(x)$ in red). If the approximation is good, the probability calculations using the $f(x)$ approximation will be approximately equal to the results we would have if had we used $f(x)$ instead.

Looking at the histogram is a practical way to find a candidate density. After knowing the main distributions, we will have a certain collection of possibilities to choose from. To propose a candidate density, we look at the histogram and try to find a match with the density shapes that we know. Of course, the histogram can have a very strange shape, which does not resemble the shapes of the basic distributions. In those cases where none of our known densities seem to fit the histogram, we can appeal to mixtures of these basic distributions (see chapter ??).

But let's see first how to look at the basic distributions in the histograms. Figure 7.5 shows the standardized histogram of samples of size 1000 composed of computer-simulated data. As they were simulated, we are sure about the generating distribution in each case. Starting from the top row and going from left to right, these distributions were exponential, log-normal, uniform, and beta, respectively. The red lines are the respective probability densities. We can see that, in fact, histograms have the same format as densities. That is, looking at the histogram should suggest the shape of the probability density.

In practice, we can guess what is the red lines as in Figure 7.5but we will not know that for sure. This is so because we do not know the true probability density that will have generated the data that we are analyzing. Knowing that by looking at the histogram gives a good indication to what is this unknown density is an easy way to learn directly from the data the hidden mechanism that is producing the observed data.

What is the theoretical justification for this similarity between the histogram shape and the unknown probability density? Let $f^*(x)$ be the true density that generated the data. In general, $f^*(x)$ is unknown. Let $f(x)$ be the density of a distribution taken from our limited catalogue of

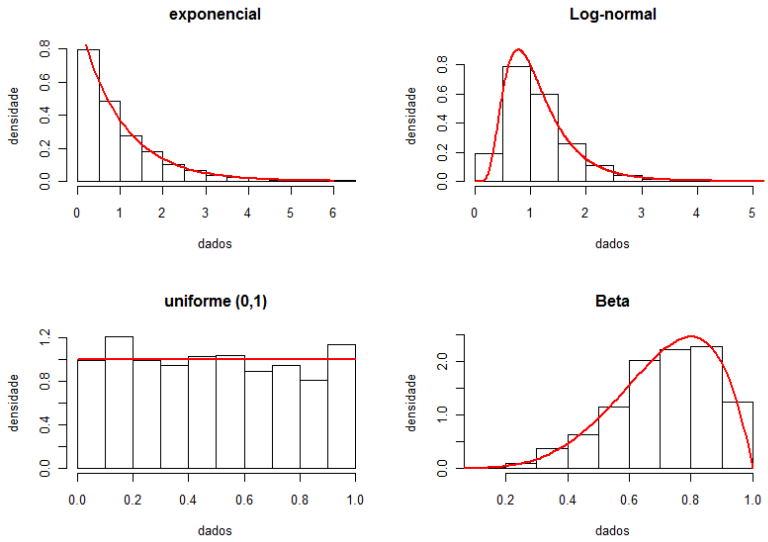


Figure 7.5: Standardized histograms of samples of distributions whose densities are shown with the red line.

known distributions, which have their own names and that are well studied. If the sample histogram is well approximated by $f(x)$ then we believe that $f(x) \approx f^*(x)$ Why?

To answer this question, let $(y_0 - \delta/2, y_0 + \delta/2)$ be one of the small intervals of the horizontal axis in the histogram, that interval centered on y_0 and of small length δ. Consider the probability that the random variable Y is observed in the range $(y_0 - \delta/2, y_0 + \delta/2)$. This probability is the area under the unknown and true density $f*(x)$ given by

$$\mathbb{P}(Y \in (y_0 - \delta/2, y_0 + \delta/2)) = \int_{y_0-\delta/2}^{y_0+\delta/2} f^*(y)\, dy$$

and illustrated in the left hand side left in Figure 7.6. We can approximate this probability or area by the rectangle area with base δ centered on y_0 and height $f^*(y_0)$:

$$\mathbb{P}(Y \in (y_0 - \delta/2, y_0 + \delta/2)) = \int_{y_0-\delta/2}^{y_0+\delta/2} f^*(y)\, dy \approx f^*(y_0)\delta\,. \tag{7.1}$$

This rectangle is on the right hand side of Figure 7.6.

Let's now use the idea of estimating a probability by the proportion of times the event happens in a large number of independent repetitions. We have not yet defined precisely what *independent* random variable means but we will go on with its intuitive idea. The same probability $\mathbb{P}(Y \in (y_0 - \delta/2, y_0 + \delta/2)$ can also be approximated by the fraction of elements Y_i from sample size n that fell in the range $(y_0 - \delta/2, y_0 + \delta/2)$:

$$P(Y \in (y_0 - \delta/2, y_0 + \delta/2) \approx \frac{\#\{Y_i's \in (y_0 - \delta/2, y_0 + \delta/2)\}}{n} \tag{7.2}$$

Equating the two approximations for the probability $P(Y \in (y_0 - \delta/2, y_0 + \delta/2)$, that in (7.1) with that in (7.2), we have

$$\frac{\#\{Y_i's \in (y_0 - \delta/2, y_0 + \delta/2)\}}{n\delta} \approx f^*(y_0)$$

The left hand side is the histogram height at the point y_0. The right side is the density curve height at the same point y_0. Thus, the height of the histogram at the midpoint y_0 of one of the intervals is

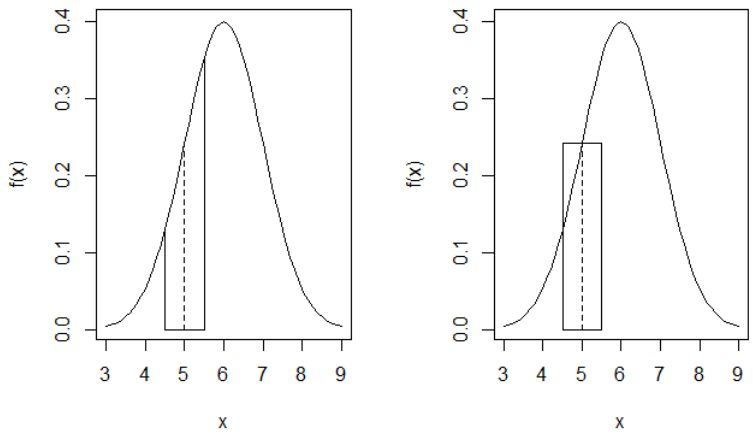


Figure 7.6: Left: area under the curve in the range $(4.5, 5.5)$ is the probability $\mathbb{P}(Y \in (5.0-0.5, 5.0+0.5))$. Right: rectangle with base $(4.5, 5.5)$ and height $f^*(5.0)$.

approximately equal to the density $f^*(y_0)$. By looking at the histogram and following its peaks and valleys is essentially the same as looking at the unknown density.

7.3 $\mathbb{F}(X)$ in the continuous case

We have already defined the cumulative probability function $\mathbb{F}(x) = \mathbb{P}(X \leq x)$. In the continuous case, the function $\mathbb{F}(x)$ associates to each value x of the real line the entire area under the density curve in the range $(-\infty, x]$.

Definition 7.3.1 If X is a continuous random variable with probability density $f(x)$, the function cumulative probability function is

$$\mathbb{F}(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(u)du\,.$$

for $x \in \mathbb{R}$.

$\mathbb{F}(x)$ is defined for the entire x of the real line, not just for the points x of the density support (the set of points in which $f(x) > 0$).

Spend some time studying the Figure **??**. On the left hand side, it shows two curves, $f(x)$ and $\mathbb{F}(x)$, in two separate plots. On the top, we have the density $f(x)$. On the bottom, the cumulative distribution function $\mathbb{F}(x)$. For any arbitrary point x on the line, we get the entire area under the density $f(x)$ up to point x. This is the shaded area in the upper plot and it equals $\mathbb{P}(X \leq x) = \mathbb{F}(x)$. As the total area under $f(x)$ is 1, for any x we have a value between 0 and 1 for $\mathbb{F}(x) = \mathbb{P}(X \leq x)$. This value is placed in the chart from below as the *height* of the curve $\mathbb{F}$ at the point x. Notice how the $\mathbb{F}(x)$ height varies as we slide the x value from the far left. $\mathbb{F}(x)$ starts out at almost zero, reflecting the fact that there is almost no area to the left of a point x on the far left. As we move to the right with x, we accumulate the left hand side area on the $f(x)$ graph and the height of $\mathbb{F}(x)$ increases towards the value 1, which is the total area under density $f(x)$. The graph on the right in Figure 7.7 shows the two curves, $f(x)$ e and $\mathbb{F}(x)$, in the same plot. The scales of the two curves are different and are shown on the two vertical axes. The $f(x)$ values are read on the left vertical axis, while the values of $\mathbb{F}(x)$ are read on the right vertical axis.

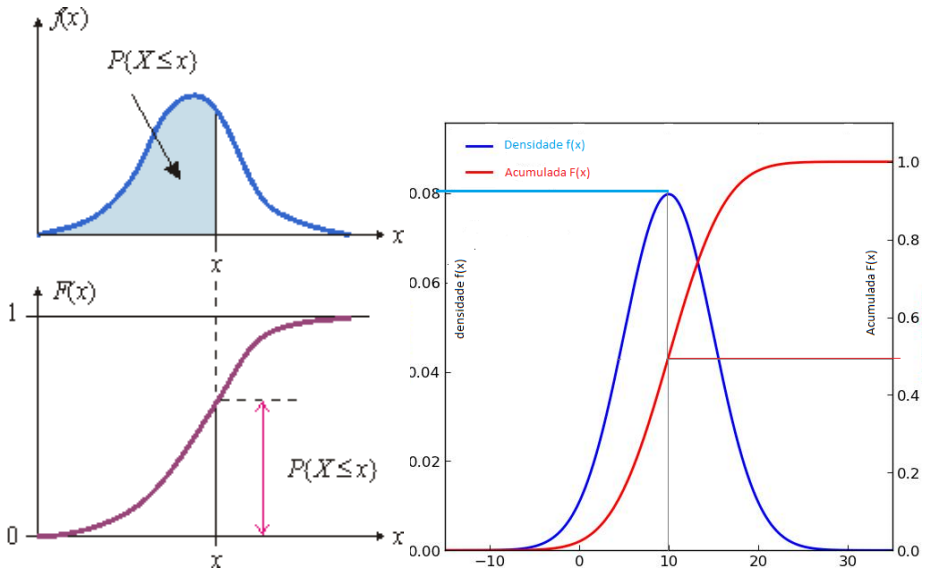


Figure 7.7: Probability density curve $f(x)$ and its cumulative distribution function $\mathbb{F}(x)$.

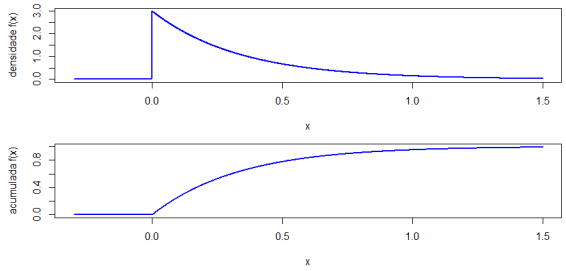


Figure 7.8: Density function $f(x)$ and correspondent cumulative distribution function $\mathbb{F}(x)$.

We can reverse the relationship shown in Definition 7.3.1 which goes from $f(x)$ to $\mathbb{F}(x)$.Using the Fundamental Theorem of Calculus, we have:

Proposition 7.3.1 If X is a continuous r.v. and $\mathbb{F}(x)$ is differentiable at x then $f(x) = \mathbb{F}'(x)$.

■ **Example 7.1 — Density and Cumulative Functions.** Suppose X is a continuous r.v. with support $\mathscr{S} = (0,\infty)$ and density $f(x) = 3e^{-3x}$, for $x > 0$. For $x \leq 0$, and therefore out of support, we have $f(x) = 0$. See the upper plot in Figure 7.8.

Starting with the simplest case, the one where $x \leq 0$, by the definition, we have

$$\mathbb{F}(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(t)\, dt = \int_{-\infty}^{x} 0\, dt = 0\,.$$

For an arbitrary $x > 0$ in the support set, we have

$$\mathbb{F}(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(t)\, dt = \int_{-\infty}^{0} 0\, dt + \int_{0}^{x} 3e^{-3t}\, dt = 1 - e^{-3x}\,.$$

For $x > 0$, if we take the derivative of $\mathbb{F}(x) = 1 - e^{-3x}$ we find $\mathbb{F}'(x) = 3e^{-3x}$, which is exactly $f(x)$. For $x < 0$, the derivative of $\mathbb{F}(x) \equiv 0$ is $\mathbb{F}'(x) = 0$, which is again the density value for $x < 0$.

■

Theorem 7.3.2 Let's write $X \sim \mathbb{F}$ or $X \sim f$ to denote that the r.v. has the cumulative distribution function $\mathbb{F}$ or the probability density function f.

7.4 $\mathbb{E}(X)$ in the continuous case

We have already studied the definition and the empirical meaning of the expected value $\mathbb{E}(X)$ of a discrete r.v. X in sections 6.4.4 and 6.4.5, respectively. Assume that X is a discrete r.v. whose possible values are those in the support set $\mathscr{S} = \{a_1, a_2, \ldots\}$. Then

$$\mathbb{E}(X) = \sum_i a_i \mathbb{P}(X = a_i)\,.$$

We have a large random sample $\{X_1, X_2, \ldots, X_n\}$ of size n of this discrete r.v. The semantic connection between the definition and the sample with n randomly observed values in this discrete case $\mathscr{S}$ is that the theoretical value $\mathbb{E}(X)$ should be approximately equal to the arithmetic mean or average $\overline{X} = (X_1 + \ldots + X_n)/n$ of the n elements in the sample.

Definition 7.4.1 — $\mathbb{E}(X)$ in the continuous case. Let X be a v.a. continuous with support set $\mathscr{S}$ on the real line and probability density $f(x)$. The expected value of X is equal to

$$\mathbb{E}(X) = \int_{\mathscr{S}} x f(x) dx = \int_{\mathbb{R}} x f(x) dx\,.$$

The two integrals are the same because $f(x) = 0$ for $x \notin \mathscr{S}$.

Occasionally, for some distributions, this integral may not exist or not be defined. In these cases, the r.v. does not have an expected value. For the most common distributions in the practice of data analysis, this expected value exists without any problem.

The continuous case is the discrete version pushed to the limit. We can reason intuitively just as in the discrete case. Partition the entire real line axis into small intervals (or bins) of length Δ centered on $\ldots, x_{-2}, x_{-1}, x_0, x_1, x_2, \ldots$. In each small bin $(x_i - \Delta/2, x_i + \Delta/2)$, the integral of any function $h(x)$ is approximated by the rectangle with base Δ and height $h(x_i)$:

$$\int_{\text{bin}_i} h(x) dx = \int_{x_i - \Delta/2}^{x_i + \Delta/2} h(x) dx \approx h(x_i)\Delta\,.$$

In fact, the Riemann integral is *defined* by taking the sum over all bins and taking it to the limit as the length of each bin goes to zero and the number of bins goes to infinity :

$$\int_{\mathbb{R}} h(x) dx = \lim_{\Delta \to 0} \sum_i h(x_i)\Delta\,.$$

Figure 7.9 shows a red curve representing the function $h(x)$. The area under the curve is the integral $\int h(x)dx$ and it is approximated by the sum of the rectangles. Each rectangle has a base of length Δ and height $h(x_i^*)$, equal to the value of the function $h(x)$ at the midpoint x_i^* of the rectangle basis.

In the case of the expectation of a continuous r.v. with density $f(x)$, if we make the generic function $h(x)$ equal to $xf(x)$ (that is, $h(x) = xf(x)$) then the above approximation is

$$\int_{\text{bin}_i} x f(x) dx = \int_{x_i - \Delta/2}^{x_i + \Delta/2} x f(x) dx \approx x_i f(x_i)\ Delta$$

But we also have an approximation of probabilities as areas under the density curve so that

$$\mathbb{P}(X \in (x_i - \Delta/2, x_i + \Delta/2)) = \int_{x_i - \Delta/2}^{x_i + \Delta/2} f(x) dx \approx f(x_i)\Delta$$

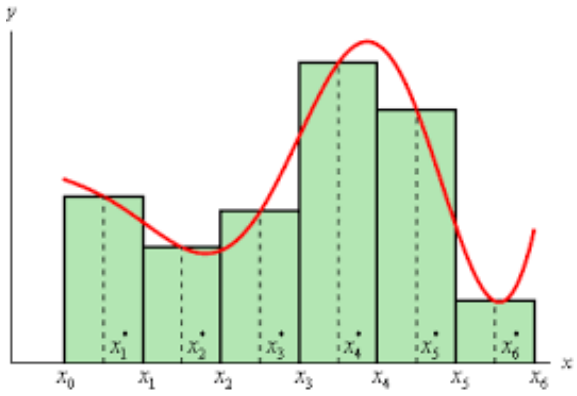


Figure 7.9: Integral $\int h(x)dx$ where $h(x)$ is the red curve. This integral is approximated by the sum of the areas of rectangles with base Δ and height equal to the value of $h(x)$ at the center point x_i^* of the rectangle basis.

Thus, we can obtain the following approximation for $\mathbb{E}(X)$ in the continuous case:

$$\begin{aligned}
\mathbb{E}(X) &= \int_{-\infty}^{\infty} xf(x)dx \\
&= \ldots + \int_{x_{-1}-\Delta/2}^{x_{-1}+\Delta/2} xf(x)dx + \int_{x_0-\Delta/2}^{x_0+\Delta/2} xf(x)dx + \int_{x_1-\Delta/2}^{x_1+\Delta/2} xf(x)dx + \int_{x_2-\Delta/2}^{x_2+\Delta/2} xf(x)dx + \ldots \\
&\approx \ldots + x_{-1}f(x_{-1})\Delta + x_0 f(x_0)\Delta + x_1 f(x_1)\Delta + x_2 f(x_2)\Delta + \ldots \\
&\approx \ldots + x_{-1}\mathbb{P}(X \in \text{ bin}_{-1}) + x_0\mathbb{P}(X \in \text{ bin}_0) + x_1\mathbb{P}(X \in \text{ bin}_1) + x_2\mathbb{P}(X \in \text{ bin}_2) + \ldots
\end{aligned}$$

This last expression is equal to the expectation of a discrete r.v. that assumes the possible values $\ldots, x_{-2}, x_{-1}, x_0, x_1, x_2, \ldots$ with probabilities $\mathbb{P}(X \in \text{ bin}_i)$. That is, the definition $\int xf(x)dx$ of the expectation in the continuous case is just the same expression of the discrete case taken to the limit.

7.5 Uniform Distribution

From this section, we are going to build our small catalog of continuous distributions. We are going to present some of the most important continuous distributions with some very simple illustrations about their use in the practice of data analysis. We start with the simplest of all, the Uniform distribution.

Definition 7.5.1 — Uniform Distribution. The uniform distribution over an interval (a,b), with $a < b$, is defined by the density

$$f(x) = \frac{1}{(b-a)}$$

if $x \in (a,b)$. See Figure 7.10. The support set of this distribution is the interval $\mathscr{S} = (a,b) \subset \mathbb{R}$. In case we want to extend the definition of $f(x)$ to all $x \in \mathbb{R}$, we can define the density so that it is equal to zero outside the support set:

$$f(x) = \begin{cases} \frac{1}{(b-a)}, & \text{if } x \in [a,b] \\ 0, & \text{otherwise} \end{cases}$$

Notation 7.1. *If X has a uniform distribution in the range (a,b) we write $X \sim U(a,b)$.*

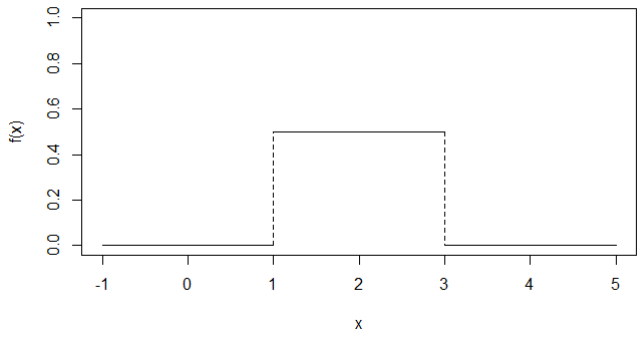


Figure 7.10: Density function of the Uniform(1,2) probability distribution.

The constant height $1/(b-a)$ of uniform density $f(x)$ for x in the support set (a,b) depends on the length $b-a$ of the support interval. For example, if $X \sim U(3,10)$ then $f(x) = 1/7$ for $x \in (3,10)$.

The most famous uniform distribution is $X \sim U(0,1)$, defined in the range $(0,1)$. In this case, $f(x) = 1$ for $x \in (0,1)$. The probability that $X \sim U(0,1)$ falls in an interval (a,b) contained in $(0,1)$ is the area under uniform density :

$$\mathbb{P}(X \in (a,b)) = \int_a^b f(x)\,dx = \int_a^b 1\,dx = b-a = \text{ length of } (a,b).$$

Thus, in the case of $U(0,1)$ the probability of an interval (a,b) contained in $(0,1)$ is its length. For example, $\mathbb{P}(X \in (1/2,1)) = 1/2$ and $\mathbb{P}(X \in (0.75,0.78)) = 0.03$.

The uniform distribution $U(0,1)$ is crucial in the Monte Carlo simulation methods (see chapter 11). Algorithms, called random number generators, provide a succession of values that are similar in their statistical behavior This is the successive and independent observations obtained from a distribution $U(0,1)$. For example, if we generate a large number of values $U(0,1)$ using the `runif` function in R we get the picture below. It presents the relative frequency f_i among 100 thousand random numbers generated by R that belong to the range $I_i = (0.i, 0.(i+1)] \quad i = 0,\ldots,9$. For example, f_5 is the proportion of values that fall in $I_5 = (0.5, 0.6]$. See how the proportion of numbers in each interval is approximately 0.1, as it really should be if the data follows a $U(0,1)$. The code used is below.

```
> 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
```

The expectation of a r.v. $X \sim U(a,b)$ with a uniform distribution in (a,b) is easily obtained:

$$\mathbb{E}(X) = \int_a^b x\,\frac{1}{b-a}\,dx = \frac{1}{b-a}\int_a^b x\,dx = \frac{1}{b-a}\left(\frac{b^2}{2} - \frac{a^2}{2}\right) = \frac{(b-a)(b+a)}{2(b-a)} = \frac{a+b}{2}$$

The expectation of a r.v. $X \sim U(a,b)$ is the midpoint of the (a,b) support set range.

■ **Example 7.2** If $X \sim U(0,1)$ then $\mathbb{E}(X) = 1/2$. If $X \sim U(90,100)$ then $\mathbb{E}(X) = 95$. ■

The cumulative probability distribution function is also very easy to be obtained:

$$\mathbb{F}(x) = \mathbb{P}(X \leq x) = \begin{cases} 0, & \text{if } x < a \\ x/(b-a), & \text{if } a \leq x \leq b \\ 1, & \text{if } x > b \end{cases}$$

■ **Example 7.3** The uniform distribution can serve to approximate a continuous distribution on a small interval. For example, actuaries often approximate the distribution of an individual age at death using this trick. Suppose it is somehow known that an individual has completed his birthday of k years but died before reaching age $k+1$. Let X be the exact (continuous) age at death of this individual. Note that X is the exact age at death *conditioned* on the fact that $k < X < k+1$. We can approximate the conditional distribution of X in the interval $[k, k+1)$ assuming that the moment of death throughout the year occurs according to a uniform distribution: $X \sim U(k, k+1)$. That is, he can die at any time throughout the year with equal chance. No day or month throughout the year of death would be more likely to happen. This is just an approximation but it usually allows several calculations that would otherwise be impossible. See [7]. ■

■ **Example 7.4 — Measuring paper thickness.** This example is from the classic Yule book [37]. Five hundred specimens of a new type of paper under test were covered with a specific liquid polymer and then dried and distributed according to their weights in five categories. The thickness of the papers in the central category was measured to determine if they were evenly distributed. The thickness of each sheet was calculated by taking an arithmetic average of five points, one in the center and in the 4 corners. The data in the table below represent the thickness of the 116 specimens in this central class. The measurement is in coded units, representing downward or upward deviations in relation to a norm.

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

Figure 7.11 is a histogram of the 116 thickness values from the table above along with the graph of the constant function $f(x) = 1/12$ which would be the probability density of a uniform distribution. The expected value of the height of a histogram bar would be the red line if the distribution of X were uniform. To test whether the hypothesis that the distribution is not uniform, we need techniques such as the chi-square test, to be taught in Chapter ??. Advancing this topic, we can say that, despite the apparent discrepancy between the observed data and the uniform distribution, the data are perfectly compatible with a uniform distribution. That is, there is no evidence in the data that the uniform distribution did not generated these data. The differences between the bars and the theoretical line are perfectly natural and plausible for a sample of size 116. The code to this barplot is below.

```
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
```

■

7.6 Beta Distribution

There are several phenomena whose variables of interest have their values limited above and below by known numbers a and b. A typical example constitutes data appearing in the form of a random proportion:

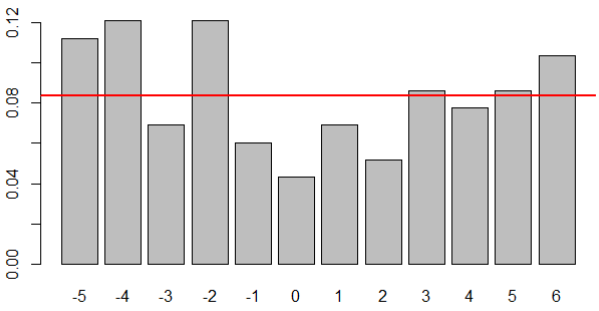


Figure 7.11: Barplot of the proportion of papers that fell into different intervals of the same length according to their average thickness.

- in each company, we measure the proportion of monthly expenditure on wages with respect to the total expenditure. We can also record, for example, the proportion of energy expenditure in the total production expenditure in each month.
- the ratio between the femur length and the total length of an individual's leg.

Note that these proportions assume values in the continuous range $[0,1]$. They differ from other ratios based on counts such as the ratio of successes in 10 tosses of a coin. In this latter case, the proportion always assumes one of the values $\{0.0, 0.1, 0.2, \ldots, 0.9, 1.0\}$ and thus the proportion is a discrete random variable. Compare this discrete case with the continuous situation in the examples above.

A distribution class that is rich enough to provide models for many random variables with values constrained into a finite length interval is the beta distribution. The standard beta distribution has its values concentrated in the range $[0,1]$.

Definition 7.6.1 — Standard Beta Distribution. The random variable X has a standard beta distribution with parameters α and β where $\alpha > 0$ and $\beta > 0$ if X has density given by

$$f(x) = Cx^{\alpha-1}(1-x)^{\beta-1}$$

for $x \in [0,1]$. The constant C is such that the density function integrates to 1 into the interval $[0,1]$ and it is equal to $C = \gamma(\alpha + beta)/(\Gamma(\alpha)\Gamma(\beta))$.

Notation 7.2. *If X has a beta distribution in $[0,1]$ with parameters $\alpha > 0$ and $\beta > 0$ we write $X \sim Beta(\alpha, \beta)$.*

For example, with $\alpha = 3$ and $\beta = 5$ we have $f(x) = Cx^2(1-x)^4$ (see left graph in Figure 7.12). With $\alpha = 8$ and $\beta = 2$, we have $f(x) = C^* x^7(1-x)$ (see right graph in Figure 7.12). We write the constant as C^* in this last example just to emphasize that the constant in the first case is different from the constant in the second case. Based on the Beta(3.5) density function, the region that concentrates the most probability is quite wide, ranging from 0.1 to 0.7. Considering the other Beta$(8,2)$ density function, it is well concentrated in a much narrower band, shifted to the upper end of the range, between 0.6 and 1.0.

Analysing the expression of the beta distribution density function, we see that $f(x) = Cx^{\alpha-1}(1-x)^{\beta-1}$ is the result of the product of two monomials, the first being $x^{\alpha-1}$ and the second being $(1-x)^{\beta-1}$. For example, with with $\alpha = 3$ and $\beta = 5$ we have $f(x) = Cx^2(1-x)^4$ being obtained

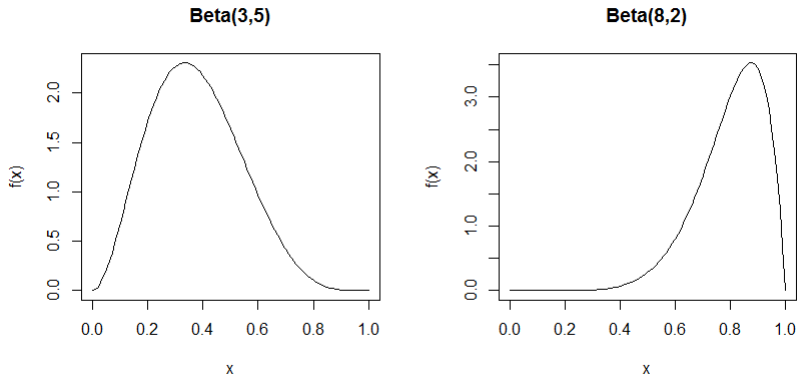


Figure 7.12: Probability density function of the Beta(3.5) (left) and Beta(8.2) (right) distributions.

with the product of x^2 by $(1-x)^4$. Note that the first factor, x^2, increases with x, while the second factor, $(1-x)^4$, decreases with x. As the two factors reach zero at one end, their product é zero at $x = 0$ and at $x = 1$. Density í the product of these two factors and it will go up and then down along the range $(0,1)$. Its maximum point and the shape of its rise and fall are controlled by the parameters α and β. This reasoning í valid as long as $\alpha > 1$ and $\beta > 1$. The other cases, with values of $\alpha \leq 1$ or $\beta \leq 1$, will be discussed shortly.

The probability density constant C is obtained such that the area below $f(x)$ is equal to 1:

$$\begin{aligned} 1 &= \int_0^1 Cx^{\alpha-1}(1-x)^{\beta-1}dx \\ &= C\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx \end{aligned}$$

from which we conclude that

$$C = \frac{1}{\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx}$$

It can be shown that this integral in the denominator can be expressed in terms of a mathematical function known as the gamma function and denoted by $\Gamma(z)$:

$$C = \frac{1}{\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \tag{7.3}$$

The gamma function $\Gamma(z)$ is a generalization of the factorial of a positive integer. When z is a positive integer, we have $\Gamma(z) = (z-1)!$. For z between two successive integers k and $k+1$, we have $\Gamma(k) = (k-1)! < \Gamma(z) < k! = \Gamma(k+1)$. The function $\Gamma(z)$ interpolates smoothly between the factorials of integers. For example, for any $z \in (4,5)$, we have $\Gamma(4) = 3! < \Gamma(z) < 4! = \Gamma(5)$. Although not relevant in this book, the definition of the gamma function is as follows:

$$\Gamma(z) = \int_0^\infty y^{z-1}e^{-y}dy\,. \tag{7.4}$$

The other fundamental property of the gamma function is that $\Gamma(z+1) = z\Gamma(z)$ for all $z > 0$. It is implemented in R and can be called with the command `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
```

In general, the integral in the denominator of 7.3) must be obtained numerically unless α and β are positive integers. In this particular case, we have

$$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!}.$$

The beta distribution includes the uniform distribution because, if $\alpha = 1$ and $\beta = 1$, we have $f(x) = Cx^{1-1}(1-x)^{1-1} = C$ to $x \in (0,1)$. Since the integral in $(0,1)$ must be 1, we have $C = 1$. Thus, $f(x) = 1$ for $x \in (0,1)$, which is the uniform distribution density function.

When $0 < \alpha < 1$ or $0 < \beta < 1$ we have a density with a special format. In this case, the density goes to infinity at least in one of the two extremes, 0 or 1. For example, if $\alpha = 1/2$ and $\beta = 4$ we have $f(x) = C(1-x)^3/\sqrt{x}$. When $x \to 0$ we have $f(x) \to \infty$. Despite this, the total area under the function $g(x) = (1-x)^3/\sqrt{x}$ is finite and therefore the constant C can be calculated and a density $f(x)$ does exist with the choices of $\alpha = 1/2$ and $\beta = 4$.

The Figure 7.13 shows the great diversity of forms taken by the density function $f(x)$ of the beta distribution as we vary the parameters $\alpha > 0$ and $\beta > 0$. The curves in the left upper plot are examples with $\alpha > 1$ and $\beta > 1$. They have a single well-defined maximum at the position $x = \alpha/(\alpha+\beta)$. There is a symmetry in the parameters α and β. For example, the density curves of Beta$(15,5)$ and Beta$(5,15)$ are the mirror reflection of each other around the center point $x = 1/2$. The curves in the right upper plot are examples with their maximum point fixed at the position $\alpha/(\alpha+\beta) = 0.75$ but with both parameters growing. The effect of making α and β grow while keeping the ratio $\alpha/(\alpha+\beta)$ fixed is to make the density increasingly concentrated around its maximum point. The density curves at the left lower plot are examples with $\alpha = \beta$. Now, the densities are symmetric around their maximum at $x = 1/2$. Finally, the curves in the right lower plot are examples with one of the parameters smaller than 1 (but positive) and also the case $\alpha = \beta = 1$, which is equivalent to the $U(0,1)$ uniform distribution. In the case where $\alpha = \beta = 1/2$, the curve asymptotes towards infinity at both ends of the range $(0,1)$ and has a more or less uniform shape over most of the middle of the range. This $\alpha = \beta = 1/2$ distribution is very important in several modern Bayesian models such as the *Latent Dirichlet Allocation* model for text processing. The beta distribution with only one of the parameters smaller than 1, like Beta$(0.5,5)$ in this graph, has an asymptote for ∞ at $x = 0$ and decreases to zero when x grows to 1. The case $\beta < 1$ and $\alpha > 1$ follows a mirrored symmetry. The R code for this figure is below.

```
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");
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"))
```

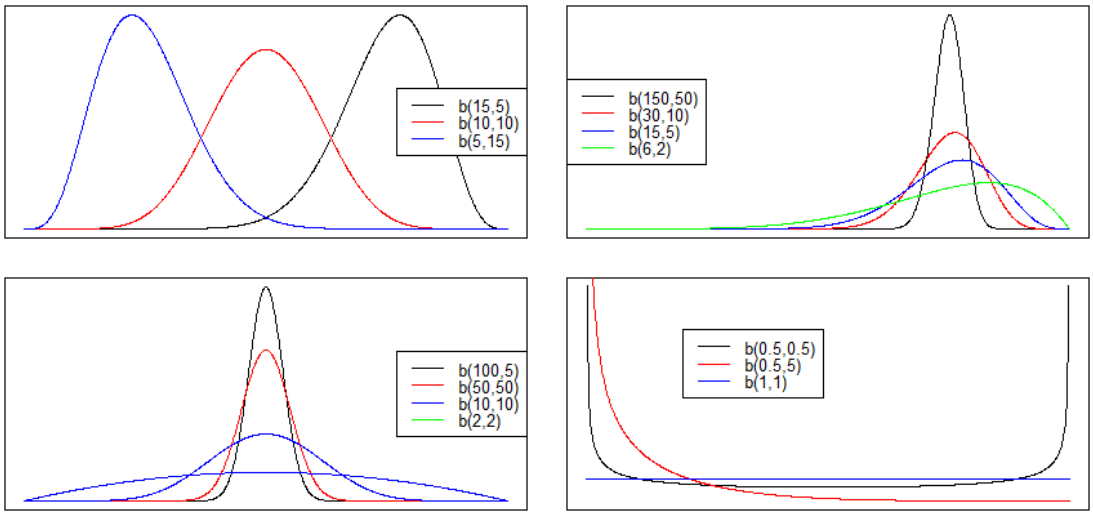


Figure 7.13: Graphs of the density function Beta(α,β) varying the parameters α and β. Right, above: examples with $\alpha > 1$ and $\beta > 1$; Left, above: effect of making α and β grow while keeping the $\alpha/(\alpha+\beta)$ ratio fixed; Right, bottom: $\alpha = \beta > 1$; Left, bottom: $\alpha < 1$.

```
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 — Stock Market.** Let's represent by X the random variable that measures the proportion of the variation in the daily average price of shares when these prices fall. That is, if the average share price in a given day is p_1 and $p_2 < p_1$ in the next day, then we calculate $X = (p_1 - p_2)/p_1$. If the price goes up from one day to the next, the X r.v. is not defined. A total of 2314 stocks with prices that dropped overnight were observed. The continuous data were grouped into bins and are summarized in the form of a frequency distribution f_i in the table below. Figure 7.14 presents a histogram of these data together with the density of a beta distribution with parameters $\alpha = 1.04$ and $beta = 10.63$. We use the method of moments to set the values for these parameters and this method is the subject of Chapter **??**. From a purely visual point of view, it seems that the beta distribution with these parameters provides a reasonable model for the variable X since the density of the beta fits reasonably well to the histogram.

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

The R code to produce this figure is below. There is a trick to placing a line graph or dots over a barplot in R. Saving the return value of the command `barplot` we have `df.bar`, a matrix object with a single column containing the values that are used by the barplot on the x axis. Adjusting your line to this scale, everything works out in the end. See the code.

```
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")
```

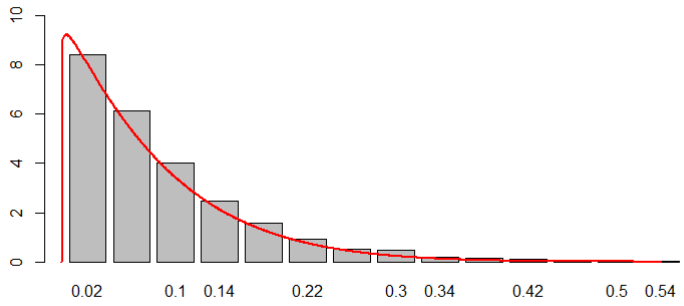


Figure 7.14: Standardized histogram with data on the proportion of decline in the daily value of 2314 shares that saw their price falling from one day to the next. The curve in red is the density of a Beta$(1.04, 10.63)$.

■

The expectation of a r.v. $X \sim$ Beta(α, β) is obtained by calculating the integral:

$$\begin{aligned} \mathbb{E}(X) &= \int_0^1 x\, f(x)\, dx = C \int_0^1 x\, x^{\alpha-1}(1-x)^{\beta-1}\, dx \\ &= C \int_0^1 x^{\alpha}(1-x)^{\beta-1}\, dx \end{aligned}$$

We stated earlier that the constant C of a Beta(α, β) can be written as $C = \Gamma(\alpha+\beta)/(\Gamma(\alpha)\Gamma(\beta))$ where $\Gamma(z)$ is the gamma function, which has the recursive property $\Gamma(z+1) = z\,\Gamma(z)$.

Note now that a Beta$(\alpha+1, \beta)$ would have density $C^* x^{\alpha+1-1}(1-x)^{\beta-1}$, which corresponds to the core within the last integral in the above development. As the area under any density function is 1, that integral must be equal to $1/C^*$ and therefore

$$\mathbb{E}(X) = \frac{C}{C^*} = \frac{\Gamma(\alpha+\beta)/(\Gamma(\alpha)\Gamma(\beta))}{\Gamma(\alpha+1+\beta)/(\Gamma(\alpha+1)\ Gamma(\beta))} = \frac{\alpha}{\alpha+\beta}$$

In conclusion, in the case of $X \sim$ Beta(α, β) distribution, the expectation is $\mathbb{E}(X) = \alpha/(\alpha+\beta)$.

The cumulative probability distribution does not have a closed formula, an analytic expression. However, it is well studied numerically and can be used in probability calculations without difficulty.

7.7 Exponential Distribution

The exponential distribution is often a good model for the waiting time *between* random events that occur continuously in time and at a λ constant rate of occurrence. This rate represents the average number of occurrences per unit of time. This exponential distribution was pioneered in 1909 by Erlang, a Danish mathematician, showing that the random waiting time between telephone incoming calls on a server was distributed as an exponential distribution with a certain parameter value. Let's say that the unit of time is hours and that, at that time, the λ rate was 9.4 calls per hour, on average. Sometimes there were more calls than 9.4 in an hour, sometimes fewer calls than 9.4. But, on average, within an hour, we had 9.4 calls.

Other examples of waiting time between random occurrences that have a more or less constant rate are the following:

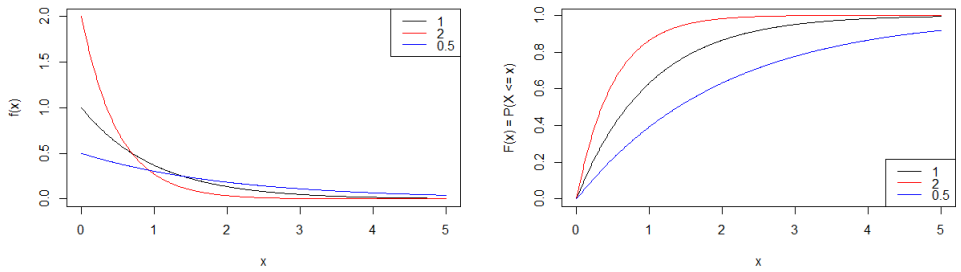


Figure 7.15: Left: Density function of the exponential distribution with λ equal to 1, 2 and 0.5. Right: Graph of the cumulative distribution function $\mathbb{F}(x)$ in the case of exponential density with λ equal to 1, 2 and 0.5.

- The waiting time for the next radioactive decay of an atomic mass (time between counts of a Geiger counter)
- distance between mutations in a DNA strand
- time between communications on social networks (see [2]).
- time between accesses on a web page

In practice, the assumption of a constant rate of occurrence of events may not be realistic if we did not impose some constraints. For example, the rate of incoming phone calls changes throughout the day. However, if we consider only a limited range of time, such as between 14 and 16 hours of a working day, the rate may be considered approximately constant and the exponential distribution can be a good model for the time between calls during this period.

Definition 7.7.1 — Exponential Distribution. A v.a. X supported $\mathscr{S} = (0,\infty)$ is called exponential if its probability density is given by $f(x) = \lambda e^{-\lambda x}$ where $\lambda > 0$ is the parameter of the distribution.

Notation 7.3. *If the v.a. X follows the exponential distribution with parameter λ we write $X \sim exp(\lambda)$.*

The left side of Figure 7.15 shows three examples of the exponential density function. There is not much diversity. It is always a simple exponential decay starting from the origin. The rate of decay is dictated by the parameter λ, the rate of occurrence of events. The higher the value of λ, the faster an event usually occurs and the waiting time tends to decrease. Thus, very short waiting times become more common and faster and faster is the density decay. Therefore, the density function becomes more concentrated around zero and closer to zero the value of X tends to be.

Reflecting this control of the parameter λ in the X values, the expectation of X can be obtained by integrating by parts with $u = \lambda x$ and $dv = e^{-\lambda x}$:

$$\mathbb{E}(X) = \int_0^\infty x\,\lambda e^{-\lambda x}\,dx = -\frac{x\lambda e^{-\lambda x}}{\lambda}\,|_0^\infty + \int_0^\infty e^{-\lambda x}\,dx = -(0-0) + \left(-\frac{e^{-\lambda x}}{\lambda}\,|_0^\infty\right) = \frac{1}{\lambda}\,.$$

The cumulative distribution function $\mathbb{F}(x) = \mathbb{P}(X \leq x)$ is easily obtained. As the exponential distribution support set is the semi-axis $(0,\infty)$, we have $\mathbb{F}(x) = \mathbb{P}(X \leq x) = 0$ for all $x < 0$. For $x \geq 0$, we have:

$$\mathbb{F}(x) = \mathbb{P}(X \leq x) = \int_0^x \lambda e^{-\lambda x} dx = 1 - e^{-\lambda x}\ :\,.$$

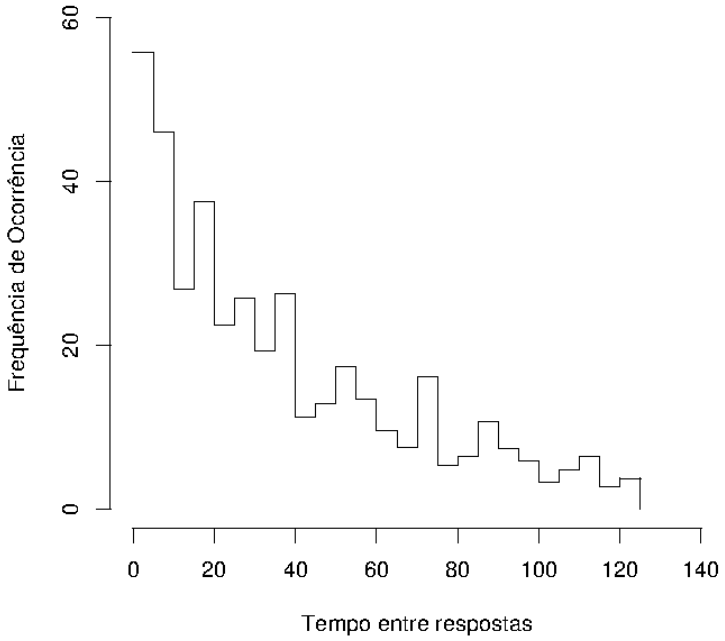


Figure 7.16: Non-standardized histogram of the time between impulses in the spontaneous activity of a spinal cord neuron.

The graph of this accumulated function with different values for λ is on the right side of Figure 7.15.

■ **Example 7.6 — Electric impulses.** Psychologists and biophysicists are interested in the time X between successive electrical impulses in the spinal cord of several mammals. W. J. Megill made several measurements of the time elapsed between impulses in a cat's medulla. 391 intervals were recorded and the histogram of these data in Figure 7.16 suggests the adoption of an exponential model.

■

■ **Example 7.7 — Air conditioning in Boeings.** [29] recorded the times between successive failures of the air conditioning systems of a fleet of 13 Boeing 730 jets. These times between failures are the lengths of time that are listed in the table below. Each plane is in a column indicated by the header and the times between failures on that plane are displayed along the column. Thus, plane number 7907 had a failure of its air conditioning system after 194 hours of service, a second failure only 15 hours after the first one, and a third failure 41 service hours after the second failure. More or less after 2000 hours of service, each plane received a general inspection. If a time interval between two air conditioning system failures included the general inspection, the length of this time interval was not recorded (and therefore is not in the table below) as their magnitude may have been affected by possible repairs made during the general inspection.

ID number												
7907	7908	7909	7910	7911	7912	7913	7914	7915	7916	7917	8044	8045
194	413	90	74	55	23	97	50	359	50	130	487	102
15	14	10	57	320	261	51	44	9	254	493	18	209
41	58	60	48	56	87	11	102	12	5		100	14
29	37	186	29	104	7	4	72	270	283		7	57
33	100	61	502	220	120	141	22	603	35		98	54
181	65	49	12	239	14	18	39	3	12		5	32
	9	14	70	47	62	142	3	104			85	67
	169	24	21	246	47	68	15	2			91	59
	447	56	29	176	225	77	197	438			43	134
	184	20	386	182	71	80	188				230	152
	36	79	59	33	246	1	79				3	27
	201	84	27	NA	21	16	88				130	14
	118	44	NA	15	42	106	46					230
	NA	59	153	104	20	206	5					66
	34	29	26	35	5	82	5					61
	31	118	326		12	54	36					34
	18	25			120	31	22					
	18	156			11	216	139					
	67	310			3	46	210					
	57	76			14	111	97					
	62	26			71	39	30					
	7	44			11	63	23					
	22	23			14	18	13					
	34	62			11	191	14					
		NA			16	18						
		130			90	163						
		208			1	24						
		70			16							
		101			52							
		208			95							

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.

■

7.8 Distribuição normal ou gaussiana

A distribuição normal ou gaussiana é a mais famosa distribuição de probabilidade. O requerimento mínimo para adotaremos 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

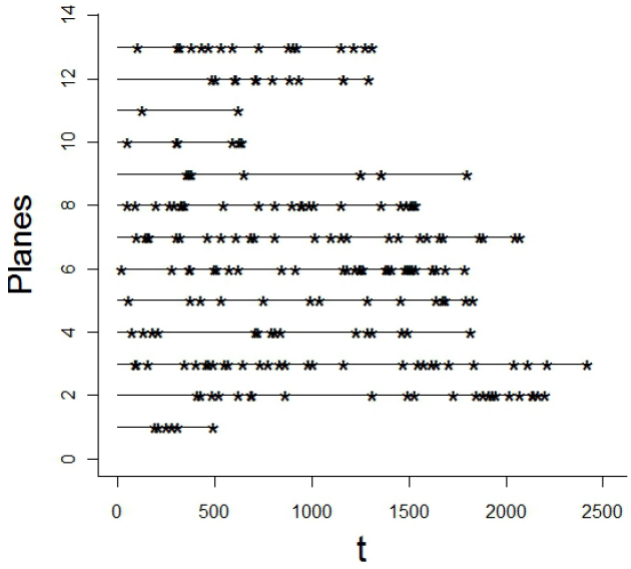


Figure 7.17: Visualization of the air conditioning repair times on Boeings. Each line is an airplane. Each star marks the time of a repair.

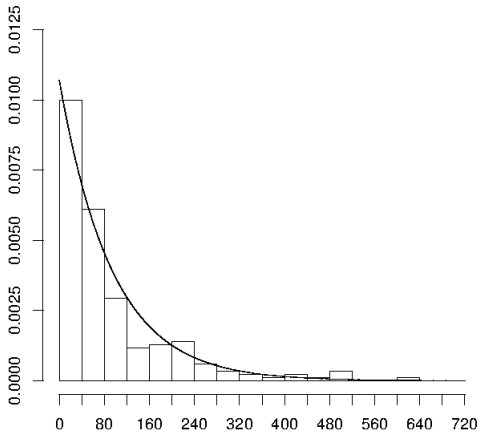


Figure 7.18: Histograma padronizado e densidade exponencial usando λ igual ao inverso da média aritmética dos tempos na amostra.

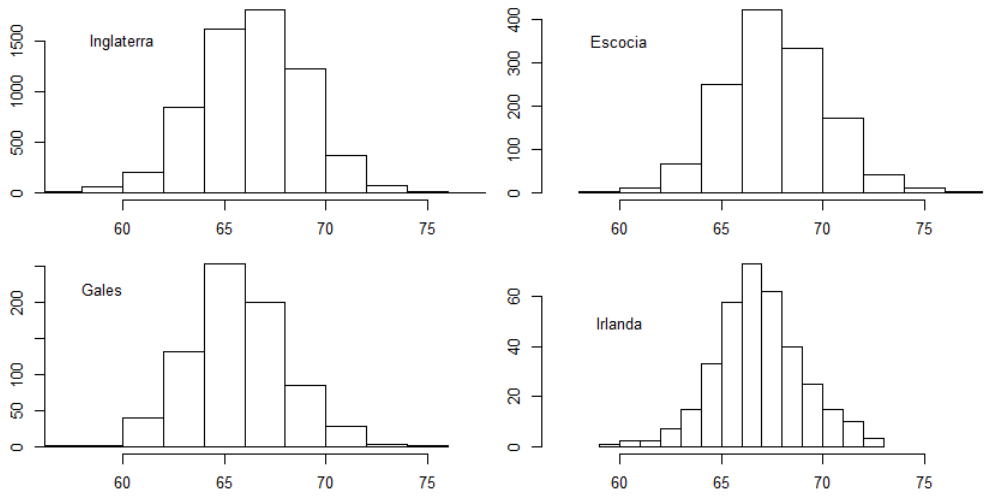


Figure 7.19: Histogramas de alturas (em polegadas) de amostras de indivíduos adultos do sexo masculino da Grã-Bretanha em 1883.

gaussiana se ajusta bem.

Definition 7.8.1 — Distribuição normal ou gaussiana. Uma v.a. X com suporte na reta real $\mathbb{R} = (-\infty, \infty)$ possui distrubuição 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 $X \sim \mathcal{N}(\mu, \sigma^2)$.*

■ **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).

■

■ **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 sadios 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 [20]. 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 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.

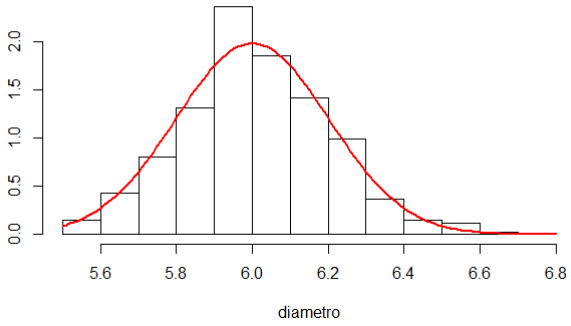


Figure 7.20: Histogramas dos diâmetros transversais da cabeça de 1000 estudantes da Universidade de Cambridge

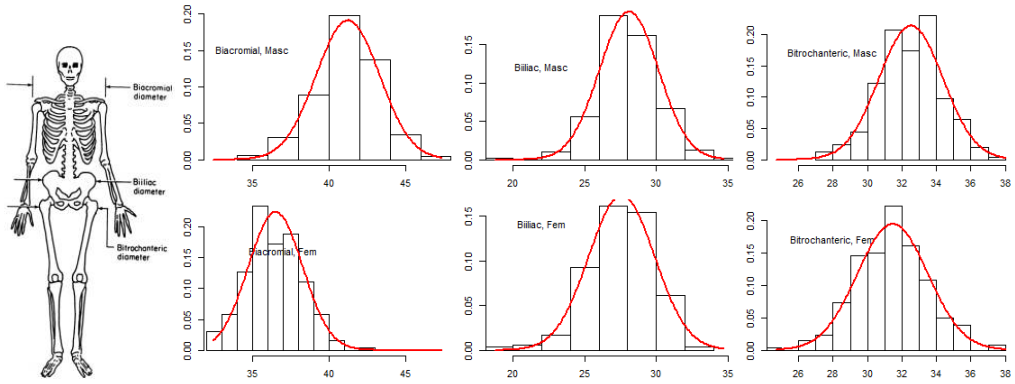


Figure 7.21: Histogramas para homens (linha superior) e mulheres (linha inferior) com os diâmetros Biacromial, Biiliac e Bitrochanteric, respectivamente, da esquerda para a direita.

```
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="Biiliac", 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 19. 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

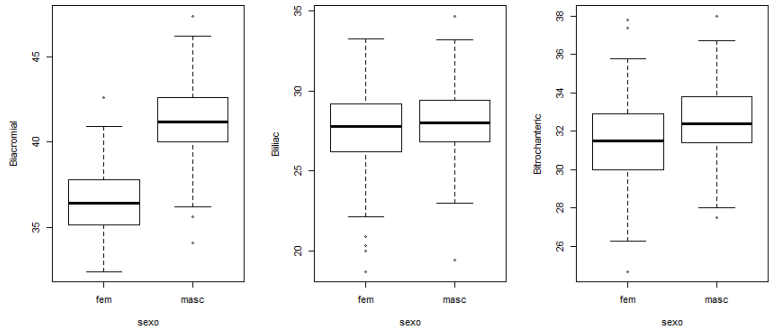


Figure 7.22: Histogramas para homens (linha superior) e mulheres (linha inferior) com os diâmetros Biacromial, Biiliac e Bitrochanteric, respectivamente, da esquerda para a direita.

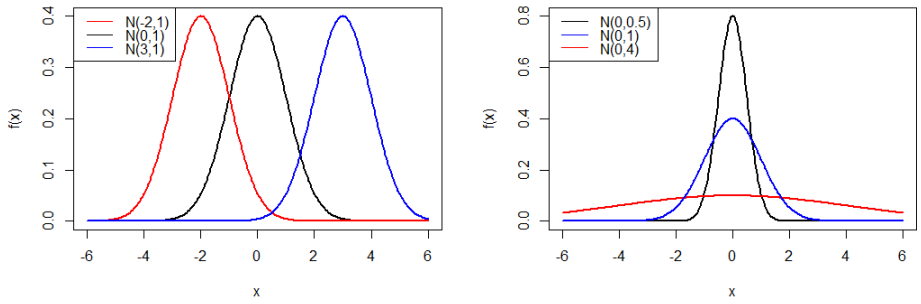


Figure 7.23: Ilustração do efeito de variar μ e σ no caso de uma densidade da distribuição normal $\mathcal{N}(\mu, \sigma^2)$.

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)
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)
```

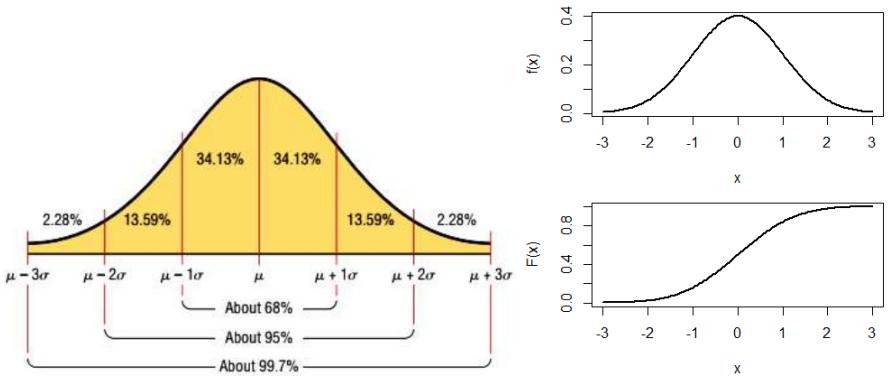


Figure 7.24: Ilustração da regra de 2σ e de 3σ no caso de uma distribuição normal $\mathcal{N}(\mu, \sigma^2)$.

```
legend("topleft",c("N(0,0.5)", "N(0,1)", "N(0,4)"), lty=1,
          col=c( "black","blue" ,"red"))
```

Definition 7.8.2 — 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 fómula fechada:$C = 1/(\sqrt{2\pi\sigma^2}$. Está além dos objetivos deste texto demonstrar este fato. O leitor interessado deve consultar [22]. 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 qe 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, 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 7.8.1

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\, dx = 1.$$

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 7.8.2 $\mathbb{E}[X] = \mu$.

Proof:

$$\begin{aligned} \mathbb{E}[X] &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x e^{-(x-\mu)^2/2\sigma^2}\, \text{x} \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x-\mu) e^{-(x-\mu)^2/2\sigma^2}\, \text{x} + \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \mu e^{-(x-\mu)^2/2\sigma^2}\, \text{x}. \end{aligned}$$

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 7.8.3 $\mathbb{V}(X) = \sigma^2$.

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

$$\begin{aligned} \mathbb{V}(Z) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2}\, \text{z} \\ &= \left[-\frac{1}{\sqrt{2\pi}} z e^{-z^2/2} \right]_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2}\, \text{z} \\ &= 0 + 1 \\ &= 1 \end{aligned}$$

So $\mathbb{V}X = \sigma^2$.

7.9 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 7.9.1 — 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

$$f(x) = C x^{\alpha - 1} e^{-\beta x}$$

para $x > 0$. A constante C é obtida para garantir que a área sob a densidade é iguala 1.

Notation 7.5. *Se a v.a. X segue a distribuição gama com parâmetros α e β escrevemos* $X \sim \Gamma(\alpha, \beta)$.

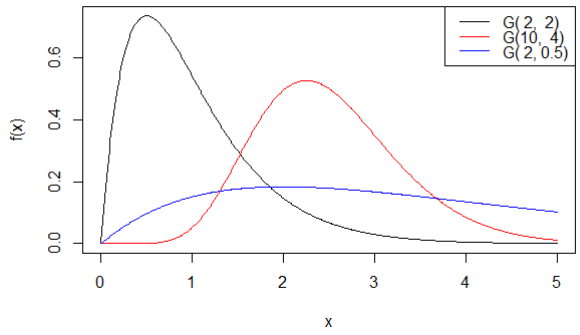


Figure 7.25: Exemplos de Gama.

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 $x^9 \to \infty$ e $e^{-4x} \to 0$. O produto desses dois termos ficaráo próximo de zero pois o decrescimento exponencial domina qualquer crescimento polinomial. Assim, $f(x) \approx 0$ tanto se $x \approx 0$ quanto se $x \to \infty$. Para x 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 $T_1, T_2, \ldots, T_k$ são v.a.'s $\exp(\lambda)$ e independentes então $X = T_1 + T_2 + \ldots + T_k \sim \Gamma(k,\lambda)$. Este caso especial da gama, com α igual a um inteiro positivo, aparece com tanta frequência em aplicações que acabou ganhando o nome de dstribuição de Erlang, em homenagem a Agner Krarup Erlang, um matemático dinarmaquês 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 análie 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.

[10] estudou matematica e empiricamente o crescimento de populações desses insetos sob diversas condições. Eles apresentaram justificativas ecológicas para afirmar que o número de

Figure 7.26: Foto de besouro da espécie *Tribolium castaneum* e amostra de milho infestado por eles cercado de milhos sadios.

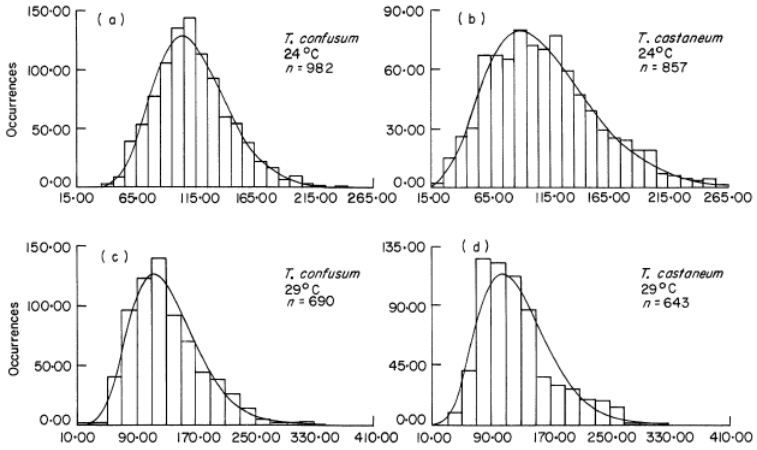


Figure 7.27: Histogramas dos tamanhos de n populações de besouros crescidas sob diferentes condições e ajuste da distribuição gama em cada caso. Parâmetros da gama foram obtidos por máxima verossimilhança, assunto do capítulo ??.

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 estudo têm mostrado que a distribuição gama costuma ser um bom ajuste para a abundância de espécies por razões teóricas ([13]) ou empíricas. Por exemplo, [32] mostraram que a distribuição gama foi a melhor distribuição para ajustar dados de tamanho de comunidades de 128 dentre 136 diferentes 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 $\beta\ dx = dy$ 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 **??**. 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.

7.10 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 $\mathbb{P}(t < T < t+\Delta t \mid T > t)$. 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 $\mathbb{P}(t < T < t+\Delta t \mid T > t) = \mathbb{P}(60 < T < 61 \mid T > 60)$.

Agora queremos saber a chance de sobreviver o próximo 1 minuto dado que o componente já funcionou por 5 anos seguidos. Quanto deveria ser $\mathbb{P}(2628000 < T < 2628001 \mid T \geq 2628000)$? 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 qualqer 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 ito 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 $\mathbb{P}(t < T < t+\Delta t \mid T > t)$ 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 longuínquo.

Naturalmente, temos $\lim_{\Delta t\to\infty} \mathbb{P}(t < T < t+\Delta t \mid T > t) = 0$. 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 *mili*segundos é menor ainda. Uma maneira de de se fazer este estudo sem ficar preso neste resultado óbvio é padronizar a probabilidade $\mathbb{P}(t < T < t+\Delta t \mid T > t)$ calculando-a *por unidade de tempo* e levando-a ao limite quando Δt vai a zero.

Definition 7.10.1 — Taxa de Falha Instantânea. A taxa de falha instantânea de uma v.a. T é definida como

$$\lambda(t) = \lim_{\Delta t\to\infty} \frac{P(t < T < t+\Delta t \mid T > t)}{\Delta t} \tag{7.5}$$

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 7.10.1 — 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:

$$f(t) = Ct^{\alpha-1}e^{(\frac{t}{\beta})^{\alpha}} \tag{7.6}$$

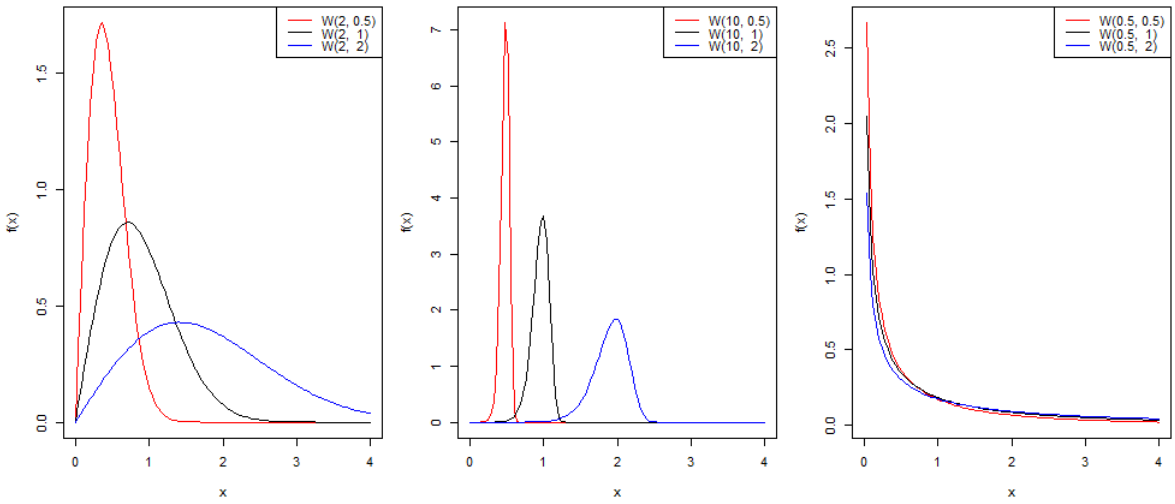


Figure 7.28: Densidade da distribuição Weibull $\mathscr{W}(\alpha,\beta)$. Esquerda: $\alpha = 2$ e $\beta = 0.5, 1, 2$. Centro: $\alpha = 10$ e $\beta = 0.5, 1, 2$. Direita: $\alpha = 0.5$ e $\beta = 0.5, 1, 2$.

para $t > 0$. Esta distribuição é chamada de Weibull com parâmetros α e β.

A constante $\beta > 0$ em (7.6) 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 $X \sim \mathscr{W}(\alpha,\beta)$.*

Como você imaginar, a forma da densidade de uma Weibull $\mathscr{W}(\alpha,\beta)$ depende dos parâmetros α e β. A Figura 7.28 mostra o gráfico da função densidade de uma Weibull $\mathscr{W}(\alpha,\beta)$ 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 parãmetro 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. $X \sim \mathscr{W}(\alpha,\beta)$ envolve a função gama $\Gamma(Z)$: $\mathbb{E}(X) = \beta\Gamma((\alpha+1)/\alpha)$. A função distribuição acumulada $\mathbb{F}(x)$ tem uma forma funcional simples. Para $x > 0$ temos

$$\mathbb{F}(x) = \int_0^x C\, x^{\alpha-1}\, e^{-(x/\beta)^{\alpha}}\, dx = 1 - e^{-(x/\beta)^{\alpha}}$$

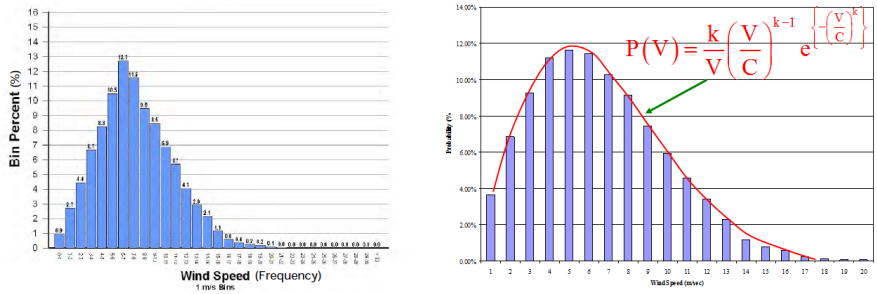


Figure 7.29: Histograma de dados de velocidade do vento (esquerda). Outro histograma com o ajuste de uma Weibull (direita).

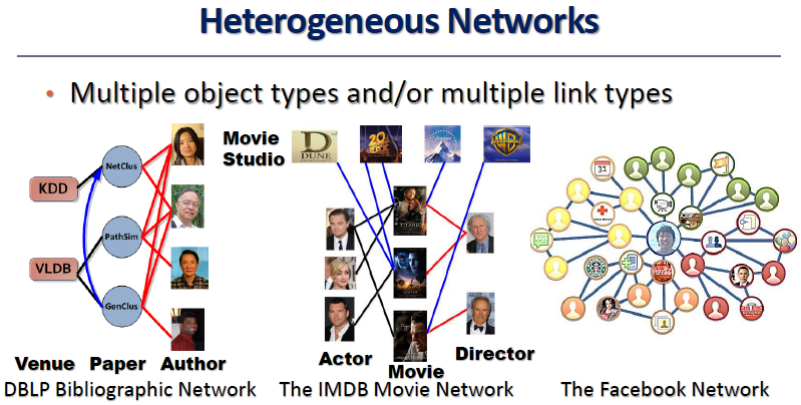


Figure 7.30: Redes heterogêneas. Extraído de [33].

■ **Example 7.12 — Weibull em redes sociais.** [33] 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-comércio, 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 do modelos lineares generalizados, a ser estudado no capítulo **??**.

■

7.11 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 varição mas uma certa proporção não desprezível tem valores ordens de grandeza maior que o valor esperado.

Definition 7.11.1 — 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 $\mathscr{S} = (x_o, \infty)$ e densidade de probabilidade da forma

$$f(x) = C\,\frac{1}{x^{\alpha+1}}$$

para $x > x_o$.

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 $C = \alpha\, x_o^{\alpha}$:

$$\begin{aligned} 1 &= \int_{x_o}^{\infty} C\,\frac{1}{x^{\alpha+1}}\,dx = C\left(\frac{x^{1-\alpha-1}}{-\alpha}\,|_{x_o}^{\infty}\right) \\ &= \frac{C}{\alpha}\left(-\frac{1}{\infty^{\alpha}} - \frac{-1}{x_o^{\alpha}}\right) \\ &= \frac{C}{\alpha\, x_o^{\alpha}} \end{aligned}$$

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* $X \sim \texttt{Pareto}(\alpha, x_o)$.

■ **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 clásssico do estatístico inglês G. U. Yule (ver [37]).

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 resseguros? A Swiss Re, a maior companhia européia de resseguros, fez um estudo. Nos casos de

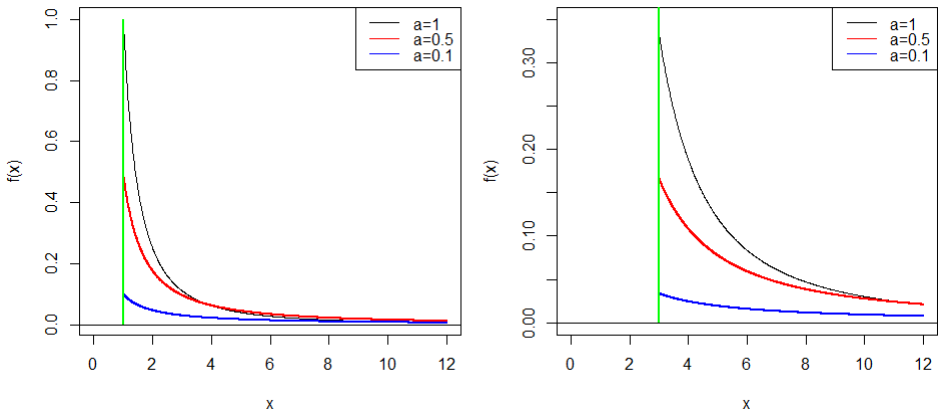


Figure 7.31: Exemplos de densidades de Pareto. Casos a esquerda têm $x_o = 1$ e $\alpha = 1, 0.5, 0.1$. Casos a direita têm $x_o = 3$ e os mesmos valores para α.

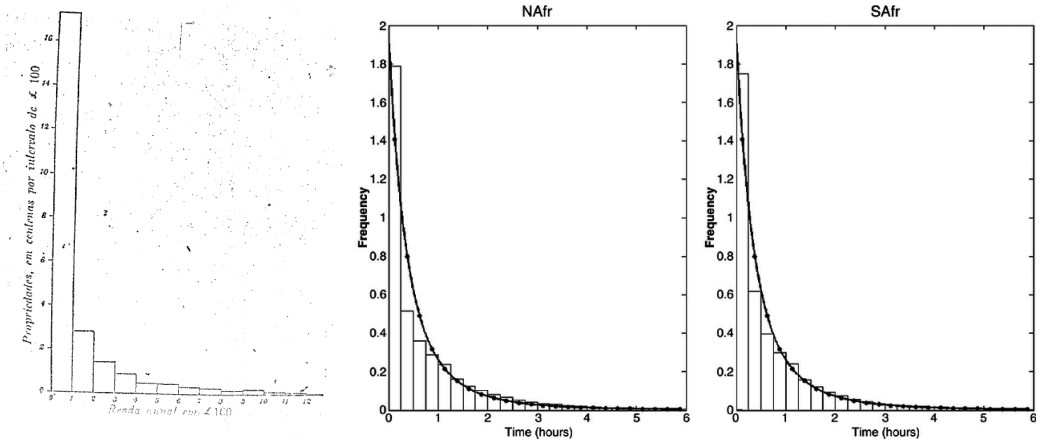


Figure 7.32: Número de propriedades de terra na Inglaterra em 1717. Ajuste de distribuição de Pareto a dados da duração temporal dos incêndios florestais no Sul e no Norte da África.

