Modern Birkhäuser Classics

Manfred Denker
Wojbor Woyczynski

Introductory Statistics and Random Phenomena

Uncertainty, Complexity and Chaotic Behavior in Engineering and Science

Birkhäuser

Modern Birkhäuser Classics

Many of the original research and survey monographs, as well as textbooks, in pure and applied mathematics published by Birkhäuser in recent decades have been groundbreaking and have come to be regarded as foundational to the subject. Through the MBC Series, a select number of these modern classics, entirely uncorrected, are being re-released in paperback (and as eBooks) to ensure that these treasures remain accessible to new generations of students, scholars, and researchers.

Introductory Statistics and Random Phenomena

Uncertainty, Complexity and Chaotic Behavior in Engineering and Science

Manfred Denker
Wojbor Woyczynski

Reprint of the 1997 Edition

Birkhäuser

Manfred Denker
Department of Mathematics
Pennsylvania State University
State College, PA, USA

Wojbor Woyczynski
Math, Applied Math, & Stat, Yost 229
Case Western Reserve University
Cleveland, OH, USA

ISSN 2197-1803 ISSN 2197-1811 (electronic)
Modern Birkhäuser Classics
ISBN 978-3-319-66151-3 ISBN 978-3-319-66152-0 (eBook)
DOI 10.1007/978-3-319-66152-0

Library of Congress Control Number: 2017952061

Mathematics Subject Classification (2010): 62-01, 62P30

Printed on acid-free paper

This book is published under the trade name Birkhäuser, www.birkhauser-science.com
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Introductory Statistics and Random Phenomena

Uncertainty, Complexity and Chaotic Behavior in Engineering and Science

with *Mathematica® Uncertain Virtual Worlds™*
by Bernard Ycart

Manfred Denker
Wojbor A. Woyczyński

Springer Science+Business Media, LLC

Manfred Denker
Georg-August-Universität
Göttingen, GERMANY

Wojbor A. Woyczyński
Case Western Reserve University
Cleveland, OH

Bernard Ycart
Université Joseph Fourier
Grenoble, FRANCE

Library of Congress Cataloging-in-Publication Data
Denker, Manfred, 1944-
Introductory statistics and random phenomena : uncertainty, complexity, and chaotic behavior in engineering and science / M. Denker and W.A. Woyczyński ; with Mathematica Uncertain Virtual Worlds by B. Ycart.
p. cm. -- (Statistics for industry and technology)
Includes bibliographical references and index.
ISBN 978-0-8176-4031-6 ISBN 978-1-4612-2028-2 (eBook)
DOI 10.1007/978-1-4612-2028-2
1. Mathematical statistics. 2. Probabilities. 3. Mathematica (computer file) I. Woyczyński, W. A. (Wojbor Andrzej), 1943- II. Ycart, B. (Bernard), 1958- . III. Title. IV. Series.
QA276.12.D45 1998
519.5--dc21 98-4735
CIP

Printed on acid-free paper

ISBN 978-0-8176-4031-6

Mathematica® is a registered trademark of Wolfram, Inc.,
100 Trade Center Drive, Champaign, IL 61820-7237
Typeset by T & T Techworks Inc., Coral Springs, FL

9 8 7 6 5 4 3 2 1

Contents

To our wives,
JEANNE, LIZ, and TILLU

Preface

The present book is based on a course developed as part of the large NSF-funded Gateway Coalition Initiative in Engineering Education which included Case Western Reserve University, Columbia University, Cooper Union, Drexel University, Florida International University, New Jersey Institute of Technology, Ohio State University, University of Pennsylvania, Polytechnic University, and University of South Carolina. The Coalition aimed to restructure the engineering curriculum by incorporating the latest technological innovations and tried to attract more and better students to engineering and science. Drafts of this textbook have been used since 1992 in statistics courses taught at CWRU, Indiana University, Bloomington, and at the universities in Göttingen, Germany, and Grenoble, France.

Another purpose of this project was to develop a courseware that would take advantage of the Electronic Learning Environment created by CWRUnet—the all fiber-optic Case Western Reserve University computer network, and its ability to let students run *Mathematica* experiments and projects in their dormitory rooms, and interact paperlessly with the instructor.

Theoretically, one could try to go through this book without doing *Mathematica* experiments on the computer, but it would be like playing Chopin's *Piano Concerto in E-minor*, or Pink Floyd's *The Wall*, on an accordion. One would get an idea of what the tune was without ever experiencing the full richness and power of the entire composition, and the whole ambience would be miscued.

Acknowledgments

Thanks are due to several groups of students that have taken different versions of this course over the last six years. They patiently and consistently found mistakes, produced good graphics, and came up with interesting data sets from their own disciplines. Their individual contributions are acknowledged in the text. We also appreciate help from Tom Ryan and Jiming Jiang of the CWRU Statistics Department, who read various portions of the manuscript and pointed out places that could be improved upon. We thank Steve Pinkus of Yale University and Burt Singer of Princeton University for discussing with us their recent work on computable

framework for randomness. Piotr Biler of Wrocław University kindly agreed to read several chapter of the manuscript with his usual sharp eye towards details and inconsistencies, and we are grateful for his help. Leszek Sczaniecki of Wolfram Research, Inc., took an early interest in our project, put us on the list of *Mathematica* developers, and kept us current on the latest *Mathematica* developments. We thank him for his support. Neepa Mukherjee, a CWRU statistics graduate student, assisted us in preparation of the final versions of non-*Mathematica* figures and working with her was a pleasant experience. Finally, we want to acknowledge a patient and benevolent guidance and encouragement from Wayne Yuhasz and Lauren Lavery, our Birkhäuser editors.

Authors

Manfred Denker studied Mathematics and Physics at the University of Erlangen, Germany, and Warwick University, England, and received the Ph.D. in 1972 from Erlangen University. Since 1974, he has been Professor of Mathematics at the Georg-August-University of Göttingen, Germany, where he served as Chairman and Dean for several years. He has supervised many master's and doctoral dissertations and written/or edited 5 books on probability theory, statistics and dynamical systems. He has been Visiting Professor at many universities in North America, Asia and Europe. His research interest and teaching experience includes Applied Statistics, Nonparametric Statistics, Linear Models, Probability Theory, Fractal Geometry, Ergodic Theory and Dynamical Systems with applications to Biometrics, Biology, and Physics. At present he is directing research programs in the Graduate School on 'Turbulence and Instabilities' at Göttingen University, the National Research Program on 'Ergodic Theory, Dynamical Systems and Numerical Simulations' and a bilateral Research Program in Statistics on 'Bioequivalence Studies and Model Validation'.

Wojbor A. Woyczynski received his B.S./M.Sc. in Electrical and Computer Engineering from Wrocław Polytechnic in 1966 and a Ph. D. in Mathematics in 1968 from Wrocław University, Poland. He moved to the U.S. in 1970, and since 1982 has been Professor of Mathematics and Statistics at Case Western Reserve University in Cleveland, and served as chairman of the department from 1982 to 1991. Previously he held tenured faculty positions at Wrocław University, Poland, and at Cleveland State University, and visiting appointments at Carnegie-Mellon University, Northwestern University, University of North Carolina, University of South Carolina, University of Paris, Göttingen University, Aarhus University, Nagoya University, University of Minnesota, University of Tokyo, and the University of New South Wales in Sydney. He is also (co-)author and/or editor of nine books on probability theory, harmonic and functional analysis, and applied mathematics, and serves as a member of editorial boards of the *Annals of Applied Probability, Probability Theory and Mathematical Statistics,* and the *Stochastic Processes and Their Applications.* His research interests include probability theory, stochastic models, functional analysis, and partial differential equations and their applications in statistics, statistical physics, surface chemistry, and hydrodynamics. He is currently Director of the CWRU Center for Stochastic and Chaotic Processes in Science and Technology.

Bernard Ycart received his Ph.D. in Mathematics in 1983 from Toulouse University, France. Since 1984 he has been Professor of Mathematics at the University of Pau, France, and in 1992 was appointed Professor of Applied Mathematics at the University of Grenoble. Subsequently he became the head of research groups in Statistics and Stochastic Modeling, and in Random Models in Computer Science. His visiting appointments include a year at Case Western Reserve University and stints at the University of Santiago, Chile, University of Zaragoza, Spain, and Oxford University, England. His research interests lie in applied probability and he is the author of several stochastic processes simulation packages.

Introduction

Goal and Audience

The present book is intended as a text for an introductory level statistics course. It addresses the phenomenon of uncertainty, which appears in most of the engineering and scientific problems for various reasons, and which can be modeled in several, basically different, ways. The book's novelty is integration of ideas about statistics of random phenomena stemming from three distinct viewpoints:

algorithmic/computational complexity,
classical probability theory, and
chaotic behavior in nonlinear systems.

Given an elementary level of the textbook and an anticipated preparation of the targeted audience, the exposition depends heavily on the *Mathematica*[1] computer experimentation and simulations by the students. Here, we would like to think about instruction proceeding in an environment of *Uncertain Virtual Worlds (UVW)*, and we provide some *Mathematica* tutoring as we move along. The goal is to give engineering and science students a forward looking alternative to the usual introductory statistics courses, an alternative that we feel will become the norm of the future as pressures to incorporate a study of algorithmic complexity and chaos-induced uncertainty increases in the already crowded curriculum. Such a course can be comfortably and profitably taken by either upper division undergraduate students or graduate engineering and science students who have never had a statistics course before.

Prerequisites include a typical engineering/science 2-3 semester calculus sequence (including some differential equations and linear algebra) in addition to a basic programming course in computer science (generally taken during the student's first year). The course can serve both as an important technical engineering/science statistics elective and, possibly, as a mathematics or statistics curricular requirement.

[1] Or any other symbolic manipulation software, such as, e.g., *Maple*.

Typically, a course like this would be taught out of a Statistics Department. However, in many schools, departments of Mathematics, Mathematics and Statistics, Applied Mathematics, or even some non-mathematical sciences departments (such as Industrial Engineering, Systems Engineering and Operations Research) could be responsible for this course.

Although the primary deliverers of this course would be statisticians, the course should be fun as well to teach for mathematicians and broader-minded engineers. It goes beyond the orthodox beginning statistical offering (same for, more or less, the last 50 years) to some mathematically thrilling territory, while maintaining a fairly introductory level accessible to broad student audiences.

Philosophy

Persi Diaconis is fond of saying that "statistics is a physics of numbers" and our philosophy is not too far from that statement. Loosely speaking, the book will emphasize statistics as a science (as opposed to a formal abstract theory and a branch of probability theory) concerned with all facets of handling large numerical data sets. It very much subscribes to the standard scientific methodology: proceed from experiment to inductive inference. It is woven around themes like data collection, compression, representation and analysis, modeling of random phenomena, model identification and design of experiments. We emphasize that all the data actually collected in today's computerized environment are discrete. Continuous models are then a convenient analytical abstraction- that is how Gaussian distribution was initially perceived by de Moivre, before the central limit theorem was proved. Examples from actual engineering and science studies are plentiful and are an integral part of the exposition.

In 1992, just as we started putting our ideas together on paper and in the classroom, we found out that French mathematicians and physicists David Ruelle and Ivar Ekeland, published two volumes popularizing a position that was also ours. We couldn't have hoped for a better preparation of the public for the appearance of our textbook. Ruelle's book *Chance and Chaos* was published by Princeton University Press, and Ekeland's *Au Hasard: La Chance, La Science, Le Monde* appeared in Paris at Le Seuil. We used Ruelle's book as a mandatory additional reading assignment for students enrolled in the *Uncertainty* course, and an essay-style project related to the book was routinely assigned.

In the perennial "Bayesians vs. frequentists" debate we come squarely on the frequentist side, but only as a more effective pedagogical approach. Recent studies in the psychology of learning showed that "the mind is a frequentist device". This may be a result of the way the human brain evolved through environmental pressures. The above claim even found its way into the popular press. *The Economist*, in the 4th of July 1992 issue, argued, in a piece[2] entitled "A critique of pure rea-

[2] Reproduced at the end of this book as Appendix C.

son", that the psychologists' findings show that "merely rephrasing a problem in frequentist rather than Bayesian terms generally increases the number of people who can solve it".

In real life, modern applied statistics takes advantage of powerful software packages. However, we felt that the pedagogical benefits of using them from the start are limited because they do not give students sufficient insight into the nature of algorithms and do not let students experiment with random phenomena. The latter have to be simulated and that simulation methodology is now widespread in the engineering and science research and design communities. It is therefore a crucial topic to explain randomness from the algorithmic and computational viewpoint.

Although random phenomena have always struck peoples' imagination and affected their lives, until recently, students of randomness came to the subject with a limited experience usually acquired by playing the games of chance. However, to detect the laws of random behavior, the data sets have to be so large that one cannot easily see or grasp these regularities in everyday circumstances. The situation was thus basically different from, say, mechanics or calculus where students' intuition is formed by a lifetime of experiences of walking, throwing baseballs, swimming, and sledding. In the past, the usual path to the discovery of the laws of randomness, even in our elementary school programs, was by logical understanding and abstract computations. However, with the advent of computers and, especially, very flexible symbolic manipulation programs, such as *Mathematica* and *Maple*, it became possible to obtain a reasonably quick insight and develop sophisticated intuition about randomness by doing computer experiments. The software enables students to handle large sets of data in a relatively simple fashion. It seemed obvious to us that such an approach has to be built-in in any modern statistics course. The textbook takes advantage of this development permitting students independent exploration and self-paced instruction.

Also, during the last decade or so, the chaotic behavior in nonlinear systems emerged as an omnipresent source of randomness in real physical nonlinear dynamical systems. So, it was obvious to us that its study should be incorporated in the statistics curriculum from the very beginning.

Finally, it is worth observing that this book could have the subtitle *The Kolmogorov's Legacy*. Indeed, it is Andrei Nikolaevich Kolmogorov (1903 - 1987), a Russian mathematician, whose life's intellectual journey is being retraced on the pages of this volume. In his 1933 treatise he laid the rigorous mathematical foundation for probability theory (see Chapter 5) and statistics which blossomed afterwards into major areas of the scientific enterprise. Then, after World War II, he made major breakthroughs in the study of nonlinear dynamical systems in his work on turbulence and development of the concept of entropy, and conducted fundamental studies of the idea of randomness in terms of algorithmic (computational) complexity. This book could not have been written without Kolmogorov's seminal contributions.

Organization

Major topics included in the book are

I. Descriptive Statistics-Compressing Data. This part includes chapters on numerical and graphical representation of data, statistical functions, analytic representation of discrete data, and introduces the concepts of fractals and random fractals in association with image compression. The topic of computer generation of "random sequences" is also discussed.

II. Modeling Uncertainty. Here models arising via simple mathematical recursive relations but nevertheless exhibiting random behavior are introduced. Relationships between randomness and algorithmic complexity, so important in computer science and engineering, are studied, pseudo-random numbers and questions of validity of the Monte-Carlo methods are discussed. This material is followed by the concept of statistical independence and the classical Kolmogorovian probability theory. The part ends with an exposition on basic properties of chaotic dynamical systems and a discussion on how uncertainty appears in the real physical systems.

III. Statistical Inference-Selecting a Model. This part introduces the common, and generally accepted, methodology for designing experiments and making inferences on the basis of their outcomes. General principles of experimental design and data collection are explained, as well as the principal statistical functions (estimators) on which inference is based. The two types of statistical inference, confidence intervals and test procedures, are developed in the framework of normal models. In particular, the one- and two-sample models, regression, and the analysis of variance for one- and two-factor completely randomized designs are studied.

The textbook specifically addresses needs of engineering and science students by a selection of examples of statistical problems arising in real-life industrial and scientific lab situations. They form a constant background for our discussions as we proceed through the material in a spiral-like fashion, starting at each level with real-life examples, followed by a simulated computer exploration, and then a formulation of formal analytical principles. The examples have been collected from engineering and scientific literature and through direct interaction with practicing engineers and scientists. In particular, sets of experimental data for statistical analysis are made accessible to the students on Internet.

A series of student projects should be an essential part of the course and play a major role in students' evaluation. At CWRU we encouraged students to work in small groups of 2 to 3 people. Except for the projects for Chapters 3 and 5 which are analytic in nature and individual, all the projects are *Mathematica* intensive and students should be required to turn in the code, explanations, analysis, and plentiful graphics. Figures 0.0.1 and 0.0.2 present some of the graphics obtained by the students in projects that involved Gaussian approximation in the central limit theorem (Fig. 0.0.1a), analysis of the algorithmic complexity of binary representations (Fig. 0.0.1b), or simulation of the invariant density for the logistic chaotic mapping of the unit interval (Fig. 0.0.2a). Sometimes, students explorations resulted in

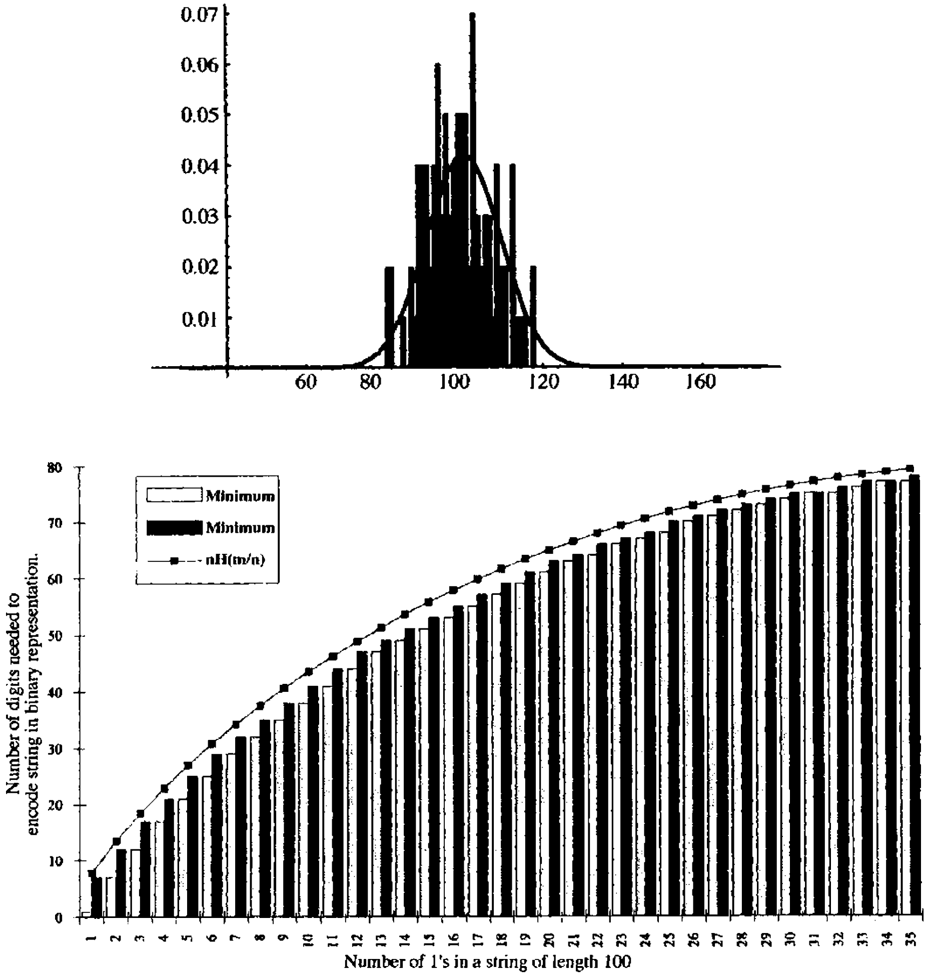


FIGURE 0.0.1

interesting insights: Bill Dickinson produced an unorthodox 3-dimensional orbit diagram for the logistic dynamical system which, in addition to the usual representations of bifurcations, also displayed relative frequencies of visits to different states (Fig. 0.0.2b). The *Mathematica* projects usually engendered a lot of enthusiasm and independent work by the students. The length of many of the reports could be a mixed blessing to the instructor though as they easily run into 40 to 50 pages each; more recently we gave students page limitations.

The last, nontechnical individual project was an essay on the theme of the above mentioned book by David Ruelle from the vantage point of the material learned

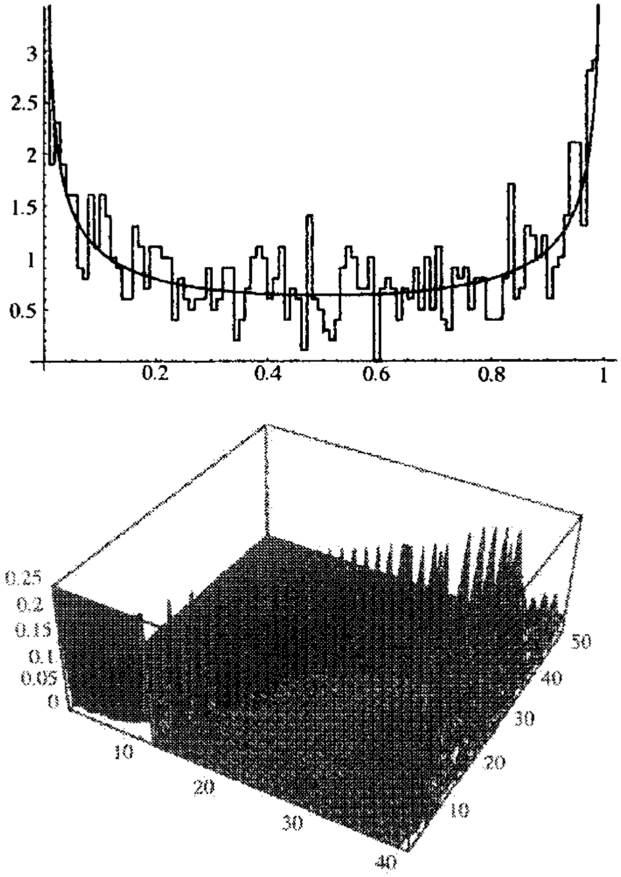


FIGURE 0.0.2

in the course. The students were asked to select a chapter from Ruelle's book that they found most stimulating or provocative (whether they agreed or disagreed with it) and provide their own commentary to it. The project emphasized good writing skills and the students often displayed an amazing maturity and sophistication. They wrote with flair on self-selected topics that ranged from predictably technical, such as "The Bell Inequality in Quantum mechanics", to "Life, Intelligence, Uncertainty", "Determinism and the Orthodox Judaism", "Determinism, Free Will and Choice", "True Meaning of Sex" and a Platonian dialogue on the question of randomness.

Realistically, given the amount of the material, it is impossible to go through the whole book in great detail in a one-term course. We have successfully taught three somewhat different courses using early drafts of the textbook:

1. A more elementary version based on Parts I and III, emphasizing concrete algorithmic skills.

2. A more advanced version that would cover Parts I and II and include more theoretical material on algorithmic complexity, statistical independence modeling, and chaotic behavior in dynamical systems.

3. A selection of sections from all three parts, with some other sections assigned as independent reading.

The book is complemented by data sets and interactive `UVW` *Mathematica* packages written by Bernard Ycart of the Grenoble University; the latter are described in detail in Appendix E. These electronic materials are extensively used throughout this book in the *Mathematica* experiments, and can be downloaded by the reader from the `UVW Web Site` which can be accessed at the Internet address

`http://www.birkhauser.com/book/isbn/0-8176-4031-2` .

The `UVW Web Site` is an integral part of our book. In the future, we plan to develop it into a fully interactive, electronic version of this text.

Notation and Abbreviations

a	$=$	$b \pmod n$ means that a and b have the same remainder when divided by n
a	$\gg$	b means that a is much greater than b
a_t	$\sim$	b_t $(t \to s)$ means that $a_t/b_t \to 1$ as $t \to s$
a	$\approx$	b means that a is approximately equal to b
$\#A$	—	the number of elements (cardinality) of set A
$\mathcal{A}^*$	—	the set of all finite strings written in alphabet $\mathcal{A}$
$\mathbf{1}_A(x)$	$=$	1, if $x \in A$, and $= 0$ if $x \notin A$, the indicator function of the interval (or, in general, set) A
$\mathbf{C}$	—	the set of all complex numbers
$\operatorname{cov}(\boldsymbol{x}, \boldsymbol{y})$	—	covariance of paired samples $\boldsymbol{x}$ and $\boldsymbol{y}$
$\operatorname{Cov}(X, Y)$	—	covariance of random quantities X and Y
CLT	—	Central Limit Theorem
$E(X)$	—	expected value of random quantity X
$f^{-1}A$	—	inverse image of set A under function (map) f, i.e., the set of points x that are mapped by f into A
$f_X(x)$	—	probability density function (p.d.f.) of random quantity X
$F_X(x)$	—	cumulative distribution function (c.d.f.) of random quantity X
$H(x)$	$=$	$\mathbf{1}_{[0,\infty)}(x)$, the Heaviside unit step function
LLN	—	Law of Large Numbers
$\operatorname{med}(\boldsymbol{x})$	—	median of sample $\boldsymbol{x}$
$m_k(\boldsymbol{x})$	—	the k-th moment of sample $\boldsymbol{x} = (x_1, \ldots, x_n)$
M^T	—	transpose of matrix M
$\mathbf{N}$	—	the set of all natural numbers (positive integers)
$\Pr\{A\}$	—	probability of event A
$P(.)$	—	countably additive probability measure in axiomatic probability theory

$\mathbf{Q}$	—	the set of all rational numbers
$q(\alpha, \boldsymbol{x})$	—	α-quantile of sample $\boldsymbol{x} = (x_1, \ldots, x_n)$
$Q(\alpha, X)$	—	α-quantile of random quantity X
$\mathbf{R}$	—	the set of all real numbers
$\mathbf{R}^d$	—	the set of d-dimensional vectors
rng $(\boldsymbol{x})$	—	range of sample $\boldsymbol{x}$
SFL	—	Stability of Fluctuations Law
$s^2(\boldsymbol{x})$	=	unbiased variance of sample $\boldsymbol{x} = (x_1, \ldots, x_n)$
std $(\boldsymbol{x})$	—	$\sqrt{\text{var}\,(\boldsymbol{x})}$, standard deviation of sample $\boldsymbol{x} = (x_1, \ldots, x_n)$
`UVW`	—	Uncertain Virtual Worlds Packages/Web Site `http://www.birkhauser.com/book/isbn/0-8176-4031-2`
$\lvert\boldsymbol{x}\rvert$	=	$\sqrt{x_1^2 + x_2^2 + \ldots + x_d^2}$
$x_{(i)}$	—	i-th order statistic of sample $\boldsymbol{x} = (x_1, \ldots, x_n)$
X, Y	—	random quantities, random variables
var $(\boldsymbol{x})$	—	variance of sample $\boldsymbol{x} = (x_1, \ldots, x_n)$
Var (X)	=	$\sigma^2(X)$, variance of quantity X
$\bar{\boldsymbol{x}}$	—	sample mean of sample $\boldsymbol{x} = (x_1, \ldots, x_n)$
$\mathbf{Z}$	—	the set of all integers
$\lceil\alpha\rceil$	—	least integer greater than or equal to α
$\lfloor\alpha\rfloor$	—	greatest integer less than or equal to α
$\Gamma(s)$	=	$\int_0^\infty e^{-t} t^{s-1} dt$, the Gamma function
$\mu_k(X)$	=	$E(X^k)$, the k-th moment of random quantity X
ϕ, ψ	—	test functions
$\Phi(x)$	—	c.d.f. of the standard Gaussian (normal) random quantity
$:=$	—	defining equality
$\mapsto$	—	maps to
$\overset{\circ}{\longmapsto}$	—	partial map, perhaps defined only for some arguments
$\to$	—	converges to

Part I

DESCRIPTIVE STATISTICS–COMPRESSING DATA

Chapter 1

Why One Needs to Analyze Data

In this chapter you will find a collection of examples of phenomena where the randomness plays an essential role. Browse through them at your leisure, experiment with the data provided, and use this opportunity to ease your way into *Mathematica*. The idea is to get a general feel for the issues to be discussed later in the book in greater detail.

1.1 Coin tossing, lottery, and the stock market

Coin tossing is a proverbial and generic example one associates with randomness. One tosses a (fair) coin repeatedly and the observed outcome is either heads or tails. For the sake of simplicity we will code heads as "1" and tails as "0", so the outcome of an experiment consisting of multiple tosses of a coin can be encoded as a *string*[1] of zeros and ones, a long word in the alphabet consisting of two "letters" 0 and 1. Here are a few examples of such strings:

(a) 1111111111111111111111111
(b) 10101010101010101010101010
(c) 10010011100100111001001110010011
(d) 011011100101110111100010011010
(e) 10111001011111001000000110101001

Their length n are, respectively, 25, 26, 32, 30, and 32.

Intuitively, not all of them seem equally random. If the coin is "fair", one could test its "fairness" (or the lack of preference for either side— arguably, an attribute of randomness) by comparing *relative frequencies* of appearance of heads and tails. Recall that for a string

$$x_1, x_2, x_3, \ldots, x_n \tag{1}$$

[1] Note that in *Mathematica* "strings" are a special `InputForm` not to be confused with algebraic expressions and lists. All of them, though, are strings in the sense of this chapter.

M. Denker, W. Woyczynski, *Introductory Statistics and Random Phenomena*,
Modern Birkhäuser Classics, DOI 10.1007/978-3-319-66152-0_1

consisting of 0s and 1s, the relative frequency of either of these letters is defined as

$$f_0 = \frac{\#\{i : x_i = 0\}}{n}, \tag{2}$$

$$f_1 = \frac{\#\{i : x_i = 1\}}{n}. \tag{3}$$

The notation $\#A$ means the number of elements (cardinality) of set A, so that the numerator $\#\{i : x_i = 0\}$ reads: the number of indices i for which the string element x_i is equal to 0. In all of the above strings, except the first one, both frequencies are equal

$$f_0 = f_1 = 1/2. \tag{4}$$

So, the violation of equality (4) seems to be sufficient grounds to eliminate the first string as random: we would like to believe that the "fortune" is blind. (On the other hand, coin tossing is a dynamical system subject to the Newtonian laws of mechanics and, given sufficient initial data about each toss, we should be able to precisely determine the outcome of each toss. Or should we?[2])

However, the fact that strings (b-e) satisfy the "*equipartition*" rule (4) does not make them look totally random to us. As a matter of fact, we can detect a vague increasing degree of randomness in these strings as we move from the first to the last. How can we express this intuition more formally and provide a framework for a quantitative analysis of this *uncertainty*?

The first, and not unreasonable, hunch could be that the *randomness* (or the *uncertainty*, but let us stick to the first term for the time being) in each of these strings has something to do with the *complexity* of each sequence or, more exactly, with our inability to encode the strings perceived as random in simple terms or, to rephrase it one more time, to provide short descriptions for them. However, such an approach—define a string as random if it has no short description—requires some caution because of what is known as the *Richard -Berry Paradox*.[3] The description

THE SMALLEST NUMBER THAT CAN NOT BE DEFINED
WITH LESS THAN ONE THOUSAND LETTERS

has itself less than one thousand letters!!!

So, let us return for a moment to the *equipartition* idea and see if we can exploit it in a more sophisticated fashion. Observe that although the strings (b-e) have similar relative frequency of zeros and ones, the situation changes dramatically if we start inspecting the relative frequency of consecutive blocks of letters of length more than one. In the case of blocks of length 2, we get

[2] This and related issues of the chaotic behavior in dynamical systems will be discussed in detail in Chapter 6.

[3] This paradox and the related algorithmic complexity issues will be discussed at length in Chapter 4.

(a) 11 11 11 11 11 11 11 11 11 11 11 11 11 11 11 11 11 11 11 11 11 11 11 11
(25 − 1 = 24 blocks)

(b) 10 01 10 01 10 01 10 01 10 01 10 01 10 10 01 10 01 10 01 10 01 10 01 10
(25 blocks)

(c) 10 00 01 10 00 01 11 11 10 00 01 10 00 01 11 11 10 00 01 10 00 01 11 11
10 00 01 10 00 01 11 (31 blocks)

(d) 01 11 10 01 11 11 10 00 01 10 01 11 11 10 01 11 11 11 10 00 00 01 10 00
01 11 10 01 10 (29 blocks)

(e) 10 01 11 11 10 00 01 10 01 11 11 11 11 10 00 01 10 00 00 00 00 00 01 11
10 01 10 01 10 00 01 (31 blocks)

Now, both (a) and (b) fail this *equipartition test of order 2* as, out of four possible blocks 00, 01, 10, 11 of length 2, block 11 is favored in the first string and blocks 10 and 01 in the second. Think about them as letters of an alphabet consisting of four letters: 00, 01, 10, 11. Neither of them should be favored if a string is to be called random.

This suggests the following hierarchy of tests of randomness for the binary (zero or one) strings $x_1, x_2, \dots, x_n$:

1. *Test of order 1.* Check the relative frequencies of 0s and 1s. If they are equal (or, in practical situations, close) to 1/2, then we can say that the string passes the 1st order test of randomness.

String (a) fails this test but strings (b-e) pass it.

2. *Test of order 2.* Check the relative frequencies of blocks 00, 01, 10, 11, computed as follows:

$$f_{00} = \frac{\#\{i : (x_i, x_{i+1}) = (0,0)\}}{n-1}, \qquad f_{01} = \frac{\#\{i : (x_i, x_{i+1}) = (0,1)\}}{n-1}, \tag{5}$$

$$f_{10} = \frac{\#\{i : (x_i, x_{i+1}) = (1,0)\}}{n-1}, \qquad f_{11} = \frac{\#\{i : (x_i, x_{i+1}) = (1,1)\}}{n-1}. \tag{6}$$

If they are all close to 1/4, then we can say that the string passes the 2nd order test of randomness.

As observed above, strings (a) and (b) fail the test of order 2. Strings (c) and (e), practically, pass it as for both, $f_{00} = f_{01} = f_{10} = 8/31$, $f_{11} = 7/31$. The case of string (d) is more debatable because then $f_{00} = 4/29$, $f_{01} = f_{10} = 8/29$, $f_{11} = 9/29$, but more on this string below. Note that for shorter strings it may be impossible to achieve a perfect frequency balance between different blocks; our concepts are more appropriate for very long strings.

At this stage it is clear how to proceed further. The 3rd order test would check that the frequencies of blocks of length three are about 1/8, and the k-th order test would check that the frequencies of blocks of length k stay close to $1/2^k$, and so on.

Definition 1.1.1 Equipartition Property.
A long string is said to enjoy the equipartition property if it passes the tests of randomness of all orders $k = 1, 2, 3, \ldots$

In our examples, string (c) fails the test of randomness of order 8 because it is periodic. This becomes obvious if we write it in the form

(c) 10010011 10010011 10010011 10010011.

There are $2^8 = 256$ different possible blocks of length 8, and each of them should have the same frequency 1/256. String (c) contains $32 - 7 = 25$ blocks of length 8. The block 10010011 appears with the relative frequency 4/25, the seven blocks 00100111, 01001110, 10011100, 00111001, 01110010, 11100100, 11001001 with the frequency 3/25, and other blocks have frequency 0.

So, that leaves open the question: How random are strings (d) and (e)? Well, on a closer inspection one discovers that string (d) can be rewritten in the form

(d) 0 1 10 11 100 101 110 111 1000 1001 1010 ...

which we immediately recognize as the binary representation of the decimal string representing the so-called *Champernowne number*

(d') 0. 1 2 3 4 5 6 7 8 9 10 11 12 13 14 ...

—hardly a sequence anybody would call random. However, it turns out that (d), as an infinite sequence, passes the tests of randomness of arbitrary order although the proof of its equipartition property is not easy. So, what is really going on? Despite the equipartition property, the string is perfectly *predictable*, violating another obviously desirable attribute of randomness, the *unpredictability*. Indeed, if the previous term of string (d') is known, one can produce a very deterministic formula to produce the next term:

$$x_{n+1} = g(x_n) \qquad \text{where} \qquad g(x) = x + 1.$$

Clearly, equipartition property cannot be equated with randomness although it should be implied by the latter. And what about the sequence (e)? It was produced by a "random" number generator on a computer. We will return to computer generation of "random" numbers later.

The above informal discussion gives a taste of foundational problems with which we are faced when we try to formalize the notion of uncertainty or randomness. They will be addressed in greater depth and detail in the remainder of this book. Of course, the questions of randomness routinely arise in engineering, science,

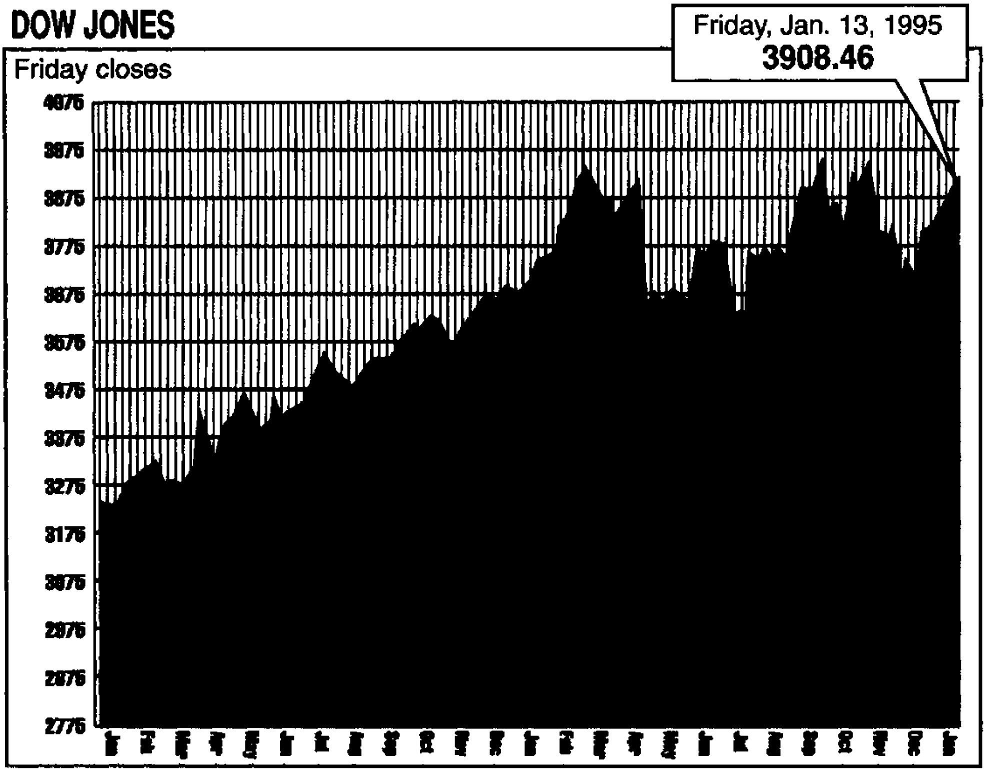


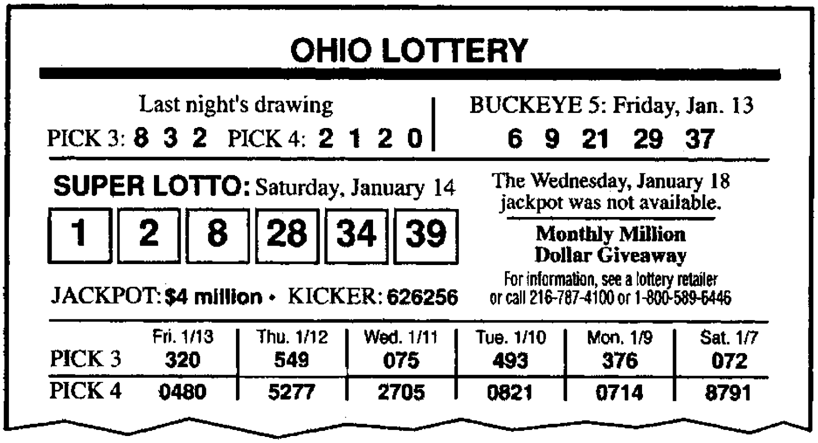
FIGURE 1.1.1
The Plain Dealer of January 15, 1995.

economics, and daily life. We open our local paper to inspect if our latest picks in the stock market are panning out (Fig. 1.1.1), or to check if we won in last night's lottery drawing (Fig. 1.1.2) — both, perplexingly uncertain events.

In the next few sections we will go through a number of real-life examples where uncertainty is an important aspect of the phenomenon.

Computer experiments and projects are an important ingredient of this book. We conduct them in the *Mathematica* symbolic manipulation software environment, but any other similar language, such as *Maple*, would do. To facilitate your introduction to *Mathematica*, we will go through a series of beginner-friendly tutorials which, however, should not replace your independent and systematic familiarization with *Mathematica* using any of the excellent books listed in the Bibliographical Notes at the end of this chapter. The larger data with which you are asked to experiment are supplied on the Internet `UVW Web Site`; no need to keyboard them manually.

Mathematica Experiment 1. Zeros and Ones. Mathematica makes it possible



OHIO LOTTERY

Last night's drawing
PICK 3: **8 3 2** PICK 4: **2 1 2 0**

BUCKEYE 5: Friday, Jan. 13
6 9 21 29 37

SUPER LOTTO: Saturday, January 14

1 2 8 28 34 39

JACKPOT: **$4 million** • KICKER: **626256**

The Wednesday, January 18
jackpot was not available.

Monthly Million
Dollar Giveaway
For information, see a lottery retailer
or call 216-787-4100 or 1-800-589-6446

	Fri. 1/13	Thu. 1/12	Wed. 1/11	Tue. 1/10	Mon. 1/9	Sat. 1/7
PICK 3	320	549	075	493	376	072
PICK 4	0480	5277	2705	0821	0714	8791

FIGURE 1.1.2
The Plain Dealer of January 15, 1995.

to manipulate data written in the format `list={a,b,...,}` where `a,b,...,z` are arbitrary numbers. Here `list` is just a name given to the string `a,b,...,z`. The following command lines for *Mathematica* help produce various frequencies for a specific list:

`Length[list]`
Usage: returns the length of the `list`.

`Sum[list[[i]],{i,1,Length[list]}]/Length[list]`
Usage: returns the relative frequency of 1s in the `list`.

`lil1[i_]:=list[[i]]*list[[i+1]] ; Sum[lil1[i], {i,1,Length[list]-1}]/(Length[list]-1)`
Usage: returns the relative frequency of two consecutive 1s.

`li10[i_]:=list[[i]]*(1-list[[i+1]]) ; Sum[ li10[i],{i,1,Length[list]-1}]/(Length[list]-1)`
Usage: returns the relative frequency of blocks 10.

`N[expression]`
Usage: gives the numerical value of `expression`.

Now, the relative frequencies in string (e) can be computed as follows:

```
In[1]:= bernE={1,0,1,1,1,0,0,1,0,1,1,1,1,1,0,0,1,0,0,0,
                          0,0,0,1,1,0,1,0,1,0,0,1}
Out[1]= {1, 0, 1, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1,
          0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0,1 }
In[2]:= Length[bernE]
Out[2]= 32
In[3]:= Sum[bernE[[i]],{i,1,31}]/32
```

```
Out[3]= {16/32}
In[4]:= N[%]
Out[4]= 0.5
In[5]:= N[Sum[bernE[[i]],{i,1,Length[bernE]}]/Length[bernE]]
Out[5]= 0.5
In[6]:= bernE01[i_]:=(1-bernE[[i]])*bernE[[i+1]]; N[Sum[
           bernE01[i], {i,1,Length[bernE]-1}]/(Length[bernE]-1)]
Out[6]= 0.258064
```

So, the frequency of 1s in string (e) is 0.5 and the frequency of blocks 01 is 0.258064. Further examples of this type, using much longer strings supplied on the `UVW Web Site`, are provided in Section 1.14: Experiments, Exercises, and Projects.

1.2 Inventory problems in management

A hardware distribution company has to prepare its inventory (say, for tax purposes, annual report, bankruptcy proceedings, etc.). In the process, the number of items in each category (nails, rat traps, snow shovels, memory chips, etc.) has to be determined. Who is doing the checking and how the checking is to be done is often a point of contention (see, e.g., Huff's *How to Lie with Statistics* quoted in the Bibliographical Notes), and one would like to have fair and sound auditing procedures written into the law of the land.

Obviously, counting all of the items one by one would be too expensive and time consuming. A more reasonable alternative, assuming that the items are stored in bins of the same size, would be to take a *small sample* of bins and on that basis determine the number of each item in the whole *population*. If this is the procedure on which we settle, then we immediately face a number of practical questions:

How to take the sample? Clearly the sample has to be representative in the sense that every source of bias has to be removed.

How to judge the degree to which the sample is representative?

- How large or small should it be?
- What procedure should be used to determine the total population from the sample count?

How to organize the storage system to optimize the above sampling process, and permit an optimization of the statistical techniques used?

The statistical techniques depend on the assumed underlying empirical (or theoretical) relative frequency distributions which can be different for different companies and product types. How can they be determined?

Similar issues arrise in many other areas where testing the whole population is impossible, e.g., in public opinion polls.

1.3 Battery life and quality control in manufacturing

The following hypothetical (the data are simulated) example represents the typical situation.

Example 1.3.1 Batteries Are Not Forever.
Lifetimes of 50 batteries have been tested at the manufacturing company. The test lasted 10 hours and by the end of the test period 41 batteries failed, with their lifetimes (in hours) being

0.33, 5.71, 2.23, 3.41, 1.83, 3.01, 0.71, 3.95, 4.37, 0.90, 0.30, 1.94,
8.31, 5.15, 3.25, 0.06, 2.89, 6.99, 2.15, 6.58, 5.28, 0.78, 1.70, 6.68,
4.73, 5.94, 4.26, 7.23; 8.31, 2.52, 1.35, 2.66, 1.30, 0.71, 2.41, 3.66,
9.69, 0.43, 4.41, 8.77, 9.66

The remaining 9 batteries were still going strong at the end of the 10-hour testing period. The company has to make a decision about the advertised and guaranteed battery lifetime on the basis of the above, *censored*, data. It plans to replace batteries that do not meet advertised specifications. A number of natural questions arise:

- Given the unit manufacturing cost, what should be the warranty period for the company to break even?
- Given that the warranty duration is given (say, forced by competition), what should be the price of the item for the company to break even?
- Which parameters computed from the above data are essential for answering the above questions? Equivalently, how should the above data be compressed to preserve what is really important?

Example 1.3.2 Bombs Away.
A Cleveland company manufactures bases for fragmentation bombs (no peace dividens here yet). Measurements (in inches) of 75 bomb bases' heights are provided

below, with more complete data on 145 bases (provided by Ravi Jayaraman, a graduate student in Systems Engineering) to be found on the `UVW Web Site` in the BOMBS file:

0.831 0.829 0.836 0.840 0.826 0.834 0.826 0.831 0.831 0.831
0.836 0.826 0.831 0.822 0.816 0.833 0.831 0.835 0.831 0.833
0.830 0.831 0.831 0.833 0.820 0.829 0.828 0.828 0.832 0.841
0.835 0.833 0.829 0.830 0.841 0.818 0.838 0.835 0.834 0.830
0.841 0.831 0.831 0.833 0.832 0.832 0.828 0.836 0.832 0.825
0.831 0.838 0.844 0.827 0.826 0.831 0.826 0.828 0.832 0.827
0.838 0.822 0.835 0.830 0.830 0.815 0.832 0.831 0.831 0.838
0.831 0.833 0.831 0.834 0.832

The randomness found in the above data is encountered in most manufacturing processes due to the variability of materials, machinery, conditions, and human factors. The manufacturer's goal is to reduce this randomness but those efforts have to be balanced against increased costs. For mass produced items, the usual quality control procedure is to measure the variability in a given batch and reject the whole batch if that variability is too large. The ways to assess this variability will be discussed later.

Mathematica Experiment 1. Batteries Are Not Forever. This is a good opportunity to introduce additional *Mathematica* commands:

`Position[list, number]`

Usage: shows the position of the number in the `list`.

`Delete[list,{{a},{b},...,{z}}]`

Usage: deletes the elements with numbers `a,b, ...,z` from the `list`.

`Sort[list]`

Usage: sorts the elements of the `list` in the increasing order.

`Floor[number]`

Usage: returns the largest integer less than or equal to `number`.

The file BATTERY on the `UVW Web Site` contains the lifetimes of batteries from Example 1.3.1, including those that survived the 10 hour test. The first data manipulating step in our experiment is to remove all the 10s from the data set. Then we compare the averages of the new list created by this deletion with the original list in the BATTERY file. Finally, we want to find the number such that 20% of the observed lifetimes in the original data set fall below that number.

```
In[1]:=  battery={0.33,  5.71, . . . ,8.77, 9.66}
In[2]:= Position[battery,10.0]
Out[2]= {{4},{14},{17},{24},{32},{38},{43},{44},{47}}
In[3]:= Delete[battery, %]
```

```
      Remark: Here, % stands for {{4},{14},{17},{24},{32},{38},
      {43},{44},{47}}, i.e., the data from the preceding line
      which, therefore, need not be rekeyed.
Out[3]= {0.33, 5.71, 2.33, 3.41, 1.83, 3.01, 0.71, 3.95, 4.37,
        0.90, 0.30, 1.94, 8.31, 5.15, 3.25, 0.06, 2.89, 6.99,
        2.15, 6.58, 5.28, 0.78, 1.70, 6.68, 4.73, 5.94, 4.26,
        7.23, 8.31, 2.52, 1.35, 2.66, 1.30, 0.71, 2.41, 3.66,
        9.69, 0.43, 4.41, 8.77, 9.66}
In[4]:= battery1= %;
      Remark: This names the deleted list as battery1
In[5]:= Sum[battery1[[i]],{i,1,Length[battery1]}]/Length[battery1]
Out[5]= 3.82073
In[6]:= Sum[battery[[i]],{i,1,Length[battery]}]/Length[battery]
Out[6]= 4.933
In[7]:= Sort[battery]
Out[7]= {0.06, 0.3, 0.33, 0.43, 0.71, 0.71, 0.78, 0.9, 1.3, 1.35,
         1.7, 1.83, 1.94, 2.15, 2.33, 2.41, 2.52, 2.66, 2.89,
         3.01, 3.25, 3.41, 3.66, 3.95, 4.26, 4.37, 4.41, 4.73,
         5.15, 5.28, 5.71, 5.94, 6.58, 6.68, 6.99, 7.23, 8.31,
         8.31, 8.77, 9.66, 9.69, 10., 10., 10., 10., 10., 10.,
         10., 10., 10.}
In[8]:= battery2= %;
In[9]:= battery2[[Floor[Length[battery]*0.2]]]
Out[9]= 1.35
In[10]:= Quit
```

See Section 1.14 for more experiments and projects of this nature.

1.4 Reliability of complex systems

Example 1.3.1 illustrates a common situation where only statistical, and often censored, data about the lifespans of manufactured components is available. These items, incorporated into a more complex device, may then be put to work under different (sometimes extreme) conditions and the ability of the device to function properly depends on the ability of its individual components to survive, or—in other words—on their *reliability*.

Development of the reliability theory was largely spurred by the electronics industry where, typically, devices are built of hundreds or thousands of not perfectly reliable parts, and there is a need to evaluate the reliability of the whole instrument based on information about its components. For a fixed component, the *reliability* r could be measured as the proportion of components of the same type that are likely to work without failure for a given period of time.

Then, for example, we could inquire about the reliability of a large device consisting of many components of known reliability interconnected in a particular fashion. In the simplest case, the reliability of a device consisting of n components $C_1, C_2, \ldots, C_n$ in series (Fig. 1.4.1) of reliability $r_1, r_2, \ldots, r_n$, respectively, can be shown to be equal to the product $r_1 \cdot r_2 \cdot \ldots \cdot r_n$. Hence, it is clear that the reliability of a serial device can never be better than the reliability of its *worst* component.

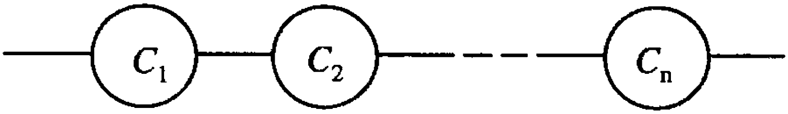


FIGURE 1.4.1
A serial device.

On the other hand, for a parallel device (Fig. 1.4.2), the reliability turns out to be

$$1 - (1 - r_1) \cdot (1 - r_2) \cdot \ldots \cdot (1 - r_n). \tag{1}$$

So, the reliability of a parallel device is never worse than the reliability of its *best* component.

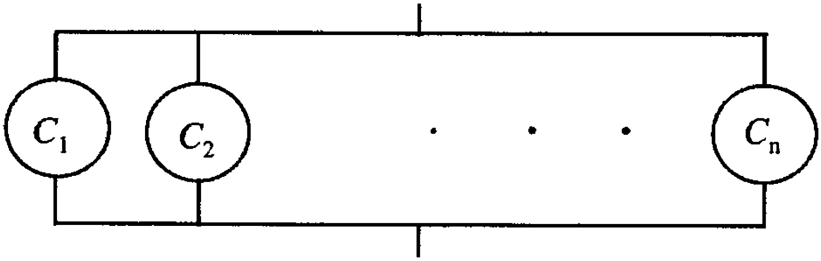


FIGURE 1.4.2
A parallel device.

The serial and parallel devices represent the simplest device structures. Determination of the reliability of more complex devices like, for instance, the one shown in Fig. 1.4.3, may be quite difficult.

Mathematica Experiment 1. Reliability. In this experiment we will compute:

(1) Reliability of a serial device consisting of 35 elements with individual reliabilities $r_m = 1/m$, $m = 1, \ldots, 35$.

(2) Reliability of a parallel device consisting of 35 elements with individual reliabilities $r_m = 1/m$, $m = 1, \ldots, 35$.

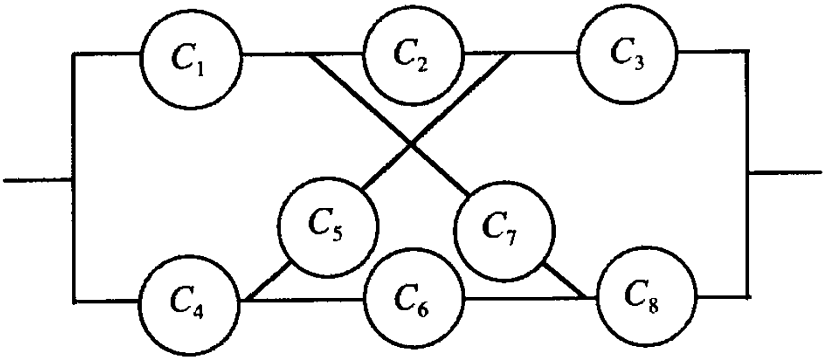


FIGURE 1.4.3
A more complex device.

It is worthwhile to compare the results of (1) and (2). Are there any surprises? The experiment will be concluded by computation of the reliability of serial and parallel devices for any number of individual components with individual reliabilities r_m given by an arbitrary list.

You will find the following new *Mathematica* commands useful:

```
Product[function[i],{i,minimum,maximum}]
```

Usage: returns the product of the numbers `function[argument]` for the values of the argument between `minimum` and `maximum`.

```
Table[expression[i],{i,minimum,maximum}]
```

Usage: makes a list of elements where the i-th element is `expression[i]`.

```
Save["name.m", definition1, definition2,...]
```

Usage: saves definitions into a file called `name.m`.

```
In[1]:= f[i_]:= 1/i
          Remark: This defines the function f(x)=1/x.
In[2]:= rel1= Product[f,{i,1,35}]
Out[2]= {1/ 10333147966386144929666651337523200000000}
In[3]:= g[i_]:= 1-1/i
In[4]:= rel2=1-Product[g,{i,1,35}]
Out[4]= 1
In[5]:= serial[l_]:= Product[l[[i]],{i,1,Length[l]}]
In[6]:= probab=Table[1/m, {m,1,35}]
Out[6]= {1, {1/2}, {1/ 3}, {1/ 4}, {1/ 5}, {1/ 6},
        {1/ 7}, {1/ 8}, {1/ 9}, {1/ 10}, {1/ 11}, {1/ 12},
        {1/ 13}, {1/ 14}, {1/ 15}, {1/ 16}, {1/ 17}, {1/ 18},
        {1/ 19}, {1/ 20}, {1/ 21}, {1/ 22}, {1/ 23}, {1/ 24},
        {1/ 25}, {1/ 26}, {1/ 27}, {1/ 28}, {1/ 29}, {1/ 30},
        {1/ 31}, {1/ 32}, {1/ 33}, {1/ 34}, {1/ 35}}
In[7]:= serial[probab]
```

```
Out[7]= {1/ 10333147966386144929666651337523200000000}
In[8]:= parallel[l_]:= 1-Product[1-l[[i]],{i,1,Length[l]}]
In[9]:= parallel[probab]
Out[9]= 1
In[10]:= Save["device.m", serial, parallel]
In[11]:= !!device
            Remark: Press enter here!
         serial[l_]:= Product[l[[i]],{i,1,Length[l]}]
         parallel[l_]:= 1-Product[1-l[[i]],{i,1,Length[l]}]
In[12]:= Quit
```

In real life, the reliability depends on time. The number $N(t)$ of units surviving at time t in a population that started with $N(0)$ units at time $t = 0$ is decreasing with time and a typical *survival curve* is pictured in Fig. 1.4.4.

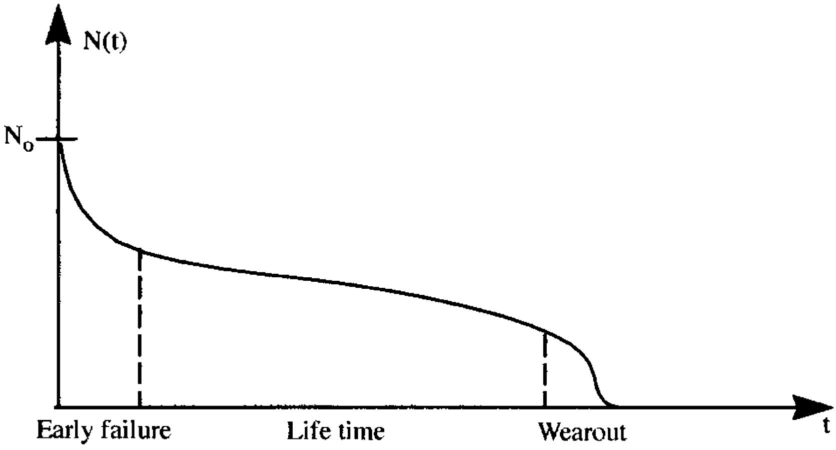


FIGURE 1.4.4
A typical survival curve.

Then, the time-dependent reliability $r(t)$ of the device can be defined as

$$r(t) = \frac{N(t)}{N(0)}, \tag{2}$$

the proportion of surviving units at time t. The function r is equal to 1 at $t = 0$ and decays monotonically to 0 as t increases, following the general shape in Fig. 1.4.4.

Often, the quantity one wants to watch is the failure rate $\lambda(t)$ of the device (per unit time and per unit device), which can be expressed via the formula

$$\lambda(t) = \frac{N(t) - N(t + \Delta t)}{\Delta t\, N(t)} \approx -\frac{d}{dt} \log r(t). \tag{3}$$

A typical graph of the failure rate time-dependence, corresponding to the survival curve of Fig. 1.4.4 is shown in Fig. 1.4.5. It reflects the typical reliability history of a device: the initial high failure rate due to the presence of "bugs", the constant failure rate during the intermediate "utility" period of the device's lifetime, and the steadily increasing failure rate as the device wears out with age.

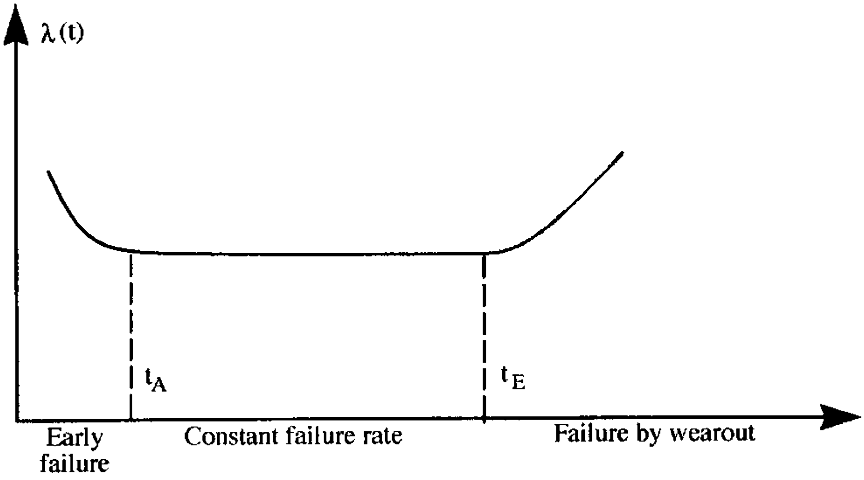


FIGURE 1.4.5
A typical failure rate curve.

Notice that if the failure rate $\lambda(t) = \lambda$ is constant in the time interval $[t_0, t_1]$, then solving the simple differential equation

$$\frac{d}{dt}\log r(t) = -\lambda, \tag{4}$$

one obtains that, in that interval, the survival curve decays exponentially and

$$N(t) = N(t_0)e^{-\lambda(t-t_0)}. \tag{5}$$

The reliability of serial and parallel devices assembled from components with time dependent reliability can then be determined via formulas discussed in the first part of this section.

1.5 Point processes in time and space

There are many situations where the experimental data or observations form a sequence of random singular point signals spread over time or space. Workstations' connection times to the server, particle arrivals registered in the Geiger counter, locations where gold deposits were found, or recorded outbursts of a mad cow disease, or arrival times of customers in a queue are typical examples here. Technically, they are called point processes in time or in space, depending on the context.

Example 1.5.1 Bright Stars.
The set of data shown on the next page was supplied by Jacqueline Monkiewicz, a CWRU astronomy major. The data includes the magnitudes and sidereal positions of the stars brighter than 2.5 magnitude in the year 2000. More complete data, for all the stars brighter than the 3rd magnitude, are provided on the `UVW Web Site`. The stars' coordinates are given by their declinations (angle from the celestial equator) and right ascensions (angular distance from the vernal equinox, measured in hours). Also included is the basic spectral class of each star.

What is the explanation for this particular, seemingly random, distribution of stars? Could it be derived from the Big Bang hypothesis? Perhaps from simpler geometric arguments?

Example 1.5.2 Water Drips.
The set of data provided below represents the time intervals (in seconds) between consecutive water drips from a nozzle.

0.1822 0.1962 0.1342 0.1035 0.1551 0.2327 0.2023
0.1289 0.1868 0.2265 0.2611 0.1917 0.1376 0.1483
0.2227 0.1605 0.1378 0.0952 0.2457 0.1738 0.2581
0.1893 0.2542 0.2246 0.2615 0.1095 0.2203 0.1014
0.1969 0.1281 0.1359 0.1005 0.2558 0.1404 0.2556
0.1352 0.2519 0.2531 0.2565 0.0720 0.2222 0.1065
0.2308 0.1430 0.1203 0.0757 0.2835 0.1340 0.2535
0.1360 0.1596 0.2041 0.2544 0.1051 0.2245 0.1085
0.2314 0.1876 0.1481 0.1376 0.2255 0.1429 0.2121
0.1243 0.1705 0.2637 0.2244 0.1357 0.2210 0.1485

What is the source of randomness in this data set? The mechanical system used in the experiment remained unchanged for the duration of the experiment. The data set was collected by Bill Dimmock, a CWRU physics major.

Table 1.5.1 Magnitudes and sidereal positions of stars

Magn.	R.A.	Declin.	Cl.	Magn.	R.A.	Declin.	Cl.
2.06	0 8 23.2	29 5 26	B	2.27	0 0 10.6	59 8 59	F
2.39	0 26 12.1	-43 40 48	K	2.04	0 43 35.3	-17 59 12	K
2.47	0 56 42.4	60 43 0	B	2.06	1 9 43.9	35 37 14	M
0.46	1 37 42.9	-57 14 12	B	2.02	2 31 50.5	89 15 51	F
2.00	2 7 10.3	23 27 45	K	2.26	2 3 53.9	42 19 47	K
1.79	3 24 19.3	49 51 41	F	2.12	3 8 10.1	40 57 21	B
1.65	5 26 17.5	28 36 27	B	0.85	4 35 55.2	16 30 33	K
2.23	5 32 0.3	-0 17 57	B	0.08	5 16 41.3	45 59 53	G
0.12	5 14 32.2	-8 12 6	B	1.64	5 25 7.8	6 20 59	B
1.70	5 36 12.7	-1 12 7	B	2.05	5 40 45.5	-1 56 34	O
2.06	5 47 45.3	-9 40 11	B	0.5	5 55 10.3	7 24 25	M
1.9	5 59 31.7	44 56 51	A	1.98	6 22 41.9	-17 57 22	B
-0.72	6 23 57.2	-52 41 44	F	1.93	6 37 42.7	16 23 57	A
-1.46	6 45 8.9	-16 42 58	A	1.5	6 58 37.5	-28 58 20	B
2.06	14 6 40.8	-36 22 12	K	-0.04	14 15 39.6	19 10 57	K
2.31	14 35 30.3	-42 9 28	B	-0.01	14 39 36.2	-60 50 7	G
1.33	14 39 36.2	-60 50 7	K	2.3	14 41 55.7	-47 23 17	B
2.08	14 56 46	-11 24 35	K	2.23	15 34 41.2	26 42 53	A
2.32	16 0 19.9	-22 37 18	B	2	15 59 30.1	25 55 13	B
0.96	16 29 24.4	-26 25 55	M	1.92	16 48 39.9	-69 1 40	K
2.29	16 50 9.7	-34 17 36	K	2.43	17 10 22.6	-15 43 29	A
1.63	17 33 36.4	-37 6 13	B	1.87	17 37 19	-42 59 52	F
2.08	17 34 56	-38 3 56	A	2.41	17 42 29.1	-39 1 48	B
2.23	17 56 36.6	51 29 20	K	1.85	18 24 10.3	-34 23 5	B
0.03	18 36 20.9	38 47 1	A	2.02	18 55 15.8	-26 17 48	B
2.06	14 6 40.8	-36 22 12	K	1.84	7 8 23.4	-26 23 35	F
2.45	7 24 5.6	-29 18 11	B	1.58	7 34 35.9	31 53 18	A
1.59	7 34 35.9	31 53 18	A	0.38	7 39 18.1	5 13 30	F
1.14	7 45 18.9	28 1 34	K	2.25	8 3 35	-40 0 11	O
1.78	8 9 31.9	-47 20 12	W	1.86	8 22 30.8	-59 30 34	K
1.96	8 44 42.2	-54 42 30	A	2.21	9 7 59.7	-43 25 57	K
1.68	9 13 12.1	-69 43 2	A	2.25	9 17 5.4	-59 16 31	A
2.5	9 22 6.8	-55 0 38	B	1.98	9 27 35.2	-8 39 31	K
1.35	10 8 22.3	11 58 2	B	2.37	11 1 50.4	56 22 56	A
1.79	11 3 43.6	61 45 3	K	2.14	11 49 14.8	16 14 34	A
2.44	11 53 49.8	53 41 41	A	1.58	12 26 35.9	-63 5 56	B
2.09	12 26 36.5	-63 5 58	B	1.63	12 31 9.9	-57 6 47	M
2.17	12 41 30.9	-48 57 34	A	1.25	12 47 43.3	-59 41 19	B
1.77	12 54 1.7	55 57 35	A	2.27	13 23 55.5	54 55 31	A
0.98	13 25 11.5	-11 9 41	B	2.3	13 39 53.2	-53 27 59	B
1.86	13 47 32.3	49 18 48	B	0.61	14 3 49.4	-60 22 22	B
0.77	19 50 46.9	8 52 6	A	1.94	20 25 38.8	-56 44 7	B
2.2	20 22 13.6	40 15 24	F	1.25	20 41 25.8	45 16 49	A
2.46	20 46 12.6	33 58 13	K	2.44	21 18 34.7	62 35 8	A
2.39	21 44 11.1	9 52 30	K	1.74	22 8 13.9	-46 57 40	B
2.1	22 42 40	-46 53 5	M	1.16	22 57 39	-29 37 20	A
2.42	23 3 46.4	28 4 58	M	2.49	23 4 45.6	15 12 19	B

Example 1.5.3 DNA Sequences.
DNA and protein data are stored world-wide in huge databases (like the U.S. Human Genome project) to provide easy access to information needed in a rapidly developing research area. In fact the information about each DNA "sequence" is encoded as a (unique) word written in a fixed alphabet. To discover similarities between different proteins much work has been done on methods of comparison of such sequences. For example, similarities were found between src proteins of the bovine cyclic AMP dependent kinase and the Roussavian and Maloney murine sarcoma virus.

Such comparisons involve uncertainty because the above mentioned encoding is one-to-one only with a certain (unknown) probability. Hence, one has to devise qualitative statistical criteria for the matching decision process. The work on this problem is still in progress and the dynamic programming methods have also been used for this purpose.

If $A_1, \ldots, A_n$ and $B_1, \ldots, B_m$, are two such sequences, then one considers the "diagonals" (A_i, B_j), where the difference $i - j$ is a fixed number, and one studies the k-word matches on these diagonals by looking at the match-indicators $x_i = 0$ or 1. If, for $1 \leq i \leq k$, $A_i = B_j$, then one sets $x_i = 1$ (and 0 otherwise). For a q such that $1 \leq q \leq k$, a q-match is said to hold if at least q of the x_is are equal to 1. For long strings of A_is and B_is one obtains a sequence of k-words consisting of 0s and 1s, which are strongly dependent. Therefore, their analysis has to use tools different than those that have been developed for studying, say, repeated independent experiments.

Mathematica Experiment 1. Bright Stars. The file `STARS` on the `UVW Web Site` contains expanded data from Example 1.3.1. To carry out our experiment, we need *Mathematica* commands to find elements, rows, and columns of a matrix, and to represent matrix data in the graphical form:

```
ListPlot[{{a,b}, {c,d}, ...}]
```
Usage: plots points in two dimensions.

```
TableForm[list]
```
Usage: gives a table of the data in the `list`.

```
Prolog->AbsolutePointSize[n]
```
Usage: plots point of size `n` times the basic unit.

```
Axes->False
```
Usage: suppresses axes in the graph.

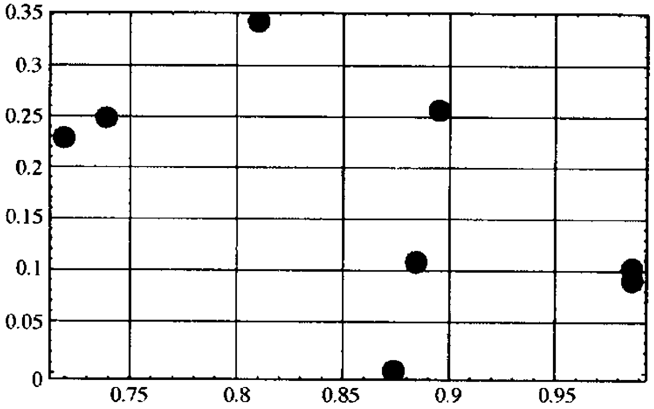
```
Cos[x], Sin[x]
```
Usage: cosine and sine functions.

```
Transpose[matrix]
```
Usage: interchanges columns and rows.

```
In[1]:= star={{2.06, 0, 8, 23.2, 29, 5, 26, B},
         {0.12, 5, 14, 32.2, -8, 12, 6, B},
         {1.5, 6, 58, 37.5, -28, 58, 20, B},
         {1.87, 17, 37, 19, -42, 59, 52, F},
         {1.16, 22, 57, 39, -29, 37, 20, A},
         {0.03, 18, 36, 20.9, 38, 47, 1, A},
         {0.5, 5, 55, 10.3, 7, 24, 25, M},
         {2.32, 16, 0, 19.9, -22, 37, 18, B}};
In[2]:= TableForm[%]
Out[2]//TableForm=

2.06   0    8    23.2   29    5    26   B
0.12   5    14   32.2   -8    12   6    B
1.5    6    58   37.5   -28   58   20   B
1.87   17   37   19     -42   59   52   F
1.16   22   57   39     -29   37   20   A
0.03   18   36   20.9   38    47   1    A
0.5    5    55   10.3   7     24   25   M
2.32   16   0    19.9   -22   37   18   B

In[3]:= star[[5]]
Out[3]= {1.16, 22, 57, 39, -29, 37, 20, A}
In[4]:= star[[5,4]]
Out[4]:= 39
In[5]:= Transpose[star]
Out[5]= {{2.06, 0.12, 1.5, 1.87, 1.16, 0.03, 0.5, 2.32},
         {0, 5, 6, 17, 22, 18, 5, 16},
         {8, 14, 58, 37, 57, 36, 55, 0},
         {23.2, 32.2, 37.5, 19, 39, 20.9, 10.3, 19.9},
         {29, -8, -28, -42, -29, 38, 7, -22},
         {5, 12, 58, 59, 37, 47, 24, 37},
         {26, 6, 20, 52, 20, 1, 25, 18},
         {B, B, B, F, A, A, M, B}}
In[6]:= %[[8]]
Out[6]= {B, B, B, F, A, A, M, B}
In[7]:= ListPlot[ Table[
        {Cos[(star[[i,2]]+star[[i,3]]/60+star[[i,4]]/3600) Degree]
         * Cos[(star[[i,5]]+star[[i,6]]/60+star[[i,7]]/3600) Degree],
           Sin[(star[[i,2]]+star[[i,3]]/60+star[[i,4]]/3600) Degree]
         * Cos[(star[[i,5]]+star[[i,6]]/60+star[[i,7]]/3600) Degree]},
           {i,1,Length[star]}],
           Axes->False, Prolog->AbsolutePointSize[8],
           Frame->True, GridLines->Automatic]
Out[7]= Graphics
```



Mathematica Experiment 2. Queue Arrivals. This experiment uses the Uncertain Virtual Worlds package UVW'TimeRep' which simulates random arrivals to a single queue (so-called Poisson process and M/M/1 queue). The precise significance of the quantitative information obtained here will become clear in later chapters.

```
In[1]:= <<UVW'TimeRep'
In[2]:= <<Statistics'ContinuousDistributions'
In[3]:= arr=Table[Random[ExponentialDistribution[1.]],{100}];
In[4]:= Geiger[arr]
In[5]:= CumulatedTimes[arr]
In[6]:= ser=Table[Random[ExponentialDistribution[1.2]],{100}];
In[7]:=Queue[arr,ser]
```

1.6 Polls-social sciences

Complex social phenomena often produce random and uncertain results when subject to empirical study. You have seen political pollsters presenting opinions or electoral predictions based on random sampling.

The table on pages 24 and 25 presents the rate (per 100,000 resident population) of sentenced prisoners in state and federal institutions on Dec. 31, of the years 1971 to 1991 (by region and jurisdiction). Is there any regularity in the table? The data depend both on time and geographical location. Are the rates for different states correlated over the years? And how? The data was provided by Ramona Myers, a CWRU chemical engineering graduate student, who also does volunteer work with the prisoners.

Mathematica Experiment 1. Vive le Quebec? The 1941 Canada Census compared family sizes depending on the age at marriage (0=15–19 years; 1=20–24 years), years of schooling (U=0–6; E=7+), income class (L=low income, H=high income), and language (F=French-speaking, M=mixed), among others. The list CENSUS on the UVW Web Site contains a sample of data for eight different groups. The last two variables are, respectively, the average number of children in that group and the number of families examined.

To carry out this experiment it will be necessary to sort data, select sublists, merge lists, etc. Finally, we will extract a sublist consisting of those groups that are French-speaking. Again, the idea is to ease your way into manipulating data using *Mathematica*, and the following commands will be useful:

Sort[list]

Usage: sorts a list in lexicographical ordering.

Drop[list,{m,n}]

Usage: drops the elements list[[m]],...,list[[n]].

Intersection[list1,list2,...,listk]

Usage: forms the list consisting of the elements which are included in all lists list1,...,listk

If[statement for numerical data, out1 if true, out2 if false, out3 if neither true nor false]

Usage: returns one of the three out expressions (1, 2, or 3 here but, in general, not necessarily numerical), according whether the statement is true, false, or neither.

Do[expression[i], {i,1,n}]

Usage: evaluates expression[i] for each i beginning with i=1 and ending with i= n.

```
In[1]:= census= {{0, U, L, F, 7.4, 5}, {0, E, L, M, 11.3, 7},
          {1, E, H, M, 8.8, 12}, {1, U, L, F, 8.3,10},
          {1, E, H, F, 10.3, 28}, {0, E, H, F, 8.7, 15},
          {1, E, L, F, 6.7, 37}, {1, U, L, M, 9.7, 3}};
In[2]:= TableForm[%]
Out[2]// TableForm= 0   U   L   F   7.4    5
                    0   E   L   M   11.3   7
                    1   E   H   M   8.8    12
                    1   U   L   F   8.3    10
                    1   E   H   F   10.3   28
                    0   E   H   F   8.7    15
                    1   E   L   F   6.7    37
                    1   U   L   M   9.7    3
In[3]:= Sort[census]
Out[3]= {{0, E, H, F, 8.7, 15}, {0, E, L, M, 11.3, 7},
         {0, U, L, F, 7.4, 5}, {1, E, H, F, 10.3, 28},
        {1, E, H, M, 8.8, 12}, {1, E, L, F, 6.7, 37},
        {1, U, L, F, 8.3, 10}, {1, U, L, M, 9.7, 3}}
In[4]:= TableForm[%]
Out[4]// TableForm= 0   E   H   F   8.7    15
                   0   E   L   M   11.3   7
                   0   U   L   F   7.4    5
                   1   E   H   F   10.3   28
                   1   E   H   M   8.8    12
                   1   E   L   F   6.7    37
                   1   U   L   F   8.3    10
                   1   U   L   M   9.7    3




In[5]:= Sum[census[[i,6]],{i,1,Length[census]}]
Out[5]= 117
          Remark: number of families in the study
In[6]:= TableForm[Drop[census, {2,4}]]
Out[6]// TableForm= 0   U   L   F   7.4    5
                   1   E   H   F   10.3   28
                   0   E   H   F   8.7    15
                   1   E   L   F   6.7    37
                   1   U   L   M   9.7    3
In[7]:= f[M]=0; d[i_]:=If[f[census[[i,4]]]==0,
        Drop[census,{i,i}],census,census]
        Remark: defines a function d which returns a dropped or
        original list. f[M]=0 sets M to a numerical value for the
        test procedure.
In[8]:= d[2]
Out[8]= {{0, U, L, F, 7.4, 5}, {1, E, H, M, 8.8, 12},
        {1, U, L, F, 8.3, 10}, {1, E, H, F, 10.3, 28},
        {0, E, H, F, 8.7, 15},  {1, E, L, F, 6.7, 37},
        {1, U, L, M, 9.7,3}}
In[9]:= Intersection[ d[1], d[2], d[3]]
Out[9]= {{0, E, H, F, 8.7, 15}, {0, U, L, F, 7.4, 5},
        {1, E, H, F, 10.3, 28}, {1, E, L, F, 6.7, 37},
        {1, U, L, F, 8.3, 10}, {1, U, L, M, 9.7, 3}}
In[10]:=french[1]=d[1]; Do[
        french[i+1]= Intersection[french[i], d[i+1]],
            {i,1,Length[census]-1}]
In[11]:=TableForm[french[Length[census]]]
Out[11]// TableForm= 0   E   H   F   8.7    15
                    0   U   L   F   7.4    5
                    1   E   H   F   10.3   28
                    1   E   L   F   6.7    37
                    1   U   L   F   8.3    10
```

Table 1.6.1 The rate (per 100,000 residents) of sentenced prisoners

Region	'71	'72	'73	'74	'75	'76	'77	'78	'79	'80
NORTHEAST	56	57	60	63	70	73	77	82	84	87
CT	63	59	54	48	59	62	53	70	69	68
ME	45	46	44	50	60	57	61	53	58	61
MA	38	32	34	38	42	46	48	49	50	56
NH	28	31	35	27	31	30	26	32	35	35
NJ	73	72	74	72	77	78	78	74	76	76
NY	65	64	71	79	89	98	108	114	120	123
PA	45	53	55	57	60	56	56	65	67	68
RI	41	36	43	49	41	53	56	56	63	65
VT	47	30	40	52	51	64	57	76	62	67
MIDWEST	73	66	63	69	84	95	108	104	105	109
IL	52	50	50	56	73	87	95	96	95	94
IN	83	73	63	58	73	79	80	82	98	114
IA	54	46	49	52	63	66	70	70	72	86
KS	91	74	61	64	76	91	97	98	95	106
MI	107	94	87	95	119	137	151	162	163	163
MN	40	35	36	35	42	41	44	49	51	49
MO	77	75	79	88	92	105	111	116	113	112
NE	69	63	66	68	80	93	83	80	71	89
ND	21	29	25	21	27	26	30	21	19	28
OH	85	77	72	87	107	117	120	122	125	125
SD	58	51	35	37	49	70	76	74	77	88
WI	55	45	47	56	65	71	72	73	73	85
SOUTH	124	125	128	135	150	161	169	181	198	188
AL	110	104	105	110	121	83	94	144	141	149
AR	84	80	82	100	102	115	111	115	132	128
DE	33	49	57	76	100	118	120	173	181	183
DC	349	341	324	289	326	334	330	383	433	426
FL	136	139	133	138	183	211	221	239	220	208
GA	146	174	173	191	204	225	224	216	224	219
KY	94	90	89	92	100	107	106	97	105	99
LA	113	92	108	128	126	120	152	184	190	211
MD	125	139	144	155	169	192	198	193	187	183
MS	83	83	76	92	103	91	67	110	141	132
NC	153	160	184	207	210	214	234	223	240	244
OK	144	140	120	109	114	133	129	146	147	151
SC	118	121	130	158	198	230	239	243	237	238
TN	86	82	84	91	109	114	127	134	151	153
TX	141	136	147	141	154	167	176	189	196	210
VA	109	107	108	105	110	126	142	157	158	161
WV	60	59	61	57	65	71	67	63	66	64
WEST	82	79	86	94	84	91	92	99	101	105
AK	66	61	57	57	56	63	75	127	133	143
AZ	74	77	81	97	118	125	129	146	139	160
CA	87	84	99	106	81	85	80	88	93	98
CO	86	81	78	79	80	87	89	93	90	96
HI	34	39	37	39	42	39	44	57	58	65
ID	49	50	558	66	71	82	87	91	92	87
MT	35	40	44	46	50	73	81	87	96	94
NV	124	121	135	130	136	156	187	204	224	230
NM	61	56	66	81	86	105	126	123	112	106
OR	94	84	75	88	108	122	122	117	122	120
UT	53	51	45	46	54	60	64	69	68	64
WA	82	77	77	86	96	109	118	122	113	106
WY	78	76	77	74	80	87	98	102	95	113

Table 1.6.1 in state and federal institutions on Dec. 31, of the years 1971 to 1991.

'81	'82	'83	'84	'85	'86	'87	'68	'89	'90	'91
103	115	127	136	145	157	169	186	215	232	248
95	114	114	119	127	135	144	146	194	238	263
71	69	75	72	83	106	106	100	116	118	123
65	77	79	84	88	92	102	109	122	132	143
42	47	50	57	68	76	81	93	103	117	132
92	107	136	138	149	157	177	219	251	271	301
145	158	172	187	195	216	229	248	285	304	320
78	88	98	109	119	128	136	149	169	183	192
72	82	92	92	99	103	100	118	146	157	173
76	84	72	74	82	81	91	98	109	117	124
121	130	135	144	161	173	184	200	225	239	255
113	119	135	149	161	168	171	181	211	234	247
138	152	164	165	175	181	192	202	217	223	226
88	93	92	97	98	98	101	107	126	139	144
116	129	152	173	192	217	233	232	222	227	231
165	162	159	161	196	227	259	298	340	366	388
49	50	52	52	56	58	60	64	71	72	78
131	147	162	175	194	203	218	236	269	267	305
104	99	91	95	108	116	123	129	141	140	145
33	47	51	54	55	53	57	62	62	67	68
139	160	155	174	194	209	219	243	279	289	324
97	109	115	127	146	160	160	143	175	187	191
93	96	102	105	113	119	126	130	138	149	157
201	224	225	231	236	248	255	266	292	316	333
183	215	243	256	267	283	307	300	328	370	394
143	166	179	188	195	198	227	230	261	277	317
208	250	273	263	281	311	326	331	333	323	344
467	531	558	649	738	753	905	1,078	1,132	1,148	1,221
224	261	235	242	247	272	265	278	307	336	344
220	247	259	254	251	265	282	281	300	327	342
114	110	127	128	133	142	147	191	222	241	262
216	251	290	310	308	316	346	370	396	427	462
218	244	277	285	279	280	282	291	323	348	366
177	210	211	229	237	249	256	277	293	307	330
248	255	233	246	254	257	250	249	250	265	269
169	201	212	236	250	288	296	323	361	381	416
251	270	276	284	294	324	344	369	416	451	473
171	173	187	154	149	157	156	157	213	207	227
210	237	221	226	226	228	231	240	257	290	297
165	177	177	185	204	215	217	230	263	279	311
80	77	83	82	89	77	77	78	84	85	83
119	139	152	166	176	197	214	234	256	277	287
170	194	219	252	288	306	339	355	361	348	345
184	209	223	247	256	268	307	328	350	375	396
114	135	150	162	181	212	231	257	283	311	318
92	108	109	104	103	115	145	174	207	209	249
77	88	103	124	134	142	141	136	142	150	153
99	107	121	127	133	144	144	157	180	190	205
104	114	104	121	136	135	147	158	165	176	183
245	301	354	380	397	447	432	452	438	444	439
100	126	142	133	144	154	174	180	178	196	191
124	146	157	170	165	176	200	215	235	223	228
73	77	77	84	98	108	110	115	137	142	149
125	148	155	156	156	147	134	124	142	162	182
117	135	138	143	148	168	190	199	216	237	237

1.7 Time series

Natural phenomena are often observed over a period of time at discrete time intervals and then recorded as a *time series*. Typical examples would be daily recordings of rainfall or maximum temperature at a given location, or stock price fluctuations on the New York Stock Exchange.

Example 1.7.1 SAMS in Space.

The Table 1.7.1 below (and the file SHUTTLE on the UVW Web Site) provides data extracted from the Space Accelerometer Measurement System (SAMS) that was present aboard the space shuttle *Columbia* during STS-50. The acceleration ruins the diffusion controlled aggregation experiments that were supposed to be conducted in the near-zero gravity environment; so it was important to keep track of it. The data period begins on day 007, hour 22 (MET). Sampling rate is 12.5 samples/second. No significant peaks are present in this sample, so the data can be interpreted as random background noise and accelerations. The data were provided by Milton Moskowitz, a graduate student in the Materials Science Department, and were obtained as part of a CWRU Microgravity Lab project.

Often to detect regularities, or irregularities, it is more convenient to present the time series in the graphical form.

Example 1.7.2 EKG on Soaps.

Fig. 1.7.1 shows successive R-R intervals (in seconds) of a normal, resting dog. A major upward "blip" on an electrocardiogram (EKG) is called an *R-wave*, and the R-R interval is an interval between two consecutive R-waves. You may have watched many a soap opera emergency room where a confident young doctor would casually opine to an emaciated patient: "See, your R-R interval correlation dimension has fallen from the normal of 2-3 to near 1, so things look very grim." But what did the good doctor really mean? You will find out later on. Note that presented at a different time scale, things may look quite different (see Fig. 1.7.2), and sharp "blips" do not look that sharp anymore.

Figs. 1.7.3 and 1.7.4 show portions of an electroencephalogram (EEG) of a normal, waking adult. There is no way to add regular "blips" to an EEG, so these rhythms never quite make it to prime time TV. The above sample is 2 minutes long, sampled at 200 Hz. The data were provided by Mark D. Bej, of the Cleveland Clinic Foundation.

Sometimes the periodicities and time-correlations in the signal are quite obvious as in the following example.

Table 1.7.1 Acceleration aboard space shuttle Columbia sampled at the rate of 12.5 samples/second.

0	-1.393413e-04	1	-1.347195e-04	2	-1.100744e-04
3	-9.775257e-05	4	-1.008320e-04	5	-1.100709e-04
6	-1.285549e-04	7	-1.424222e-04	8	-1.408843e-04
9	-1.331803e-04	10	-1.239331e-04	11	-1.193078e-04
12	-1.069854e-04	13	-9.928829e-05	14	-1.270203e-04
15	-1.424276e-04	16	-1.254827e-04	17	-9.620895e-05
18	-7.001450e-05	19	-7.001350e-05	20	-9.928802e-05
21	-1.393437e-04	22	-1.455037e-04	23	-1.085298e-04
24	-8.388260e-05	25	-1.162321e-04	26	-1.439624e-04
27	-1.224023e-04	28	-1.147032e-04	29	-1.362659e-04
30	-1.516663e-04	31	-1.347214e-04	32	-1.100748e-04
33	-1.193151e-04	34	-1.532005e-04	35	-1.562799e-04
36	-1.100700e-04	37	-9.466843e-05	38	-1.193158e-04
39	-1.393405e-04	40	-1.316371e-04	41	-1.331771e-04
42	-1.377995e-04	43	-1.285593e-04	44	-1.208582e-04
45	-1.347212e-04	46	-1.470435e-04	47	-1.316405e-04
48	-1.193169e-04	49	-1.116138e-04	50	-1.285562e-04
51	-1.501199e-04	52	-1.377996e-04	53	-1.085368e-04
54	-1.085385e-04	55	-1.177798e-04	56	-1.316410e-04
57	-1.362619e-04	58	-1.270204e-04	59	-1.362619e-04
60	-1.516642e-04	61	-1.331787e-04	62	-9.928946e-05
63	-1.054490e-04	64	-1.300945e-04	65	-1.300968e-04
66	-1.193166e-04	67	-1.085363e-04	68	-1.008375e-04
69	-1.054583e-04	70	-1.008344e-04	71	-9.158860e-05
72	-9.620688e-05	73	-9.312755e-05	74	-1.069962e-04
75	-1.131623e-04	76	-1.008200e-04	77	-9.464904e-05
78	-1.069712e-04	79	-1.146803e-04	80	-1.193076e-04
81	-1.085253e-04	82	-1.131418e-04	83	-1.115988e-04
84	-1.239222e-04	85	-1.193024e-04	86	-1.254620e-04
87	-1.162172e-04	88	-1.162160e-04	89	-1.285403e-04
90	-1.424036e-04	91	-1.192985e-04	92	-1.100573e-04
93	-1.300821e-04	94	-1.454869e-04	95	-1.470287e-04
96	-1.454863e-04	97	-1.439431e-04	98	-1.239184e-04
99	-1.085183e-04	100	-1.285449e-04		

Example 1.7.3 Breathing Patterns.

Fig. 1.7.5 shows the activation times of a neural cell in the brain (top graphs in each of the modes 1-4). The selected cell is responsible for inspiration. Modes 1 and 2 show, respectively, spontaneous activity of the cell during wakefulness and non-rapid-eye-movement (non-REM) sleep. Modes 3 and 4 show, respectively, intense activity of the cell during a smoke-induced apnea (suspended breathing,

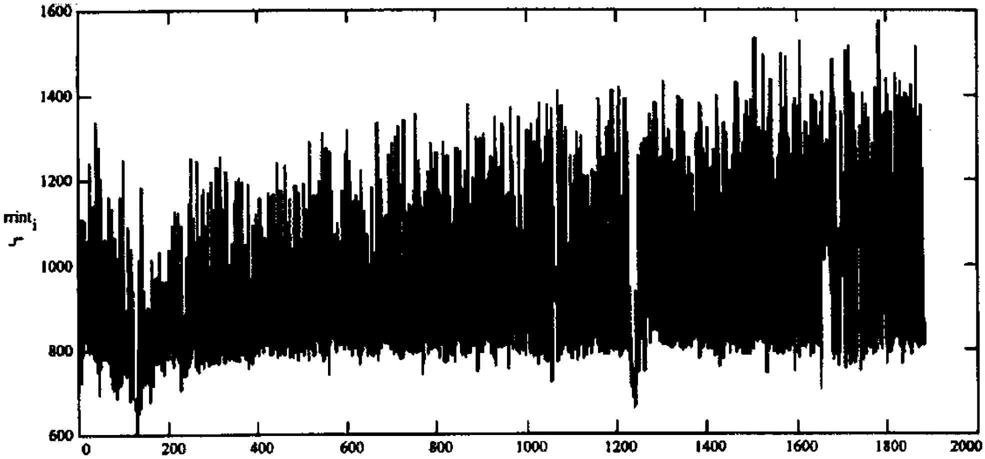


FIGURE 1.7.1
Successive R-R intervals on an EKG.

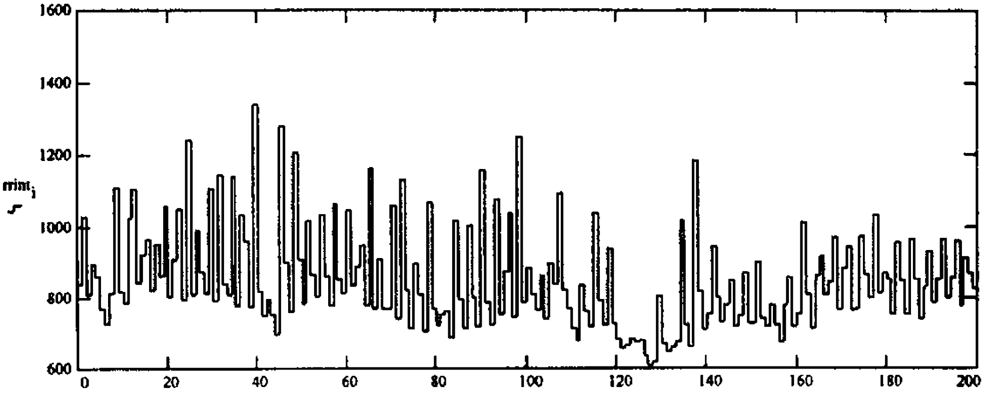


FIGURE 1.7.2
A portion of Fig. 1.7.1.

seen as a long pause in tracing 3) and in response to the conditioning stimulus. In each mode, the tracing below the cell activation times (action potentials) is the intratracheal pressure, with negative pressures (inspiration) indicated by upward deflections. These data were supplied by Sharmila Kopanathi, a bio-engineering graduate student.

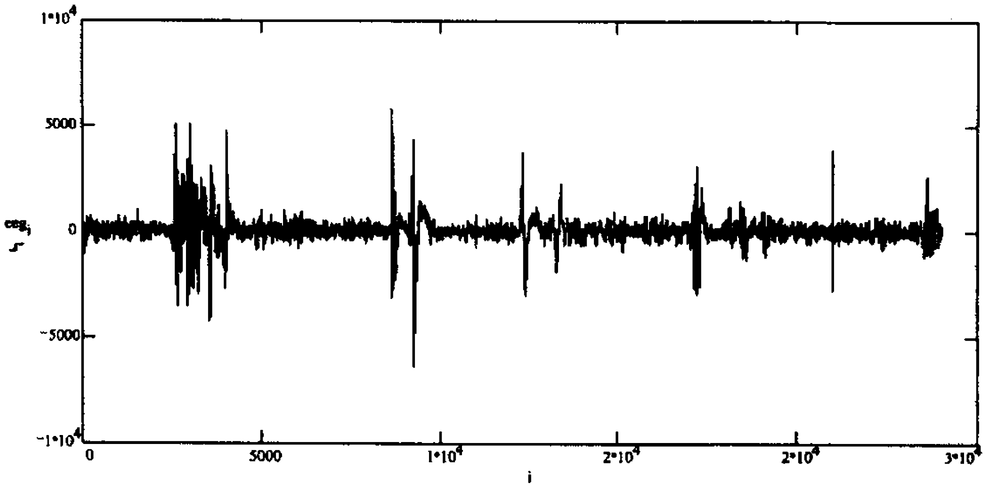


FIGURE 1.7.3
EEG of a normal, waking adult.

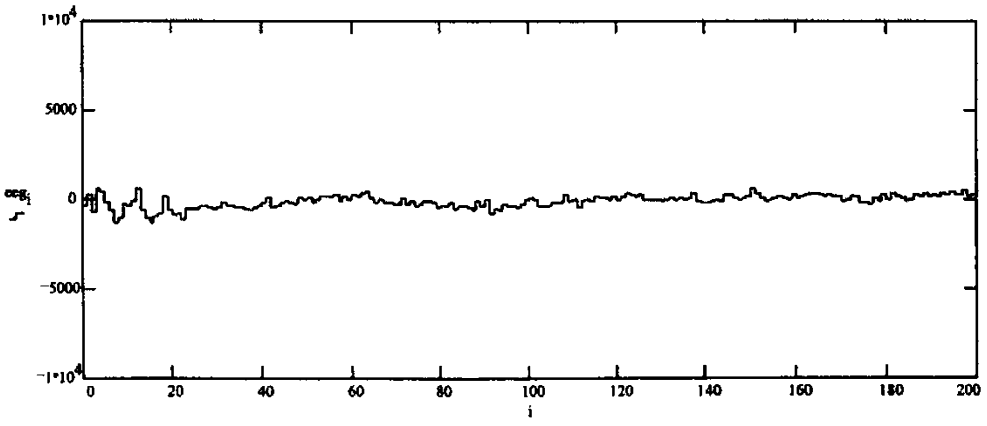


FIGURE 1.7.4
A portion of Fig. 1.7.3.

1.8 Repeated experiments and testing

Repeated (and repeatable) experiments are the mainstay of engineering and scientific research. The famous Lord Rutherford's advice "If your experiment needs statistics, you ought to have done a better experiment" has, obviously, only limited applicability.

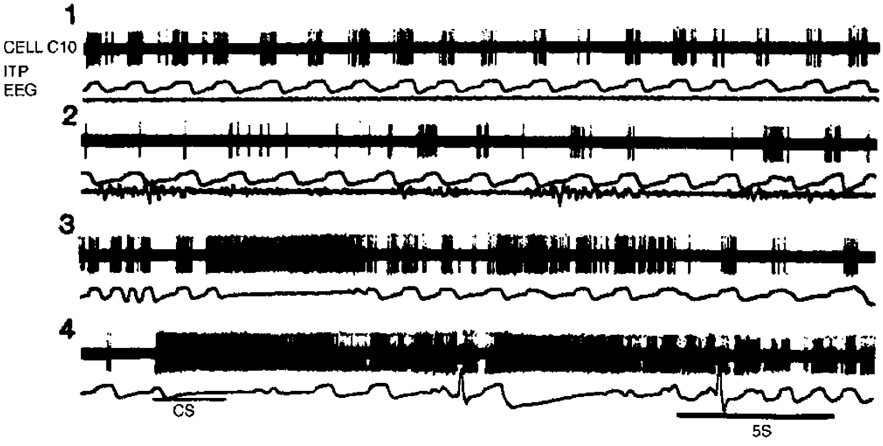


FIGURE 1.7.5
Activation pattern in an inspirational neuron.