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

$$\mathbb{F}(x) = \int_{x_o}^{x} C\, \frac{1}{t^{1+\alpha}}\, dt = \underbrace{\alpha\, x_o^\alpha}_{C} \left(\frac{1}{-\alpha}\, t^\alpha \Big|_{x_o}^{x} \right) = 1 - \left(\frac{x_o}{x} \right)^\alpha \tag{7.7}$$

7.11.1 Simulando uma Pareto

No capítulo ?? veremos lgumas 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 7.7 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 `Pareto`(x_o, α). 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)
```

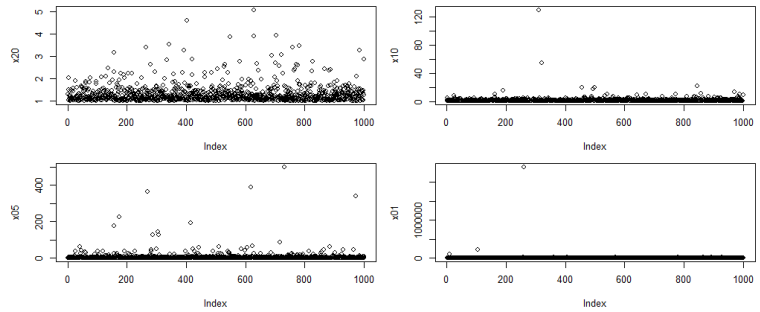


Figure 7.33: Amostras de 1000 valores Pareto com $x_o = 1$ e α igual a 4 (canto superior esquerdo), 2 (canto superior direito), 1 (canto inferior esquerdo) e 0.5 (canto inferior direito).

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 dles chegam a valores superiores a a 100 mil. Compare com a faixa $(0, 100)$ onde encontram-se 91% deles.

7.11.2 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 $\texttt{Pareto}(x_o = 1, \alpha = 2)$. 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 estamos olhando todos os dados mas apenas aqueles que não ultrapassaram um limiar, e portanto atemos apenas uma informação parcial. Na linha inferior, com $\texttt{Pareto}(x_o = 1, \alpha = 1)$, esta situação se repete.

Uma maneira mais eficiente de visualizar a possível adequação do modelo Pareto é usando a função $\mathbb{F}(x)$. Trataremos disso no próximo capítulo.

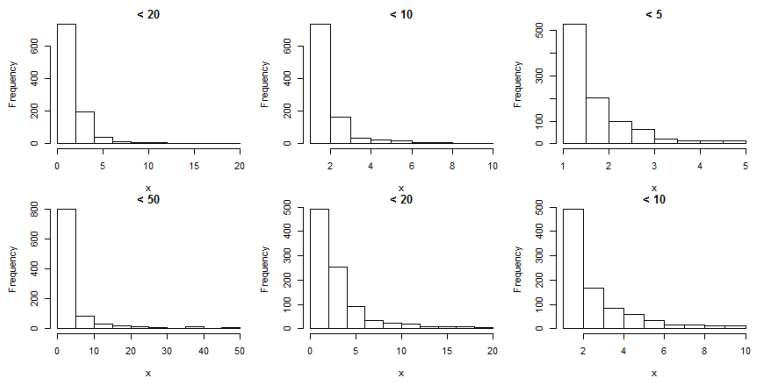


Figure 7.34: Amostras de 1000 valores Pareto com $x_o = 1$ e α igual a 4 (canto superior esquerdo), 2 (canto superior direito), 1 (canto inferior esquerdo) e 0.5 (canto inferior direito).

8. Independence and Transformation

8.1 Independence of r.v.'s

We previously defined the concept of independent events: A and B, both contained in Ω, are independent events if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Equivalently, we can prove that events A and B are independent if, and only if, $\mathbb{P}(A|B) = \mathbb{P}(A)$. We can extend this concept to random variables instead of just events. Intuitively, we'll say that two random variables X and Y are independent random variables if knowing the value of one random variable doesn't change the probabilities associated with the possible values of the other random variable. What's the difference between independence of events and random variables? Stating that random variables X and Y are independent is equivalent to stating that any event associated with X (for example, the event $X \in (0.0, 0.1)$) is independent of any other event defined from just Y (for example, $Y \in (0.9, 1.0)$).

Lembre-se que as v.a.'s são representações matemáticas das colunas da matriz de dados. Considere k dessas colunas e as v.a.'s correspondentes $X_1, X_2, \ldots, X_k$. Estamos interessados nos valores dessas variáveis *numa mesma linha da tabela de dados*. Isto é, para um dado resultado ω do espaço amostral, observamos os valores das v.a.'s $X_1(\omega), X_2(\omega), \ldots, X_k(\omega)$. Todos são valores medidos no mesmo resultado ω. Por causa disso, é possível que os valores $X_1(\omega), X_2(\omega), \ldots, X_k(\omega)$ estejam associados, um valor dando alguma informação sobre o valor dos demais. Na maioria das vezes será de fato assim. Mas existem situações em que as v.a.'s não estão associadas.

Considere um exemplo extremo para enfatizar o conceito. Imagine que ω representa um indivíduo escolhido ao acaso de certa população humana. Sejam $X_1(\omega)$ o seu nível de colesterol LDL (colesterol ruim), $X_2(\omega)$ um indicador binário de que o indivíduo é obeso, $X_3(\omega)$ um indicador binário de que o indivíduo é fumante, $X_4(\omega)$ um indicador binário de que seu primeiro nome começa com uma das letras $A, \ldots, M$ ou se começa com $N, \ldots, Z$, e $X_5(\omega)$ é sua altura. Intuitivamente, podemos esperar que as variáveis X_1, X_2 e X_3 não sejam independentes umas das outras. Vários estudos clínicos mostram que pessoas obesas são mais propensas a terem altos níveis de colesterol LDL. Eles também mostram que o fumo prejudica as paredes arteriais tornando-as mais suscetíveis ao acúmulo de colesterol LDL. Por outro lado, é difícil imaginar como a primeira letra do nome de um indivíduo (X_4) ou sua altura (X_5) podem estar associadas com X_1, X_2 e X_3.

Outro exemplo onde obviamente as v.a.'s são independentes é a observação de n lançamentos de

uma moeda. Seja ω a sequência dos n lançamentos. O resultado ω é uma n-upla com C (cara) ou $\tilde{C}$ (coroa) em cada entrada. O espaço de probabilidade $(\Omega, \mathscr{A}, \mathbb{P})$ é bem conhecido: $\mathscr{A}$ é o conjunto de todos os subconjunto de Ω e, sendo Ω discreto, basta especificarmos $\mathbb{P}$ em cada resultado atômico ω. Por exemplo, para a sequência ω de n caras ou coroas, temos $\mathbb{P}(\omega) = \theta^k(1-\theta)^{n-k}$ onde θ é a probabilidade de um sucesso num único lançamento e k é o número total de caras em ω, com $k = 0, 1, \ldots, n$. Seja $X_i(\omega)$ uma v.a. binária indicando se o i-ésimo lançamento foi cara ($X_i(\omega) = 1$) ou coroa ($X_i(\omega) = 0$). Em condições usuais, os lançamentos não guardam qualquer relação com os resultados prévios ou futuros. Assim, intuitivamente, $X_1, X_2, \ldots, X_n$ seriam v.a.s independentes.

Neste último exemplo, é importante conectar uma tabela de dados com os resultados ω. COMPLETAR ??

Sejam $X_1, X_2, \ldots, X_n$ v.a.'s medidas num mesmo espaço de probabilidade $(\Omega, \mathscr{A}, \mathbb{P})$. Informalmente, as v.a.'s $X_1, X_2, \ldots, X_n$ são independentes se, e somente se, quaisquer eventos determinados por qualquer grupo de variáveis distintas formam eventos independentes. Por exemplo, se as v.a.'s são independentes então os eventos $[X_1 \leq 2]$ e $[X_2 > 4]$ são eventos independentes; $[X_1 > 4]$ e $[X_2 > 4]$ são independentes também, mesmo que o número 4 apareça nos dois eventos; $[X_1 \leq 2]$, $[X_2 > 4$ e $X_3 > 7]$ e $[X_4 < 0$ ou $X_5 > 10]$ são eventos independentes; etc.

Definition 8.1.1 — V.A.s independentes. As v.a.'s $X_1, X_2, \ldots, X_n$ são v.a.s independentes se

$$\mathbb{P}(X_1 \in B_1, X_2 \in B_2, \ldots, X_n \in B_n) = \mathbb{P}(X_1 \in B_1)\mathbb{P}(X_2 \in B_2)\ldots\mathbb{P}(X_n \in B_n) \tag{8.1}$$

para todo conjunto B_i da reta real, $i = 1, \ldots, n$. Se as variáveis não forem independentes, dizemos que elas são dependentes.

Se $X_1, X_2, \ldots, X_n$ são v.a.s independentes então qualquer subconjunto delas também é formado de v.a.'s independentes (propriedade hereditária da independência). Por exemplo, X_1 e X_2 são independentes se $X_1, X_2, \ldots, X_n$ forem v.a.s independentes pois

$$\begin{aligned}
\mathbb{P}(X_1 \in B_1, X_2 \in B_2) &= \mathbb{P}(X_1 \in B_1, X_2 \in B_2, X_3 \in \mathbb{R}, \ldots, X_n \in \mathbb{R}) && (8.2)\\
&= \mathbb{P}(X_1 \in B_1)\mathbb{P}(X_2 \in B_2)\mathbb{P}(X_3 \in \mathbb{R})\ldots\mathbb{P}(X_n \in \mathbb{R}) && (8.3)\\
&= \mathbb{P}(X_1 \in B_1)\mathbb{P}(X_2 \in B_2) \times 1 \ldots \times 1 && (8.4)\\
&= \mathbb{P}(X_1 \in B_1)\mathbb{P}(X_2 \in B_2) && (8.5)
\end{aligned}$$

■ **Example 8.1 — Lançamentos de uma moeda.** De volta ao exemplo dos n lançamentos de uma moeda com probabilidade de cara em um único lançamento igual a $\theta \in (0,1)$ e com $X_i(\omega) = 1$ se o i-ésimo lançamento foi cara e $X_i(\omega) = 0$, caso contrário. Se $X_1, \ldots, X_n$ são independentes então o cálculo de vários eventos fica muito simples. Por exemplo,

$$\begin{aligned}
\mathbb{P}(X_1 = 1, X_2 = 0) &= \mathbb{P}(X_1 = 1, X_2 = 0, X_3 \in \{0,1\}, \ldots, X_n \in \{0,1\}) && (8.6)\\
&= \mathbb{P}(X_1 = 1)\ \mathbb{P}(X_2 = 0)\ \mathbb{P}(X_3 \in \{0,1\})\ \ldots\ \mathbb{P}(X_n \in \{0,1\}) && (8.7)\\
&= \theta(1-\theta) \times 1 \ldots \times 1 = \theta(1-\theta) && (8.8)
\end{aligned}$$

■

Definition 8.1.2 — Subvetores de v.a.s independentes. Sejam as v.a.'s $X_1, X_2, \ldots, X_n$ medidas num mesmo espaço de probabilidade $(\Omega, \mathscr{A}, \mathbb{P})$. Dizemos que o subconjunto X_1, X_2 é independente do subconjunto $X_3, \ldots, X_n$ se

$$\mathbb{P}(X_1 \in B_1, X_2 \in B_2, X_3 \in B_3, \ldots, X_n \in B_n) = \mathbb{P}(X_1 \in B_1, X_2 \in B_2) \times \mathbb{P}(X_3 \in B_3, \ldots, X_n \in B_n)$$

para todo conjunto B_i da reta real, $i = 1, \ldots, n$.

Obviamente, a definição estende-se para qualquer partição do conjunto de v.a.'s e não apenas em dois subconjuntos. Isto é, podemos particionar o conjunto de n variáveis em subconjuntos independentes mas com possível dependência interna a cada um deles. Se todas as v.a.'s forem dependentes entre si, a partição trivialmente seria o próprio conjunto de todas as variáveis.

▪ **Example 8.2 — De volta ao colesterol.** Considere novamente o exemplo discutido acima em que ω representa um indivíduo escolhido ao acaso. Temos: $X_1(\omega)$ é o seu nível de colesterol LDL; $X_2(\omega) = 1$ se obeso e 0, caso contrário; $X_3(\omega) = 1$ se fumante e 0, caso contrário; $X_4(\omega) = 1$ se o primeiro nome começa com $A, \ldots, M$ e 0, caso contrário; $X_5(\omega)$ é sua altura. Suponha que as v.a.'s (X_1, X_2, X_3) sejam dependentes entre si mas independentes das variáveis X_4 e X_5. Isto é, particionamos o conjunto de 5 variáveis em três subconjuntos independentes e portanto a probabilidade $\mathbb{P}(X_1 > 200\text{mg/dl}, X_2 = 1, X_3 = 0, X_4 = 1, X_5 > 1.80\text{m})$ seria igual a $\mathbb{P}(X_1 > 200\text{mg/dl}, X_2 = 1, X_3 = 0)\ \mathbb{P}(X_4 = 1)\ \mathbb{P}(X_5 > 1.80\text{m})$. ▪

Definition 8.1.3 — V.A.'s i.i.d.. Sejam as v.a.'s $X_1, X_2, \ldots, X_n$ medidas num mesmo espaço de probabilidade $(\Omega, \mathscr{A}, \mathbb{P})$. Dizemos que as v.a.'s são independentes e identicamente distribuídas quando elas forem coletivamente independentes e tiverem, cada uma delas, a mesma distribuição de probabilidade. Abreviamos dizendo que elas são i.i.d.

Definition 8.1.4 — Amostra aleatória. Sejam $X_1, X_2, \ldots, X_n$ v.a.'s i.i.d. com a mesma distribuição de probabilidade de uma v.a. X com distribuição acumulada $\mathbb{F}$ e densidade $f(x)$ (caso contínuo) ou função de probabilidade $\mathbb{P}(X = x_i)$. Dizemos que o vetor aleatório $(X_1, \ldots, X_n)$ é uma *amostra aleatória* da v.a. X. Alternativamente, dizemos que o vetor é uma amostra de $\mathbb{F}$ ou de $f(x)$ ou de $\mathbb{P}(X = x_i)$.

8.2 Como saber quando as v.a.'s são independentes?

Como sabemos que v.a.s são independentes? Em princípio verificando a definição. Por exemplo, no caso de apenas duas v.a.'s X_1 e X_2 deveríamos verificar que

$$\mathbb{P}(X_1 \in B_1, X_2 \in B_2) = \mathbb{P}(X_1 \in B_1)\mathbb{P}(X_2 \in B_2)$$

para todo par de conjuntos B_1 e B_2 da reta real. Na prática, existem duas maneiras de concluir que um conjunto de v.a.'s é composto por v.a.'s independentes. A primeira delas é por suposição. Refletindo sobre as condições do problema específico nós *assumimos* que as v.a.'s são independentes. Neste caso, basta sabermos a distribuição de cada v.a. individualmente, $\mathbb{P}(X_i \in B_i)$, para obtermos as probabilidades envolvendo todas as v.a.'s. Isto é, ao invés de especificar a distribuição conjunta das v.a.'s

$$\mathbb{P}(X_1 \in B_1, X_2 \in B_2, \ldots, X_n \in B_n)\,, \tag{8.9}$$

nós *afirmamos* que ela é igual ao produto das distribuições individuais das v.a.'s,

$$\mathbb{P}(X_1 \in B_1)\mathbb{P}(X_2 \in B_2)\ldots\mathbb{P}(X_n \in B_n)\,. \tag{8.10}$$

Por exemplo, no caso da amostra com dados sobre colesterol e outros atributos, podemos assumir que X_4 (se o primeiro nome começa com $A, \ldots, M$ ou não) é independente do nível de colesterol X_1. Começamos obtendo as distribuições $\mathbb{P}(X_i \in B_i)$ de cada coluna-variável separadamente. A seguir, *afirmamos* que $\mathbb{P}(X_1 \in B_1, X_4 \in B_4)$ é dado pelo produto $\mathbb{P}(X_1 \in B_1)\mathbb{P}(X_4 \in B_4)$. Assim, a condição de independência é obtida por suposição. E quando isto não bater com a realidade? E quando a suposição de independência não for válida? Existem métodos para testarmos se duas ou mais variáveis são independentes. Ver capítulo ??.

A segunda maneira pela qual checamos a indepência de v.a.'s é verificando matematicamente que a condição (8.1) é correta. Neste caso, a probabilidade conjunta $\mathbb{P}(X_1 \in B_1, X_2 \in B_2, \ldots, X_n \in B_n)$ é fornecida. De alguma forma, temos um modelo para as probabilidades envolvendo todas as variáveis. Devemos obter cada probabilidade individual $\mathbb{P}(X_i \in B_i)$ e então mostrar que a probabilidade conjunta é igual ao produto $\mathbb{P}(X_1 \in B_1)\mathbb{P}(X_2 \in B_2)\ldots\mathbb{P}(X_n \in B_n)$,

■ **Example 8.3 — Sequência de caras e coroas.** Seja ω a sequência dos n lançamentos lançamentos de uma "moeda". A moeda é apenas uma abstração para uma sequência de resultados binários. O resultado ω é uma n-upla com C (cara) ou $\tilde{C}$ (coroa). Suponha que $\mathbb{P}(\omega) = \theta^k(1-\theta)^{n-k}$ onde k é o número total de caras em ω, com $k = 0, 1, \ldots, n$. Seja $X_i(\omega)$ a v.a. binária indicando se a i-ésima entrada na n-upla ω foi cara ($X_i(\omega) = 1$) ou coroa ($X_i(\omega) = 0$).

Como *consequência* deste modelo, nós podemos *provar* que as variáveis $X_1, X_2, \ldots, X_n$ são independentes (de fato, i.i.d.). Para ver isto, vamos considerar apenas o caso $n = 2$. O caso geral segue raciocínio idêntico. Temos

$$\begin{aligned}
\mathbb{P}(X_1 = 1) &= \mathbb{P}(X_1 = 1 \text{ e } X_2 \in \{0,1\}) && (8.11)\\
&= \mathbb{P}(X_1 = 1, X_2 = 0) + \mathbb{P}(X_1 = 1, X_2 = 1) && (8.12)\\
&= \theta(1-\theta) + \theta\theta \qquad \text{pela probabilidade conjunta} && (8.13)\\
&= \theta\,(1-\theta+\theta) = \theta && (8.14)
\end{aligned}$$

É claro que então $\mathbb{P}(X_1 = 0) = 1 - \mathbb{P}(X_1 = 1) = 1 - \theta$. De forma idêntica, encontramos $\mathbb{P}(X_2 = x)$ para $x = 0, 1$ e descobrimos que $X_1 \sim X_2$ (possuem a mesma distribuição).

Agora verifique que $\mathbb{P}(X_1 = x_1, X_2 = x_2) = \mathbb{P}(X_1 = x_1)\mathbb{P}(X_2 = x_2)$ para toda combinação de x_1 e x_2 em $\{0,1\}$. Por exemplo, pelo modelo de probabilidade conjunta temos $\mathbb{P}(X_1 = 1, X_2 = 0) = \theta(1-\theta)$ e multiplicando as probabilidades individuais temos $\mathbb{P}(X_1 = x_1)\mathbb{P}(X_2 = x_2) = \theta(1-\theta)$, o mesmo valor nos dois casos. ■

■ **Example 8.4 — Moedas dependentes.** Considere o mesmo contexto do exemplo anterior mas agora suponha que o modelo de probabilidade conjunta $\mathbb{P}(\omega)$ seja o seguinte:

ω	$\mathbb{P}(\omega)$
CC	$\theta\sqrt{\theta}$
$C\tilde{C}$	$\theta(1-\sqrt{\theta})$
$\tilde{C}C$	$\theta(1-\theta)$
$\tilde{C}\tilde{C}$	$(1-\theta)^2$
Total	1

Você deve verificar que, repetindo os cálculos da forma que fizemos no exemplo anterior, obtemos $\mathbb{P}(X_1 = 1) = \theta$, $\mathbb{P}(X_1 = 0) = 1-\theta$, $\mathbb{P}(X_2 = 1) = \theta(1-\theta+\sqrt{\theta})$ e $\mathbb{P}(X_2 = 0) = 1+\theta^2-\theta(1+\sqrt{\theta})$. Esta "moeda" não tem lançamentos independentes. Basta checar que para pelo menos uma das combinações de resultados nã vale a condição (8.1). Por exemplo,

$$\mathbb{P}(X_1 = 1, X_2 = 1) = \theta\sqrt{\theta} \neq \mathbb{P}(X_1 = 1)\mathbb{P}(X_2 = 1) = \theta^2(1-\theta+\sqrt{\theta})\,.$$

■

Só existe uma maneira de um conjunto de v.a.'s ser independente: satisfazendo a condição (8.1) para todo conjunto B_i da reta. Entretanto, existem infinitas maneiras delas serem dependentes. Assim, dizer que um conjunto de variáveis é dependente diz muito pouco acerca de como é esta dependência. Este assunto será explorado de forma mais substancial no capítulo ??.

■ **Example 8.5 — Independência de normais.** UK adult male heights are normally distributed with mean 70" and standard deviation 3". In the Netherlands, these figures are 71" and 3".

What is $\mathbb{P}(Y > X)$, where X and Y are the heights of randomly chosen UK and Netherlands males, respectively?

We have $X \sim N(70, 3^2)$ and $Y \sim N(71, 3^2)$. Then (as we will show in later lectures) $Y - X \sim N(1, 18)$.

$$\mathbb{P}(Y > X) = \mathbb{P}(Y - X > 0) = \mathbb{P}\left(\frac{Y - X - 1}{\sqrt{18}} > \frac{-1}{\sqrt{18}}\right) = 1 - \Phi(-1/\sqrt{18}),$$

since $\frac{(Y-X)-1}{\sqrt{18}} \sim N(0,1)$, and the answer is approximately 0.5931.

Now suppose that in both countries, the Olympic male basketball teams are selected from that portion of male whose hight is at least above 4" above the mean (which corresponds to the 9.1% tallest males of the country). What is the probability that a randomly chosen Netherlands player is taller than a randomly chosen UK player?

For the second part, we have

$$\mathbb{P}(Y > X \mid X \geq 74, Y \geq 75) = \frac{\int_{x=74}^{75} \phi_X(x)\, dx + \int_{x=75}^{\infty} \int_{y=x}^{\infty} \phi_Y(y)\phi_X(x)\, dy\, dx}{\int_{x=74}^{\infty} \phi_X(x)\, dx \int_{y=75}^{\infty} \phi_Y(y)\, dy},$$

which is approximately 0.7558. So even though the Netherlands people are only slightly taller, if we consider the tallest bunch, the Netherlands people will be much taller on average. ∎

8.3 Transformação de uma v.a.

Em probabilidade e análise de dados, é fundamental trabalhar com transformações de v.a.'s. Isto é, Y é uma v.a. transformada por uma função matemática a partir da v.a. X. Temos $Y = h(X)$. Por exemplo, $Y = h(X) = X^2$ ou $Y = h(X) = \sqrt{X} + \log(X)$. A v.a. Y passa a ter uma distribuição de probabilidade $f(y)$ sobre seus valores possíveis induzida pela distribuição de probabilidade de X (representada pela densidade $f(x)$) e pela transformação h. O cálculo de probabilidades de eventos associados com a ocorrência de $Y = h(X)$ pode ser reduzido ao cálculo de probabilidades de eventos associados com a ocorrência de X.

∎ **Example 8.6 — Quadrado aleatório.** Por exemplo, seja X o lado de um quadrado aleatório. A v.a. X é selecionada de uma distribuição Unif$(0,1)$. A probabilidade de que X caia num intervalo (a,b) contido em $(0,1)$ é o comprimento $b-a$ do intervalo. A área do quadrado aleatório formado com lado X é a v.a. $Y = X^2$.

Qual a distribuição de Y? Como os valores possíveis de X formam o intervalo $(0,1)$, os valores possíveis de $Y = X^2$ também formam o intervalo $(0,1)$. Entretanto, apesar do suporte das v.a.'s X e Y serem iguais, as probabilidades associadas da v.a. X e de Y são bem diferentes. Enquanto X tem uma distribuição $U(0,1)$, a distribuição de Y não é uniforme. Por exemplo, o intervalo $(0, 0.1)$ e o intervalo $(0.9, 1)$ possuem probabilidades diferentes sob a distribuição de Y embora eles tenham o mesmo comprimento. Outra consequência é que, em geral, teremos $\mathbb{E}(Y) = \mathbb{E}(X^2) \neq (\mathbb{E}(X))^2$. Em palavras mais memoráveis: a esperança de $g(X)$ não é g da esperança de X. Ou, em geral, $\mathbb{E}(g(X)) \neq g(\mathbb{E}(X))$.

A Figura 8.1 mostra do lado esquerdo cinco quadrados com lados aleatórios $U(0,1)$. O lado direito mostra o resultado de simular 10 mil valores da v.a. $X \sim U(0,1)$ e depois tomar o seu quadrado $Y = X^2$. O histograma desse 10 mil valores indica o formato da sua densidade de probabilidade, que é mostrada pela curva vermelha. Vamos ver como obter esta densidade na seção 8.4. Por ora, basta notar que os valores mais prováveis para a área Y são aqueles mais próximos de zero do que do outro extremo, 1. É um tanto surpreendente a falta de balanço entre os dois extremos, com a densidade bem concentrada ao redor de 0, gerando tipicamente quadrados de área próxima do mínimo. Existe uma chance muito menor de gerarmos quadrados de área grande, próxima de 1, o máximo possível.

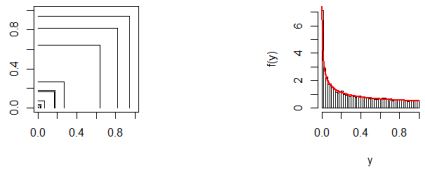


Figure 8.1: Esquerda: Cinco quadrados com lados aleatórios $U(0,1)$. Direita: Densidade $f(y)$ da área $Y = X^2$ de um quadrado de lado aleatório $X \sim U(0,1)$. A área não tem uma distribuição uniforme entre 0 e 1.

```
x = runif(10000)
y = x^2
xx = seq(0,1,by=0.001)
yy = 1/(2*sqrt(xx))

par(mfrow=c(1,2)); n = 10; set.seed(12)
s = runif(n)
plot(c(0,1), c(0,1), type="n", xlab="", ylab="",pty="s")
segments(c(s,rep(0,n)), c(rep(0,n),s), c(s,s), c(s,s))
hist(y, prob=T, n=40, ylab="f(y)", main="")
lines(xx, yy, lwd=2, col="red")
```

A distribuição de uma v.a. $Y = h(X)$, transformação matemática de outra v.a. X, é o resultado de combinar dois fatores: de um lado, a distribuição de probabilidade de X, representada pela densidade sua $f(x)$ (supondo X contínua); de outro, a função h que transforma a v.a. X. Veremos como estes dois fatores, a densidade $f(x)$ de X e a função h, são combinados para produzir a distribuição de Y na seção 8.4.

■

■ **Example 8.7 — Valor presente de um seguro de vida.** Outro exemplo relevante: seja T o tempo de vida a partir de um indivíduo adulto a partir de seus 22 anos. A distribuição de T é conhecida pelos atuários a partir de dados estatísticos da idade ao morrer de grandes populações. Por exemplo, a distribuição de Gompertz costuma ser um bom modelo para este tipo de dado. O indivíduo quer adquirir um seguro de vida que pague 100 mil reais a beneficiários (esposa e filhos) no momento de seu falecimento no futuro. Qual o valor que a seguradora deve cobrar hoje desse indivíduo para cobrir os seus custos com o pagamento futuro aos beneficiários? O pagamento é certo pois o invidivíduo vai falecer com certeza em algum momento futuro. Vamos assumir também que pelo menos um beneficiário estará vivo neste momento do falecimento[1]. Entretanto, o *momento desse pagamento* no futuro é aleatório. Vamos ignorar as despesas e lucros da seguradora, queremos apenas cobrir o custo do pagamento do benefício no momento de morte.

Isto complica as coisas pois dinheiro tem valor no tempo. Suponha que a taxa de juros anual estimada para os próximos anos seja δ. Por exemplo, $\delta = 0.05$ significa um taxa de juros de 5% ao ano. Um capital de C unidades auferindo juros anuais de δ acumula um total de $Ce^{\delta t}$ ao final de t anos. Para pagar 100 mil reais daqui a t anos, este tipo de cálculo financeiro permite à seguradora saber quando deveria cobrar hoje de um indivíduo que vai viver t anos. Para pagar 100 dentro de t anos ela deve cobrar Y de modo que Y posto a juros por t anos forneça 100. Isto é, $Ye^{\delta t} = 100$, ou $Y = 100e^{-\delta t}$. A quantidade $\exp(-\delta)$ é chamada de fator de desconto. O valor que é necessário

[1] Se todos os beneficiários tiverem morrido, a seguradora deverá fazer uma doação a uma entidade especificada. Assim, o pagamento é certo.

hoje para acumular um certo total no futuro (sob a taxa de juros δ) é chamado de *valor presente* deste total futuro. Por exemplo, com $\delta = 0.05$, o valor presente de 100 mil reais a serem pagos daqui a 27 anos é igual a $100e^{-0.05 * 27} = 25.9240$, ou 26 mil reais aproximadamente. Isto é, 100 mil reais daqui a 27 anos é equivalente, do ponto de vista financeiro, a 26 mil reais hoje (desde que a taxa de juros permaneça igual a 5% ao ano.

Dinheiro pode ser deslocado no tempo. Pode-se trazer dinheiro do futuro para ser consumido hoje. Pode-se guardar um dinheiro que temos hoje para ser usufrído apenas no futuro. Mas nestes dois casos, ao deslocar o dinheiro no tempo, não faz sentido mantê-lo com o mesmo valor *nominal* (a não ser que você vá deixá-lo debaixo do colchão). A taxa de juros é o valor que você precisa pagar se quiser trazer o dinheiro do futuro para consumir hoje. Ou é o valor que você vai receber por decidir não consumir seu dinheiro hoje (e portanto outra pessoa o usa pagando a você por este uso). Veja o lindo livro de Eduardo Gianetti, *O Valor do Amanhã*, para aprender que juros é o valor que damos ao tempo e que ele está em todos os aspectos da nossa vida, não apenas na vida financeira, mas desde a vida sexual até as decisões mais importantes tais como devo casar? com quem? que fazer da minha vida? [18].

Se o tempo de vida futuro do indivíduo fosse conhecido aos 22 anos de idade, o cálculo do valor presente dos 100 mil reais a serem pagos no futuro seria trivial. Mas sabendo quanto ainda vai viver, o indivíduo não teria interesse em fazer seguro. Se fosse viver por muito tempo, não precisaria fazer seguro. Se fosse viver pouco tempo, teria de pagar praticamente os mesmos 100 mil reias que deseja que sua família receba.

Mas a situação mais realista em que o tempo de vida futuro é desconhecido pela seguradora e por cada indivíduo envolvido exige que a seguradora pense estatisticamente. De um indivíduo com tempo de vida T, ela precisa cobrar pelo menos o valor presente $Y = 100e^{-\delta T}$. Mas como T é aleatório, o valor presente Y também é aleatório. Isto é, $Y = h(T)$ onde h é a função $h(t) = 100e^{-\delta t}$.

A seguradora calcula o valor esperado de Y e cobra este valor de todos os indivíduos na mesma situação. Isto é, a seguradora vai cobrar $\mathbb{E}(Y) = \mathbb{E}\left(100e^{-\delta T}\right)$ de cada indivíduo de 22 anos do sexo masculino que procurar seus serviços neste tipo de seguro. Ela sabe que, para alguns dos indivíduos, aqueles que viverão pouco tempo, ela estará cobrando menos do que deveria. Por outro lado, ela sabe que vai cobrar mais do que precisa de um indivíduo que viverá muitos anos, muito além do tempo médio de vida. Pela idéia frequentista de probabilidade, sabemos que o valor presente médio de um grande número de indivíduos é aproximadamente igual ao valor teórico $\mathbb{E}(Y) = \mathbb{E}\left(100e^{-\delta T}\right)$ e portanto as contas da seguradora devem ficar equilibradas.

■

8.4 Distribuição de $Y = h(X)$

Se $Y = h(X)$ e X é uma v.a. então Y também é uma v.a. Portanto, ela possui duas "listas": a de valores possíveis e a de probabilidades associadas. Com base nisso, podemosfazer cálculos de probabilidade relacionados com Y. Podemos querer saber, por exemplo, a probabilidade de ter um valor $Y \leq 2$ ou qualquer outro valor arbitrário: $\mathbb{P}(Y \leq 2)$ ou, em termos mais gerais, $\mathbb{F}_Y(y) = \mathbb{P}(Y \leq y)$ para qualquer y. Como Y é uma função de X e a distribuição de X é conhecida, devemos ser capazes de obter a distribuição de Y a partir daquela de X e da expressão da função h.

Como agora temos duas v.a.'s, X e $Y = h(X)$, vamos diferenciar mais explicitamente as duas distribuições de probabilidades. Por exemplo, as função de distribuição acumulada serão denotadas por $F_X(x)$ e $F_Y(y)$ e as densidades por $f_X(x)$ e $f_Y(y)$. Assim, $f_X(x)$ é o valor da densidade da v.a. X no ponto x da reta real. Se neste instante parece redundante ter os sub-índice e o argumento dessas funções iguais, veja abaixo a utilidade dessa notação.

Definition 8.4.1 — Distribuição de $Y = h(X)$. A função de distribuição acumulada $\mathbb{F}_Y(y)$ de $Y = h(X)$ é definida em termos de $\mathbb{F}_X(x)$ da seguinte forma

$$\begin{aligned}\mathbb{F}_Y(y) &= \mathbb{P}(Y \le y) = \mathbb{P}(h(X) \le y)\\ &= \mathbb{P}\left(X \in \{x \text{ tal que } h(x) \le y\}\right)\\ &= \mathbb{P}\left(X \in \{x \text{ tal que } x \le h^{-1}(y)\}\right)\\ &= \mathbb{F}_X(h^{-1}(y))\end{aligned}$$

Existem duas maneiras básicas de obter a distribuição de $Y = h(X)$.

(A) Obtenha a função de distribuição acumulada $\mathbb{F}_Y(y)$ de Y e a seguir, se Y for contínua, tome a derivada ou, se discreta, obtenha as alturas e os seus pontos de salto.

(B) Use o método geral e mais ou menos automático do jacobiano, a ser explicado mais a frente.

■ **Example 8.8 — Distribuição de área de □ aleatório.** Para entender o método (A), vamos começar com um exemplo. Considere a área $Y = X^2$ do nosso quadrado aleatório formado com um lado $X \sim U(0,1)$. Como $X \sim U(0,1)$, temos $\mathbb{F}_X(x) = x$ para $x \in (0,1)$. A v.a. Y é contínua e seu suporte é o intervalo $(0,1)$. Considere um dos valores possíveis de Y, um valor $y \in (0,1)$ arbitrário. Por exemplo, vamos fixar $y = 0.37$. Queremos $\mathbb{F}_Y(0.37) = \mathbb{P}(Y \le 0.37)$. O evento $[Y \le 0.37]$ é igual ao evento $[X \le \sqrt{0.37}]$. Por quê? Porque o evento

$$\begin{aligned}[Y \le 0.37] &= \{\omega \in \Omega \text{ tais que } Y(\omega) \le 0.37\}\\ &= \{\omega \in \Omega \text{ tais que } (X(\omega))^2 \le 0.37\}\\ &= \{\omega \in \Omega \text{ tais que } X(\omega) \le \sqrt{0.37}\}\\ &= [X \le \sqrt{0.37}]\end{aligned}$$

Se os conjuntos (ou eventos) $[Y \le 0.37]$ e $[X \le \sqrt{0.37}]$ são os mesmos, é que suas probabilidades também são as mesmas:

$$\mathbb{F}_Y(0.37) = \mathbb{P}(Y \le 0.37) = \mathbb{P}(X \le \sqrt{0.37}) = \mathbb{F}_X(\sqrt{0.37}) = \sqrt{0.37}\,.$$

Este cálculo pode ser refeito para um valor $y \in (0,1)$ qualquer e produz

$$\mathbb{F}_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(X \le \sqrt{y}) = \mathbb{F}_X(\sqrt{y}) = \sqrt{y}\,.$$

Para obter a densidade de probabilidade $f_Y(y)$, nós derivamos a sua função distribuição acumulada $F_Y(y)$:

$$f_Y(y) = \frac{d}{dy}\mathbb{F}_Y(y) = \frac{d}{dy}\sqrt{y} = \frac{1}{2\sqrt{y}}\,.$$

O gráfico dessa função densidade é a curva vermelha na Figura 8.1.

■

O uso desse resultado depende de nossa habilidade em escrever o evento $[Y \le y]$ em termos de um evento envolvendo a v.a. X que nós saibamos calcular a probabilidade. Quando a funaão h é inversível no conjunto suporte $\mathscr{S}_X$ isto é fácil. Por exemplo, imagine que h seja uma função crescente de X, como no exemplo $Y = h(X) = X^2$ com $X \sim U(0,1)$. Neste caso, $[Y \le y] = [h(X) \le y] = [X \le h^{-1}(y)]$ e portanto $\mathbb{F}_Y(y) = \mathbb{F}_X(h^{-1}(y))$.

■ **Example 8.9 — Distribuição de $Y = h(X) = e^{-0.05X}$.** Por exemplo, se $Y = h(X) = e^{-0.05X}$ então o suporte de Y está contido no intervalo $(0,\infty)$ e, para $y > 0$, temos $h^{-1}(y) = -\log(y)/0.05$ e

portanto

$$\begin{aligned}\mathbb{F}_Y(y) &= \mathbb{P}(Y \le y) = \mathbb{P}(h(X) \le y) \\ &= \mathbb{P}\left(X \le h^{-1}(y)\right) \\ &= \mathbb{P}(X \le -\log(y)/0.05) \\ &= \mathbb{F}_X(-\log(y)/0.05)\end{aligned}$$

■

Às vezes, a função $h(x)$ não é inversível no suporte de X. Por exemplo, suponha que $X \sim U(-1,1)$, um número real escolhido ao acaso no intervalo $(-1,1)$. Temos $\mathbb{F}_X(x) = (1+x)/2$ Se $Y = h(X) = X^2$, não existe a função inversa h^{-1} já que existem dois valores, $x = \sqrt{y}$ e $x = -\sqrt{y}$, tais que $h(x) = y$. Assim, devemos ter um pouco mais de cuidado para obter $\mathbb{F}_Y(y)$:

$$\begin{aligned}\mathbb{F}_Y(y) &= \mathbb{P}(Y \le y) = \mathbb{P}(X \le y) \\ &= \mathbb{P}\left(X \in \{x \text{ tal que } x^2 \le y\}\right) \\ &= \mathbb{P}(X \in (-\sqrt{y}, \sqrt{y})) \\ &= \mathbb{F}_X(\sqrt{y}) - \mathbb{F}_X(-\sqrt{y}) \\ &= \sqrt{y},\end{aligned}$$

para $y > 0$.

A regra (A) para obter a distribuição de Y é intuitiva e facilmente aplicada nos casos em que $h(x)$ é simples. Entretanto, existem casos em que a inversão não é simples ou é impossível. Por exemplo, se $X \sim U(-\pi,\pi)$ e $Y = h(X) = \log(1+\sin^2(x))$, esta inversão só é em segmentos do eixo x que depois devem ser colados de alguma forma. Ver o resultado geral em [22].

Existe uma razão adicional para estudarmos outro método, a regra (B), para obter para obter a distribuição de Y. Ele é um método mais geral e mais ou menos automático baseado no jacobiano da função $h(x)$. A importância dessa forma de cálculo é que ele pode ser estendido para o caso em que Y é função não de uma única v.a., mas de várias v.a.'s. Por exemplo, ele pode ser aplicado se $Y = XW + Z$ onde X, Z e W são v.a.'s. No caso univariado que estamos estudando neste capítulo, em que $Y = h(X)$, o método (B) fornece uma regra geral quando X e Y são v.a.'s contínuas. o jacobiano é simplesmente a derivada da função inversa.

Theorem 8.4.1 — Densidade de $Y = h(X)$. Seja $Y = h(X)$. Suponha que no suporte de X a função h seja inversível com $g = h^{-1}$ e portanto $x = g(y)$. Então a densidade de Y no ponto y é dada por

$$f_Y(y) = f_X(g(y)) \left|\frac{dg(y)}{dy}\right| = f_X(g(y)) \left|\frac{dh^{-1}(y)}{dy}\right| \tag{8.15}$$

O segundo fator do lado direito é o valor absoluto do jacobiano dessa transformação.

■ **Example 8.10 — Disco aleatório.** O raio de um disco é escolhido ao acaso de acordo com uma distribuição uniforme no intervalo $(0,1)$. O disco aleatório produzido é varrido completamente. Seja $A = \pi R^2$ a área desse disco aleatório gerado. Obtenha a lista de valores possíveis da v.a. A e sua densidade de probabilidade. A seguir obtenha sua esperança. É verdade que $\mathbb{E}(A) = \pi\left(\mathbb{E}(R)\right)^2$?

Como $R \in (0,1)$ temos a densidade $f_R(r) = 1$ para todo $r \in (0,1)$. A lista de valores possíveis de A é dada por $A \in (0,\pi)$. Como $A = h(R) = \pi R^2$, temos $R = h^{-1}(A) = \sqrt{\frac{A}{\pi}}$. A densidade de A neste intervalo é dada por (8.15):

$$f_A(a) = f_R\left(\sqrt{\frac{A}{\pi}}\right)\left|\frac{d}{dA}\sqrt{\frac{A}{\pi}}\right| = 1 \times \frac{1}{2\sqrt{\pi A}}$$

Como temos a densidade de A podemos calcular

$$\mathbb{E}(A) = \int_0^\pi A \, \frac{1}{\sqrt{\pi A}} \mathrm{d}A = \frac{1}{2\sqrt{\pi}} \left(\frac{1}{2\pi} 2/3 A^{3/2} \, |_0^\pi \right) = \frac{\pi}{3} .$$

Como $\mathbb{E}(R) = 1/2$, temos $\pi/3 = \mathbb{E}(A) \neq \pi/4 = \pi\,(\mathbb{E}(R))^2$. A área esperada é maior que a área de um disco que usa um raio igual ao raio esperado. ■

■ **Example 8.11 — Intensidade da luz.** A lei do inverso do quadrado descreve a intensidade da luz a diferentes distâncias de uma fonte de luz. A intensidade da luz numa certa posição é inversamente proporcional ao quadrado da distância entre esta posição e a fonte luminosa. Isto significa que à medida que a distância de uma fonte de luz aumenta, a intensidade da luz é igual a c/X^2 onde c é uma constante que depende do tipo de fonte e X é a distância até a fonte. Esta lei vale para a luz visível bem como para a parte não visível do espectro (ondas de rádio, microondas, luz infravermelha e ultravioleta, raios X e raios gama).

Suponha que um sensor seja colocado ao acaso de acordo ao longo de um corredor linear de comprimento 12 metros. Sua posição é tal que ele tende a ficar posicionado longe do centro do corredor. Mais especificamente, se X é a distância aleatória entre o sensor e o centro do corredor então a densidade da v.a. X é $f(x) = 0.08x$ para $x \in (1,6)$. Uma fonte luminosa está no centro desse corredor. A intensidade captada pelo sensor é $I = 25/X^2$. Obtenha a esperança e a distribuição de I.

A distância entre o sensor e o centro segue uma distribuição com densidade $f(x) = 0.08x$ para $x \in (0,5)$ e assim

$$\begin{aligned} \mathbb{E}(I) &= \int_1^6 \frac{35}{x^2} \, f(x) \, \mathrm{d}x = (35 \times 0.08) \int_1^6 \frac{x}{x^2} \, \mathrm{d}x \\ &= 2.8 \int_1^6 \frac{1}{x} \, \mathrm{d}x = 2.8 \log(x) \, |_1^6 = 2.8(\log(6) - \log(1)) = 2.8 \log(6) \end{aligned}$$

A densidade de probabilidade da intensidade da luz é obtida a partir da densidade de X. Como $I = h(X) = 25/X^2$, a função inversa é $g(y) = h^{-1}(y) = 5/\sqrt{y}$ e portanto

$$f_I(i) = f_X\left(\frac{5}{\sqrt{i}}\right) \left| \frac{d}{di} 5/\sqrt{i} \right| = 0.08 \frac{5}{\sqrt{i}} \left| -\frac{5}{2} \frac{1}{i\sqrt{i}} \right| = \frac{1.0}{i^2}$$

para $i \in (0.694, 25)$. ■

8.5 Esperança de $Y = h(X)$

Muitas vezes, apenas o valor esperado de $Y = h(X)$ é suficiente, sem precisar conhecer toda a distribuição de Y. Felizmente, o cálculo de $\mathbb{E}(Y) = \mathbb{E}(h(X))$ é relativamente simples e não requer o conhecimento da distribuição de Y. Assim, podemos passar sem os cálculos da seção anterior quando o interesse for apenas em $\mathbb{E}(Y) = \mathbb{E}(h(X))$..

Theorem 8.5.1 — Esperança de $Y = h(X)$. Suponha que X seja uma v.a. com suporte $\mathscr{S}$ e que $Y = h(X)$. Se X for uma v.a. contínua com densidade $f(x)$,

$$\mathbb{E}(h(X)) = \int_{\mathscr{S}} h(x) f(x) \mathrm{d}x .$$

Se X for uma v.a. discreta com função de probabilidade $\mathbb{P}(X = x_i)$,

$$\mathbb{E}(h(X)) = \sum_{x_i \in \mathscr{S}} h(x_i) \mathbb{P}(X = x_i) .$$

Por exemplo, se $Y = h(X) = X^2$, então

$$\mathbb{E}(h(X)) = \mathbb{E}(X^2)) = \int_{\mathbb{R}} x^2 f(x) \mathrm{d}x$$

Terminar o cálculo depende de conhecer a densidade $f(x)$ mas agora este cálculo é um problema puramente matemático.

Se X for uma v.a.discreta com valores possíveis $\{x_1, x_2, \ldots\}$ e probabilidades associadas $\{p(x_1), p(x_2), \ldots\}$ temos

$$\mathbb{E}(h(X)) = \sum_{x_i} h(x_i) p(x_i)$$

Por exemplo, se $h(x) = \sqrt{x}$ e se $X \sim \text{Poisson}(\lambda = 2.3)$ então

$$\mathbb{E}(h(X)) = \mathbb{E}(\sqrt{X}) = \sum_{k=0}^{\infty} \sqrt{k} \frac{2.3^k e^{-2.3}}{k!}$$

Não existe fórmula conhecida para esta série, ela tem de ser calculada numericamente.

8.6 Probabilidades e variáveis aleatórias indicadoras

Toda probabilidade pode ser vista como a esperança de uma v.a. indicadora. Seja A um evento qualquer e $I_A(\omega)$ uma v.a. binária dada por

$$I_A(\omega) = \begin{cases} 1, & \text{se } \omega \in A \\ 0, & \text{caso contrário} \end{cases}$$

Esta v.a. é apenas a função indicadora de que o evento A ocorreu. Para cada resultado do experimento ela retorna 1 se A ocorreu e 0 se A náo ocorreu.

O evento $[I_A = 1]$ é o conjunto de elementos $\omega \in \Omega$ tais que $I_A(\omega) = 1$. Mas isto é exatamente o conjunto de $\omega \in A$. Ou seja, temos a igualdade $[I_A = 1] = A$ de dois eventos em Ω. Portanto, a probabilidade dos dois eventos é a mesma e $\mathbb{P}(I_A = 1) = \mathbb{P}(A)$.

Como a v.a. I_A é binária temos

$$\mathbb{E}(I_A) = 1 \times \mathbb{P}(I_A = 1) + 0 \times \mathbb{P}(I_A = 0) = \mathbb{P}(I_A = 1) = \mathbb{P}(A)$$

Assim, para todo evento A, podemos escrever sua proabilidade $\mathbb{P}(A)$ como a esperança $\mathbb{E}(I_A)$ da v.a. aleatória indicadora da ocorrência de A. Parece um jeito muito complicado de obter uma probabilidade mas este é um truque muito útil em situações um pouco mais complicadas. A utilidade vem da possibilidade de usar algumas propriedades conhecidas do operador $\mathbb{E}$, e isto torna às vezes mais fácil de manipular que o operador $\mathbb{P}$. Uma dessas propriedades é a linearidade da esperança: $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.

Eventos podem ser definidos em termos de variáveis aleatórias. Por exemplo,

- $[X = 3] = \{\omega \in \Omega \text{ tais que } X(\omega) = 3\}$
- $[X \leq 3] = \{\omega \in \Omega \text{ tais que } X(\omega) \leq 3\}$
- $[1 \leq X \leq 3] = \{\omega \in \Omega \text{ tais que } X(\omega) \in [1,3]\}$
- se $S \subset \mathbb{R}$ temos $[X \in S] = \{\omega \in \Omega \text{ tais que } X(\omega) \in S\}$

Assim, podemos estender a ideia de escrever como esperança a probabilidade de um evento associado a uma v.a. Por exemplo, seja A o evento $A = [X = 3]$ associado com a v.a. X. Então

$$\mathbb{P}(X = 3) = \mathbb{P}(A) = \mathbb{E}(I_A) = \mathbb{E}(I_{[X=3]})$$

De maneira geral, se $S \subset \mathbb{R}$ temos $\mathbb{P}(X \in S) = \mathbb{E}(I_{[X \in S]})$.

Vamos adotar outra notação alternativa:

$$I_{[X\in S]} = I_S(X)$$

Veja que $I_S(X)$ é uma v.a. Ela recebe como input um resultado $\omega \in \Omega$ e retorna um valor binário $I_{[X\in S]}(\omega) = I_S(X(\omega))$. Esta v.a. é uma transformação $h(X)$ de X. De fato, temos

$$I_S(X(\omega)) = h(X(\omega)) = \begin{cases} 1, & \text{se } X(\omega) \in S \\ 0, & \text{caso contrário} \end{cases}$$

Portanto, se tivermos a densidade $f(x)$ de uma v.a. contínua X, usando o que aprendemos no início deste texto, podemos escrever

$$\begin{aligned} \mathbb{P}(X \in S) &= \mathbb{E}(I_S(X)) \\ &= \mathbb{E}(h(X)) \\ &= \int_{\mathbb{R}} h(x)f(x)\mathrm{d}x \\ &= \int_{\mathbb{R}} 1_S(x)f(x)\mathrm{d}x \\ &= \int_S 1\times f(x)\mathrm{d}x + \int_{\mathbb{R}-S} 0\times f(x)\mathrm{d}x \\ &= \int_S f(x)\mathrm{d}x \end{aligned}$$

Eu sei que parece estarmos dando voltas em torno do mesmo ponto mas, acredite, isto é útil. Por exemplo, suponha que $X \sim N(0,1)$, uma gaussiana padrão que possui densidade

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$$

Suponha que, por alguma razão queiramos calcular a probabilidade de que $exp(-|X|) > |X|$. Manipulação algébrica mostra que $exp(-|x|) > x$ se, e somente se, $x \in (-a,a)$ onde $a \approx 0.5671$. Assim,

$$\mathbb{P}(\exp(-|X|) > |X|) = \int_{(-a,a)} f(x)\mathrm{d}x = \int_{\mathbb{R}} h(x)f(x)\mathrm{d}x = \mathbb{E}(h(X))$$

onde $h(X) = I_S(X)$ e S é o evento $[\exp(-|X|) > |X|]$.

8.7 Multiple random variables - OPTIONAL READING

Suppose $X_1, X_2, \cdots, X_n$ are random variables with joint pdf f. Let

$$\begin{aligned} Y_1 &= r_1(X_1, \cdots, X_n) \\ Y_2 &= r_2(X_1, \cdots, X_n) \\ &\vdots \\ Y_n &= r_n(X_1, \cdots, X_n). \end{aligned}$$

For example, we might have $Y_1 = \frac{X_1}{X_1+X_2}$ and $Y_2 = X_1 + X_2$.

Let $R \subseteq \mathbb{R}^n$ such that $\mathbb{P}((X_1, \cdots, X_n) \in R) = 1$, i.e. R is the set of all values (X_i) can take.

Suppose S is the image of R under the above transformation, and the map $R \to S$ is bijective. Then there exists an inverse function

$$\begin{aligned} X_1 &= s_1(Y_1, \cdots, Y_n) \\ X_2 &= s_2(Y_1, \cdots, Y_n) \\ &\vdots \\ X_n &= s_n(Y_1, \cdots, Y_n). \end{aligned}$$

For example, if X_1, X_2 refers to the coordinates of a random point in Cartesian coordinates, Y_1, Y_2 might be the coordinates in polar coordinates.

Definition 8.7.1 — Jacobian determinant. Suppose $\frac{\partial s_i}{\partial y_j}$ exists and is continuous at every point $(y_1, \cdots, y_n) \in S$. Then the *Jacobian determinant* is

$$J = \frac{\partial(s_1, \cdots, s_n)}{\partial(y_1, \cdots, y_n)} = \det \begin{pmatrix} \frac{\partial s_1}{\partial y_1} & \cdots & \frac{\partial s_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial s_n}{\partial y_1} & \cdots & \frac{\partial s_n}{\partial y_n} \end{pmatrix}$$

Take $A \subseteq \mathbb{R}$ and $B = r(A)$. Then using results from IA Vector Calculus, we get

$$\begin{aligned} \mathbb{P}((X_1, \cdots, X_n) \in A) &= \int_A f(x_1, \cdots, x_n)\, dx_1 \cdots dx_n \\ &= \int_B f(s_1(y_1, \cdots y_n), s_2, \cdots, s_n)|J|\, dy_1 \cdots dy_n \\ &= \mathbb{P}((Y_1, \cdots Y_n) \in B). \end{aligned}$$

So

Proposition 8.7.1 $(Y_1, \cdots, Y_n)$ has density

$$g(y_1, \cdots, y_n) = f(s_1(y_1, \cdots, y_n), \cdots s_n(y_1, \cdots, y_n))|J|$$

if $(y_1, \cdots, y_n) \in S$, 0 otherwise.

■ **Example 8.12** Suppose (X, Y) has density

$$f(x,y) = \begin{cases} 4xy & 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

We see that X and Y are independent, with each having a density $f(x) = 2x$.

Define $U = X/Y$, $V = XY$. Then we have $X = \sqrt{UV}$ and $Y = \sqrt{V/U}$.

The Jacobian is

$$\det \begin{pmatrix} \partial x/\partial u & \partial x/\partial v \\ \partial y/\partial u & \partial y/\partial v \end{pmatrix} = \det \begin{pmatrix} \frac{1}{2}\sqrt{v/u} & \frac{1}{2}\sqrt{u/v} \\ -\frac{1}{2}\sqrt{v/u^3} & \frac{1}{2}\sqrt{1/uv} \end{pmatrix} = \frac{1}{2u}$$

Alternatively, we can find this by considering

$$\det \begin{pmatrix} \partial u/\partial x & \partial u/\partial y \\ \partial v/\partial x & \partial u/\partial y \end{pmatrix} = 2u,$$

and then inverting the matrix. So

$$g(u,v) = 4\sqrt{uv}\sqrt{\frac{v}{u}}\frac{1}{2u} = \frac{2v}{u},$$

if (u,v) is in the image S, 0 otherwise. So

$$g(u,v) = \frac{2v}{u} I[(u,v) \in S].$$

Since this is not separable, we know that U and V are not independent. ∎

In the linear case, life is easy. Suppose

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = A \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} = A\mathbf{X}$$

Then $\mathbf{X} = A^{-1}\mathbf{Y}$. Then $\frac{\partial x_i}{\partial y_j} = (A^{-1})_{ij}$. So $|J| = |\det(A^{-1})| = |\det A|^{-1}$. So

$$g(y_1, \cdots, y_n) = \frac{1}{|\det A|} f(A^{-1}\mathbf{y}).$$

■ **Example 8.13** Suppose X_1, X_2 have joint pdf $f(x_1, x_2)$. Suppose we want to find the pdf of $Y = X_1 + X_2$. We let $Z = X_2$. Then $X_1 = Y - Z$ and $X_2 = Z$. Then

$$\begin{pmatrix} Y \\ Z \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = A\mathbf{X}$$

Then $|J| = 1/|\det A| = 1$. Then

$$g(y,z) = f(y-z, z)$$

So

$$g_Y(y) = \int_{-\infty}^{\infty} f(y-z,z)\, dz = \int_{-\infty}^{\infty} f(z, y-z)\, dz.$$

If X_1 and X_2 are independent, $f(x_1,x_2) = f_1(x_1) f_2(x_2)$. Then

$$g(y) = \int_{-\infty}^{\infty} f_1(z) f_2(y-z)\, dz.$$

∎

Non-injective transformations

We previously discussed transformation of random variables by injective maps. What if the mapping is not? There is no simple formula for that, and we have to work out each case individually.

■ **Example 8.14** Suppose X has pdf f. What is the pdf of $Y = |X|$?

We use our definition. We have

$$\mathbb{P}(|X| \in x(a,b)) = \int_a^b f(x) + \int_{-b}^{-a} f(x)\, dx = \int_a^b (f(x) + f(-x))\, dx.$$

So

$$f_Y(x) = f(x) + f(-x),$$

which makes sense, since getting $|X| = x$ is equivalent to getting $X = x$ or $X = -x$. ∎

■ **Example 8.15** Suppose $X_1 \sim \mathscr{E}(\lambda), X_2 \sim \mathscr{E}(\mu)$ are independent random variables. Let $Y = \min(X_1, X_2)$. Then

$$\begin{aligned}\mathbb{P}(Y \geq t) &= \mathbb{P}(X_1 \geq t, X_2 \geq t)\\ &= \mathbb{P}(X_1 \geq t)\mathbb{P}(X_2 \geq t)\\ &= e^{-\lambda t}e^{-\mu t}\\ &= e^{-(\lambda+\mu)t}.\end{aligned}$$

So $Y \sim \mathscr{E}(\lambda+\mu)$. ■

Given random variables, not only can we ask for the minimum of the variables, but also ask for, say, the second-smallest one. In general, we define the *order statistics* as follows:

Definition 8.7.2 — Order statistics. Suppose that $X_1, \cdots, X_n$ are some random variables, and $Y_1, \cdots, Y_n$ is $X_1, \cdots, X_n$ arranged in increasing order, i.e. $Y_1 \leq Y_2 \leq \cdots \leq Y_n$. This is the *order statistics*.

We sometimes write $Y_i = X_{(i)}$.

Assume the X_i are iid with cdf F and pdf f. Then the cdf of Y_n is

$$\mathbb{P}(Y_n \leq y) = \mathbb{P}(X_1 \leq y, \cdots, X_n \leq y) = \mathbb{P}(X_1 \leq y) \cdots \mathbb{P}(X_n \leq y) = F(y)^n.$$

So the pdf of Y_n is

$$\frac{d}{dy}F(y)^n = nf(y)F(y)^{n-1}.$$

Also,

$$\mathbb{P}(Y_1 \geq y) = \mathbb{P}(X_1 \geq y, \cdots, X_n \geq y) = (1-F(y))^n.$$

What about the joint distribution of Y_1, Y_n?

$$\begin{aligned}G(y_1, y_n) &= \mathbb{P}(Y_1 \leq y_1, Y_n \leq y_n)\\ &= \mathbb{P}(Y_n \leq y_n) - \mathbb{P}(Y_1 \geq y_1, Y_n \leq y_n)\\ &= F(y_n)^n - (F(y_n) - F(y_1))^n.\end{aligned}$$

Then the pdf is

$$\frac{\partial^2}{\partial y_1 \partial y_n}G(y_1, y_n) = n(n-1)(F(y_n) - F(y_1))^{n-2}f(y_1)f(y_n).$$

We can think about this result in terms of the multinomial distribution. By definition, the probability that $Y_1 \in [y_1, y_1+\delta)$ and $Y_n \in (y_n-\delta, y_n]$ is approximately $g(y_1, y_n)$.

Suppose that δ is sufficiently small that all other $n-2$ X_i's are very unlikely to fall into $[y_1, y_1+\delta)$ and $(y_n-\delta, y_n]$. Then to find the probability required, we can treat the sample space as three bins. We want exactly one X_i to fall into the first and last bins, and $n-2$ X_i's to fall into the middle one. There are $\binom{n}{1,n-2,1} = n(n-1)$ ways of doing so.

The probability of each thing falling into the middle bin is $F(y_n) - F(y_1)$, and the probabilities of falling into the first and last bins are $f(y_1)\delta$ and $f(y_n)\delta$. Then the probability of $Y_1 \in [y_1, y_1+\delta)$ and $Y_n \in (y_n-\delta, y_n]$ is

$$n(n-1)(F(y_n) - F(y_1))^{n-2}f(y_1)f(y_n)\delta^2,$$

and the result follows.

We can also find the joint distribution of the order statistics, say g, since it is just given by

$$g(y_1,\cdots y_n) = n!f(y_1)\cdots f(y_n),$$

if $y_1 \le y_2 \le \cdots \le y_n$, 0 otherwise. We have this formula because there are $n!$ combinations of $x_1,\cdots,x_n$ that produces a given order statistics $y_1,\cdots,y_n$, and the pdf of each combination is $f(y_1)\cdots f(y_n)$.

In the case of iid exponential variables, we find a nice distribution for the order statistic.

■ **Example 8.16** Let $X_1,\cdots,X_n$ be iid $\mathscr{E}(\lambda)$, and $Y_1,\cdots,Y_n$ be the order statistic. Let

$$\begin{aligned} Z_1 &= Y_1 \\ Z_2 &= Y_2 - Y_1 \\ &\vdots \\ Z_n &= Y_n - Y_{n-1}. \end{aligned}$$

These are the distances between the occurrences. We can write this as a $\mathbf{Z} = A\mathbf{Y}$, with

$$A = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

Then $\det(A) = 1$ and hence $|J| = 1$. Suppose that the pdf of $Z_1,\cdots,Z_n$ is, say h. Then

$$\begin{aligned} h(z_1,\cdots,z_n) &= g(y_1,\cdots,y_n)\cdot 1 \\ &= n!f(y_1)\cdots f(y_n) \\ &= n!\lambda^n e^{-\lambda(y_1+\cdots+y_n)} \\ &= n!\lambda^n e^{-\lambda(nz_1+(n-1)z_2+\cdots+z_n)} \\ &= \prod_{i=1}^{n}(\lambda i)e^{-(\lambda i)z_{n+1-i}} \end{aligned}$$

Since h is expressed as a product of n density functions, we have

$$Z_i \sim \mathscr{E}((n+1-i)\lambda).$$

with all Z_i independent. ■

8.8 Transformações não-explícitas algebricamente

Suponha que um vetor aleatório m-dimensional $\mathbf{Y} = (Y_1,\ldots,Y_m)$ seja o resultado de uma transformação matemática de um outro vetor aleatóorio n-dimensional $\mathbf{X} = (X_1,X_2,\ldots,X_n)$. Isto é, cada componente Y_j é uma função matemática g_j do vetor $\mathbf{X}$ de modo que:

$$\begin{aligned} Y_1 &= g_1(X_1,\ldots,X_n) \\ Y_2 &= g_2(X_1,\ldots,X_n) \\ &\vdots \\ Y_m &= g_m(X_1,\cdots,X_n). \end{aligned}$$

Observe que podemos ter $n \neq m$.

8.8.1 Distribuição dos autovalores de matriz aleatória

Algumas vezes, as variáveis em $\mathbf{Y}$ são o resultado de uma função mas não existe uma fórmula matemática simples conectando $\mathbf{Y}$ e $\mathbf{X}$. Por exemplo, gere n^2 v.a.'s i.i.d. $N(0,1)$ formando o vetor $\mathbf{X}$. Organize os n^2 valores numa matriz quadrada $\mathbf{M}$. A seguir, obtenha os n autovalores aleatórios $\mathbf{Y} = (Y_1, \ldots, Y_n)$ da matriz aleatória $\mathbf{M}$. Estes autovalores são as n raízes do polinômio característico $p(\lambda) = \det(\mathbf{I} - lambda\mathbf{M})$ e portanto função das n^2 v.a.'s em $\mathbf{X}$. mas não existe uma fórmula algébrica fechada em que imputamos os valores $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ e a saída sejam os autovalores. Temos algoritmos clássicos que produzem aproximações numéricas para os autovalores. mas, sendo raízes de polinômios de grau n, sabemos que não existem fórmulas para obter as raízes no caso geral de polinômios de grau $n > 5$.

8.8.2 Distribuições em alta dimensão

`https://bit.ly/2Gw8NGt` blog como texto abaixo.

John Cook recentemente escreveu um post interessante (em `https://bit.ly/2GuF1le`) sobre vetores aleatórios e projeções aleatórias. No post, ele afirma dois fatos surpreendentes da geometria de alta dimensão e dá alguma intuição para o segundo fato. Neste post, vou fornecer o código R para demonstrar os dois.

Fato 1: Dois vetores escolhidos aleatoriamente em um espaço de alta dimensão são muito provavelmente quase ortogonais.

Cook não discute esse fato, pois é "bem conhecido". Deixe-me demonstrar isso empiricamente. Abaixo, a primeira função gera um vetor unitário p-dimensional uniformemente ao acaso. A segunda função aceita dois vetores p-dimensionais, x1 e x2, e calcula o ângulo entre eles. (Para mais detalhes, consulte a postagem no blog de Cook.)

```
genRandomVec <- function(p) {
    x <- rnorm(p)
    x / sqrt(sum(x^2))
}

findAngle <- function(x1, x2) {
    dot_prod <- sum(x1 * x2) / (sqrt(sum(x1^2) * sum(x2^2)))
    acos(dot_prod)
}
```

Em seguida, usamos a função replicate para gerar 100.000 pares de vetores de 10.000 dimensões e plotar um histograma dos ângulos que eles criam:

```
simN <- 100000  # no. of simulations
p <- 10000
set.seed(100)
angles <- replicate(simN, findAngle(genRandomVec(p), genRandomVec(p)))
angles <- angles / pi * 180  # convert from radians to degrees
hist(angles)
```

Observe a escala do eixo x: os ângulos estão muito próximos de 90 graus, como reivindicado. Esse fenômeno só acontece para dimensões "altas". Se mudarmos o valor de p acima para 2, obtemos um histograma muito diferente:

Quão "alta" a dimensão tem que ser antes de vermos este fenômeno? Bem, depende de quão bem agrupados queremos que os ângulos sejam em torno de 90 graus. O histograma abaixo é a mesma simulação, mas para p = 20 (observe a escala do eixo x mais larga):

Parece que a curva em forma de sino já começa a aparecer com p = 3!

Fato 2: Gere 10.000 vetores aleatórios em 20.000 espaços dimensionais. Agora, gere outro vetor aleatório nesse espaço. Então o ângulo entre este vetor e sua projeção no vão dos primeiros 10.000 vetores é muito provável que seja muito próximo de 45 graus.

Cook apresenta uma explicação intuitiva muito legal deste fato que eu recomendo. Aqui, apresento evidências de simulação do fato.

A dificuldade nesta simulação é computar a projeção de um vetor no espaço de muitos vetores. Pode-se mostrar que a projeção de um vetor v no espaço de coluna de uma matriz (completa) é dada por $\text{proj}_A(v) = A(A^TA)^{-1}A^Tv$. Em nosso problema, A é uma matriz de 20.000×10.000, então a computação de $(A^TA)^{-1}$ vai demorar proibitivamente.

Não sei outra maneira de calcular a projeção de um vetor no intervalo de outros vetores. Felizmente, com base em minhas simulações no Fato 1, esse fenômeno provavelmente também irá contribuir para dimensões muito menores!

Primeiro, vamos escrever duas funções: uma que tenha um vetor v e uma matriz A e retorne a projeção de v para o espaço de coluna de A:

```
projectVec <- function(v, A) {
    A %*% solve(t(A) %*% A) %*% t(A) %*% v
}
```

e uma função que executa a simulação. Aqui, p é a dimensionalidade de cada um dos vetores, e eu suponho que estamos olhando para o espaço de vetores p / 2:

```
simulationRun <- function(p) {
    A <- replicate(p/2, genRandomVec(p))
    v <- genRandomVec(p)
    proj_v <- projectVec(v, A)
    findAngle(proj_v, v)
}
```

O código abaixo executa 10.000 simulações para p = 20, demorando cerca de 2 segundos no meu laptop:

```
simN <- 10000 # no. of simulations
p <- 20     # dimension of the vectors
set.seed(54)
angles <- replicate(simN, simulationRun(p))
angles <- angles / pi * 180  # convert from radians to degrees
hist(angles, breaks = seq(0, 90, 5))
abline(v = 45, col = "red", lwd = 3)
```

Já podemos ver o agrupamento em torno de 45 graus:

A simulação para p = 200 leva menos de 2 minutos no meu laptop, e vemos um agrupamento mais apertado em torno de 45 graus (observe a escala do eixo x).

9. Variance and Inequalities

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 aletório 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 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?

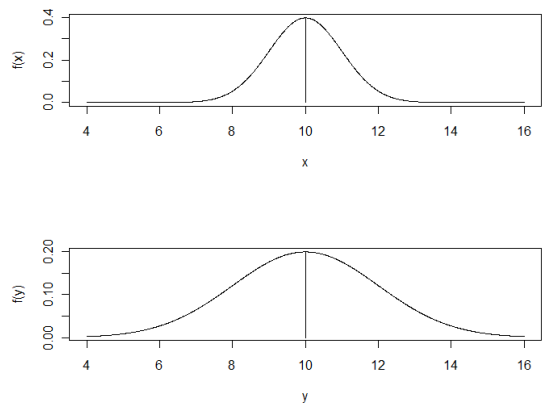


Figure 9.1: Densidades de X e Y com mesmo valor esperado: $\mathbb{E}(X) = \mathbb{E}(Y) = 10$.

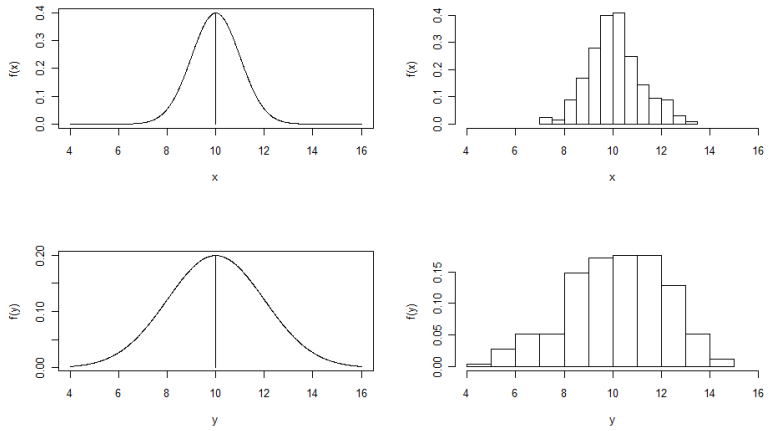


Figure 9.2: Histogramas de amostras e densidades de X e Y com mesmo valor esperado: $\mathbb{E}(X) = \mathbb{E}(Y) = 10$. Qual das amostras varia mais em torno do seu valor esperado?

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 masi 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

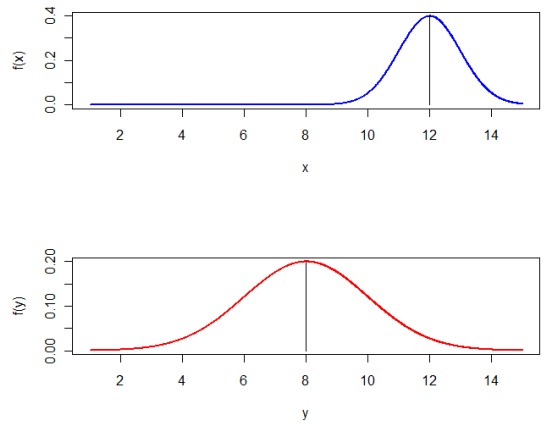


Figure 9.3: Densidades de X e Y com diferentes valores esperados: $\mathbb{E}(X) \neq \mathbb{E}(Y)$. Qual das amostras varia mais em torno do seu valor esperado?

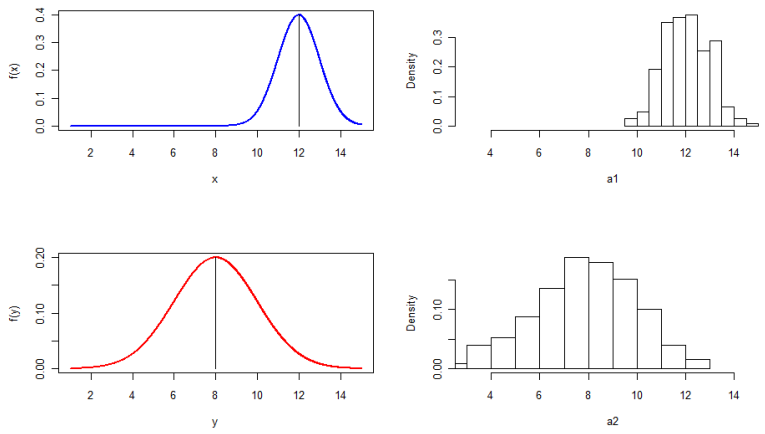


Figure 9.4: Histogramas de amostras e densidades de X e Y com diferentes valores esperados: $\mathbb{E}(X) \neq \mathbb{E}(Y)$. Qual das amostras varia mais em torno do seu valor esperado?

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.

[1] Estamos usando duas densidades gama aqui, com $\alpha = \beta$.

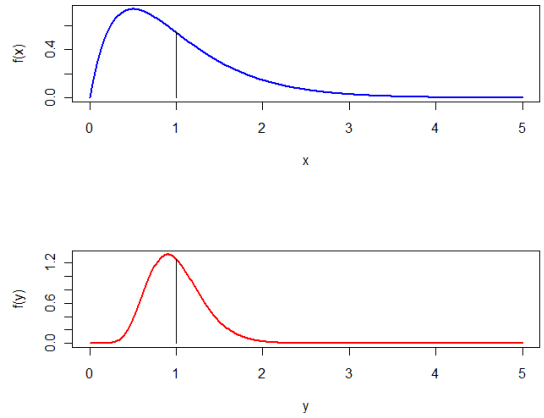


Figure 9.5: Densidades assimétricas de X e Y mas mesmo valor esperado: $\mathbb{E}(X) = \mathbb{E}(Y) = 1$.

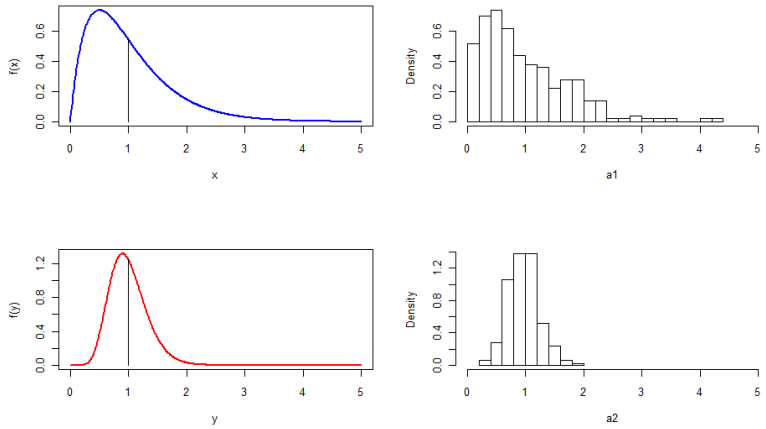


Figure 9.6: Histogramas de amostras e densidades assimétricas de X e Y com mesmo valor esperado: $\mathbb{E}(X) = \mathbb{E}(Y) = 1$. Qual das amostras varia mais em torno do seu valor esperado?

Novamente considerando que éreas 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 funccões 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 trono de $\mathbb{E}(X) = 1.1$, indicando que valores da v.a. X tendem a se afastar pouco de seu valor

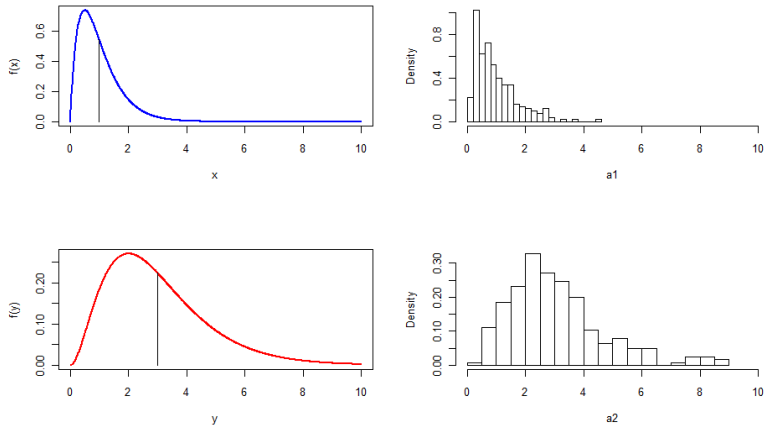


Figure 9.7: Histogramas e densidades assimétricas de X e Y com $1 = \mathbb{E}(X) \neq \mathbb{E}(Y) = 3$. Qual das amostras varia mais em torno do seu valor esperado?

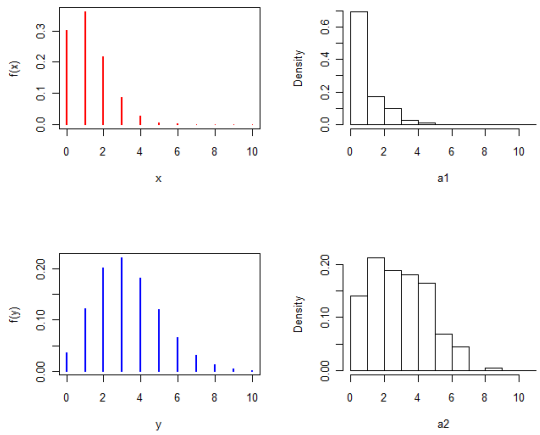


Figure 9.8: Histogramas e funções de probabilidade de duas Poisson com $1.2 = \mathbb{E}(X) \neq \mathbb{E}(Y) = 3.3$. Qual das amostras varia mais em torno do seu valor médio?

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 seráo 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 as 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 $\mathscr{S} = \{|1-1.63|, |2-1.63|, |3-1.63|, |4-1.63|\} = \{|-0.63|, 0.37, 1.37, 2.37\}$. Vamos denotar por $d \in \mathscr{S}$ 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 $\mathscr{S} = (0, \infty)$ e densidade $f(y) = 3\exp(-3y)$. Temos

$$\mathbb{E}(Y) = \int_0^\infty y\, f(y)\, \mathrm{d}y = \int_0^\infty y\, 3e^{-3y}\, \mathrm{d}y = \frac{1}{3}$$

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 $f(y)$:

$$\begin{aligned}\mathbb{V}(Y) &= \mathbb{E}(\,|Y-\mu|^2\,) = \mathbb{E}(\,|Y-1/3|^2\,) \\ &= \int_0^\infty |y-1/3|^2\, f(y)\, \mathrm{d}y \\ &= \int_0^\infty |y-1/3|^2\, 3e^{-3y}\, \mathrm{d}y \\ &= 1/3^2 \qquad \text{após manipulações de cálculo.}\end{aligned}$$

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 $\mathscr{S}$ e valor esperado μ, a

variância $\mathbb{V}(Y) = \sigma^2 = E(\ |Y-\mu|^2\ )$ pode ser obtida da seguinte maneira:

$$\text{Caso discreto: } \sum_{y_i \in \mathscr{S}} (y_i - \mu)^2\ \mathbb{P}(Y = y_i). \tag{9.1}$$

$$\text{Caso contínuo: } \int_{y \in \mathscr{S}} (y-\mu)^2\ f(y)\ \mathrm{d}y \tag{9.2}$$

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 maultiplicador 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 e ela estáo 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: $\mathbb{P}(|Y-\mu| > 4\sigma) \leq 1/4^2 = 0.06$. 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 consierar aqpena o caso contínuo. O caso discreto ésimilar 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,

$$\begin{aligned}
\mathbb{P}(|Y-\mu| > k\sigma) &= \mathbb{P}([Y < \mu - k\sigma] \cup [Y > \mu + k\sigma]) && (9.3)\\
&= \mathbb{P}([Y < \mu - k\sigma]) + \mathbb{P}([Y > \mu + k\sigma]) && (9.4)\\
&= \int_{-\infty}^{\mu-k\sigma} f(y)\,\mathrm{d}y + \int_{\mu+k\sigma}^{\infty} f(y)\,\mathrm{d}y && (9.5)
\end{aligned}$$

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

$$\int_{\mu+k\sigma}^{\infty} 1 \times f(y)\,\mathrm{d}y \leq \int_{\mu+k\sigma}^{\infty} \frac{(y-\mu)^2}{k^2\sigma^2} \times f(y)\,\mathrm{d}y$$

De maneira análoga, podemos também limitar a primeira integral:

$$\int_{-\infty}^{\mu-k\sigma} 1 \times f(y)\,\mathrm{d}y \leq \int_{-\infty}^{\mu-k\sigma} \frac{(y-\mu)^2}{k^2\sigma^2} \times f(y)\,\mathrm{d}y$$

Assim, usando estes dois limites superiores para as duas integrais, temos

$$\mathbb{P}(|Y-\mu| > k\sigma) \leq \int_{-\infty}^{\mu-k\sigma} \frac{(y-\mu)^2}{k^2\sigma^2} \times f(y)\,\mathrm{d}y + \int_{\mu+k\sigma}^{\infty} \frac{(y-\mu)^2}{k^2\sigma^2} \times f(y)\,\mathrm{d}y$$

Como $(y-\mu)^2/(k^2\sigma^2) \geq 0$ para todo y na reta real, teremos

$$\int_{\mu-k\sigma}^{\mu+k\sigma} \frac{(y-\mu)^2}{k^2\sigma^2} \times f(y)\,\mathrm{d}y \geq 0$$

e portanto

$$\begin{aligned}
\mathbb{P}(|Y-\mu| > k\sigma) &\leq \int_{-\infty}^{\mu-k\sigma} \frac{(y-\mu)^2}{k^2\sigma^2} \times f(y)\,\mathrm{d}y + \int_{\mu+k\sigma}^{\infty} \frac{(y-\mu)^2}{k^2\sigma^2} \times f(y)\,\mathrm{d}y + \int_{\mu-k\sigma}^{\mu+k\sigma} \frac{(y-\mu)^2}{k^2\sigma^2} \times f(y)\,\mathrm{d}y\\
&= \int_{-\infty}^{\infty} \frac{(y-\mu)^2}{k^2\sigma^2} \times f(y)\,\mathrm{d}y\\
&= \frac{1}{k^2\sigma^2} \int_{-\infty}^{\infty} (y-\mu)^2 \times f(y)\,\mathrm{d}y\\
&= \frac{1}{k^2\sigma^2}\,\sigma^2 = \frac{1}{k^2}
\end{aligned}$$

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 desiguldade 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 $DP = \sigma = 1/k$. 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 desiguldade 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. Fitting Distributions to Data

Neste capítulo, vamos aprender a verificar se uma amostra aleatória da v.a.'s i.i.d. $X_1, X_2, \ldots, X_n$ 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 distibuições 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 varição 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 distrubuição é aceita como um modelo compatível com os dados. O teste possui duas vantagens principais: ele pode ser usado com distribuições conínuas 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 $Y_1, Y_2, \ldots, Y_n$ 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 práica, 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 especificaçãoes 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 distrbuição (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 $Y_1, Y_2, \ldots, Y_n$ 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: $Y = 0$, $Y = 1, \ldots, Y = 10$ 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 altas.

O passo 2 do teste qui-quadrado é calcular o número de elementos da amostra $Y_1, Y_2, \ldots, Y_n$ 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 cm 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-categoria $[a_k, b_k)$ 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 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 Poisson(λ) 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: $Y = 0$, $Y = 1, \ldots, Y = 5$, 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 $k = 0, \ldots, 4$. Para a útima 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 distrbuição particular sob estudo (binmomial, 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

$$\chi^2 = \frac{(2.26)^2}{226.74} + \frac{(-0.39)^2}{211.39} + \ldots + \frac{(-0.50)^2}{1.5} = 1.13$$

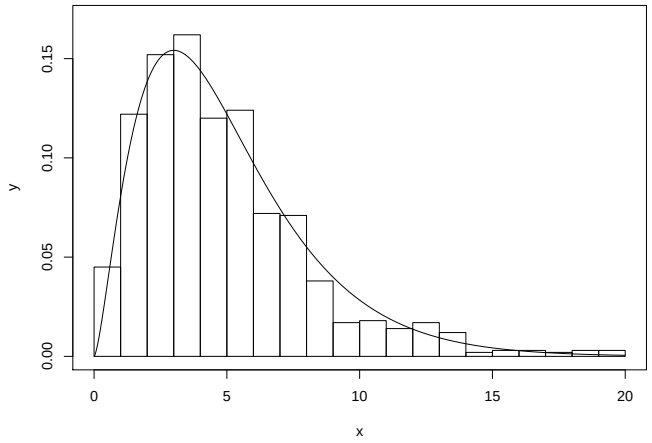


Figure 10.1: Histograma de 1000 simulações e densidade de Qui-quadrado com 5 g.l. superimposta

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 $X = 0$, $X = 1, \ldots, X = 4$, 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 $X_1, \ldots, X_{576}$ iid Poisson($\lambda = 0.9323$)
  - Conte o número N_k de X_i's iguais a $0, 1, \ldots, 4, \geq 5$
  - 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 posíveis (o eixo positivo $(0, \infty)$) e probabilidades associadas. O histograma dá uma ideia de quais regiões do eixo $(0, \infty)$ são mais prováveis equai 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

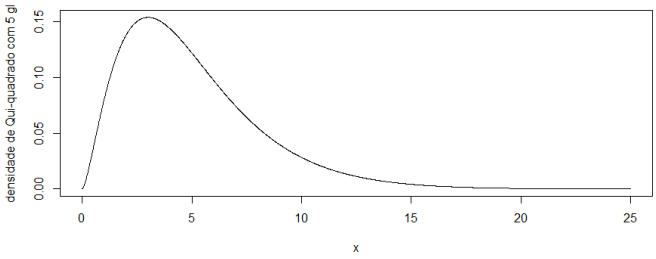


Figure 10.2: Densidade de uma distribuição qui-quadrado com $\nu = 5$ g.l.

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 $\mathbb{F}(x)$ 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 $\mathbb{F}(x)$ que precisaram ser estimados. Nesta parte inicial do texto supomos que todos os parâmetros do modelo teórico $\mathbb{F}(x)$ 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 $\mathbb{F}(x)$ é verdadeiro (de fato, está gerando os dados observados),então χ^2 varia aproximadamente como uma distribuição qui-quadrado com ν graus de liberdade onde ν é ao número de categorias menos 1, caso não existam parâmetros desconhecidos em $\mathbb{F}(x)$. No caso das bombas em Londres, temos $\nu = 6 - 1 = 5$ graus de liberdade (abreviado como g.l. daqui por diante).

Quais são os valores típicos de uma qui-quadrado com $\nu = 5$ 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 $\nu = 5$ g.l.

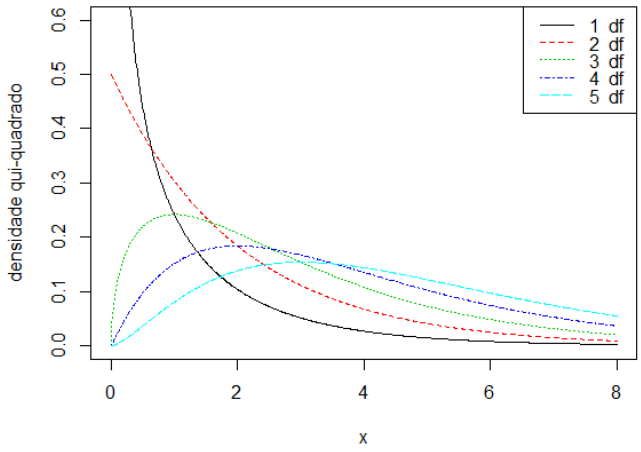


Figure 10.3: Densidades da distribuição qui-quadrado com $k = 1,2,3,4,5$

Na verdade, a distribuição qui-quadrado com ν graus de liberdade é um caso particular da distribuição gama (ver seção ?? no capítulo ??). Uma qui-quadrado com ν graus de liberdade é o mesmo que uma Gama$(\nu/2, 1/2)$, que possui densidade de probabilidade dada por

$$f(x) = \frac{1}{2^{\nu/2}\Gamma(k/2)}\, x^{\nu/2-1} e^{-x/2} = (\text{cte.})\, x^{\nu/2-1} e^{-x/2}$$

Assim, a densidade da qui-quadrado é o produto de $x^{\nu/2-1}$, uma potência crescente de x, por um decrescimento exponencialmente rápido em x, $e^{-x/2}$. Mesmo com um número ν de graus de liberdade grande, o decrescimento exponencial eventualmente domina o produto de modo que $x^{\nu/2-1}e^{-x/2} \to 0$ quando $x \to \infty$. A Figura 10.3 mostra as densidades $f(x)$ com diferentes valores para os graus de liberdade ν.

Mas, como usar o teste qui-quadrado? Vamos voltar ao caso das bombas de Londres. Os valores típicos de uma χ^2 com $\nu = 5$ 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 $\nu = 5$ 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)`.

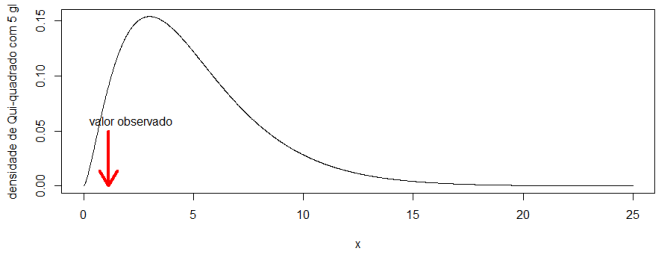


Figure 10.4: 1.13 é o valor observado de χ^2 no caso das bombas em Londres. Gráfico da densidade de uma qui-quadrado com 5 g.l.

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 o 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 usamos parãmetros 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 ν 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 $E_0 = 200\times\mathbb{P}(X=0) = 200exp(-0.61) = 108.6702$. 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 $\nu = 5-1-1 = 3$ 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-quadrdo 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?

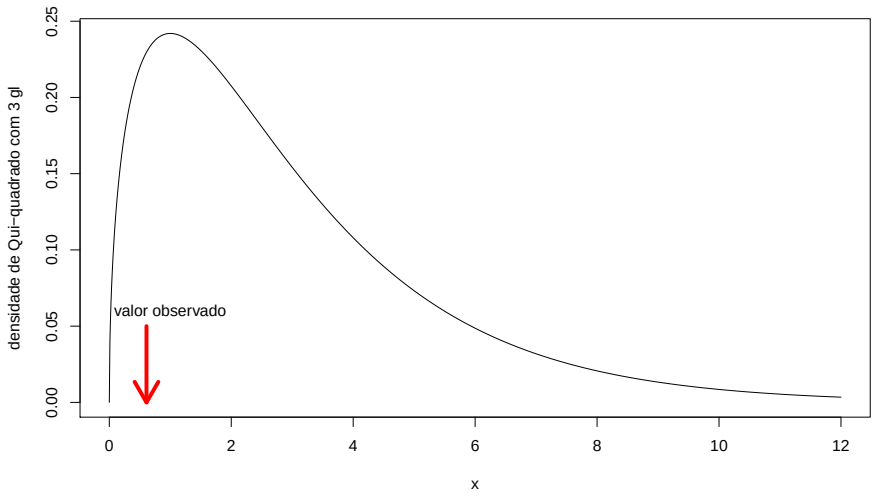


Figure 10.6: Valor observado $\chi^2 = 0.61$ e densidade de qui-quadrado com 3 g.l.

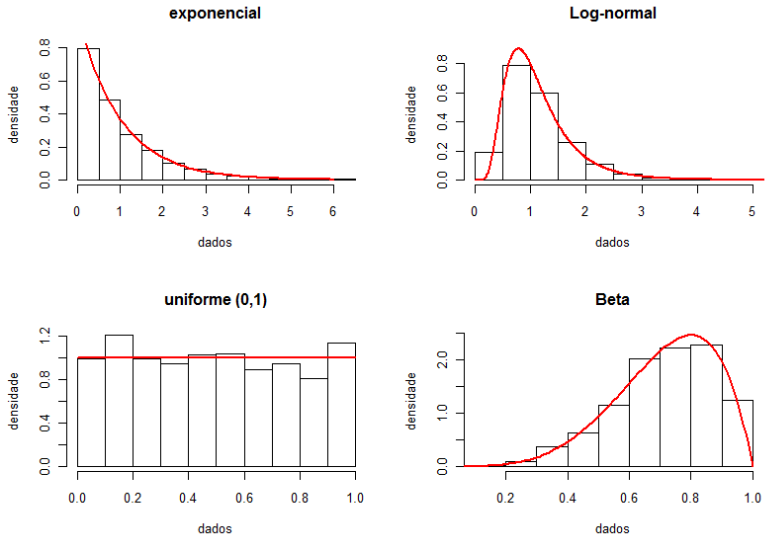


Figure 10.7: Histogramas de amostras das distribuições exp(1), log-normal $(0, 0.5)$, uniforme $U(0,1)$ e beta$(5,2)$ com as densidades correspondentes.

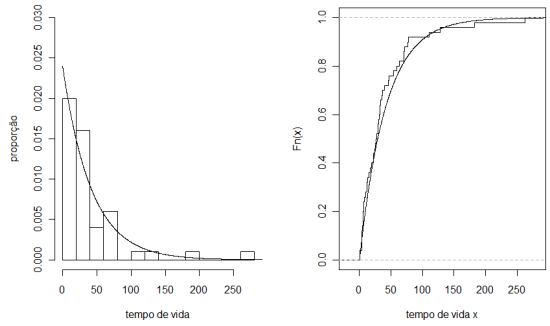


Figure 10.8: Esquerda: Gráfico de histograma de dados de tempos de vida de componentes eletrônicos e densidade de uma exponencial com parâmetro $\lambda = 0.024$. Direita: Função distribuição acumulada empírica e teórica de uma exponencial

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 $\hat{\mathbb{F}}_n(x)$ dos dados de tempos de vida e a curva contínua e suave é a função distribuição acumulada $\mathbb{F}(x) = 1 - \exp(-0.024x)$ 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 $x_1, x_2, \ldots, x_n$ um conjunto de números reais. A função distribuição acumulada empírica $\hat{\mathbb{F}}_n(x)$ é uma função $\hat{\mathbb{F}}_n : \mathbb{R} \to [0,1]$ tal que, para qualquer $x \in \mathbb{R}$ temos

$$\hat{\mathbb{F}}_n(x) = \frac{\#\{x_i \leq x\}}{n} = \text{Proporção dos } x_i \text{ que são} \leq x$$

A função $\hat{\mathbb{F}}_n(x)$ 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 $\hat{\mathbb{F}}_n(x)$ 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 $\hat{\mathbb{F}}_n(x)$ 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 $\mathbb{F}(x)$. 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 $\hat{\mathbb{F}}_n(x)$. Se tivermos $\hat{\mathbb{F}}_n(x) \approx \mathbb{F}(x)$ para todo x na reta, como na Figura 10.10, nós concluímos que o modelo adotado ajusta-se bem aos dados. Como saber se $\widehat{F_n}(y) \approx F(y)$? Veremos a resposta na próxima seção.

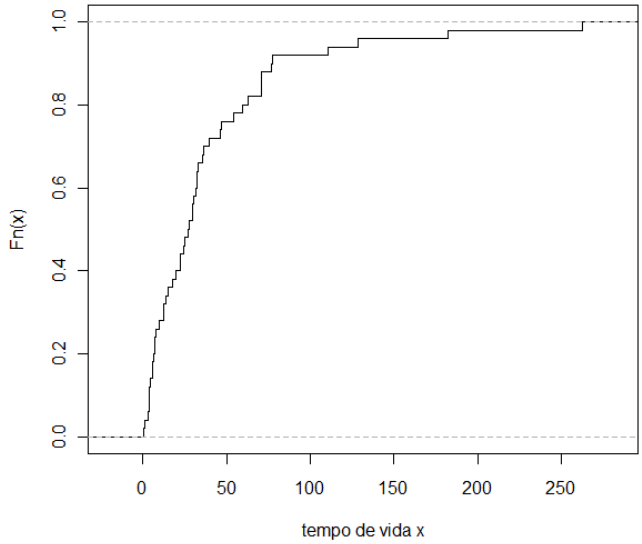


Figure 10.9: Função distribuição acumulada empírica $\mathbb{F}_n(x)$ dos dados de tempos de vida.

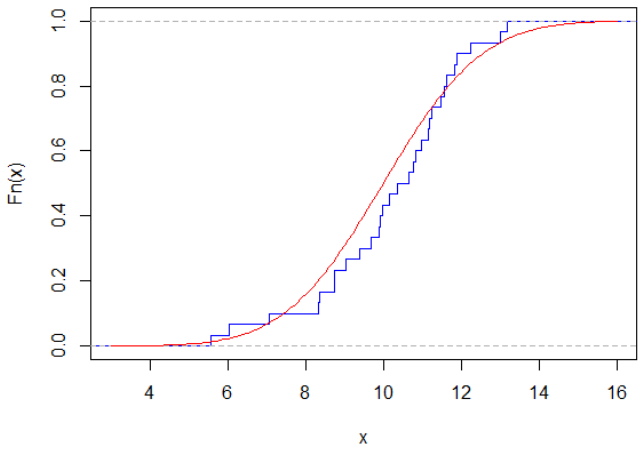


Figure 10.10: Empírica $\hat{\mathbb{F}}_n(x)$ e a teórica $\mathbb{F}(x)$.

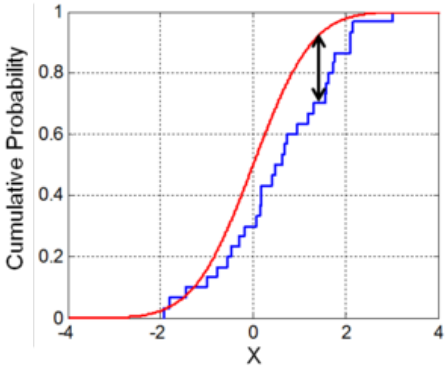


Figure 10.11: Empírica $\hat{\mathbb{F}}_n(x)$ e a teórica $\mathbb{F}(x)$ com a distância de Kolmogorov D_n.

10.7 Distância de Kolmogorov

Para cada ponto x na reta, olhe a distância vertical $|\hat{\mathbb{F}}_n(x) - \mathbb{F}(x)|$ entre as duas curvas $\hat{\mathbb{F}}_n(x)$ e $\mathbb{F}(x)$. 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 $D_n = \max_x |\hat{\mathbb{F}}_n(x) - \mathbb{F}(x)|$, a maior distância vertical entre as duas curvas $\hat{\mathbb{F}}_n(x)$ e $\mathbb{F}(x)$.

A Figura 10.11 mostra o ponto x em que as curvas $\hat{\mathbb{F}}_n(x)$ e $\mathbb{F}(x)$ 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 $\hat{\mathbb{F}}_n(x)$ é uma função aleatória. A Figura 10.12 mostra a função $\hat{\mathbb{F}}_n(x)$ obtida com uma amostra de tamanho $n = 20$ de uma gaussiana $N(0,1)$. A seguir, vemos outra função $\hat{\mathbb{F}}_n(x)$, contruída 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 $\hat{\mathbb{F}}_n(x)$ 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=
 for(i in 2:9) lines(ecdf(rnorm(20)), do.p=F, verticals=T)
```

Apesar de aleatório, podemos afirmar algumas coisas sobre $\hat{\mathbb{F}}_n(x)$. Suponha que $\mathbb{F}(x)$ é o verdadeiro modelo gerador dos dados. Na Figura 10.13 usei o modelo $N(0,1)$. Pode-se mostrar matematicamente que, apesar de $\hat{\mathbb{F}}_n(x)$ (e portanto, D_n também) flutuar com a amostra, temos D_n convergindo para zero quando n cresce: $D_n \to 0$ se $n \to \infty$, qualquer que seja modelo contínuo $\mathbb{F}(x)$.