Example 1.8.1 Less Pain.
Bob Ruff, Ted Carroll and Richard Welser, faculty members at the CWRU Medical School, tested a new medication regimen to reduce pain in 21 terminal cancer patients over a certain period of time. Patients were of different ages and tumors were located at different sites and the dosage for both old and new regimens were different for different patients. The results of the probe were recorded as subjective patient evaluations on a scale from 0 to 10 of the current, best, and worst pain suffered during pre-regimen and post-regimen periods. The results are presented in Table 1.8.1 below.

How effective was the new regimen of medication in comparison to the old one? How can this statement be quantified? What was the dependence of the effectiveness of the new regimen on the dosage and on the location of the tumor? Was the patient sample sufficient to draw any firm conclusions? Did the researchers have any control over the sample size? There could have been only so many patients available. What about a control population?

Here is an example of an engineering testing problem.

Example 1.8.2 Cracks Propagate.
Mechanical and civil engineers often face the problem of component failure in situations when the latter are subject to cyclic loading which may lead to creation and propagation of fatigue cracks. Remember that fateful flight of Aloha Airlines when the whole top of the plane came off in midair. The later investigation by the Federal Aviation Administration showed that the fuselage, submitted to periodic pressurizing and depressurizing during, respectively, takeoffs and landings, finally succumbed to the excessive fatigue cracking.

Table 1.8.1 Pain rating in terminal cancer patients.

	Pain rating					
Patient	Pre-regimen			Post-regimen		
No.	Now	Best	Worst	Now	Best	Worst
1	4	1	8	1	0	3
2	6	6	9	0	0	2
3	5	3	8	2	0	4
4	7	5	9	0	0	3
5	4	4	4	2	0	4
6	7	5	9	1	0	3
7	4	2.5	9	2	2	4
8	6	4	8	0	0	0
9	5	5	8	0	0	4
10	6	3	9	1	0	3
11	5	3	7	2	0	4
12	4	4	7	0	0	2
13	4	2	8	2	0	5
14	4	2	6	2	1	4
15	7	2	10	2	0	4
16	2	0	5	0	0	2
17	8	5	10	0	0	2
18	5	3	8	2	0	4
19	8	5	10	0	0	5
20	2	2	9	0	0	2
21	2-3	2-3	8	0	0	4

The crack is usually initiated at a (random) site of unavoidable material defect, often microscopic, where stresses are particularly high. Then, periodic loading causes accumulation of damage in the micro-structure of the material and the crack propagates. The lab data are usually collected via periodic inspection of the trajectory of the crack and consist of recording the crack length a corresponding to the total number of load cycles $N(a)$ up to that time. Even for fairly uniform samples, the data show a lot of randomness. In a study by P. Goel, a statistician at Ohio State University, and D. Virkler, a mechanical engineer at Purdue University (see Fig. 1.8.1), 68 replicate tests were conducted under identical loading. The specimens were aluminum panels, and the constant-amplitude load was cycled at 20 Hz.

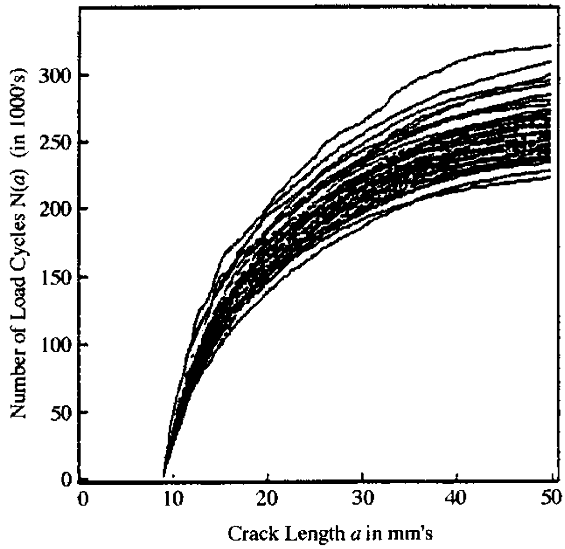


FIGURE 1.8.1
The time-evolution of crack lengths for 68 identical centercracked aluminum panels subject to 20 Hz cyclical loading.

1.9 Simple chaotic dynamical systems

Surprisingly, random effects can arise in seemingly simple dynamical systems with only deterministic and well-controlled ingredients. When this happens we often talk about the system's *chaotic behavior.*

Example 1.9.1 Billiard vs. Pinball.

Consider a ball moving on a rectangular billiard table. Assume, idealizing the situation, that there is no friction and no spin, and that the ball moves with constant unit speed along straight line intervals between reflections, and obeys the law of equal incidence and reflection angles on collision with the boundaries.

The trajectories of such a ball can be of different nature: periodic, sweeping perhaps only part of the billiard table, or aperiodic which may sweep the whole table surface. It is clear that the single ball's trajectory depends on the ball's initial position and velocity (angle), and on the relationship between the sizes of the ball and the table. Theoretically, with an ideal point-size ball, when the initial angle is a rational multiplicity of π the trajectory is periodic and, when it is irrational, the trajectory is aperiodic (see a related discussion on irrational rotations in Chapter 6). In computer experiments, the question of irrationality is delicate; only rational

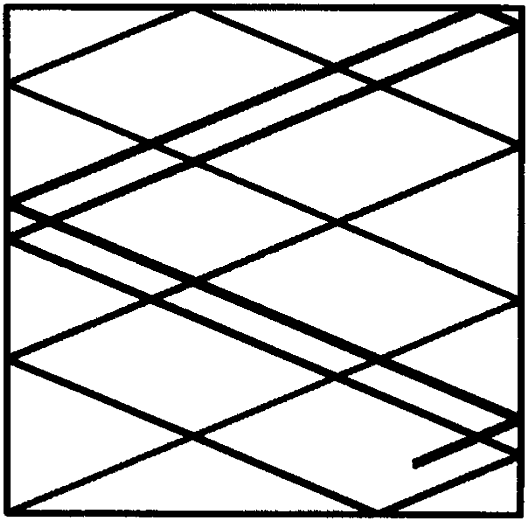


FIGURE 1.9.1
Billiard table. The initial angle between the trajectories of two balls is 0.001. The trajectories are almost indistinguishable.

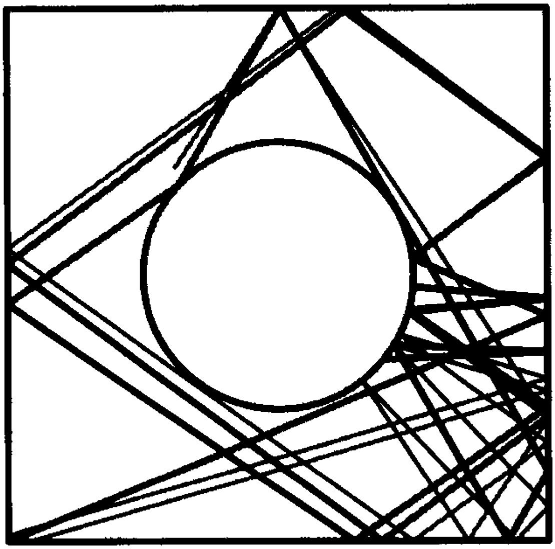


FIGURE 1.9.2
Pinball table. The initial angle between the trajectories of two balls is 0.001. The trajectories quickly diverge.

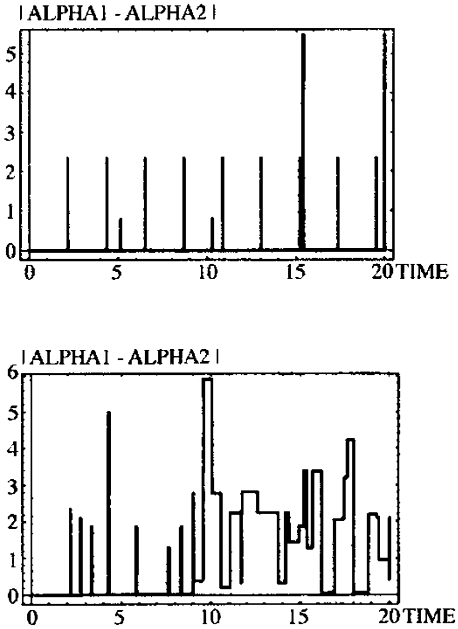


FIGURE 1.9.3

The time-evolution of the angle between trajectories of two balls shot at slightly different initial angles. Top: On the billiard table pictured in Fig. 1.9.1. Bottom: On the pinball table with a round obstacle pictured in Fig. 1.9.2.

numbers can be produced, although, perhaps, with a lesser or greater degree of complexity.

Now, consider trajectories of two balls shot from the same point but at slightly different angles (Fig. 1.9.1). The time-evolution of the angle between the velocities of two balls is shown in the top half of Fig. 1.9.3. Notice that, except for short-duration "blips" due to boundary effects, the angle α between the trajectories of the balls is being preserved. In the same situation the distance between two balls increases linearly with time (see Fig. 1.9.4, top).

Next, let us analyze what happens if we put a round obstacle in the middle of the billiard table thus creating a sort of simple pinball table. For a single ball, periodic trajectories are still possible, and so are the aperiodic trajectories, sweeping the whole pinball table and hitting the obstacle infinitely often. However, if one looks now at the trajectories of two balls shot from the same point at close initial angles (Fig. 1.9.2.) the situation differs dramatically from the one we encountered in the case of the billiard table.

After the first collision with the obstacle, the angle between the two trajectories is (see Fig. 1.9.5)

$$\alpha^1 = \alpha + 2\beta$$

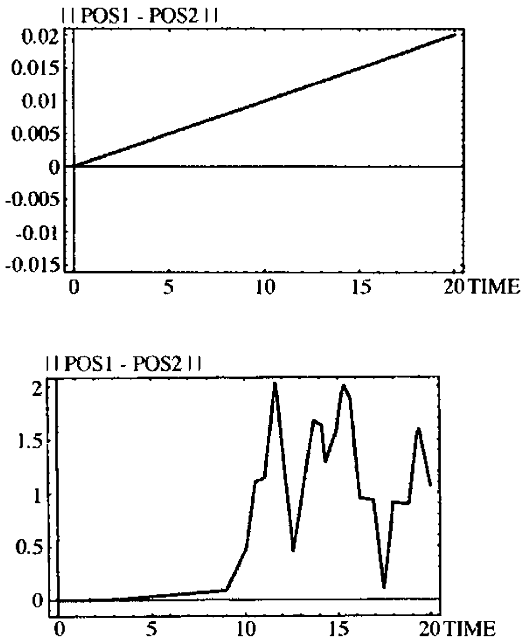


FIGURE 1.9.4

The time-evolution of the distance between two balls shot at slightly different initial angles. Top: On the billiard table pictured in Fig. 1.9.1. Bottom: On the pinball table with a round obstacle pictured in Fig. 1.9.2. Distances between the positions of two billiard balls in the case of the billiard table without an obstacle and the billiard table with an obstacle.

and since, for a small initial angle α, the angle β is proportional to α (say $\beta = \alpha$), we see that

$$\alpha^1 = 3\alpha.$$

In other words, the initial angle between trajectories is tripled after the first collision with the obstacle. The same will happen after the second collision so that

$$\alpha^2 = 3^2\alpha = 9\alpha.$$

Again, were it not for the boundary, each collision with the round obstacle would triple the angle between the trajectories, causing it to grow exponentially. This phenomenon is called a *sensitive dependence on initial conditions* and it does not occur in the billiard table without obstacles. Also, if one looks at the behavior of the distance between the two balls on the pinball table there seems to be no regularity there; we have discovered *chaotic behavior* in a very simple, otherwise

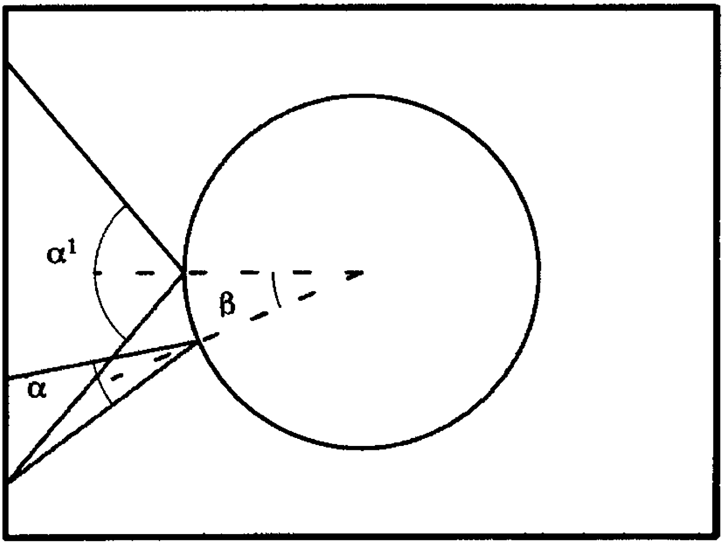


FIGURE 1.9.5
The angle between the trajectories of two balls triples after each collision with the obstacle.

deterministic, dynamical system. In 1970, the above heuristic arguments have been made rigorous by a Russian mathematician Yakov G. Sinai.

Example 1.9.2 Brazilian Butterfly.
Consider a dynamical system $(x(t), y(t), z(t))$ evolving in a three-dimensional space, depending on the continuous time t, and described by a system of three ordinary nonlinear differential equations:

$$\frac{dx}{dt} = -\sigma x + \sigma y,$$

$$\frac{dy}{dt} = -xz + rx - y \qquad (1)$$

$$\frac{dz}{dt} = xy - bz.$$

The system was proposed in 1963 by Edward N. Lorenz as a "toy" atmospheric circulation model but it helped jump-start the modern theory of chaotic behavior in the physical sciences. Despite its simplicity it displays a sensitive dependence on initial conditions (see Fig. 1.9.6). The model gave rise to the well-popularized chaos-theory image: a butterfly flapping its wings in the Brazilian rain forest can

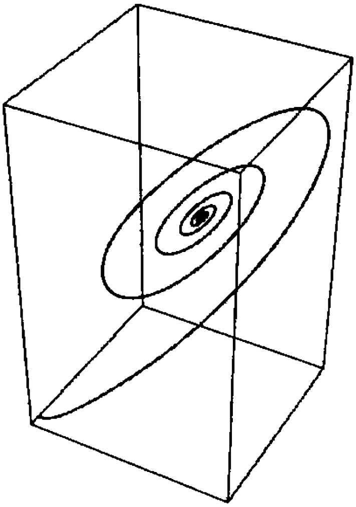


FIGURE 1.9.6

The trajectory of the points (x, y, z) *corresponding to the solutions of (1) with initial conditions near* $(0, 0, 0)$ *and with* $\sigma = 5, b = 7, r = 2$.

cause a typhoon a few weeks later in the China Sea. In 1990 Lorenz was awarded the Kyoto Prize ($500,000) for his contribution.

The nonlinearity is essential for the complex behavior of the Lorenz model. The equations are obtained by truncation of the Navier-Stokes equations (see also Section 1.10 on complex dynamical systems) that describe the conservation laws of the fluid flow. Fig. 1.9.6 and 1.9.7 show *Mathematica* simulations of the trajectories of the Lorenz system which used the package `UVW'Lorenz`. Depending on the parameter values the behavior of the system may be quite different.

Example 1.9.3 Iterations of Quadratic Maps.

To conclude this section we will take a quick look at a simple dynamical system determined by iterations of the function $f(x) = ax(1 - x)$, $0 \leq x \leq 1$, which will play a role later on in Chapter 6. The time $t = 0, 1, 2, \ldots$ is assumed to be discrete. Starting with an $x_0 \in [0, 1]$, the successive states $x_1, x_2, \ldots \in [0, 1]$ of the system are produced according to the following recursive formula

$$x_t = f(x_{t-1}).$$

Depending on the value of the coefficient a and the starting point x_0 the system displays a whole variety of behaviors from asymptotically stable, to periodic, to chaotic.

Mathematica Experiment 1. Billiard vs. Pinball. The above analysis of the

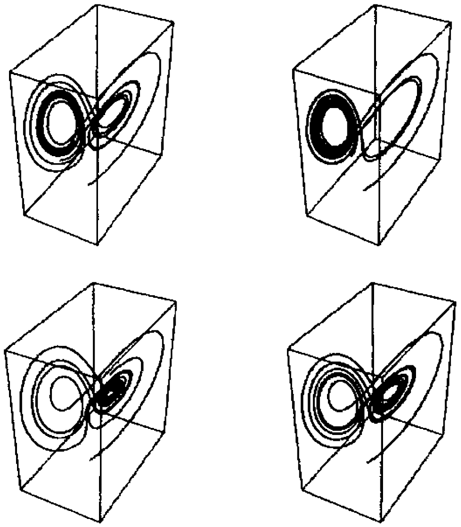


FIGURE 1.9.7
The trajectory of the points (x, y, z) *corresponding to the solutions of (1) with initial conditions near* $(0, 0, 0)$ *and parameters* $\{\{\{3, 26.5, 1\}, \{3, 25, 1\}\}, \{\{4, 26.5, 1\}, \{4, 25, 1\}\}\}$.

billiard and pinball was conducted via the *Mathematica* package `UVW‘Billiard‘` which is a part of the `Uncertain Virtual Worlds (UVW)` packages that can be found on the `UVW Web Site`. Before you start experimenting with it, the package has to be loaded using the command `<<UVW‘Billiard‘`.[4]

```
In[1]:= <<UVW‘Billiard‘
In[2]:= BilliardAnimate[{0.2},20]
In[3]:= BilliardTrajectories[{0.2,0.3,0.4,0.5},10]
In[4]:= BilliardDifferences[0.4,0.001,20]
In[5]:= BilliardDifferences[0.4,0.001,20,0.]
```

Mathematica Experiment 2. Iterations of Quadratic Maps. In this experiment we will need the following *Mathematica* commands:

`NestList[f, x, n]`

Usage: Produces a list of `n` successively nested (iterated) functions f, i.e. `x, f[x], f[f[x]], ... , f[f[f....f[x]]]`

`Random[Real, {0,1}, n]`

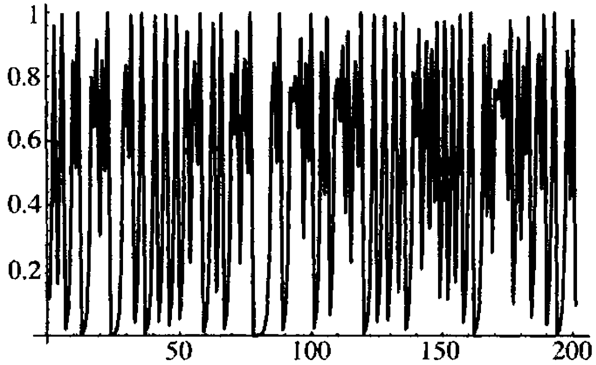
Usage: Produces an n-digit pseudorandom number in the range 0 to 1.

[4] Instructions for installation of UVW packages can be found in Appendix E. Also see *Mathematica* bibliography at the end of this chapter.

The following experiment will produce a (joined) graph of 200 iterations of the quadratic (logistic) function $f(x) = 4x(1-x)$ starting with a random point between 0 and 1 determined with the precision of 200 digits. You may further experiment with this system by changing the coefficient 4 to any number a, $0 < a < 4$. Note that for $a > 4$ the system no longer maps the unit interval state space into itself. We will take a closer look at its complex behavior in Chapter 6.

```
In[1]:= f[x_] := 4 x(1-x)
In[2]:= ListPlot[NestList[f, Random[Real, {0,1}, 200]], 200],
          PlotJoined -> True]

Out[2]:=  -Graphics-
```



The output of the graphics is random because the starting point was selected randomly. Every time you run the above experiment the trajectory will be slightly different.

1.10 Complex dynamical systems

Loosely speaking, by a complex dynamical system we mean a system with a very large, or even infinite, number of degrees of freedom. As an example consider the system of gas particles in the classroom. The instantaneous state of the system is described by the vector

$$(x_1, y_1, z_1, \dot{x}_1, \dot{y}_1, \dot{z}_1, \ldots, x_N, y_N, z_N, \dot{x}_N, \dot{y}_N, \dot{z}_N),$$

where (x_i, y_i, z_i) are the coordinate vectors of the ith particle and $(\dot{x}_i, \dot{y}_i, \dot{z}_i)$ is its velocity vector and $i = 1, 2, \ldots, N$. The number of degrees of freedom in this system is $6 \cdot N$, with N, roughly speaking, of the order of the Avogadro's number, i.e., $N \approx 6.0225 \cdot 10^{23}$. Another example is given below by following the continuum fluid flow dynamical system, which, as a matter of fact, can be obtained from the above particle system by taking the number N of particles to infinity.

Example 1.10.1 Turbulent Flows and Diffusions.
The motion of an incompressible fluid is described by the following system of nonlinear partial differential equations

$$\partial_t \boldsymbol{\xi} + (\boldsymbol{u} \cdot \nabla)\boldsymbol{\xi} - (\boldsymbol{\xi} \cdot \nabla)\boldsymbol{u} = R^{-1}\Delta\boldsymbol{\xi},$$

$$\operatorname{div} \boldsymbol{u} = 0, \qquad \boldsymbol{\xi} = \operatorname{curl} \boldsymbol{u}, \tag{1}$$

where

$$\boldsymbol{u}(\boldsymbol{x}) = \Big(u_1(x_1, x_2, x_3), u_2(x_1, x_2, x_3), u_3(x_1, x_2, x_3)\Big)$$

is the velocity field, $\boldsymbol{\xi}$ is the vorticity field, t is the time, and

$$\nabla = \left(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \frac{\partial}{\partial x_3}\right), \qquad \Delta = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial x_3^2}$$

are, respectively, the gradient and the Laplacian differential operators. Equations (1) are called *Navier-Stokes equations* and their solutions are vector fields depending on time and space location, and can be thought of as points in an infinite dimensional space of vector-valued functions of three variables. Depending on initial and boundary conditions, and the value of the Reynolds number R which is proportional to the characteristic scale and velocity of the flow, and inversely proportional to the viscosity of the fluid, the behavior of the solutions can vary from laminar flows to turbulent ones (see Fig. 1.10.1).

What is often important in the study of atmospheric and oceanic flows is how a passive tracer, that is light particles that are carried by the fluid flow but do not affect the flow itself, is transported in turbulent and more generally, random velocity flows. This is the problem of *turbulent diffusion* that until this day is not completely understood.

Fig. 1.10.2 shows the density distribution of a passive tracer at $t > 0$. At the initial time $t = 0$ it was uniformly distributed in space and then was carried by a random potential velocity flow.

Fig. 1.10.3 shows the contour of constant density of the passive tracer carried by an incompressible flow. The contour initially was a circle indicating a radial symmetric distribution of the tracer density. One can demonstrate that the length of such a contour grows exponentially in time.

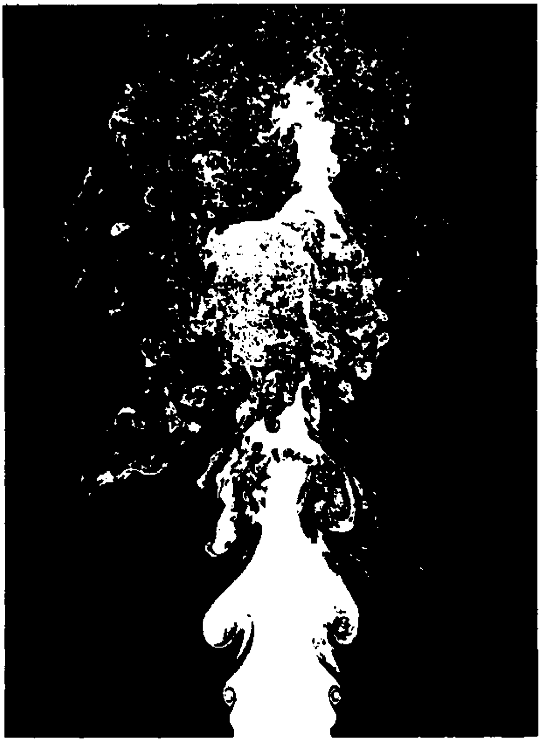
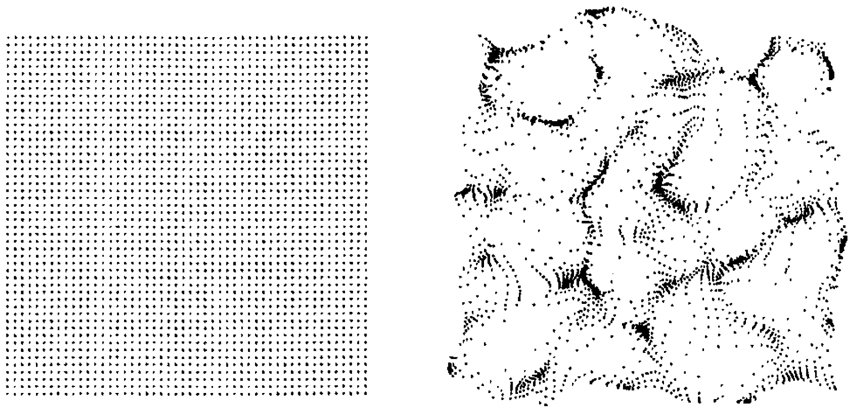


FIGURE 1.10.1
Radial section of a turbulent flow from an axisymmetric jet.

The turbulent velocity field shown in Fig. 1.10.1 appears random and not likely to appear again in exactly the same way even if the experiment is repeated under the same conditions. So, a fluid dynamicist would study a more stable object connected with it, namely the distribution (histogram) of the velocity components in a measurement taken over a certain period of time, see Fig. 1.10.4.

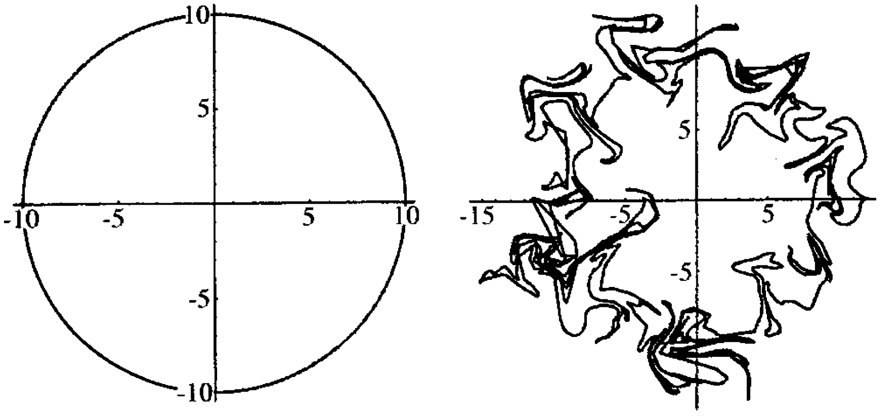
1.11 Coin tossing revisited: pseudorandom number generators and the Monte-Carlo methods

Simulation of random phenomena, often generically called the *Monte-Carlo Method*, introduced by Stanisław Ulam, John von Neumann and Nicolas Metropolis in the late 1940s, is now a routine technique in engineering and the physical sciences. It depends on the computer's ability to produce a random sequence of



Time 0.0 *Time 0.5*

FIGURE 1.10.2
Distribution of the passive tracer in the random potential velocity flow. The initial distribution was uniform in space.



Time 0.00 *Time 4.00*

FIGURE 1.10.3
Contour of constant passive tracer density in an incompressible fluid. The initial contour was a circle.

numbers. However, the `random` command should not be used uncritically because its execution is always a result of the deterministic code. What kind of "randomness" can actually be expected from such a procedure? As it turns out, to get the computer to reproduce a coin-tossing experiment is not an easy matter.

Actually, there are only a few truly distinct methods used by computer random

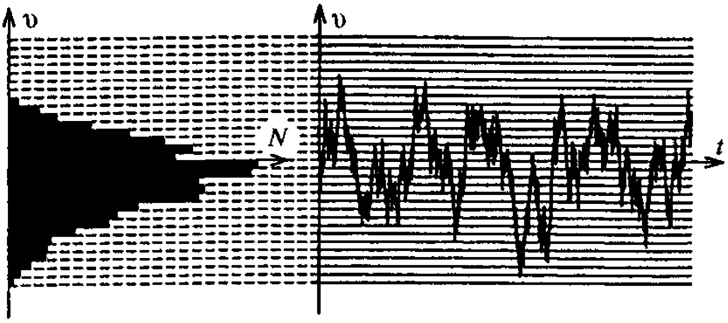


FIGURE 1.10.4
The turbulent signal on the right appears random and not repeatable, but the distribution (histogram) of its values shown on the left is a more stable object. (From U. Frisch, Turbulence, Cambridge University Press, 1996.)

number generators, and none of them produce perfectly random sequences because they are all using deterministic algorithms. For that reason it is safer to call them *pseudorandom number generators.* It is clear that one expects from such a generator more than just satisfaction of the equipartition property discussed in Section 1.1. After all, in the Champernowne number

$$C = 0.1234567891011121314151617181920212223\ldots$$

all blocks of the same length have identical frequencies but nobody would propose it as a random, or even pseudorandom number.

Example 1.11.1 Midsquare Method.
The oldest computer method for producing pseudorandom numbers is due to John von Neumann and is called the *midsquare method.* It works as follows: Take a four digit number, say 6514, compute its square 42432196 and take the middle four digits 4321 as the next pseudorandom number. Then repeat the procedure to obtain the next pseudorandom number 6710 and so on. One hopes that with a clever choice of the initial seed number one gets a sequence uniformly distributed among ten thousand four-digit numbers. However, it is easy to see that this is not so and the method, although simple in execution, has serious statistical flaws.

Example 1.11.2 Fibonacci Sequence.
It is also called the *additive congruential method.* The prescription is as follows: Pick the first two integers x_0, x_1 arbitrarily and then proceed recursively by defining

$$x_i = x_{i-1} + x_{i-2} \pmod m)$$

for a choice of m. So,

$$1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, 110, 487, 97, 84, 181, \ldots$$

is the Fibonacci sequence for $m = 500$. The problem with the Fibonacci sequence is that it can never produce triples x_i, x_{i-1}, x_{i-2} satisfying inequalities

$$x_{i-1} < x_{i+1} < x_i, \quad x_i < x_{i+1} < x_{i-1},$$

and such triples should appear with frequency 1/6. Right? Try to explain why.

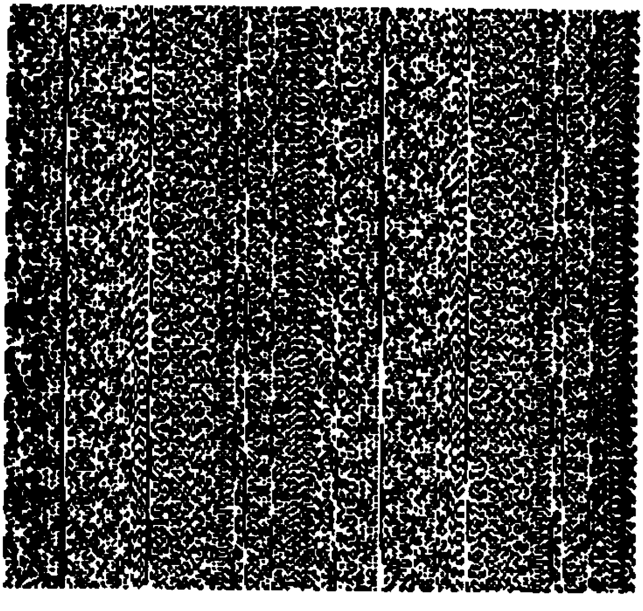


FIGURE 1.11.1
The Turbo Pascal random number generator shows an obvious lack of randomness. (From D. Griffeath, in Statistical Science, Vol. 8(1993).)

Example 1.11.3 Bold Stripes: Linear Congruential Method.
The prescription is somewhat similar to the one considered above. One picks x_0 arbitrarily, and for a choice of fixed parameters a, c, and m, one computes pseudorandom numbers by the recursive formula

$$x_i = ax_{i-1} + c \pmod{m}.$$

This method is commonly used in modern computers. The choice $m = 2^{31} - 1 = 2147483647$, $a = 16807$, and $c = 0$ is implemented in some computer languages.

However, take a look at Fig. 1.11.1 which plots (starting with the top row, left to right) pseudorandom numbers (0s represented by white dots and 1s by black dots) generated by TURBO PASCAL on a rectangular grid of 256×240 pixels. The obvious nonrandom pattern that shows up should shatter anybody's absolute faith in random number generators. On the other hand, as we will see later, devising a perfect random number generator is impossible, so one has to use what we have, understanding its shortcomings and limitations. The so-called Minimal Standard 32-bit generator is based on the recursive formula

$$x_{n+1} = 16807 x_n \pmod{2147483647},$$

and you can easily implement it yourself in *Mathematica.*

Recently, George Marsaglia of the Florida State University produced a *Random Number CDROM* including the *Diehard Battery of Tests of Randomness* which is based on mixed techniques incorporating some data collected from observations of inherently random quantum effects. This looks like a promising avenue in developing new pseudorandom number generators.

1.12 Fractals and image reconstruction

Some complex images can be encoded in a simple fashion although what is meant by "simple encoding" can require additional explanations. A good example of such a situation is fractals and random fractals (the term is used here in a colloquial sense). The first, presented on Fig. 1.12.1 has a very simple description in terms of the angle between the branches, the length of the branches, and the number of levels.

The picture presented on Fig. 1.12.2 is similar, yet subtly different. Some of the branches are missing, but there is no simple pattern to how they were dropped. This is a random fractal. To reconstruct it exactly would take a long description. However, one can easily reconstruct it "statistically" by specifying the above parameters of the deterministic fractal and, in addition, provide a probability with which the branch in each generation is dropped. This is often the approach taken in the practically important area of image reconstruction (see Bibliographical Notes).

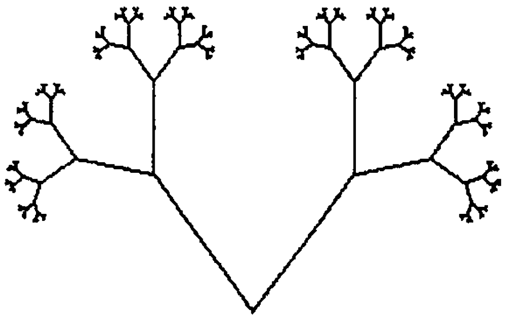


FIGURE 1.12.1
A deterministic fractal.

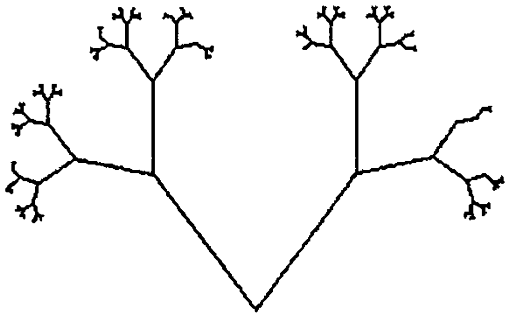


FIGURE 1.12.2
A random fractal.

1.13 Coding and decoding, unbreakable ciphers

When one tries to encrypt a confidential message, a high complexity of the encryption system could be a desirable goal. Consider a simple fixed code
It assigns to each letter a certain fixed sequence of numbers. If the message is long enough, the code can be broken by analyzing the frequencies of different symbols. During World War I the AT&T employee Gilbert S. Vernam and Major Joseph O. Mauborgne of the U.S. Army Signal Corps developed a different coding system, called the one-time pad system, that subsequently was demonstrated to be unbreakable and is commonly used in clandestine communications.

A good illustration of how it works is provided in Fig. 1.13.2, which shows a photograph of a sheet of paper found in 1967 on the body of the Latin American revolutionary Ché Guevara after he was captured (with CIA help) and executed by the Bolivian Army. It contains an encoded massage Guevara prepared for the Cuban President Fidel Castro who was supporting the insurrection.

$A \mapsto 6 \quad B \mapsto 38 \quad C \mapsto 32 \quad D \mapsto 4 \quad E \mapsto 8 \quad F \mapsto 30$
$G \mapsto 36 \quad H \mapsto 34 \quad I \mapsto 39 \quad J \mapsto 31 \quad K \mapsto 78 \quad L \mapsto 72$
$M \mapsto 70 \quad N \mapsto 76 \quad O \mapsto 9 \quad P \mapsto 79 \quad Q \mapsto 71 \quad R \mapsto 58$
$S \mapsto 2 \quad T \mapsto 0 \quad U \mapsto 52 \quad V \mapsto 50 \quad W \mapsto 56 \quad X \mapsto 54$
$Y \mapsto 1 \quad Z \mapsto 56$

FIGURE 1.13.1
An example of a fixed code that translates each letter of the alphabet into a one- or two-digit decimal number. It was used in the message shown in Fig. 1.13.2.

The message (in Spanish) was first encoded using a fixed code shown in Fig. 1.13.1 which transformed the original text into a sequence of decimal digits written out in the first line of each three line paragraph. For convenience the encoded message was broken into five-digit blocks.

The second line of each paragraph contained a sequence of random (pseudorandom) numbers known only to Guevara and Castro, used only once to encode this message and then destroyed. The third line contains the sums (written without carries) of the two digits appearing in the first two lines directly above them. It was only this line that was transmitted over open shortwave radio and then decoded by the reverse procedure in Havana. Because of the encryption procedure the cryptogram itself is a pseudo-random sequence; the more random, the better.

Physicists and cryptographers discuss current coding methods based on quantum effects (see Bibliographical Notes).

Mathematica Experiment 1. Union Jack in London Fog. Electronic transmission of messages over long distances (for example, from a spacecraft) unavoidably introduces some errors. In this experiment we will examine how random errors can influence the perception of the message, in this case the digitized picture of the Union Jack which is stored in file UJACK on the UVW Web Site.

The following *Mathematica* commands will be used:

```
Show[Graphics[Raster[n× n matrix]]]
```

Usage: produces a two-dimensional picture with (unmarked) rasters (i,j) where the intensity of the grey is given by the value of the matrix element at position (i,j). 0 stands for black and 1 for white.

```
<< Statistics‘DiscreteDistributions‘
```

Usage: loads the package Statistics’DiscreteDistributions’.

```
Random[BernoulliDistribution[p]]
```

Usage: returns a string of 0s and 1s. In the long run, 1s will have the relative frequency close to p.

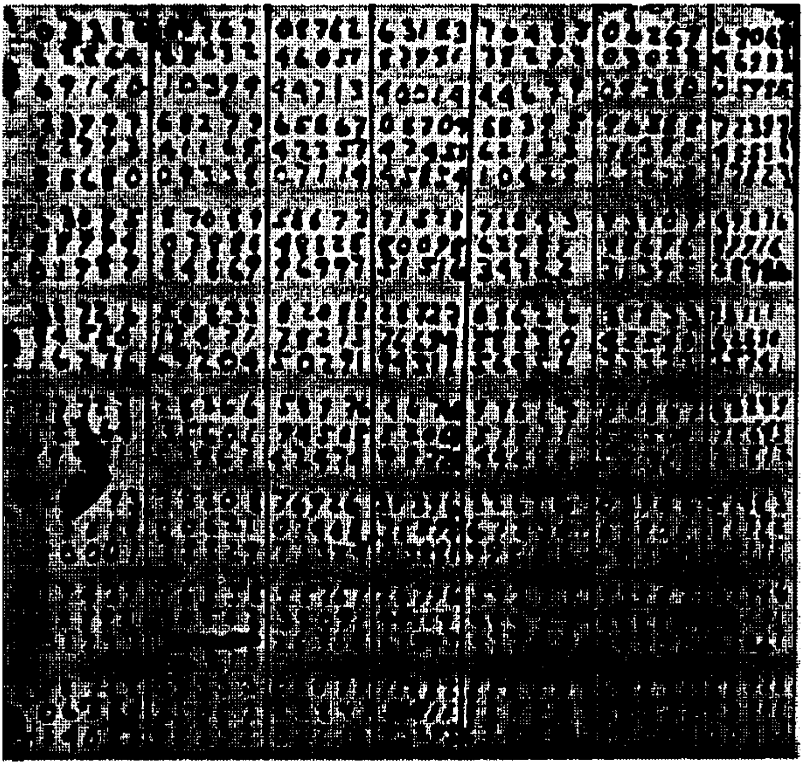


FIGURE 1.13.2
Latin American revolutionary Ché Guevara prepared this encoded message to Cuban President Fidel Castro in 1967 just before he was captured and executed by the Bolivian Army. He used the unbreakable, one-time pad, Vernam-Mauborgne cipher. (From C.H. Bennet et al., Scientific American, October 1992, pp. 50–57.)

`Mod[k,n]`

Usage: the remainder when dividing `k` by `n`.

A typical session follows.

```
In[1]:= <<Statistics'DiscreteDistributions'
In[2]:= Mod[2,2]
Out[2]= 0
In[3]:= TableForm[Table[{i,Mod[i,2]},{i,0,5}]]
Out[3]// TableForm=  0   0
                     1   1
                     2   0
                     3   1
                     4   0
                     5   1
```

```
In[4]:= Random[BernoulliDistribution[0.5]]
Out[4]= "randomly 0 or 1"
In[5]:= Table[Random[BernoulliDistribution[0.5]],{12}]
Out[5]= "list of twelve  random 0's and  1's"
In[6]:= Table[Random[BernoulliDistribution[0.1]],{12}]
Out[6]= "list of twelve  0's and  1's with the frequency of 1's
           equal to 1/10"
In[7]:= ran[p_]:=Table[Table[Random[BernoulliDistribution[p]],
           {30}],{30}]
In[8]:= ran[0.3]
Out[8]= "30 by 30 matrix with random 0's and 1's"
In[9]:= ujack={...};
In[10]:=Show[Graphics[Raster[ujack]]]
Out[10]= "Graphics"
In[11]:= co[p_]:=Table[Mod[r[p][[i,j]]+ujack[[i,j]],2],
           {i,1,30},{j,1,30}]
In[12]:= r[0.01]=ran[0.01]; Show[Graphics[Raster[co[0.01]]]
Out[12]= "Graphics"
In[13]:= r[0.12]=ran[0.12]; Show[Graphics[Raster[co[0.12]]]
Out[13]= "Graphics"
In[14]:= Quit
```

1.14 Experiments, exercises, and projects

1. *Mathematica Experiment 1.1.1 continued.* The file `ZEROONE1` on the `UVW Web Site` contains a list of 0 and 1. It should be loaded and examined. The colon in the following command suppresses the statement `Out[7]`.

```
In[1]:= zeroone1={ . . . };
In[2]:= Length[zeroone1]
Out[3]= 500
In[4]:= N[Sum[zeroone1[[i]],{i,1,Length[zeroone1]}]/
          Length[zeroone1]]
Out[4]= 0.48
In[5]:= Quit
```

1a) Compute the frequencies of the blocks 01 in the list `zeroone1.m`.

1b) Compute the frequencies of 1 in the lists (a)-(d).

1c) Compute the frequencies of all blocks of length one, two and three in the lists `zeroone1, zeroone 2, zeroone 3`.

1d) Find a string of 0s and 1s of length nine for which the frequencies of all the blocks of length two are $1/4$. Check your findings. (*Solution*: 000110110.)

1e) Repeat Exercise 1d for strings of length 26, blocks of length 3, and frequencies 1/ 8. (*Solution*: 00001010011100101110111100.)

2. *Mathematica Experiment 1.3.1 continued.*

2a) Find the 5th, 23rd and 52nd largest number in the list contained in the `BATTERY` file on the `UVW Web Site`.

2b) The file `REFRIGER` on the `UVW Web Site` contains the list of lifetimes for two brands of refrigerators. For each list compute the k-th largest where k is a multiple of Length[list]/10. Compare the two sequences and discuss how different they are.

2c) The data in file `RIVET` on the `UVW Web Site` contains the measurements of rivet heads. Find the number of measurements and determine the range of the measurements, i.e., the smallest and the largest measurements.

3. *Mathematica Experiment 1.4.1 continued.*

3a) Find the reliability of serial and parallel devices with reliabilities of individual components $r_j = 1/3 + j^2/j!$ for $j = 1, 2, \ldots, 15$. Use the definitions of `serial` and `parallel` saved from the `Mathematica` Experiment 1.4.1 (you need to do that experiment first, quit and restart again to do this exercise).

3b) Find the reliability of a device consisting of three components, the first two in parallel and the third in series with the first two. Take $r_j = 1/j$, $j = 1, 2, 3$.

4. *Project. (Mathematica Experiment 1.5.1 continued).* Draw a map of the stars contained in the file `STARS` on the `UVW Web Site`. Select one of four different dot sizes to indicate the star's brightness.

5. *Mathematica Experiment 1.6.1 continued.*

5a) Select the list of all educated groups in the list `CENSUS`.

5b) Compute the average number of children in the group of all uneducated and low income families in the list `CENSUS`.

5c) The file `PRISON` in the `UVW Web Site` contains data on the rate of sentenced prisoners from the table in Section 1.6. Order the list by region as in the printed table.

5d) Order the list `PRISONER` alphabetically by states.

5e) Extract the list of all states in `PRISONER` where, in 1991, the rate of sentenced prisoners per 100.000 residents was above 250.

6. *Project (Mathematica Experiment 1.6.1 continued).* For each year, make a list containing the name of the states with the lowest and the highest rate of sentenced prisoners. Which state had the highest overall rate?

7. *Mathematica Experiment. Lorenz equations.* Experiment with the `UVW Web Site` package `UVW'Lorenz'` varying the parameters. Note those that give rise to a chaotic behavior. A sample session is quoted below.

```
In[1]:= <<UVW'Lorenz'
In[2]:= Lorenz[3,26.5, 1]
In[3]:= para={{{3,26.5,1},{3,25,1}},{{4,26.5,1},{4,25,1}}};
In[4]:= LorenzArray[para]
```

8. *Project. Midsquare generator.* Use the `UVW'PsedoGene'` package to generate pseudorandom strings via the midsquare algorithm with different seeds. A sample session follows.

```
In[1]:= <<UVW'PsedoGene'
In[2]:= MidsquareGenerator[1234, 100]
In[3]:= MidsquareLoop[4578]
In[4]:= MidsquareLoop[9854]
In[5]:= MidsquareLoop[1245]
```

Then, using the tools developed in *Mathematica* experiments in Section 1.1.1, investigate its equipartition properties. Analyze the structure of the package itself.

9. *Project. Congruential generator.* Use the `UVW'PsedoGene'` package to generate pseudorandom strings via the congruential algorithm with various seeds. A sample session follows.

```
In[1]:= <<UVW'PsedoGene'
In[2]:= samp=CongruGenerator[0.23, 181,0,16384,2000]
In[3]:= <<UVW'DataRep'
In[4]:= RegularHisto[samp,0,1,10]
In[5]:= LargeNumbers[samp]
In[6]:= CentralLimit[samp,0.5,Sqrt[1./12],6]
In[7]:= samp2=Partition[samp,2];
In[8]:= SamplePlot2D[samp2]
In[9]:= CongruentialLoop[10,181,0,16384];
In[10]:= Length[%]
```

Then, using the tools developed in *Mathematica* experiments in Section 1.1.1, investigate its equipartition properties. Analyze the structure of the package itself.

10. *Project. Mathematica Experiment 1.13.1, continued.* Write a file of 2,500 zeros and ones producing a picture of your choice using the Raster command (here, one can also use a standard scanner equipment). Analyze and document changes in the picture and its perception as random errors are introduced with parameter p ranging in the interval [0, 1]. Discuss your findings.

11. Mathematica Experiment 1.6.1, an alternative. Use *Mathematica* package `Statistics'DataManipulation'` to repeat Experiment 1.6.1 on the database of the number of children in French-speaking Canadian families.

1.15 Bibliographical notes

With online *Mathematica* help that now comes with the software one can do without any hard copy manuals. However, we have found the following books helpful:

[1] S. Wolfram, *The Mathematica Book*, Wolfram Media, Champaign, IL, 1996.

[2] W.T. Shaw and J. Tigg, *Applied Mathematica*, Addison-Wesley, Reading, MA, 1994.

[3] R. Maeder, *Programming in Mathematica*, Addison-Wesley, Reading, MA, 1991.

[4] T.B. Bahder, *Mathematica for Scientists and Engineers*, Addison-Wesley, reading, MA, 1995.

[5] E. Martin, Ed., *Mathematica 3.0 Standard Add-on Packages*, Wolfram Media, Cambridge University Press, Champaign, IL, 1996.

There are standard catalogs of celestial objects, and the one used in Section 1.5 was

[6] D. Hoffleit, *The Fourth Revised Edition of The Bright Star Catalogue*, Yale University Observatory, 1982.

One of the successful nonlinear models of mass clustering in the universe is discussed in

[7] S.F. Shandarin, Three-dimensional Burgers' equation as a model for the large-scale structure formation in the universe, *Stochastic Models*

in Geosystems, S.A. Molchanov and W.A. Woyczynski, Eds., Springer-Verlag, New York, 1997.

[8] W.A. Woyczynski, *Göttingen Lectures on Burgers Turbulence*, Springer-Verlag, New York, 1998.

The following are good surveys in their respective areas:

[9] C.H. Bennet, G. Brassard, and A.K. Ekert, Quantum Cryptography, *Scientific American*, October 1992, pp. 50-57.

[10] M.F. Barnsley and L.P. Hurd, *Fractal Image Compression*, AKPeters Ltd., Wellesley, MA. 1993.)

[11] H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods*, SIAM, Philadelphia, 1992.

[12] J.M. Hammersley and D.C. Hanscomb, *Monte Carlo Methods*, Chapman and Hall, London, 1964.

The last position, although dated, still reads very well.

If you want to read an amusing but also informative account of the perils of (ab)using statistics, see the classic

[13] D. Huff, *How to Lie with Statistics*, Norton, New York, 1954,

which became an honest-to-goodness bestseller with over half-a-million copies sold.

Chapter 2

Data Representation and Compression

The principal question addressed in this chapter is how to present in a readable fashion (often) large sets of data, extracting their essential features in a compact and digestible form which, for instance, would permit an easy comparison of different data sets, discern trends, facilitate management and engineering decisions, or predict future behavior. This is what we call the problem of data representation and compression.

2.1 Data types, categorical data

Data compression can take different forms, such as graphical representation (bar charts, histograms, etc.), condensing the information to a single number, or a few numbers, characterizing some features of the data (say, median, mean, variance), or an analytic or algorithmic representations discussed in the next chapter and used in modeling and in, for example, fractal image compression techniques.

The material is organized depending on the type of data:

- categorical data
- numerical data
  multidimensional data
- fractal data

However, one has to recognize that many of our examples involve data that are of mixed nature, say, both categorical and numerical, and that the above classification is intended to systematize for the reader the tools that are available for representation and compression in each grouping. Each of these categories can also display a time dependence (i.e., *trends*), and the associated *time series* need to be represented as well.

Also, data can be collected in different ways and, depending on the data collection method, different statistical techniques have to be utilized for their analysis. This consideration gives rise to another classification into

M. Denker, W. Woyczynski, *Introductory Statistics and Random Phenomena*,
Modern Birkhäuser Classics, DOI 10.1007/978-3-319-66152-0_2

systematically collected data, and
- random samples.

The first type, like the number of prisoners tallied by state and year in Section 1.6, comes from a complete and systematic collection of information about different categories. The legitimate question is how to represent them best but their compression always involves a loss of information. Calculation of the average number of prisoners per 1,000 of population in the Midwest is an example of such a compression.

On the other hand, the random samples, usually numerical, can be meaningfully compressed without any loss of information. For example, their relative frequency distributions or cumulative distribution functions contain all the information about them.

Many popular publications, such as *U.S.A. Today*, the *Economist*, the *Scientific American*, have long recognized that data displays and related graphics have a decorative value, especially if full color is included. Such an "artistic" approach may or may not improve the transmission of information to the reader, but has to be considered seriously as a way to improve data presentation.

The treatment of *categorical data*, that is data in which each observation of the sample belongs to one of the finite number of categories, has to be, naturally, quite different than the treatment of numerical data. The positive or negative outcome of a medical treatment in a group of patients can be tabulated depending on their blood type which can be one of the four types: O, A, B, or AB. The result is categorical data that record the number of patients in each of the four categories who positively responded to the treatment. Data on positions of bright stars in Example 1.5.1 were numerical as far as the magnitude, right ascensions, and declinations were concerned, but categorical as far as their spectral class was concerned. The data on rates of sentenced prisoners in Section 1.6 were categorized by state, but for each state they formed a time series as a function of the year.

Further subclasses of categorical data can be distinguished such as

- nominal data
- ordinal data

For *nominal categorical data*, different categories are assigned different numbers in an arbitrary fashion. There is no mathematical relationship between those numbers which could be interpreted in terms of the characteristic properties of each category. Examples of such nominal data are car license plate numbers, blood types, gender, and color.

If $\alpha, \beta, \dots, \omega$ are categories in our data set of size n, and we assign arbitrarily numbers $1, 2, \dots, m$ to these categories, no algebraic or order-related manipulation of these data would be meaningful. However, computation of frequencies makes perfect sense. If $n_1, n_2, \dots, n_m$ are, respectively, numbers of data points in each category $\alpha, \beta, \dots, \omega$, then the corresponding relative frequencies are

$$p_\alpha = \frac{n_1}{n}, p_\beta = \frac{n_2}{n}, \dots, p_\omega = \frac{n_m}{n},$$

and we can represent them graphically by a bar chart or a pie chart.

Ordinal categorical data is a further subclass of nominal categorical data in which different categories are assigned different numbers (ranks) the ordering thereof has a meaningful interpretation in terms of the characteristic property of each category. Examples of such ordinal data are the degree of response to a new drug treatment (low, medium, high), students' grades, social ranking, and weekly NCAA basketball and football rankings.

In this case, in addition to the frequency tabulation, bar charts, and pie charts available for general nominal data, we can also perform order-related manipulations such us finding the j-th smallest category, sort the data according to their rankings, or even to split them into quantilelike subgroups which are described in detail in Section 2.2.

In this section we show how the categorical (or mixed) data can be represented graphically. More and much richer information of the subject can be found in the sources quoted in the Bibliographical Notes at the end of this chapter.

Example 2.1.1 Telephone Charges.
The cost of three-minute international calls in various countries was surveyed by the National Utility Services in February 1995. The results appeared in the *Economist*, March 25, 1995, in the form of two bar charts presented in Figure 2.1.1. Note the

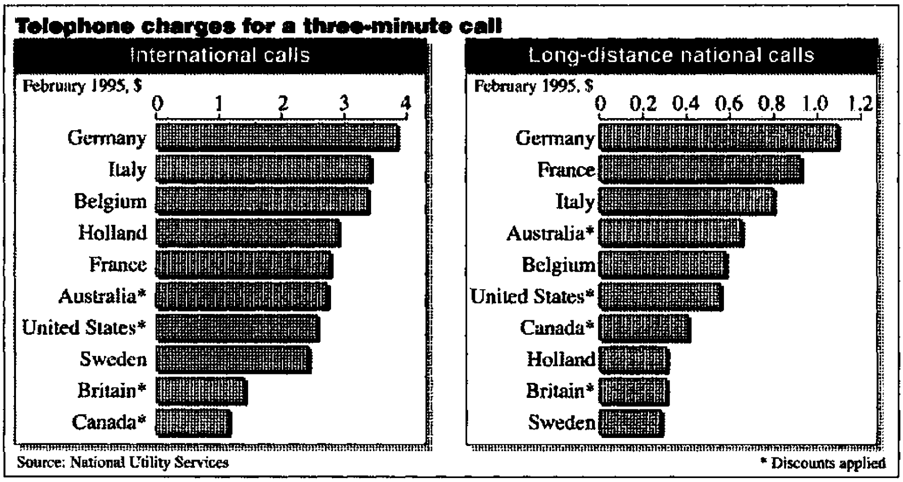


FIGURE 2.1.1
International and long-distance national telephone charges for a three-minute call. (From the Economist, March 25, 1995.)

horizontal position of the bars; it is much easier to label them than the vertical bars. Also, other pertinent data such as the source of the information, the date, and the units are meticulously (but unobtrusively) displayed. For a clearer understanding of relationships of charges in different countries, the corresponding bars were arranged in the decreasing order of magnitude. The data were collected systematically and they contain a categorical (country, long-distance national vs.

international) and numerical (rates) components. They do not form a random sample although it is quite obvious that the numbers came via some kind of compression of a random sample of, say, long distance calls to selected locations. One can think of countries as being ordinal data if you insist on putting them in alphabetical order.

Example 2.1.2 Japanese Technical Citations.

The CWRU Mathematics Department has a foreign language reading requirement for its Ph.D. students. French, German, and Russian have been ruled to be acceptable "major foreign languages". Barbara Margolius, a graduate student who studied Japanese for a number of years, seeking a waiver of departmental rules, presented data showing the growing trend of Japanese language citations in the mathematical technical literature, and asked that the department accept Japanese as a "major foreign language". The trend, shown in Fig. 2.1.2, shows the ratio of Japanese citations to the German, Russian, and French citations. Note that the length of time intervals over which the data were aggregated varied, reflecting different sample sizes in different time periods. This is an example of an effective presentation of aggregate time-dependent categorical data. Needless to say, Barbara's request was granted. She collected the data by a systematic and exhaustive computer search of the *MathSci Index*. They are categorical (nominal) by country, categorical (ordinal, trend displaying) by year groupings, and numerical (proportional) in terms of relative citation numbers.

Different hatchings of bars corresponding to different categories are not always the most fortunate graphical technique to differentiate categorical data, as they often produce unwanted Moiré effects. Different grayscale levels are preferable.

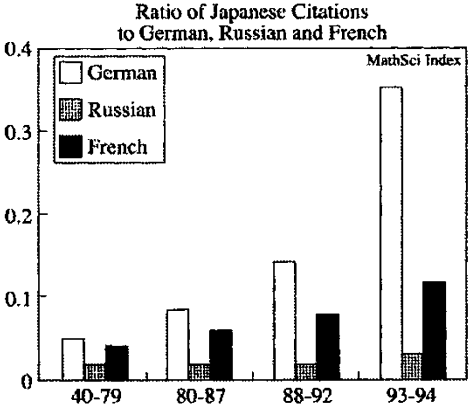


FIGURE 2.1.2
The trend in the ratio of Japanese citations in mathematical literature, to German, Russian, and French citations.

Example 2.1.3 Emergency Calls.

The times of 39,939 emergency calls to the 6th District of the Cleveland Police Department were recorded from Dec 28, 1993 to July 31, 1994, and stored as computerized data in each of the four priority categories. Priority 1 (the highest) calls are the calls requiring immediate response (felony assault in progress, etc.). Priority 4 calls are the lowest priority calls (abandoned vehicle, blocked driveway, etc.). Fig. 2.1.3 shows the time dependence of the average (over the whole data set) number of calls per hour (intensity of call arrivals) over the 24-hour time period for each of the four priorities but in a cumulative (stacked up) fashion, starting with Priority 1 at the bottom of the graph. This continuous (in reality the time step was discrete and equal to 1/60 h) *multivariate time series* was obtained by calculating the average of the number of calls for each priority within the moving time-window frame of size equal to 1 hour. The data were collected systematically and they contain a categorical ordinal (priority) component, and the numerical component (call intensity). The data also display time dependence. Some compression was already done (averaging) and resulted in loss of information but improved the clarity of the representation.

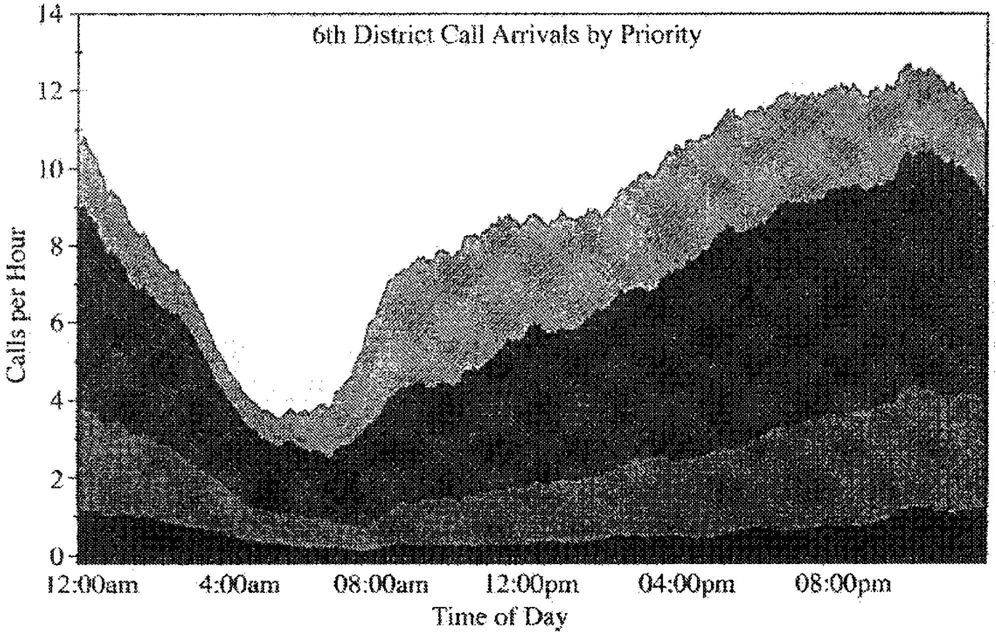


FIGURE 2.1.3

Police emergency call arrivals' intensity. (From B. Margolius, Time-Dependent Multiserver and Priority Queues, Ph.D. Dissertation, CWRU, Cleveland, 1996.

Our final example shows an interesting and surprising property of the decimal expansion of the number π.

Example 2.1.4 Random Number Generation: To Pi or Not to Pi.

Mariya Tikhunova, a Computer Engineering junior at CWRU, investigated a possibility of using the decimal expansion of the number π as a random number generator. The *Mathematica* command `N[Pi,10000]` produced a string of 10,001 digits

$$3.14159265358979323846264338327950288419716939937511\ldots37568,$$

and the starting point of the project (see, Section 4.6 for further developments) was to check the equipartition property (frequencies) for 10 single digits, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, and for 100 pairs of digits, 00, 01, ..., 98, 99. These frequencies are shown in Fig. 2.1.4, where they are compared with the results of similar experiments for the decimal expansion of the number $e = 2.7182818284\ldots$, and 10,000 pseudorandom digits generated in C++.

Remark 2.1.1 *Psychology of Graphical Representation.* The human perception of relations of quantities represented by different graphical techniques varies. Thus, the pie charts (not to mention the distortion-prone but popular 3-D pie charts) convey the relation between quantities less accurately than bar charts. Empirical studies (W.S. Cleveland and R. McGill, Graphical perception; Theory, experimentation, and application to the development of graphical methods, *J. Amer. Stat. Asso.* 79, 531-554,1984) have demonstrated that the absolute error in judging percentage differences in two slices of a pie was greater than in judging differences in two bars. This could be related to the fact that the perception of the area size is not linear but scales as the actual area raised to an exponent of about 0.8 (you can try to determine it yourself by polling your class about relative area sizes of two irregular shapes that you have measured in advance). The area perception dominates in the pie chart, whereas in a bar chart it is principally the bar length that dominates the perception.

Mathematica Experiment 1. Manipulation of Categorical Data: Party Allegiance. A random sample of 39 voters was asked about their political preferences: Democratic, Republican, and Perotistas. Their responses are included in the file `VOTERS` on the `UVW Web Site`. This is a purely categorical and nominal set of data. We compute the frequency of each party's supporters and represent the data as a pie chart.

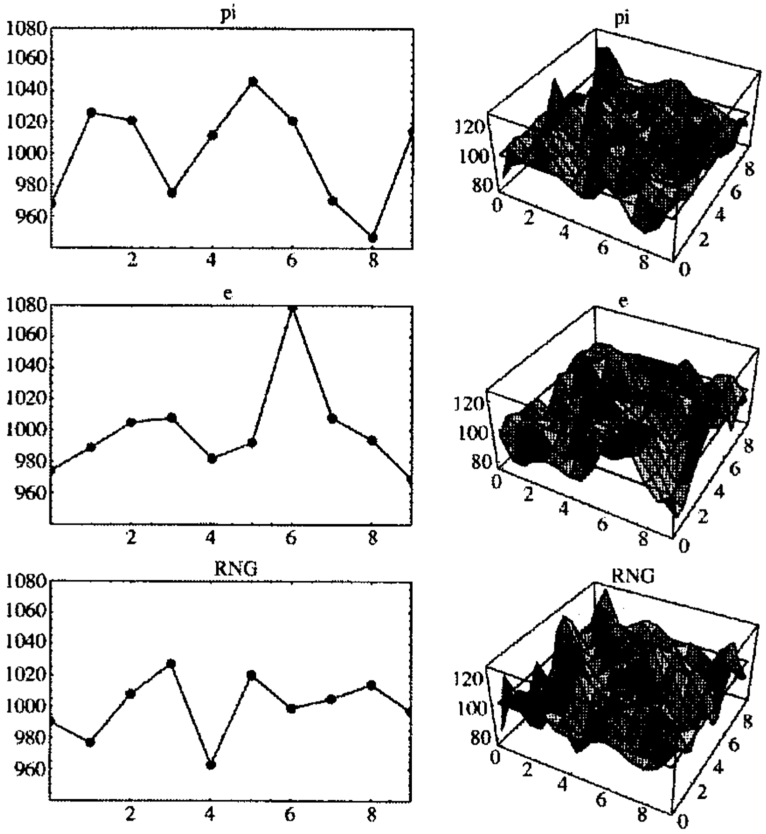
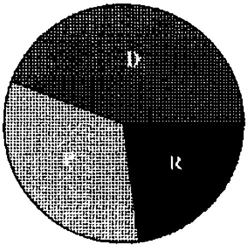


FIGURE 2.1.4

Frequencies of single digits and pairs of digits in the first 10,000 decimals of π (top), e (middle), and a C++ generated random number (bottom).

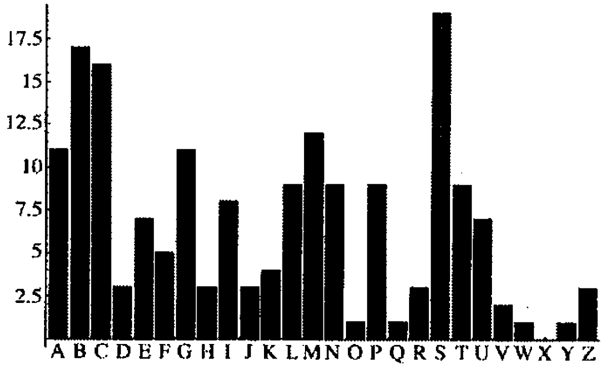
```
In[1]:= <<Statistics'DataManipulation'
In[2]:= <<Graphics'
In[3]:= voters= {{1,D},{2,D},{3,R},{4,P},{5,D},{6,D},{7,R},
    {8,D},{9,R},{10,P}, {11,D},{12,D},{13,R},{14,D},{15,P},
    {16,P},{17,P},{18,P},{19,R},{20,D},{21,P},{22,D},{23,R},
    {24,R},{25,P},{26,P},{27,P},{28,D},{29,R},{30,P},{31,D},
    {32,D},{33,D},{34,P},{35,D},{36,D},{37,P},{38,R},{39,D}}
Out[3]=
   {{1,D},{2,D},{3,R},{4,P},{5,D},{6,D},{7,R},{8, D},{9,R},
   {10,P},{11,D},{12,D},{13,R},{14,D},{15,P},{16,P},{17,P},
   {18,P},{19,R},{20,D},{21,P},{22,D},{23,R},{24,R},{25,P},
   {26,P},{27,P},{28,D},{29,R},{30,P},{31,D},{32,D},{33,D},
   {34,P},{35,D},{36,D},{37,P},{38,R},{39,D}}
In[4]:= Frequencies[Column[voters,2]]
```

```
Out[4]= {{17,D},{13,P},{9,R}}
In[5]:= Length[voters]
Out[5]= 39
In[6]:= PieChart[%4]
Out[6]:= Graphics
```



Mathematica Experiment 2. Countries of the World. The file `World` in the *Mathematica* package `WorldPlot` contains names of 174 countries. We consider this data set as categorical assigning each country to 1 of the 26 categories A,B,. . .,Z depending on its initial. The percentage of countries in each category is then represented by a bar chart. The pie chart would not be legible and informative in this case because the number of categories is too large.

```
In[1]:= <<Miscellaneous'WorldPlot'
In[2]:= <<Graphics'Graphics'
In[3]:= World
Out[3]= {Afghanistan, Albania, . . . . . . . . . ,Zambia,Zimbabwe}
In[4]:= Length[World]
Out[4]= 174
In[5]:= Map[FromCharacterCode,Range[97,122]]
Out[5]= {a,b,c,d,e,f,g,h,i,j,k,l,m,n,o,p,q,r,s,t,u,v,w,x,y,z}
In[6]:= Map[FromCharacterCode, Range[65,90]]
Out[6]= {A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z}
In[7]= s[x_]:=Sum[If[StringMatchQ[StringTake[World[[i]],1],
           FromCharacterCode[x]],1,0],{i,1,174}]
In[8]:= tabl=Table[{s[x],FromCharacterCode[x]},{x,65,90}]
Out[8]= {{11,A},{17,B},{16,C},{3,D},{7,E},{5,F},{11,G},{3,H},
        {8,I},{3,J},{4,K},{9,L},{12,M},{9,N},{1,O},{9,P},{1,Q},
        {3,R},{19,S},{6,T},{7,U},{2,V},{1,W},{0,X},{1,Y},{3,Z}}
In[9]:= BarChart[tabl]
Out[9]= -Graphics-
```



2.2 Numerical data: order statistics, median, quantiles

In this section we assume that the data

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \tag{1}$$

are numerical, that is form a finite sequence (vector) of real numbers called *sample points*, and in our examples they will come often from random samples. The positive integer n is called the *sample size*. We will often use the bold face $\boldsymbol{x}$ to denote sample (1).

The simplest operation that introduces some organization into numerical data is reordering the sample in the increasing order of sample points. In other words, if (1) is the original sample, then there exists a permutation

$$\pi(1), \pi(2), \ldots, \pi(n),$$

of indices

$$1, 2, \ldots, n,$$

such that

$$x_{\pi(1)} \leq x_{\pi(2)} \leq \cdots \leq x_{\pi(n)}.$$

The ordered sample points

$$x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(n)}, \tag{2}$$

are called the *order statistics* (first through n-th) of the sample (1), and traditionally denoted by

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}.$$

The permutation π is usually not mentioned explicitly.

Example 2.2.1 A Die is Cast.
A die was rolled eight times and the resulting sample was

$$x_1 = 1, x_2 = 3, x_3 = 2, x_4 = 5, x_5 = 2, x_6 = 6, x_7 = 5, x_8 = 5.$$

The ordered sample was then

$$x_{(1)} = 1, x_{(2)} = 2, x_{(3)} = 2, x_{(4)} = 3, x_{(5)} = 5, x_{(6)} = 5, x_{(7)} = 5, x_{(9)} = 6.$$

Once the sample has been reordered, it is relatively easy to determine a number of useful and informative numerical characteristics of the sample. For example, the first order statistic

$$x_{(1)} = \min_{1 \le i \le n} x_i,$$

and the n-th order statistic

$$x_{(n)} = \max_{1 \le i \le n} x_i,$$

are, respectively, the smallest and the largest sample points. The interval

$$[x_{(1)}, x_{(n)}] = [\min_{1 \le i \le n} x_i, \max_{1 \le i \le n} x_i]$$

is called the *sample interval*, and its length

$$\operatorname{rng}(\boldsymbol{x}) = x_{(n)} - x_{(1)} = \max_{1 \le i \le n} x_i - \min_{1 \le i \le n} x_i,$$

gives the *sample range*. Thus, in Example 2.2.1, the sample interval is $[1, 6]$ and the sample range is 5.

The sample point that is located in the middle of the reordered sample, or the middle order statistic of the sample, is called the *sample median*. More precisely, the median $\operatorname{med}(\boldsymbol{x})$ of sample $\boldsymbol{x}$ is determined by the condition

$$\#\{i : x_i \le \operatorname{med}(\boldsymbol{x})\} = \#\{i : x_i \ge \operatorname{med}(\boldsymbol{x})\}. \tag{3}$$

Recall, that the notation $\#A$ means the number of elements of the set A, so that $\#\{i : x_i \le \operatorname{med}(\boldsymbol{x})\}$ reads: the number of indices i for which $x_i \le \operatorname{med}(\boldsymbol{x})$. Hence, if the sample size n is odd, the median is exactly the middle element in the reordered sample, that is,

$$\operatorname{med}(\boldsymbol{x}) = x_{((n+1)/2)}.$$

However, when the sample size n is even, any number between $x_{(n/2)}$ and $x_{((n/2)+1)}$ satisfies condition (3), so that median is not uniquely defined (unless, of course, $x_{(n/2)} = x_{((n/2)+1)}$). Traditionally, one chooses the midpoint between two middle elements, so that, for even-sized samples,

$$\text{med}\,(\boldsymbol{x}) = \frac{1}{2}(x_{(n/2)} + x_{((n/2)+1)}).$$

In a similar spirit one could define three *quartiles* Q_1, Q_2, Q_3 of sample $\boldsymbol{x}$ as numbers that divide the ordered sample into four groups with the same number of elements. In other words,

$$\#\{i : x_i \le Q_1\} = \#\{i : Q_1 \le x_i \le Q_2\}$$

$$= \#\{i : Q_2 \le x_i \le Q_3\} = \#\{i : Q_3 \le x_i\}. \qquad (4)$$

Obviously the second quartile is just the median:

$$Q_2 = \text{med}\,(\boldsymbol{x}),$$

and the first and third quartiles can be obtained by finding the medians of the left and right halves of the ordered data. In a similar fashion, *percentiles* would then divide the sample into 100 equal groups.

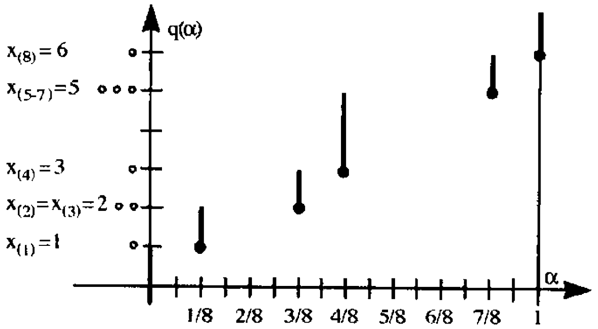


FIGURE 2.2.1

The graph of the multi-valued quantile function $q(\alpha)$ as defined by (5), for the set of eight data points from Example 2.2.1 marked as circles to the left of the vertical axis. The function is well defined only for $\alpha = 0, 1/8, 3/8, 4/8, 7/8, 1$. The bottom dots on each vertical bar indicate a unique selection of the version of quantile corresponding to formula (6).

Quartiles and percentiles are special cases of the general concept of a *quantile*, but defining quantiles $q(\alpha)$ for an arbitrary $0 < \alpha < 1$ (rather than, say, $\alpha = 1/4, 1/2$, and $3/4$, as we have done for quartiles) and a finite sample $x_1, \ldots, x_n$, of size n, is a little more tricky. If the real number α is of the form $\alpha = \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, \frac{n}{n} = 1$, then one can define the *α-quantile* as a number $q = q(\alpha)$ such that

$$q\left(\frac{k}{n}\right) = x_{(k)}. \tag{5}$$

Like the median, the α-quantile $q(\alpha)$ is a *multivalued function*, that is, there can be many acceptable values of $q(\alpha)$ for each α. The typical situation is displayed in Fig. 2.2.1, where the dot-plot of eight data points from Example 2.2.1 is marked on the vertical axis together with the labeling of the five possible values $v_1, \ldots, v_5$ that the data can assume.

The ambiguity embedded in the above definition of the quantile can be avoided if one specific realization of the quantile multivalued function is selected. For example, one can uniquely define

$$q(\alpha) = \min q, \tag{6}$$

where the minimum is taken over all qs satisfying the defining condition (5). Such a selection is marked by dots at the bottom of vertical bars on Fig. 2.2.1.

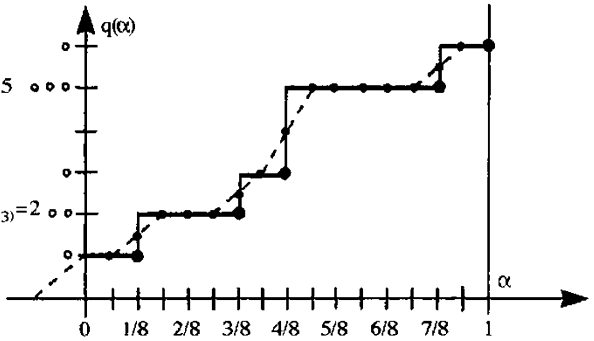


FIGURE 2.2.2

The plot of the piecewise constant extension of the quantile function $q(\alpha)$ (continuous line), and the plot of the linearly interpolated extension corresponding to formula (7) (dotted line). The data are those of Example 2.2.1.

Even if one selects a unique version of $q(\alpha)$ for $\alpha = \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, \frac{n}{n} = 1$, there is no unique way to extend the definition of $q(\alpha)$ for other $0 < \alpha < 1$. Two common choices are:

Piecewise-constant extension: In this case, for any $0 < \alpha < 1$, we select as $q(\alpha)$ the value $q(k/n)$ from formula (5) for k/n immediately to the right of the number α. For data from Example 2.2.1, this selection of the quantile function is shown in Fig. 2.2.2 by the continuous line.

Linearly interpolated extension: In practice, to avoid any ambiguity, often one defines the quantile $q(\alpha)$ for any $< \alpha < 1$ via the interpolation formula

$$q_{\text{int}}(\alpha) = x_{(\lfloor n\alpha+1/2\rfloor)} + \Big(n\alpha + 1/2 - \lfloor n\alpha + 1/2\rfloor\Big)\Big(x_{(\lfloor n\alpha+1/2\rfloor+1)} - x_{(\lfloor n\alpha+1/2\rfloor)}\Big). \tag{7}$$

Recall, that the symbol $\lfloor x \rfloor$ denotes the *floor* or the *integer part* of the number x, i.e., the largest integer $\leq x$. The number $q(\alpha)$ cuts then, approximately, the sample into two parts; $\alpha \cdot 100\%$ of sample points are below $q(\alpha)$ and $(1-\alpha)\cdot 100\%$ of sample points are above $q(\alpha)$. In particular, $q(0.5)$ is the sample median, and $q(0.25)$, $q(0.75)$ correspond to the sample first and third quartiles. For data from Example 2.2.1, the interpolated quantile function is marked in Fig. 2.2.2 by the dotted line.

Notice that the quantile function $q(\alpha)$ defined by (6) completely determines the (ordered) data set. Once it is known, one has the complete information about what are the possible values in the data set and how many times each of these values appears in the data set.

Mathematica Experiment 1. A Die is Cast. The piecewise-constant extension is taken as a definition of quantiles in *Mathematica* under command `Quantile[data, α]`. *Mathematica* also provides a command `InterpolatedQuantile [data, α]` which computes linearly interpolated quantiles. So, for data from Example 2.2.1, we can proceed as follows:

```
In[1]:= <<Statistics'DescriptiveStatistics'
In[2]:= data= {1,3,2,5,2,6,5,5}
Out[2]= {1,3,2,5,2,6,5,5}
In[3]:= Quantile[data,0.25]
Out[3]= 2
In[4]:= Median[data]
Out[4]= 4
In[5]:= Quantile[data,0.75]
Out[5]= 5
In[6]:= Quantile[data,0.33]
Out[6]= 2
In[7]:= InterpolatedQuantile[data,0.33]
Out[7]= 2.14
```

In practice, one often summarizes the quantile characteristics graphically in the form of the *box plots* (or, *box-and whiskers plots*) introduced by John Tukey (see Bibliographical Notes). We present a typical application in the next example. A package creating a box-and-whiskers plot is included on the `UVW Web Site`.

Example 2.2.2 Salaries of New Ph.D.s.
The American Mathematical Society annually publishes a salary survey for new recipients of doctorates in the mathematical sciences (includes mathematics, applied mathematics, and statistics) who took positions in teaching, research, government, or business and industry. The results of the 1992-1993 survey for those employed in business and industry are summarized in Fig. 2.2.3.

Twelve-Month Salaries

Ph.D. Year	Min	Q_1	Median	Q_3	Max	Reported Median in 1992 $
	BUSINESS AND INDUSTRY (33 men + 10 women)					
1960	78		110		150	512
1965	100		136		180	580
1970	96		170		235	685
1975	114		187		240	460
1980	190		284		400	480
1985	260	360	400	420	493	513
1990	320	438	495	533	700	529
1991	235	480	510	573	830	525
1992	208	450	530	620	1000	530
1993	270	480	560	600	1100	—
1990M	320	443	490	533	630	
1990F	390	440	500	525	700	
1991M	330	500	520	587	830	
1991F	235	420	481	554	720	
1992M	300	440	520	625	1000	
1992F	208	528	549	591	850	
1993M	270	500	560	600	1100	
1993F	424	475	568	600	670	
One year or less experience (17 men + 7 women)						
1993M	270	480	543	600	700	
1993F	424	458	584	595	600	

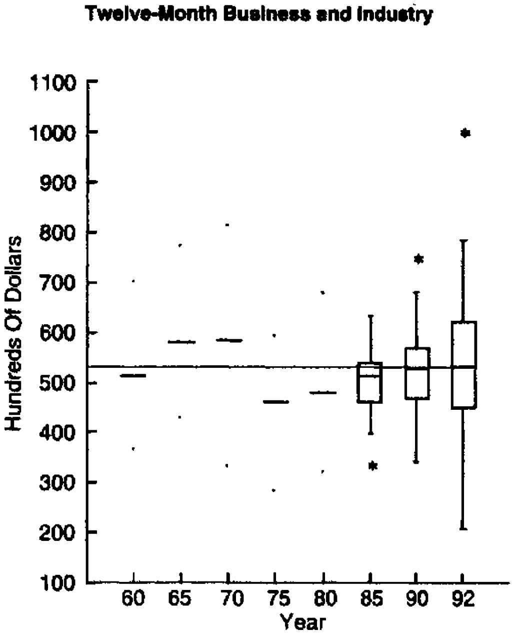


FIGURE 2.2.3
Starting business and industry, 12-month salaries of mathematical sciences Ph.Ds. (From A.M.S. Notices 40 (9), 1993.)

The table on the left summarizes the actual numerical values of the characteristics, while the graph on the right shows a variant of the box plot with inflation-adjusted data expressed in (hundreds of) 1992 dollars, using the price deflator published annually by the Bureau of Economic Analysis, U.S. Department of Commerce.

The *box-and-whiskers plots* provide the graphical representation of the same summarized data. The horizontal line shows the 1992 median salary. For a given year, the box incorporates the first and third quartiles and the median salary. Prior to 1975 the quartiles were not available and the median is depicted by a horizontal stroke. The "wiskers" give additional information about the spread of data, extending to the *extreme values* that are between zero and 1.5 times the interquartile distance from the edge of the box. The data points that are between 1.5 and 3 times the interquartile distance from the edge of the box are called *outliers*, and those that are beyond three times the interquartile distance from the edge of the box are called *extreme outliers*. Usually, different symbols, like dots and asterisks, are used to identify the two types of outliers.

Q-Q plots. Plotting quantiles $q(\boldsymbol{x}, \alpha)$ and $q(\boldsymbol{y}, \alpha)$ of two different samples $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$ on the $(q(\boldsymbol{x}), q(\boldsymbol{y}))$-plane creates the so-called *Q-Q plot* that can be used to compare their quantile functions, and thus the distributions of their values. To create such a plot one simply marks points with coordinates

$$(q(\boldsymbol{x}, k/n), q(\boldsymbol{y}, k/n)), \quad k = 1, 2, \ldots, n \tag{8}$$

on the $(q(\boldsymbol{x}), q(\boldsymbol{y}))$-plane, selecting an unambiguous definition of the quantiles. An example of the Q-Q plot is shown in Fig. 2.2.4.

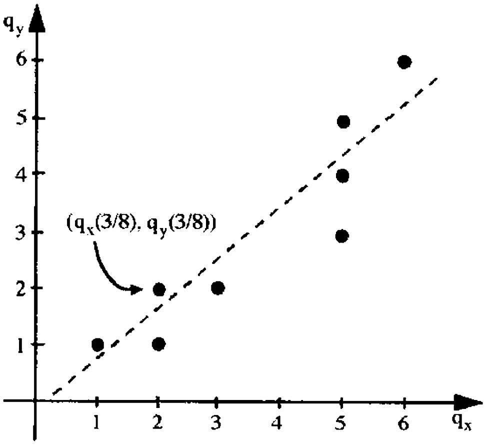


FIGURE 2.2.4

Q-Q plot for the data set from Example 2.2.1 paired with the data set $\boldsymbol{y} = (1, 1, 2, 2, 3, 4, 5, 6)$.

An approximate alignment of points along the straight line $y = ax + b$ is evidence that, up to a linear transformation, the two data sets have identical quantile functions.

If data sets $\boldsymbol{x}$ and $\boldsymbol{y}$ are of different sizes, say n_x and n_y, then one usually picks $n = \min\{n_x, n_y\}$,

$$\alpha = \frac{k - 1/2}{n}, \qquad k = 1, 2, \ldots, n,$$

and one finds the largest value of $q(\boldsymbol{x}) = q(\boldsymbol{x}, (k - 1/2)/n)$ and $q(\boldsymbol{y}) = q(\boldsymbol{y}, (k -$

$1/2)/n$) among $q(\boldsymbol{x})$ and $q(\boldsymbol{y})$ satisfying conditions

$$\frac{\#\{i : x_i \leq q(\boldsymbol{x})\}}{n_x} \leq \frac{k-1/2}{n}, \qquad \frac{\#\{i : y_i \leq q(\boldsymbol{y})\}}{n_y} \leq \frac{k-1/2}{n}, \tag{9}$$

and then one plots points

$$\left(q\left(\boldsymbol{x}, \frac{k-1/2}{n}\right), q\left(\boldsymbol{y}, \frac{k-1/2}{n}\right)\right), \qquad k = 1, 2, \ldots, n, \tag{10}$$

on the Q-Q plot.

2.3 Numerical data: histograms, means, moments

Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ be a sample of size n, where sample points take numerical values. In computing percentiles (or general quantiles) in the previous section we split the *ordered* sample into subsamples of equal (or prescribed) size In this section we take a different approach to summarizing the sample data by counting the number of sample points that take a prescribed value, or fall within given intervals of a partition of the sample range. The results of such a count are then plotted in the form of a *histogram.*

Let us start with the situation where the sample points can take only finitely many values (sample taken from a *discrete set*)

$$v_1, v_2, \ldots, v_N.$$

Then, the *frequency distribution function (d.f.)* $\phi(v)$ of sample $\boldsymbol{x}$ counts how many times each of the possible values v appeared in the samples $\boldsymbol{x}$. Clearly, it can be nonzero only for vs from the allowable set v_i, $i = 1, 2, \ldots, N$. More formally,

$$\phi(v_1) = \#\{i : x_i = v_1\}, \tag{1}$$

$$\phi(v_2) = \#\{i : x_i = v_2\},$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots,$$

$$\phi(v_N) = \#\{i : x_i = v_N\},$$

and

$$\phi(v) = 0, \quad \text{if} \quad v \neq v_i,\ i = 1, \ldots, N.$$

Its plot, often represented in the bar chart form (Fig. 2.3.1) is called the sample histogram.

Mathematica Experiment 1. A Die is Cast. In Example 2.2.1, where the sample points represent the outcome of rolling a die, it is natural to take $N = 6$ and $v_1 = 1, \ldots, v_6 = 6$. The frequencies of data are computed by the `Frequencies[data]` command and the bar chart of those frequencies can be produced by using the `BarChart[Frequencies[data]]` command. Both are within the `Statistics'DataManipulation'` and `'Graphics'` package.

```
In[1]:= <<Statistics'DataManipulation'
In[2]:= <<Graphics'Graphics'
In[3]:= data={1,3,2,5,2,6,5,5}
Out[3]= {1, 3, 2, 5, 2, 6, 5, 5}
In[4]:= Frequencies[data]
Out[4]= {{1, 1}, {2, 2}, {1, 3}, {3, 5}, {1, 6}}
In[5]:= Insert[%, {0,4},4]
Out[5]= {{1, 1}, {2, 2}, {1, 3}, {0,4}, {3, 5}, {1, 6}}
In[6]:= BarChart[%]
Out[6]= -Graphics-
```

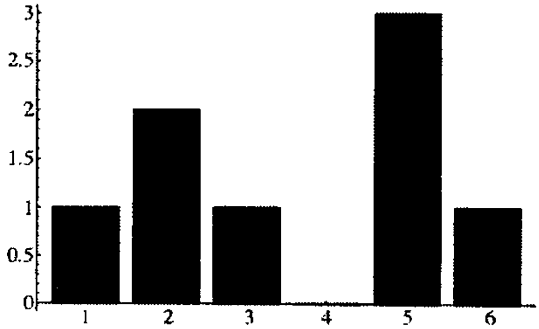


FIGURE 2.3.1

The frequency distribution function ϕ, normalized by the sample size n,

$$f(v) = \frac{1}{n}\phi(v), \tag{2}$$

is called the *relative frequency distribution function* and is often more convenient to use. Its plot is called the *normalized histogram.* Advantage of the normalization is that, of course,

$$\sum_{j=1}^{N} f(v_j) = 1. \tag{3}$$

If the relative frequency d.f. $f(x)$ of a sample is known, then a number of numerical characteristics of the sample, compressing the information contained in a sample to a single number, are easily calculated. In particular, the fundamental *location characteristic*, the *sample mean*

$$\bar{\boldsymbol{x}} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i = \sum_{j=1}^{N} v_j f(v_j), \tag{4}$$

which is simply an arithmetic average of sample points, but a *weighted average* of possible values v, with the relative frequency d.f. $f(v)$ providing the weights.

The sample mean provides only the coarsest information about the location of the sample and none about its spread, or dispersion, around the mean value. To measure the latter, the first impulse would be to look at the *sample deviations from the mean*

$$x_1 - \bar{\boldsymbol{x}},\ x_2 - \bar{\boldsymbol{x}},\ \ldots,\ x_n - \bar{\boldsymbol{x}}, \tag{5}$$

and compute their mean, to get a single dispersion parameter. However, a simple calculation shows that the mean of the sample deviations from the mean is always zero. Hence, this quantity is not a suitable measure of dispersion.

A better idea is to look at the *mean absolute deviation* (mean distance) of sample points from the mean

$$\frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}| = 0. \tag{6}$$

This turns out to be a respectable choice, but analytical calculations with absolute values are notoriously unpleasant, and that is why this is not the first choice of statisticians. Traditionally, they measure the spread of the sample points around the mean by computing the *theoretical sample variance*

$$\text{var}\,(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{\boldsymbol{x}})^2, \tag{7}$$

that is, the mean of the squares of deviations of sample points from the sample mean. The sample variance is easily computable if the relative frequency d.f. f is given. Indeed,

$$\text{var}\,(\boldsymbol{x}) = \sum_{j=1}^{N}(v_j - \bar{\boldsymbol{x}})^2 f(v_j). \tag{8}$$

Also, notice that the following formula

$$\text{var}\,(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{\boldsymbol{x}} + \bar{\boldsymbol{x}}^2) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 = \overline{\boldsymbol{x}^2} - \bar{\boldsymbol{x}}^2, \tag{9}$$

is more economical computationally than the original definition of the variance (why?). It expresses the sample variance var $(\boldsymbol{x})$ in terms of the sample mean $\bar{\boldsymbol{x}}$ and the *sample second moment*

$$m_2(\boldsymbol{x}) = \overline{\boldsymbol{x}^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2. \tag{10}$$

Higher, *k-th sample moments*

$$m_k(\boldsymbol{x}) = \overline{\boldsymbol{x}^k} = \frac{1}{n}\sum_{i=1}^{n} x_i^k, \tag{11}$$

are defined in a similar fashion, for any positive integer k.

Remark 2.3.1 *Scaling Properties of Mean and Variance.* The sample mean *scales* linearly, that is if the sample $\boldsymbol{x}$ is rescaled by a numerical factor a:

$$a\boldsymbol{x} = (ax_1, ax_2, \ldots, ax_n),$$

then

$$\overline{a\boldsymbol{x}} = \frac{ax_1 + \ldots + ax_n}{n} = a \cdot \bar{\boldsymbol{x}}. \tag{12}$$

However, the theoretical sample variance does not scale linearly with the magnitude of the sample points because

$$\text{var}\,(a\boldsymbol{x}) = a^2 \text{var}\,(\boldsymbol{x}). \tag{13}$$

To remedy this flaw one often considers the *theoretical standard deviation* of the sample

$$\text{std}\,(\boldsymbol{x}) = \sqrt{\text{var}\,(\boldsymbol{x})} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{\boldsymbol{x}})^2} \tag{14}$$

of the sample, which is the square root of the sample variance, and which grows linearly with the growth of the sample points' amplitude. Indeed,

$$\text{std}\,(a\boldsymbol{x}) = |a|\,\text{std}\,(\boldsymbol{x}). \tag{15}$$

Mathematica Experiment 1. A Die is Cast. In *Mathematica* the package `Statistics‘DescriptiveStatistics‘` contains all the needed commands. Thus, the sample mean is obtained via Mean[data], the theoretical sample variance by VarianceMLE[data]. The MLE stands for "maximum likelihood estimate" and the command Variance[data] is reserved for the *unbiased sample variance*

$$s^2(\boldsymbol{x}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \boldsymbol{x})^2. \tag{16}$$

The significance of the unbiased sample variance will be explained later on. The command MeanDeviation[data] stands for the mean absolute deviation, and the RootMeanSquare[data] —for the root mean square of the sample, i.e., $\sqrt{m_2(\boldsymbol{x})}$.

```
In[1]:= <<Statistics'DescriptiveStatistics'
In[2]:= data={1,3,2,5,2,6,5,5}
Out[2]= {1, 3, 2, 5, 2, 6, 5, 5}
In[4]:= Mean[data]
Out[4]= 29/8
In[5]:= N[%]
Out[5]= 3.625
In[6]:= Variance[data]
Out[6]= 191/56
In[%]:= N[%]
Out[6]= 3.41071
In[6]:= VarianceMLE[data]
Out[6]= 191/64
In[%]:= N[%]
Out[6]= 2.98437
In[6]:= RootMeanSquare[data]
Out[6]= Sqrt[129/8]
In[%]:= N[%]
Out[6]= 4.01559
```

In the case of a sample drawn from a *continuous set*, or from a very large discrete set, the histogram based on frequencies calculated for each possible value v may turn out to be difficult to interpret because most of these values may appear only once or never. In such a case, a much more informative representation of data is obtained by counting frequencies of sample points falling into *bins* of prescribed size. One has to remember though that some information is lost in the process and that the *binned histogram* provides a coarser description of data than the full frequency d.f.

More precisely, the *binned histogram* is now determined by a partition

$$t_0 < t_1 < t_2 < \ldots < t_p \tag{17}$$

of the sample interval

$$[\min x_i, \ \max x_i],$$

(or, of a larger interval) into p bins

$$B_1 = [t_0, t_1],\; B_2 = (t_1, t_2], \ldots, B_p = (t_{p-1}, t_p], \tag{18}$$

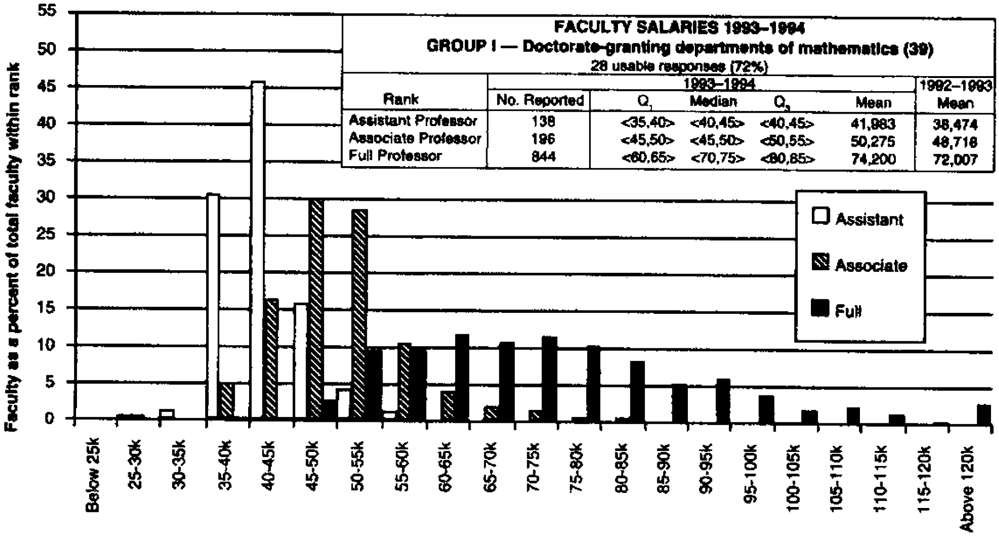
which usually are taken of equal size (step) $(\max x_i - \min x_i)/p$. Then the ***histogram*** of the sample $\boldsymbol{x}$ corresponding to the partition (17) is the graph of the function

$$h(x) = \frac{1}{n}\#\{i : x_i \in B_j\}, \quad for \quad x \in B_j, j = 1, 2, \ldots, p. \tag{19}$$

Outside the union of bins $B_1, \ldots, B_p$, we set $h(x) \equiv 0$. Within a given bin, the histogram function, which is piecewise constant, simply counts the number of sample points that fall within that bin.

Example 2.3.1 Faculty Salaries.

The American Mathematical Society annually gathers data about the faculty salaries in mathematical sciences. The summary of the 1993-1994 faculty salary survey (*AMS Notices*, November 1993) in 39 top mathematical sciences departments is shown in Fig. 2.3.2.