O que acontece com D_n quando n cresce se estivermos usando o modelo teórico *incorreto*? Suponha que $\mathbb{F}(x)$ não seja o modelo que gera os dados da amostra. Na Figura 10.14, eu uso o modelo $\mathbb{F}(x) \sim N(0,1)$ mas, na verdade, os dados são gerados de uma $N(0.3,1)$. Então, pode-se

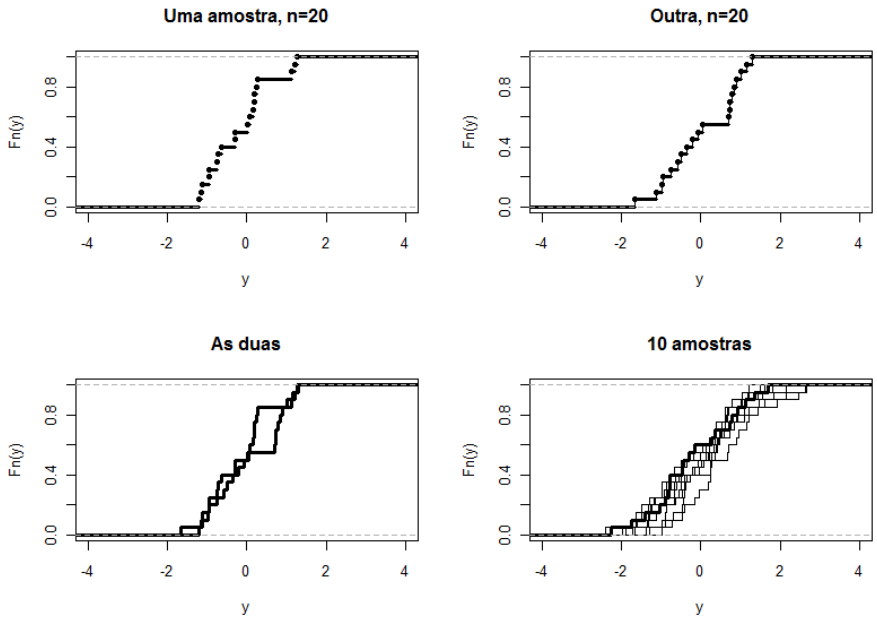


Figure 10.12: Mostrando o caráter aleatório de $\hat{\mathbb{F}}_n(x)$.

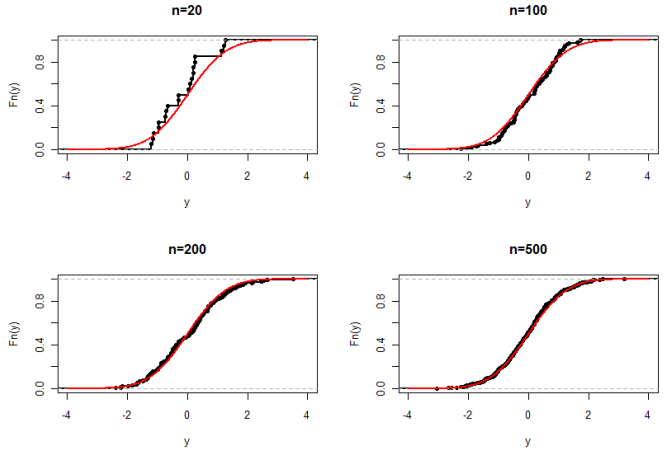


Figure 10.13: $D_n \to 0$ se o modelo é correto

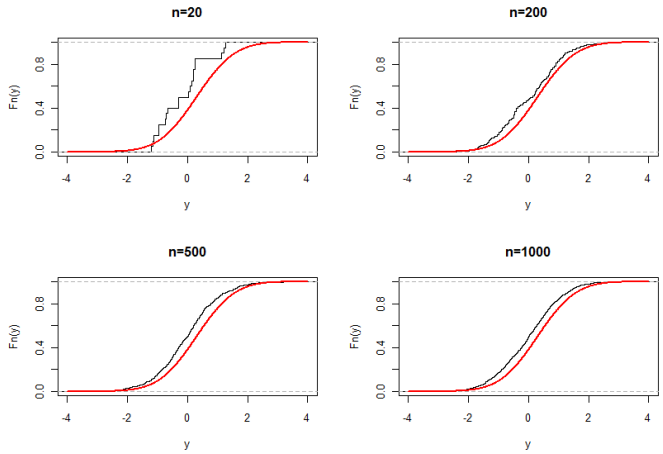


Figure 10.14: $D_n \nrightarrow 0$ se o modelo é correto

mostrar que D_n converge para um valor maior que zero.

Resumindo:

- Suponha que $\mathbb{F}(x)$ é o modelo verdadeiro. Então $D_n \to 0$ se $n \to \infty$.
- Se $\mathbb{F}(x)$ não é o modelo verdadeiro, $D_n \to a > 0$.

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 $D_n \to 0$ se $n \to \infty$. 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 $D_n \to 0$ se $n \to \infty$. Com que rapidez ele decresce em direção a 0? Kolmogorov mostrou que:

- $nD_n \to \infty$ (degenera).
- $\log(n)D_n \to 0$ (degenera).
- $\sqrt{n}D_n \nrightarrow 0$ e também $\nrightarrow \infty$.
- $\sqrt{n}D_n$ fica (aleatoriamente) estabilizado. Qualquer outra potência diferente de $-1/2$ leva a resultados degenerados.
- $n^{0.5+\varepsilon}D_n \to \infty$.
- $n^{0.5-\varepsilon}D_n \to 0$.

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

$$\sqrt{n}D_n \to K$$

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 $f(x) = 8x\sum_{k=1}^{\infty}(-1)^{k+1}k^2e^{-2k^2x^2}$ para $x > 0$. O gráfico da Figura 10.15 mostra esta densidade

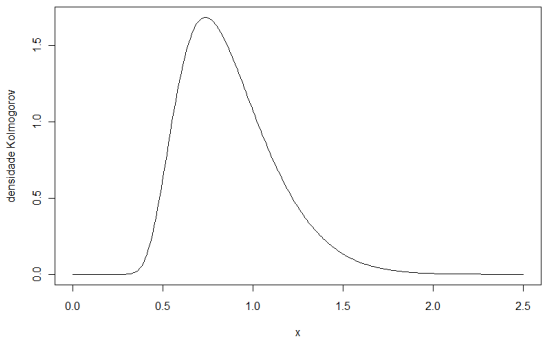


Figure 10.15: Densidade de $K \approx \sqrt{n}D_n$

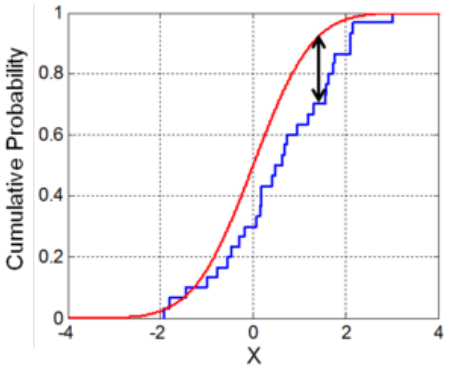


Figure 10.16: Empírica $\hat{F}_n(y)$ e a teórica $F(y)$.

$f(x)$. 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 $\sqrt{n}D_n \to \infty$.

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: $x_1, x_2, \ldots, x_n$. 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 $\hat{\mathbb{F}}_n(x)$. Calcule a distância de Kolmogorov $D_n = \max_y |\hat{\mathbb{F}}_n(x) - \mathbb{F}(x)|$ (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 Teste de Kolmogorov com parâmetros estimados da amostra

O teste de Kolmogorov é válido apenas quando a distribuição do modelo teórico é *completamente especificada*. Isto é, a distribuição não possui parâmetros desconhecidos que necessiatam ser

estimados da mesma amostra que é usada no teste. Por exemplo, contraste a situação em que o modelo teórico é que os dados vieram da distribuição $\mathbb{F}(y) \sim N(10,4)$ com aquela situação em que o modelo teórico é que os dados vieram da distribuição $\mathbb{F}(y) \sim N(\mu, \sigma^2)$ *sem especificar qual é o valor de* μ *e de* σ^2.

O que acontece se os parâmetros forem estimados? Vamos sofrer um problema sempre presente na ciência dos dados e aprendizado de máquina: o problema de *over-fitting*. O teste de Kolmogorov é pensado para a situação em que o modelo teórico é fixo, imutável, não variando em função dos dados da amostra. Por exemplo, fixamos $\mathbb{F}(y) \sim N(10,4)$ e fazemos o teste de Kolomogorov. Suponha agora que dizemos apenas que o modelo teórico é uma gaussiana $\mathbb{F}(y) \sim N(\mu, \sigma^2)$ mas não especificamos um valor fixo para μ e σ^2. Ao invés disso, estimamos um valor para μ e σ^2 a partir da própria amostra $\mathbf{y} = (y_1, y_2, \ldots, y_n)$. Assim, μ e σ^2 passam a ser funções dos dados. Tipicamente, faríamos $\hat{\mu}(\mathbf{y}) = \bar{y} = (y_1 + \ldots + y_n)/n$ e $\hat{\sigma^2}(\mathbf{y}) = \sum_i (y_i - \bar{y})^2/n$. Assim, ao invés de um modelo teórico imutável com respeito à amostra (tao como $\mathbb{F}(y) \sim N(10,4)$), teremos um modelo teórico que flutua e ajusta-se aos dados observados, com $\mathbb{F}(y) \sim N(\hat{\mu}(\mathbf{y}), \hat{\sigma^2}(\mathbf{y}))$. A consequência é que, curiosamente, esta acomodação aleatória do modelo teórico aos dados vai subtrair parte da flutuação casual que a estatística de teste $\sqrt{n}D_n$ tem no caso em que $\mathbb{F}(y)$ é completamente especificada. Assim, $\sqrt{n}D_n$ tende a estar mais concentrada perto de zero do que Kolmogorov obteve.

Vamos exemplificar com uma simulação. Suponha que o modelo verdadeiro que gera os dados de uma amostra com tamanho $n = 200$ seja uma gaussiana $\mathbb{F}(y) \sim N(10,4)$. Vamos simular no computador 10 mil amostras, todas de tamanho $n = 200$ e calcular a estatística de teste $\sqrt{n}D_n = \sqrt{200}D_{200}$ para cada uma delas. Fazemos isto usando o modelo teórico $\mathbb{F}(y) \sim N(10,4)$. Em seguida, vamos fazer um histograma dos 10 mil valores obtidos e compará-lo com a densidade contínua obtida por Kolomogorov e discutida na seção 10.8. O resultado está no gráfico da esquerda na Figura ??. Vemos que, de fato, a estatística $\sqrt{200}D_{200}$ flutua de amostra para amostra de acordo com a teoria prevista nos resultados de Kolmogorov.

No gráfico à direita na Figura ?? vemos o que acontece quando dizemos apenas que $\mathbb{F}(y) \sim N(\mu, \sigma^2)$ e estimamos os parâmetros da gaussiana a partir dos dados de cada uma das 10 mil amostras. O que podemos ver é o hstograma de $\sqrt{n}D_n = \sqrt{200}D_{200}$, as medidas de afastamento não são bem representadas pela distribuição encontrada por Kolmogorov. Elas estão concentradas entre 0.5 e 1.0 quando os resultados de Kolmogorov (assumindo que o modelo teórico é fixo, completamente especificado) esperam que estas medidas possam ir até 1.5.

Assim, usar a teoria do teste de Kolmogorov quando o modelo teórico requer a estimação de parâmetros contrai a distribuição $\sqrt{n}D_n = \sqrt{200}D_{200}$ e faz com que os p-valores obtidos dificlmente sejam próximos de zero. Mas, espera um momento: isto não é bom? Estaremos dificlmente rejeitando o modelo gaussiano e esta é a decisão correta!

COMPLETAR considerando o poder do teste.

```
nsample = 200
niter = 10000
sqrtn = sqrt(nsample)
dados = matrix(0,nrow=niter,col=nsample)
ks = rep(0, niter)
for(i in 1:niter){
  x = rnorm(nsample, 10, 4)
  m = mean(x)
  s = sd(x)
  ks[i] = sqrtn*(ks.test(x, "pnorm", m, s)$stat)
}
```

```
hist(ks, breaks=30, xlim=c(0, 2.5), prob=T)
x <- seq(0, 2.5, by=0.01); y <- kolmpdf(x)
lines(x, y, type="l", ylab = "densidade Kolmogorov")
```

Como corrigir este problema? Quando o teste de Kolmogorov precisar estimar parâmetros, devemos obter um p-valor por meio de simulação Monte Carlo.

TEXTO EXTRAIDO de resposta de Greg Snow em stacksexchange: `https://stats.stackexchange.com/questions/45033/can-i-use-kolmogorov-smirnov-test-and-estimate-distribution-paramete`

A simple example using the Normal distribution as the null hypothesis:

```
tmpfun <- function(x, m=0, s=1, sim=TRUE) {
  if(sim) {
    tmp.x <- rnorm(length(x), m, s)
  } else {
    tmp.x <- x
  }
  obs.mean <- mean(tmp.x)
  obs.sd <- sd(tmp.x)
  ks.test(tmp.x, 'pnorm', mean=obs.mean, sd=obs.sd)$statistic
}

set.seed(20200319)
x <- rnorm(25, 100, 5)

out <- replicate(1000, tmpfun(x))

hist(out)
abline(v=tmpfun(x, sim=FALSE))
mean(out >= tmpfun(x, sim=FALSE))
```

The function will either compute the KS test statistic from the actual data (sim=FALSE) or simulate a new dataset of the same size from a normal distribution with specified mean and sd. Then in either case will compute the test statistic comparing to a normal distribution with the same mean and sd as the sample (original or simulated).

The code then runs 1,000 simulations (feel free to change and rerun) to get/approximate the distribution of the test statistic under the NULL (but with estimated parameters) then finally compares the test statistic for the original data to this NULL distribution.

We can simulate the whole process (simulations within simulations) to see how it compares to the default p-values:

```
tmpfun2 <- function(B=1000) {
  x <- rnorm(25, 100, 5)
  out <- replicate(B, tmpfun(x))
  p1 <- mean(out >= tmpfun(x, sim=FALSE))
  p2 <- ks.test(x, 'pnorm', mean=mean(x), sd=sd(x))$p.value
  return(c(p1=p1, p2=p2))
}

out <- replicate(1000, tmpfun2())
```

```
par(mfrow=c(2,1))
hist(out[1,])
hist(out[2,])
```

For my simulation, the histogram of the simulation based p-values is fairly uniform (which is should be since the NULL is true), but the p-values for the ks.test function are bunched up much more against 1.0.

You can change anything in the simulations to estimate power by having the original data come from a different distribution, or using a different Null distribution, etc. The normal is probably the simplest since the mean and variance are independent, more tuning may be needed for other distributions.

Simulations used to be much more difficult and time consuming, so the tests were developed to be quicker/easier than simulation, some of the early tables were created by simulation. Many tests can now easily be replaced by simulation, but will probably be with us for a while longer due to tradition and simplicity.

10.11 Kolmogorov versus Qui-quadrado

Dados $X_1, X_2, \ldots, X_n$ 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 $X_1, X_2, \ldots, X_n$ seguem uma $N(\mu, \sigma^2)$ mas não sabemos o valor de μ e σ^2, a bela teoria de Kolomogorov 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.

Comment: RAMON pergunta: os testes podem divergir do resultado quando aplicados no mesmo cenário? Exemplo: suponha uma amostra iid de N(0,1). É possível x2 e kolmogorov apresentar conclusões diferentes? Fazer talvez por simulacao e contra a proporcao de vezes em que coincidem. Bolar exercicios desse tipo?

10.12 Teste de Kolmogorov-Smirnov

completar ??

10.13 Teste de Anderson-Darling

Cramer-von Mises é um caso especial quando $w(x) = 1$.

10.14 Exemplos da literatura

10.15 Duas colunas

Exemplo de duas colunas com com código em R e em python:

Original Code

```
class C {
  public X x = new X();

  public void f() {
  ...
  }
}
```

After Extract Local

```
class c {
  public X x = new X();

  public void f(){
  ...
  }
}
```

Original Code

```
class C {
  public X x = new X();

  public void f() {
  ...
  }
}
```

After Extract Local

```
class c {
  public X x = new X();

  public void f(){
  ...
  }
}
```

10.16 Testes de ajustes de distribuição na prática de análise de dados

É difícil que uma única distribuição se ajuste a dados reais. Heterogeneidade do ambiente. Uma classe, ok. Uma mesma distribuição, menos provável. As populações reais são misturas de diferentes distribuições. Os parâmetro variem de acordo com diferentese aspectos.

Exemplo: altura de populações. Depende da idade. Para uma idade x, teremos $N(\mu(x), \sigma^(x))$. Achar a equação para $\mu(x)$ na literatura.

Peso de animal com idade?

árvore versus ??

10.17 Prova do teste de Kolmogorovn e Kolmogorov-Smirnov

completar??

10.18 Como provar o resultado de Pearson? Um esboço

Esta seção é opcional e pode ser omitida numa primeira leitura sem prejuízo do restante do livro.

Preciso melhorar MUITO esta parte . Muitos typos aqui.

Se $Z_1, Z_2, \ldots, Z_k$ são i.i.d. $N(0,1)$ então $Y = Z_1^2 + \ldots + Z_k^2 \sim \chi_k^2$, uma qui-quadrado com k g.l. Temos

$$X^2 = \sum_k \frac{(N_k - E_k)^2}{E_k}$$

Porque X^2 segue uma qui-quadrado? N_k é a contagem dos elementos da amostra que caem na categoria k $N_k \sim \text{Bin}(n, \theta)$ onde $\theta = \mathbb{P}(X \in \text{categoria k})$ Se n é grande, pelo Teorema Central do Limite,

$$\frac{N_k - n\theta}{n\theta(1-\theta)} = \frac{N_k - E_k}{E_k(1-\theta)} \approx N(0,1)$$

Se $(1-\theta) \approx 1$ então

$$\frac{N_k - E_k}{E_k} \approx N(0,1) \longmapsto \frac{(N_k - E_k)^2}{E_k} \approx N^2(0,1) = \chi_1^2$$

Somando sobre as categorias, teremos uma qui-quadrado (estou omitindo vários detalhes e sutilezas).

11. Monte Carlo Simulation

11.1 O que é uma simulação Monte Carlo

O verbo *simular* quer dizer fazer aparecer como real uma coisa que não o 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 resposta oferecidas pelo método matemático.

Estes números aleatórios gerados no computador possuem uma distribuição de probabilidade de interesse. Pode ser a distribuição normal (gausssiana), 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. Todos os métodos conhecidos começam gerando uma variável aleatória U com distribuição $U(0,1)$, a distribuição 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 os números reais do intervalo $(0,1)$. No computador, os números possuem uma representação finita e não podem expressar de forma exata números tais como os irracionais ($\sqrt{2}$ ou π, por exemplo) ou números fracionários com dízima periódica $1/3 = 0.3333\ldots$). Vamos ignorar esta finitude da representação computacional neste livro. Uma limitação mais relevant é o fato de que os números 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 $U_1, U_2, \ldots$. Para todos os efeitos práticos, eles podem ser considerados i.i.d. com distribuição uniforme em $(0,1)$.

Não veremos em detalhes os geradores de números com distribuição uniforme no intervalo $(0,1)$. Este é um assunto bastante técnico e especializado. Como ele é de pouco uso na prática da análise de dados. Vamos dar apenas um ligeira ideia de como os geradores de $U(0,1)$ funcionam apresentando 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 $0, 1, \ldots, 6$.

De forma geral, n pode ser escrito de forma única como $n = kp + r$ onde k é inteiro e $r = 0, \ldots, p-1$. O resto é o valor r igual a um dos valores $0, \ldots, p-1$. 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 $x_1, x_2, \ldots$ por meio da fórmula:

$$ax_{i-1} + b \equiv x_i \pmod{p}$$

onde a, b, e p são inteiros positivos e x_i é um dos inteiros $0, 1, \ldots, p-1$. Além disso, p é maior que a e b. 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 $x_1, x_2, \ldots$ são os restos da divisão inteira módulo 32777. Portanto, todos eles são inteiros no conjunto $\{0, 1, 2, \ldots, 32775, 32776\}$. 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 $u_1 = 0.91466577$, $u_2 = 0.38944992$, $u_3 = 0.09549379$, $u_4 = 0.32626537$, $u_5 = 0.86466120$, $u_6 = 0.78957806$, $u_7 = 0.89190591, \ldots$

A sequência

$$u_1 = x_1/p, u_2 = x_2/p, \ldots$$

é uma aproximação *determinística* para uma sequência de valores de variáveis *aleatórias independentes* e com distribuição uniforme em $(0,1)$. A qualidade desta aproximação é atestada pela incapacidade de vários testes estatísticos em detectar os padrões não aleatórios presentes 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, 32777 restos x_i distintos: os inteiros $0,1,\ldots,32776$. Assim, existem apenas 32777 números u_i no intervalo $(0,1)$ que podem ser gerados por este procedimento:

$$0/32777, 1/32777, 2/32777, \ldots, 32776/32777$$

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 $u_1, u_2, \ldots$ 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 $a = 3$, $b = 0$, $p = 30$ e $x_0 = 1$, teremos a sequência $\{3,9,27,21,3,9,27,21,3,9,27,21,3,9,27,21,3,\ldots\}$. Com probabilidade 1, depois de certo tempo, obtém-se um valor x_i igual a algum valor x_{i-k} já obtido anteriormente. Existem p restos possíveis e portanto, com certeza, ao gerar o $(p+1)$-ésimo número pseudo-aleatório ele já terá sido gerado anteriormente já que existem apenas p valores distintos com este gerador congruencial. A partir do momento em que acontecer a primeira repetição, teremos o resto da sequência repetindo-se com $x_{i+j} = x_{i-k+j}$. O número de passos k necessário até obter-se uma repetição numa sequência é determina *período* do gerador. Um gerador de números aleatórios com um período mais longo geralmente é preferível.

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 $2^{57} \approx 1.44 \times 10^{17}$. Bons geradores tem períodos tão grandes que os ciclos podem ser ignorados na prática.

A semente x_0 que dá início ao algoritmo de gereção 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 todas as vezes que a mesma semente for usada. Esta garantia é importante para tornar possível a replicação de resultados de simulação. De qualquer forma, a semete x_0 é um número arbitrário para iniciar o processo.

Este gerador congruencial 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.

■ **Example 11.1 — Gerando** $U(a,b)$. Nem sempre queremos gerar pontos ao acaso do intervalo $(0,1)$. Para gerar $X \sim U(a,b)$, uma distribuição uniforme num intervalo genérico (a,b), basta transformarmos $U \sim U(0,1)$. Seja $X = a + (b-a) * U$ onde $U \sim U(0,1)$. É fácil mostrar que $X \sim U(a,b)$. Primeiro, os valores possíveis para X são os pontos no intervalo (a,b). Depois, a

função distribuição acumulada num ponto $x \in (a,b)$ é

$$\mathbb{F}(x) = \mathbb{P}(X \leq x) = \mathbb{P}(a + (b-a) * U \leq x) = \mathbb{P}\left(U \leq \frac{x-a}{b-a}\right) = \frac{x-a}{b-a}$$

e portanto a densidade de probabilidade é constante e igual a $f(x) = 1/(b-a)$ para $x \in (a,b)$.

Por exemplo, se quisermos gerar 10 mil valores aleatórios com distribuição uniforme no intervalo $(3,7)$ basta fazermos `x = 3 + 4*runif(10000)`. Na verdade, em R, basta digitarmos `runif(10000, 3, 7)`. ∎

∎ **Example 11.2 — É difícil ganhar da sorte.** Em [3], Aoki, Assunção e Vaz-de-Melo (2017) fizeram um estudo do papel da sorte e da habilidade em esportes. Coletamos e analisamos todos os jogos de 198 ligas esportivas compreendendo 1503 temporadas de 84 países de 4 modalidades diferentes: basquete, futebol, vôlei e handebol. Medimos a competitividade por países e esportes encontrando que o futebol é realmente uma caixinha de surpresas. Quantificamos o peso da sorte e da habilidade das equipes através de um coeficiente que mede a distância entre os resultados finais observados das ligas esportivas e as competições perfeitamente equilibradas idealizadas em termos de habilidade. Também identificamos em cada temporada quais times deveriam ser removidos de suas ligas para que restasse um torneio completamente aleatório. Surpreendentemente, não são necessários muitos deles. Se interessado, veja o delicioso vídeo criado por Raquel Aoki para este artigo em `https://www.kdd.org/kdd2017/papers/view/luck-is-hard-to-beat-the-difficulty-of-sports-prediction`.

Esta combinação de sorte e habilidade foi discutida pelo YouTuber Derek Muller em seu canal Veritassium no vídeo *The Success Paradox*. (ver `https://www.youtube.com/watch?v=3LopI4YeC4I&t=223s`). Ele simula o processo de formação dos astronautas pela NASA. Em 2017, dentre mais de 18300 candidatos, apenas 11 foram selecionados e se formaram como astronautas. Quanto de sorte pode estar presente num processo desses? Para responder, podemos usar a simulação Monte Carlo. Vamos supor que os astronautas são selecionados principalmente por suas habilidades, experiência e esforço mas também, digamos, por 5% de sorte, ou seja, circunstâncias favoráveis. Seja $n = 18300$ o número de candidatos. Vamos gerar as habilidades de cada candidato, $H_1, \ldots, H_n$, como v.a.'s i.i.d. com distribuição $U(0, 0.95)$. A sorte $S_1, \ldots, S_n$ de cada um deles é gerada independentemente com distribuição $U(0, 0.05)$. Os 11 candidatos selecionados são aqueles com os maiores valores da v.a. $R_i = H_i + S_i$. Quantos dos candidatos escolhidos entraram devido à sorte? Isto é, quantos desses 11 escolhidos teriam sido preteridos caso a seleção tivesse se baseado apenas nas suas hablidades H_i? O código abaixo simula todo o processo 10 mil vezes, com 18300 candidatos em cada replicação, representando 10 mil seleções diferentes.

```
nsim = 10000
n = 18300
comuns = rep(0, nsim)
for(k in 1:nsim){
  habilidade = runif(n, 0, 0.95)
  sorte = runif(n, 0, 0.05)
  escore = habilidade + sorte
  selecao1 = order(escore)[18289:18300]
  selecao2 = order(habilidade)[18289:18300]
  comuns[k] = length(intersect(selecao1, selecao2))
}
mean(comuns); sum(comuns >= 4)/nsim
[1] 1.8126
[1] 0.0822
```

Em média, apenas 1.8 selecionados estariam nas duas listas ao mesmo tempo. Assim, aproximadamente 11-2 = 9 dos candidatos escolhidos passaram para o topo da lista na frente de outros candidatos mais habilidosos devido aos 5% de sorte. De todas as 10 mil simulações apenas em 8.2% delas tivemos 4 ou mais indivíduos nas duas listas. Assim, a sorte tem um papel muito grande na seleção final. Este resultado pode ser visto de duas maneiras. Uma delas é reconhecer que quando a competição é intensa, ser talentoso e trabalhar duro é fundamental mas não é suficiente para garantir sucesso. A outra é inspirada pelo que eu ouvicerta vez de um colega: só vale a pena concorrer pelas coisas que você pode não ganhar. Ele quis dizer que, se as suas chances de sucesso são muito altas, é porque você está mirando muito baixo. ∎

∎ **Example 11.3 — Gerando uniforme num retângulo.** Queremos gerar pontos (X,Y) no retâgulo $[a,b]\times[c,d]$ com uma distribuição uniforme. Como veremos no capítulo 12, basta gerarmos as coordenadas X e Y independentemente uma da outra com $X\sim U(a,b)$ e $Y\sim U(c,d)$. E para gerar pontos uniformemente numa figura geométrica diferente? Por exemplo, pontos ao acaso no mapa do Brasil ou pontos ao acaso num círculo? A maneira mais simples será usar o método de aceitação-rejeição, a ser visto na seção 11.14. ∎

11.3 Simulação de v.a. Bernoulli

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 $U\leq p$, diga que o evento $[X=1]$ ocorreu. Se $U>p$, diga que o evento $[X=0]$ ocorreu.

Com este procedimento, qual a probabilidade de gerarmos o evento $[X=1]$? Temos

$$\mathbb{P}(X=1)=\mathbb{P}(U\in(0,p))=p-0=p\,,$$

exatamente a probabilidade desejada. Por exemplo, para gerar uma 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`

∎ **Example 11.4 — Passeio aleatório e a lei do arco-seno.** No livro clássico de probabilidade escrito por William Feller [16], um exemplo lindo enche os olhos ao mostrar como nossa intuição pode ser facilmente enganada em problemas de probabilidade. Um jogo simples consiste em lançar uma moeda honesta repetidamente e independetemente n vezes e ganhar um real cada vez que sair cara e perder um real se sair coroa. Seja $S_t=X_1+\ldots+X_t$ a soma acumulada dos primeiros t resultados com $X_t=+1$, se sair cara no t-ésimo lançamento e $X_t=-1$, se sair coroa. A frequência relativa de caras converge para 1/2 se n for grande e assim o ganho líquido deve convergir para 0: $S_n/n\to 0$. Este resultado é consequência da Lei dos Grandes Números (ver 19), é bastante intuitivo e correto, e todo mundo o aceita sem problemas. Em cada instante de tempo t, o mais provável é que $S_n\neq 0$ de forma que um dos jogadores esteja ganhando do outro. Como após muitas jogadas

é praticamente certo que o ganho médio S_n/n seja praticamente zero, somos levados a pensar que haverá muitos momentos nos quais vai ocorrer a troca de liderança, com S_t trocando de sinal frequentemente. Pois bem, esta última intuição não corresponde aos fatos e está errada. Em um longo jogo de cara ou coroa, é bem provável que um dos dois jogadores permaneça praticamente o tempo todo do lado vencedor e o outro jogador no lado perdedor. Isto é surpreendente e pode ser provado rigorosamente com probabilidade elementar (sem conhecimentos avançados) mas requerendo um trabalho razoável. Vamos usar simulação para verificar que os resultados teóricos realmente são corretos e que nossa intuição inicial está errada.

Observe que, se $S_k > 0$, o jogador favorecido por caras (jogador-caras) está liderando no instante k. Caso $S_k < 0$, o outro jogador é o líder. Se $S_k = 0$, os jogadores estarão empatados. Vamos criar a variável indicadora (binária) $I[S_k > 0]$ que vale 1 se $S_k > 0$, e vale 0, caso contrário. Se calcularmos $p_n = \sum_{k=1}^{n} I[S_k > 0]/n$ estaremos obtendo a proporção de jogadas ou tempo em que o jogador-caras esteve liderando na primeira n jogadas. Esta proporção é aleatória. Ela varia com n ao longo de um jogo particular. Ela também varia se um novo é iniciado. O gráfico do lado esquerdo na Figura 11.1 mostra 10 linhas reprsentando 10 jogos distintos, cada um deles com 1000 lançmentos da moeda. Temos as 10 trajetórias aleatórias do ganho S_k do jogador-caras nos primeiros 1000 lançamentos da moeda.

```
# As 10 trajetorias
amostra = matrix(0, nrow=1000, ncol=10)
for(k in 1:10){
  n = 1000 # Lancar a moeda n vezes
  x = -1 + 2 * (runif(n) < 0.5)  # ganho: -1 ou 1
  amostra[, k] = cumsum(x) # soma acumulada dos ganhos
}
matplot(amostra, type="n", xlab="n", ylab="S(n)")
matlines(amostra, lty=1:10)
abline(h = 0, lwd=3)
```

Ao invés de olhar apenas 10 jogos, repetimos o experimento 10 mil vezes independentemente, sempre jogando a moeda 1000 vezes em sequência. Registramos o valor final da proporção p_{1000} do tempo em que o jogador-caras liderou ao longo do jogo de 1000 lançamentos. O lado direito Figura 11.1 mostra o histograma desses 10 mil valores aleatórios de p_{1000}. O que observamos é que a proporção do tempo em que o jogador-caras lidera *não está* concentrada em torno de 1/2 como nossa intuição poderia sugerir. Ao contrário, são nos dois extremos, próximos a zero e a 1, que o histograma possui mais massa. Isto quer dizer que podemos esperar num jogo longo de cara-coroa que o *menos provável* seja um equilíbrio entre os dois jogadores: a maior parte do tempo, um deles deve estar na liderança e não é improvável que um deles esteja liderando quase que todo o jogo. A curva azul é o gráfico da densidade de probabilidade do valor limite aleatório de p_n. Quando n cresce para infinito, o valor aleatório p_n converge para um valor $p \in (0,1)$ que segue a distribuição com densidade dada por

$$f(p) = \frac{1}{\pi} \frac{1}{\sqrt{p(1-p)}}.$$

```
# Histograma e densidade
# Repetir nim vezes
nsim = 10000
res = rep(0, nsim)
for(k in 1:nsim){
  # Lancar a moeda n vezes
```

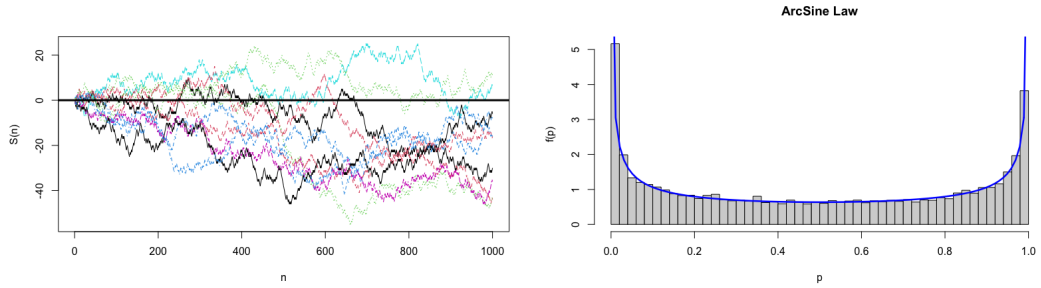


Figure 11.1: Esquerda: Dez trajetórias de S_n para $n = 1, 2, \ldots, 1000$ do passeio aleatório. Direita: Histograma de 10 mil valores simulados de p_{1000} gráfico da densidade da distribuição arco-seno.

```
  n = 1000
  x = -1 + 2 * (runif(n) < 0.5)  # ganho: -1 ou 1
  s = cumsum(x) # soma acumulada dos ganhos
  res[k] = sum( s > 0 ) / n # proporcao do tempo em que liderou
}
hist(res, breaks= 50, xlim=c(0,1), prob=T, xlab="p", ylab="f(p)", main="ArcSine Law")
curve(1/(pi*sqrt(x*(1-x))), from = 0.001, to = 0.999,
      add=T, col="blue", lwd=2.5)
```

Esta é a famosa (primeira) lei do arco-seno derivada pela probabilista francês Paul Lévy em 1939. O nome aparece porque a função de distribuição acumulada dessa densidade é $\mathbb{F}(p) = \frac{2}{\pi}\arcsin(\sqrt{p})$ para $p \in (0,1)$. Existem vários resultados surpreendentes no livro [16] de Feller. Por exemplo, não apenas a linha S_k cruza o eixo do horizontal raramente mas, quando o número de jogadas n é grande, os tempos de espera até que uma troca de liderança ocorra tende a aumentar também. Assim, se você está no jogo há muito tempo e está perdendo, não espere que vai passar a ganhar no futuro próximo. E isto é num jogo de pura sorte, com uma moeda honesta.

■

■ **Example 11.5 — Modelos de aprendizagem comportamental.** Nos primórdios da teoria da aprendizagem comportamental nas décadas de 50 e 60, foram conduzidos vários experimentos para entender como animais aprendiam a executar com sucesso certas tarefas. Existiam várias possibilidades. A aprendizagem podia ser por insight, um súbito entendimento de como executar as tarefas através de um processo de tentativas aleatórias. Outras hipóteses consideravam uma aprendizagem gradual, em que cada tentativa aumentava de alguma maneira a chance de executar corretamente a tarefa na próxima tentativa. Bush e Mosteller (1955) [8] estudaram oito diferentes modelos de aprendizagem usando dados de um experimento de aprendizagem comportamental com 30 cães em que cada um deles executou uma uma sequência de 25 tentativas. Em cada tentativa, o cão era colocado numa caixa de transporte (ver lado esquerdo da Figura 11.2). Ela é uma câmara contendo dois compartimentos retangulares divididos por uma barreira que o cão seria capaz de saltar com certo esforço. Em cada tentativa, uma luz era acesa por 10 segundos e, em seguida, um choque elétrico era aplicado através de placas metálicas no chão da gaiola em estava o cão. Em cada tentativa, o cão tinha a oportunidade, depois da luz se acender, de pular para uma gaiola adjacente e assim evitar o choque. Nas tentativas iniciais, os cães receberam o choque mas, ao longo das tentativas seguintes, todos eles eventualmente aprenderam a evitar o choque.

Esse experimento foi um exemplo de crueldade contra os animais e é considerado antiético hoje em dia, não sendo mais conduzido. Ignorando a crueldade envolvida, o lado direito da Figura 11.2 mostra os dados experimentais para os 30 cães, ordenados pelo tempo da última tentativa

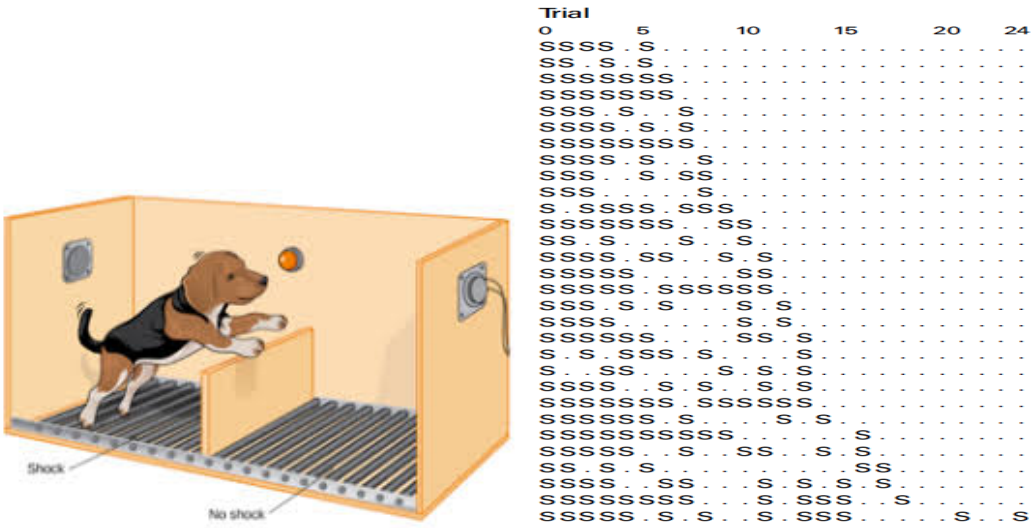


Figure 11.2: Esquerda: Esquema do experimento. Direita: Sequência de choques (S) e esquivas (.) para 25 tentativas em cada um dos 30 cães no experimento analisado por Bush e Mosteller (1955). Os cães foram ordenados de acordo com a número da tentativa em que receberam o último choque. Aqueles que demoraram mais para aprender a evitar o choque estão listados por último.

em que receberam um choque. Entre os oito modelos considerados por Bush e Mosteller (1955), um deles considerava os fatores que afetaram o aprendizado dos cães: o que acelerava mais a aprendizagem, uma evitação bem sucedida ou um choque? O modelo, chamado de Two-Operator Linear Model (TOL), definia q_n como sendo a probabilidade de um cão tomar choque na tentativa n. A probabilidade de evitar o choque nesta tentativa é $1-q_n$. O modelo TOL supõe que a probabilidade de choque q_n diminuia com o tempo n de forma que o cão aprendia a medida que executava as tentativas. A rapidez dessa aprendizagem dependia das ações e consequências vividas pelo cão. O método de máxima verossimilhança (ver capítulo **??**) foi usado para estimar os parâmetros envolvidos. Um deles é $\alpha_1 = 0.80$ e indica a proporção da redução da probabiliadde de choque se o cão consegue evitá-lo. Isto é, $q_{n+1} = 0.80 q_n$ se o cão consegue evitar o choque na tentativa n. O outro parâmetro é $\alpha_2 = 0.92$ e indica que a chance de levar choque diminui *menos* se ele leva um choque: $q_{n+1} = 0.92 q_n$ se o cão leva choque na tentativa n. Com estes paâmetros, vamos simular um grande conjunto de stat-cães e calcular aspectos característicos das sequências simuladas para compará-los com os aspectos realmente observados nos dados reais. Veja que tudo se passa como se os cães jogassem moedas sucessivamente (um lançamento de moeda para cada tentativa) mas as probabilidades de cara variassem de acordo com o que ocorreu previamente. O valor $q_1 = 1$ pois, na primeira tentativa, todos os cães receberam o choque uma vez que eles não sabiam ainda o significado do sinal.

O código abaixo simula 10 mil stat-cães, cada um deles fazendo 25 tentativas de acordo com o modelo TOL. A seguir, obtemos algumas estatísticas descritivas para comparar com o que observamos nos dados reais. As estatísticas que calculamos são as mesmas que foram usadas em [8]:

- Número de tentativas antes da primeira evitação (lembre-se que o cão sempre leva um choque na primeira tentativa).
- Número de tentativas antes da segunda evitação.
- Número total de choques nas 25 tentativas de cada cão.
- Número de tentativas antes do último choque.
- Número de revezamentos entre choques e evitações.

- Duração da série mais longa de choques sucessivos.
- Número de tentativas antes da primeira série de 4 evitações sucessivas.

Mosteller e Bush (1955) foram pioneiros neste tipo de compração dos dados reais com dados gerados por simulação usando um modelo teórico. Eles escreveram: "In the testing of a scientific model or theory, one rarely has a general measure of goodness-of-fit, a universal yardstick by which one accepts or rejects the model. Indeed, science does not and should not work this way; a theory is kept until a better one is found. One way that science does work is by comparing two or more theories to determine their relative merits in handling relevant data. In this paper we present a comparison of eight models for learning by using each to analyze the data from the same experiment.1 A primary goal of any learning model is to predict correctly the learning curve-proportions of correct responses versus trials. Almost any sensible model with two or three free parameters, however, can closely fit the curve, and so other criteria must be invoked when one is comparing several models. A criterion that has been used in recent years is the extent to which a model can reproduce the fine-grain structure of the response sequences. Many properties can be and have been invented for this purpose. Fourteen such properties are used in this paper. A summary index of how well one model fits the fine-grain detail of data compared with another model is the likelihood ratio. There are three objections to this measure, however. First, for many models it is very difficult to compute. Second, its use obscures the particular strengths and weaknesses of a model and so fails to suggest why the model is inadequate. Third, it may be especially sensitive to uninteresting differences between the model and the experiment. Therefore we do not use likelihood ratios in this paper."

```
## Read data
## Data are at http://www.stat.columbia.edu/~gelman/arm/examples/dogs

y1 <- as.matrix (read.table ("dogs.txt"), nrows=30, ncol=25)
y <- ifelse (y1[,]=="S",1,0)

nsim = 1000 # number of simulations
ntrials = 25 # number of trials for each dog
dogs.sim = matrix(NA, nrow=nsim, ncol=ntrials)
dogs.sim[,1] = 1   # first trial is always a shock
q1 = 1; alpha1 = 0.80; alpha2 = 0.92 # parameters
for(k in 1:nsim){
    qnew = q1
    for(i in 2:ntrials){
      shock = runif(1) < qnew # in trial k: shock or avoidance?
      dogs.sim[k, i] = shock  # 1=shock, 0=avoidance
      qnew = qnew * ifelse(shock, alpha2, alpha1) # update q = Pr(shock)
    }
}

# Summary Statistics
first_avoidance = numeric(nsim); second_avoidance = numeric(nsim)
total_shocks = numeric(nsim); trials_until_last_shock = numeric(nsim)
alternations = numeric(nsim); longest_run_shocks = numeric(nsim)
trial_before_4run = numeric(nsim)

for(k in 1:nsim){
  aux = which(dogs.sim[k,] == 0)
```

```
  first_avoidance[k] = aux[1]
  second_avoidance[k] = aux[2]
  total_shocks[k] = sum(dogs.sim[k,])
  aux = dogs.sim[k,] == 1
  trials_until_last_shock[k] = max(which(dogs.sim[k,] == 1))
  alternations[k] = sum( (dogs.sim[k,2:25] - dogs.sim[k,1:24]) != 0 )
  x.rle = rle( dogs.sim[k,] )
  longest_run_shocks[k] = max( x.rle$lengths[x.rle$values == 1] )
  z = which( x.rle$values == 0 & x.rle$lengths >= 4 )[1]
  if(is.na(z) | z ==1) trial_before_4run[k] = NA
    else trial_before_4run[k] = sum(x.rle$lengths[1:(z-1)])
}

mean(first_avoidance); sd(first_avoidance)
mean(second_avoidance); sd(second_avoidance)
mean(total_shocks); sd(total_shocks)
mean(trials_until_last_shock); sd(trials_until_last_shock)
mean(alternations); sd(alternations)
mean(longest_run_shocks); sd(longest_run_shocks)
mean(trial_before_4run, na.rm=T); sd(trial_before_4run, na.rm=T)
```

A Tabela 11.1 mostra os valores médios e os desvios-padrão das estatísticas que resumem aspectos da sequência de choques e evitações que cada cão vivencia. Mostramos os valores calculados com os 30 cães reais e os valores calculados com os stat-cães simulados com o modelo TOL.

Table 11.1: Comparação entre o modelo TOL e os dados reais no experimento de aprendizagem com cães.

Statistic		Real	TOL
# Trials before first avoidance	$\bar{x}$	4.50	5.29
	SD	2.25	2.24
# Trials before second avoidance	$\bar{x}$	6.47	7.46
	SD	2.62	2.27
Total number of shocks	$\bar{x}$	7.80	8.67
	SD	2.52	2.33
# Trials before last shock	$\bar{x}$	11.33	14.56
	SD	4.36	4.73
Number of alternations	$\bar{x}$	5.47	5.62
	SD	2.72	2.34
Length of longest run of shocks	$\bar{x}$	4.73	5.43
	SD	2.03	2.14
# Trials before first run of 4 avoidances	$\bar{x}$	9.70	10.53
	SD	4.14	3.72

■

Exerc: Suponha que um certo cão teve a seguinte sequência se choques (S) e evitações (A) no experimento descrito no exemplo **Modelos de aprendizagem comportamental** do capítulo de Simulação Monte Carlo: SSASSASASAAA. Obtenha a probabilidade dessa sequência ocorrer usando os símbolos α_1 e α_2. Obtenha um valor numérico substituindo $\alpha_1 = 0.80$ e $\alpha_2 = 0.92$.

Simular outro modelo

11.4 Simulação de v.a. Binomial

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 e somar o número de sucessos. 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. A explicação é simples. Imagine que n é grande (por exemplo, $n = 10000$) e p é pequeno (tal como $p = 0.01$). Uma distribuição $\text{Bin}(10000, 0.01)$ tem valor esperado igual a $np = 100$ e desvio-padrão $\sqrt{np(1-p)} = 9.95$. A maioria de seus valores vai ficar em torno de 2-3 desvios-padrão de seu valor esperado. Assim, a grande maioria dos valores binomiais nãO passa de 120 ou 130. Talvez não seja necessário gerar todos os 10 mil ensaios de Bernoulli pois a maioria deles será um frscasso. 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 Bin(`size`, `prob`). WARNING: No HELP do *R*, o argumento n refere-se a quantos valores binomiais Bin(`size`, `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 Bin(size,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
```

■ **Example 11.6** Completar - Feller? ■

11.5 Simulação de v.a. Multinomial

Completar.

```
import numpy as np
probabilities = [0.3, 0.4, 0.3]  # Probabilities for 3 categories
n = 1000  # Number of trials
```

```
# Uses random.multinomial from the NumPy library
samples = np.random.multinomial(n, probabilities)
# return value is a list, each element is #times of the category.
# Display the results
for i, count in enumerate(samples):
    print(f"Category {i+1}: {count} occurrences")
```

11.6 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 $\{x_1, x_2, \ldots\}$. Temos $\mathbb{P}(X = x_i) = p_i > 0$ para $i = 0, 1, \ldots$ 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 $i = 0, 1, 2, \ldots$.

A distribuição de X é dada pelas duas primeiras colunas na tabela 11.2

x_i	$P(x = x_i) = p_i$	$\sum_{j=1}^{i} p_j$
x_1	p_1	p_1
x_2	p_2	$p_1 + p_2$
x_3	p_3	$p_1 + p_2 + p_3$
$\vdots$	$\vdots$	
Total	$\sum_i p_i = 1$	-

Table 11.2: Distribuição de probabilidade discreta: valores possíveis (primeira coluna), probabilidades associadas (segunda coluna), probabilidades acumuladas (terceira coluna).

Na terceira coluna da tabela acima, nós acumulamos as probabilidades obtendo $\mathbb{F}(x_i) = P(X \leq x_i) = \sum_{j=1}^{i} p_j$:

$$\begin{aligned} \mathbb{F}(x_1) &= p_1 \\ \mathbb{F}(x_2) &= p_1 + p_2 \\ \mathbb{F}(x_3) &= p_1 + p_2 + p_3 \text{ Etc.} \end{aligned}$$

O algoritmo é muito simples:

- Se $0 \leq U < \mathbb{F}(x_1) = p_1$, 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:

$$X = g(U) = \begin{cases} x_1, & \text{se } U < p_1 \\ x_2, & \text{se } p_1 \leq U < p_1 + p_2 \\ x_3, & \text{se } p_1 + p2 \leq U < p_1 + p_2 + p_3 \\ \ldots & \ldots \\ x_i, & \text{se } \sum_{j=1}^{i-1} p_j \leq U < \sum_{j=1}^{i} p_j \\ \ldots & \ldots \end{cases}$$

De forma mais resumida, podemos escrever que $X = g(U) = x_j$ se $\mathbb{F}(x_{j-1}) \leq U < \mathbb{F}(x_j)$. Esta variável aleatória possui a distribuição de probabilidade desejada, como mostraremos após o exemplo seguinte.

■ **Example 11.7 — 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 probabiliade 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 $0.25 \leq 0.4897 < 0.60$. ■

Por que este algoritmo gera X com a distribuição desejada e apresentada na Tabela 11.2? Claramente, o algoritmo gera números da lista de valores possíveis dessa distribuição. Qual a probabilidade de que obtermos $X = x_i$ por este algoritmo? Teremos $X = x_i$ se U cair no intervalo $[a,b)$ onde $a = p_1 + \ldots + p_{i-1}$ e $b = p_1 + \ldots + p_{i-1} + p_i \leq 1$, um intervalod com comprimento igual a p_i. Assim, $\mathbb{P}(X = x_i) = \mathbb{P}(U \in (a,b]) = p_i$. Desta forma, estamos gerando X com a distribuição de probabilidade desejada.

11.7 Gerando Poisson

A distribuição de Poisson É discreta e portanto, um caso especial do que vimos na seção anterior. Para o caso de $X \sim \text{Poisson}(1.61)$, temos $\mathbb{P}(X = i) = (1.61)^i/i!e^{-1.61}$. Acumulando estas probabilidades, teríamos o seguinte algoritmo:

$$X = g(U) = \begin{cases} 0 & \text{se } U < 0.1998876 \\ 1 & \text{se } 0.1998876 \leq U < 0.5217067 \\ 2 & \text{se } 0.5217067 \leq U < 0.7807710 \\ \ldots & \ldots \\ i & \text{se } 0.1998876 \sum_{k=1}^{i-1} (1.61)^i/i! \leq U < 0.1998876 \sum_{k=0}^{i} (1.61)^i/i! \\ \ldots & \ldots \end{cases}$$

Neste exemplo da Poisson, existe uma dificuldade aparente: 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:

$$p_{i+1} = \frac{\lambda^{i+1}}{(i+1)!} e^{-\lambda} = \frac{\lambda}{i+1} \frac{\lambda^i}{i!} e^{-\lambda} = \frac{\lambda}{i+1} p_i \,.$$

O código em R para este procedimento com $\lambda = 1.61$ seria:

```
lambda <- 1.61; x <- -1; i <- 0; p <- exp(-lambda); F <- p
u <- runif(1)
while(x == -1){
 if(u < F) x <- i
  else{
    p <- lambda*p/(i+1)
    F <- F + p
    i <- i+1
  }
 }
```

The same in python:

```
import random,  math
lambda = 1.61; x = -1; i = 0; p = math.exp(-lambda); F = p
u = random.random()
while x == -1:
    if u < F:
        x = i
    else:
        p = (lambda * p) / (i + 1)
        F += p
        i += 1

print("x:", x)
```

■ **Example 11.8 — 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 aleatórios correspondentes da distribuição de Poisson com $\lambda = 1.7$.

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, 1, 2, respectivamente. ■

■ **Example 11.9 — 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.

```
x <- -1
i <- 0; c <- p/(1-p); pr <- (1-p)^n; F <- pr
u <- runif(1)
while(x == -1){
 if(u < F) x <- i
  else{
```

```
    pr <- ((c*(n-i))/(i+1))*pr
    F <- F + pr
    i <- i+1
  }
 }
```

O mesmo, em python:

```
import random

x = -1; i = 0
p = 0.5  # Update with your desired probability
c = p / (1 - p)
n = 10  # Update with your desired value
pr = (1 - p) ** n
F = pr
u = random.random()

while x == -1:
    if u < F:
        x = i
    else:
        pr = ((c * (n - i)) / (i + 1)) * pr
        F += pr
        i += 1

print("x:", x)
```

■

11.8 Método da transformada inversa

Para variáveis contínuas, um dos métodos mais simples e eficientes é o da transformada inversa, criada por Stanislaw Ulam logo após a Segunda Guerra Mundial (ver um pouco da fascinante história da ciração dos métodos Monte Carlo na seção 11.17). Suponha que certa variável aleatória contínua X possua função distribuição acumulada dada por $\mathbb{F}(x)$. Por exemplo, se $X \sim \exp(3)$ então $\mathbb{F}(x) = 1 - \exp(-3x)$ para $x \geq 0$ e $\mathbb{F}(x) = 0$ se $x < 0$. Para gerar X com esta distribuição, vamos gerar uma variável uniforme $U \sim U(0,1)$ e transformá-la usando a função matemática $\mathbb{F}^{-1}$. Isto é, vamos gerar U e aseguir obter $Y = \mathbb{F}^{-1}(U)$. A variável aleatória Y possui a mesma distribuição que X pois, como veremos, a função distribuição acumulada de Y é exatamente $\mathbb{F}$.

A Figura 11.3 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^{-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 aleaória $Y = \mathbb{F}^{-1}(U)$ é $\mathbb{F}$.

Por exemplo, se $X \sim \exp(3)$ então $\mathbb{F}(x) = 1 - \exp(-3x)$ para $x \geq 0$. Calcule a inversa de $\mathbb{F}$. Basta igualar a expressão de $\mathbb{F}(x)$ a um valor u e inverter isolando x como função de u. Se

$$u = 1 - \exp(-3x)$$

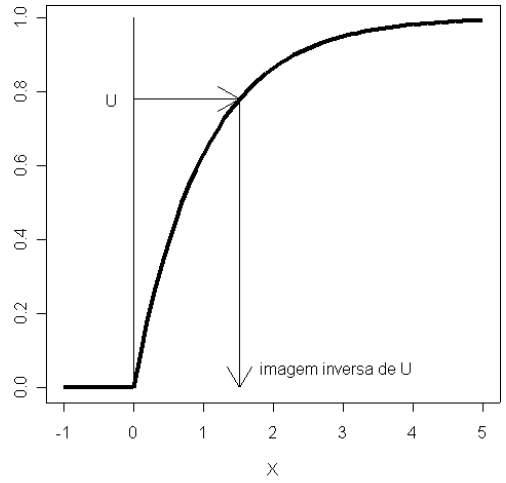


Figure 11.3: Gráfico da função distribuição acumulada $\mathbb{F}(x)$ de certa variável aleatória. U é um número aleatório com distribuição $U(0,1)$. O gráfico mostra sua imagem inversa $\mathbb{F}^{-1}(U)$ através da distribuição acumulada de X. Esta imagem é aleatória e possui distribuição igual a de X.

então

$$1-u = \exp(-3x)\,.$$

Tomando log dos dois lados, temos

$$\log(1-u) = -3x$$

Isolando x do lado esquerdo, encontramos

$$x = -\frac{1}{3}\log(1-u)\,.$$

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 num ponto y é exatamente $\mathbb{F}(y) = 1-\exp(-3y)$. De fato, se $y > 0$, nós temos

$$\begin{aligned}
\mathbb{P}(Y \leq y) &= \mathbb{P}(-1/3\log(1-U) \leq y) \\
&= \mathbb{P}(\log(1-U) \geq -3y) \\
&= \mathbb{P}\left(1-U \geq e^{-3y}\right) \qquad \text{tomando exp dos dois lados} \\
&= \mathbb{P}(U \leq 1-e^{-3y}) \\
&= \mathbb{P}(U \in (0, 1-e^{-3y})) \\
&= 1-e^{-3y} \qquad \text{pois } U \sim U(0,1)
\end{aligned}$$

Repetindo o argumento acima, pode-se gerar uma exponencial com qualquer parâmetro λ. Se $Y \sim \exp(\lambda)$ use a transformação $Y = -1/\lambda\ \log(1-U)$.

R 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. Portanto, ao invés de usar $Y = -1/\lambda \log(1-U)$ podemos usar $Y = -1/\lambda \log(U)$. O valor aleatório Y terá distribuição $\exp(\lambda)$.

A mesma prova dada acima pode ser usada para mostrar o resultado de forma geral, para qualquer v.a. contínua com c.d.f $\mathbb{F}$ e para a qual tenhamos a inversa $\mathbb{F}^{-1}$. 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,

$$\begin{aligned}
\mathbb{F}_Y(x) &= \mathbb{P}(Y \leq x) \\
&= \mathbb{P}(\mathbb{F}^{-1}(U) \leq x) \\
&= \mathbb{P}(\mathbb{F}(\mathbb{F}^{-1}(U)) \leq \mathbb{F}(x)) \\
&= \mathbb{P}(U \leq \mathbb{F}(x)) \\
&= \mathbb{F}(x)
\end{aligned}$$

11.9 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 de indivíduos que não falecem nos primeiros anos de vida. Na maioria das vezes, estes são indivíduos adultos são aqueles para os quais os seguros de vida são emitidos. Seja X a v.a. que representa o tempo total de vida a partir do nascimento de um indivíduo escolhido ao acaso de uma população. Seu suporte é o intervalo $(0,\infty)$. Para x neste intervalo, temos a densidade de probabilidade dada por

$$f(x) = Bc^x e^{-B(c^x-1)/\log(c)}$$

onde c e B são duas constantes conhecidas com $c > 1$ 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 $\mathbb{F}(x)$ é

$$\mathbb{F}(x) = 1 - \exp\left(-\frac{B}{\log(c)}(c^x - 1)\right)$$

onde $c > 1$ e $B > 0$. A Figura 11.4 mostra a função densidade $f(x)$ (acima) e a distribuição acumulada $\mathbb{F}(x)$ (abaixo) de uma v.a. Gomperz com parâmetros $c = 1.09$ e $B = 0.000102$. Esta figura foi feita usando o seguinte código em R:

```
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)")
```

The same in python:

```
import numpy as np
import matplotlib.pyplot as plt

ce = 1.09; B = 0.000102; k = B / np.log(ce)
```

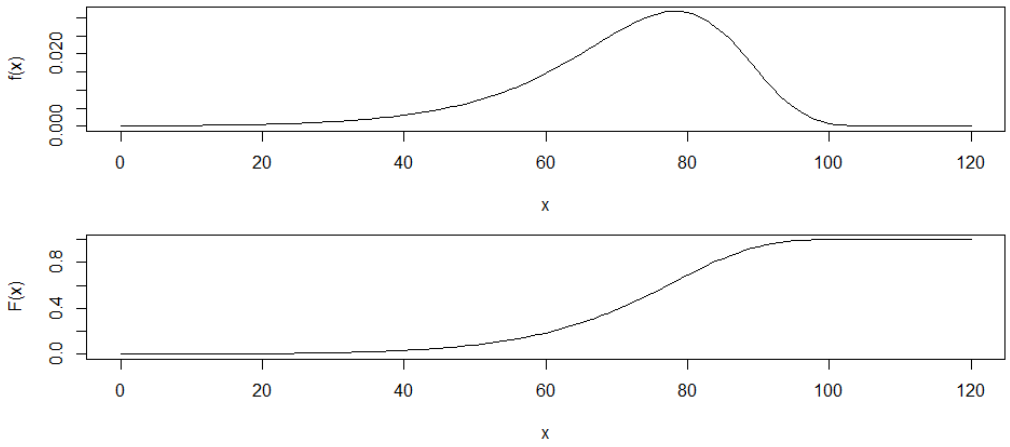


Figure 11.4: Densidade $f(x)$ (acima) e distribuição acumulada $\mathbb{F}(x)$ de uma v.a. Gomperz com parâmetros $c = 1.09$ e $B = 0.000102$.

```
eixox = np.arange(0, 121, 1)
dens = B * ce ** eixox * np.exp(-k * (ce ** eixox - 1))
Fx = 1 - np.exp(- (B / np.log(ce)) * (ce ** eixox - 1))

plt.figure(figsize=(8, 8))
plt.subplot(2, 1, 1)
plt.plot(eixox, dens, 'b-')
plt.xlabel('x')
plt.ylabel('f(x)')

plt.subplot(2, 1, 2)
plt.plot(eixox, Fx, 'r-')
plt.xlabel('x')
plt.ylabel('F(x)')

plt.tight_layout()
plt.show()
```

A transformada inversa de uma Gomperz é facilmente obtida:

$$\mathbb{F}^{-1}(u) = \log\left(1 - \log(c)\log(1-u)/B\right)/\log(c)$$

Com isto obtemos a amostra (ver Figura 11.5). 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)
```

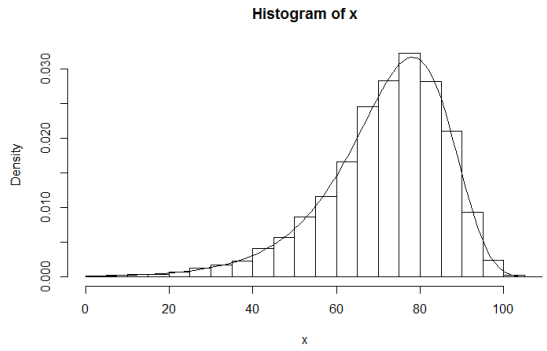


Figure 11.5: Histograma dos valores gerados de uma Gomperz com parâmetros $c = 1.09$ e $B = 0.000102$ e densidade de probabilidade.

```
dens <- B * ceˆeixox * exp(-k * (ceˆeixox - 1))
lines(x,y)
```

In python:

```
import numpy as np
import matplotlib.pyplot as plt

ce = 1.09; B = 0.000102; k = B / np.log(ce)

u = np.random.uniform(size=10000)
x = (1 / np.log(ce)) * (np.log(1 - np.log(1 - u) / k))

plt.hist(x, density=True)
eixox = np.arange(0, 121, 1)
dens = B * ce ** eixox * np.exp(-k * (ce ** eixox - 1))
plt.plot(eixox, dens, 'r-')

plt.show()
```

11.10 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. 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. Suponha que ocorra uma perda monetária de valor x para um segurado. Se $x \leq x_0$, a seguradora não cobre nada da perda (em geral, ela nem é notificada pelo segurado acerca perda). Se $x > x_0$, a seguradora cobre o valor excedente $x - x_0$.

O segundo parâmetro da distribuição Pareto é representado por α e é positivo. Ele 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

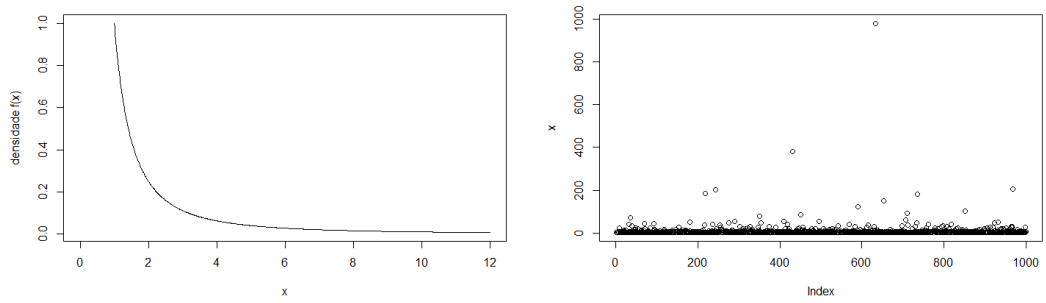


Figure 11.6: (Left) Densidade da Pareto com $x_0 = 1$ e $\alpha = 1$. (Right) Amostra de 1000 valores i.i.d. de uma Pareto com $x_0 = 1$ e $\alpha = 1$.

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 no lado esquerdo da Figura 11.6. No lado direito, temos a sequência de 1000 valores i.i.d. simulados dessa densidade de probabilidade.

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. 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, $\alpha \in (1, 2.5)$. 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 no lado direito da Figura 11.6

O efeito do parâmetro α na geração de valores extremos fica mais claro na Figura 11.7. 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 amosra da Pareto, $i = 1, \ldots, 100$. 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).

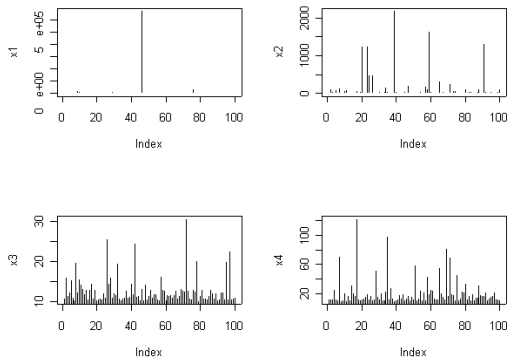


Figure 11.7: Amostras de 100 valores Pareto com (x_0, α) igual a $(1.3, 0.25)$ (canto superior esquerdo), $(1.3, 0.5)$ (canto superior direito), $(10, 5)$ (canto inferior esquerdo) e $(10, 2)$ (canto inferior direito).

11.11 Summary of the Inverse Transform Method

Add details about flat F, discrete F, non-existing F, Table with F's and F-inverses.

Let $\mathbb{F} : (a, b) \to (0, 1)$ be the cumulative distribution function of a random variable X. The interval (a, b) do not have to be finite, as we have seen in the previous examples using the Pareto and Gomperz distributions. Assume that $\mathbb{F}$ is invertible with inverse function $\mathbb{F}^{-1} : (0, 1) \to (a, b)$.

Theorem 11.11.1 — Monte Carlo Inverse Transform Method. If $U \sim U(0, 1)$, then the random variable $\mathbb{F}^{-1}(U)$ has the same distribution as X. That is, $X \overset{d}{\sim} \mathbb{F}^{-1}(U)$.

Proof. To do ■

There is a reverse theorem that is extremely useful. We will have many opportunities in the rest of this book to make use of it. This reverse theorem says that if you transform the random variable X using its own cumulative distribution $\mathbb{F}$ then we obtain a uniform distribution. As before, assume that $\mathbb{F}$ is invertible with inverse function $\mathbb{F}^{-1} : (0, 1) \to (a, b)$.

Theorem 11.11.2 — Reverse Transform Method. If $Y = \mathbb{F}(X)$ then $Y \sim U(0, 1)$.

Proof. $Y = \mathbb{F}(X)$ is a random variable. Its c.d.f. is given by

$$\mathbb{P}(Y \leq y) = \mathbb{P}(\mathbb{F}(X) \leq y) = \mathbb{P}(X \leq \mathbb{F}^{-1}(y)) = \mathbb{F}\left(\mathbb{F}^{-1}(y)\right) = y$$

As Y has the same c.d.f. of an uniform distribution in $(0, 1)$ and as the c.d.f. determines uniquely the probability distribution, we conclude that $Y \sim U(0, 1)$. ■

This result is not very useful in the context of this chapter but it will come out several times along this book. In my teaching experience, students tend to be more surprised by this reverse theorem than the previous one.

■ **Example 11.10** Rather than using our own functions to generate random variables with a given probability distribution, we use directly python libraries to do that. We will simulate a large number of Gaussian random variables with mean 0 and standard deviation 10. Therefore, the values should be concentrated around 0 and mostly ranging between -20 and 20. We use the native function ?? that implements the c.d.f. $\mathbb{F}$ of a Gaussian distribution. A histogram of the transformed values in

seen in the left plot of Figure ??. The right plot in this Figure shows ?? the same experiment with ?? distribution. The code is below. ∎

```
To do
```

R O caso de uma v.a. discreta X usa o método da transformada inversa. Neste caso, a função acumulada de probabilidades $\mathbb{F}(x)$ é descontínua, com saltos nos pontos x que correspondem aos valores discretos possíveis x_i. Ela não possui a inversa usual. Entretanto, se definirmos

$$\mathbb{F}^{-1}(U) = \inf\{x \in \mathbb{R} : \mathbb{F}(x) \geq U\}$$

teremos o método descrito anteriormente para v.a.'s discretas.

11.12 Gerando v.a. gaussiana ou normal

A distribuição normal é muito especial pois ela é a aproximação para a soma de variáveis independentes (Teorema Cenral 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 disribuição 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 já sabe gerar uniformes e exponenciais, você pode usar este resultado para gerar normais padronizadas. Em *R*, basta criar a função abaixo:

```
minharnorm = function(n){
 sqrt(rexp(n, 0.5)) * cos(runif(n, 0, 2 * pi))
}

In python:
import numpy as np
def minharnorm(n):
 v = np.random.exponential(scale=0.5, size=n))
 u = np.random.uniform(0, 2 * np.pi, size=n)
    return np.sqrt(v *  np.cos(u))

# Exemplo de uso
n = 1000
result = minharnorm(n)
```

Geramos uma amostra com o código R abaixo e mostramos o resultado na Figura 11.8.Sobrepomos ao histograma a função densidade de probabilidade de uma $N(0,1)$.

```
  set.seed(123)
  minharnorm = function(n){ sqrt(rexp(n, 0.5)) * cos(runif(n, 0, 2 * pi))}
  hist(minharnorm(1000), prob=T)
  plot(dnorm, -3,3, add=T)
```

Como gerar gaussianas $N(\mu,\sigma^2)$, centrada em μ e com desvio-padrão σ? Usamos uma propriedade da distribuição gaussiana: se $Z \sim N(0,1)$ então $X = \mu + \sigma Z \sim N(\mu,\sigma^2)$. Veremos a prova dessa propriedade no Capítulo 13. 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 $Z \sim N(0,1)$ e em seguida tomar $X = 10 + \sqrt{4}\,Z$. Em *R*:

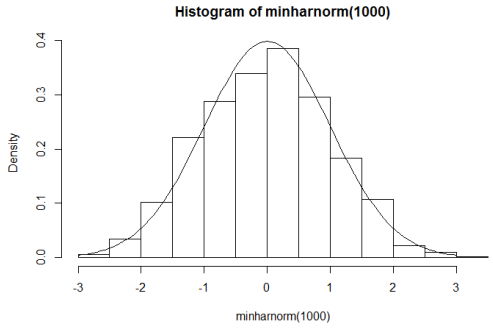


Figure 11.8: Histograma padronizado de 1000 valores $N(0,1)$ gerados com função `minhanorm` com a densidade $f(x)$ sobreposta.

```
x = 10 + sqrt{4} * minhanorm(100)
```

É claro que, sendo a gaussiana uma distribuição muito importante, *R* já possui um gerador de gaussianas: `rnorm(100, mean=0, sd=1)`.

11.13 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 $U_1, U_2, \ldots, U_n$ são i.i.d. $U(0,1)$ então as v.a.'s $Y_1 = g(U_1), Y_2 = g(U_2), \ldots, Y_n = g(U_n)$ também são i.i.d. com esperança θ. Pela Lei dos Grandes Números (ver capítulo 19), se $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n} g(U_i) \to E[g(U)] = \theta$$

Assim, se n é grande, θ é aproximadamente a média aritmética dos valores simulados $g(u_i)$.

■ **Example 11.11 — 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:

$$u_1 = 0.4886415, u_2 = 0.1605763, u_3 = 0.8683941, \ldots, u_{1000} = 0.3357509$$

Calculamos então

$$\begin{aligned}\hat{\theta} &= \left(u_1^2 + u_2^2 + \ldots + u_{1000}^2\right)/1000 \\ &= \left((0.4886415)^2 + (0.1605763)^2 + \ldots + (0.3357509)^2\right)/1000 \\ &= 0.33406 \approx \theta\end{aligned}$$

Se fizermos uma nova geração das U_i, com uma semente diferente, vamos produzir $\hat{\theta}$ ligeiramente diferente. Outros 1000 valores da uniforme produzem $\hat{\theta} = 0.32468$. Aumentando o tamanho da amostra a variação de uma simulação para outra diminui. A escolha do tamanho da amostra para obter certa precisão desejada necessita do uso do Teorema Central do Limite que que veremos no capítulo 19. ■

■ **Example 11.12 — 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 100000 valores i.i.d. de uma $U(0,1)$ e calcule $(y_1 + y_2 + \ldots + y_{1000})/1000$ 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 100000 valores: 0.3414839, 0.3413451, 0.3411779, 0.3412634. Com uma amostra de tamanho 100 mil, estamos tendo alguma varição na quarta decimal. Comparando com $\theta = 0.3413447$, os erros de estimação são pequenos. ■

11.13.1 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 11.13 — Integral com limites genéricos.** Calcule o valor aproximado de

$$\theta = \int_3^9 \log(2 + |\sin(x)|)e^{-x/20}\, dx$$

Uma amostra $U_1, U_2 \ldots$ 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

$$(9-3)\bar{W} = 6\, \frac{W_1 + \ldots + W_{100000}}{100000}$$

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. ■

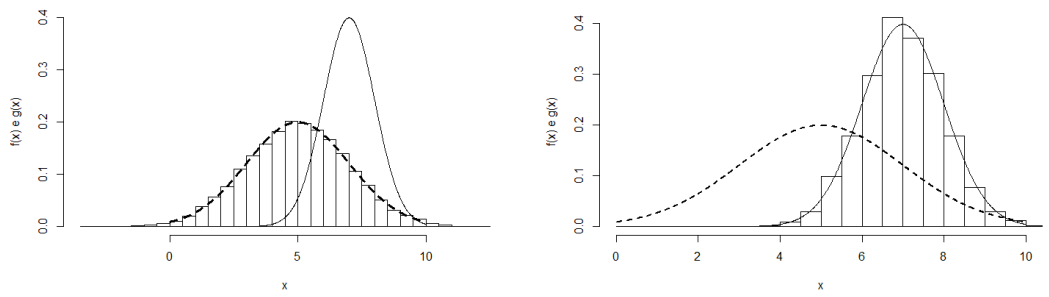


Figure 11.9: Left: Linha contínua: densidade $f(x)$ de onde queremos amostrar. Linha tracejada: densidade $g(x)$ de onde sabemos amostrar. histograma de amostra de 20000 elementos de $f(x)$. Histograma dos 3696 elementos da amostra anterior que restaram após rejeitar seletivamente 16304 dos elementos gerados. Right: 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 $f(x)$.

11.14 Método da aceitação-rejeição

Queremos gerar amostra de densidade $f(x)$. Suponha que não somos capazes ou é impossível obter $\mathbb{F}(x)$ ou $\mathbb{F}^{-1}(u)$ analiticamente. 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 após o descarte possui a distribuição com densidade $f(x)$.

A essência dessa ideia está no lado esquerdo Figura 11.9. Suponha que sabemos gerar com facilidade da densidade $g(x)$ (linha tracejada). Uma amostra gerada a partir de $g(x)$ produz o histograma visto na figura. Mas queremos amostra de $f(x)$, a densidade em linha contínua. Eliminamos de forma seletiva alguns dos valores gerados de $g(x)$. 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 no lado direito 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 $f(x)$ pode gerar um valor x então $g(x)$ também deveria ser capaz de gerar este x. Ou seja, se $f(x) > 0$ então devemos ter $g(x) > 0$. Isto significa que o suporte de g deve conter o suporte de f:

$$\mathscr{S}_f = \{x : f(x) > 0\} \subseteq \mathscr{S}_g = \{x : g(x) > 0\}$$

Podemos ter $g(x)$ gerando valores impossíveis sob $f(x)$. Isto não será um problema pois eles serão eliminados da amostra final. Mas não podemos permitir que valores possíveis sob $f(x)$ sejam impossíveis de serem simulados 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 $M = 2$, comparamos o valor de $f(x)$ com $2g(x)$, duas vezes a altura da densidade g no ponto x. Devemos ter $f(x) \leq Mg(x)$.

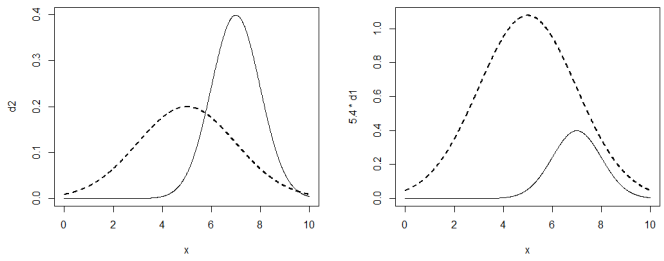


Figure 11.10: $f(x)$ e $Mg(x)$.