FACULTY SALARIES 1993–1994
GROUP I — Doctorate-granting departments of mathematics (39)
28 usable responses (72%)

Rank	1993–1994					1992–1993
	No. Reported	Q_1	Median	Q_3	Mean	Mean
Assistant Professor	138	<35,40>	<40,45>	<40,45>	41,983	38,474
Associate Professor	196	<45,50>	<45,50>	<50,55>	50,275	48,718
Full Professor	844	<60,65>	<70,75>	<80,85>	74,200	72,007

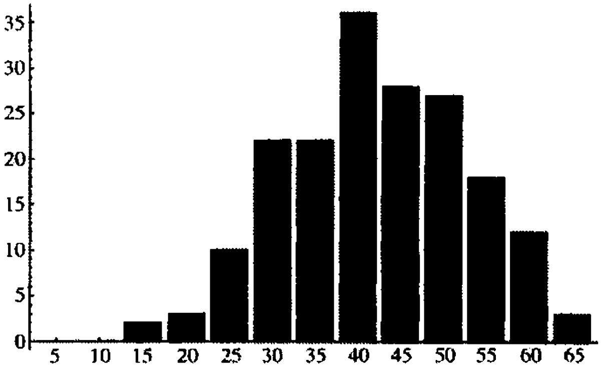
FIGURE 2.3.2

The normalized binned histogram of 1993-94 faculty salaries. The bin sizes were selected to be 5k$. (From A.M.S. Notices, November 1993.)

In addition to the histogram of what is essentially three-dimensional continuous data, the graph also summarizes the quartiles and means in all three faculty categories. The departments were asked to report the number of faculty whose 1993-1994 academic-year salaries fell within given salary intervals. Reporting salary data in this fashion eliminated some of the concerns about confidentiality but did not permit determination of actual quartiles. What could be determined were the salary intervals in which the quartiles occurred; they are denoted by $< a, b >$.

Mathematica Experiment 2. Rivets. This experiment explores different ways of constructing histograms for the data (rivet length measurements in millimeters) contained in the file RIVET which can be found on the UVW Web Site.

```
In[1]:= <<Statistics‘DescriptiveStatistics‘
In[2]:= <<Graphics‘Graphics‘
In[3]:= <<Statistics‘DataManipulation‘
In[4]:= rivet={13.39, 13.43, ... , 13.58, 13.38}
Out[4]= {13.39, 13.43, ... , 13.58, 13.38}
In[5]:= Length[rivet]
Out[5]= 184
In[5]:= freq=BinCounts[rivet,{13.025,13.675, 0.05}]
Out[5]= {0, 0, 2, 3, 10, 22, 22, 36, 28, 27, 18, 12, 3}
In [6]:= midpoints=Table[5+5k, {k,0,12}]
Out[6]= {5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65}
In[7]:= trans= Transpose[{freq,midpoints}]
Out[7]= {{0, 5}, {0, 10}, {2, 15}, {3, 20}, {10, 25}, {22, 30},
         {22, 35}, {36, 40}, {28, 45}, {27,50}, {18, 55},
        {12, 60}, {3, 65}}
In[8]:= BarChart[trans]
Out[8]= -Graphics-
```



Optimizing Histograms. A few of practical pointers on construction of histograms are in order:

(i) The bin boundaries should be selected in such a fashion that no sample points are on these boundaries. One way to achieve it is as follows: if data were collected with four-digit accuracy, say 13.39, 13.43,..........,13.58, 13.38, then pick as bin boundaries points 13.385, 13.395,.......,13.575, 13.585. That is how we proceeded in the *Mathematica* Experiment 1.

(ii) Bin size has to be such that at least 5 to 7 data points fall within each bin. If the bin size is too small, then the histogram values are too random, and create too many false modes (local maxima), to draw any conclusions.

(iii) The number of bins, or the resolution (step, bin size) of the histogram representation, has to be selected for optimum conveying of the information contained in the sample. The *Sturges' rule* says that for a sample of size n the number p of bins should be of the order $p \approx 1 + \log_2 n$. Such a selection is justified by the Stability of Fluctuations Law of Section 3.6; also, see the Bibliographical Notes at the end of this chapter.

2.4 Location, dispersion, and shape parameters

In this section we will return to some of the characteristic parameters introduced in Sections 2.2 and 2.3, introduce some new ones, and provide their comparison from the viewpoint of the type of information they provide.

Location Parameters. The sample mean $\bar{x}$ and the sample median med (x) are obviously the prime location parameters indicating where the sample is centered. They, however, need not coincide. In a sample of family incomes of an urban population, the median can be quite small given that more than half of the people can be quite poor. However, the mean, due to a few billionaires residing in the city, can easily be much higher. In public discussions one can often observe a tendentious selection of the parameters used depending on the agenda of the selector. Several other location parameters are commonly used.

(a) The *mode* is the value in the data set which corresponds to the local maximum of the frequency d.f., or, equivalently, of the histogram. A data set can have several modes. The *principal mode* is a mode that corresponds to the global maximum of the frequency d.f. It need not be unique, either.

(b) As we observed in Section 2.3 (Remark 2.3.1), the sample mean scales linearly and it is always within the *sample interval*, that is

$$\min_{1\le i\le n} x_i \le \bar{x} \le \max_{1\le i\le n} x_i.$$

However, it is not the only function of sample points with the above properties. For example, the *weighted sample mean*,

$$\frac{w_1x_1 + w_2x_2 + \ldots + w_nx_n}{w_1 + w_2 + \ldots + w_n} = \frac{\sum_{i=1}^n w_ix_i}{\sum_{i=1}^n w_i}$$

with different weights

$$w_1, w_2, \ldots, w_n \geq 0$$

assigned to different sample points (say, because of lesser reliability of a part of the data) is another example of such a "mean" location characteristic.

(c) Sometimes, for practical reasons, one may opt for another version of the "mean" called *censored mean*. Suppose we are measuring the lifetimes of light bulbs in a sample of size n. Ideally, one would want to wait until the last light bulb burns out and obtain a complete sample

$$t_1 \leq t_2 \leq \ldots \leq t_n$$

of failure times for the whole light bulb population. However, such an approach may involve an unacceptably long wait, so one often stops the experiment after a certain fixed time T, by which time the first $k < n$ light bulbs fail. This leads to consideration of the censored mean:

$$\frac{1}{n}\left(\sum_{i=1}^k t_i + (n-k)t_k\right).$$

Such a censored mean is useful in reliability studies.

(d) Another location parameter compressing the data is the so-called *harmonic mean* hmean $(\boldsymbol{x})$ satisfying the condition

$$\frac{1}{\text{hmean}\,(\boldsymbol{x})} = \frac{1}{n}\sum_{i=1}^n \frac{1}{x_i}.$$

Remark 2.4.1 *Mean and Median Under One Umbrella.* For more theoretically minded readers, we would like to mention that both mean and median are special cases of the so-called *M-estimators* which are produced from the sample by means of a weight function $\psi(x)$ which is assumed to be increasing. Then the sample ψ-mean is defined as a number $m_\psi(\boldsymbol{x})$ such that

$$\sum_{i=1}^n \psi(m_\psi(\boldsymbol{x}) - x_i) = 0.$$

For $\psi(x) = x$, the ψ-sample mean is the usual sample mean. For $\psi(x) = sgn\,(x)$ (defined as equal to $+1$ for $x > 0$, and -1 for $x < 0$, and 0 at 0), the ψ-mean is the sample median.

Obviously, quartiles, percentiles, and other quantiles can also be considered as more subtle location parameters of the data sets.

Remark 2.4.2 *A General Concept of the Mean.* A number m is said to be a mean of sample $\boldsymbol{x} = (x_1, \ldots, x_n)$ ($x_k > 0$, say) with respect to function $f(x_1, \ldots, x_n)$ if

$$f(x_1, \ldots, x_n) = f(m, \ldots, m).$$

In other words, replacing the sample points by the sample mean does not affect the value of the function. In mechanics of rigid bodies one uses the analogous concept of the barycenter. The system evolves as if the whole mass of the body were concentrated at the barycenter. The usual sample mean corresponds to the selection

$$f(x_1, \ldots, x_n) = (x_1 + \ldots + x_n)/n,$$

the harmonic mean to

$$f(x_1, \ldots, x_n) = [(1/x_1 + \ldots + 1/x_n)/n]^{-1},$$

and the geometric mean to

$$f(x_1, \ldots, x_n) = (x_1 \cdot \ldots \cdot x_n)^{1/n}.$$

A mean is called *associative* if it is not affected by the replacements of some subsets of sample points by their mean. The above three means are all associative. Nagumo and Kolmogorov proved that all associative means are (increasing) transforms of arithmetic weighted means. More precisely, if $m(\boldsymbol{x})$ is such a mean then one can find an increasing function $\gamma(x)$ and the weights $w_1, \ldots, w_k > 0$, $\sum_{k=1}^{n} w_k = 1$, such that

$$m(\boldsymbol{x}) = m(\boldsymbol{x}; \gamma) = \gamma^{-1}\left(\sum_{k=1}^{n} w_k \gamma(x_k)\right).$$

In other words, they are all obtained by changing the scale of sample points via application of function γ, calculating the (weighted) arithmetic average, and then reverting to the original scale by applying the inverse function γ^{-1}. For example, the geometric mean corresponds to $\gamma(x) = \log x$, with $\gamma^{-1}(y) = \exp y$.

The γ-mean is greater than the arithmetic (linear) mean if the function γ is concave upwards. Fig. 2.4.1 suggests the obvious way to prove it.

One can also check that the means increase with the quantity γ''/γ', which measures the local concavity upwards. For power functions $\gamma(x) = x^c$ we have $\gamma''/\gamma' = (c-1)x$, which increases with c. The geometric mean corresponds to the case of $\gamma(x) \approx (x^c - 1)/c,\ c \to 0$. In this fashion one can establish that the various means satisfy the following inequalities:

$$\textit{harmonic} < \textit{geometric} < \textit{arithmetic} < \textit{quadratic} < \textit{cubic} < \dots$$

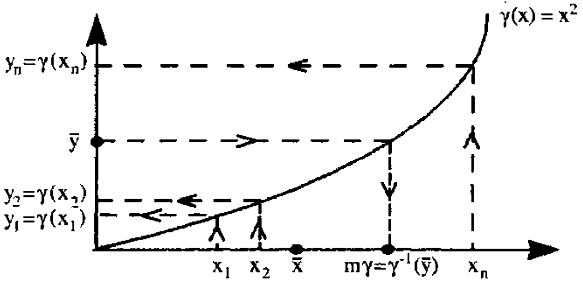


FIGURE 2.4.1
The illustration shows that if $\gamma(x)$ is concave upwards, then the γ-mean $m(\boldsymbol{x}, \gamma)$ is greater than the arithmetic mean $\bar{\boldsymbol{x}}$ (with the same weights).

Dispersion Parameters. The most often used parameters compressing information about the dispersion of sample points are the sample variance var $(\boldsymbol{x})$ (or the unbiased sample variance defined by formula (2.3.16)), the standard deviation std $(\boldsymbol{x})$, and the sample range

$$\operatorname{rng}(\boldsymbol{x}) = x_{(n)} - x_{(1)}.$$

Since all the deviations $|x_i - \bar{x}| \le \operatorname{rng}(\boldsymbol{x})$, we always have that

$$\operatorname{std}(\boldsymbol{x}) \le \operatorname{rng}(\boldsymbol{x})$$

That is, the standard deviation is always estimated from the above by the range. The opposite inequality is, clearly, not valid. However, we have always the following useful and universal

Chebyshev's Law: *At least the fraction $1 - (1/k^2)$ of the sample points are located within k standard deviations σ of the sample mean $\bar{x}$.*

In particular, with $k = 3$, we get that in any sample, at least $8/9 \approx 91\%$ of sample points are located within the interval $[\bar{x} - 3\,\mathrm{std}\,(\boldsymbol{x}), \bar{x} + 3\,\mathrm{std}\,(\boldsymbol{x})]$. For a more rigorous treatment of the Chebyshev's Law, see Theorem 5.4.1.

Another, related, dispersion parameter is the *interquartile distance* $q(.75) - q(.25)$ which already was used in the construction of the box-and-whisker plots. By definition, 50% of sample points are located in the interval $[q(.25), q(.75)]$.

Shape Parameters. Several parameters can give compressed information about the general shape of the data distribution.

The r-th central sample moments

$$cm_r(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r$$

can be used to detect *skewness* in the frequency d.f., which is formally defined as the ratio cm_3/s^3 of the third sample central moment and the third power of unbiased standard deviation. Clearly, for frequency distributions symmetric about their means the skewness parameter is zero, and its size can be viewed as a measure of the asymmetry of the histogram. Another useful parameter is *kurtosis excess* $\mu_4/s^4 - 3$. In general, the fast growth of higher moments indicates that the frequency distribution is *heavy-tailed* and *intermittent*, the latter meaning that there are patchy pockets of high values of the frequency d.f. interspersed with intervals where there are no data.

Mathematica Experiment 1. Rivets. This experiment explores different ways of compressing the data (rivet length measurements are in millimeters) contained in the file `RIVET` which can be found on the `UVW Web Site`. The *Mathematica* command `Mode[data]` finds the principal modes of the `data`.

```
In[1]:= <<Statistics‘DescriptiveStatistics‘
In[2]:= rivet ={13.39, 13.43,...,13.58, 13.58}
Out[2]= {13.39, 13.43,...,13.58, 13.58}
In[3]:= <<Graphics‘Graphics‘
In[4]:= Mean[rivet]
Out[4]= 13.4216
In[5]:= Median[rivet]
Out[5]= 13.42
In[6]:= Mode[rivet]
Out[6]= 13.4
In[7]:= LocationReport[rivet]
Out[7]= Mean ->13.4216, HarmonicMean -> 13.4207, Median -> 13.42}
In[8]:= Quartiles[rivet]
Out[8]= {13.34, 13.42, 13.5}
In[9]:= Quantile[rivet, 0.7]
Out[9]= 13.48
```

```
In[10]:= SampleRange[rivet]
Out[10]= 0.56
In[11]:= Variance[rivet]
Out[11]= 0.0118738
In[12]:= StandardDeviation[rivet]
Out[12]= 0.108967
In[13]:= DispersionReport[rivet]
Out[13]= {Variance ->  0.0118738, StandardDeviation -> 0.108967,
          SampleRange ->0.56, MeanDeviation -> 0.0882053,
          MedianDeviation -> 0.08, QuartileDeviation -> 0.08}
In[14]:= Skewness[rivet]
Out[14]= -0.0656664
In[15]:= KurtosisExcess[rivet]
Out[15]= -0.390325
In[16]:= ShapeReport[rivet]
Out[16]= {Skewness ->-0.0656664, QuartileSkeweness -> 0.,
          KurtosisExcess -> -0.390325}
```

2.5 Probabilities: a frequentist viewpoint

Plotting the relative frequency d.f. for samples from continuous or very large data sets has its drawbacks related to possible intermittency in the data set, selection of bin size and location, loss of information in the process of producing binned histograms, etc. They can be partly circumvented by introduction of the sample *cumulative distribution function*

$$F(x) = F(t; \boldsymbol{x}) = \frac{1}{n}\#\{i : x_i \leq x\}. \tag{1}$$

Notice that it is a nondecreasing function, well defined for all x and positive within the sample interval, independently of the number N and location of the possible values $v_1, \ldots v_N$. For data from Example 2.2.1, the cumulative d.f. is plotted in Fig. 2.5.1.

It is also easy to see that for any data set $\boldsymbol{x} = (x_1, \ldots, x_n)$, the multivalued sample cumulative distribution function $F(x)$ and the sample quantile function $q(\alpha)$ introduced in Section 2.2 are inverses of each other. The inverse function has to be understood here in a generalized sense (for example, as a reflection in the diagonal of the graph of the original function) since neither of the two functions is, in general, one-to-one. In particular, the cumulative d.f. shown in Fig. 2.5.1 is the inverse function of the quantile function (for the same data set) shown on Fig. 2.2.2.

The relative frequency d.f. and the cumulative d.f. are easily computable from

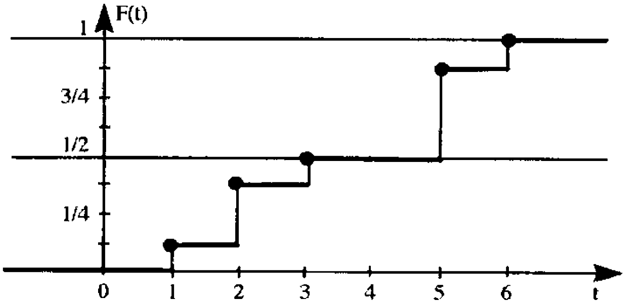


FIGURE 2.5.1

Cumulative distribution function $F(x) = F(x;\mathbf{x})$ for the data set $\mathbf{x}$ from Example 2.2.1. It is the inverse function of the quantile function $q(\alpha)$ shown in Fig. 2.2.2.

each other via the following formulas:

$$F(x) = F(x;\mathbf{x}) = \sum_{j:v_j \le x} f(v_j) \text{ and } f(v_j) = F(v_j) - \lim_{x \to v_j -} F(x), \tag{2}$$

where, as before, $v_1, \ldots, v_N$, are all the values appearing in the data set $\mathbf{x}$, the limit is taken for x approaching the value v_j from the left (see Fig. 2.5.1). Observe that cumulative d.f. F(x) is right continuous and has the left limits.

In the case of a sample of size n taken from a discrete set $v_1, \ldots, v_N$, one can hope that the normalized histogram, that is, the plot of its relative frequency d.f. $f(v) = f_n(v)$, stabilizes as n becomes very large. If that is the case (in general, the limit need not exist), one could define a probability $p(v)$ of value v as the limit

$$p(v) = \lim_{n \to \infty} f_n(v). \tag{3}$$

Intuitively, such an informal *frequentist* definition of probabilities on a discrete set makes perfect sense. Formally, however, it raises a number of difficulties, the foremost being the question of independence of the probability distribution function $p(v)$ of the selected sample $x_1, x_2, \ldots$ Also, in practice, one never deals with infinite samples, so the question is: How large a sample is necessary to make the approximation error "negligible"? We will address these issues in Chapters 3 and 5.

In the case of samples taken from a continuous population, the situation is even more complex. For a fixed partition, one would hope for the stability of the binned histograms as the sample size n goes to infinity. But in an effort to get a more and more complete information from the histograms, one may also want to increase their resolution, that is to decrease the bin size to zero. This double limit passage, if it works, could produce a "frequentist" probability d.f. on a continuous set of values. Actually, this problem is even crisper if one thinks in terms of the limit of the corresponding cumulative d.f. F, which in the double limit (if they existed in some formal sense) would become a cumulative d.f. on a continuous set of values. This latter approach will prove useful. It will be formalized in Chapters 3 and 5, and utilized throughout the rest of this book.

Example 2.5.1 Survival Curves and Cumulative D.F.
Functions similar to sample cumulative distribution functions also appear naturally in several other contexts. Let us return to the survival curves introduced in Section 1.4. If $N(0)$ units (say, cars, computer chips, etc.) are put in use at time $t = 0$, we denote by $N(t)$ the number of units still in use at time t. The graph of the function $N(t)$ is called the survival curve. The reliability $r(t) = N(t)/N(0)$ is the fraction of units still in use at time t and it decreases with time. The complementary function

$$F(t) = 1 - \frac{N(t)}{N(0)}, \tag{4}$$

represents the fraction of units that failed by the time t. The function $F(t)$ has all the features of a cumulative distribution function. It is nondecreasing, with values $F(-\infty) = 0$ and $F(+\infty) = 1$. Its derivative

$$f(t) = \frac{dF(t)}{dt}, \tag{5}$$

may be interpreted as the relative failure rate.

Mathematica Experiment 1. Spinning Yarn. The breaking strength data (in kilograms) for a batch of yarn is shown in Table 2.5.1, and also can be found in the file `COTTON` on the `UVW Web Site`. To produce a *Mathematica* code that would show the cumulative d.f. $F(x)$ for any data set $\boldsymbol{x} = (x_1, \ldots, x_n)$ of size n, note that we can write

$$F(x) = \frac{1}{n}\sum_{i=1}^{n} H(x - x_i) \tag{6}$$

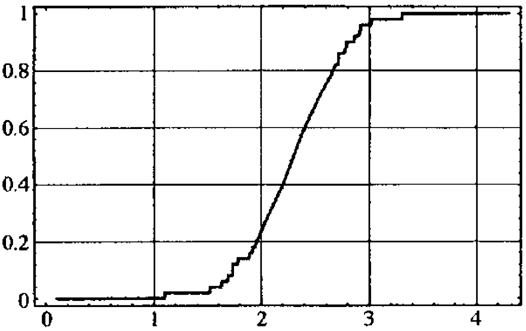
where $H(x) = 0$, for $x < 0$, and $= 1$, for $x \geq 0$, is the usual Heaviside step function. You can think of formula (6) for the cumulative d.f. as an algorithm that tells you to scan the real line from $-\infty$ to $+\infty$, and each time you encounter one of the sample points $x_1, \ldots, x_n$ you accumulate (add) an extra $1/n$ to the value of $F(x)$. So, you start with value 0 at $-\infty$ and by the time you get to $+\infty$ you will have added n $1/n$s, that is you end up with 1.

The *Mathematica* command `If[cond, a, b]` produces `a` if condition `cond` is satisfied and `b` if it is not. Hence `H[x]:=If[x<0,0,1]` defines the Heaviside unit step function. The command `list[[i]]` selects the i-th element of the `list`.

Table 2.5.1 The breaking strengths of 50 cotton threads.

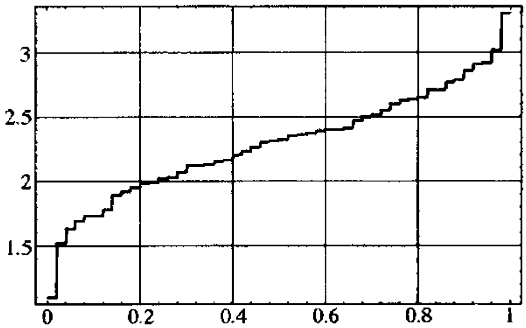
No.	Breaking strength in kg	No.	Breaking strength in kg	No.	Breaking strength in kg
i	$x_{(i)}$	i	$x_{(i)}$	i	$x_{(i)}$
1	1,10	18	2.13	35	2.50
2	1.52	19	2.15	36	2.52
3	1.63	20	2.16	37	2.55
4	1.69	21	2.20	38	2.60
5	1.73	22	2.23	39	2.63
6	1.73	23	2.26	40	2.64
7	1.78	24	2.30	41	2.65
8	1.89	25	2.31	42	2.71
9	1.92	26	2.32	43	2.71
10	1.95	27	2.35	44	2.77
11	1.98	28	2.36	45	2.79
12	1.99	29	2.37	46	2.86
13	2.02	30	2.39	47	2.91
14	2.03	31	2.40	48	2.92
15	2.07	32	2.40	49	3.02
16	2.12	33	2.41	50	3.30
17	2.12	34	2.47		

```
In[1]:= <<Statistics‘DescriptiveStatistics'
In[2]:= <<Graphics‘Graphics‘
In[3]:= cotton={ 1.10, ............. ,3.30}
Out[3]= { 1.10, ............ ,3.30}
In[4]:= n=Length[cotton]
Out[4]= 50
In[5]:= H[x_]:=If[x<0,0,1]
In[6]:= F[x_]:=(1/n) Sum[H[x- cotton[[i]] ],{i,1,n}]
In[7]:= Plot[F[x],{x, cotton[[1]] -1, cotton[[n]] +1},
                  Frame ->True, GridLines ->Automatic ]
Out[7]= -Graphics-
```



Alternatively, we can get the same information from the quantile function $x = q(\alpha)$ for the same data, which is the inverse function for the cumulative d.f. $\alpha = F(x)$.

```
In[1]:= <<Statistics‘DescriptiveStatistics‘
In[2]:= <<Graphics‘Graphics‘
In[3]:= cotton={ 1.10, .............,3.30}
Out[3]= { 1.10, ............,3.30}
In[4]:= f[x_]:= Quantile[cotton,x]
In[5]:= Plot[f[x],{x,0.001, 0.999},Frame->True,
        GridLines->Automatic]
Out[5]= -Graphics-
```



The above cumulative d.f. curve $F(x)$ shows the fraction of the sample with breaking strength $\leq x$. A customer shopping for yarn with tensile strength $x_0 = 1.9$ or better will know immediately that he can expect about 16% of the batch to be "bad".

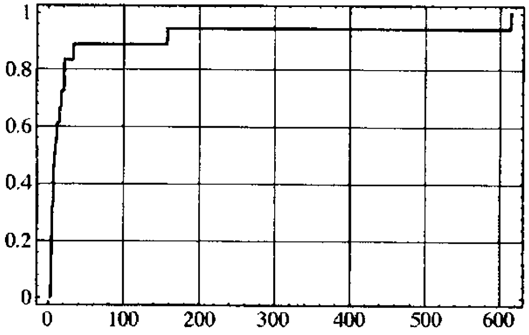
Mathematica Experiment 2. Companies, Small Town, U.S.A. The number of employees in $n = 18$ companies located in Small Town, U.S.A., is listed in the file

COMPANY on the UVW Web Site. If the number of employees serves as a measurement of the company size, one can get information about the distribution of company sizes by counting the number $N(x)$ of companies with $\leq x$ employees. Then,

$$F(x) = \frac{N(x)}{n},$$

where N is the total number of companies surveyed gives a cumulative d.f. of the sample.

```
In[1]:= <<Statistics‘DescriptiveStatistics‘
In[2]:= <<Graphics‘Graphics‘
In[3]:= company={3,3,4,4,4, 5, 6,6, 8,9,11,14,17,21,21,33,
        157,614}
Out[3]= { 3,3,4,4,4, 5, 6,6, 8,9,11,14,17,21,21,33,157,614}
In[4]:= n=Length[company]
Out[4]= 18
In[5]:= H[x_]:=If[x<0,0,1]
In[6]:= F[x_]:=(1/n) Sum[H[x- company[[i]] ],{i,1,n}]
In[7]:= Plot[F[x],{x, company[[1]] -1, company[[n]] +1},
                Frame ->True, GridLines ->Automatic ]
Out[7]= -Graphics-
```



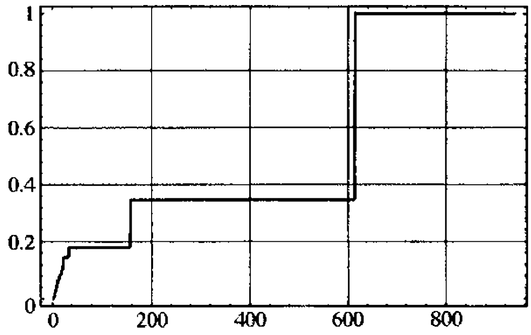
On the other hand, if $E(x)$ denotes the number of employees employed by all companies with $\leq x$ employees then one can consider the cumulative d.f.

$$G(x) = \frac{E(x)}{E},$$

where E is the total number of employees of all surveyed companies.

```
In[8]:= G[x_]:=(1/Sum[company[[i]],{i,1,n}])
```

```
                 Sum[company[[i]]H[x- company[[i]] ],{i,1,n}]
In[9]:= Plot[G[x],{x, company[[1]] -1, Sum[company[[i]],
        {i,1,n}] +1}, Frame->True, GridLines->Automatic ]
Out[9]= -Graphics-
```



One immediately finds that 90% of all companies have $\leq$ 50 employees, but they employ only 20% of the total work force.

2.6 Multidimensional data: histograms and other graphical representations

The question of representation of multidimensional data, that is data in which each sample point is a vector with d components, has come up on several occasions in the preceding sections. We can introduce for such data multivariate analogs of one-dimensional notions of the relative frequency distribution functions, histograms, cumulative distribution functions, etc., which are then functions of d variables. Obviously, graphing them is an impossible task with the exception of 2-D data, where their graphs become surfaces.

To be more precise, let us consider a 2-D sample of size n (you can think about it as an n-D vector in which each component is a 2-D vector)

$$(x_1, y_1), \ldots, (x_n, y_n),$$

and assume that sample $\boldsymbol{x} = (x_1, \ldots, x_n)$ consisting of first components takes values from the set of values

$$v_1, \ldots, v_N,$$

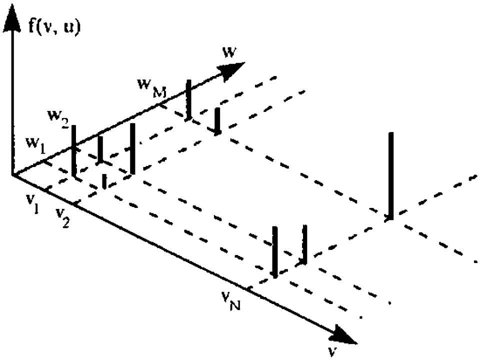


FIGURE 2.6.1
A schematic example of a 2-D relative frequency d.f. for a discrete sample from 2-D data.

and sample $\boldsymbol{y} = (y_1, \ldots, y_n)$ from the set of values

$$w_1, \ldots, w_M.$$

Then the *joint 2-D relative frequency d.f.* $f(v_j, w_k)$ counts the relative frequencies of appearance in our 2-D sample of all possible pairs of values (v_j, w_k), $j = 1, \ldots, N, k = 1, \ldots, M$. In other words,

$$f(v_j, w_k) = \frac{1}{n}\#\{i : (x_i, y_i) = (v_j, w_k)\}. \tag{1}$$

A schematic example of the graph of a 2-D relative frequency d.f. for discrete data is given in Fig. 2.6.1 in the form of a stick graph.

It is also easy to see that the relative frequency d.f.s of 1-D component samples $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$ are easily obtainable from the joint 2-D relative frequency d.f. via the following formulas

$$f(v_j) = \sum_{k=1}^{M} f(v_j, w_k), \tag{2}$$

and

$$f(w_k) = \sum_{j=1}^{N} f(v_j, w_k). \tag{3}$$

Thus, the sample means and variances of the component samples $\boldsymbol{x}$ and $\boldsymbol{y}$ are easily available. There are, however, other characteristics of multidimentional samples

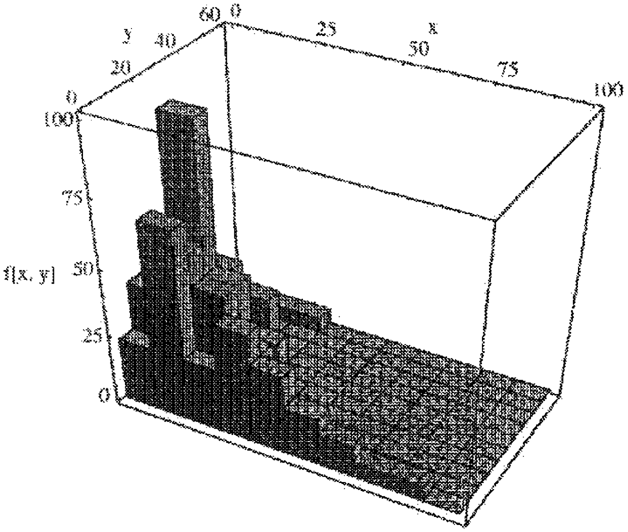
$(\boldsymbol{x}, \boldsymbol{y})$ computable from the joint relative frequency d.f. $f(x, y)$ that could not be obtained if only 1-D relative frequency d.f.s of component samples $\boldsymbol{x}$ and $\boldsymbol{y}$ were available. One of them is the covariance between the two component data defined as

$$\operatorname{cov}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j=1}^{N} \sum_{k=1}^{M} (v_j - \bar{x})(w_k - \bar{y}) f(v_j, w_k). \tag{4}$$

We will see its usefulness and interpretation in the next section.

Mathematica Experiment 1. Heart Trouble. The file `HOSP/HEART` on the `UVW Web Site` contains heart transplant data from all the organ transplant centers in the United States. The quantities listed are the median waiting times W, one-year mortality rates M (that is, percentage of patients dying within one year of the operation), and the average annual number of transplant V at that center for a 4-year period beginning in October 1987. The data are three-dimensional. In this experiment we will produce the 2-D histogram of the paired data (M, V) pairing the mortality and volume at each organ transplant center. The command `Histogram2D[list,xmin,xmax,nx,ymin,ymax,ny]` of the `UVW‘DataRep‘` package produces a 2D histogram of a 2D `list`, within the `[xmin,xmax]` interval on the x-axis and `[ymin,ymax]` interval on the y-axis. The number of bins on the x-axis is `nx` and the number of bins on the y-axis is `ny`.

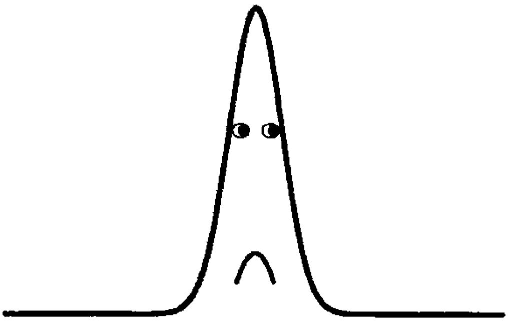
```
In[1]:= <<Graphics‘Graphics3D‘
In[2]:= <<UVW‘datarep‘
In[3]:= heart={{17.9,27}, {23.1,4},...,{17.6,26},{19.6,14}}
Out[3]= {{17.9,27}, {23.1,4},...,{17.6,26},{19.6,14}}
In[4]:= Length[heart]
Out[4]= 134
In[5]:= Histogram2D[heart,0,100, 10,0,60, 10]
Out[5]= -Graphics-
```



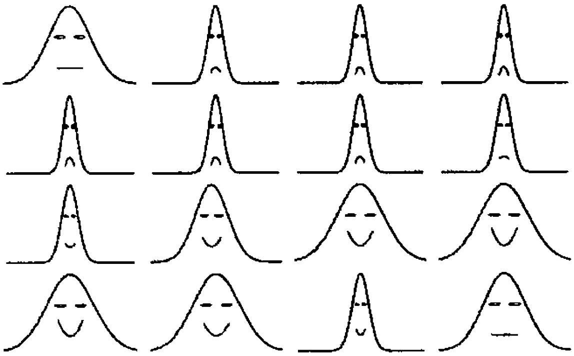
The above concept of the joint 2-D relative frequency d.f. can be extended easily to representations of 3- or higher-dimensional data but it is quite clear that the graphical representation becomes more and more difficult as the dimension of the data set increases. There are, however, other ingenious ways to represent multivariate data and one of them is due to Hermann Chernoff. He suggested associating multidimensional data with several features of the human face—an object humans are especially apt to recognize in its multiplicity of (multidimensional) features. Our light-hearted version of Chernoff's idea brings you `StoGho`—the quintessential stochastic ghost. Play with him in the next *Mathematica* experiment. The curvature of his lip, the eye shape, pupils' position and the flatness of the Gaussian-shaped head encode four-dimensional information.

Mathematica Experiment 2. StoGho Lives. Be patient here as it takes time. Also, `StoGho` can be coupled with `Animate[]` if there is enough memory in your computer. The command `Random[Real, {-3,3}]` produces a pseudorandom random number uniformly distributed in the interval (-3,3).

```
In[1]:= <<UVW‘StoGho‘
In[2]:= StoGho[Random[Real,{-3,3}]]
Out[2]= -Graphics-
```



```
In[3]:= moods=Partition[Range[-Pi,Pi,2 Pi/15],4];
In[4]:= GalleryOfPortraits[moods]
Out[4]= -GraphicsArray-
```



2 7 2-D data: regression and correlations

Consider a 2-D numerical data set

$$((x_1, y_1), \ldots, (x_n, y_n)) = (\boldsymbol{x}, \boldsymbol{y})^T$$

of size n. When graphed in the 2-D plane as dots, the data produce the so-called *scatter plot.*

Example 2.7.1 Current and Conductivity.
To verify the Ohm's Law an experimenter applied a fixed voltage V to $n = 7$ different passive electrical circuits and measured the current intensity I and the circuits conductivity (inverse resistance) $1/R$. The results are tabulated in Table 2.7.1, and the corresponding scatter plot is shown in Fig. 2.7.1.

Table 2.7.1 Current intensity I vs. inverse resistance $1/R$ experiment

$X = 1/R$ (in $1/\Omega$)	1/10	1/20	1/50	1/100	1/300	1/500	1/1000
I (in mA)	4.95	2.52	0.98	0.50	0.16	0.102	0.052

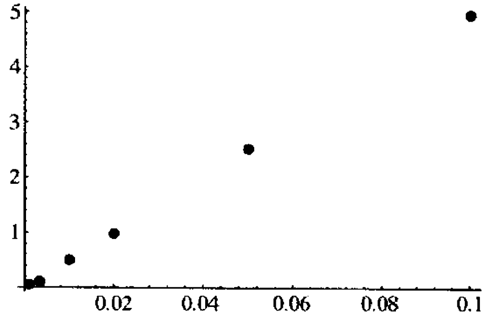


FIGURE 2.7.1
Scatter plot of 2-D data from Table 2.7.1. It suggests a strong linear relationship between the two components, current intensity I and conductivity $1/R$.

The scatter plot suggests an almost perfect linear relationship between the two components, current intensity I and conductivity $1/R$.

Whenever there is a suspicion that there is a linear relationship between the components of 2-D data, the obvious goal is to find coefficients α and β which would make the straight line

$$y = \alpha + \beta x \tag{1}$$

the best possible approximation for the scatter plot representing the data. In other words, the job is to find a compressed representation of the 2-D data in the form of an optimally selected straight line. Such a line is traditionally called the *regression line* for 2-D data. An historical explanation for the use of the term *regression* in this context can be found in Section 8.5.

Obviously, to make the above task meaningful we have to decide on the optimality criterion for choosing α and β in the formula (1). The usual choice is

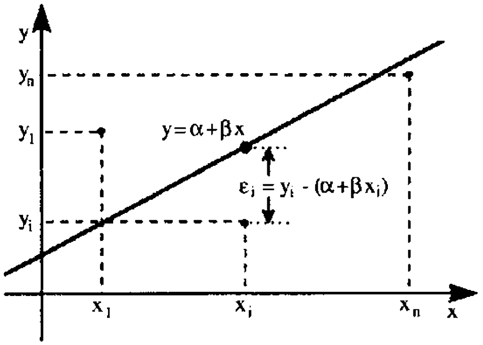


FIGURE 2.7.2
A schematic illustration for the regression line selection algorithm for 2-D data.

minimization of the total quadratic error

$$\text{Err}\ (a, b) = \sum_{i=1}^{m} \epsilon_i^2 = \sum_{i=1}^{n} \Big(y_i - (a + bx_i)\Big)^2 \tag{2}$$

of approximation of the scatter plot by the regression line over all possible choices of real numbers a and b. The quantities $\epsilon_i = y_i - (a + bx_i)$, $i = 1, 2, \ldots, n$, represent the individual, sample point by sample point, errors of approximation and are also often called *fit residuals*; the optimization method itself is called Gauss' *least squares method.*

Since function Err (a, b) is a nonnegative quadratic function of variables a and b, its minimum is achieved at the points $(a, b) = (\alpha, \beta)$ satisfying equations

$$\frac{\partial}{\partial\alpha}\text{Err}\ (\alpha, \beta) = -2\sum_{i=1}^{n} y_i - (\alpha + \beta x_i) = 0,$$

$$\frac{\partial}{\partial\beta}\text{Err}\ (\alpha, \beta) = -2\sum_{i=1}^{n} \Big(y_i - (\alpha + \beta x_i)\Big)x_i = 0,$$

which are also called the *normal equations* for the regression problem. Finding explicit solutions of a system of two linear equations with two unknowns is not difficult, but a lucid notation helps to see what is happening. So, observe that the two normal equations can be rewritten as follows:

$$\sum_{i=1}^{n} y_i = \alpha n + \beta \sum_{i=1}^{n} x_i,$$

$$\sum_{i=1}^{n} x_i y_i = \alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2.$$

The first equation is recognizable as $\bar{y} = \alpha + \beta\bar{x}$, wherefrom, immediately,

$$\alpha = \bar{y} - \beta\bar{x}. \tag{3}$$

Substituting this α into the second normal equation gives

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

so that, finally

$$\beta = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}. \tag{4}$$

This formula, although good for numerical computations (why?) can be made more transparent if we recognize the quantity in the denominator as the theoretical sample variance of $\boldsymbol{x}$ (see formula (2.3.7)) multiplied by n, and the numerator as the *theoretical sample covariance* (see formula (2.6.4)) multiplied by n:

$$n \operatorname{cov}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} x_i y_i - \bar{\boldsymbol{x}}\bar{\boldsymbol{y}} = \sum_{i=1}^{n} (x_i - \bar{\boldsymbol{x}})(y_i - \bar{\boldsymbol{y}}). \tag{5}$$

In this notation,

$$\beta = \frac{\operatorname{cov}(\boldsymbol{x}, \boldsymbol{y})}{\operatorname{var}(\boldsymbol{x})}. \tag{6}$$

Notice that the covariance of a 2-D data $(\boldsymbol{x}, \boldsymbol{x})^T$ is just the variance of the 1-D data $\boldsymbol{x}$:

$$\operatorname{cov}(\boldsymbol{x}, \boldsymbol{x}) = \operatorname{var}(\boldsymbol{x}). \tag{7}$$

Often, one normalizes the covariance by the standard deviations (2.3.14) of samples $\boldsymbol{x}$ and $\boldsymbol{y}$ to obtain what is called the *correlation coefficient* between 1-D samples $\boldsymbol{x}$ and $\boldsymbol{y}$:

$$\operatorname{corr}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\operatorname{cov}(\boldsymbol{x}, \boldsymbol{y})}{\operatorname{std}(\boldsymbol{x}) \operatorname{std}(\boldsymbol{y})}. \tag{8}$$

In view of the Schwarz Inequality (see Project 1 at the end of this chapter)

$$-1 \leq \operatorname{corr}(\boldsymbol{x}, \boldsymbol{y}) \leq 1. \tag{9}$$

That is why the normalization of the covariance was useful.

Note that, taking into account (3),(4),(6), and (8), the regression line equation (1) can now be written in the following elegant nondimensional form:

$$\frac{y - \bar{y}}{\text{std}\,(y)} = \text{corr}\,(x, y)\frac{x - \bar{x}}{\text{std}\,(x)}. \tag{10}$$

Its physical interpretation is clear and intuitive: The regression line passes through the point with coordinates $(\bar{x}, \bar{y})$ and, after normalizing the scales of x and y by division by their respective standard deviations, the regression line's slope is equal to the correlation coefficient between x and y. The correlation coefficient is nondimensional, and so are the normalized quantities $(x - x)/\text{std}\, x$ and $(y - y)/\text{std}\, y$.

Example 2.7.2 Extreme Correlations.
It is illuminating to calculate values of the correlation coefficient of three sets of 2-D data with the scatter plots presented on Fig. 2.7.2:

- If $y_i = \beta x_i, \beta < 0$, then $\text{corr}\,(x, y) = -1$.
- If $y_i = \beta x_i, \beta > 0$, then $\text{corr}\,(x, y) = +1$.
- If $x = (-1, -1, +1, +1)$ and $y = (-1, +1, -1, +1)$, then $\text{corr}\,(x, y) = 0$.

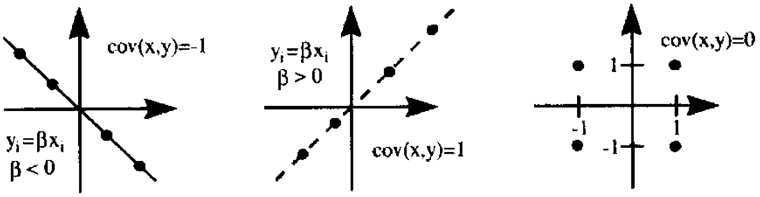
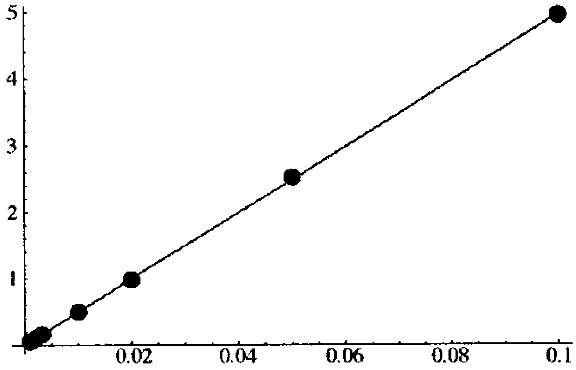


FIGURE 2.7.3
Three extreme cases of the correlation coefficient.

In Chapters 5 and 8 we will see that the above three examples (please, do all the calculations by hand) are of more than passing significance.

Mathematica Experiment 1. Current and Conductivity. We'll work with 2-D data from Example 2.7.1. The *Mathematica* command `Fit[data, {1,x}, x]` produces the least squares linear fit of the `data`.

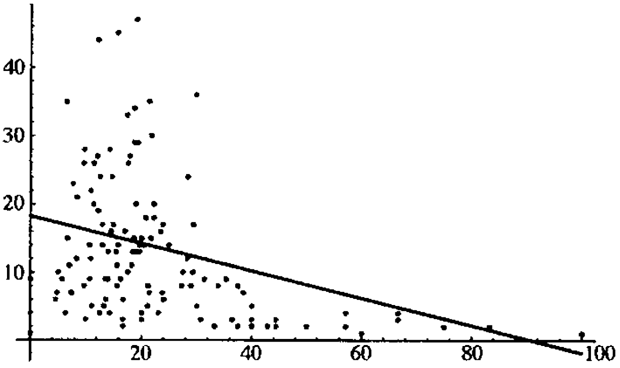
```
In[1]:= <<Graphics'Graphics'
In[2]:= <<Statistics'LinearRegression'
In[3]:= current={{1/10,4.95},{1/20,2.52},{1/50,0.98},
            {1/100,0.50},{1/300,0.16},{1/500,0.102},
            {1/1000,0.052}}
Out[3]= {{1/10,4.95},{1/20,2.52},{1/50,0.98},
            {1/100,0.50},{1/300,0.16},{1/500,0.102},
             {1/1000,0.052}}
In[4]:= linear=Fit[current,{1,x},x]
Out[4]= 0.00243603 + 49.6258 x
In[5]:= DisplayTogether[ListPlot[current],Plot[linear,{x,0,1}]]
Out[5]= -Graphics-
```



You will notice that the linear fit is almost perfect confirming the validity of Ohm's law.

Mathematica Experiment 2. Heart Trouble. Finally, let us take a look at a much larger data set from the 2-D data `heart` from Mathematica Experiment 2.6.1.

```
In[1]:= <<Graphics`Graphics`
In[2]:= <<Statistics`LinearRegression`
In[3]:= heart={{17.9,27}, {23.1,4},...,{17.6,26},{19.6,14}}
Out[3]= {{17.9,27}, {23.1,4},...,{17.6,26},{19.6,14}}
In[4]:= linear=Fit[heart,{1,x},x]
Out[4]= 18.2703 -0.201084
In[5]:= DisplayTogether[ListPlot[heart],Plot[linear,{x,0,100}]]
Out[5]= -Graphics-
```

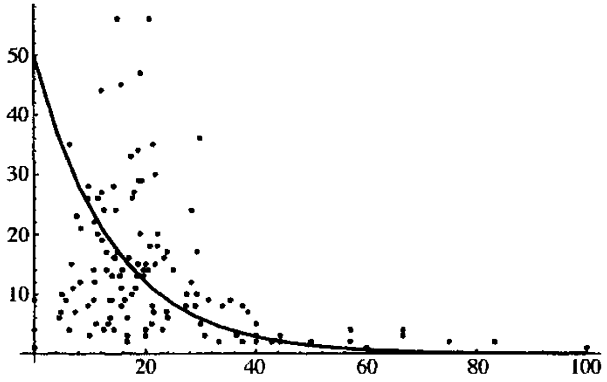


One notices immediately that the linear fit does not compress the data well. Moreover, the negative values on the fitted regression line do not make much sense. So, one could wonder if a nonlinear fit would be better suited here for the data compression job. Given the shape of the data we will try to find a fit by the function (model)

$$50\exp[-0.01\alpha x]$$

which contains the parameter α to be optimally selected by the least-squares method. The explicit analytic solution cannot be found as easily as in the linear model but *Mathematica* provides a command `NonlinearFit[data, model, variables, parameters]` which fits the `data` to the `model` with the named `variables`, returning the model evaluated at the parameter values achieving the least-squares fit. You will notice that the computer takes much longer to find a nonlinear fit than a linear one.

```
In[6]:= <<Statistics`NonlinearFit`
In[7]:= exponential=NonlinearFit[heart,  50 *
                     Exp[-0.01  *alpha  *x], {x}, {alpha }]
Out[7]= {alpha  -> 7.1603}
In[8]:= DisplayTogether[ListPlot[heart],Plot[50 *
                     Exp[-0.01 *7.1603 *x], {x,0,100}]]
Out[8]= -Graphics-
```



2.8 Fractal data

Fractal (fractional dimension) data are generated by physical phenomena governed by chaotic dynamics, in interacting particle systems, flows in porous media, and in many other situations. Fig. 2.8.1 shows the bacteria *Baccillus subtilis* culture growing in the Petri dish on the surface of agar plates. The mathematical models which give rise to such data will be discussed in Chapter 6.

In this section we will address only the question of computing the fractal dimension in data sets that are suspected to be fractal, such as in Example 1.5.1 of time intervals between water drops, Example 1.7.2 of the EKG time series showing the onset of seizure, or Fig. 1.10.2 and 1.10.3 showing the passive tracer density in a random velocity flow.

There are several and, in general, nonequivalent definitions of the fractal dimension of a subset of a d-dimensional Euclidean space. They all coincide, however, for some simple sets.

The simplest definition, due to Hermann Minkowski, relies on the fact that if you have a "solid" object in $\mathbf{R}^d$ (such as an interval in $\mathbf{R}^1$, square in $\mathbf{R}^2$, or a cube in $\mathbf{R}^3$) then its "natural" dimension d coincides with the "coverage exponent" which can be explained as follows:

For a "solid" set $A \subset \mathbf{R}^d$ consider a coverage by d-dimensional volume elements such as balls (or d-dimensional cubes) of radius (or edge size) ϵ. Fig. 2.8.2 shows coverage of a 2-dimensional square by 2-dimensional discs of radius ϵ.

Then, it is intuitively clear that the smallest number $N(\epsilon)$ of such volume elements needed for a coverage of A is equal to $C\epsilon^{-d}$, where C is a certain constant. Solving this equality for d, and taking the limit $\epsilon \to 0$ to free ourselves from the dependence on an unknown constant C, we obtain the following formula of the

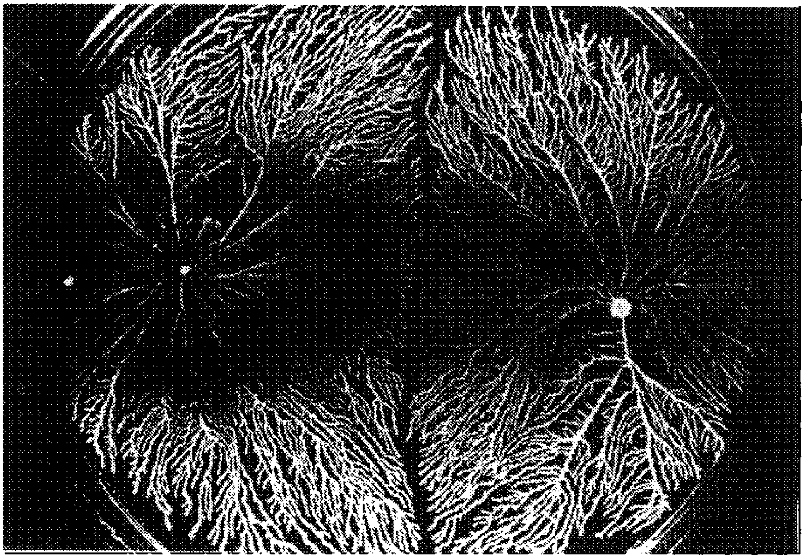


FIGURE 2.8.1

Bacteria colonies of Baccillus subtilis growing in a Petri dish show a fractal structure. (Courtesy of M. Matsushita, Chuo University.)

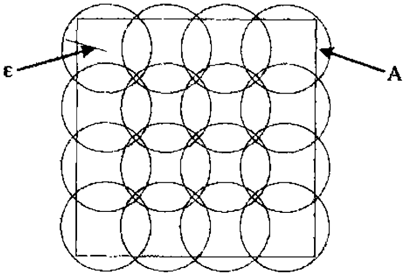


FIGURE 2.8.2

Coverage of a 2-D square by discs of radius ε.

dimension of set A :

$$d_{cap}(A) = \lim_{\epsilon \to 0} \frac{\ln N(\epsilon)}{\ln(1/\epsilon)}, \qquad (1)$$

which, if applied to an arbitrary subset $A \subset \mathbf{R}^d$ (with possible replacement of the limit by lim sup if the former does not exist), serves as a definition of the *capacity dimension* of A which can take noninteger values.

Example 2.8.1 Cantor Set C.

Here is a constructive algorithm for C. The numbering ((1), (2.0), (2.2), . . . , ($n.k$))

of steps corresponds to the construction process itself to make it easier to see what is going on. Begin with the unit interval [0, 1].

(1) Remove the interval $(1/3, 2/3)$ from $[0, 1]$ to obtain the intervals $[0, 1/3]$ and $[2/3, 1]$.

(2.0) Remove the "middle third" interval $(1/9, 2/9)$ from $[0, 1/3]$ to obtain the intervals $[0, 1/9]$ and $[2/9, 3/9]$.

(2.2) Remove the interval $(7/9, 8/9)$ from $[2/3, 1]$ to obtain the intervals $[6/9, 7/9]$ and $[8/9, 1]$.

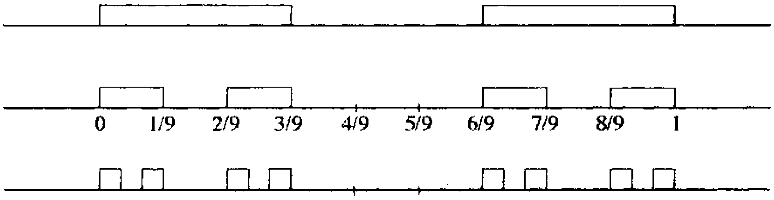


FIGURE 2.8.3
Repeated "middle-third-removed" construction of the Cantor set.

Then continue in the same fashion. The general recursive recipe is as follows:

$(n.k)$ Suppose we have obtained in the n-th step the interval $[k3^{-n}, (k+1)3^{-n}]$ for an integer k in whose triadic representation (i.e., using three digits: 0,1,2) only digits 0 or 2 appear. Then, in the next step, remove the middle third interval $((3k+1)3^{-n-1}, (3k+2)3^{-n-1})$ from $[k3^{-n}, (k+1)3^{-n}]$ to obtain the intervals $[(3k)3^{-n-1}, (3k+1)3^{-n-1}]$ and $[(3k+2)3^{-n-1}, (3k+3)3^{-n-1}]$. Note, that both $3k$ and $3k+2$ have again triadic representations containing only digits 0 or 2.

The set of points that are in the intersection of all these "middle-third-removed" sequence of sets constructed above is the classical Cantor set. The fact that the Cantor set is nonempty is a deep mathematical theorem. Its existence is closely related to some other basic mathematical foundational facts such as the existence of irrational numbers. The Cantor set is an example of a fractal set, which means that its dimension is not an integer. Indeed, a direct application of the definition of the capacity dimension to the Cantor set C gives

$$d_{cap}(C) = \frac{\ln 2}{\ln 3} = 0.6309\ldots$$

The original Felix Hausdorff[1] *dimension* is more complicated to introduce and to compute but, in view of its historical and theoretical importance, we will define it formally below.

A cover of a set $S \subset \mathbf{R}^d$ is a family $\mathcal{A}$ of sets A_i such that each point from S lies in at least one set of the covering family. The diameter of a set A is the maximum of the possible distances of two points from A, and will be denoted $\delta(A)$. A cover is called an ϵ-cover if all the sets of the covering family have diameters less than ϵ.

Fix a $d > 0$ and define number

$$\mathcal{N}(S, d, \mathcal{A}) = (\delta(A_1))^d + (\delta(A_2))^d + (\delta(A_3))^d + \ldots$$

where $\mathcal{A}$ is a cover of S. A number $d_H(S)$ is said to be the *Hausdorff dimension* of S if it satisfies the following two conditions:

(a) For every $d > d_H(S)$, there exists a sequence of ϵ_n-covers $\mathcal{A}_n$, $\epsilon_n \to 0$, such that $\sup_n \mathcal{N}(S, d, \mathcal{A}_n) < \infty$, and

(b) For every $d < d_H(S)$, there exists an unbounded sequence of numbers $M_\epsilon \to +\infty$ as $\epsilon \to 0$ such that, for every ϵ-cover $\mathcal{A}$, $\mathcal{N}(S, d, \mathcal{A}) \geq M_\epsilon$.

In other words,

$$d_H(S) = \inf\left\{d > 0 : \exists \epsilon_n\text{–covers } \mathcal{A}_n, \epsilon_n \to 0, \text{s.t. } \mathcal{N}(S, d, \mathcal{A}_n) < \infty\right\}. \quad (2)$$

Example 2.8.2 Hausdorff Dimension of the Unit Interval.
Let $S = [0, 1]$, and let $\mathcal{A}_n = \{[k/n, (k+1)n] : 0 \leq k < n - 1\}$ be a $1/n$-cover. Then, for $d > 0$,

$$\mathcal{N}(S, d, \mathcal{A}_n) = n\left(\frac{1}{n}\right)^d = n^{1-d}.$$

So, if $d > 1$, then $\sup_n \mathcal{N}(S, d, \mathcal{A}_n) = 1$, and if $d < 1$, then $\mathcal{N}(S, d, \mathcal{A}_n) \to \infty$ as $n \to \infty$. It follows that $d_H([0, 1]) = 1$.

Example 2.8.3 Hausdorff Dimension of the Cantor Set.
Fix a number $d > 0$ and the diameter $\epsilon = 3^{-n}$. The intervals in the n-th step of the construction of C have length ϵ, cover C, and there are exactly 2^n of them. For this cover

$$\mathcal{N}(C, d, \mathcal{A}) = 2^n 3^{-nd}.$$

Since

$$\lim_{n\to\infty} 2^n 3^{-dn}$$

[1] Hausdorff was also a writer, publishing fiction under the pseudonym Paul Mongré

remains bounded if and only if $2/3^d \leq 1$, and that is possible only if the coefficient $\log 2 - d \log 3 \leq 0$, or, equivalently, $d \geq \log 2/\log 3$, we conclude that the Hausdorff dimension of the Cantor set is

$$d_H(C) \leq \frac{\log 2}{\log 3}.$$

To prove the equality is not so easy and omitted. It is always true that

$$d_H(S) \leq d_{cap}(S),$$

and since $d_{cap}(C) = \log 2/\log 3$, the Hausdorff dimension of the Cantor set is equal to its capacity dimension.

The Hausdorff and capacity dimensions are not the most appropriate (or easiest to compute) quantities for fractal data that arise as time series because they do not take into account the frequency of visits to the same "state". This difficulty is overcome by the Grassberger and Procaccia's (1983) definition of the *correlation dimension* d_{cor} which, for a discrete data set $S = (x_1, x_2, \ldots, x_n)$, and a given (small) resolution $\epsilon > 0$, is defined by the formula

$$d_{cor}(S, \epsilon) = \frac{\ln C(\epsilon)}{\ln \epsilon}, \tag{3}$$

where

$$C(\epsilon) = \frac{\#\{(x_i, x_j) : |x_i - x_j| < \epsilon,\ 1 \leq i, j \leq n\}}{n^2}. \tag{4}$$

One can show that, for infinite data sets S,

$$\lim_{\epsilon \to 0} d_{cor}(S, \epsilon) \leq d_H(S) \leq d_{cap}(S), \tag{5}$$

and that, for many classes of sets, the above three concepts of fractal dimension coincide.

Mathematica Experiment 1. Water Drips. We will use the water drips data provided in Example 1.5.2.

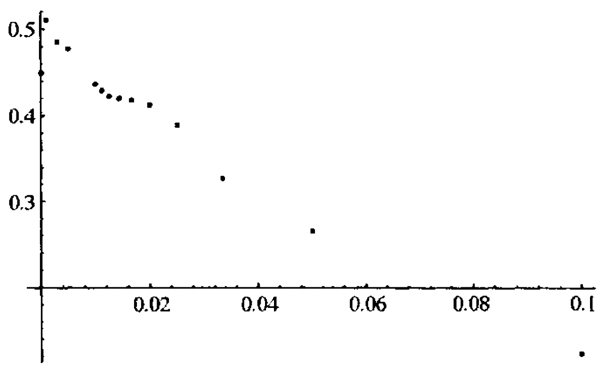
```
In[1]:= drip={0.1822, 0.1962,  ....... , 0.2210, 0.1485}
Out[1]= {0.1822, 0.1962,  ....... , 0.2210, 0.1485}
In[2]:= n=Length[drip]
Out[2]= 70
In[3]:= dripdiff=Table[drip[[i]]-drip[[j]],{i,n},{j,n}]
Out[3]= {{0.,-0.014, 0.048, ... , -0.0388, 0.0337},
          . . . . . . . . . . . . . . . . . . .
          {-0.0337,-0.0477, ... , -0.0725, 0. }}
```

The last command creates the 70 × 70 matrix of differences $x_i - x_j$ needed in formula (4). To see what is the reasonable selection of resolution ϵ we will check the maximum and minimum of absolute values of the (nonzero) terms in the above matrix. The command `Range [n]` produces the list $\{1, 2, \ldots, n\}$ providing enumeration of all the rows of the matrix `dripdiff`.

```
In[4]:= Max[ Abs[dripdiff[[Range[n]]] ]]
Out[4]= 0.2115
In[5]:= zeros=Position[dripdiff,0.]
Out[5]= {{1, 1}, {2, 2}, ... , {69, 69}, {70, 70}}
In[6]:= nozeros=ReplacePart[dripdiff,1,zeros]
Out[6]= {{1,-0.014, 0.048, ... , -0.0388, 0.0337},
          . . . . . . . . . . . . . . . . . . . . .
         {-0.0337,-0.0477, ... , -0.0725, 1}}
In[7]:= Min[ Abs[nozeros[[Range[n]]] ]]
Out[7]:=0.0001
```

The above result indicates that the values of the function $d_{cor}(\epsilon)$ for $\epsilon < 0.0001$ are not of much interest as $C(\epsilon)$ remains constant in that domain. Let us check the values of $d_{cor}(\epsilon)$ for a selection of $\epsilon > 0.0001$.

```
In[8]:= H[x_]:=If[x<0,0,1]
In[9]:= d[epsilon_]:=(1/Log[epsilon])* Log[ (1/n^2)*
        Sum[H[epsilon-Abs[dripdiff [[i]][[j]] ] ],
           {i,1,n},{j,1,n}]]
In[10]:= Table[{0.1/k,d[0.1/k]},{k,10}]
Out[10]= {{0.1, -0.434294 Log[922/1225]},...,
          {0.01, -0.217147 Log[164/1225]}}
In[11]:= cordim1=N[%]
Out[11]= {{0.1, 0.123405}, {0.05, 0.26549},
           {0.0333333, 0.327238},
           {0.025, 0.389185}, {0.02, 0.412454},
           {0.0166667, 0.418268}, {0.0142857, 0.420781},
           {0.0125, 0.42291}, {0.0111111, 0.42929},
           {0.01, 0.436646}}
In[12]:=  {{0.005,d[0.005]}, {0.003,d[0.003]},
           {0.001,d[0.001]} ,{0.0001,d[0.0001]},
Out[12]= {{0.005, -0.188739 Log[39/490]}, . . .}
In[13]:= cordim2=N[%]
Out[13]= {{0.005, 0.477669}, {0.003, 0.485482},
          {0.001, 0.510611} ,{0.0001, 0.449525}}
In[14]:= Join[cordim1, cordim2]
Out[14]= {{0.1, 0.123405}, {0.05, 0.26549},  .... ,
          {0.001, 0.510611} ,{0.0001, 0.449525}}
In[15]:= ListPlot[Join[cordim1, cordim2]
Out[15]= - Graphics-
```



Given the outcome of the above *Mathematica* experiment it would be reasonable to say that the correlation dimension of the above data is in the neighborhood of 0.5. For further discussion of these issues, see Section 2.10, and Chapters 6 and 7.

2.9 Measuring information content: entropy

To avoid confusion caused by many colloquial interpretations of the word *information*, we should make it clear at the very beginning that we are not seeking here the measure of information as measure of meaning or semantic content, but only as measure of content of information *transmitted* from a known pool of possible messages. The semantic aspects of communication, or the questions of the truth of messages, are totally irrelevant to our mathematical formulation.

Example 2.9.1 Hot, Warm, and Cold.

The weather reports in the Cleveland *Plain Dealer* provide five-day forecasts and one of the predicted items is temperature which is described as hot (H), warm (W), or cold (C). The past records show that during the Spring season the relative frequency of hot and cold weather was much smaller than the warm weather. How much information does tomorrow's forecast carry? Clearly, if the forecast says W, then the amount of new information provided to us is smaller than if the forecast says C, because on past evidence we already know that warm weather is more likely in the spring than cold weather.

To settle on a particular quantitative measure of information content of a received message, let us take a look at another familiar example.

Example 2.9.2 Hard Disk.

The amount of information carried by the data obviously depends on how many "bits" are necessary to transmit the data and on the frequency of different symbols appearing in the data. Thus, intuitively speaking, the 2-megabyte hard disk should be able to carry twice the amount of information on a 1-megabyte disk, and the n-megabyte hard disk should carry n times the information on a 1-megabyte disk.

A received message that is one of the 10,000,000 equally likely possible messages (here, think about your favored state lottery) is more valuable and carries more information than the message that comes from the pool of 100 equally likely messages. In other words, the measure of information content should be an increasing function of the pool size from which the messages come, assuming, on the past experience, that they all are equally likely. So, the information content has to be tied monotonically to the number of possible states of our data set.

For a hard disk, the number of possible states N grows exponentially with the disk information storage capacity: 1-bit binary storage can store 2 messages, 0 and 1, but the n-bit binary storage can store $N = 2^n$ strings of length n. Thus, the disk capacity n required to store one of possible N states of our data set is

$$n = \log_2 N. \tag{1}$$

For a general pool of N possible messages

$$H = \ln N \tag{2}$$

is usually called the *Hartley information capacity*. It is measured in *bits*. The choice of the natural logarithm is somewhat arbitrary and corresponds to the selection of a specific measurement unit. For binary messages, the choice of the logarithm to the base 2 would be more appropriate but the selection of the natural logarithm makes our approach uniform. Notice that the above formula can be written in the form

$$H = -\sum_m \frac{1}{N} \ln \frac{1}{N} = -\sum_m f(m) \ln f(m), \tag{3}$$

where the summation is over all possible messages m, each appearing with the *a priori* relative frequency $f(m) = 1/N$.

On the other hand, if the disk has been destroyed and has 0s permanently recorded on all its bytes, then its information capacity is obviously zero.

This leads us to the question: What is the information content of a message received from the message pool if we know that the possible messages are not equally likely? Formula (3) suggests a measure of information content in this case as well. If $m_1, m_2, \ldots, m_N$, are possible messages in our pool, with *prior* relative frequencies $f_1, \ldots, f_N$, then the following natural generalization of (3),

$$H = H(f_1, \ldots, f_N) = -\sum_{i=1}^{N} f_i \ln f_i, \tag{4}$$

measures how much "uncertainty" is involved when we receive any particular message from this pool. The quantity H is called the *Shannon entropy* of the pool of messages (or data sets) and is traditionally also measured in bits. Notice that it depends just on the *prior* relative frequency d.f. f_i.

Example 2.9.2 Continued. Hot, Warm, and Cold.
Assume that the records show that in the past, for a Spring day, W was forecast with relative frequency $f_W = 0.5$, H with frequency $f_H = 0.2$, and C with frequency $f_C = 0.3$. Then the Shannon entropy of this message pool is

$$H = H(0.5, 0.3, 0.2) = -0.5 \ln 0.5 - 0.3 \ln 0.3 - 0.2 \ln 0.2 = 1.02965$$

If the past record indicated that all three forecasts appeared with the same relative frequency $f_W = f_H = f_C = 1/3$, then the Shannon entropy of this message pool would be $H(1/3, 1/3, 1/3) = -\ln(1/3) = 1.09861$, larger than in the non-identically distributed case above.

If there are just two possible messages, say 0 and 1, with relative frequencies f and $1 - f$ then

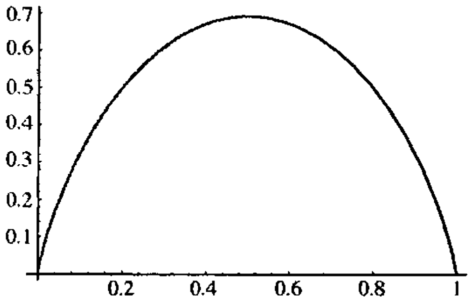
$$H = H(f, 1 - f) = -f \ln f - (1 - f) \ln(1 - f) \tag{5}$$

Notice that its maximum is attained for $f = 1/2$, that is when both messages are equally likely. Indeed, this is confirmed by finding that, at $f = 1/2$, the derivative

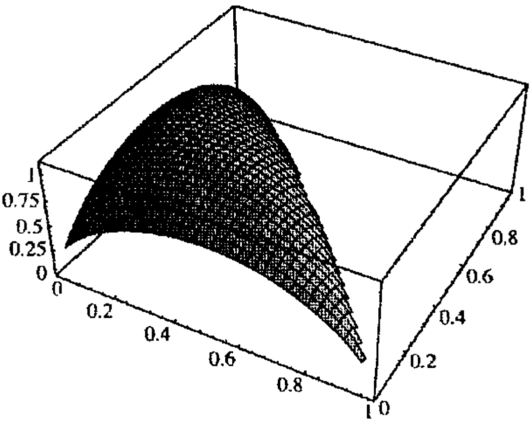
$$\frac{dH}{df} = -\ln f + \ln(1 - f) - 1 + 1 = 0. \tag{6}$$

Mathematica Experiment 1. Shannon Entropy. We will graph entropy as a function of frequencies in the case of pools of messages consisting of two and three messages, that is, function $H(f, 1 - f)$ of one variable f, $0 \le f \le 1$ and function $H(f_1, f_2, 1 - f_1 - f_2)$ of two variables f_1, f_2, with $0 \le f_1, f_2 \le 1$, $f_1 + f_2 \le 1$.

```
In[1]:= H[f_]:=-f Log[f]-(1-f) Log[(1-f)]
In[2]:= Plot[H[f],{f,0,1}]
Out[2]= -Graphics-
```



```
In[3]:= H[f1_,f2_]:=-f1 Log[f1]-f2 Log[f2]-(1-f1-f2) Log[(1-f1-f2)]
In[4]:= Plot3D[H[f1,f2], {f1,0,1}, {f2,0,1}, PlotPoints ->40]
Out[4]= -SurfaceGraphics-
```



A calculation similar to (6) (see, Experiments, Exercises, and Projects at the end of this chapter) proves the first of the following general properties of Shannon entropy:

Property 1. The maximum of the Shannon entropy function $H(f_1, \ldots, f_N)$ is achieved for $f_1 = \ldots = f_N = 1/N$. Again, this is not surprising, as the selection from the most random source of information carries with it most information.

Property 2. $H = 0$ if, and only if, for one of the $i = 1, \ldots, N$, we have $f_i = 1$ (other frequencies are then 0). Intuitively, no information is carried if there is no uncertainty about the message.

Property 3. Any perturbation of the situation described in Property 2 towards the equalized one described in Property 1 will increase the information content of a message from the pool of messages. More precisely, if $f_1 < f_2$ and if we decrease the distance between f_1 and f_2, then H will increase. In more generality, any averaging operation performed on f_is increases H, that is, if

$$f_i' = \sum_{j=1}^{N} a_{ij} f_j, \qquad a_{ij} \geq 0, \qquad \sum_i a_{ij} = \sum_j a_{ij} = 1, \tag{7}$$

then

$$H(f_1', \dots, f_N') \geq H(f_1, \dots, f_N). \tag{8}$$

This is a simple consequence of the convexity of the function $g(x) = -x \log x$. Obviously, entropy does not change if the above operation just leads to permutation of frequencies.

Remark 2.9.1 *Rigorous Derivation of the Entropy Function.* The Shannon form (4) of the entropy function $H(f_1, \dots, f_N)$ is not as arbitrary as it may initially seem. As a matter of fact, it can be rigorously derived on the basis of the following three natural postulates only:

(i) $H(f, 1-f)$ is a continuous function of variable f.

(ii) If $H(f_1, \dots, f_n)$ is a symmetric function of its variables.

(iii) If one message in our pool of messages is split into two possible messages with certain frequency weights, then the corresponding weighted split holds true for the respective entropy. More precisely, if $f_N = g_1 + g_2 > 0$, then

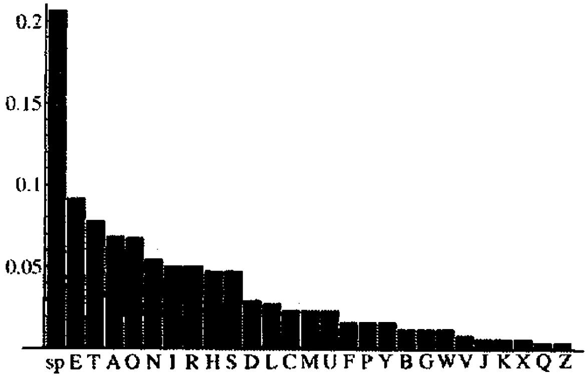
$$H(f_1, \dots, f_{N-1}, g_1, g_2) = H(f_1, \dots, f_N) + f_N H(g_1/f_N, g_2/f_N). \tag{9}$$

Mathematica Experiment 2. Approximations to English. Any English-language text is written in the alphabet of 27 symbols, 26 letters A,B,...,Z plus space sp. If we assumed, naively, that all the symbols appear with equal frequency, then, using the pseudo-random number generator, we could produce a simulated English text of length as follows

```
XFOML RXKHRJFFJUJ ZLP........
```

We could call it the *zero-order approximation* to the English language. If the text is, say, 200 letters long, the entropy (per symbol) is $H = \ln 27^{200}/200 = \ln 27 = 3.29584$. However, in the natural English language the frequencies of different letters are different, given, for a typical newspaper text, in the file LETTERFREQ on the UVW Web Site, and their histogram is shown below. Also, the Shannon entropy is computed.

```
In[1]:= Graphics'Graphics'
In[2]:= letterfreq={{sp,0.206},{E,0.091},{T,0.077},{A,0.068},
        {O,0.067},{N,0.054},{I,0.050},{R,0.050}, {H,0.047},{S,0.047},
        {D,0.029},{L,0.027},{C,0.023},{M,0.023},{U,0.023},{F,0.016} ,
        {P,0.016},{Y,0.016},{B,0.012},{G,0.012},{W,0.012},{V,0.008},
        {J,0.006},{K,0.006},{X,0.006},{Q,0.004},{Z,0.004}}
Out[2]= {{sp,0.206},{E,0.091}, ... ,{Z,0.004}}
In[3]:= Sum[letterfreq[[i]][[2]],{i,1,27}]
Out[3]= 1.
In[4]:= freq=Part[letterfreq[[Range[27],2]]]
Out[4]= {0.206, 0.091, ... , 0.004}
In[5]:= H=-Sum[ freq[[i]] Log[freq[[i]]], {i,1,27} ]
Out[5]= 2.84258
In[6]:= letters=Part[letterfreq[[Range[27],1]]]
Out[6]= {sp, E, T, A, O, N, I, R, H, S, D, L, C, M, U, F, P, Y, B,
                G, W, V, J, K, X, Q, Z}
In[7]:= BarChart[freq, BarLabels ->letters ]
Out[7]= -Graphics-
```



Thus, the frequency of E is .091, and the frequency of W is .012. A sample of an artificial text (let's call it the *first-order approximation* to the English language) produced with the help of a pseudo-random number generator is

```
OCRO HLI RGWR NMIELWIS .....
```

The entropy of such a text (per letter) is

$$H = -f_{sp}\ln f_{sp} - f_A \ln f_A - f_B \ln f_B - \ldots - f_Z \ln f_Z = 2.84258, \qquad (10)$$

less than for a uniformly random selection of letters. Although the sample visually feels more familiar than the zero-order approximation (one feels that some randomness has been removed from the text), it still does not look like an English text. To improve on our simulation of the English language, we would have to look at the blocks of two letters (all 27^2 of them) and their frequencies. These (and the frequencies for groups of three symbols) are given for the natural English language in *Secret and Urgent. The Story of Codes and Ciphers,* by Fletcher Pratt, Indianapolis-New York 1939, and reproduced on the attached UVW Web Site. Using them and the pseudo-random number generator produces a simulated *second-order approximation* to the English text

```
ON IE ANSOUTINYS ARE T INCTORE ST BE S DEAMY....
```

Looks better, doesn't it? It's entropy expressed by the formula

$$H = \frac{1}{2}(-f_{AA}\ln f_{AA} - f_{AB}\ln f_{AB} - \ldots - f_{ZY}\ln f_{ZY} - f_{ZZ}\ln f_{ZZ}). \qquad (11)$$

The *third-order approximation* would look like this:

```
IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID....
```

with the corresponding entropy H.

Then, instead of increasing the group size one can switch to the *first order word approximation* that would mimic the word frequency of the English language (given, e.g., in *Relative Frequency of English Speech Sounds* by G. Dewey, Harvard University Press, 1923), as in the simulated example

```
REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME
CAN DIFFERENT NATURAL HERE ...
```

or, in

```
THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH
WRITER THAT THE CHARACTER OF THIS POINT IS
THEREFORE ANOTHER METHOD FOR THE LETTERS.....
```

which gives a second-order word approximation.[2]

Remark 2.9.2 *Entropy vs. Complexity.* Shannon's entropy does not take into account the compressibility of information. Another measure of information, based on the algorithmic complexity, will address this issue in Chapter 4.

2.10 Experiments, exercises, and projects

1. Classify each of the data sets provided in Chapter 1 as categorical, numerical, mulitivariate, time dependent, etc.
2. *Mathematica Experiment: Telephone Rates.* Manipulate the phone rates data from Example 2.1.1 to order them by: (a) the international call rates, (b) the long distance national call rates. Produce the correspondingly ordered bar charts.

[2] The idea of different order approximations to the English language was borrowed from S. Shannon and W. Weaver, *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, 1949.

3. Explore *Mathematica* and write your own *Mathematica* formulas for various statistical functions:

   a. Sample means and weighted sample means.

   b. Ordered sample.

   c. Median, quartiles, and percentiles.

   d. Variance and standard deviation (biased and unbiased).

   e. Censored mean.

   f. Relative frequency distribution function.

   Test the above formulas and data representation techniques introduced in this chapter (whichever are appropriate) on data from the following examples (available as files on the `UVW Web Site`).

   A. Fragmentation bombs bases from Example 1.3.2.

   B. Positions of bright stars from Example 1.5.2.

   C. Time intervals between water drops from Example 1.5.1.

   D. Accelerometer data from Example 1.7.1.

   Notice that some these functions (and many others) are available as part of the *Mathematica* `Statistics` packages and our own `UVW` packages provided on the `UVW Web Site`. Compare your code with that of those packages.

4. Prove the Schwarz inequality: for any real numbers $x_1, \dots, x_n, y_1, \dots, y_n$,

$$\left(\sum_{i=1}^{n} x_i y_i\right)^2 \leq \sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2.$$

   Hint: Consider the nonnegative quadratic polynomial $\sum(x_i + \xi y_i)^2$in ξ and check its discriminant.

5. Write *Mathematica* formulas for

   a. sample covariance

   b. regression coefficients α and β

   c. scatter plot and regression line

   Test the above formulas to

   A. Determine relation between normal stress (x variable) and the shear resistance of soil (y variable) given the following data (in kN/m^2):

$x=$	11	13	15	17	19	21
$y=$	15.2	17.7	19.3	21.5	23.9	25.4

B. Test correlations between rates of sentenced prisoners from Section 1.6 for different states and regions of the United States. For each state (region), consider the rates for years 1971–1991 as a single 21-dimensional data. Analyze the whole data set from this perspective.

Compare your code with *Mathematica* `Statistics` and `UVW` packages. Use the latter for additional information.

6. Use data on mortality rates, volume, and waiting period for transplants of kidneys, liver, heart and lungs, and pancreas provided on the `UVW Web Site` to produce relevant scatter plots, 2-D histograms, correlations, and regression line. Draw conclusions.

7. Make a Q-Q plot of two selected data sets to judge similarity of their frequency distributions. You can use the single selected data set split into two subsets and test the frequency distribution of one part against the other. What are the implications of such an experiment?

8. Produce a more complete graph of the function $d_{cor}(\epsilon)$ from *Mathematica* Experiment 2.8.1.

9. Analyze the correlation dimension of the space shuttle accelerometer data set from Example 1.7.1.

10. Compare the Shannon entropy (use $\log_2$ base) of a message written in a four letter alphabet (say 00, 01, 10, 11) with letters appearing with the same frequencies 1/4, with that of a message written in the same alphabet but with letter frequencies 1/2, 1/4, 1/8, 1/8.

11. Check that the Shannon entropy function $H(f_1, \ldots, f_N)$ attains the maximum for $f_1 = \ldots = f_N = 1/N$. Remember that it is a constrained maximum of a function of N variables with the additional condition $f_1 + f_2 + \ldots + f_N = 1$. Verify Properties 1–3, and (i)–(iii) in Section 2.9.

12. Adjust the random number generator to simulate a sample text of 500 symbols (26 letters plus space) with

    (a) the equally distributed symbols

    (b) symbols distributed as in the natural English language

    (c) pairs of symbols distributed as in the natural English language

    (d) triples of symbols distributed as in the natural English language

    Use the frequencies provided on the `UVW Web Site` and coding from *Mathematica Experiment* 2.1.2 and 2.9.2. For each case, calculate the entropy per letter. Devise a method to produce such simulations if you were not given these frequency tables and had to get these frequencies by analysis of concrete texts.

13. *Mathematica Experiment. Entropy Olympics.* Calculate entropy of selected English, French, German and Spanish texts, a scanned picture, and classical or rock music. Could this information serve as a tool for the linguistic study of quantitative relationships between languages, or help decide which language was derived from which? Draw your own conclusions.

14. *Mathematica Experiment. Estimating Fractal Dimension via Linear Regression.* The formal definition of the Grassberger-Procaccia correlation dimension for general sets will be postponed to Section 6.5 (see, also, Section 7.4).

Let $S = (x_1, x_2, ..., x_n)$ be a finite data set of d-dimensional vectors. For a fixed (small) resolution $\epsilon > 0$, define the correlation sum

$$C^n(S,\epsilon) = C^n(\epsilon) = \frac{\#\{(x_i, x_j) : |x_i - x_j| < \epsilon, 1 \le i, j \le n\}}{n^2}, \qquad (1)$$

where $|x - y|$ denotes the d-dimensional distance of vectors x and y.

Assume, for the moment, that the correlation sum, as a function of ϵ, is of the form

$$C^n(\epsilon) = K\epsilon^\nu. \qquad (2)$$

Taking logarithms on both sides of (2), we obtain

$$\ln C^n(\epsilon) = \ln K + \nu \ln \epsilon. \qquad (3)$$

In other words, the relationship between $\ln C^n(\epsilon)$ and $\ln\epsilon$ is here linear, so, in view of formula (2.8.3), the coefficient ν is the correlation dimension of the finite set S. In reality, formula (2) can never be established rigorously, but can be taken as an approximation. In such a case, to estimate ν we can use the usual linear regression techniques developed in Section 2.7 (see, also, Chapter 8). This method forms a basis for the command `CorrelationDimension` included in the package `UVW‘DynSyst‘`.

Thus, we will proceed as follows: For any finite set of m (different) resolutions $\{\epsilon_1, \epsilon_2, \ldots, \epsilon_m\}$ chosen by the experimenter, consider the set of paired data

$$\{(\ln\epsilon_1, \ln C^n(\epsilon_1)), (\ln\epsilon_2, \ln C^n(\epsilon_2)), \ldots, (\ln\epsilon_m, \ln C^n(\epsilon_m))\}, \qquad (4)$$

and find the best linear fit for it. The slope $\hat{d}_{cor}$ of the regression line will be called the *correlation dimension of S based on resolutions* $\{\epsilon_1, \ldots, \epsilon_m\}$. From Section 2.7 we deduce that

$$\hat{d}_{cor} = \frac{\sum_{i=1}^{m} \ln\epsilon_i \ln C^n(\epsilon_i) - m\bar{\epsilon}\,\overline{\ln C^n(\epsilon)}}{\sum_{i=1}^{m} \epsilon_i^2 - m\bar{\epsilon}^2}, \qquad (5)$$

where

$$\bar{\epsilon} = \frac{1}{m}\sum_{i=1}^{m} \epsilon_i, \qquad \text{and} \qquad \overline{\ln C^n(\epsilon)} = \frac{1}{m}\sum_{i=1}^{m} \log C^n(\epsilon_i).$$

As an example, consider the 4-dimensional data set `iris` included on the `UVW Web Site`. The four components provide the lengths and widths of petals and sepals of the iris flower. In this experiment we will take just the first ten data points and estimate their correlation dimension using $m = 5$ and $\epsilon_i = i/10$, $i = 1, 2, \ldots, 5$. The relevant error analysis is discussed in Chapter 8.

```
In[1]:= <<Statistics‘LinearRegression‘
In[2]:= iris={ . . . . . . . . . . . . . . }
In[3]:= c[r_]:= (1/10.)^2 Sum[ Sum[ If[ Sum[
                 (iris[[i]][[k]]-iris[[j]][[k]])^2,
                 {k,1,4}]<r^2,1,0],
                 {j,1,10}],{i,1,10}]
In[4]:= c[.2]
Out[4= -0.16
In[5]:= reg= Table[{Log[i/10.], Log[c[i/10.]]}, {i,1,5}]
Out[5]= {{-2.30259, -2.30259}, {-1.60944, -1.83258},
        {-1.20397, -1.42712}, {-0.916291, -1.02165},
        {-0.693147, -0.1693147}}
In[6]:= Fit[reg, {1,x},x]
Out[6]= -0.125049 + 0.989057 x
In[7]:= Quit
```

So the correlation dimension of `iris` is estimated to be 0.989057. Try the same technique on the `drip` data from the *Mathematica* Experiment 2.8.1.

15. *Mathematica Experiment. Entropy of a Finite Data Set.* Let $S = (x_1, x_2, ..., x_n)$ be a finite data set which may consist of numbers, d-dimensional vectors, or some other abstract objects (e.g., strings of letters). Let $\{m_1, ..., m_k\}$ denote the set of different elements from this data set. Then, the entropy of S is

$$H = -\frac{1}{n}\sum_{i=1}^{k}\sum_{l=1}^{n}\mathbf{1}_{m_i}(x_l)\ln\frac{1}{n}\sum_{l=1}^{n}\mathbf{1}_{m_i}(x_l),$$

where $1_m(x) = 1$ if $x = m$, and 0 otherwise.

As an example, we will estimate the entropy of the data set `rivet` which can be found on the `UVW Web Site`. The data are first binned into bins of size 0.2 mm.

```
In[1]:= rivet={. . . . . .}
In[2]:= Min[rivet]
Out[2]= 13.13
In[3]:= Max[rivet]
Out[3]= 13.69
In[4]:= Length[rivet]
Out[4]= 200
In[5]:= freq[s_]:=
          (1/200.) Sum[If[(s-1)*.2<rivet[[i]]]]-13.119<
          s*.2, 1,0], {i,1,200}]
In[6]:= b=-Sum[freq[s] Log[freq[s]], {s,1,3}]
Out[6]=  0.91488
In[7]:= Quit
```

2.11 Bibliographical notes

A relatively new source on how to graph data, addressed mainly to social scientists, is

[1] G.T. Henry, *Graphing Data; Techniques for Display and Analysis*, Sage Publications, Inc., Thousand Oaks, London, 1995.

It contains an interesting analysis of the human perception of different methods of graphical data representation; see, also

[2] W.S. Cleveland and R. McGill, Graphical perception; Theory, experimentation, and application to the development of graphical methods, *J. Amer. Stst. Assoc.* 79(1984), 531-554.

[3] John W. Tukey, *Exploratory Data Analysis*, Addison-Wesley, Reading, MA, 1977.

is a classic of statistical literature and contains a wealth of information, both for a professional statistician and a practical user.

[4] E.R. Tufte, *Envisioning Information*, Graphic Press, Cheshire, CT, 1982, and

[5] E.R. Tufte, *The Visual Display of Quantitative Information*, Graphic Press, Cheshire, CT, 1990

contain a very imaginative exposition of how to represent complex information (not only statistical in nature). According to a review in the *Computer*: "A remarkable range of examples for the idea of visual thinking, with beautifully printed pages."

A nice mathematical introduction to the issues of fractional dimension can be found in

[6] G.A. Edgar, *Measure, Topology and Fractal Geometry*, Springer-Verlag, New York, 1990.

The February 1992 issue of the journal *Statistical Science* was devoted, in part, to the statistics of dynamical systems and their fractality, and included articles by S. Chatterjee and M. Yilmaz, by L.M. Berliner, and comments by other researchers working in the area. It is also a good source of more detailed references.

The small volume

[7] C. Shannon and W. Weaver, *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, 1949,

still remains a lucidly argued classic in the area. The idea of different order approximations to the English language was borrowed from it. For a more modern and more mathematical treatment of information theory, see, e.g.,

[8] A. Feinstein, *Foundations of Information Theory*, McGraw-Hill, New York, 1958.

[9] S. Guiasu, *Information Theory with Applications*, McGraw-Hill, New York, 1977.

The former contains a proof of Remark 2.9.1.

[10] Y. Bar-Hillel, *Language and Information*, Addison-Wesley, Reading, MA, 1964,

and also discusses issues of the semantic content of information. The frequency tables for the English language were taken from

[11] F. Pratt, *Secret and Urgent. The Story of Codes and Ciphers*, The Bobbs Merril Co., Indianapolis-New York, 1939.

[12] G. Dewey, *Relative Frequency of English Speech Sounds*, Harvard University Press, Cambridge, MA, 1923.

A discussion of mathematical models of word frequencies in a natural language can be found in

[13] B. Mandelbrot, On the theory of word frequencies and on related Markovian models of discourse, in *Structure of Language and Its Mathematical Aspects*, Amer. Math. Soc. , Providence, RI, 1961, pp. 190-219.

Chapter 3

Analytic Representation of Random Experimental Data

In *Mathematica* Experiment 2.5.1 we observed that the cumulative d.f. and histograms of large samples drawn with finer and finer resolution or, in other words, with smaller and smaller bin size, often seem to smooth out and assume a form that is almost begging to be compressed into a single analytic formula. These various idealized limit relative frequency d.f.s, called *probability density functions*, and the related *cumulative probability distribution functions*, will be studied in this chapter. We will also learn how to simulate data sets with an *a priori* given probability density function.

We begin by discussing stability of frequencies and fluctuations as *laws of nature* and then move on to analytic formulas for 1-D discrete probability distributions, introduce the concept of changing scale and location in data description, and then move on to probability densities for continuous data. A section on multivariate probability distributions, both discrete and continuous, follows. The analytic compression of fractal random objects is then briefly discussed.

3.1 Repeated experiments and the law of large numbers

Everybody has an intuitive notion of what a scientific experiment is. Our first step is to make this concept more precise.

Example 3.1.1 Galileo on the Leaning Tower of Pisa.

In order to measure the gravity constant, in 1627 Galileo performed a series of experiments, repeatedly dropping various bodies from the Leaning Tower of Pisa. Outcomes of these repeated experiments were not identical but they showed remarkable stability; a typical sequence of 10 measurements (in kg/m^3 units) could

M. Denker, W. Woyczynski, *Introductory Statistics and Random Phenomena*,
Modern Birkhäuser Classics, DOI 10.1007/978-3-319-66152-0_3

have looked as follows:

$$9.8102, \quad 9.8107, \quad 9.8098, \quad 9.8101, \quad 9.8109,$$

$$9.8092, \quad 9.8157, \quad 9.8131, \quad 9.8097, \quad 9.8095$$

The experiments produced random outcomes, the randomness arising from inaccuracies and other uncertainties of the experimental process, and the gravity constant was measured with what we now call *statistical errors*. They were relatively small.

Example 3.1.2 Accidents Happen (Randomly).
Monthly number of accidents at the Best Co. was recorded over a period of 20 months, producing the following data set:

$$3, 5, 7, 9, 10, 18, 6, 14, 11, 9, 5, 11, 15, 6, 11, 17, 12, 15, 8, 4.$$

Each monthly survey can be thought of as an experiment. So, the above data set represents outcomes of 20 experiments. Now, the data set no longer looks like a representation of a constant affected by statistical errors. The data set displays large statistical fluctuations.

Both of the above examples displayed some random behavior, although the mechanisms that produced them may have been of different nature.

Example 3.1.3 Double-Blind Medical Test.
A researcher studying acute leukemia would like to test the effectiveness of a new drug. His expectation is that the drug prolongs the duration of the illness' remission. For that purpose he performs a double-blind experiment: a random sample of 10 patients from the population diagnosed with the illness is split (again, randomly) into two equal groups. Patients are not aware of which group they were assigned. Then, those in the first group are given a dose of the new drug, while those in the second group are given a neutral placebo. This is being done to eliminate the so-called placebo effects (usually improvement) that the administration of any drug has on some patients. Then a physician, who also does not know to which group a particular patient belongs, questions the patients as to the duration of the remission. The collected data (in weeks) for five patients in each category are given below:

$$Placebo : 1, 22, 3, 12, 8 \qquad Drug : 10, 7, 32, 23, 22$$

Example 3.1.4 Drawing Balls at Random.
An urn contains balls of different colors: red, blue, yellow, etc. The experiment consists in drawing a ball at random and recording its color. "Drawing at random"

means here that the drawing mechanism is blind and does not give preference to any particular ball; the chance of any ball to be drawn is the same and equal to

$$\frac{1}{\text{number of balls in the urn}}$$

The outcomes of this experiment form a nominal categorical data set. We can transform it into an ordinal categorical data by labeling different colors with numbers $1, 2, \ldots$

In general, an experiment is performed on a physical "device" which produces data as an output. That physical "device" can be a measuring instrument used in a certain concrete situation, a pollster or a physician querying people, or a computer producing a string of numbers via its pseudorandom number generator. The data describing the outcome of an experiment, as we have seen in Chapter 2, can be either quantitative (numerical, vector, fractal) or qualitative (categorical). Another typical feature of the experiment is that, if repeated independently, it may yield a different set of data.

A finite set of d experiments (conducted simultaneously or consecutively), each yielding (say) numerical data, can be thought of as a single *grand* experiment producing d-dimensional vector data.

Symbolically, an experiment with random outcomes will be denoted by a capital letter, typically X, Y, Z, or $X_1, X_2, \ldots$, and called a *random quantity* if the outcomes are real numbers (resp. *random vectors, functions, fields, etc.* in other situations). In the case when outcomes are categorical, we will speak about *random entities*.

An independent n-fold repetition of an experiment described by the random quantity X results in a new *grand experiment* described by the random vector $\boldsymbol{X} = (X_1, X_2, ..., X_n)$. Single experimental random quantities $X_1, \ldots, X_n$ serve as components of $\boldsymbol{X}$. A particular run of a series of n experiments will produce a sequence of real numbers (vectors, etc.) $x_1, \ldots, x_n$, a concrete *realization* of independent random quantities $X_1, X_2, ..., X_n$.

The basic description of a random quantity X, i.e., the randomly varying outcomes of an experiment, is via, already encountered in Chapter 2, relative frequency distribution function measured over multiple (ideally, infinite) independent repetitions of the experiment under the same circumstances. The requirement that the experiments be *independently repeatable* is an important postulate in experimental sciences.

However, compared to our simple definition of the relative frequency d.f. for a fixed finite set $v_1, \ldots, v_N$, of possible experimental outcomes (Section 2.3) we will proceed slightly differently to permit analysis of experiments with any numerical outcomes. Our approach will be similar to that of the creation of a binned histogram with an arbitrary size and location of the bin. So, given a realization $x_1, \ldots, x_n$ of

n independent experiments involving the random quantity X, and an interval R on the real line, the real number

$$f_n(R; X) = \frac{\text{number of } x_i \in R}{n} = \frac{1}{n}\#\{i : x_i \in R\}, \tag{1}$$

tells us the relative frequency with which the outcomes of n independent experiments involving the random quantity X appear in the interval R. A similar quantity can be introduced in the case when each of the random quantities X_i is itself vector-valued. In that case, the interval R has to be replaced by a rectangular box in the space of appropriate dimensionality.

It is an empirical fact that for practically all intervals R, the relative frequencies $f_n(R; X)$ stabilize as the number n of independently repeated experiments increases. This phenomenon, which we will call the Stability of Frequencies Law (SFL), is a law of nature like any other law of nature one learns about in physics. It can be summarized as follows:

Stability of Frequencies Law (SFL). *Suppose that the experiment X is independently repeated and a sequence of outcomes $x_1, x_2, \ldots, x_n$, is observed. Then the relative frequencies f_n stabilize as n becomes large, i.e., there exists a real number* $\Pr(R; X)$ *such that*

$$f_n(R; X) \longrightarrow \Pr(R; X), \qquad \text{as} \quad n \to \infty, \tag{2}$$

for almost all intervals R. The limit $\Pr(R; X)$ *of the relative frequencies will be called the probability that the random quantity X takes values in R, and is also denoted* $\Pr\{X \in R\}$.

Notice that the above law was phrased somewhat cautiously and talks about the limit (2) existing only for "almost all" intervals R. The reason is that our measuring instruments are never perfectly precise and the condition $x_i \in R$, with its sharply defined cut-off points at the ends of the interval R, requires infinitely precise measurement to decide whether or not it is satisfied.