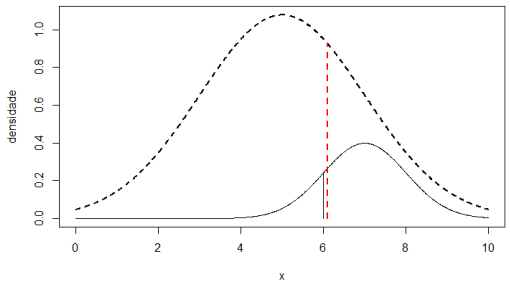


Figure 11.11: Em $x = 6.0$, temos as alturas $f(6.0)$ (linha contínua) e a altura $5.4g(6.0)$ (tracejada).

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 $5.4 * g(x)$. 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 11.14. No ponto $x = 6.0$ temos a altura $f(6.0)$ (contínua) e a 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 $x_1, x_2, \ldots$ os elementos da amostra de $g(x)$. Quais destes valores vamos reter e quais vamos rejeitar? Calcule $r(x_1), r(x_2), \ldots$ 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, vamos reter 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 bem menor que n devido à rejeição de vários elementos.

Seja Y um valor inicialmente gerado a partir de $g(x)$ e X um dos valores aceitos no final do processo.

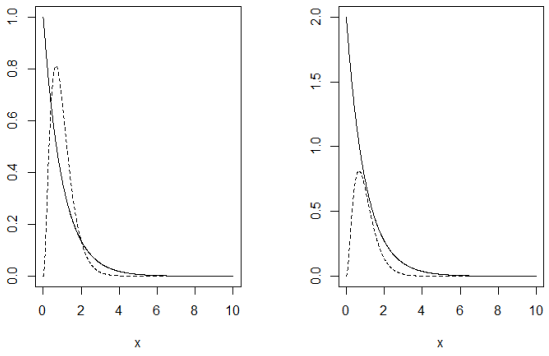


Figure 11.12: Esquerda: Densidade-alvo $f(x)$ (linha tracejada) e densidade $g(x)$ de onde sabemos gerar (linha contínua). Direita: Densidade $f(x)$ e a função $2g(x)$.

Algorithm 1 Método da Rejeição.

1: $I \leftarrow$ `True`
2: **while** I **do**
3: Gere $Y \sim g(y)$
4: Gere $U \sim \mathscr{U}(0,1)$
5: **if** $U \leq r(Y) = f(Y)/Mg(Y)$ **then**
6: $X \leftarrow Y$
7: $I =$ `False`
8: **end if**
9: **end while**

■ **Example 11.14 — Gerando de uma gama.** Queremos gerar $X \sim \text{Gamma}(3,3)$ com densidade:

$$f(x) = \begin{cases} 0 & , \text{ se } x \leq 0 \\ \frac{27}{2}x^2e^{-3x} & , \text{ se } x \geq 0 \end{cases} \tag{11.1}$$

Esta densidade tem suporte $\mathscr{S}_f = (0,\infty)$. Sabemos gerar $W \sim exp(1)$ que possui o mesmo suporte que f. Para isto, basta tomar $W = -\log(1-U)$ onde $U \sim \mathscr{U}(0,1)$. A densidade de W é:

$$g(x) = \begin{cases} 0, & \text{se } x < 0 \\ e^{-x}, & \text{se } x \geq 0 \end{cases} \tag{11.2}$$

As duas densidades estão na Figura 11.12. Então:

$$0 \leq \frac{f(x)}{g(x)} = \frac{\frac{27}{2}x^2e^{-3x}}{e^{-x}} = \frac{27}{2}x^2e^{-2x} \tag{11.3}$$

Derivando 11.3 em relaç ao a x 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)
```

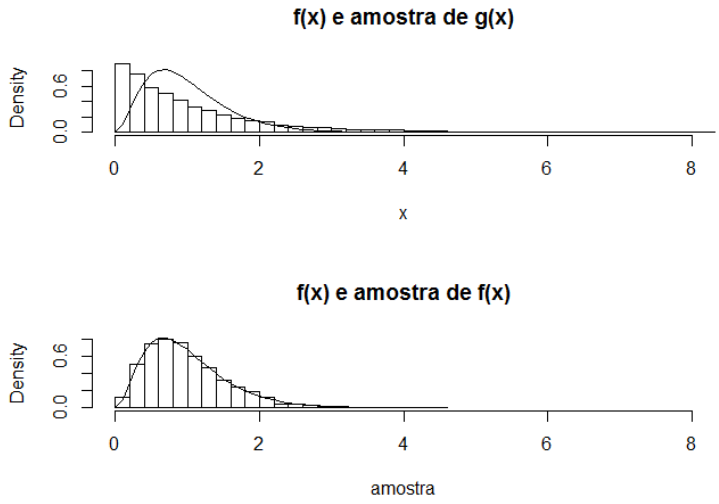


Figure 11.13: Amostra de 10 mil valores de uma $g(x) = \exp(1)$; rejeitando aproximadamente 5000 valores terminamos com amostra de $f(x) = \text{Gama}(3,3)$.

```
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)) ]
```

Pseudo-code:

1: $I \leftarrow true$
2: **while** I **do**
3: Selecione $U \sim \mathscr{U}(0,1)$
4: Selecione $U^* \sim \mathscr{U}(0,1)$
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**

■

11.14.1 Dois teoremas

Temos dois teoremas para este método.

Theorem 11.14.1 — Aceitação-Rejeição gera valores de $f(x)$. A variável aleatória X gerada pelo método de aceitação-rejeição possui densidade $f(x)$.

Prova: Leitura opcional na seção 11.21.

Theorem 11.14.2 — Impacto de M. O número de iterações necessários até que um valor seja aceito possui distribuição geométrica com valor esperado M.

Prova: Leitura opcional na seção 11.21.

11.14.2 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 à condição. Se rodarmos o método em paralelo com 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, teremos $f(x) = g(x)$. 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. Em geral, $M > 1$ e muitas vezes $M >> 1$.

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. Assim, embora teoricamente $g(x)$ possa ser quase qualquer densidade desde que $\mathscr{S}_f \subseteq \mathscr{S}_g$, na prática procuramos de alguma forma ter uma ideia de $f(x)$ para escolhermos $g(x)$ similar a f.

11.15 Importance Sampling

This is an important method that can be used when we must simulate a vector of non-independent random variables. However, since we have studied joint distributions yet, we will consider only a univariate random variable. In the acceptance-rejection method, we: (a) select from a working density $g(x)$ from which we know how to sample from. (b) we keep some simulated values and delete the others; (c) this is done in such a way that the kept values follow the desired target density $f(x)$. In the importance sampling method, we also start by sampling from $g(x)$. However, we keep everything that has been generated, we do not delete anything. Then we assign an appropriate weight to each generated point. In the end, these weights correct the distortion of not having a sample of f(x).

11.15.1 The intuition

Let $f(x)$ in the left hand side of Figure 11.15.1 be the target density, the distribution from which we want to sample from. The "carpet" of points in the horizontal axis represent a sample extracted from $f(x)$. All of them have the same weight, equal to 1.

However, we do not have this sample because we do not know how to sample directly from $f(x)$. Instead, we have a sample from $g(x)$, a density we know how to sample from. This other

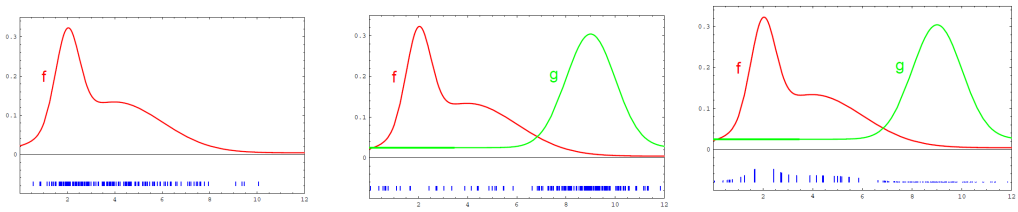


Figure 11.14: The points shown in horizontal axis show a random sample from $f(x)$ (right) and from $g(x)$ (center). The right plot show the sample from $g(x)$ weighted according to $w(x) = f(x)/g(x)$.

$g(x)$-based sample is shown in the middle of Figure 11.15.1 as the points forming the "carpet"in the horizontal axis. They again have the equal weights. The next step is assign weights to these points such that that look like having been generated from $f(x)$. Intuitively, how to do that? Which points gets larger weights, and which ones gets the smaller weights?

Let's look at the answer directly and then discuss the intuition behind it. The right hand side plot in Figure 11.15.1 shows the weights for the sample from $g(x)$ in the middle plot. The two densities distribute their probability masses very differently, with $f(x)$ having higher values in regions different from where $g(x)$ has higher values. Consider initially the leftmost region on the horizontal axis, where $g(x)$ is much smaller than $f(x)$. The intuitive idea is to give larger importance (or larger weight) to those rarely g-sampled points in these regions. It is as if we had each g-rare point in this region copied multiple times so that the g-sample is increased in these regions to reflect what we expect from a sample of $f(x)$. Points sampled from g in this low f region can get copied 5, 6 times. Conversely, looking at the rightmost region in the horizontal axis, we have $g(x)$ large and $f(x)$ small. The many points sampled from g in this region must be reduced in some way to reflect their low chance of appearing under f. Hence, each point in this region may be reduced to only 1/5 so that we need five g-sampled points to stand for a single sampled point from f in this region.

The exact formula behind this intuition is that we need to assign weight equal to $w(x_i) = f(x_i)/g(x_i)$ to each x_i element sampled from $g(x)$. The weight is depends on the ration between $f(x)$ and $g(x)$. If $f(x)$ is large but $g(x)$ is equally large and with approximately the same value (that is, if $f(x_i) \approx g(x_i)$), then we have $w(x_i) \approx 1$. If $f(x_i) << g(x_i)$ then $w(x_i) >> 1$ while $w(x_i) << 1$ when $g(x_i) << f(x_i)$. The entire sample of g, not rejecting anyone but duly weighted by $w(x)$, can be used to make inferences about $f(x)$.

11.15.2 Interest quantities as expectations

What we want to know about $f(x)$? We want a Monte Carlo sample to estimate (to know approximately) some aspects of a random variable X with target distribution $f(x)$. For example, we might want to know the following:

- $\theta_1 = \mathbb{E}_f(X)$, the expectation of $X \overset{d}{\sim} f(x)$, without having to calculate the integral, which can be a very difficult task. It will be important to make explicit which density are we using this expectation, as you will see soon.
- $\theta_2 = \mathbb{V}_f(X) = \mathbb{E}_f(X^2) - (\mathbb{E}_f(X))^2$, the variance of $X \overset{d}{\sim} f(x)$.
- $\theta_3 = \mathbb{P}_f(X > 2)$, the probability of observing X greater than 2, a threshold-value important in some application.
- $\theta_4 = \mathbb{P}_f(e^{-|X|} > |X|)$, a probabilitic calculation that is difficult or almost impossible for most densities $f(x)$.
- $\theta_5 = \mathbb{P}_f(X \in A)$, where A is some convoluted set.

The trick to obtain Monte Carlo estimates of these quantities is to write each of them as an

expectation of certain random variable $h(X)$.

Let us consider each of the cases above. Starting the simplest one, $\theta_1 = \mathbb{E}_f(X)$. As this is already in form of an expectation, we only need to take $h(x) = x$, the identity function. With $h(x) = x$ we have trivially $\theta_1 = \mathbb{E}_f(X) = \mathbb{E}_f(h(X))$.

The second case is more interesting and uses θ_1 defined in the previous case:

$$\theta_2 = \mathbb{V}_f(X) = \mathbb{E}_f(X^2) - (\mathbb{E}_f(X))^2 = \mathbb{E}_f(X^2) - \theta_1^2$$

If $h(X) = X^2$ then $\theta_2 = \mathbb{E}_f(h(X)) - \theta_1^2$ and this will be estimated after we obtain a Monte Carlo estimate of θ_1.

Turning to $\theta_3 = \mathbb{P}_f(X > 2)$, it is possible to write this probability as an expectation of a certain random variable $h(X)$. Let $h(X) = I_{[X>2]}$. This is a binary random variable that indicates, for each realization ω, if the measurement $X(\omega)$ is greater than 2:

$$h(X)(\omega) = I_{[X(\omega)>2]} = \begin{cases} 1, & \text{if } X(\omega) > 2 \\ 0, & \text{if } X(\omega) \leq 2 \end{cases}$$

This is an important trick: the probability of a certain event may be seen as the expectation of a indicator variable, the binary random variable indicating either the event occurs or not. We apply this trick to write θ_4 as an expectation:

$$\theta_4 = \mathbb{P}(e^{-|X|} > |X|) = \mathbb{E}_f(h(X))$$

where $h(X) = I_{[X \in A]}$ with $A = \{x \text{ tais que } e^{-|x|} > |x|\}$. In simpler terms,

$$h(X)(\omega) = I_{[X(\omega) \in A]} = \begin{cases} 1, & \text{if } e^{-|X(omega)|} > |X(omega)| \\ 0, & \text{otherwise} \end{cases}$$

We also apply the same trick in the more general situation of $\theta_5 = \mathbb{P}_f(X \in A)$: we can write $\theta_5 = \mathbb{E}_f(I_A)$, where A is a complicated set.

11.15.3 Interest quantities as expectations

Suppose that we want to know approximately the value of a certain quantity θ which is expressed as the expectation of $h(X)$ where $X \overset{d}{\sim} \mho$. By the frequentist idea, we know that $\mathbb{E}(X)$ can be approximated by the arithmetic average of a large sample of X: (?? ja defini o que eh uma amostra de X??)

$$\mathbb{E}(X) \approx \frac{1}{n} \sum_{i=1}^{n} X_i$$

By the same reasoning, if we want to estimate the expected value $\mathbb{E}(h(X))$ of a new random variable $h(X)$ that is a transformation $h(X)$ of X, we can use the arithmetic mean of $h(X_i)$:

$$\mathbb{E}(h(X)) \approx \frac{1}{n} \sum_{i=1}^{n} h(X_i)$$

For example, if $h(X) = X^2$, we have

$$\mathbb{E}(X^2) \approx \frac{1}{n} \sum_{i=1}^{n} X_i^2$$

Estimando esperanças

- Simples: se quiser conhecer o valor esperado de X^2, tome uma amostra de X, aplique a função quadrática a cada valor e tome a sua média aritmética.
- Para estimar o valor esperado de qualquer função $h(X)$, transforme cada valor de uma grande amostra de $X \sim f(x)$ e tome sua média aritmética.
- Problema: não conseguimos gerar $X \sim f(x)$ desejada.
- Sabemos gerar de OUTRA distribuição $g(x)$.
- Aceitação-rejeição joga fora seletivamente vérios elementos da amostra de modo a terminar com uma amostra de $f(x)$: é como dar pesos iguais a 0 ou 1 a cada valor.
- Amostragem por importância pondera TODOS os valores amostrados de $g(x)$ com pesos mais flexíveis.

Esperança sob QUAL densidade, g ou f?

- Queremos o valor esperado de $h(X)$ onde $X \sim f(x)$ com suporte $\mathscr{S}$.
- Isto é, queremos $\theta = \mathbb{E}_f(h(X))$.
- SUB-INDICE f para indicar a distribuição de X. A partir de agora,X pode ter densidade $g(x)$ ou $f(x)$ e queremos distinguir isto na notação.
- Sabemos gerar apenas de $g(x)$, com suporte maior ou igual a $\mathscr{S}$.
- Vamos mostrar que $\theta = \mathbb{E}_f(h(X))$ pode ser visto como a esperança de OUTRA funąão $h^*(X)$ quando X tem densidade g.
- Isto é, vamos mostrar que

$$\theta = \mathbb{E}_f(h(X)) = \theta = \mathbb{E}_g(h^*(X))$$

Por que o algoritmo funciona

- O truque mais barato da matemática: multiplique e divida por um mesmo valor...

$$\begin{aligned}\theta = \mathbb{E}_f(h(X)) &= \int_{\mathbb{R}} h(x)\, f(x)\, \mathrm{d}x \\ &= \int_{\mathbb{R}} \left(h(x)\, \frac{f(x)}{g(x)} \right) g(x) \mathrm{d}x \\ &= \int_{\mathbb{R}} h^*(x)\, g(x) \mathrm{d}x\end{aligned}$$

  onde $h^*(x) = h(x)f(x)/g(x) = h(x)w(x)$ é uma nova função.
- Assim, podemos reconhecer a última expressão como uma nova esperança: o valor esperado de $h^*(X)$ quando X segue a densidade $g(x)$!!

Por que o algoritmo funciona

- Isto é,

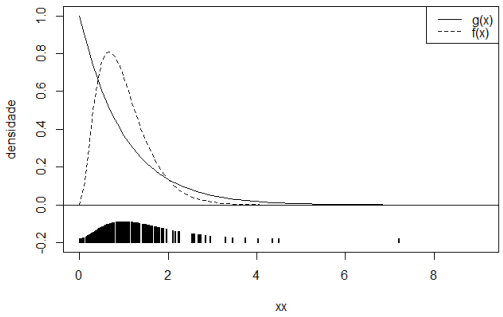
$$\theta = \mathbb{E}_f(h(X)) = \mathbb{E}_g\left(h^*(X)\right) = \mathbb{E}_g\left(h(X)\frac{f(X)}{g(X)}\right) = \mathbb{E}_g\left(h(X)w(X)\right)$$

- Note que, na última esperança, a v.a. X possui distribuição $g(x)$ e não mais $f(x)$!!
- Tudo se resume a multiplicar e dividir por um mesmo valor dentro da integral e reconhecer que a nova integral é uma esperança de uma v.a. $h^*(X)$ onde X tem OUTRA distribuição $g(x)$.

O truque

- Repetindo

$$\theta = \mathbb{E}_f(h(X)) = \mathbb{E}_g\left(h(X)\frac{f(X)}{g(X)}\right) = \mathbb{E}_g\left(h(X)w(X)\right)$$



- A esperança de $h(X)$ com $X \sim f(x)$ é igual à esperança de $h^*(X) = h(X)w(X)$ onde $X \sim g(x)$.
- Como isto pode ser útil?
- Como sabemos amostrar de $X \sim g(x)$, a última esperança $\mathbb{E}_g\left(h(X)w(X)\right)$ pode ser estimada facilmente.

Exemplo

- Desejamos $\mathbb{E}_f(X)$ onde $X \sim \text{Gama}(3,3)$. Neste caso, $h(X) = X$.
- Geramos 200 valores de uma $\exp(1)$
- Para cada um dos 200 valores $x_1, x_2, \ldots x_{200}$ calculamos os pesos

$$w(x_i) = \frac{f(x_i)}{g(x_i)} = \frac{\frac{27}{2}x_i^2 e^{-3x_i}}{e^{-x_i}} = \frac{27}{2}x_i^2 e^{-2x_i}$$

- Com estes pesos, estimamos

$$\mathbb{E}_f(X) = \mathbb{E}_g(h(X)\ w(X)) = \mathbb{E}_g(X\ w(X)) \approx \frac{1}{200}\sum_{i=1}^{200} x_i w(x_i)$$

- Com o script a seguir, obtive uma estimativa igual a 1.102366 quando o valor exato é igual a 1.

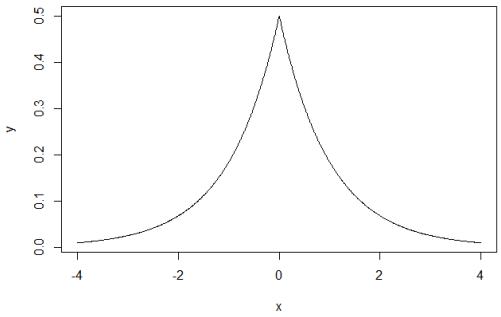
Script R

```
set.seed(123)
nsim = 200
x = rexp(nsim, 1)
wx = dgamma(x, 3, 3)/dexp(x, 1)
theta1 = mean(x*wx)
par(mfrow=c(1,1))
xx = seq(0, 9.1, by=0.1)
fx = dgamma(xx, 3, 3)
gx = dexp(xx, 1)
plot(xx, gx, type="l", ylim=c(-0.2, 1), ylab="densidade")
lines(xx, fx, lty=2)
abline(h=0)
segments(x, -0.2, x, -0.18+wx/20, lwd=2)
legend("topright",lty=1:2,c("g(x)", "f(x)") )
```

Saída do script R

Exemplo

- Queremos gerar de uma $N(0,1)$ sem usar Box-Muller.
- Precisamos simular de uma distribuição com suporte na reta real.
- Sabemos gerar com facilidade uma v.a. Y com distribuição $\exp(1)$: $Y = -\log(U)$ onde $U \sim U(0,1)$.



- Problema: exp(1) possui suporte $(0,\infty)$ e normal possui suporte na reta inteira.
- Truque: selecionamos Y exp(1). A seguir, jogue uma moeda com probabilidade 1/2: se cara, tome Y; se coroa, tome $-Y$.
- Esta distribuição é chamada de exponencial dupla ou distribuição de Laplace.
- Ver `http://en.wikipedia.org/wiki/Laplace_distribution`

Laplace ou exponencial dupla

- Densidade de Laplace padrão ($\mu = 0$ e $b = 1$):

$$g(x) = \frac{1}{2}e^{-|x|}$$

Exemplo

- Queremos calcular $\mathbb{E}_f(h(X))$ onde $X \sim f(x) = N(0,1)$
- Sabemos gerar de Laplace padrão.
- Queremos estimar:
  - $0 = \theta_1 = \mathbb{E}_f(X)$ onde $h(X) = X$
  - $1 = \theta_2 = \mathbb{V}_f(X) = \mathbb{E}_f(X^2) - (\mathbb{E}_f(X))^2$: Se $h(X) = X^2$ então $\theta_2 = \mathbb{E}_f(h(X)) - \theta_1^2$, se tivermos uma estimativa de θ_1
  - $0.02275 = \theta_3 = \mathbb{P}_f(X > 2) = \mathbb{E}_f(h(X))$ onde $h(X) = I_{[X>2]}$, a função indicadora do evento $X > 2$.
  - $\theta_4 = \mathbb{P}(e^{-|X|} > |X|) = E(h(X))$ onde $h(X) = I_A(X)$ onde $A = \{x \text{ tais que } e^{-|x|} > |x|\}$

Exemplo

- Simule $X_1, X_2, \ldots, X_B$ de uma Laplace.
- Calcule os pesos

$$w(x_i) = \frac{f(x_i)}{g(x_i)} = \frac{\frac{1}{\sqrt{2\pi}}e^{-x_i^2/2}}{\frac{1}{2}e^{-|x_i|}}$$

- Em seguida, estime

$$\theta_1 \approx \widehat{\theta_1} = \frac{1}{B}\sum_i x_i w_i$$

$$\theta_2 \approx \widehat{\theta_2} = \frac{1}{B}\sum_i x_i^2 w_i - \left(\widehat{\theta_1}\right)^2$$

$$\theta_3 \approx \widehat{\theta_3} = \frac{1}{B}\sum_i I_{[x_i>2]} w_i = \text{média dos } w_i \text{ em que } x_i > 2$$

$$\theta_4 \approx \widehat{\theta_4} = \frac{1}{B}\sum_i I_{[e^{-|x_i|}>x_i]} w_i = \text{média dos } w_i \text{ em que } e^{-|x_i|} > x_i$$

Script R

```
B = 10000
x <- (2*(runif(B) > 0.5)-1) * rexp(B) # gerando de uma exponencial dupla
hist(x)
## estimativa via importance sampling
peso <- dnorm(x)/(exp(-abs(x))/2)
a1 <- mean(x * peso)
a2 <- mean( (x^2 *peso) ) - a1^2
a3 <- mean( (x > 2) * peso )
a4 <- mean( (exp(-abs(x)) > abs(x)) * peso )
c(a1, a2, a3, a4) # [1] -0.02186748  1.00444225  0.02259828  0.42595014
## Refazendo com B maior
B = 50000
x <- (2*(runif(B) > 0.5)-1) * rexp(B) # gerando de uma exponencial dupla
peso <- dnorm(x)/(exp(-abs(x))/2)
a1 <- mean(x * peso); a2 <- mean( (x^2 *peso) ) - a1^2;
a3 <- mean( (x > 2) * peso ); a4 <- mean( (exp(-abs(x)) > abs(x)) * peso )
c(a1, a2, a3, a4) # [1] 0.0004117619 0.9879179922 0.0222652043 0.4345303436
```

11.16 Self-Normalized Importance Sampling

What justifies SIR? Let $f(x) = c_f f^*(x)$ and $g(x) = c_g g^*(x)$ and $w^*(x) = f^*(x)/g^*(x)$. Also, let $X_1, \ldots, X_n$ be i.i.d. generated from $g(x)$. By the Law of Large Numbers, we have

$$\begin{aligned}
\frac{1}{n}\sum_i w^*(X_i) &= \frac{1}{n}\sum_i \frac{f^*(X_i)}{g^*(X_i)} \\
&\to \mathbb{E}_g\left(\frac{f^*(X)}{g^*(X)}\right) \\
&= \int \frac{f^*(x)}{g^*(x)}\, g(x)\, dx = \int \frac{f^*(x)}{g^*(x)}\, c_g g^*(x)\, dx \\
&= \frac{c_g}{c_f} \underbrace{\int c_f f^*(x)\, dx}_{=1, \text{ because it is a probability density}} \\
&= \frac{c_g}{c_f}
\end{aligned}$$

Now, define

$$w(x) = \frac{f(x)}{g(x)} = \frac{c_f f^*(x)}{c_g g^*(x)} = \frac{c_f}{c_g} w^*(x) \approx \frac{w^*(x)}{\left(\frac{1}{n}\sum_i w^*(X_i)\right)}$$

Reamostragem da amostragem por importância

- Sampling importance resampling (SIR)
- Para usar amostragem por importância, precisamos conhecer as densiaddes $f(x)$ e $g(x)$, incluindo as suas constantes de integração c_1 e c_2:

$$\begin{aligned} f(x) &= c_1 f_0(x) \\ g(x) &= c_2 g_0(x) \end{aligned}$$

- O algoritmo SIR dispensa o conhecimento de c_1 e c_2
- Isto nã é muito relevante nos casos de v.a. unidimensionais mas se quisermos gerar VETORES de v.a.'s correlacionadas, este problema aparece como uma grande dificuldade.

Outra opção

- Outra opcao é simplesmente tomar a soma ponderada dos valores x com pesos iguais aos pesos NORMALIZADOS
- Isto é diferente de importance sampling, pois não é a média de x vezes o peso não-normalizado.

```
c(sum(x*peso), sum((x-a1)^2 * peso), sum((x > 2)*peso),
    sum((exp(-abs(x))>abs(x))*peso))
```

Algoritmo SIR

- Simule uma amostra $X_1, X_2, \ldots, X_B$ de $g(x)$
- Calcule os pesos $w_i = f_0(x_i)/g_0(x_i)$
- Normalize os pesos $w_i \leftarrow w_i/S$ onde $S = \sum_k w_k$
- REAMOSTRE os B dados da amostra original com reposição e com pesos w_i gerando $X_1^*, X_2^*, \ldots, X_n^*$
- Cada X_j^* assume um dos valores $X_1, X_2, \ldots, X_B$ com probab $w_1, w_2, \ldots, w_B$.

SIR

- Mostra-se que a distribuição de X_j^* tem densidade aproximadamente igual a $f(x)$.
- SIR começa gerando de $g(x)$ como importance sampling.
- Ao invés de reter todos os valores gerados atribuindo um peso...
- SIR REAMOSTRA os valores gerados com um peso.
- Voltando ao exemplo anterior, queremos estimar $\theta_1 = \mathbb{E}_f(X)$, $\theta_2 = \mathbb{V}_f(X)$, $\theta_3 = \mathbb{P}_f(X > 2)$, $\theta_4 = \mathbb{P}_f(e^{-|X|} > |X|)$. onde $X \sim f(x) = N(0,1)$.
- Temos $f(x) = (2\pi)^{-1/2} e^{-x^2/2} \propto e^{-x^2/2}$
- Vamos SUPOR que não conhecemos a constante $(2\pi)^{-1/2}$
- Amostramos $X_1, \ldots X_B$ de $g(x)$, uma Laplace padrão.

SIR

- Reamostramos m elementos
- Reamostra $X_1^*, X_2^*, \ldots, X_m^*$ i.i.d com

$$X_j^* = \begin{cases} X_1 & \text{com probab } w_1 \\ \vdots & \\ X_B & \text{com probab } w_B \end{cases}$$

- No final, calculamos uma média aritmética simples de $h(X_j^*)$:

$$\widehat{\theta} = \frac{1}{m} \sum_{j=1}^{m} h(X_j^*)$$

- Ver script R

Script R

```
  B = 20000
## amostra de exponencial dupla (ou Laplace)
x <- (2*(runif(B) > 0.5)-1) * rexp(B)
## estimativa via SIR - sampling importance resampling
peso <- exp(-x^2/2)/(exp(-abs(x))/2)
peso <- peso/sum(peso)
xstar <- sample(x, 10000, replace=T, prob=peso)
a1 <- mean(xstar)
a2 <- mean(xstar^2) - a1^2
a3 <- mean(xstar > 2)
a4 <- mean(exp(-abs(xstar)) > abs(xstar))
c(a1, a2, a3, a4)
```

Escolha de $g(x)$

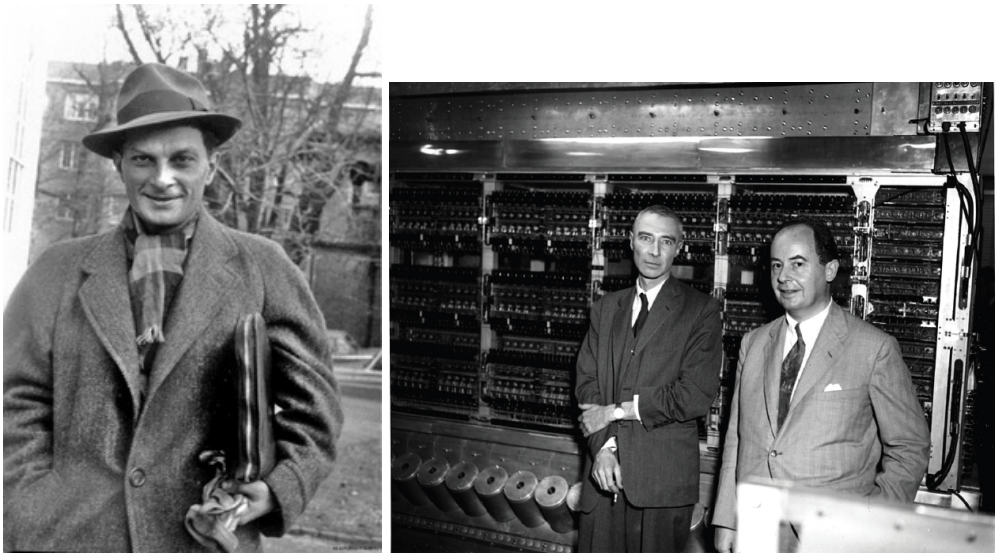


Figure 11.15: Esquerda: Stanislaw Ulam na Polônia. Direita: John von Neumann (à direita) e Robert Oppenheimer em frente a um pequeno pedaço do ENIAC, um dos primeiros computadores do mundo.

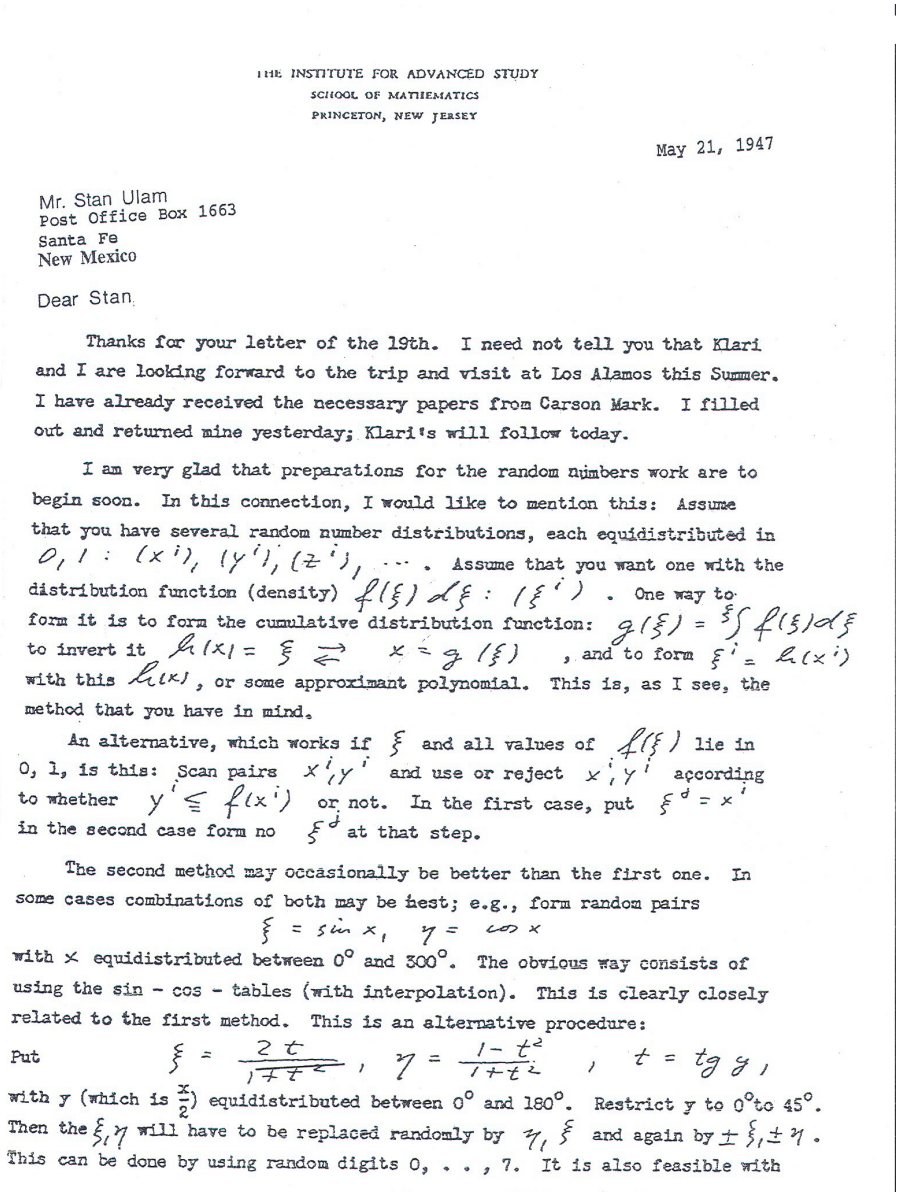
- Nos métodos de aceitação-rejeição, importance sampling e SIR geramos de $g(x)$ mas o objetivo é estimar quantidades associadas com $f(x)$.
- Como deve ser escolhida $g(x)$?
- Ela deve ter um suporte maior ou igual a $f(x)$.
- Além disso, ela deve ser o mais parecida possível com $f(x)$.
- Uma má escolha para $g(x)$ põe muita massa de probabilidade numa região (de onde amostramos frequentemente) e esta região tem baixa probabilidade sob $f(x)$.
- Pior: região onde $f(x)$ põe massa de probabilidade tem pouca chance de ser selecionada sob $g(x)$

11.17 História do método Monte Carlo

Os dois métodos principais que vimos neste capítulo, o da transformada inversa e o da aceitação-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 11.15 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 aceitação-rejeição. Esta carta sobreviveu e está na Figura 11.16 e 11.17. Note que, nas última linhas da carta, von Neumann menciona que seria útil manter

seu método em mente, especialmente depois que o ENIAC estivesse diponível. Nós mantemos este método na nossa mente até hoje, muito tempo depois que o ENIAC desapareceu.



THE INSTITUTE FOR ADVANCED STUDY
SCHOOL OF MATHEMATICS
PRINCETON, NEW JERSEY

May 21, 1947

Mr. Stan Ulam
Post Office Box 1663
Santa Fe
New Mexico

Dear Stan,

Thanks for your letter of the 19th. I need not tell you that Klari and I are looking forward to the trip and visit at Los Alamos this Summer. I have already received the necessary papers from Carson Mark. I filled out and returned mine yesterday; Klari's will follow today.

I am very glad that preparations for the random numbers work are to begin soon. In this connection, I would like to mention this: Assume that you have several random number distributions, each equidistributed in $0, 1 : (x^i), (y^i), (z^i), \cdots$. Assume that you want one with the distribution function (density) $f(\xi)\,d\xi : (\xi^i)$. One way to form it is to form the cumulative distribution function: $g(\xi) = \int^{\xi} f(\xi)\,d\xi$ to invert it $h(x) = \xi \rightleftarrows x = g(\xi)$, and to form $\xi^i = h(x^i)$ with this $h(x)$, or some approximant polynomial. This is, as I see, the method that you have in mind.

An alternative, which works if ξ and all values of $f(\xi)$ lie in 0, 1, is this: Scan pairs x^i, y^i and use or reject x^i, y^i according to whether $y^i \leqq f(x^i)$ or not. In the first case, put $\xi^j = x^i$ in the second case form no ξ^j at that step.

The second method may occasionally be better than the first one. In some cases combinations of both may be best; e.g., form random pairs

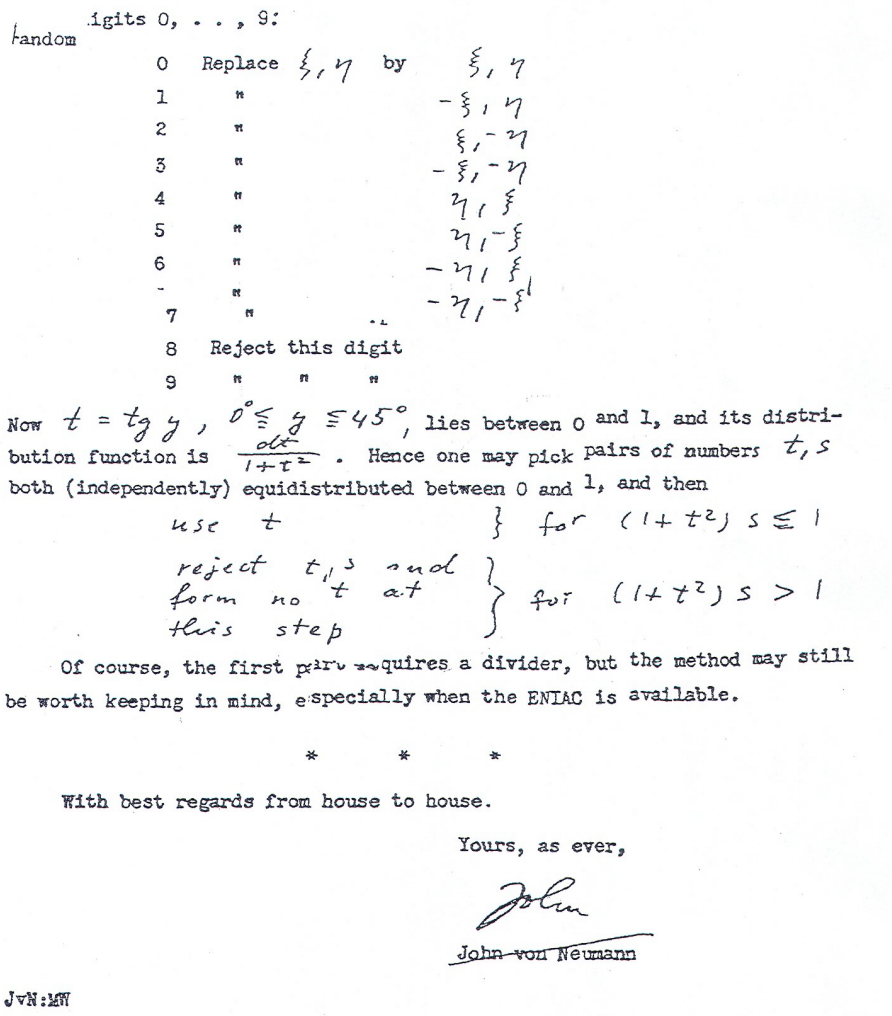
$$\xi = \sin x, \quad \eta = \cos x$$

with x equidistributed between 0° and 300°. The obvious way consists of using the sin - cos - tables (with interpolation). This is clearly closely related to the first method. This is an alternative procedure:

Put $\xi = \frac{2t}{1+t^2}$, $\eta = \frac{1-t^2}{1+t^2}$, $t = \operatorname{tg} \gamma$,

with γ (which is $\frac{x}{2}$) equidistributed between 0° and 180°. Restrict γ to 0° to 45°. Then the ξ, η will have to be replaced randomly by η, ξ and again by $\pm\xi, \pm\eta$. This can be done by using random digits 0, . . , 7. It is also feasible with

Figure 11.16: Primeira página da carta de von Neumann a Stan Ulam.



Random digits 0, . . . , 9:

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 $t = \operatorname{tg} y$, $0° \leqq y \leqq 45°$, 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 $(1+t^2)\, s \leqq 1$

reject t, s and form no t at this step } for $(1+t^2)\, s > 1$

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

John von Neumann

JvN:MW

Figure 11.17: Segunda página da carta de von Neumann a Stan Ulam.

11.18 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 $t_1, t_2, \ldots, t_n$ i.i.d. com distribuição $\exp(\lambda)$. A seguir tome a média aritmética dos valores $w_1, w_2, \ldots, w_n$ 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 importância. Não veremos estes métodos neste livro. □

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

$$\theta = P(65 < X \leq 85) = E\left(I_{(55,85]}(T)\right)$$

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. □

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 $X_1, X_2, \ldots, X_n$ seja uma amostra aleatória de variáveis aleatórias i.i.d. com distribuição X. Como $X_1^2, X_2^2, \ldots, X_n^2$ é uma amostra de v.a.'s i.i.d., pela Lei dos Grandes Números,

$$\frac{1}{n}\left(X_1^2 + X_2^2 + \ldots + X_n^2\right) \to m_2 = E(X^2)$$

quando $n \to \infty$.

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 $\widehat{m_2} = (X_1^2 + X_2^2 + \ldots + X_n^2)/n$ onde $X_1, X_2, \ldots, X_n$ é uma grande amostra da variável X.

A variância de X é $\sigma^2 = \mathrm{Var}(X) = E(X^2) - (E(X))^2 = m_2 - m_1^2$ onde $m_1 = E(X)$. Assim, uma estimativa para σ^2 é

$$\widehat{\sigma^2} = (X_1^2 + X_2^2 + \ldots + X_n^2)/n - ((X_1 + X_2 + \ldots + X_n)/n)^2$$

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:

- $Y = 10\,v^T I_{[5,25]}(T)$, 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 lineramente com o tempo).
- $$Y = \begin{cases} 30v^T, & \text{se } T < 10 \\ 30(1-(T-10)/30)v^T, & \text{se } 10 \le T < 20 \\ 20v^T, & \text{se } T \ge 20 \end{cases}$$

11.19 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. $Z_1, \ldots, Z_{100}$ 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 inetgral 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 $t_1, \ldots, t_{10000}$ e calcule a média

$$\hat{\mu} = \frac{1}{10000} \sum_{k=1}^{10000} 10 \exp(-0.06 t_k)$$

Numa simulação em meu computador obtive $\hat{\mu} = 1.638743$. Com estes mesmos 10 mil números, calcula-se uma estimativa do desvio padrão:

$$\hat{\sigma} = \frac{1}{10000} \sum_{k=1}^{10000} (10 \exp(-0.06 t_k)) - (\hat{\mu})^2 = 2.509787$$

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 dependeu 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 $t_1, t_2, \ldots$ em que benefícios foram pagos antes de t e ele pode ser calculado da seguinte maneira:

$$C(t) = \begin{cases} 175\, e^{0.06t}, & \text{se } t < t_1 \\ (175\, e^{0.06t_1} - 10) e^{0.06(t-t_1)}, & \text{se } t_1 \leq t < t_2 \\ \left((175\, e^{0.06t_1} - 10) e^{0.06(t_2-t_1)} - 10\right) e^{0.06(t-t_2)}, & \text{se } t_2 \leq t < t_3 \\ \ldots & \ldots \end{cases}$$

Fazendo as multiplicações em cada segmento de tempo e simplificando, obtemos

$$C(t) = \begin{cases} 175\,e^{0.06t}, & \text{se } t < t_1 \\ (175 - 10\,e^{0.06t_1})e^{0.06t}, & \text{se } t_1 \le t < t_2 \\ (175 - 10e^{0.06t_1} - 10e^{0.06t_2})e^{0.06t}, & \text{se } t_2 \le t < t_3 \\ \ldots & \ldots \end{cases}$$

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 11.18:

```
## 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 11.18 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

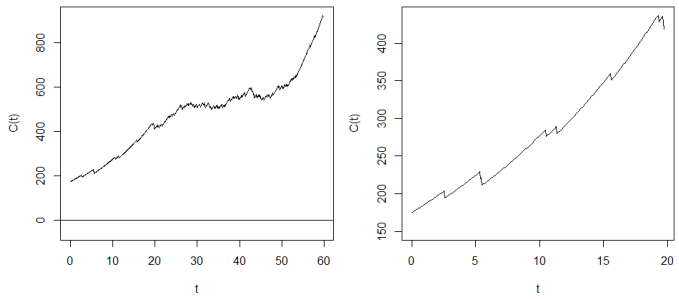


Figure 11.18: Esquerda: Gráfico de C(t) versus t. Direita: Gráfico de C(t) versus t até o pagamentos dos 10 primeiros benefícios. Ver texto para detalhes.

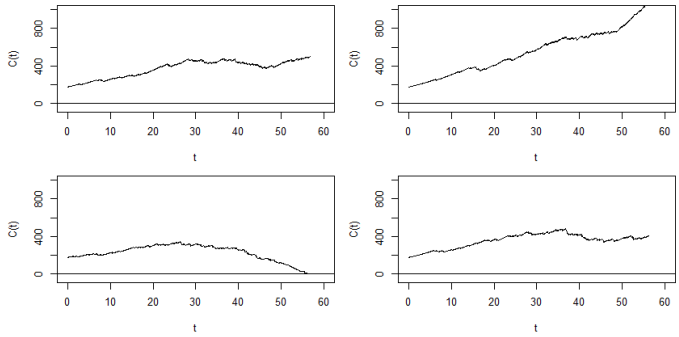


Figure 11.19: Gráficos com 4 desenvolvimentos independentes do fundo, todos gerados nas mesmas condições . O gráfico da segunda linha à esquerda mostra uma situação em que o fundo fica insolvente por volta de $t = 57$.

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 11.19 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 possuia 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 psoitivo.

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)){
```

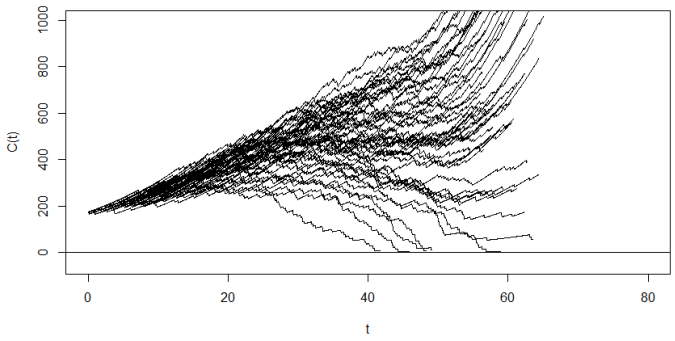


Figure 11.20: Gráfico com 50 desenvolvimentos independentes do fundo $C(t)$, todos gerados nas mesmas condições . Aqueles que ficaram insolventes tiveram suas linhas interrompidas imediatamente antes de tornarem-se negativos.

```
    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])
}
```

11.20 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-iésimo sinistros. Por convenção, $T_1=X_1$ é o tempo entre a origem $t=0$ e o primeiro evento. Então $X_1,X_2,X_3,\ldots$ são i.i.d. $\exp(\lambda)$.

Sabemos como gerar os X_i's pelo método da tranformação inversa: $X_i=-1/lambda\log(U_i)$ onde $U_1,U_2,\ldots$ 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 11.21.

```
lambda <- 2; tf <- 12; t <- 0; i <- 0
s <- numeric(); n <- numeric()
```

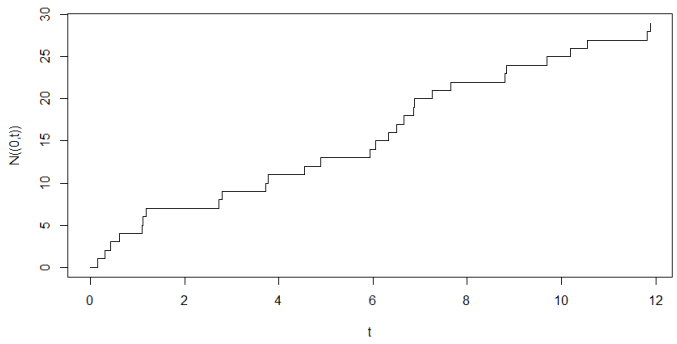


Figure 11.21: Gráfico com uma realização de um processo de Poisson de taxa $\lambda = 2$ em $(0,12)$. Para cada tempo t, o eixo vertical mostra o número total $N([0,t))$ de eventos que ocorreram até o instante t.

```
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 11.21 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 ? □.

11.20.1 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* $T_1, T_2, \ldots, T_n$ 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))")
```

11.20.2 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 $E(N([3.3,3.4))) \approx 5 0.1 = 0.5$, 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 $E(N([3.3,3.4)))/0.1 \approx \lambda(3.3) = 5$.

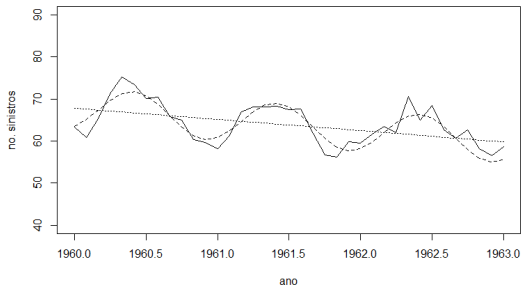


Figure 11.22: Gráfico com dados mensais de número de acidentes com motocicletas ao longo do tempo. Ver texto para detalhes.

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 11.22 é adaptada de [5] 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

$$\lambda(t) = 68 - 0.22t + 5\cos(\pi(1+t/6)) \tag{11.4}$$

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 $0 < t_1 < t_2 < \ldots < t_n < t_F$. 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 demostração em livros de processos estocásticos).

Assim, a geração pode ser feita por meio dos seguintes comandos em R, com o resultado mostrado na Figura 11.23:

```
t <- 0; i <- 0; tf <- 3; s <- 0; k <- 72
lambdat <- function(t){
   68 - 0.22*t +  5*cos(pi * (t + 1))
}
```

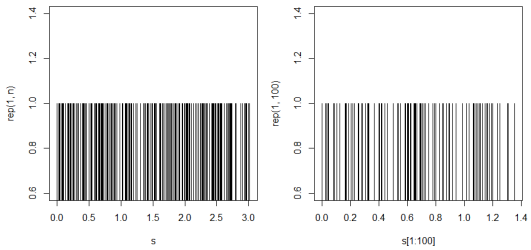


Figure 11.23: Processo pontual de Poisson com posterior afinamento.

```
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.

11.21 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 $Y_1, Y_2, \ldots$ os valores gerados a partir de $g(x)$ e por X o primeiro valor aceito a partir da sequência de Y's. O algoritmo de aceitação-rejeição é o seguinte:

Algorithm 2 Método da Rejeição.

1: $I \leftarrow$ `True`
2: **while** I **do**
3: Gere $Y \sim g(y)$
4: Gere $U \sim \mathscr{U}(0,1)$
5: **if** $U \leq r(Y) = f(Y)/Mg(Y)$ **then**
6: $X \leftarrow Y$
7: $I =$ `False`
8: **end if**
9: **end while**

Assuma que a distribuição acumulada associada com $f(x)$ e $g(x)$ seja igual a $\mathbb{F}(x)$ e $\mathbb{G}(x)$, respectivamente. Isto é, $f(x) = \mathbb{F}'(x)$ e $g(x) = \mathbb{G}'(x)$.

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 11.21.1 — 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

$$\begin{aligned} \mathbb{P}(Y \leq t \mid Y \text{ gerado é aceito }) &= \mathbb{P}\left(Y \leq t \;\middle|\; U \leq \frac{f(Y)}{Mg(Y)}\right) \\ &= \frac{\mathbb{P}\left(Y \leq t \quad \text{and} \quad U \leq \frac{f(Y)}{Mg(Y)}\right)}{\mathbb{P}\left(U \leq \frac{f(Y)}{Mg(Y)}\right)} \end{aligned} \tag{11.5}$$

Denote por B o evento $B = [\; Y \leq t \quad \text{and} \quad U \leq \frac{f(Y)}{Mg(Y)} \;]$ do numerador. Aplicando a regra da probabilidade total a este evento, temos

$$\mathbb{P}\left(Y \leq t \quad \text{and} \quad U \leq \frac{f(Y)}{Mg(Y)}\right) = \int_{-\infty}^{\infty} \mathbb{P}\left(Y \leq t \quad \text{and} \quad U \leq \frac{f(Y)}{Mg(Y)} \;\middle|\; Y = z\right) g(z)\, dz$$

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

$$\mathbb{P}\left(Y \leq t \quad \text{and} \quad U \leq \frac{f(Y)}{Mg(Y)} \;\middle|\; Y = z\right) = \begin{cases} 0, & \text{se } t < z \\ \mathbb{P}\left(U \leq \frac{f(z)}{Mg(z)}\right) = \frac{f(z)}{Mg(z)}, & \text{se } t \geq z \end{cases}$$

Assim, o numerador de 11.5 é igual a

$$\begin{aligned} \int_{-\infty}^{\infty} \mathbb{P}\left(Y \leq t \quad \text{and} \quad U \leq \frac{f(Y)}{Mg(Y)} \;\middle|\; Y = z\right) g(z)\, dz &= \int_{-\infty}^{t} \frac{f(z)}{Mg(z)} g(z)\, dz \\ &= M^{-1} \int_{-\infty}^{t} f(z)\, dz \\ &= M^{-1} \mathbb{F}(t) \end{aligned}$$

O denominador de 11.5 é calculado de modo semelhante:

$$\begin{aligned} \mathbb{P}\left(U \leq \frac{f(Y)}{Mg(Y)}\right) &= \int_{-\infty}^{\infty} \mathbb{P}\left(U \leq \frac{f(Y)}{Mg(Y)} \;\middle|\; Y = z\right) g(z)\, dz \\ &= \int_{-\infty}^{\infty} \frac{f(z)}{Mg(z)} g(z)\, dz \\ &= \frac{1}{M} \int_{-\infty}^{\infty} f(z)\, dz \\ &= \frac{1}{M} \quad \text{pois } f(x) \text{ é uma densidade e integra 1} \end{aligned}$$

Portanto, retornando a 11.5, podemos escrever

$$\mathbb{P}(Y \leq t \mid Y \text{ gerado é aceito }) = \frac{\mathbb{F}(t)/M}{1/M} = \mathbb{F}(t)$$

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 11.21.2 — 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 $i = 1, 2, \ldots$. 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 $I_1, I_2, \ldots$ i.i.d. com probabilidade de sucesso constante e igual a $1/M$. Por definição, esta é a distribuição geométrica com com valor esperado M. ■

Em conclusão, ao adotarmos um valor M para este parâmetro, vamos selecionar, em média, M valores de g para cada valor eventualmente aceito.

12. Random Vectors

12.1 Introdução

Neste capítulo vamos lidar um vetor de v.a.'s, e não com uma única v.a. Teremos $\mathbf{X} = (X_1, \ldots, X_k)$, 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 $\mathscr{A}$ de Ω. Ω é "complexo": cada resultado do experimento aleatório pode ter muitas características de interesse: $X_1, X_2, X_3, \ldots$ Coletamos estas várias medidas num vetor aleatório $\mathbf{X} = (X_1, \ldots, X_n)$. As variáveis $X_1, X_2, X_3, \ldots$ são medições feitas no *mesmo resultado* $\omega \in \Omega$ do experimento. Ver Figura 12.1.

■ **Example 12.1 — Imagem como vetor.** Seja $\Omega = \{$ 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, $\mathbf{X} = (X_{11}, X_{12}, \ldots, X_{nm})$ 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:

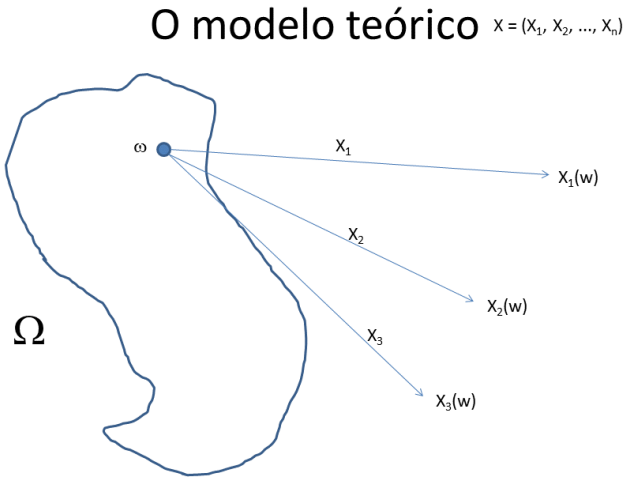


Figure 12.1: O arcabouço teórico para vetores aleatórios.

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;
- Crânios 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 crânios, 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 detreminando 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 aleatórioé $\mathbf{X} = (Y, X_1, X_2, X_3, X_4, X_5)$. 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.

$$(Y | X_1 = \text{ Sion }, X_2 = \text{ 10 anos }, X_3 = 200m^2, X_4 = 4, X_5 = \text{ não})$$

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 $X_1 = \text{ Sion }, X_2 = \text{ 10 anos}, X_3 = 200m^2, X_4 = 4, X_5 = \text{ não}$? 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 interelações entre as variáveis componentes do vetor. Se $\mathbf{X} = (X_1, X_2, \ldots, X_k)$, 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 $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ 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 $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ terá $m_1 \times m_2 \times \ldots m_k$ 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 otorrinolaringologista 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 padão-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 válidade 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 $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ 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 $\mathbf{X} = (X_1, X_2, \ldots, X_k)$? 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.

Definition 12.3.2 — Distribuição Marginal - 2. Se $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ é composto apenas de v.a.'s discretas, a *distribuição marginal* de uma v.a. X_i é dada por

$$\mathbb{P}(X_i = x) = \sum_S \mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_i = x, \ldots, X_k = x_k)$$

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,

Definition 12.4.1 — Independência no caso discreto. As v.a.'s discretas X e Y são independentes se

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\ \mathbb{P}(Y = y) \tag{12.1}$$

para todo par (x, y).

Theorem 12.4.1 — Definição equivalente de independência. Se X e Y são independentes se, e somente se,

$$\mathbb{P}(X = x | Y = y) = \mathbb{P}(X = x) \tag{12.2}$$

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 concordãncia al'em 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:

$$\begin{aligned}\mathbb{P}(X=0) &= \mathbb{P}(X=0\wedge Y=0)+\mathbb{P}(X=0\wedge Y=1)=0.075+0.225=0.300\\ \mathbb{P}(Y=0) &= \mathbb{P}(Y=0\wedge X=0)+\mathbb{P}(Y=0\wedge X=1)=0.075+0.175=0.250\end{aligned}$$

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

$$\begin{aligned}\mathbb{P}(X_1=0) &= \sum_{i,j,k}\mathbb{P}(X_1=0,X_2=i,X_3=j,X_4=k)\\ &= (4.6+5.7+4.1+5.3+5.2+6.6+\\ &\quad +1.6+1.8+4.9+0.2+3.1+4.4)/100\\ &\quad = 0.475\end{aligned}$$

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

$$\begin{aligned}\mathbb{P}(X_3=1) &= \sum_{i,j,k}\mathbb{P}(X_1=i,X_2=j,X_3=1,X_4=k)\\ &= (4.6+6.7+1.6+1.8\\ &\quad +1.1+6.8+3.6+3.7)/100\\ &\quad = 0.299\end{aligned}$$

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 $X_3 = 2$:

$$\begin{aligned}\mathbb{P}(X_3 = 2) &= \sum_{i,j,k} \mathbb{P}(X_1 = i, X_2 = j, X_3 = 2, X_4 = k) \\ &= (4.1 + 5.3 = 4.9 + 0.2 + \\ &\quad + 0.5 + 4.0 + 6.6 = 6.8)/100 \\ &\quad = 0.324\end{aligned}$$

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 $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ 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 $n_2 \times n_3 \times \ldots \times n_k$ 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:

$$\mathbb{P}(X_1 = x_{n_1}) = 1 - \mathbb{P}(X_1 = x_1) - \ldots - \mathbb{P}(X_1 = x_{n_1 - 1})$$

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

$$\begin{aligned}\mathbb{P}(X_1=0,X_2=1) &= \sum_{i,j}\mathbb{P}(X_1=0,X_2=1,X_3=i,X_4=j)\\ &= (1.6+1.8+4.9+0.2+3.1+4.4)/100\\ &= 0.16\end{aligned}$$

Definition 12.5.1 — Independência de v.a's discretas. Seja $\mathbf{X}=(X_1,\ldots,X_k)$ um vetor aleatório composto de v.a.'s discretas. Elas são *independentes* se

$$\mathbb{P}(X_1=x_1,\ldots,X_k=x_k)=\mathbb{P}(X_1=x_1)\ldots\mathbb{P}(X_k=x_k)$$

para qualquer configuração de valores possíveis $(x_1,\ldots,x_k)$.

Theorem 12.5.1 — Independência: outra maneira. Se o vetor $\mathbf{X}$ é composto de v.a.'s independentes então

$$\mathbb{P}(X_1=x_1|X_2=x_2,\ldots,X_k=x_k)=\mathbb{P}(X_1=x_1)$$

para qualquer configuração de valores possíveis $(x_1,\ldots,x_k)$.

Este resultado é válido se X_1 trocar de posição com qualquer outra v.a.

12.6 Simulação de 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	0.19	$\mathbb{P}(TO=0) = 0.40+0.19=0.51$
$T0=1$	0.03	0.38	$\mathbb{P}(TO=1) = 0.03+0.38=0.49$

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 inércia 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 inércia? Como comparar diferentes sociedades quanto ao seu grau de mobilidade social? Este é um assunto fascinante e estudado por vários autores [21].

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.

	SF: status do indivíduo em 1988.					
SP: status do pai	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:

	SF: status do indivíduo em 1988 .						
SP: status do pai	Baixo Inf.	Baixo Sup.	Médio Inf.	Médio	Médio Sup.	Alto	TOTAL
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 como obter a distribuição de uma v.a. condicionada nos valores de outras no vetor no caso discreto [1]. 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 distibuição 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	0	2	1	5.3
0	1	2	0	4.9
0	1	2	1	0.2
1	0	2	0	0.5
1	0	2	1	4.0
1	1	2	0	6.6
1	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	$100\% \times \mathbb{P}(\ldots \mid X_3 = 2)$
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 distibuição 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	$100\% \times \mathbb{P}(X_1 = i, X_4 = j \mid X_2 = 0, X_3 = 2)$
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 distibuição 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 o 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 e 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 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, dividi-los pela sua soma $\mathbb{P}(SF = 6) = \sum_i \mathbb{P}(SP = i, SF = 6)$): para que somem 1:

ALTO
1.0
1.3
1.5
0.9
0.5
0.9
$\sum = 6.1$

$\longrightarrow$

ALTO
.8 0.16
.8 0.21
.8 0.25
.8 0.15
.8 0.08
.8 0.15
$\sum = 1$

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? Ests 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

	j=1	j=2	j=3	j=4	j=5	j=6
$\mathbb{P}(SF = j\|SP = 1)$	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 populaçã 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 $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ 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 $X_2 = 0, \ldots, X_k = 2$.

$$(X_1 | X_2 = 0, \ldots, X_k = 2) \sim ??$$

O que é a distribuição de uma v.a. discreta? Duas coisas...

- Uma lista $\{a_1, \ldots, a_m\}$ dos valores possíveis de X_1 *quando* $X_2 = 0, \ldots, X_k = 2$
- Uma lista com as probabilidades associadas quando $X_2 = 0, \ldots, X_k = 2$:

$$\mathbb{P}(X_1 = a_i | X_2 = 0, \ldots, X_k = 2)$$

Ao mudar os valores condicionados de $X_2, \ldots, X_k$ esta distibuição também muda. Por exemplo, as probabilidades de $\mathbb{P}(X_1 = a_i | X_2 = 1, \ldots, X_k = 0)$ usualmente são diferentes das de $\mathbb{P}(X_1 = a_i | X_2 = 0, \ldots, X_k = 2)$. A distribuição é função dos valores em que estamos condicionando as demais variáveis $X_2, \ldots, X_k$.

Vamos nos fixar em obter $\mathbb{P}(X_1 = a_i | X_2 = 0, \ldots, X_k = 2)$. Para estes valores fixos $X_2 = 0, \ldots, X_k = 2$ das variáveis condicionantes, a distribuição é encontrada pela fórmula de probabilidade condicional:

$$\mathbb{P}(X_1 = a_i | X_2 = 0, \ldots, X_k = 2) = \frac{\mathbb{P}(X_1 = a_i, X_2 = 0, \ldots, X_k = 2)}{\mathbb{P}(X_2 = 0, \ldots, X_k = 2)}$$

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

$$\frac{\mathbb{P}(X_1 = a_i | X_2 = 0, \ldots, X_k = 2)}{\mathbb{P}(X_1 = a_j | X_2 = 0, \ldots, X_k = 2)} = \frac{\frac{\mathbb{P}(X_1 = a_i, X_2 = 0, \ldots, X_k = 2)}{\mathbb{P}(X_2 = 0, \ldots, X_k = 2)}}{\frac{\mathbb{P}(X_1 = a_j, X_2 = 0, \ldots, X_k = 2)}{\mathbb{P}(X_2 = 0, \ldots, X_k = 2)}} = \frac{\mathbb{P}(X_1 = a_i, X_2 = 0, \ldots, X_k = 2)}{\mathbb{P}(X_1 = a_j, X_2 = 0, \ldots, X_k = 2)}$$

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 [11]. 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											
Apple	0	1	2	3	4	5	6	7	8	9	10	θ
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

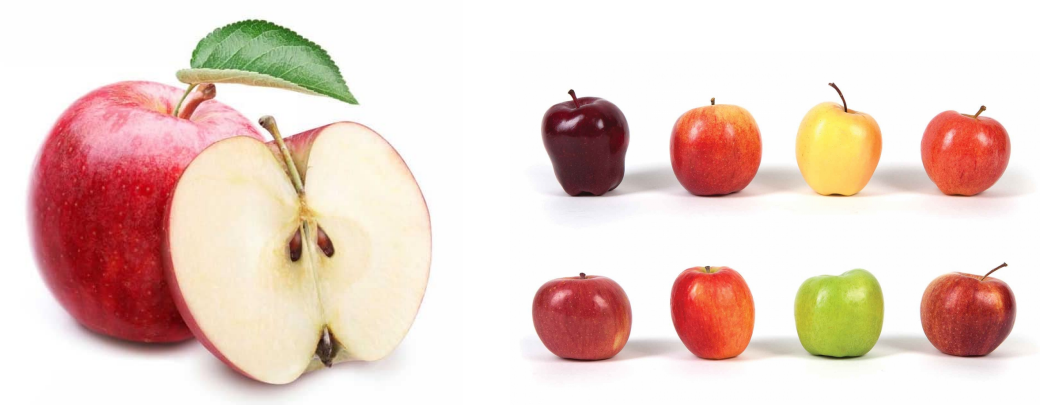


Figure 12.4: O número de sementes numa maçã depende da sua variedade.

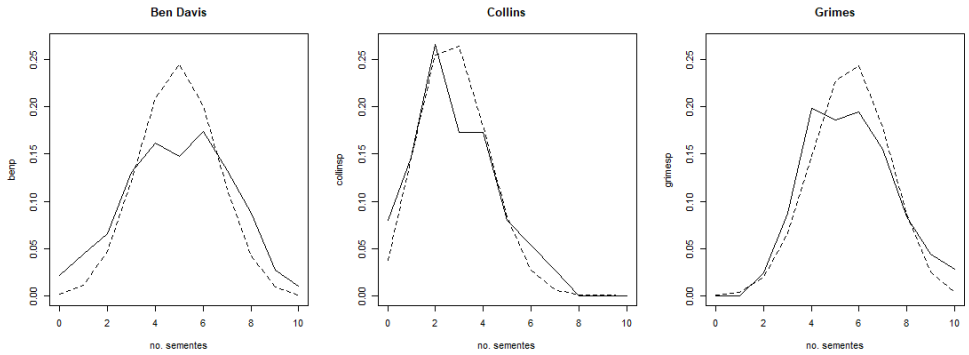


Figure 12.5: Frequência relativa (linha sólida) e teórica (linha tracejada, derivada de um modelo binomial) para o número de sementes de três variedades de maçãs.

Seja C a classe de maçã considerada, com $C \in \{Ben,\ Col,\ Gri\}$. Estamos sugerindo que

$$\begin{aligned}(X \mid C = Ben) &\sim \text{Bin}(10, 0.495)\\ (X \mid C = Col) &\sim \text{Bin}(10, 0.280)\\ (X \mid C = Gri) &\sim \text{Bin}(10, 0.562)\end{aligned}$$

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 $mathbbP(X = x)$. A linha sólida é a estimativa empírica simples, a proporção de maçãs com x semesntes em cada variedade. A linha tracejada são as probabilidade 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)
```

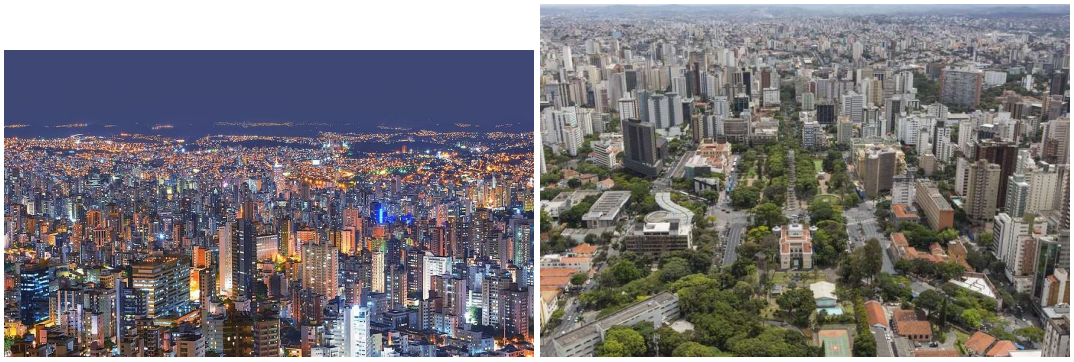


Figure 12.6: ...

```
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 refriadas nas semanas 1 e 2, respectivament.

Começamos encontrando a distribuição marginal de W_1. Esta variável conta o número de infectados (sucessos) na primeira semana. Temos dois indivíduos, ambos cm probabilidade de infecção 0.1 e independentes. Portanto, $W_1 \sim \text{Bin}(2,0.1)$ e assim encontramos $\mathbb{P}(W_1 = k) = 2/(k!(2-k)!)0.1^k 0.9^{2-k}$. 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.** [6] 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	$\text{Bin}(60, p_{76.5})$

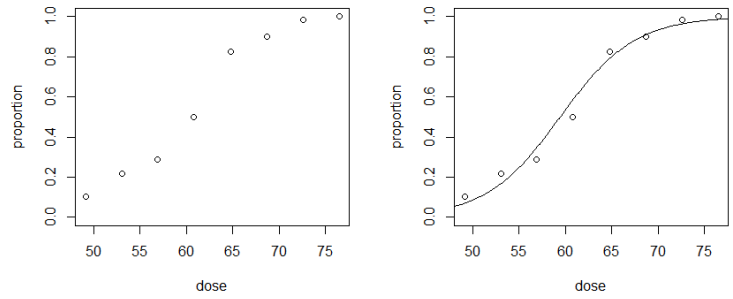


Figure 12.7: Esquerda: Proporção de besouros mortos versus dose aplicada. Direita: Curva obtida por modelo de regressão logística fornecendo uma estimativa da probabilidade de morte para qualquer dose d num intervalo contínuo.

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-se às sua 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 $(28,35]$, respetivamente. Assim, temos o vetor (Y,X) medido no mesmo caranguejo-ferradura fêmea. Nosso 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)
```

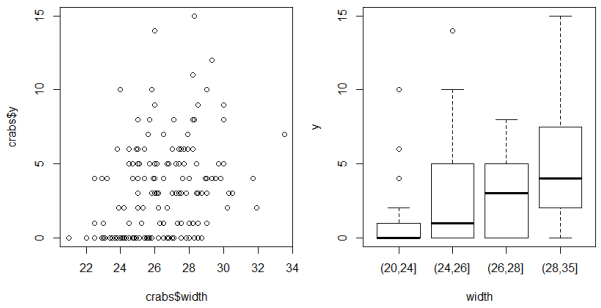


Figure 12.9: Esquerda: Gráfico do número Y de machos satélites versus a largura da carapaça da fêmea. Direita: Boxplots de Y versus a categoria X de carapaça.

```
# [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 $Y = (Y_1, Y_2, \ldots, Y_n)$. Mostramos como obter a distribuição condicional de Y_1 dados os valores de $Y_2, \ldots, Y_p$. 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 $Y_2, \ldots, Y_p$. É simplesmente a definição usual de esperança de v.a.'s discretas mas agora usando a a distribuição condicional: a soma dos valores possíveis vezes as probabilidades condicionais associadas.

Definition 12.13.1 — Esperança condicional discreta. Seja $Y = (Y_1, Y_2, \ldots, Y_n)$ um vetor aleatório de variáveis discretas. O valor esperado de Y_1 condicionado nos valores $Y_2 = a_2, \ldots, Y_p = a_p$ das demais v.a.'s é definido como

$$\mathbb{E}(Y_1 | Y_2 = a_2, \ldots, Y_p = a_p) = \sum_y y \mathbb{P}(Y_1 = y | Y_2 = a_2, \ldots, Y_p = a_p) \,.$$

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 $\mathbb{P}(Y_1 = y | Y_2 = a_2, \ldots, Y_p = a_p)$ 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 $\mathbb{E}(Y_1|Y_2=a_2,\ldots,Y_p=a_p)$ condicionada em valores das outras v.a.s $(Y_2,\ldots,Y_p)$:

$$\mathbb{V}(Y_1|Y_2=a_2,\ldots,Y_p=a_p) = \sum_y (y-m)^2 \mathbb{P}(Y_1=y|Y_2=a_2,\ldots,Y_p=a_p)$$

onde $m = \mathbb{E}(Y_1|Y_2=a_2,\ldots,Y_p=a_p)$ é a esperança condicional. Pode-se mostrar que

$$\mathbb{V}(Y_1|Y_2=a_2,\ldots,Y_p=a_p) = \mathbb{E}(Y_1^2|Y_2=a_2,\ldots,Y_p=a_p) - m^2$$

■ **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. continua: 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):

$$(y_{11},y_{12}),(y_{21},y_{22}),(x_{31},y_{32}),\ldots,(y_{n1},y_{n2})$$

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

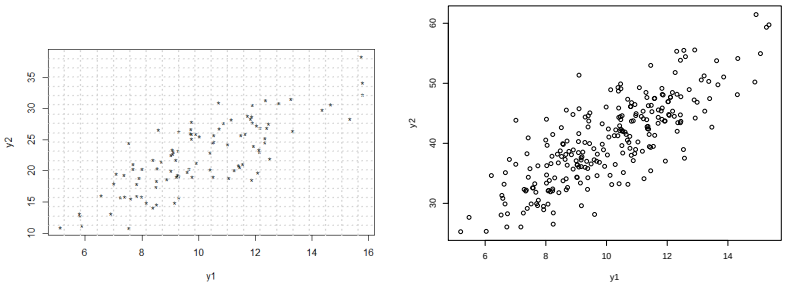


Figure 12.10: Esquerda: Amostra de 100 instâncias do vetor aleatório (Y_1, Y_2) e grade regular sobreposta. Direita: Histograma tri-dimensional baseado em amostra de vetor (X, Y).

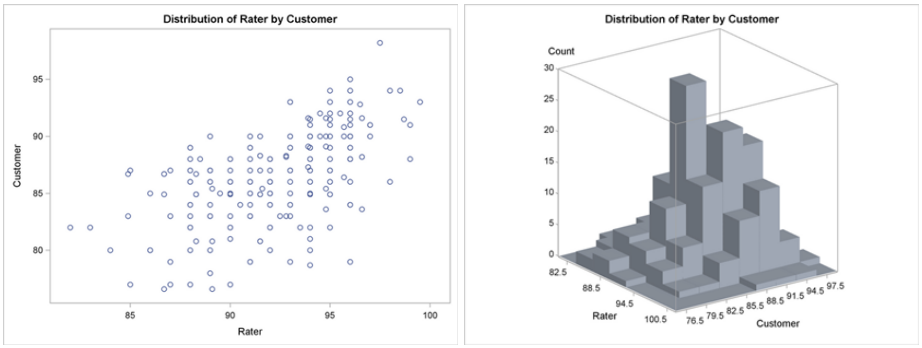


Figure 12.11: Amostra de n pontos do vetor aleatório (X, Y) e histograma tri-dimensional baseado nesta amostra.

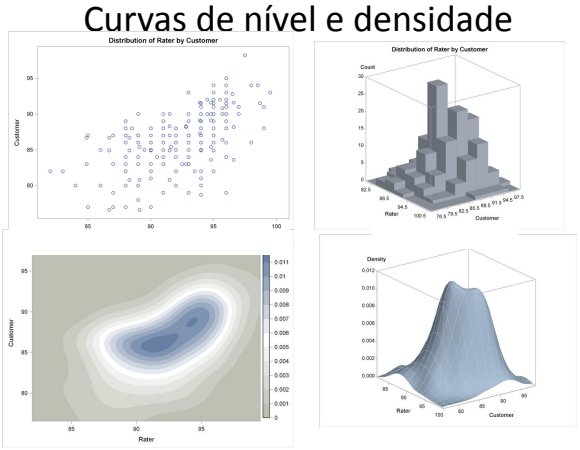


Figure 12.12: Amostra de 250 dados de (Y_1, Y_2) com histograma 3d, densidade $f(y_1, y_2)$ e suas curvas de nível

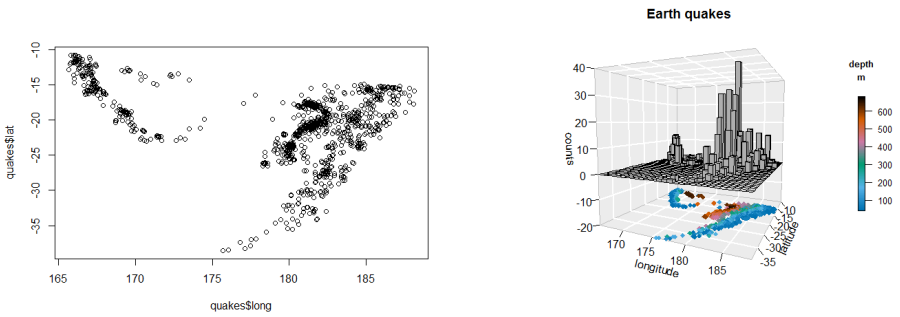


Figure 12.13: Esquerda: longitude e latitude do epicentro de 1000 terremotos. Direita: Histograma construído com `hist3D(x = xmid, y = ymid, z = xy)`.

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 duraccão 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 ti-diemnsionais 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:

- $f(y_1, y_2) \geq 0$
-

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(y_1, y_2)dy_1dy_2 = 1$$

No caso uni-dimensional, probabilidades s ao áreas debaixo da curva-densidade $f(x)$. No caso bi-dimensional, probabilidades são volumes debaixo da superfície-densidade $f(x,y)$. A probab do

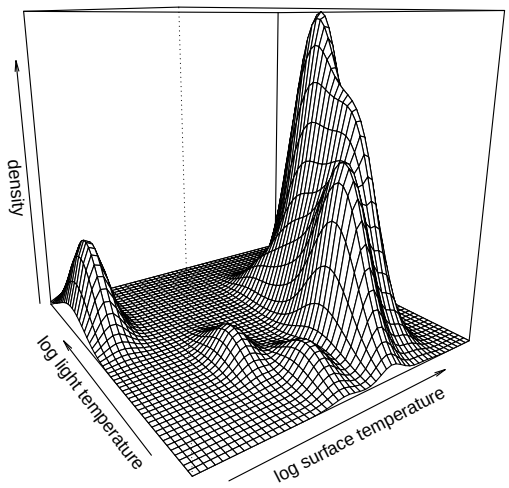


Figure 12.14: Densidade do vetor com $Y_1 =$ Intensidade da luz e $Y_2 =$ Temperatura de estrelas.

Figure 12.15: Old Faithful geyser.

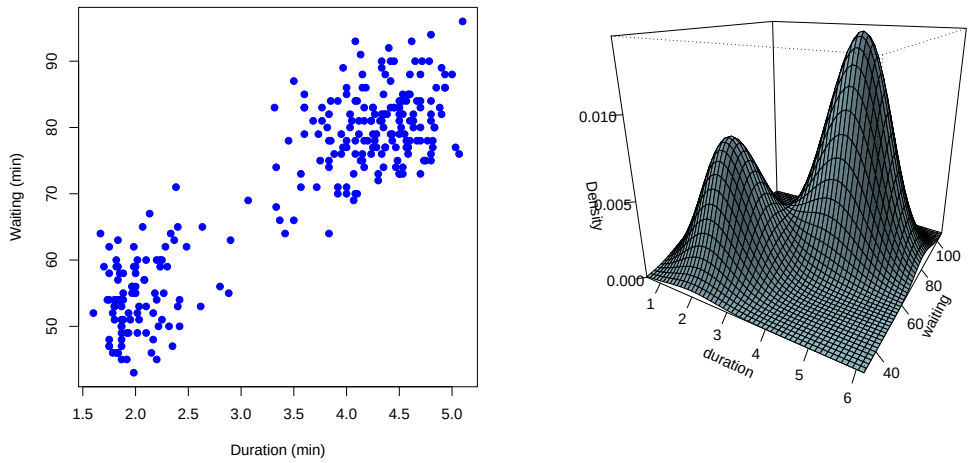


Figure 12.16: Old Faithful geyser: waiting time and eruption duration

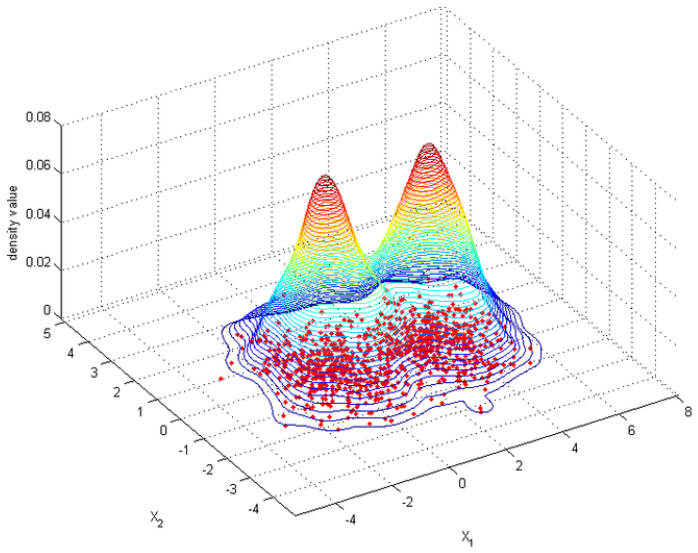


Figure 12.17: Old Faithful geyser: tempo de espera e duração de erupção.