To avoid this difficulty one usually introduces a more practical concept of a bounded and smooth test function $\psi(x)$ which represents a realistic measuring device. In this context, the general Law of Tested Averages (LTA) can be formulated as follows:

Law of Tested Averages (LTA). *Suppose that an experiment X is independently repeated and a sequence of outcomes $x_1, x_2, \ldots, x_n, \ldots$ is observed. Then the averages of outcomes*

$$\mathrm{Av}_n(\psi(X)) := \frac{\psi(x_1) + \ldots + \psi(x_n)}{n}, \tag{3}$$

measured via a bounded and smooth test function $\psi(x)$ converge, as n increases, to a constant, say $\mu(\psi(X))$, i.e.,

$$\mathrm{Av}_n(\psi(X)) \longrightarrow \mu(\psi(X)), \qquad \text{as} \quad n \to \infty. \tag{4}$$

The limit $\mu(\psi(X))$ will be called the mean of random quantity X tested via the test function ψ.

Mathematica Experiment 1. Smooth Approximation of Discontinuous Test Function. The somewhat idealistic Stability of Frequencies Law can then be viewed as the special (in the limit) case of the Law of Tested Averages, where the test function is the *discontinuous* indicator function of the interval R:

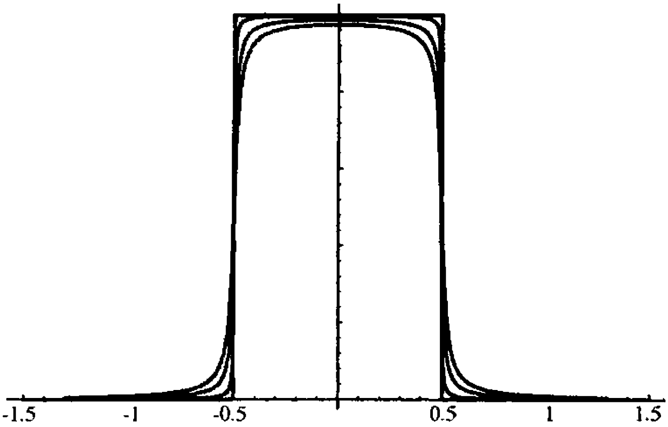
$$\mathbf{1}_R(x) := \begin{cases} 1, & \text{if } x \in R; \\ 0, & \text{if } x \notin R. \end{cases} \tag{5}$$

Indeed, with $\psi(x) = \mathbf{1}_R(x)$, the tested average (3) becomes the relative frequency in (1), i.e.,

$$f_n(R, X) = \frac{\mathbf{1}_R(x_1) + \ldots + \mathbf{1}_R(x_n)}{n}. \tag{6}$$

On the other hand, the indicator function $\mathbf{1}_R(x)$ can be approximated by smooth test functions ψ. In the *Mathematica* experiment that follows, we have selected $R = [-1/2, 1/2]$.

```
In[1]:= H[x_]:=If[x<0,0,1]
In[2]:= Indicator[x_]:=H[-(x-0.5)]*H[x+0.5]
In[3]:= Psi[x_,a_]:=(1/Pi)(Pi-((ArcTan[(a(x-0.5))]+Pi/2)+
                    (ArcTan[(-a(x+0.5))]+Pi/2)))
In[4]:= Plot[{Indicator[x], Psi[x,50],Psi[x,100],Psi[x,500]},
                                        {x,-1.3,1.3}]
Out[5]= -Graphics-
```



If the experimental outcomes $x_1, \ldots, x_n, \ldots$ come from a bounded interval (independent of n), as is often the case in practice, one can take as the test function $\psi(x) \equiv x$, the unboundedness of which at large xs being irrelevant. In this case, the Law of Tested Averages becomes the celebrated

Law of Large Numbers (LLN). *Suppose that an experiment X is independently repeated and a sequence of outcomes $x_1, x_2, \ldots, x_n, \ldots$ is observed. Then the sample means $\bar{x}$ of outcomes converge, as n increases, to a constant, say $\mu = \mu(X)$, i.e.,*

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n} \longrightarrow \mu = \mu(X), \qquad \text{as} \quad n \to \infty. \tag{7}$$

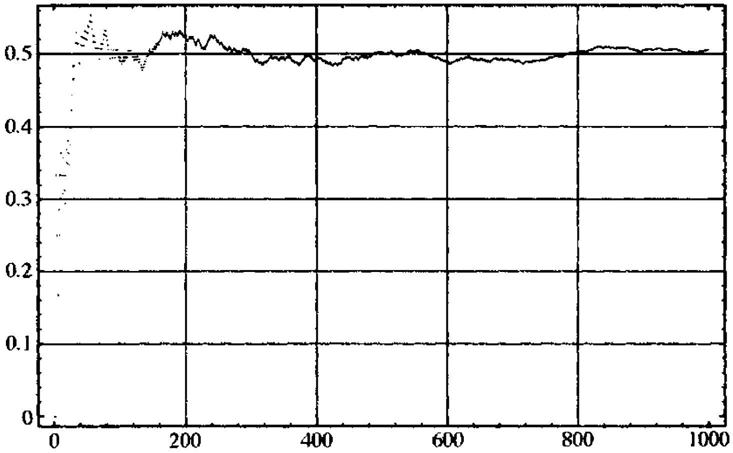
The constant $\mu(X)$ is called the mean of the random quantity X.

The above laws, introduced here as laws of nature, will be recovered as *mathematical theorems*, with precise assumptions, within the framework of Kolmogorov's axiomatic probability theory discussed in Chapter 5. They also permit a verification of independence of repeated experimental ensembles.

Remark 3.1.1 *Limitations on LLN.* There are some limitations on the applicability of the LLN. The *Mathematica* Experiment on Cauchy distributed random quantities in Section 3.8 shows an example of the situation where the LLN fails.

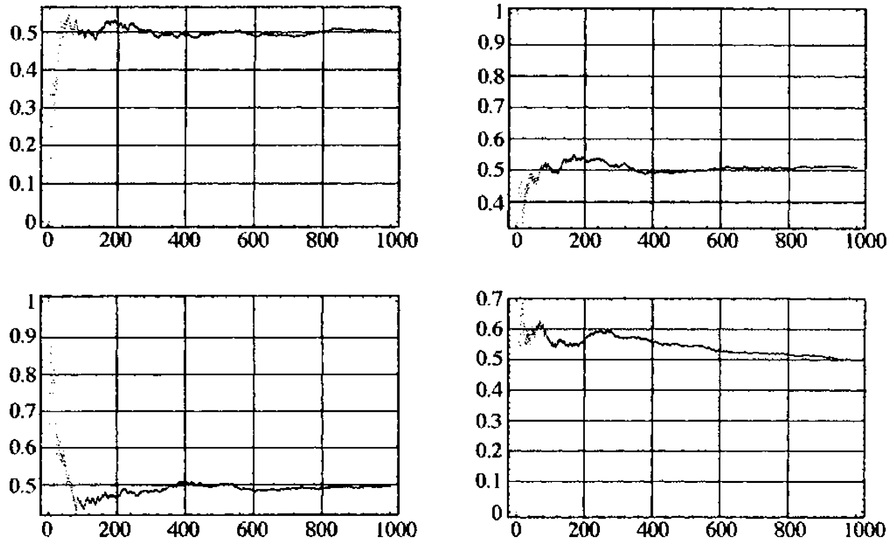
Mathematica Experiment 2. Law of Large Numbers. In this experiment we take the computer as a physical device that produces various random numerical outcomes. The command `Random[Integer]` produces a pseudorandom number equal to either 0 or 1. Then the command `Table[Random[Integer],{i,1,n}]` will produce a sequence of n pseudorandom zeros and ones. The command `SeedRandom[ ]` reseeds the pseudorandom number generator with the time of day (measured in small fractions of a second) to make sure that two different sessions will give different and hopefully independent pseudorandom strings. Finally the command `LargeNumbers[List]` of the `UVW'DataRep'` will plot the successive averages of the data from the `List`.

```
In[1]:= << UVW'DataRep'
In[2]:= SeedRandom[ ]
In[3]:= T1 =Table[Random[Integer],{i,1,1000}]
Out[3]= {0, 1, 1, 1, 0, . . . . , 1, 0, 1, 1, 0, 1, 0, 1, 1}
In[4]:= LargeNumbers[T1]
Out[4]= -Graphics-
In[5]:= S1=Show[%,Frame ->True, GridLines->Automatic]
Out[5]= -Graphics-
```



You will notice that the averages stabilize around 1/2 but the fluctuations around that value remain. Four repetitions of the same experiment give four different realizations of the data sets but the asymptotic behavior of their successive averages is similar.

```
. . . . . . . . . . . . . . . . .
In[16]:= LargeNumbers[T4]
Out[16]= -Graphics-
In[17]:= S4=Show[%,Frame ->True, GridLines->Automatic]
Out[17]= -Graphics-
In[18]:= Show[GraphicsArray[{{S4,S2},{S3,S1}}]]
Out[18]= -GraphicsArray-
```



Independent random quantities. Suppose two series of experiments X and Y were conducted with outcomes $x_1, \ldots, x_n, \ldots$ and $y_1, \ldots, y_n, \ldots$ According to the Stability of Frequencies Law, for almost all intervals R, S, the relative frequencies $f_n(R; X)$, $f_n(S; Y)$, for X, Y respectively, stabilize at probabilities $\Pr(R; X)$ and $\Pr(R; Y)$, but the same phenomenon also happens for these data considered as an outcome of a single 2-D vector experiment (X, Y) with outcomes $(x_i, y_i), i = 1, \ldots, n, \ldots$. In other words,

$$f_n(R \times S; (X, Y)) := \frac{\text{number of } (x_i, y_i) \in R \times S, 1 \le i \le n}{n}$$

$$\longrightarrow \Pr\{R \times S; (X, Y)\} \equiv \Pr\{(X, Y) \in R \times S\}, \tag{8}$$

as $n \to \infty$. Then, the relative frequency $f_n(R; X|S; Y)$ of outcomes of experiment X being in R given the extra information that the outcomes of experiment Y are in S (that is the fraction of pairs (x_i, y_i) in the rectangle $R \times S$ in the universe of pairs with the second coordinate $y_i \in S$) can be expressed via the formula

$$f_n(R; X|S; Y) = \frac{f_n(R \times S; (X, Y))}{f_n(S; Y)}. \tag{9}$$

The number $f_n(R; X|S; Y)$ is called the *conditional relative frequency of X being in R given Y being S.*

Now, to say that the experiment X with outcomes $x_1, \ldots, x_n$, is independent of experiment Y with outcomes $y_1, \ldots, y_n$, is equivalent to the statement that, for any intervals R and S, the conditional relative frequency $f_n(R; X|S; Y) \approx f_n(R; X)$, that is, it becomes independent of the extra information about the outcomes of Y. This independence condition can be rephrased, in view of definition (9), as a more symmetric independence condition

$$f_n(R \times S; (X, Y)) \approx f_n(R; X) f_n(S; Y). \tag{10}$$

If the number of experiments n increases, the Stability of Frequencies Law assures that the conditional frequencies $f_n(R; X|S; Y)$ stabilize (unless $\Pr(Y \in S) = 0$) at the conditional probabilities

$$\Pr(R; X|S; Y) = \Pr\{X \in R | Y \in S\} = \frac{\Pr\{X \in R, Y \in S\}}{\Pr\{y \in S\}}, \tag{11}$$

This leads to the following.

Criterion of Independence of Experimental Random Quantities. *The experimental random quantities X and Y are independent if*

$$\Pr(X \in R, Y \in S) = \Pr\{X \in R\} \cdot \Pr\{Y \in S\}, \tag{12}$$

for any intervals R and S.

Mathematica Experiment 3. Independence of Experiments. Let us check that four repetitions of experiments in the above *Mathematica* Experiments 2, which took advantage of the `SeedRandom[ ]` command, satisfy approximately the above criterion of independence. Since there are only four possible values for (X, Y), namely, $(0, 0)$, $(0, 1)$, $(1, 0)$, $(1, 1)$, it is not necessary to check the condition (12) [or, in practice, condition (10)] for all possible rectangles $R \times S$, and it suffices to verify that for any $i = 0, 1$ $j = 0, 1$,

$$\Pr(X = i, Y = j) \approx \Pr(X = i) \cdot \Pr(Y = j).$$

```
In[1]:= <<Statistics‘DataManipulation‘
In[2]:= SeedRandom[ ]
In[3]:= X =Table[Random[Integer],{i,1,1000}]
Out[3]= {1,1,1,0,1,0, ... , 0,1,1,1,0,0}
In[4]:= SeedRandom[ ]
In[5]:= Y =Table[Random[Integer],{i,1,1000}]
Out[5]= {1,1,0,1,1,0, ... , 1,1,0,0,1,0}
In[6]:= (XcomaY)=Table[{X[[n]],Y[[n]]},{n,1000}]
Out[6]= {{1,1},{1,1},{1,0},{0,1}, ... ,{1,0},{0,1},{0,0}}
In[7]:= FrX=Frequencies[X]
Out[7]= {{493, 0}, {507, 1}}
```

```
In[8]:= FrY=Frequencies[Y]
Out[8]= {{474, 0}, {526, 1}}
In[9]:= Fr(XcomaY)=Frequencies[XcomaY]
Out[9]= {{229, {0, 0}}, {264, {0, 1}}, {245, {1, 0}},
         {262, {1, 1}}}
In[10]:= PrXtimesPrY=
          N[ Table [{FrX[[m]][[1]]*FrY[[k]][[1]]/(10)^6,
           {FrX[[m]][[2]], FrY [[k]][[2]]}},{m,2},{k,2}],3]
Out[10]= {{0.234, {0, 0}}, {0.259, {0, 1.}},
           {0.240, {1., 0}}, {0.267, {1., 1.}}}
In[11]:= Pr(XcomaY)=N[Table[{(XcomaY)[[m]][[1]]/1000,
          (XcomaY)[[m]][[2]]},{m,4}]]
Out[11]= {{0.229, {0, 0}}, {0.264, {0, 1.}},
           {0.245, {1., 0}},  {0.262, {1., 1.}}}
```

A comparison of `Out[10]` and `Out[11]` indicates that the claim of independence of X and Y is relatively well founded. Note that whenever you repeat the above experiments the particular output is going to be different. However, the relevant probabilities will be similar.

3.2 Characteristics of experiments: distribution functions, densities, means, variances

In practice, instead of all the probabilities $\Pr(R; X)$ for all the possible intervals R, one often operates with a one-parameter family of probabilities of the events that the experimental random quantity $X \leq x$, i.e., with the *cumulative distribution function* of X:

$$F(x; X) := \Pr\{X \in R = (-\infty, x]\}, \qquad -\infty < x < \infty. \tag{1}$$

It is an analogue of the sample cumulative distribution function $F(x, \mathbf{x})$ introduced in Chapter 2. Note that $F(-\infty; X) = 0$, $F(+\infty; X) = 1$ and that $F(x; X)$ is nondecreasing. Sometimes, for the sake of better typography, we will write $F_X(x)$ instead of $F(x; X)$. A typical picture of a cumulative distribution function is shown in Fig. 3.2.1.

Now, for a pair of experimental random quantities represented by the random vector (X, Y), the independence condition (3.1.12) can also be written in terms of the distribution functions:

$$F_{(X,Y)}(x, y) = F_X(x)F_Y(y), \qquad x, y \in \mathbf{R}, \tag{2}$$

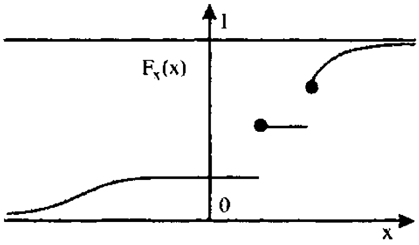


FIGURE 3.2.1

A typical picture of the cumulative d.f. $F_X(x)$ of a random quantity X. Notice that $F_X(-\infty) = 0$, $F_X(+\infty) = 1$ and that $F_X(x)$ is nondecreasing. Intervals of constancy as well as jumps upwards are also possible.

where the *joint 2-D distribution function*

$$F_{(X,Y)}(x, y) := \Pr\{X \leq x, Y \leq y\}. \tag{3}$$

Means of tested random quantities as integrals with respect to cumulative d.f. In the particular case of data $x_1, \dots, x_n$ uniformly distributed over the interval $[a, b]$, that is, with

$$\Delta x_i = x_i - x_{i-1} = \frac{b-a}{n}, \tag{4}$$

and for any bounded and continuous test function $\psi(x)$, it is clear that the tested average $\mathrm{Av}_n(\psi)$ defined in (3.1.3) is nothing but a discrete approximation to the Riemann integral of $\psi(x)$ over the interval $[a, b]$, or more precisely

$$\mathrm{Av}_n(\psi) = \frac{\psi(x_1) + \dots + \psi(x_n)}{n} = \frac{1}{b-a} \sum_{i=1}^{n} \psi(x_i)\, \Delta x_i \tag{5}$$

which converges, as $n \to \infty$, to

$$\mu(\psi) = \frac{1}{b-a} \int_a^b \psi(x)\, dx. \tag{6}$$

Also, in this case, the distribution function $F(x)$ grows linearly in the interval $[a, b]$, i.e.,

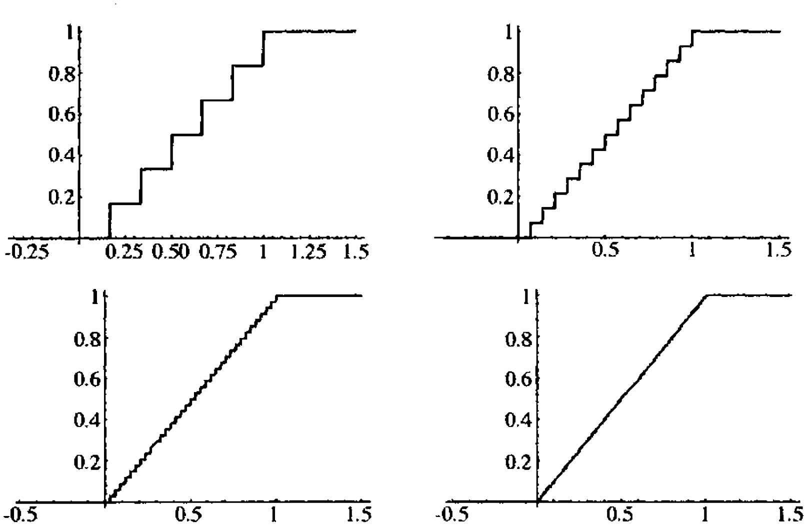
$$F(x) = \begin{cases} 0, & \text{for } -\infty < x \leq a; \\ (x-a)/(b-a), & \text{for } a \leq x \leq b; \\ 1, & \text{for } b \leq x < +\infty, \end{cases} \tag{7a}$$

with

$$dF(x) = \begin{cases} 0, & \text{for } -\infty < x \leq a; \\ dx/(b-a), & \text{for } a \leq x \leq b; \\ 0, & \text{for } b \leq x < +\infty. \end{cases} \tag{7b}$$

Hence, we can symbolically write

$$\mu(\psi) = \int_{-\infty}^{\infty} \psi(x)\, dF(x). \tag{8}$$



Mathematica Experiment 1. Cumulative d.f. for Uniform Data. We will illustrate the limit passage for (5) to (6) using the tools developed in *Mathematica* Experiment 2.5.1. The interval $[a, b]$ is taken to be $[0, 1]$.

```
In[1]:= <<Statistics‘DescriptiveStatistics‘
In[2]:= <<Graphics‘Graphics‘
In[3]:= n=6;
In[4]:= uniformdata= (1/n)Range[n]
Out[4]= {1/6, 1/3, 1/2, 2/3, 5/6, 1}
In[5]:= H[x_]:=If[x<0,0,1]
In[6]:= F[x_]:=(1/n) Sum[H[x- uniformdata[[i]] ],{i,1,n}]
In[7]:= cdf6=Plot[F[x],{x,uniformdata[[1]]-1/2,
        uniformdata[[n]]+1/2 } ]
```

Repeating the procedure for, say, $n = 14, 34$, and 83, and using the command

```
In[23]:= Show[GraphicsArray[{{cdf6, cdf14}, {cdf34, cdf83}}]]
```

produces the `GraphicsArray` shown above.

The above discussion suggests that, perhaps, for any experimental random quantity X and a bounded continuous test function ψ we could view the tested mean as a kind of integral with respect to the cumulative d.f. $F_X(x)$:

$$\mu(\psi(X)) = \lim_{n\to\infty} \mathrm{Av}_n(\psi) = \lim_{n\to\infty} \sum_{j=1}^{n} \psi(x_{(j)} \Delta F(x_{(j)}) = \int_{-\infty}^{\infty} \psi(x)\, dF_X(x). \tag{9}$$

The integral on the right-hand side, called the *Riemann-Stieltjes integral* of ψ with respect to F_X, is defined via the limit on the left-hand side whose existence and uniqueness (that is independence of any particular realization of a series of repeated independent experiments X) is assured by the Law of Tested Averages. In the intermediate discrete approximation formula, of course,

$$\Delta F(x_{(i)}) = F(x_{(i)}) - F(x_{(i-1)})$$

corresponds to the jump of the data $\boldsymbol{x}$ cumulative d.f. $F(x, \boldsymbol{x})$ at the data point $x = x_{(i)}$.

In the above sense, the cumulative d.f. $F_X(x)$ of an experimental random quantity X acts on test functions ψ as an operation

$$\psi \longmapsto \mu(\psi(X)) = \int_{-\infty}^{\infty} \psi(x)\, dF_X(x) \tag{10}$$

which enjoys the following properties:

(i) It is positive, i.e.,

$$\psi(x) \leq 0 \qquad \Longrightarrow \qquad \int_{-\infty}^{\infty} \psi(x)\, dF_X(x) \leq 0; \tag{11}$$

(ii) It scales homogeneously, i.e., for any real number a,

$$\int_{-\infty}^{\infty} a\psi(x)\, dF_X(x) = a \int_{-\infty}^{\infty} \psi(x)\, dF_X(x) \tag{12}$$

(iii) It is additive on superpositions of test functions, i.e.,

$$\int_{-\infty}^{\infty} (\psi_1(x) + \psi_2(x))\, dF_X(x) = \int_{-\infty}^{\infty} \psi_1(x)\, dF_X(x) + \int_{-\infty}^{\infty} \psi_2(x)\, dF_X(x). \tag{13}$$

Thus, the Riemann-Stieltjes integral $\int \psi \, dF_X$ is just the *mean* of the experimental random quantity X tested by the test function ψ. For the special choice of the test function $\psi(x) \equiv x$,

$$\mu(X) = \int x \, dF_X(x) \tag{14}$$

is simply called the *mean* of the experimental random quantity X [or, its cumulative d.f. $F_X(x)$]. Notice that since $\psi(x) = x$ is an unbounded test function, the existence of the mean is not guaranteed for every cumulative d.f. $F_X(x)$. It has to be understood as the limit of $\int x \mathbf{1}_{[a,b]}(x) \, dF_X(x)$, for $a \to -\infty, b \to +\infty$, which may or may not exist. The indicator function $\mathbf{1}_{[a,b]}(x)$ is equal to 1 for xs inside interval $[a, b]$, and 0 outside that interval.

In this context, the probability

$$\Pr\{a < X \le b\} = F_X(b) - F_X(a), \tag{15}$$

and the variance of an experimental random quantity X with the cumulative d.f. $F_X(x)$ is defined by the formula

$$\sigma^2(X) = \int_{-\infty}^{\infty} (x - \mu(X))^2 \, dF_X(x), \tag{16}$$

expressing the mean square deviation of the experimental random variable X from its mean $\mu(X)$. Observe that, in general, finiteness of the variance is not guaranteed either. Using the above properties *(i-iii)* of the Riemann-Stieltjes integral one can easily check that

$$\operatorname{Var} X = \sigma^2(X) = \int x^2 \, dF_X(X) - \mu^2(X). \tag{17}$$

The quantity

$$\mu_2(X) = \int x^2 \, dF_X(x) \tag{18}$$

is called the second order moment of the random quantity X, or, equivalently, of its cumulative d.f. F_X. By analogy, the k-th order moment are

$$\mu_k(X) = \int x^k \, dF_X(x). \tag{19}$$

All of these characteristics of the cumulative distribution functions (and thus experimental random quantities) are analogous to the corresponding finite numerical data characteristics introduced in Chapter 2. They will be revisited within the framework of the formal mathematical model of probability theory in Chapter 5.

For the purposes of the present discussion of analytic representations of random experimental quantities, we will note two classes of cumulative d.f.s. Examples from these two classes will fill the rest of this chapter.

Discrete Distributions. Let us assume that the experimental random quantity X assumes values only from a fixed finite (or infinite) discrete set $v_1, \ldots, v_N$ of real numbers, and that in the series of n independently repeated experiments the outcomes form a random sample $x_1, \ldots, x_n$ with relative frequencies

$$f_n(v_1; X) = \frac{\#\{i : x_i = v_1\}}{n}, \quad \ldots, \quad f_n(v_N; X) = \frac{\#\{i : x_i = v_N\}}{n}. \tag{20}$$

Then, by the Stability of Frequencies Law, as $n \to \infty$,

$$f_n(v_1; X) \to \Pr\{X = v_1\} = p_1, \ldots, f_n(v_N; X) \to \Pr\{X = v_N\} = p_N,$$

and by the Law of Tested Averages, for any test function ψ

$$\mathrm{Av}_n(\psi(X)) \to p_1\psi(v_1) + \ldots + p_N\psi(v_N) = \int \psi(x)\, dF_X(x), \tag{21}$$

where the cumulative d.f. $F_X(x)$ can now be identified as a function constant at all points x except $x = v_i$ where it has jumps upwards of size p_i. Such random quantities and their cumulative d.f.s are called *discrete*.

For discrete random quantities, the Riemann-Stieltjes integral formulas for mean, variance, moments, etc. become just sums (finite or infinite). For example,

$$\mu = \mu(X) = \sum_{i=1}^{N} v_i p_i, \quad \sigma^2(X) = \sum_{i=1}^{N} (v_i - \mu)^2 p_i. \tag{22}$$

Absolutely Continuous Distributions, Densities. In some cases the cumulative d.f. $F_X(x)$ is differentiable (at all except, say, some discrete points) and

$$F_X(x) = \int_{-\infty}^{x} f_X(y)\, dy. \tag{23}$$

Such cumulative d.f.s are called *absolutely continuous* and their derivatives

$$f_X(x) = \frac{dF_X(x)}{dx} \tag{24}$$

are called the *probability density function (d.f.)* of the random quantity X. In view of the monotonicity of $F_X(x)$ and the property $F_X(+\infty) = 1$ of the cumulative d.f., any density function $f_X(x)$ must satisfy the following two properties

(i) Positivity:

$$f_X(x) \geq 0, \qquad x \in \mathbf{R}, \tag{25}$$

(ii) Normalization:

$$\int_{-\infty}^{\infty} f_X(x)\, dx = 1. \tag{26}$$

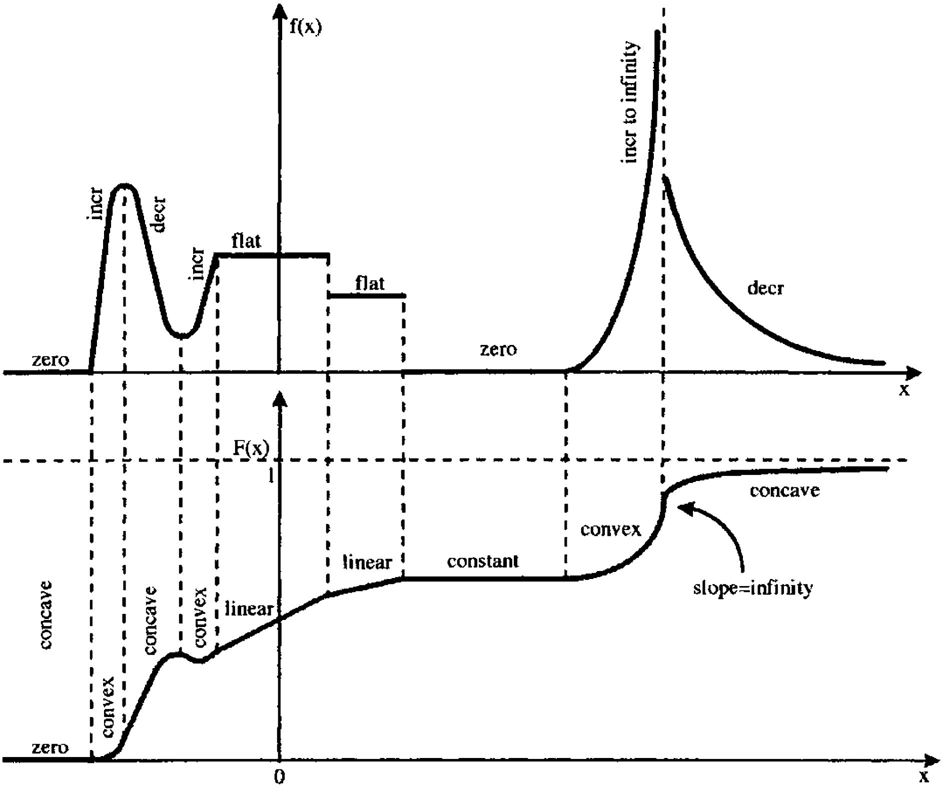


FIGURE 3.2.2

A typical graph of a probability density function $f_X(x)$ (above) and the corresponding cumulative distribution function $F_X(x)$ (below). Notice that the density can have intervals where it is zero and singular points where it is infinity.

For an absolutely continuous distribution $F_X(x)$, the Stieltjes integral formulas for probability, mean, variance, moments, etc., become just the usual Riemann integrals, because for them $dF_X(x) = f_X(x)\,dx$. For example,

$$\Pr\{a < X \le b\} = \int_a^b f_X(x)\,dx, \tag{27}$$

and

$$\mu = \mu(X) = \int_{-\infty}^{\infty} x f_X(x)\,dx, \quad \sigma^2(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)\,dx. \tag{28}$$

Concrete examples of these types of distributions will be provided in the remainder of this chapter. Also, not surprisingly, there are distribution functions that are mixtures of discrete and absolutely continuous distribution functions. A more striking fact is the existence of continuous cumulative d.f.s of the "devil's staircase" type, which are not absolutely continuous. We will discuss them briefly in Chapter 5.

Calculation of tested means, especially using *Mathematica*, can sometimes be simplified. *If $F(x)$ is a cumulative d.f. concentrated on* $[0,\infty)$ *(i.e.,* $F(0-)=0$*), and the test function $\psi(x)$ is continuously differentiable, and either $\psi(0)=0$ or $F(0-)=F(0)$, then*

$$\int \psi(x)F(dx) = \int_0^{\infty} (1-F(x))\psi'(x)\,dx. \tag{29}$$

We just check this useful identity for an absolutely continuous cumulative d.f. $F(x)$ with the density $f(x)$. Then, since $d(1-F(x))/dx = -f(x)$, integrating by parts we get

$$\int \psi(x)F(dx) = \int_0^{\infty} \psi(x)f(x)dx = [-(1-F)\psi]_0^{\infty} + \int_0^{\infty} \psi'(x)(1-F(x))dx.$$

Often, we speak generically of the *probability distribution of a random experimental quantity* X (or, simply, the *distribution of* X), by which we mean either the cumulative d.f. $F_X(x)$, or the discrete probabilities p_i of values v_i taken by X, or the probability density function $f_X(x)$, whichever is appropriate or handy in any given case.

Inverse Distribution Function; Quantile Function. If a cumulative d.f. $F(x)$ is strictly increasing, then there is obviously a function G satisfying $G(F(u)) = u = F(G(u))$. Function G is called the inverse function of F and is denoted by $F^{(-1)}$. Similarly to the sample quantiles $q(\alpha)$ discussed in Section 2.2 for sample cumulative d.f.s, $F^{-1}(\alpha)$ is called the α-th quantile of the distribution function F, and F^{-1} is called the *quantile function*. Number $F^{-1}(1/2)$ is called the *median* and

$F^{-1}(1/4)$, $F^{-1}(2/4)$, $F^{-1}(3/4)$ are called the *first, second, and third quartiles*, respectively. For cumulative d.f.s that are not strictly increasing, one also defines a generalized inverse function (quantile) following the ideas explained in Section 2.2, e.g, via the formula

$$F^{-1}(z) = \max\{F(x) : F(x) \leq z\}, \qquad 0 \leq z \leq 1. \tag{30}$$

3.3 Uniform distributions, simulation of random quantities, the Monte Carlo method

A random quantity U is said to have a *continuous uniform distribution* on the interval $[a, b]$ if its density has the form

$$f_U(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{for } x \text{ outside the interval } [a, b]. \end{cases} \tag{1}$$

There are two parameters a and b, $a < b$. The corresponding cumulative d.f.

$$F_U(x) = \begin{cases} 0 & \text{for } x \leq a; \\ \frac{x-a}{b-a} & \text{for } a < x < b; \\ 1 & \text{for } b \leq x. \end{cases} \tag{2}$$

A simple calculation shows that the mean

$$\mu(U) = \frac{a+b}{2}, \tag{3}$$

and the variance

$$\sigma^2(U) = \frac{1}{b-a}\int_a^b \left(x - \frac{a+b}{2}\right)^2 dx = \frac{(b-a)^2}{12}. \tag{4}$$

In this case, the standard deviation σ is simply $1/(2\sqrt{3})$ times the size of the interval $b - a$; a good illustration of the intuitive meaning of the notion of variance.

A random quantity U is said to have a *discrete uniform distribution* on values $v_1, \ldots, v_N$ if

$$\Pr\{U = v_1\} = \ldots = \Pr\{U = v_N\} = 1/N. \tag{5}$$

The continuous uniform density and the corresponding cumulative d.f. are pictured in Fig. 3.3.1.

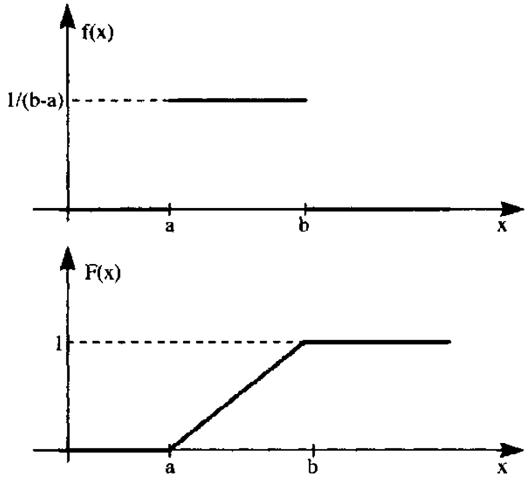
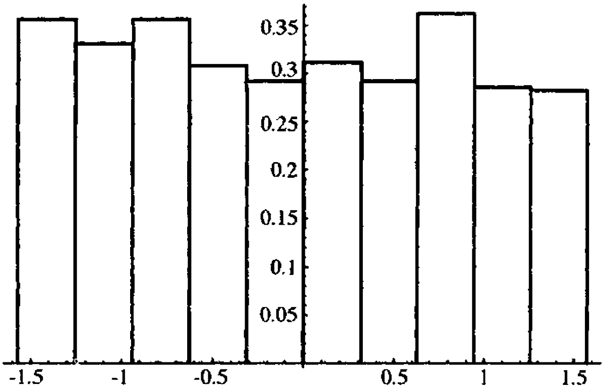


FIGURE 3.3.1
Uniform density with parameters a, b, and the corresponding cumulative d.f.

The problem of generation of the simulated data with a prescribed probability distribution is of fundamental importance in any computer-aided study of random phenomena. We have already discussed the pseudorandom number generators which produce sequences of zeros and ones with (almost) identical distribution of frequencies of blocks of different (reasonable) length. Considering these blocks to be binary representations of numbers in the interval (0, 1), we thus obtain a way to generate data with the uniform distribution in that interval. As a matter of fact, this is a function that is explicitly provided in *Mathematica* via command `Random[ ]`.

Mathematica Experiment 1. Uniformly Distributed Pseudorandom Numbers. In Experiments of Section 1 we have seen how to generate uniformly distributed pseudorandom integers in the discrete set $\{1, 2, \ldots, n\}$. The command `Random[ ]` generates a pseudorandom number between 0 and 1, `Random[Real, { xmin, xmax }]` produces a pseudorandom number between `xmin` and `xmax`. The `UVW'DataRep'` package command `RegularHisto[ list, xmin, xmax, nx]` produces a histogram of the data contained in the `listofdata` with `nx` bins between `xmin` and `xmax`.

```
In[1]:= <<UVW'Datarep'
In[2]:= T1=Table[Random[Real,{-N[Pi/2],N[Pi/2]}],{1000}]
Out[2]= {1.45002, 0.429825, 0.417296, -1.0639, ... ,
              0.92243, 0.0658026, -1.25003, 0.20874}
In[3]:= RegularHisto[T1,-N[Pi/2],N[Pi/2], 10]
Out[3]= -Graphics-
```



Now, assume that we have generated a pseudorandom data set

$$u_1, u_2, \dots, u_n, \tag{6}$$

approximately uniformly distributed in the unit interval $(0, 1)$ and we want to produce a simulated data set with a given cumulative d.f. $F(x)$. The idea of how to proceed is suggested by the discussion in Sections 2.2, 2.4, and 3.2 on quantile functions and their relationship with the cumulative distribution functions. Indeed, if $F^{-1}(u)$, $u \in (0, 1)$, denotes the (generalized) inverse of the cumulative distribution function $F(x)$, then the transformed data set

$$x_1 = F^{-1}(u_1),\ x_2 = F^{-1}(u_2), \dots, x_n = F^{-1}(u_n), \tag{7}$$

will have the cumulative relative frequency distribution $F(x)$. Indeed,

$$\frac{\#\{i : x_i \leq x\}}{n} = \frac{\#\{i : F^{-1}(u_i) \leq x\}}{n} = \frac{\#\{i : u_i \leq F(x)\}}{n} \approx F(x)$$

in view of the monotonicity of the cumulative distribution function $F(x)$ and its inverse, and the uniform distribution on $(0, 1)$ of the data $u_1, \dots, u_n$.

The above observation provides an obvious algorithm for simulation random data with prescribed probability distribution and is implemented in *Mathematica* packages `Statistics` and `UVW`. It is also the basis of the so-called Monte-Carlo method of numerically calculating integrals over very complex and high-dimensional domains. In practice, one can also replace the above simple algorithm with more sophisticated numerical methods that provide faster convergence of simulated data histograms to theoretical probability densities.

The simulation of a sequence $X_1, \dots, X_n$ of independent random experimental quantities can then be accomplished via the general:

Monte-Carlo Law: *If $U_1, \ldots, U_n$ are independent uniformly distributed on $[0, 1]$ random quantities, and if F is a given cumulative d.f., then the random quantities*

$$X_1 = F^{-1}(U_1), \ldots, X_n = F^{-1}(U_n)$$

are independent, each with the cumulative d.f. F.

The Monte-Carlo Law is easily verified, say, for $n = 1$, as follows: Let x be an arbitrary real number. Then

$$F_X(x) = \Pr\{X \leq x\} = \Pr\{F^{-1}(U) \leq x\}$$

$$= \Pr\{U \leq F(x)\} = F(x),$$

since F is monotone (the third equality), and U is uniformly distributed on $[0, 1]$ (the fourth equality). For $n = 2, 3, \ldots$, the claim is proved similarly by noticing that the independence of U_1 and U_2 implies the independence of $F^{-1}(U_1)$ and $F^{-1}(U_2)$.

Concrete examples of the above procedures will be given in the next few sections. Also, remember that in practical applications the independence can be simulated by the use of the *Mathematica* `SeedRandom[ ]` command, see Section 3.1.

3.4 Bernoulli and binomial distributions

The family of *Bernoulli distributions* describes discrete random quantities X with only two possible values: $v_0 = 0$ and $v_1 = 1$, appearing with probabilities and $f_X(1) = \Pr\{X = 1\} = p$ and $f_X(0) = \Pr\{X = 0\} = 1 - p$. This can be written as a formula

$$f_X(v) = \begin{cases} 1 - p, & \text{if } v = 0; \\ p, & \text{if } v = 1. \end{cases} \tag{1}$$

The mean of the Bernoulli distribution is

$$\mu(X) = 0 \cdot f_X(0) + 1 \cdot f_X(1) = 0 \cdot (1 - p) + 1 \cdot p = p, \tag{2}$$

and its variance

$$\sigma^2(X) = (0 - p)^2 \cdot (1 - p) + (1 - p)^2 \cdot p = p(1 - p). \tag{3}$$

The family depends on only one parameter p, $0 \leq p \leq 1$. If there is reason to suspect that neither 0 nor 1 are favored in the data, i.e., $p = 1/2$, then the distribution is called *symmetric* Bernoulli distribution.

Let us now perform the following experiment. Toss a fair coin ($p = 1/2$) n times. The i-th toss is described by the symmetric Bernoulli random quantity X_i and the random quantities $X_1, \ldots, X_n$, are independent. Suppose that each time the coin comes up heads ($X_i = 1$) you win one dollar and you win nothing if it comes up tails ($X_i = 0$). Your total win after n tosses is a random quantity

$$S_n = X_1 + \ldots + X_n, \tag{4}$$

and we are interested in determining its probability distribution.

For a small n, say $n = 2$, we clearly see that in two tosses there are four possible outcomes $(X_1, X_2) \in \{(0,0), (0,1), (1,0), (1,1)\}$ and because of the coin's symmetry they are all equally probable with probability 1/4. Hence, the distribution of the sum S_2 is easily determined:

$$\Pr\{S_2 = 0\} = \Pr\{(X_1, X_2) = (0,0)\} = 1/4$$

$$\Pr\{S_2 = 1\} = \Pr\{(X_1, X_2) = (0,1)\} + \Pr\{(X_1, X_2) = (1,0)\} = 1/2$$

$$\Pr\{S_2 = 2\} = \Pr\{(X_1, X_2) = (1,1)\} = 1/4$$

Of course, such a pedestrian approach will have to be adjusted if we are to solve our problem for larger ns. Before we approach it theoretically let us conduct the following.

Mathematica Experiment 1. Repeated Bernoulli Experiments. Our basic experiment consists of tossing the fair coin $n = 7$ times. Repeat this experiment independently N times. The raw outcome of the experiment is a sequence

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N,$$

where each sample point $\boldsymbol{x}_i$, $i = 1, \ldots, N$, has the structure of a 7-dimensional vector:

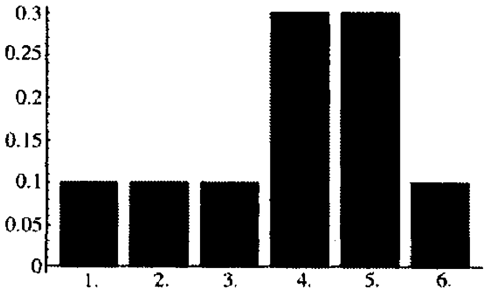
$$x_i^1, x_i^2, \ldots, x_i^7,$$

where x_i^js are either 0 or 1. By the Stability of Frequencies Law, we can approximate the distribution of S_7 by the relative f.d. of

$$s_i = x_i^1 + x_i^2 + \ldots + x_i^7, i = 1, 2, \ldots, N,$$

which simply sum all your wins in each basic experiment. The possible values of Y_i are clearly nonnegative integers from 0 to 7. We take this opportunity to introduce some simple *Mathematica* programming. Let us begin with $N = 10$ repetitions of our basic 7-toss experiment. We use `rn` (repetition number) instead of `N` because the latter is a protected symbol in *Mathematica*.

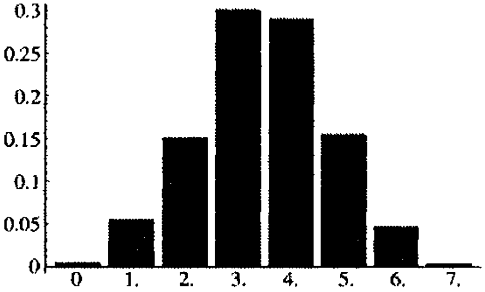
```
In[1]:= <<Graphics'Graphics'
In[2]:= <<Statistics'DataManipulation'
In[3]:= <<Statistics'DescriptiveStatistics'
In[4]:= <<Statistics'DiscreteDistributions'
In[5]:= n=7
Out[5]= 7
In[6]:= rn=10
Out[6]= 10
In[7]:= Do[
         r7=Table[{0},{rn}];
         For[i=1, i<=rn, i++,  r7[[i]]=Table [Random[Integer],{n}]]
         ]
In[8]:= r7
Out[8]= {{1, 0, 0, 1, 1, 0, 1}, {1, 1, 0, 0, 1, 0, 0},
         {0, 0, 1, 0, 0, 1, 0}, {0, 0, 1, 0, 1, 1, 0},
         {0, 0, 1, 1, 1, 1, 0}, {0, 1, 1, 0, 0, 1, 1},
         {1, 1, 1, 1, 1, 1, 0}, {1, 1, 0, 1, 1, 1, 1},
         {1, 0, 0, 0, 0, 0, 0}, {1, 0, 0, 0, 1, 1, 1}}
In[9]:= Do[
         r7sum=Table[{0},{rn}];
         For[i=1,i<=rn, i++,  r7sum[[i]]=Apply[Plus,r7[[i]]]]
         ]
In[10]:= r7sum
Out[10]= {5, 4, 5, 5, 6, 3, 4, 4, 1, 2}
In[11]]:= freq7=Frequencies[r7sum]
Out[11]= {{1, 1}, {1, 2}, {1, 3}, {3, 4}, {3, 5}, {1, 6}}
In[12]:= relfreq7=N[{Column [freq7,1]/rn,Column [freq7,2] }]
Out[12]= {{0.1, 0.1, 0.1, 0.3, 0.3, 0.1}, {1., 2., 3., 4., 5., 6.}}
In[13]:= hist7=BarChart[Transpose[relfreq7]]
Out[13]= -Graphics-
```



So, after 10 independent repetitions of our basic 7-toss experiment we obtained a relative frequency d.f. which, however, does not show any regularities; nothing to write home about. Moreover, the possible values 0 and 7 have not appeared among our 10 repetitions at all. Obviously, to approximate the distribution of S_7

well, to take advantage of the SFL, we need many more repetitions. Redoing the above *Mathematica* session with `rn=1000` (of course, you can leave steps `In[8]` and `In[10]` out) we obtain the following much more symmetric relative frequency d.f.

```
In[22]:= relfreq7=N[{Column [freq7,1]/rn,Column [freq7,2] }]
Out[22]= {{0.004, 0.054, 0.15, 0.3, 0.289, 0.154, 0.046, 0.003},
          {0, 1., 2., 3., 4., 5., 6., 7.}}
In[23]:= hist7=BarChart[Transpose[relfreq7]]
Out[23]= -Graphics-
```



Now, let us try to discover analytically a formula for the probability distribution of the random quantity S_n defined in (4). The derivation will be based on the assumption that in each n-toss series, all the possible outcomes are equally likely. Since there are 2^n strings of n 0s and 1s of length n, the probability of each

$$\Pr(X_1 = x_1, \dots, X_n = x_n) = \frac{1}{2^n}, \qquad x_i = 0 \quad \text{or} \quad 1. \tag{5}$$

Among all of these strings there are $\binom{n}{k}$ strings which have exactly k 1s, as that is the number of ways in which you can choose k sites out of n positions. This gives the probability of a string with exactly k 1s to be

$$\Pr(S_n = k) = \binom{n}{k}\frac{1}{2^n}, \qquad k = 0, 1, 2, \dots, n, \tag{6}$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{7}$$

is the binomial coefficient.

If the coin is biased, with the Bernoulli probability p of 1 appearing in each toss, then via a similar reasoning, the probability of winning k dollars in n tosses becomes

$$\Pr(S_n = k) = b(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, 2, \ldots, n. \tag{8}$$

Not surprisingly, the above probability distribution is called the *binomial distribution.* Because of the binomial formula (see Experiments, Exercises, and Projects), the above binomial probability distribution satisfies the normalization condition

$$\sum_{k=0}^{n} b(k; n, p) = 1. \tag{9}$$

Notice that the binomial distribution has two parameters: the Bernoulli probability $p, 0 \leq p \leq 1$, and the integer parameter n. Sample points from the population with the binomial distribution $b(k; n, p)$ can take values $k = 0, 1, 2, \ldots, n$ with probabilities given by formula (8).

The above reasoning can be summarized in the following.

Binomial Principle. *If n binary experiments, i.e., each of them with two possible outcomes (success/failure), are performed independently, then the probability p_k of exactly k successes is*

$$p_k = b(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, 2, \ldots, n,$$

where p is the probability of success in a single trial.

The mean μ of the population with the binomial distribution $b(k; n, p)$ is

$$\mu(S_n) = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} \tag{10}$$

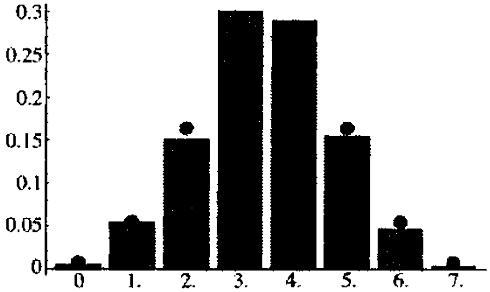
$$= np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!((n-1)-(x-1))!} p^{x-1} (1-p)^{(n-1)-(x-1)} = np$$

in view of the same binomial formula we used before. The formula (10) also immediately follows from the definition (4) of S_n as the sum of n Bernoulli random quantities, each with mean value p. Similarly, one obtains the sample variance

$$\sigma^2(S_n) = np(1-p). \tag{11}$$

Mathematica Experiment 1 Continued. Repeated Bernoulli Experiments. The package `Statistics‘DiscreteDistributions‘` contains the command `BinomialDistribution[n,p]`. We will compare it with the experimental data obtained earlier.

```
In[24]:= bdist7=BinomialDistribution[7,p]
Out[24]= BinomialDistribution[7, 0.5]
In[25]:= tablebdist7=N[Table[PDF[bdist7, x], {x,0,7}], 3]
Out[25]= {0.00781, 0.0547, 0.164, 0.273, 0.273, 0.164, 0.0547,
           0.00781}
In[26]:= relfreq7
Out[26]= {{0.004, 0.054, 0.15, 0.3, 0.289, 0.154, 0.046, 0.003},
          {0, 1., 2., 3., 4., 5., 6., 7.}}
In[27]:= plotbdist7=ListPlot[tablebdist7, PlotStyle ->
         {GrayLevel[0],PointSize[0.03]}]
Out[27]= -Graphics-
In[28]:= Show[hist7,plotbdist7]
Out[28]= Graphics-
```



Means, variances, quantiles, and other parameters can be readly evaluated. The command `Random[dist]` produces a pseudorandom number with probability distribution `dist]`. A similar command `RSDiscreteDistribution[freq, n]` in `UVW‘DiscSamp‘` produces a pseudorandom sample of size `n` with prescribed frequencies `freq`.

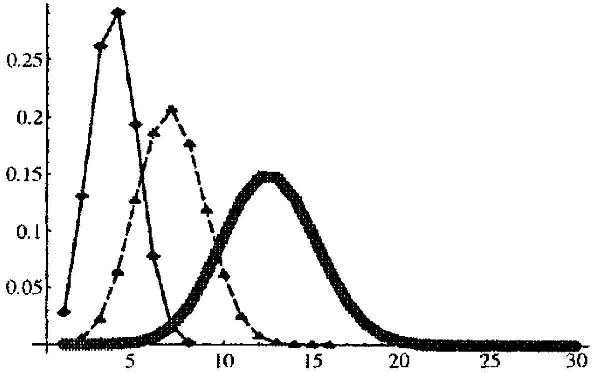
```
In[29]:= Mean[bdist7]
Out[29]= 3.5
In[30]:= Variance[bdist7]
Out[30]= 1.75
In[31]:= Quantile[bdist7,0.75]
Out[31]= 4
In[32]:= Table[Random[bdist7],{50}]
Out[32]= {1, 3, 4, 4, 4, 4, 3, 5, 4, 2, 2, 6, 5, 3, 3, 5, 6, 4, 7,
          6, 1, 1, 5, 4, 3, 3, 5, 3, 3, 5, 4, 5, 4, 4, 2, 2, 2, 2,
          5, 2, 5, 4, 5, 3, 1, 4, 6, 1, 5, 3}
```

3.5 Rescaling probabilities: Poisson approximation

If parameter n in the binomial distribution $b(k; n, p)$ increases to infinity, then both the mean $\mu = np \to \infty$, and the variance $\sigma^2 = np(1 - p) \to \infty$ (see, (3.4.10-11)). The distribution itself sort-of escapes to infinity while getting flatter and more and more spread out.

Mathematica Experiment 1. From Binomial to Poisson Distribution. We will plot the values of binomial probabilities $b(k; n, p), k = 0, 1, \ldots, n$, for three values $n = 7, 15, 29$, and the Bernoulli probability $p = 0.4$. The latter choice makes the graphs asymmetric.

```
In[1]:= <<Graphics`MultipleListPlot`
In[2]:= <<Statistics`DiscreteDistributions`
In[3]:= p=0.4
Out[3]= 0.4
In[4]:= bdist7=BinomialDistribution[7,p]
Out[4]= BinomialDistribution[7, 0.4]
In[5]:= tablebdist7= Table[PDF[bdist7, x], {x,0,7}]
Out[5]= {0.0279936, 0.130637, 0.261274, 0.290304, 0.193536,
          0.0774144, 0.0172032, 0.0016384}
In[6]:= bdist15 =BinomialDistribution[15,p]
Out[6]= BinomialDistribution[15, 0.4]
In[7]:= tablebdist15= Table[PDF[bdist15, x], {x,0,15}]
Out[7]= {0.000470185, 0.00470185, 0.021942, 0.0633879, 0.126776,
           0.185938, 0.206598, 0.177084, 0.118056, 0.0612141,
           0.0244856, 0.00741989, 0.00164886, 0.000253672,
           0.0000241592, 1.07374  .10^ (-6) }
In[8]:=  bdist29 = .........
In[10]:= MultipleListPlot[tablebdist7, tablebdist15, tablebdist29,
                         PlotJoined ->True]
Out[10]= -Graphics-
```



To remedy this "escape to infinity" problem we need to keep the mean bounded and the easiest way to accomplish this is by *rescaling* the Bernoulli probability

$$p = \frac{1}{n} \tag{1}$$

so that, by (3.4.10), the mean μ_n of the rescaled binomial distributions $b(x; n, 1/n)$, $n = 1, 2, \ldots$, is

$$\mu_n = n \cdot \frac{1}{n} = 1 \tag{2}$$

for any $n = 1, 2, \ldots$. Also, it so happens that, by (3.4.11), the variance of so rescaled binomial distributions

$$\sigma_n^2 = n\frac{1}{n}\left(1 - \frac{1}{n}\right) \to 1 \tag{3}$$

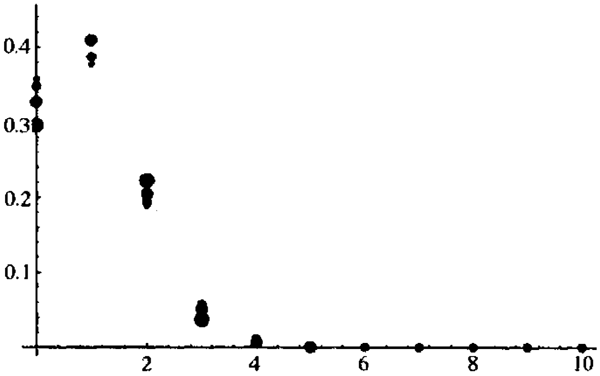
as $n \to \infty$. So, as n grows, the rescaled binomial distributions $b(x; n, 1/n)$, $n = 1, 2, \ldots$, stabilize their means and variances at 1.

In terms of our original n-coin toss experiment, the above rescaling operation accomplished the following. The probability of 1$ win in each toss was reduced to $1/n$ so that in a series of n tosses the mean win remains constant at 1$.

So, we managed to stabilize the means and variances of distributions $b(x; n, 1/n)$, $n = 1, 2, \ldots$, but what about the probabilities themselves?

Mathematica Experiment 1 Continued. From Binomial to Poisson Distribution. We will plot the probabilities of the rescaled binomial distribution $b(x; n, 1/n)$ for increasing values $n = 3, 5, 10, 20$.

```
In[1]:= <<Graphics`Graphics`
In[2]:= <<Statistics`DiscreteDistributions`
In[3]:= tablebdist[n_]:= N[Table[ {x,
          PDF[BinomialDistribution[n,1/n], x]}, {x,0,Min[n,10]}] ]
In[4]:= DisplayTogether[
          ListPlot[tablebdist[3], PlotStyle -> PointSize[0.025]],
          ListPlot[tablebdist[5], PlotStyle -> PointSize[0.02]],
          ListPlot[tablebdist[10], PlotStyle -> PointSize[0.015]],
          ListPlot[tablebdist[20], PlotStyle -> PointSize[0.01]]
          ]
Out[4]= -Graphics-
```



Why the domain of x was restricted to $\min(n, 10)$ is clear from the pictures; for x larger than 10 the values are practically zero. But the trend is clear: as n increases the probabilities $b(x; n, 1/n)$ themselves seem to stabilize.

So, it is not surprising to see the analytic proof of the convergence discovered experimentally above. For each fixed k, $0 \le k \le n$,

$$\begin{aligned} b(k; n, 1/n) &= \frac{n!}{k!(n-k)!}\left(\frac{1}{n}\right)^k\left(1-\frac{1}{n}\right)^{n-k} \\ &= \frac{n(n-1)\dots(n-k+1)}{k!n^k}\left(1-\frac{1}{n}\right)^{n-k} \\ &= \frac{1(1-1/n)\dots(1-(k-1)/n))}{k!}\left(1-\frac{1}{n}\right)^{n-k}. \end{aligned}$$

As $n \to \infty$, the numerator in the last expression clearly goes to 1. On the other hand, in view of the standard calculus formula $\lim_{x\to\infty}(1+1/x)^x = e$,

$$\left(1-\frac{1}{n}\right)^{n-k} = \left[\left(1-\frac{1}{n}\right)^n\right]\left(1-\frac{1}{n}\right)^{-k} \to e^{-1}\cdot 1. \tag{4}$$

Therefore, as $n \to \infty$, for any $k = 0, 1, 2, \dots,$

$$b(k; n, 1/n) \to e^{-1}\frac{1}{k!}. \tag{5}$$

The limit distribution is called the Poisson distribution (with parameter 1). The possible values of a Poisson random quantity X (with parameter 1) are all nonnegative integers $k = 0, 1, 2, 3, \ldots$, taken with probabilities

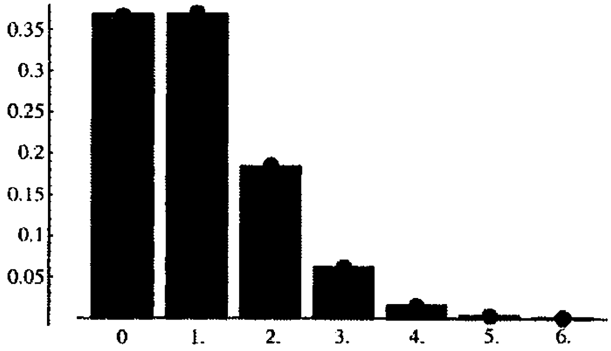
$$\Pr\{X = k\} = p(k;\, 1) = e^{-1}\frac{1}{k!}, \qquad k = 0, 1, 2, \ldots. \tag{6}$$

Clearly, formula (3) defines a probability distribution as

$$\sum_{k=0}^{\infty} p(k;\, 1) = e^{-1}\sum_{k=0}^{\infty}\frac{1}{k!} = 1. \tag{7}$$

Mathematica Experiment 1 Continued. From Binomial to Poisson Distribution. To illustrate the above approximation of the binomial distribution $b(x;\, n,\, 1/n)$ for large n, by the Poisson distribution $p(x;\, 1)$ let us compare their numerical values.

```
In[1]:= <<Graphics`Graphics`
In[2]:= <<Statistics`DiscreteDistributions`
In[3]:= tablebdist100= N[Table[
          {PDF[BinomialDistribution[100,1/100],x], x},{x,0,6}], 3 ]
Out[3]= {{0.366, 0},{0.37, 1.}, {0.185, 2.}, {0.061, 3.},
         {0.0149, 4.}, {0.0029, 5.}, {0.000463, 6.}}
In[4]:= tablepdist = N[Table[
         {PDF[PoissonDistribution[1],x], x }, {x,0,6}], 3 ]
Out[4]= {{0.368, 0}, {0.368, 1.}, {0.184, 2.}, {0.0613, 3.},
         {0.0153, 4.}, {0.00307, 5.}, {0.000511, 6.}}
In[5]:= t= N[Table[{x+1,
         PDF[BinomialDistribution[100,1/100],x   ]}, {x, 0,6}],3 ]
Out[5]= {{1., 0.366}, {2., 0.37}, {3., 0.185}, {4., 0.061},
         {5., 0.0149}, {6., 0.0029}, {7., 0.000463}}
In[6]:= DisplayTogether[ ListPlot[t,
         PlotStyle->PointSize[0.03]], BarChart[tablepdist] ]
Out[6]= -Graphics-
```



The above analytical arguments and experiments can be repeated (see Experiments, Exercises, and Projects) with the rescaling condition (1) replaced by a more general condition

$$p = \frac{\mu}{n}, \tag{8}$$

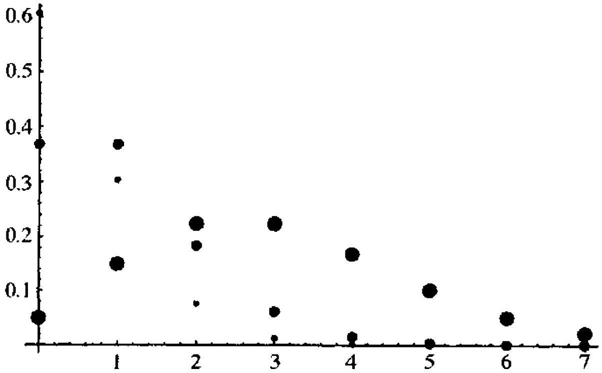
where $\mu > 0$ is an arbitrary constant. As a result, as $n \to \infty$, for any $k = 0, 1, 2, \ldots,$

$$b(k; n, \mu/n) \to e^{-\mu}\frac{\mu^k}{k!} \equiv p(k; \mu). \tag{9}$$

A random quantity X with this limit probability distribution is called the *Poisson random quantity with parameter* μ, and $\Pr\{X = k\} = p(k; \mu)$, for any $k = 0, 1, 2, \ldots.$

Mathematica Experiment 2. Poisson Distributions. The graphs of Poisson probability distributions for parameter values $\mu = 0.5, 1, 3$ are shown below. The larger dots correspond to the larger values of μ.

```
In[1]:= <<Graphics‘Graphics'
In[2]:= <<Statistics‘DiscreteDistributions'
In[3]:= DisplayTogether[
     ListPlot[N[Table[ {x,PDF[PoissonDistribution[1],x]  },
     {x,0,7}],3 ], PlotStyle->PointSize[0.0175] ],
     ListPlot[ N[Table[ {x,PDF[PoissonDistribution[0.5],x]  },
     {x,0,7}],3 ], PlotStyle->PointSize[0.01] ],
     ListPlot[ N[Table[ {x,PDF[PoissonDistribution[ 3],x]  },
     {x,0,7}],3 ], PlotStyle->PointSize[0.025] ]
     ]
Out[3]= -Graphics-
```



Since the probability distribution $p(k; \mu)$ of a Poisson random quantity X is the limit of binomial distributions $b(x; n, \mu/n)$ with means equal to $p \cdot (\mu/n) = \mu$ and variances $n(\mu/n)(1 - (\mu/n))$ converging to μ one would suspect that it itself has the mean and variance equal to μ. This is indeed the case, and can be verified by a direct calculation:

$$\mu(X) = \sum_{k=0}^{\infty} kp(k;\mu) = e^{-\mu}\sum_{k=1}^{\infty} k\frac{\mu^k}{k!}$$

$$= e^{-\mu}\mu\sum_{k=1}^{\infty}\frac{\mu^{k-1}}{(k-1)!} = \mu,$$

$$\sigma^2(X) = \sum_{k=0}^{\infty} k^2 p(k;\mu) - \mu^2$$

$$= e^{-\mu}\sum_{k=2}^{\infty} k(k-1)\frac{\mu^k}{k!} + \mu - \mu^2$$

$$= \mu^2 + \mu - \mu^2 = \mu.$$

So, the parameter μ represent both the mean and the variance of the Poisson distribution.

The Poisson distribution is often said to model *rare events* because it approximates the binomial distribution with vanishingly small Bernoulli probability of success p.

Other common discrete distributions. Introduced in the preceding sections, Bernoulli, binomial, and Poisson distributions are but a few examples of the large supply of analytically expressible probability distributions $f(k)$ of discrete random quantities which can take integer values, and often appear in applications. The only requirements for $f(k)$s is that

$$f(k) \geq 0 \qquad \text{for all } k, \tag{10}$$

and that

$$\sum_{\text{all } k} f(k) = 1. \tag{11}$$

In the remainder of this section we will provide additional examples of discrete probability distributions, indicating the type of physical situations in which they arise.

Example 3.5.1 Geometric Distribution.
In a series of independently repeated Bernoulli trials, the random quantity in which we are interested is the number k of trials until the first success, that is until the first 1 appears. This random quantity takes values $k = 1, 2, \ldots$, with probabilities

$$f(k; p) = (1 - p)^{k-1} p, \tag{12}$$

where p is the probability of success in a single Bernoulli trial. Formula (12) reflects that fact that if it took k trials to achieve the first success, then this first success had to be preceded by $k - 1$ failures, each occurring with probability $1 - p$. Clearly, formula (12) defines a probability distribution as

$$\sum_{k=1}^{\infty} f(k; p) = \sum_{k=1}^{\infty} p(1 - p)^{k-1} = \frac{p}{1 - (1 - p)} = 1.$$

Example 3.5.2 Negative Binomial Distribution.
In a series of independently repeated Bernoulli trials, the random quantity we are interested in is the number of trials until a total of r successes are accumulated. The possible values of this random quantity are $k = r, r + 1, r + 2, \ldots$, and they are taken with corresponding

$$f(k; r, p) = \binom{k - 1}{r - 1} p^r (1 - p)^{k-r}.$$

Example 3.5.3 Hypergeometric Distribution.
Suppose that a sample of size n is to be randomly chosen (without replacements) from a collection of N items, of which K are classified as defective and $N - K$ as good. We are interested in the random quantity representing the number k of good items in the sample. This random quantity can take values $0, 1, \ldots, \min(n, K)$ with probability distribution

$$f(k; n, K, N) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

Example 3.5.4 Discrete Pareto Distribution.
This distribution often occurs in economic applications. The possible values of the corresponding random quantity are integers $k = 1, 2, \ldots$, and they are taken with

probabilities

$$f(k;\alpha)=\frac{k^{-\alpha}}{\sum_{n=1}^{\infty}n^{-\alpha}},$$

where α is an arbitrary parameter greater than 1.

3.6 Stability of Fluctuations Law: Gaussian approximation

In Chapter 2 we considered the effects of changing scale and location of data on their compressed characteristics such as the mean and variance. We will pursue these ideas here, initially, in application to the binomial data, to obtain another universal approximation for the binomial distributions $b(x;n,p)$ for large n. But the same approach will yield a much more general Stability of Fluctuations Law.

Consider, as in the preceding two sections, random quantities $X_1,\ldots,X_n$, describing a series of n independent Bernoulli trials with $\Pr\{X_i=1\}=p=1-\Pr\{X_i=0\}$, and the corresponding binomial random quantity $S_n=X_1+\ldots+X_n$.

In this section we will take a look at the binomial random quantity which is both shifted by β and rescaled at the rate α, i.e., at the new random quantity

$$T_n=\alpha(S_n+\beta)=\sum_{i=1}^{n}\alpha\left(X_i+\frac{\beta}{n}\right). \tag{1}$$

If you liked our gambling interpretation of S_n, i.e., the total winnings in the game of n coin tosses which pays \$1 each time 1 (heads) come up and \$0 if 0 comes up, you will immediately see that the formula (1) just changes the payout scheme in the same game. Indeed, the second part of (1) indicates that in an n-toss game played according to the new scheme, the payout is $\alpha(1+\beta/n)$ if 1 comes up and $\alpha\beta/n$ if 0 comes up. For example, with $n=20,\beta=-10,\alpha=4$, the payout is \$2 if 1 comes up and \$$-2$ if 0 comes up.

The probability distribution of the random quantity T_n is easy to determine. Since the possible values of S_n were $0,1,2,\ldots,n$, the possible values of T_n are

$$\alpha(k+\beta),\qquad k=0,1,2,\ldots,n, \tag{2}$$

and the corresponding probabilities are binomial, i.e.,

$$\Pr\{T_n=\alpha(k+\beta)\}=b(k;n,p),\qquad k=0,1,2,\ldots,n. \tag{3}$$

Knowing the distribution of T_n permits, in turn, an immediate determination of its mean and variance:

$$\mu(T_n) = \sum_{k=0}^{n} \alpha(k+\beta)b(k; n, p) = \alpha(np+\beta), \tag{4}$$

and

$$\sigma^2(T_n) = \alpha^2 np(1-p). \tag{5}$$

The above calculation was facilitated by remembering that the mean and variance of binomial S_n is np and $np(1-p)$, respectively.

Now, if you recall our past struggles to stabilize the mean and variance of the binomial distribution, an obvious opportunity opens up. If we select

$$\beta = -np, \qquad \alpha = \frac{1}{\sqrt{np(1-p)}}, \tag{6}$$

that is, if we consider

$$Y_n = \frac{S_n - np}{\sqrt{np(1-p)}}, \tag{7}$$

then its mean is going to be 0 and its variance 1:

$$\mu(Y_n) = 0, \qquad \sigma^2(Y_n) = 1. \tag{8}$$

As before, we are curious if this rock-solid stability of means and variances of Y_ns does anything to stabilize the behavior of the distributions of Y_ns themselves. The interpretation of the random quantity Y_n is obviously as that of *fluctuations* of

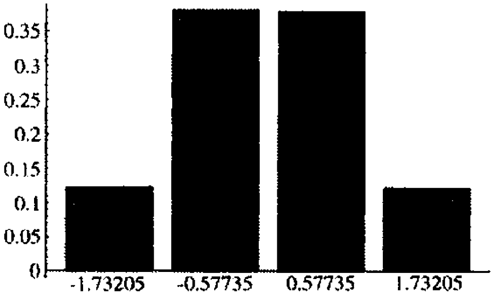
$$S_n - \mu(S_n) = S_n - np \tag{9}$$

of the binomial random quantities S_n about their means, resized by their natural scale, i.e., their standard deviation $\sigma(S_n) = \sqrt{np(1-p)}$.

Mathematica Experiment 1. Stability of Fluctuations Law. We will use the tools developed in Mathematica Experiment 3.4.1 related to repeated Bernoulli experiments to observe histograms of Y_n as n increases. We will take $p = 1/2$ and produce rn=1000 repetitions of each series of $n = 3, 5, 10$, and 20 tosses.

```
In[1]:= <<Graphics`Graphics`
In[2]:= <<Statistics`DataManipulation`
In[3]:= <<Statistics`DescriptiveStatistics`
In[4]:= n=3
Out[4]= 3
```

```
In[5]:= rn=1000
Out[5]= 1000
In[6]:= Do[ r3=Table[{0},{rn}];
         For[i=1, i<=rn, i++, r3[[i]]=Table [Random[Integer],
            {n}]]]
In[7]:= r3
Out[7]= {{1, 0, 0},  {1, 1, 0},  ... ,  { 0, 0, 0},
         {1, 1, 1}}
In[8]:= Do[ r3sum=Table[{0},{rn}];
         For[i=1,i<=rn, i++,  r3sum[[i]]=Apply[Plus,
            r3[[i]]]] ]
In[9]:= r3sum
Out[9]= {1, 2, ... , 2, 0, 3}
In[10]:= freq3=Frequencies[r3sum]
Out[10]= {{122, 0}, {380, 1}, {377, 2}, {121, 3}}
In[11]:= relfreq3=N[{Column [freq3,1]/rn, Column [freq3,2] }]
Out[11]= {{0.122, 0.38, 0.377, 0.121}, {0.,1., 2., 3.}}
In[12]:= relfreq3SFL={Table[relfreq3[[1]][[i]],{i,n+1}],
         Table[((relfreq3[[2]][[i]]-n*0.5)/N[Sqrt[n*0.25]]),
              {i,n+1}]}
Out[12]= {{0.122, 0.38, 0.377, 0.121},
          {-1.73205, -0.57735, 0.57735, 1.73205}}
In[13]:= hist3SFL=BarChart[Transpose[relfreq3SFL]]
Out[13]= -Graphics-
```



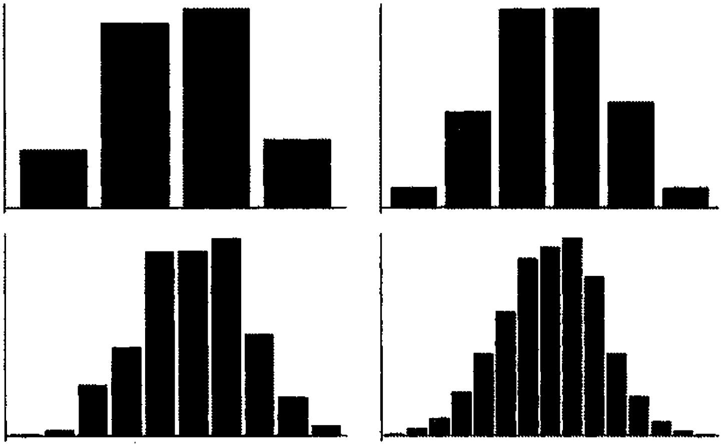
Repeating the above experiment for $n = 5, 10, 20$, and putting the resulting histograms together gives the pictures shown below. You will notice that, e.g., for $n = 20$, not all 21 possible values of T_{20} appeared in our repeated sampling although the number of repetitions was large; their probability decays very fast when we move away from the mean value 0. For example, it follows from formula (3) that the probability of extreme values of T_{20}:

$$\Pr\{T_{20} = \pm 4.47\} = b(0; 20, 0.5) = \Pr\{X_1 = 0, \ldots, X_{20} = 0\}$$

$$= (1/2)^{20} = 9.53674 \cdot 10^{-7}$$

is very small, and even in a thousand repetitions these values are unlikely to appear. In this context, one has to pay attention to the term `n+1` in line `In[12] :=` of the above code, and adjust it as necessary.

```
In[34] := Show[ GraphicsArray[{{hist3SFL, hist5SFL},
                {hist10SFL, hist20SFL}}]]
Out[34]= - GraphicsArray-
```



The above experiments suggest the existence, as $n \to \infty$, of the limiting probability distribution of the random quantities T_n, which is bell shaped, continuous, and almost totally concentrated on the interval $(-3.5, 3.5)$. Armed with this intuition we will find its shape analytically.

To avoid possible high fluctuations in the histograms, we will analyze the limit behavior of the cumulative d.f. $F_n(x)$ of the random quantity Y_n. The summation [remember formula (2.5.6)] contained in the definition of the cumulative d.f. tends to smooth it out and make it easier to deal with analytically than the histograms themselves. So, in view of (3) and (7)

$$\begin{aligned} F_n(z) \equiv \Pr\{Y_n \le z\} &= \Pr\left\{\frac{S_n - n/2}{\sqrt{n}/2} \le z\right\} \\ &= \Pr\left\{S_n \le (z\sqrt{n}+n)/2\right\} \\ &= \sum_{k=0}^{(z\sqrt{n}+n)/2} \binom{n}{k}\left(\frac{1}{2}\right)^n \end{aligned}$$

By Stirling's formula

$$n! \sim \sqrt{2\pi n}n^n e^{-n}, \tag{10}$$

where $\sim$ means that the ratio of the two quantities approaches 1 as $n \to \infty$. Hence, the above cumulative d.f.

$$F_n(z) \sim \sum_{k=0}^{(z\sqrt{n}+n)/2} \frac{1}{\sqrt{2\pi}} \frac{2}{\sqrt{n}} \exp\left[-\frac{1}{2}\left(\frac{2k-n}{\sqrt{n}}\right)^2\right]$$

which we recognize as the Riemann sum approximation the integral

$$\int_{-\sqrt{n}}^{z} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)\, dx.$$

Thus, as $n \to \infty$,

$$F_n(z) \longrightarrow \Phi(z) \equiv \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)\, dx. \tag{11}$$

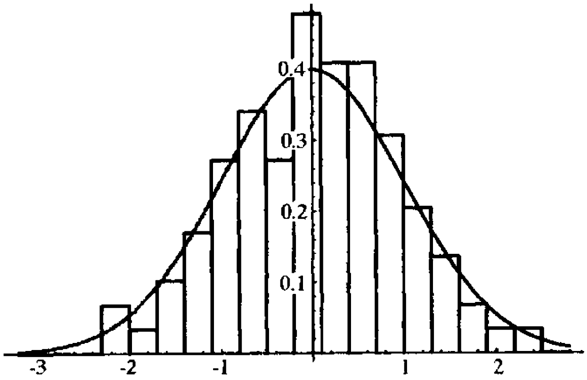
Formula (11) immediately gives the limit density as

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2). \tag{12}$$

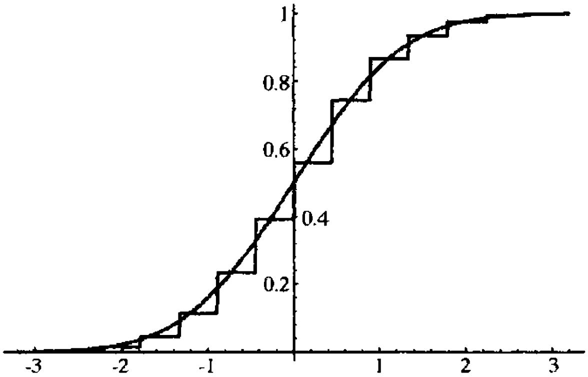
The continuous cumulative d.f. $\Phi(z)$ is called the *standard Gaussian (or, normal) cumulative d.f.*, and $\phi(z)$ is called the *standard Gaussian (normal) probability density function.*

Mathematica Experiment 1 Continued. Stability of Fluctuations Law. Let us compare the above theoretical Gaussian limit distribution (both the density and the cumulative d.f.) with the corresponding objects for the rescaled and shifted experimental binomial random quantities Y_{20}. Note that the bar charts used in the first part are not suitable for the comparison since (for esthetic reasons) they have gaps between vertical bars, so that their areas are distorted. For that reason we will use a `RegularHisto` command provided in the `UVW'DataRep'` package.

```
In[34]:= <<UVW'DataRep'
In[35]:= r20sumSFL=Table[((r20sum[[i]]-n*0.5)/N[Sqrt[n*0.25]]),
                                          {i,1000}];
In[36]:= RegHisto=RegularHisto[r20sumSFL,-3.2,3.2,15]
Out[36]= -Graphics-
In[37]:= GaussPDF[x_]:=(1/Sqrt[2Pi])  Exp[-x^2/2]
In[38]:= Show[RegHisto, GaussPDF[x],{x,-3.2,3.2}]
Out[38]= -Graphics-
```



```
In[39]:= Fr20=Transpose[relfreq20]
Out[39]= {{0.001, -3.1305}, {0.003, -2.68328}, {0.011, -2.23607},
          {0.031, -1.78885}, {0.068, -1.34164}, {0.121, -0.894427},
          {0.158, -0.447214}, {0.167, 0.}, {0.184, 0.447214},
          {0.122, 0.894427}, {0.067, 1.34164}, {0.042, 1.78885},
          {0.019, 2.23607}, {0.005, 2.68328}, {0.001, 3.1305}}
In[40]:= H[x_]:=If[x<0,0,1]
In[41]:= CumDiFun[x_]:= Sum[F20[[i]][[1]]*
           H[x-F20[[i]][[2]] ],{i,1,Length[F20]} ]
In[42]:= GaussCDF[y_]:=(1/Sqrt[2Pi])
          NIntegrate[Exp[-x^2/2],{x,-Infinity,y}]
In[43]:= Plot[CumDiFun[x],  GaussCDF[x], {x, -3.2, 3.2}]
Out[44]= -Graphics-
```



A random quantity Z with standard Gaussian probability d.f. $\Phi(z)$ is called the *standard Gaussian (or normal) random quantity.* Since its distribution is a limit of distributions of the random quantities Y_n which have zero means and variances equal to one, one would suspect that

$$\mu(Z) = 0, \qquad \sigma^2(Z) = 1. \tag{13}$$

This can be verified directly by integration by parts and use of the normalization condition (14) below; see also the approach through the gamma function discussed later in this chapter.

Although the indefinite integral (11) cannot be expressed in terms of elementary functions, we still can check the normalization condition

$$\begin{aligned}\left(\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz\right)^2 &= \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-s^2/2}ds\\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}e^{-(t^2+s^2)/2}dt\,ds \qquad (14)\\ &= \frac{1}{2\pi}\int_{0}^{\pi}\int_{0}^{\infty}e^{-r^2/2}r\,dt\,d\theta = 1,\end{aligned}$$

by changing to the polar coordinate system.

By changing the scale of a standard Gaussian random quantity Z by σ and shifting its location by μ, we obtain the whole family of random quantities

$$Z_{\mu,\sigma^2} = \sigma Z + \mu \qquad (15)$$

with absolutely continuous cumulative d.f.

$$\begin{aligned}\Phi(z;\mu,\sigma^2) &= \Pr\{\sigma Z+\mu \le z\} = \Pr\{Z \le (z-\mu)/\sigma\}\\ &= \int_{-\infty}^{(z-\mu)/\sigma}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx\\ &= \int_{-\infty}^{z}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[\frac{(x-\mu)^2}{2\sigma^2}\right]dx, \qquad (16)\end{aligned}$$

and the densities

$$\phi(z;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[\frac{(z-\mu)^2}{2\sigma^2}\right]. \qquad (17)$$

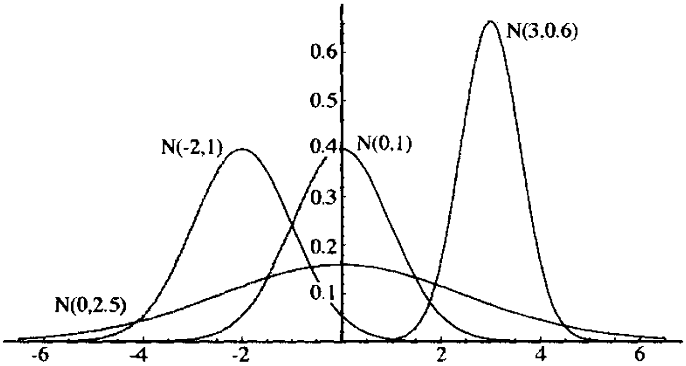
Clearly, in view of (13) and (15),

$$\mu(Z_{\mu,\sigma^2}) = \mu, \qquad \sigma^2(Z_{\mu,\sigma^2}) = \sigma^2, \qquad (18)$$

and the random quantity Z_{μ,σ^2}, or the corresponding cumulative d.f. $\Phi(z;\,\mu,\sigma^2)$ and the probability d.f. $\phi(z;\,\mu,\sigma^2)$, are called *Gaussian (normal) with mean μ and variance σ^2*, or, compactly, $N(\mu,\sigma^2)$ random quantities, cumulative d.f.s, and densities.

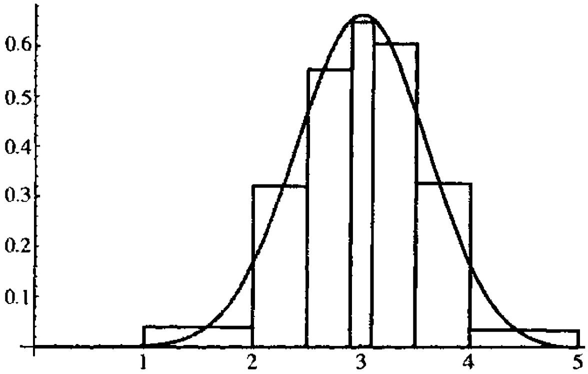
Mathematica Experiment 2. Gaussian Densities. The *Mathematica* package `Statistics'ContinuousDistributions'` contains the commands: `PDF[NormalDistribution [mu,sigma], x]` which provides probability d.f. of $N(\mu,\sigma^2)$ at x, `CDF[NormalDistribution [mu,sigma], x]` which provides cumulative d.f. of $N(\mu,\sigma^2)$ at x, and `Random[NormalDistribution [mu,sigma]]` which produces a pseudorandom number with $N(\mu,\sigma^2)$ distribution. In the process we will also see how one can place text within *Mathematica* graphics at any location prescribed by the coordinates of text's center.

```
In[1]:= <<Statistics'ContinuousDistributions'
In[2]:= <<UVW'DataRep'
In[3]:= phi[x_, mu_, sigma_] :=
          PDF[NormalDistribution[mu, sigma], x]
In[4]:= p1 = Plot[{phi[x, 0, 1], phi[x, -2, 1],
         phi[x, 0, 2.5], phi[x, 3, 0.6]},
         {x, -6.5 ,6.5}, PlotRange -> {0, 0.7},
         AspectRatio -> 0.5];
         p2 = Graphics[Text["N(0,1)", {1.2, 0.42}]];
         p3 = Graphics[Text["N(-2,1)", {-3, 0.42}]];
         p4 = Graphics[Text["N(0, 2.5)", {-5.2, 0.08}]];
         p5 = Graphics[Text["N(3, 0.6)", {5, 0.5}]];
         Show[p1,p2,p3,p4,p5]
Out[4]= -Graphics-
```



To simulate, say, an $N(3,\,(0.6)^2)$ random quantity and to compare the histogram of the simulated data with the density $\phi(x,3,(0.6)^2)$ we will use the command `Histogram[data, listofbounds]` of the `UVW'DataRep'` package, which permits selection of bin locations to be matched to where the data are concentrated. Ideally, all the bins should contain the same amount of data points.

```
In[5]:= nd=NormalDistribution[3,0.6]
Out[5]= NormalDistribution[3, 0.6]
In[6]:= tr=N[Table[Random[nd],{1000}],3]
Out[6]=  {2.19, 3.72, ... , 2.29, 2.13, 3.28}
In[7]:= ph = Plot[PDF[NormalDistribution[3, 0.6], x], {x,0,5}];
          lb = {1, 2, 2.5, 2.9, 3.1, 3.5, 4, 5};
          hist = Histogram[tr, lb]; Show[ph, hist]
Out[7]= -Graphics-
```



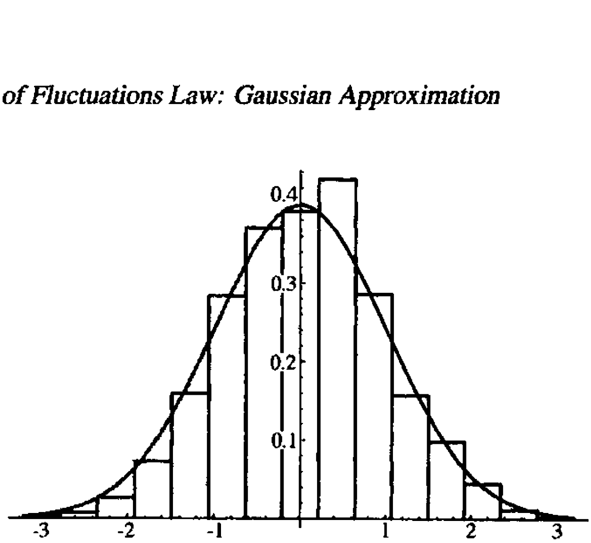
It turns out that the Gaussian limit behavior (or the Gaussian approximation if you will) is not restricted to rescaled and shifted sums of Bernoulli random quantities, or even to discrete random variables.

Mathematica Experiment 1 Continued. Stability of Fluctuations Law. Consider independent random quantities $X_1, \ldots, X_n$, uniformly distributed on the interval [0,1] with the density given by the formula (3.3.1). Its mean and variance, according to the formulas (3.3-4) are, respectively, 1/2 and 1/12. Therefore, the random quantities

$$Y_n = \frac{X_1 + \ldots + X_n - n/2}{\sqrt{n/12}}$$

have means 0 and variance 1. To make their simulation and comparison with the $N(0, 1)$ probability d.f. easier we will use the specially written command `CentralLimit [ listofdata, mu, sigma, n]` of the `UVW'DataRep'` package which takes consecutive groups of `n` data in `listofdata`. Then the sum of each group is centered by `n *mu` and then divided by `Sqrt[n]*sigma`. The results are represented on a regular histogram. In the mathematical literature, the name Central Limit Theorem is used for the Stability of Fluctuations Law (see Chapter 5).

```
In[1]:= <<UVW'DataRep'
In[2]:= unidata= N[Table[Random[],{2000}],2]
Out[2]= {0.15, 0.76, 0.68, ... ,0.18, 0.18, 0.58}
In[3]:= CentralLimit[unidata, 0.5, N[Sqrt[(1/12)]], 20]
Out[3]= -Graphics-
```



The above arguments can be rephrased in the form of a general

Stability of Fluctuations Law. *If the random quantities $\boldsymbol{X}_n = (X_1, X_2, \ldots, X_n)$ represent outcomes of n repeated independent random experiments, each with mean μ and variance σ^2, then for large n, the cumulative p.d. of sample averages:*

$$\mathrm{Av}\,(\boldsymbol{X}_n) = \frac{X_1 + X_2 + \ldots + X_n}{n},$$

suitably centered and rescaled, is well approximated by the $N(0, 1)$ cumulative d.f. $\Phi(z)$. More precisely, for each real number z the cumulative d.f.

$$\Pr\left\{\frac{\sqrt{n}}{\sigma}\Big(\mathrm{Av}\,(\boldsymbol{X}_n) - \mu\Big) \le z\right\} \longrightarrow \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx\,,$$

as $n \to \infty$.

Obviously, the Stability of Fluctuations Law gives an approximate probability distributions of fluctuations of sample averages around their theoretical means. It will be given a more formal treatment, with precise assumptions, in Chapter 5, where it will become the so-called Central Limit Theorem.

Note that the Gaussian densities $\phi(z; \mu, \sigma^2)$ are symmetric in z about the mean μ, positive everywhere, and that the probability that the Gaussian random quantity $Z = Z_{\mu,\sigma^2}$ takes values far away from the mean μ is very small. Indeed, measured in terms of the natural scale parameter σ, the probability of deviation from μ by more than $a\sigma$ goes to zero very fast as $a \to \infty$:

$$\Pr\{|Z - \mu| > a\sigma\} = 2(1 - \Phi(a; 0, 1)) = 2\int_a^\infty \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,dz$$

$$\le \frac{2}{\sqrt{2\pi}} \int_a^\infty \frac{z}{a} e^{-z^2/2}\,dz = \frac{2}{\sqrt{2\pi}\,a} e^{-a^2/2}$$

$$= \frac{2}{a}\phi(a; 0, 1) \tag{19}$$

This rate is much faster than the rate a^{-2} predicted by the crude but general *Chebyshev's inequality*:

$$\Pr\{|X-\mu| > a\sigma(X)\} \le \int_{\frac{|x-\mu|^2}{a^2\sigma^2(X)}>1} \frac{|x-\mu|^2}{a^2\sigma^2(X)}\,dF_X(x)$$

$$\le \int_{-\infty}^{\infty} \frac{|x-\mu|^2}{a^2\sigma^2(X)}\,dF_X(x) = \frac{1}{a^2}. \tag{20}$$

Mathematica Experiment 2 Continued. Gaussian Densities. We will compare the exact values of $\Pr\{|Z-\mu| > a\sigma\}$ with the estimates given by (19) and (20). Initially, $a = 3$, that is, we seek the probability that the Gaussian random quantity deviates from its mean μ by more than 3σ.

```
In[1]:= <<Statistics'ContinuousDistributions'
In[2]:= nd=NormalDistribution[0,1]
Out[2]=  NormalDistribution[0,1]
In[3]:= 2(1-CDF[nd, 3])
Out[3]= 0.0026998
In[4]:= (2/3)PDF[nd, 3]
Out[4]= 0.00295457
In[5]:= N[1/(3^2)]
Out[5]= 0.111111
```

So, the true value is fairly close to the estimate (19), while Chebyshev's estimate (20) is not very accurate. For $a = 4$, the analogous numbers are 0.0000633, 0.0000892, 0.0625.

The values of the $N(0, 1)$ quantile function $\Phi^{-1}(\alpha)$ that can be obtained using the `Statistics'ContinuousDistributions'` package are schematically pictured in Fig. 3.6.1.

It satisfies the symmetry condition $\Phi^{-1}(1-\alpha) = -\Phi^{-1}(\alpha)$ so that it suffices to know its values only for, say, $1/2 \le \alpha \le 1$. Also, traditionally, one often uses (see Chapters 7 through 9) the complementary normal *upper tail quantile function*

$$z_\alpha = \Phi^{-1}(1-\alpha). \tag{21}$$

Graphically, the tail quantile z_α marks the point so that the area under the graph of the standard normal density $\phi(z; 0, 1)$ to the right of z_α is exactly equal to α (see Fig. 3.6.2).

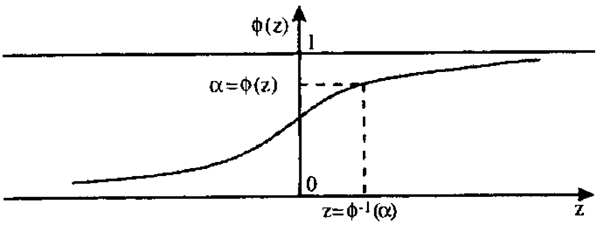


FIGURE 3.6.1

The standard normal cumulative distribution $\alpha = \Phi(z)$ and its inverse $z = \Phi^{-1}(\alpha)$ (quantile function).

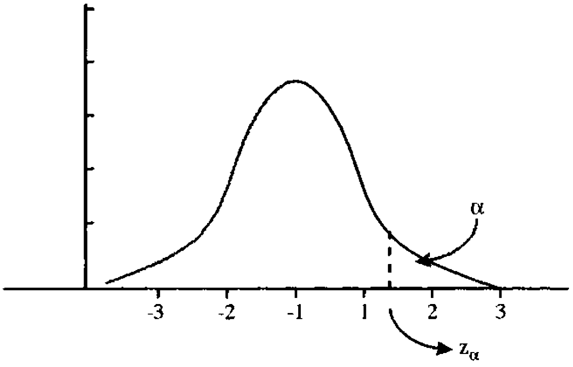


FIGURE 3.6.2

The upper tail quantile z_α marks the point so that the area under the graph of the standard normal density $\phi(z; 0, 1)$ to the right of z_α is exactly equal to α.

3.7 How to estimate p in Bernoulli experiments

In this section we provide a simple application of the Stability of Fluctuations Law to statistical inference concerning parameter p in the Bernoulli distributions. In general, it is the goal of statistics to retrieve properties of unknown distributions on the basis of experimental data; the subject will be further developed in Chapter 7 through 9.

In the case of a sequence $\boldsymbol{x} = (x_1, \ldots, x_n)$ of outcomes of independently repeated experiments $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ with two possible outcomes, e.g., success or failure, the unknown distribution is in the class of Bernoulli distributions which are parametrized by the single parameter $p \in [0, 1]$—the probability of success in a single trial. Its value is unknown and our goals are

1) To *estimate* the value of p.

2) To *test the hypothesis* whether p belongs to a certain subset H_0 of $[0, 1]$.

By the Law of Large Numbers of Section 3.1, the sample average

$$\mathrm{Av}\,(\boldsymbol{X}_n) = \frac{X_1 + \ldots + X_n}{n} = \hat{p} \tag{1}$$

is a *consistent estimator* for p, i.e., $\hat{p}$ approaches, for large n, the correct value of p. Note that the estimator $\hat{p}$ is a random quantity depending on a particular realization $\boldsymbol{x} = (x_1, \ldots, x_n)$ of outcomes of independently repeated experiments $\boldsymbol{X}_n = (X_1, \ldots, X_n)$. The notation $\hat{p}$ for an estimator of a parameter p is traditional in statistics.

Also notice that the random quantity $\hat{p}$ is an *unbiased* estimator for p, which means that the theoretical mean of the estimator $\hat{p}$ of the parameter p is equal to p itself:

$$\mu(\hat{p}) = \mu\left(\frac{X_1 + \ldots + X_n}{n}\right) = p. \tag{2}$$

We can summarize the above discussion as follows: *In independently repeated Bernoulli trials, the sample mean $\bar{x}$ is a consistent and unbiased estimator of the probability p of success.*

The random quantity $n\hat{p} = S_n$ has a binomial $b(k; n, p)$ probability d.f., see Section 3.5. Thus, we can calculate the probability that the estimator $\hat{p}$ approximates the parameter p with accuracy better than, say, a:

$$\begin{aligned} \Pr\{|\hat{p} - p| \le a\} &= \Pr\left\{n(p-a) \le n\hat{p} \le n(p+a)\right\} \\ &= \Pr\left\{n(p-a) \le S_n \le n(p+a)\right\} \\ &= \sum_{k:n(p-a)\le k\le n(p+a)} \binom{n}{k} p^k (1-p)^{n-k}. \end{aligned} \tag{3}$$

For any selected accuracy level $a > 0$, if the number n of repetitions is large, the probability (3) is close to 1, say, $1 - \alpha$, with small α. This is a side effect of the Stability of Fluctuations Law and Chebyshev's type estimates (3.6.19-20); see also the Weak Law of Large Numbers of Chapter 6.

The surprising main consequence of the Stability of Fluctuations Law is that given the accuracy level a, the probability α (or, conversely, given the probability α, the accuracy level a) can be chosen (almost) independently of the underlying

Bernoulli distribution, that is unknown value of parameter p. Indeed, for large n,

$$\begin{aligned}\Pr\{|\hat{p}-p|\leq a\} &= \Pr\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_i - p\right|\leq a\right\}\\ &= \Pr\left\{\frac{\sqrt{n}}{\sqrt{p(1-p)}}|\mathrm{Av}\,(\boldsymbol{X}_n)-p|\leq\frac{a\sqrt{n}}{\sqrt{p(1-p)}}\right\}\\ &\approx \int_{-\frac{a\sqrt{n}}{\sqrt{p(1-p)}}}^{\frac{a\sqrt{n}}{\sqrt{p(1-p)}}}\frac{1}{\sqrt{2\pi}}e^{-u^2/2}\,du. \end{aligned}\tag{4}$$

Now, fix α and choose a depending on p, so that

$$z_{\alpha/2}=\frac{a\sqrt{n}}{\sqrt{p(1-p)}},\tag{5}$$

where $z_{\alpha/2}$ is the $1-\alpha/2$-quantile of the $\mathcal{N}(0,1)$ distribution, that is

$$1-\alpha=\int_{-z_{\alpha/2}}^{z_{\alpha/2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}\,du.\tag{6}$$

Then we can rewrite (4) in the form

$$\Pr\{|\hat{p}-p|\leq a\}=\Pr\left\{\frac{\sqrt{n}}{\sqrt{p(1-p)}}|\hat{p}-p|\leq z_{\alpha/2}\right\}\approx 1-\alpha.\tag{7}$$

Since $\hat{p}$ approaches p for large n, we can replace p by $\hat{p}$ in the denominator (as long as p and $\hat{p}$ are not too close to 0 or 1), and get

$$\Pr\left\{\frac{\sqrt{n}}{\sqrt{\hat{p}(1-\hat{p})}}|\hat{p}-p|\leq z_{\alpha/2}\right\}\approx 1-\alpha,\tag{8}$$

The inequality

$$\frac{\sqrt{n}}{\sqrt{\hat{p}(1-\hat{p})}}|\hat{p}-p|\leq z_{\alpha/2}\tag{9}$$

can be solved for p to give

$$\hat{p}_L\leq p\leq\hat{p}_U,\tag{10}$$

where the lower bound

$$\hat{p}_L = \hat{p} - z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \tag{11}$$

and the upper bound

$$\hat{p}_U = \hat{p} + z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}. \tag{12}$$

The above analysis can be summarized as

Parameter Estimation Procedure via Confidence Intervals: *Let* $\boldsymbol{X} = (X_1, \ldots, X_n)$ *be a random vector representing n independently repeated Bernoulli experiments with an unknown probability of success p. Select* $0 < \alpha < 1$. *Then the sample average* $\hat{p} = \mathrm{Av}\,(\boldsymbol{X})$ *is an unbiased, consistent estimator of the parameter p. Moreover, with probability* $1-\alpha$, *the true value of parameter p lies in a random interval* $[\hat{p}_L, \hat{p}_U]$. *Such an interval is called a confidence interval with confidence level* $1-\alpha$.

In other words, with probability $1-\alpha$, *the true value of p is within the distance*

$$\hat{l} = z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \tag{13}$$

of the sample average $\hat{p}$ *(for better confidence intervals, see Sec. 8.2).*

The second goal we set for ourselves at the beginning of this section is related to the first which was achieved by the construction of confidence intervals at a given confidence level. In the simplest case, we would like to have a procedure to decide whether the unknown parameter p in the Bernoulli experiments equals a fixed given value $p_0 \in [0, 1]$ (e.g., whether a coin is fair, or whether the proportion of defective items on the assembly line is 2%). This is called the *hypothesis testing problem.*

By analogy with (3) and (4), assuming that the hypothesis $H_0 : p = p_0$ is true,

$$\Pr\{|\hat{p} - p_0| \le a\} = \Pr\Big\{n(p_0 - a) \le n\hat{p} \le n(p_0 + a)\Big\}$$

$$= \sum_{n(p_0-a)\le k\le n(p_0+a)} \binom{n}{k} p_0^k (1-p_0)^{n-k}, \tag{14}$$

and, for large n, we have

$$\Pr\{|\hat{p} - p_0| \le a\} \approx \int_{-\frac{a\sqrt{n}}{\sqrt{p_0(1-p_0)}}}^{\frac{a\sqrt{n}}{\sqrt{p_0(1-p_0)}}} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\,du, \tag{15}$$

in view of the Stability of Fluctuations Law. This means that provided the hypothesis $H_0 : p = p_0$ is valid, the unbiased estimator $\hat{p}$ differs from p_0 by less than a. On the other hand, if the hypothesis $H_0 : p = p_0$ is false and an alternative $H_1 : p = p_1 \neq p_0$ holds true, then the estimator $\hat{p}$ should, with probability close to 1, be close to p_1 and far away from p_0 or, equivalently, the event that $\hat{p}$ and p_0 are close should be a rare, small probability event. Thus, we arrive at the following.

Hypothesis Testing Procedure. *Choose confidence level α at a preassigned level between 0 and 1, and then select a so that*

$$\Pr\{|\hat{p} - p_0| \leq a\} = \alpha.$$

If $|\hat{p} - p_0| > a$, reject the hypothesis $H_0 : p = p_0$ and conclude that the true parameter value p is different from p_0. In the opposite case $|\hat{p} - p_0| \leq a$, the hypothesis H_0 is not rejected.

Note that in the case $|\hat{p} - p_0| \leq a$, we are not claiming that the hypothesis $H_0 : p = p_0$ is to be accepted; the true value of p may be close to p_0 and yet not equal to it. The construction of confidence intervals and hypothesis testing procedures will be discussed in greater depth in Chapters 7 and 8.

Example 3.7.1 Statistical Quality Control.

A batch of N items (light bulbs, capacitors, computer memory chips, etc.) is mass manufactured at a plant and needs to be tested before shipment to customers. Usually, one takes a random sample of size $n \ll N$ from the whole batch, and then either tests each item in the random sample under working conditions until it fails (destructive testing), or one measures some important parameter of each item in the random sample without destroying it (nondestructive testing).

It could be the customer's policy, say, to accept the batch of N items only if no item from the sample of size n fails before time T prescribed by the contract. In this case, the probability of acceptance of the batch is computed as follows: Suppose that the batch of N items contains B bad and G good ones, so that $B + G = N$. Consequently,

$$\Pr\{\text{acceptance}\} = \frac{G}{N}\frac{G-1}{N-1}\cdots\frac{G-(n-1)}{N-(n-1)}$$

$$= \left(1 - \frac{B}{N}\right)\left(1 - \frac{B}{N-1}\right)\cdot\ldots\cdot\left(1 - \frac{B}{N-(n-1)}\right). \tag{16}$$

If the number B of bad items is small and the sample size n is small compared to the batch size N, then the acceptance probability is high; the chance of discovering any bad items in the batch is small.

In this context, it seems that allowing some bad items in the random sample would not be a bad idea. The new procedure would call for accepting the batch if the number of bad items in the random sample does not exceed an integer c. If p is the probability that a randomly chosen item from the batch is bad, then the *operational characteristic*

$$L(c; n, p) = \sum_{k=0}^{c} \binom{n}{k} p^k (1-p)^{n-k}, \qquad (17)$$

which is just a cumulative d.f. for the binomial distribution, gives the probability of acceptance of the batch, based on the *test plan* (n, c). It is easy to check that:

$$L(c; n, 0) = 1, \qquad (17)$$

that is, if there are no bad items in the batch then the acceptance probability is 1;

$$L(c; n, 1) = 0, \qquad (18)$$

that is, if there are no good items in the batch then the acceptance probability is 0;

$$\begin{aligned} &L(c; n, p_1) \geq L(c; n, p_2), && \text{if} \quad p_1 \leq p_2; \\ &L(c; n_1, p) \geq L(c; n_2, p), && \text{if} \quad n_1 \leq n_2; \qquad (19) \\ &L(c_1; n, p) \leq L(c_2; n, p), && \text{if} \quad c_1 \leq c_2. \end{aligned}$$

By the Stability of Fluctuations Law, for a large sample size n, we have that

$$L(c; n, p) = \Pr\{n\hat{p} \leq c\} \sim \int_{-\infty}^{\frac{c-pn}{\sqrt{np(1-p)}}} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du. \qquad (20)$$

Of course, to be able to select a large random sample, the batch size has to be very large.

Ideally, it would be desirable to choose p_0 such that the acceptance probability $L(c; n, p) = 1$ for $p \leq p_0$, and $= 0$ for $p > p_0$ (why?). This is, however, impossible in view of (20). The way out of this dilemma is the following compromise quality control standard:

(1) If the probability p of a bad item in the batch is small, say, $\leq p_\alpha$, then the probability of acceptance must be $\geq \alpha$, where (p_α, α) sets the quality control standard (often, $p_\alpha = .02, \alpha = .9$).

(2) If the probability p of a bad item in the batch is large (say $\geq p_\beta$), then the probability of acceptance must be $\leq \beta$, where (p_β, β) sets another quality control standard (often, $p_\beta = .05, \beta = .1$).

Heuristically, our standards demand that if the batch is really good ($p \leq p_\alpha$), then it should be accepted with a very high probability, and if it is really bad, then its acceptance probability should be low. In this fashion, with a high probability, a really bad decision will be avoided. Of course, there will be a price to pay for using such standards: when the true probability p stays within the range (p_α, p_β), we are not going to be able to say how good our quality control procedure is, but it would not matter anyway.

To implement the above standards, for given (p_α, α) and (p_β, β), we have to find a test plan (n, c) such that

$$L(c; n, p) \geq L(c; n, p_\alpha) = \alpha, \qquad \text{for all} \qquad p \leq p_\alpha, \tag{21}$$

and

$$L(c; n, p) \leq L(c; n, p_\beta) = \beta, \qquad \text{for all} \qquad p \geq p_\beta. \tag{22}$$

In view of (20), this leads (asymptotically in $n \to \infty$) to the equations

$$\alpha = \int_{-\infty}^{\frac{c-p_\alpha n}{\sqrt{np_\alpha(1-p_\alpha)}}} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du \tag{23}$$

and

$$\beta = \int_{-\infty}^{\frac{c-p_\beta n}{\sqrt{np_\beta(1-p_\beta)}}} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du. \tag{24}$$

Denoting by $\Phi^{-1}(\alpha)$ the inverse of the $N(0, 1)$ cumulative d.f. $\Phi(x)$, the equations (23) and (24) can be rewritten in the form

$$\Phi^{-1}(\alpha) = \frac{c - p_\alpha n}{\sqrt{np_\alpha(1 - p_\alpha)}}, \tag{25}$$

$$\Phi^{-1}(\beta) = \frac{c - p_\beta n}{\sqrt{np_\beta(1 - p_\beta)}}. \tag{26}$$

The latter can be solved easily for n and c, giving

$$\sqrt{n} = \frac{\sqrt{p_\alpha(1 - p_\alpha)}\Phi^{-1}(\alpha) - \sqrt{p_\beta(1 - p_\beta)}\Phi^{-1}(\beta)}{p_\beta - p_\alpha} \tag{27}$$

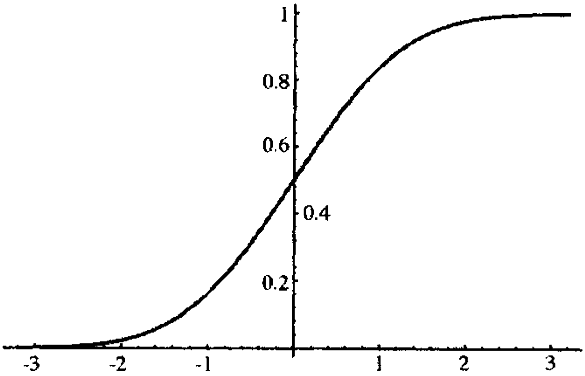
and

$$c = np_\alpha + \sqrt{np_\alpha(1 - p_\alpha)}\Phi^{-1}(\alpha). \tag{28}$$

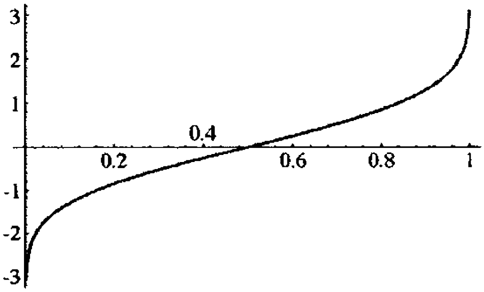
Note that, finally, n has to be rounded to the next larger integer, and c to the next lower integer!

Mathematica Experiment 1. Quality Control Plan. We will implement the above test plan for $\alpha = 0.9$, $\beta = 0.1$, $p_\alpha = 0.02$, $p_\beta = 0.05$, illustrating on the way other aspects of our discussion.

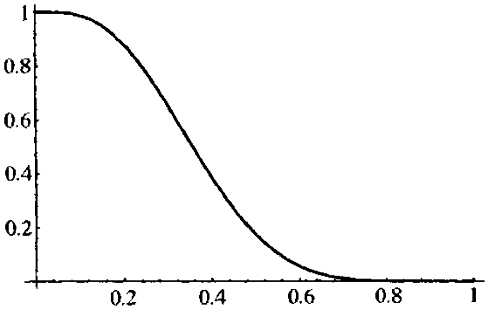
```
In[1]:= <<Statistics'ContinuousDistributions'
In[2]:= Plot[CDF[NormalDistribution[0,1],x],{x,-3,3}]
Out[2]= -Graphics-
```



```
In[3]:= PhiInverse[x_]:= Quantile[NormalDistribution[0,1],x]
In[4]:= Plot[PhiInverse[x],{x,.001,.999}]
Out[4]= -Graphics-
```



```
In[5]:= L[c_,n_,p_]:= Sum[Binomial[n,k] p^k(1-p)^{n-k},{k,0,c}]
In[6]:= Plot[L(3,10,p),{x,.001,.999}]
Out[6]=  -Graphics-
```



```
In[7]:= u=N[(Sqrt[.02 * .98] PhiInverse[.9]-Sqrt[.05 * .95] *
             PhiInverse[.1])/.03]
Out[7]= 15.2908
In[8]:= n=Ceiling[u^2]
Out[8]= 234
In[9]:= c=Floor[ N[ 234* .02 + Sqrt[234*.02 * .98]*
             PhiInverse[.9]]]
Out[9]= 7
```

3.8 Other continuous distributions; Gamma function calculus

Thus far, we have encountered only two types of probability density functions: uniform, and Gaussian. In this section we will provide a number of other examples that are of importance in applied problems.

Recall that the one-dimensional probability density function $f(x)$ has to satisfy two conditions:

$$f(x) \geq 0, \qquad x \in \mathbf{R} \tag{1}$$

$$\int_{-\infty}^{\infty} f(x)\,dx = 1. \tag{2}$$

The corresponding cumulative distribution function

$$F(x) = \int_{-\infty}^{x} f(t)\,dt \tag{3}$$

is nondecreasing,

$$F(-\infty) = 0, \qquad F(\infty) = 1. \tag{4}$$

For the random quantity X with probability d.f. $f(x)$, the probability that X has values within the interval $(a, b]$ is

$$F(b) - F(a) = \int_a^b f(x)\,dx. \tag{5}$$

Also, by the Fundamental Theorem of Calculus,

$$f(x) = \frac{d}{dx}F(x). \tag{6}$$

The mean and variance of the random quantity X are

$$\mu(X) = \int_{-\infty}^{+\infty} xf(x)\,dx \tag{7}$$

and

$$\sigma^2(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)\,dx. \tag{8}$$

Example 3.8.1 Exponential Distributions.
The lifetime T of many devices is often a random quantity. For some of them, the experiments show that if the device survives up to time t then its remaining lifetime has the same probability distribution as that of a new device, displaying what we call the *memoryless behavior*. In terms of the cumulative distribution function $F_T(t) = \Pr\{T \le t)$, or, more conveniently, in terms of the corresponding *reliability* or *survival function* (upper tail distribution) $R(t) = 1 - F(t) = \Pr(T > t)$, the memoryless behavior can be written as the condition:

$$\frac{R(t+s)}{R(t)} = R(s), \qquad t, s > 0, \tag{9}$$

which can be rewritten in the form $R(t+s) = R(t)R(s)$. The latter equation, differentiatied with respect to t, yields $R'(t+s) = R'(t)R(s)$. Letting, $t \to 0$ we obtain a simple differential equation

$$R'(s) = -\lambda R(s), \qquad \lambda = -R'(0) > 0, \tag{10}$$

since $R(s)$ is a decreasing function. The obvious solution is $R(s) = e^{-\lambda s}$, $s, \lambda > 0$, and the corresponding (normalized) density is of the form

$$f_T(t) = \begin{cases} 0 & \text{for } t < 0; \\ \lambda e^{-\lambda t} & \text{for } t \ge 0. \end{cases} \tag{11}$$

The *exponential family* of probability d.f.s (11) is parametrized by the single parameter $\lambda > 0$, which is often called the *intensity* of the exponential distribution. Exponential distributions also appear as probability distributions of waiting times between random events (such as log-ons to the server in a local area network) whose number in a given time-interval has the Poisson distribution introduced in Section 3.5. The corresponding cumulative distribution function

$$F_T(t) = \begin{cases} 0 & \text{for } t < 0; \\ 1 - e^{-\lambda t} & \text{for } t \geq 0. \end{cases} \tag{12}$$

The mean and the second moment, via the integration by parts formula, are

$$\mu(T) = \int_0^\infty t\lambda \exp[-\lambda t]dt = \frac{1}{\lambda} \tag{13}$$

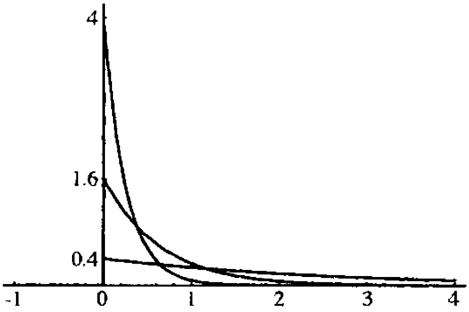
and the second moment

$$m_2(T) = \int_0^\infty t^2\lambda \exp[-\lambda t]dt = \frac{2}{\lambda^2}, \tag{14}$$

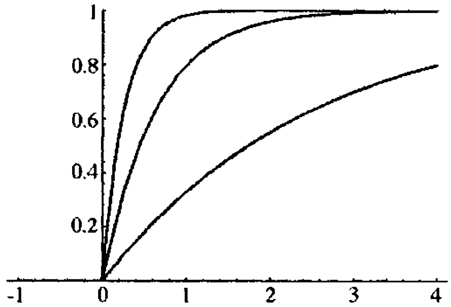
so that the variance $\sigma^2(T) = \lambda^{-2}$.

Mathematica Experiment 1. Exponential Distributions. We shall plot the graphs of exponential densities and cumulative d.f. for $\lambda = 0.4$, 1.6, 4.

```
In[1]:= <<Statistics`ContinuousDistributions`
In[2]:= H[x_]:=If[x<0,0,1]
In[3]:= dens [x_,l_]:=  H[x]*l*Exp[-l*x]
In[4]:= Plot[{dens[x,0.4],dens[x,1.6],dens[x,4]},{x,-1,4},
          PlotRange->{0,4.2}, Ticks->{Automatic,{0.4, 1.6, 4}}]
Out[4]= -Graphics-
```



```
In[5]:= cumdf[x_,l_]:=H[x]*(1-Exp[-x*l])
In[6]:= Plot[{cumdf[x,0.4],cumdf[x,1.6],cumdf[x,4]} ,{x,-1,4}]
Out[6]= -Graphics-
```



Computation of higher moments of the exponential distribution requires calculations with integrals of the type $\int x^k e^{-x}\,dx$, which gives us the opportunity to introduce the *gamma function* $\Gamma(x)$. The gamma function "calculus" is a very convenient tool in analysis of many probability d.f.s.

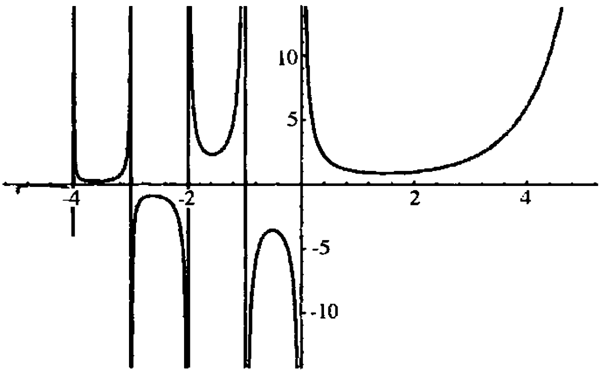
The gamma function is defined by the formula

$$\Gamma(\alpha) := \int_0^\infty x^{\alpha-1} e^{-x}\,dx. \tag{15}$$

Notice that the integral cannot be evaluated in closed form for most values of α. So, the gamma function $\Gamma(\alpha)$ is a new special transcendental function. It is easy to see that the integral (15) is well defined for $\alpha > 0$.

Mathematica Experiment 2. Gamma Function. We will obtain values of the gamma function and graph it. Actually, definition of the gamma function can be extended also to noninteger negative real numbers. Although a mathematical justification of this fact is beyond the scope of this book, we can explore the problem using *Mathematica*.

```
In[1]:= {Gamma[-2.9999], Gamma[3.32], Gamma[4], Gamma[14.3]}
Out[[1]= {-1666,88, 2.73975,  6,  1.3641 10^10  }
In[2]:= Plot[ Gamma[x], {x,-5, 5}]
Out[2]=  -Graphics-
```



There are few cases when one can obtain by analytic means the precise value for $\Gamma(\alpha)$:

$$\Gamma(1/2)^2 = \left(\int_0^\infty x^{-1/2} e^{-x} dx\right)^2 = \int_0^\infty \int_0^\infty (xy)^{-1/2} e^{-x-y} dx dy.$$

Substituting $u^2 = x$ and $v^2 = y$, we arrive at

$$\Gamma(1/2)^2 = 4 \int_0^\infty \int_0^\infty e^{-u^2-v^2}\, du dv$$

Finally, using polar coordinates (you have seen a similar "trick" before in calculations with Gaussian densities),

$$\Gamma(1/2)^2 = 4 \int_0^\infty \int_0^{\pi/2} e^{-r^2}\, d\phi r\, dr = \pi$$

Hence,

$$\Gamma(1/2) = \sqrt{\pi}. \tag{16}$$

Another useful identity is obtained by integration by parts:

$$\alpha\Gamma(\alpha) = [x^\alpha e^{-x}]_0^\infty + \int_0^\infty x^\alpha e^{-x}\, dx = \Gamma(\alpha + 1) \tag{17}$$

so that the gamma function behaves like the factorial but is defined for all positive real numbers rather than just for integers. Moreover, that connection is very direct as

$$\Gamma(1) = \int_0^\infty \exp[-x] dx = 1,$$

and, in view of (17),

$$\Gamma(n) = (n-1)! = (n-1)(n-2)...3 \cdot 2 \cdot 1. \tag{18}$$

Here are two applications of our gamma function calculus that provide alternative calculations for the Gaussian and exponential densities.

Example 3.8.2 Gaussian Densities Revisited.
For the $N(\mu, \sigma^2)$ random quantity X, we have

$$\int_{-\infty}^\infty \phi(z)\, dz = 2 \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(t-\mu)^2}{2\sigma^2}}\, dt$$

$$= \sqrt{\frac{1}{\pi}} \int_0^\infty s^{-1/2} e^{-s}\, ds = \sqrt{\frac{1}{\pi}} \Gamma(1/2) = 1.$$

and, substituting $y = (x - \mu)^2/(2\sigma^2)$,

$$\sigma^2(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \phi(x) dx = 2 \int_0^\infty \frac{2\sigma^2 y}{\sqrt{2\pi\sigma^2}} e^{-y} \frac{\sigma^2 dy}{\sqrt{2\sigma^2 y}}$$

$$= \frac{2\sigma^2}{\sqrt{\pi}} \int_0^\infty y^{1/2} e^{-y}\, dy$$

$$= \frac{2\sigma^2}{\sqrt{\pi}} \Gamma(3/2) = \frac{2\sigma^2}{\sqrt{\pi}} (1/2)\Gamma(1/2) = \sigma^2.$$

Example 3.8.3 Exponential Densities Revisited.
For the exponential distribution, the mean

$$\mu = \int_0^\infty x\lambda e^{-\lambda x}\, dx = \frac{1}{\lambda}\int_0^\infty y^{2-1} e^{-y}\, dy = \frac{1}{\lambda}\Gamma(2) = \frac{1}{\lambda},$$

and the second moment

$$m_2 = \int_0^\infty x^2 \lambda e^{-\lambda x}\, dx = \frac{1}{\lambda^2}\int_0^\infty y^{3-1} e^{-y}\, dy = \frac{1}{\lambda^2}\Gamma(3) = \frac{2}{\lambda^2}.$$

Other interesting and more complex examples of the gamma function calculus' applications can be found in Chapter 5.

Example 3.8.4 Reliability Analysis; Weibull and Rayleigh Distributions.
Recall that the reliability (survival) function of a device is defined by the formula

$$R(t) = \Pr(T > t), \qquad t \geq 0, \tag{19}$$

where T is the random quantity describing the lifetime of the device. Let $T_1, T_2, \ldots,$ be a sequence of independent random quantities with the same distribution as T. For a fixed t and small Δt we have, as $n \to \infty$,

$$\frac{\#\{i \leq n : T_i \in [t, t+\Delta t]\}}{\#\{i \leq n : T_i > t\}} \to \frac{R(t) - R(t+\Delta t)}{R(t)} \approx \frac{R'(t)}{R(t)}\Delta t = \Lambda(t)\Delta t,$$

where the hazard function

$$\Lambda(t) := \frac{R'(t)}{R(t)}, \tag{20}$$

need not be constant as was the case for the memoryless exponential distribution case of Example 3.8.1. It gives the infinitesimal rate of change of the failure rate once the device survived until time x. The differential equation (20) can be immediately solved to yield

$$R(t) = \exp\left[-\int_0^t \Lambda(s)\,ds\right]. \tag{21}$$

In the special case when the hazard function is of the power form

$$\Lambda(t) = \frac{\alpha}{\beta}t^{\alpha-1}, \qquad \alpha, \beta > 0, \tag{22}$$

we obtain

$$R(t) = \exp[-t^\alpha/\beta], \qquad x > 0, \tag{23}$$

with the corresponding *Weibull*, or *stretched exponential*, probability d.f.

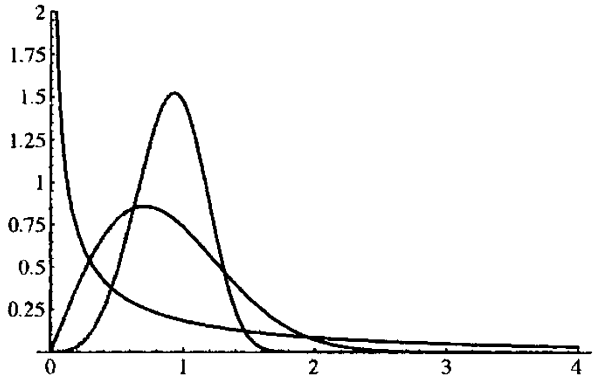
$$f(t; \alpha, \beta) = \begin{cases} (\alpha/\beta)t^{\alpha-1}\exp[-t^\alpha/\beta], & \text{for } t > 0; \\ 0, & \text{for } t \le 0. \end{cases} \tag{24}$$

If $\alpha = 1$, the Weibull probability d.f. $f(t; 1, \beta)$ is just the familiar exponential distribution with intensity $\lambda = 1/\beta$. In the case $\alpha = 2$, the Weibull distribution is known as the *Rayleigh distribution*. Note that the Rayleigh distribution corresponds to the linear hazard function

$$\Lambda(x) = 2x/\beta. \tag{25}$$

Mathematica Experiment 3. Weibull Densities. We plot Weibull densities with parameters $\alpha = 0.5, 2, 4$, and $\beta = 1$. It is clear that for bigger α, the density $f(t; \alpha, \beta)$ decays faster as $t \to \infty$.

```
In[1]:= <<Statistics'ContinuousDistributions'
In[2]:= WeiDen[t_,a_,b_]:= PDF[WeibullDistribution[a,b], t]
In[3]:= Plot[{WeiDen[t,0.5,1], WeiDen[t,2,1], WeiDen[t,4,1]},
          {t, 0.01, 4}, PlotRange->{0, 2}]
Out[3]= -Graphics-
```



A more general version of the Weibull distribution can be obtained by shifting the origin to the point v. The extended family of Weibull densities $f(t;\alpha,\beta,v)$ is zero for $t \le v$, and for $t > v$, it is given by the formula

$$f(t) = \left(\frac{\beta}{\alpha}\right)\left(\frac{t-v}{\alpha}\right)^{\beta-1} \exp\left[-\left(\frac{t-v}{\alpha}\right)^{\beta}\right] \tag{26}$$

where $v \in \mathbf{R}$, $\alpha, \beta > 0$ are parameters. The corresponding cumulative distribution function is again 0 for $x \le v$, and for $x > v$ it is given by

$$F(t;\alpha,\beta,v) = 1 - \exp\left[-\left(\frac{t-v}{\alpha}\right)^{\beta}\right]. \tag{27}$$

Example 3.8.5 Detection of Particles; Cauchy Densities.

Consider a source located at the point with coordinates $(0,\eta)$ emitting particles in the half-plane with uniformly distributed random directions (angles) $\Theta \in [-\pi/2, \pi/2]$ (see Fig. 3.8.1). The particles are being detected by a flat panel device D (represented by the vertical line $x = \tau$) at the distance τ from the source. What is the distribution of the random quantity representing the position Y particles on the detecting device? Clearly,

$$\begin{aligned} F_Y(y;\eta,\tau) &= \Pr\{Y \le y\} = \Pr\{\tan\Theta \le (y-\eta)/\tau\} \qquad (28) \\ &= \Pr\{\Theta \le \arctan((y-\eta)/\tau)\} = \frac{1}{2} + \frac{1}{\pi}\arctan((y-\eta)/\tau). \end{aligned}$$

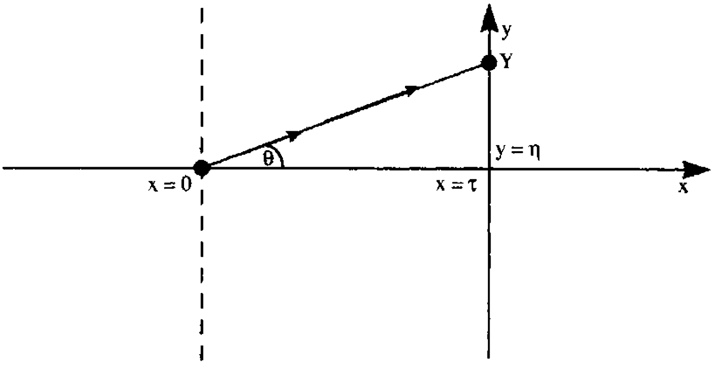


FIGURE 3.8.1
From radially uniform to Cauchy distribution: detection, on a flat panel, of particles emitted by a point source.

The corresponding *Cauchy density* with the *location parameter* η and the *scale parameter* τ is given by the formula

$$f_Y(x;\eta,\tau)=\frac{1}{\pi\tau(1+[(y-\eta)/\tau]^2)} \tag{29}$$

for any real y. In the physical sciences, Cauchy densities are often called *Lorentz densities*. The location parameter η, the median, plays for the Cauchy distribution the role similar to the mean of a Gaussian distribution. The mean of the Cauchy distribution, however, does not exist. Indeed, the function $x/(1+x^2)$ is not integrable over the whole real line. The scale parameter τ measures the dispersion of the Cauchy distribution around its location parameter η but it is not its standard deviation. The latter does not exist either.

Finally, observe that the tail probabilities

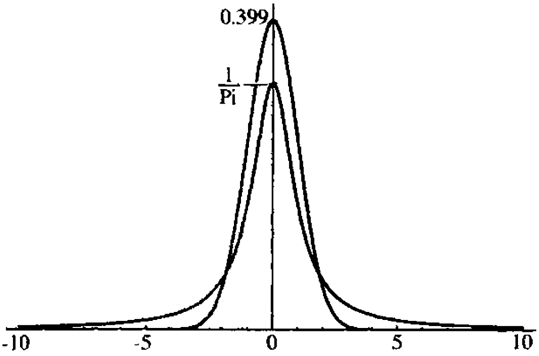
$$\Pr\{Y>a\}=\int_a^\infty \frac{\tau}{\pi(\tau^2+(y-\eta)^2)}\,dy\sim\frac{C}{a} \tag{30}$$

decay much slower, as $a\to\infty$, than those of Gaussian distributions, which decay at the rate $e^{-a^2/2}/a$, despite a superficial similarity of the graphs of two densities. For that reason, Cauchy distributions are often described as "heavy-tailed".

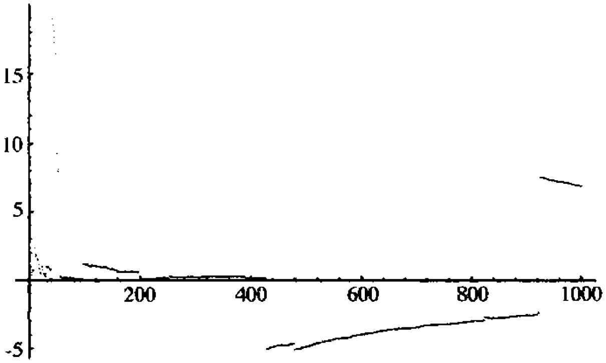
Mathematica Experiment 4. Cauchy Densities. We shall begin by comparing the densities of the Gaussian and Cauchy distributions, both with parameters 0 and 1. Around the origin, the Cauchy density has a sharper peak than the Gaussian density, but is much, much flatter far away from the origin. The lack of mean and variance raises an interesting question about the validity of the Law of Large

Numbers and the Stability of Fluctuations Law for independently repeated Cauchy experiments.

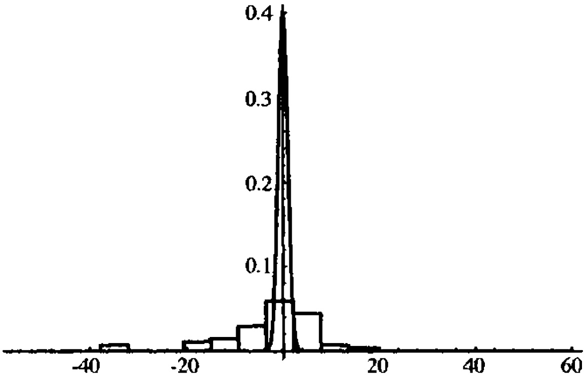
```
In[1]:= <<Statistics'ContinuousDistributions'
In[2]:= UVW'DataRep'
In[3]:= Plot[{PDF[CauchyDistribution[0.,1.],x],
          PDF[NormalDistribution[0.,1.],x]},{x,-10,10},
          PlotRange->{0,0.42}, Ticks->{Automatic,{Pi^(-1), 0.399}}]
Out[3]= -Graphics-
```



```
In[4]:= {PDF[CauchyDistribution[0.,1.],10],
          PDF[NormalDistribution[0.,1.],10]}
Out[4]= {0.00315158,  7.6946 10^(-23)}
In[5]:= data=Table[Random[CauchyDistribution[0.,1.]],{1000}]
Out[5]= {2.23794, -2.43299, 0.463945, 7.96682,  ... ,
           -24.9201,   -0.782545, -2.06564, 1.38234, -0.832504}
In[6]:= LargeNumbers[data]
Out[6]= -Graphics-
```



```
In[7]:= CentralLimit[Data,0,1,10]
Out[7]= -Graphics-
```



Note the dramatic difference in values of the Gaussian and Cauchy probability d.f.s at $x = 10$. The experiments (and they should be repeated several times to get a feel for the variety of pictures one can get here) indicate that neither LLN nor SFL holds true. The averages oscillate wildly, and the centered and rescaled by $\sqrt{n}$ averages just produce flatter and flatter histograms which do not seem converge to the $N(0, 1)$ density. This is marked on the histograms but sometimes it is so concentrated around zero in comparison with the former that it is almost invisible. This is a warning that neither LLN nor SFL should be expected to hold without any precondition; some assumptions (like those mentioned in previous sections) are necessary.

The distributions described in the next five examples will find applications in Chapters 5, 7, and 8.

Example 3.8.5. Continuous Pareto Distributions. The density is given by the formula

$$f(x) = \begin{cases} 0 & \text{for } x < 1; \\ x^{-\alpha} / \int_1^\infty u^{-\alpha} du & \text{for } x \geq 1. \end{cases} \tag{31}$$

The parameter α has to be > 1 to guarantee that the integral in (31) remains finite. These distributions appeared first in economics applications. The mean does not exist for $1 < \alpha \leq 2$.

Example 3.8.6 Gamma Distribution.

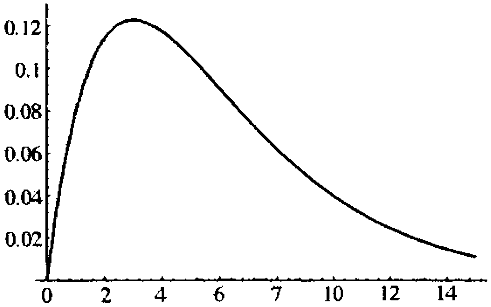
A gamma distribution is an absolutely continuous distribution with density

$$f(x; \alpha, \beta) = \begin{cases} (\beta^\alpha \Gamma(\alpha))^{-1} x^{\alpha-1} \exp[-x/\beta], & \text{for } x > 0; \\ 0, & \text{for } x \leq 0. \end{cases} \tag{32}$$

Parameters α, β are positive numbers and the distribution is concentrated on the positive half-axis. Substituting $y = x/\beta$ we find that $\int_0^\infty x^{\alpha-1} \exp[-x/\beta]\, dx = \beta^\alpha \Gamma(\alpha)$, so that $f(x; \alpha, \beta)$ satisfies the normalization condition $\int f(x)\, dx = 1$.

Mathematica Experiment 5. Gamma Density.

```
In[1]:= f[x_, a_, b_]:=x^(a-1)*Exp[-x/b]/((b^a) *  Gamma[a])
In[2]:= Plot[f[x,2.0,3.0],{x,0.0,15}]
Out[2]= -Graphics-
```



The mean

$$\mu = \int_0^\infty x\, f(x;\alpha,\beta)\, dx = \frac{\beta\Gamma(\alpha+1)}{\Gamma(\alpha)} \int_0^\infty f(x;\alpha+1,\beta)\, dx = \alpha\beta, \qquad (33)$$

and the second moment

$$m_2 = \int_0^\infty x^2 f(x;\alpha,\beta)\, dx = \frac{\beta^2\Gamma(\alpha+2)}{\Gamma(\alpha)} \int_0^\infty f(x;\alpha{+}2,\beta)\, dx = \alpha(\alpha{+}1)\beta^2, \qquad (34)$$

so that the variance

$$\sigma^2 = \alpha(\alpha+1)\beta^2 - \alpha^2\beta^2 = \alpha\beta^2. \qquad (35)$$

Probabilities for the gamma distribution are computed using the *incomplete gamma function*

$$\Gamma(\alpha, z) := \int_z^\infty x^{\alpha-1} \exp[-x]\, dx \qquad (36)$$

which in *Mathematica* is called `Gamma[alpha,z]`. If the random quantity X has a gamma distribution with parameters α and β, then

$$\Pr\{u < X \le v\} = \int_u^v \frac{1}{\beta^\alpha\Gamma(\alpha)} t^{\alpha-1} \exp[-t/\beta]\, dt$$

$$= \int_{u/\beta}^{v/\beta} \frac{1}{\Gamma(\alpha)} s^{\alpha-1} \exp[-s]\, ds \qquad (37)$$

$$= \frac{\Gamma(\alpha, u/\beta) - \Gamma(\alpha, v/\beta)}{\Gamma(\alpha)}$$

Gamma density with parameters $\alpha = 1$ and $\beta = 1/\lambda$ is just the exponential density.

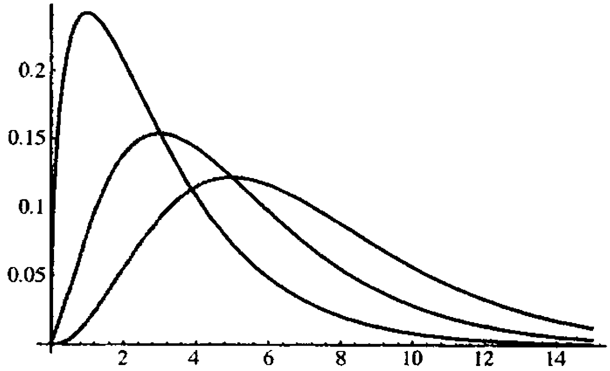
Example 3.8.7 The Chi-Square Distribution.
The density of the χ^2-distribution

$$f(x) = f(x; n) := \begin{cases} \frac{(1/2)^{n/2} x^{(n/2)-1}}{\Gamma(n/2)} e^{-x/2}, & \text{for } x > 0; \\ 0 & \text{for } x \le 0. \end{cases} \quad (38)$$

$n = 1, 2, \ldots,$ is a gamma density with parameters $\alpha = n/2$ and $\beta = 2$. The parameter n is called the *number of degrees of freedom* of the χ^2-distribution.

Mathematica Experiment 6. Chi-Square Distribution.

```
In[1]:= <<Statistics'ContinuousDistributions'
In[2]:= f[x_,n_]:= PDF[ChiSquareDistribution[n],x]
In[3]:= Plot[{f[x,3], f[x,5], f[x,7]}, {x,0,15}]
Out[3]= -Graphics-
```



It follows from Example 3.8.6 that its mean and variance are, respectively,

$$\mu = n, \qquad \sigma^2 = 2n. \quad (39)$$

As we will see in Chapter 5, the sum of squares $X_1^2 + \ldots + X_n^2$ of n independent $N(0, 1)$ random quantities has the χ^2-distribution with n degrees of freedom.

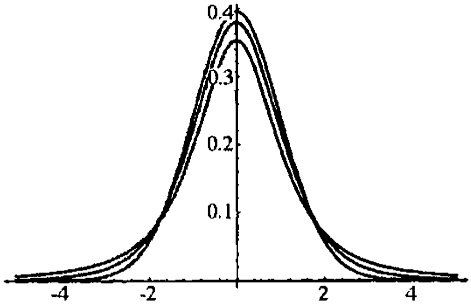
Example 3.8.8 The Student t-Distribution.
The density is given by the formula

$$f(x;n)=\frac{\Gamma\left((n+1)/2\right)}{\sqrt{n\pi}\,\Gamma(n/2)}\left(1+\frac{x^2}{n}\right)^{(n+1)/2}, \qquad x\in\mathbf{R}. \tag{40}$$

The parameter n is called the *number of degrees of freedom* of the t-distribution. It is the distribution of the random quantity $X/\sqrt{Y_n/n}$, where X is a $N(0,1)$ random quantity and Y_n is an independent random quantity with χ^2-distribution with n degrees of freedom.

Mathematica Experiment 7. Student t-Distribution.

```
In[1]:= <<Statistics‘ContinuousDistributions'
In[2]:= f[x_,n_]:= PDF[StudentTDistribution[n],x]
In[3]:= Plot[{f[x,2], f[x,5], f[x,35]}, {x,-5,5}]
Out[3]= -Graphics-
```



For larger n, the Student t-distribution's density clearly approaches $N(0,1)$ density.

Example 3.8.9 Fisher's F-Distribution.
The density of an F-distribution with n, m degrees of freedom is concentrated on the positive half-line and defined there by the formula

$$f(x;n,m)=\frac{(n/m)^{n/2}}{B(n/2,m/2)}x^{(n/2)-1}\left(1+\frac{n}{m}x\right)^{-(n+m)/2}, \qquad x>0, \tag{41}$$

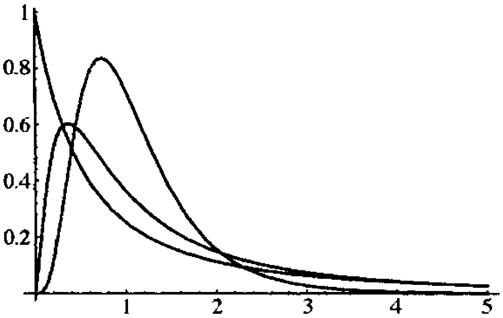
where

$$B(n/2,m/2)=\frac{\Gamma(n/2)\Gamma(m/2)}{\Gamma((n+m)/2)}.$$

It is the distribution of the ratio $(X_n/n)/(Y_m/m)$, where X_n is a random quantity with the χ^2-distribution with n degrees of freedom, and Y_m is an independent random quantity with the χ^2-distribution with m degrees of freedom.

Mathematica Experiment 8. Fisher F-Distribution.

```
In[1]:= <<Statistics‘ContinuousDistributions‘
In[2]:= f[x_,n_,m_]:= PDF[FRatioDistribution[n,m],x]
In[3]:= Plot[{f[x,2,2], f[x,5,3], f[x,10,20]}, {x,0.01,5}]
Out[3]= -Graphics-
```



3.9 Testing the fit of a distribution

Suppose a random sample $x_1, \dots, x_n$, has been obtained from n independent repeated random experiments $X_1, \dots, X_n$, with the common cumulative distribution function $F(x) = \Pr\{X \le x\}$ which is unknown to us. If we want to find $F(x)$, the first thought is to approximate it by the cumulative relative frequencies

$$\frac{\#\{i \le n : x_i \le x\}}{n} \approx \Pr(X \le x) = F(x), \tag{1}$$

in the spirit of the Law of Large Numbers. Actually, the random quantities

$$\hat{F}_n(x) := \frac{\#\{i \le n : X_i \le x\}}{n} = \frac{1}{n} \sum_{i=1}^{n} H(x - X_i), \tag{2}$$

where

$$H(x) = \begin{cases} 0, & \text{for } x < 0; \\ 1, & \text{for } x \ge 0. \end{cases}$$

is the Heaviside unit step function, are called the *empirical distribution functions* and one can prove the following crucial fact.

Glivenko-Cantelli Law. *For large sample size n, the empirical distribution $\hat{F}_n(x)$ uniformly approximates the true distribution $F(x)$, i.e.,*

$$\lim_{n\to\infty} \max_{-\infty<x<\infty} |\hat{F}_n(x) - F(x)| = 0.$$

The obvious next question is: How good is this approximation in the statistical sense? The surprising answer is provided by the following important result. For another goodness-of-fit test, see Section 8.6. Note that the random quantity $|\hat{F}_n(x) - F(x)|$ takes values in the interval $[0, 1]$.

Kolmogorov-Smirnov Distribution. *For any continuous d.f. $F(x)$, the distribution of the nonnegative random quantity*

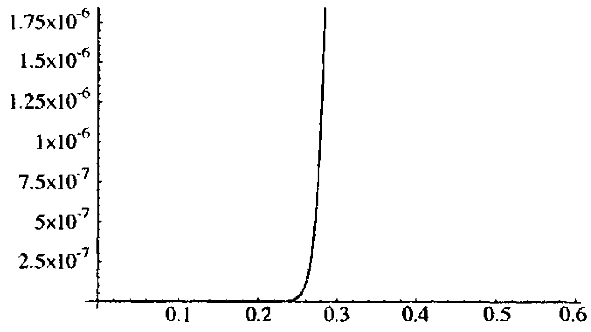
$$D_n := \max_{-\infty<x<\infty} |\hat{F}_n(x) - F(x)|$$

is independent of $F(x)$ and, for every $z \geq 0$,

$$\lim_{n\to\infty} \Pr\{D_n\sqrt{n} \leq z\} = K(z) \equiv 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}\exp[-2k^2z^2].$$

Mathematica Experiment 1. Kolmogorov-Smirnov Distribution. The evaluation of the Kolmogorov-Smirnov cumulative d.f., which is represented by an infinite series, can be implemented in *Mathematica*; the necessary computations take, however, a long time on the average platform, so it is worth it to experiment with $K(z)$ a little bit more cautiously. Just trying to plot it does not give good results and you will see why.

```
In[1]:= K[z_]:= 1-2*Sum[(-1)^(k-1)*Exp[-2*k^2*z^2],{k,1,Infinity}]
In[2]:= Plot[K[z],{z,0,1}]
Out[2]= -Graphics-
```

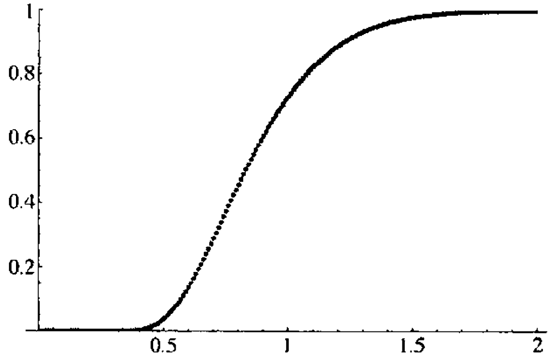


Clearly, *Mathematica* has a problem with huge differences in scales of different values of $K(z)$. So, let us explore the individual values at discrete points.

```
In[3]:= KS1={K[0.1], K[0.2], K[0.3], K[0.4], K[0.5], K[0.6]}
Out[3]= {-3.90334 10^(-15), 5.05041 10^(-13), 9.3058 10^(-6) ,
         0.00280767, 0.0360548,  0.135717}
```

So, the values of $K(z)$ become significantly positive only around $z = 0.3$. Now, instead of trying to graph $K(z)$ as a continuous function, let us just `ListPlot` its values at 0.01 intervals.

```
In[4]:= KS2=Table[{K[0.01*k], 0.01*k},{k,1,199}]
Out[4]= {{0.01,-2.22 10^(-16)}, ... ,{1.99, 0.999}}
In[5]:= ListPLot[KS2]
Out[5]= -Graphics
```



Of course, the initial tiny negative numbers appeared only because of round-off errors in machine arithmetic. Actually, the first one has a special name, `$MachineEpsilon`, and it gives the distance between 1.0 and the closest number which has a distinct binary representation.

The information contained in the Kolmogorov-Smirnov Theorem can be used in two different ways: to construct confidence intervals for true $F(t)$ and to test hypotheses about potential candidates for $F(t)$.

In the first mode, we can select z_α so that $K(z_\alpha) = \alpha$ and claim that the random strip

$$[\hat{F}_n(x) - z_\alpha/\sqrt{n}, \hat{F}_n(x) + z_\alpha/\sqrt{n}] \tag{3}$$

around the empirical distribution $\hat{F}(x)$ is an $\alpha \times 100$ percent confidence region for the true distribution function $F(x)$.

In the second mode, given $\alpha \in (0, 1)$ we can check if a candidate distribution function $G(x)$ lies inside the confidence region (3). If it does not, then we can reject the hypothesis that $G(x)$ is true. The probability that $G(x)$ does not lie inside the confidence region, while it is the true distribution function, is at most $1 - \alpha$, so that if α is selected close to 1, then probability of this type of error is small. Note that in case $G(x)$ is inside the confidence region, no decision is made.

3.10 Random vectors and multivariate distributions

A random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ is a random quantity assuming vector values. It has an absolutely continuous distribution, if it possesses a multivariate probability d.f., that is, a function

$$f(\boldsymbol{x}) = f(x_1, \ldots, x_d) \geq 0 \tag{1}$$

which is a nonnegative function of d-variables, normalized so that

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_d)\, dx_1 \ldots dx_d = 1, \tag{2}$$

and such that the probabilities of events concerning $\boldsymbol{X}$ can be calculated in terms of multiple integrals of $f(\boldsymbol{x})$. More precisely, if A is a subset of the d-space $\mathbf{R}^d$, then

$$\Pr\{\boldsymbol{X} \in A\} = \int \cdots \int_A f(x_1, \ldots, x_d)\, dx_1 \ldots dx_d. \tag{3}$$

The mean of the i-th component X_i in the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ is calculated via the formula

$$\mu(X_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_i f(x_1, \ldots, x_N)\, dx_1 \ldots dx_N. \tag{4}$$

Their variances are

$$\sigma^2(X_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (x_i - \mu_i)^2 f(x_1, \ldots, x_N)\, dx_1 \ldots dx_N. \tag{5}$$

and covariances between the component random quantities X_i and X_j are

$$\operatorname{Cov}(X_i, X_j) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (x_i - \mu_i)(x_j - \mu_j) f(x_1, \ldots, x_N)\, dx_1 \ldots dx_N. \tag{6}$$

They form a $d \times d$ matrix. It is not sufficient to know the one-dimensional densities of the individual components X_i to calculate the covariances. Also notice that by integrating out all the variables in the density $f(x_1, \ldots, x_d)$ except the i-th one, we obtain the probability d.f. (*marginal density*) of the random component X_i:

$$f_{X_i}(x) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_N) \times$$

$$dx_1 \dots dx_{i-1}\, dx_{i+1} \dots dx_N. \tag{7}$$

The components $X_1, \dots, X_d$ of the random vector X with multivariate density $f_{\boldsymbol{X}}(\boldsymbol{x})$ are independent, if and only if

$$f_{\boldsymbol{X}}(x_1, \dots, x_d) = f_{X_1}(x_1) \cdot \dots \cdot f_{X_d}(x_d). \tag{8}$$

Obviously, in view of (6) and (8), if $X_1, \dots, X_d$ are independent then their covariances are zero.

Similar definitions and formulas apply to the multivariate discrete probability distribution, with the integrals replaced by finite or infinite sums.

Example 3.10.1 Multinomial Distribution.
Each component $k_1, k_2, \dots, k_d$ of an d-dimensional vector $\boldsymbol{k} = (k_1, \dots, k_d)$ can take nonnegative integer values from 0 to n under the additional constraint that their sum equals n. The probabilities are distributed according to the multinomial formula, i.e.,

$$f(\boldsymbol{k}) = \frac{n!}{k_1!k_2!\dots k_d!} p_1^{k_1} \cdot p_2^{k_2} \cdot \dots \cdot p_d^{k_d} \tag{9}$$

if $k_1 + \dots + k_d = n$, and is 0 otherwise, where $p_1, p_2, \dots, p_d$ are parameters such that

$$p_1 + p_2 + \dots + p_d = 1.$$

The fact that all these probabilities add up to 1 follows from the multinomial formula which is an extension of the binomial formula.

Example 3.10.2 Bivariate Normal Distribution.
The general two-dimensional density of a normal random vector (X_1, X_2) is given by the formula

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times \tag{10}$$

$$\exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\right]\right).$$

A linear change of variables can reduce the double integral of this density to a product of single integrals and prove its normalization. More tedious calculations on double integrals show that the coefficients $\mu_1 = \mu(X_1), \mu_2 = \mu(X_2)$ are means of the first and second components, respectively, and that $\sigma_1^2 = \sigma^2(X_1), \sigma_2^2 = \sigma^2(X_2)$ are corresponding variances. The components are themselves one-dimensional Gaussian quantities. The parameter ρ turns out to be the *correlation coefficient* between the components X_1 and X_2, that is

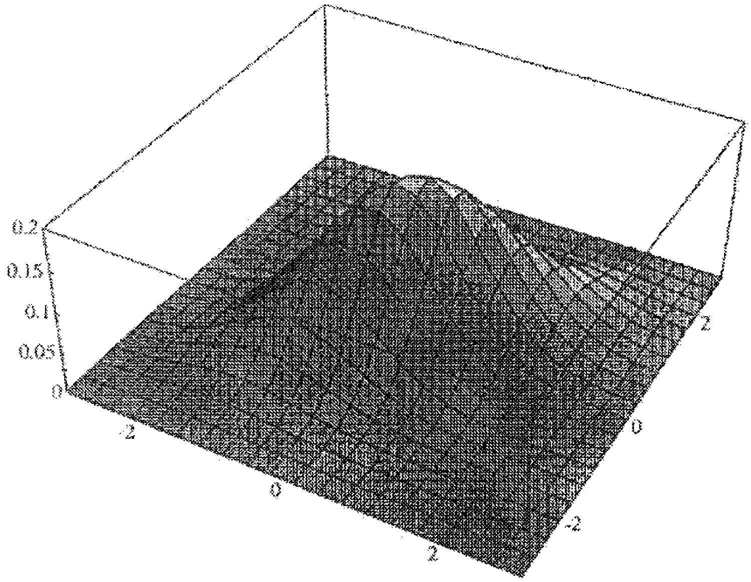
$$\mathrm{Cov}\,(X_1, X_2) = \rho\sigma(X_1)\sigma(X_2). \tag{11}$$

A quick check of (10) and (7) shows that the components X_1, X_2, of a Gaussian random vector $\boldsymbol{X}$ are independent *if and only if* the correlation coefficient $\rho = 0$. With the correlation coefficient ρ defined in general by the equality (11), this is not necessarily true if the random vector is not Gaussian.

The graph of the surface representing the bivariate normal density is obtained in the next *Mathematica Experiment.* Note that the level curves are ellipses with half-axes σ_1, σ_2 and correlation parameter ρ determines the angle by which these half-axes are rotated with respect to the x_1, x_2 axes.

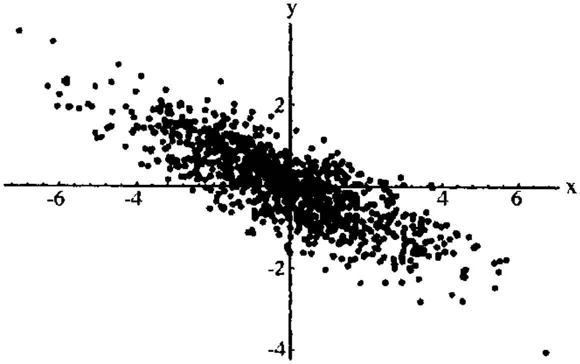
Mathematica Experiment 1. Bivariate Normal Distributions. The bivariate (and, in general, multivariate) normal distribution is specified by the command `MultinormalDistribution[mu, sigma]` where `mu` is the mean vector (μ_1, μ_2), and `sigma` is the covariance matrix $(\rho\sigma_i\sigma_j)$.

```
In[1]:= <<Statistics`MultinormalDistribution`
In[2]:= f[x1_,x2_,m_,s_]:=PDF[MultinormalDistribution[m,s],
         {x1,x2}]
In[3]:= m={0,0}
Out[3]= {0,0}
In[4]:= s={{1, 1/Sqrt[3]},{1/Sqrt[3],1}}
Out[4]= {{1, 1/Sqrt[3]},{1/Sqrt[3],1}}
In[5]:= Plot3D[f[x1,x2,m,s],{x1,-3,3},{x2,-3,3},
         PlotRange->{0, 0.2}]
Out[5]= -Graphics-
```

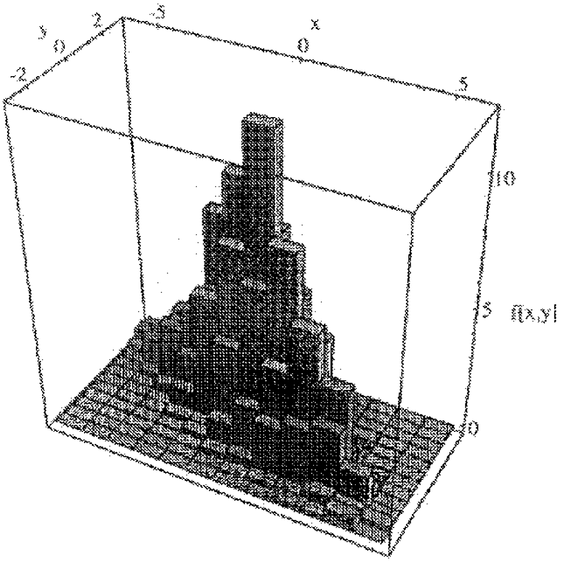


Next, we will create a simulated random sample of size $n = 1000$ of a Gaussian random vector with means $\mu_1 = 0, \mu_2 = 0$, variances $\sigma_1 = 2, \sigma_2 = 1$, and the correlation coefficient $\rho = -0.8$, taking advantage of the `RSNormal2D[sigma1, sigma2,rho, n]` command of the `UVW'ContSamp'` package. Then the sample will be pictured on the scatter plot and its two dimensional histogram will be produced. Finally, we will check that the projection of the data on the line with direction vector (0.5, 1.6) are approximately Gaussian.

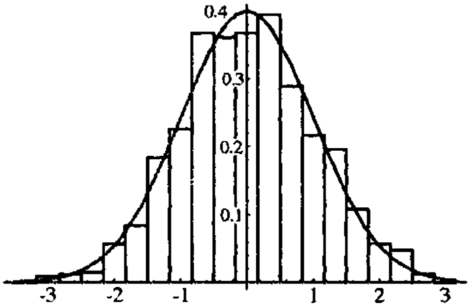
```
In[1]:= <<UVW'ContSamp'
In[2]:= <<UVW'DataRep'
In[3]:= gauss2d=RSNormal2D[2,1,-0.8,1000]
Out[3]= {{4.78287, -1.83111}, ... , {-3.68319, 0.967511}}
In[4]:= SamplePlot2D[gauss2d]
Out[4]= -Graphics-
```



```
In[5]:= Histogram2D[gauss2d,-6,6,12,-3,3,12]
Out[5]= -Graphics3D-
```



```
In[6]:= proj=gauss2d.Transpose[{0.5,1.6}]
Out[6]= {-0.538344, ... , 1.47689}
In[7]:= CentralLimit[proj,0,1,1]
Out[7]= -Graphics-
```



Example 3.10.3 Uniform Multivariate Distributions.

A d-dimensional random vector $\boldsymbol{X}$ is said to have a uniform distribution over a set $A \subset \mathbf{R}^d$ if, for any set $B \subset \mathbf{R}^d$,

$$\Pr\{\boldsymbol{X} \in B\} = |A \cap B|, \tag{12}$$

where $|C|$ stands for the d-dimensional measure (area, volume, etc.) of a set C. In other words, the density of X is

$$f(\boldsymbol{x}) = \frac{1}{|A|}\mathbf{1}_A(\boldsymbol{x}) = \begin{cases} 1/|A|, & \text{for } \boldsymbol{x} \in A, \\ 0, & \text{for } \boldsymbol{x} \notin A. \end{cases} \tag{13}$$

Of course, the uniform distribution over A makes sense only if the measure $|A| < \infty$.

Distributions of various functions of the random vector $\boldsymbol{X}$, themselves random quantities, may then be calculated. For example, the cumulative d.f. $F_R(r)$ of the distance from the origin

$$R = \sqrt{X_1^2 + \ldots + X_d^2} \tag{14}$$

is, for $\boldsymbol{X}$ uniformly distributed over A, of the form

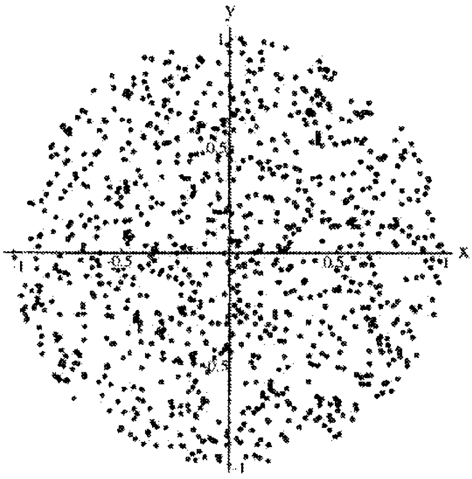
$$\begin{aligned} F_R(r) &= \Pr\{R \le r\} = \Pr\{X_1^2 + \ldots + X_d^2 \le r^2\} \\ &= \int\int\int_{\{|\boldsymbol{x}| \le r\}} \mathbf{1}_A(\boldsymbol{x})\, dx_1 \cdot \ldots \cdot dx_d = |B_d(r) \cap |A|, \end{aligned} \tag{15}$$

where $B_d(r)$ is a d-dimensional ball centered at the origin with radius r. In the particular case where the $\boldsymbol{X}$ is uniformly distributed over $B_d(1)$, the d-dimensional unit ball,

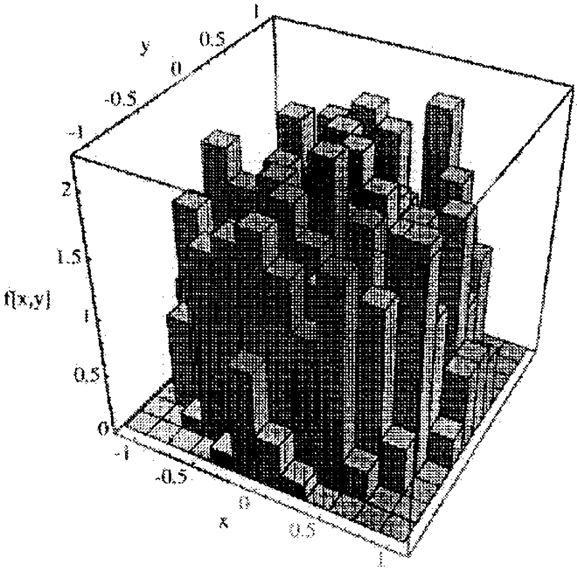
$$F_R(r) = \begin{cases} 0, & \text{if } r \le 0; \\ |B_d(r)|/|B_d(1)| = r^d, & \text{of } 0 \le r \le 1; \\ 1, & \text{if } r \ge 1. \end{cases} \tag{16}$$

Mathematica Experiment 2. Uniform Distribution on the Unit Ball. We shall produce two random samples of size $n = 1000$ from the uniform distribution on the unit balls of dimensions 2 and 4, and then check the histograms of their distances from the origin, i.e., their norms R. These should be compared with the result in formula (16) which gives the density of R to be $f_R(r) = 2r$ in dimension 2, and $f_R(r) = 4r^3$ in dimension 4, both, of course, concentrated on the interval $[0, 1]$

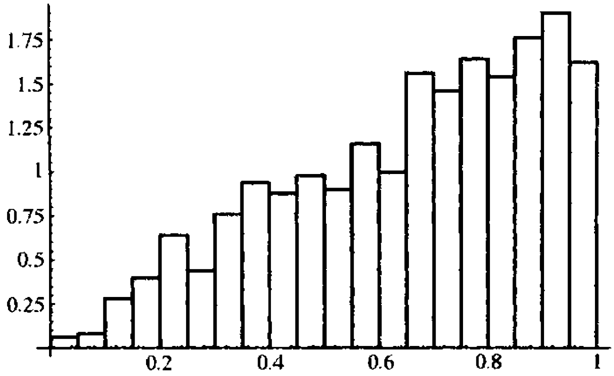
```
In[1]:= <<UVW'ContSamp'
In[2]:= <<UVW'DataRep'
In[3]:= ball2d=RSUnitBall[2,1000]
Out[3]= {{-0.208536, -0.292938}, ... , {-0.355194, -0.0873841}}
In[4]:= SamplePlot2D[ball2d,AspectRatio->1]
Out[4]= -Graphics-
```



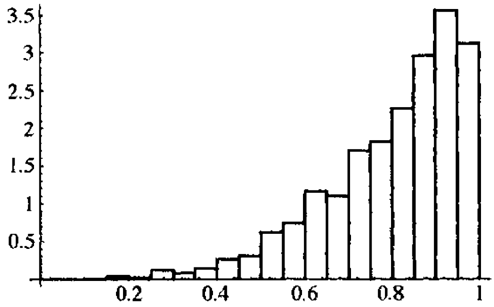
```
In[5]:=  Histogram2D[ball2d,-1.1 ,1.1 ,12,-1.1 ,1.1 ,12]
Out[5]= -Graphics3D-
```



```
In[6]:= norms=Sqrt[Table[ball2d[[i]].ball2d[[i]],{i,1000}]]
Out[6]= {0.359583, ... , 0.365785}
In[7]:= RegularHisto[norms,0,1,20]
Out[7]= -Graphics-
```



```
In[8]:= ball4d=RSUnitBall[4,1000];
In[9]:= norms=Sqrt[Table[ball4d[[i]].ball4d[[i]],{i,1000}]]
In[10]:= RegularHisto[norms,0,1,20]
Out[10]= -Graphics-
```



Example 3.10.4 Multivariate Normal Distribution.

The density f of an d-dimensional normal random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ (written here as a column) is now a function of the column vector $\boldsymbol{x} = (x_1, \ldots, x_d)^T$, and is given by the formula

$$f(\boldsymbol{x}) = f(x_1, \ldots, x_d) = \tag{17}$$

$$\frac{1}{\sqrt{(2\pi)^d \det[C(\boldsymbol{X})]}} \exp\Big(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T [C(\boldsymbol{X})]^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\Big),$$

where

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$$

is the column vector of means of the component random quantities $X_i, i = 1, \ldots, d$. and the $d \times d$ matrix $[C(\boldsymbol{X})]$ is the *covariance matrix* of the above components, i.e.,

$$[C(\boldsymbol{X})] = \left[\operatorname{cov}(X_i, X_j)\right]_{1 \le i, j \le d}. \tag{18}$$

For a Gaussian random vector $\boldsymbol{X}$ with independent $N(0, 1)$ components, that is with the covariance matrix $C(\boldsymbol{X})$ which is the identity d-dimensional matrix with ones on the diagonal and zeros off the diagonal, the distance R from the origin expressed by the formula (14), has the chi-square distribution with d degrees of freedom (see Example 3.8.7).

3.11 Experiments, exercises, and projects

1. In 1,000 Bernoulli experiments, the sample mean $\bar{x} = \hat{p} = 0.48$. Calculate the confidence level of the interval 0.48 ± 0.03.

2. For the same experiments, calculate the size of the confidence interval given the confidence level 0.95.

3. Find the number n of Bernoulli experiments needed to obtain for p a confidence interval of size 0.02, at confidence level 0.99.

4. Random quantity X has the *Laplace density* $f_X(x; \mu, s) = ce^{-|x-\mu|/s}$ defined on the whole real line. Calculate c. Using *Mathematica*, graph this probability d.f. for different values of parameters μ and $s > 0$, and the corresponding cumulative d.f. $F_X(x; \mu, s)$. Calculate the mean, variance, and the n-th moment of X using the gamma-function calculus.

5. Calculate the mean, variance, and the cumulative d.f. for the generalized Pareto density

$$f(x; \alpha, c) = \begin{cases} c^{\alpha-1}(\alpha - 1)x^{-\alpha}, & \text{if } c \le x < \infty; \\ 0, & \text{elsewhere.} \end{cases}$$

What restrictions are necessary on parameters α and c?

6. *Mathematica Experiment. Simulation with a Given Density.* Use `UVW` packages to simulate and analyze random samples with a prescribed density. A trial run is given below. Read first the detailed description of the packages given in the Appendix.

```
In[1]:= UVW‘ContSamp‘
In[2]:= UVW‘DataRep‘
In[3]:= f[x_]:=1+Sin[x]
In[4]:= samp= RSContinuousDistribution[f,0,10,2000]
In[5]:= g1 = RegularHisto[samp, 0,10,20, DisplayFunction->Identity]
In[6]:= integral=NIntegrate[f[x],{x,0,10}]
In[7]:= fnorm[x_]:= f[x]/integral
In[8]:= g2 = Plot[fnorm[x],{x,0,10}, DisplayFunction->Identity]
In[9]:= Show[g1,g2, DisplayFunction->$DisplayFunction]
```

7 *Mathematica Experiment. Simulation with a Given Discrete Distribution.* Use `UVW‘DiscSamp‘` package to simulate and analyze random samples with a prescribed discrete distribution. Follow the lines of Experiment 6. Select a discrete distribution yourself.

8. *Mathematica Experiment. Simulation with a Given Bivariate Density.* Use `UVW` packages to simulate and analyze two-dimensional random samples with independent components with a prescribed density. A trial run is given below. Read first the detailed description of the packages given in the Appendix.

```
In[1]:= UVW‘ContSamp‘
In[2]:= UVW‘DataRep‘
In[3]:= f[x_]:=x^2
In[4]:= para2d= RSIndependent2D[f,-1,1,f,-1,1, 2000]
In[5]:= SamplePlot2D[para2d,Frame->True]
In[6]:= Histogram2D[para2d, -1,1,8,-1,1,8]
In[7]:= marge1=transpose[para2d][[1]];
In[8]:= RegularHisto[marge1,-1,1,10]
```

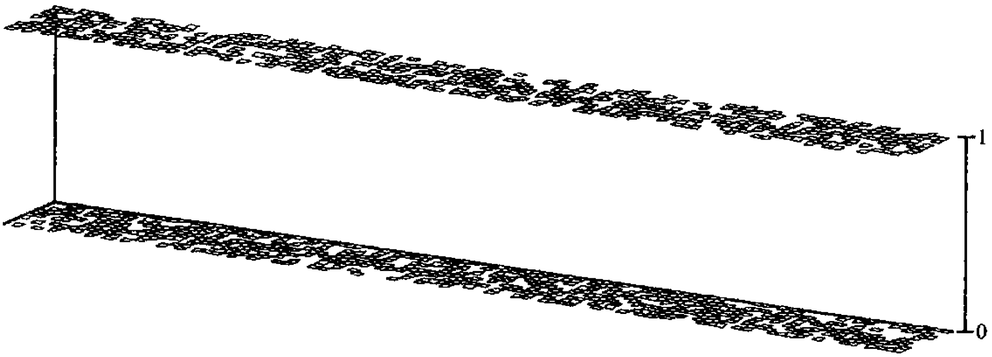
9. Simulate a random sample of size $n = 1000$ with an exponential distribution and plot its histogram against the exponential density. Verify the Law of Large Numbers and the Stability of Fluctuations Law for these data. Use `UVW` and/or `Statistics` packages.

10. Adjust the proof of the Poisson limit behavior (3.5.5) of the binomial distribution to show its generalization (3.5.9) for arbitrary intensities μ.

11. Simulate sums of independent Cauchy random quantities but normalize them by n^{-1} and $n^{-3/2}$ instead of the usual SFL $\sqrt{n}$. Draw the corresponding histograms. Comment on what you see.

12. Implement the Kolmogorov-Smirnov test of fit for the exponential distribution (find its mean first) and the water drops data contained in the file `DROPS` located on the `UVW Web Site`. Do the Q-Q plots as well.

13. Devise the *Mathematica* code to produce the following display of the outcomes of 100 repetitions of Bernoulli series of length 10.



14. Calculate the mean and variance of the geometric and negative binomial distributions of Section 3.5. Use *Mathematica* if your analytic tools fail you.

15. In the experiments supporting the SFL, draw the smoothed histograms of rescaled sums by shifting them by $\Delta x = k/10$th of the bin size, for $k = 0, 1, \ldots, 9$ and then averaging the 10 shifted histograms. Compare graphically the smoothed histograms with the $N(0, 1)$ density.

3.12 Bibliographical notes

For more theoretical details on the Law of Large Numbers and the Stability of Fluctuations Law (Central Limit Theorem) consult Chapter 5 and/or any mathematically rigorous probability text such as

[1] P. Billingsley, ***Probability and Measure***, Wiley-Interscience, New York, 1987,

which also contains a proof of the Glivenko-Cantelli Law. The latter and the Kolmogorov-Smirnov distribution are discussed in great depth in the monograph

[2] G.R. Shorack and J.A. Wellner, *Empirical Processes and Applications to Statistics*, Wiley-Interscience, New York, 1986.

A good source on multivariate distributions is

[3] T.W. Anderson, *An Introduction to Multivariate Statistical Analysis*, Second Edition, Wiley-Interscience, New York, 1984.

The Monte Carlo method is exhaustively discussed in

[4] G. Fishman, *Monte Carlo*, Springer-Verlag, New York, 1996.

An incomparable encyclopedia of all kinds of discrete and continuous distributions is the four-volume compendium:

[5] N. Johnson, S. Kotz, and A. Kemp, *Univariate Discrete Distributions*, Wiley-Interscience, New York, 1992.

[6] N. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions 1*, 2nd ed., Wiley-Interscience, New York, 1994.

[7] N. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions 2*, 2nd ed., Wiley-Interscience, New York, 1995.

[8] N. Johnson and S. Kotz, *Continuous Multivariate Distributions*, Wiley-Interscience, New York, 1972.

Part II

MODELING UNCERTAINTY

Chapter 4

Algorithmic Complexity and Random Strings

In this chapter we will try to get to the heart of the notion of randomness by showing its fundamental connection with several concepts of algorithmic and computational complexity. Although the discussion illuminates the philosophical underpinnings of the concept of randomness for a concrete string of data, the conclusions are sobering: perfectly random strings cannot be produced by any finite algorithms (read, computers). A practical way out of this dilemma is suggested.

4.1 Heart of randomness: when is random – random ?

When one begins to contemplate the notion of randomness it is not unreasonable to start with consulting the relevant entry in the *American Heritage Dictionary of the English Language*. Here is what it says:

> **ran·dom** *adj.* **1.** Having no specific pattern or objective; lacking causal relationships; haphazard. **2.** *Statistics.* **a.** Of or designating a phenomenon that does not produce the same outcome or consequences every time it occurs under identical circumstances. **b.** Of or designating an event having a relative frequency of occurrence that approaches a stable limit as the number of observations of the event increases to infinity. **c.** Of or designating a sample drawn from a population so that each member of the population has an equal chance to be drawn. **d.** Of or pertaining to a member of such a sample: *a random number.* —See Synonyms at **chance.** —**at random.** Without definite method or purpose; unsystematically: *"accusa-*

M. Denker, W. Woyczynski, *Introductory Statistics and Random Phenomena*,
Modern Birkhäuser Classics, DOI 10.1007/978-3-319-66152-0_4

> *tions are not made at random, but form part of a coherent whole*" (Denis Baly). [Middle English *randoun*, from Old French *randon*, haphazard, from *randir*, to run, from Frankish *rant*†(unattested), a running.] — **ran′dom·ly** *adv.*

Colloquially, randomness means a complete lack of discernible "rules" governing a given phenomenon and the lack of "predictability". However, such a broad and loosely worded definition is unlikely to produce a rigorous theory of random phenomena. To make our thinking on the subject more precise, let us begin by being more specific about what we mean by a "rule".

Let us return to the five binary strings

(a) 1111111111111111111111111

(b) 10101010101010101010101010

(c) 10010011100100111001001110010011

(d) 011011100101110111100010011010

(e) 0110111001011101111000100110101

which were introduced at the beginning of Chapter 1. To what extent do we perceive them as random? Or more exactly, how would we order them depending on their perceived degree of randomness? To answer questions like this, we need to analyze what is so special about each of them that makes them feel more or less random.

The following three properties of *binary strings* (i.e., finite *words* in the alphabet $\mathcal{A}$ consisting of two letters 0 and 1) seem intuitively acceptable as fundamental attributes of "randomness":

(ML) The string has no exceptional features that stand out in the sense of belonging to a large majority (reasonably defined) of strings, or perhaps, more cautiously, of being part of a large majority in the universe of strings of a given type. Such strings will be called here **typical**, or Per *Martin-Löf random*, as he was the person who crystallized this concept.

(vM) Frequencies of 0s and 1s in the string are stable under (admissible) rules of subsequence selection. Strings satisfying this property will be called here Richard *von Mises random*.

(K) The string has a complex minimal description in the sense of not being easily described via easily discernible rules governing the alternation of 0's and 1's. Such strings will be called here Andrei Nikolaevich *Kolmogorov random* or **computationally complex**. The computational complexity

> of a problem is its intrinsic difficulty as measured by the time, space or other quantity *required* for its solution. In other words, the computational complexity is the minimal cost of an algorithm which solves the problem.

An effort to formalize the idea of a complex sequence which has no simple description rules might go as follows. Intuitively, the string of length n can always be described by n bits of information, that is, by a complete listing of its bits. But shorter descriptions may be possible. However, defining the computational complexity of a string $x_0, \ldots, x_{n-1} \in \mathcal{A}^*$, where $\mathcal{A}^*$ denotes the family of all finite strings in the alphabet $\mathcal{A}$, as the length of its shortest description in the natural language leads to the Richard-Berry Paradox discussed in Section 1.1. The expression

> *The smallest number that cannot be defined in less than twenty words*

itself contains less than 20 words.

The initial suspicion is that the problem lies in the "vagueness" of the natural English language. However, an example provided below shows that the natural language is not the main culprit here.

Example 4.1.1 Richard-Berry Paradox in a Formal Programming Language. Consider a programming language **PL** strong enough to define all natural numbers. A description of a natural number n in **PL** is a pair $D = (P, m)$, where P is a *program* in PL and the *input m* is a natural number. The program P would take m as an input, produce n as an output, and would stop (see Fig. 4.1.1) □

$$\xmapsto{m} \quad P \quad \xmapsto{n} \quad PRINT \quad \longmapsto \quad STOP$$

FIGURE 4.1.1
Description of n in a formal language.

For any natural number, there exists a trivial description

$$D_n = (P_n, n), \tag{1}$$

where P_n is the following program:

```
BEGIN                                                    (2)
READ n
```

```
PRINT $n$
STOP
```

By definition, the length of a description $D = (P, m)$ is the sum of the number of characters in program P and the number of bits in input m. So, for the trivial description (1)

$$\text{length } D_n = 20 + \log_2 n + 1, \tag{3}$$

where 2 is the base in the binary representation for n.

Now, the question of the computational complexity of a natural number n reduces to the optimization problem: For a given $n \in \mathbf{N}$, find in **PL** a description $D = (P, m)$ of n such that the length of D is minimal. Clearly, the trivial description D_n is not always optimal. Indeed, take the integer of the form $n = 2^k$, $k \geq 35$, and consider the following program R_n:

```
BEGIN                                                   (4)
READ $k$
$i = 1$
$n = 2$
WHILE $i \neq k$ DO: $n = n \cdot 2$, $i = i + 1$
PRINT $n$
STOP
```

The program has 49 characters, 29 more than P_n, but, since for $k \geq 35$ we have $k < 2^{k-29}$, for large k the length of the trivial description (D_n, n) is greater than that of the description (R_n, k).

It turns out that even in the formal programming language **PL** *the optimization problem has no solution* as long as **PL** is sufficiently strong to satisfy the following conditions:

(a) It contains a program which, for every given program P in **PL**, computes the number of characters in program P and halts.

(b) It accepts subroutines.

(c) It performs some basic algorithms, including the WHILE construction.

(d) It uses a finite alphabet.

(e) It works with natural numbers written in a base $p \geq 2$.

Each natural number $n \in \mathbf{N}$ has at least one definition, for example the trivial definition (D_n, n) , and since the alphabet $\mathcal{A}$ used by **PL** is finite, the number of all definitions of length $\leq$ length (D_n, n) is finite. So, the optimization problem has (at least) one solution. However, the machine using **PL** would not be able to find it. Indeed, assume to the contrary that there exists a program $\mathcal{P}$ in **PL** which produces such a shortest definition $(\mathcal{P}(n), m(n))$ using input $m(n)$ (see Fig. 4.1.2).

$$\overset{m}{\longmapsto} \quad \mathcal{P} \quad \overset{n}{\longmapsto} \quad \begin{array}{c} \mathit{SHORTEST} \\ \mathit{DEFINITION} \\ \mathit{OF}\ \ n \end{array} \quad \longmapsto \quad \mathit{STOP}$$

FIGURE 4.1.2
Program producing the shortest definition.

Then consider, for all natural $l \in \mathbf{N}$, the following program Q_l:

```
BEGIN                                                        (5)
READ $l$
$y = 0$
$z = 0$
WHILE $z < l$ DO:
  { CALL $\mathcal{P}(y)$,
  $z = \text{length}\,(\mathcal{P}(y)) + \text{length}\,(m(y))$
  $y = y + 1$}
PRINT $y$
STOP
```

In view of the assumptions, it is a correct program. It prints the smallest natural number requiring definition of length $> l$. But, on the other hand, for some constant $c > 0$

$$\text{length}(Q_l) = \text{length}(l) + c < \log_2 l + 2 + c < l, \qquad (6)$$

for l large enough. A contradiction.

It turns out that the resolution of this foundational difficulty can be achieved by considering the shortest description of a string for a *given* descriptive process (computer). Before we make these ideas more precise, we need to introduce the concept of a *computable string*, or a *computable function*.

4.2 Computable strings and the Turing machine

In this section we will try to formalize the idea of an *effective description*, or a *computable string*. The need for such a precaution follows from the discussion of the previous section. The device we choose is a symbolic computer called the *Turing machine*. It was proposed by the English mathematician Alan Turing, of the *Enigma* decoding machine fame; the latter was credited with turning the tide in the World War II Battle of Britain.

The Turing machine is a symbolic computing device that consists of the following two components:

(a) A control unit (CU), with a finite number of possible states $s_1, \ldots, s_m \in \mathcal{S}$, one of which is called the halting state;

(b) A (possibly) infinite memory tape divided into square (memory) cells, each cell containing one symbol of a finite alphabet $a_1, \ldots, a_z \in \mathcal{A}$ or a blank B (see Fig. 4.2.1).

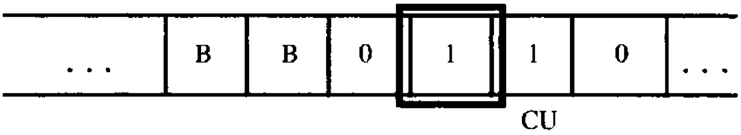


FIGURE 4.2.1
The Turing machine with a binary alphabet.

The machine operates in discrete steps as follows:

- At step 0 there is a finite contiguous (i.e., surrounded by blanks B) input written on the tape in the alphabet $\mathcal{A}$. CU is positioned over the left-most cell containing the input.
- At any step, CU is in one of the states $s_1, \ldots, s_m \in \mathcal{S}$, scans the cell directly underneath, noting which character of the alphabet $a_1, \ldots, a_z \in \mathcal{A}$ appears in the memory cell. Then, depending on its own state and the character in the currently scanned memory cell, it performs one of the following four actions:

(1) Changes its own state according to a finite matrix

$$\begin{array}{c|cccc|} & a_1 & a_2 & \ldots & a_z \\ s_1 & s_{11} & s_{12} & \ldots & s_{1z} \\ s_2 & s_{21} & s_{22} & \ldots & s_{2z} \\ & \ldots & \ldots & \ldots & \ldots \\ s_m & s_{m1} & s_{m2} & \ldots & s_{mz} \end{array}. \qquad (1)$$

(2) If the new state is the halting state, the machine halts. If its new state is not the halting state, then the machine:

(3) Moves one cell to the right (R) or to the left (L) according to a finite matrix

$$\begin{array}{cc} & \begin{array}{cccc} a_1 & a_2 & \dots & a_z \end{array} \\ \begin{array}{c} s_1 \\ s_2 \\ \\ s_m \end{array} & \begin{pmatrix} d_{11} & d_{12} & \dots & d_{1z} \\ d_{21} & d_{22} & \dots & d_{2z} \\ \dots & \dots & \dots & \dots \\ d_{m1} & d_{m2} & \dots & d_{mz} \end{pmatrix} \end{array} \tag{2}$$

where $d_{ij} = R \text{ or } L$;

(4) Changes the symbol on the currently scanned cell according to a finite matrix

$$\begin{array}{cc} & \begin{array}{cccc} a_1 & a_2 & \dots & a_z \end{array} \\ \begin{array}{c} s_1 \\ s_2 \\ \\ s_m \end{array} & \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1z} \\ a_{21} & a_{22} & \dots & a_{2z} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mz} \end{pmatrix} \end{array}. \tag{3}$$

Note that each Turing machine can be defined as a partial (i.e., not defined for all its arguments; this fact is indicated by the circle placed on top of the arrow) function

$$T : \mathcal{S} \times \mathcal{A} \ni (s, a) \quad \overset{\circ}{\longmapsto} \quad T(s, a) \in \mathcal{A} \cup \{B, L, R\}, \tag{4}$$

with the understanding that the control unit halts when it faces the pair (s, a) for which the above function T is not defined.

Every Turing machine generates a partial function

$$M_T : \mathcal{A}^* \ni x \overset{\circ}{\longmapsto} M_T(x) \in \mathcal{A}^*, \tag{5}$$

where, as usual, $\mathcal{A}^*$ denotes the set of all finite strings in the alphabet $\mathcal{A}$, by assigning a finite binary string $M_T(x) \in \mathcal{A}$ to a finite binary string $x \in \mathcal{A}$ as follows:

(i) Write the string x on a blank tape;

(ii) Place CU over the left-most letter of the string symbol of x and run it until it halts;

(iii) Select as $M_T(x)$ the maximal string (surrounded by blanks B) of which some letter is scanned when the machine comes to a halting state.

The Turing machine permits us to introduce the correct definition of a *computable* (*recursive, effective*) function.

Definition 4.2.1 Computable Functions.
The function f which assigns finite strings $f(x) \in \mathcal{A}^*$ to certain finite strings $x \in \mathcal{A}^*$ is said to be computable if there exists a Turing machine T such that $f(x) = M_T(x)$.

Remark 4.2.1 *Computational Complexity.* In practice, it is important to know how long it takes to compute the computable function for an input of given length, or how much space on the memory tape it would take to complete such a computation. It is simply not feasible to expect answers, even for reasonable input length, to problems that take an ***exponential time*** to solve, i.e., such that for an input of length n it takes $\approx 10^n$ steps to find the value of f. Even for $n = 100$, and a processor taking a billion steps per second, it would take about 10^{91} seconds to solve. With about $3 \cdot 10^7$ seconds in a year, that is some $3 \cdot 10^{83}$ years, much more than the age of the Universe.

For that reason, it is important to distinguish between problems in the *class $\mathcal{P}$-time* that can be solved in polynomial time (i.e., in time less then n^c, where n is the input length and c is a certain positive constant), or problems in the *class $\mathcal{P}$-space* that use a polynomial amount of memory, and other problems which are not thought to be feasible. There are many important problems that do not appear to be in the class $\mathcal{P}$ but can be demonstrated to have feasible (polynomial) algorithms for a nondeterministic Turing machine which permits a random outcome. Such problems are said to be in the *class $\mathcal{NP}$*. It is known that the famous *Hamilton circuit problem*, that is a problem of deciding whether a graph G of n vertices has a path that visits each vertex exactly once and returns to the starting point, is in the class $\mathcal{NP}$. It is known that

$$\text{class}\,\mathcal{P}\text{–time} \quad \subseteq \quad \text{class}\,\mathcal{NP} \quad \subseteq \quad \text{class}\,\mathcal{P}\text{–space}, \tag{6}$$

but it remains an open question whether or not these inclusions are proper.

Remark 4.2.2 *Martians and the Universal Turing Machine.* There exists a *universal* Turing machine U such that for any Turing machine T there exists a binary string p_T (a compiler of T in the language of U) such that for all strings $x \in \mathcal{A}^*$

$$M_T(s) = M_U(p_T s), \tag{7}$$

where $p_T s$ is a concatenation of strings p_T and s. Intuitively speaking, p_T gives a program for machine T on the universal machine U. Such a universal Turing machine can be constructed effectively (see references in the Bibliographical Notes at the end of this chapter). It follows that Martians, humans, and computers will all approximately agree on the intrinsic complexity of n bits (for large n) of *War and Peace*, the Mona Lisa, and a Bernoulli sequence with parameter p. Experimentally, the first two have been determined to be of the order $n/3$ and $n/10$, respectively, and the last one can be computed to be $n(-p\log p - (1-p)\log(1-p))$—the entropy of the Bernoulli sequence (see Section 2.8 and Kolmogorov's work on the complexity of works of art described in the *Annals of Probability* volume quoted in Bibliographical Notes).

Remark 4.2.3 *Undecidability of the Halting Problem.* The number of all Turing machines is effectively denumerable, that is it is possible to provide an effective (computable) one-to-one pairing

$$\mathbf{N} \ni n \longleftrightarrow T_n \tag{8}$$

between natural numbers and Turing machines. For this reason, the question "which machine computations eventually terminate with a definite result and which machine computations go on forever without a definite conclusion?" is *undecidable*. This surprising result can be formulated rigorously as:

Turing Lemma. *There is no computable function f such that for all $n \in \mathbf{N}$ and $\boldsymbol{x} \in \mathcal{A}^*$, we have $f(n, \boldsymbol{x}) = 1$ if $M_{T_n}(\boldsymbol{x})$ is defined, and $f(n, \boldsymbol{x}) = 0$ otherwise.*

PROOF Suppose to the contrary, and define a partial recursive function $\psi(\boldsymbol{x})$ by $\psi(\boldsymbol{x}) = 1$ if $f(\boldsymbol{x}, \boldsymbol{x}) = 0$, and $\psi(\boldsymbol{x})$ is undefined otherwise (remember that f is totally recursive). Let ψ have an index k in the fixed enumeration (1) of partial recursive functions, i.e., $\psi = M_{T_k}$. Then $M_{T_k}(k)$ is defined if and only if $f(k, k) = 0$, according to ψ's definition. But this contradicts the assumption of existence of f as defined in the statement of the lemma. ∎

The above proof depends on the "diagonalization" argument invented by Georg Cantor to prove that the set of all real numbers is not denumerable. The Turing Lemma itself is directly related to Kurt Gödel's famous 1931 *incompleteness theorem* stating that there are statements of (Peano's) arithmetic that are unprovable within such a system.

Remark 4.2.4 *Formal Definition of Recursive Functions.* More formally, the set of *recursive functions* can be defined as the smallest set of functions containing

(A) Successor functions

$$\mathrm{Succ}_i^{\mathcal{A}} : \mathcal{A}^* \ni \boldsymbol{x} \longmapsto \mathrm{Succ}_i^{\mathcal{A}}(x) = a_i\boldsymbol{x} \in \mathcal{A}^*;$$

(B) Constant functions

$$C_y^{\mathcal{A}} : (\mathcal{A}^*)^n \ni (x_1, \ldots, x_n) \longmapsto C_y^{\mathcal{A}}(x_1, \ldots, x_n) = y \in \mathcal{A}^*;$$

(C) Projection functions

$$P_i^{\mathcal{A}} : (\mathcal{A}^*)^n \ni (x_1, x_2, \ldots, x_n) \longmapsto P_i^{\mathcal{A}}(x_1, \ldots, x_n) = x_i \in \mathcal{A}^*;$$

and which is closed under function composition and primitive recursion. The latter is defined as follows: $f : (\mathcal{A}^*)^{n+1} \mapsto \mathcal{A}^*$ is obtained by primitive recursion from $g : (\mathcal{A}^*)^n \mapsto \mathcal{A}^*$ and $h_i : (\mathcal{A}^*)^{n+2} \to \mathcal{A}^*, i = 1, \ldots, p$ if

$$f(x_1, \ldots, x_n, \lambda) = g(x_1, \ldots, x_n),$$

$$f(x_1, \ldots, x_n, \mathrm{Succ}_i^{\mathcal{A}}(y)) = h_i(x_1, \ldots, x_n, y, f(x_1, \ldots, x_n, h))$$

for all $i = 1, \ldots p,\ x_1, \ldots, x_n, y \in \mathcal{A}^*$. A concatenation $\mathrm{con}_2(\mathcal{A}^*)^2 \to \mathcal{A}^*$ defined by the formula $\mathrm{con}_2(x, y) = xy$ is a good example here.

4.3 Kolmogorov complexity and random strings

Having established in the previous section the formal notion of computable (recursive) functions, we are now prepared to introduce the correct notion of complexity of a fixed string. Consider a finite alphabet

$$\mathcal{A} = \{a_1, \ldots, a_z\} \tag{1}$$

consisting of $z \geq 2$ letters and a (partial) computable function

$$\phi : \mathcal{A}^* \times \mathbf{N} \ni (x, n) \longmapsto \phi(x|n) \in \mathcal{A}^*, \tag{2}$$

where, as before, $\mathcal{A}^*$ is the set of all finite words in the alphabet $\mathcal{A}$ (including the empty word $\emptyset$).

Definition 4.3.1 Kolmogorov Complexity.
The Kolmogorov complexity induced by ϕ is a function

$$K_\phi : \mathcal{A}^* \times \mathbf{N} \ni (\boldsymbol{x}, n) \longmapsto K_\phi(\boldsymbol{x}|n) \in \mathbf{N} \cup \{\infty\}, \tag{3}$$

defined by the formula

$$K_\phi(\boldsymbol{x}|m) = \begin{cases} \min\{\mathrm{length}\,(\boldsymbol{y}) : \boldsymbol{y} \in \mathcal{A}^*,\ \phi(\boldsymbol{y}|m) = \boldsymbol{x}\}, & \text{if } \boldsymbol{x} = \phi(\boldsymbol{y}|m) \\ & \text{for some } \boldsymbol{y} \in \mathcal{A}^*; \\ \infty, & \textit{otherwise.} \end{cases}$$

In other words, for each positive integer n, we have an *effective* dictionary $\phi(..|n)$ translating some finite words y into the finite word x, and the related Kolmogorov complexity of x is the length of the shortest word y that the dictionary would translate into x or $+\infty$ if there are no words translatable into x. To gain an intuitive understanding of this concept, let us go through a number of examples.

Example 4.3.1 Trivial Dictionary.
Let $\phi : \mathcal{A}^* \times \mathbf{N} \to \mathcal{A}^*$ be defined via a single, independent of n, trivial computable function $\phi(x|n) = x$, for every $x \in \mathcal{A}^*$ and $n \in \mathbf{N}$. Then

$$K_\phi(x|m) = \text{length}\,(x), \qquad x \in X^*, n \in \mathbf{N}. \tag{4}$$

□

Example 4.3.2 Single Nontrivial Dictionary.
Let $f : \mathbf{N} \overset{\circ}{\mapsto} \mathcal{A}^*$ be a (partial) computable function. In other words, we have a single dictionary containing effectively a denumerated list of finite words. Define

$$\phi_f(x|n) = f(n), \qquad \text{for all } x \in \mathcal{A}^*, n \in \mathbf{N}. \tag{5}$$

If $x = f(n)$ for some n, then $\phi_f(\emptyset|n) = x$, where $\emptyset$ is the empty word. Consequently,

$$K_{\phi_f}(x, m) = \begin{cases} 0, & \text{if } x = f(m) \\ \infty, & \text{otherwise.} \end{cases} \tag{6}$$

Simply stated, if a word is listed in our dictionary, then its complexity is 0, and if it is not listed, then its complexity is infinite. □

Example 4.3.3 Series of Dictionaries, Each Containing Words of Fixed Length.
Consider a partial recursive function $\phi : \mathcal{A}^* \times \mathbf{N} \overset{\circ}{\to} \mathcal{A}^*$ given by the formula

$$\phi(x|n) = \begin{cases} x, & \text{if length}\,(x) = n \\ \infty, & \text{otherwise.} \end{cases} \tag{7}$$

Then

$$K_\phi(x|n) = \begin{cases} \text{length}\,(x), & \text{if length}\,(x) = n \\ \infty, & \text{otherwise.} \end{cases} \tag{8}$$

□

Example 4.3.4 "Optimal" Coding of Integers.
Consider positive integers written in the binary form (i.e., finite strings written in the alphabet $\mathcal{A} = \{0, 1\}$, with $z = 2$). In the set $\mathcal{A}_n^*$ of all strings of fixed length n introduce the following special enumeration: first, group strings in order of increasing number of 1s in each string, and then, within each group, order strings lexicographically. For instance, the $2^4 = 16$ strings of length 4 would thus be ordered as follows:

0000
0001, 0010, 0100, 1000,
0011,0101, 0110, 1001,1010,1100,
0111,1011,1101,1110,
1111.