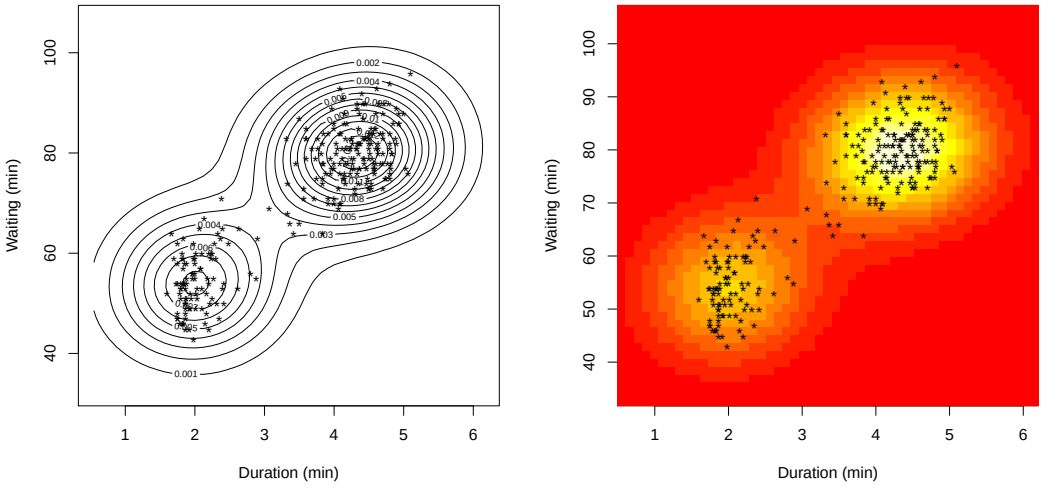


Figure 12.18: Old Faithful geyser: waiting time and eruption duration

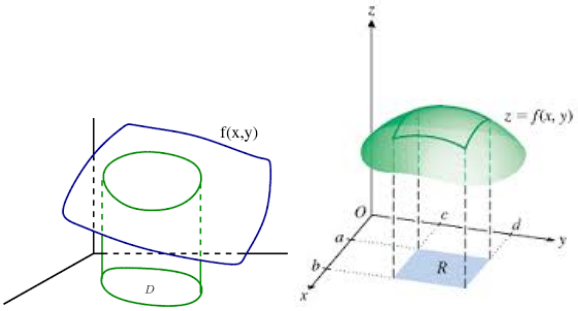


Figure 12.19: Probabilidade de (X,Y) cair em D é igual ao volume sob a superfície.

vetor (X,Y) cair numa região D do plano é

$$\mathbb{P}((X,Y) \in D) = \int\int_D f(x,y)dxdy$$

A Figura 12.19 ilustra a situação.

No caso geral de um vetor aleatório k-dimensional $\mathbf{Y} = (Y_1, \ldots, Y_k)$ com k v.a.'s contínuas, a densidade de probabilidade é qualquer função tal que:

- $f(\mathbf{y}) = f(y_1, \ldots, y_k) \geq 0$ para todo ponto $\mathbf{y} \in \mathbb{R}^k$
-

$$1 = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(y_1, \ldots, y_k)dy_1 \ldots dy_k$$

A probabilidade do vetor $\mathbf{Y}$ cair numa região D de $\mathbb{R}^k$ é dada por

$$\mathbb{P}((Y_1, \ldots, Y_k) \in D) = \int \ldots \int_D f(y_1, \ldots, y_k)dy_1 \ldots dy_k$$

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

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(X \leq x, Y < \infty) = F(x,\infty) = \lim_{y\to\infty} F(x,y)$$

Definition 12.16.2 — Jointly distributed random variables. We say $X_1, \cdots, X_n$ are *jointly distributed continuous random variables* and have *joint pdf* f if for any set $A \subseteq \mathbb{R}^n$

$$\mathbb{P}((X_1, \cdots, X_n) \in A) = \int_{(x_1,\cdots x_n)\in A} f(x_1, \cdots, x_n) \, \mathrm{x}_1 \cdots \mathrm{x}_n.$$

where

$$f(x_1, \cdots, x_n) \geq 0$$

and

$$\int_{\mathbb{R}^n} f(x_1, \cdots, x_n) \, \mathrm{x}_1 \cdots \mathrm{x}_n = 1.$$

In the case where $n = 2$,

$$F(x,y) = \mathbb{P}(X \leq x, Y \leq y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(x,y) \, \mathrm{x} \, \mathrm{y}.$$

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

$$\begin{aligned}\mathbb{P}(X \in A) &= \mathbb{P}(X \in A, Y \in (-\infty, +\infty)) \\ &= \int_{x\in A}\int_{-\infty}^{\infty} f(x,y) \, \mathrm{y} \, \mathrm{x} \\ &= \int_{x\in A} f_X(x) \, \mathrm{x}\end{aligned}$$

So

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y) \, \mathrm{y}$$

is the (marginal) pdf of X.

Definition 12.16.3 — Independent continuous random variables. Continuous random variables $X_1, \cdots, X_n$ are independent if

$$\mathbb{P}(X_1 \in A_1, X_2 \in A_2, \cdots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1)\mathbb{P}(X_2 \in A_2) \cdots \mathbb{P}(X_n \in A_n)$$

for all $A_i \subseteq \Omega_{X_i}$.

If we let F_{X_i} and f_{X_i} be the cdf, pdf of X, then

$$F(x_1, \cdots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$$

and

$$f(x_1, \cdots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$

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 $Y_1 \le Y_2$, then they are not independent, since $f(x_1, x_2) = 2I[Y_1 \le Y_2]$, 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

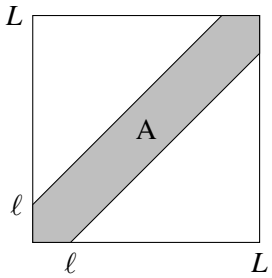
Often, when doing probability problems that involve geometry, we can visualize the outcomes with the aid of a picture.

■ **Example 12.16** Two points X and Y are chosen independently on a line segment of length L. What is the probability that $|X - Y| \le \ell$? 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

$$\int_A f(x, y) \, \dot{x} \, \dot{y} = \frac{2L\ell - \ell^2}{L^2}.$$

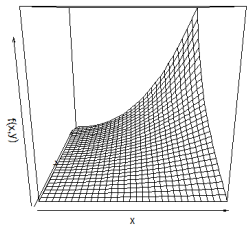


Figure 12.20: Gráfico de $f(x,y) = 60/13\ (x^2y + x^3y^4)$ no suporte $[0,1] \times [0,1]$.

■

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 conínuo, 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:

$$\begin{aligned} 1 &= \int\int_{[0,1]^2} k(x^2y + x^3y^4)\, dxdy = k \int_{[0,1]} \left(\frac{x^2}{2} + \frac{x^3}{5} \right) dx \\ &= k \left(\frac{1}{6} + \frac{1}{20} \right) = \frac{13k}{60} \end{aligned}$$

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 $f_Y(0.1) = 5/13\ (4(0.1) + 30.1^2) = 0.165$ enquanto que, no ponto $y = 0.9$, temos $f_Y(0.9) = 5/13\ (4(0.9) + 30.9^2) = 2.319$.

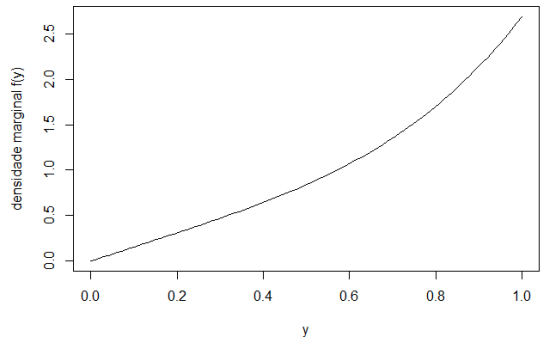


Figure 12.21: Gráfico de $f_Y(y) = 5/13(4y + 3y^4)$ no suporte $[0,1]$.

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: $f(x,y) = 60/13\ (x^2y + x^3y^4)$ 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 $Y = 0.2$:

$$f_{X|Y}(x|y = 0.2) = \frac{f_{XY}(x,0.2)}{f_Y(0.2)} = \frac{60/13\ (0.2\,x^2 + 0.2^4\,x^3)}{5/13\ (4\,0.2 + 3\,0.2^4)} = \frac{12}{0.8048}(0.2x^2 + 0.0016x^3)$$

Exercício básico

Comparando duas densidades condicionais de X: condicionada ao evento $Y = 0.20$ e ao evento $Y = 0.95$.

$$\begin{aligned} f_{X|Y}(x|y = 0.95) &= \frac{f_{XY}(x,0.95)}{f_Y(0.95)} \\ &= \frac{60/13\ (0.95\,x^2 + 0.95^4\,x^3)}{5/13\ (4\,0.95 + 3\,0.95^4)} = \frac{60}{31.217}(0.95\,x^2 + 0.95^4\,x^3) \end{aligned}$$

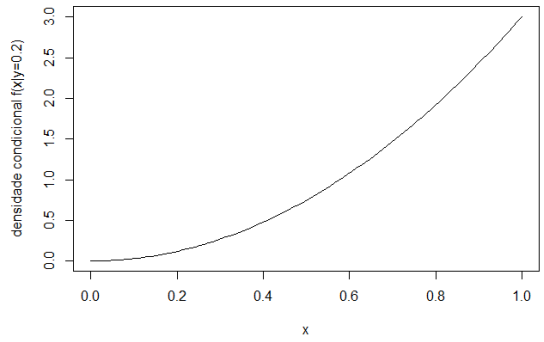


Figure 12.22: Gráfico de $f_{X|Y}(x|y=0.2) = 12/0.8048\ (0.2x^2 + 0.0016x^3)$ no suporte $[0,1]$.

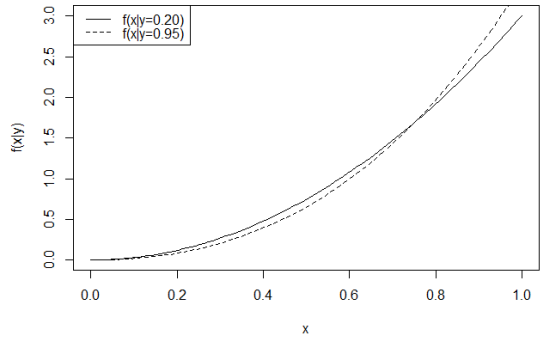


Figure 12.23: $f_{X|Y}(x|y=0.20)$ e $f_{X|Y}(x|y=0.95)$.

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 $f_Y(y) = 1/\sqrt{2\pi}\exp\left(-y^2/2\right)$ para $y \in \mathbb{R}$.
Densidade $(X|Y=-2)$:

$$\begin{aligned} f_{X|Y}(x|y=-2) &= \frac{f_{XY}(x,-2)}{f_Y(-2)} = \frac{\frac{1}{2\pi\sqrt{0.51}}\exp\left(-\frac{x^2+y^2-1.4xy}{1.02}\right)}{1/\sqrt{2\pi}\exp\left(-(-2)^2/2\right)} \\ &= 1/\sqrt{1.02\pi}\exp\left(-\frac{(x+1.4)^2}{1.02}\right) \end{aligned}$$

De forma similar, obtemos $f_{X|Y}(x|y=+2)$. Gráficos abaixo.
Vendo a condicional na conjunta

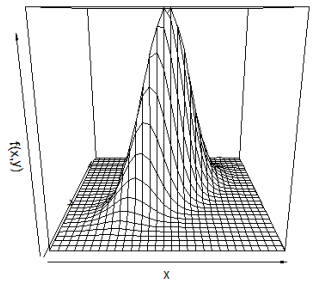


Figure 12.24: Densidade gaussiana bivariada $f_{XY}(x,y)$.

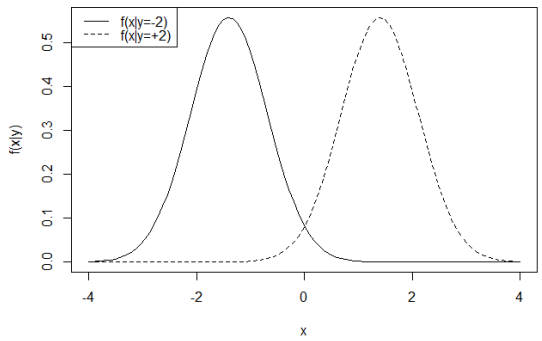


Figure 12.25: Gráfico de $f_{X|Y}(x|y=-2)$ e $f_{X|Y}(x|y=+2)$.

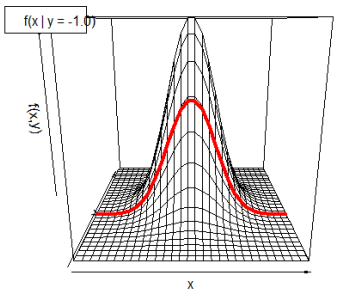


Figure 12.26: Gráfico de $f_{XY}(x,-1.0)$, que é proporcional a $f_{X|Y}(x|y=-1.0)$.

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 quisemos 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 $Y=(Y_1,Y_2,\ldots,Y_p)$ Calculamos a distribuição condicional de Y_1 dados os valores de $Y_2,\ldots,Y_p$ Podemos calcular o valor esperado de Y_1 dados (ou fixados) os valores de $Y_2,\ldots,Y_p$ É simplesmente como na definição usual de esperança de v.a.'s discretas:

$$\mathbb{E}(Y_1|Y_2=a_2,\ldots,Y_p=a_p) = \int y\, f_{Y_1|Y_2\ldots Y_p}(y|y_2=a_2,\ldots,y_p=a_p)\, dy$$

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 $(Y_2,\ldots,Y_p)$:

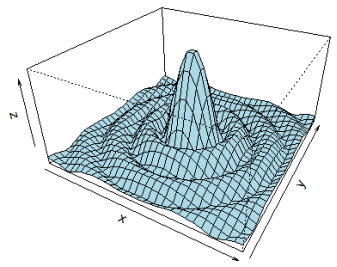


Figure 12.27: Gráfico de $f_{XY}(x,y)$, de onde queremos simular.

$$\mathbb{V}(Y_1|Y_2=a_2,\ldots,Y_p=a_p)=\int (y-m)^2 f_{Y_1|Y_2\ldots Y_p}(y|y_2=a_2,\ldots,y_p=a_p)\,dy$$

onde $m=\mathbb{E}(Y_1|Y_2=a_2,\ldots,Y_p=a_p)$ é a esperança condicional.

Pode-se mostrar que

$$\mathbb{V}(Y_1|Y_2=a_2,\ldots,Y_p=a_p)=\mathbb{E}(Y_1^2|Y_2=a_2,\ldots,Y_p=a_p)-m^2$$

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 $f(x,y)\leq Mg(x,y)$ 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

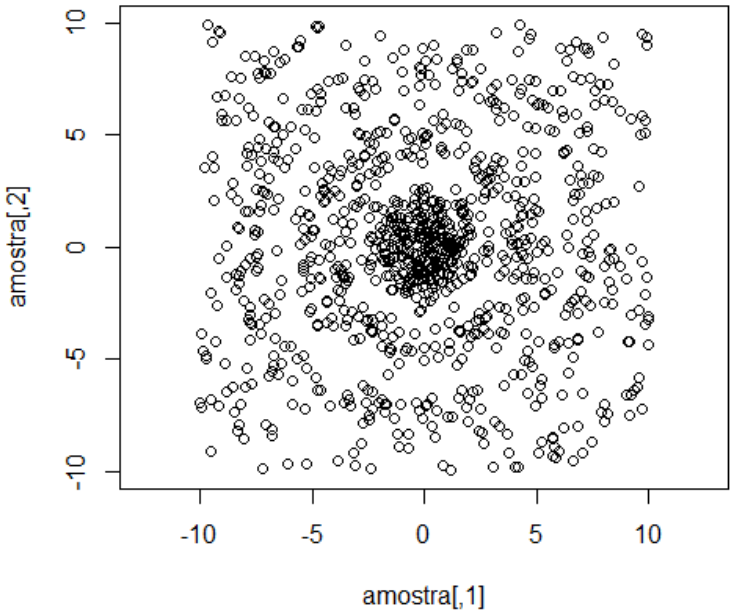


Figure 12.28: Amostra gerada de $f(x,y)$ por aceitação-rejeição

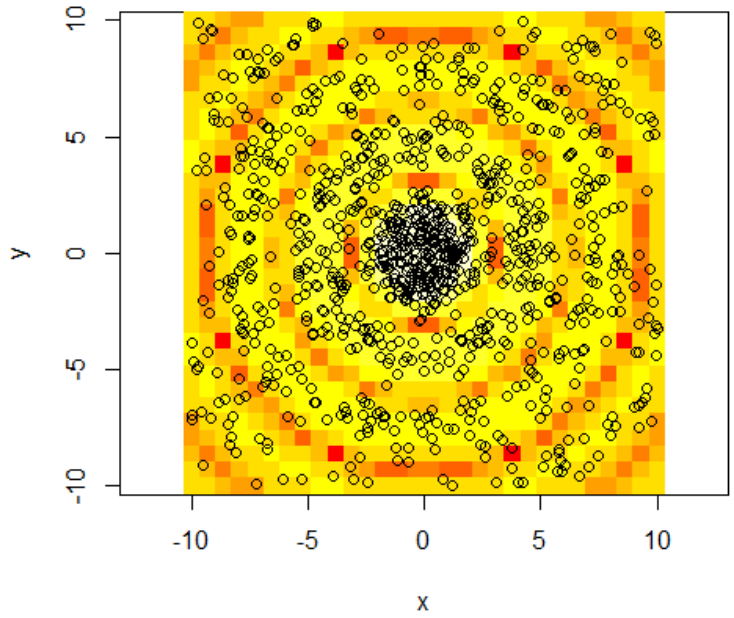


Figure 12.29: Amostra de $f(x,y)$e imagem heatmap da densidade.

13. Multivariate Gaussian Distribution

13.1 Normal bivariada: introdução

A distribuição gaussiana multivariada é extremamente importante para a análise de dados. Esta relevância ficará mais clara após estudarmos o Teorema Central do Limite no capítulo 19. Vamos começar estudando o caso bi-dimensional, em que temos um vetor aleatório $\mathbf{Y} = (Y_1, Y_2)$. Cada uma das v.a's segue uma distribuição gaussiana univariada com sua própria esperança μ_j e variância σ_j^2. Isto é, $Y_1 \sim N(\mu_1, \sigma_1^2)$ e $Y_2 \sim N(\mu_2, \sigma_2^2)$, como vimos na seção 7.8. No caso de uma gaussiana bivariada, as amostras do vetor bivariado $\mathbf{Y}$ formam nuvens de pontos no plano em forma de elipses centradas em $\mu = (\mu_1, \mu_2)$. Caso Y_1 e Y_2 sejam v.a.'s independentes, basta o conhecimento dessas duas distribuições marginais para conhecermos a distribuição conjunta pois, neste caso, $f_{Y_1,Y_2}(y_1, y_2) = f_{Y_1}(y_1) f_{Y_2}(y_2)$. Entretanto, o mais comum na análise de dados é elas não sejam v.a.'s independentes. A distribuição de Y_2 *muda se soubermos o valor da v.a.* Y_1 e nesta situação teremos $f_{Y_1,Y_2}(y_1, y_2) \neq f_{Y_1}(y_1) f_{Y_2}(y_2)$. A boa notícia é que mesmo no caso de v.a.'s dependentes, a gaussiana bivariada é bem simples. Toda a estrutura de dependência é controlada por um único parâmetro $\rho \in [-1, 1]$. Este parâmetro ρ é o índice de correlação ou associação linear entre Y_1 e Y_2 e ele será definido na seção 13.3.

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 i.i.d. do vetor aleatório $\mathbf{Y} = (Y_1, Y_2)$ extraídas desta distribuição. Isto é, vemos a amostra composta pelos $n = 100$ pares de vetores bi-dimensionais $\mathbf{Y}_i = (Y_{i1}, Y_{i2})$ com $i = 1, \ldots, 100$.

```
#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")
```

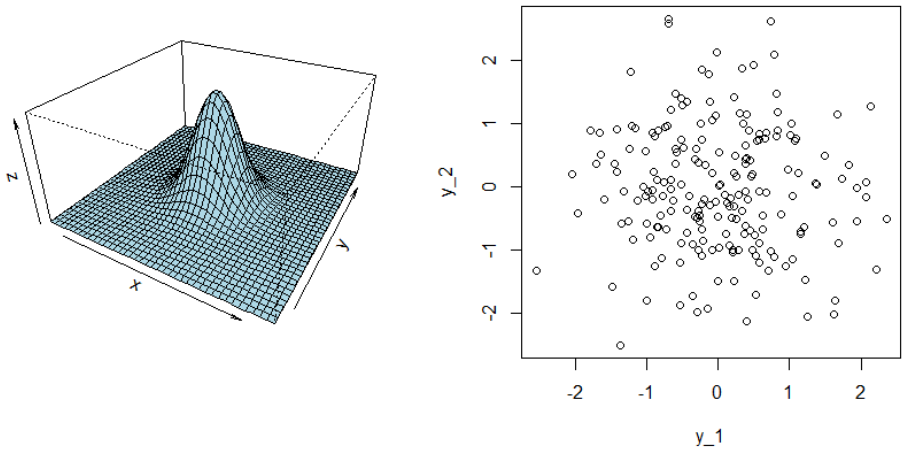


Figure 13.1: Esquerda: Densidade $f(y_1, y_2)$ de uma distribuição normal bivariada com $Y_1 \sim N(\mu_1 = 0, \sigma_1^2 = 1)$ e $Y_2 \sim N(\mu_2 = 0, \sigma_2^2 = 1)$ e com correlação $\rho = 0$. Direita: amostra com $n = 100$ instâncias do vetor Y.

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 mantendo as mesmas distribuições marginais da Figura 13.1 mas agora com correlação linear $\rho = 0.7$. No lado 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. 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^
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 da projeção ortogonal dos pontos ao longo de cada um dos dois eixos coordenados do gráfico da Figura 13.2 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 $\sqrt{\mathbb{V}(Y_2)} = \sigma_2 \approx 15$.

13.1.1 A distribuição condicional $(Y_2|Y_1 = y)$

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 $\mathbf{Y}$ é da forma $\mathbf{Y} = (Y_1, Y_2) = (14, Y_2)$. 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 $\mu_2 = 50$.

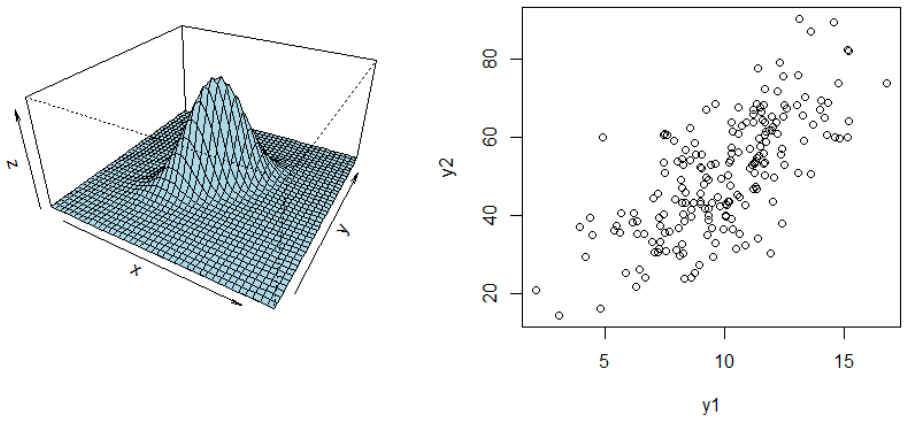


Figure 13.2: Esquerda: Densidade $f(y_1, y_2)$ de uma distribuição normal bivariada com $Y_1 \sim N(\mu_1 = 0, \sigma_1^2 = 1)$ e $Y_2 \sim N(\mu_2 = 0, \sigma_2^2 = 1)$ e com correlação $\rho = 0.7$. Direita: amostra com $n = 100$ instâncias do vetor $\mathbf{Y} = (Y_1, Y_2)$.

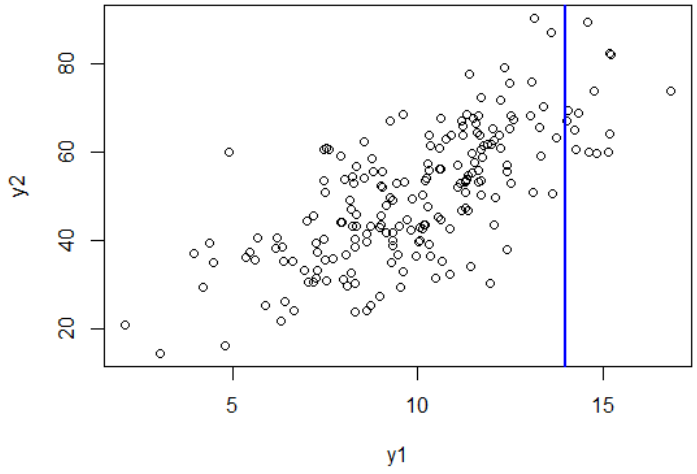


Figure 13.3: Amostra de normal bivariada. Se soubermos que o valor da v.a. Y_1 é igual a 14, o vetor vetor aleatório $\mathbf{Y}$ é da forma $\mathbf{Y} = (14, Y_2)$ com a primeira coordenada fixada no valor 14 e toda a aleatoriedade concentrada na segunda coordenada.

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 $\mathbf{Y}$ 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 $\mathbb{E}(Y_2)$ ou, em símbolos, $\mathbb{E}(Y_2|Y_1 = 14) > \mathbb{E}(Y_2)$.

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 da distribuição 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, usando os valores aproximados que obtivemos no olhômetro, teremos $(Y_2|Y_1 = 14) \approx N(70, 7.5^2)$.

Vamos encontrar os valores exatos dessa distribuição condicional. Para isto, vamos considerar a distribuição da v.a. $(Y_2|Y_1 = y)$ com um valor y genérico. Esta fórmula depende do coeficiente de correlação ρ, a ser definido na seção 13.3 e que neste exemplo vale $\rho = 0.7$. Temos

$$(Y_2|Y_1 = y) \sim N(\mu_{Y_2|Y_1=y}, \sigma^2_{Y_2|Y_1=y})$$

com

$$\mu_{Y_2|Y_1=y} = \mathbb{E}(Y_2|Y_1 = y) = \mu_2 + \frac{\rho\sigma_2}{\sigma_1}(y - \mu_1) \tag{13.1}$$

e

$$\sigma_{Y_2|Y_1=y} = \sqrt{\mathbb{V}(Y_2|Y_1 = y)} = \sigma_2\sqrt{1 - \rho^2}\,. \tag{13.2}$$

Por exemplo, para $Y_1 = 14$ (isto é, tomando y igual a 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

$$(Y_2|Y_1 = 14) \sim N(69.2, 10.71^2)$$

Nossa aproximação anterior, obtida de forma grosseira ao extrair os momentos condicionais a partir do gráfico da Figura 13.3, foi $(Y_2|Y_1 = 14) \approx N(70, 7.5^2)$.

13.1.2 A intuição para os momentos condicionais

Vamos tentar entender em detalhes a fórmula para a esperança condicional mostrada em (13.1). Suponha que ρ seja um valor positivo no intervalo $(0, 1)$. Como veremos na seção 13.3, este valor positivo da correlação indica que instâncias do vetor $\mathbf{Y} = (Y_1, Y_2)$ em que o valor observado de Y_1

está acima de seu valor esperado $\mathbb{E}(Y_1)$ tendem a ter valores da v.a. Y_2 também acima de seu valor esperado $\mathbb{E}(Y_2)$. Vamos quebrar a fórmula (13.1) em pedaços identificados pelas letras abaixo:

$$\mathbb{E}(Y_2|Y_1=y) = \underbrace{\mu_2}_{A} + \underbrace{\rho}_{E}\underbrace{\sigma_2}_{D}\underbrace{\frac{1}{\sigma_1}}_{C}\underbrace{(y-\mu_1)}_{B} \tag{13.3}$$

Sem conhecer o valor de Y_1, nós esperamos que Y_2 seja um valor ao redor de μ_2, indicado pela letra A em (13.3). Como o valor y assumido pela v.a. Y_1 afeta o que podemos esperar para a v.a. Y_2? Esta modificação é dada pelo termo (positivo ou negativo) que estamos adicionando a μ_2 e composto pelo produto dos termos indicados nas letras B,C,D,E. De alguma forma somos informados que o valor de Y_1 é igual a y. Olhamos quão afastado de $\mu_1 = \mathbb{E}(Y_1)$ é este valor y observado para Y_1. Este é o valor de B em (13.3). Se y estiver muito afastado de μ_1 (um valor extremo para a v.a. Y_1) e se as v.a.'s Y_1 e Y_2 forem correlacionadas, podemos esperar um grande impacto no valor que podemos esperar para Y_2. Entretanto Y_1 e Y_2 são variáveis que podem ter escalas completamente diferentes tais como Y_1 sendo quantidade de chuva e Y_2 sendo o preço da tonelada de café. Como passar o desvio $y-\mu_1$ em B para a escala de Y_2? Primeiro, nós padronizamos esse desvio usando C, o desvio-padrão de Y_1. Assim, o desvio y em relação à μ_1 é medido agora em unidades de desvio-padrão.

Suponhamos que $(y-\mu_1)/\sigma_1$ seja igual a 2. Isto é, Y_1 desviou-se acima de seu valor esperado por 2 desvios-padrão. Quanto devemos somar ao valor esperado para Y_2? O termos D multiplica este desvio-padronizado (igual a 2) pelo desvio-padrão σ_2, letra D em (13.3). Assim, se não houvesse o fator ρ (em E) o acréscimo a μ_2 seria a mesma quantidade de desvios de σ_2. Isto é, se y desvia-se de μ_1 por $2*\sigma_1$, ignorando ρ, teríamos também o mesmo acréscimo de $2*\sigma_2$ sendo somado a μ_1. Todo desvio em Y_1 seria transferido para o valor esperado de Y_1.

Entretanto, esse repasse não é feito integralmente. Como veremos na seção 13.3, na prática teremos $|\rho|<1$. Assim, o acréscimo $\sigma_2*(y-\mu_1)/\sigma_1$ é reduzido (em valor absoluto) pela multiplicação por ρ. Quando $\rho\approx 0$, quase nada do desvio de Y_1 é repassado para o valor esperado condicional de Y_2. Entretanto, se $\rho\approx 1$ ou $\rho\approx -1$, quase todo o desvio padronizado observado em Y_1 é transferido para Y_2. O fato de não ser transferido todo o desvio mas apenas uma parcela ρ é chamado de *Feeito de regressão para a média* e será discutido no capítulo **??**, que trata de regressão linear.

Quanto à fómula (13.2) para o desvio-padrão da distribuição condicional de $(Y_2|Y_1=y)$, ela também pode ser interpretada intuitivamente. Primeiro, note que o valor y não aparece nesta fórmula e portanto este desvio-padrão condicional é o mesmo, qualquer que seja o desvio observado na v.a. Y_1. Depois, note que, como $|\rho|<1$, então $\sqrt{1-\rho^2}<1$ e portanto o desvio-padrão condicional $\sigma_{Y_2|Y_1=y}$ é menor que $\sqrt{\mathbb{V}(Y_2)}=\sigma_2$, o desvio-padrão marginal de Y_2. Isto significa que, ao condicionarmos no valor de Y_1, a variabilidade na outra variável Y_2 diminui. A informação de que $Y_1=y$ faz com que possamos esperar um valor de Y_2 mais concentrado em torno de $\mathbb{E}(Y_2|Y_1=y)$, o novo valor esperado para Y_2. De quanto é esta redução? Depende do valor de ρ. Se $|\rho|\approx 1$, a redução é grande. Se $\rho\approx 0$, a redução é pequena. Assim, conhecer o valor de uma v.a. Y_1 faz com que possamos predizer o valor de outra v.a. Y_2 altamente correlacionada (isto é, com $|\rho|\approx 1$) com a primeira usando o valor esperado $\mathbb{E}(Y_2|Y_1=y)$. O valor realmente observado de Y_2 vai desviar-se pouco (em média, apenas por $\sigma_2\sqrt{1-\rho^2}\approx 0$) deste valor esperado condicional.

13.1.3 A densidade conjunta bivariada de $\mathbf{Y}=(Y_1,Y_2)$

Como conhecemos as fórmulas (13.1) e (13.2)? Simplesmente fazendo o cálculo matemático da densidade condicional:

$$f_{Y_2|Y_1}(y_2|y_1=a) = \frac{f_{\mathbf{Y}}(a,y_2)}{f_{Y_1}(a)} \tag{13.4}$$

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

$$\sum = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \tag{13.5}$$

onde ρ é a correlação com valor no intervalo $[-1,1]$, e onde σ_1 e σ_2 são os desvios padrões marginais de cada uma das duas variáveis. 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

$$f_{\mathbf{Y}}(\mathbf{y}) = \text{ cte} \times \exp\left(-\frac{1}{2}\ d^2(\mathbf{y}, \mu)\right) \tag{13.6}$$

onde $d^2(\mathbf{y}, \mu)$ é uma medida de distância (ao quadrado) 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

$$d^2(\mathbf{y}, \mu) = (\mathbf{y} - \mu)' \Sigma^{-1} (\mathbf{y} - \mu) \tag{13.7}$$

Vamos estudá-la detalhadamente na seção 13.8.

O restante desta seção é opcional e pode ser lido de maneira ligeira, sem focar excessivamente nos detalhes algébricos. Assumindo $|\rho| < 1$, vamos usar a definição da matriz Σ em (13.5) para invertê-la e fazer as multiplicações matriciais. Um cálculo tedioso mas muito simples leva finalmente ao resultado final. Este resultado está abaixo para referência mas ele não será usado posteriormente neste livro:

$$\begin{aligned} d^2(\mathbf{y}, \mu) &= (\mathbf{y} - \mu)' \Sigma^{-1} (\mathbf{y} - \mu) \\ &= (y_1 - \mu_1, y_2 - \mu_2) \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \end{bmatrix} \\ &= -\frac{1}{1-\rho^2} \left[\left(\frac{y_1 - \mu_1}{\sigma_1} \right)^2 - 2\rho \frac{(y_1 - \mu_1)(y_2 - \mu_2)}{\sigma_1\sigma_2} + \left(\frac{y_2 - \mu_2}{\sigma_2} \right)^2 \right] \end{aligned}$$

Observe que $d^2(\mathbf{y}, \mu)$ soma o desvio padronizado $(y_1 - \mu_1)/\sigma_1$ ao quadrado, bem como o desvio padronizado $(y_2 - \mu_2)/\sigma_2$, também ao quadrado. Além desses dois termos, d^2 envolve também um termo que multiplica os dois desvios padronizados. Este termo só aparece se $\rho \neq 0$.

A constante que aparece na expressão da densidade normal bivariada em (13.6) é igual a $(2\pi\sigma_1\sigma_2\sqrt{1-\rho^2})^{-1}$. Portanto, a densidade conjunta da normal bivariada no ponto (a, y_2) e escrita de forma bem menos condensada que em (13.6) é igual a

$$f_{\mathbf{Y}}(y_1, y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)} \left[\left(\frac{y_1 - \mu_1}{\sigma_1} \right)^2 - 2\rho \frac{(y_1 - \mu_1)(y_2 - \mu_2)}{\sigma_1\sigma_2} + \left(\frac{y_2 - \mu_2}{\sigma_2} \right)^2 \right] \right)$$

(13.8)

Para obter a densidade conjunta de $(Y_2|Y_1 = a)$ como mostrada em (13.4), precisamos do denominador. Isto é, precisamos integrar a densidade conjunta com respeito a y_2 e avaliar o resultado no valor $y_1 = a$. Este é um cálculo longo e que, no final, produz uma densidade que podemos reconhecer como sendo uma gaussiana com esperança μ_1 e desvio-padrão σ_1:

$$f_{Y_1}(a) = \int f_{\mathbf{Y}}(a, y_2)dy_2 = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{1}{2}\frac{(a-\mu_1)^2}{\sigma_1^2}\right)$$

Finalmente, dividindo a expressão em (13.8) por aquela em (**??**) e fazendo uma série de manipulações algébricas acabam levando à expressão da densidade condicional $(Y_2|Y_1 = a)$ que é a densidade de uma gaussiana com a esperança e desvio-padrão condicionais mostrados em (13.1) e em (13.2).

13.2 O desvio padronizado

Podemos *resumir* a distribuição de probabilidade de uma v.a. Y com os valores numéricos e teóricos representados pela esperança $\mu_Y = \mathbb{E}(Y)$ e o desvio-padrão $\sigma_Y = \sqrt{\mathbb{V}(Y)}$. Estes resumos μ_Y e σ_Y não dependem de dados estatísticos. São resultados de cálculos matemáticos feitos com a densidade de probabilidade $f(y)$ no caso contínuo ou com a função de probabilidade $p(y) = \mathbb{P}(Y = y)$ no caso discreto. Eles servem para resumir toda a distribuição teórica de uma v.a. em dois números. O primeiro deles, $\mathbb{E}(Y)$, é o valor em torno do qual os dados tendem a oscilar e o segundo, $\sigma_Y = \sqrt{\mathbb{V}(Y)}$, é o valor típico do desvio dos dados em relação a $\mu_Y = \mathbb{E}(Y)$.

Vamos agora passar a olhar os dados estatísticos. Suponha que temos uma amostra aleatória de Y. Isto é, v.a.'s $Y_1, Y_2, \ldots, Y_n$ 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. Como já sabemos, a forma do histograma segue aproximadamente a densidade $f(y)$.

Existe a contraparte empírica dos resumos: podemos estimar os resumos *teóricos* $\mu_Y = \mathbb{E}(Y)$ e $\sigma_Y = \sqrt{\mathbb{V}(Y)}$ a partir dos dados. Pela Lei dos Grandes Números (ver capítulo 19), se o tamanho n da amostra é grande, temos a média aritmética $\overline{Y} = (Y_1 + \ldots + Y_n)$ aproximadamente igual a $\mathbb{E}(Y)$. Além disso, o desvio-padrão *amostral* $S = \sqrt{\sum_i \left(Y_i - \overline{Y}\right)^2 / n} \approx \sigma$. Às vezes, define-se o DP amostral S usando $n-1$ no denominador. A razão para isto ficará clara no capítulo **??** mas a diferença entre usar n ou $n-1$ é mínima a não ser que n seja muito pequeno. Assim, na prática, tanto faz usar n ou $n-1$ se a amostra não for muito pequena. Note que $\overline{Y} \neq \mathbb{E}(Y)$ e $S \neq \sigma$. As estatísticas $\overline{Y}$ e S dependem dos dados e variam de amostra para amostra, mesmo que o mecanismo gerador das amostras não mude.

Em geral, não olhamos diretamente o desvio $Y - \mu$ de uma observação aleatória Y em relação ao seu valor esperado $\mu = \mathbb{E}(Y)$. Costumamos preferir olhar o desvio paronizado, definido a seguir.

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 $Y - \mu$ 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 desvios-padrão em relação a μ. Pela desigualdade de Tchebyshev, vimos que, qualquer que seja a distribuição de Y, temos que o evento $Z > 4$ é raro, tendo probabilidade de ocorrer menor que $1/16 \approx 0.06$.

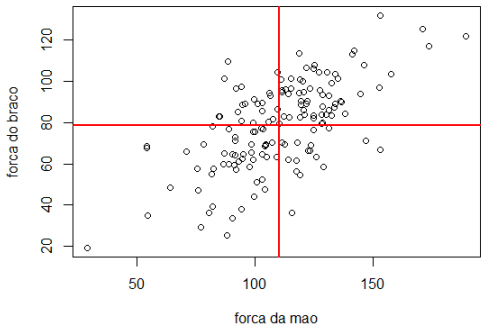


Figure 13.4: Relação entre força de preensão (do aperto de mão) e força do braço para 147 pessoas que trabalham em empregos fisicamente extenuantes.

13.3 O índice ρ de correlação de Pearson

Como medir a associação entre duas variáveis Y_1 e Y_2 medidas num mesmo elemento aleatório ω? Estas duas variáveis podem 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 tendẽncia 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 foram desenhadas nas posiccões determinadas pelas médias aritméticas $\bar{y}_1$ e $\bar{y}_2$, respectivamente. Elas 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. Assim, quando Y_1 está acima de seu valor esperado μ_1 (isto é, $Z_1 > 0$), em geral, temos também Y_2 acima de seu valor esperado μ_2 (ou seja, $Z_2 > 0$). E quando $Z_1 < 0$, costumamos ter $Z_2 < 0$.

Existem várias formas intuitivas de medir a associação entre Y_1 e Y_2. Por exemplo, uma forma simples seria medir a associação como a proproção dos pontos nos quadrantes 1 e 3 menos a proporção nos quadrantes 2 e 4. Entretanto, nós preferimos uma outra medida que é *não intuitiva* mas que tem excelentes propriedades teóricas: o índice de correlação de Pearson. Considere o produto dos desvios padronizados:

$$Z_1 Z_2 = \frac{Y_1 - \mu_1}{\sigma_1} \times \frac{Y_2 - \mu_2}{\sigma_2} \tag{13.9}$$

Se desvios grandes e positivos de Y_1 tendem a ocorrer com desvios grandes e positivos de Y_2, seu também será grande. Ao mesmo tempo, se os desvios grandes e negativos de Y_1 tendem a ocorrer com desvios grandes e negativos de Y_2, seu produto também será grande e positivo. 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 $Z_1 Z_2$. Como $\mathbf{Y} = (Y_1, Y_2)$ é um vetor aleatório, com os valores das duas v.a.'s medidas no mesmo item, o produto $Z_1 Z_2$ em (13.9) é uma variável aleatória. 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 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. $Z_1 Z_2$. Como resumir esta v.a.

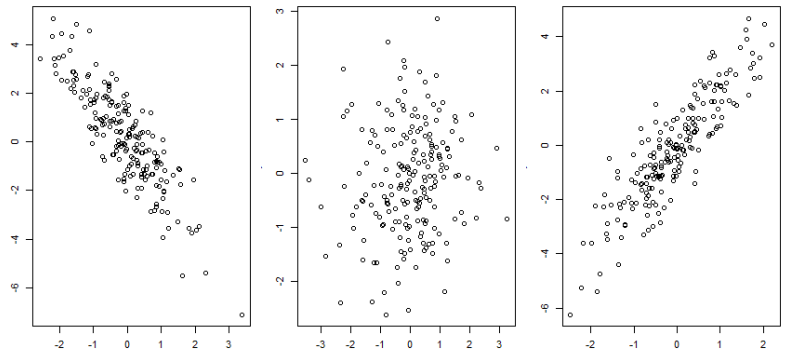


Figure 13.5: Tipicamente, em média, o produto dos desvios padronizados $Z_1 Z_2$ é positivo (esquerda), próximo de zero (centro) e negativo (direita).

num único número? Já sabemos fazer isto com qualquer v.a.: tomamos o seu valor esperado. Isto é, vamos calcular

$$\rho = \text{Corr}(Y_1, Y_2) = \mathbb{E}(Z_1 Z_2) = \mathbb{E}\left(\frac{Y_1 - \mu_1}{\sigma_1} \times \frac{Y_2 - \mu_2}{\sigma_2}\right)$$

Este resumo é o índice de correlação de Pearson.

13.4 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:

$$\text{Corr}(Y_1, Y_2) = \mathbb{E}(Z_1 Z_2) = \text{Corr}(Y_2, Y_1)$$

- 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.

13.5 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 $(Y_1, Y_2, \ldots, Y_p)$ de v.a.'s Podemos fazer uma matriz $p \times p$ de correlação. Na posição (i, j) teremos

$$\rho_{ij} = \text{Corr}(Y_i, Y_j) = \mathbb{E}\left(\frac{Y_i - \mu_i}{\sigma_i} \times \frac{Y_j - \mu_j}{\sigma_j}\right)$$

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.

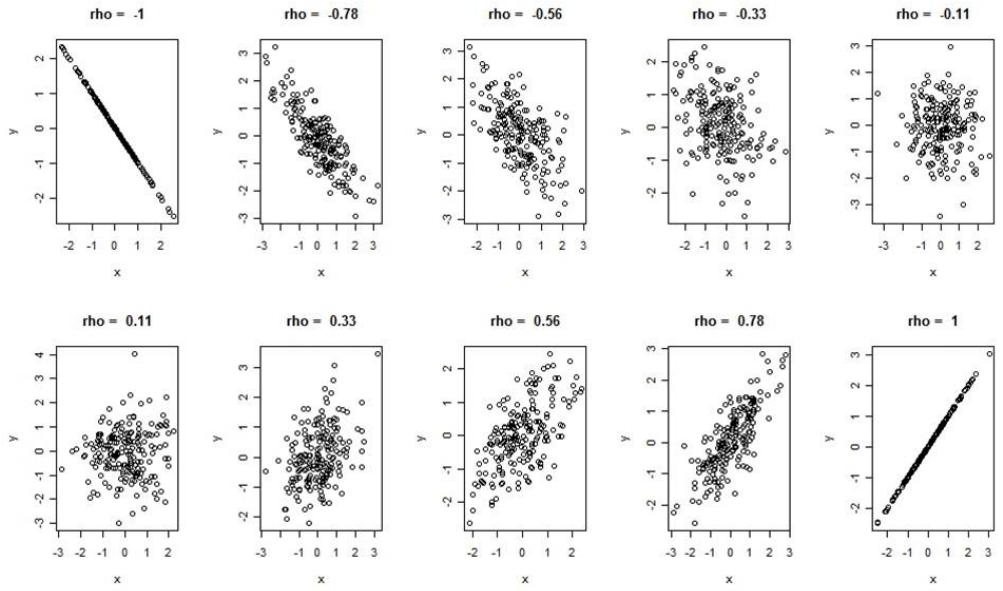


Figure 13.6: Mostrando como a associação entre as variáveis muda quando ρ muda de valor.

Por exemplo, vamos considerar um vetor aleatório $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_9)$ 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 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 variáes são

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.

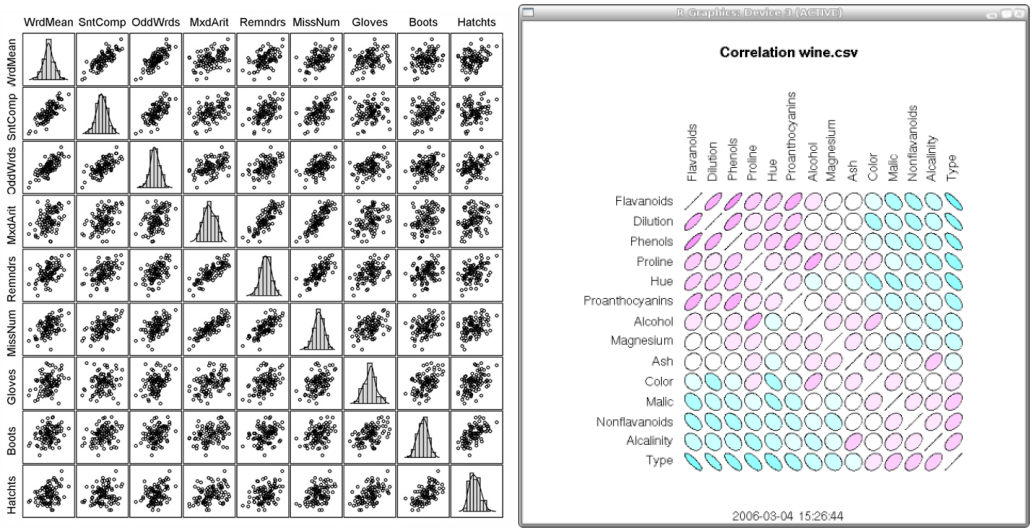


Figure 13.8: Esquerda: Matriz de scatterplots entre os pares formados a partir das 9 medidas de um teste de personalidade. Direita: Matriz estilizada dos scatterplots. A partir dos dados amostrais mostra-se o formato da nuvem de pontos de uma amostra de vinhos com 14 variáveis medidas em cada um dos vinhos. Gráfico feito com o pacote `rattle`.

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 lo 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 póximas 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 a 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 [17] desenvolveu um modelo discriminante linear (ver capítulo 17) 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. 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

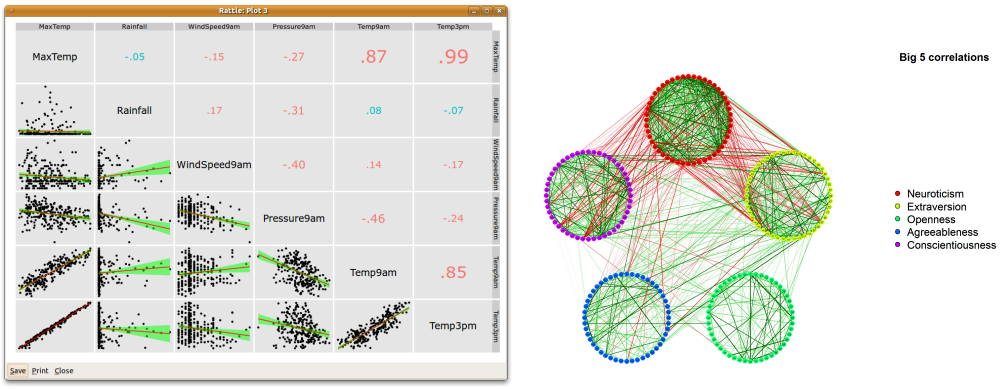


Figure 13.9: Esquerda: Mais uma visualização da matriz de correlação com R + rattle. Direita: Uma visualização com `qgraph`. V.a.'s são vértices e correlações são arestas. Uma aresta verde significa uma correlação positiva e enquanto vermelha significa uma correlação negativa. As arestas mais grossas e saturadas tem $|\rho|$ grande.

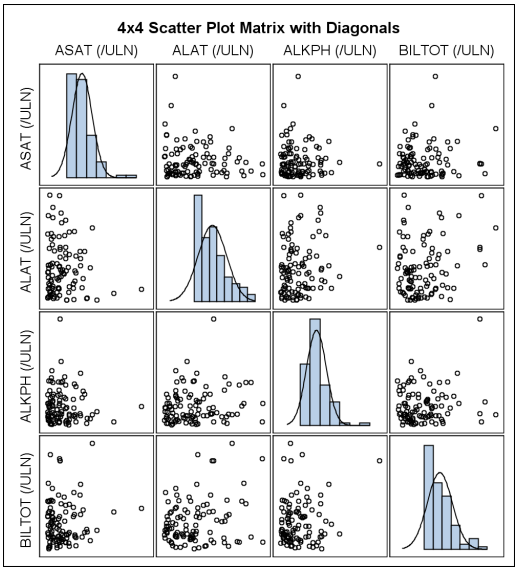


Figure 13.10: Scatterplot matrix of 4 lab variables to test liver functioning commonly used in clinical research

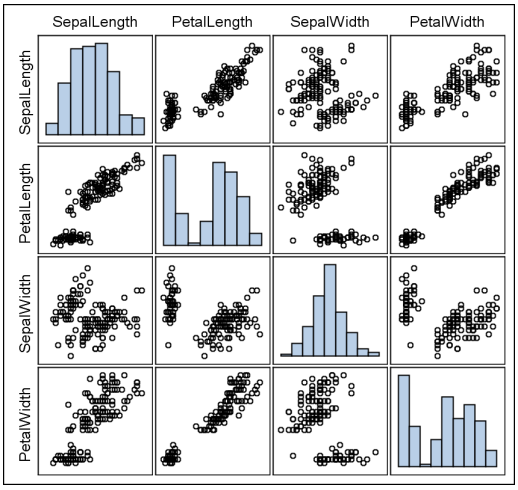


Figure 13.11: Scatterplot matrix de 4 variáveis medidas numa flor: comprimento de pétala, largura de pétala, comprimento de sétala, largura de sétala. Três espécies distintas misturadas. Relação entre as variáveis é diferente, ela depende da espécie.

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 consideramos 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 **??**. Lá, nós voltaremos a tratar os modelos de mistura e suas complicações.

13.6 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:

$$\text{Corr}(Y_1, Y_3) = \text{Corr}(Y_1, 32 + 1.8Y_2) = \text{Corr}(Y_1, Y_2)$$

13.7 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 $\frac{1}{n}(Y_1 + \ldots + Y_n) = \overline{Y} \approx \mathbb{E}(Y)$ e como $S = \sqrt{\sum_i \left(Y_i - \overline{Y}\right)^2 / n} \approx \sigma$ podemos aproximar

$$\rho = \mathbb{E}\left(\frac{Y_1 - \mu_1}{\sigma_1} \times \frac{Y_2 - \mu_2}{\sigma_2}\right) \approx \mathbb{E}\left(\frac{Y_1 - \overline{Y_1}}{S_1} \times \frac{Y_2 - \overline{Y_2}}{S_2}\right)$$

onde $\overline{Y_1}$ é a média aritmética dos n valores da variável 1, etc. Isto é, $\overline{Y_1}$ 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 $y_{11}, \ldots, y_{n1}$ da coluna 1 da tabela, calcule uma nova coluna d comprimento n formada por

$$z_{i1} = \frac{y_{i1} - \overline{y_1}}{s_1}$$

Faça o mesmo para a coluna 2, criando uma outra coluna de desvios padronizados empíricos:

$$z_{i2} = \frac{y_{i2} - \overline{y_2}}{s_2}$$

A seguir, multiplique as duas colunas de desvios padronizados e tire a sua média aritmética calculando

$$r = \frac{1}{n}\sum_{i=1}^{n} z_{i1} z_{i2} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_{i1} - \overline{y_1}}{s_1}\right)\left(\frac{y_{i2} - \overline{y_2}}{s_2}\right)$$

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.

13.8 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 aleatáorio $\mathbf{Y} = (Y_1, Y_2)$. 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 $\mu_2 = 80$.

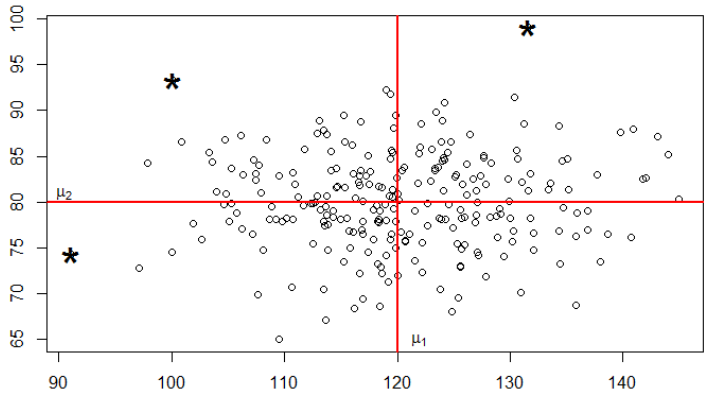


Figure 13.12: Amostra de $\mathbf{y}_i = (y_{i1}, y_{i2})$ com $i = 1, 2, \ldots, 250$.

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 náo 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 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 $\sigma = 10$

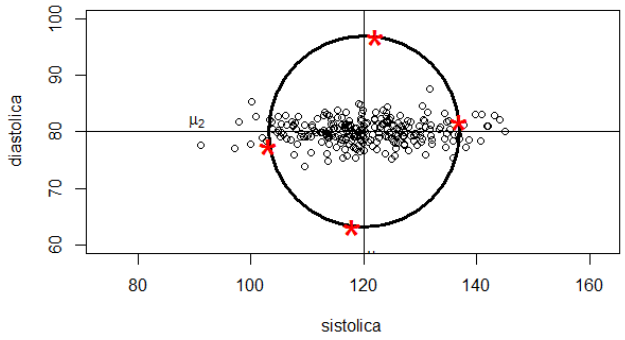


Figure 13.13: Amostra de $\mathbf{y}_i = (y_{i1}, y_{i2})$ com $i = 1, 2, \ldots, 250$. A distância euclidiana continua sendo uma medida de distância estatística razoável?

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)

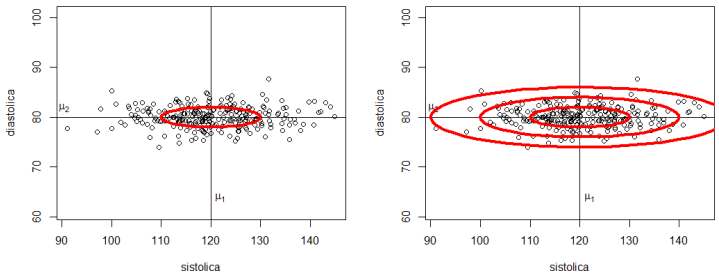


Figure 13.14: Esquerda: Lugar geométrico dos pontos a igual distância $c = 1$ do centro: uma elipse. Estes são pontos (y_1, y_2) que estão a uma distância c igual a 1 do centro (μ_1, μ_2). Direita: variando $c = 1, 2, 3$ obtemos elipses concêntricas. Isto é, os pontos de cada elipse satisfazem $d(\mathbf{y}, \mu) = c$ 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

$$d(y, \mu) = \sqrt{\left(\frac{y_1 - 120}{10}\right)^2 + \left(\frac{y_2 - 80}{2}\right)^2} = 1$$

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

$$d(y, \mu) = \sqrt{\left(\frac{y_1 - 120}{10}\right)^2 + \left(\frac{y_2 - 80}{2}\right)^2} = c$$

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 **??**. 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?

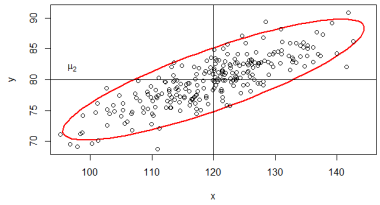


Figure 13.15: Amostra de vetor bivariado $\mathbf{Y} = (Y_1, Y_2)$ em que as variáveis aleatórias possuem correlação positiva.

$$
\begin{aligned}
d^2(\mathbf{y},\mu) &= \left(\frac{y_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{y_2-\mu_2}{\sigma_2}\right)^2 \\
&= (y_1-\mu_1, y_2-\mu_2) \begin{bmatrix} 1/\sigma_1^2 & 0 \\ 0 & 1/\sigma_2^2 \end{bmatrix} \begin{pmatrix} y_1-\mu_1 \\ y_2-\mu_2 \end{pmatrix} \\
&= (y_1-\mu_1, y_2-\mu_2) \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}^{-1} \begin{pmatrix} y_1-\mu_1 \\ y_2-\mu_2 \end{pmatrix} \\
&= \begin{pmatrix} y_1-\mu_1 \\ y_2-\mu_2 \end{pmatrix}' \Sigma^{-1} \begin{pmatrix} y_1-\mu_1 \\ y_2-\mu_2 \end{pmatrix} \\
&= (\mathbf{y}-\mu)'\Sigma^{-1}(\mathbf{y}-\mu)
\end{aligned}
$$

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

$$\sum = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

então $d^2(\mathbf{y},\mu) = (y-\mu)'\Sigma^{-1}(y-\mu) = c$ é a equação de uma elipse centrada no vetor $\mu = (\mu_1, \mu_2)$. Quando, como acima, a matriz $\sum$ 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 $\mathbb{E}(\mathbf{Y}) = \mu = (\mu_1, \mu_2)$. Pontos estatisticamente equidistantes de μ 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:

$$d^2(y,\mu) = (y-\mu)'\Sigma^{-1}(y-\mu)$$

Ela é a mesma expressão matricial de distância que usamos antes, mas a matriz Σ não é mais diagonal. Quem é $\sum$? A matriz $\sum$ é uma matriz 2 x 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 $\sum$ 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 17). Depois desta revisão, vamos retornar à distância estatística e à distribuição normal multivariada.

13.9 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.

13.9.1 Formas quadráticas

Começamos definindo forma quadrática. Seja $\mathbf{v} = (v_1, \ldots, v_p)'$ um vetor em $\mathbb{R}^p$. Um vetor será sempre um vetor-coluna neste livro. Seja Σ uma matriz $p \times p$.

Definition 13.9.1 — 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

$$(v_1, v_2) \, \Sigma \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \Sigma_{11} v_1^2 + \Sigma_{12} v_1 v_2 + \Sigma_{21} v_2 v_1 + \Sigma_{22} v_2^2$$

A forma quadrática envolveas 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 razã 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).

13.9.2 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 *bsA* no lugar de Σ^{-1}, a distância estatística entre um vetor $\mathbf{y}$ e a origem (ou o vetor nulo) *bs*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 13.9.2 — 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

$$0 < \mathbf{v}' \, \mathbf{A} \, \mathbf{v} = \begin{bmatrix} v_1 & \dots & v_n \end{bmatrix} \mathbf{A} \begin{bmatrix} v_1 \\ \dots \\ v_n \end{bmatrix} = \sum_{i,j} \mathbf{A}_{i,j} v_i v_j$$

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 sua entradas positivas, por exemplo, não é um critério válido. Para confirmar isto, considere o caso abaixo

$$\mathbf{v}'\ \mathbf{A}\ \mathbf{v} = \begin{bmatrix} 1 & -2 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = -3$$

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 13.9.1 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.

13.9.3 Autovetores e autovalores

Definition 13.9.3 — 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 $\sqrt{\mathbf{v}'\,\mathbf{v}} = \sqrt{v_1^2 + v_2^2 + \ldots + v_p^2}$. 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 qe 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 númers complexos. Este não é o caso quando Σ for uma matriz simétrica.

Theorem 13.9.2 — 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 13.9.3 — 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

$$\mathbf{v}'\,\Sigma\,\mathbf{v} = \lambda\,\mathbf{v}'\mathbf{v} = \lambda\,\|\mathbf{v}\|^2 \tag{13.10}$$

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 (13.10), $\lambda\,\|\mathbf{v}\|^2 > 0$. Como $\mathbf{v} \neq \mathbf{0}$ e a norma $\|\mathbf{v}\|^2 0$ 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

$$\Sigma^{-1}\Sigma\,\cdot\,\mathbf{v} = \Sigma^{-1}\,(\lambda\,\mathbf{v})$$

ou seja

$$\mathbf{v} = \lambda\Sigma^{-1}\,\mathbf{v}$$

ou ainda, como $\lambda > 0$,

$$\frac{1}{\lambda}\,\mathbf{v} = \lambda\Sigma^{-1}\,\mathbf{v}$$

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 13.9.4 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.

13.9.4 Teorema Espectral

Seja Σ uma matriz $p \times p$ simétrica e positiva definida. Existem p autovalores associados com A. 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 $\mathbf{P}^t\mathbf{P} = \mathbf{I}$ 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 $\Sigma = \mathbf{P}\,\mathbf{D}\,\mathbf{P}^t$

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 diagonalizavel.

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^t. Lembre-se que $\mathbf{P}^t\mathbf{P} = \mathbf{I}$.

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.

13.10 Densidade da normal multivariada

Finalmente podemos apresentar a densidade da distribuição normal multivariada. Seja $\mathbf{Y} = (Y_1, \ldots, Y_p)$ um vetor aleatório de v.a.'s contínuas. Seja $\mu = (\mu_1, \ldots, \mu_p)$ o vetor esperado $\mathbb{E}(\mathbf{Y}) = (\mathbb{E}(Y_1), \ldots, \mathbb{E}(Y_p))$. Seja Σ a matriz $p \times p$ de covariância do vetor $\mathbf{Y}$.

Definition 13.10.1 — 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

$$d^2(\mathbf{y}, \mu) = (\mathbf{y} - \mu)'\Sigma^{-1}(\mathbf{y} - \mu)$$

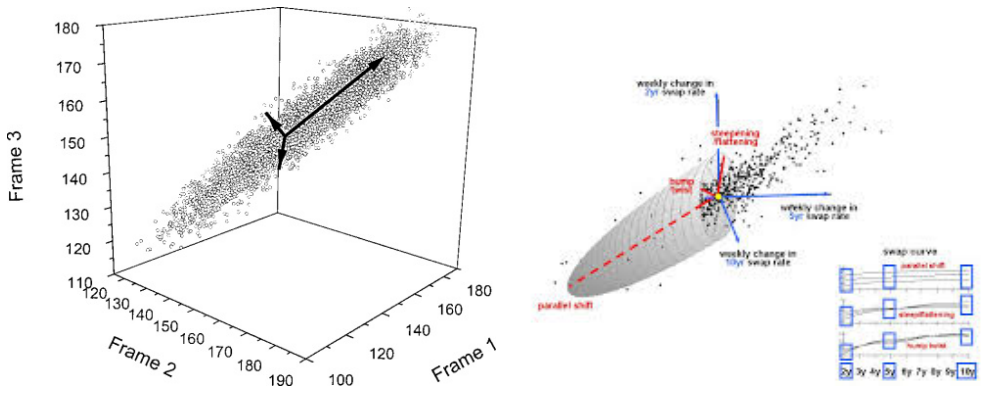


Figure 13.16: Nuvem de pontos de uma normal tri-dimensional $\mathbf{Y} \sim N_3(\mu, \sum)$.

é a distância estatística (de Mahalanobis) entre $\mathbf{y}$ e μ. Notação: $\mathbf{Y} \sim N_k(\mu, \sum)$

A densidade decresce com d^2. As superfícies de nível da densidade são elipsoides concêntricos centrados em μ. Os eixos do elipsóide estão na direção dos autovetores de $\sum$ 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. Principal Component Analysis

14.1 Introdução

Falta passar os slides para o livro-texto.

14.2 Teoria

Falta passar os slides para livro-texto.

Ler os slides para a teoria.

14.3 Reconhecimento de faces com componentes principais

O arquivo `DadosYaleFaces.rar` contem diretórios com arquivos de fotos de 15 indivíduos, um diretório para cada indivíduo. Dentro deles, temos 11 fotos de cada indivíduo. As fotos variam de acordo com aspectos tais como iluminação, expressão (sorrindo, sério, triste), pela presença de óculos ou não. A Figura 14.1 exibe uma amostra dessas fotos. Cada linha de fotos corresponde a um indivíduo. As imagens são normalizadas para alinhar os olhos e bocas. Eles aparecem mais ou menos no mesmo local na imagem.

Vamos fazer uma análise das fotos via componentes principais com o objetivo de reconhecer estes rostos. Imagine que você tenha uma base de dados com várias fotos de um conjunto de indivíduos. Você vai especificar um sistema de vigilância para uma companhia e apenas os 15 indivíduos destas fotos podem entrar num certo local. A checagem é feita automaticamente com uma nova foto tirada no momento da tentativa de entrada. Vamos usar o PCA para criar um sistema para classificar esta nova foto a uma das 15 classes representadas pelos diferentes indivíduos. Assim,o problema é: Chega uma nova foto. Queremos encontrar o rosto mais parecido com a nova foto no banco de dados. Se o rosto mais parecido não estiver próximo o suficiente da nova foto, a entrada não é permitida.

O método das *autofaces* foi proposto por [34] and [35]. Ele é principalmente um método de redução de dimensionalidade, podendo representar muitos indivíduos com um conjunto relativamente pequeno de dados. A idéia é representar a foto de um rosto como uma soma de um rosto médio

Figure 14.1: Amostra de 5 fotos de 4 indivíduos.

mais uma combinação linear de um pequeno número de pseudo-fotos, que são as autofaces. Estas autofaces são fotos embaçadas que capturam aspectos importantes da composição de uma face.

Podemos imaginar cada foto sendo aproximadamente obtida como representado na Figura 14.2. As fotos à esquerda são aproximadamente iguais a uma mesma foto média (a primeira do lado direito) mais quatro autofaces, cada uma delas multiplicada por um peso w_{ij} que é específico do indivíduo. Diferentes indivíduos vão variar apenas nos 4 pesos w_{ij} que cada autoface recebe. As autofaces são fixas e as mesmas para todos os indivíduos considerados, bem como a face média, ue também é a mesma para todos.

Neste exercício, você deve reproduzir a análise abaixo em `R`. Primeiramente, leia as as fotos no `R`. Vou usar o pacote `imager` que fornece algumas funcionalidades para processamento de imagens no `R`. Informações em `http://dahtah.github.io/imager/gettingstarted.html`. O pacote `imager` is baseado em CImg, uma biblioteca em C++ criada por David Tschumperle (CNRS). `imager` está no CRAN e pode ser instalada a partir da linha de comando (se conectado a web). Depois de instalar, carregue o pacote para a sessão de trabalho.

```
install.packages("imager")
library(imager)
```

A função `load.image` lê imagens em arquivos ou URLs. Formatos de imagens que têm suporte atualmente são os seguintes: JPEG, PNG e BMP. Outros formatos pedem a instalação de `ImageMagick`. Vamos mudar o diretório de trabalho para aquele que contem as imagens. usando `setwd("dir_com_imagens")`. A leitura da foto `s5.jpg` no diretório `Faces3` é feita com o comando `load.image`. A seguir, visualize a foto e imprima uma informação básica sobre a foto:

```
im <- load.image('Faces3/s5.jpg')
plot(im)   # O eixo vertical corre na direcao contraria ao usual
im
# A saida eh o que vai abaixo:
# Image. Width: 100 pix Height: 100 pix Depth: 1 Colour channels: 3
```

O objeto `im` é um objeto do tipo imagem com 100×100 pixels (width e height). Depth indica quantos frames tem a imagem. Se depth > 1 então a imagem é um vídeo. `im` possui três canais de cores, o usual sistema RGB (red-green-blue channels).

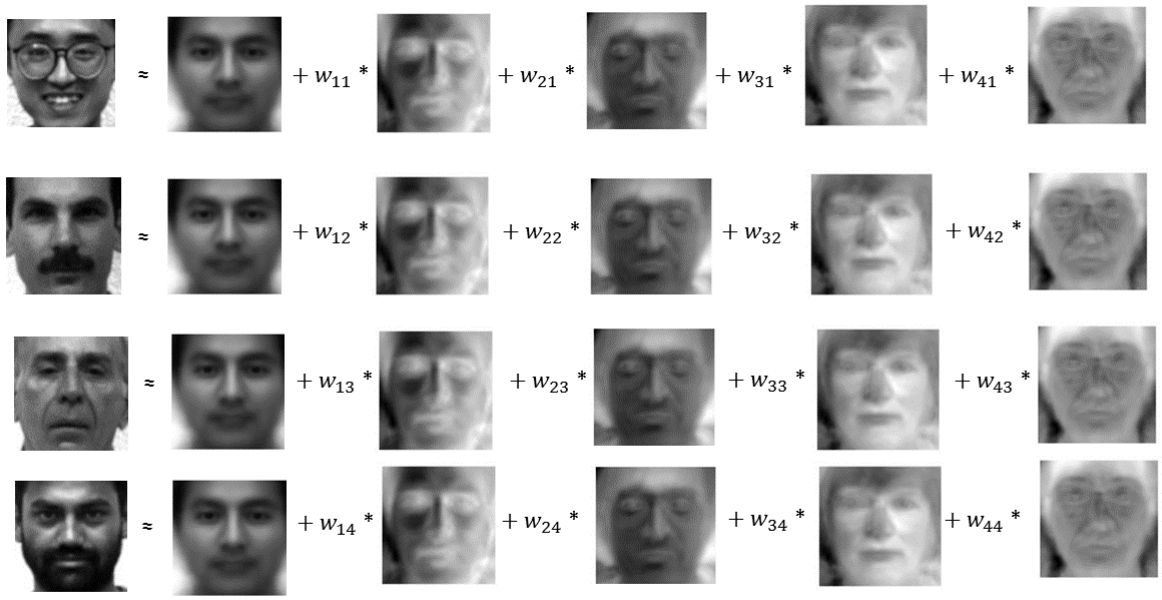


Figure 14.2: Uma foto (a esquerda) é aproximandamente a soma de uma foto média mais quatro autofaces multiplicada por pesos w_{ij} específicos do indivíduo. Diferentes indivíduos vão variar apenas nos 4 pesos w_{ij} que cada autoface recebe.

O comando `is.array(im)` retorna T mostrando que, na prática, `im` é apenas um array 3-dim para permitir a leitura de fotos coloridas com os 3 canais rgb. Cada slice do array armazena uma das cores-canais fundamentais (3 canais, red-green-blue ou rgb channels com intensidades em cada pixel). Entretanto, os 3 slices são idênticospois nossas imagens não são coloridas. Elas são apenas pixels com diferentes intensidades de cinza. Vamos checar que os 3 slices do array 3-dim sao os mesmos verificando apenas uma pequena parte:

```
im[1:5, 1:5, 1] == im[1:5, 1:5, 2]
# Como os 3 slices do array sao identicos, reduzimos a uma matriz 2-dim
im_mat = im[, , 1]
is.matrix(im_mat)
dim(im_mat) # Agora temos uma matriz 100 x 100
plot(im_mat) # im_mat ja nao eh mais uma imagem
image(im_mat) # heat map da matriz im_mat.
# O comando image eh do R basico e faz um heat map de matriz numerica
# Histograma dos tons de cinza dos 100*100 pixels da matriz im_mat
hist(im_mat)
# O tom de cinza eh um real entre 0 e 1 (ao inves do inteiro de 0 a 255)
# Uma amostra aleatoria de 5 desses pixels.
sample(im_mat,5)
```

Vamos agora ler todas as fotos e começar de fato a análise. Vamos declarar uma lista para receber as fotos de todos os individuos: `fotos = list()`. Esta é uma lista de tamanho 15 e cada elemento desta lista é uma outra lista que contem 11 fotos de um mesmo indivíduo. As fotos estão em 15 diretórios, um para cada indivíduo, e nomeados como `Faces1, Faces2`, ...,

Faces15. Dentro de um diretório Facesi encontramos as 11 fotos, em formato .bmp e .jpg, nomeadas s1.bmp, s2.bmp, ..., s11.bmp e s1.jpg, s2.jpg, ..., s11.jpg. Cada uma dessas fotos correspondem a diferentes poses de um mesmo indivíduo. Vamos ver como fazer a leitura das fotos e armazenamento na lista fotos.

```
fotos = list()
for(i in 1:15){
  fotos[[i]]=list() # elemento i da lista individuos eh uma lista tambem
  for(j in 1:11){
    # Leitura dos arquivos de fotos do individuo i
    # Sao 11 fotos no diretorio Faces#i
    fotos[[i]][[j]] <- load.image(sprintf("Faces%i/s%i.jpg",i,j))
  }
}
plot(fotos[[9]][[7]]) # Exibe a foto 7 do individuo 9
```

Vamos ver algumas fotos aleatórias. Cada linha de fotos será um indivíduo distinto. Para que todas as fotos caibam na janela gráfica, vamos eliminar o espaço deixado (como default) nas margens dos gráficos. Para isto, vamos alterar os parâmetros gráficos. Mas antes, vamos guardar uma cópia dos parâmetros gráficos default para restaurá-los no final:

```
opar <- par()  # salve parametros graficos
## elimine os espacos brancos nas margens e prepare a janela
## grafica para receber 4*5=20 fotos
par(mfrow=c(4,5), mar=c(0,0,0,0))
## plot as 5 primeiras fotos dos individuos 1, 3, 8 e 10
for(i in c(1, 3, 8, 10)){
  for(j in 1:5) plot(fotos[[i]][[j]], axes=F)
}
## restaure as opcoes graficas default
par(opar)
```

Para a análise de componentes principais, vamos converter as fotos em matrizes e colocá-las numa lista fotosmat:

```
fotosmat = list()
for(i in 1:15){
  fotosmat[[i]]=list()
  for(j in 1:11){
    fotosmat[[i]][[j]] = fotos[[i]][[j]][ , , 1]
  }
}
# Checando se todas as 11 fotos do individuo 5 sao matrizes
sapply(fotosmat[[5]], is.matrix)
# todas as 11 matrizes-fotos sao de dimensao 100 x 100
sapply(fotosmat[[5]], dim)
```

Vamos empilhar as colunas de cada matriz formando uma única coluna. Com isto, perdemos a dimensão espacial dos pixels da imagem. A seguir, coletamos as colunas numa grande matriz. O comando stack faz isto mas ele funciona apenas com dataframes, não funciona com matrizes. Assim, transformamos a matriz em um dataframe e usamos o comando stack para empilhar.

O comando `stack` retorna uma matriz com *duas* colunas: numa, ficam os valores das colunas empilhados; na outra, ficam os índices das colunas na matriz original. Procure entender usando com este exemplo simples: `stack(as.data.frame(matrix(1:6, ncol=2)))`. Assim, precisamos da *primeira* coluna da saida de `stack`. Montamos a matriz com os vetores empilhados.

```
mat_pixels = matrix(0,nrow=(100*100), ncol=11*15)
for(i in 1:15){
  for(j in 1:11){
    mat_pixels[,j+(i-1)*11] = stack(as.data.frame(fotosmat[[i]][[j]]))[,1]
  }
}
```

Vamos separar uma foto de cada indivíduo para fazer sua classificação. Isto vai constituir o conjunto de teste, onde vamos avaliar o nosso método de classificação como se estas fotos separadas fossem novas fotos. Ficaremos com uma matriz com $150 = 15 \times 11 - 15$ colunas pois existem 15 indivíduos com 11 fotos cada um. Vamos escolher uma foto ao acaso de cada indivíduo. Comece fixando a semente de números aletórios:

```
set.seed(123)
## indice do numero da foto, dentro de cada individuo
ind = sample(15, 1:11, replace=T)

## pegando agora os indices das colunas de cada foto em mat_pixels
indcol = ind + ((1:15) - 1) *11

# separando as fotos para teste posterior. Elas estao numa matriz
# com 15 colunas ordenadas de acordo com o indice dos individuos.
mat_teste = mat_pixels[ , indcol]
dim(mat_teste)

## Retirando as colunas de teste da matriz onde vamos aplicar PCA:
mat_pixels = mat_pixels[,-indcol]
```

Precisamos centrar todas as fotos do conjunto de treinamento subtraindo de cada foto a foto média de todo o conjutnode fotos. Esta foto média é simplesmente a "foto" obtida tirando a média aritmética sobre o conjunto de fotos em cada pixel. Isto é, para um pixel localizado numa certa posição, tiramos a média de todos os valores observados naquela posição nas diferentes fotos do conjunto de treinamento. Em `R`, isto é muito simples:

```
mat_media = apply(mat_pixels, 1, mean)
mat_centrada = mat_pixels - mat_media
```

Vamos desempilhar esta foto média e visualizá-la. Para desempilhar, nós quebramos a coluna `mat_centrada` em 100 colunas de tamanho 100, gerando uma "imagem" com a mesma quantidade de pixels que as fotos originais.

```
foto_media = as.cimg(mat_media, x=100, y=100)
par(mfrow=c(1,1))
plot(foto_media, axes=F)
```

Obtemos agora os componentes principais da matriz *transposta* `t(mat_centrada)` com 150 itens e 10000 atributos. A matriz Σ de covariância dos atributos (os pixels, neste caso) é uma matriz de dimensão 10000×10000. Devemos portanto obter 10000 autovetores e autovalores. Vamos usar a função `princomp` do `R`.

```
pca_pixels = princomp(t(mat_centrada))

Error in princomp.default(t(pca_pixels = princomp(mat_centrada))) :
  'princomp' can only be used with more units than variables
```

A matriz `mat_centrada` possui mais atributos-colunas ($p = 10000$) que itens-linhas ($n = 150$). Nesta situação, `princomp` não funciona. O comando sabe que a matriz de covariâancia empírica de dimensão $p \times p$ baseada numa matriz de dados de dimensão $n \times p$ possui posto igual ou menor ao mínimo entre n e p. Isto significa que a matriz de covariância da matriz de dads 150×10000 é de dimnsão 10000×10000 e com posto150. Isto implica que existem $10000 - 150$ autovalores exatamente nulos e o algoritmo de `princomp` não vai rodar. Uma saída simples é usar outro comando, `prcomp`, que usa a decomposicao do valor singular (SVD) e não se incomoda com as dimensões da matriz de dados. Uma vantagem adicional é que o algoritmo SVD paraencontrar autovaloes e autovetores é mais estável numericamente e deveria ser preferido mesmo quando a matriz tem mais linhas-itens que colunas-atributos.

```
pca_pixels = prcomp(t(mat_centrada))
summary(pca_pixels)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7
Standard deviation     8.5911 7.9668 6.8476 4.96998 4.10745 3.83001 3.27323
Proportion of Variance 0.2041 0.1755 0.1297 0.06831 0.04666 0.04057 0.02963
Cumulative Proportion  0.2041 0.3796 0.5093 0.57759 0.62425 0.66481 0.69444
                           PC8    PC9    PC10    PC11    PC12    PC13    PC14
Standard deviation     3.00237 2.8962 2.52647 2.35629 2.23969 2.14522 2.06362
Proportion of Variance 0.02493 0.0232 0.01765 0.01535 0.01387 0.01273 0.01178
Cumulative Proportion  0.71937 0.7426 0.76022 0.77557 0.78944 0.80217 0.81394
                          PC15    PC16    PC17    PC18    PC19    PC20    PC21
Standard deviation     1.84285 1.83284 1.74649 1.66961 1.63947 1.58556 1.48814
Proportion of Variance 0.00939 0.00929 0.00844 0.00771 0.00743 0.00695 0.00612
Cumulative Proportion  0.82334 0.83263 0.84106 0.84877 0.85620 0.86315 0.86928
...
```

Veja a variância explicada por cada um dos componentes. Os 10 primeiros PCs explicam 76% da variação total. Acrescentar mais 10, ficando com 20 PCs, leva a 86% A saída do comando `prcomp` mostra apenas os 150 primeiros componentes pois os outros $10000 - 150$ componentes principais têm autovalor igual a zero. Os 150 primeiros autovetores, de dimensão 10000 cada um, formam as colunas da matriz `pca_pixels$rot`.

```
dim(pca_pixels\$rot)

## Grafico scree com os 10 primeiros autovalores (ou
## variancias dos 10 1os PCAs)
plot(pca_pixels)
```

```
## Como sao muitos autovalores, defaultdo plot usa apenas os 10 1os
## Para ver TODOS os 150, podems extrai-los do objeto sdev
autovalores = (pca_pixels$sdev)^2

## Barplot das variancias acumuladas indicando a escolha de poucos
## PCAs devem representar bem os 10000 atributos (pixels)
barplot(cumsum(autovalores))

## Vamos normalizar o eixo vertical dividindo pela soma total
aux = cumsum(autovalores)/sum(autovalores)
barplot(aux)

## Procurando ver a parte inicial do grafico com mais detalhes
barplot(aux[1:30], ylim=c(0,1))

## Vamos usar os 20 primeiros autovetores.
## Eles sao vetores de dimensao 10000 = 100 * 100

autovetores = pca_pixels$rot[ , 1:20]
```

Vamos criar agora as autofaces. Vamos desempilhar os autovetores quebrando a coluna de cada um deles em 100 colunas de tamanho 100 cada uma. Com isto, cada autovetor dá origem a uma "imagem" com a mesma quantidade de pixels que as fotos originais. Vamos chamar estas pseudo-fotos de *autofaces*

```
## Criando as autofaces
auto_face=list()

for(i in 1:20){
  auto_face[[i]] = as.cimg(pca_pixels\$rot[,i], x=100, y=100)
  #100x100, grayscale image
}

## Vendo estas 20 autofaces
par(mfrow=c(4,5), mar=c(0,0,0,0))
for(i in 1:20) plot(auto_face[[i]], axes=F)
```

Vamos visualizar uma face em `mat_teste` e sua aproximação usando as k primeiras autofaces. Para isto precisamos escrever uma foto como aproximadamente uma soma da foto média mais uma combinação linear das primeiras k autofaces (todos como vetores). Em seguida, desempilhamos os vetores e mostramos as faces.

Por exemplo, usando a face na coluna 5 de `mat_teste`. Vamos obter os coeficientes b_k da combinação linear

$$mboxfoto \approx \text{fotomedia} + b_1\mathbf{v}_1 + b_2\mathbf{v}_2 + \ldots + b_k\mathbf{v}_k$$

Sabemos que b_j é o produto interno dos vetores formados pela foto centrada com o autovetor $\mathbf{v}_k$. Então

```
coef = t(autovetores) %*% (mat_teste[, 5] - mat_media)
dim(coef)
# [1] 20  1
```

A matriz `coef` é uma *matriz* 20 x 1 com os coeficientes $(b_1, b_2, \ldots, b_{20})$. Este vetor-coluna é a representação da foto no espaço gerado pelos 20 primeiros autovetores. Nesta base de autovetores, o vetor `coef` representa a foto aproximadamente. Os comandos abaixo geram uma aproximacao da foto 5 em `mat_teste` usando 2, 3, 4 até 20 autovetores resultando na Figura 14.3.