Now, the idea is to encode sparse strings $x \in \mathcal{A}_n^*$ more economically by their number, No.(x), in the above enumeration. The number No.(x) of a string x with small frequency m of 1s will be relatively small, so the length of its description in the new enumeration would be small as well. Given our discussion of the information contents in Section 2.8, it is not surprising to see that if the relative frequency $p = m/n$ of 1s in string x is $\leq 1/2$, then the number of binary digits required to encode No.(x) is approximately (on the average) equal to

$$nH(m/n), \tag{9}$$

where H is the binary entropy function

$$H(p) = -p\log_2 p - (1-p)\log_2(1-p), \qquad 0 < p < 1. \tag{10}$$

If our description of the new enumeration system itself requires, say, c binary digits, then the length of the new description of a string $x \in \mathcal{A}_n^*$ containing exactly m 1s is, approximately,

$$nH(m/n) + c. \tag{11}$$

Observe that for large n and small frequencies m/n of 1s this number is certainly smaller than n—the length of string x. Take, for example, two binary strings of length 32:

00001001100000010100000000100000

01001110100111101000001100101101

and note that for the first one the ratio $m/n = 6/32$ and for the second one it is 1/2. So, the first has a shorter definition. □

Let us put this example in the context of Kolmogorov complexity. For each $n \in \mathbf{N}$, define

$$\phi(\text{No.}(x)|n) = x,$$

that is the number No.(x) (say, expressed as a string in the binary form) of the string x in the new enumeration is mapped into x. For example, in the case of $n = 3$, we get the following table of values of the (computable) function ϕ:

$$\begin{pmatrix} x \\ \phi(x|3) \end{pmatrix} = \begin{pmatrix} 0 & 1 & 10 & 11 & 100 & 101 & 110 & 111 \\ 000 & 001 & 010 & 100 & 011 & 101 & 110 & 111 \end{pmatrix}$$

If we denote by $m(x)$ the number of 1s in the string x, then the result of the above intuitive analysis can be rewritten in the form: for all $x \in \mathcal{A}^*$

$$K_\phi(x|\text{length}\,(x)) \approx \text{length}\,(x) \cdot H\left(\frac{m(x)}{\text{length}\,(x)}\right) \tag{12}$$

whenever

$$\frac{m(x)}{\text{length}\,(x)} \leq \frac{1}{2}. \tag{13}$$

You will be asked to confirm this result experimentally in the Experiments, Exercises, and Projects section at the end of this chapter.

At this point we are ready to define a Kolmogorov random string as a string of maximal, or close to maximal, Kolmogorov complexity. More formally, we have:

Definition 4.3.2
A string $x \in \mathcal{A}^*$ is called Kolmogorov random if

$$K(x|\text{length}\,(x)) \geq \text{length}\,(x). \tag{14}$$

Given an integer m, $0 < m <$ length(x), a string $x \in \mathcal{A}^*$ is called Kolmogorov m-random if

$$K(x|\text{length}\,(x)) \geq \text{length}\,(x) - m. \tag{15}$$

It is natural to ask the question: *Are there any random or m-random strings?* The answer is yes and it depends on a couple of simple estimates that are rigorously proved below. They also give a taste of what can be accomplished within the rigorous complexity theory. Recall that the general assumption is that the strings $x \in \mathcal{A}^*$ are written in the finite alphabet $\mathcal{A} = \{a_1, \ldots, a_z\}$, $z \geq 2$.

Proposition 4.3.1
For all positive integers $n, r \in \mathbf{N}$, $n \geq r$, there are at most z^r strings of length n and Kolmogorov complexity r, i.e.,

$$\#\left\{x \in \mathcal{A}^* : \text{length}\,(x) = n,\, K_\phi(x|n) = r\right\} \leq z^r. \tag{16}$$

PROOF Denote

$$A = \{x \in \mathcal{A}^* : \text{length}(x) = n, K_\phi(x|n) = r\}$$

and define a mapping $D : A \to$ (subsets of $\mathcal{A}^*$) by the following formula:

$$D(x) = \{y \in \mathcal{A}^* : \text{length}(y) = r, \phi(y, \text{length}(x)) = x\}.$$

Then, for any $x \in A$, the set $D(x) \neq \emptyset$, and if $x_1 \neq x_2$ then $D(x_1) \cap D(x_2) = \emptyset$ so that $D(x_1) \neq D(x_2)$. Thus, by a simple counting argument,

$$\# A = \# D(A) \leq \# \mathcal{A}^r = z^r. \quad \blacksquare$$

The above proposition immediately yields the following:

Proposition 4.3.2
For any integers $n, m \in \mathbf{N}$, $0 \leq m < n$, there are at most $(z^{n-m} - 1)/(z - 1)$ strings of length n and Kolmogorov complexity smaller than $n - m$, i.e.,

$$\#\left\{x \in \mathcal{A}^* : \text{length}(x) = n, K_\phi(x|n) < n - m\right\} \leq \frac{z^{n-m} - 1}{z - 1}. \tag{17}$$

PROOF Add sidewise the inequalities from Proposition 4.3.1 for $r = 0, 1, 2, \ldots, n - m - 1$. Then, an application of the formula for the sum of the geometric progression gives us the estimate (17). $\blacksquare$

Applying the above upper estimate to the complementary set of strings with complexity at least $n - m$, and recalling that the total number of strings of length n is z^n, we get the following:

Corollary 4.3.1
For any $n, m \in \mathbf{N}$, $0 \leq m < n$, the fraction of all strings of length n which are m-random is strictly positive. More precisely, denote by $c(n, m)$ the number of strings of length n with the Kolmogorov complexity within m of the maximal complexity of such strings, i.e.,

$$c(n, m) := \#\left\{x \in \mathcal{A}^* : \text{length}(x) = n, \ K_\phi(x|n) \geq n - m\right\}. \tag{18}$$

Then

$$\frac{c(n, m)}{z^n} > \left(1 - \frac{z^{-m} - z^{-n}}{z - 1}\right) > \left(1 - \frac{z^{-m}}{z - 1}\right) \geq 0. \tag{19}$$

In particular, the fraction of Kolmogorov random strings is positive as

$$\frac{c(n,0)}{z^n} > \frac{z-2}{z-1} \geq 0. \tag{20}$$

The last estimate shows that there is at least one Kolmogorov random string of each length, that is a string x of length n whose complexity satisfies condition $K_\phi(x|n) \geq n$. It also shows a drastic difference between the properties of binary strings and strings written in longer alphabets: If $z \geq 3$, then, in view of (20), $\lim_{n\to\infty} c(n,0) = \infty$ that is the number of Kolmogorov random strings of length n increases to infinity as $n \to \infty$. As a matter of fact, in such a case, *more than half of all the strings are Kolmogorov random.* However, inequality (20) does not give such assurance for $z = 2$ in which case it only states that $c(n,0) > 0$ (or $c(n,0) > 1$, if one uses the stronger first inequality from Corollary 4.3.1). On the other hand, if $m > 0$, then the number of m-random binary strings also increases to infinity with n. For example, if $z = 2, m = 10$, then more than 99.8% of all strings of length greater then 10 are 10-random.

Remark 4.3.1 *Gödel's Incompleteness Theorem.* The above proofs of existence are *non-effective.* The Kolmogorov complexity function $K(x)$ is not computable. Indeed, if $K(x)$ were computable, then we could define a string of high complexity with a short program – the program would make use of the algorithm to compute $K(x)$.

In other words, although we have proved the existence of Kolmogorov random strings, the statement "x `is Kolmogorov-random`" is not provable within any consistent[1] formal deductive system (i.e., consisting of definitions, axioms, rules of inference) for all but finitely many strings.

Indeed, fix a formal system $\mathcal{F}$ (say, described completely by f bits) in which the statement "x `is Kolmogorov-random`" is expressible. Assume to the contrary that the proof of such statement, for all random x of any length n, is contained within the system $\mathcal{F}$. Then, take a random string of length $n \gg f$ and print the proof that it is random.[2] The whole proof uses only $\log n + f < n$ bits of data—impossible if the system $\mathcal{F}$ is consistent.

Results of this type, although in a different context, are known as Gödel incompleteness results. Here, although initially it might sound pessimistic and look like a reincarnation of the Berry-Richard paradox, it is understandable if one contemplates the notion of randomness at a deeper level.

[1] Recall that a formal system is said to be consistent if no statement is expressible in this system that can be proved to be both true and false.

[2] The notation $a \gg b$ means that a is much greater than b.

In our present context, the existence of the universal Turing machine mentioned in Section 4.2 also can be put on a more formal ground. The next result is fundamental for the theory of algorithmic complexity.

Kolmogorov's Universality Theorem. *There exists a partial recursive function* $\omega : \mathcal{A}^* \times \mathbf{N} \overset{\circ}{\to} \mathcal{A}^*$ *such that for any* $\phi : \mathcal{A}^* \times \mathbf{N} \overset{\circ}{\to} \mathcal{A}^*$ *one can effectively find a* $C (= C(\omega, \phi))$ *such that*

$$K_\omega(x|m) \leq K_\phi(x|m) + C \quad \forall\, x \in \mathcal{A}^*, m \in \mathbf{N}.$$

The importance of the result rests, of course, on the fact that the constant C is independent of the string x and its length m, which means that the complexity measured by different devices is essentially the same (up to a constant C). For very long strings, the differences are negligible. So, for them all the previous considerations in this section can be assumed to be conducted in the context of a fixed universal Kolmogorov complexity $K = K_\omega$. Then, in view of the Kolmogorov's Universality Theorem and Example 4.3.1 (where $\phi(x, m) = x$) there exists a $c \in \mathbf{N}$ such that

$$K(x|\text{length }(x)) \leq \text{length }(x) + c, \qquad \forall x \in \mathcal{A}^*,$$

In particular, if the number z of characters in the alphabet $\mathcal{A}$ is > 2, then there exists a sequence $(m_n) \subset \mathbf{N}$ such that

$$n \leq m_n \leq n + c,$$

and such that

$$\lim_{n \to \infty} \# \left\{ x \in \mathcal{A}^* : \text{length }(x) = n,\, K(x|n) = m_n \right\} = \infty.$$

4.4 Typical sequences: Martin-Löf tests of randomness

In this section we return to the idea of a *typical* or Martin-Löf *random* string first brought up in Section 4.1. In the context of random strings it is often beneficial to think about very long, perhaps even infinite, strings. How do we define "small", "negligible", "exceptional", or "atypical", families of such strings? One natural way to introduce a "measure" measuring the size of families of, say binary, strings written in the alphabet $\mathcal{A} = \{0, 1\}$ is to proceed as follows:

Denote by $\mathcal{A}^\infty$ the set of all infinite binary strings and by $\mathcal{A}^*$, as before, the set of all finite binary sequences. Each finite string $x = (x_1, x_2, \ldots, x_n) \in \mathcal{A}^*$

determines what is called a *n-dimensional elementary cylindrical set* $\Gamma_{\mathbf{x}}$ of infinite strings that begin with the finite string $\mathbf{x}$ and have no restrictions imposed on the digits $x_{n+1}, x_{n+2}, \ldots$, i.e.,

$$\Gamma_{\mathbf{x}} := \{\mathbf{y} \in \mathcal{A}^{\infty} : y_1 = x_1, \ldots y_n = x_n, y_{n+1}, y_{n+2}, \ldots \text{ arbitrary}\}. \quad (1)$$

The name "cylindrical" is an obvious analogy to cylindrical sets in Euclidean spaces for which some coordinates are subject to constraints while others remain unconstrained. Since there are 2^n n-dimensional elementary cylindrical sets in $\mathcal{A}^{\infty}$, and they are all *mutually disjoint*, we will assign to each of them the same measure

$$\Pr(\Gamma_{\mathbf{x}}) = 2^{-\text{ length }(\mathbf{x})}. \quad (2)$$

It corresponds to the symmetric Bernoulli probability distribution on the binary strings of length n. So, for example, there are $2^4 = 16$ 4-dimensional elementary cylinders, starting with the cylinder the string therein being of the form

0000..........

The strings in the second, third, and so on, cylinders are of the form

0001.........

0010.........

0011.........

....

1101.........

1110.........

1111.........

Each of these cylinders is assigned probability = 1/16.

In this context it is tempting to try to define a "negligible" family of infinite strings N as any subset of $\mathcal{A}^{\infty}$ which has the probability (2) zero. This means that N would have to be contained in a finite collection of cylinders (and we already know how to measure those via the formula (2)) of arbitrarily small total probability. More formally, we have:

Definition 4.4.1 Sets of Infinite Strings of Probability Zero.
Let N be a set of infinite binary strings. We shall say that

$$\Pr(N) = 0$$

if, for any $\varepsilon > 0$ *there exist an integer* $k \in \mathbf{N}$ *and finite strings* $\boldsymbol{x}^1, \dots, \boldsymbol{x}^k \in \mathcal{A}^*$ *such that*

(i) $\bigcup_{i=1}^{k} \Gamma_{\boldsymbol{x}^i} \supset N$,

(ii) $\sum_{i=1}^{k} \Pr(\Gamma_{\boldsymbol{x}^i}) < \varepsilon$.

Then the next step would be to call a string "atypical" if it belongs to at least one "negligible" set of probability zero and, finally, to call a string "typical" if it is not "atypical", i.e., the family T of "typical" strings would be defined by the formula

$$T = \mathcal{A}^\infty \setminus \bigcup_{\Pr(N)=0} N.$$

As attractive as this line of reasoning might look, it is not successful since, as it is easy to check, with this definition the set T of "typical" strings would be empty.

The above approach was salvaged by Per Martin-Löf who has shown how the above definition of the family of strings of probability zero can be fixed.

Following his approach, we shall say that a set in infinite strings is *effectively* of probability zero if the cylinder bases in Definition 4.4.1 can be selected effectively. More precisely, we have:

Definition 4.4.2 Sets of Infinite Strings Effectively of Probability Zero. *A set* $N \subset \mathcal{A}^\infty$ *is said to be effectively of probability zero, in short*

$$\Pr(N) \overset{\text{eff}}{=} 0,$$

if, for any $\varepsilon > 0$*, there exist an integer* $k \in \mathbf{N}$ *and computable strings* $\boldsymbol{x}^1, \dots \boldsymbol{x}^k$ *which satisfy conditions (i) and (ii) of Definition 4.4.1.*

Now, the set T of *typical strings*

$$T := \mathcal{A}^\infty \setminus \bigcup_{\Pr(N) \overset{\text{eff}}{=} 0} N$$

is nonempty, and moreover

$$\Pr(T) \overset{\text{eff}}{=} 1.$$

This is the contents of the Martin-Löf's Theorem. For its proof, and a more detailed analysis of the concept of Martin-Löf randomness, we refer to the Bibliographical Notes at the end of this chapter. In this context, a string $\boldsymbol{x} \in \mathcal{A}^\infty$ is called *Martin-Löf random* if $\boldsymbol{x} \in T$.

Example 4.4.1 A Typical State of the Glass of Water.[3]
In large systems, such as statistical mechanical ensembles, the property of being a member of a large majority is often related to having a certain measurable "typical property". Consider a gas G of n molecules of mass 2, with three component velocities v_i^1, v_i^2, v_i^3, $i = 1, 2, \ldots, n$, which describe the kinetic state of the system. Our "universe" consists of systems G for which the total kinetic energy is bounded by a certain constant, say 1, i.e., such that

$$\text{KinEn}\ (G) := \sum_{i=1}^{n} \left[(v_i^1)^2 + (v_i^2)^2 + (v_i^3)^2 \right] \leq 1.$$

In other words, in the phase space $\mathbf{R}^{3n} \ni (v_i^1, v_i^2, v_i^3)$, $i = 1, 2, \ldots, n$, our "universe" is simply a $3n$-dimensional ball of radius 1, with a $3n$-dimensional volume equal to

$$V_n = \frac{\pi^{3n/2}}{\Gamma(1 + 3n/2)}.$$

We assume that the energies are randomly uniformly distributed over this unit ball (see *Mathematica* Experiment 3.10.2). Now, the volume of the part of our "universe" that consists of states with kinetic energy less than $1 - \varepsilon$ is equal to $(1 - \varepsilon)^{3n} V_n$ so that the fraction of our "universe" with energies within an ε of the maximum energy 1 is

$$1 - (1 - \varepsilon)^{3n}.$$

This number is extremely close to 1 for large systems, that is for large values of n.

For example, even with only $n = 3000$ molecules, 99.99% of our "universe" has approximately the same (thus the typical) kinetic energy equal to 1. More precisely,

$$\frac{\text{Volume}\,\{G : .999 \leq \text{ KinEn } (G) \leq 1\}}{\text{Volume}\,\{G : \text{ KinEn } (G) \leq 1\}} \approx .9999.$$

In other words, the probability that the kinetic energy of a molecule is within 0.001 of 1, is about 0.9999. Just imagine how close to 1 that probability would be with a more realistic $n = 10^{23}$ molecules. □

The above simple calculation explains an intuitively obvious thermodynamic fact that although different molecules *can and do* have different and "randomly" distributed velocities, it is extremely unlikely to observe in a glass of water a spontaneous formation of subregions (say bottom and top halves of the glass) with drastically different temperatures (say, 32° F and 212° F.)

[3] This striking example was borrowed from David Ruelle's book.

The notion of a typical or Martin-Löf random string permits us to construct *tests of randomness*. So, consider a finite binary string $x = x_1 x_2 \dots x_n$ written in the alphabet $\mathcal{A} = \{0, 1\}$, and remember that the quantity

$$x_1 + \dots + x_n, \tag{3}$$

simply counts the total number of 1s in the string x. The sum (3) is familiar from the development of the binomial distribution in Section 3.1. Remembering the construction of measure (2) on cylindrical sets of strings, and the fact that the equipartion of 0s and 1s was a fundamental desirable feature of random strings (see Section 1.1), we can propose the following algorithm of testing the hypothesis[4] that "x is random".

Example 4.4.2 A Basic Test of Randomness.
Select the *significance level* $\varepsilon = 2^{-m}, m = 1, 2, \dots, n$, and find the smallest constant $\delta = \delta(m, n)$ such that the number of strings y of length n for which the inequality

$$\left| \frac{y_1 + \dots + y_n}{n} - \frac{1}{2} \right| \geq \delta$$

holds is $< 2^{n-m}$ or, in other words, the number δ has to satisfy the inequality

$$\Pr \left\{ y \in \mathcal{A}^n : \left| \frac{y_1 + \dots + y_n}{n} - \frac{1}{2} \right| \geq \delta \right\} < \varepsilon.$$

Then:

(i) Reject the hypothesis of randomness of x (at the significance level ε), if

$$\left| \frac{x_1 + \dots + x_n}{n} - \frac{1}{2} \right| \geq \delta;$$

(ii) Do not reject it, in the opposite case.

For example, for strings of length $n = 10$ and significance level $\varepsilon = 2^{-5}$, that is $m = 5$, it is easy to see that $\delta = \delta(5, 10) = 0.4$, since exactly 22 strings y out of the total of $2^{10} = 1024$ satisfy the inequality

$$\left| \frac{y_1 + \dots + y_{10}}{10} - \frac{1}{2} \right| \geq \frac{4}{10}.$$

[4] The general idea of statistical hypothesis testing will be studied at length in Part 3 of this book.

Indeed, these are the strings containing either zero, one, nine, or ten 1s. Hence,

$$\Pr\left\{y \in \mathcal{A}^{10} : \left|\frac{y_1 + \ldots + y_{10}}{10} - \frac{1}{2}\right| \geq \frac{4}{10}\right\} = \frac{22}{1024} \approx 0.215 < 2^{-5}.$$

Thus our test would not reject, at significance level $2^{-5} \approx 3\%$ (meaning that the small minority of at most 3% of all strings of length 10 would fail this tests), as random all the strings for which the relative frequency of 1s differs from 1/2 by less than 0.4, i.e., all the strings containing at least two, and at most eight, 1s. □

Of course the above test is just a coarse example but its spirit is correct, and Martin-Löf used this simple idea to propose a general concept of a test (or rather a series of tests with improving significance levels) of randomness of strings written in an alphabet $\mathcal{A} = \{a_1, \ldots, a_z\}$.

Definition 4.4.3 Martin-Löf Test of Randomness.
A non-empty effectively enumerable set $V \subset \mathcal{A}^* \times (\mathrm{N} \setminus 0)$ is called a Martin-Löf test if, for each $n \in \mathrm{N}$ and $m = 1, 2, \ldots,$

$$V_{m+1} \subset V_m := \{x \in \mathcal{A}^* : (x, m) \in V\}, \tag{4}$$

and

$$\#\{x \in \mathcal{A}^* : \text{length}\ (x) = n, x \in V_m\} < \frac{z^{n-m}}{z-1}. \tag{5}$$

It is an obvious observation that if V is a Martin-Löf test and $(x, m) \in V$, then necessarily

$$\text{length}\ (x) > m \geq 1.$$

Now, let us consider a number of illuminating examples. The first one will connect the notion of the Martin-Löf test of randomness with the notions of Kolmogorov complexity and randomness considered in Section 4.3.

Example 4.4.3 Testing Randomness via Kolmogorov Complexity.
Consider a computable function $\phi : \mathcal{A}^* \times \mathrm{N} \to \mathcal{A}^*$. Then

$$V^\phi = \Big\{(x, m) : x \in \mathcal{A}^*, m = 1, 2, \ldots, \text{ and}$$

$$K_\phi(x \,|\, \text{length}\ (x)) < \text{length}\ (x) - m\Big\}$$

is a Martin-Löf test. This follows directly from Proposition 4.3.2. □

Example 4.4.4

Select an $q \in \mathbf{N}$ and a finite string $x \in \mathcal{A}^*$ such that length $(x) > q \geq 1$. Then, the set

$$H(x,q) = \{(x,1),(x,2),\dots(x,q)\}$$

defines a Martin-Löf test. In fact, $V_m = \{(x,m)\}$ for $m \leq q$, and $= \emptyset$ for $m > q$. It just declares one particular string to be non-random. We need to check only the second condition (5). This is done as follows: Because

$$\#\{y \in \mathcal{A}^* : \text{ length } (y) = n,\ (y,m) \in H(x,q)\} \tag{6}$$

is equal to 1 if $n =$ length (x), and $1 \leq q \leq q$, and 0, otherwise, we have that the number in (6) is less than $z^{n-m}/(z-1)$, since length $(x) > q > m$, i.e., $z^{n-m}/(z-1) > 1$. □

Example 4.4.5

Take x, q as in Example 4.4.4. Then,

$$\tilde{H}(x,q) = \Big\{(y,n) : y \in \mathcal{A}^*, n \in \mathbf{N}, 1 \leq n \leq q, y \supset x\Big\},$$

is a Martin-Löf test. The notation $y \supset x$ means that y is a concatenation of x and some other string $w \in \mathcal{A}^*$, i.e., $y = xw$. Again, we will verify only the second part of Definition 4.4.3. Take $n, m \in \mathbf{N}$ with $m \geq 1$. Then,

$$\#\{y \in \mathcal{A}^* : \text{ length } (y) = n, (y,m) \in \tilde{H}(x,q)\}$$

$$= \begin{cases} \#\{y \in \mathcal{A}^* : \text{length}(y) = n, y \supset x\}, & \text{if } 1 \leq m \leq q,\ n \geq \text{length}(x), \\ 0, & \text{otherwise,} \end{cases}$$

$$= \begin{cases} z^{n-\text{length}(x)}, & \text{if } 1 \leq m \leq q,\ n \geq \text{ length } (x), \\ 0, & \text{otherwise,} \end{cases}$$

$$< z^{n-m}/(z-1),$$

because length $(x) > q > m$. □

The question of how to select proper tests of randomness is not easy. Testing equipartition is one idea suggested by our discussion in previous chapters. Other tests are suggested by the theoretical models of randomness such as those arising in the context of the notion of statistical independence and the Kolmogorov's axiomatic probability theory which will be developed in Chapter 5. They could involve, for instance, testing the Gaussianness of deviations from the mean (the

Central Limit Theorem) or other, more subtle phenomena, such as the law of the iterated logarithm.

The set of all Martin-Löf tests is effectively enumerable. However, a finite union of Martin-Löf tests of type $H(x, q)$ (see Example 4.4.4) need not be a Martin-Löf test. For example, take $z = 2, a_1 = 0, a_2 = 1, x_1 = 00, x_2 = 01, x_3 = 10$. Then

$$H = H(x_1, 1) \cup H(x_2, 1) \cup H(x_3, 1)$$

is not a Martin-Löf test as the second condition in Definition 4.4.3 is violated. On the other hand, there exists an analog of the Kolmogorov's Universality Theorem of Section 4.3.

Martin-Löf's Universality Theorem. *There exists a universal Martin-Löf test U such that for every Martin-Löf test V one can effectively find a positive integer $c = c(U, V)$ such that, for all $m \geq 1$,*

$$V_{m+c} \subset U_m.$$

A universal Martin-Löf test U has the following property: if a string is random with respect to U, then it is random with respect to any other test (with perhaps a change of the significance levels).

Definition 4.4.4
The critical level induced by an Martin-Löf test V is given by a function

$$m_V : \mathcal{A}^* \to \mathbf{N}$$

defined by the formula

$$m_V(x) = \begin{cases} \max\{m \geq 1 : x \in V_m\}, & \textit{if } x \in V_1, \\ 0, & \textit{otherwise.} \end{cases}$$

In other words, the critical level is the smallest level of significance (i.e., z^{-m}) at which the randomness hypothesis is rejected. In terms of the critical level, the definition of the universal test U can be rephrased as follows: U is a universal Martin-Löf test if for every Martin-Löf test V there exists a constant $c = c(U, V)$ such that for all $x \in \mathcal{A}^*$

$$m_V(x) \leq m_U(x) + c.$$

Also, there exists a relation between the fixed Kolmogorov's universal complexity discussed in Section 4.3 and the critical level of a universal Martin-Löf test.

Namely, one can find a positive integer q such that for all $x \in \mathcal{A}^*$

$$\left|\text{length }(x) - K\big(x \mid \text{length }(x)\big) - m_U(x)\right| \leq q.$$

The proof of this result is quite difficult and will not be given here (see Bibliographical Notes). However, given the above result, it is not hard to see that a Martin-Löf test associated with the universal Kolmogorov complexity is universal.

Corollary 4.4.1
Fix a universal Kolmogorov algorithm $\omega : \mathcal{A}^* \times \mathrm{N} \to \mathcal{A}^*$. Then every Kolmogorov-random string withstands the universal Martin-Löf test V^ω.

Less formally, Kolmogorov-random strings possess almost all conceivable statistical properties of randomness. Also, every m-random string withstands the universal Martin-Löf test V^ω. Indeed, for some

$$m < \text{length }(x), \ K\big(x \mid \text{length }(x)\big) \geq \text{length }(x) - m$$

so that $(x, m) \notin V^\omega$.

On the other hand, asymptotically random strings are *not constructable*, i.e., there is no effective algorithm for generating (m)-random strings. More exactly, the function

$$\tilde{K} : \mathcal{A}^* \to \mathrm{N}, \quad \tilde{K}(x) = K(x| \text{length }(x))$$

is not recursive. More generally, if $f : \mathrm{N} \overset{\circ}{\to} \mathcal{A}^*$ satisfies, for n in the domain of f, the condition

$$K_\omega(f(n)|n) \geq \alpha(n),$$

and $\alpha : \mathrm{N} \mapsto \mathrm{N}$ is such that $\lim_{n\to\infty} \alpha(n) = \infty$ (e.g., $\alpha(n) = n$, $\alpha(n) = \lfloor \log_2 n \rfloor$, α need not be recursive), then f is not a computable function. In particular, *the critical level induced by an Martin-Löf test V is not computable, and the universal Kolmogorov algorithm ω is not computable.*

4.5 Stability of subsequences: von Mises randomness

In this section we briefly return to the notion of *von Mises randomness* to complete the discussion of Section 4.3. The idea is that the frequencies of letters in a string should be stable under the operation of substring selection. This is clearly related to the equipartition ideas introduced in Section 1.1.

To illustrate what we have in mind, let us consider a string of 0s and 1s of length n containing exactly m 1s. For any "method" of splitting a random string into two substrings of lengths n_1 and n_2, with m_1 and m_2 1s, respectively (so that $n = n_1 + n_2$, and $m = m_1 + m_2$), one would like to see the frequency of 1s in the original string preserved in the substrings, i.e., one would like the quantity

$$\left| \frac{m_1}{n_1} - \frac{m_2}{n_2} \right|$$

to be small. More precisely, for random strings of increasing length, one would like to see, for each $\varepsilon > 0$,

$$\Pr\left\{ \boldsymbol{x} \in \mathcal{A}^n : \left| \frac{m_1}{n_1} - \frac{n_2}{n_2} \right| < \varepsilon \right\} \to 1, \quad \text{as} \quad n_1, n_2 \to \infty, \tag{1}$$

where Pr is the uniform measure on the cylindrical sets of strings introduced in Section 4.4.

Clearly, not every "method" of selecting substrings is admissible. If we say: select the subsequence which contains only 1s then, for sure, the stability of frequencies will be ruined. So, one of the principal tasks here is to describe *admissible algorithms* of selecting substrings.

For simplicity, let us consider an infinite binary string $\boldsymbol{x} = (x_1, x_2, \ldots,)$ written in the alphabet $\mathcal{A} = \{0, 1\}$. The following definition of an *admissible algorithm* φ was devised in the late thirties by Alonzo Church.

Definition 4.5.1

The Church-admissible substring selection algorithm is a computable function

$$\varphi : \mathcal{A}^* \to \{Yes, No\},$$

such that the decision Yes *to include the x_k in a substring depends only on the values of $x_1, \ldots, x_{k-1}$, and such that the selected digits appear in the substring in the same order as in the original one. To be more precise, the input of φ is $= x_1, \ldots, x_{k-1}$, and the output $=$* Yes *for inclusion of x_k in the substring, or* No *for exclusion of that digit.*

The subsequent development of the algorithmic complexity theory demonstrated, however, that such a concept of admissible selection algorithm is somewhat too strict when placed in the context of Kolmogorov and Martin-Löf randomness, and the resulting class of Church-random sequences—too large. All the Kolmogorov-random strings turn out to be Church-random but there are Church-random strings that fail some basic Martin-Löf tests of randomness (such as the tests based on the so-called Law of Iterated Logarithm, see Chapter 5).

In response to this crisis, Andrei Nikolaevich Kolmogorov introduced the following broader selection rule:

Definition 4.5.2
The Kolmogorov-admissible substring selection algorithm is a computable function for which the inputs are, for each $k \in \mathbf{N}$,

$$n_1, n_2, \ldots, n_k, x_{n_1}, x_{n_2}, \ldots, x_{n_k},$$

and the output is $n_{k+1} (\neq n_1, \ldots, n_k)$ and Yes *or* No, *that is, a decision on whether or not to include $x_{n_{k+1}}$ in the substring. In this definition, the order of the terms in the selected substring need not be the same as in the original string.*

This definition, and the older von Mises' ideas of "collectives", permit introduction of a usable concept of *von Mises randomness.*

Definition 4.5.3
An infinite binary string is said to be von Mises-random if the frequency of 1s of any of its Kolmogorov-admissible substrings is equal to 1/2.

It is known that any Kolmogorov-random sequence is also von Mises-random. However, the problem of whether these two classes of random sequences are the same remains open.

4.6 Computable framework of randomness: degrees of irregularity

As we have seen in this chapter, an analysis of the concept of randomness of a fixed binary string very quickly leads to subtle, or even philosophical considerations. Although all of this analysis impacts the computer generation of pseudorandom numbers used in simulation, Monte Carlo methods and secure ciphers (see Sections 1.13, 2.8, etc.), the latter has its own demands and requirements such as ease of coding, set-up and running time, memory requirements, and portability.

The concept of Kolmogorov randomness as maximal computational complexity, although intellectually so appealing, and obviously of great relevance for cryptographic purposes, is, by definition, the worst one for computer implementation and suffers from fundamental incomputability problems. So, in practice, as we mentioned in Chapter 1, one uses relatively simple recipes, such as the linear congruential methods or quadratic residue (QR) methods, containing few parameters and one concentrates on optimal selection of these parameters from the viewpoint of a battery of tests one can run on them and that are theoretically suggested by the notion of Martin-Löf's hierarchy of tests. As a matter of fact, different purposes for which the pseudorandom numbers are generated may require different batteries of tests.

Practically, one always deals with periodic random strings, and, of course, one wants the periods to be as large as possible. One also wants to avoid any obvious intrinsic structure (such as lattice structure, see Fig. 1.11.1) in a pseudorandom string. Then one runs some of the statistical tests on it such as the uniformity test checking the equiparition properties (based on the Kolmogorov-Smirnov law of Section 3.9), frequency test (based on the Law of Iterated Logarithm of Section 5.7), gap test, run test, permutation test, or the test for serial correlation that probes the interdependencies within the pseudorandom sequence. What we mean by some of them will become clearer after we develop appropriate probability theory and statistics tools in subsequent chapters. Here are some simple examples complementing the Kolmogorov-Smirnov goodness-of-fit test from Section 3.9.

Example 4.6.1 Chi-Square Test.
This test applies to the general discrete random quantity taking values $v_1, \ldots, v_N$ with distribution F. The null hypothesis H_0 is " the sequence $x_1, \ldots, x_n$, is a sample of independent random quantities with common distribution $F = (p_1, \ldots, p_N)$". Let f_i be sample frequencies of values v_i, $i = 1, 2, \ldots, N$ (binning can be employed for an absolutely continuous distribution). Then, asymptotically, for large n, the random quantity

$$\chi^2_{N-1} := \sum_{i=1}^{N} \frac{(f_i - np_i)^2}{np_i} \tag{1}$$

has the chi-square distribution with $N - 1$ degrees of freedom. The approximation becomes reasonable if the minimum frequency of the possible values is at least five. The hypothesis H_0 is rejected, at a prescribed significance level, for values of χ^2_{n-1} larger than the corresponding critical value. □

This general test can be applied to test the particular uniformity hypothesis, of interest in this section, in several ways. One of the possibilities is shown below. For more details on this test see Chapter 8.

Example 4.6.2 Birthday Spacing Test.
The null hypothesis H_0 is "the sequence $u_1, \ldots, u_k$, is a sample of independent random quantities uniformly distributed on the unit interval $[0, 1)$". Select a positive integer d. Sort the sequence of integers $j_1 = \lfloor du_1 \rfloor, \ldots, j_k = \lfloor du_k \rfloor$ in the

nondecreasing order to obtain the order statistics

$$0 \leq j_{(1)} \leq j_{(2)} \leq \dots \leq j_{(k)} \leq d-1.$$

Denote by X the random quantity equal to k minus the number of distinct spacings among

$$j_{(2)} - j_{(1)}, \quad j_{(3)} - j_{(2)}, \quad \dots, j_{(k-2)} - j_{(k-1)}, \quad j_{(1)} + d - j_{(k)}.$$

It is known that under H_0, asymptotically, for large d, the random quantity X has the Poisson distribution

$$\Pr\{X = i\} = e^{k^3/4d} \frac{(k^3/4d)^i}{i!}, \qquad i = 1, 2, \dots. \tag{2}$$

Now, independently repeating this procedure n times, we can apply the chi-square test of Example 4.6.1. □

Example 4.6.3 Autocorrelation Test.
The null hypothesis H_0 is again "the sequence $u_1, \dots, u_k$, is a sample of independent random quantities uniformly distributed on the unit interval [0, 1)". Define the *autocorrelation*, with delay d, by the formula

$$\mathrm{AC}(d) := \frac{1}{k-d} \sum_{i=1}^{k-d} (u_i - 1/2)(u_{i+d} - 1/2). \tag{3}$$

Under the hypothesis H_0, the random quantity AC(d) has mean value 0 and variance $1/(144(k-d))$. One can show that the random quantity $T = 12\sqrt{k-d}\,\mathrm{AC}(d)$, has asymptotically, for k much larger than d, the $N(0, 1)$ distribution. This fact now can be used in a standard way to reject hypothesis H_0 at a given significance level, selecting the critical rejection interval from the tables of the standard normal distribution, like in Section 3.7. □

Degrees of Irregularity. In this section, we would also like to describe a recent effort by Steve Pincus, Burton H. Singer, and Rudolf E. Kalman (see references in the Bibliographical Notes section), to produce a *computable framework for randomness* based on the concept of *approximate entropy.*

Consider a binary string $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$ of length n and its $n - m + 1$ blocks

$$\boldsymbol{x}(i) = (x_i, x_{i+1}, \dots, x_{i+m-1}), \qquad i = 1, \dots, n-m+1, \tag{4}$$

of length $m \leq n$ each. The *distance* $d(\boldsymbol{x}(i), \boldsymbol{x}(j))$ between the blocks $\boldsymbol{x}(i)$ and $\boldsymbol{x}(j)$ is defined as

$$d\Big(\boldsymbol{x}(i), \boldsymbol{x}(j)\Big) := \max_{k=1,2,\ldots,m} |x_{i+k-1} - x_{j+k-1}| \tag{5}$$

It is either 0 if the two blocks are the same, and 1 if they are different. For a fixed real number $0 < \epsilon < 1$, the quantity

$$C_i^m := \frac{\#\big\{j : d(\boldsymbol{x}(i), \boldsymbol{x}(j)) \leq \epsilon, \ 1 \leq j \leq n - m + 1\big\}}{n - m + 1} \tag{6}$$

measures the fraction of blocks of length m which are exactly the same. The role of ϵ is not essential here, but it is introduced so that nonbinary strings could also be considered within this framework.

Definition 4.6.1 Approximate Entropy.

Consider a fixed binary string $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ of length n, and its blocks of length m. The approximate entropy of $\boldsymbol{x}$ measures the logarithmic frequency with which blocks of length m that remain identical for blocks augmented by one position. More precisely,

$$\mathrm{AE}_n^m(\boldsymbol{x}) := \Phi^m - \Phi^{m+1}, \qquad m \geq 1, \tag{7}$$

where

$$\Phi^m := \frac{1}{n - m + 1} \sum_{i=1}^{n-m+1} \ln C_i^m, \tag{8}$$

with $\mathrm{AE}_n^0(\boldsymbol{x}) := -\Phi^1$.

The above definition is, obviously, a cousin of the concepts of the Grassberger and Procaccia's correlation dimension and of Shannon's entropy, introduced in Sections 2.8 and 2.9. Large values of $\mathrm{AE}(\boldsymbol{x})$ imply strong fluctuations and irregularities in the string $\boldsymbol{x}$.

Definition 4.6.2 Irregular Strings.

A binary string $\boldsymbol{x}_*^{(n)}$ of length n is said to be $\{m, n\}$-irregular if

$$\mathrm{AE}_n^m(\boldsymbol{x}_*^{(n)}) = \max_{\boldsymbol{x}} \mathrm{AE}_n^m(\boldsymbol{x}), \tag{9}$$

where the maximum is taken over all 2^n binary strings of length n. It is called n-irregular if it is $\{m, n\}$-irregular for $m = 0, 1, 2, \ldots m_{crit}(n)$, where

$$m_{crit}(n) := \max\{m : 2^{2^m} \leq n\}. \tag{10}$$

The selection of m_{crit} is motivated by the fact that for a "typical" Bernoulli random string $\boldsymbol{x}_n$ the limit

$$\lim_{n\to\infty} \mathrm{AE}_n^{m_{crit}(n)}(\boldsymbol{x}_n) = h$$

is equal to the entropy of the Bernoulli ensemble.

Example 4.6.4 Irregular Strings of Length $n = 5$.
In this case, $m_{crit}(n) = 1$, and the string $\boldsymbol{x} = (x_1, x_2, x_3, x_4, x_5)$ is 5-irregular if both

$$\mathrm{AE}_5^0(\boldsymbol{x}) = \max_{\boldsymbol{x}} \mathrm{AE}_5^0(\boldsymbol{x}) \approx 0.673, \tag{11}$$

and

$$\mathrm{AE}_5^1(\boldsymbol{x}) = \max_{\boldsymbol{x}} \mathrm{AE}_5^1(\boldsymbol{x}) \approx 0.7133. \tag{12}$$

Out of $2^5 = 32$ binary strings of length 5, the $\{0, 5\}$-irregular strings satisfying condition (11) are those with three 0s and two 1s, or two 0s and three 1s. There are 20 of them. Of these, only four,

(1,1,0,0,1),
(1,0,0,1,1),
(0,0,1,1,0),
(0,1,1,0,0),

satisfy the condition (12) as well, that is are also $\{1, 5\}$-irregular. Note in each of the 5-irregular strings, each of the four blocks of length two, (0,0), (0,1), (1,0), and (1,1), occurs once, the property not enjoyed by 16 strings that are (0,5)-irregular, but not (1,5)-irregular. □

Remark 4.6.1 As the length n of the strings increases, the fraction of those that are n-irregular decreases. It is easy to see for n-irregular strings of even length $n = 2k$, since they have to have exactly k 0s and 1s, and the fraction of those among all 2^n strings of length n is, in view of Stirling's formula,

$$\frac{\binom{2k}{k}}{2^{2k}} \approx \frac{(2k/e)^{2k}\sqrt{4\pi k}}{(k/e)^{2k} 2\pi k} \frac{1}{2^{2k}} = \frac{1}{\sqrt{\pi k}}. \tag{13}$$

For infinite binary strings $x = (x_1, x_2, \ldots)$, for which the limit

$$\mathrm{AE}^m(x) := \lim_{n \to \infty} \mathrm{AE}_n^m(x) \tag{14}$$

exists, we can introduce the following definition of *computational randomness*:

Definition 4.6.3 Infinite Computationally Random Strings.
An infinite binary string x is said to be computationally random if $AE^m(x) = \ln 2 \approx 0.693$, for all $m \geq 0$.

Clearly, $\ln 2$ is selected because it gives the maximal entropy rate for Bernoulli random sequences, with the maximum information per digit carried. One can show that if the approximate entropy is less than $\ln 2$, then the finite prior history block biases the subsequent observations, resulting in a degree of predictability for the string.

Example 4.6.5 Deficits from Maximal Irregularity of the Champernowne number, π, e, $\sqrt{2}$, and $\sqrt{3}$.
We can study the proximity to maximal irregularity of a finite string $x = (x_1, \ldots, x_n)$ of length n by considering the quantities

$$\mathrm{DEF}^m(x) := \max_{y} \mathrm{AE}_n^m(y) - \mathrm{AE}_n^m(x), \tag{15}$$

which we will call the *deficit from maximal irregularity.* □

For the initial string of length 20 of the base 2 Champernowne number

$$ç = (0,\ 1,\ 1,\ 0,\ 1,\ 1,\ 0,\ 0,\ 1,\ 1,\ 0,\ 1,\ 1,\ 1,\ 1,\ 0,\ 0,\ 0,\ 1,\ 0,\ )$$

checking the blocks of length 2 (i.e., for $m = 1$), gives the deficit from maximal regularity

$$\mathrm{DEF}^m(ç) := \max_{y} \mathrm{AE}_{20}^1(y) - \mathrm{AE}_{20}^1(ç) \approx 0.693 - 0.677 = 0.016, \tag{16}$$

whereas, for the periodic

$$p = (0,\ 1,\ 0,\ 1,\ 0,\ 1,\ 0,\ 1,\ 0,\ 1,\ 0,\ 1,\ 0,\ 1,\ 0,\ 1,\ 0,\ 1,\ 0,\ 1,\ )$$

we have the deficit

$$\mathrm{DEF}^{m}(p) := \max_{y} \mathrm{AE}^{1}_{20}(y) - \mathrm{AE}^{1}_{20}(p) \approx 0.693 - 0.000 = 0.693. \qquad (17)$$

Recently, Pincus and Kalman (see Bibliographical Notes) computed the deficits from maximal irregularity for initial strings of length n of the binary expansions of the numbers π, e, $\sqrt{2}$, and $\sqrt{3}$, checking the approximate entropy for 1-blocks and 3-blocks, that is computing DEF^0_n and DEF^2_n. The computations were carried out for $n = 1, 2, \ldots, 300000$ and the results are shown in Fig. 4.6.1.

The difference between the expansions of π and e are quite dramatic, with $\sqrt{2}$ somewhere inbetween. Clearly, for the binary expansion of π the deficit from maximal irregularity is much smaller than for the binary expansion of e. Remarkably, this difference almost totally disappears if instead of binary expansions one studies the decimal expansions. The results are shown in Fig. 4.6.2.

Linear Complexity Profiles. As we have already noticed, practical use of the Kolmogorov complexity is difficult because one cannot easily calculate the minimum size of a universal Turing machine program that produces a given string. In the area of cryptography, the following substitute concept of *linear complexity* is popular. For details, see Shu Tezuka's monograph quoted in the Bibliographical Notes section.

Definition 4.6.4 Linear Complexity.
The linear complexity $\lambda(x|n)$ of a binary string $x = (x_1, \ldots, x_n)$ of length n is the minimum degree $r < n$ of a polynomial

$$p(z) = z^r + a_{r-1}z^{r-1} + \ldots + a_1 z + a_0, \qquad (18)$$

with binary coefficients (with mod 2 multiplication), such that

$$x_{i+r} = a_{r-1}x_{i+r-1} + \ldots + a_1 x_{i-1} + a_0 x_i \qquad (\text{mod } 2), \qquad (19)$$

for $i = 1, \ldots, n - r$.

Clearly, for a binary string of period T, the linear complexity of its initial piece of length n is at most n if $n < T$, and for larger n it is at most T. Also, the algorithm computing the linear complexity of any binary string of length n, in about $n^2 \ln n$ binary steps, is known.

It is more difficult to show that the mean value of the linear complexity of the random symmetric Bernoulli sequence $X = (X_1, \ldots, X_n)$ of length n is equal to $n/2 + c_n$, where the constants c_n are known to be in the interval $[0, 5/18]$. The variance of this random quantity is about 86/81. The linear complexity is a serious

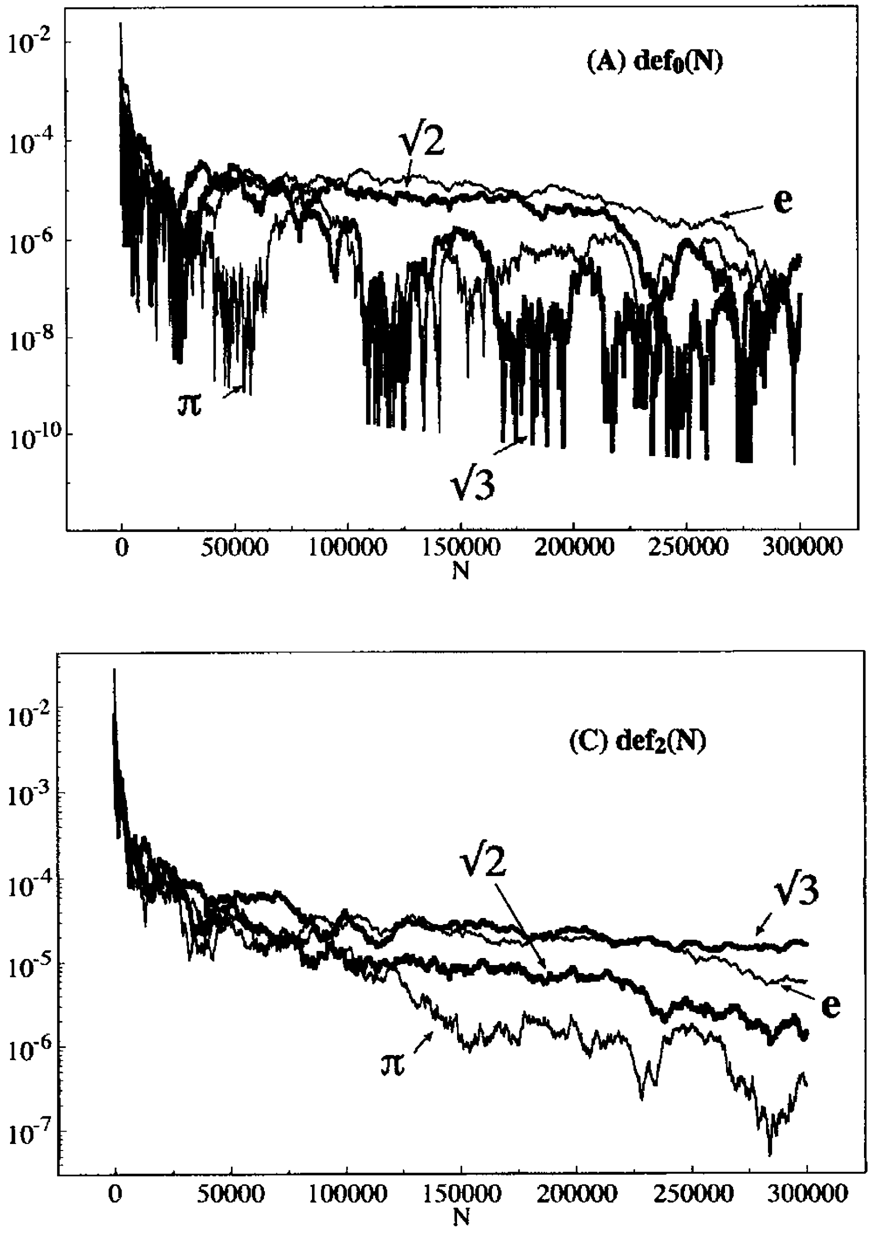


FIGURE 4.6.1

Deficit from maximal irregularity for initial strings of length n of the binary expansions of the numbers π, e, $\sqrt{2}$, and $\sqrt{3}$. Top: $\mathrm{DEF}^0(x)$. *Bottom:* $\mathrm{DEF}^2(x)$. *From Pincus and Kalman (1997).*

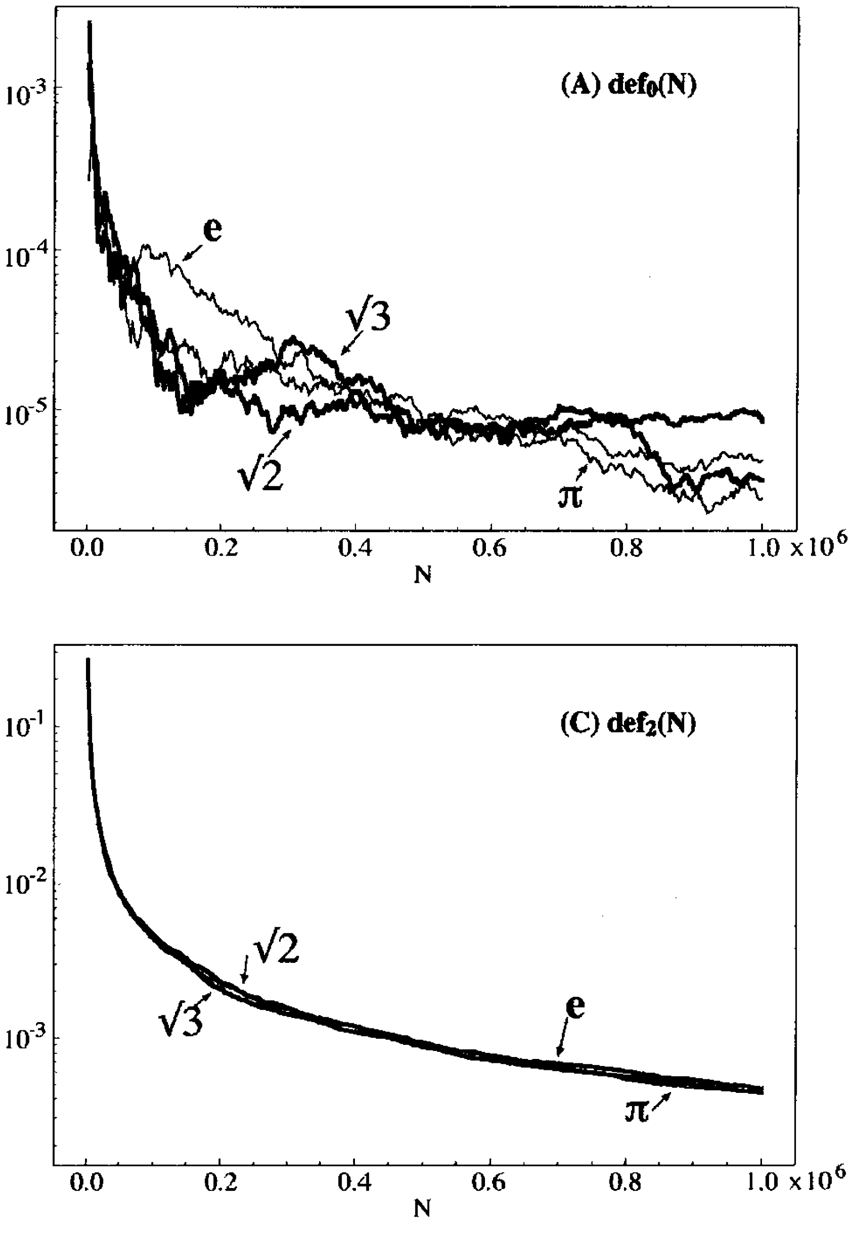


FIGURE 4.6.2

Deficit from maximal irregularity for initial strings of length n of the decimal expansions of the numbers π, e, $\sqrt{2}$, and $\sqrt{3}$. Top: $\mathrm{DEF}^0(x)$. *Bottom:* $\mathrm{DEF}^2(x)$. *From Pincus and Kalman (1997).*

alternative to the Kolmogorov complexity $K(\boldsymbol{x}|n)$ of Section 4.3 because of the following asymptotic estimate: for any $\epsilon \in (0, 1)$,

$$\lim_{n\to\infty} \Pr\left\{(1-\epsilon)\lambda(\boldsymbol{X}|n) \le \frac{K(\boldsymbol{X}|n)}{2} \le (1+\epsilon)\lambda(\boldsymbol{X}|n)\right\} = 1. \qquad (20)$$

In other words, as the Bernoulli string's length increases to infinity, the probability that the random quantities $\lambda(\boldsymbol{X}|n)$ and $K(\boldsymbol{X}|n)$ remain arbitrarily close to each other approaches 1.

One then is tempted to say that an infinite binary string $\boldsymbol{x} = (x_1, x_2, \ldots)$ has a *perfect linear complexity profile* if

$$\lambda(\boldsymbol{x}|n) = \frac{n+1}{2}, \qquad \text{for} \qquad n = 1, 2, \ldots \qquad (21)$$

However, it turns out that the strings with perfect linear complexity profiles are just those that satisfy the string of recurrence relations

$$x_{2i+1} = x_{2i} + x_i \qquad (\text{mod } 2), \qquad \text{for } i = 1, 2, \ldots, \qquad (22)$$

with $x_1 = 1$. Thus, for the first $2n$ bits of such strings there are only 2^n rather than 2^{2n} choices, not a good cryptographic property. Linear complexity profiles of well-known pseudorandom number generators, such as those based on Fibbonacci sequences, have been studied.

4.7 Experiments, exercises, and projects

1. The length of a finite binary string $\boldsymbol{x}$, interpreted here as a natural number in a binary representation, is defined as the number of bits it contains. Show that the length of $\boldsymbol{x}$ is equal to $\lfloor \log_2(\boldsymbol{x}+1) \rfloor$.

2. (a) Find a probability function $f(x)$ defined on positive integers $x \in \mathbf{N}$ with a binary representation of length $l(x) \le n$ such that for each $k \le n$ the conditional probability $\Pr\{X = x | l(x) = k\} = 2^{-k}$, that is, it is uniform. Show several examples of such $f(x)$.

(b) Show that the following two probability distribution functions

$$f(x) = 2^{-2l(x)+1}, \qquad f(x) = \frac{6}{\pi^2 l^2(x)} 2^{-l(x)},$$

defined on the set of all positive integers $x \in \mathbf{N}$ give uniform conditional probabilities for integers with a given length of the binary representation, i.e., for each $k \in \mathbf{N}$ the conditional probability $\Pr\{X = x | l(x) = k\} = 2^{-k}$. Give some other examples of such p.d.f.s.

3. Construct a Turing machine, that is matrices (4.2.1-3), which translates any binary word written in the alphabet $\{0, 1\}$ into the word in which the roles of 0 and 1 are interchanged.

4. Construct a Turing machine that would replace any standard English text written in the alphabet of 26 letters plus "space" by the text in which all the vowels are dropped. Use *Mathematica* for this project.

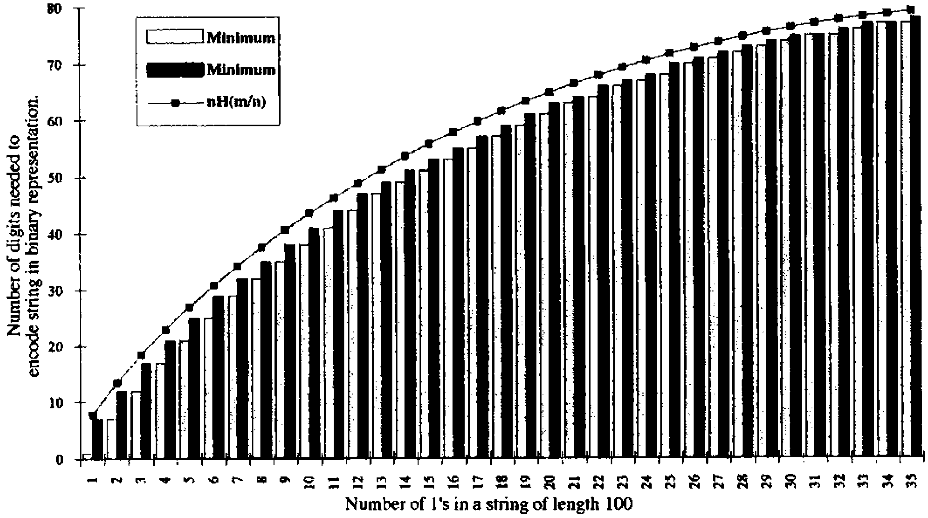
5. *Mathematica Project.* Using the Kolmogorov-Smirnov test, and the chi-square test, check the equipartition property for singles, pairs, etc., at different significance levels, and for substrings, for pseudorandom generators of Chapter 1, the generator provided by *Mathematica*, and for pieces of George Marsaglia pseudorandom numbers provided on the `UVW Web Site`.

6. *Mathematica Experiment.* This experiment is designed to illustrate the conclusions reached in Example 4.3.4 which dealt with "optimal" coding of sparse binary integers. Use the `UVW'ZeroOne'` package to make your job easier.

   (i) Produce a graph of the entropy function $H(p)$, $0 < p < 1$.

   (ii) Write a *Mathematica* code computing the number k of a string of length n in the special ordering of Example 4.3.4. Then, produce a code computing the number K of binary digits needed to encode k.

   (iii) Show, by experimenting with a large number, say 1000, of random strings of length $n = 30, 50$, and 100, that if the string is sparse, i.e., the number m of 1s in the string is small (say, less than 10), then K is approximately $n \cdot H(m/n)$. Present your results graphically to obtain something like the graphics shown below.



7. List all the 6-irregular and 7-irregular binary strings following the analysis of Example 4.6.1. Use *Mathematica* to help, if necessary. Find common characteristics of these strings.

8. *Mathematica Project.* Design a *Mathematica* experiment reproducing results of Example 4.6.5. Find the deficit from maximal irregularity DEF_n^1 for 2-blocks of binary and decimal expansions π, e, $\sqrt{2}$, $\sqrt{3}$, for several values of n, say $n = 100, 1000, 10000, 10000$. Test the hypothesis of uniformity for these expansions using the autocorrelation test and the birthday spacings tests. Produce your own versions of Fig. 4.6.1 and 4.6.2.

9. *Mathematica Project.* Find the deficit from maximal irregularity for 1-, 2-, and 3-blocks of binary and decimal expansions for the Euler constant γ, and $\sqrt{7}$, for several values of n, say $n = 100, 1000, 10000, 10000$. Test the hypothesis of uniformity for these expansions using the autocorrelation test and the birthday spacings tests. Recall, that the Euler constant can be defined as

$$\gamma = \int_0^1 \frac{1 - \cos s}{s}\,ds - \int_1^\infty \frac{\cos s}{s}\,ds \approx 0.57721566490.$$

For information on applications of this constant see, e.g., A.I. Saichev and W.A. Woyczynski, *Distributions in the Physical and Engineering Sciences. Volume 1. Distributional and Fractional Calculus, Integral Transforms and Wavelets,* Birkhäuser-Boston, 1997.

4.8 Bibliographical notes

Among several general sources on algorithmic complexity we quote two

[1] C. Calude, *Theories of Computational Complexity*, North Holland, Amsterdam, 1988

[2] M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*, Springer-Verlag, New York, 1993

which both contain exhaustive bibliographies and historical commentaries. The first one is more mathematical. The second is more in the spirit of computer science. It also contains a very interesting chapter on physics and computation. Another book, with a different emphasis,

[3] G. J. Chaitin, *Algorithmic Information Theory*, Cambridge University Press, Cambridge, 1987

was written by one of the pioneers of the complexity theory. Some of the classic works in the area of algorithmic complexity are

[4] A. Church, On the concept of a random sequence, *Bull. Amer. Math. Soc.* 46(1940), 130-135.

[5] A.N. Kolmogorov and V.A. Uspensky, Algorithms and randomness, *SIAM Journal in Probability Theory and Applications* 32(1987), 389-412 (appeared after Kolmogorov's death in 1987).

[6] P. Martin-Löf, The definition of random sequences, *Information and Control* 9(1966), 602-619.

[7] R. von Mises, *Probability, Statistics and Truth*, MacMillan, New York, 1939.

Issue 17.3 (1989) of the *Annals of Probability* was devoted to the in-depth analysis of Kolmogorov's contributions to probability theory, computational complexity, and dynamical systems theory. The physics of the coin tossing were discussed in

[8] J. Ford, How random is a random coin toss, *Physics Today* 36(1983), 40-47 (April).

As always, the multi-volume

[9] D. E. Knuth, *The Art of Computer Programming*, Volumes 1-3, Addison-Wesley, Reading, MA, 1973,

is an invaluable source for anything related to computing. The second volume contains a good discussion of the pseudorandom number generation problem. More recent sources in this area are

[10] H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods*, SIAM, Philadelphia, 1992.

[11] S. Tezuka, *Uniform Random Numbers. Theory and Practice*, Kluwer, Boston, 1997.

The equipartition property of the Champernowne number has been proved in

[12] D.G. Champernowne, The construction of decimals normal in the scale of ten, *J. London Math. Soc.* 8 (1933), 254-260.

A pseudorandom number generator using the digits of number π has been recently proposed in

[13] Y. Dodge, A natural random number generator, *Int. Stat. Review* 64(1996), 329-344

and nonlinear methods in pseudorandom number generation are discussed in

[14] J. Eichenauer-Herrmann, Pseudorandom number generation by nonlinear methods, *Int. Stat. Review* 63(1995), 247-255.

A compact disc

[15] G. Marsaglia, The Marsaglia Random Number CD ROM Including the Diehard Battery of Tests of Randomness

distributed in 1995 as freeware by the Department of Statistics and Supercomputer Computations Research Institute at the Florida State University, is very relevant to the discussions of this chapter.

Finally, Section 4.6 is based on two very recent articles:

[16] S. Pincus and B.H. Singer, Randomness and degrees of irregularity, *Proc. Natl. Acad. Sci. USA* 93 (1996), 2083-2088.

[17] S. Pincus and R.E. Kalman, Not all (possibly) "random" sequences are created equal, *Proc. Natl. Acad. Sci. USA* 94 (1997), 3513-3518.

Chapter 5

Statistical Independence and Kolmogorov's Probability Theory

Independently repeated experiments with random outcomes have a formal counterpart in the mathematical concept of statistical independence. The cleanest way to introduce the latter can be found within the framework of Kolmogorov's axiomatic probability theory which, since the 1930s, became the standard, and by far most widespread, mathematical model of randomness.

5.1 Description of experiments, random variables, and Kolmogorov's axioms

In this section we will consider experiments with several (or, infinitely many) possible random outcomes which cannot be precisely predicted given the experimental conditions. It is convenient to have a special term for this kind of empirical situation: we will call such experiments *random trials*. Many diversified examples of random trials can be found in Chapter 1.

Example 1.3.2 gives measurements of diameters of bases of fragmentation bombs in a Cleveland factory. Although the bases were manufactured under "identical" (as much as the manufacturer can reasonably control them) conditions, the outcomes of the measurement process were not uniquely determined. Each measurement can thus be viewed as a random trial.

In Example 1.7.1 we provided the space shuttle *Columbia* accelerometer readings taken at different times. Again, the values of the gravitational constant at that point in space are scattered, the fluctuations impossible to determine exactly by taking into account the physical conditions. When, in the 17th century, Galileo kept dropping different objects from the Leaning Tower of Pisa to determine the gravitational constant on the surface of the Earth, he was just conducting random trials.

M. Denker, W. Woyczynski, *Introductory Statistics and Random Phenomena*,
Modern Birkhäuser Classics, DOI 10.1007/978-3-319-66152-0_5

Obviously, our thought experiments of random coin tossing and dice rolling, or blindly selecting 6 out of the 45 balls in the Ohio Lottery drawing, are also convenient examples of random trials. In these three cases the expressions "random tosses" or "blind selection" referred to the assumption that equal chances (thought of as relative frequencies in long runs of the experiment) were assigned to different outcomes of the experiment; often this assumption was implicitly justified by built-in symmetries of the physical phenomenon under study.

We will begin a description of the A.N. Kolmogorov's probability theory, which was designed to provide a mathematical model of random trials, by introducing a formal labeling of their possible outcomes.

Definition 5.1.1 Probability Space.

A mathematical model of a random trial is a probability space $(\Omega, \mathcal{B}, P)$, *where* Ω, *the sample space, consists of all possible simple outcomes* ω *of the trial which are called sample points.* $\mathcal{B}$ *is the family of all composite outcomes* $B \subset \mathcal{B}$, *called random events, for which the probability measure* $P(B)$ *is defined a priori. The latter is assumed to satisfy the following three requirements (axioms):*

(i) Positivity Axiom: For any random event $B \in \mathcal{B}$,

$$0 \leq P(B) \leq 1,$$

(ii) Normalization Axiom:

$$P(\Omega) = 1,$$

(iii) Additivity Axiom: For any disjoint random events $A, B \in \mathcal{B}$, $A \cap B = \emptyset$,

$$P(A \cup B) = P(A) + P(B).$$

Obviously, the probability space axioms are motivated by the corresponding properties of the relative frequencies for empirical data. By induction, the Additivity Axiom (iii) immediately implies the *finite additivity* of probability measure P:

$$P(B_1 \cup \ldots \cup B_n) = P(B_1) + \ldots + P(B_n) \qquad (1)$$

for any pairwise-disjoint random events $B_1, \ldots, B_n \in \mathcal{B}$, $B_i \cap B_j = \emptyset, i \neq j$.

A Mathematical Aside: Countable Additivity. For infinite sample spaces Ω, the Additivity Axiom (iii) is usually replaced by the condition of *countable additivity*:

(iii') Countable Additivity Axiom: For any pairwise-disjoint sequence of random events $B_1, B_2, \ldots \in \mathcal{B}$, $B_i \cap B_j = \emptyset, i \neq j$,

$$P(\bigcup_{i=1}^{\infty} B_i) = \sum_{i=1}^{\infty} P(B_i). \tag{2}$$

The infinite union of sets appearing on the left-hand side is understood as follows:

$$\bigcup_{i=1}^{\infty} B_i := \{\omega \in \Omega : \omega \in B_i, \text{ for some } i = 1, 2, \ldots\} \tag{3}$$

Compared to (1), this is more than just a cosmetic change. Actually, the condition (2) is equivalent to the assumption that probability measures on infinite sample spaces are *continuous* as set functions. Indeed, the Countable Additivity Axiom is satisfied if and only if for every sequence of descending random events $B_1 \supset B_2 \supset \ldots$,

$$\lim_{i\to\infty} B_i := \bigcap_{i=1}^{\infty} B_i := \{\omega \in \Omega : \omega \in B_i, \text{ for every } i = 1, 2, \ldots\},$$

we have

$$P(\lim_{i\to\infty} B_i) = \lim_{i\to\infty} P(B_i). \tag{4}$$

The Countable Additivity Axiom also makes selection of the random events family $\mathcal{B}$ more poignant. It turns out that if one chooses as the sample space Ω the unit interval [0, 1] and as the random events family $\mathcal{B}$ the collection of *all* subsets of the sample space, then there exists no nontrivial countably additive probability P which would provide a way to measure probabilities of all $B \in \mathcal{B}$. This is a fairly deep mathematical result and shows that, for richer sample spaces, one has to be careful about what one calls random events.

The question of extension of a finitely additive probability to a countably additive probability is quite subtle. Contemplate for few minutes the following example:

Take as Ω the *countable* set $\mathbf{Q}$ of all rational numbers in the interval [0,1]. Introduce, on intervals of rational numbers $[a, b]$, $a \leq b$, $a, b \in \mathbf{Q}$, the probability measure $P([a, b]) := b - a$ which defines the probability of finding a rational number in the rational interval $[a, b]$ as its colloquially understood length. Such a probability is obviously finitely additive but it has no countably additive extension. Indeed, by definition, the probability of each simple event $P([a, a]) = 0$, so if P were countably additive then we would have

$$P(\Omega) = \sum_{a \in Q} P([a, a]) = 0,$$

a contradiction since, by the Normalization Axiom, $P(\Omega) = 1$.

Of course, if one accepts the Countable Additivity Axiom, then it is not possible either to have a probability distribution giving equal probabilities to every one of the natural numbers. Yet, intuitively, such a distribution seems to be quite natural and useful. This issue is discussed in a recent article "Using finitely additive probability: uniform distributions on the natural numbers" by J. B. Kadane and A. O'Hagan, *J. Amer. Stat. Asso.* 90(1995), 626-631.

To summarize the above discussion we emphasize that the subject of the axiomatic probability theory is the probability space $(\Omega, \mathcal{B}, P)$, where the probability measure P is selected *up-front* by other applied probability considerations taking into account the physical nature of the phenomena under study. How to fine-tune this selection to real-life experimental data is the subject matter of *statistics* which will be discussed in Part 3 of this book.

The need for such a rigorous approach became clear towards the end of the 19th century when a number of probabilistic "paradoxes" baffled the experts. All of those "paradoxes" were caused by differing interpretation of what the word "random" meant in terms of the probability measure P.

Example 5.1.1 Bertrand Random Chord Paradox.
What is the probability P that a randomly selected chord is shorter than the side S of an equilateral triangle inscribed in the circle?

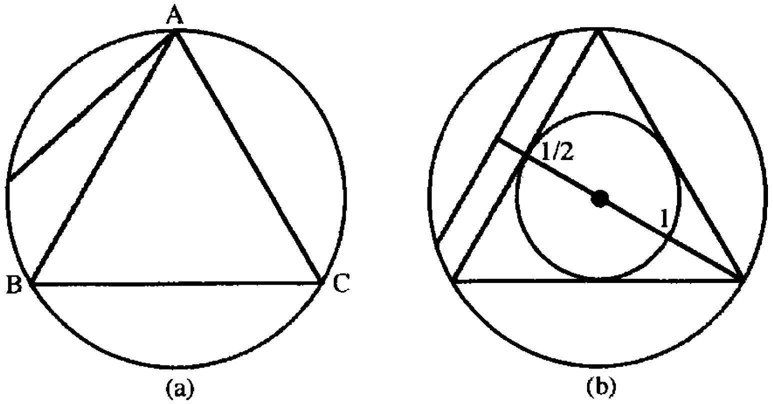


FIGURE 5.1.1
Different, and "equally" justified, ways of selecting "uniform" probability P in the Bertrand random chord paradox.

Here are two possible solutions corresponding to illustrations in Fig. 5.1.1:

(a) A chord is determined by its two end points. Fix one of them to be A. For the chord to be shorter than the side S, the other end point must be chosen on either the arc AB or the arc CA, and each of them is subtended by an angle of 120°. Thus, $P = 2/3$.

(b) A chord is completely determined by its center. For the chord to be shorter than the side S, the center must lie outside the circle of radius equal to the half of the radius of the original circle and the same center. Hence, the probability P equals the ratio of the annular area between two circles and the area of the original circle, which is 3/4.

This "paradox", still hotly debated in 1907 when Jean Bertrand lectured on probability theory at the Paris Sorbonne, can be resolved only by an *a priori* imposition of the probability measure P, the recipe Kolmogorov recommended in the axiomatic theory introduced in his 1933 book. The answer to the question "Which of the two solutions is correct?" cannot be provided within the framework of probability theory. However, it can be phrased as an empirical (and thus statistical) question, or an applied probability question about the physical mechanism of the chord selection.

We will illustrate the formal probability theory framework by returning to the generic examples of random phenomena that were previously discussed.

Example 5.1.2 Single Coin Toss.

The sample space Ω here can be chosen to consist of only two sample points; say the first sample point is H (for Heads), and the second sample point is T (for Tails). The random events family $\mathcal{B}$ consists of the empty set $\emptyset$, the simple random events $\{H\}$ and $\{T\}$ consisting of single sample points, and one composite event $\{H, T\} = \Omega$ which happens to coincide with the whole sample space. If we want to model a fair coin toss, then we have to assign the probability measure to random events (notice that for finite sample spaces it suffices to define it on events consisting of single sample points) as follows: $P(\emptyset) = 0$, $P(\{H\}) = P(\{T\}) = 1/2$, and $P(\Omega) = 1$. If a biased coin toss is to be modeled, then we have to select a number p, $0 < p < 1$, and impose the probabilities $P(\emptyset) = 0$, $P(\{H\}) = p$, $P(\{T\}) = 1 - p$, and $P(\Omega) = 1$.

Example 5.1.3 Multiple Coin Toss.

For n coin tosses, the sample space Ω can be selected to consist of sample points $\omega = (\eta_1, \dots, \eta_n)$, where each of η_i is either equal to H or to T. There are 2^n sample points in this sample space. The family $\mathcal{B}$ of random events is again the family of all subsets of the sample space, including the empty set $\emptyset$ and the whole sample space Ω—there are 2^{2^n} random events in $\mathcal{B}$ (check it for small ns first). For $n = 10$, the collection of all the 10-toss series in which exactly 2 heads came up constitutes a random event (call it B_2^{10}); it consists of $\binom{10}{2} = 45$ sample points which easily can be written out explicitly. You may want to do it by hand as a warm-up exercise.

In the fair coin model, *assuming* equal probabilities of all sample points, we have no choice but to impose the probability P on simple random events by the condition $P(\{\omega\}) = 2^{-n}$ for all $\omega \in \Omega$, and by extending it to other (composite) random events via the finite additivity property of P. So, for $n = 10$, we have $P(B_2^{10}) = 45 \cdot 2^{-10} \approx 0.04394$.

In the biased coin model, we must select a number p, $0 < p < 1$, and then we can impose the following probabilities on simple random events consisting of a single sample point $\omega = (\eta_1, \dots, \eta_n)$ with exactly k, $0 \le k \le n$, of η_is equal to H [like, e.g., $\omega = (T, H, T, T, T, T, H, T, T, T)$, where $k = 2$]:

$$P(\{\omega\}) = p^k(1-p)^{n-k}. \tag{5}$$

So, for example, for $n = 10$ and $p = 4/10$, we get $P(B_2^{10}) = 45 \cdot 0.4^2 \cdot 0.6^8 \approx 0.75583$.

That all of these probabilities add up to 1, thus satisfying the Normalization Axiom (ii), is a direct consequence of the binomial theorem:

$$\sum_{\omega\in\Omega} P(\{\omega\}) = \sum_{k=0}^{n} \sum_{\omega:\#\{i:\eta_i=H\}=k} p^k(1-p)^{n-k} = \sum_{k=0}^{n} \binom{n}{k} p^k(1-p)^{n-k} = 1. \tag{6}$$

An experimenter who has found (5) as the relative frequencies of the appearance of k heads in n tosses, obviously knows that (6) has to hold true: no other outcomes are possible and one of them has to occur.

Example 5.1.4 Random Numbers from the Unit Interval.
Select as the sample space Ω the set of all real numbers from the unit interval $[0, 1]$, and as the family $\mathcal{B}$ of random events the intervals $(a, b]$, $0 \le a < b \le 1$, and all the other subsets of the unit interval that can be produced effectively (in the sense of Chapter 4) from the above intervals using the operations of union, intersection, and complement, and their limits (plus the usual empty set and the whole sample space). The sample points ω are just real numbers in the unit interval. If our desire is to model random numbers uniformly distributed on the unit interval, then we have to impose the probabilities $P([a, b]) = b - a$ on random events that are intervals and extend it to other random events using the additivity property of P. Thus, the event $\{\omega : |\omega - 1/3| > 1/9\}$ has probability

$$P\left(\left\{\omega : \left|\omega - \frac{1}{3}\right| > \frac{1}{9}\right\}\right) = P\left(\left[0, \frac{2}{9}\right]\right) + P\left(\left[\frac{4}{9}, 1\right]\right) = \frac{2}{9} + \frac{5}{9} = \frac{7}{9}.$$

The sample space here (and hence the family of random events) is infinite and simple random events consisting of single sample points have probability 0.

Example 5.1.5 A System of n Molecules.

As in Example 4.4.1, consider a model for the 3-dimensional gas consisting of n molecules of mass 2 located in a unit cube and of total kinetic energy bounded by 1. Then, as the sample space Ω we can select the set of all phase space points

$$\omega = (x_1^1, x_1^2, x_1^3, v_1^1, v_1^2, v_1^3, \dots, x_n^1, x_n^2, x_n^3, v_n^1, v_n^2, v_n^3) \in \mathbf{R}^{6n}, \tag{7}$$

such that

$$0 \le x_i^k \le 1, \quad k = 1, 2, 3, \quad i = 1, \dots, n, \tag{8}$$

and

$$\text{KinEn}\,(\omega) = \sum_{i=1}^{n} \Big((v_i^1)^2 + (v_i^2)^2 + (v_i^3)^2 \Big) \le 1. \tag{9}$$

You can try to visualize this sample space as a unit cube in $3n$ position dimensions and a ball of radius 1 in the remaining $3n$ velocity (momentum) dimensions. As the family $\mathcal{B}$ of random events, we can select the subsets of Ω that are produced from $6n$-dimensional parallelepipeds via procedures described in Example 3.1.1. The probability measure P can be selected in many different ways dictated by the physics of the situation, and one choice is to choose that probability to be the normalized $6n$-dimensional volume on subsets of Ω, i.e., for a $B \in \mathcal{B}$, define

$$P(B) = \frac{\text{Volume } B}{\text{Volume } \Omega},$$

the number in the denominator being equal to just the volume of the $3n$-dimensional Euclidean ball of radius 1 because the d-dimensional volume of the unit d-dimensional cube is equal to 1 anyway. Recall that the uniform distribution on a d-dimensional ball was simulated in the *Mathematica* Experiment 3.10.2.[1]

In theoretical models of experiments one is often interested not in the original labeling ω of the experimental outcomes themselves, but in some other numerical or vector quantities that depend, that is, are functions of the outcomes. Functions for which the probability distributions of their values are computable, at least in principle, are called *random variables* or *random vectors*, if they take vector values. They are usually denoted by capital letters $X, Y, \dots$ More formally, we have the following:

[1] Recall that the unit sphere in $\mathbf{R}^d$ has the $(d-1)$-dimensional surface measure $s_d \equiv 2\pi^{d/2}/\,\Gamma(d/2)$, and that the d-dimensional volume of the unit ball is s_d/d.

Definition 5.1.2 Random Variables.
A real-valued function

$$X : \Omega \ni \omega \longmapsto X(\omega) \in \mathbf{R} \tag{10}$$

is called a random variable on the probability space $(\Omega, \mathcal{B}, P)$ *if, for every* $x \in \mathbf{R}$, *its cumulative distribution function (d.f.)*

$$F_X(x) := P(\{\omega : X(\omega) \leq x\}) \tag{11}$$

is well defined, i.e., if

$$\{\omega : X(\omega) \leq x\} \in \mathcal{B}. \tag{12}$$

Similarly, the vector-valued function

$$\boldsymbol{X} : \Omega \ni \omega \longmapsto \boldsymbol{X}(\omega) = (X_1(\omega), \ldots, X_n(\omega)) \in \mathbf{R}^n \tag{13}$$

is called a random vector on the probability space $(\Omega, \mathcal{B}, P)$ *if for every* $x_1, \ldots, x_n \in \mathbf{R}$, *its joint cumulative distribution function*

$$F_{\boldsymbol{X}}(x_1, \ldots, x_n) := P(\{\omega : X_1(\omega) \leq x_1, \ldots, X_n(\omega) \leq x_n\}) \tag{14}$$

is well defined.

Obviously, components of a random vector are random variables, and the cumulative d.f. (see also a more elementary discussion in Chapter 3) enjoys the following properties that are direct consequences of the three axioms of Definition 5.1.1:

(i)

$$0 \leq F_X(x) \leq 1; \tag{15}$$

(ii)

$$\lim_{x \to -\infty} F_X(x) = 0, \qquad \lim_{x \to \infty} F_X(x) = 1; \tag{16}$$

(iii) $F_X(x)$ is nondecreasing and continuous on the right, i.e.,

$$\lim_{x \to x_0+} F_X(x) = F_X(x_0). \tag{17}$$

In view of the additivity property of the probability P, we have also

$$P(a < X \leq b) = F_X(b) - F_X(a). \tag{18}$$

Note that the sample point label ω, the argument of the function $X(\omega)$ is being suppressed in our notation. It is not essential. As a matter of fact, in most of the probabilistic problems, cumulative d.f.s and/or related distributional descriptors such as densities, are the only objects of interest. The underlying probability space $(\Omega, \mathcal{B}, P)$ is of no direct interest and remains invisible in calculations. Obviously, any of the cumulative distribution functions discussed in Chapter 3 on analytic representation of data can serve as an example of cumulative probability d.f. within the framework of axiomatic probability theory. The reader should review that material before proceeding any further.

The typical picture of a simple cumulative probability d.f. is shown in Fig. 5.1.2 (see also Fig. 3.2.2).

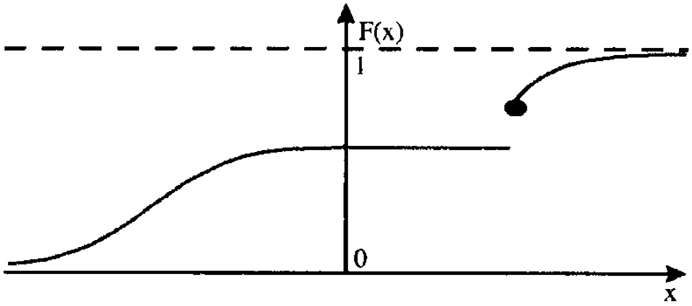


FIGURE 5.1.2

Example of a simple cumulative distribution function $F_X(x)$. The intervals I where $F_X(x)$ is flat carry no probability mass, that is, the probability that the random variable X takes values in I is zero. Jumps occur at points x such that the probability of the random event $\{X = x\}$ is strictly positive.

Example 5.1.6 Single Coin Toss.

If we model a game in which heads result in winning \$1 and tails in winning \$0, then the random variable of interest, defined on the probability space described in Example 5.1.1, would be

$$X(\omega) = \begin{cases} 1, & \text{if } \omega = H; \\ 0, & \text{if } \omega = T. \end{cases} \tag{19}$$

The corresponding cumulative d.f.

$$F_X(x) = \begin{cases} 0, & \text{if } x < 0; \\ 1/2, & \text{if } 0 \le x < 1; \\ 1, & \text{if } 1 \le x. \end{cases} \tag{20}$$

This random variable has what we recognize as the Bernoulli distribution with $p = 1/2$. If we model a game in which heads result in winning \$1 and tails in loosing \$1, then the appropriate random variable defined on the same probability space $\Omega = \{H, T\}$, would be defined by

$$Y(\omega) = \begin{cases} 1, & \text{if } \omega = H; \\ -1, & \text{if } \omega = T. \end{cases} \tag{21}$$

The corresponding cumulative d.f.

$$F_Y(x) = \begin{cases} 0, & \text{if } x < -1; \\ 1/2, & \text{if } -1 \leq x < 1; \\ 1, & \text{if } 1 \leq x. \end{cases} \tag{22}$$

Both cumulative d.f.'s are pictured in Fig. 5.1.3.

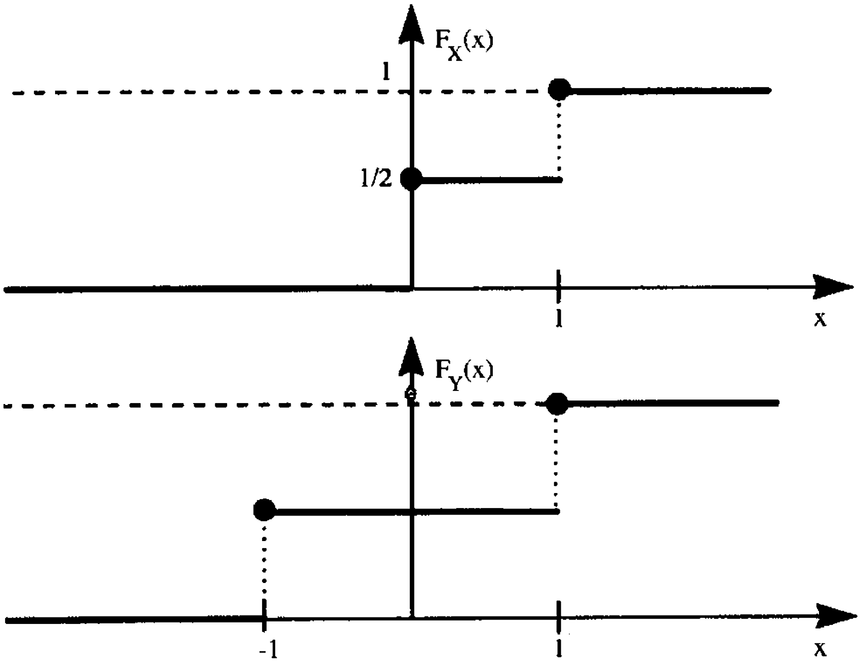


FIGURE 5.1.3
Graphs of cumulative distribution functions of random variables X and Y from Example 5.1.5.

Example 5.1.7 Total Wins in a Series of Coin Tosses.
The probability space is that of Example 5.1.2. The corresponding random variable

$$Z(\omega) := \sum_{i=1}^{n} X(\eta_i), \tag{23}$$

where $\omega = (\eta_1, \ldots, \eta_n)$, and X is the Bernoulli random variable introduced in Example 5.1.5 above. Now, it is easy to see (see Section 3.4) that the corresponding cumulative d.f. is binomial, and for general p, $1 < p < 1$, given by the formula

$$F_Z(x) = \begin{cases} 0, & \text{if } x < 0; \\ \sum_{i=0}^{\lfloor x \rfloor} \binom{n}{i} p^i (1-p)^{n-i}, & \text{if } 0 < x \le n; \\ 1, & \text{if } x \ge n. \end{cases} \tag{24}$$

The notation $\lfloor x \rfloor$ stands for the "floor" of the number x, i.e., the largest integer less than or equal to x. This cumulative d.f. is pictured, for $n = 10$, and $p = 1/2$, in Fig. 5.1.4.

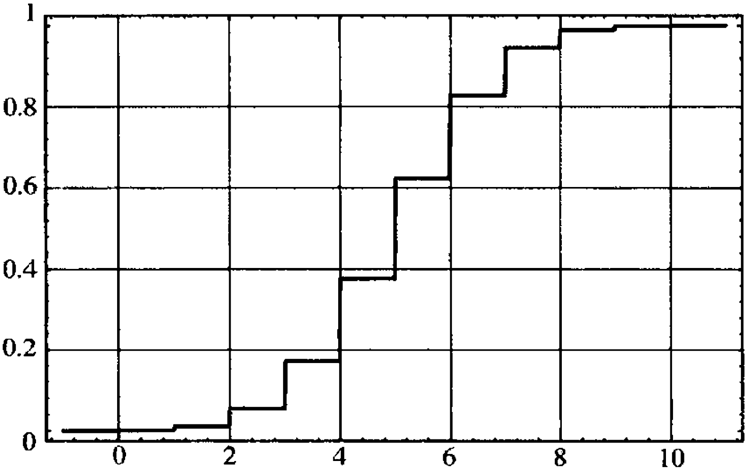


FIGURE 5.1.4
The cumulative distribution function of the binomial random variable with parameters $n = 10$, $p = 1/2$.

Example 5.1.8 System of Molecules Revisited.
Define the random variable

$$X(\omega) = \text{KinEn}\,(\omega), \tag{25}$$

as introduced in Example 5.1.4. Its cumulative d.f.

$$F_X(x) = \frac{\text{Volume of } 3n\text{-dimensional ball of radius } x}{\text{Volume of } 3n\text{-dimensional ball of radius } 1} = x^{3n}, \tag{26}$$

for $0 \leq x \leq 1$, and $F_X(x) = 0$ for $x \leq 0$, and $= 1$ for $x \geq 1$. For several values of the dimension n, the cumulative p.d. is pictured in Fig. 5.1.5. For n larger than 100 it is practically indistinguishable from a function that is 0 for $x < 1$ and 1 for $x \geq 1$, which corresponds to the probability mass almost totally concentrated at $x = 1$; see, comments in Example 4.4.1 and also *Mathematica* Experiment 3.10.2, where the vector random quantity uniformly distributed over the unit n-dimensional ball was simulated.

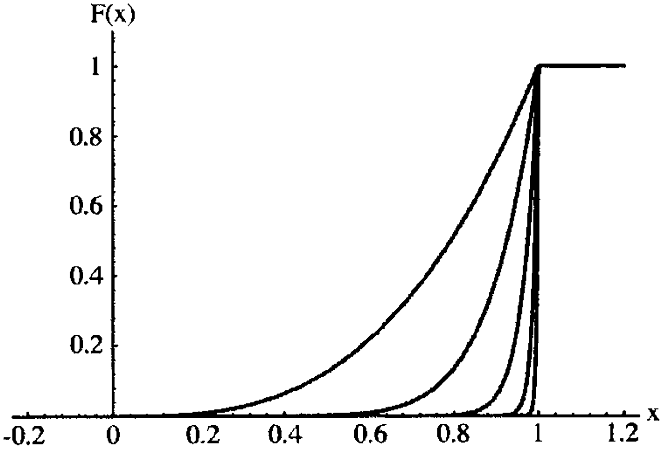


FIGURE 5.1.5
Cumulative distribution functions from Example 5.1.7 for dimensions $n = 1, 3, 9, 27, 81$ (from left to right).

There is a standard way to produce a random variable with a given cumulative d.f. $F(x)$ satisfying the above condition (15-17) which is an analogue of the quantile function method of Section 3.3. Take as the probability space $(\Omega, \mathcal{B}, P)$ the unit interval $[0, 1]$, and $\mathcal{B}$ and the uniform probability P as specified in Example 5.1.3. Since

$$F : \mathbf{R} \ni x \longmapsto F(x) \in [0, 1] = \Omega \tag{27}$$

is nondecreasing, the (generalized) inverse function

$$F^{-1} : \Omega \ni \omega \longmapsto F^{-1}(\omega) = \min\{x : F(x) \geq \omega\} \in \mathbf{R} \tag{28}$$

defines a random variable $X = F^{-1}$ on the standard probability space [0, 1] with cumulative d.f. equal to $F(x)$. Indeed,

$$P(X \leq x) = P(\{\omega : F^{-1}(\omega) \leq x\}) = P(\{\omega : \omega \leq F(x)\}) = F(x). \quad (29)$$

We have already employed the same idea in Section 3.3, where we devised a method to simulate a random quantity with the prescribed relative frequency distribution. A graph of the inverse function $F^{-1}(\omega)$ is simply the reflection in the diagonal $\omega = x$ of the graph of the original cumulative d.f. $F(x)$, and is nothing but the graph of the quantile function $Q(q)$ of the cumulative d.f. $F(x)$ (see Fig. 5.1.6).

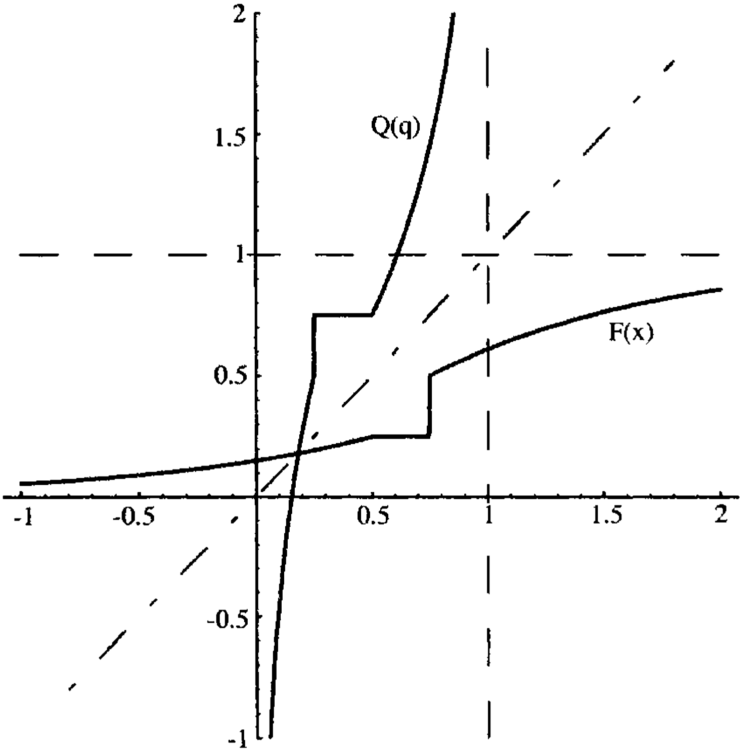


FIGURE 5.1.6
Graph of the standard representation of a random variable $F^{-1}(\omega) = Q(q)$, with a prescribed cumulative d.f. $F(x)$ is obtained by reflecting the graph of $F(x)$ in the diagonal.

As we have already observed in the more elementary context of Chapter 3, there are two basic types of cumulative d.f.s: discrete and (absolutely) continuous, and they correspond to the classification introduced in Chapter 3, where we used both of them as analytical approximations to experimental frequency distributions.

Discrete distributions. The discrete random variable X can only take a finite or a countable number of values $x_1, x_2, \ldots \in \mathbf{R}$ (or $\mathbf{R}^d$) with positive probability, and the corresponding probability distribution is determined by a sequence of discrete probabilities

$$p_i := P(X = x_i) \tag{30}$$

which have to satisfy the condition

$$\sum_i p_i = 1. \tag{31}$$

In this case, for any set $A \subset \mathbf{R}$,

$$P(X \in A) = \sum_{\{i:x_i \in A\}} p_i. \tag{32}$$

In particular, for a discrete random variable X, the cumulative d.f.

$$F_X(x) = \sum_{\{i:x_i \leq x\}} p_i, \tag{33}$$

and it is piecewise-constant with jumps upwards of size p_i at points x_i. Several examples of discrete probability distributions (Bernoulli, binomial, Poisson, multinomial) were given in Chapter 3.

Absolutely continuous distributions. An (absolutely) continuous random variable X takes an uncountable number of values x and its cumulative probability d.f. is determined via the formula

$$F_X(x) = \int_{-\infty}^{x} f_X(y)\,dy, \tag{34}$$

where the *density function* $f_X(x)$ is nonnegative (≥ 0), and must satisfy the normalization condition

$$\int_{-\infty}^{\infty} f_X(x)\,dx = 1. \tag{35}$$

A d-dimensional (absolutely) continuous *random vector* $\boldsymbol{X} = (X_1, X_2, \ldots, X_d)$ has the joint cumulative probability d.f. determined by the analogous formula:

$$F_{\boldsymbol{X}}(x_1, \ldots, x_d) = \int_{-\infty}^{x_d} \cdots \int_{-\infty}^{x_1} f_{\boldsymbol{X}}(x_1, \ldots, x_d)\,dx_1 \ldots dx_d, \tag{36}$$

where the joint density function $f_{\boldsymbol{X}}(x_1, x_2, \dots, x_d) \geq 0$ satisfies the normalization condition

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\boldsymbol{X}}(x_1, \dots, x_d)\, dx_1\, dx_2 \dots dx_d = 1. \tag{37}$$

The corresponding probabilities are computed by integrating the densities over the desired region:

$$P(X \in A) = \int_A f_X(x)\, dx \tag{38}$$

in the one-dimensional case, with a similar formula in the d-dimensional case. Notice that, in view of the fundamental theorem of calculus, for absolutely continuous cumulative probability d.f.s

$$\frac{dF_X(x)}{dx} = f_X(x). \tag{39}$$

A similar formula involving partial derivatives holds true for the d-dimensional case.

Several examples of absolutely continuous distributions (uniform, exponential, Gaussian, multivariate normal, Weibull, Cauchy, Pareto, etc.) were given in Section 3.5. Other natural examples will arise later in this chapter.

Singular distributions. Although it is not apparent at the first sight, the discrete and continuous cumulative probability distributions and their mixtures do not exhaust the realm of all possible distributions. Consider the so-called *devil's staircase* cumulative d.f. $F(x)$ obtained with the help of the Cantor set (see Section 2.7)) as indicated in Fig. 5.1.7.

Such a function is indeed a cumulative probability d.f. since $F(x) = 0$ for $x \leq 0$, and $= 1$ for $x \geq 1$, and since it is nondecreasing and right-continuous. As a matter of fact, one can check that it is *continuous* everywhere, but its derivative is zero wherever it is defined. So, obviously, $F(x)$ cannot be the indefinite integral of its derivative, i.e., the cumulative d.f. F does not have a density. In other words, it is "continuous" but not "absolutely continuous"; hence, the need to distinguish between those two concepts.

Cumulative probability d.f.s that are neither absolutely continuous, nor discrete, nor their mixtures, are called *singular*.