Figure 14.3: Um indivíduo e sucessivas aproximações usando 2, 3, até 20 autovetores.

```
foto_teste5 = list()
foto_teste5[[1]] = as.cimg(mat_teste[,5], x=100, y=100)
for(i in 2:20){
  aprox_vetor = mat_media + autovetores[, 1:i] %*% coef[1:i,1]
  foto_teste5[[i]] = as.cimg(as.numeric(aprox_vetor), x=100, y=100)
}

## guarde uma copia dos parametros graficos default
opar <- par()
## elimine os espacos brancos nas margens e prepare para 4*5=20 fotos
par(mfrow=c(4,5), mar=c(0,0,0,0))
## plot as fotos
for(i in 1:20) plot(foto_teste5[[i]], axes=F)
## restaure as opcoes graficas default
par(opar)
```

A primeira imagem é a imagem real. As seguintes fornecem as aproximações usando sucessivamente 2, 3, até 20 autofaces. Parece que usar 20 autofaces talvez seja excessivo. Visualmente, não existe uma diferença relevante entre a aproximação obtida usando apenas os 12 primeiros autovetores ou aquela com 20 autovetores. Apesar disto, vamos repetir esta aproximação usando 20 autofaces para todas as fotos-colunas em `mat_teste`, salvando os coeficientes numa matriz `coef` de dimensão 20×15:

```
coef = t(autovetores) %*% (mat_teste - mat_media)
dim(coef)
# [1] 20 15
```

Como classificar cada uma destas 15 fotos em uma das categorias disponíveis, as categorias sendo os 15 indivíduos? Qual será a taxa de acerto deste sistema de classificação? Vamos primeiro obter a representação de cada uma das 150 fotos de treino nos 20 PCAs, exatamente como fizemos com a fotos de teste

```
coef_treino = t(autovetores) %*% (mat_pixels - mat_media)
## coef_treino eh matriz 20 x 150
```

Podemos ver como as 150 fotos do treino ficam dispostas no plano determinado apenas pelas duas primeiras componentes principais. Supostamente, os dois primeiros componentes principais não deve ser suficiente para discriminar bem os indivíduos.

```
par(mfrow=c(1,1))
colface = rainbow(15)[rep(1:15,rep(10,15))]
plot(coef_treino[1,], coef_treino[2,], pch=21, bg=colface)
```

Vamos obter uma representação *média* de cada um dos 15 indivíduos. Vamos tirar a média dos coeficientes das 10 fotos de cada indivíduo.

```
coefmedio = matrix(0, ncol=15, nrow=20)
for(i in 1:15){
  coefmedio[,i] = apply(coef_treino[ ,(1+(i-1)*10) : (i*10)], 1, mean)
}
## Esta eh uma matriz 20 x 15

## Vamos ver como os 15 perfis medios do treino no plano
## determinado apenas pelas duas primeiras componentes principais

par(mfrow=c(1,1))
colface = rainbow(15)
plot(coefmedio[1,], coefmedio[2,], pch=21, bg=rainbow(15))
```

Cada foto de `mat_teste`, na representação dos 20 primeiros PCAs, é uma coluna da matriz `coef` que tem dimensão 20×15. A matriz `coefmedio` também é de dimensão 20×15. Para cada foto (isto é, cada coluna de `coef`), vamos encontrar a coluna de `coefmedio` que é a mais próxima. A ordem desta coluna mais próxima é o indivíduo mais próximo.

```
indproximo = numeric()
for(j in 1:15){
  indproximo[j] = which.min( apply((coefmedio - coef[,j])^2, 2, mean) )
}

indproximo
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
## Voila!! Classificacao perfeita!!!
```

O vetor `indproximo` indica que a primeira foto em `mat_teste` (coluna 1) foi classificada ao primeiro indivíduo, o que é a decisão correta. Ele também indica que a segunda foto em `mat_teste` (coluna 2) foi classificada ao segundo indivíduo, o que também é correto. Olhando o resto do vetor,fica claro que a classificação foi a correta para todas as 10 fotos do conjunto de teste.

Em conclusão, uma versão aproximada de uma foto em escala de cinza de um rosto humano pode ser obtida como uma combinação linear de umas poucas autofaces (*eigenfaces*, em inglês):

$$\text{foto} = \text{média geral} + c_1\mathbf{v}_1 + \ldots c_k\mathbf{v}_k$$

Os autovetores ou PCAs $\mathbf{v}_1, \ldots, \mathbf{v}_k$ da matriz de covariância Σ da distribuição conjunta dos pixels formam as autofaces. É impressionante que apenas algumas poucas, as primeiras k, autofaces ou PCAs sejam suficientes para obter uma boa semelhança dos rostos da maioria das pessoas. As autofaces se parecem com um rosto humano médio, sem muitos traços distintivos.

O método de autofaces foi proposto por [34] and [35]. Ele é principalmente um método de redução de dimensionalidade, podendo representar muitos indivíduos com um conjunto relativamente pequeno de dados. De fato, no nosso problema, temos uma base de 10 fotos de 15 indivíduos, cada foto armazenada como uma matriz com 10^4 elementos. Assim, temos, ao todo, $10 \times 15 \times 10^4 = O(10^6)$ bytes para armazenar. Vamos imaginar uma base maior com n indivíduos implicando em $O(n10^5)$. Ao chegar uma nova foto precisamos compará-la com as $10n$ fotos por um procedimento ingênuo (naive). Com as autofaces, guardamos apenas 20 pseudo-fotos, as autofaces (possivelmente, apenas 12 seriam suficientes), e os coeficientes médios de cada indivíduo. Veja que o número de autofaces não varia muito com n. Mesmo que n seja muito grande, teremos apenas um número pequeno de autofaces. Supondo que sejam 20 autofaces, isto significa guardar $(20 \times 10^4) + (20 \times n)$. Por exemplo, com $n = 1000$ e 10 fotos para cada um e com 20 autofaces teríamos que guardar e manipular 2.2×10^5 enquanto que uma comparação ingênua requer armazenar 10^8, ou 1000 vezes mais espaço.

No entanto, o método de autofaces pode ter um desempenho muito ruim se existir muita diferença entre as imagens na base de treinamento e as novas imagens [25]. Uma imagem de um indivíduo sob iluminação frontal pode ter coeficientes muito diferentes do mesmo indivíduo, na mesma pose, sob iluminação lateral intensa. Assim, as novas fotos devem ter representantes similares na base de treino.

Autofaces é uma técnica antiga e já existem muitas variações e melhorias, bem como outra abordagens completamente diferentes para o mesmo problema. Para quem quiser se aprofundar sobre as muitas outras técnicas envolvidas com o reconhecimento de faces, visite o website `http://www.face-rec.org/general-info/`.

■ **Example 14.1 — Exercício: Reconhecimento de dígitos.** Este exercício é praticamente a mesma coisa que foi feita acima para as faces. Ele foi extraído da página web do livro página do livro *The Elements of Statistical Learning*, por Hastie, Tibshirani e Friedman. Este excelente (e avançado) livro está disponível para download gratuito e legal na página `http://statweb.stanford.edu/~tibs/ElemStatLearn/`.

O objetivo é construir um algoritmo em `R` para a classificação de dígitos escritos á mão. Os dados são uma parte da base *US Postal Service Database* e correspondem à digitalização de números de CEP escritos a mão em correspondências enviadas pelo correio americano. Estes dados estão na página do livro, onde é chamado de `ZIP code` (é o último da lista de datasets).

O conjunto de dados refere-se a dados numéricos obtidos a partir da digitalização de dígitos escritos à mão a partir dos envelopes pelo Serviço Postal dos EUA. Imagens em preto e branco foram normalizadas em termos de seu tamanho de forma a caber em uma caixa de pixels 20×20, preservando a sua razão de aspecto (aspect ratio). As imagens resultantes contêm níveis de cinza como um resultado da técnica de anti-aliasing usada pelo algoritmo de normalização. As imagens foram centradas em uma imagem 28×28 calculando o centro de massa dos pixels e traduzindo a imagem de modo a posicionar este ponto no centro da matriz 28×28. O resultado final são imagens 28×28 em tons de cinza.

A Figura 14.4 mostra os dígitos 4 da base de dados. O objetivo do exercício é inteiramente análogo ao de reconhecimento de faces. Queremos um método de classificação de novas imagens

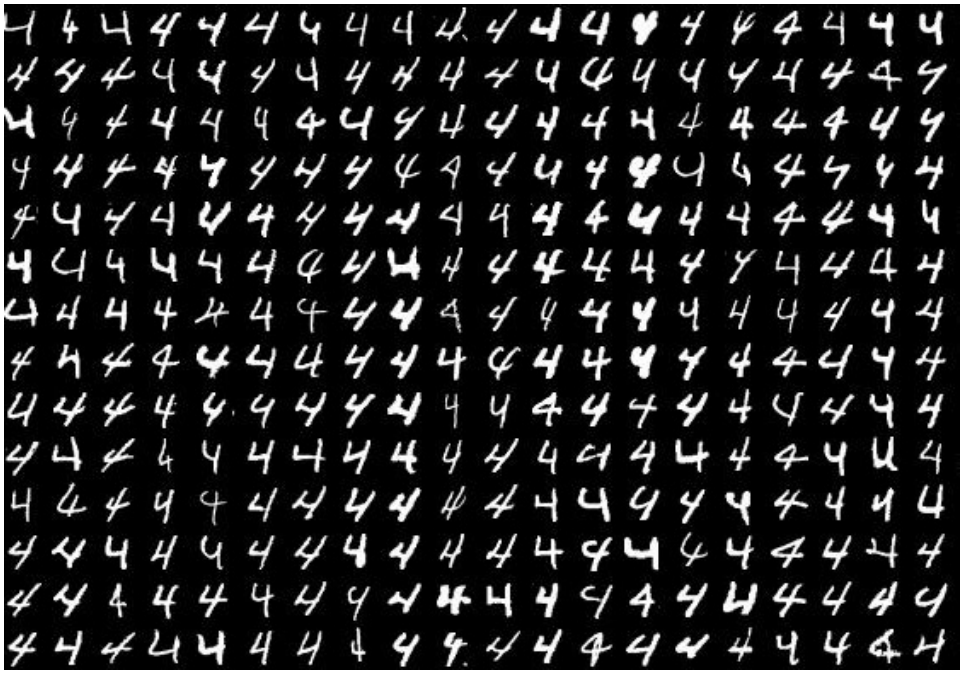


Figure 14.4: Imagens dos dígitos 4 da base USPS.

k	precisão média	revocação média
5	??	??
6	??	??
⋮	⋮	⋮
20	??	??

Table 14.1: Precisão e revocação do método de classificação de dígitos como função do número k de autovetores.

de dígitos manuscritos. Assim, você deverá:

- Usando um conjunto de treinamento, criar uma regra de classificação de novas imagens de dígitos. Use os primeiros k autovetores da matriz de covariância entre os pixels para fazer esta regra de classificação. Você deve fazer seus cálculos com $k = 5, 10, 15, 20$.
- Usando apenas a amostra de TESTE, crie uma tabela de contingência 10×10 de confusão C. Nesta matriz C as linhas representam a classe verdadeira do dígito (de 0 a 9) e a coluna a classe em que ele foi alocado. Na entrada C_{ij} você deve colocar o número de itens (ou imagens) que caíram naquela categoria cruzada. Crie esta tabela com os quatro valores distintos de $k = 5, 10, 15, 20$.
- Calcule a proporção total das imagens da amostra de teste que caem na diagonal principal. Esta é uma medida global de classificação correta do método. Para qual valor de k esta proporção foi máxima?
- Preencha uma tabela como a que está abaixo:
  Precisão média é a média aritmética da precisão das 10 classes e definida como:

$$pm = \frac{1}{10} \sum_{i=0}^{9} \frac{C_{ii}}{C_{i+}}$$

  com C_{i+} sendo a soma da linha i na matriz de confusão. Revocação média é a média

aritmética da revocação das 10 classes e definida como:

$$rm = \frac{1}{10}\sum_{i=0}^{9}\frac{C_{ii}}{C_{+i}}$$

com C_{+i} sendo a soma da coluna i na matriz de confusão. Mais detalhes sobre precisão (precision) e revocação (recall) podem ser vistos no verbete *Precision and recall* na wikipedia. Ver também `http://www.text-analytics101.com/2014/10/computing-precision-and-recall-for.html`.

■

15. Factor Analysis

A melhor maneira de apresentar o modelo fatorial é começar com um exemplo. Suponha que estamos estudando o desempenho escolar de uma população através das notas ou escores em um conjunto de 16 disciplinas. Cada estudante possui um vetor $\mathbf{X}$ contendo as suas notas em 16 assuntos. As coordenadas do vetor $\mathbf{X} = (X_1, \ldots, X_{16})$ são as seguintes:

X_1 = Gramática	X_9 = Lógica
X_2 = Literatura	X_{10} = Química
X_3 = Redação	X_{11} = Física
X_4 = História	X_{12} = Biologia
X_5 = Geografia	X_{13} = Inglês
X_6 = Filosofia	X_{14} = Sociologia
X_7 = Álgebra	X_{15} = Espanhol
X_8 = Geometria	X_{16} = Educação Física

O desempenho individual nestes 16 assuntos é muito variado:

Alguns vão bem em todas as disciplinas; Alguns vão bem apenas em algumas e tem um desempenho médio nas outras; Alguns vão muito bem em algumas e muito mal em outras; Alguns vão mal em todas.

Como entender essa diversidade?

Psicólogos se perguntaram se esta variabilidade na capacidade cognitiva não poderia ser explicada pela existência de uns poucos traços latentes, não observados.

Por exemplo,

- FV = Fator refletindo habilidade verbal
- FQ = Fator refletindo habilidade lógico-quantitativa

Os fatores tem uma escala centrada em zero.

Se

$F = 0 \Rightarrow$ Indivíduo tem habilidade média no fator.

$F > 0 \Rightarrow$ Habilidade acima da média.

$F \gg 0 \Rightarrow$ Habilidade muito acima da média.

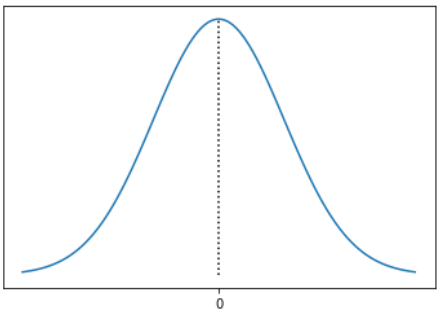
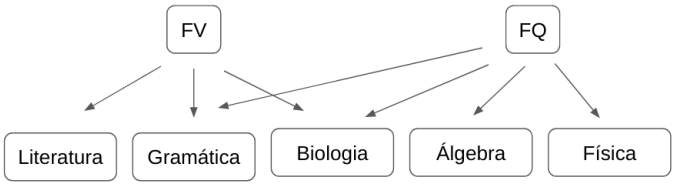


Figure 15.1: Densidade da distribuição de um fator na população.



$F < 0 \Rightarrow$ Habilidade abaixo da média.

Cada indivíduo recebe uma "dose" de FV e uma "dose" de FQ. Doses de FV e FQ são independentes. Por exemplo, alguns recebem muito de FV. Dentre estes, metade recebe $FQ+$. A outra metade recebe $FQ-$

As notas nos 15 assuntos são reflexos e combinações desses dois fatores, além de um ruído causado por outros fatores que não levamos em conta.

Modelo hierárquico para a geração das 15 notas de um mesmo indivíduo.

- Passo 1: Recebe doses independentes de FV e FQ. $FV-$ $FQ+$.
- Passo 2: Estas doses afetam as notas dos 15 assuntos.

Ausência de aresta $\Rightarrow$ sem influência do fator.

Peso ou importância da aresta representada por grossura da aresta.

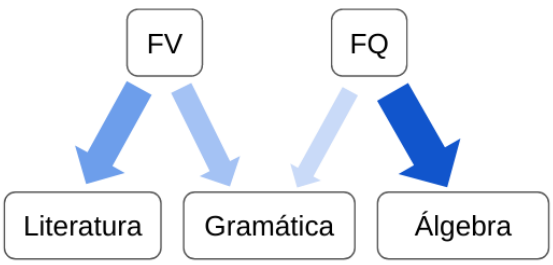
Peso da aresta é a carga do fator (factor loading).

15.0.1 Representação algébrica da análise fatorial

- Gramática = $X_1 \approx \mu_{Gram} + \ell_{GV} * FV + \ell_{GQ} * FQ$
- Literatura = $X_2 \approx \mu_{Lit} + \ell_{LV} * FV + \ell_{LQ} * FQ$
- $\vdots$

Carga dos fatores.

- Gramática = $X_1 \approx \mu_{Gram} + \ell_{GV} * FV + \ell_{GQ} * FQ$



- Literatura = $X_2 \approx \mu_{Lit} + \ell_{LV} * FV + \ell_{LQ} * FQ$
- $\vdots$

Fator verbal do indivíduo. O mesmo para todos os assuntos.
Mais formalmente:

- Gramática = $X_1 = \mu_{Gram} + \ell_{GV} * FV + \ell_{GQ} * FQ + \varepsilon_{Gram}$
- Literatura = $X_2 = \mu_{Lit} + \ell_{LV} * FV + \ell_{LQ} * FQ + \varepsilon_{Lit}$
- $\vdots$
- Espanhol = $X_{15} = \mu_{Esp} + \ell_{EV} * FV + \ell_{EQ} * FQ + \varepsilon_{Esp}$

Erros ou fatores não observados.
Um exemplo esquemático:

- Assuntos muito associados com FV:
  - $X_1 = \mu_1 + (1.0) * FV + (0.1) * FQ + \varepsilon_1$
  - $X_2 = \mu_2 + (0.8) * FV + (0.01) * FQ + \varepsilon_2$
- Nenhuma das duas habilidades é muito relevante:
  - $X_3 = \mu_3 + (0.03) * FV - (0.02) * FQ + \varepsilon_3$
- Precisa ser bom em FQ e muito ruim em FV (difícil de pensar num exemplo)
  - $X_4 = \mu_4 - (0.3) * FV + (0.9) * FQ + \varepsilon_4$
- Bom em FQ e pouco relevante em FV.
  - $X_5 = \mu_5 - (0.1) * FV + (1.2) * FQ + \varepsilon_5$

Representação matricial para *um único* indivíduo:

$$\mathbf{X} = \underset{\sim}{\mu} + \underset{pxm}{L} \cdot \underset{mx1}{\underset{\sim}{F}} + \underset{px1}{\underset{\sim}{\varepsilon}}$$

L = matriz de carga dos fatores (loading).
F = vetor dos fatores comuns (às p variáveis).
ε = vetor dos erros ou fatores específicos (de cada variável).

Vamos imaginar dois indivíduos com seus dois vetores instanciados:

$$\mathbf{X}^{(1)} = \begin{pmatrix} X_1^{(1)} \\ . \\ . \\ X_{15}^{(1)} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ . \\ . \\ \mu_{15} \end{pmatrix} + \begin{bmatrix} \ell_{11} & \ell_{12} \\ . & . \\ . & . \\ \ell_{15,1} & \ell_{15,2} \end{bmatrix} \begin{bmatrix} FV^{(1)} \\ FQ^{(1)} \end{bmatrix} + \begin{bmatrix} \varepsilon_1^{(1)} \\ . \\ . \\ \varepsilon_{15}^{(1)} \end{bmatrix}$$

$$\mathbf{X}^{(2)} = \begin{pmatrix} X_1^{(2)} \\ . \\ . \\ X_{15}^{(2)} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ . \\ . \\ \mu_{15} \end{pmatrix} + \begin{bmatrix} \ell_{11} & \ell_{12} \\ . & . \\ . & . \\ \ell_{15,1} & \ell_{15,2} \end{bmatrix} \begin{bmatrix} FV^{(2)} \\ FQ^{(2)} \end{bmatrix} + \begin{bmatrix} \varepsilon_1^{(2)} \\ . \\ . \\ \varepsilon_{15}^{(2)} \end{bmatrix}$$

Vamos imaginar dois indivíduos com seus dois vetores instanciados:

$$\mathbf{X}^{(1)} = \underset{\sim}{\mu} + L \cdot \underset{\sim}{F}^{(1)} + \underset{\sim}{\varepsilon}^{(1)}$$
$$\mathbf{X}^{(2)} = \underset{\sim}{\mu} + L \cdot \underset{\sim}{F}^{(2)} + \underset{\sim}{\varepsilon}^{(2)}$$

$\underset{\sim}{\mu}$ e $L \Rightarrow$ iguais para todos os indivíduos.

$\underset{\sim}{F}^{(i)} = \begin{bmatrix} FV^{(i)} \\ FQ^{(i)} \end{bmatrix}$ = doses dos fatores recebidos pelos indivíduos.
$\underset{\sim}{\varepsilon}^{(i)}$ = erros do indivíduo (i)

Vamos imaginar dois indivíduos com seus dois vetores instanciados:

$\mathbf{X}^{(1)} = \underset{\sim}{\mu} + L \cdot \underset{\sim}{F}^{(1)} + \underset{\sim}{\varepsilon}^{(1)}$

$\mathbf{X}^{(2)} = \underset{\sim}{\mu} + L \cdot \underset{\sim}{F}^{(2)} + \underset{\sim}{\varepsilon}^{(2)}$

São iguais. Específicos do indivíduo.

Só observamos o vetor $\mathbf{X}$ em vários indivíduos. Queremos entender como $\mathbf{X}$ varia. Basta entender como os fatores FV e FQ variam de indivíduo para indivíduo. As notas dos 15 assuntos apenas refletem e combinam estes fatores através da matriz L. Um pequeno ruído ε para cada disciplina é adicionado para levar em conta os demais fatores que estamos ignorando. Como podemos inferir L a partir dos dados? E os escores dos fatores $\begin{bmatrix} FV^{(i)} \\ FQ^{(i)} \end{bmatrix}$ de cada indivíduo, como obtemos?

Indivíduo (i) com $m = 2$ fatores.

$$\underset{px1}{\mathbf{X}^{(i)}} = \underset{px1}{\underset{\sim}{\mu}} + \underset{px2}{L} \cdot \underset{2x1}{\underset{\sim}{F}^{(i)}} + \underset{px1}{\underset{\sim}{\varepsilon}^{(i)}}$$

$\underset{\sim}{\mu}$ e L são comuns a todos os indivíduos, não aleatórios. $\underset{\sim}{F}^{(i)}$ e $\underset{\sim}{\varepsilon}^{(i)}$ variam de indivíduo para indivíduo, são aleatórios. $\underset{2x1}{\mathbb{E}(\underset{\sim}{F}^{(i)})} = \mathbb{E}\begin{pmatrix} FV^{(i)} \\ FQ^{(i)} \end{pmatrix} = \begin{pmatrix} \mathbb{E}(FV^{(i)}) \\ \mathbb{E}(FQ^{(i)}) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \underset{\sim}{0}$ $\underset{2x2}{Cov(\underset{\sim}{F}^{(i)})} = \begin{bmatrix} Var(FV^{(i)}) & Cov(FV^{(i)}, FQ \\ Cov(FV^{(i)}, FQ^{(i)}) & Var(FQ^{(i)}) \end{bmatrix}$

$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

15.0.2 Resumo até aqui...

Seja $\mathbf{X} \sim N_p(\mu, \Sigma)$ um vetor aleatório p-dimensional. Imagine que existem *duas* variáveis F_1 e F_2 que *não são* diretamente observáveis. Dizemos que F_1 e F_2 são variáveis *latentes*, ocultas. São chamadas também de *fatores latentes* ou, às vezes, apenas de *fatores*. Assumimos que os fatores F_1 e F_2 são variáveis aleatórias independentes. Cada uma das variáveis X_i no vetor observado $\mathbf{X}$ é aproximadamente uma combinação linear de F_1 e F_2. Isto é,

$$\begin{aligned} X_1 &\approx \mu_1 + \ell_{11}F_1 + \ell_{12}F_2 \\ X_2 &\approx \mu_2 + \ell_{21}F_1 + \ell_{22}F_2 \\ &\vdots \\ X_p &\approx \mu_p + \ell_{p1}F_1 + \ell_{p2}F_2 \end{aligned}$$

Os símbolos ℓ_{ij} indicam constantes desconhecidas. Eles são chamados de *carga dos fatores* ou, em inglês, *factor loadings*.

15.0.3 Suposições do modelo fatorial

A opção de tomar a variância de cada fator igual a 1 (e portanto, tomar o DP de cada fator = 1) é baseada no seguinte argumento:

Cada fator (FV ou FQ) terá a "mesma escala" indo de -2 a +2 aproximadamente ao variar dos menos habilidosos aos mais habilidosos. Se o fator F_k afetar muito uma nota X_j isto será refletido numa carga ℓ_{jk} muito positiva (ou muito negativa). Mas a escala de todos os fatores é a mesma (em DP's): vai de -2 a +2, aproximadamente.

A covariância $Cov(FV^{(i)}, FQ^{(i)}) = 0$ pois estamos supondo fatores independentes.

Mais suposições, agora sobre $\underset{\sim}{\varepsilon}^{(i)}$:

$$\underset{15x1}{\mathbb{E}(\underset{\sim}{\varepsilon}^{(i)})} = \mathbb{E}\begin{pmatrix}\varepsilon_1^{(i)}\\ .\\ .\\ \varepsilon_{15}^{(i)}\end{pmatrix} = \begin{pmatrix}\mathbb{E}(\varepsilon_1^{(i)})\\ .\\ .\\ \mathbb{E}(\varepsilon_{15}^{(i)})\end{pmatrix} = \begin{pmatrix}0\\ 0\\ .\\ .\\ 0\end{pmatrix} = \underset{\sim}{0}$$

Mais suposições, agora sobre $\underset{\sim}{\varepsilon}^{(i)}$:

$$\underset{15x15}{Cov(\underset{\sim}{\varepsilon})} = \begin{bmatrix} Var(\varepsilon_1^{(i)}) & Cov(\varepsilon_1^{(i)},\varepsilon_2^{(i)}) & . & . & Cov(\varepsilon_1^{(i)},\varepsilon_{15}^{(i)})\\ Cov(\varepsilon_2^{(i)},\varepsilon_1^{(i)}) & Var(\varepsilon_2^{(i)}) & . & . & Cov(\varepsilon_2^{(i)},\varepsilon_{15}^{(i)})\\ . & . & . & . & .\\ . & . & . & . & .\\ Cov(\varepsilon_{15}^{(i)},\varepsilon_1^{(i)}) & Cov(\varepsilon_{15}^{(i)},\varepsilon_2^{(i)}) & . & . & Var(\varepsilon_{15}^{(i)}) \end{bmatrix}$$

$$= \begin{bmatrix} \psi_1 & & & & \\ & \psi_2 & & & \\ & & . & & \\ & & & . & \\ & & & & \psi_{15} \end{bmatrix} = diag(\psi_1,...,\psi_{15}) = \psi$$

É razoável deixar que $Var(\varepsilon_j^{(i)})$ varie. Podemos ter $\psi_j \neq \psi_k$. A razão é que a nota de redação, digamos, pode ter muito mais variabilidade que a nota de matemática devido a fatores não relacionados com FV ou FQ. A subjetividade do corretor da redação, a variação da qualidade da redação como fruto do conhecimento do aluno sobre o tema, entre outras causas, pode gerar mais variação na nota da redação do que a variação induzida pela diversidade de FV e FQ.

A covariância entre os erros de assuntos distintos, $Cov(\varepsilon_j^{(i)},(\varepsilon_k^{(i)})$, provavelmente <u>não é</u> zero, mas deve ser pequena. Por isto façamos todas iguais a <u>zero</u> no modelo. Assim, adotamos $\underset{\sim}{Cov(\varepsilon)} =$

$$\begin{pmatrix} \psi_1 & & & \\ & . & \underset{\sim}{0} & \\ & & . & \\ & \underset{\sim}{0} & & \\ & & & \psi_{15} \end{pmatrix} = \text{diagonal.}$$

Uma última suposição: $\underset{\sim}{F}$ e $\underset{\sim}{\varepsilon}$ são independentes. Isso implica que:

$$\underset{15x2}{Cov(\underset{\sim}{\varepsilon},\underset{\sim}{F})} = \begin{bmatrix} Cov(\varepsilon_1,F_1) & Cov(\varepsilon_1,F_2)\\ Cov(\varepsilon_2,F_1) & Cov(\varepsilon_2,F_2)\\ . & .\\ . & .\\ Cov(\varepsilon_{15},F_1) & Cov(\varepsilon_{15},F_2) \end{bmatrix} = \begin{bmatrix} 0 & 0\\ 0 & 0\\ . & .\\ . & .\\ 0 & 0 \end{bmatrix}$$ Note que, como $\underset{15x1}{\mathbb{E}(\underset{\sim}{\varepsilon})} = \underset{\sim}{0}$ e $\underset{2x1}{\mathbb{E}(\underset{\sim}{F})} = \underset{\sim}{0}$, temos $Cov(\underset{\sim}{\varepsilon},\underset{\sim}{F}) = \mathbb{E}(\underset{\sim}{\varepsilon}\underset{\sim}{F}') = \underset{\sim}{0}$ e também $\mathbb{E}(\underset{\sim}{F}\underset{\sim}{\varepsilon}') = \underset{\sim}{0}$. Podemos agora obter a estrutura de convariância das observações $\mathbf{X}$.

$\mathbb{E}(\mathbf{X}) = \mathbb{E}(\underset{\sim}{\mu} + L\underset{\sim}{F} + \underset{\sim}{\varepsilon})$
$= \underset{\sim}{\mu} + \mathbb{E}(L\underset{\sim}{F}) + \mathbb{E}(\underset{\sim}{\varepsilon})$
$= \underset{\sim}{\mu} + L\mathbb{E}(\underset{\sim}{F}) + \mathbb{E}(\underset{\sim}{\varepsilon})$
$= \underset{\sim}{\mu} + L\underset{\sim}{0} + \underset{\sim}{0}$
$= \underset{\sim}{\mu}$ É isto mesmo, o valor esperado das notas é o vetor $\underset{\sim}{\mu}$, que representa a média da população de interesse. $\underset{15x15}{Cov(\mathbf{X})} = \Sigma = \mathbb{E}((\mathbf{X}-\underset{\sim}{\mu})(\mathbf{X}-\underset{\sim}{\mu})') = \mathbb{E}((L\underset{\sim}{F}+\underset{\sim}{\varepsilon})(L\underset{\sim}{F}+\underset{\sim}{\varepsilon})')$
$= \mathbb{E}((L\underset{\sim}{F})(L\underset{\sim}{F})' + \underset{\sim}{\varepsilon}(L\underset{\sim}{F})' + (L\underset{\sim}{F})\underset{\sim}{\varepsilon}' + \underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}')$
$= \mathbb{E}(L\underset{\sim}{F}\underset{\sim}{F}'L') + \mathbb{E}(\underset{\sim}{\varepsilon}(L\underset{\sim}{F})') + \mathbb{E}((L\underset{\sim}{F})\underset{\sim}{\varepsilon}') + \mathbb{E}(\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}')$
$= L\mathbb{E}(\underset{\sim}{F}\underset{\sim}{F}')L' + \underset{\sim}{0} + \underset{\sim}{0} + \psi$

$= L\mathbb{I}_2 L' + \psi$
$= LL' + \psi$
Isto é, $Cov(\mathbf{X}) = \Sigma = LL' + \psi$.

No nosso exemplo com 15 notas e dois fatores:

$$\underset{15x15}{\Sigma} = \begin{bmatrix} \ell_{11} & \ell_{12} \\ \ell_{21} & \ell_{22} \\ . & . \\ . & . \\ \ell_{15,1} & \ell_{15,2} \end{bmatrix} \begin{bmatrix} \ell_{11} & \ell_{21} & . & . & \ell_{15,1} \\ \ell_{12} & \ell_{22} & . & . & \ell_{15,2} \end{bmatrix} + \begin{bmatrix} \psi_1 & & & & \\ & . & & & \\ & & . & & \\ & & & . & \\ & & & & \psi_{15} \end{bmatrix} =$$

$$\begin{bmatrix} \ell_{11}^2 + \ell_{12}^2 + \psi_1 & \ell_{11}\ell_{12} + \ell_{12}\ell_{22} & \ell_{11}\ell_{31} + \ell_{12}\ell_{32} & . & . & \ell_{11}\ell_{15,1} + \ell_{12}\ell_{15,2} \\ \ell_{11}\ell_{21} + \ell_{12}\ell_{22} & \ell_{21}^2 + \ell_{22}^2 + \psi_2 & \ell_{21}\ell_{31} + \ell_{22}\ell_{32} & . & . & \ell_{21}\ell_{15,1} + \ell_{22}\ell_{15,2} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & \ell_{15,1}^2 + \ell_{15,2}^2 + \psi_{15} \end{bmatrix}$$

Outra maneira de pensar sobre Σ é perceber que, se olharmos a matriz $\underset{15x2}{L}$ como um conjunto de 15 vetores-cargas,

$$L = \begin{bmatrix} \ell_{11} & \ell_{12} \\ \ell_{21} & \ell_{22} \\ . & . \\ . & . \\ \ell_{15,1} & \ell_{15,2} \end{bmatrix} = \begin{bmatrix} \underset{\sim}{\ell}_1 \\ \underset{\sim}{\ell}_2 \\ . \\ . \\ \underset{\sim}{\ell}_{15} \end{bmatrix}$$

onde $\underset{\sim}{\ell}_j =$ cargas da disciplina j.

Então,

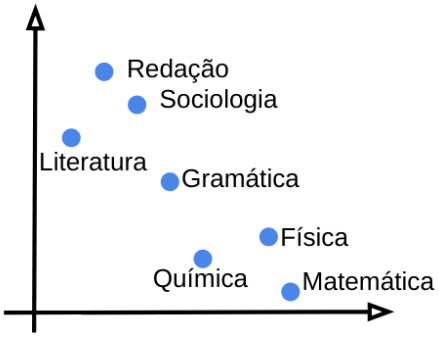
$$\underset{15x15}{\Sigma} = \begin{bmatrix} \langle \underset{\sim}{\ell}_1, \underset{\sim}{\ell}_1 \rangle + \psi_1 & \langle \underset{\sim}{\ell}_2, \underset{\sim}{\ell}_1 \rangle & \langle \underset{\sim}{\ell}_3, \underset{\sim}{\ell}_1 \rangle & . & . & \langle \underset{\sim}{\ell}_{15}, \underset{\sim}{\ell}_1 \rangle \\ \langle \underset{\sim}{\ell}_1, \underset{\sim}{\ell}_2 \rangle & \langle \underset{\sim}{\ell}_2, \underset{\sim}{\ell}_2 \rangle + \psi_2 & \langle \underset{\sim}{\ell}_3, \underset{\sim}{\ell}_2 \rangle & . & . & \langle \underset{\sim}{\ell}_{15} \underset{\sim}{\ell}_2 \rangle \\ \langle \underset{\sim}{\ell}_1, \underset{\sim}{\ell}_3 \rangle & \langle \underset{\sim}{\ell}_2, \underset{\sim}{\ell}_3 \rangle & \langle \underset{\sim}{\ell}_3, \underset{\sim}{\ell}_3 \rangle + \psi_3 & . & . & \langle \underset{\sim}{\ell}_{15} \underset{\sim}{\ell}_3 \rangle \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \langle \underset{\sim}{\ell}_1, \underset{\sim}{\ell}_{15} \rangle & \langle \underset{\sim}{\ell}_2, \underset{\sim}{\ell}_{15} \rangle & \langle \underset{\sim}{\ell}_3, \underset{\sim}{\ell}_{15} \rangle & . & . & \langle \underset{\sim}{\ell}_{15}, \underset{\sim}{\ell}_{15} \rangle + \psi_{15} \end{bmatrix}$$

$$= \begin{bmatrix} \|\underset{\sim}{\ell}_1\|^2 + \psi_1 & \langle \underset{\sim}{\ell}_2, \underset{\sim}{\ell}_1 \rangle & . & . & \langle \underset{\sim}{\ell}_{15}, \underset{\sim}{\ell}_1 \rangle \\ \langle \underset{\sim}{\ell}_2, \underset{\sim}{\ell}_1 \rangle & \|\underset{\sim}{\ell}_2\|^2 + \psi_2 & . & . & \langle \underset{\sim}{\ell}_{15} \underset{\sim}{\ell}_2 \rangle \\ . & . & . & . & . \\ . & . & . & . & . \\ \langle \underset{\sim}{\ell}_{15}, \underset{\sim}{\ell}_1 \rangle & \langle \underset{\sim}{\ell}_{15}, \underset{\sim}{\ell}_2 \rangle & . & . & \|\underset{\sim}{\ell}_{15}\|^2 + \psi_{15} \end{bmatrix}$$

Assim,

$$Var(X_i) = \Sigma_{ii} = \|\underset{\sim}{\ell}_i\|^2 + \psi_i = \ell_{i1}^2 + \ell_{i2}^2 + \psi_i$$

$\|\underset{\sim}{\ell}_i\|^2 \Rightarrow$ comumalidade
$\psi_i \Rightarrow$ variância específica Se os dois fatores latentes não possuem impacto na disciplina i (por exemplo, se a disciplina for educação física), então $\ell_{i1}^2 + \ell_{i2}^2 \approx 0$ e toda a variância da nota é devida aos fatores específicos ε_i e diferentes dos fatores latentes. Suponhamos que a disciplina X_i tenha uma carga grande do fator verbal ($\ell_{i1}^2 \gg 0$), mas uma carga pequena do fator quantitativo ($l_{i2}^2 \approx 0$). Então $Var(X_i) = \ell_{i1}^2 + \ell_{i2}^2 + \psi_i \approx \ell_{i1}^2 + \psi_i$. Toda a variabilidade das notas entre os alunos é devida às diferenças do fator verbal. Alunos apenas com o fator quantitativo FQ muito diferentes não terão notas muito distintas.



15.0.4 Interpretando as cargas dos fatores

$$L = \begin{bmatrix} \ell_{11} & \ell_{12} \\ \ell_{21} & \ell_{22} \\ . & . \\ . & . \\ \ell_{15,1} & \ell_{15,2} \end{bmatrix} = \begin{bmatrix} \underset{\sim}{\ell}_1 \\ \underset{\sim}{\ell}_2 \\ . \\ . \\ \underset{\sim}{\ell}_{15} \end{bmatrix}$$

Podemos plotar as linhas de L num gráfico planar. A primeira coordenada (fator 1) no eixo horizontal e a segunda coordenada no eixo vertical (fator 2).

Esta representação mostra que as disciplinas
Física, Química, Matemática $\Rightarrow$ possuem cargas altas no Fator 1 e cargas baixas no Fator 2.
Redação, Literatura, Sociologia $\Rightarrow$ pouca carga do Fator 1 e muita cargado Fator 2.

Isto implica que, nesta disposição de colunas da matriz L, a primeira coluna (ou a primeira coordenada das linhas) representa o fator quantitativo. A segunda coluna de L representa o fator verbal. Observe que "Gramática" ficou a meio caminho, com carga mediana nos dois fatores. Para ter nota alta em "Gramática" é preciso ter "doses" razoáveis dos dois fatores OU uma "dose" bem grande de um dos fatores, qualquer um deles.

15.0.5 Métodos de estimação

- Máxima verossimilhança
- Componentes principais

Veremos agora apenas o segundo método. O primeiro requer o conhecimento do capítulo **??** Pelo teorema espectral, $\Sigma = P \wedge P'$, onde $P = \begin{bmatrix} \underset{\sim}{y}_1 & \ldots & \underset{\sim}{y}_p \end{bmatrix}$ são os autovetores de Σ e $\wedge = \begin{pmatrix} \lambda_1 & & \\ & . & \\ & & \lambda_p \end{pmatrix}$ é matriz diagonal com os autovalores. Manipulação matricial permite escrever

$$\Sigma = L^*(L^*)' = \begin{bmatrix} \sqrt{\lambda_1}\underset{\sim}{\vec{y}}_1 & \ldots & \sqrt{\lambda_p}\underset{\sim}{\vec{y}}_p \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1}\underset{\sim}{y}'_1 \\ . \\ \sqrt{\lambda_p}\underset{\sim}{y}'_p \end{bmatrix}$$

Suponha que os últimos autovalores sejam ≈ 0. Isto implica que as últimas colunas de L^* são aproximadamente nulas e podem ser ignoradas. Mais formalmente, suponha que a soma dos k primeiros autovalores seja praticamente igual à soma de todos os p autovalores:

$$\frac{\lambda_1 + \lambda_2 + \ldots + \lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_k + \lambda_{k+1} + \ldots + \lambda_p} \approx 1$$

Ignorando as últimas colunas da matriz $\underset{pxp}{L^*}$ ficamos com uma matriz $\underset{pxk}{L}$:

$$\Sigma \approx LL' = \begin{bmatrix} \sqrt{\lambda_1}\underset{\sim}{v}_1 & \dots & \sqrt{\lambda_k}\underset{\sim}{v}_k \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1}\underset{\sim}{v}_1' \\ \cdot \\ \cdot \\ \sqrt{\lambda_k}\underset{\sim}{v}_k' \end{bmatrix}$$

Para completar o modelo fatorial, estimamos a matriz diagonal

$$\psi = \begin{pmatrix} \psi_1 & & & \\ & \cdot & & \\ & & \cdot & \\ & & & \psi_p \end{pmatrix} = diag[\Sigma - LL']$$

Isto é, $\psi_i = \Sigma_{ii} - (LL')_{ii}$

15.0.6 Resumo prático

Matriz de dados $\underset{nxp}{X}$ Obtenha $S = Cov(X)$ Alternativamente, se os s_{ii} forem muito distintos, usamos a matriz $R = cor(x)$, a matriz de correlação. Obtermos os autovalores ordenados e os autovetores de S. Calcule a soma acumulada $\frac{\lambda_1+\dots+\lambda_k}{\lambda_1+\dots+\lambda_k+\lambda_{k+1}+\dots+\lambda_p} = a_k$ Se $a_k \approx 1$ com k pequeno, então o modelo fatorial pode ser usado pois vai simplificar a estrutura dos dados.

Use os primeiro k autovetores (tal que $a_k \approx 1$) para criar a matriz de cargas

$L = \begin{bmatrix} \sqrt{\lambda_1}\underset{\sim}{v}_1 & \dots & \sqrt{\lambda_k}\underset{\sim}{v}_k \end{bmatrix}$ e $\psi = \begin{bmatrix} \psi_1 & & & \\ & \cdot & & \\ & & \cdot & \\ & & & \psi_p \end{bmatrix}$ onde $\psi_i = S_{ii} - (LL')_{ii}$ Um bom critério de escolha de k é verificar que a soma das entradas ao quadrado da matriz $(S - (LL' + \psi)) \leq \lambda_{k+1}^2 + \dots + \lambda_p^2$

Assim, se $\lambda_{k+1}^2 + \dots + \lambda_p^2 \approx 0 \Rightarrow S \approx LL' + \psi$ e o modelo fatorial é um bom ajuste.

15.0.7 Exemplos de Johnson & Wichern

15.0.8 Identificabilidade do modelo

Existe um problema de identificabilidade na determinação do modelo fatorial. O problema é que a matriz de cargas L só pode ser conhecida a menos de uma rotação. Seja $\underset{2x2}{T}$ uma matriz ortogonal. Isto é, $TT' = T'T = \mathbb{I}_2 =$ identidade $= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ De álgebra de matrizes, sabemos que matrizes de rotação ríida dos eixos coordenados são matrizes ortogonais.

Isto significa que uma matriz $\underset{2x2}{T}$ da seguinte forma:

$T = \begin{bmatrix} cos(\phi) & sin(\phi) \\ -sin(\phi) & cos(\phi) \end{bmatrix} \Rightarrow$ rotação clockwise

ou $T = \begin{bmatrix} cos(\phi) & -sin(\phi) \\ sin(\phi) & cos(\phi) \end{bmatrix} \Rightarrow$ counter-clockwise

para $\phi \in [0, 2\pi]$ é tal que $TT' = T'T = \mathbb{I}_2$. Estas matrizes correspondem a rotações no plano.

Seja $\underset{2x1}{\vec{V}} = \begin{bmatrix} x \\ y \end{bmatrix} \in \mathbb{R}^2$ e $\underset{2x2}{T} = \begin{bmatrix} cos(\phi) & sin(\phi) \\ -sin(\phi) & cos(\phi) \end{bmatrix}$ Então $\underset{2x2}{T}\underset{2x1}{\vec{V}}$ é um novo ponto no $\mathbb{R}^2$ obtido rotacionando $\vec{V}$ pelo ângulo ϕ na direção do relógio:

Example 9.3 (Factor analysis of consumer-preference data) In a consumer-preference study, a random sample of customers were asked to rate several attributes of a new product. The responses, on a 7-point semantic differential scale, were tabulated and the attribute correlation matrix constructed. The correlation matrix is presented next:

Attribute (*Variable*)		1	2	3	4	5
Taste	1	1.00	.02	(96)	.42	.01
Good buy for money	2	.02	1.00	.13	.71	(85)
Flavor	3	.96	.13	1.00	.50	.11
Suitable for snack	4	.42	.71	.50	1.00	(79)
Provides lots of energy	5	.01	.85	.11	.79	1.00

It is clear from the circled entries in the correlation matrix that variables 1 and 3 and variables 2 and 5 form groups. Variable 4 is "closer" to the (2, 5) group than the (1, 3) group. Given these results and the small number of variables, we might expect that the apparent linear relationships between the variables can be explained in terms of, at most, two or three common factors.

The first two eigenvalues, $\hat{\lambda}_1 = 2.85$ and $\hat{\lambda}_2 = 1.81$, of $\mathbf{R}$ are the only eigenvalues greater than unity. Moreover, $m = 2$ common factors will account for a cumulative proportion

$$\frac{\hat{\lambda}_1 + \hat{\lambda}_2}{p} = \frac{2.85 + 1.81}{5} = .93$$

of the total (standardized) sample variance. The estimated factor loadings, communalities, and specific variances, obtained using (9-15), (9-16), and (9-17), are given in Table 9.1.

Table 9.1

Variable	Estimated factor loadings $\tilde{\ell}_{ij} = \sqrt{\hat{\lambda}_i}\,\hat{e}_{ij}$ F_1	F_2	Communalities $\tilde{h}_i^2$	Specific variances $\tilde{\psi}_i = 1 - \tilde{h}_i^2$
1. Taste	.56	.82	.98	.02
2. Good buy for money	.78	−.53	.88	.12
3. Flavor	.65	.75	.98	.02
4. Suitable for snack	.94	−.10	.89	.11
5. Provides lots of energy	.80	−.54	.93	.07
Eigenvalues	2.85	1.81		
Cumulative proportion of total (standardized) sample variance	.571	.932		

Now,

$$\tilde{\mathbf{L}}\tilde{\mathbf{L}}' + \tilde{\boldsymbol{\Psi}} = \begin{bmatrix} .56 & .82 \\ .78 & -.53 \\ .65 & .75 \\ .94 & -.10 \\ .80 & -.54 \end{bmatrix} \begin{bmatrix} .56 & .78 & .65 & .94 & .80 \\ .82 & -.53 & .75 & -.10 & -.54 \end{bmatrix}$$

$$+ \begin{bmatrix} .02 & 0 & 0 & 0 & 0 \\ 0 & .12 & 0 & 0 & 0 \\ 0 & 0 & .02 & 0 & 0 \\ 0 & 0 & 0 & .11 & 0 \\ 0 & 0 & 0 & 0 & .07 \end{bmatrix} = \begin{bmatrix} 1.00 & .01 & .97 & .44 & .00 \\ & 1.00 & .11 & .79 & .91 \\ & & 1.00 & .53 & .11 \\ & & & 1.00 & .81 \\ & & & & 1.00 \end{bmatrix}$$

nearly reproduces the correlation matrix $\mathbf{R}$. Thus, on a purely descriptive basis, we would judge a two-factor model with the factor loadings displayed in Table 9.1 as providing a good fit to the data. The communalities (.98, .88, .98, .89, .93) indicate that the two factors account for a large percentage of the sample variance of each variable.

We shall not interpret the factors at this point. As we noted in Section 9.2, the factors (and loadings) are unique up to an orthogonal rotation. A rotation of the factors often reveals a simple structure and aids interpretation.

Example 9.4 (Factor analysis of stock-price data) Stock-price data consisting of $n = 103$ weekly rates of return on $p = 5$ stocks were introduced in Example 8.5. In that example, the first two sample principal components were obtained from $\mathbf{R}$. Taking $m = 1$ and $m = 2$, we can easily obtain principal component solutions to the orthogonal factor model. Specifically, the estimated factor loadings are the sample principal component coefficients (eigenvectors of $\mathbf{R}$), scaled by the square root of the corresponding eigenvalues. The estimated factor loadings, communalities, specific variances, and proportion of total (standardized) sample variance explained by each factor for the $m = 1$ and $m = 2$ factor solutions are available in Table 9.2. The communalities are given by (9-17). So, for example, with $m = 2$, $\tilde{h}_1^2 = \tilde{\ell}_{11}^2 + \tilde{\ell}_{12}^2 = (.732)^2 + (-.437)^2 = .73$.

Table 9.2

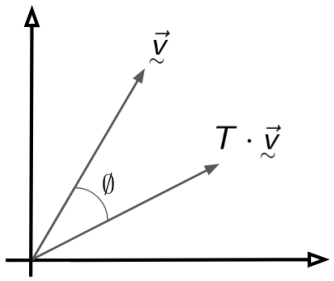
Variable	One-factor solution: Estimated factor loadings F_1	One-factor solution: Specific variances $\tilde{\psi}_i = 1 - \tilde{h}_i^2$	Two-factor solution: Estimated factor loadings F_1	Two-factor solution: Estimated factor loadings F_2	Two-factor solution: Specific variances $\tilde{\psi}_i = 1 - \tilde{h}_i^2$
1. J P Morgan	.732	.46	.732	−.437	.27
2. Citibank	.831	.31	.831	−.280	.23
3. Wells Fargo	.726	.47	.726	−.374	.33
4. Royal Dutch Shell	.605	.63	.605	.694	.15
5. ExxonMobil	.563	.68	.563	.719	.17
Cumulative proportion of total (standardized) sample variance explained	.487		.487	.769	

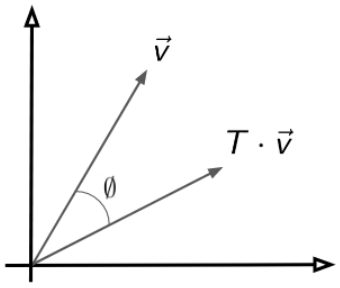
The residual matrix corresponding to the solution for $m = 2$ factors is

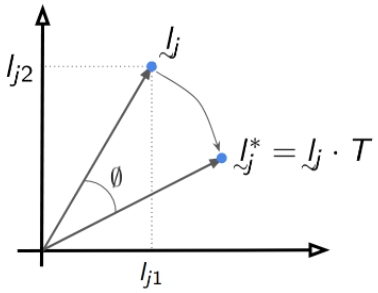
$$\mathbf{R} - \tilde{\mathbf{L}}\tilde{\mathbf{L}}' - \tilde{\Psi} = \begin{bmatrix} 0 & -.099 & -.185 & -.025 & .056 \\ -.099 & 0 & -.134 & .014 & -.054 \\ -.185 & -.134 & 0 & .003 & .006 \\ -.025 & .014 & .003 & 0 & -.156 \\ .056 & -.054 & .006 & -.156 & 0 \end{bmatrix}$$

The proportion of the total variance explained by the two-factor solution is appreciably larger than that for the one-factor solution. However, for $m = 2$, $\tilde{\mathbf{L}}\tilde{\mathbf{L}}'$ produces numbers that are, in general, larger than the sample correlations. This is particularly true for r_{13}.

It seems fairly clear that the first factor, F_1, represents general economic conditions and might be called a *market factor*. All of the stocks load highly on this factor, and the loadings are about equal. The second factor contrasts the banking stocks with the oil stocks. (The banks have relatively large negative loadings, and the oils have large positive loadings, on the factor.) Thus, F_2 seems to differentiate stocks in different industries and might be called an *industry factor*. To summarize, rates of return appear to be determined by general market conditions and activities that are unique to the different industries, as well as a residual or firm specific factor.







Considerando vetores-linha:

$$(T\vec{v})' = \vec{v}'T' = (x,y)\begin{bmatrix} cos(\phi) & -sin(\phi) \\ sin(\phi) & cos(\phi) \end{bmatrix}$$

A situação geométrica continua a mesma de antes, representar o vetor como linha ou coluna não altera o resultado. Vamos trabalhar com as linhas da matriz L.

Considere a matriz L das cargas

$$L = \begin{bmatrix} \underset{\sim}{\ell}_1' \\ . \\ . \\ \underset{\sim}{\ell}_{15}' \end{bmatrix} = \begin{bmatrix} \ell_{11} & \ell_{12} \\ \ell_{21} & \ell_{22} \\ . & . \\ . & . \\ \ell_{15,1} & \ell_{15,2} \end{bmatrix} \text{ e } T = \text{matriz ortogonal.}$$

O produto $\underset{15x2}{L} \cdot \underset{2x2}{T}$ pode ser pensado linha a linha

$$\underset{15x2}{L} \cdot \underset{2x2}{T} = \begin{bmatrix} \underset{\sim}{\ell}_1' \\ \underset{\sim}{\ell}_2' \\ . \\ . \\ \underset{\sim}{\ell}_{15}' \end{bmatrix} = \begin{bmatrix} \underset{\sim}{\ell}_1' \cdot \underset{2x2}{T} \\ \underset{\sim}{\ell}_2' \cdot T \\ . \\ . \\ \underset{\sim}{\ell}_{15}' \cdot T \end{bmatrix} = L^*$$

As linhas de L^* são as linhas de L rotacionadas de certo ângulo $\emptyset$ associado à matriz T.

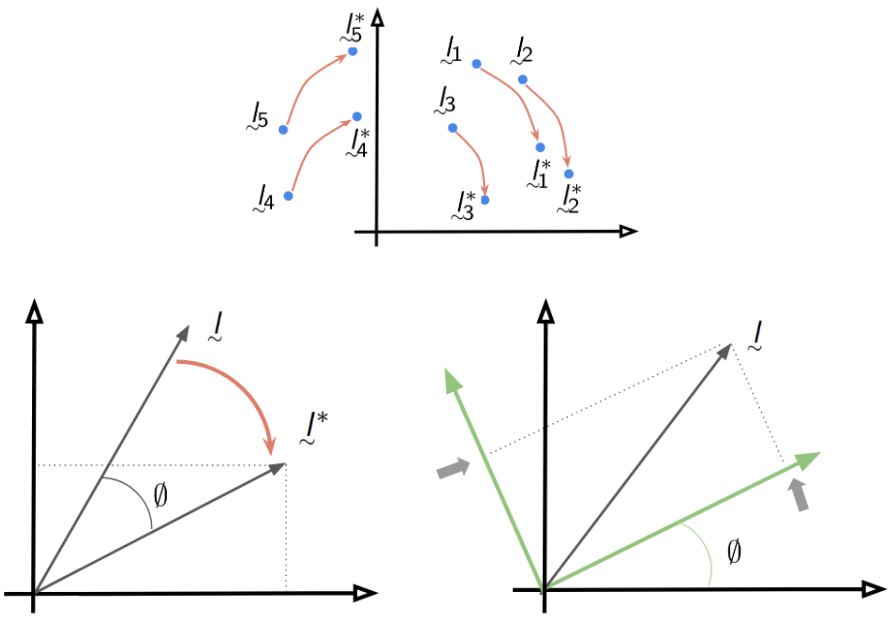
Rodando uma linha de L Rodando cinco linhas de L

OK, o que tudo isto quer dizer? Suponha que o modelo fatorial é correto e que realmente podemos escrever ou decompor a matriz de covariância de $\mathbf{X}$ como:

$Var(\mathbf{X}) = \underset{15x5}{\Sigma} = \underset{15x2}{L} \cdot \underset{2x15}{L'} + \underset{15x15}{\psi} \Rightarrow$ diagonal Seja $\underset{2x2}{T}$ qualquer matriz ortogonal de rotação no

plano.

Então podemos escrever



$$\Sigma = LL' + \psi = LTT'L' + \psi = (LT)(LT)' + \psi = \underset{15x2}{L^*}\underset{2x15}{(L^*)'} + \psi$$
$$\mathbb{I}_2 \qquad L^*$$

Isto significa que, se tivermos apenas Σ, teremos

$$LL' + \psi = \Sigma = L^*(L^*)' + \psi$$

onde $L^* = LT$ é diferente de L. As linhas de L^* são as linhas de L rotacionadas de um ângulo $\emptyset$. Como T é arbitrária (pode ser qualquer T) isto significa que podemos rodar L à vontade, com qualquer ângulo $\emptyset$, que sempre teremos uma representação de Σ da forma $\Sigma = L^*(L^*)' + \psi$.

Mas como interpretar os números que aparecem em L? Qual a L "correta"? Não é possível determinar uma única L tal que $\Sigma = LL' + \psi$ Existem infinitos L com esta propriedade. Qualquer $L^* = LT$ (isto é, L rotacionada) terá a mesma propriedade. Todas as matrizes de carga L^* obtidas a partir de uma matriz L inicial terão a mesma capacidade de reproduzir a matriz de covariância Σ.

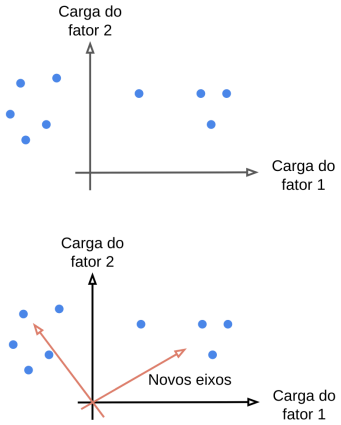
Ao invés disso se tornar um problema, transformamos os limões numa limonada. Caso uma matriz de cargas L inicialmente obtida por algum método de estimação não fornecer uma boa interpretação para os fatores, nós procuramos uma versão rotacionada $L^* = LT$ tal que as novas cargas sejam mais interpretáveis. É comum sermos capazes de terminar com uma estrutura mais simples que a matriz L inicial. Qual é esta estrutura mais simples?

Idealmente, nós gostaríamos de ver um padrão em que cada variável tenha uma carga alta num dos fatores e uma carga ≈ 0 nos demais. O objetivo é procurar uma rotação dos eixos de forma que as novas cargas fiquem o mais pròximo possível deste ideal. OBS: Se temos 15 pontos no plano (as cargas $\underset{\sim}{\ell}_j$) e rodamos todas elas de um ângulo $\emptyset$, isto é o mesmo que rodar os dois eixos do plano de $-\emptyset$ e deixar os "pontos intactos".

Coordenadas nos novos eixos são as mesmas de $\underset{\sim}{\ell}^*$nos eixos antigos.

Se nossas cargas $\underset{px2}{L} = \begin{bmatrix} . & . \\ . & . \\ . & . \end{bmatrix}$ são assim:

procuramos rodar os eixos até que as cargas sejam próximas do ideal
Nos novos eixos, as cargas são ≈ 0 exceto em um único fator.



15.0.9 Procedimento VARIMAX

Defina $\tilde{\ell}^*_{ij} = \frac{\ell^*_{ij}}{h_i} = \frac{\ell^*_{ij}}{\ell^*_{i1}+\ell^*_{i2}+\ldots+\ell^*_{ip}} \in [0,1]$ as cargas dos fatores rotacionados e normalizados. Busque a rotação T tal que maximize

$$V = \frac{1}{p}\sum_{j=1}^{m}[\sum_{i=1}^{p}\tilde{\ell}^{*\,4}_{ij} - \frac{(\sum_{i=1}^{p}\tilde{\ell}^{*\,2}_{ij})^2}{p}]$$

$\propto \sum_{j=1}^{m}$(variância das $(\text{cargas})^2$ normalizadas do fator j)

Maximizar V significa espalhar as $(\text{cargas})^2$ o máximo possível, com valores altos em alguns fatores e valores ≈ 0 em outros. Tendo estimado a matriz de cargas, podemos estimar o valor dos fatores de cada indivíduo da amostra.

Suponha que o i-ésimo indivíduo tenha o vetor $\mathbf{X}_i$ e que tenhamos estimado $\underset{\sim}{\mu}$ (a média das variáveis sobre a amostra) e tenhamos também a matriz de cargas L (talvez rotacionada). O vetor $\underset{\sim}{F_i}$ deste indivíduo é estimado pela minimização da diferença entre $\mathbf{X}_i$ e $\underset{\sim}{\mu} + L\underset{\sim}{F}$.

Isto é, procuramos um vetor $\underset{\sim}{F_i}$ tal que ele minimize o comprimento

$$||X_i - \mu - L\underset{\sim}{F_i}||^2$$

Veja a lista de exercícios (beer example) para um exemplo.

16. Classification

Neste capítulo, vamos estudar o problema de classificação supervisionada ou, abreviadamente, simplesmente classificação. Esta é uma das tarefas mais estudadas e bem sucedidas na análise de dados. Já estudamos uma das metodologias para este problema, as àrvores de classificação, no capítulo 5. Neste capítulo vamos aprofundar o estudo desse problema. Vamos abordar o problema aprendendo alguns de de seus conceitos básicos e o resultado fundamental do classificador ótimo de Bayes.

Vamos começar com a situação mais simples em que temos duas classes. Indivíduos (ou itens ou exemplos) são amostrados de uma certa população. Esta população é particionada em duas classes disjuntas: pop_1 (denotada π_1) e pop_2 (denotada π_2). As duas classes representam uma partição da população:

- Todo indivíduo pertence a uma das duas subpopulações.
- Nenhum indivíduo pertence a duas classes ao mesmo tempo.

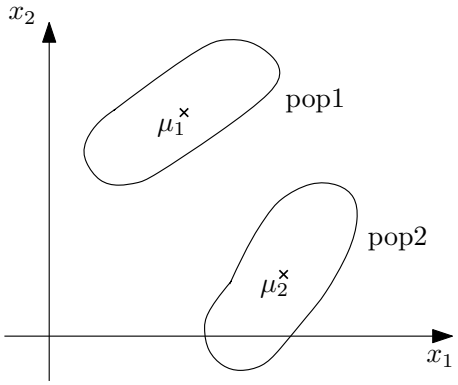
Vamos ver alguns exemplos dessa situação.

- Risco de Crédito: Empresas tomadoras de crédito em um banco: $\pi_1 \rightarrow$ créditos bons; $\pi_2 \rightarrow$ créditos ruins
- Crânios em um sítio arqueológico: $\pi_1 \rightarrow$ homens; $\pi_2 \rightarrow$ mulheres
- Saúde: Pessoas com úlcera (π_1) e pessoas sem úlcera (π_2)
- Saúde: Mulheres com (π_1) ou sem (π_2) câncer de mama
- Análise de textos de dois participantes do movimento de independência dos EUA: π_1 : James Madison ou π_2 : Alexander Hamilton.
- Duas espécies da flor Iris: π_1: Iris setosa; π_2: Iris virginica.
- Usuários de um website: π_1: aqueles que clicam num certo anúncio e π_2, aqueles que não clicam
- Alunos de um curso online: π_1 : aqueles que evadem e π_2: aquels que completam o curso

Preditores ou Features: Em cada indivíduo, medimos um conjunto de p variáveis preditoras (ou features):

$$\mathbf{X} = (X_1, X_2, \ldots, X_p)$$

Com base nas medições em $\mathbf{X}$, queremos *inferir ou predizer* se $\mathbf{X} \in \pi_1$ ou se $\mathbf{X} \in \pi_2$. Queremos



descobrir (ou aprender) uma regra matemática $g(\mathbf{X})$ que prediga se o indivíduo pertence à classe π_1 ou a π_2. Esta regra será usada para predizer a classe de novos itens para os quais sabemos $\mathbf{X}$ mas não sabemos a sua classe.

Para construir uma *regra de classificação* de novos itens, usamos uma amostra com as classes *conhecidas* (amostra rotulada com a classe):

Item	Classe ou População	Variáveis X_1 X_2 ... X_p
1	π_1	$X_{1,1}$ $X_{1,2}$ $X_{1,3}$ ... $X_{1,p}$
2	π_1	$X_{2,1}$ $X_{2,2}$ $X_{2,3}$ ... $X_{2,p}$
$\vdots$	$\vdots$	$\vdots$
m_1	π_1	$X_{m_1,1}$ $X_{m_1,2}$ $X_{m_1,3}$ ... $X_{m_1,p}$
1	π_2	$X_{m_1+1,1}$ $X_{m_1+1,2}$ $X_{m_1+1,3}$ ... $X_{m_1+1,p}$
2	π_2	$X_{m_1+2,1}$ $X_{m_1+2,2}$ $X_{m_1+2,3}$ ... $X_{m_1+2,p}$
$\vdots$	$\vdots$	$\vdots$
m_2	π_2	$X_{m_1+m_2,1}$ $X_{m_1+m_2,2}$ $X_{m_1+m_2,3}$ ... $X_{m_1+m_2,p}$
Novo Item	?????	X_1^* X_2^* X_3^* ... X_p^*

Chega um novo item. conhecemos $\mathbf{X}$ desse novo item mas não conhecemos a sua classe. X_1^* X_2^* X_3^* ... X_p^* têm seus valores conhecidos, são efetivamente observados. O símbolo **?????** na tabela 16 indica que queremos inferir a classe desse novo item Vamos retomar os exemplos anteriores com exemplos de possíveis na Tabela 16.

Por que precisamos predizer a classe de um item novo? Existem algumas razões práticas para estabelecer esta tarefa:

- A classe poderá ser conhecida apenas no futuro mas seria importante ter pelo menos uma predição desta classe desconhecida no presente. Por exemplo, no momento em que o crédito é solicitado, não sabemos se o Indivíduo vai honrar ou não o contrato do empréstimo.
- Informação sobre a classe não é conhecida com certeza. Por exemplo, no problema dos crânios arqueológicos danificados.
- Obter a classe pode implicar em destruir o item. Por exemplo, queremos classificar um paciente chegando ao pronto socorro com lesão na cabeça como UTI ou não-UTI, com base em algumas medidas rápidas. Esperar para saber com certeza se deve ir para UTI pode significar esperar demais.

Cada uma das duas populações possui uma distribuição conjunta para as p variáveis preditoras representadas pelo vetor $\mathbf{X} = (X_1, X_2, ..., X_p)$. Na população 1, temos $(\mathbf{X}| \in \pi_1) \sim f_1(\mathbf{x})$. Na população 2, temos $(X| \in \pi_2) \sim f_2(\mathbf{x})$. A Figura 16 mostra as regiões onde os dados seriam observados caso tivéssemos apenas $p = 2$ variáveis.

Por exemplo, poderíamos ter $f_1(\mathbf{x}) = N_p(\underset{px1}{\mu_1}, \underset{pxp}{\Sigma_1})$ e $f_2(\mathbf{x}) = N_p(\underset{px1}{\mu_2}, \underset{pxp}{\Sigma_2})$. Esta escolha é apenas ilustrativa. A distribuição gaussiana não é um re uisito na teoria de classificação. As densidades

Populações π_1 e π_2	Variáveis X_1 ... 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 - n^o 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 ou sem úlcera	- 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
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	- Nota do exame de entrada no curso - Medidas de motivação a partir de questionário na entrada - Renda familiar - Idade

$f_1(\mathbf{x})$ e $f_2(\mathbf{x})$ são completamente arbitrárias e podem ser qualquer coisa.

Com base na amostra rotulada (com as classes conhecidas), podemos obter estimativas dos valores esperados μ_1 e μ_2 das distribuições $f_1(\mathbf{x})$ e de $f_2(\mathbf{x})$ simplesmente tomando a mé dia aritmética de cada variável dentro de cada classe. Isto é apresentado de forma esquemática na Tabela 16.

Item	Classe ou População	Variáveis X_1 X_2 ... X_p
1	π_1	$X_{1,1}$ $X_{1,2}$ $X_{1,3}$ ... $X_{1,p}$
2	π_1	$X_{2,1}$ $X_{2,2}$ $X_{2,3}$ ... $X_{2,p}$
$\vdots$	$\vdots$	$\vdots$
m_1	π_1	$X_{m_1,1}$ $X_{m_1,2}$ $X_{m_1,3}$ ... $X_{m_1,p}$
	Médias das p vars	$\bar{x}_{11}$ $\bar{x}_{12}$ $\bar{x}_{13}$ ... $\rightarrow$ vetor $\bar{\mathbf{x}}_1 = \hat{\mu}_1$
1	π_2	$X_{m_1+1,1}$ $X_{m_1+1,2}$ $X_{m_1+1,3}$ ... $X_{m_1+1,p}$
2	π_2	$X_{m_1+2,1}$ $X_{m_1+2,2}$ $X_{m_1+2,3}$ ... $X_{m_1+2,p}$
$\vdots$	$\vdots$	$\vdots$
m_2	π_2	$X_{m_1+m_2,1}$ $X_{m_1+m_2,2}$ $X_{m_1+m_2,3}$ ... $X_{m_1+m_2,p}$
	Médias das p vars	$\bar{x}_{21}$ $\bar{x}_{22}$ $\bar{x}_{23}$ ... $\rightarrow$ vetor $\bar{\mathbf{x}}_2 = \hat{\mu}_2$

Podemos também estimar as matrizes $p \times p$ de covariância Σ_1 e Σ_2 com as amostras rotuladas. Por exemplo, para a classe 1, estimamos as p variâncias $\sigma^2_{11}, \sigma^2_{1,2}, \ldots, \sigma^2_{1,p}$ através das variâncias amostrais $s^2_{11}, s^2_{1,2}, \ldots, s^2_{1,p}$. A covariância c_{12} entre a variável i e a variável j (da classe 1) é estimada por $s_{1i}s_{1j}r_{1,ij}$ usando os desvios-padrão de cada variável e a correlação $r_{1,ij}$.

O gráfico à esquerda na Figura 16.1 mostra a situação básica do problema de classificação usando apenas duas variáveis preditoras. Temos um novo exemplo $\mathbf{X} = (X_1, X_2)$ para o qual

Item	Classe ou População	Variáveis X_1 X_2 ... X_p
1	π_1	$X_{1,1}$ $X_{1,2}$ $X_{1,3}$ ... $X_{1,p}$
2	π_1	$X_{2,1}$ $X_{2,2}$ $X_{2,3}$ ... $X_{2,p}$
$\vdots$	$\vdots$	$\vdots$
m_1	π_1	$X_{m_1,1}$ $X_{m_1,2}$ $X_{m_1,3}$ ... $X_{m_1,p}$
	Variância amostral das p vars	$s^2_{11}, s^2_{1,2}, \ldots, s^2_{1,p}$

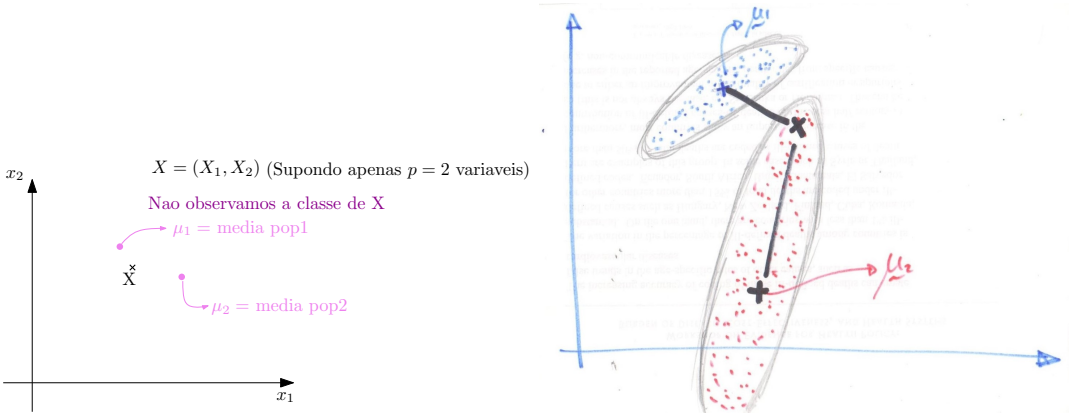


Figure 16.1: Colocar caption aqui

queremos predizer a classe: população 1 ou 2? Uma primeira abordagem seria olhar a distância do novo item $\mathbf{X}$ aos vetores μ_1 e μ_2. Parece razoável alocar $\mathbf{X}$ à população 1, pois a distância euclidiana entre $\mathbf{X}$ e μ_1 é menor que entre $\mathbf{X}$ e μ_2. Entretanto, o gráfico à direita na Figura 16.1 mostra que este raciocínio pode levar a erros. Vendo agora toda a distribuição dos dados em torno de suas médias, vemos que $\mathbf{X}$ parece pertencer à população 2, e não à população 1. Não basta olharmos apenas as distâncias euclidianas entre $\mathbf{X}$ e os vetores μ_1 e μ_2. Nós precisamos levar em conta as correlações entre as variáveis preditoras. Precisamos olhar a distância estatística ou a distância de Mahalanobis do novo item $\mathbf{X}$ a cada um dos centros μ_1 e μ_2.

De fato, nós já vimos na seção ?? que a distância euclidiana não é a melhor maneira de medir distâncias entre vetores aleatórios. Os pontos coloridos de vermelho e azul no lado esquerdo da Figura 16 possuem a mesma distância euclidiana ao centro da nuvem de pontos estatísticos. O centro da nuvem de pontos representa o valor médio de cada variável, o perfil "médio" desta população estatística. Intuitivamente, os pontos vermelhos estão estatisticamente mais distantes do centro da nuvem que os pontos azuis. Como criar uma medida de distância matemática que incorpore esta intuição? A resposta é a Distância de Mahalanobis, que leva em conta as variâncias e correlações de cada variável. Os pontos na elipse inclinada estão à mesma distância de Mahalanobis do centro da nuvem. As correlações entre as variáveis $\rightarrow$ inclinação da elipse. Vamos lembrar que os eixos das elipses são os componentes principais da matriz de covariância.

A distância de Mahalanobis entre um ponto $\mathbf{X}$ com p variáveis e o seu vetor esperado $\mathbb{E}(X) = \underset{px1}{\mu}$ (vetor com os valores esperados de cada uma das p variáveis) usa a matriz $p \times p$ $\mathbb{V}(X) = \underset{pxp}{\Sigma}$, a matriz de variâncias e covariâncias do vetor $\mathbf{X}$. Ela é dada por

$$d^2_{\Sigma}(\mathbf{X}, \mu) = (\mathbf{X} - \mu)^t \, \Sigma^{-1} \, (\mathbf{X} - \mu)$$

Assim, uma regra de classificação que parece mais razoável é a seguinte:

- Calcule a distância estatístca entre $\mathbf{X}$ e μ_1:

$$d^2_1 = d^2_{\Sigma_1}(\mathbf{X}, \mu_1) = (\mathbf{X} - \mu_1)^t \Sigma_1^{-1} (\mathbf{X} - \mu_1)$$

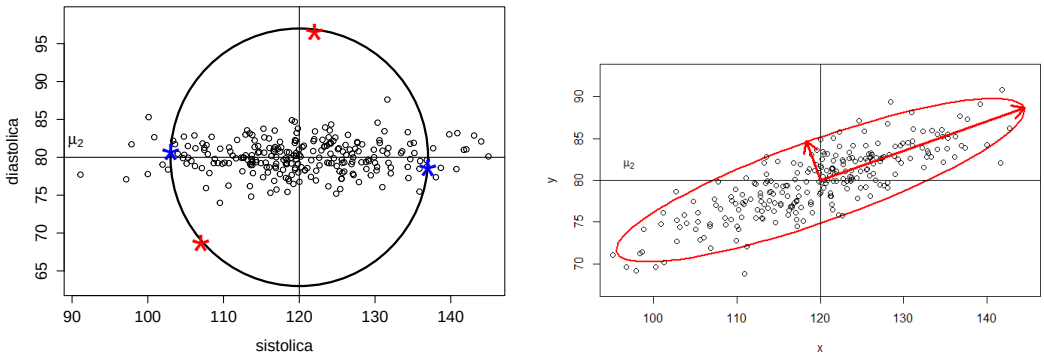


Figure 16.2: Colocar caption aqui

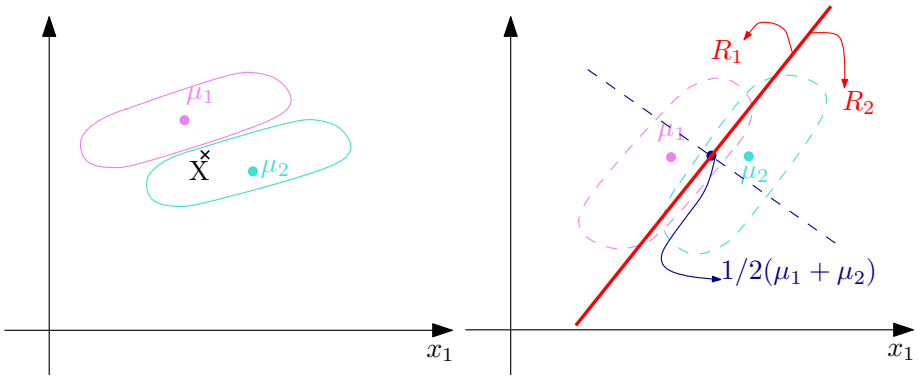


Figure 16.3: Colocar caption aqui

- Calcule a distância estatístca entre $\mathbf{X}$ e μ_1:

$$d_2^2 = d_{\Sigma_2}^2(\mathbf{X}, \mu_2) = (\mathbf{X} - \mu_2)^t \Sigma_2^{-1} (\mathbf{X} - \mu_2)$$

- Aloque $\mathbf{X}$ à população com a menor distância de Mahalanobis d^2.

Isto é,

- Se $d_{\Sigma_1}^2(\mathbf{X}, \mu_1) < d_{\Sigma_2}^2(\mathbf{X}, \mu_2) \Rightarrow$ aloque $\mathbf{X}$ à pop1;
- Caso contrário, aloque $\mathbf{X}$ à pop2.

Agora vamos olhar para esta regra de classificação de uma maneira que vai ser muito importante no restante do capítulo. O espaço $\mathbb{R}^p$ é particionado em duas regiões:

- $R_1 = \left\{ x \in \mathbb{R}^p \mid d_{\Sigma_1}^2(\mathbf{X}, \mu_1) < d_{\Sigma_2}^2(\mathbf{X}, \mu_2) \right\}$
- $R_2 = \mathbb{R}^p - R_1 =$ pontos que serão alocados à pop2.

Quais são essas duas regiões? Ela depende do método de classificação adotado. A Figura 16 mostra duas possíeis separações entre as duas regiões. Na esquerda, temos uma separação linear entre as duas regiões. Na direita, temos uma separação não-linear. A linha que separa as duas regiões é chamada fronteira de decisão (*decision boundary*, em inglês).

16.0.1 De Mahalanobis para razão de densidades

Existe uma outra maneira de ver a regra de classificação baseada na comparação das distâncias de Mahalanobis. Sejam $f_1(\mathbf{x})$ a densidade de probabilidade do vetor $\mathbf{X}$ para os indivíduos que pertecem à população π_1. Similarmente, $f_2(\mathbf{x})$ é a densidade do vetor $\mathbf{X}$ se ele estivar em π_2. Vamos supor que, dentro de cada classe, os dados $\mathbf{X}$ sigam uma distribuição gaussiana:

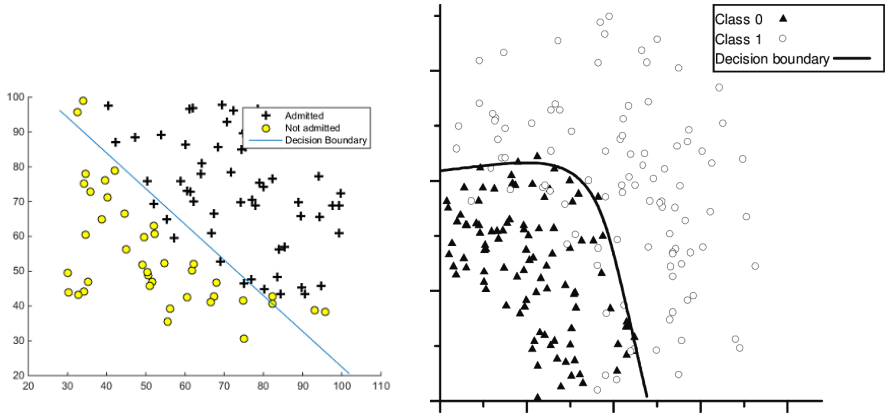


Figure 16.4: Colocar caption aqui

$$\mathbf{X} \sim \begin{cases} N(\mu_1, \Sigma_1), \ se \in \pi_1 \\ N(\mu_2, \Sigma_2), \ se \in \pi_2 \end{cases}$$

No caso gaussiano, temos

$$f_1(\mathbf{x}) = \underbrace{\left[\frac{1}{(2\pi)^{p/2}|\Sigma_1|^{1/2}}\right]}_{\text{constante em } \mathbf{x}:\ c_1} \exp\left(-\frac{1}{2}\underbrace{(\mathbf{x}-\mu_1)^t\Sigma_1^{-1}(\mathbf{x}-\mu_1)}_{\text{distância de Mahalanobis}}\right)$$

$$f_2(\mathbf{x}) = \underbrace{\left[\frac{1}{(2\pi)^{p/2}|\Sigma_2|^{1/2}}\right]}_{\text{constante em } \mathbf{x}:\ c_2} \exp\left(-\frac{1}{2}\underbrace{(\mathbf{x}-\mu_2)^t\Sigma_2^{-1}(\mathbf{x}-\mu_2)}_{\text{distância de Mahalanobis}}\right)$$

Note que estamos chamando de c_1 e c_2 os dois fatores constantes em $\mathbf{x}$ que multiplicam a expressão envolvendo a função exponencial. Tomando a razão das duas densidades no mesmo ponto $\mathbf{x}$, temos

$$\begin{aligned} \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} &= \frac{c_1 \ \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_1)^t\Sigma_1^{-1}(\mathbf{x}-\mu_1)\right)}{c_2 \ \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_2)^t\Sigma_2^{-1}(\mathbf{x}-\mu_2)\right)} \\ &= \frac{c_1}{c_2} \ \exp\left(-\frac{1}{2}(d^2_{\Sigma_1}(\mathbf{x},\mu_1) - d^2_{\Sigma_2}(\mathbf{x},\mu_2)\right) \end{aligned}$$

Assim:

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > 1 \iff \underbrace{d^2_{\Sigma_1}(\mathbf{x},\mu_1) < d^2_{\Sigma_2}(\mathbf{x},\mu_2)}_{\text{condição para alocar a } \pi_1}$$

Isto é, no caso gaussiano, o conjunto R_1 dos pontos $\mathbf{x} \in \mathbb{R}^p$ tais que $f_1(\mathbf{x}) > f_2(\mathbf{x})$ é o mesmo conjunto de pontos em que $d^2_{\Sigma_1}(\mathbf{x},\mu_1) < d^2_{\Sigma_2}(\mathbf{x},\mu_2)$ Assim, no caso gaussiano, definir a região de classificação à população π_1 usando a razão de densidades é matematicamente equivalente a definir usando as distâncias de Mahalanobis.