If one knows the joint cumulative probability d.f. of a random vector, then it is easy to recover from it the cumulative probability d.f.s of its 1-dimensional components which are called *marginal cumulative probability d.f.s*. Indeed, if, say $\boldsymbol{X} = (X, Y)$, then

$$F_X(x) = P(X \leq x) = P(X \leq x, Y < \infty) = F_{\boldsymbol{X}}(x, \infty), \tag{40}$$

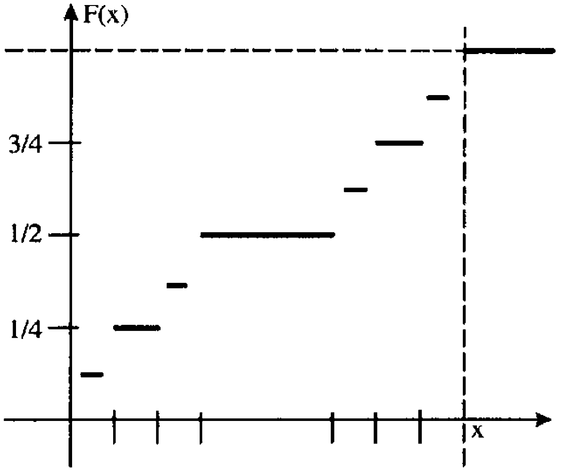


FIGURE 5.1.7

A step in the construction of the devil's staircase cumulative probability d.f. We start with setting $F(x) = 2^{-1}$ on the "middle-third" interval of length 3^{-1} in the interval [0,1], then set $F(x) = 2^{-2}$ and, respectively, $3 \cdot 2^{-2}$, on the next generation two "middle-third" intervals of length 3^{-2}, and continue this process indefinitely.

and, in the absolutely continuous case, the *marginal density* of the first component

$$f_X(x) = \frac{dF_X(x)}{dx} = \frac{dF_{\mathbf{X}}(x,\infty)}{dx} = \int_{-\infty}^{\infty} f_{\mathbf{X}}(x,y)\,dy \tag{41}$$

is obtained by integrating out the second variable in the joint density.

5.2 Uniform discrete distributions and counting

Uniform discrete distributions on finite sets are imposed in view of various symmetry considerations. We have encountered such distributions (e.g., symmetric Bernoulli) in Chapter 3. Calculation of related probabilities requires techniques to count the number of sample points in random events of interest. The area of mathematics that studies such problems (often, very involved) is called *combinatorics*. In this section we will review a few combinatorial tools that are helpful in determining discrete probabilities.

A general model of the uniform discrete distribution assumes a finite sample space $\Omega = \{\omega_1, \ldots, \omega_n\}$, the family of random events $\mathcal{B}$ consisting of all 2^n

subsets of Ω, and the probability of any $A \in \mathcal{B}$

$$P(A) = \frac{k}{n}, \tag{1}$$

where $k = |A|$ is the number of sample points in A. This model is sometimes called the *Laplace probability space*. Here are a few special cases where the methods are standard.

Multiplication of choices. If sets $A_1, \ldots, A_k$ contain, respectively, $n_1, \ldots, n_k$, points, then the number of different k-tuples (vectors) $(x_1, \ldots, x_k)$, where $x_i \in A_i$, $i = 1, \ldots, k$ is equal to

$$n_1 \cdot \ldots \cdot n_k. \tag{2}$$

Hence, there are $2 \cdot 2 \cdot \ldots \cdot 2$ (10 *times*) $= 2^{10} = 1024$, different outcomes of 10 coin tosses, $6 \cdot 6 \cdot 6 = 216$ different outcomes of a roll of three dice, and $30 \cdot 29 \cdot 28 = 24360$ ways of selecting the president, vice-president, and treasurer of the *Phi Gamma Delta* fraternity with a membership of 30. The multiplication rule leads to another useful formula:

Permutations. Let $S = \{s_1, s_2, \ldots, s_n\}$ be a set of n different objects. The numbers of ways k objects, $k \leq n$, can be selected from S in a particular order (and without replacement) is

$$n(n-1)(n-2) \cdot \ldots \cdot (n-k+1) = \frac{n!}{(n-k)!}. \tag{3}$$

The derivation of the formula is clear: the first object can be selected in n ways, the second in $(n-1)$ ways, until, finally, the k-th object can be selected in $(n-k+1)$ ways. At this point, one applies the multiplication of choices principle. The above number is called the number of *permutations* of k objects selected from a set of n objects. It is important to remember that the orderings count: (s_1, s_3, s_2) and (s_2, s_3, s_1), say, are different permutations.

For example, the top 6 downhill *NASTAR* racers can place in the field of 13 in $13!/(13-6)! = 1,235,520$ ways, and there are $13! = 6,227,020,800$ ways all 13 racers can place (i.e., $13!$ is the number of permutations of 13 objects).

Example 5.2.1 Two People With the Same Birthday.

A party is attended by n revelers. What is the probability P_n that at least two people have the same birthday? It is easier to calculate the complementary probability, that is, the probability of the event that all the n people have different birthdays. There are $365!/(365-n)!$ ways different birthdays can be selected out of the total of 365^n ways the birthdays can occur in a group of n. Thus, the sought probability is

$$P_n = 1 - \frac{365!}{(365-n)!365^n}.$$

Check that $P_{23} \approx 0.5$, $P_{50} \approx 0.97$ (use *Stirling formula* $n! \approx \sqrt{2\pi}n^{n+1/2}e^{-n}$ to approximate the factorials).

Combinations. Let $S = \{s_1, s_2, \dots, s_n\}$ be again a set of n different objects. If one selects objects from a set, then we say that a *combination* of k objects was selected from S if they were picked at once, and the order of selection does not matter: (s_1, s_3, s_2) and (s_2, s_3, s_1), say, are the same combinations. Obviously, the number of combinations of k objects out of n is smaller than the number of permutations, since $k!$ permutations count as a single combination. Thus, the number of combinations of k objects out of a set of n objects (or, as one often says, the *number of combinations of n objects taken k at a time*) is

$$\frac{n!}{(n-k)!} \cdot \frac{1}{k!} = \binom{n}{k}, \tag{4}$$

the familiar binomial coefficient.

Example 5.2.2 Drawing Balls With Replacement.
We have n boxes, and each contains w white balls and r red balls. We draw a ball from each box (this is equivalent to drawing n times from the same box, replacing the ball after each draw). What is the probability P_k of drawing k white balls and $n-k$ red balls? There are $(w+r)^n$ different equiprobable n draws of $w+r$ balls, and out of those there are $\binom{n}{k}w^k r^{n-k}$ ways to select exactly k white balls. Thus, the sought probability is

$$P_k = \binom{n}{k}\frac{w^k r^{n-k}}{(w+r)^n} = \binom{n}{k}p^k(1-p)^{n-k},$$

where $p = w/(w+r)$. It is the familiar *binomial distribution.*

Example 5.2.3 Drawing Balls Without Replacement.
We draw, without replacement, $n \le \min(w, r)$ balls from a box containing w white balls and r red balls. What is the probability P_k of drawing k white balls, $k \le n$, and $n-k$ red balls? There are $(w+r)(w+r-1) \cdot \ldots \cdot (w+r-n+1)$ possible selections out of which

$$\binom{n}{k}w(w-1)\cdots(w-k+1)r(r-1)\ldots(r-(n-k)+1)$$

have exactly k white balls. Thus the desired probability

$$P_k = \frac{\binom{w}{k}\binom{r}{n-k}}{\binom{w+r}{n}}.$$

Example 5.2.4 A Politically Incorrect Committee.

A faculty committee of 4 is to be selected at random from a group of 5 men and 11 women. What is the probability P that the committee consists of 3 men and 1 woman? Out of the total number of $\binom{16}{4} = 1820$ possibilities, there are $\binom{5}{3}\binom{11}{1} = 110$ ways to select a committee of 3 men and 1 woman. Thus, $P = 110/1820 = 0.06$.

Example 5.2.5 Distributing Christmas, Hanukkah, and Kwanzaa Presents Among Relatives.

We distribute n presents among m relatives and all the m^n arrangements are assumed to be equally probable. Then, the probability that for each $i = 1, 2, \ldots, m$, the i-th relative receives k_i presents, $k_1 + \ldots + k_m = n$, is

$$\frac{n!}{k_1!k_2!\cdots k_m!}\cdot\frac{1}{m^n},$$

which corresponds to the ***multinomial distribution*** introduced earlier.

5.3 Statistical independence as a model for repeated experiments

Introduced in Example 5.1.3 probabilistic model for multiple, fair, coin tosses has one curious property: The imposed joint probability distribution of the vector $(X_1, \ldots, X_n)$ and its marginal distributions, that is the distributions of the component real-valued random variables $X_1, \ldots, X_n$, are connected by the multiplicative formula:

$$P(X_1 = x_1, \ldots, X_n = x_n) = \frac{1}{2^n} = \left(\frac{1}{2}\right)^n = P(X_1 = x_1)\cdots P(X_n = x_n). \tag{1}$$

As it turns out, this multiplicative property of joint distributions can serve as a general model of randomness within the Kolmogorov's axiomatic probability theory.

Definition 5.3.1 Random Vectors With Independent Components.

Components $X_1, \ldots, X_n$ of a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ are said to be statistically independent random variables, if the joint cumulative probability d.f. of $\boldsymbol{X}$ is equal to the product of the marginal cumulative probability d.f.s of its component random variables, that is, if for every n-tuple of real numbers $x_1, \ldots, x_n$,

$$F_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = F_{X_1}(x_1)\cdots F_{X_n}(x_n). \tag{2}$$

In practice, one often casually speaks of random variables being independent, but one has to remember that such a notion is meaningless unless one has the joint distributions of these random variables. Notice that for discrete random variables, the definition of their independence can be written as the condition that for every n-tuple of real numbers $x_1, \ldots, x_n$,

$$P(X_1 = x_1, \ldots, X_n = x_n) = P(X_1 = x_1) \cdot \ldots \cdot P(X_n = x_n). \tag{3}$$

For absolutely continuous random variables, independence simply means that their joint density is a product of marginal densities, that is,

$$f_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdot \ldots \cdot f_{X_n}(x_n). \tag{4}$$

Besides recalling Example 5.1.2, we will provide here two additional examples of independent random variables.

Example 5.3.1 Bivariate Normal Random Vectors.
A Gaussian random vector (X, Y) with the joint density function

$$f_{(X,Y)}(x, y) = \frac{1}{2\pi} \exp(-\|(x, y)\|^2/2) \tag{5}$$

has independent components since, for every $x, y \in \mathbf{R}$,

$$\frac{1}{2\pi} \exp\left(-\frac{\|(x, y)\|^2}{2}\right) = \frac{1}{2\pi} \exp\left(-\frac{x^2 + y^2}{2}\right) = \frac{e^{-x^2/2}}{\sqrt{2\pi}} \cdot \frac{e^{-y^2/2}}{\sqrt{2\pi}}.$$

Example 5.3.2 A Generic Construction of Independent Random Variables.
Consider the unit square $\Omega = [0, 1] \times [0, 1] \ni \omega = (\omega_1, \omega_2)$ with probability defined as the standard planar area measure. Then, any random vector of the form

$$\big(X(\omega), Y(\omega)\big) = \big(X(\omega_1), Y(\omega_2)\big)$$

in which the first component depends only on the first variable in the square and the second depends only on the second variable, has statistically independent components. This follows from the properties of the planar area measure (see Fig. 5.3.1).

Indeed, the set

$$\{\omega : X(\omega) \le x, Y(\omega) \le y\} = \{\omega_1 : X(\omega_1) \le x\} \times \{\omega_2 : Y(\omega_2) \le y\}$$

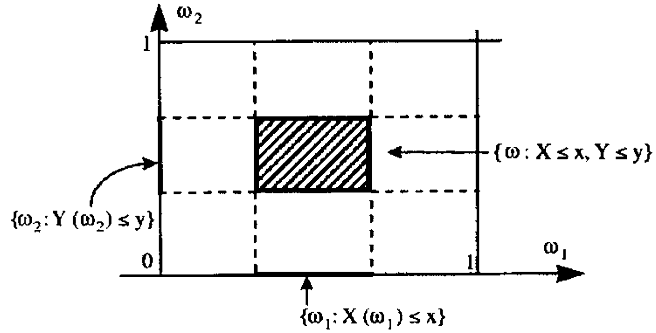

ω_2
1
$\{\omega : X \le x, Y \le y\}$
$\{\omega_2 : Y(\omega_2) \le y\}$
0
1
ω_1
$\{\omega_1 : X(\omega_1) \le x\}$

FIGURE 5.3.1
A generic construction of independent random variables.

is a rectangle, so its area (=probability)

$$P(\omega : X(\omega) \le x, Y(\omega) \le y) = P(\omega_1 : X(\omega_1) \le x) \cdot P(\omega_2 : Y(\omega_2) \le y).$$

Using the n-dimensional cube instead of the square, one can similarly construct n independent random variables with prescribed marginal distributions. A construction of the random variable on the standard probability sample space $\Omega = [0, 1]$, and with a prescribed cumulative probability d.f. $F(x)$, was shown in Fig. 5.1.6.

Modeling of more involved random phenomena within the probability theory framework requires a study of *infinite* sequences of statistically independent random variables.

Definition 5.3.2 Infinite Sequences of Independent Random Variables.
Random variables $X_1, X_2, \ldots$ are said to form a sequence of independent random variables if for each finite collection of indices $i_1, \ldots, i_n$, the random vector $(X_{i_1}, \ldots, X_{i_n})$ has statistically independent components.

A Mathematical Aside: Do Infinite Sequences of Independent Random Variables Exist? The question of existence of an infinite sequence of independent random variables with prescribed distributions is somewhat delicate, and it was exactly the issue Kolmogorov had to address to construct his successful probability theory. It is clear that what is needed is an extension of the generic construction of finitely many independent random variables provided in Example 5.3.2, to infinitely many random variables. Then the natural sample space to consider would be an infinite dimensional analogue of our familiar n-dimensional cube. The formal proof of the existence theorem can be found in the probability theory textbooks quoted in the Bibliographical Notes. However, in some cases it is easy to see that the infinite dimensional cube is not needed.

Example 5.3.3 Infinite Sequence of Independent Bernoulli Random Variables. Consider the standard sample space $\Omega = [0, 1]$, with standard linear Lebesgue measure taken as probability P, and an infinite sequence $X_1(\omega), X_2(\omega), \ldots,$ of random variables on Ω defined by the formulas

$$X_i(\omega) := \frac{\operatorname{sgn} \sin(2\pi 2^i \omega) + 1}{2}, \qquad i = 1, 2, \ldots, \tag{6}$$

The first four random variables $X_1(\omega), X_2(\omega), X_3(\omega), X_4(\omega)$ are shown in Fig. 5.3.2.

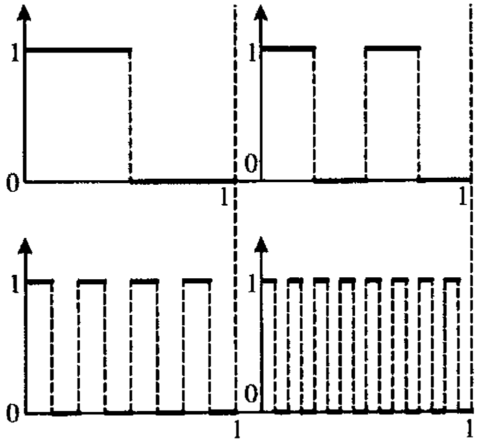


FIGURE 5.3.2

The first four functions representing an infinite sequence of independent Bernoulli random variables defined on the standard sample space $\Omega = [0, 1]$, with the Lebesgue measure taken as probability measure P.

Clearly, for each i, the probability $P(X_i = 0) = P(X_i = 1) = 1/2$ and one can check that the random variables $X_1(\omega), X_2(\omega), \ldots$ form an independent sequence.

The notion of statistical independence can be introduced in terms of random events. The random events $A, B \in \mathcal{B}$ are said to be *independent* if

$$P(A \cap B) = P(A) \cdot P(B). \tag{7}$$

Note that this equation can be rewritten in the form

$$P(B) = \frac{P(A \cap B)}{P(A)}, \tag{8}$$

where the quantity on the right-hand side is usually called the *conditional probability of B given A* and denoted $P(B|A)$. Its meaning is obvious: the conditional probability $P(B|A)$ measures the probability of the random event B but is restricted to the new probability space $\Omega' = A$. In other words, we assume that we know for sure that the event A has occurred. In this context, the statistical independence of random events A and B simply means that

$$P(B|A) = P(B), \quad \text{and} \quad P(A|B) = P(A), \tag{9}$$

that is, the conditional probability of B given condition A is independent of that condition, and vice versa.

This point of view also can be reinterpreted in terms of the repeated experimental data and their relative frequency distributions. The condition that the data sets $x_1, \dots, x_n$ and $y_1, \dots, y_n$, were obtained from "independent" experiments can be written as a condition

$$\frac{\#\{i : x_i \in R, y_i \in S\}/n}{\#\{i : x_i \in R\}/n} \approx \frac{\#\{i : y_i \in S\}}{n}. \tag{10}$$

for the joint relative frequencies, which has to be satisfied for all intervals R and S. In other words, the outcome of the first experiment did not affect the outcome of the second experiment.

5.4 Expectations and other characteristics of random variables

5.4.1 Expectations.

In our probability model, the role of the sample mean is played by the *(mathematical) expectation* of a random variable X, which in the discrete case is defined by the formula

$$E(X) = \sum_i x_i P(X = x_i), \tag{1}$$

and in the absolutely continuous case, by the formula

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx \tag{2}$$

In other words, the expectation is just the "weighted average" of the values taken by the random variable with weights provided by its probability distribution. It is important to remember that the expectation of a random variable depends *only* on its distribution and not on a particular selection of the probability space $(\Omega, \mathcal{B}, P)$ or a particular realization $X(\omega)$ of the random variable.

The following exposition is presented in terms of the absolutely continuous random variables, the formulas in the discrete case being analogous; simply, the integrals have to be replaced by summations. The case of random variables that are neither discrete nor absolutely continuous will be briefly addressed at the end of this section.

Example 5.4.1 Random Variable Taking 3 Values.
For a random variable X with distribution

$$P(X = 1) = 0.1, \quad P(X = 2) = 0.4, \quad P(X = 3) = 0.5,$$

the expectation

$$E(X) = 1 \cdot 0.1 + 2 \cdot 0.4 + 3 \cdot 0.5 = 2.4.$$

Example 5.4.2 Poisson Random Variable.
For the Poissonian random variable X with parameter λ,

$$E(X) = \sum_{i=0}^{\infty} i e^{-\lambda} \frac{\lambda^i}{i!} = e^{-\lambda} \lambda \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!} = \lambda.$$

Example 5.4.3 Exponential Random Variable.
For a random variable X with exponential distribution with parameter λ,

$$E(X) = \int_0^{\infty} x\lambda e^{-\lambda x} dx = x(-e^{-\lambda x}) \Big|_0^{\infty} + \int_0^{\infty} e^{-\lambda x} dx = \frac{1}{\lambda}.$$

Note that the integral is restricted to the positive halfline because the exponential density is zero for negative real numbers.

Remark 5.4.1 *Warning: Expectations Need Not Exist.* One should remember that, for a general random variable, the expectation need not be well defined. Indeed, the series (resp. the integral) in formula (1) [resp. (2)] may diverge. For example, the discrete random variable with the distribution

$$P(X = i) = \frac{C}{i^2}, \quad C = \left(\sum_{i=1}^{\infty} \frac{1}{i^2} \right)^{-1},$$

has no expectation because the harmonic series diverges:

$$\sum_{i=1}^{\infty} i\frac{C}{i^2} = C\sum_{i=1}^{\infty}\frac{1}{i} = \infty.$$

The Cauchy distribution introduced in Chapter 3 has no expectation either. Calculations of expectations for other probability distributions are provided in Chapter 3, where we discussed densities as a means of compression of experimental data.

5.4.2 Expectations of functions of random variables. Variance.

In more generality, if $g(x)$ is a function and X is a random variable, then the expectation of the random variable $g(X)$ is defined by the formula

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x)\,dx. \qquad (3)$$

In particular, selecting $g(x) = (x - E(X))^2$ we get the *variance*

$$\text{Var}\,(X) \equiv \sigma^2(X) = E(X - E(X))^2 = \int_{-\infty}^{\infty} (x - E(X))^2 f_X(x)\,dx \qquad (4)$$

of the random variable X. In other words, the variance of X measures the average square deviation of the random variable X from its expectation $E(X)$. The bigger the variance, the more spread-out the distribution is. The following result provides an estimate of the probability of such a "deviation from expectation" in terms of the variance. It will play a critical role later in this chapter in our study of the Law of Large Numbers.

Theorem 5.4.1 Chebyshev's inequality.
For any random variable X, and any $\epsilon > 0$,

$$P\big(|X - E(X)| > \epsilon\big) \leq \frac{\text{Var}\,(X)}{\epsilon^2}. \qquad (5)$$

PROOF In the absolutely continuous case, if $|x - E(X)| > \epsilon$, then

$$f_X(x) \leq \frac{(x - E(X))^2}{\epsilon^2} f_X(x).$$

Hence,

$$P\,(|X - E(X)| > \epsilon) = \int_{\{x:|x-E(X)|>\epsilon\}} f_X(x)\,dx$$

$$\leq \int_{\{x:|x-E(X)|>\epsilon\}} \frac{(x-E(X))^2}{\epsilon^2} f_X(x)\,dx \leq \frac{1}{\epsilon^2}\int_{-\infty}^{\infty}(x-E(X))^2 f_X(x)\,dx$$

$$= \frac{\operatorname{Var}(X)}{\epsilon^2}.$$

A similar argument gives the Chebyshev inequality for discrete random variables.
∎

Example 5.4.4 A Universal "Three Sigma" Estimate.
For any random variable X with expectation μ and variance σ^2, the Chebyshev inequality immediately yields the estimate

$$P\left(|X - E(X)| > 3\sigma\right) \leq \frac{\sigma^2}{9\sigma^2} = \frac{1}{9}.$$

In other words, the probability that any random variable differs from its expectation by more than three standard deviations $\sigma = \sigma(X) = \sqrt{\operatorname{Var}(X)}$, is at most 1/9. Expressed differently,

$$P\left(|X - E(X)| \leq 3\sigma\right) \geq 1 - \frac{1}{9} = \frac{8}{9} = 0.8888, \tag{6}$$

the probability that any random variable takes values within three standard deviations σ of its expectation is at least 8/9. For particular random variables, the Chebyshev inequality (5) may give a somewhat crude estimate. For example, if X is a standard Gaussian random variable, then the actual probability on the left-hand side of (6) is .9987, the value that can be verified in *Mathematica* or in the table at the end of this book. However, the value of the Chebyshev inequality lies in its universal applicability.

Expectations scale linearly, that is, rescaling the random variable X leads to the identical rescaling of its expectation. Indeed,

$$E(\alpha X) = \int_{-\infty}^{\infty} \alpha x f_X(x)\,dx = \alpha E(X). \tag{7}$$

5.4.3 Expectations of functions of vectors. Covariance.

The notion of expectation is also applicable to random vectors $X = (X_1, \ldots, X_n)$ by defining the expectation componentwise:

$$EX = (EX_1, \ldots, EX_n).$$

To avoid too many brackets we will often write EX instead of $E(X)$. On the other hand, if $g(x_1, \ldots, x_n)$ is a real-valued function of n variables, then the expectation of the random variable $g(\boldsymbol{X})$ is defined by the formula

$$E(g(\boldsymbol{X})) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_n) f_{\boldsymbol{X}}(x_1, \ldots, x_n)\, dx_1 \cdots dx_n. \qquad (8)$$

Taking the function of two variables $g(x, y) = (x - EX)(y - EY)$ produces the *covariance*

$$\begin{aligned}\text{Cov}\,(X, Y) &= Eg(X, Y) = E(X - EX)(Y - EY) \qquad (9)\\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - EX)(y - EY) f_{(X,Y)}(x, y)\, dx\, dy\end{aligned}$$

of random variables X and Y. Its normalized version

$$\text{Corr}\,(X, Y) = \frac{\text{Cov}\,(X, Y)}{\sigma(X)\sigma(Y)} \qquad (10)$$

is called the *correlation coefficient* of random variables X and Y [compare (10) with the Section 2.6 definition of the correlation coefficient for finite experimental data], and satisfies the inequality

$$-1 \le \text{Corr}\,(X, Y) \le 1. \qquad (11)$$

Indeed, in view of the *Schwarz inequality*, for any real-valued functions g and h,

$$E|g(X)h(Y)| \le \sqrt{E(g^2(X))} \cdot \sqrt{E(h^2(Y))}, \qquad (12)$$

which gives (11) by substituting $g(x) = x - EX,\ h(y) = y - EY$.

5.4.4 Expectation of the product. Variance of the sum of independent random variables.

For independent random variables we have

Theorem 5.4.2
If X and Y are independent random variables and $g(x)$ and $h(y)$ are real functions, then

$$E\big(g(X)h(Y)\big) = Eg(X) \cdot Eh(Y),$$

as long as the expectations are well defined. In particular,

$$E(XY) = EX \cdot EY,$$

that is, for independent random variables the expectation of the product is the product of expectations.

PROOF Indeed, in view of the independence of X and Y, the joint density $f_{(X,Y)}(x, y) = f_X(x) f_Y(y)$, so that

$$E\big(g(X)h(Y)\big) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x)h(y) f_{(X,Y)}(x, y)\, dx\, dy$$

$$= \int_{-\infty}^{\infty} g(x) f_X(x)\, dx \cdot \int_{-\infty}^{\infty} h(y) f_Y(y)\, dy$$

$$= Eg(X) \cdot Eh(Y). \quad \blacksquare$$

This striking property of independent random variables (notice that it essentially claims that the integral of the product of two functions is the product of integrals—not a common occurrence), immediately implies that the covariance of two independent random variables disappears:

$$\mathrm{Cov}\,(X, Y) = E(X - EX)(Y - EY) = E(X - EX) \cdot E(Y - EY) = 0\,. \quad (13)$$

This justifies the interpretation of the correlation coefficient Corr (X, Y) as a numerical measure of the degree of independence of two random variables. Its values are always between -1 and $+1$, with the minimal value -1 taken for negatively linearly dependent random variables $X = -\alpha Y, \alpha > 0$, maximal value $+1$ taken for positively linearly dependent random variables $X = \alpha Y, \alpha > 0$, and the value 0 taken for independent random variables X, Y. Note, however, that two random variables can be uncorrelated without being independent.

The expectations not only scale linearly but are also *additive* so that, in general, they act as a *linear functional* on random variables, and we have:

Theorem 5.4.3
If X and Y are random variables with finite expectations, then

$$E(\alpha X + \beta Y) = \alpha EX + \beta EY.$$

PROOF Let $f_{(X,Y)}(x, y)$ be the joint density of random vector (X, Y). Then

$$\begin{aligned}
&E(\alpha X + \beta Y)\\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (\alpha x + \beta y) f_{(X,Y)}(x, y)\, dx\, dy\\
&= \alpha \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{(X,Y)}(x, y)\, dy\, dx + \beta \int_{-\infty}^{\infty} y \int_{-\infty}^{\infty} f_{(X,Y)}(x, y)\, dx\, dy\\
&= \alpha \int_{-\infty}^{\infty} x f_X(x)\, dx + \beta \int_{-\infty}^{\infty} y f_Y(y)\, dy\\
&= \alpha EX + \beta EY.
\end{aligned}$$

For discrete random variables, the proof is analogous. ∎

The above theorem combined with Theorem 5.4.2 gives another striking property of variances of independent random variables (which, remember, are quadratic functionals of random variables):

Theorem 5.4.4
If X and Y are independent random variables with finite variances, then

$$\mathrm{Var}\,(X + Y) = \mathrm{Var}\,(X) + \mathrm{Var}\,(Y),$$

that is, for independent random variables the variance of the sum is the sum of variances.

PROOF Indeed

$$\begin{aligned}
&\mathrm{Var}\,(X + Y)\\
&= E\Big((X + Y) - E(X + Y)\Big)^2
\end{aligned}$$

$$= E\Big((X - EX) + (Y - EY)\Big)^2$$

$$= E(X - EX)^2 + E(Y - EY)^2 + 2E(X - EX)(Y - EY)$$

$$= \text{Var }(X) + \text{Var }(Y)$$

since, in view of Theorem 5.4.2,

$$E(X - EX)(Y - EY) = 0. \quad \blacksquare$$

5.4.5 Moments and the moment generating function.

The above numerical characteristics are often complemented by so-called ***higher order moments*** of random variables. By definition, the *n-th order moment* of the random variable X is defined by the formula

$$E(X^n) = \int_{-\infty}^{\infty} x^n f_X(x)\, dx. \tag{14}$$

The knowledge of moments is, in general, not sufficient to uniquely characterize a probability distribution. For that reason, one introduces the *Laplace transform* of the random variable (whenever it exists):

$$\varphi_X(u) = E\left(e^{uX}\right) = \int_{-\infty}^{\infty} e^{ux} f_X(x)\, dx, \tag{15}$$

and the density $f_X(x)$ can then be recovered from $\varphi_X(u)$ by the inverse Laplace transform procedure. In view of the linearity and continuity of the expectation functional, one obtains the following useful formulas which express all the moments of random variable X in terms of the derivatives of its Laplace transform:

$$\varphi'_X(0) = E(Xe^{uX})\Big|_{u=0} = E(X),$$

$$\varphi''_X(0) = E(X^2 e^{uX})\Big|_{u=0} = E(X^2),$$

so that, for example,

$$\text{Var }(X) = EX^2 - (EX)^2 = \varphi''_X(0) - (\varphi'_X(0))^2.$$

In general, the Laplace transform of X contains information about moments of X of all orders, since

$$EX^n = \varphi^{(n)}(0), \tag{16}$$

For that reason, $\varphi_X(u)$ is also called the *moment generating function of X*.

Example 5.4.5 Moments of the Normal Random Variable.
A calculation of the Laplace transform of the standard normal random variable X gives

$$\varphi_X(u) = \int_{-\infty}^{\infty} e^{ux} \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx = e^{u^2/2} \int_{-\infty}^{\infty} \frac{e^{-(x-u)^2/2}}{\sqrt{2\pi}}\, dx = e^{u^2/2}, \tag{17}$$

since the above integral integrates out to 1 as a Gaussian density. Formulas (16-17) immediately permit calculation of any moment of the Gaussian random variable. In particular,

$$EX = (e^{u^2/2})' \Big|_{u=0} = (ue^{u^2/2}) \Big|_{u=0} = 0,$$

$$EX^2 = (e^{u^2/2})'' \Big|_{u=0} = (u^2e^{u^2/2} + e^{u^2/2}) \Big|_{u=0} = 1,$$

$$EX^3 = (e^{u^2/2})''' \Big|_{u=0} = (u^3e^{u^2/2} + 3ue^{u^2/2}) \Big|_{u=0} = 0,$$

$$EX^4 = (e^{u^2/2})^{(4)} \Big|_{u=0} = (u^4e^{u^2/2} + 6u^2e^{u^2/2} + 3e^{u^2/2}) \Big|_{u=0} = 3,$$

etc.

Example 5.4.6 Poisson Random Variable.
Let X be a Poisson random variable with parameter λ. Then

$$\varphi_X(t) = \sum_{n=0}^{\infty} e^{tn} e^{-\lambda} \frac{\lambda^n}{n!} = e^{-\lambda} \sum_{n=0}^{\infty} \frac{(e^t\lambda)^n}{n!} = e^{-\lambda} e^{\lambda e^t}.$$

Example 5.4.7 Exponential Random Variable.
Let X be an exponential random variable with parameter λ. Then

$$\varphi_X(t) = \int_0^{\infty} e^{tx} \lambda e^{-\lambda x} dx = \frac{\lambda}{\lambda - t},$$

and is defined only for $t < \lambda$.

In view of the multiplicative property of expectations of products of independent random variables, the Laplace transform has another important property which will be formulated as a theorem. In the next three sections, it will help us identify distributions of sums and arithmetic averages of independent random variables.

Theorem 5.4.5 Laplace Transform of a Sum of Independent Random Variables.
If X and Y are independent random variables, then

$$\varphi_{X+Y}(u) = \varphi_X(u)\varphi_Y(u), \tag{18}$$

that is the Laplace transform of the sum of independent random variables is the product of the Laplace transform of the summands.

PROOF The proof is immediate from Theorem 5.4.2, by selecting $g(x) = h(x) = e^{ux}$. Indeed, then

$$\varphi_{X+Y}(u) = Ee^{u(X+Y)} = E(e^{uX}e^{uY}) = Ee^{uX} \cdot Ee^{uY} = \varphi_X(u)\varphi_Y(u). \quad \blacksquare$$

Of course, the power of the Laplace transform comes from the fact that it can be inverted, that is, that we can find the function with a given Laplace transform. This is not always easy to do analytically by hand , but quite automatic with *Mathematica*. Just load the package `Calculus'LaplaceTransform'`, and command `LaplaceTransform[expr, x, u]` will give you the Laplace transform of `expr` as a function of `u`. The command `InverseLaplaceTransform[expr, u, x]` will get you the inverse transform as a function of `u`.

In this context, Theorem 5.4.5 is usually applied to obtain an explicit formula for the probability d.f. $f_{X+Y}(x)$ of the sum of two (or more) independent random variables. The former is obtained by taking the inverse Laplace transform of the product of the Laplace transforms of $f_X(x)$ and $f_Y(x)$. Some examples of this procedure are included in the Experiments, Exercises, and Projects section.

Remark 5.4.2 *Laplace vs. Fourier Transform.* The Laplace transform tool has its limitations; not all probability d.f.s have the Laplace transform. For example, if X is the Cauchy random variable, then its Laplace transform does not exist since the integral $\int_{-\infty}^{\infty} e^{ux}(\pi(1+x^2))^{-1}\,dx$ is not well defined for all $u \neq 0$. The remedy is to use the general *Fourier transform*

$$\phi_X(u) = Ee^{iuX} = \int_{-\infty}^{\infty} e^{iux} f_X(x)\,dx, \tag{19}$$

which is always well defined since the complex factor e^{iux} is always bounded, unlike the real exponential e^{ux} in the Laplace transform. As a matter of fact, $|e^{iux}| = 1$. The Fourier transform has properties similar to the Laplace transform, but its values are, in general, complex numbers; see the Experiments, Exercises, and Projects section.

5.4.6 Expectations of general random variables.

Here are a few comments on expectations of general random variables which do not have to be either discrete or continuous.

If $X \geq 0$ is an arbitrary non-negative random variable with the cumulative probability d.f. $F_X(x) = 0$ for $x < 0$, then one can approximate its expectation as a limit of an increasing sequence of discrete random variables X_n defined as follows:

Pick a sequence of points

$$x_i^n = \frac{i}{2^n}, \quad i = 0, 1, \dots, n2^n,$$

partitioning the interval $[0, n]$ into subintervals of length $1/2^n$ and set,

$$X_n(\omega) = \begin{cases} (i-1)/2^n, & \text{if } (i-1)2^{-n} \leq X(\omega) < i2^{-n}, i = 1, \dots, n2^n; \\ 0, & \text{if } X(\omega) \geq n, \end{cases}$$

for $n = 1, 2, \dots$ Each of the random variables X_n is bounded by n and takes $n2^n$ values. The expectations of discrete random variables X^n are well defined:

$$EX_n = \sum_{i=1}^{n2^n} \frac{i-1}{2^n}\left(F(i2^{-n}) - F((i-1)2^{-n})\right)$$

and they increase with n. Thus, we can define

$$EX = \lim_{n\to\infty} EX_n,$$

since the limit of a monotonically increasing sequence always exists (if $+\infty$ is a permissible value).

If X takes both positive and negative values, then we can always write it as a difference

$$X = X^+ - X^-$$

of two nonnegative random variables and, if both of them have finite expectations defined by the above procedure, set

$$EX \equiv \int_{-\infty}^{\infty} x\, dF_X(x) = EX^+ - EX^-.$$

This method gives the usual Riemann integral if the cumulative probability d.f. $F_X(x)$ has a density, and it produces the appropriate sums for expectations of discrete random variables. However, it is also able to handle mixed discrete-continuous distributions, and even distributions of the devil's staircase type which are of neither type. One can develop calculus rules for such expectations (integrals). For example, if $F_X(x)$ is a distribution function vanishing on $(-\infty, 0]$ and $g(x)$ is a differentiable function with continuous derivative, then the following computationally useful version of the *integration-by-parts formula* is valid:

$$E(g(X)) \equiv \int_0^{\infty} h(x) dF_X(x) = \int_0^{\infty} h'(x)(1 - F_X(x))\, dx.$$

In particular,

$$E(X) = \int_0^{\infty} (1 - F_X(x))\, dx.$$

The above, loosely sketched construction corresponds to the general construction of the so-called *Lebesgue integral* which can be found in the monographs quoted in the Bibliographical Notes.

5.5 Averages of independent random variables

In the earlier chapters, arithmetic averages of random data (sample means) played a critical role in verifying presence of randomness. In the next few sections, we will see how a similar object can be studied within the rigorous Kolmogorov's probability theory. For that purpose, consider the arithmetic averages

$$T_n = \frac{X_1 + \ldots + X_n}{n}, \qquad n = 1, 2, \ldots, \tag{1}$$

of an infinite sequence of independent random variables $X_1, X_2, \ldots$. The fact that the sequence is infinite, and that the random variables are independent play a critical role in our model.

The influence of multiplication by the constant $1/n$ on the distribution of a random variable is easy to establish. Indeed,

$$F_{\alpha X}(x) = P(\alpha X \le x) = P(X \le x/\alpha) = F_X(x/\alpha), \qquad \alpha > 0. \tag{2}$$

So, in this section we shall concentrate on the study of the numerator in (1), that is, on the partial sums

$$S_n = X_1 + \ldots + X_n, \qquad n = 1, 2, \ldots, \tag{3}$$

of an infinite sequence of independent random variables $X_1, X_2, \ldots$.

Writing the formula for the cumulative probability d.f. of a sum of two independent absolutely continuous random variables X and Y is not difficult: after a change of variables $x = t - y$ in the integral,

$$F_{X+Y}(z) = P(X + Y \le z) = \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f_X(x) f_Y(y)\, dy\, dx \tag{4}$$

$$= \int_{-\infty}^{z} \left(\int_{-\infty}^{\infty} f_X(t - y) f_Y(y)\, dy \right) dt,$$

which implies that the sum $X + Y$ has a density which is the *convolution* of densities of random variables X and Y:

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(z - x) f_Y(x)\, dx \equiv (f_X * f_Y)(z). \tag{5}$$

Then, by induction, for the sum of n independent random summands,

$$f_{X_1+\ldots+X_n}(z) = (f_{X_1} * \ldots * f_{X_n})(z). \tag{6}$$

However, as one may remember from the calculus classes, evaluation of convolutions, especially the multiple convolutions required to obtain densities of sums of more than two independent random variables, is notoriously unpleasant and cumbersome. Their discrete series counterparts are even more tricky. Of course, in some special cases it can, and should, be done with the help of *Mathematica*. In most cases, the method of Laplace transform (or other related methods, like the Fourier transform) are preferable, see Theorem 5.4.5.

Example 5.5.1 Sums of Independent, Uniform Random Variables.
Let X and Y be independent random variables with identical uniform densities

$$f_X(x) = f_Y(x) = \begin{cases} 1, & \text{if } 0 \le x \le 1, \\ 0, & \text{elsewhere.} \end{cases} \tag{6}$$

Then, the sum $X + Y$ has the density

$$f_{X+Y}(z) = \begin{cases} x, & \text{if } 0 \le x \le 1; \\ 2 - x, & \text{if } 1 \le x \le 2; \\ 0, & \text{elsewhere.} \end{cases} \tag{7}$$

The formula is obtained by inserting the definition of $f_X(x)$ into the formula (5) and performing the integration (watch for the limits of the integral).

Example 5.5.2 Gamma Distribution and Sums of Independent Exponential Random Variables.
Gamma distribution is an absolutely continuous distribution with the density

$$f(x;\alpha,\beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \qquad \text{for} \quad x \ge 0, \tag{8}$$

and equal to 0 on the negative half-axis $(x < 0)$. It depends on two parameters, α and β. Substituting $y = x/\beta$ we find

$$\int_0^\infty x^{\alpha-1} e^{-x/\beta} dx = \beta^\alpha \Gamma(\alpha),$$

so that $f(x;\alpha,\beta)$ is indeed a density function (see Section 3.8 for the gamma function calculus). The gamma-distribution with parameters $\alpha = 1, \beta = 1/\lambda$, is simply the exponential distribution with parameter λ.

The expectation and the second moment of a gamma-distributed random variable X are easily calculated:

$$EX = \int_0^\infty x f(x,\alpha,\beta) dx = \frac{\beta\Gamma(\alpha+1)}{\Gamma(\alpha)} \int_0^\infty f(x;\alpha+1,\beta) dx = \alpha\beta, \tag{9}$$

and

$$\int_0^\infty x^2 f(x;\alpha,\beta) dx = \frac{\beta^2\Gamma(\alpha+2)}{\Gamma(\alpha)} \int_0^\infty f(x;\alpha+2,\beta) dx = \alpha(\alpha+1)\beta^2, \tag{10}$$

which, in turn, yields the variance

$$\operatorname{Var} X = \alpha(\alpha+1)\beta^2 - \alpha^2\beta^2 = \alpha\beta^2. \tag{11}$$

If X and Y are independent and have gamma densities with parameters α_1, β and α_2, β, respectively, then their sum $X+Y$ has a gamma-density with parameters $\alpha_1+\alpha_2$, β. Indeed, according to (5),

$$\begin{aligned}
f_{X+Y}(t) &= \int_{-\infty}^{\infty} f_X(t-x) f_Y(x)\,dx \\
&= \frac{1}{\beta^{\alpha_1+\alpha_2}\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^t (t-x)^{\alpha_1-1} x^{\alpha_2-1} e^{-t/\beta}\,dx \\
&= \frac{1}{\beta^{\alpha_1+\alpha_2}\Gamma(\alpha_1)\Gamma(\alpha_2)} t^{\alpha_1+\alpha_2-2} e^{-t/\beta} \int_0^t (1-x/t)^{\alpha_1-1}(x/t)^{\alpha_2-1}\,dx \\
&= \frac{1}{\beta^{\alpha_1+\alpha_2}\Gamma(\alpha_1)\Gamma(\alpha_2)} t^{\alpha_1+\alpha_2-1} e^{-t/\beta} \int_0^1 (1-y)^{\alpha_1-1} y^{\alpha_2-1}\,dy \\
&= \frac{1}{\beta^{\alpha_1+\alpha_2}\Gamma(\alpha_1+\alpha_2)} t^{\alpha_1+\alpha_2-2} e^{-t/\beta},
\end{aligned}$$

where we used the identity[2]

$$\int_0^1 (1-y)^{\alpha_1-1} y^{\alpha_2-1} dy = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}.$$

So, in particular, if $X_1, \ldots, X_n$, are independent exponential random variables with parameter $\lambda = 1/\beta$, then their sum has a gamma density with parameters $\alpha = n$, $\beta = 1/\lambda$, that is

$$f_{X_1+\ldots+X_n}(x) = f(x; n, 1/\lambda). \tag{12}$$

[2] The left-hand side defines a new special transcendental *beta function* $B(\alpha_1, \alpha_2)$ which appears naturally in fractal calculus. For a proof see A.I. Saichev and W.A. Woyczynski, *Distributions in the Physical and Engineering Sciences, Volume 1, Distributional and Fractal Calculus, Integral Transforms and Wavelets*, Birkhäuser-Boston, 1997, Sections 6.8 and 6.9.

Example 5.5.3 Sums of Independent Gaussian Random Variables.
For sums of independent Gaussian random variables, it is easier to use the Laplace transform to determine their distribution and Theorems 5.4.3 and 5.4.4 to determine their expectations and variances. Thus, we get that if $X_1, \ldots, X_n$, are independent random variables with Gaussian distributions with expectations $\mu_1, \ldots, \mu_n$, and variances $\sigma_1^2, \ldots, \sigma_n^2$, respectively, then their sum,

$$S_n = X_1 + \ldots + X_n,$$

is also a Gaussian random variable with expectation

$$E(S_n) = \mu_1 + \ldots + \mu_n, \tag{13}$$

variance

$$\mathrm{Var}\,(S_n) = \sigma_1^2 + \ldots + \sigma_n^2, \tag{14}$$

and the density

$$f_{S_n}(x) = \frac{1}{\sqrt{2\pi(\sigma_1^2 + \ldots + \sigma_n^2)}} \exp\left[-\frac{x - (\mu_1 + \ldots + \mu_n)}{2(\sigma_1^2 + \ldots + \sigma_n^2)}\right]. \tag{15}$$

This result can also be obtained by calculations with the gamma functions, similar to those in Example 5.5.2.

In the following example we shall identify the distribution of a more complicated, quadratic function of independent Gaussian random variables which is of importance in the statistical applications in Chapters 7–9.

Example 5.5.4 Chi-Square Distribution and Sums of Squares of Independent Gaussian Random Variables.
By definition, a random variable Y is said to have a *chi-square distribution* with parameter $n = 1, 2, 3, \ldots$ (which is often called the number of degrees of freedom) if its density is

$$\chi_n^2(x) := \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}, \qquad \text{for} \quad x \geq 0, \tag{16}$$

and 0 on the negative half-line. Obviously, this is the gamma density with parameters $\beta = 2$ and $\alpha = n/2$. Hence, the expectation and the variance are

$$EY = \int_0^\infty x\chi_n^2(x)\,dx = n, \qquad \mathrm{Var}\ Y = 2n. \tag{17}$$

The importance of the χ^2 distribution comes from its relation to the normal distribution. If $X_1, \ldots, X_n$, are independent standard Gaussian random variables (with expectations 0 and variance 1), then the random variable

$$Y = X_1^2 + X_2^2 + ... + X_n^2 \tag{18}$$

has the χ^2 distribution with parameter n. Indeed, this follows immediately from the formula for the density of the sum of independent gamma random variables and from the fact that, for each i, the random variable X_i^2 has a χ^2 density with parameter $n = 1$.

Graphs of the gamma and chi-square densities are shown, using *Mathematica*, in Section 3.8.

5.6 Laws of large numbers and small deviations

One of the first attributes of randomness discussed and confirmed experimentally in the preceding chapters, is the stability of sample means as the sample size becomes larger and larger. In terms of consecutively and independently repeated experiments with outcomes $x_1, \ldots, x_n$, one would like to see the (time) averages

$$\frac{x_1 + \ldots + x_n}{n}$$

converge, as the number of experiments $n \to \infty$. This, in particular, also gives the related desirable attribute of randomness: the stability of relative frequencies.

The goal of this section is to establish rigorously, within the model of Kolmogorov's probability theory, that if $X, X_1, X_2, \ldots$, is a sequence of independent, identically distributed random variables, then

$$\lim_{n\to\infty} \frac{X_1 + \ldots + X_n}{n} = EX, \tag{1}$$

whenever EX is well defined. In other words, the arithmetic "time" averages of the sequence converge to their common expectation, i.e., probability "space" mean (average). In probability theory, results of this type are called the *Laws of Large Numbers (LLNs)*. Of course, we have not yet specified in what sense the convergence in (1) takes place. Here, there are several possibilities and we will explore two of them below.

In our first law of large numbers we will measure the difference of the left-hand and right-hand sides in (1) (i.e., the distance between the "time" average and the "space" average) in terms of the expected mean square distance. For that reason, the next theorem is often called the *Law of Large Numbers in the Mean (Square).*

Theorem 5.6.1 Law of Large Numbers in the Mean.
If $X, X_1, X_2, \ldots$ is a sequence of independent and identically distributed random variables with finite variances, then

$$\lim_{n\to\infty} E\left(\frac{X_1+\ldots+X_n}{n} - EX\right)^2 = 0$$

PROOF In view of Theorem 5.4.4, the variance of the sum of independent random variables is the sum of their variances. Hence,

$$\lim_{n\to\infty} E\left(\frac{(X_1-EX_1)+\ldots+(X_n-EX_n)}{n}\right)^2 = \lim_{n\to\infty}\frac{n\operatorname{Var}X}{n^2} = 0. \qquad (2)$$

∎

Our second law of large numbers asserts that for large n, the probability of even smallest deviations of the "time" average from the "space" average becomes negligible. For that reason, the following result is called the *Law of Large Numbers in Probability*, or, sometimes, the *Weak Law of Large Numbers.*

Theorem 5.6.2 Weak Law of Large Numbers.
If $X, X_1, X_2, \ldots$ is a sequence of independent and identically distributed random variables with finite variances, then, for any $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{X_1+\ldots+X_n}{n} - EX\right| > \epsilon\right) = 0.$$

PROOF In view of the Chebyshev inequality (Theorem 5.4.1), and the calculation (2) in the proof of LLN in the Mean, we have that for any $\epsilon > 0$,

$$P\left(\left|\frac{X_1+\ldots+X_n}{n} - EX_1\right| > \epsilon\right) \le \frac{\operatorname{Var}\left((X_1+\ldots+X_n)/n\right)}{\epsilon^2} = \frac{\operatorname{Var}X}{n\epsilon^2}, \qquad (3)$$

which, for any fixed $\epsilon > 0$, does converge to 0 as $n \to \infty$. ∎

Inequality (3), a direct consequence of Chebyshev's inequality, is very useful in its own right and can be used to obtain various estimates.

Example 5.6.1 One Million Coin Tosses.

Estimate the probability that in one million tosses of a fair coin one obtains between 490,000 and 510,000 heads. The answer can be obtained by an application of the inequality (3). Indeed , take $X_1, \ldots, X_{1,000,000}$, to be independent random variables with identical symmetric Bernoulli distribution

$$P(X = 0) = P(X = 1) = 1/2.$$

Then, $EX = 1/2$ and $\text{Var } X = 1/4$. Hence,

$$\begin{aligned}
&P\Big(490,000 < X_1 + \ldots + X_{1,000,000} < 510,000\Big)\\
&= P\Big(-10,000 < (X_1 + \ldots + X_{1,000,000}) - 500,000 < 10,000\Big)\\
&= P\left(-\frac{1}{100} < \frac{(X_1 + \ldots + X_{1,000,000})}{1,000,000} - \frac{1}{2} < \frac{1}{100}\right)\\
&= P\left(\left|\frac{(X_1 + \ldots + X_{1,000,000})}{1,000,000} - \frac{1}{2}\right| < \frac{1}{100}\right)\\
&= 1 - P\left(\left|\frac{(X_1 + \ldots + X_{1,000,000})}{1,000,000} - \frac{1}{2}\right| \geq \frac{1}{100}\right)\\
&\geq 1 - \frac{1/4}{1,000,000 \cdot (1/100)^2} = \frac{399}{400}.
\end{aligned}$$

In a similar fashion one can answer another, similar question: How many times do we need to toss a fair coin so that the relative frequency of heads approximates 1/2 with accuracy better than .00001 with probability at least .99?

An alert reader must have already noticed that our LLN in the Mean and Weak LLN require that the variances of random variables involved are finite. That is a somewhat more restrictive assumption than the existence of the expectation EX that we hoped for in (1). An even stronger result, called the Strong LLN, and requiring that just $E|X| < \infty$, is indeed true, and can be obtained also via a more complicated theoretical machinery. You can find the proof in any of the probability theory monographs cited in the Bibliographical Notes.

5.7 Central limit theorem and large deviations

Another—you could say, second level—attribute of randomness was discovered experimentally in Section 3.6: the stability, for large sample sizes, of the probability distributions of fluctuations around the sample means. More precisely, we have found out those universal asymptotic distributions to be Gaussian. In this section we will establish this rigorously within the framework of the Kolmogorov model of independent random variables.

In the preceding section we proved the Laws of Large Numbers, which establish that for independent, identically distributed random variables $X, X_1, X_2, \dots,$

$$\frac{X_1 + \dots + X_n}{n} - EX \longrightarrow 0, \qquad \text{as} \quad n \to \infty, \tag{1}$$

where the convergence of the difference between the "time" average and the "space" average to zero was meant either in the mean square (Theorem 5.6.1), or in probability (Theorem 5.6.2). The direct study of the distributions of the differences on the left-hand side of (1) does not promise much; indeed their distributions collapse to 0 since, by LLN in the Mean,

$$\operatorname{Var} \frac{X_1 + \dots + X_n}{n} = \frac{\sigma^2}{n} \to 0, \quad \text{as} \quad n \to \infty, \tag{2}$$

where $\sigma^2 = \operatorname{Var} X$. However, the experiments of Section 3.6 suggest that a proper scaling can produce a nontrivial limit distribution. In our present probability theory model, in view of (2), the correct rescaling of (1) is

$$\begin{aligned} Y_n &= \sqrt{\frac{n}{\sigma^2}} \left(\frac{X_1 + \dots + X_n}{n} - EX \right) \qquad (3) \\ &= \frac{(X_1 - EX_1)/\sigma + \dots + (X_n - EX_n)/\sigma}{\sqrt{n}}, \end{aligned}$$

so that, for all $n = 2, \dots,$ the new rescaled random variables Y_n all have means = 0 and variances = 1. They are the proper object of the fluctuations' distribution study. Indeed, the next result, traditionally called the *Central Limit Theorem (CLT)*, shows that the cumulative probability d.f.s of Y_ns converge, as $n \to \infty$, to the standard Gaussian cumulative probability d.f.

Theorem 5.7.1 Central Limit Theorem.
If $X, X_1, X_2, \dots$ is a sequence of independent and identically distributed random

variables with common expectation μ and variance $\sigma^2 < \infty$, then, for each real number z,

$$P\left(\frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} \le z\right) \to \int_{-\infty}^{z} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx \equiv \Phi(z), \tag{4}$$

as $n \to \infty$.

PROOF We will sketch the proof only in the case when the Laplace transform φ_X is well defined. In view of (3), it suffices to consider the case $\mu = 0$ and $\sigma = 1$. The obvious tool to study the distributions of averages of independent random variables is the Laplace transform introduced in Section 5.4. Actually, it will be more convenient here to use the logarithm of the Laplace transform

$$L(u) := \log \varphi_X(u).$$

Then, by formula (5.4.16),

$$L(0) = 0, \qquad L'(0) = \mu = 0, \qquad L''(0) = 1,$$

and by Theorem 5.4.5, and the calculus' L'Hospital's rule applied twice,

$$\lim_{n\to\infty} \log \varphi_{Y_n}(u) = \lim_{n\to\infty} \frac{L(u/\sqrt{n})}{n^{-1}} = \lim_{n\to\infty} \frac{L'(u/\sqrt{n})u}{2n^{-1/2}}$$

$$= \lim_{n\to\infty} L''(u/\sqrt{n})(u^2/2) = u^2/2.$$

Hence, the limit cumulative probability d.f. $F(x)$ has the Laplace transform $\varphi(u) = e^{u^2/2}$ and the result in Example 5.4.6 identifies it as the standard normal cumulative probability d.f. ▮

Observe that the scaling $1/\sqrt{n}$ in the Central Limit Theorem depends on the assumption that the random variables $X, X_1, X_2, \ldots,$ have finite variance. The alert reader should have also noticed that the proof used a continuity property of the Laplace transforms: The limit cumulative probability d.f. has the Laplace transform that is the limit of Laplace transforms. This analytical result has not been formally established in this book but can be found in sources quoted in the Bibliographical Notes.

One can show, via integration by parts, that the asymptotics of the tails of the standard Gaussian cumulative probability d.f. $\Phi(y)$ is

$$1 - \Phi(y) = \frac{1}{\sqrt{2\pi}} \int_y^\infty e^{-x^2/2} dx \sim \frac{1}{\sqrt{2\pi}} \frac{e^{-x^2/2}}{x}, \tag{5}$$

as $x \to \infty$. Thus, a formal application of the CLT for independent, identically distributed random variables with mean = 0, and variance = 1, suggests that, for the "time" averages Y_n,

$$P(Y_n \geq a_n) \sim \frac{1}{\sqrt{2\pi}} \frac{e^{-a_n^2/2}}{a_n} = e^{-a_n^2(1+b_n)/2}, \tag{6}$$

as long as $b_n \to 0$, and $a_n \to \infty$. Actually, this *large deviation* result can be proved rigorously for discrete random variables with finitely many values, and a_ns such that $a_n/\sqrt{n} \to 0$.

A Mathematical Aside. The Law of the Iterated Logarithm. For more general random variables the situation is more subtle and other, less obvious and intuitive, large deviation results can be found in the literature. A related result, called the *Law of the Iterated Logarithm (LIL),* is quoted below and sometimes used for higher-level tests of randomness.

Theorem 5.7.2 Law of the Iterated Logarithm.

Let $X_1, X_2, \ldots,$ be independent, identically distributed random variables with mean 0 and variance 1. Then

$$P\left(\limsup_{n\to\infty} \frac{X_1 + \ldots + X_n}{\sqrt{2n \ln \ln n}} = 1\right) = 1. \tag{7}$$

Its effects can be seen only for very large n because the iterated logarithm function grows very slowly; indeed, observe that the iterated (decimal) logarithm of 10^{100} is equal to only 2.

The CLT displayed a universal limit law (Gaussian) in the context of independent identically distributed random variables $X_1, X_2, \ldots$. There are many other results of this type (see also simulations with the Cauchy limit distribution in Chapter 3). As another example we shall quote the so-called *Arcsine Law* which describes the limit distribution for the fraction of time the sum $S_n = X_1 + \ldots + S_n$ (with $EX_i = 0$, $\operatorname{Var} X_i = 1$) remains positive:

$$\lim_{n\to\infty} P\left(\frac{\#\{k : S_k > 0\}}{n} \leq x\right) = \begin{cases} 0, & \text{for } x \leq 0; \\ \frac{2}{\pi} \arcsin \sqrt{x}, & \text{for } 0 \leq x \leq 1; \\ 1, & \text{for } x \geq 1. \end{cases} \tag{8}$$

This limit law is also often used in high-level tests of *pseudorandom number generators*.

5.8 Experiments, exercises, and projects

1. How many times do you need to roll a die to guarantee that the frequency of 5s is within ±0.01 of 1/6, with probability more than 9/10?

2. A closet contains 12 pairs of shoes. If 6 shoes are randomly selected, what is the probability that there will be

   a) no complete pair;

   b) exactly one complete pair;

   c) exactly two complete pairs?

3. If the odds are 5 to 3 that event M will not occur, 2 to 1 that event N will occur, and 4 to 1 that they will not both occur, are the two events M and N independent? This exercise introduces the notion of the odds.

4. Using the Chebyshev inequality, construct a table showing the upper bounds for the probabilities that a random variable differs from its mean by at least 1,2, and 3 standard deviations. Find the corresponding exact probabilities for the binomial distribution with $n = 16$ and $p = 1/2$.

5. Explain the connection between the appearance of the binomial coefficient $\binom{n}{k}$ in the binomial distribution, and the coefficients in the expansion of the binomial $(a+b)^n$, generalizing the familiar formula $(a+b)^2 = a^2 + 2ab + b^2$.

6. To illustrate the Law of Large Numbers, find the probabilities that the proportion of heads will be anywhere from 0.49 to 0.51, when a balanced coin is flipped (a) 1,000 times, (b) 10,000 times.

7. Find the mean and the standard deviation of a random variable X, with gamma distribution with parameters $\alpha = 2$ and $\beta = 2$. Recall that

$$f(x) = \begin{cases} (\beta^\alpha \Gamma(\alpha))^{-1} x^{\alpha-1} e^{-x/\beta} & \text{for } x > 0; \\ 0 & \text{elsewhere,} \end{cases}$$

   and $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$.

8. Suppose that events A, B, and C are independent (that is, the multiplicative law holds for all pairs and triples) with probabilities 1/2, 1/3, and 1/5, respectively. Find:

(a) $P(A \cap B \cap C)$,

(b) $P(A \cup B \cup C)$,

(c) P(exactly one of the three events occurs).

9. A pollster wishes to know the percentage p of people in a population to vote for candidate Z. How large must a random sample be in order to be 99% sure that the sample percentage is within 1% of p?

10. Let X, Y be independent, discrete random variables uniformly distributed on $1, 2, \ldots, n$. Find:

(a) $P(X = Y)$,

(b) $P(X \leq Y)$,

(c) $P(\min(X, Y) = k)$, for $k = 1, 2, \ldots, n$.

11. What is the expected number of 6s appearing on three die rolls? What is the expected number of odd numbers?

12. A random variable X has expectation 10 and standard deviation 5.

(a) Find the smallest upper bound for $P(X \geq 20)$.

(b) Could X be a binomial random variable?

13. In a certain population 5% of the people are poor, 1% are downtrodden and 0.1% are poor and downtrodden. Find the:

(a) probability that a person is not poor,

(b) probability that a person is poor but not downtrodden,

(c) probability that a person is either poor or downtrodden.

14. The four major blood types are present in the following proportions in the population of the U.S.: A —42%, B—10%, AB—4%, and O—44%. These are all separate types. If two people are picked at random, what is the chance that their blood is of the same type? Of different types?

15. Suppose that each week you buy an Ohio Lottery ticket which gives you a chance of one in ten million of a win. What is the chance that you get $k = 0, 1, 2$ wins during the year?

16. Suppose that each of 300 patients has the probability of 1/3 of being helped by a treatment independent of its effects on the other patients. Find approximately the probability that more than 120 patients are helped by the treatment.

17. A hat contains n coins, f of which are fair and b of which are biased to land heads with probability 2/3. A coin is drawn from the hat and tossed twice. The first time it lands heads, and the second time it lands tails. Given this information, what is the probability that it is a fair coin?

18. In the World Series, the Cleveland Indians and the Atlanta Braves play until one team wins four games. Suppose that all games are independent, and that in each game the probability that the Indians beat the Braves is 2/3.

(a) Find the probability that the Indians win in four games.

(b) Find the probability that the Indians win the World Series given that the Braves won Games 1 and 2.

19. Suppose that X, Y are two independent random variables with the same density function $f(x) = x \exp(-x^2/2)$. Find:

(a) the density of $Z = \min(X, Y)$;

(b) EZ^2.

20. For X, Y independent and uniformly distributed on $[-1, 1]$ find:

(a) $P(|X + Y| \leq 1)$,

(b) $E|X + Y|$.

21. Suppose that X, Y have the joint density function

$$f_{(X,Y)}(x, y) = \begin{cases} C/x^3, & \text{for } x > y > 1, \\ 0, & \text{otherwise,} \end{cases}$$

where C is a constant.

(a) Find C;

(b) Find the marginal density of X.

22. Let X, Y by independent with the Laplace density $f_X(x) = f_Y(x) = \alpha e^{-\beta|x|}$. Given β, find α and then find $f_{X+Y}(x)$.

23. Use the CLT, the Arcsine Law, and the Law of the Iterated Logarithm to test pseudorandom number generators of your choice.

24. Use *Mathematica*'s commands `LaplaceTransform[expr, x, u]`, `FourierTransform[expr, x, u]`, and the `InverseLaplaceTransform[expr, x, u]`, `InverseFourierTransform[expr, x, u]` commands to calculate the probability d.f.s of the following sums of independent random variables. The Fourier transform commands are in the package `Calculus'FourierTransform'`. Produce the graphs (for specific ns) of their probability d.f.s, and cumulative probability d.f.s.

(a) Sum of 3, 4, 5, 10, and, in general, n independent random variables uniformly distributed on the interval [0,1].

(b) Sum of 3, 4, 5, 10, and, in general, n independent random variables with the Cauchy distribution.

(c) Sum of 3, 4, 5, 10, and, in general, n independent random variables with the chi-squared distribution with the 2 degrees of freedom.

(d) Sum of the independent standard normal and the exponential (with parameter 1) random variables.

25. *Conditional Probabilities and Bayes' Formula.* If a random event B has positive probability P, then the *conditional probability* of a random event A given B is defined by the formula $P(A|B) = P(A \cap B)/P(B)$.

(a) Prove that if random events $B_1, B_2, \ldots, B_n$ form a partition of the sample space Ω, i.e., $B_1 \cup B_2 \cup \ldots \cup B_n = \Omega$, $B_i \cap B_j = \emptyset$, $i \neq j, i, j = 1, \ldots, n$, then, for any random event A,

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i).$$

This formula is known as the *total probability* formula.

(b) Under the same assumption as in (a), prove the *Bayes formula* for reverse conditional probabilities: For each $i = 1, \ldots, n$,

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A|B_1)P(B_1) + \ldots + P(A|B_n)P(B_n)}.$$

(c) A channel transmits binary symbols 0 and 1 with random errors. The probability that the symbols 0 and 1 appear at the input of the channel are, respectively, 0.45 and 0.55. Given that the symbol 0 was transmitted, the probability of receiving 0 is equal to 0.95. For the symbol 1 the analogous conditional probability is 0.9. Find the probability that the symbol 1 was transmitted, given that the symbol 1 was received.

26. *Weibull distribution and related extremal distributions.* Suppose $T_1, \ldots, T_n$, are independent real-valued random variables. Define two new random variables by setting

$$T_{\min} = \min_{1 \le i \le n} T_i \quad \text{and} \quad T_{\max} = \max_{1 \le i \le n} T_i.$$

Such random quantities play a role in many practical situations, for example, in reliability problems. If a device consists of n components in series, and all of them have to function for the whole device to function, then the time of failure of the device is the minimum of times to failure of the individual components. If the components are in parallel (redundant) then the corresponding time is the maximum of individual failure times. A landing gear of an aircraft consists of several components which are subject, during each ground-to-air cycle, to stresses caused by the dynamic loads during taxing, take-off, landing impact, and landing runs. The stresses are random in nature due to the runway unevenness, and wind conditions, and together with non-homogeneity of the material and manufacturing processes they cause crack formation at random times. If we consider random variables describing the times of the first crack appearance in each component of the landing gear, then the time of the first crack appearance in the whole landing gear (of importance, for example, in determining the time intervals between inspections) is the minimum value of these variables.

If T_i's are identically distributed with the cumulative distribution function $F(t)$, then

$$\Pr(T_{\max} \leq t) = \Pr(T_i \leq t \text{ for all } i) = F^n(t).$$

It turns out that for large n, under some technical conditions on F, a version of the *Central Limit Theorem* holds true, and the distribution of $T_{\max}$ converges to the *Weibull distribution*

$$f(t, \alpha, \beta) = \alpha\beta(t/\beta)^{\alpha-1}\exp[-(t/\beta)^{\alpha}], \qquad t \geq 0,$$

where α is the shape parameter and β is the scale parameter. This is one of the so-called *extremal distributions*.

Find the expectation, the variance, and the moment generating function of a random variable with the Weibull distribution.

27. For random variables X, Y, with the joint density function

$$f(x, y) = \begin{cases} x + y, & 0 < x, y < 1; \\ 0, & \text{elswhere,} \end{cases}$$

find the correlation coefficient between X and Y.

28. Let $P(X = x_1) = 1 - P(X = x_2) = p < 1$ and $P(Y = y_1) = 1 - P(Y = y_2) = q < 1$. Show that X and Y are independent if and only if the correlation coefficient between them is zero.

5.9 Bibliographical Notes

The rigorous mathematical foundations for probability theory were laid in a brief treatise

[1] A.N. Kolmorogov, *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Springer-Verlag, Berlin, 1933.

A two-volume

[2] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. I and II, Wiley, New York, 1968 (3rd ed.) and 1971 (2nd ed.),

has become a classic, and still reads extremely well. More modern, and rigorous but concise expositions are

[3] L. Breiman, *Probability*, Addison-Wesley, Reading, MA, 1968.

[4] P. Billingsley, *Probability and Measure*, Wiley, New York, 1986.

Recent monographs with an abstract, mathematical orientation and complete expositions are

[5] R.M. Dudley, *Real Analysis and Probability*, Chapman and Hall, New York, 1989.

[6] S. Kwapień and W.A. Woyczyński, *Random Series and Stochastic Integrals: Single and Multiple*, Birkhäuser-Boston, Cambridge, MA, 1992.

The former contains extensive and meticulous historical commentaries. The latter concentrates on the behavior of sums of independent random variables and their continuous counterparts, stochastic integrals.

Chapter 6

Chaos in Dynamical Systems: How Uncertainty Arises in Scientific and Engineering Phenomena

The two preceding chapters analyzed the phenomenon of randomness from the viewpoint of algorithmic and computational complexity of a fixed string of data, and in the context of the formal mathematical probability theory based on Kolmogorov's concept of a sequence of statistically independent random variables. We complete this picture in the present chapter by demonstrating that certain, seemingly deterministic, dynamical systems also exhibit some attributes of randomness such as stability of frequencies and fluctuations. The essential features here are *nonlinearity* and/or *sensitive dependence on initial conditions.*

6.1 Dynamical systems: general concepts and typical examples

We shall begin with an introduction of some basic terminology and examples of dynamical systems. The former will facilitate our more rigorous discussion of the behavior of such systems, and the latter will provide motivation for their study. Several real-life examples of dynamical systems were mentioned in Chapter 1 (water dripping from a faucet, turbulent flows, etc.). Others were experimented with in computer simulations (billiards). On the other hand, some of the examples described below have a character of greatly simplified cartoons of real systems. Their value is in their transparency and the relative ease of their analysis. Although simple, they are not simple-minded; it took experts quite a bit of time to discover some of them, or to realize their usefulness.

The dynamical system evolves on a *state space* which we will denote by S, with individual states denoted by s, x, y, etc. The state space S can be very simple and very small; for a single coin toss, it can consist of two states, 0 and 1. For a complex system, such as a gas of molecules, it can be a huge ca. $6 \cdot 10^{23}$-dimensional

M. Denker, W. Woyczynski, *Introductory Statistics and Random Phenomena*,
Modern Birkhäuser Classics, DOI 10.1007/978-3-319-66152-0_6

Euclidean space, see Examples 4.4.1 and 5.1.5. In studies of physical systems one often selects the state space to coincide with the phase space, but the same physical system may be given different dynamical system descriptions depending on the level of accuracy desired. For example, the coin toss could be described also as a motion of the rigid 3-dimensional body in a rich phase space with its several degrees of freedom, rather then as a simplistic evolution in a two-point state space. At a superficial level, the state space S also resembles the sample space Ω introduced in the Kolmogorov probability theory to label possible outcomes. Their roles are, however, different; for the sample space Ω, its structure was of little consequence, any other labeling would suffice, and the unit interval could always serve as a universal sample space. What mattered was the distribution of a random variable. On the other hand, the structure of a dynamical system's state space is of great significance; the systems on a unit interval may display behavior very different from those on a two-dimensional torus.

In this chapter we will discuss mainly dynamical systems with *discrete time* $t = 0, 1, 2, \ldots$ As a description of real-life phenomena it is often a simplification, although one can argue that data collected via computerized measuring tools are always discrete-time. Evolution of a system with the state space S is determined by a mapping (function)

$$f : S \ni s \longmapsto f(s) \in S, \tag{1}$$

which, colloquially, is often referred to as the "dynamics" of the system. In one time-step, the system, originally in state s, is transformed into another state $f(s)$ in the same state space S. This process is then iterated to determine the future states of the system at discrete times. So, starting from $s \in S$, we can observe time evolution of the initial state s under consecutive iterations of the map f:

$$s \quad \overset{f}{\longmapsto} \quad f(s) \quad \overset{f}{\longmapsto} \quad f(f(s)) \quad \overset{f}{\longmapsto} \quad f(f(f(s))) \quad \overset{f}{\longmapsto} \ldots. \tag{2}$$

For simplicity, we shall denote the n-th iteration of the function $f(s)$ by $f^n(s)$ (not to be confused with the n-th power of a real-valued function). The *orbit* of the initial state s in the system, whose dynamics is determined by the mapping f, is the sequence of states

$$s_0 = s, \quad s_1 = f(s), \quad s_2 = f^2(s), \quad \ldots, s_n = f^n(s), \ldots. \tag{3}$$

Fig. 6.1.1 visualizes evolution of the dynamical system.

Discrete vs. continuous-time systems. The dynamics of a discrete dynamical system can also be described equivalently via its increments:

$$x_n - x_{n-1} = f(x_{n-1}) - x_{n-1}. \tag{4}$$

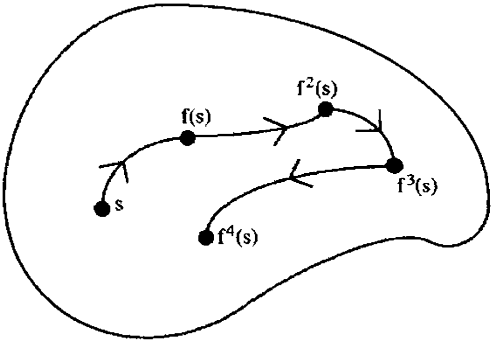


FIGURE 6.1.1
A schematic representation of the orbit of the initial state s, generated by the map f on the state space S.

Denoting

$$\Delta x_n = x_n - x_{n-1}, \qquad \text{and} \qquad F(x) = f(x) - x, \tag{5}$$

evolution of our discrete dynamical system can be described by the difference equation

$$\Delta x_n = F(x_{n-1}) \tag{6}$$

which, if we denote the unit (in our case) time increment $\Delta n = n - (n - 1) = 1$, takes the form

$$\frac{\Delta x_n}{\Delta n} = F(x_{n-1}). \tag{7}$$

This description provides a connection with the usual description of continuous-time dynamical systems via differential equations. Replacing, if possible, the discrete time n by the continuous time t permits one to pass to the limit $\Delta n = \Delta t \to 0$ in the equation (7), which leads to the differential equation

$$\frac{dx(t)}{dt} = F(x(t)) \tag{8}$$

with the initial condition $x(0) = x_0$. If the dynamics is permitted to change with time (a similar generalization is, of course, possible in discrete time as well), then the equation (8) takes the form

$$\frac{dx(t)}{dt} = F(x(t), t), \tag{9}$$

familiar from physics and engineering courses. On the other hand, for numerical purposes, the continuous-time differential equation (9) are routinely approximated by discrete difference schemes

$$x(t+h)-x(t)\approx F(x(t),t)\cdot h, \tag{10}$$

with the time-step h.

For a discrete system, the following two types of problems are of interest:

(1) Determine the system's behavior a few time-steps ahead into the future, assuming that its recent past is known. In other words, given $x_0, x_1, ..., x_{n-1}$, find the behavior of $x_n, x_{n+1}, ..., x_{n+k}$, for small values of k. Often this problem boils down to finding good approximations for otherwise cumbersome formulas.

(2) Determine the system's behavior in the far future, assuming that its whole, long past is known. In other words, for large values of k and n, given $x_0, x_1, ..., x_{n-1}$, find the behavior of x_{n+k}. Here, two basic phenomena can be observed. Either the system is relatively *stable* and small changes of the initial conditions lead to small changes in large-time behavior, or the system is *chaotic* and small variations in the past behavior may lead to large changes in the future behavior. In the first case, reasonable deterministic predictions about the future can be made; in the second case, they are practically impossible and one has to resort to statistical tools.

Before proceeding with the theoretical analysis any further, let us take a look at a number of examples.

Example 6.1.1 A Finite-Dimensional Linear Map.

Here, the state space S is the d-dimensional Euclidean space $\mathbf{R}^d$, with states s represented by d-dimensional vectors (points) $\boldsymbol{x}=(x_1,x_2,\dots,x_d)$. A map $f:\mathbf{R}^d\to\mathbf{R}^d$ is said to be linear if, for any real numbers α,β, and any vectors $\boldsymbol{x}$, $\boldsymbol{y}$,

$$f(\alpha\boldsymbol{x}+\beta\boldsymbol{y})=\alpha f(\boldsymbol{x})+\beta f(\boldsymbol{y}). \tag{11}$$

A linear map f on $\mathbf{R}^d$ always has a representation of the form

$$f(\boldsymbol{x})=A\boldsymbol{x}', \tag{12}$$

where $A=(a_{ij})_{1\le i,j\le d}$, is a $d\times d$ matrix, $\boldsymbol{x}'$ is the transpose of vector $\boldsymbol{x}$, and $A\boldsymbol{x}'$ stands for the matrix multiplication. In other words, if $\boldsymbol{y}=A\boldsymbol{x}'$, then its coordinates

$$y_i=\sum_{k=1}^{d}a_{ik}x_k, \qquad i=1,2,...,d. \tag{13}$$

Suppose, for instance, that $d = 2$ and

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

Then the point $\boldsymbol{x} = (1/2, 1/5)$ is mapped to $\boldsymbol{y} = (9/10, 7/10)$. The iterations of the linear map $f(\boldsymbol{x}) = A\boldsymbol{x}'$, which determine the orbit of the system, correspond to the matrix multiplication, so that

$$\boldsymbol{x}_n = f^n(\boldsymbol{x}_0) = A^n \boldsymbol{x}_0', \qquad n = 0, 1, 2, \ldots, \tag{14}$$

where A^n is the n-th power of matrix A.

Example 6.1.2 Rotation of the Unit Circle.
The state space $\mathbf{T}$ is here the unit circle $\{(x, y) : x^2 + y^2 = 1\}$ in the plane. It is convenient to write it in the complex-plane form

$$\mathbf{T} = \{z \in \mathbf{C} : z = e^{i\theta}, \theta \in [0, 2\pi)\} = \{z \in \mathbf{C} : |z| = 1\}, \tag{15}$$

where $\mathbf{C} \ni z = x + iy$, $i = \sqrt{-1}$. The map f corresponds to the rotation of the circle at the fixed angular rate α per time-step (see Fig. 6.1.2).
In other words,

$$f(e^{i\theta}) = e^{i(\theta+\alpha)}. \tag{16}$$

Its iterations, and thus the orbit of a given starting point on the unit circle, are easily determined since

$$f^n(e^{i\theta}) = e^{i(\theta+n\alpha)}, \qquad n = 0, 1, 2, \ldots. \tag{17}$$

They track the trajectory of the initial point $x_0 = e^{i\theta}$ as it is rotated counterclockwise at the angular velocity of α radians per unit time-step.

Example 6.1.3 Rotation As a Map of the Unit Interval.
Rotation of the unit circle can be written as a map of the state space $S = [0, 1)$, where one thinks of the unit interval as being wrapped around, so that its endpoints, 0 and 1, are identified. In this setting, the rotation-of-the-circle map corresponds to the additive shift transformation, modulo 1:

$$f(x) = (x + a) \pmod 1 := (x + a) - \lfloor x + a \rfloor, \qquad x \in [0, 1], a \in \mathbf{R}, \tag{18}$$

where $\lfloor b \rfloor$ is the integer part of a real number b. A plot of this map is shown on Fig. 6.1.3.

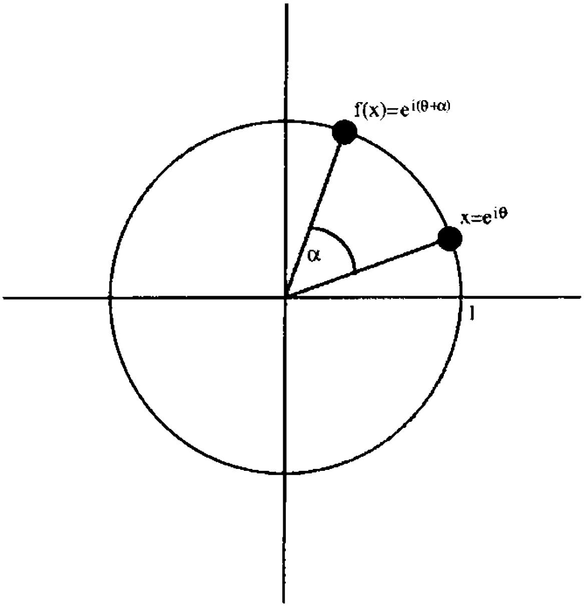


FIGURE 6.1.2
Rotation map of the unit circle.

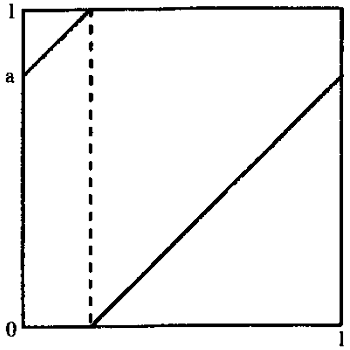


FIGURE 6.1.3
A plot of the additive shift map, modulo 1, on the state space $S = [0, 1)$.

Example 6.1.4 Multiplication Modulo 1, and Independent Bernoulli Random Variables.

The state space S is again the unit interval $[0, 1)$, and the map is defined by the formulas

$$f(x) = 2x \pmod 1) = \begin{cases} 2x, & \text{if } 0 \le x < 1/2; \\ 2x - 1, & \text{if } 1/2 \le x < 1. \end{cases} \tag{19}$$

A plot of this map is shown in Fig. 6.1.4.

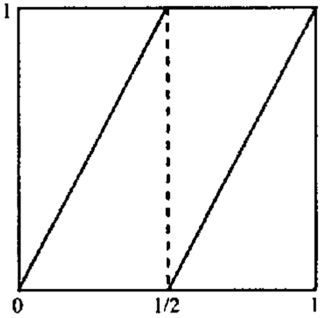


FIGURE 6.1.4
A plot of the multiplication-by-2, modulo 1, map of the unit interval.

It is relatively easy to determine iterations of this map. Indeed, the second iteration, see Fig. 6.1.5,

$$f^2(x) = f(f(x)) = \begin{cases} 4x, & \text{if } 0 \le x < 1/4; \\ 4x - 1, & \text{if } 1/4 \le x < 2/4; \\ 4x - 2, & \text{if } 2/4 \le x < 3/4; \\ 4x - 3, & \text{if } 3/4 \le x < 1. \end{cases} \tag{20}$$

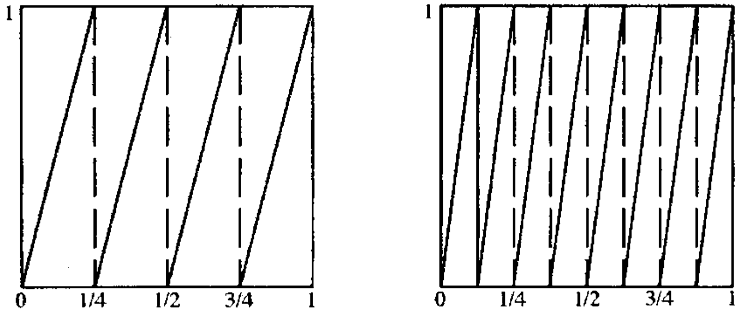


FIGURE 6.1.5
Graphs of the second ($n = 2$) and third ($n = 3$) iterations of the multiplication-by-2, modulo 1, map of the unit interval.

The n-th iteration clearly produces a function with the slope equal to 2^n, "wrapped around" the unit interval 2^n times. Rather than writing the explicit formulas for $f^n(x)$, we produce their graphs for $n = 2$, and 3, see Fig. 6.1.5.

These graphs should bring back memories of independent Bernoulli random variables considered in Example 5.3.3. Indeed, let us *test* our dynamical system (19) via the *test function*

$$\phi(x) = \begin{cases} 0, & \text{if } 0 \leq x < 1/2; \\ 1, & 1/2 \leq x < 1; \end{cases} \tag{21}$$

which simply probes if the state of our dynamical system is above the level 1/2 or below it. Plots of the tested via ϕ successive states of the orbit of (19),

$$\phi(x), \quad \phi(f(x)), \quad \phi(f^2(x)), \quad \phi(f^3(x)), \ldots \tag{22}$$

are shown in Fig. 6.1.6.

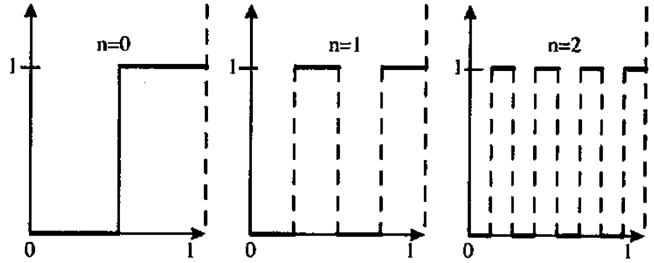


FIGURE 6.1.6

Graphs of the tested iterations of the multiplication-by-2-modulo-1 map of the unit interval.

They are obviously identical with plots of the independent Bernoulli random variables on the standard sample space $\Omega = [0, 1]$ considered in Example 5.3.3. In light of the results of Chapter 5, this is the first indication that one could expect to see some randomness effects in a purely deterministic dynamical system. Also, the notion of a test function introduced above will be useful later on. The particular test function $\phi(x)$ is also called the *indicator function* of the set [0, 1/2) and denoted $1_{[0,1/2)}(x)$.

Example 6.1.5 Shifting Binary Strings and Independent Bernoulli Random Variables.

In this example, the state space S consists of all infinite binary strings

$$\boldsymbol{x} = (x_1, x_2, \ldots), \tag{23}$$

with the digits $x_i = 0$, or 1. The map f consists of shifting all the digits to the left and dropping the left-most one. More precisely

$$f(\boldsymbol{x}) = (x_2, x_3, \ldots). \tag{24}$$

The graphical representation of this dynamical system becomes easy if one remembers that one can identify each string $\boldsymbol{x} = (x_1, x_2, \ldots)$, with the real number

$$x = \frac{x_1}{2^1} + \frac{x_2}{2^2} + \frac{x_3}{2^3} + \ldots \tag{25}$$

from the interval $[0, 1]$. In other words, the string (23) is a binary representation of the number x from (25). In this interpretation it becomes immediately clear that the shift map is identical with the multiplication-by-2, modulo 1, map of Example 6.1.4. Testing it again with the indicator function $\mathbf{1}_{[1/2,)}(x)$ brings back independent Bernoulli random variables.

Example 6.1.6 The Logistic Function in Population Dynamics.
The state space here is the unit interval $S = [0, 1]$ and for a point $x \in [0, 1]$

$$f(x) = ax(1 - x). \tag{26}$$

As long as $0 \leq a \leq 4$, the quadratic function f maps the unit interval into itself, and its graph is shown in Fig. 6.1.7 together with a sample orbit (notice a convenient way of constructing graphically orbits of maps of the interval).

The system describes a typical behavior of the growth rate in population dynamics. The population size is assumed not to exceed a certain maximal value M. The quantity x represents the size of the population as a fraction of the maximum population, hence $x \in [0, 1]$. If the size of the population is small, its growth is not restricted by the food supply, space, and other environmental limitations. After the population reaches its maximal size, the same environmental factors cause the rate of growth to decline.

Mathematica Experiment 1. Iterations of Logistic Function. The iterates $f^n(x)$ of the function (26) are obviously polynomials of degree $2n$, and can be drawn easily using *Mathematica*.

```
In[1]:= f[a_,x_]:=a*x*(1-x)
In[2]:=Plot[{f[2,x], f[3,x], f[3.75,x], f[4,x]}, {x,0,1},
          Frame->True, AspectRatio->1, GridLines->Automatic]
Out[2]= -Graphics-
```

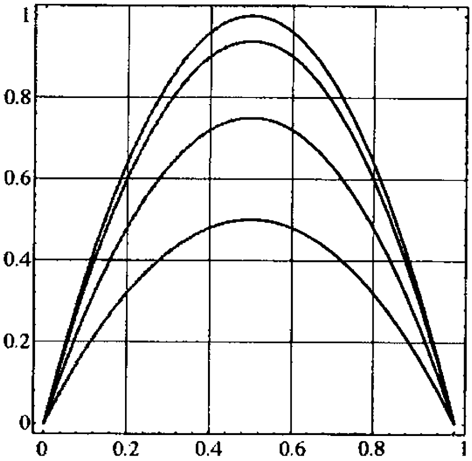
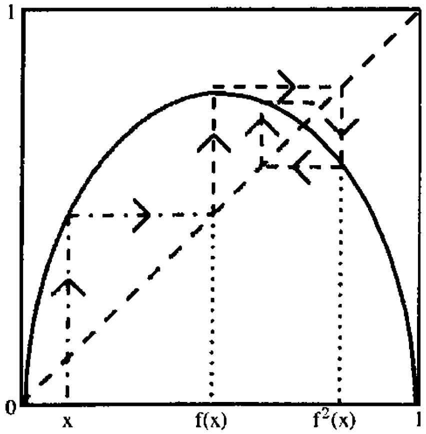
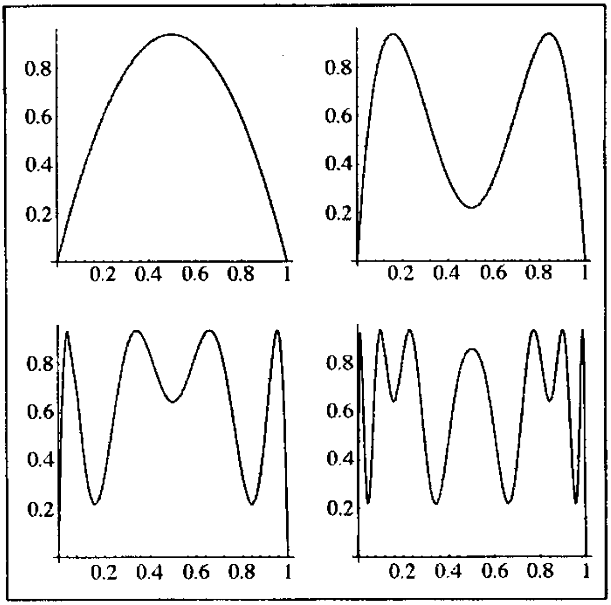


FIGURE 6.1.7
A plot of the logistic map $f(x) = ax(1 - x)$, and its sample orbit.



```
In[3]:= f2[a_,x_]:=f[a,f[a,x]]
In[4]:= f3[a_,x_]:=f[a,f2[a,x]]
In[5]:= f4[a_,x_]:=f[a,f3[a,x]]
In[6]:= p1=Plot[f[3.75,x],{x,0,1}, AspectRatio->1]
In[7]:= p2=Plot[f2[3.75,x],{x,0,1}, AspectRatio->1]
In[8]:= p3=Plot[f3[3.75,x],{x,0,1}, AspectRatio->1]
In[9]:= p4=Plot[f4[3.75,x],{x,0,1}, AspectRatio->1]
In[10]:= Show[GraphicsArray[{{p1,p2}, {p3,p4}}],
          Frame->True, FrameTicks->None]
Out[10]= -GraphicsArray-
```



Example 6.1.7 Anemia in Rabbits.

A laboratory study of hemolytic anemia in rabbits called for a model of the time-evolution of the number x_n of red blood cells per unit blood volume. Initially, J. Williams proposed[1] that the logistic model (26) be used with an experimentally fitted parameter a. Although this model correctly reflected some features of the system's evolution, it did not lead to a good prediction of experimental red blood cell counts. A better model, described by the discrete dynamics

$$x_{n+1} = (1 - L)x_n + Kx_n^s e^{-x_n}, \tag{27}$$

where L, K, and s are certain parameters, has been proposed by Andrzej Lasota in 1977. A particular example of the function governing Lasota's dynamics is shown in Fig. 6.1.8.

Continuous-Time Dynamics in Physical Systems. In the following few examples the continuous-time dynamics is provided by differential equations.

Example 6.1.8 A Particle in the 3-Space.

A particle is moving in the 3–dimensional space with velocity

$$v(t) = (u(t), v(t), w(t))$$

[1] *Journal of Mathematical Biology* 3(1976), 1–5.

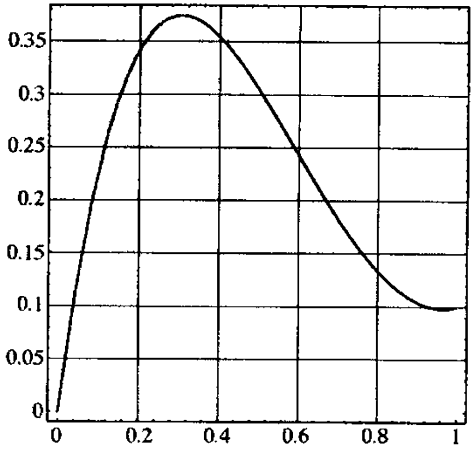


FIGURE 6.1.8

A special case of Lasota's dynamics for hemolytic anemia in rabbits corresponding to $f(x) = 3x - 2.9x^2 \exp(1 - x)$.

at time t. If we denote its position at time t by

$$\boldsymbol{x}(t) = (x(t), y(t), z(t)),$$

then $\boldsymbol{x}(t)$ and $\boldsymbol{v}(t)$ are tied together by the differential equation

$$\dot{\boldsymbol{x}} = \boldsymbol{v}(t). \tag{28}$$

For each initial position (x_0, y_0, z_0), its solution is given by the formulas

$$x(t) = x_0 + \int_0^t u(s)\,ds,$$

$$y(t) = y_0 + \int_0^t v(s)\,ds,$$

$$z(t) = z_0 + \int_0^t w(s)\,ds.$$

In particular, after time $t = 1$, its new position is

$$f((x_0, y_0, z_0)) = (x(1), y(1), z(1)),$$

and the above function f defines a map

$$f : \mathbf{R}^3 \to \mathbf{R}^3.$$

The velocity function $\boldsymbol{v}$ could also depend on the location, that is $\boldsymbol{v} = \boldsymbol{v}(t, \boldsymbol{x})$, defining a more general dynamical system. The corresponding nonlinear differential equation

$$\dot{\boldsymbol{x}} = \boldsymbol{v}(t, \boldsymbol{x}) \tag{29}$$

is, however, much harder to solve, so that, in general, we can define the corresponding map f only implicitly. As a matter of fact, in certain cases, the last equation may have no solutions at all.

Example 6.1.9 Hamiltonian Systems in Classical Mechanics.
A major program of study of physical dynamical systems was launched between 1892 and 1897 by French mathematician Henri Poincaré in connection with his work on celestial mechanics. For the past 100 years, those investigations had a major impact on development of the theory of dynamical systems. Simple celestial mechanics models describing the motion of planets around the sun, go back to Isaac Newton and Johannes Kepler. Newton's second law of dynamics says that a force **F** acting on a moving particle with mass m is directly proportional to its acceleration. This law is expressed by a second order differential equation

$$\boldsymbol{F} = m\frac{d^2\boldsymbol{x}}{dt^2}, \tag{30}$$

where $\boldsymbol{x} = (x_1, x_2, x_3)$ denotes the particle's position. If the particle is subject to a gravitation force field exerted by another material point of mass M, where, say, M is much larger than m, and located at the origin of the coordinate system, then

$$\boldsymbol{F} = -g\frac{mM\boldsymbol{x}}{||\boldsymbol{x}||^3}, \tag{31}$$

where $||\boldsymbol{x}||$ is the distance of the particle from the origin. The above two vector equations lead to three scalar differential equations:

$$\frac{d^2x_1}{dt^2} = -gM\frac{x_1}{||x||^3},$$

$$\frac{d^2x_2}{dt^2} = -gM\frac{x_2}{||x||^3},$$

$$\frac{d^2x_3}{dt^2} = -gM\frac{x_3}{\|x\|^3},$$

which have solutions describing the particle's motion along a conical curve (parabola, ellipse, or hyperbola). This idealized model, applied to the real planetary system (like the Earth moving in the Sun's gravitational field), provides a good approximation of what really happens only on short astronomical time-scales. To obtain good long-range predictions, one has to adjust the model to take into account gravitational fields of other planets, influence of other masses inside our Galaxy, etc. This leads to a much more complex system of differential equations describing the so-called *n-body problem* of classical mechanics. At this stage of our knowledge we do not know how to solve it, or even how to decide if it is stable. The n-body problem is one of the most important outstanding problems in astrophysics and the theory of dynamical systems.

Notice that the above model of Newtonian motion in the central gravitational force field can be rewritten as follows:

$$\frac{d}{dt}(m\dot{\boldsymbol{x}}) + \frac{\partial}{\partial \boldsymbol{x}}E_p(\boldsymbol{x}) = 0, \tag{32}$$

where

$$E_p(\boldsymbol{x}) = -gmM\frac{1}{\| \boldsymbol{x} \|}$$

is the potential energy of the gravitational force field, and

$$\frac{\partial}{\partial \boldsymbol{x}} = \left(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \frac{\partial}{\partial x_3}\right)$$

is the gradient operator. If we denote the kinetic energy of the particle by

$$E_k(\dot{\boldsymbol{x}}) = \frac{1}{2}m\|\dot{\boldsymbol{x}}\|^2, \tag{33}$$

then the Newton equation of motion can be rewritten, one more time, in terms of the so-called *Lagrangian*

$$L = L(\boldsymbol{x}, \dot{\boldsymbol{x}}) = E_k - E_p, \tag{34}$$

to obtain

$$\frac{d}{dt}\left(\frac{\partial}{\partial \dot{\boldsymbol{x}}}L\right) - \frac{\partial}{\partial \boldsymbol{x}}L = 0. \tag{35}$$

The motion minimizes the *action functional* $\int L(\mathbf{x}, \dot{\mathbf{x}}, t)dt$. So, if we introduce the system's *Hamiltonian*

$$H(p, q, t) = p\dot{q} - L(q, \dot{q}, t) = E_k + E_p, \tag{36}$$

where q and p are interpreted as a generalized position $q = x$ and momentum $p = m\dot{x}$, then the Newton equation can be written as a pair of equations

$$\frac{dq}{dt} = \frac{\partial H}{\partial p},$$

$$\frac{dp}{dt} = -\frac{\partial H}{\partial q},$$

which is known as the system of *Hamiltonian equations*. As it turns out, very general dynamical systems are governed by equations of this type. They are called Hamiltonian systems and their general solutions are not known except in very special cases. In most practical cases, a numerical approximation is used. If a Hamiltonian system has a solution, then it generates a dynamical system by setting $f^t(x_0)$ to be the position of the solution after time t assuming that the initial position at time $t = 0$ was x_0.

6.2 Orbits and fixed points

Evolution of the dynamical system generated by the map $f : S \mapsto S$, and starting at the state $x_0 \in S$, is described by its *orbit*

$$x_0, \quad x_1 = f(x_0), \quad x_2 = f(x_1), \quad \dots \quad , x_n = f(x_{n-1}), \dots, \tag{1}$$

which is a sequence of consecutive states of the system in the state space S. In this section we will take a look at different types of orbits and their various properties.