E quando $\mathbf{X}$ não seguir uma gaussiana? O que devemos usar para definir R_1? Faz diferença? Suponha que conhecemos a densidade de $\mathbf{X}$ em cada classe: $f_1(\mathbf{x})$ e $f_2(\mathbf{x})$. Vamos pensar em dois métodos:

- **Método 1:** Imitando o que descobrimos no caso gaussiano, podemos escolher R_1 como sendo o seguinte conjunto de pontos do $\mathbb{R}^p$:

$$R_1 = \{\mathbf{x} \in \mathbb{R}^p \text{ tais que } f_1(\mathbf{x}) > f_2(\mathbf{x})\}$$

- **Método 2:** Outra opção, quando as amostras forem grandes:
  - com as amostras rotuladas, obtemos boas aproximições para os vetores de valores esperados μ_1 e μ_2.
  - Obtemos boas estimativas das matrizes $p \times p$ de covariância Σ_1 e Σ_2.
  - Para cada pontos $\mathbf{x} \in \mathbb{R}^p$ podemos calcular as duas distâncias de Mahalanobis: $d^2_{\Sigma_1}(\mathbf{x}, \mu_1)$ e $d^2_{\Sigma_2}(\mathbf{x}, \mu_2)$.
  - Escolhemos R_1 como sendo o seguinte conjunto de pontos:

$$R_1 = \{\mathbf{x} \in \mathbb{R}^p \text{ tais que } d^2_{\Sigma_1}(\mathbf{x}, \mu_1) < d^2_{\Sigma_2}(\mathbf{x}, \mu_2)\}$$

Como vimos, os dois métodos acima coincidem no caso gaussiano, gerando a mesma região R_1. Entretanto, se as densidades não forem gaussianas, eles geralmente não serão equivalentes. Qual dos dois métodos é o melhor no caso não-gaussiano? É melhor usar a razão de densidades ou a distância de Mahalanobis? Existiria um terceiro método (árvore de classificação, por exemplo, estudadas no capítulo 5) melhor que estes dois métodos? Talvez este terceiro método possa ser usado até no caso gaussiano também. Existirá um método imbatível, insuperável, o melhor de todos os possíveis e imagináveis, por mais criativos que sejamos?

Pelo tom altamente retórico das perguntas anteriores, você já deve ter antecipado corretamente que devemos ter repostas para elas e que devem ser surpreendentes. De fato, podemos responder essas questões. Em particular, vamos responder SIM sobre a existência de um método imbatível. Mais legal ainda, saberemos que método ótimo é este e ele é bem simples!

16.0.2 O caso geral para classificação

Na verdade, o problema que vamos resolver é mais geral do que o que consideramos até agora. Queremos levar em conta os seguintes pontos que serão explicados em seguida:

- O custo de classificação errada pode variar de acordo com a classe.
- Uma das populações pode ser muito mais frequente do que a outra
- A distribuição pode não ser gaussiana

Para entender cada um desses pontos, vamos começar com o primeiro deles no contexto de um exemplo específico. Pense no problema de risco de crédito. Um cliente solicita empréstimo no banco. Queremos saber, no momento do empréstimo, se ele é um bom risco (se pagará no prazo) ou um mau risco. Nos baseamos em várias características (features) medidas no momento do empréstimo tais como:

- idade, sexo, tempo como cliente, saldo médio,
- % do empréstimo em relação ao saldo,
- já pegou empréstimo antes e pagou no prazo? Sim ou Não

Os custo de classificação podem ser colocados numa matriz onde as linhas representam as condições reais e as colunas as decisões de classificação tomadas pelo seu algoritmo favorito:

		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 $= 0$

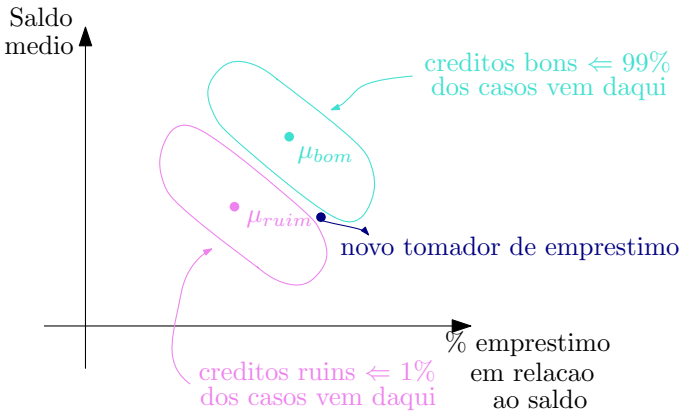


Figure 16.5: Colocar caption aqui

Se uma decisão correta é tomada (um indivíduo é classificado pelo algoritmo na sua classe correta), o cursto será zero. Decisões erradas terão um custo. Vamos representar por $c(2|\in\pi_1)$ o custo de classificar como mau crédito um bom pagador. Vamos negar-lhe o crédito com base na decisão do algoritmo e com isto o custo será aquele de perder um bom cliente. Isto é, perder o pequeno ganho a ser obtido por juros do empréstimo. Vamos representar por $c(1|\in\pi_2)$ o custo de classificar como bom crédito um mau pagador. Neste caso, é o custo de perder todo dinheiro emprestado ou recuperá-lo apenas após um processo judicial longo e custoso. Assim, em geral, neste problema $c(1|\in\pi_2)$ será muito maior que $c(2|\in\pi_1)$.

Um outro exemplo típico aparece na área de saúde. Um paciente entra no pronto-socorro com um traumatismo craniano causado por uma queda (comum entre idosos e crianças), um acidente com moto ou na prática de esportes ou ainda como resultado de uma agressão física. Esta não é uma situação rara. Ocorrem 50 casos por 10 mil habitantes nos EUA por ano, com 2,5 milhões atendimentos em pronto-socorros, 282 mil internações hospitalares e 56 mil mortes[1]. A decisão mais importante e urgente é se devemos levar o paciente imediatamente para a UTI ou se ele deve ficar sob observação. As primeiras horas após a lesão ocorrer são decisivas. Os custos de uma decisão errada são bem diferentes:

- Levar para a UTI imediatamente mas sem necessidade gasta recursos do hospital que poderiam ser usados de outra forma.
- Deixar sob observação um paciente que necessitava de tratamento intensivo pode significar sua morte ou lesão permanente.

Os custos muito diferentes têm impacto numa regra de classificação: se quisermos minimizar o custo esperado de uma decisão ruim, devemos levar em conta esses custos muito diferentes. Como fazer isso?

O segundo ponto que queremos considerar é o tamanho desbalanceado das duas populações No mercado de risco de crédito, maus pagadores são muito mais raros do que bons pagadores. Para simplificar a explicação, suponha que os custos de classificação incorreta sejam iguais: $c(1|\in\pi_2) = c(2|\in\pi_1)$. Se $d^2_{\Sigma_1}(\mathbf{x},\mu_1) = d^2_{\Sigma_2}(\mathbf{x},\mu_2)$, estamos dizendo que não existe evidência nas variáveis em $\mathbf{x}$ para saber se $\in\pi_1$ ou se $\in\pi_2$. Veja a Figura 16.0.2. O ponto $\mathbf{x}$ está igualmente Mahalanobis-distante das duas populações.

Resta uma informação *a priori*, que não está no novo caso $\mathbf{x}$. É que, com alta probabilidade (0.99), o novo caso $\mathbf{x}$ vem de π_1. A chance de um novo casos qualquer vir de π_2 é muito pequena (1% apenas). Então:

- se os custos são os mesmos

[1] REF: `https://msdmnls.co/3i1d303`

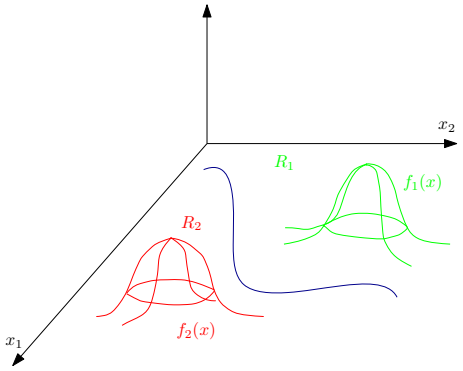


Figure 16.6: Colocar caption aqui

- se o novo caso está igualmente distante de π_1 e π_2
- parece razoável usar a informação *adicional* de que existem muito mais casos em π_1 do que em π_2 e alocar $\mathbf{x}$ em π_1.

Como misturar custos e probabilidades *a priori* no caso geral?

O tamanho n e m das amostras de cada grupo não tem de ser proporcional às probabilidades a priori π_1 (???) e π_2. Por exemplo, a classe 1 pode ser constitu{ída de indivíduos que não quitam o débito no banco no prazo estipulado e a classe 2 são os demais que quitam. Temos $\pi_1 \approx 0$. Na análise de dados, podemos buscar na base de dados todos os n casos de clientes que não pagaram o débito dentro do prazo. Para a amostra da classe 2, pegamos um número m arbitrário de clientes. Podemos tomar $m = n$ ou $m = 3n$. Assim, a razão n/m não precisa ser igual à razão π_1/π_2.

Agora um terceiro e último ponto a ser considerado: a distribuição dos dados pode não ser gaussiana. No caso gaussiano, como

$$f(\mathbf{x}) = \text{ cte } \cdot \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_1)^t \Sigma^{-1}(\mathbf{x}-\mu_1)\right),$$

comparar distâncias de Mahalanobis é equivalente a comparar duas densidades de probabilidades. Os dois métodos são, na verdade, um só. Mas não sabemos se existe outro método melhor que estes dois, nem se algum deles funcionaria bem num caso não-gaussiano. Vamos considerar o caso de distribuições $f_1(\mathbf{x})$ e $f_2(\mathbf{x})$ arbitrárias. Não estaremos restritos a densidades gaussianas nem vamos nos limitar a olhar apenas as distâncias de Mahalanobis. E vamos descobrir a melhor regra de classificação: não existe nada melhor que este novo classificador. Ele é o *classificador ótimo de Bayes* (optimal Bayes classifier). Como sempre, precisamos especificar: ótimo em que sentido? Melhor para quê? Ele será ótimo no sentido de minimizar o custo esperado de classificação incorreta. Vamos definir esta quantidade a seguir.

16.0.3 Expected cost of misclassification (ECM)

Temos duas densidades, $f_1(\mathbf{x})$ e $f_2(\mathbf{x})$, e uma regra de classificação que particiona o espaço das features em duas regiões: $\mathbb{R}^p = R_1 \cup R_2$. As regiões R_1 e R_2 são definidas por alguma regra de classificação que não é necessariamente boa. Na Figura 16.0.3 visualizamos as duas densidades e a fronteira de decisão.

Veja que:

(a) estabelecer uma partição de $R_1 \cup R_2 = \mathbb{R}^p$, com $R_2 = \mathbb{R}^p - R_1$, implica em criar uma regra de classificação:
Regra: Se $\mathbf{x} \in R_1$, aloque $\mathbf{x}$ a π_1. Else, aloque $\mathbf{x}$ a π_2.

(b) estabelecer uma regra de classificação qualquer implica em criar uma partição de $\mathbb{R}^p$:
$R_1 = \{\mathbf{x} \in \mathbb{R}^p |$ a regra aloca $\mathbf{x}$ a $\pi_1\}$

$R_2 = \mathbb{R}^p - R_1$

Assim, estabelecer uma regra de classificação baseada em $\mathbf{x} \in \mathbb{R}^p$ é equivalente a estabelecer uma partição $R_1 \cup R_2 = \mathbb{R}^p$.

O classificador (ou regra de classificação) é uma função matemática, determinística. Por exemplo, seja $\mathbf{x} = (x_1, x_2, x_3) = (\text{age, sex, income})$. Suponha que $\mathbf{x}_i = (37, \text{FEM}, 25) = \mathbf{x}_j$, duas pessoas i e j possuindo os mesmos três atributos. O classificador não muda de valor (ou de classe) diante desses dois exemplos. A classe atribuída será a mesma para os dois exemplos. O classificador é uma função matemática de $\mathbf{x}$:

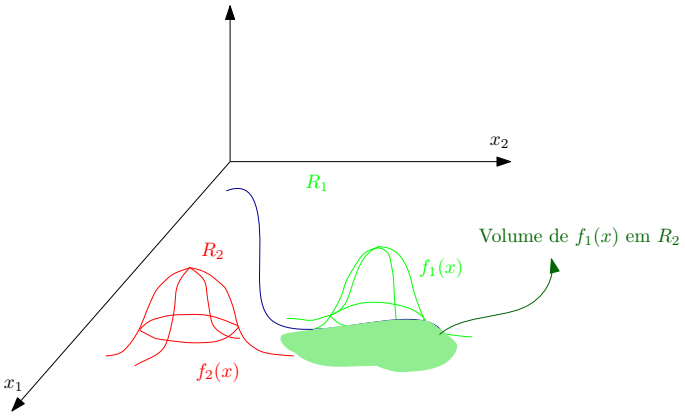
$$g(\mathbf{x}) = \begin{cases} \pi_1 & \text{if } \mathbf{x} \in R_1 \\ \pi_2 & \text{if } \mathbf{x} \in R_2 = \mathbb{R}^p - R_1 \end{cases}$$

Assim, dado um certo exemplo $\mathbf{x}$, a regra vai alocá-lo a uma das duas classes. Se tivermos outro exemplo $\mathbf{x}^*$ cujas variáveis tenham os mesmos valores que $\mathbf{x}$, a classe atribuída a $\mathbf{x}^*$ será a mesma da classe atribuída a $\mathbf{x}$.

Vamo obter a probabilidade condicional de classificar um objeto em π_2 quando, de fato, ele é de π_1. Ela é igual a

$$\mathbb{P}(Class.\ em\ \pi_2 | \in \pi_1) = \mathbb{P}(\mathbf{X} \in R_2 | \in \pi_1) = \int_{R_2} f_1(\mathbf{x}) d\mathbf{x}$$

e corresponde ao volume verde sob a densidade $f_1(\mathbf{x})$ na região R_2 (ver Figura 16.0.3).



Similarmente, a probabilidade de classificar erradamente em π_2 dado que ele é de π_2 é dada por:

$$\mathbb{P}(Class.\ em\ \pi_1 | \in \pi_2) = \mathbb{P}(\mathbf{X} \in R_1 | \in \pi_2) = \int_{R_1} f_2(\mathbf{x}) d\mathbf{x}$$

A probabilidade desse segundo erro de classificação é o volume (integral) sob $f_2(\mathbf{x})$ na região R_1.

Fica mais fácil visualizar estas duas regiões associadas com as probabilidades das decisões incorretas se olharmos para o caso uni-dimensional, quando temos apenas uma única feature contínua X. As densidades $f_1(x)$ e $f_2(x)$, e R_1 e R_2 podem ser vistas no lado esquerdo da Figura 16.0.3. As duas probabilidades de classificação incorreta podem ser vistas no lado direito desta mesma Figura. A área em vermelho é a probabilidade de alocar a π_1 um exemplo vindo de $f_2(x)$ (e portanto, vindo de π_2). A área em verde é a probabilidade de alocar a π_2 um exemplo vindo de $f_1(x)$ (e portanto, vindo de π_1). Quando procuramos diminuir $\mathbb{P}(Class.\ em\ \pi_1 | \in \pi_2)$ estaremos aumentando $\mathbb{P}(Class.\ em\ \pi_2 | \in \pi_1)$. Existe um trade-off entre essas probabilidades de classificação incorreta. Como escolher uma boa partição R_1 e R_2 do espaço $\mathbb{R}^p$ das features? Como os dois erros possuem custos diferentes, nós vamos minimizar o *custo esperado de má classificação*.

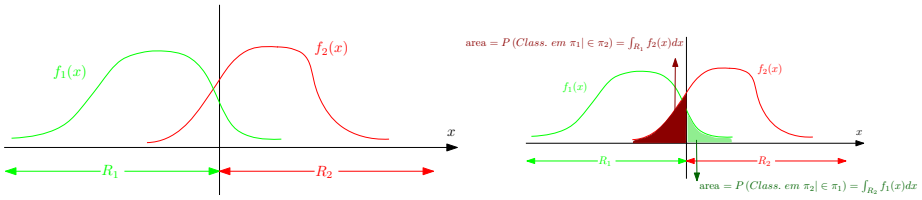


Figure 16.7: Colocar caption aqui

Nós temos também as probabilidades a priori de que os objetos venham de π_1 ou π_2: Seja $P_1 = \mathbb{P}(\in \pi_1)$ e $p_2 = \mathbb{P}(\in \pi_2) = 1 - \mathbb{P}(\in \pi_1) = 1 - p_1$. A Figura 16.0.3 mostra o quadro geral com os custos e as probabilidades de classificação.

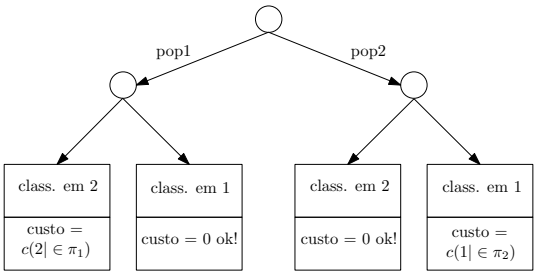


Figure 16.8: Colocar caption aqui

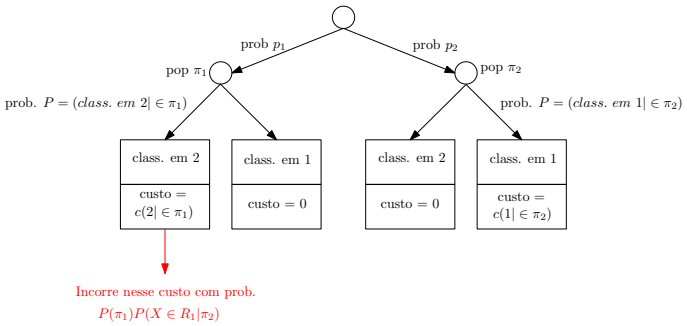


Figure 16.9: TEXTO VERMELHO ERRADO: Deveria ser $P(\pi_1)P(X \in R_2|\in\pi_1)$.

Casos novos chegam: alguns nós clasificamos corretamente (com custo zero); outros, classificamos incorretamente (com custo > 0). Podemos ter $c(2|\in\pi_1) \neq c(1|\in\pi_2)$ Nos casos em que erramos, às vezes caímos no custo mais elevado; às vezes, no custo menor. É impossível (nos casos realistas) ter uma regra baseada num vetor $\mathbf{x}$ que nunca erre. Queremos uma regra de classificação que, em geral (ou, em média) leve a um custo pequeno $\Rightarrow$ queremos um custo médio (ou esperado) pequeno.

16.0.4 EMC: Expected misclassification cost

Custo esperado (ou custo médio) de má-classificação. Custo é variável aleatória e possui três valores possíveis: 0, $c(2|\in\pi_1)$ e $c(1|\in\pi_2)$ Estes custos aleatórios acontecem com certas probabilidades. Qual seu valor esperado?

$$\begin{aligned} EMC &= \mathbb{E}(\text{cost}) \\ &= 0 \times \mathbb{P}(\text{acertar}) + \text{cost}_1 \times \mathbb{P}(\text{erro 1}) + \text{cost}_2 \times \mathbb{P}(\text{erro 2}) \\ &= c(2|\in\pi_1)\mathbb{P}(\text{vir de } \pi_1 \text{ e errar}) + c(1|\in\pi_2)\mathbb{P}(\text{vir de } \pi_2 \text{ e errar}) \\ &= c(2|\in\pi_1)\mathbb{P}(\mathbf{X}\in R_2|\in\pi_1)\mathbb{P}(\pi_1) + c(1|\in\pi_2)\mathbb{P}(\mathbf{X}\in R_1|\in\pi_2)\mathbb{P}(\pi_2) \end{aligned}$$

Assim, $EMC \rightarrow$ é custo esperado de má classificação. (expected misclassification cost). Queremos achar as regiões R_1 e R_2 que minimizam o EMC. Isto é equivalente a encontrar o classificador que torna o EMC o menor possível. Solução:

$$R_1 = \left\{ \mathbf{x}\in\mathbb{R}^p \text{ tais que } \underbrace{\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}}_{(1)} \geq \underbrace{\frac{c(1|\in\pi_2)}{c(2|\in\pi_1)}}_{(2)} \cdot \underbrace{\frac{p_2}{p_1}}_{(3)} \right\}$$

Veja que:
- (1): razão das densidades das duas classes
- (2): razão de custos
- (3): razão de probabilidades a priori

Prova: Queremos R_1 e R_2 que minimizem EMC:

$$EMC = c(2|\in\pi_1)\underbrace{\mathbb{P}(\mathbf{X}\in R_2|\in\pi_1)}_{\int_{R_2} f_1(\mathbf{x})\,d\mathbf{x}}\mathbb{P}(\pi_1) + c(1|\in\pi_2)\underbrace{\mathbb{P}(\mathbf{X}\in R_1|\in\pi_2)}_{\int_{R_1} f_2(\mathbf{x})\,d\mathbf{x}}\mathbb{P}(\pi_2)$$

Como $R_1 \cup R_2 = \mathbb{R}^p$ então

$$1 = \int_{\mathbb{R}^p} f_1(\mathbf{x})\,d\mathbf{x} = \int_{R_1} f_1(\mathbf{x})\,d\mathbf{x} + \int_{R_2} f_1(\mathbf{x})\,d\mathbf{x}.$$

Podemos escrever a primeira integral em EMC (em azul) da seguinte forma:

$$\int_{R_2} f_1(\mathbf{x})\,d\mathbf{x} = 1 - \int_{R_1} f_1(\mathbf{x})\,d\mathbf{x}$$

Vamos agora substituir a integral azul em EMC pela expressão em vermelho. Temos

$$EMC = c(2|\in\pi_1)\left(1 - \int_{R_1} f_1(\mathbf{x})\,d\mathbf{x}\right)\mathbb{P}(\pi_1) + c(1|\in\pi_2)\int_{R_1} f_2(\mathbf{x})\,d\mathbf{x})\,\mathbb{P}(\pi_2)$$

Agora, as duas integrais possuem a mesma região R_1 de integração e portanto os dois integrandos podem ser colocados sob o mesmo sinal de integral. Isto implica que

$$EMC = c(2|\in\pi_1)\mathbb{P}(\pi_1) + \int_{R_1} \left(c(1|\in\pi_2)\,\mathbb{P}(\pi_2)\,f_2(\mathbf{x}) - c(2|\in\pi_1)\,\mathbb{P}(\pi_1)\,f_1(\mathbf{x})\right)\,d\mathbf{x})$$

Queremos escolher R_1 de forma que EMC seja mínimo.

$$EMC = c(2|\in\pi_1)\mathbb{P}(\pi_1) + \int_{R_1} \underbrace{\left(c(1|\in\pi_2)\,\mathbb{P}(\pi_2)\,f_2(\mathbf{x}) - c(2|\in\pi_1)\,\mathbb{P}(\pi_1)\,f_1(\mathbf{x})\right)}_{h(\mathbf{x})}\,d\mathbf{x}$$

O 1^o termo não envolve R_1. Vamos olhar o 2^o termo. Escolher R_1 é escolher a região em que vamos integrar ("somar') $h(\mathbf{x})$. A expressão $h(\mathbf{x})$ não envolve R_1. Para alguns $\mathbf{x}$, teremos $h(\mathbf{x}) > 0$; para outros pontos $\mathbf{x}$, teremos $h(\mathbf{x}) < 0$. Para minimizar EMC, devemos tornar a integral o mais negativa possível. Conseguimos isto escolhendo $R_1 = \{\mathbf{x} \in \mathbb{R}^p \text{ tais que } h(\mathbf{x}) \leq 0\}$. Isso minimiza EMC!!

ECM é minimizado se escolhermos

$$R_1 = \{\mathbf{x} \in \mathbb{R}^p \text{ tais que } h(\mathbf{x}) \leq 0\}$$

$$\begin{aligned} EMC &= c(2|\in\pi_1)\mathbb{P}(\pi_1) + \int_{R_1} \underbrace{(c(1|\in\pi_2)\,\mathbb{P}(\pi_2)\,f_2(\mathbf{x}) - c(2|\in\pi_1)\,\mathbb{P}(\pi_1)\,f_1(\mathbf{x}))}_{h(\mathbf{x})}\,d\mathbf{x} \\ &= c(2|\in\pi_1)\mathbb{P}(\pi_1) + I \qquad \text{onde } I = \int_{R_1} h(\mathbf{x})\,d\mathbf{x} \end{aligned}$$

onde $I = \int_{R_1} h(\mathbf{x})\,d\mathbf{x}$.

Veja que $h(\mathbf{x}) \leq 0$ é o mesmo que

$$c(1|\in\pi_2)\,\mathbb{P}(\pi_2)\,f_2(\mathbf{x}) - c(2|\in\pi_1)\,\mathbb{P}(\pi_1)\,f_1(\mathbf{x}) \leq 0$$

Passando o segundo termo para o outro lado da desigualdade e as posições, temos

$$c(2|\in\pi_1)\,\mathbb{P}(\pi_1)\,f_1(\mathbf{x}) \geq c(1|\in\pi_2)\,\mathbb{P}(\pi_2)\,f_2(\mathbf{x})$$

ou ainda, após rearranjar os termos,

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c(1|\in\pi_2)}{c(2|\in\pi_1)}\,\frac{\mathbb{P}(\pi_2)}{\mathbb{P}(\pi_1)}$$

Para ficar em paz com esta afirmação, defina $R_1 = \{\mathbf{x} \in \mathbb{R}^p \text{ tais que } h(\mathbf{x}) \leq 0\}$ e seja E_1 o valor do ECM com esta regra de classificação . Se $\mathbf{x} \notin R_1$ então $h(\mathbf{x}) > 0$. Seja $R_1^* = A \cup R_1$ uma nova regra de classificação (com $A \cap R_1 = \emptyset$) com ECM dado por E_1^* Veremos que $E_1 \leq E_1^*$.

Temos

$$\begin{aligned} E_1^* &= c(2|\in\pi_1)\mathbb{P}(\pi_1) + \int_{R_1^*} h(\mathbf{x})\,d\mathbf{x} \\ &= c(2|\in\pi_1)\mathbb{P}(\pi_1) + \int_{R_1\cup A} h(\mathbf{x})\,d\mathbf{x} \\ &= c(2|\in\pi_1)\mathbb{P}(\pi_1) + \int_{R_1} h(\mathbf{x})\,d\mathbf{x} + \int_A \underbrace{h(\mathbf{x})}_{>0}\,d\mathbf{x} \\ &\geq c(2|\in\pi_1)\mathbb{P}(\pi_1) + \int_{R_1} h(\mathbf{x})\,d\mathbf{x} = E_1 \end{aligned}$$

Assim, aumentar a região R_1 com qualquer outra região A lnunca será capaz de fazer o ECM ser menor que aquele de R_1. Um argumento análogo, mostra que definir uma nova região para a classe 1 subtraindo uma área qualquer de R_1 também nunca leva a um ECM menor (exercício).

Resumo: Optimal Bayes Classifier

Em cada caso, observamos o vetor aleatório $\mathbf{X}$. Existem duas populações: π_1 e π_2. Um novo caso vem da pop π_1 com probabilidade p_1 e da pop π_2 com probabilidade $p_2 = 1 - p_1$. As densidades de $\mathbf{x}$: $f_1(\mathbf{x})$ e $f_2(\mathbf{x})$. Existem dois custos de classificação errada: $c(2|\in\pi_1)$ e $c(1|\in\pi_2)$ Baseado em $\mathbf{x}$, queremos predizer a sua classe: 1 ou 2. Cada regra de classificação tem seu ECM =

custo esperado de má-classificação (custo médio se classificarmos vários itens). Dentre todas as regras possíveis, aquela que torna mínimo o ECM é: aloque o caso a π_1 caso

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c(1|\in\pi_2)}{c(2|\in\pi_1)} \frac{\mathbb{P}(\pi_2)}{\mathbb{P}(\pi_1)}$$

Regra ótima de Bayes:

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \underbrace{\frac{c(1|\in\pi_2)}{c(2|\in\pi_1)} \frac{\mathbb{P}(\pi_2)}{\mathbb{P}(\pi_1)}}_{\text{cte. em } \mathbf{x}}$$

A regra é bem intuitiva, gera um algoritmo muito simples. Recebemos um novo caso com atributos no vetor $\mathbf{x}$. Qual sua classe? 1 ou 2? Calcule $f_1(\mathbf{x})/f_2(\mathbf{x})$, a razão de densidades no ponto $\mathbf{x}$ (aprox, é a razão das "probabilidades" de observar $\mathbf{x}$ em 1 ou 2). Se esta razão for "grande", aloque a 1. Razoável, não? De fato, se $f_1(\mathbf{x})/f_2(\mathbf{x}) \approx 7$, então a chance de observar $\mathbf{x}$ em 1 é aprox 7 vezes maior que a mesma chance em 2. Parece razoável alocar a 1. Mas...

Por quê não alocar a 1 simplesmente se tivermos $f_1(\mathbf{x})/f_2(\mathbf{x}) > 1$? Isto é, se $f_1(\mathbf{x}) > f_2(\mathbf{x})$, devemos alocar a 1? Nem sempre. Queremos levar em conta os custos e diferentes frequências das classes na população total. Como fazer isto? Basta calcular: (a) a razão de custos, (b) a razão de probabilidades *a priori*, e multiplicá-las. Este valor não depende de $\mathbf{x}$, é uma constante. A beleza não é porque a regra é muito simples. Qualquer um pode bolar uma regra simples. A beleza é que a regra é muito simples *e é a melhor possível e imaginável*. Nada pode ser melhor que ela (para reduzir o ECM).

■ **Example 16.1 — Exemplo uni-dimensional.** Imagine que π_1 é uma classe rara: $p_1 = 0.02$ É muito pior errar quando o item é de π_1: $c(2|\in\pi_1) = 40c(1|\in\pi_2)$. A regra é então alocar a 1 toda vez que $f_1(\mathbf{x})/f_2(\mathbf{x}) \geq (1/40) \times (0.98/0.02) = 1.22$. Ou seja, se $f_1(\mathbf{x}) \geq 1.22\ f_2(\mathbf{x})$, aloque a 1. Suponha que as duas densidades sejam como a seguir. Como encontrar a região de alocação a π_1? Veja a Figura 16.1. Na esquerda, temos o plot de $f_1(x)$ (preto) e $f_2(x)$ (azul). Na direita, temos $f_1(x)$ e $1.22 \times f_2(x)$. Aloque a 1 toda vez que a curva preta for maior que a curva azul neste plot da direita.

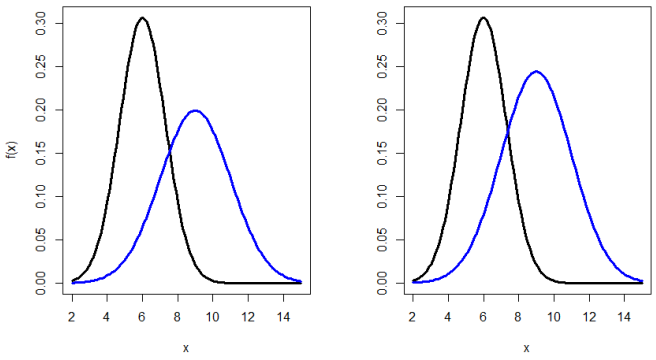


Figure 16.10: Completar caption

```
k = 1/40 * (1-0.02)/0.02

x = seq(2, 15, by=0.1)
```

```
y1 = dnorm(x, 6, 1.3)
y2 = dnorm(x, 9, 2)
par(mfrow=c(1,2))
plot(x, y1, type="l", lwd=3, ylab="f(x)")
lines(x, y2, col="blue", lwd=3)

plot(x, y1, type="l", lwd=3, ylab="")
lines(x, k*y2, col="blue", lwd=3)
```

■

16.0.5 Classificação ótima com duas gaussianas

Densidade de uma $N_p(\mu,\Sigma)$ no ponto $\mathbf{x}\in\mathbb{R}^p$:

$$\begin{aligned} f(\mathbf{x}) &= \underbrace{\left[(2\pi)^{-p/2}|\Sigma|^{-1/2}\right]}_{\text{constante em } \mathbf{x}} \exp\left(-\frac{1}{2}\underbrace{(\mathbf{x}-\mu)^t\,\Sigma^{-1}\,(\mathbf{x}-\mu)}_{\text{Mahalanobis}}\right) \\ &= k\,\exp\left(-\frac{1}{2}d^2_\Sigma(\mathbf{x},\mu)\right) \end{aligned}$$

Queremos a razão de duas densidades gaussianas, $f_1(\mathbf{x})\sim N_p(\mu_1,\Sigma_1)$ e $f_2(\mathbf{x})\sim N_p(\mu_2,\Sigma_2)$, no mesmo ponto $\mathbf{x}$:

$$\begin{aligned} \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} &= \frac{k_1\,\exp\left(-\frac{1}{2}d^2_{\Sigma_1}(\mathbf{x},\mu_1)\right)}{k_2\,\exp\left(-\frac{1}{2}d^2_{\Sigma_2}(\mathbf{x},\mu_2)\right)} \\ &= \frac{k_1}{k_2}\exp\left(-\frac{1}{2}(d^2_{\Sigma_1}(\mathbf{x},\mu_1)-d^2_{\Sigma_2}(\mathbf{x},\mu_2))\right) \end{aligned}$$

A regra ótima é: aloque a π_1 se

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c(1|\in\pi_2)}{c(2|\in\pi_1)}\,\frac{\mathbb{P}(\pi_2)}{\mathbb{P}(\pi_1)}$$

Portanto, aloque a π_1 se

$$\frac{k_1}{k_2}\exp\left(-\frac{1}{2}(d^2_{\Sigma_1}(\mathbf{x},\mu_1)-d^2_{\Sigma_2}(\mathbf{x},\mu_2))\right) \geq \frac{c(1|\in\pi_2)}{c(2|\in\pi_1)}\,\frac{\mathbb{P}(\pi_2)}{\mathbb{P}(\pi_1)}$$

Tomando logs dos dois lados e mudando de lado alguns termos, temos que alocar a π_1 se:

$$d^2_{\Sigma_1}(\mathbf{x},\mu_1) \leq d^2_{\Sigma_2}(\mathbf{x},\mu_2) + 2\left[\log\left(\frac{c(2|\in\pi_1)}{c(1|\in\pi_2)}\right) + \log\left(\frac{\mathbb{P}(\pi_1)}{\mathbb{P}(\pi_2)}\right) + \log\left(\frac{k_2}{k_1}\right)\right]$$

Vamos entender um pouco melhor esta fórmula.

Repetindo, alocar a π_1 se:

$$\underbrace{d^2_{\Sigma_1}(\mathbf{x},\mu_1)}_{\text{Mahalanobis}} \leq \underbrace{d^2_{\Sigma_2}(\mathbf{x},\mu_2)}_{\text{Mahalanobis}} + \underbrace{2\log\left(\frac{c(2|\in\pi_1)}{c(1|\in\pi_2)}\right)}_{\text{cost}} + \underbrace{2\log\left(\frac{\mathbb{P}(\pi_1)}{\mathbb{P}(\pi_2)}\right)}_{\text{prioris}} + \underbrace{2\log\left(\frac{k_2}{k_1}\right)}_{\text{covariances}}$$

Ideia: aloque a π_1 se a distância de Mahalanobis de $\mathbf{x}$ a μ_1 for menor que a distância de Mahalanobis a μ_2 mais ou menos "alguma coisa". O "alguma coisa" leva em conta os custos, prioris e estruturas

de covariância de cada população. Esta fórmula mostra a *melhor* maneira de levar estes aspectos em conta. Por exemplo, os custos devem ser analisados em função de sua diferença *relativa* e numa escala log. Por exemplo, não é a diferença $c(2|\in\pi_1)-c(1|\in\pi_2)$ que nos interessa, mas sim $c(2|\in\pi_1)/c(1|\in\pi_2)$.

Repetindo, alocar a π_1 se (ETC): O termo envolvendo os custos dos dois erros desaparece se eles forem iguais. Aquele envolvendo as probabilidades *a priori* também desaparece se $\mathbb{P}(\pi_1)=\mathbb{P}(\pi_2)$. Do mesmo modo, se $\Sigma_1=\Sigma_2$, o último termo desaparece. Neste caso em que todos estes termos desaparecem, a regra ótima simplesmente compara as distâncias de Mahalanobis": alocar a π_1 se

$$d^2_\Sigma(\mathbf{x},\mu_1)\leq d^2_\Sigma(\mathbf{x},\mu_2)$$

Vamos começar a introduzir os termos adicionais, um de cada vez, para entender seu efeito. Suponha que os custos sejam diferentes e que um deles é 100 vezes maior que o outro: $c(2|\in\pi_1)=100c(1|\in\pi_2)$. Isto é, é 100 mais pior alocar um caso de π_1 erradamente que alocar errado um caso de π_2. Deveríamos ser *menos propensos* então a alocar um caso a π_2. De fato, neste caso, a regra ótima é alocar $\mathbf{x}$ a π_1 se

$$d^2_\Sigma(\mathbf{x},\mu_1)\leq d^2_\Sigma(\mathbf{x},\mu_2)+2\log(100)$$

Agora ficou mais fácil alocar um caso a π_1. A distância de Mahalanobis a π_1 nem precisa ser a menor delas agora. Lembre-se: esta é a regra ótima, a melhor possível.

Do mesmo modo, suponha que a classe 1 seja 100 vezes mais frequente que a classe 2: $\mathbb{P}(\pi_1)=100\mathbb{P}(\pi_2)$. Isto é, quando um caso qualquer aparece, sem considerar o valor de $\mathbf{x}$, sabemos que é 100 mais provável que ele seja de π_1 do que de π_2. Novamente, deveríamos ser *menos propensos* então a alocar um caso a π_2. Como antes, a regra ótima é alocar $\mathbf{x}$ a π_1 se

$$d^2_\Sigma(\mathbf{x},\mu_1)\leq d^2_\Sigma(\mathbf{x},\mu_2)+2\log(100)$$

Por último, para ver o efeito do terceiro termo, vamos imaginar que a variável $\mathbf{x}$ é unidimensional. Assim, $\Sigma_1=\sigma_1^2$ e $\Sigma_2=\sigma_2^2$, as variâncias de X em cada população. Além disso, a distância de Mahalanobis reduz-se a $d^2_\Sigma(\mathbf{x},\mu)=((x-\mu)/\sigma)^2$, o desvio padronizado. Suponha que uma das populações tenha variância muito maior que a outra: $\sigma_2^2=(10)^2\sigma_1^2$. Isto é, pontos de π_2 espalham-se em torno de sua média μ_2 muito mais que pontos de π_1 em torno de μ_1. A regra ótima é alocar $\mathbf{x}$ a π_1 se

$$\left(\frac{x-\mu_1}{\sigma_1}\right)^2\leq\left(\frac{x-\mu_2}{\sigma_2}\right)^2+\log(100)$$

Assim, penalizamos a população com maior dispersão. Pense assim, se as duas distâncias padronizadas forem iguais, o melhor é alocar à π_1, a menos dispersa.

Caso gaussiano com $\Sigma_1=\Sigma_2$ Alocar a π_1 se:

$$d^2_{\Sigma_1}(\mathbf{x},\mu_1)-d^2_{\Sigma_2}(\mathbf{x},\mu_2)\leq 2\log\left(\frac{c(2|\in\pi_1)\mathbb{P}(\pi_1)}{c(1|\in\pi_2)\mathbb{P}(\pi_2)}\right)+2\log\left(\frac{k_2}{k_1}\right)$$

Se $\Sigma_1=\Sigma_2$, os seus determinantes também são iguais e $k_2=k_1$. Vamos denotar a constante (em $\mathbf{x}$) do lado esquerdo, envolvendo as probabilidades a priori e os custos de K:

$$K=\log\left(\frac{c(2|\in\pi_1)\mathbb{P}(\pi_1)}{c(1|\in\pi_2)\mathbb{P}(\pi_2)}\right)$$

A regra ótima no caso gaussiano com covariâncias iguais é alocar a π_1 se

$$d^2_{\Sigma_1}(\mathbf{x},\mu_1)-d^2_{\Sigma_2}(\mathbf{x},\mu_2)\leq 2K$$

Alocar a π_1 se $d^2_{\Sigma_1}(\mathbf{x},\mu_1) - d^2_{\Sigma_2}(\mathbf{x},\mu_2) \leq 2K$ Expandimos a expressão da distância:

$$\begin{aligned} d^2_{\Sigma_1}(\mathbf{x},\mu_1) &= (\mathbf{x}-\mu_1)^t\ \Sigma_1^{-1}\ (\mathbf{x}-\mu_1) \\ &= \mathbf{x}^t\Sigma_1^{-1}\mathbf{x} + \mu_1^t\Sigma_1^{-1}\mu_1 - 2\mathbf{x}^t\Sigma_1^{-1}\mu_1 \end{aligned}$$

Do mesmo modo, expandimos a outra distância. Cancelamos o termo $\mathbf{x}^t\Sigma_1^{-1}\mathbf{x}$ encontrando: aloque $\mathbf{x}$ a π_1 se

$$\begin{aligned} 0 &\leq \mathbf{x}^t \underbrace{\Sigma_1^{-1}(\mu_1-\mu_2)}_{p\times 1,\quad \beta} + \underbrace{(\mu_2^t\Sigma_2^{-1}\mu_2 - \mu_1^t\Sigma_1^{-1}\mu_1 - K)}_{1\times 1,\quad \alpha} \\ &= \alpha + \beta^t\mathbf{x} \end{aligned}$$

usando que $\beta^t\mathbf{x} = \mathbf{x}^t\beta$.

Vamos escrever $\lambda(\mathbf{x}) = \alpha + \beta^t\mathbf{x}$. O conjunto de pontos $\mathbf{x} \in \mathbb{R}^p$ tais que $\lambda(\mathbf{x}) = 0$ constitui a fronteira de decisão (decision boundary). No caso em que $\mathbf{x} \in \mathbb{R}^2$, esta fronteira é uma linha reta. Se $\mathbf{x} \in \mathbb{R}^3$, a fronteira é um plano. Passar para o notebook python para ilustrar.

Caso gaussiano com $\Sigma_1 \neq \Sigma_2$ E o caso gaussiano com $\Sigma_1 \neq \Sigma_2$? Manipulação matricial da regra ótima leva à conclusão de que a fronteira de decisão é uma parábola, e não mais uma reta. A fórmula geral do caso gaussiano, como já sabemos, alocar a π_1 se:

$$d^2_{\Sigma_1}(\mathbf{x},\mu_1) - d^2_{\Sigma_2}(\mathbf{x},\mu_2) \leq 2\log\left(\frac{c(2|\in\pi_1)\mathbb{P}(\pi_1)}{c(1|\in\pi_2)\mathbb{P}(\pi_2)}\right) + 2\log\left(\frac{k_2}{k_1}\right)$$

Expandindo as fórmulas quadráticas das distâncias, exatamente como fizemos antes, leva a uma expressão simples.

Alocar $\mathbf{x}$ a π_1 se:

$$\mathbf{x}^t\mathbf{A}\mathbf{x} - 2\beta^t\mathbf{x} + \alpha \leq 0$$

com $\mathbf{A}$ sendo uma matriz $p \times p$, β sendo um vetor coluna $p \times 1$ e α sendo um escalar (um número real). Mais especificamente,

$$\mathbf{A} = \Sigma_1^{-1} - \Sigma_2^{-1}$$

$$\beta = \Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2$$

A constante α é uma expressão um pouco mais longa:

$$\alpha = \left(\mu_1^t\Sigma_1^{-1}\mu_1 - \mu_2^t\Sigma_2^{-1}\mu_2\right) - 2\log\left(\frac{c(2|\in\pi_1)\mathbb{P}(\pi_1)}{c(1|\in\pi_2)\mathbb{P}(\pi_2)}\right) - 2\log\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right)$$

Caso gaussiano bi-dimensional, com $\Sigma_1 \neq \Sigma_2$ No caso em que $\mathbf{x} \in \mathbb{R}^2$, a fronteira ótima de Bayes é uma forma quadrática em x_1 e x_2. Isto é, a fronteira de decisão $\lambda(\mathbf{x}) = 0$ (o conjunto de pontos que separa as duas classes) será uma expressão do seguinte tipo:

$$c_1x_1^2 + c_2x_2^2 + c_3x_1x_2 + c_4x_1 + c_5x_2 + c_6 = 0$$

onde as constantes c_j são determinadas pelos parâmetros das duas distribuições, pelos custos e pelas probabilidades a priori. A fronteira $\lambda(\mathbf{x}) = 0$ costuma ser uma curva que lembra o formato de uma parábola (mas não é exatamente uma parábola). Ver notebook python para exemplos.

Pros e Cons of Optimal Bayes Classifier

- Precisa conhecer a densidades.
- Se não conhecer, precisa estimá-las e então terá erro de estimacao
- Em princípio: estimar via kde (kernel). Entao, a dificuldade será encontrar a regiao em espacos multi-dimensionais.
- Eh otima apenas para ECM. Não otimiza para outras funções objetivo.
- Vantagens: otima; simples; intuitiva;

17. Linear Discriminant Analysis

Em 1936, (Sir) Ronald A. Fisher, o maior estatístico que já existiu e um dos maiores geneticistas do mundo, criou uma regra de classificação muito popular até hoje. É chamada de LDA: Linear Discriminant Analysis. Mas, espere um pouco... Se temos a regra ótima (Optimal Bayes Classifier), por quê aprender uma regra antiga e diferente? Por dois motivos: (a) o LDA está conectado com a regra ótima. (b) ele fornece uma abordagem bem diferente para o nosso problema de classificação: ele o vê como um problema de *separação* de populações.

Linear Discriminant Analysis (LDA) para duas classes. Seja $\mathbf{X} = (X_1, X_2, \ldots, X_p)$. Fisher: Vamos criar um índice univariado (escalar) calculando

$$\begin{aligned} Y &= \ell^t \mathbf{X} \\ &= \ell_1 X_1 + \ell_2 X_2 + \ldots + \ell_p X_p \end{aligned}$$

onde $\ell^t = (\ell_1, \ell_2, \ldots, \ell_p)$ é um vetor de constantes. Queremos escolher o vetor ℓ de forma que:

$$\begin{cases} Y's \text{ de pop1} \\ Y's \text{ de pop2} \end{cases} \longrightarrow \text{o mais separado possivel}$$

Suponha que $\|\ell\|^2 = 1$ (vetor de comprimento 1) Estamos buscando direção ℓ em que $\ell^t \mathbf{X} = Y$ dos dois grupos sejam maximalmente separados O vetor ℓ acima é uma má escolha!

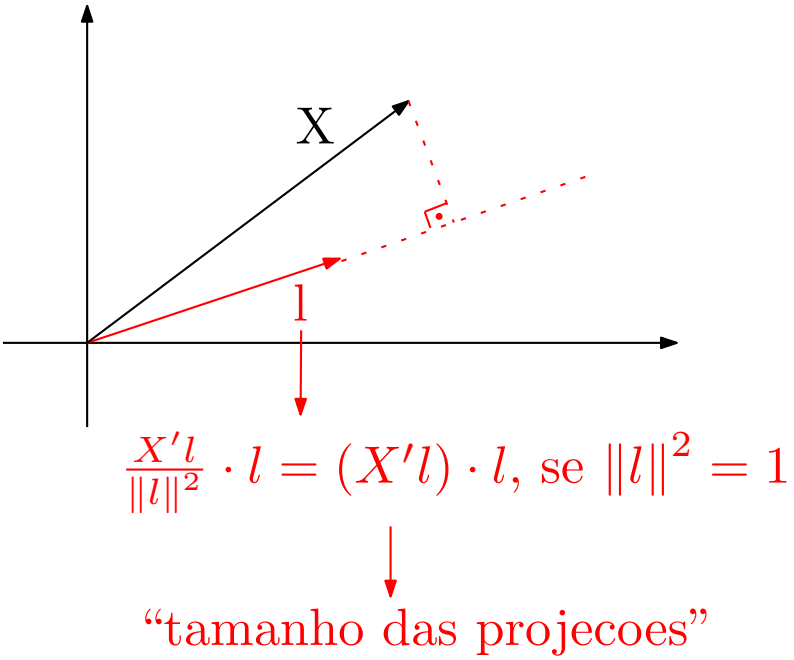
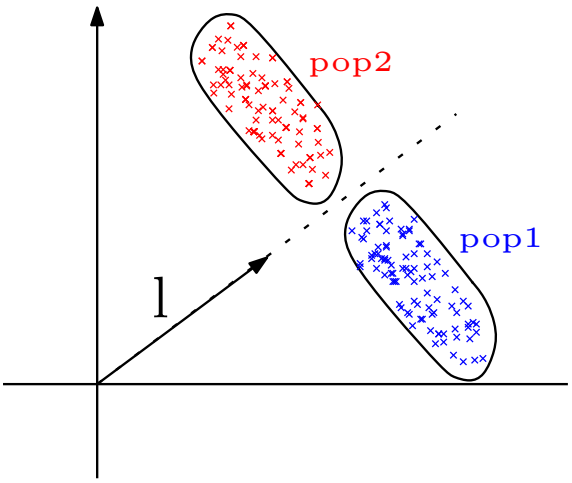


Figure 17.1: Projeção ortogonal de $\mathbf{X}$ em ℓ:



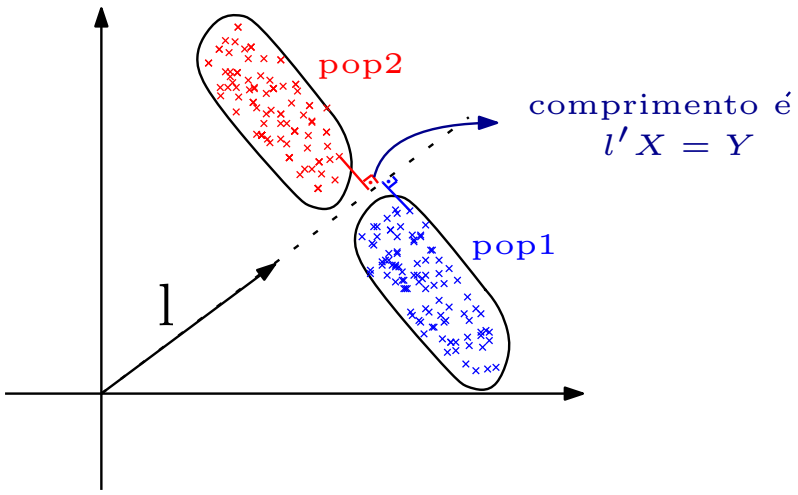


Figure 17.2: Projeção $Y = \ell^t \mathbf{X}$ de dois vetores $\mathbf{X}$: um de pop1, outro de pop2.

18. Iterated and Conditional Expectation

18.1 Introduction

In this chapter, we cover several important topics. We begin by showing how to transform the conditional expectation $\mathbb{E}(Y \mid X = x)$ of Y given $X = x$ into a random variable. In the usual setting, the random variable X is held fixed at a particular value x, and from this we derive $\mathbb{E}(Y \mid X = x)$. However, our real interest lies in examining the behavior of Y conditioned on X *without fixing X to a specific value*. In that situation, the conditional expectation itself becomes a random variable.

This viewpoint opens the door to a range of techniques for handling problems defined hierarchically, such as those involving sequential random experiments. We will dedicate the first three sections to laying out these foundational ideas. Although their immediate utility might not be evident, I encourage you to continue; in the later sections, we will present two major results, often referred to as the theorems of iterated expectations and iterated variances. They give this chapter its name.

We start by reviewing conditional expectation and variance. In Sections 12.13 and 12.14, we explored the definitions of conditional expectation and variance for discrete random variables. The continuous case was addressed in sections 12.19 and 12.20. If we have a bivariate continuous random vector (X,Y), we define the conditional distribution of Y given that X takes the value x, denoted as the distribution of the random variable $(Y|X = x)$, by its probability density function:

$$f_{Y|X=x}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}.$$

The discrete case has a similar formula (see Section 12.13). For example, given that $X = 4$ was observed, we have

$$f_{Y|X=4}(y|4) = \frac{f_{XY}(4,y)}{f_X(4)}.$$

Note that this conditional density with $X = 4$ is an univariate function depending only on y. The conditional expectation $m(x) = \mathbb{E}(Y|X = x)$ in the continuous case is given by:

$$m(x) = \mathbb{E}(Y|X = x) = \int y f_{Y|X=x}(y|x)dy.$$

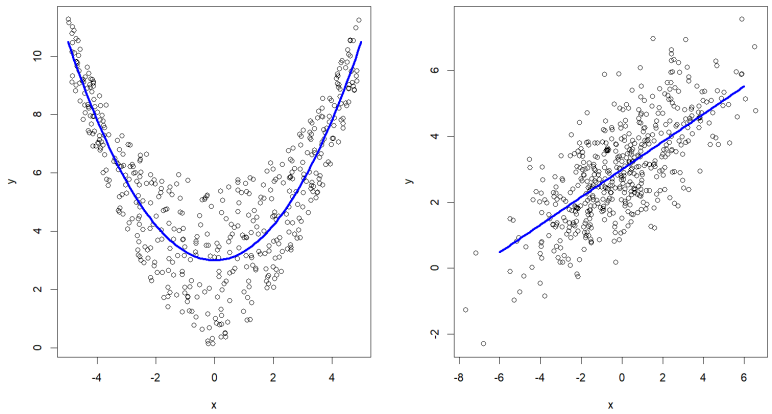


Figure 18.1: *Samples from two different joint density $f(x,y)$. The blue line represents the conditional expectation $m(x) = \mathbb{E}(Y|X=x)$.*

For example, with $X = 4$, we have

$$m(4) = \mathbb{E}(Y|X=4) = \int y f_{Y|X=4}(y|4)dy\,.$$

Likewise, the conditional variance $v(x) = \mathbb{E}(Y|X=x)$ in the continuous case is given by:

$$v(x) = \mathbb{V}(Y|X=x) = \int (y-m(x))^2 f_{Y|X=x}(y|x)dy = \int y^2 f_{Y|X=x}(y|x)dy - m^2(x)\,.$$

For example, with $X = 4$, we have

$$v(4) = \mathbb{V}(Y|X=4) = \int (y-m(4))^2 f_{Y|X=4}(y|4)dy = \int y^2 f_{Y|X=4}(y|4)dy - m^2(4)\,.$$

The calculation of conditional distributions and of conditional expectations and variances are difficult in general. The easy calculation in the multivariate Gaussian case is a rather special situation, rarely seen in other distributions. However, when we have a large random sample of the random vector (X,Y), we can relatively easily obtain these conditional distributions, along with their conditional mean and variance. For example, consider the two scatterplots of random samples of two continuous random vector (X,Y) in Figure 18.1. In the left-hand side plot we observe that, for each specific x value, the random variable Y appears to be uniformly distributed within an interval whose center and width depends on x. Given that the expected value of an uniform distribution is the center of its support interval, we guess that $\mathbb{E}(Y|X=2.0) \approx 4.0$ and $\mathbb{E}(Y|X=4.0) \approx 8.0$. The blue line represents the curve $m(x) = \mathbb{E}(Y|X=x)$. Hence, $m(2) \approx 4$ and $m(4) \approx 8$.

Not only the conditional expectation $m(x) = \mathbb{E}(Y|X=x)$ is a function of x, but so is the standard deviation $SD(x) = \sqrt{v(x)} = \sqrt{\mathbb{V}(Y|X=x)}$. The dispersion of the random variable $(Y|X=x)$ around its conditional mean $m(x)$ is much larger when $x = 0$ than at the extremes, $x \approx -5$ and $x \approx 5$. We have learned before how to guess the standard deviation based on a random sample. Applying that to this case, we can say that $SD(0) \approx 2$ while $SD(-5) \approx 0.3 \approx SD(5)$.

In the right-hand plot of Figure 18.1, the data was generated using a bivariate Gaussian distribution:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}\right) = \mathcal{N}\left(\begin{pmatrix} 0 \\ 3 \end{pmatrix}, \begin{pmatrix} (2.5)^2 & 0.7(2.5)(1.5) \\ 0.7(2.5)(1.5) & (1.5)^2 \end{pmatrix}\right)$$

As discussed in Chapter 13, any conditional distribution $(Y|X=x)$ of a bivariate Gaussina distribution follows an univariate Gaussian distribution:

$$(Y|X=x) \sim \mathcal{N}\left(\mu_y + \frac{\rho\sigma_y}{\sigma_x}(x-\mu_x), \quad \sigma_y^2(1-\rho^2)\right).$$

The blue line represents the conditional mean

$$m(x) = \mathbb{E}(Y|X=x) = \mu_y + \frac{\rho\sigma_y}{\sigma_x}(x-\mu_x) = 3 + \frac{(0.7)(1.5)}{2.5}x = 3 + 0.42x.$$

Unlike the left-hand plot, this is a linear function of x. Also in contrast to the first case, the conditional variance (and standard deviation) does not vary with x. It is equal to $\mathbb{V}((Y|X=x) = \sigma_y^2(1-\rho^2) = (0.82)(1.5)^2$ for all x.

There are two main results in this chapter. The first links the marginal expectation $\mathbb{E}(Y)$ to the conditional expectations $m(x) = \mathbb{E}(Y \mid X=x)$. In Section 18.5, we will see that $\mathbb{E}(Y)$ can be expressed as a weighted average of the $m(x)$ values, with the weights given by the probabilities of the different X values.

The second main result demonstrates how to derive the marginal variance $\mathbb{V}(Y)$ from the conditional expectation $m(x)$ and the conditional variance $v(x) = \mathbb{V}(Y \mid X=x)$. In essence, $\mathbb{V}(Y)$ is the sum of how much $m(x)$ varies with x and of what we can expect of $v(x)$. This less obvious formula will be addressed in Section 18.6.

18.2 Conditional expectation as a random variable

This introductory discussion is to prepare a simple but very important step. To take this step, consider a very simple example. Let (X,Y) be a discrete random vector with X representing the count of plant leaves showing signs of disease or damage, which could be related to the plant's overall health. Y represents the count of insect larvae found on the plant, which may be influenced by the health of the plant or the extent of infestation in the field. Plants with more discolored or damaged leaves (larger values of X) might be more susceptible to insect infestation, resulting in a higher number of insect larvae Y. To simplify, we will assume that X is limited to only three possible values: $X=0$: No discolored or damaged leaves, $X=1$: One discolored or damaged leaf, and $X=2$: Two or more discolored or damaged leaves. The random variable Y has five possible values: $Y=0$: No larvae found; $Y=1$: One larva found, up to $Y=4$ representing four larvae found (no leaf could contain more than four larvae).

The marginal distribution $p_x = \mathbb{P}(X=x)$ of X is given in Table 18.1, while Table 18.2 shows the conditional probabilities $p_{y|x} = \mathbb{P}(Y=y|X=x)$ of the conditional distributions $(Y|X=x)$:

x	p_x
0	$p_0 = 0.80$
1	$p_1 = 0.15$
2	$p_2 = 0.05$
sum	1

Table 18.1: Marginal, X

y	$p_{(y\|0)}$	$p_{(y\|1)}$	$p_{(y\|2)}$
0	0.50	0.20	0.05
1	0.30	0.20	0.05
2	0.15	0.20	0.20
3	0.05	0.20	0.35
4	0.00	0.20	0.35
sum	1	1	1

Table 18.2: Conditional $p_{y|x} = \mathbb{P}(Y=y|X=x)$

We can easily obtain the expected value $m(x) = \mathbb{E}(Y|X=x)$ of Y conditioned on each of the number of diseased leaves:

- $m(0) = 0*0.50+1*0.30+2*0.15+3*0.05+4*0.00 = 0.75$. On average, we find 0.75 larvae when the plant has no diseased leaves.
- $m(1) = 0*0.20+1*0.20+2*0.20+3*0.20+4*0.20 = 2.00$. On average, we find 2 larvae when the plant has one diseased leaf.
- $m(2) = 0*0.05+1*0.05+2*0.20+3*0.35+4*0.35 = 2.90$. On average, we find almost three larvae when the plant has two diseased leaves.

We know that the expectation of a random variable is simply a real number. For example, if $\mathbb{E}(X) = 0 \times 0.80 + 1 \times 0.15 + 2 \times 0.05 = 0.25$, then $\mathbb{E}(X) = 0.25$. Because $\mathbb{E}(X)$ is a single number, not a random variable, it does not come with its own set of possible values or associated probabilities.

Similarly, the conditional expectation $m(x) = \mathbb{E}(Y \mid X = x)$ is also just a real number. However, unlike $\mathbb{E}(X)$, $m(x)$ depends on x. In the example from the table, we have $m(0) = 0.75$, $m(1) = 2.00$, and $m(2) = 2.90$. This makes sense: because the value taken by X usually influences the entire distribution of Y, it naturally affects the mean of Y.

Once X is fixed at x, $\mathbb{E}(Y|X = x) = m(x)$ is just a single numeric value. To evaluate this function $m(x) = \mathbb{E}(Y|X = x)$, we:

- Fix X at its realized value x,
- Compute the expected value of Y under the condition $X = x$.

However, we can also think of the function $m(x)$ evaluated at a random value of X. That is, we can conceive of the random value $m(X)$. The difference between $m(x)$ and $m(X)$ is that $m(X)$ is a random variable. This implies that $m(X)$ has two lists: a list of possible values and a list of associated probabilities. You can imagine that the value of the random variable X will soon be known and will take on a specific value x. When this happens, you will refer to the expected value $m(x) = \mathbb{E}(Y|X = x)$ for Y under the condition that $X = x$ has occurred. Thus, we can conceive of an experiment in which the random outcome is an expected value for Y as long as we know the value that X has taken. However, before the value of X is known, we can think of the possible values of the expected mean. Let us consider the previous example in detail.

In the discrete example we are considering, the random variable $m(X)$ has two lists. The first list, of possible values, consists of the three values $m(0) = 0.75$, $m(1) = 2.00$, and $m(2) = 2.90$ from $m(x) = \mathbb{E}(Y|X = x)$. Before the value of X is known, we can say that, if we eventually observe $X = 0$, we can expect $m(0) = 0.75$ for Y. However, if $X = 1$ happens, we can expect $m(1) = 2.00$ for Y. Finally, $m(2) = 2.90$ if $X = 2$.

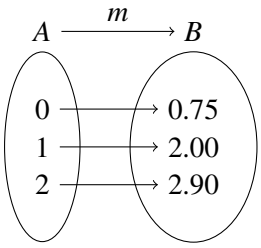
These three possible values of the conditional expectation $\mathbb{E}(Y|X = x)$ do not occur with equal probability. Because $\mathbb{P}(X = 0) = 0.80$ is much larger than $\mathbb{P}(X = 1) = 0.15$ or $\mathbb{P}(X = 2) = 0.05$, it is far more likely that our experiment will result in a conditional expectation $m(x) = \mathbb{E}(Y \mid X = x)$ equal to $m(0)$ rather than $m(1)$ or $m(2)$.

The random variable $m(X) = E(Y \mid X)$ is discrete, taking the three possible values $m(0)$, $m(1)$, and $m(2)$. Each of these values has an associated probability, as shown here:

$$\mathbb{E}(Y \mid X) = \begin{cases} m(0) = 0.75 & \text{with probability } 0.80, \\ m(1) = 2.00 & \text{with probability } 0.15, \\ m(2) = 2.90 & \text{with probability } 0.05. \end{cases} \tag{18.1}$$

Note that these probabilities are simply the marginal probabilities $\mathbb{P}(X = x)$.

Let us emphasize, at the risk of being overly repetitive, the important conceptual shift behind this seemingly small notational difference. The objects $m(x) = \mathbb{E}(Y|X = x)$ and $m(X) = \mathbb{E}(Y|X)$ are very different. The former represents a mathematical function with the domain $A = \{0, 1, 2\}$ and the image range $B = \{0.75, 2.00, 2.90\}$.



The notation $m(X)$, using a capital X instead of a lowercase x, represents a very different object. $m(X)$ is a random variable and, following our mantra, is described by two lists. These two lists, possible values and associated probabilities, are represented in (18.1).

18.3 Continuous Case

These ideas naturally extend to the continuous case. Let us revisit the two scatterplots of the random sample (X,Y) in Figure 18.1. The blue curves in these plots depict $m(x) = \mathbb{E}(Y \mid X = x)$. To generate the left-hand plot, we proceed in two steps. First, we sample $X \sim U(-5,5)$. Then, given the realized value of $X = x$, we sample $(Y \mid X = x)$ from a uniform distribution that depends on x. Specifically,

$$(Y \mid X = x) \sim U\big(m(x) - \Delta_x,\, m(x) + \Delta_x\big),$$

where $m(x) = 3 + 0.3x^2$ and $\Delta_x = 3 - 2|x|/5$. Thus, if $X = 2$ is drawn, then $(Y \mid X = 2) \sim U(4.2 \pm 2.2)$. If $X = 4.5$ occurs, then $(Y \mid X = 4.5) \sim U(9.1 \pm 1.2)$.

Next, consider the random variable $m(X) = \mathbb{E}(Y \mid X)$. Its possible values are the ordinates of the curve $m(x)$, ranging over $[3 + 0.3 \cdot 0^2,\ 3 + 0.3 \cdot 5^2] = [3,\ 10.5]$. To find the probabilities associated with these values, fix a point t in this interval. We derive the cumulative distribution of the random variable $m(X)$ and then differentiate to obtain its probability density function:

$$\mathbb{F}_{m(X)}(t) \;=\; \mathbb{P}\big(m(X) \le t\big) \;=\; \mathbb{P}\big(3 + 0.3X^2 \le t\big) \;=\; \mathbb{P}\Big(X \in \Big[-\sqrt{\tfrac{t-3}{0.3}},\, \sqrt{\tfrac{t-3}{0.3}}\Big]\Big) \;=\; \frac{2}{10}\sqrt{\frac{t-3}{0.3}},$$

because $X \sim U(-5,5)$. For t outside this interval, we have $\mathbb{F}_{m(X)}(t) = 0$ if $t < 3$ and $\mathbb{F}_{m(X)}(t) = 1$ if $t > 10.5$. Differentiating with respect to t yields

$$\mathfrak{O}_{m(X)}(t) \;=\; \frac{1}{3}\sqrt{\frac{0.3}{t-3}} \quad \text{for } t \in [3,\, 10.5],$$

and $\mathfrak{O}_{m(X)}(t) = 0$ otherwise. The left-hand plot in Figure 18.2 shows this density function over its support.

The right-hand scatterplot in Figure 18.1 was simulated from the bivariate Gaussian distribution:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}\right) = \mathcal{N}\left(\begin{pmatrix} 0 \\ 3 \end{pmatrix}, \begin{pmatrix} (2.5)^2 & 0.7(2.5)(1.5) \\ 0.7(2.5)(1.5) & (1.5)^2 \end{pmatrix}\right)$$

and therefore $(Y \mid X = x)$ is also a Gaussian with conditional mean

$$m(x) = \mathbb{E}(Y|X = x) = \mu_y + \frac{\rho\sigma_y}{\sigma_x}(x - \mu_x) = 3 + \frac{(0.7)(1.5)}{2.5}x = 3 + 0.42x\,.$$

Turning our attention to the *random variable* $m(X) = \mathbb{E}(Y \mid X) = 3 + 0.42X$, we want to determine its distribution. Because X is a continuous random variable, $m(X)$ is also continuous. In

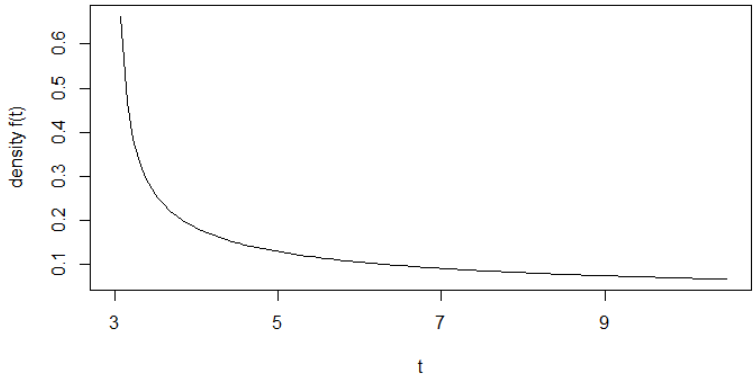


Figure 18.2: *Probability density function of the random variables $m(X) = \mathbb{E}(Y|X)$ for the distributions illustrated in Figure 18.1.*

fact, since $m(X)$ is a linear transformation of the Gaussian random variable X, it too follows also a Gaussian distribution. Its parameters are straightforward to calculate:

$$\mathbb{E}[m(X)] \;=\; \mathbb{E}[3+0.42X] \;=\; 3+0.42\,\mathbb{E}(X) \;=\; 3$$

and

$$\mathbb{V}[m(X)] = \mathbb{V}[3+0.42X] = (0.42)^2\,(2.5)^2 = (1.05)^2\,.$$

Hence, $m(X) \sim \mathcal{N}\big(3,\,(1.05)^2\big)$.

18.4 The expectation of the conditional expectation

Since $m(X)$ is a random variable, it is legitimate to ask about its properties. For instance, what is the expected value of the random variable $m(X)$? In other words, what is $\mathbb{E}[m(X)] = \mathbb{E}\big[\mathbb{E}(Y \mid X)\big]$? Although we will clarify the notation shortly, it might be surprising that we are led back to the original variable Y:

$$\mathbb{E}[m(X)] = \mathbb{E}\big[\mathbb{E}(Y \mid X)\big] = \mathbb{E}(Y)\,.$$

The expected value of $\mathbb{E}(Y \mid X)$ is just the expected value $\mathbb{E}(Y)$. This result is formally proved in Theorem 18.5.1 in the next section. Before delving deeper into it, we will verify its validity through a few examples. Let us start with the extended example from Section 18.2, which involves diseased plants' leaves.

■ **Example 18.1 — Diseased Plants.** In Section 18.2, we calculated $m(x) = \mathbb{E}(Y|X=x)$ for $x = 0, 1, 2$ finding:

- $m(0) = \mathbb{E}(Y|X=0) = 0.75$
- $m(1) = \mathbb{E}(Y|X=1) = 2.00$
- $m(2) = \mathbb{E}(Y|X=2) = 2.90$

Now, consider the discrete random variable $m(X) = \mathbb{E}(Y|X)$. As a function of X, it also has only three possible values and the associated probability list is given in (18.1) and repeated here for convenience:

$$\mathbb{E}(Y|X) = \begin{cases} 0.75, & \text{with probab } 0.80 \\ 2.00, & \text{with probab } 0.15 \\ 2.90, & \text{with probab } 0.05 \end{cases}$$

As a discrete random variable, its expected value is simply the weighted average of its possible value:

$$\mathbb{E}[\mathbb{E}(Y|X)] = 0.75 * 0.80 + 2.00 * 0.15 + 2.90 * 0.05 = 1.045$$

We will show that this is exactly equal to $\mathbb{E}(Y)$. For that, we need the marginal distribution of Y. We will use the general formula:

$$\begin{aligned}\mathbb{P}(Y=y) &= \mathbb{P}(Y=y|X=0)\mathbb{P}(X=0) + \mathbb{P}(Y=y|X=1)\mathbb{P}(X=1) + \mathbb{P}(Y=y|X=2)\mathbb{P}(X=2) \\ &= 0.80 * \mathbb{P}(Y=y|X=0) + 0.15 * \mathbb{P}(Y=y|X=1) + 0.05 * \mathbb{P}(Y=y|X=2)\,.\end{aligned}$$

We have

- $\mathbb{P}(Y=0) = 0.80 * \mathbb{P}(Y=0|X=0) + 0.15 * \mathbb{P}(Y=0|X=1) + 0.05 * \mathbb{P}(Y=0|X=2) = 0.80 * 0.50 + 0.15 * 0.20 + 0.05 * 0.05 = 0.4325$
- $\mathbb{P}(Y=1) = 0.80 * \mathbb{P}(Y=1|X=0) + 0.15 * \mathbb{P}(Y=1|X=1) + 0.05 * \mathbb{P}(Y=1|X=2) = 0.80 * 0.30 + 0.15 * 0.20 + 0.05 * 0.05 = 0.2725$
- $\mathbb{P}(Y=2) = 0.80 * \mathbb{P}(Y=2|X=0) + 0.15 * \mathbb{P}(Y=2|X=1) + 0.05 * \mathbb{P}(Y=2|X=2) = 0.80 * 0.15 + 0.15 * 0.20 + 0.05 * 0.20 = 0.16$
- $\mathbb{P}(Y=3) = 0.80 * \mathbb{P}(Y=3|X=0) + 0.15 * \mathbb{P}(Y=3|X=1) + 0.05 * \mathbb{P}(Y=3|X=2) = 0.80 * 0.05 + 0.15 * 0.20 + 0.05 * 0.35 = 0.0875$
- $\mathbb{P}(Y=4) = 0.80 * \mathbb{P}(Y=4|X=0) + 0.15 * \mathbb{P}(Y=4|X=1) + 0.05 * \mathbb{P}(Y=4|X=2) = 0.80 * 0.00 + 0.15 * 0.20 + 0.05 * 0.35 = 0.0475$

Having this marginal distribution of Y it is very simple to obtain its expected value:

$$\mathbb{E}(Y) = \sum_{y=0}^{4} y\mathbb{P}(Y=y) = 0 * 0.4325 + 1 * 0.2725 + 2 * 0.16 + 3 * 0.0875 + 4 * 0.0475 = 1.045\,,$$

which is the same number obtained when we calculated the expected value of $m(X) = \mathbb{E}(Y|X)$. We verified in this example that, considering the conditional expectation $m(X) = \mathbb{E}(Y|X)$ as a random variable, we have that the expectation of the conditional expectation is equal to expectation of Y. That is, $\mathbb{E}(Y) = \mathbb{E}(m(X)) = \mathbb{E}[\mathbb{E}(Y|X)]$. ■

■ **Example 18.2 — Different Learning.** A population is divided into two groups: those who learn fast (F) and those who tend to learn more slowly (L). Of this population, 20% are of type F, and the remaining 80% are of type L. We randomly select one individual from this population, and let the random variable X represent the type (F or L) of the selected person.

Next, we conduct a series of independent trials until the first success is recorded, and let Y be the number of trials required. Individuals differ in their probability of succeeding in a single trial. Specifically, those in group F have $\theta_1 = 0.8$ while those in group L have $\theta_2 = 0.4$.

In summary:

$$\mathbb{P}(X=F) = 0.20 \quad \text{and} \quad \mathbb{P}(X=L) = 0.80.$$

Once an individual is chosen, we observe Y, the number of trials (all independent and identically repeated) until the first success. The *conditional* distribution $(Y \mid X)$ is straightforward to identify:

- $(Y \mid X=F) \sim \text{Geom}(\theta_1) = \text{Geom}(0.8)$,
- $(Y \mid X=L) \sim \text{Geom}(\theta_2) = \text{Geom}(0.4)$.

The geometric distribution has well-known properties. In particular, its expected value is the reciprocal of its success probability (see Chapter 6). Hence,

$$m(X) = \mathbb{E}(Y \mid X) = \begin{cases} \mathbb{E}(Y \mid X=F) = 1/\theta_1 = 1/0.8 = 1.25, & \text{with probability } 0.20, \\ \mathbb{E}(Y \mid X=L) = 1/\theta_2 = 1/0.4 = 2.5, & \text{with probability } 0.80. \end{cases}$$

Before selecting an individual, we do not know what to expect for the number of trials until the first success. On average, it could be only 1.25 if we pick an F-type individual but could rise to 2.5 for an L-type. As L is more common than F, a value closer to 2.5 is more typical. Indeed, taking the average over the possibilities for X, we get

$$\mathbb{E}[m(X)] = \mathbb{E}\big[\mathbb{E}(Y \mid X)\big] = \mathbb{E}(Y \mid X = F)\cdot 0.20 + \mathbb{E}(Y \mid X = L)\cdot 0.80 = 1.25\cdot 0.20 + 2.5\cdot 0.80 = 2.25.$$

We will verify that $\mathbb{E}(Y)$ equals the previously obtained value $\mathbb{E}(Y) = \mathbb{E}[\mathbb{E}(Y \mid X)] = 2.25$. For this, we need the marginal distribution of Y. The support of Y is the set of positive integers: $\mathscr{S} = \{1,2,3,\ldots\}$. For $y \in \mathscr{S}$, we have

$$\begin{aligned}
\mathbb{P}(Y = y) &= \mathbb{P}(Y = y \mid X = F)\mathbb{P}(X = F) + \mathbb{P}(Y = y \mid X = L)\mathbb{P}(X = L) \\
&= \mathbb{P}(Y = y \mid X = F) * (0.20) + \mathbb{P}(Y = y \mid X = L) * (0.80) \\
&= \big((0.2)^{y-1}0.8\big) * 0.20 + \big((0.6)^{y-1}0.4\big) * 0.80 \quad \text{(using geometric probabilities)} \\
&= (0.2)^{y-1}(0.16) + (0.6)^{y-1}(0.32)
\end{aligned}$$

To finish the calculation, we recall a standard result from infinite series: for any $\theta \in (0,1)$,

$$\begin{aligned}
\sum_{y=1}^{\infty} y\theta^{y-1} &= 1 * \theta^0 + 2 * \theta^1 + 3 * \theta^2 + 4 * \theta^3 + \ldots \\
&= \frac{d}{d\theta}\left(\theta + \theta^2 + \theta^3 + \ldots\right) \\
&= \frac{d}{d\theta}\left(\frac{1}{1-\theta} - 1\right) \\
&= \frac{1}{(1-\theta)^2}
\end{aligned}$$

To finish the calculation, we recall a standard result from infinite series: for any $\theta \in (0,1)$,

$$\sum_{y=1}^{\infty} y\,\theta^{y-1} = 1\cdot\theta^0 + 2\cdot\theta^1 + 3\cdot\theta^2 + \cdots = \frac{d}{d\theta}\Big(\theta + \theta^2 + \theta^3 + \ldots\Big) = \frac{d}{d\theta}\Big(\tfrac{1}{1-\theta} - 1\Big) = \frac{1}{(1-\theta)^2}.$$

We apply this to compute $\mathbb{E}(Y)$:

$$\begin{aligned}
\mathbb{E}(Y) &= \sum_{y=1}^{\infty} y\mathbb{P}(Y = y) \\
&= \sum_{y=1}^{\infty} y\left((0.2)^{y-1}(0.16) + (0.6)^{y-1}(0.32)\right) \\
&= 0.16\sum y(0.2)^{y-1} + 0.32\sum y(0.6)^{y-1} \\
&= 0.16\frac{1}{(1-0.2)^2} + 0.32\frac{1}{(1-0.6)^2} \\
&= 2.25\,,
\end{aligned}$$

which matches our simpler calculation $\mathbb{E}(m(X))$. ∎

The key takeaway is that calculating $\mathbb{E}(Y)$ can be simplified by using $\mathbb{E}[m(X)] = \mathbb{E}[\mathbb{E}(Y \mid X)]$. The idea is to find another random variable X such that, when conditioning on X, the distribution of Y is straightforward to compute.

This approach is often used in sequential experiments, where the next action depends on the previous step's outcome. We may want to determine the average value of the final outcome Y by considering all possible outcomes of the previous steps (represented by X). This is precisely what $\mathbb{E}[\mathbb{E}(Y \mid X)]$ captures.

■ **Example 18.3 — Randomly stopped sum: collective risk.** This is an important example, known as the *randomly stopped sum model*, which appears in many applications. First, consider the simpler case of a fixed number n of i.i.d. random variables:

$$Y = X_1 + \cdots + X_n.$$

If $\mathbb{E}(X_i) = \mu$ and $\mathbb{V}(X_i) = \sigma^2$, finding $\mathbb{E}(Y)$ and $\mathbb{V}(Y)$ is straightforward:

$$\mathbb{E}(Y) \;=\; \mathbb{E}(X_1 + \cdots + X_n) \;=\; \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n) \;=\; \mu + \cdots + \mu \;=\; n\mu,$$

and, using the independence of the X_is,

$$\mathbb{V}(Y) \;=\; \mathbb{V}(X_1 + \cdots + X_n) \;=\; \mathbb{V}(X_1) + \cdots + \mathbb{V}(X_n) \;=\; n\sigma^2.$$

However, in many applications, the number n is not fixed but *random*. Here is a simple model from actuarial science, arising in *collective risk theory*: the total amount of an insurance company's in a given time period is modeled by

$$Y = \sum_{i=1}^{N} X_i = X_1 + \ldots + X_N,$$

where N is a discrete *random variable* representing the total number of claims in that period. The variables $X_1, X_2, \ldots$ are nonnegative, independent, identically distributed random variables representing the individual claim amounts, and they are independent of N. For example, a random number of N vehicle accidents could result in the claims $X_1, X_2, \ldots, X_N$. In such scenarios, Y is called a *randomly stopped sum* because the summation of the X_i's stops at the random index N.

The property $\mathbb{E}(Y) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n)$ holds when n is a fixed, non-random number. However, when N is random, the sum $Y = X_1 + \ldots + X_N$ has two sources of variability: (i) the randomness of the individual X_i values and (ii) the variability in the number of terms being summed. In this case, it is incorrect to assume that we can simply exchange the expectation $\mathbb{E}$ with the sum as before and write something such as $\mathbb{E}(Y) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_N)$. This is not valid.

Fortunately, this issue is easily resolved by first conditioning on N and then integrating over its possible random values. If $N = n$, then the randomness in N is eliminated, and we obtain:

$$\mathbb{E}(Y|N = n) = \mathbb{E}(X_1 + \ldots X_N|N = n) = \mathbb{E}(X_1 + \ldots X_n) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n) = n\mu\,.$$

Thus, the conditional expectation of Y given N is $m(N) = \mathbb{E}(Y|N) = N\mu$. Applying the theorem of iterated expectation in section 18.5, we obtain:

$$\mathbb{E}(Y) = \mathbb{E}[\mathbb{E}(Y \mid N)] = \mathbb{E}[N\mu] = \mu\mathbb{E}[N]$$

This result may seem somewhat disappointing, as it confirms what intuition already suggests: the expected value of a sum of a random number of terms is simply the expected number of terms times the expected value of each term. However, do not be too quick to dismiss the importance of formal rigor. In Section 18.6, we will see that the variance of such random sums does not behave as intuitively as their expectation. We will revisit this example there. ■

■ **Example 18.4 — Random stick breaking.** Consider a rod of length L. First, break the rod at a random point X, where $X \sim U(0, L)$. Retain the left piece and break it again at Y, where $(Y \mid X) \sim U(0, X)$. We want to calculate $\mathbb{E}(Y)$. Since Y depends on two successive steps, we must account for all possible outcomes of both steps.

To simplify the calculation, we break it into two stages. First, we condition on the outcome of the initial break and compute $m(x) = \mathbb{E}(Y \mid X = x)$. This is straightforward: as $(Y \mid X = x) \sim U(0, x)$, its expectation is the midpoint of the interval $[0, x]$. Thus, $m(X) = \mathbb{E}(Y \mid X) = \frac{X}{2}$.

Next, we take the expected value of $m(X)$. Since $X \sim U(0,L)$, we compute:

$$\mathbb{E}(Y) = \mathbb{E}[\mathbb{E}(Y \mid X)] = \mathbb{E}\left[\frac{X}{2}\right] = \frac{\mathbb{E}(X)}{2} = \frac{L}{4}.$$

■

18.5 Law of iterated expectation

In this section we will state the theorem we have been using and provide a formal proof.

Theorem 18.5.1 — Law of iterated expectation. Suppose X and Y are random variables and Y has finite expectation. Then we have

$$\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(Y).$$

Proof. Consider initially the continuous cases. Let

$$m(x) = \mathbb{E}(Y|X = x) = \int y f_{Y|X}(y|x)dy\,.$$

We transform this function into the random variable $m(X)$ and obtain its expected value $\mathbb{E}(m(X))$:

$$\begin{aligned}
\mathbb{E}[\mathbb{E}(Y|X)] &= \mathbb{E}(m(X)) \\
&= \int m(x) f_X(x)dx \\
&= \int \left(\int y f_{Y|X}(y|x)dy \right) f_X(x)dx \\
&= \int \int y f_{Y|X}(y|x) f_X(x)dxdy \\
&= \int \int y \frac{f_{(X,Y)}(x,y)}{f_X(x)} f_X(x)dxdy \\
&= \int \int y f_{(X,Y)}(x,y)dxdy \\
&= \int y \left(\int f_{(X,Y)}(x,y)dx \right) dy \\
&= \int y f_Y(y)dy \\
&= \mathbb{E}(Y).\square
\end{aligned}$$

The discrete case has a completely analogous proof and it is left as an exercise.

■

Does this result make intuitive sense? Yes, it does. First, we compute the conditional expectation $m(x) = \mathbb{E}(Y \mid X = x)$, which represents the expected value of Y for each possible realization of $X = x$. Then, we take the expectation of this conditional value, averaging over all possible values of X, weighted by their associated probabilities $\mathbb{P}(X = x)$. This process brings us back to the original expectation of Y, confirming that $\mathbb{E}(Y)$ remains unchanged.

■ **Example 18.5 — Mouse in a maze.** This example is adapted from the lecture notes of Rachel Fewster in `https://www.stat.auckland.ac.nz/~fewster/325/notes.php`. A mouse is trapped in a central room with three exits inside a maze:

- Exit 1 leads directly outside the maze after 3 minutes.

- Exit 2 leads back to the central room after 5 minutes.
- Exit 3 leads back to the central room after 7 minutes.

The mouse keeps trying until it eventually leaves the maze. Each time the mouse chooses an exit, it selects one at random with equal probability. If the mouse returns to the central room, the process resets, repeating the same decision mechanism. Given the probabilities shown in Figure 18.3, what is the expected time for the mouse to exit the maze?

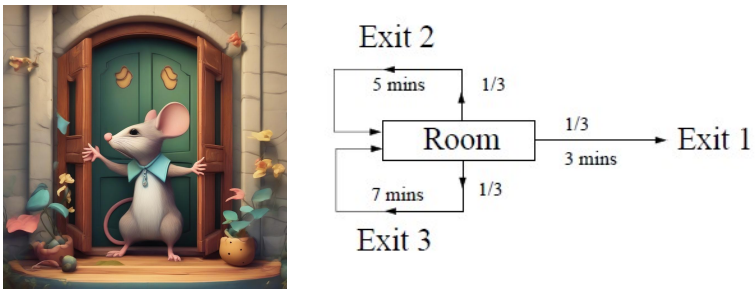


Figure 18.3: Mouse in a maze: diagram with probabilities.

Solution: Let T be the total time required for the mouse to escape the maze, starting from the central room. Define D as the first exit chosen by the mouse. The key observation is that if the mouse selects exit 2, it will take 5 minutes to return to the central room, at which point the process restarts from the beginning. Thus, given that the first choice is $D = 2$, the expected time to exit is the 5 minutes spent wandering plus the original expected time $\mathbb{E}(T)$.

We compute the conditional expectations:

$$\begin{aligned}
&\mathbb{E}(T \mid D = 1) = 3 \text{ minutes},\\
&\mathbb{E}(T \mid D = 2) = 5 + \mathbb{E}(T),\\
&\mathbb{E}(T \mid D = 3) = 7 + \mathbb{E}(T).
\end{aligned}$$

Applying the law of iterated expectation:

$$\begin{aligned}
\mathbb{E}(T) &= \mathbb{E}[\mathbb{E}(T \mid D)]\\
&= \mathbb{E}(T \mid D = 1)\mathbb{P}(D = 1) + \mathbb{E}(T \mid D = 2)\mathbb{P}(D = 2) + \mathbb{E}(T \mid D = 3)\mathbb{P}(D = 3)\\
&= 3 \cdot \frac{1}{3} + (5 + \mathbb{E}(T)) \cdot \frac{1}{3} + (7 + \mathbb{E}(T)) \cdot \frac{1}{3}.
\end{aligned}$$

Rearranging the equation:

$$\begin{aligned}
&\mathbb{E}(T) - \frac{1}{3}\mathbb{E}(T) - \frac{1}{3}\mathbb{E}(T) = 3 \cdot \frac{1}{3} + 5 \cdot \frac{1}{3} + 7 \cdot \frac{1}{3},\\
&\frac{1}{3}\mathbb{E}(T) = 15 \cdot \frac{1}{3},\\
&\Rightarrow \mathbb{E}(T) = 15 \text{ minutes}.
\end{aligned}$$

Additionally, we can determine the possible values of the conditional expectation $\mathbb{E}(T \mid D)$:

$$\begin{aligned}
&\mathbb{E}(T \mid D = 1) = 3 \text{ minutes},\\
&\mathbb{E}(T \mid D = 2) = 5 + \mathbb{E}(T) = 20 \text{ minutes},\\
&\mathbb{E}(T \mid D = 3) = 7 + \mathbb{E}(T) = 22 \text{ minutes}.
\end{aligned}$$

■

18.6 Law of iterated variance

We have defined the *conditional variance* $v(x) = \mathbb{V}(Y \mid X = x)$ in the discrete case (section 12.14) and continuous case (section 12.20). The function $v(x)$ quantifies the dispersion of $(Y|X = x)$ around its conditional mean $m(x) = \mathbb{E}(Y \mid X = x)$. Formally,

$$\begin{aligned} v(x) &= \mathbb{V}(Y \mid X = x) \\ &= \mathbb{E}\left[(Y - \mathbb{E}(Y \mid X = x))^2 \mid X = x\right] \\ &= \mathbb{E}\left[(Y - m(x))^2 \mid X = x\right] \\ &= \int (y - m(x))^2 f_{Y|X}(y \mid x)\, dy. \end{aligned}$$

As before, we can convert the function $v(x)$ into a random variable $v(X)$ by replacing the fixed value x with the random variable X.

Just as the law of iterated expectation relates $\mathbb{E}(Y)$ to the conditional expectation $\mathbb{E}(Y \mid X)$, there is also a *law of iterated variance*, which connects $\mathbb{V}(Y)$ to the expected value of the conditional variance $\mathbb{V}(Y \mid X)$. However, unlike the law of iterated expectation, this result includes an additional term that accounts for the variance of $\mathbb{E}(Y \mid X)$. Consequently, the law of iterated variance is more intricate and, in many ways, more insightful than its expectation counterpart.

Theorem 18.6.1 — Law of iterated variance. Suppose X and Y are random variables and Y has finite variance. Then we have

$$\mathbb{V}(Y) = \mathbb{V}(\mathbb{E}(Y|X)) + \mathbb{E}(\mathbb{V}(Y|X)).$$

We will postpone the proof of this theorem until after we have examined its meaning and explored a few examples of its application.

This result decomposes the total variability of Y into two components. Return to the left-hand plot of Figure 18.1. The marginal variance $\mathbb{V}(Y)$ can be visualized by projecting the data points onto the vertical axis. The spread of these projected y points along the vertical axis represents the total variance of Y. Theorem 18.6.1 says that the marginal variance $\mathbb{V}(Y)$ is the sum of two components. The first one, $\mathbb{V}(\mathbb{E}(Y|X))$, captures how much the blue curve $m(x) = \mathbb{E}(Y \mid X = x)$ varies as x changes. Ignore the spread of the points in the scatterplot and focus only on the blue line $m(x)$. The first component measures how much this $m(x)$ line varies as a function of x. More formally, this first term captures the variation (or variance) of the random variable $\mathbb{E}(Y \mid X)$. If $m(x)$ is a horizontal line, meaning that $m(x)$ does not vary with x, we have this first term equal zero. Intuitively, this measures the variability of the conditional mean across different values of X. However, not all values of x contribute equally. More likely values of X have greater influence when computing this variability.

The second component, $\mathbb{E}(\mathbb{V}(Y \mid X))$, measures the average value of the random conditional variance $v(X) = \mathbb{V}(Y \mid X)$. The possible values $v(x)$ of the random variable $v(X)$ represent the spread of the random variable $(Y|X = x)$ around $m(x)$, the blue line at x. In the first plot on the left-hand side, we observe that $v(x)$ is large when $x \approx 0$, and it decreases as x moves toward the extremes. The second component $\mathbb{E}(\mathbb{V}(Y \mid X))$ computes the average value of these different $v(x)$ values, again weighted by how frequently different values of X occur.

Thus, the marginal variance $\mathbb{V}(Y)$ is composed of the sum of two contributions:

- $\mathbb{V}(\mathbb{E}(Y \mid X))$, which quantifies how much the random variable $m(X) = \mathbb{E}(Y \mid X)$ varies.
- $\mathbb{E}(\mathbb{V}(Y \mid X))$, which represents the expected value of the random variable $v(X) = \mathbb{V}(Y \mid X)$, capturing the average spread of Y around its conditional mean.

Together, these two components explain how much Y varies both due to changes in its conditional mean and due to its dispersion around that mean.

■ **Example 18.6 — Cont: Random stick breaking.** Continuing the example 18.4, we will calculate $\mathbb{V}(Y)$ using the law of iterated variance. Since $(Y|X=x) \sim U(0,x)$, we can use the formula for the variance of an uniform distribution to find $\mathbb{V}(Y|X) = X^2/12$. We then directly apply the law of iterated variance:

$$\begin{aligned}
\mathbb{V}(Y) &= \mathbb{V}[\mathbb{E}(Y|X)] + \mathbb{E}[\mathbb{V}(Y|X)] = \mathbb{V}\left[\frac{X}{2}\right] + \mathbb{E}\left[\frac{X^2}{12}\right] \\
&= \frac{1}{4}\mathbb{V}[X] + \frac{1}{12}\mathbb{E}[X^2] = \frac{1}{4}\frac{L^2}{12} + \frac{1}{12}[\mathbb{V}(X) + (\mathbb{E}(X))^2] \\
&= \frac{L^2}{48} + \frac{1}{12}\left[\frac{L^2}{12} + (L/2)^2\right] = \frac{L^2}{48} + \left(\frac{L^2}{48*3} + \frac{L^2}{48}\right) \\
&= \frac{L^2}{48}\left(1 + \left(\frac{1}{3} + 1\right)\right)
\end{aligned}$$

It is useful to assess the contribution of each component to the total variance $\mathbb{V}(Y)$. The last line above shows that the two components contribute approximately equally but the second component is slightly larger than the first and accounts for $(1+1/3)/(2+1/3) \approx 0.57$, or 57% of the total variance. ■

■ **Example 18.7 — Cont: Different learning.** Continuing the example 18.2, we will calculate $\mathbb{V}(Y)$. Consider first the random variable $\mathbb{E}(Y|X)$. As $(Y \mid X = x)$ is an exponential random variable, the random variable $\mathbb{E}(Y|X)$ is a discrete binary random variable with the possible values $1/\theta_1 = 1/0.8$ and $1/\theta_2 = 1/0.4$ with respective probabilities $\mathbb{P}(X=F) = 0.20$ and $\mathbb{P}(X=L) = 0.80$. We have already calculated its expected value: $\mathbb{E}[\mathbb{E}(Y|X)] = 2.25$. Therefore, its variance is given by

$$\mathbb{V}[\mathbb{E}(Y|X)] = (1/0.8 - 2.25)^2 * 0.20 + (1/0.4 - 2.25)^2 * 0.80 = 0.25\,.$$

For the other component, we use the variance formula for a geometric distribution. We have $\mathbb{V}(Y|X=F) = (1-\theta_1)/\theta_1^2 = 0.3125$ and $\mathbb{V}(Y|X=L) = (1-\theta_2)/\theta_2^2 = 3.75$. These possible values occur with probabilities $\mathbb{P}(X=F) = 0.20$ and $\mathbb{P}(X=L) = 0.80$, respectively. Therefore, its expected value is

$$\mathbb{E}[\mathbb{V}(Y|X)] = 0.3125 * 0.20 + 3.75 * 0.80 = 3.0625\,.$$

Putting these two pieces together, we have

$$\mathbb{V}(Y) = \mathbb{V}[\mathbb{E}(Y|X)] + \mathbb{E}[\mathbb{V}(Y|X)] = 0.25 + 3.0625 \approx 3.31\,.$$

The second component is the most relevant because $\mathbb{E}(Y|X)$ does not vary much as X changes its values. In consequence, the first component $\mathbb{V}[\mathbb{E}(Y|X)]$ has a small value. ■

■ **Example 18.8 — cont: Mouse in a maze.** Using the same description given in the Example 18.5, find the variance of the random time for the mouse to leave the maze.

We have

$$\mathbb{V}(T) = \mathbb{V}[\mathbb{E}(T|D)] + \mathbb{E}[\mathbb{V}(T|D)]$$

We start by obtaining the first term. As seen in the Example 18.5, the random variable $\mathbb{E}(T|D)$ has the possible values 3, 20, and 22. We already know that the expected value of this random variable is $\mathbb{E}(T) = 15$ because $\mathbb{E}(T) = \mathbb{E}[\mathbb{E}(T|D)]$. However, we can also verify this directly:

$$\mathbb{E}[\mathbb{E}(T|D)] = 3 \times \frac{1}{3} + 20 \times \frac{1}{3} + 22 \times \frac{1}{3} = 15$$

The variance of this random variable can be easily obtained:

$$\begin{aligned}\mathbb{V}[\mathbb{E}(T|D)] &= \mathbb{E}(\mathbb{E}(T|D)-\mathbb{E}(T))^2 \\ &= (3-15)^2\frac{1}{3}+(20-15)^2\frac{1}{3}+(22-15)^2\frac{1}{3} \\ &= 72.7\end{aligned}$$

To obtain the second term, $\mathbb{E}[\mathbb{V}(T|D)]$, we start by obtaining the possible values of the random variable $\mathbb{V}(T|D)$. The first one is $\mathbb{V}(T|D=1)=0$. This is so because T does not vary in this case: we always have $T=3$ When $D=1$, with no variation. The second one is $\mathbb{V}(T|D=2)$. As the system is restored when the mouse returns after exiting door 2, we have $(T|D=2)$ with the same distribution as $5+T$ and therefore $\mathbb{V}(T|D=2)=\mathbb{V}(5+T)=\mathbb{V}(T)$. Likewise, we obtain $\mathbb{V}(T|D=2)=\mathbb{V}(T)$. We conclude that

$$\begin{aligned}\mathbb{E}[\mathbb{V}(T|D)] &= \mathbb{V}(T|D=1)\times\frac{1}{3}+\mathbb{V}(T|D=2)\times\frac{1}{3}+\mathbb{V}(T|D=3)\times\frac{1}{3} \\ &= 0\times\frac{1}{3}+\mathbb{V}(T)\times\frac{1}{3}+\mathbb{V}(T)\times\frac{1}{3} \\ &= \frac{2\mathbb{V}(T)}{3}\end{aligned}$$

Now we can conclude the calculation

$$\begin{aligned}\mathbb{V}(T) &= \mathbb{V}[\mathbb{E}(T|D)]+\mathbb{E}[\mathbb{V}(T|D)] \\ &= 72.7+\frac{2\mathbb{V}(T)}{3} \\ \frac{\mathbb{V}(T)}{3} &= 72.7 \\ \mathbb{V}(T) &= 3\times 72.7 = 218\end{aligned}$$

Note that, in the variance decomposition, the second term is twice the value of the first term:

$$\begin{aligned}\mathbb{V}(T) &= \mathbb{V}[\mathbb{E}(T|D)]+\mathbb{E}[\mathbb{V}(T|D)] \\ &= 72.7+(2\times 72.7)\end{aligned}$$

■

18.7 Proof of the law of iterated variance

Proof. To prove Theorem 18.6.1, start by noticing that, for all fixed x, the usual definition of variance of a random variable allows us to write

$$v(x)=\mathbb{V}(Y\mid X=x)=\mathbb{E}(Y^2\mid X=x)-(\mathbb{E}(Y|X=x))^2\,.$$

Since this is valid for all fixed x, we have the same formula for the random variables:

$$\mathbb{V}(Y\mid X)=\mathbb{E}(Y^2\mid X)-(\mathbb{E}(Y\mid X))^2\,.$$

These are functions of X. If we take the expectation with respect to X in both sides of the last equation and use the law of integrated expectation (Theorem 18.5.1), we have:

$$\begin{aligned}\mathbb{E}[\mathbb{V}(Y\mid X)] &= \mathbb{E}\left[\mathbb{E}(Y^2\mid X)\right]-\mathbb{E}\left[(\mathbb{E}(Y\mid X))^2\right] \\ &= \mathbb{E}\left[Y^2\right]-\mathbb{E}\left[(\mathbb{E}(Y\mid X))^2\right]\end{aligned}$$

Now, consider the random variable $\mathbb{E}(Y \mid X)$ which has expected value $\mathbb{E}(Y)$ by Theorem 18.5.1. Then, by definition, its variance is given by

$$\begin{aligned}\mathbb{V}[\mathbb{E}(Y \mid X)] &= \mathbb{E}\left[(\mathbb{E}(Y \mid X))^2\right] - (\mathbb{E}[\mathbb{E}(Y \mid X)])^2 \\ &= \mathbb{E}\left[(\mathbb{E}(Y \mid X))^2\right] - (\mathbb{E}(Y))^2\end{aligned}$$

Summing up the two expressions found, we have:

$$\begin{aligned}\mathbb{E}[\mathbb{V}(Y \mid X)] + \mathbb{V}[\mathbb{E}(Y \mid X)] &= \mathbb{E}\left[Y^2\right] - \mathbb{E}\left[(\mathbb{E}(Y \mid X))^2\right] + \mathbb{E}\left[(\mathbb{E}(Y \mid X))^2\right] - (\mathbb{E}(Y))^2 \\ &= \mathbb{E}\left[Y^2\right] - (\mathbb{E}(Y))^2 \\ &= \mathbb{V}(Y)\end{aligned}$$

This completes the proof. ■

19. Limit Theorems

19.1 Introduction

So far we have been informally using the idea that the probability $\mathbb{P}(B)$ of a certain event B can be approximated empirically. Simply repeat the random experiment independently a large number n of times and obtain the k/n proportion of times that event B occurs. Then $\mathbb{P}(B) \approx k/n$. In this chapter, we will obtain the formal, mathematical, justification for this result. It can be deduced mathematically as a consequence of the definitions and preliminary results that we have obtained so far.

To obtain this mathematical justification for the $\mathbb{P}(B) \approx k/n$ approximation, we need to be more specific and more precise about limiting operations involving random variables. What we will learn in this chapter is that the concept of an asymptotic limit of random variables is much more subtle and more elaborate than the usual concept of a limit of real numbers. We will see that there are three very different definitions (or ways of thinking) about limits in probability. We will also appreciate the need to make use of the definitions and results involving σ-algebras and set measures that we briefly cited in chapter 3. There's no escape: every time uncountable infinite sets come into play in probability, we have to use those σ-algebras and set measures ideas and results.

We will also learn the most important limit result in probability and statistics, the Central Limit Theorem (CLT). Informally, it says that a sum of a large number of random variables is approximately normally distributed. There are qualifications for this result to be valid but it is general enough to cover a large number of applications. Let's start by recalling the concept of a limit of sequences of real numbers.

19.2 The limit of a sequence of real numbers

Consider a sequence of real numbers xn. When do we say that x_n converges to a finite limit L? It seems easy: look at the sequence and see if it approaches a given finite value L more and more as we progress in the sequence. For example, consider the following sequences of real numbers:

- $x_1 = 1, x_2 = 1/2, x_3 = 1/3, x_4 = 1/4, \ldots, x_n = 1/n, \ldots$: intuitively (and correctly), this sequence converges to 0.

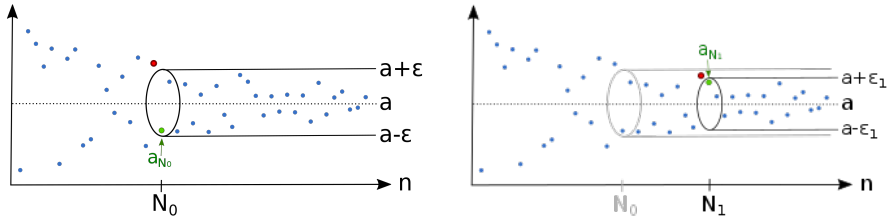
- $x_1 = 1, x_2 = 1, x_3 = 1, x_4 = 1, \ldots, x_n = 1, \ldots$: intuitively (and correctly), it converges to 1 (it never departs from 1).
- $x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4, \ldots, x_n = n, \ldots$: it does not converge to a finite real number L. The sequence increases systematically with n, with no finite limiting value.
- $x_1 = 1, x_2 = 1, x_3 = 1, x_4 = 2, x_5 = 1, x_6 = 3, x_7 = 1, x_8 = 4 \ldots, x_{2n} = n, x_{2n+1} = 1, \ldots$: it does not converge to any value L. The odd elements are constant and equal to 1 but the even elements go away to infinite.
- $x_1 = 1+1, x_2 = 1+1/2, x_3 = 1+1/3, x_4 = 1+1/4, \ldots, x_n = 1+1/n, \ldots$: converges to 1 as we move along the sequence (or, in equivalent way, as n increases).
- $x_1 = 1, x_2 = -1+1/2, x_3 = 1-1/3, x_4 = -1+1/4, x_5 = 1-1/5, x_6 = -1+1/6, \ldots, x_{2n} = -1+1/(2n), x_{2n+1} = 1-1/(2n+1), \ldots$: oscilates between positive and negative but intuitively it converges to 0.

This intuitive notion solves several problems but is far from being able to cover all cases that can occur. Quickly, we start thinking about sequences that challenge our intuitive idea. Consider the following example: $x_n = 2$ if $n = 10^k$ (that is, $x_n = 2$ if n is a power of 10 such as 10, 10^2, 10^3, 10^4, etc.), and $x_n = 1 + 1/n$, otherwise. Thus, $|x_n - 1| \to 0$ as $n \to \infty$, but for $n \neq 10, 100, 1000, 10000, 100000$, etc. However, when $n = 10, 100, 1000, 10000, 100000, \ldots$, etc., we will have $x_n = 2$. So, less frequently, x_n deviates significantly from 1 and equals 2. In this way, $|x_n - 1| = |2 - 1| = 1$, but it takes more and more time to occur: it takes 10 steps; then, it takes 100 steps to occur again; then, 1000 steps, and so on. Although the event of $|x_n - 1| = 1$ becomes rarer, we will *never* have $|x_n - 1|$ small (less than 1) for every n from a certain value onwards. There will always be some future n in the form of 10^k, even if it is far in the sequence, to make the difference $|x_n - 1|$ not arbitrarily small.

To move away from intuition and make the definition of a limit rigorous, mathematicians spent a substantial amount of time. It was only in 1870 that the definition we use today was popularized by Karl Weierstrass.

Definition 19.2.1 — Limit of sequences of real numbers. Let $x_1, x_2, x_3, \ldots$ be a sequence of real numbers. We say that x_n converges to the limit L as n approaches infinity if the following property holds: for every $\varepsilon > 0$, there exists an integer n_0 such that, for every $n > n_0$, we have $|x_n - L| < \varepsilon$.

The definition can be seen as a kind of game. One player proposes a maximum distance $\varepsilon > 0$ between the limit L and the elements of the sequence. That is, we want $|x_n - L| < \varepsilon$. Initially, for some small values of n, this may not be true and occasionally we may see $|x_n - L| \geq \varepsilon$. However, if the sequence converges to L, there exists a certain point n_0 from which every subsequent element of the sequence must be within the interval $(L - \varepsilon, L + \varepsilon)$. In other words, there must exist a specific n_0 such that $x_{n_0+k} \in (L - \varepsilon, L + \varepsilon)$ or, equivalently, $|x_{n_0+k} - L| < \varepsilon$. The graph on the left in Figure 19.2, extracted from Wikipedia, illustrates this definition. The horizontal axis represents the index n, and the vertical axis represents the values x_n of the sequence of real numbers. With a fixed ε, a horizontal band $(a - \varepsilon, a + \varepsilon)$ is created around the limit value a. The goal is to find a value n_0 such that, from that n_0 onwards, all the elements x_n remain within the band. This value n_0 depends on the chosen ε. The graph on the right in Figure 19.2 shows the situation with a smaller value ε_1 compared to the ε value in the graph on the left. The band $(a - \varepsilon_1, a + \varepsilon_1)$ around a is narrower than the one on the left. In this case, several elements x_n of the sequence no longer fall within the new band. However, if the sequence x_n truly converges to a, it will be possible to find another integer n_1 such that, from that point onward, *all* elements of the sequence will be within the narrower band. We emphasize the word "all" because it is crucial. If the sequence converges, we do not want many or most of the subsequent values to be close. The definition is strict. It requires that all the values x_{n_0+k} of the sequence from the cutoff point n_0 onwards be close to the limit L (i.e., within a distance



smaller than ε from L). The important thing is that this cutoff point n_0, from which $|x_{n_0+k} - L| < \varepsilon$, must be found for any ε, no matter how small. Thus, the cutoff point n_0 depends on the chosen ε, and strictly speaking, it should be written as $n_0(\varepsilon)$. For a very small ε, we may need to find a very large n_0, while for a large epsilon, a not-so-large n_0 may suffice. This definition of a limit is complicated but necessary to encompass all possibilities in the creation of sequences.

19.3 Convergence of sequences of random variables

If the definition of limit and convergence is complicated in the case of real numbers, it becomes even more complicated and subtle in the case of a sequence $S_1, S_2, \ldots$ of random variables. When can we say that S_n converges? What sense or meaning do we want or can we give to this convergence? We will talk extensively about coins in this chapter. It is a very simple situation, without many complications or mitigating considerations, and one that everyone understands. Following the example from Patrick Breheny's lecture notes at `https://myweb.uiowa.edu/pbreheny/4120/s20/notes/3-05.pdf`, let's describe a coin experiment conducted in 1940. John Kerrich was a mathematics professor from South Africa who was trapped in Nazi-occupied Denmark. He flipped a regular coin 10,000 times under approximately equal conditions (launch angle, torque, initial velocity, etc.). John kept an accurate record of the successive outcomes he obtained. Until now, we have informally used the idea that the number of heads should balance out with the number of tails if we flip the coin a very large number of times. The table 19.3 shows John Kerrich's record of coin flips. The second column shows the number of heads, and the third column shows the difference between the number of heads and its expected value, which is half of the total number of flips. Given the expectation of a balance between heads and tails as n increases, we might expect the numbers in the third column to decrease toward zero. However, what we see is that instead of decreasing, the number of heads after n flips does not get increasingly closer to its expected value of $n/2$.

The red line on the left side of Figure 19.3 shows the sequence of values from this third column (horizontal axis on a logarithmic scale). The lack of convergence in this experiment is not a coincidence. By independently repeating the same experiment 50 times on the computer, we generated an additional 50 curves in this graph. We can see that the lines spread around zero without a clear tendency to converge to zero.

Actually, what we can expect is that the proportion of heads converges to $1/2$ as the number of flips n increases. The graph on the right side of Figure 19.3 shows the same 51 sequences of flips from the left side, considering the proportions of heads in each sequence up to n. In this case, it seems that there is convergence to the value $1/2$. The difficulty and subtlety in the probabilistic case is that the sequence is not deterministic. What guarantees that each of these lines will get closer and closer to $1/2$? To be more formal, let's define $X_n = 1$ if the n-th flip is heads and $X_n = 0$ if it's tails. Let $Y_n = (X_1 + \ldots + X_n)/n$ be the proportion of heads in the first n flips. What do we expect to see as n increases? Naturally, we expect to see Y_n approach $1/2$, right? But the sequence $1,1,1,1,1,1,1,1$ is just as likely to occur as any other sequence of length 8. Each of them has the same probability of $1/2^8$ of occurring. And in constant sequences, such as $1,1,1,1,1,1,1,1$, the random variable Y_n does not seem to converge to $1/2$. Let's simulate three independent repetitions

n # Trials	# Heads	# Heads - $n/2$
10	4	-1
100	44	-6
500	255	5
1000	502	2
2000	1013	13
3000	1510	10
4000	2029	29
5000	2533	33
6000	3009	9
7000	3516	16
8000	4034	34
9000	4538	38
10000	5067	67

Table 19.1: Sequence of coin flips made by John Kerrich.

of this experiment on the computer, looking at Y_n for the first 200 flips (see graph on the left in Figure 19.3). It is not obvious that the red line is converging to $1/2$ after the first 200 flips. Looking a bit further, after 2000 flips, the red line in the graph on the right in Figure 19.3 appears to be converging to $1/2$. However, the blue line, which seemed very close to $1/2$ on the left with $n = 200$, deviates from this potential limit in the graph on the right. What guarantees that the blue line will approach it again and get closer and closer to $1/2$?

One of the difficulties in answering these questions is the fact that each Y_n is a random variable with a different distribution for each n. If Y_n is a random variable, it has a list of possible values and a list of associated probabilities. The possible values of Y_n are: $\{0/n, 1/n, 2/n, \ldots, (n-1)/n, n/n\}$, which depend on n. The associated probabilities also vary with n and are obtained through the binomial distribution: $\mathbb{P}(Y_n = k/n) = \mathbb{P}(X_1 + \ldots + X_n = k) = n!/(k!(n-k)!)1/2^n$. The sequence Y_n represents a sequence of random variables with their two lists (values and probabilities) changing with n. When we want to emphasize that we are dealing with a sequence of random variables, we usually write $\{Y_n : n = 1, 2, \ldots\}$ instead of just Y_n.

One way to think about limits associated with the sequence $\{Y_n : n = 1, 2, \ldots\}$ is to consider the realization of the variables $X_1, X_2, \ldots$ and the specific numerical sequence that forms as the X_i's are instantiated. For example, the red line in Figure 19.3 had $Y_1 = 0$, $Y_2 = 1/2$, $Y_3 = 1/3$, $Y_4 = 1/4$, $Y_5 = 1/5$, $Y_{10} = 3/10$, $Y_{20} = 7/20$, $Y_{100} = 40/100$, and $Y_{200} = 87/200$. We can ask whether this specific sequence, this instantiation of the experiment, will eventually converge to $1/2$. With what probability can this convergence occur? When can we say that Yn converges to something? Will there exist a random limit $\lim_{n\to\infty} Y_n$? Will this limit be a fixed, constant value? Random? Does it always exist?

These probability questions are delicate. Notice, for example, that the sequence Y_n is not formed by independent random variables. The possible values of the random variable Y_{51} are the list $\{0, 1/51, 2/51, \ldots, 50/51, 51/51\}$, and they have associated probabilities, all of which are non-zero. However, if you tell me that $Y_{50} = 23/50$, then I know that the possible values of Y_{51} are reduced to only two possible values: $23/51$ or $24/51$, each with a probability of $1/2$. The other originally possible values, such as $0/51, 1/51, 2/51$, now have a probability of zero occurrence (conditionally on the event $Y_{50} = 23/50$).

One way to approach the problem would be to ask whether the probability of generating a convergent sequence to $1/2$ is high. Is the probability exactly 1, or is it possible for the probability

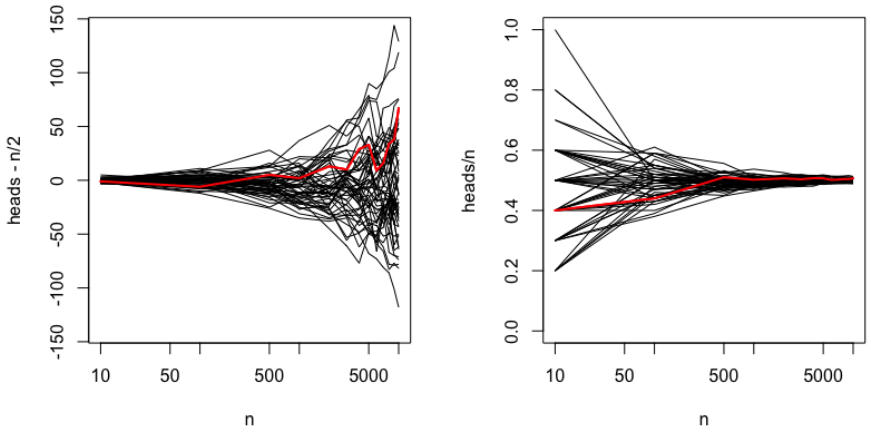


Figure 19.1: 50 other computer simulated sequences of coin flips plus the John Kerrich's sequence (red line).

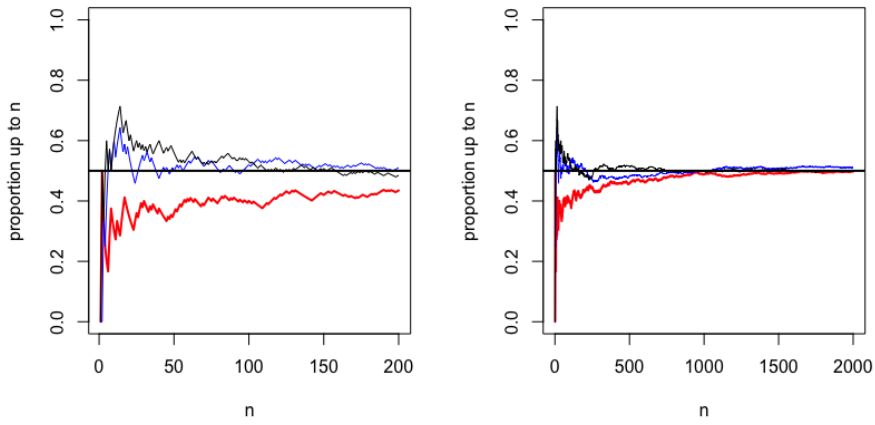


Figure 19.2: Three simulated sequences of tosses of a fair coin.

to be less than 1? How much less than 1? After all, we have several sequences where the proportion of heads does not converge to $1/2$ as n increases. Consider, for example, the following sequences that exhibit a regular pattern:

- $1,1,1,1,1,1,1,1,1,1,1,1,1\ldots$
- $0,0,0,0,0,0,0,0,0,0,0,0,0\ldots$
- $0,0,0,1,1,1,1,1,1,1,1,1,1\ldots$
- $0,0,1,0,0,1,0,0,1,0,0,1,0,0,1\ldots$
- $0,1,0,0,1,0,0,0,1,0,0,0,0,1\ldots$

These sequences are no less likely than a sequence like $0,1,0,0,1,1,1,0,1,0,0,\ldots$ They all have the same chance of occurring. There are infinitely many sequences for which the proportion of heads does not converge to $1/2$. For example, simply changing a single element of the sequence $1,1,1,1,1,1,1,1,1,1,1,1,1,\ldots$ generates the sequences

- $0,1,1,1,1,1,1,1,1,1,1,1,1\ldots$
- $1,0,1,1,1,1,1,1,1,1,1,1,1\ldots$
- $1,1,0,1,1,1,1,1,1,1,1,1,1\ldots$
- $1,1,1,0,1,1,1,1,1,1,1,1,1\ldots$
- $1,1,1,1,0,1,1,1,1,1,1,1,1\ldots$ and etc.

I believe it is not difficult for you to imagine many other (infinite) possibilities of sequences of 0's and 1's that do not converge to $1/2$.

19.4 The Strong Law of Large Numbers

Let's denote Ω as the set of all possible outcomes of the experiment, which is a potentially infinite sequence of independent tosses of a fair coin. Each possible outcome $\omega \in \Omega$ is an infinite sequence of 1's and 0's. Thus,

$$\Omega = \{\omega = (a_1, a_2, \ldots) : a_n = 0 \text{ or } a_n = 1, n = 1,2,\ldots\}. \tag{19.1}$$

This set Ω has infinitely many elements. The random variables $X_1, X_2, X_3, \ldots$ represent the outcome of the coin in successive tosses. That is, $X_n(\omega) = a_n$, the binary indicator (0 or 1) for heads or tails. The sequence of random variables $Y_n = (X_1 + \ldots + X_n)/n$ is evaluated for each element $\omega \in \Omega$ in a similar manner. That is,

$$Y_n(\omega) = (X_1(\omega) + \ldots + X_n(\omega))/n = (a_1 + \ldots + a_n)/n.$$

Note that all the variables $X_1, X_2, \ldots$ as well as $Y_1, Y_2, \ldots$ are defined on the same set Ω. Let A be the event composed of all sequences ω of 1's and 0's in which the proportion $Y_n(\omega)$ of heads converges to $1/2$ as n increases. That is,

$$A = \left\{\omega = (a_1, a_2, \ldots) \in \Omega \text{ such that } \lim_{n\to\infty} \frac{\sum_{j=1}^{n} a_j}{n} = \frac{1}{2}\right\} = \left\{\omega \text{ such that } Y_n(\omega) \to \frac{1}{2}\right\}. \tag{19.2}$$

We usually use an even simpler notation for this event A. Typically, we simply write

$$A = [\lim Y_n \to 1/2]$$

This is just an abbreviated notation. Its meaning is the subset of results ω expressed in (19.2).

For any infinite sequence of 0's and 1's, there are only two possibilities: either it converges to $1/2$ and belongs to the event A, or it does not converge to $1/2$ and belongs to the complementary event $A^c = \Omega - A$. Thus, we can partition the sample space Ω into two infinite and disjoint subsets: $\Omega = A \cup A^c$. Now, although containing infinite sequences, the set A^c is of negligible size. Its size is so small that it can be rigorously shown that $\mathbb{P}(A^c) = 0$ and $\mathbb{P}(A) = 1$!! Although it is considered

poor writing style, the two exclamation marks preceding are there to express how astonished I was when I first saw this result. It can be rigorously shown that

$$\mathbb{P}(A) = \mathbb{P}\left(\{\omega \text{ such that } \lim_{n\to\infty} Y_n(\omega) \to 1/2\}\right) = 1\,.$$

Using the simplified notation, we say that

$$\mathbb{P}(\lim_n Y_n \to 1/2) = 1$$

and therefore

$$\mathbb{P}(\lim_n Y_n \not\to 1/2) = \mathbb{P}(A^c) = 1\mathbb{P}(A) = 0$$

This is surprising. Despite the existence of infinite ω sequences for which we do not have convergence of $Y_n(\omega)$, they have a zero probability of occurring. It means that, when repeatedly flipping the coin as shown in 19.3, one of the sequences (the red line in the left graph) may deviate from the limit $1/2$. However, with a longer waiting time (as shown in the right graph), that red line will eventually approach $1/2$ *and it will never more deviates any substantial amount from 1/2*. More precisely, for any small distance $\varepsilon > 0$ (arbitrary but fixed), there will be a finite time after which this arbitrary red sequence will forever be confined to the narrow band of ε width around 1/2. As we can make ε as small as we want, this means that we are have no chance (it has probability 0) of seeing a sequence that belongs to A^c.

The proof of this result is lengthy, highly technical, and beyond the scope of this book. It can be found in Section 5.3 of the excellent book by Barry James [22]. Another elementary proof is presented by [15] and by [23]. An elementary proof means that it does not rely on advanced concepts, definitions, or results, but only on basic results. However, elementary does not necessarily mean easy.

In order to obtain powerful results like this, we need the formal definitions of $\sigma-$algebras that we have seen in Chapter 3. The abstract theory of measures (sizes) of sets is essential for the rigorous and more advanced study of probability. Indeed, there is a subtle and technical aspect that we are glossing over here. We know how to calculate probabilities associated with a finite number n of coin tosses but how to define probabilities for outcomes $\omega \in \Omega$ defined in (19.1)? How to calculate the probability of an event such as A in (19.2), which involves the infinite sequence of tosses? We need the Carathéodory's extension theorem, an advanced result in measure theory, that shows that the simple and easy sequence of probability models defined for finite n, with $\Omega_n = \{(a_1, a_2, \ldots, a_n) : a_i = 0 \text{ or } a_i = 1, i = 1, \ldots, n\}$ has a unique extension to a probability model defined for Ω in (19.1).

Once again, it is the measure of non-enumerable infinite sets that underlies the results above. The set A^c of non-convergent sequences, in a precise sense, is of an infinitely smaller order than the set A. It corresponds to the set of real numbers in the interval $[0,1]$ for which, when considering their binary expansion in (**??**), we have $\sum_n b_n/n \not\to 1/2$. This set is of such a small order that it has a probability of occurring equal to zero.

This result is not limited to coin experiments alone. It can be applied to variables with any probability distribution. The general theorem is called the Strong Law of Large Numbers (or SLLN, for short). It concerns the arithmetic mean

$$\bar{X}_n = (X_1 + X_2 + \ldots + X_n)/n$$

of a large number n of independent and identically distributed (i.i.d.) random variables with expected value $\mathbb{E}(X_i) = \mu$. The SLLN states that, with probability 1, $\bar{X}_n \to \mu$. Therefore, if n is large enough for convergence to have taken effect, we should have $\bar{X}_n \approx \mu$.

The SLLN was obtained under various assumptions and in different versions during the 1920s and 1930s by Borel (French), Cantelli (Italian), and the Russians Kolmogorov and Khintchin, among other probabilists. We will present two versions of the SLLN without proofs. In both versions, the variables X_i must have a finite expected value $\mathbb{E}(X_i) = \mu < \infty$. Indeed, there are some distributions whose expected value is infinite or undefined. However, the vast majority of distributions of practical interest have a defined expected value. Two distributions of practical relevance that do not have an expected value are the Pareto distribution (for certain values of its parameter) and the Cauchy distribution. However, for most practically relevant distributions, we have the theorem assumption satisfied.

Theorem 19.4.1 — First Strong Law of Large Numbers by Kolmogorov. Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathbb{E}(X_i) = \mu < \infty$. Then,

$$\mathbb{P}\left(\lim_n \bar{X}_n \to \mu\right) = 1$$

Prova: See Barry James [22, page 210]. ■

■ **Example 19.1 — Tossing unfair coins.** We have seen the example of the proportion of heads after n independent tosses of fair coins (i.e., with $\mathbb{P}(\text{heads}) = 1/2$). For this situation, the random variables $X_1, X_2, \ldots$ in Theorem 19.4.1 are i.i.d. with a Bernoulli distribution, where $X_i = 1$ with probability $\theta = \mathbb{P}(\text{cara}) = 1/2$ and $X_i = 0$ with probability $1/2$. Thus, $\mu = \mathbb{E}(X_i) = 1 * (1/2) + 0 * (1/2) = 1/2$. Since X_n is the proportion of heads in the first n tosses, we have that, with probability 1, this proportion of heads will converge to $\mu = 1/2$:

$$\mathbb{P}\left(\lim_n \bar{X}_n \to \frac{1}{2}\right) = 1$$

Now suppose that the coin is unfair. That is, $\theta = \mathbb{P}(\text{cara}) \neq 1/2$. Then $\mu = \mathbb{E}(X_i) = \theta$, and again by Theorem 19.4.1,

$$\mathbb{P}\left(\lim_n \bar{X}_n \to \theta\right) = 1\,.$$

This means that, if you keep tossing a biased coin indefinitely and select any small $\varepsilon > 0$ width, after some n_0, we will see $\bar{X}_n$ confined within a *varepsilon* band around from θ. This n_0 value depends on the specific infinity-long sequence $\omega = (a_1, a_2, \ldots)$.

Thus, the proportion of successes in repeated trials of a binary experiment converges with certainty to the theoretical probability θ of a single success occurring, regardless of the value of θ.
■

■ **Example 19.2 — Cholesterol level.** Let $X_1, X_2, \ldots$ represent the cholesterol levels of individuals randomly and independently chosen from a large population (millions of inhabitants). The value of X_i is random since the individual is chosen randomly. Furthermore, as they are chosen by the same random mechanism (a lottery), the random variables X_i are identically distributed. Let $\mathbb{E}(X_i) = \mu$ be the expected value of cholesterol in this population. If n is large, we can expect $\bar{X}_n = (X_1 + \ldots + X_n)/n$ to be approximately equal to μ because, according to Theorem 19.4.1, we have

$$\mathbb{P}\left(\lim_n \bar{X}_n \to \mu\right) = 1\,.$$

■

A practical and relevant problem is to determine the sample size n that would be sufficient for $\bar{X}_n$ to be approximately equal to μ. We will see an answer to this problem as a simple application of the Central Limit Theorem, the limit theorem discussed in section ??.

The version of the SLLN presented in theorem 19.4.1 requires the random variables to be i.i.d. (independent and identically distributed). This may not always be the case. For example, consider a basketball player practicing shooting. The sequence of shots can be represented by the binary random variables $X_1, X_2, \ldots$. Let $\mathbb{P}(X_i = 1) = \theta_i$. Let's imagine that the shots are independent, but the success probability θ_i does not remain constant throughout the shots. It increases slightly at the beginning of the practice session due to warm-up effects. Then, during the practice, it varies according to the player's level of concentration. Thus, the binary random variables are independent but not identically distributed because θ_i varies. For this situation, we can define an average of the expected values:

$$\bar{\theta}_n = \frac{\theta_1 + \theta_2 + \ldots + \theta_n}{n}.$$

The value $\bar{\theta}_n$ is a real number, not a random variable. It may be an unknown number but it is not stochastic. It does not have a list of possible values and associated probabilities.

In contrast,

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$$

is a random variable. What is the limiting behavior of $\bar{X}_n$? As the random variables X_i are not identically distributed (the θ_i's are not constant), we can not use the The version of the SLLN presented in theorem 19.4.1. We know that $\bar{X}_n$ does not converge to a constant value $\mathbb{P}(X_i = 1) = \theta$ (which does not exist in this case). What we obtain is that $|\bar{X}_n - \bar{\mu}_n|$ converges to zero. This means that for sufficiently large n, we will have $\bar{X}_n \approx \bar{\mu}_n$. We have the following version of the SLLN that holds for any distribution of independent random random variables X_i, not just binary ones:

Theorem 19.4.2 — Second Strong Law of Large Numbers by Kolmogorov. Let $X_1, X_2, \ldots$ be independent random variables with $\mathbb{E}(X_i) = \mu_i < \infty$. Assume that

$$\sum_{n=1}^{\infty} \frac{\mathbb{V}(X_n)}{n^2} < \infty$$

Then

$$\mathbb{P}\left(\lim_n |\bar{X}_n - \bar{\mu}_n| \to 0\right) = 1$$

onde $\bar{\mu}_n = (\mu_1 + \ldots + \mu_n)/n$.

The Theorem 19.4.2 will have limited use in this introductory book, but you can expect to encounter it frequently if you delve deeper into the theory.

■ **Example 19.3 — Different coins.** Suppose the experiment consists of tossing a different coin at each moment, with a probability θ_n of obtaining a head on the n-th toss. Since $X_i \sim \text{Bernoulli}(\theta_i)$, we have $\mathbb{E}(X_i) = \theta_i$ and $\mathbb{V}(X_i) = \theta_i(1-\theta_i) \leq 1/4$ for $\theta_i \in [0,1]$. This last statement will appear in other parts of this book, so it is better to make sure it is correct. Let $g(x) = x(1-x) = x - x^2$ where $x \in [0,1]$. Then $g(x)$ is a parabola with roots at $x = 0$ and $x = 1$, and with a maximum point at $x = 1/2$, reaching the maximum value $g(1/2) = 1/4$. Therefore, $\mathbb{V}(X_i) = \theta_i(1-\theta_i) \leq 1/4$. To use Theorem 19.4.2, let's check if its assumption is valid:

$$\sum_{n=1}^{\infty} \frac{\mathbb{V}(X_n)}{n^2} = \sum_{n=1}^{\infty} \frac{\theta_i(1-\theta_i)}{n^2} \leq \sum_{n=1}^{\infty} \frac{1/4}{n^2} = \frac{1}{4}\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{1}{4}\frac{\pi^2}{6}.$$

The convergence of the series formed by the inverse of the squares of natural numbers is a famous mathematical result obtained by Leonard Euler in 1741 and known as the Basel problem. For more details, you can refer to `https://en.wikipedia.org/wiki/Basel_problem`. The book *Euler, the master of us all* by William Durham provides a beautiful description of this problem and its solution by Euler. It revisits Euler's original works and presents the proofs and reasoning followed by Euler in a didactic and enjoyable manner, using modern notation. This is a highly recommended book.

We can obtain a simple result if we assume that the values θ_i themselves are i.i.d. random variables oscillating around a value $\mu \in [0,1]$. In other words, $\mathbb{E}(\theta_i) = \mu$. Some θ_i values may be above μ, while others may be below μ. By the 19.4.1 Theorem, we can conclude that

$$\bar{\theta}_n = (\theta_1 + \ldots + \theta_n)/n \to \mu\,.$$

This happens with probability 1. Therefore, if n is sufficiently large, we will have $|\bar{\theta}_n - \mu| < \varepsilon/2$. According to the 19.4.2, Theorem, we have$|\bar{X}_n - \bar{\theta}_n| < \varepsilon/2$ for n sufficiently large. Hence,

$$|\bar{X}_n - \mu| = |\bar{X}_n - \bar{\theta}_n + \bar{\theta}_n - \mu| \leq |\bar{X}_n - \bar{\theta}_n| + |\bar{\theta}_n - \mu| \leq \varepsilon/2 + \varepsilon/2 = \varepsilon$$

We conclude that $\bar{X}_n \approx \mu$ if n is large enough.

Consider an AB experiment where n randomly chosen users are presented with a new webpage layout. We want to infer the probability of any user clicking on the advertisement. We know that users have different propensities θ_i to click on the ad. If we assume that the individuals are independently selected and that the θ_i's are i.i.d. with $\mu = \mathbb{E}(\theta_i)$. By using a (large) sample of n users and observing the proportion $\bar{X}_n$ of them who click on the ad, we know that we are obtaining an approximate value for $\mu = \mathbb{E}(\theta_i)$. ∎

∎ **Example 19.4 — Repetition learning.** Consider a learning experiment in which a task is repeated, and in each repetition, we either have success $(X_i = 1)$ or failure $(X_i = 0)$. Success or failure are random events with $\mathbb{P}(X_i = 1) = \theta_i$. The success rate systematically varies with learning. For example, it may be that θ_i monotonically increases towards a value μ. Alternatively, θ_i may initially oscillate and eventually stabilize around μ. In all these cases, we assume that the sequence θ_i converges to a value μ. The values θ_i are fixed and deterministic, so the convergence of these values follows the usual convergence of sequences of real numbers (see Section 19.2). What can we learn if we perform a long sequence $X_1, X_2, \ldots$ of binary random experiments? Using the same argument as in the previous example, Theorem 19.4.2 states that we can expect $\bar{X}_n \approx \mu$. ∎

∎ **Example 19.5 — Divergent variables.** Theorem 19.4.2 covers more delicate and less obvious situations than Theorem 19.4.1. Let $X_1, X_2, \ldots$ be independent but not identically distributed random variables. Suppose that $X_n \sim N(n,1)$, with an expected value that increases with n. What behavior can we expect for $\bar{X}_n = (X_1 + \ldots + X_n)/n$ as n grows? The distribution of $\bar{X}_n$ is Gaussian since it is a linear combination of Gaussians. We just need to find its mean and variance. We have

$$\mathbb{E}(X_n)) = \frac{1}{n}\sum_{k=1}^{n} k = \frac{1}{n}\frac{n(n+1)}{2} = \frac{n+1}{2},$$

which tends to increase without bound as n increases. By independence,

$$\mathbb{V}(\bar{X}_n) = \frac{1}{n^2}(\mathbb{V}(X_1) + \ldots + \mathbb{E}(V_n)) = \frac{1}{n^2}\sum_{k=1}^{n} 1 = \frac{1}{n}.$$

To use Theorem 19.4.2, we calculate

$$\bar{\mu}_n = \frac{1}{n}(\mathbb{E}(X_1) + \ldots + \mathbb{E}(X_n)) = \frac{n+1}{2}.$$

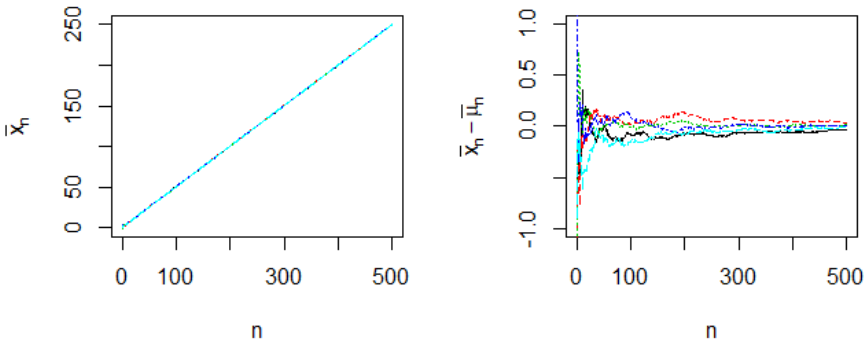


Figure 19.3: Five random sequences of $\bar{X}_n = (X_1 + \ldots + X_n)/n$ are generated, where $X_k \sim N(k,1)$ are independent random variables. The left hand side demonstrates that $\bar{X}_n$ diverges without bound. The right hand side shows that $\bar{X}_n - \bar{\mu}_n \to 0$.

Figure 19.5 displays the realization of five random sequences of $\bar{X}_n$ with $n = 1,\ldots,500$. The left hand side demonstrates that the five lines (almost indistinguishable from each other) diverge without bound. The right hand side illustrates the plot of $\bar{X}_n - \bar{\mu}_n$, which, according to Theorem 19.4.2, should converge to 0 with probability 1.

■

R The term "Law of Large Numbers" in the names of Theorems 19.4.1 and 19.4.2 refers to the number n of elements used to calculate the mean $\bar{X}_n = (X_1 + \ldots + X_n)/n$. The theorem guarantees that if n is *large* enough, we should have $\bar{X}_n \approx \mu$. We may achieve rapid convergence of $\bar{X}_n$ or we may have to wait for a long time. The theorem does not prevent us from experiencing very fast convergence, leading to excellent approximations between $\bar{X}_n$ and μ even with a small n. This can occur depending on the probability distribution of the random variables X_i. However, in other situations, this convergence may take a long time. Fortunately, the Law of Large Numbers guarantees that if we wait long enough, convergence will eventually occur. Unfortunately, the Law of Large Numbers does not provide information about the rate at which convergence will happen, as it can be either fast or slow depending on the distribution of X_i. If this information is required, we can have some approximate answer by simulating the process in the computer.

19.4.1 Limit of Sums of Functions of Random Variables

The Strong Law of Large Numbers is not as restrictive as it may seem given that it mentions only what is the limit behavior of the arithmetic mean $\bar{X}_n = (X_1 + \ldots + X_n)/n$. One way to see how it can be useful is by considering the sequence of the square elements of a given sequence. Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with $\mathbb{E}(X_i) = \mu$ and $\mathbb{V}(X_i) = \sigma^2$. Consider the new random variables $Y_1 = X_1^2$, $Y_2 = X_2^2, \ldots$. The random variables $Y_1, Y_2, \ldots$ are independent and identically distributed, and $\mathbb{E}(Y_i) = \mathbb{E}(X_i^2) = \mathbb{V}(X_i) + (\mathbb{E}(X_i))^2 = \sigma^2 + \mu^2$.

Applying Theorem 19.4.1 to the random variables $Y_1, Y_2, \ldots$, we have $\bar{Y}_n \approx \sigma^2 + \mu^2$. This means that

$$\bar{Y}_n = \frac{1}{n}\left(X_1^2 + \ldots + X_n^2\right) \approx \mathbb{E}(Y_i) = \sigma^2 + \mu^2$$

We have $\bar{X}_n \approx \mu$. Thus, if we have a large sample size n, we can guarantee that the sample variance

S^2 will be close to the theoretical variance. The reason is that

$$S^2 = \frac{1}{n}\sum_i (X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_i X_i^2 - (\bar{X}_n)^2 \to (\sigma^2 + \mu^2) - \mu^2 = \sigma^2$$

(checar notacao nos primeiro capitulos: Usei S^2??)

In general, let g be a mathematical function (such as $g(x) = x^2$) and define $Y_i = g(X_i)$. Let $X_1, X_2, \ldots$ be i.i.d. random variables. The SLLN immediately provides that

$$\bar{Y}_n = \frac{1}{n}(Y_1 + \ldots + Y_n) = \frac{1}{n}\left(g(X_1) + \ldots + g(X_n)\right) \to \mathbb{E}(Y_i) = \mathbb{E}(g(X_i))$$

■ **Example 19.6** Let $X_1, X_2, \ldots$ be i.i.d. random variables, and form the new sequence of i.i.d. random variables given by $Y_1 = e^{-X_1}$, $Y_2 = e^{-X_2}$, etc. Then,

$$\bar{Y}_n = \frac{1}{n}(Y_1 + \ldots + Y_n) = \frac{1}{n}\left(e^{-X_1} + \ldots + e^{-X_n}\right) \to \mathbb{E}(Y_i) = \mathbb{E}(e^{-X_i})$$

■

19.5 Almost surely convergence

In the SLLN, the sequence $X_1, X_2, \ldots$ does not converge to any value. It is the derived sequence of their successive arithmetic means $\bar{X}_1, \bar{X}_2, \bar{X}_3, \bar{X}_4, \ldots$ that converges. In the case of Theorem 19.4.1, the random sequence $\bar{X}_n$ converges to a constant, fixed, non-random value given by the expectation $\mu = \mathbb{E}(X_i)$ of the component random variables of the mean $\bar{X}_n$. Theorem 19.4.2 covers more subtle cases where there is no constant, fixed value to which $\bar{X}_n$ converges. There are even more subtle situations, as we will see in Example 19.7.

The Strong Law of Large Numbers is an example of a limit process involving random variables called "almost sure convergence." In the case of SLLN, we are interested in the limit value of the arithmetic mean of a sequence of random variables, and we learn that $\bar{X}_n \to \mu$. The concept of almost sure convergence deals with any sequence of random variables, not necessarily the sequence of arithmetic means of another sequence. Moreover, the limit value does not have to be a fixed, non-random number. The limit value of the sequence can be another random variable X.

Let $Y, Y_1, Y_2, \ldots$ be random variables defined on the same probability space $(\Omega, \mathscr{A}, \mathbb{P})$.

Definition 19.5.1 — Almost sure convergence. Y_n converges to Y *almost surely* if $\mathbb{P}(Y_n \to Y) = 1$, i.e., if the event $B = \{\omega \in \Omega \text{ such that } Y_n(\omega) \to Y(\omega)\}$ has probability 1 (or $\mathbb{P}(B) = 1$). We denote this by $X_n \overset{a.s.}{\to} X$.

The SLLN of Theorem 19.4.1 falls directly within this definition. If $X_1, X_2, \ldots$ are i.i.d. random variables, we simply take $Y_n = (X_1 + \ldots + X_n)/n$ and define the random variable Y to be equal to the constant $\mu = \mathbb{E}(X_i)$. That is, $Y \equiv \mu$, or $\mathbb{P}(Y = \mu) = 1$.

The SLLN of Theorem 19.4.2 also falls directly within Definition 19.5.1. Take

$$Y_n = \frac{(X_1 - \mu_1) + (X_2 - \mu_2) + \ldots + (X_n - \mu_n)}{n}$$

and, as before, let Y be equal to the constant 0.

The next example shows a situation in which Y is not a non-random constant. In this example, we will have the limiting random variable Y with a uniform distribution $U(0,1)$. This example is a bit longer and will not be used again in this book. Its purpose is to illustrate a situation where Y_n converges almost surely but not to a constant. Although it can be skipped in a first reading, it has a certain fascination. The example will demonstrate that flipping a coin infinitely many times is equivalent to choosing a real number from the interval $[0,1]$ with a uniform distribution. This is one of those highly surprising mathematical results.

■ **Example 19.7 — Borel numbers.** Consider once again the set Ω associated with the sequence of infinitely many successive tosses of a fair coin:

$$\Omega = \{\omega = (a_1, a_2, \ldots) : a_n = 0 \text{ ou } a_n = 1, n = 1, 2, \ldots\}.$$

Ω can be identified with the set of real numbers in the interval $[0,1]$. In fact, let $\omega = (a_1, a_2, \ldots) \in \Omega$. Define

$$r_n(\omega) = \sum_{j=1}^{n} \frac{a_j}{2^j} = \frac{a_1}{2} + \frac{a_2}{2^2} + \frac{a_3}{2^3} + \ldots + \frac{a_n}{2^n}.$$

First, note that $0 \leq r_n(\omega)$ for every n and every ω because $r_n(\omega)$ is the sum of non-negative terms. Secondly, for every ω, the value of $r_n(\omega)$ has a limit as $n \to \infty$. That is, the limit $\lim_{n\to\infty} r_n(\omega) = r(\omega)$ exists. To prove this, we can use the comparison test for series (see `https://en.wikipedia.org/wiki/Convergent_series`). If $0 \leq b_j \leq c_j$ and c_j forms a convergent series, then the series b_n converges. Since $a_j/2^j \leq 1/2^j$, we have

$$r_n(\omega) = \sum_{j=1}^{n} \frac{a_j}{2^j} \leq \sum_{j=1}^{n} \frac{1}{2^j} \to \sum_{j=1}^{\infty} \frac{1}{2^j} = 1.$$

Therefore, we conclude that there exists a limit $r(\omega)$ of $r_n(\omega)$ for every ω, and furthermore, $r(\omega) \in [0,1]$:

$$r_n(\omega) \to r(\omega) = \sum_{j=1}^{\infty} \frac{a_j}{2^j} \in [0,1].$$

Thus, every element $\omega \in \Omega$ can be represented by a real number in the interval $[0,1]$.

On the other hand, every real number $u \in [0,1]$ has a binary representation of the form

$$u = \sum_{j=1}^{\infty} b_j/2^j \tag{19.3}$$

where b_j is equal to 0 or 1. We can identify the digits b_j occurring in this expansion with the point $\omega(u) = (b_1, b_2, \ldots) \in \Omega$. Thus, to every real number $u \in [0,1]$, there corresponds an element $\omega(u) \in \Omega$.

There is a small technical difficulty: two elements $\omega_1 \neq \omega_2$ can have the same limit $r(\omega_1) = r(\omega_2)$. Similarly, a real number u can have two different binary expansions. It is possible to show that the only way for two different expansions to generate the same real number is when one expansion is $0.a_1a_2\ldots a_{n-1}10000\ldots$ and the other is $0.a_1a_2\ldots a_{n-1}01111\ldots$. To avoid this issue, you can take the numbers in the range $[0,1]$ which correspond to a sequence not ending with infinite many zeros or infinite many ones.

?? I need to finish this example. The remaining text in this section need to be translated and finished ??

Vamos verificar agora que esta associação é biunívoca. Os dois conjuntos (Ω e $[0,1]$) são essencialmente os mesmos. Esta identificação entre os dois conjuntos será bem útil no exemplo 19.7.

Vimos na seção 19.4 a equivalência matemática entre o o conjunto Ω de todos os resultados possíveis do experimento conceitual de infinitos lançamentos independentes de uma moeda honesta, com

$$\Omega = \{\omega = (a_1, a_2, \ldots) : a_n = 0 \text{ ou } a_n = 1, n = 1, 2, \ldots\},$$

e o intervalo $[0,1]$ de números reais. Seja $X_n(\omega) = a_n$, o indicador binário (1 ou 0) de cara ou coroa, no n-ésimo lançamento da moeda. Entretanto, não consideramos como a estrutura de probabilidade associada com o experimento da moeda é refletido na distribuição de probabilidade para o número. As variáveis aleatórias $X_1, X_2, X_3, \ldots$ representam o resultado da moeda nos lançamentos sucessivos. Isto é, $X_n(\omega) = a_n$, o indicador binário (0 ou 1) de cara ou coroa. A sequência de v.a.'s $Y_n = (X_1 + \ldots + X_n)/n$ é avaliada em cada elemento $\omega \in \Omega$ de forma similar. Isto é,

$$Y_n(\omega) = (X_1(\omega) + \ldots + X_n(\omega))/n = (a_1 + \ldots + a_n)/n\,.$$

O conjunto Ω possui infinitos elementos. Ele pode ser identificado ao conjunto de números reais no intervalo $[0,1]$. De fato, seja $\omega = (a_1, a_2, \ldots) \in \Omega$. Defina

$$r(\omega) = \sum_{j=1}^{\infty} \frac{a_j}{2^j} = \frac{a_1}{2} + \frac{a_2}{2^2} + \frac{a_3}{2^3} + \ldots$$

O número $r(\omega)$ está no intervalo $[0,1]$. Como cada $a_j = 0$ ou 1, claramente $r(\omega) \geq 0$ pois é a soma de termos não-negativos. Por outro lado, $r(\omega) \leq 1$ pois substituindo cada a_j pelo seu valor máximo (igual a 1), temos

$$0 \leq r(\omega) = \sum_{j=1}^{\infty} \frac{a_j}{2^j} \leq \sum_{j=1}^{\infty} \frac{1}{2^j} = 1\,.$$

onde usamos o conhecido valor limite de uma série geométrica. Assim, todo elemento $\omega \in \Omega$ pode ser representado por um número real no intervalor $[0,1]$. Por outro lado, todo número real $u \in [0,1]$ possui uma representação binária da forma $u = \sum_{j=1}^{\infty} b_j/2^j$ onde b_j é igual a 0 ou 1. Identifique os dígitos b_j ocorrendo nesta expansão com o ponto $(b_1, b_2, \ldots) \in \Omega$. Assim, os dois conjuntos (Ω e $[0,1]$) são essencialmente os mesmos. ∎

19.6 Convergence in probability

As we mentioned, there is more than one type of convergence of a sequence of random variables. After the almost sure convergence, the second type we are going to see is the *convergence in probability*. To understand the main difference between the two types of convergence we start with a very simple and illuminating example I read in a Quora answer by Joe Blitzstein (`https://qr.ae/pykLuQ`). Let $X_1, X_2, \ldots$ be independent binary random variables representing the outcome of a coin toss. As usual, Heads are 1, Tails are 0. They are not identically distributed. The probability $\theta_n = \mathbb{P}(X_n = 1)$ of Heads changes with n by systematically decreasing towards zero. For example, we could have $\theta_n = 1/n$, $\theta_n = 1/n^2$, or $\theta_n = e^{-n}$. In summary, we are assuming that the sequence X_n of independent Bernoulli random variables have $\theta_n \to 0$.

According to the definition that will be provided shortly, all of these sequences of random variables converge to zero in probability. As we will see, the convergence of X_n to zero in probability is equivalent to the convergence of the sequence of real numbers θ_n to 0. In the context of this example, the definition specializes to $\mathbb{P}(X_n = 1) = \theta_n \to 0$. In other words, convergence to 0 in probability implies that the occurrence of 1's becomes increasingly rare as we progress through the sequence. However, we can not be assured that after a certain point, they will never appear again; they will only be observed less frequently.

But this a natural question: as the probability of obtaining heads decreases to zero, we can consider whether the sequence of 1's will eventually become extinct. That is, whether X_n will be equal to zero from a certain point onwards. If $X_n \to 0$ almost surely, we can be certain that, with a probability of 1, the random result of this experiment is that we will no longer observe any 1 after a

certain moment n_0. The value of n_0 may vary depending on the specific realization of the sequence, sometimes occurring earlier, while other times taking longer to manifest. However, the probability of encountering an infinite number of 1's in the sequence is zero.

Convergence almost surely relates to the pointwise or individual behavior of every single realized sequence. It is concerned with and tracks the behavior of each single realization $X_1(\omega), X_2(\omega), \ldots$ for each $\omega \in \Omega$ where Ω is defined in (19.1). It says that there no positive probability, no even a tiny one, that we will see a sequence $X_1(\omega), X_2(\omega), \ldots$ that does not converge to zero.

In contrast, convergence in probability does not track the sequences $X_1(\omega), X_2(\omega), \ldots$ It only cares about the collective behavior of the possible $X_n(omega)$ values as n grows. In principle, we can have a probability greater than zero of seeing non-convergent sequences $X_1(\omega), X_2(\omega), \ldots$. Indeed, we have a very famous theorem, called Borel-Cantelli theorem, that can be applied to separate out the distinct cases of the limiting behavior of this sequence of independent Bernoulli variables. As θ_n is a non-negative number, the sum $S = \sum_{n=1}^{\infty} \theta_n$ of all these infinite numbers has only two possibilities: it either converges (and $S < \infty$) or diverges (and $S = \infty$). It depends on how fast θ_n decreases to zero. For example, if $\theta_n = 1/n$, we know from calculus that $S = \infty$. However, if $\theta_n = 1/n^2$, then $S = \infty$. Using Borel-Cantelli theorem (not stated here), we have

- if $S < \infty$, with probability 1, we will have a finite number of successes. That is, $X_n \to 0$ almost surely.
- if $S = \infty$, with probability 1, we will have an infinite number of successes. That is, X_n does not converge almost surely to zero. There is never a last 1: you can always find another 1 further down the sequence.
- However, as $\theta_n \to 0$, we have X_n converging to zero in probability in all cases, irrespective of $S < \infty$ or $S = \infty$.

After this long digression, let us state the definition of convergence in probability. Let $Y, Y_1, Y_2, \ldots$ be random variables defined on the same probability space $(\Omega, \mathscr{A}, \mathbb{P})$.

Definition 19.6.1 — Convergence in probability. Y_n converges to Y *in probability* if $\mathbb{P}(|Y_n - Y| > \varepsilon) \to 0$ as $n \to 0$ for all $\varepsilon > 0$. We denote this by $X_n \xrightarrow{p} X$.

The best example of convergence in probability is given by the Weak Law of Large Numbers theorem. One good news is that it is very simple to prove it using the Tchebyshev's inequality.

19.7 Weak Law of Large Numbers

Let's review the Chebyshev's inequality. Consider a random variable X with expected value $\mu = \mathbb{E}(X)$ and variance $\mathbb{V}(X) = \sigma^2$. For any positive real number $\varepsilon > 0$, the inequality states that:

$$P(|X - \mu| \geq \varepsilon) \leq \frac{\mathbb{V}(X)}{\varepsilon^2} .$$

It is important to note that X in the theorem can be any random variable, and ε can be any positive number.

We are now in a position to prove the first theorem. Through it, we will see in a formal and rigorous way that our intuition is correct: if $X_1, X_2, \ldots$ are i.i.d. random variables with expected value μ, then the arithmetic mean of these random variables converges to μ. In what sense does this convergence occur? Remember that the arithmetic mean is a random variable. How quickly does the convergence happen?

Theorem 19.7.1 — Law of Large Numbers. Let $X_1, X_2, \ldots, X_n$ be an independent trials process, with finite expected value $\mu = E(X_j)$ and finite variance $\sigma^2 = V(X_j)$. Let $S_n = X_1 + X_2 + \ldots + X_n$.

Then for any $\varepsilon > 0$,

$$P\left(\left|\frac{S_n}{n} - \mu\right| \geq \varepsilon\right) \to 0$$

as $n \to \infty$. Equivalently,

$$P\left(\left|\frac{S_n}{n} - \mu\right| < \varepsilon\right) \to 1$$

as $n \to \infty$.

Proof: Since $X_1, X_2, \ldots, X_n$ are independent and have the same distribution, we obtain

$$\mathbb{E}\left(\frac{S_n}{n}\right) = \frac{1}{n}\left(\mathbb{E}(X_1) + \ldots + \mathbb{E}(X_n)\right) = \frac{n\mu}{n} = \mu$$

and

$$\mathbb{V}\left(\frac{S_n}{n}\right) = \frac{1}{n^2}\left(\mathbb{V}(X_1) + \ldots + \mathbb{V}(X_n)\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

By direct application of Chebyshev's inequality with the random variable S_n/n, we have for any *fixed* $\varepsilon > 0$,

$$P\left(\left|\frac{S_n}{n} - \mathbb{E}\left(\frac{S_n}{n}\right)\right| \geq \varepsilon\right) = P\left(\left|\frac{S_n}{n} - \mu\right| \geq \varepsilon\right) \leq \frac{\mathbb{V}(S_n/n)}{\varepsilon} = \frac{\sigma^2}{n\varepsilon^2} \to 0$$

as $n \to \infty$. Equivalently, as the event

$$A_n = \{\omega \in \Omega \text{ such that } \left|\frac{S_n(\omega)}{n} - \mu\right| \geq \varepsilon\}$$

has the complementar event

$$A_n^c = \{\omega \in \Omega \text{ such that } \left|\frac{S_n(\omega)}{n} - \mu\right| < \varepsilon\}$$

we know that

$$P\left(\left|\frac{S_n}{n} - \mu\right| < \varepsilon\right) = \mathbb{P}(A_n^c) = 1 - \mathbb{P}(A_n) \to 1 - 0 = 1\,.$$

Hence, if we have a random sample of i.i.d. random variables, there is a large probability that S_n/n will be close to μ if n is large enough. What does "large enough" mean? How large is considered "large"? Is it possible to determine a sample size that guarantees a good approximation? The rate of convergence is determined by how fast the probability decreases as n increases. This rate of convergence is influenced by the properties of the underlying random variables and their distribution. In some cases, the convergence can be relatively fast, with the probability approaching zero rapidly as n increases. In other cases, the convergence may be slower, requiring larger sample sizes for the probability to decrease significantly. Hence, a precise answer depends on the specific problem we have at hand. While this vague answer is frustrating, we can antecipate that a much better one will be given after we learn the third type of convergence in section 19.9 and the Central Limite Theorem in section 19.11.

Future issues: ?? Criar exercicios mostrando como usar simulacao Monte Carlo para decidir sobre isto ?? ?? Note that this theorem is not necessarily true if σ^2 is infinite. Explicar caso da Pareto (???).

19.8 Comparison Between Weak and Strong Laws

There is a simple way to contrast the two types of convergence, almost surely and in probability. Consider the situation in which the convergence is towards a non-random constant μ, as in the the law of large numbers theorems. Fix a positive but small distance band $\varepsilon > 0$ around L. We have almost surely convergence if

$$\mathbb{P}\left(\{\omega \in \Omega; \lim_{n\to\infty} |X_n(\omega) - L| < \varepsilon\right) = 1$$

for every ε selected. Convergence in probability says that

$$\lim_{n\to\infty} \mathbb{P}\left(\{\omega \in \Omega; |X_n(\omega) - L| < \varepsilon\right) = 1$$

We pass the lim outside the probability $\mathbb{P}$.

Are they equivalent? If a sequence X_n converges almost surely to X, does it converges to the X in probability? How about the other direction? Convergence in probability to X implies convergence to the same limit X almost surely? As we saw previously, there are examples where we have one type of convergence but not the other. Indeed, it turns out that one direction is correct: almost surely convergence is stronger in the sense that if $X_n \xrightarrow{a.s.} X$ then we also have $X_n \xrightarrow{p} X$. However, the converse is not true. Below, we give another example where we have convergence in probability but not almost surely.

EXAMPLE:

PROOF THAT almost surely convergence implies in convergence in probability

19.9 Convergence in distribution

19.10 Moment Generating Function

19.11 Central Limit Theorem - CLT

19.12 Proof of the CLT

19.12.1 Moment Generating Function

19.12.2 The proof

19.13 Applications of the CLT

19.14 (The Law of the Iterated Logarithm

Bibliography

Books

[Agr15] Alan Agresti. *Foundations of linear and generalized linear models*. John Wiley & Sons, 2015 (cited on page 328).

[BPP84] Robert Eric Beard Beard, Teivo Pentikäinen, and Erkki Pesonen. *Risk Theory: The Stochastic Basis of Insurance*. 1st edition. Monographs on Statistics and Applied Probability 20. Springer Netherlands, 1984 (cited on page 300).

[Bow+97] Newton L Bowers et al. *Actuarial Mathematics, 2nd Edition*. The Society of Actuaries, Schaumburg, Illinois, 1997 (cited on page 177).

[BM06] Robert R Bush and Frederick Mosteller. *A comparison of eight models*. Springer, 2006, pages 317–332 (cited on pages 259, 260).

[Dav12] Ernest Davis. *Linear algebra and probability for computer science applications*. CRC Press, 2012 (cited on page 110).

[Fel68] William Feller. *Probability theory and its applications*. 3rd edition. Volume 1. John Wiley and Sons, Inc., 1968 (cited on pages 140, 257, 259).

[Gia12] Eduardo Giannetti. *O valor do amanhã*. Editora Companhia das Letras, 2012 (cited on page 209).

[Jam96] Barry R James. *Probabilidade: Um Curso em Nível Intermediário, 2ª Edição*. IMPA, Coleção Euclides, 1996 (cited on pages 149, 190, 211, 441, 442).

[Ric81] Frank Proschan Richard E. Barlow. *Mathematical Theory of Reliability*. SIAM, 1981 (cited on page 184).

[RCE10] Ernest Rutherford, James Chadwick, and Charles Drummond Ellis. *Radiations from radioactive substances*. Cambridge University Press, 2010 (cited on page 149).

[YK58] G. Udny Yule and Maurice Kendall. *Introduction to the Theory of Statistics*. Griffin, London, 1958 (cited on pages 177, 198).

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 157).

[Bli35] CI Bliss. “The comparison of dosage-mortality data”. In: *Annals of Applied Biology* 22.2 (1935), pages 307–333 (cited on page 326).

[Cla46] R D Clarke. “An Application of the Poisson Distribution”. In: *Journal of the Institute of Actuaries* 72 (1946), page 481 (cited on page 150).

[CD81] RF Costantino and RA Desharnais. “Gamma distributions of adult numbers for Tribolium populations in the region of their steady states”. In: *The Journal of Animal Ecology* (1981), pages 667–681 (cited on page 192).

[Cra17] Charles Spencer Crandall. “Seed production in apples”. In: *Bulletin (University of Illinois (Urbana-Champaign campus). Agricultural Experiment Station); no. 203* (1917) (cited on page 323).

[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 193).

[ER60] P Erdös and A Rényi. “On the evolution of random graphs”. In: *Publications of the Mathemafical Institute of the Hungarian Academy of Sciences* 5 (1960) (cited on pages 141, 142).

[Ete81] Nasrollah Etemadi. “An elementary proof of the strong law of large numbers”. In: *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 55.1 (1981), pages 119–122 (cited on page 441).

[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 357).

[Gol+82] Lee Goldman et al. “A computer-derived protocol to aid in the diagnosis of emergency room patients with acute chest pain”. In: *New England journal of medicine* 307.10 (1982), pages 588–596 (cited on page 122).

[Hei+03] Grete Heinz et al. “Exploring relationships in body dimensions”. In: *Journal of Statistics Education* 11.2 (2003) (cited on page 187).

[Hou83] Michael Hout. “Mobility tables.” In: (1983) (cited on page 316).

[Luz18] Nuno Luzia. “A simple proof of the strong law of large numbers with rates”. In: *Bulletin of the Australian Mathematical Society* 97.3 (2018), pages 513–517 (cited on page 441).

[McC19] Brendan McCord. “Taking Up the Revolutionary Call: Principles to Guide a Purpose-Driven AI Future”. In: *Harvard Data Science Review* 1.1 (July 2019) (cited on page 28).

[MP01] Hyeonjoon Moon and P Jonathon Phillips. “Computational and performance aspects of PCA-based face-recognition algorithms”. In: *Perception* 30.3 (2001), pages 303–321 (cited on page 380).

[Nee+07] E Shannon Neeley et al. “Quantitative temporal lobe differences: autism distinguished from controls using classification and regression tree analysis”. In: *Brain and Development* 29.7 (2007), pages 389–399 (cited on page 123).

[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 76).

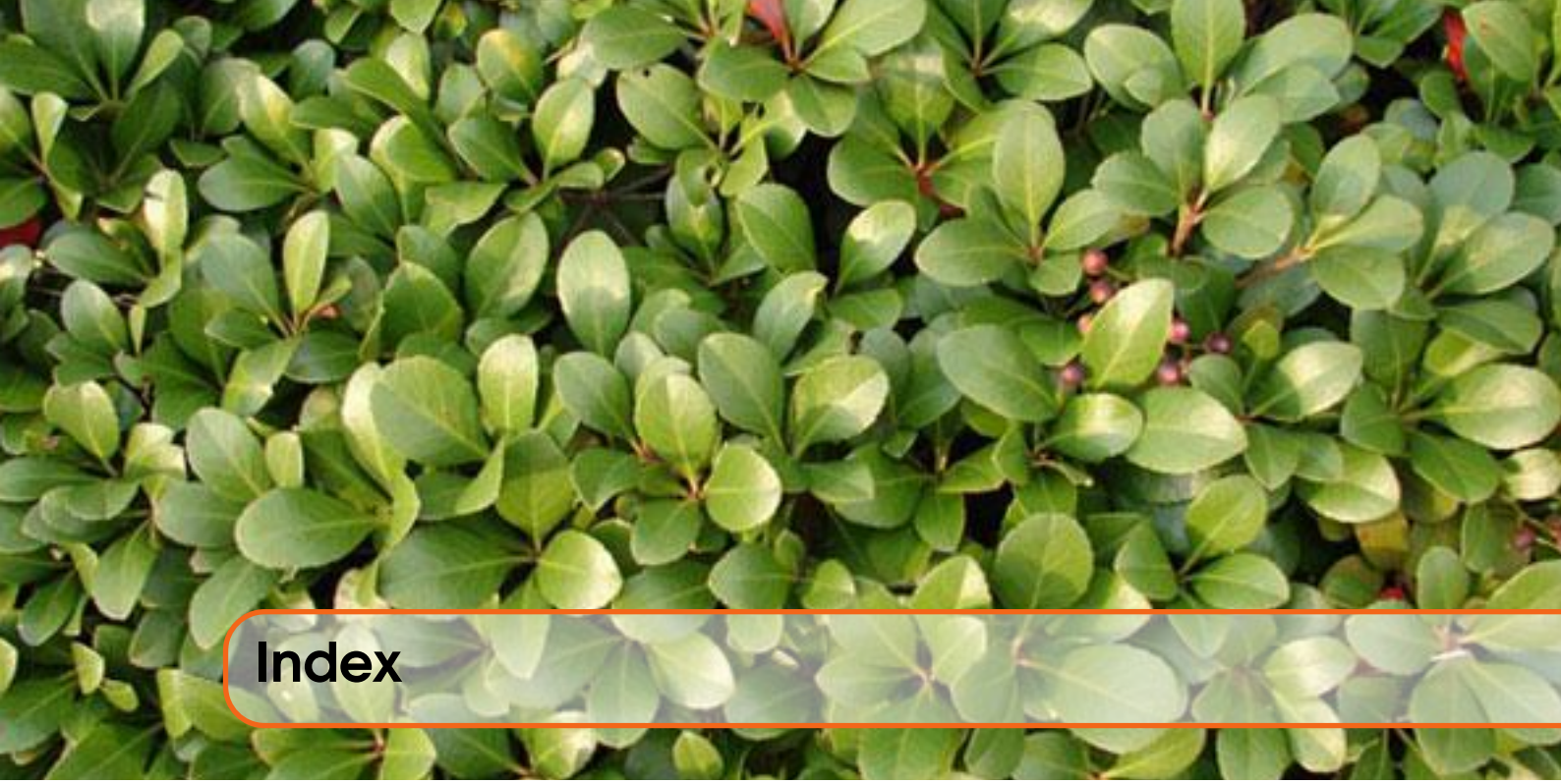
[Res+11] David N Reshef et al. “Detecting novel associations in large data sets”. In: *Science* 334.6062 (2011), pages 1518–1524 (cited on pages 50–52).

[RK04] A Rovlias and S Kotsou. “Classification and regression tree for prediction of outcome after severe head injury using simple clinical and laboratory variables”. In: *Journal of neurotrauma* 21.7 (2004), pages 886–893 (cited on page 123).

[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 193).

[TP91a] Matthew Turk and Alex Pentland. “Eigenfaces for recognition”. In: *Journal of cognitive neuroscience* 3.1 (1991), pages 71–86 (cited on pages 371, 380).

[Vel+19] Bráulio M Veloso et al. “In Search of a Stochastic Model for the E-News Reader”. In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 13.6 (2019), pages 1–27 (cited on page 160).



Index

law of total expectation, 428
law of total variance, 430