Example 6.2.1 Rectilinear Motion.
Consider a map

$$f((x_0, y_0, z_0)) = (x_0 + u, y_0 + v, z_0 + w), \tag{2}$$

on the three-dimensional state space $\mathbf{R}^3 \ni \boldsymbol{v} = (u, v, w)$, which describes a rectilinear discrete-time motion of the particle, with velocity $\boldsymbol{v}$. It is the special case of a general linear map of Example 6.1.1. Iteration of this map (see Fig. 6.2.1) yields orbit points

$$f^n((x_0, y_0, z_0)) = (x_0 + nu, y_0 + nv, z_0 + nw) \tag{3}$$

for integers $n = 1, 2, \ldots$.

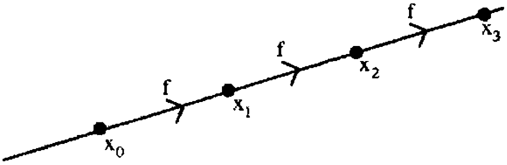


FIGURE 6.2.1
Orbit of the dynamical system described by a rectilinear discrete-time motion.

As long as the velocity vector $v \neq 0$, the orbit is ***unbounded*** and escapes to infinity. In particular, it does not approach the initial state x_0 at any future times.

The latter property of the orbit of the rectilinear motion, called ***transience***, can be formulated more precisely as follows: there is no subsequence $x_{k_1}, x_{k_2}, \ldots, k_1 < k_2 < \ldots$, of the orbit that converges to x_0. Accordingly, the state x_0 is called ***transient*** if its orbit is transient. The set of all transient states is called the ***dissipative set*** of the dynamical system.

In the opposite case, when the state x_0 is the limit point of some subsequence $x_{k_1}, x_{k_2}, \ldots, k_1 < k_2 < \ldots$, it is called a ***recurrent*** state. The set of all recurrent states is called the ***conservative set*** of the dynamical system. The simplest example of a recurrent state is any state that is an element of a finite, cyclic orbit.

In what follows we will see examples of systems in which both phases, dissipative and conservative, exist.

Definition 6.2.1 Fixed Points.
The state $s \in S$ is said to be a fixed point (or, equivalently, a stable point, or an equilibrium state) of the map f if $f(x) = x$.

Clearly, any orbit starting at a fixed point x_0 remains there forever since $x_n = f^n(x_0) = x_0$. Finding fixed points requires solving the equation $f(x) = x$.

Definition 6.2.2 Periodic Orbits.
The orbit $x_0, x_1, \ldots,$ is said to be periodic (or a cycle) if the set $\{x_0, x_1, \ldots\}$ is finite and x_0 is recurrent, or, equivalently, if there exists an integer $p > 0$, called a period, such that, for all integers k, $x_{k+p} = x_k$.

A periodic orbit visits the same point every p steps. The smallest period p is called the *principal period* of the orbit. Each state on a periodic orbit is called a *periodic state*. Each periodic state is visited infinitely many times. Finding periodic points with period p requires solving the equation $f^p(x) = x$. The orbit is called *aperiodic* if it contains no periodic states.

Example 6.2.2 Rotation of the Unit Circle Revisited.
The rotation map $f(z) = ze^{i\alpha}$ on the unit circle $T = \{z : |z| = 1\} \subset \mathbf{C}$, considered as a subset of the complex plane and introduced in Example 6.1.2, displays different kinds of recurrent orbits depending on the angular velocity parameter α.

If α is an integer multiplicity of 2π, that is,

$$\alpha = 2\pi m, \qquad m \in \mathbf{Z},$$

then each point z of the state space T is a fixed point,

$$ze^{2\pi m} = z, \qquad z \in T,$$

since rotation by an integer multiplicity of 2π will return the point back to its original position.

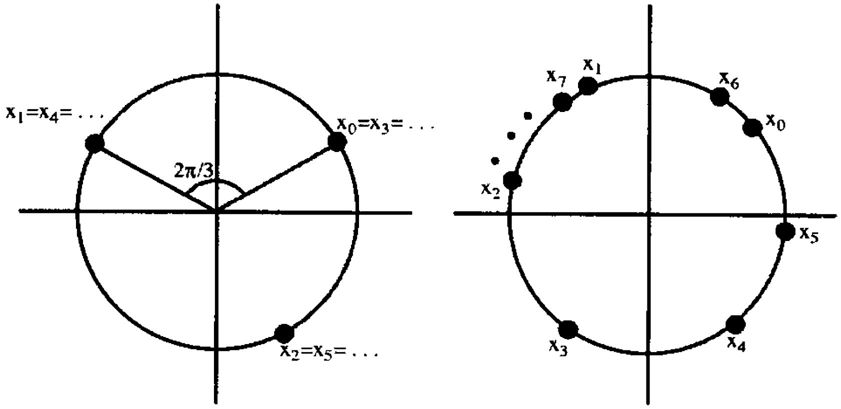


FIGURE 6.2.2
(i) Periodic orbit of rotation map with parameter $\alpha = 2\pi/3$ has period 3. (ii) The orbit of the rotation map with parameter $2\pi/\sqrt{3}$ is aperiodic.

If α is a rational multiplicity of 2π, that is,

$$\alpha = 2\pi \frac{m}{n}, \qquad m \in \mathbf{Z}, n \in \mathbf{N}$$

(assume, for simplicity, that m and n have no common divisors), then each point z of the state space T is a periodic state with period n since

$$z\left(e^{2\pi m/n}\right)^n = ze^{2\pi m} = z, \qquad z \in T,$$

and since, in view of a lack of common divisors of m and n, n is the principal period. The orbit of any point consists of exactly n points (see Fig. 6.2.2). Notice that as the denominator n increases, more and more points on the unit circle are visited by the orbit.

If α is an irrational multiplicity of 2π, that is,

$$\alpha = 2\pi\gamma,$$

where γ is an irrational number not representable as a fraction $m/n,\ m, n \in \mathbf{Z}$, then each point z of the state space T is an aperiodic (non-periodic) state since for no positive integer k

$$z\left(e^{2\pi\gamma}\right)^k = ze^{2\pi\gamma k} = z, \qquad z \in T.$$

Indeed, were it otherwise, $e^{2\pi\gamma k} = 1$ and we would have $\gamma k = l$, for an integer l, which would imply that $\gamma = l/k$ in violation of its irrationality. The orbit of an irrational rotation contains thus infinitely many different states. It can be shown that each state x_0 is recurrent. Use *Mathematica* to produce orbits for different x_0, and check this condition numerically. Remember, however, that in computer practice we almost always end up with rational approximations to irrational numbers.

The behavior of a system in a neighborhood of a periodic point is one of the central problems in the theory of dynamical systems. Here, we will encounter only two different types of behavior:

(i) The fixed point $x \in S$ is said to be *repelling* if the orbit starting in its neighborhood runs away from it (at least in the short run).

(ii) The fixed point $x \in S$ is said to be *attracting* if the orbit starting in its neighborhood converges to it.

For maps $f : \mathbf{R} \mapsto \mathbf{R}$ of the real line (and, in particular, of the unit interval) there is a simple criterion (see Fig. 6.2.3) of when the fixed point x is repelling or attracting. Namely, if

$$|f'(x)| > 1,$$

then the point is repelling, and if

$$|f'(x)| < 1,$$

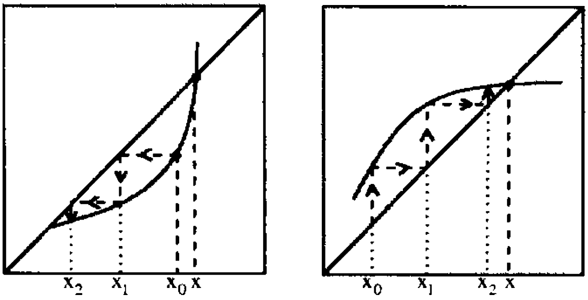


FIGURE 6.2.3

Two types of fixed points for maps of the unit interval. (i) State x is repelling since $|f'(x)| > 1$. (ii) State x is attracting since $|f'(x)| < 1$.

then the point is attracting. The fixed point x is called *super-attractive* if $f'(x) = 0$, that is, when it is also a *critical point* in the sense that the map is not locally 1-1. In the critical case when $|f'(x)| = 1$, states on one side of the fixed point can be attracted to it and points on the other side can be repelled (see Fig. 6.2.4). Such a point is called a *neutral point.*

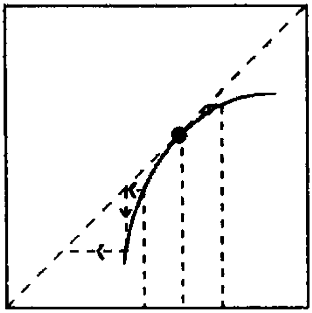


FIGURE 6.2.4

A neutral (unstable) fixed point for a map of the unit interval. Points to the right of it are attracted to it but points to the left are repelled by it.

If x is a periodic point, with principal period p, say, then x is a fixed point of the p-th iterate f^p, i.e., $f^p(x) = x$, and we shall say that x is an *attracting, repelling,* or *neutral point,* according to whether it has this property with respect to f^p. It is easy to see that for a differentiable f, the derivative of f^p is constant on $x_0, x_1, \dots, x_{p-1}$. Hence, all points in a periodic orbit are of the same type. In this case, we also say that the periodic orbit is attracting, repelling, or neutral (see Fig. 6.2.5).

The irrational rotation in Example 6.2.2 has the property of *equicontinuity,* which means here that the distance of $ze^{2\pi i \gamma}$ and $ye^{2\pi i \gamma}$ is the same as the distance of z

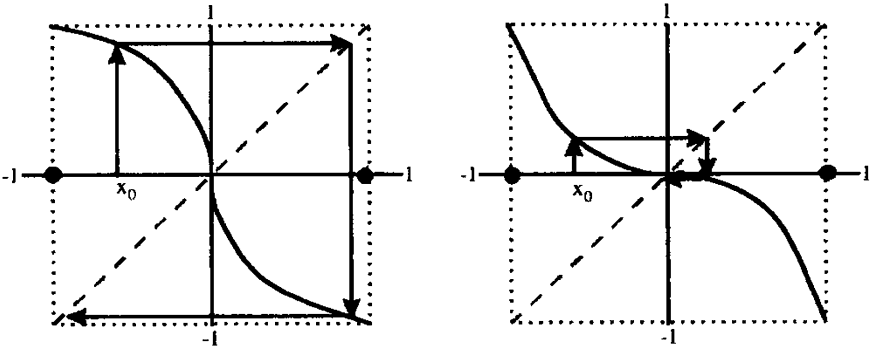


FIGURE 6.2.5
(i) Map $f(x) = -x^{1/3}$ of the interval $[-1, 1]$ has a periodic orbit $\{-1, 1\}$ which is attractive. Map $f(x) = -x^3$ has also the same periodic orbit $\{-1, 1\}$ but it is repelling.

and y. Thus, the distance relations in two different orbits do not change in time. Such a mapping is completely regular, or, in other words, deterministic. It produces no random effects.

On the opposite end of the behavior, the orbit will be called *chaotic* if is exponentially unstable, dense in the state space, and if it is neither periodic nor attracted by a periodic orbit. An orbit originating in the neighborhood of a repelling fixed point is called exponentially unstable if it moves away from it (for some, perhaps limited, period of time) at the rate $e^{\lambda n}$, with *Liapunov exponent* λ, for some $\lambda > 0$. We have seen this kind of behavior in the billiard *Mathematica* experiment in Chapter 1, when a convex obstacle was present. One can also express this property in terms of the *sensitivity to initial conditions*. Indeed, two orbits starting at nearby points diverge at an exponential rate as time progresses. This phenomenon is particularly troubling for computer simulations, where the same experiment conducted on two different computers with different error rounding mechanisms can produce dramatically different results.

Example 6.2.3 Logistic Map Revisited. Consider again the logistic map

$$f(x) = ax(1-x)$$

of the unit interval considered in Example 6.1.6. The derivative $f'(x) = a - 2ax$. It turns out that, depending on the parameter a, this family of dynamical systems on the unit interval displays a variety of different behaviors.

For $a = 1/3$, that is, for $f(x) = (1/3)x(1-x)$, the fixed point x has to solve the equation

$$x = \frac{1}{3}x(1-x).$$

Clearly, only $x = 0$ satisfies this equation inside the unit interval (the other solution $x = -2$ is outside our state space). The fixed point $x = 0$ is attractive as $f'(0) = 1/3 < 1$.

If $a < 3$, the system has an attracting fixed point. Indeed, the equation $ax(1-x) = x$ has solutions

$$x = \begin{cases} 0, & \text{if } a \leq 1 \\ 0 \quad \text{and} \quad (a-1)/a, & \text{if } a > 1. \end{cases}$$

For $a < 1$, we have $f'(0) = a < 1$, so the only fixed point $x = 0$ is attracting. For $a = 1$, the fixed point $x = 0$ although neutral, is also attracting nearby states to the right of it.

For $a > 1$, the fixed point 0 is repelling since $f'(0) = a > 1$. The other fixed point $x = (a-1)/a$, where $f'((a-1)/a) = 2 - a$, is repelling if $3 < a \leq 4$, neutral for $a = 3$, and attracting if $1 < a < 3$. However, as the parameter a crosses the critical point $x = 3$, attractive periodic orbits appear, with periods of length 2^n with $n \to \infty$ as $a \uparrow a_F =\approx 3.57\ldots$.[2]

For $a = a_F \approx 3.57\ldots$ there is a fractal attractor which, however, is not the usual attracting set because it is a limit of repelling periodic points (see Fig. 6.2.6). In this case, the Liapunov exponent λ is still $= 0$.

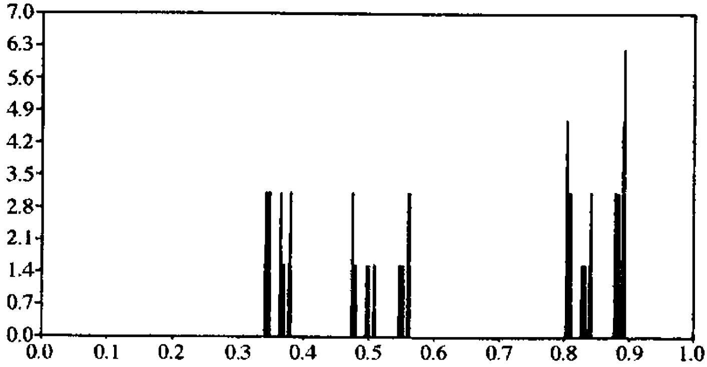


FIGURE 6.2.6

The fractal attractor of the logistic system for $a = a_F \approx 3.57\ldots$. Also marked are the relative frequencies of visits in 1000 bins which will be studied in more depth in Section 6.3.

For parameter values $a > a_F \approx 3.57\ldots$, the orbits become irregular. The whole spectrum of behaviors of the logistic maps is pictured on the orbit diagram

[2] The critical value a_F is sometimes called the Feigenbaum number in honor of the physicist who first discovered it.

in Fig. 6.2.7. The period doubling phenomenon observed on the diagram, as parameter a increases towards the critical value a_F, is called *bifurcation.* It is considered one of the typical routes to chaos in parameter-dependent dynamical systems.

For the logistic function $f(x) = 4x(1 - x)$, there are two fixed points $x = 0$, and $x = 3/4$, and both are repelling (see Fig. 6.1.7). However, in the long run, one can see (depending on the starting point) orbits visiting arbitrarily close to $x = 0$ arbitrarily often; this is the first indication that we can have a new, "chaotic" type of behavior in a simple dynamical system (see Fig. 6.2.8 and 6.2.9).

The system also shows chaotic behavior and sensitivity to initial conditions (see Fig. 6.2.9). Moreover, there are orbits that are dense in the state space $S = [0, 1]$.

Mathematica Experiment 1. Logistic Map. We would like to study orbits of the logistic system $f(x) = 4x(1 - x)$ with nearby starting points: (i) $x = \pi/10$, and (ii) $x = \pi/10 + 0.001$, and later look at their distances for the first 200 steps. There are several functional operations in *Mathematica* that will be useful for iterating maps. The command `Nest[f,x,n]` applies the function `f` nested `n` times to `x`. `NestList[f,x,n]` generates the list `{x,f[x],f[f[x]], ...}`, where `f` is nested up to `n` deep. `FixedPointList[f,x]` generates the list `{x,f[x],f[f[x]], ...}`, stopping when the elements no longer change.

```
In[1]:= f[x_]:=4*x*(1-x)
In[2]:= NestList[f, x, 4]
Out[2]= {x, 4 (1 - x) x, 16 (1 - x) x (1 - 4 (1 - x) x),
          64 (1 - x) x (1 - 4 (1 - x) x) (1 - 16 (1 - x)
          x (1 - 4 (1 - x) x)),
          256 (1 - x) x (1 - 4 (1 - x) x) (1 - 16 (1 - x)
          x (1 - 4 (1 - x) x))
          (1 - 64 (1 - x) x (1 - 4 (1 - x) x) (1 - 16 (1 - x)
          x (1 - 4 (1 - x) x)))}
In[3]:= Expand[%]
Out[3]= {x, 4 x - 4 x^2, 16 x - 80 x^2  + 128 x^3  - 64 x^4 ,
     64 x - 1344 x^2  + 10752 x^2  - 42240 x^3  + 90112 x^4
     - 106496 x^5  + 65536 x^6  - 16384 x^8,
     256 x - 21760 x^2  + 731136 x^3  - 12899328 x^4
     + 137592832 x^5  - 963149824 x^6 + 4656988160 x^7
     - 16066609152 x^8  + 40324038656 x^9  - 74281123840 x^10
     + 100327751680 x^11   - 98146713600 x^12   + 67645734912 x^13
     - 31138512896 x^14  +  8589934592 x^15   - 1073741824 x^16}
In[4]:= l1=Table[Nest[f, N[Pi/10],i], {i,200}]
Out[4]= {0.861853, 0.47625, 0.997744, 0.00900467,  . . . ,
                                0.645448, 0.915379, 0.309841}
In[5]:= lp1=ListPlot[l1, AspectRatio->1/3, PlotJoined->True]
Out[5]= -Graphics-
In[6]:= l2=Table[Nest[f, N[Pi/10+0.001],i], {i,200}]
```

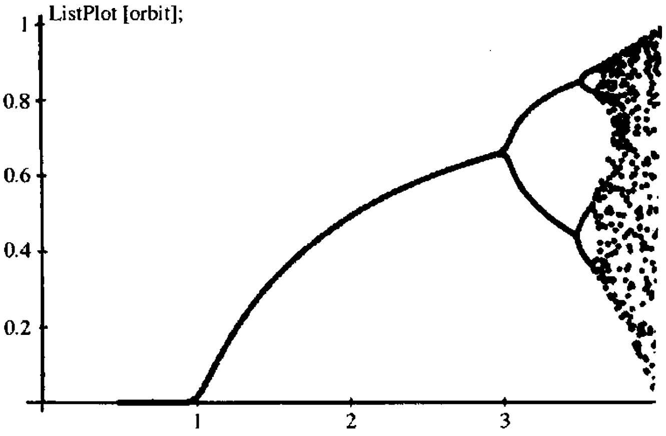


FIGURE 6.2.7
The orbit diagram of the logistic system for $0 < a \leq 4$. *The dots indicate the states visited. Attracting periodic orbits of periods* 2^n *appear for* $3 < a < a_F \approx 3.57\ldots$, *with* $n \to \infty$ *as* $a \uparrow a_F$. *For* $a_F < a \leq 4$, *the orbits are irregular.*

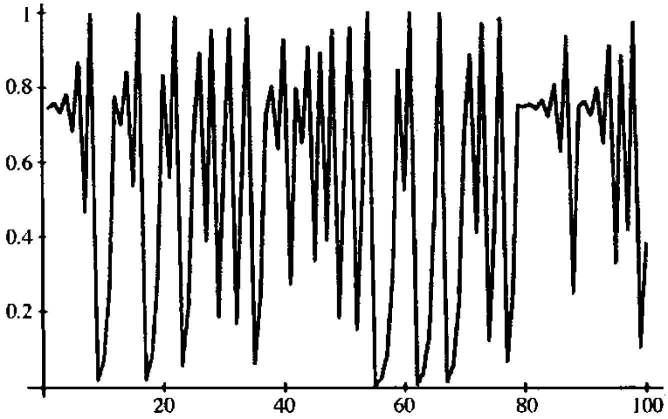


FIGURE 6.2.8
"Chaotic" behavior of a selected orbit for the logistic map $f(x) = 4x(1-x)$.

```
Out[6]= {0.863336, 0.471949, 0.996853, 0.0125502,  . . . ,
                              0.543269, 0.992511, 0.0297308}
In[7]:= lp2=ListPlot[l2, AspectRatio->1/3, PlotJoined->True]
Out[7]= -Graphics-
In[8]:= ld=Table[Abs[l1[[i]]-l2[[i]]],{i,200}]
Out[8]= {0.00148273, 0.00430102, 0.000891191, 0.0035455,  . . . ,
                              0.10218, 0.0771322, 0.28011}
```

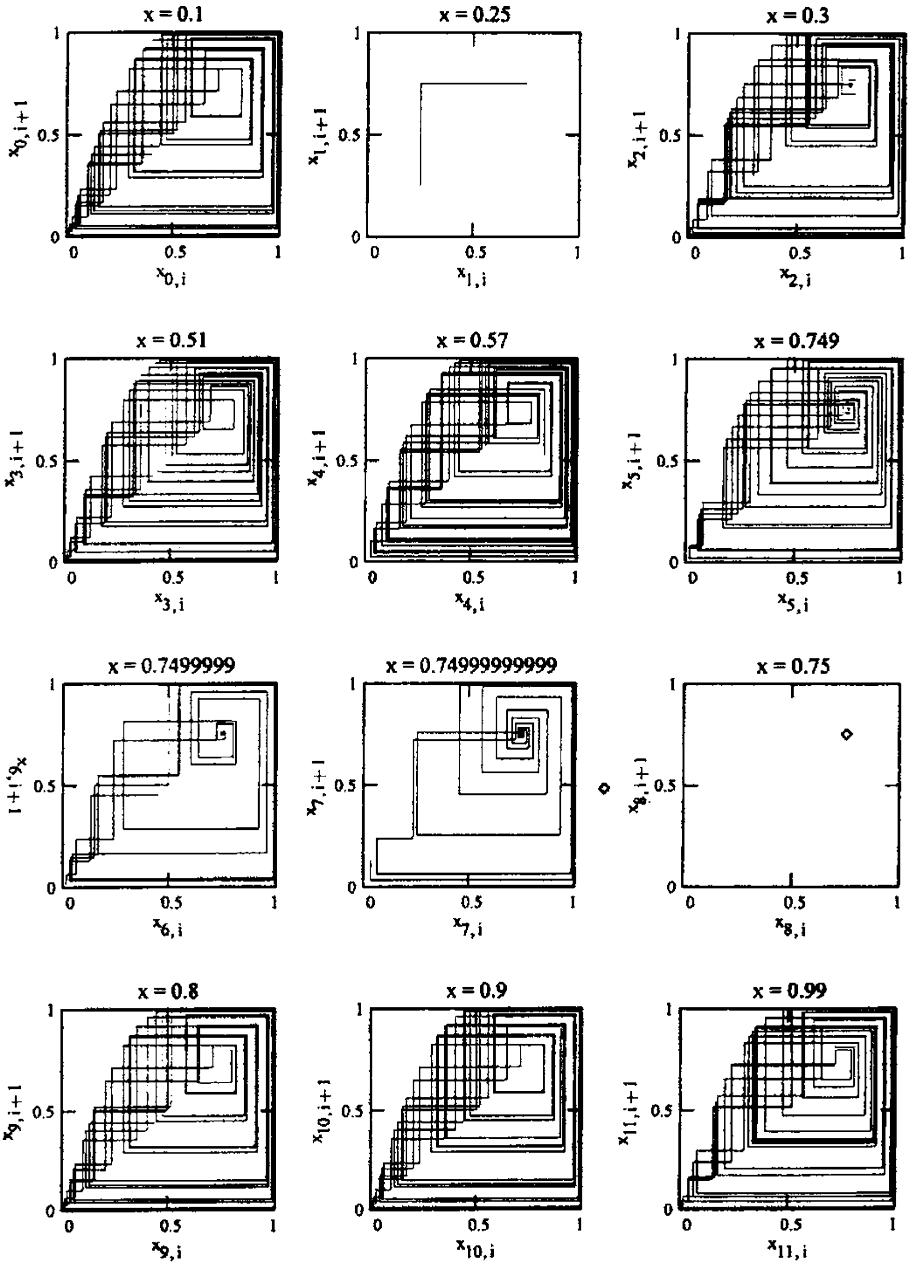
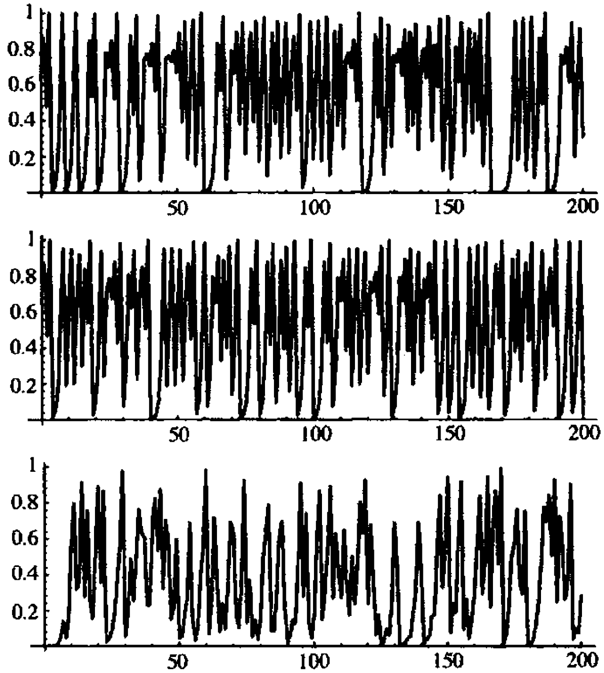


FIGURE 6.2.9

Another look at the "chaotic" behavior of selected orbits for the logistic map $f(x) = 4x(1-x)$, with starting points $x = 0.1, 0.25, 0.3, 0.51, 0.57, 0.749, 0.7499999, 0.74999999999, 0.75, 0.8, 0.9, 0.99$.

```
In[9]:= lpd=ListPlot[ld,AspectRatio->1/3,PlotJoined->True]
Out[9]= -Graphics-
In[10]:= Show[GraphicsArray[{{lp1},{lp2},{lpd}}], Frame->True]
Out[10]= -GraphicsArray-
```



Note the "exponential" initial growth of the distance between the two orbits which, of course, is later limited by the boundedness of the state space $S = [0, 1]$.

Sensitive dependence on initial conditions. The behavior observed in the Lorenz system, billiard with a convex obstacle, and the logistic system with $a = 4$, displays the so-called sensitive dependence on the initial conditions. Roughly speaking, this means that if the initial conditions differ a little, say Δx_0, then the orbits diverge exponentially as time progresses, so that, for a certain $\lambda > 0$, we have $\Delta x_n \approx e^{\lambda N}$. The number λ is called the *Liapunov exponent* (or, the *characteristic exponent*) of the system and it measures the rate of divergence of the orbits. If we operate with finite precision (resolution), as is always the case with real physical measurements, or when we use computers for simulation or data collection, the sensitive dependence means that we can start with identical (meaning, indistinguishable) initial conditions and still have wildly differing orbits. This is a new way to produce random effects in physical phenomena and a new attribute of randomness. In the rest of this chapter, we will show that chaotic dynamical systems also possess more traditional attributes of randomness such as stability of frequencies and Gaussianness of fluctuations.

A Mathematical Aside. Formally, we say that $f : S \to S$ has sensitive dependence on initial conditions if there exists $\delta > 0$ such that for any $x \in S$ and any neighborhood V of x, there exists $y \in V$ and $n > 0$ such that

$$|f^n(x) - f^n(y)| > \delta.$$

For maps of the unit interval $S = [0, 1]$, this condition means that there exists $\delta > 0$ such that, for any $x \in [0, 1]$, and any $\epsilon > 0$, there exists y and $n > 0$ such that $|x - y| < \epsilon$ but $|f^n(x) - f^n(y)| > \delta$. This, of course, means that the derivative (or, at least, the differential ratios) of some iterations of the map become very large. Indeed,

$$\Delta x_n = x_n - x_n' = f^n(x_0) - f^n(x_0') \approx \frac{df^n(x_0)}{dx}(x_0 - x_0') = \frac{df^n(x_0)}{dx}\Delta x_0.$$

Applying the chain rule to the n-fold composition of the map f we have

$$\frac{df^n(x_0)}{dx} = \frac{df(x_{n-1})}{dx} \cdot \frac{df(x_{n-2})}{dx} \cdots \frac{df(x_0)}{dx},$$

so that the average rate of growth per unit time-step can be defined as

$$\lambda = \lim_{n \to \infty} \frac{1}{n} \log \left| \frac{df^n(x_0)}{dx} \right|. \tag{1}$$

That this limit exists for "almost all" initial states x_0 is a deep mathematical theorem called the Multiplicative Ergodic Theorem. It was proved by the Russian mathematician Oseledec in 1968. We will explore this avenue further in the next section.

Example 6.2.4 Liapunov Exponents of the Logistic Family.
Using (1) or, more precisely, the formula

$$\lambda_n(a) = \frac{1}{n} \sum_{i=0}^{n-1} \log \left| \frac{df}{dx}\left(f_a^i(0)\right) \right|,$$

with $f_a(x) = ax(1 - x)$ one can easily compute the Liapunov exponents for different values of parameter a in the logistic system. The results, showing a curve of great complexity, are shown in Fig. 6.2.10

The logistic map considered above on the interval [0,1] is, after rescaling, a special case of the quadratic polynomial $z \mapsto z^2 + c$ on the complex plane. Although such mappings have a simple and transparent structure, they can produce

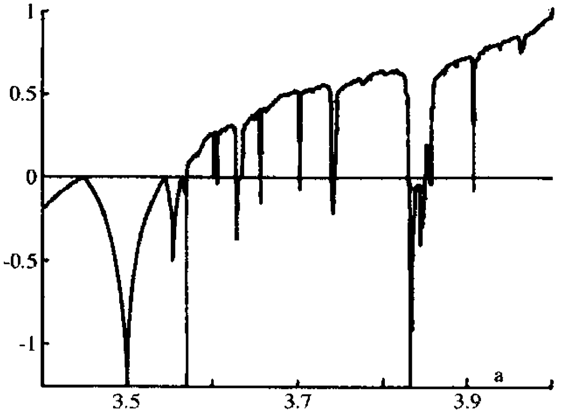


FIGURE 6.2.10

Liapunov exponent $\lambda(a)$ of the logistic map $f_a(x) = ax(1-x)$ depends on the parameter a. This dependence is shown in the above "dripping paint" graph. This particular calculation was done for $n = 100,000$ iterations, for 300 values of a, spaced 0.002 apart. (Adapted from Ruelle (1987).)

an amazing spectrum of behaviors. Answering simple questions like "What is the set of initial states for which the orbits are bounded?" or, "What is the (closure of) repelling periodic states for a given map f?" often leads to objects of daunting complexity and exhilarating richness (see Fig. 6.2.11 and 6.2.12). Other maps (see Fig. 6.2.13) are also of great interest in this context.

Fig. 6.2.11 shows the boundary of the so-called Mandelbrot set, the set of points c in the complex plane **C** for which the starting at 0 orbits of the map $f(z) = z^2+c$, do not tend to infinity as $n \to \infty$. The large cusp is at $c = 1/4$ and the left-most point is $c = -2$. Although the set contains arbitrarily small copies of itself, it is not selfsimilar, as each small copy is embelished by different "ornaments".

Mathematica Experiment 2. Sensitive Dependence on Input in Numerical Calculations. Finite precision numerical procedures, especially those that depend on "excavating" progressively less and less significant digits of the input, often show sensitive dependence on the initial condition. The following *Mathematica* experiment has been adapted from Wolfram (1996).

The map $f(x)$ we will take a look at is familiar from Example 6.1.4 the multiplication map modulo 1, that is $f(x) = ax - \lfloor ax \rfloor$, for a certain constant a or, in other words, the map f assigns the fractional part of the multiplicity ax to the input x. Initially, we take $a = 2$, and look at the orbits of close initial values. In the binary expansion terms, you may remember that the mapping is just a shift map of Example 6.1.5. We select as the starting point $x = 1/9$, for which we discover immediately that the orbit is periodic with principal period equal to 6. However, if we take various approximations to $x = 1/9$, like 0.1111 and 0.1112, the story is

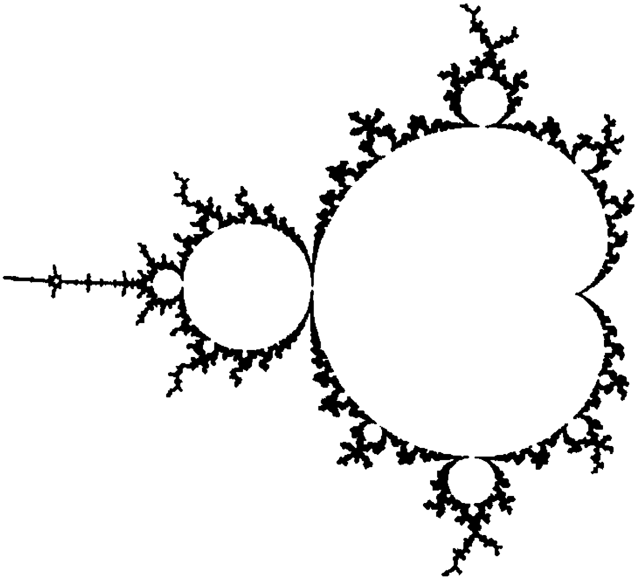


FIGURE 6.2.11

Boundary of the so-called Mandelbrot set, the set of points c in the complex plane $\mathbb{C}$ for which the starting at 0 orbits of the map $f(z) = z^2 + c$, do not tend to infinity as $n \to \infty$. The large cusp is at $c = 1/4$ and the left-most point is $c = -2$.

quite different, and, sometimes, nonsensical numbers result.

The command `RealDigits[x,b,len, n]` returns the first `len` digits starting with the coefficient of b^n, and `N[x,k]` gives the numerical value of `x` with precision `k`.

```
In[1]:= NestList[((2 #)-Floor[2 #])&, 1/9, 10]
Out[1]= {1/9, 2/9, 4/9, 8/9, 7/9, 5/9, 1/9, 2/9, 4/9,
         8/9, 7/9}
In[2]:= {0.111111, 0.222222, 0.444444, 0.888889, 0.777778,
         0.555556, 0.111111,  0.222222, 0.444444, 0.888889,
         0.777778}
In[3]:= NestList[((2 #)-Floor[2 #])&, 0.1111, 10]
Out[3]= {0.1111, 0.2222, 0.4444, 0.8888, 0.7776, 0.5552,
          0.1104, 0.2208, 0.4416, 0.8832, 0.7664}
In[4]:= NestList[((2 #)-Floor[2 #])&, 0.1112, 10]
Out[4]= {0.1112, 0.2224, 0.4448, 0.8896, 0.7792, 0.5584,
          0.1168, 0.2336, 0.4672,  0.9344, 0.8688}
In[5]:= RealDigits[Take[%,5], 2, 8, 0]
Out[5]= {{{0, 0, 0, 1, 1, 1, 0, 0}, 0},
            {{0, 0, 1, 1, 1, 0, 0, 0}, 0},
            {{0, 1, 1, 1, 0, 0, 0, 1}, 0},
```

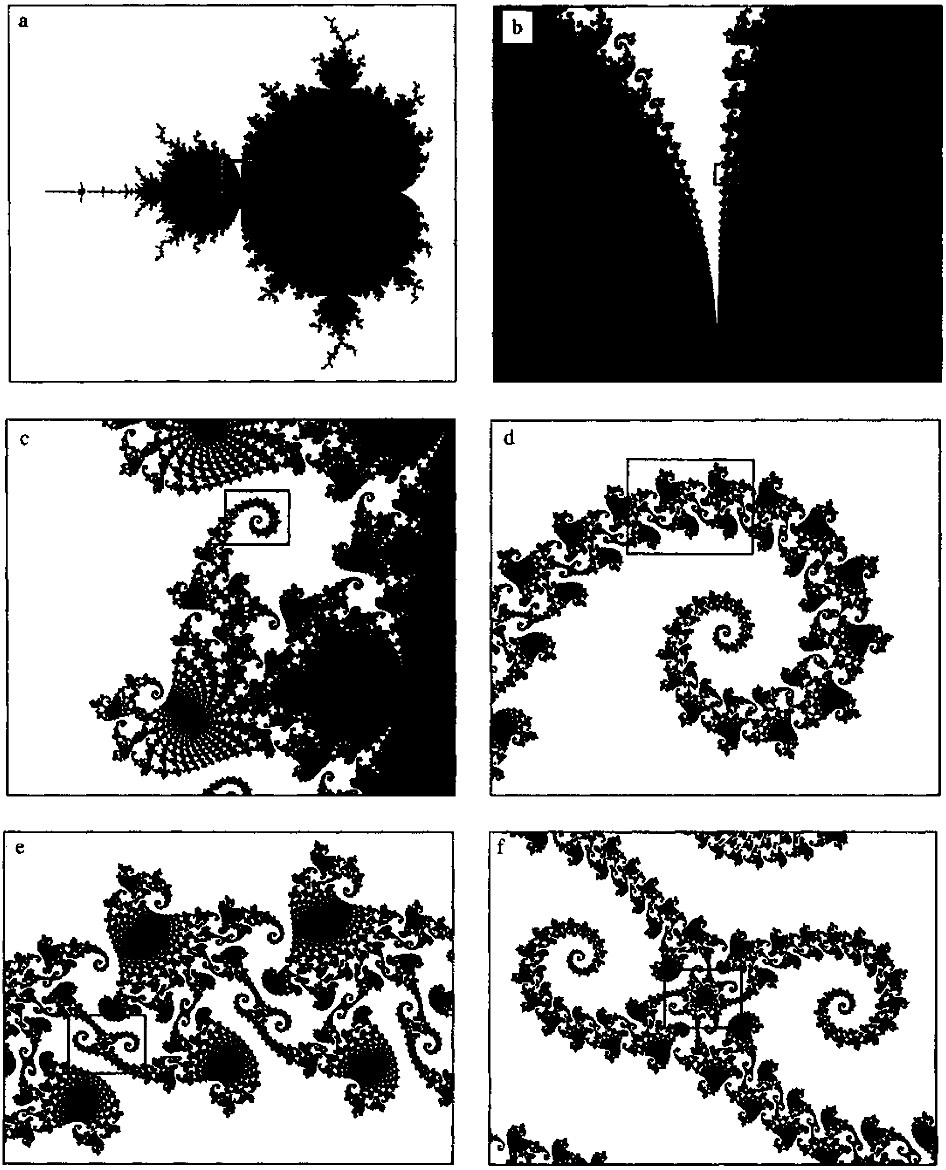


FIGURE 6.2.12

Successive blowups of the Mandelbrot set centered around $z \approx -0.1480798 + i0.6515558$.

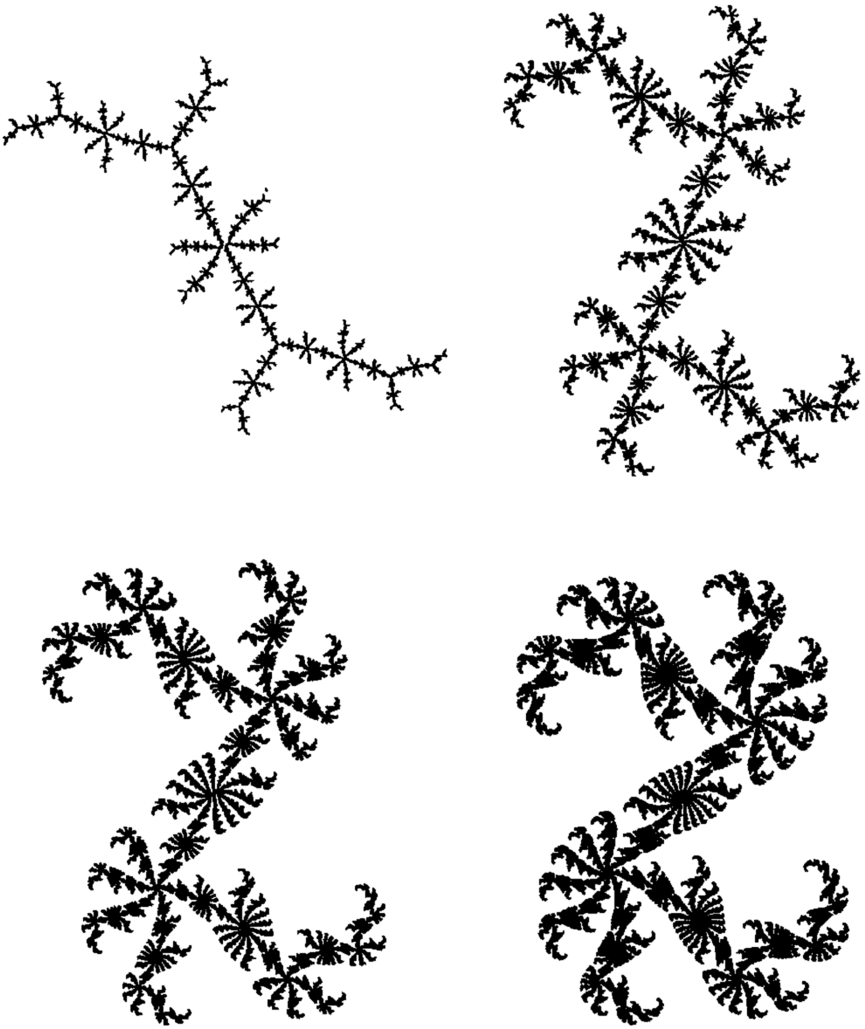


FIGURE 6.2.13

Julia sets, that is (closures of) sets of repelling periodic states. Top: For $f(z) = .36e^z$; Bottom: $f(z) = 0.66i \cos z$, $z \in \mathbf{C}$.

```
           {{1, 1, 1, 0, 0, 1, 1, 1}, 0},
           {{1, 1, 0, 0, 0, 1, 1, 1}, 0}}
In[6]:= NestList[((40 #)-Floor[40 #])&, N[1/9, 20], 20]
Out[6]= {0.11111111111111111111, 0.4444444444444444444,
           0.777777777777777778, 0.1111111111111111,
           0.444444444444444, 0.7777777777778,
           0.11111111111, 0.4444444444, 0.77777778,
           0.1111111, 0.44444, 0.778,
           0.11, 0., 0., 0., 0., 0., 0., 0., 0.}
In[7]:= NestList[((40 #)-Floor[40 #])&,  1/9 , 20]
Out[7]= {1/9, 4/9, 7/9, 1/9, 4/9, 7/9, 1/9, 4/9, 7/9,
            1/9, 4/9, 7/9, 1/9, 4/9, 7/9, 1/9, 4/9, 7/9,
             1/9, 4/9, 7/9}
```

Obviously, `Out[7]` shows periodicity with the principal period equal to 3, whereas the same computation with precision 20 completely deteriorates after 15 steps.

Mathematica Experiment 3. Sensitive Dependence on Initial Conditions for Continuous-Time Dynamical Systems. We mentioned before that differential equations are a continuous-time counterpart of discrete-time dynamical systems. They also can display sensitive dependence on initial conditions. In this experiment we will take a look at two solutions of the *Duffing nonlinear differential equation*

$$x''(t) + 2ax'(t) - \frac{1}{2}x(t)(1 - x^2(t)) = b\cos(\omega t),$$

with initial conditions $x(0) = x_0$, $x'(0) = x_1$, which describes harmonically forced oscillations in a quartic double-well potential, resulting in the cubic nonlinearity. The equation provides a rough model for the motion of a cart rolling on a track with two valleys, subject to horizontal forcing frequency, see Fig. 6.2.14. Viscous damping a is also included.

```
In[1]:= sol=NDSolve[
           {x''[t]+0.15 x'[t]-x[t]+x[t]^3 == 0.3 Cos[t],
            y''[t]+0.15 y'[t]-y[t]+y[t]^3 == 0.3 Cos[t],
            x[0] == y[0] == -1, x'[0] == 1, y'[0] == 1.001},
            {x, y}, {t, 0, 35} ]
Out[1]= {{x -> InterpolatingFunction[{{0., 35.}}, <>],
            y -> InterpolatingFunction[{{0., 35.}}, <>] }}
In[2]:= p1=Plot[Evaluate[x[t] /. sol],{t,0,35}]
Out[2]= -Graphics-
In[3]:= p2=Plot[Evaluate[y[t] /. sol],{t,0,35}]
Out[3]= -Graphics-
In[4]:= pd=Plot[Evaluate[Abs[x[t]-y[t]] /. sol],{t,0,35}]
```

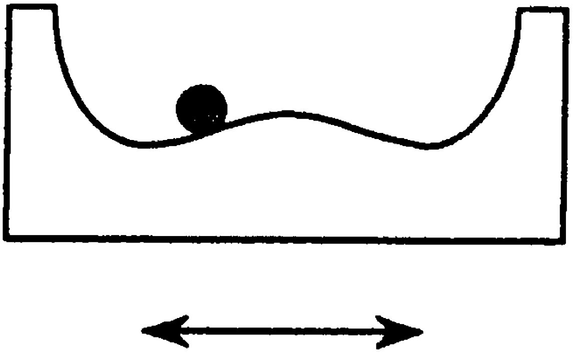
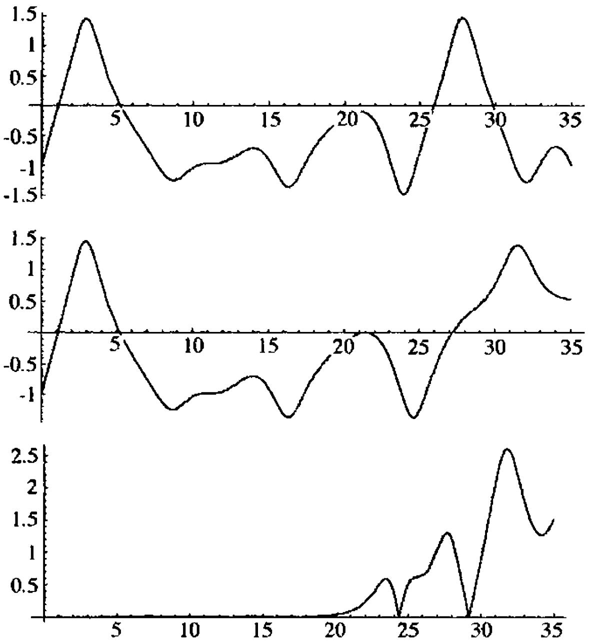


FIGURE 6.2.14

Duffing cart in a double potential well, driven by a lateral harmonic displacement of the track. The resulting differential equation has a cubic nonlinearity.

```
Out[4]= -Graphics-
In[5]:= Show[GraphicsArray[{{p1},{p2},{pd}}, Frame->True]]
Out[5]= -GraphicsArray-
```



6.3 Stability of frequencies and the ergodic theorem

In this section we will begin a systematic study of randomness in simple dynamical systems on the unit interval. Needless to say, even more striking random behavior is possible for more complex systems. The approach will be mostly through computer experimentation; theoretical mathematical proofs of the results to be presented are quite difficult and beyond the scope of this book. The reader interested in pursuing this subject on a more rigorous level can find additional references in the Bibliographical Notes section at the end of this chapter.

The first attribute of randomness discussed in this book was the stability of frequencies and we will initially take a look at this basic statistical property as applied to orbits of dynamical systems. As usual, the first step is to introduce the proper formal framework wherein our question can be discussed.

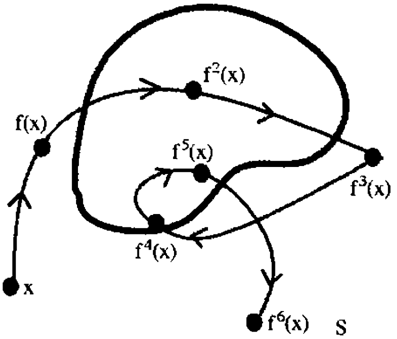


FIGURE 6.3.1

A schematic illustration of the calculation of the fraction of "time" the dynamical system starting at x spends in the given subset T of the state space S.

Consider, again, a dynamical system on the state space S, generated by the map $f : S \mapsto S$, and its orbit

$$x \mapsto f(x) \mapsto f^2(x) \mapsto \ldots \mapsto f^n(x), \tag{1}$$

originating at the state $x \in S$. Notice a little change: we dropped the subscript 0 in the designation of the starting point. Seemingly an innocent alteration, but it indicates a certain change of viewpoint in this section compared with the previous section; we will look here at the orbits with starting points varying all over the state space S. For a subset T of the state space S (see Fig. 6.3.1), the *relative frequency of visits*, during the first n time-steps, by the orbit (1) starting at the state x, is

$$A_n(T,x) = \frac{\#\{j : 1 \le j \le n, f^j(x) \in T\}}{n}$$
$$= \frac{\mathbf{1}_T(f(x)) + \ldots + \mathbf{1}_T(f^n(x))}{n}, \tag{2}$$

where the indicator function

$$\mathbf{1}_T(x) = \begin{cases} 1, & \text{if } x \in T; \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

helps to count the number of visits. Observe that $A_n(T, x)$ can also be interpreted as the fraction of (discrete) time the orbit spends inside the subset T of the state space S, and that, *a priori*, it depends on the starting point $x \in S$.

The reader may note that the discussion in this section parallels the analysis in Section 3.1. This is not an accident.

There is one immediate basic question: *Do the relative frequencies (2) of visits to T stabilize as n increases or, more formally, does the limit $A_\infty(T, x) = \lim_{n\to\infty} A_n(T, x)$ exist for every T?* If this is the case, the starting state x is called a *generic* state of the system. The limiting frequency $A_\infty(T, x)$, if it exists, possesses a number of useful properties:

(i) For any subset T of the state space S, and any starting point x,

$$0 \le A_\infty(T, x) \le 1, \qquad \text{and} \quad A_\infty(S, x) = 1. \tag{4}$$

(ii) For any two disjoint subsets $T_1, T_2 \subset S$, $T_1 \cap T_2 = \emptyset$, the relative frequency of the visits to the union $T_1 \cup T_2$ is equal to the sum[3] of frequencies of visits to each set, i.e.,

$$A_\infty(T_1 \cup T_2, x) = A_\infty(T_1, x) + A_\infty(T_2, x). \tag{5}$$

For that reason we can refer to $A_\infty(T, x)$ as the *counting measure* for the orbit starting at x.

(iii) The counting measure of each generic orbit is *invariant* under action of the map f, that is, for any subset $T \subset S$,

$$A_\infty(f^{-1}T, x) = A_\infty(T, x), \tag{6}$$

where the inverse image $f^{-1}T$ of the subset T is the set of states x that are mapped by f into T, i.e., $f^{-1}T := \{x \in S : f(x) \in T\}$.

[3] As in probability theory of Chapter 5, there remains the question of countable additivity of such measure. This property holds in many cases of interest, but we will ignore it in this book.

Indeed, property (6) can be verified as follows:

$$\begin{aligned}
A_\infty(f^{-1}T, x) &= \lim_{n\to\infty} \frac{\mathbf{1}_{f^{-1}T}(f(x)) + \ldots + \mathbf{1}_{f^{-1}T}(f^n(x))}{n} \\
&= \lim_{n\to\infty} \frac{\mathbf{1}_T(f^2(x)) + \ldots + \mathbf{1}_T(f^{n+1}(x))}{n} \\
&= \lim_{n\to\infty} \frac{\mathbf{1}_T(f(x)) + \ldots + \mathbf{1}_T(f^n(x))}{n} \\
&\quad + \lim_{n\to\infty} \frac{\mathbf{1}_T(f^{n+1}(x)) - \mathbf{1}_T(f(x))}{n} \\
&= \lim_{n\to\infty} \frac{\mathbf{1}_T(f(x)) + \ldots + \mathbf{1}_T(f^n(x))}{n} = A_\infty(T, x).
\end{aligned}$$

For that reason, we shall call $A_\infty(T, x)$ the *invariant measure* generated by the orbit starting at x. It is clear that, for any integer k, we have $A_\infty(T, x) = A_\infty(T, f^k(x))$. Hence, the invariant measure only depends on the orbit, and not on the particular starting point. In this context, the fundamental question is: *When is the invariant measure unique, i.e., independent of the starting point x?* In that case, $A_\infty(T, x) \equiv A_\infty(T)$, we can think about the dynamical system as being equipped with the unique probabilistic measure structure (S, A_∞), similar to that considered in Chapter 5; then the analogues of the Law of Large Numbers and the Central Limit Theorem (Stability of Fluctuations Law) can be investigated. Systems with unique invariant measure are also called *uniquely ergodic*.[4] The unique ergodicity implies the following property, which will play an essential role later on in this section:

Definition 6.3.1 Ergodic Invariant Measure.

The invariant measure $A_\infty(T, x)$ is said to be ergodic for the dynamical system (S, f) if, for any invariant set T, i.e., such that $f^{-1}T = T$, we have either $A_\infty(T, x) = 0$, or $A_\infty(T, x) = 1$.

In the case when the state space S is an interval, a circle, the real line, or more generally, a subset of the d-dimensional Euclidean space $\mathbf{R}^d$, we shall be interested in those unique invariant measures $A_\infty(T)$ which are described by a density, i.e.,

[4] In general, there is no unique invariant measure, not even for maps of the unit interval. However, very often there is a unique absolutely continuous invariant measure which is automatically ergodic.

an invariant measures for which there exists a nonnegative function $h(x)$ on the state space $S \subset \mathbf{R}^d$ such that, for all measurable subsets $T \subset S$,

$$A_\infty(T) = \int_T h(x)\,dx.$$

Mathematica Experiment 1. Invariant Measures for Rotations of the Unit Circle. For rational rotations

$$f(z) = ze^{i\alpha}, \qquad |z| = 1, \quad \alpha = 2\pi\frac{m}{n},$$

where m, n are integers without common divisors, all orbits are periodic with period n (see Examples 6.1.2 and 6.2.2). Since $ze^{in\alpha} = ze^{i2\pi m} = z$, an orbit starting at z visits exactly n distinct states

$$z \mapsto ze^{i\alpha} \mapsto ze^{i2\alpha} \mapsto \ldots \mapsto ze^{i(n-1)\alpha}, \tag{7}$$

each with the same relative frequency asymptotically ($n \to \infty$) equal to $1/n$. So, the invariant measure $A_\infty(B, z)$ strongly depends on the starting point z. It is a sum of discrete masses of size $1/n$ concentrated at points of the orbit (7). Often, the Dirac-delta notation is used to denote such a measure:

$$A_\infty(B, z) = \frac{1}{n}\sum_{k=0}^{n-1} \delta_{z\exp(ik\alpha)}(B), \qquad B \subset \{z : |z| = 1\}, \tag{8}$$

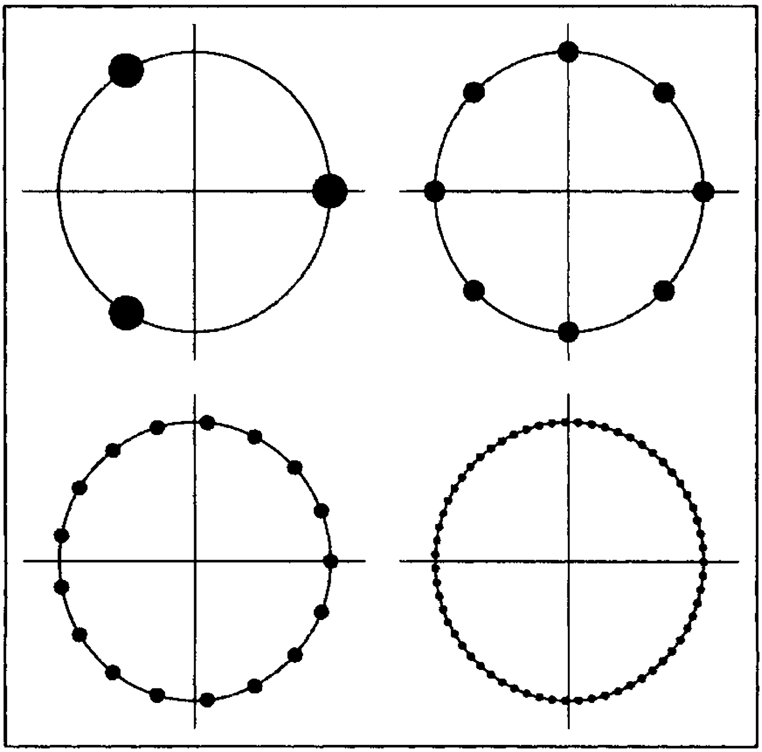
where the Dirac-delta measure $\delta_x(B)$ of a set B assigns to this set a "weight" 1 if x is in B, and "weight" 0 if point x is outside B, i.e.,

$$\delta_x(B) = \begin{cases} 1, & \text{if } x \in B; \\ 0, & \text{if } x \notin B. \end{cases} \tag{9}$$

As n increases, the corresponding invariant measures are carried by more and more points on the circle and their support becomes more and more dense in the state space, while individual points "weigh" less and less; the invariant measure becomes more diffuse.

In the experiment below, we will select the point sizes so that the total area of the discs, which indicate how much invariant measure is concentrated at each point, remains constant for different values of $n = 3, 8, 17, 63$. Remember that the semicolon at the end of a command line suppresses output.

```
In[1]:= p1= ParametricPlot[{Cos[x], Sin[x]},{x,0,2*Pi},
                  AspectRatio->1, Ticks->None]
In[2]:= p3 = ListPlot[Table[{Cos[2*Pi*i/3], Sin[2*Pi*i/3]}, {i,3}],
           AspectRatio->1,PlotStyle->PointSize[0.1]]
In[3]:= p8 = ListPlot[Table[{Cos[2*Pi*i/8], Sin[2*Pi*i/8]}, {i,8}],
           AspectRatio->1,PlotStyle->PointSize[0.1*(3/8)^(1/2)]]
In[4]:= p17=ListPlot[Table[{Cos[2*Pi*i/17],Sin[2*Pi*i/17]}, {i,17}],
           AspectRatio->1,PlotStyle->PointSize[0.1*(3/17)^(1/2)]]
In[5]:= p63=ListPlot[Table[{Cos[2*Pi*i/63],Sin[2*Pi*i/63]},{i,63}],
           AspectRatio->1,PlotStyle->PointSize[0.1*(3/63)^(1/2)]]
In[6]:= p13=Show[p1,p3]
In[7]:= p18=Show[p1,p8]
In[8]:= p117=Show[p1,p17]
In[9]:= p163=Show[p1,p63]
In[10]:= Show[GraphicsArray[{{p13,p18},{p117,p163}}]]
Out[10]= -GraphicsArray-
```



Irrational rotations can be thought of as limits of rational rotations with the denominator (period) $n \to \infty$. Indeed, any irrational number can be approximated be a sequence of rational numbers with denominators increasing to infinity. The corresponding invariant measures become uniformly washed over the whole state space (unit circle) and the limit invariant measure has a constant density $(2\pi)^{-1}$ with respect to the uniform measure on the circle:

$$A_\infty(B, z) = A_\infty(B) = (2\pi)^{-1}|B|, \tag{10}$$

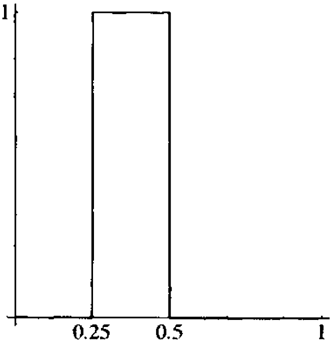
where $|B|$ denotes the arc length (Lebesgue) measure of B. In other words, for an irrational rotation map, the fraction of time an orbit spends in a set B is equal to the length of B divided by 2π. Notice that in the process, the dependence on the starting point has disappeared: the irrational rotation map on the unit circle produces a uniquely ergodic dynamical system. This result, due to the Swiss mathematician Hermann Weyl, can also be obtained as a rigorous mathematical theorem which is an example of a more general Ergodic Theorem that we will discuss later on in this section.

Equipartition Theorem for Irrational Rotations. *The dynamical system defined by an irrational rotation of the unit circle is uniquely ergodic and its invariant measure is the normalized arc length.*

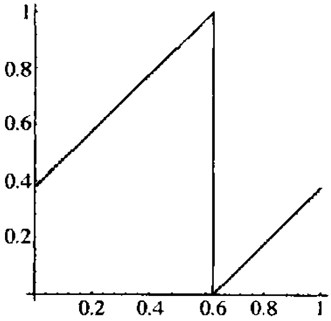
A formal proof of this theorem can be found in the literature quoted at the end of this chapter. We will verify it on a particular example of where, for a change, the rotations of the unit circle are encoded as addition map $f : x \mapsto x + \alpha$ (mod 1) on the unit interval $[0, 1)$. We will select $\alpha = 1/\sqrt{11}$ and will trace the relative frequency of visits $A_k(T, x)$ to the set $T = [1/4, 1/2)$, which is of Lebesgue measure 1/4. The starting point x is selected to be 0, and we will use formula (2) which employs the indicator function of T. It is worth remembering that *Mathematica* can calculate with arbitrary precision, and as long as possible keeps things in the symbolic form without any loss of precision. This is, of course, of great importance when doing calculations with irrational numbers.

The *Mathematica* command `Apply [Plus, list]` adds all the elements of the `list`, and `Take [list, k]` forms the list of the first `k` elements of the `list`.

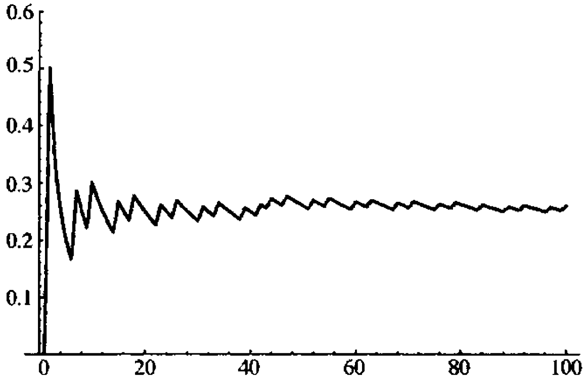
```
In[1]:= Ind[x_]:=If[x<1/4,0,1]-If[x<1/2,0,1]
In[2]:= Plot[Ind[x],{x,0,1}, AspectRatio->1]
Out[2]= -Graphics-
```



```
In[3]:= a=1/Sqrt[7]
Out[3]= 1/Sqrt[7]
In[4]:= N[%, 40]
Out[4]= 0.3779644730092272272145165362341800608158
In[5]:= g[x_]:=(x+a)-Floor[x+a]
In[6]:= Plot[g[x],{x,0,1}, AspectRatio->1]
```



```
In[7]:= l1= NestList[((N[a]+ #)-Floor[N[a]+ #])&, 0, 100]
Out[7]= {0, 0.377964, 0.755929, 0.133893, 0.511858, . . . ,
                 0.662554, 0.0405184, 0.418483, 0.796447}
In[8]:= t1= Table[ Ind[l1[[i]]], {i,100}]
Out[8]= {0, 1, 0, 0, 0,  . . . ,  0, 1, 0, 0, 1}
In[9]:= t2= Table[N[(1/k) Apply[Plus,Take[t1,k]]], {k,100}]
Out[9]= {0, 0.5, 0.333333, 0.25, 0.2,  . . . ,
             0.255102, 0.252525, 0.26}
In[10]:= ListPlot[t2, PlotJoined->True, PlotRange->{0,0.6}]
Out[10]= -Graphics-
```

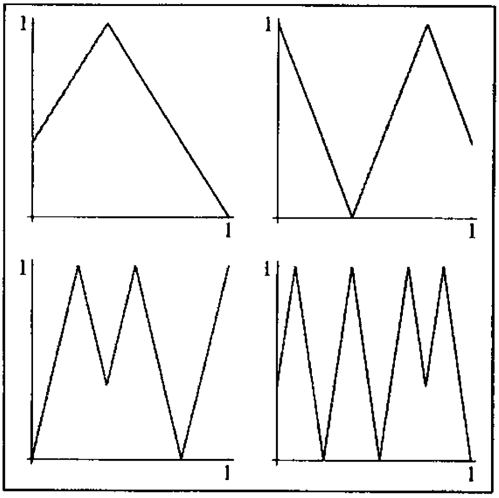


Remark 6.3.1 Observe that the rational rotations of the unit circle have other invariant measures besides (8). In particular, the normalized arc-length measure is invariant under any rotation, but only for irrational rotations it is *ergodic*.

Mathematica Experiment 2. Invariant Measure for a Tent Map. In this experiment we consider another kind of a simple map of the unit interval $S = [0, 1]$ defined by the piecewise linear function

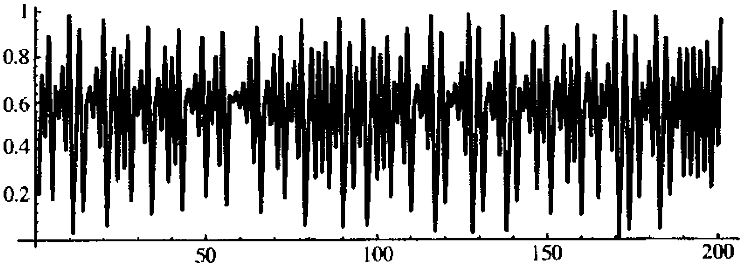
$$f(x) = 1 - a|x - 1 + 1/a|, \qquad a = (1 + \sqrt{5})/2. \tag{11}$$

```
In[1]:= f[x_]:= 1-a*Abs[x-1+1/a]
In[2]:= a=(1+Sqrt[5])/2;
In[3]:= N[%,20]
Out[3]= 1.6180339887498948482
In[4]:= N[{f[0], f[1], f[1-1/a]}]
Out[4]= {0.381966, 0, 1.}
In[5]:= p1=Plot[f[x],{x,0,1}, AspectRatio->1,
            Ticks->{{0,1},{0,1}}];
In[6]:= p2=Plot[f[f[x]]], {x,0,1}, AspectRatio->1,
            Ticks->{{0,1},{0,1}}];
In[7]:= p3=Plot[f[f[f[x]]], {x,0,1}, AspectRatio->1,
            Ticks->{{0,1},{0,1}}];
In[8]:= p4=Plot[f[f[f[f[x]]]], {x,0,1}, AspectRatio->1,
            Ticks->{{0,1},{0,1}}];
In[9]:= Show[GraphicsArray[{{p1,p2}, {p3,p4}}], Frame->True]
Out[9]= -GraphicsArray-
```

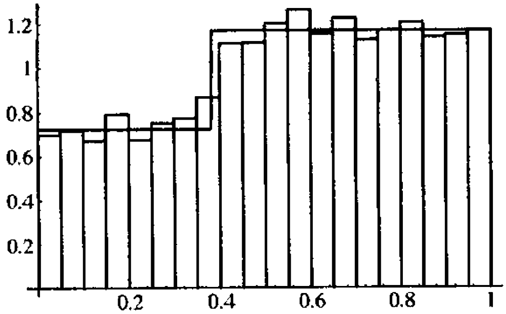


Since we have no prior intuitions about the frequencies of visits for the tent map system, we will proceed in a systematic fashion to find out what they look like for different subsets of the state space [0,1]. Since, the sets in which we are interested can be patched together from disjoint, sufficiently small intervals, and since the invariant measure is additive on disjoint sets, we will partition the unit interval into 20 small bins $T_1, \ldots, T_{20}$, and find out experimentally $A_\infty(T_k, x)$ (or, more precisely, $A_n(T_k, x)$ for $n = 5,000$) by constructing the histogram of its orbit starting at $x = 0.21$. You will also note that some orbits are periodic, but they are an exception.

```
In[1]:= <<UVW‘DataRep‘
In[2]:= a=(1+Sqrt[5])/2
In[3]:= g[x_]:= 1-N[a]*Abs[x-1+N[1/a]]
In[4]:= 10=NestList[(g[#])&, 0, 9]
Out[4]= {0, 0.381966, 1., 0., 0.381966, 1., 0., 0.381966,
         1., 0.}
In[5]:= 1021=NestList[(g[#])&, 0.21, 5000]
Out[5]= {0.21, 0.721753, 0.450213, 0.889574, 0.178673,
         0.671064, 0.532229, 0.756869, 0.393394,
         0.981509, . . . }
In[6]:= 1p021=ListPlot[Take[1021,200], AspectRatio->1/3,
              PlotJoined->True]
Out[6]= -Graphics-
```



```
In[7]:= rh=RegularHisto[1021,0,1,20]
Out[7]= -Graphics-
In[8]:= p=Plot[If[x<1-1/a,a/(2a-1),a^2/(2a-1)],{x,0,1},
                PlotRange->{0,1.3}]
Out[8]= -Graphics-
In[9]:= Show[rh,p]
Out[9]= -Graphics-
```



The outcome suggests the ergodic behavior of the tent map and the existence of the invariant measure which admits a density $h(x)$ that takes only two values, one to the left of the tip of the tent and another to the right of the tip. One can prove that, more exactly,

$$h(x) = \begin{cases} a/(2a-1), & \text{if } 0 \le x \le 1 - 1/a; \\ a^2/(2a-1), & \text{if } 1 - 1/a < x \le 1. \end{cases} \tag{12}$$

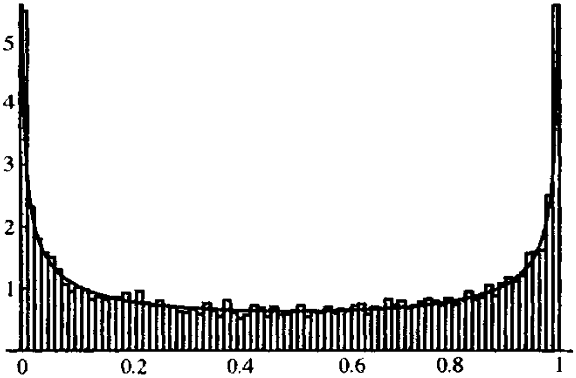
This density was superimposed on the histogram of the orbit in the above *Mathematica* experiment. The invariant measure itself of a set T is then of the form

$$A_\infty(T) = \int_T h(x)\,dx, \tag{13}$$

and one can check by direct calculation that this measure is indeed invariant under action of the tent map.

Mathematica Experiment 3. Ergodic Invariant Measure for the Logistic Map. The Experiment 2 technique used to find the invariant measure for the tent map will now be used to find a more complex invariant measure for the logistic map $f(x) = 4x(1-x)$ discussed in Examples 6.1.6 and 6.2.3. The histogram of relative frequencies of visits to 80 small bins for the orbit of length 10,000 starting at $x = 0.21$ is studied below.

```
In[1]:= <<UVW'DataRep'
In[2]:= f[x_]:= 4*x*(1-x)
In[3]:= nl=NestList[(f[#])&, 0.21, 10000]
Out[3]= {0.21, 0.6636, 0.89294, 0.382392, 0.944674,
        0.209062, 0.66142, 0.895775, 0.37345,    . . . }
In[4]:= rh=RegularHisto[nl, 0, 1, 80]
In[5]:= p=Plot[1/(Pi*Sqrt[(x(1-x))]),{x,0.001,0.999},
                              PlotRange->{0, 5.6}]
In[6]:= Show[rh, p, PlotRange->{0,5.6}]
Out[6]= -Graphics-
```



By a lucky coincidence, this density actually can be calculated explicitly if we observe that, by direct substitution, the map

$$g(x) = \psi(f(\psi^{-1}(x))), \qquad x \in [0, 1]$$

where $\psi(x) = (2/\pi)\arcsin\sqrt{x}$, is the tent map $1 - 2|x - 1/2|$, for which the invariant measure is simply the length (Lebesgue) measure on the unit interval; its density has constant value equal to 1. Changing variables back we obtain for our logistic map the invariant measure density of the form

$$h(x) = \frac{1}{\pi\sqrt{x(1-x)}}, \qquad x \in [0, 1]. \tag{14}$$

Its shape was suggested by our experiment. Moreover, this invariant measure is also ergodic.

The theoretical question of when the invariant measure is independent of the starting point is answered in the celebrated Ergodic Theorem first proved in 1930s by the American mathematician George Birkhoff.

Ergodic Theorem. *If $f : S \rightarrow S$ is a dynamical system and if the invariant measure $A_\infty(T, x_0)$ is ergodic for some $x_0 \in S$, then the set T_0 of points $x \in S$ such that $A_\infty(T, x) = A_\infty(T, x_0)$ is of full A_∞ measure, i.e., $A_\infty(T_0, x_0) = 1$.*

Conversely, if $A_\infty(T, x)$ is almost everywhere constant, then $A_\infty(T, x)$ is an ergodic measure.

As we saw earlier, the rational rotations of the unit circle are not ergodic if one takes the normalized arc-length measure as the invariant measure (see Remark 6.3.1); on the other hand, the irrational rotations are ergodic.

Remark 6.3.2 *Ergodic Hypothesis and the Recurrence of the Universe.* In general, it is very difficult to check ergodicity of many complex physical dynamical systems. Ever since the days of Boltzmann, Gibbs and Poincaré, who first posed the so-called *ergodic hypothesis* at the end of the 19th century, the general ergodic hypothesis remains unsolved. In those days the hypothesis caused fierce philosophical debates because it implies a *recurrence property* of dynamical systems. Consider two recurrence properties:

(a) for each $x \in S$ the orbit $x, f(x), f^2(x), \ldots,$ returns arbitrarily close to x, infinitely many times.

(b) There are infinitely many ns such that, for each $x \in S$, $f^n(x)$ is close to x.

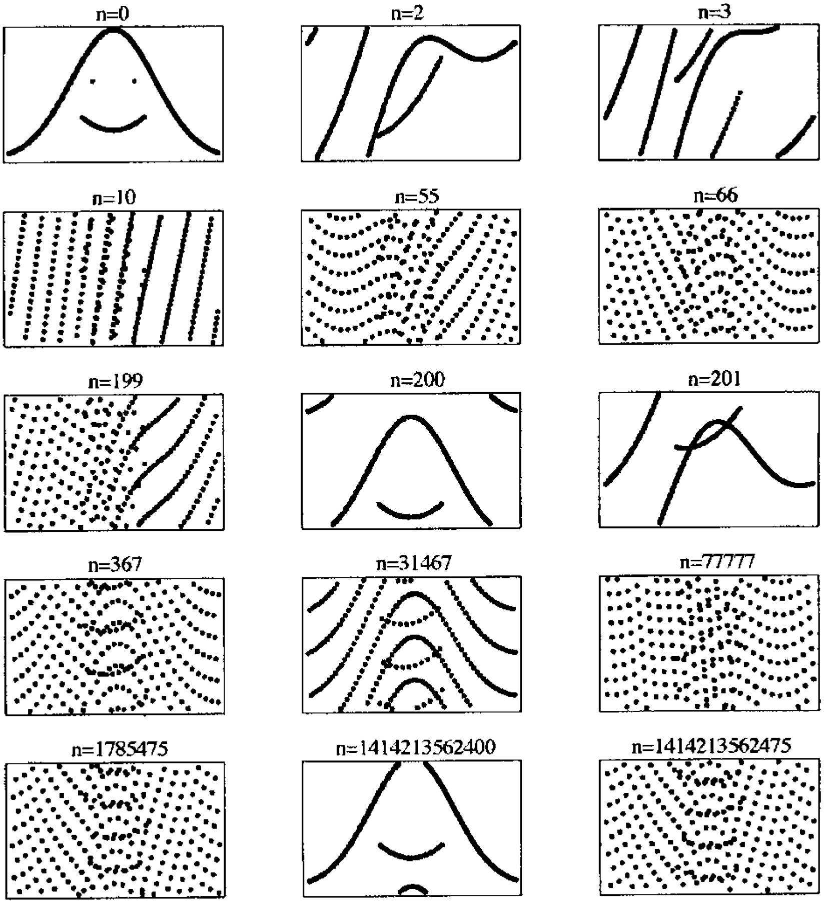
We know that the ergodic theorem implies the recurrence property (a), which is weaker than (b). If the recurrence property (b) were a consequence of the ergodic hypothesis, then the iterations of the map f itself would return infinitely often arbitrarily close to the identity map. In other words, the whole system would return infinitely often arbitrarily close to the original state—a shocking conclusion if applied to, say, the whole universe.

The irrational rotation of the unit circle satisfies the recurrence property (b). Indeed, if we start, say, at the North Pole, and select a small neighborhood arc of the North Pole of length $\epsilon > 0$, then the Ergodic Theorem guarantees that we will return infinitely often to this neighborhood and, because of its isometric property, every point will return to its ϵ-neighborhood at the same time. A point spends $\epsilon/2\pi > 0$ fraction of eternity in its ϵ-neighborhood.

Mathematica Experiment 4. StoGho Returns. This experiment illustrates the recurrence phenomenon for an irrational linear map on the unit square (mod 1) (thought of as an irrational rotation of the 2-dimensional unit torus). The starting set being mapped is marked by black pixels (it also happens that the original state produces a visage of StoGho). After n=10 iterations (say, years) StoGho is deconstructed and reaches a state that looks quite random and chaotic. However, if you wait n = 200 years, then StoGho will return to life in a state recognizably close to the original picture. He comes back over and over, always a little different, only to be dissolved into randomness again. After n = 1414213562400 years he will be back in a state pretty close to the original, although, by then, he will have lost his eyesight and grown a beard.

```
In[1]:= hair=Table[{i-1}*0.005,E^(-((i-101)*0.005)^2/(2/25)),
                                               {i,1,201]};
        mouth=Table[{0.01i,0.22+(0.1/225)(i-50)^2,{i,35,65}];
        eyes={{0.4,0.6},{0.6,0.6}};
        stogho=Join[hair,mouth,eyes];
        lp0=ListPlot[stogho,PlotStyle->PoinSize[0.02],
                Frame->True, FrameTicks->None,
                FrameLabel->{"","","n=0",""}]
Out[1]= -Graphics-
In[2]:= k=2;
        hairtrans=Table[{0.005*i,Mod[stogho[[i]][[2]]+
                        Sqrt[2]*k+k*0.005*i,1],{i,1,201}];
        mouthtrans=Table[{0.34+0.01*i,Mod[mouth[[i]][[2]]+
                        Sqrt[2]*k+ k*0.01*i,1],{i,1,31}];
        eyestrans= {{0.4,Mod[eyes[[1]][[2]]+Sqrt[2]*k+
                     k*0.01*i,1]},
                    {0.6,Mod[eyes[[2]][[2]]+Sqrt[2]*
                     k+k*0.01*i,1]};
        stoghotrans =Join[ hairtrans ,mouthtrans ,eyestrans ];
        lp2=ListPlot[stoghotrans , PlotStyle->PointSize[0.02],
                Frame->True,FrameTircks->None,
                FrameLabel->{"","","n=2",""}]

Out[2]= -Graphics-
. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
In[16]:= Show[GraphicsArray[ {
                {lp0,lp2,lp3},
                {lp10,lp55,lp66},
                {lp199,lp200,lp201},
                {lp367,lp31467,lp77777},
                {lp1485475,lp1414213562400, lp1414213562475}}]
Out[16]= -GraphicsArray-
```



A Mathematical Aside. Ergodic Theorem for Test Functions. The Ergodic Theorem has an extension to averages more general than the frequencies of visits to various sets. The generalization uses the idea of a test function, already encountered in Example 6.1.4, where the Bernoulli random variables were produced from iterations of the map $2x \pmod 1$ of the unit interval by superposing them with the test function ϕ which tested whether the orbit was above the level 1/2 or below it.

Ergodic Theorem for General Test Functions. *If $f : S \to S$ is an ergodic dynamical system with invariant measure μ and $\phi : S \to \mathbf{R}$ is a (semicontinuous[5]) test function, then the set of starting points $x \in S$ for which*

$$\lim_{n\to\infty} \frac{\phi(x) + \phi(f(x)) + \ldots + \phi(f^{n-1}(x))}{n} = \int_S \phi(x)\,\mu(dx) \qquad (15)$$

[5] i.e., a monotone limit of continuous functions.

is of measure μ equal to 1. If the invariant measure has a density $h(x)$, then, of course,

$$\lim_{n\to\infty} \frac{\phi(x) + \phi(f(x)) + \ldots + \phi(f^{n-1}(x))}{n} = \int_S \phi(x)h(x)\,dx \tag{16}$$

Selecting $\phi(x) = I_B(x)$, the indicator function of the set B, as the test function gives our original ergodic theorem for relative frequencies of visits. As an illustration of the general result, we will consider two examples.

Example 6.3.1 Law of Large Numbers for Bernoulli Random Variables.
For the map $2x \pmod 1$ of the unit interval, the invariant measure is just the length. Selecting $\phi(x) = I_{[1/2,1]}(x)$ as the test function, the above general ergodic theorem gives that

$$\lim_{n\to\infty} \frac{\phi(x) + \phi(f(x)) + \ldots + \phi(f^{n-1}(x))}{n} = \int_0^1 \mathbf{1}_{[1/2,1]}(x)\,dx = \frac{1}{2}, \tag{16}$$

for almost all (with respect to the length measure) $x \in [0, 1]$. This, of course, is a version of the Law of Large Numbers of Section 5.6[6] since the superpositions $\phi(x), \phi(f(x)), \ldots, \phi(f^{n-1}(x))$ form a sequence of independent Bernoulli random variables on the standard probability sample space [0,1].

Example 6.3.2 Shifts of Binary Strings Revisited.
Let S be the state space of binary strings (see Example 6.1.5) and σ be the shift map on S. By definition, if

$$\boldsymbol{x} = (x_0, x_1, x_2, \ldots) \in S$$

then

$$\sigma^n(x) = (x_n, x_{n+1}, x_{n+2}, \ldots).$$

Thus, each string has the whole orbit of the map embedded in it as its different tails. For example, if

$$\boldsymbol{x} = 010100010101010100000001111010101\ldots$$

then

$$\sigma^9(\boldsymbol{x}) = 101010100000001111010101\ldots$$

[6] Notice, however, that the type of convergence considered here is different than the one considered in Section 5.6.

If we define

$$\phi(x) = \begin{cases} 1 & \text{if } x_0 = 1; \\ 0 & \text{if } x_0 = 0; \end{cases}$$

then the average

$$A_n(\phi, x) = \frac{\phi(x) + \phi(f(x)) + \ldots + \phi(f^{n-1}(x))}{n} \tag{17}$$

is equal to the relative frequency of 1s in the first n bits of the string x.

If x is a string representing the Champernowne number

$$0\ 1\ 10\ 11\ 100\ 101\ 110\ 111\ 1000\ \ldots,$$

and if $n = 2 + 2^2 + 2^3 + \ldots + 2^k = 2^{k+1} - 1$, then it is easy to see that among the first n digits $x_0, \ldots, x_{n-1}$ there are exactly as many 0s as 1s, so the average

$$A_n(\phi, x) = 1/2.$$

With a little more effort one can show that, actually, $\lim_{n\to\infty} A_n(\phi, x) = 1/2$ or, in more generality, that any finite string of zeros and ones has the same frequency. If we think of the infinite binary strings as representing real numbers in the interval [0,1], then the length measure $\mu(B) = |B|$ is the invariant measure of the shift map σ, and, by the Ergodic Theorem, for almost all (with respect to μ) strings x the limit relative frequency of ones (and thus also zeros) is equal to 1/2. This is the Equipartition Theorem for binary representations of real numbers. The same can be said about other, say, decimal representations of real numbers.

6.4 Stability of fluctuations and the central limit theorem

In the previous section we established the ergodic behavior in certain simple dynamical systems: for large "times", the relative frequencies of visits to certain sets of states stabilized at what we called the invariant measures of those sets. The ergodic behavior was an analogue of the law of large numbers valid for sequences of independent random variables in probability theory. In this section, we will pursue this analogy a bit further by investigating the central limit behavior in dynamical systems, i.e., the stability of the distribution of fluctuations of the relative frequencies as they approach their ergodic limit. It is a much more rare event than the ergodic behavior, but it is still revealing that such a "second-order" attribute of randomness can be encountered at all in a simple deterministic dynamical system.

We shall begin by putting the familiar central limit theorem for Bernoulli random variables in the dynamical systems context.

Example 6.4.1 Central Limit Theorem for Bernoulli Random Variables.
Again, consider the map $f(x) = 2x \pmod 1$ of the unit interval $S = [0, 1]$ with the length as the invariant measure, and the test function $\phi(x) = \mathbf{1}_{[1/2,1]}(x)$. Considered as random variables on the generic sample space $\Omega = [0, 1]$, the superpositions $\phi(x), \phi(f(x)), \ldots, \phi(f^{n-1}(x)), \ldots$, form a sequence of independent, identically distributed Bernoulli random variables taking values 0 and 1 with probability 1/2 each. The Ergodic Theorem of Section 6.3, Example 6.3.1 (alternatively, the strong version of the Law of Large Numbers of Section 5.6) assures us that the averages

$$A_n(\phi, x) = \frac{\phi(x) + \phi(f(x)) + \ldots + \phi(f^{n-1}(x))}{n} \longrightarrow \frac{1}{2} \tag{1}$$

as $n \to \infty$, for almost all (with respect to the length measure) $x \in [0, 1]$. The limit constant 1/2 here plays a dual role: the integral $\int_S \phi(x)\,dx$ of the test function ϕ against the invariant measure (as promised by the general ergodic theorem of Section 6.3) and the common expectation μ of the Bernoulli random variables (as promised by the law of large numbers).

At this point it is the Central Limit Theorem (Theorem 5.7.1) that helps us to establish stability of fluctuations around the limiting mean. It says that if you subtract the limit $A_\infty(\phi, x) = 1/2$ from the mean $A_n(\phi, x)$ and then normalize it by its standard deviation (square root of variance), then, for large ns , the thus standardized random variable has a distribution close to $N(0, 1)$ standard Gaussian distribution, i.e., for each y, $-\infty < y < \infty$,

$$P\left(x : \frac{A_n(\phi, x) - 1/2}{\sqrt{\operatorname{var} A_\infty(\phi, x)}} < y\right) \longrightarrow \int_{-\infty}^{y} \frac{e^{-z^2/2}}{\sqrt{2\pi}}\,dz. \tag{2}$$

In this case, in view of the additivity of the variance for sums of independent random variables (Theorem 5.4.4), the variance in the above formula is easily calculated to be $(1/4)/n$, so that finally we get

$$P\left(x : \frac{A_n(\phi, x) - 1/2}{(1/2)/\sqrt{n}} < y\right) \longrightarrow \int_{-\infty}^{y} \frac{e^{-z^2/2}}{\sqrt{2\pi}}\,dz, \tag{3}$$

where the probability measure P is simply the Lebesgue length measure on subsets of the unit interval.

If one contemplates the possibility of a similar behavior for other dynamical systems on the unit interval, then the first thing of concern is that, in general, the iterates are not statistically independent, and the simple calculation of the variance of $A_n(\phi, x)$ for Bernoulli random variables has to be replaced by a more elaborate procedure. However, as it turns out, at least in certain cases, a result of the central limit theorem type of the form

$$F_n(y) = \mu\left\{x : \frac{A_n(\phi, x) - A_\infty(\phi, x)}{s_n} < y\right\} \longrightarrow \Phi(y) \equiv \int_{-\infty}^{y} \frac{e^{-z^2/2}}{\sqrt{2\pi}}\, dz. \tag{4}$$

is still valid. Here, $\mu(T) = A_\infty(T, x) = A_\infty(T)$ is the unique absolutely continuous ergodic invariant measure for the map $f : S \to S$, and

$$A_\infty(\phi) = \int_S \phi(x)\, \mu(dx), \tag{5}$$

and the variance

$$s_n^2 = \int_S (A_n(\phi, x) - A_\infty(\phi))^2\ \mu(dx) \tag{6}$$

$$= \frac{1}{n^2}\int_S \left(\phi(x) + \phi(f(x)) + \ldots + \phi(f^{n-1}(x)) - nA_\infty(\phi)\right)^2\ \mu(dx).$$

Remember that if the invariant measure μ has a density with respect to the Lebesgue measure on the unit interval, that is, $\mu(dx) = h(x)\,dx$ for a certain nonnegative function $h(x)$, then

$$F_n(y) = \mu\left\{x : \frac{A_n(\phi, x) - A_\infty(\phi, x)}{s_n} < y\right\} = \int_D h(x)\, dx, \tag{4a}$$

where the integration domain D is the set of x satisfying the condition spelled out inside the braces. Similarly,

$$A_\infty(\phi) = \int_S \phi(x)\, h(x)\, dx, \tag{5a}$$

etc.

Although simple additivity of identical variances is not valid in general, for certain systems one can prove that, asymptotically, as $n \to \infty$,

$$s_n \sim \frac{s}{\sqrt{n}} \tag{7}$$

where the constant ("asymptotic variance")

$$s^2 = \int_S (\phi(x) - A_\infty(\phi))^2 \mu(dx) + \tag{8}$$

$$+2\sum_{n=1}^{\infty}\int_S(\phi(x)-A_\infty(\phi))(\phi(f^n(x))-A_\infty(\phi))\mu(dx)<\infty.$$

For some details, see the discussion at the end of this section. So, the plausible stability-of-fluctuations effects we should look for are of the following form: for any y, $-\infty < y < \infty$, as $n \to \infty$,

$$F_n(y)=\mu\left\{x:\frac{\phi(x)+\phi(f(x))+\ldots+\phi(f^{n-1}(x))-nA_\infty(\phi)}{s\sqrt{n}}\leq y\right\}$$

$$\longrightarrow \Phi(y)\equiv\int_{-\infty}^{y}\frac{e^{-z^2/2}}{\sqrt{2\pi}}\,dz. \tag{9}$$

Mathematica Experiment 1. Central Limit Theorem for the Tent Map. Consider again the tent map

$$f(x)=1-|\beta(x-\alpha)|, \qquad \alpha=1-\beta^{-1} \tag{10}$$

on the unit interval $S=[0,1]$. It turns out that for $\beta>\sqrt{2}$, the dynamical system generated by the map f satisfies the Central Limit Theorem in the form (9). To verify this fact by numerical experimentation, we shall select, as in *Mathematica* Experiment 6.3.2,

$$\beta=a=(1+\sqrt{5})/2\approx 1.6180\ldots$$

so that $\beta^2-\beta-1=0$, and the critical points of this map have a particularly simple orbit

$$f:0\mapsto\alpha\mapsto 1\mapsto 0.$$

This fact permits explicit calculation of the invariant density, moments, and the asymptotic variance s^2 needed in the study of stability of fluctuations described by formula (9). In particular, the invariant density

$$h(x)=\begin{cases}(1+\alpha)^{-1}\approx 0.7236, & \text{if } 0\leq x\leq\alpha\approx 0.3820;\\ \beta(1+\alpha)^{-1}\approx 1.1708, & \text{if } 0.3820\approx\alpha<x\leq 1.\end{cases}$$

We will conduct the experiment for the test function $\phi(x)=x$, so, in calculation of the limit variance, we will need the first moment of the invariant density

$$A_\infty(\phi)=\int_0^1 xh(x)\,dx=\frac{\alpha^2}{2}\frac{1}{1+\alpha}+\left(\frac{1}{2}-\frac{\alpha^2}{2}\right)\frac{\beta}{1+\alpha}\approx 0.5528,$$

so that

$$(A_\infty(\phi))^2=\approx 0.3056.$$

The second moment

$$\int_0^1 x^2 h(x)\,dx = \frac{\alpha^2+1}{3} =\approx 0.3812,$$

and, finally, the individual variances

$$s_0^2 = \int_0^1 x^2 h(x)\,dx - A_\infty^2(\phi) \approx 0.0756.$$

In this experiment it is not practical to compute the asymptotic variance from (8); if you try it, you will find some limitations of this procedure and also of *Mathematica*. Instead, we use formulas (6) and (7) to obtain

$$s^2 \sim ns_n^2 = \frac{1}{n}\int_0^1 \Big(x + f(x) + f^2(x) + \ldots + f^{n-1}(x) - nA_\infty(\phi)\Big)^2 h(x)\,dx.$$

This integral can be computed numerically for small values of n (say, $n = 8, 9, \ldots$) using the *Mathematica* package `<<NumericalMath'ListIntegrate'`. One finds that

$$s^2 \approx \begin{cases} 0.0115, & \text{for } n = 8; \\ 0.0108, & \text{for } n = 9. \end{cases}$$

The algorithm we will employ is as follows:

(1) First, for $n = 1, 2, \ldots, 25$, we compute by numerical integration approximate standard deviations $s[n]$ of the averages

$$\frac{x + f(x) + \ldots + f^{n-1}(x)}{n},$$

by sampling it at 100 points x in the interval [0,1].

(2) Then, for $n = 25$, and for 100 starting points $x = 0.01, 0.02, \ldots, 0.99, 1.00$, we shall find the values of the standardized average

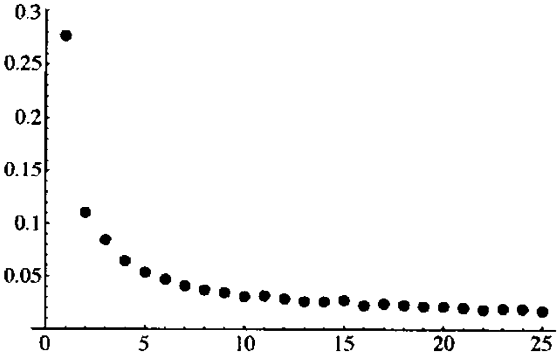
$$\text{Ave}\,(x) = \frac{n^{-1}(x + f(x) + \ldots + f^{n-1}(x)) - 0.5528}{s[25]},$$

and plot the cumulative distribution function of these data remembering the shape of the underlying invariant density $h(x)$: at the data point Ave (x) corresponding to $x \leq \alpha \approx 0.3820$ the cumulative d.f. will jump up $\approx$ $(1/100) \cdot 0.7236$, and at the data point corresponding to $x > 0.3820$, the cumulative d.f. will jump up ≈ $(1/100) \cdot 1.1708$. More precisely, the approximate cumulative d.f. will be calculated from the formula

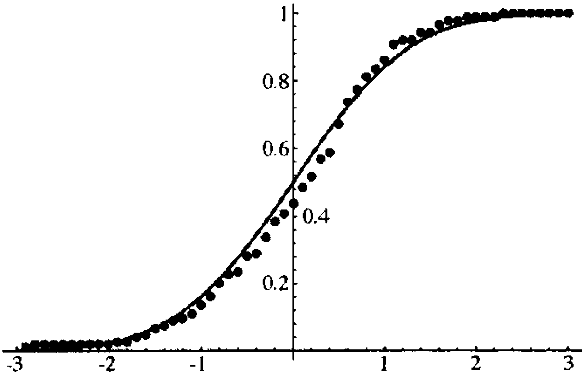
$$F(y) = \sum_{k=1}^{100} H\Big(y - \text{Ave}\,(0.01 \cdot k)\Big) \cdot \frac{h(0.01k)}{100},$$

where $H(y)$ is the Heaviside unit step function, $= 0$ for negative y and $= 1$ for positive y. Finally, the result will be compared with the standard Gaussian cumulative d.f. $\Phi(y)$.

```
In[1]:= b=N[(1+Sqrt[5])/2]
Out[2]= 1.61803
In[2]:= a=1-1/b
Out[2]= 0.381966
In[3]:= h[x_]:=If[x<a,1/(1+a),b /(1+a)]
In[4]:= NIntegrate[x*h[x], {x,0,1}]
Out[4]= 0.5527
In[5]:= f[x_]:= 1- Abs[b(x-a)]
In[6]:= Iter[x_,n_]:=NestList[(f[#])&, x, n-1]
In[7]:= Iter[0.3,10]
Out[7]= {0.3, 0.867376, 0.21459, 0.72918, 0.438197,
          0.909017, 0.147214, 0.620163,  0.61459, 0.623607}
In[8]:= s[n_]:= Sqrt[ Sum[
        (Apply[Plus,Iter[0.01*k,n]]*n^(-1)-0.5528)^2*
        h[0.01*k]*(1/100),{k,1,100}] ]
In[9]:= stab=Table[s[n],{n,1,25}];
Out[9]= {0.276467, 0.110703, 0.0846131,  . . . ,
              0.0186036, 0.0198077, 0.0192055,  0.0174124}
In[10]:= ListPlot[ stab, PlotRange->{0,0.3}]
Out[10]= -Graphics-
```



```
In[11]:= sd=s[25]
Out[11]= 0.0174124
In[12]:= Ave[x_]:= (Apply[Plus, Iter[x,25]]/25-0.5528)/sd
In[13]:= F[y_]:=Sum[ If[y-Ave[0.01*k]<0,0,1]*
                                  h[0.01*k]*(1/100) , {k,1,100}]
In[14]:= Fcdf= ListPlot[Table[{-3+0.1*k,F[-3+0.1*k]},
                             {k,1,60 }]]
Out[14]= -Graphics-
In[15]:= Phi[y_]:=(1/Sqrt[2*Pi])*NIntegrate[E^(- x^2/2),
                         {x,-Infinity,y}]
In[16]:= Phicdf=Plot[Phi[y],{y,-3,3}]
Out[16]= -Graphics-
In[17]:= Show[Fcdf,Phicdf]
Out[17]= -Graphics-
```



A Mathematical Aside. CLT for Hyperbolic Systems. Besides the above system generated by the tent map, many other ergodic dynamical systems display stability of fluctuations effects, and satisfy the Central Limit Theorem. The so-called *hyperbolic maps* of the state space S in the n-dimensional Euclidean space $\mathbf{R}^d$ are one such category. A twice differentiable map $f : S \to S$ is called *hyperbolic* in a subset $T \subset S$ if, for any point $x \in T$, the Jacobian matrix

$$\left(\frac{\partial f_i(x)}{\partial x_j}\right)_{1 \le i,j \le n}$$

has eigenvalues of modulus not equal to 1, and if f maps T into T. A general dynamical system f on S is called *hyperbolic* if there exists a compact subset $T \subset S \subset \mathbf{R^d}$ on which f is twice differentiable and hyperbolic, and such that for any point $x \notin T$,

$$\lim_{n \to \infty} f^n(x) \in T.$$

Example 6.4.2
Let

$$S = \{(x, y) : 0 \leq x < 1, 0 \leq y < 1\}$$

denote the unit square and let

$$M = \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix},$$

as in Example 6.1.1. This is a linear map on $\mathbf{R}^2$ with eigenvalues

$$\lambda_1 = \sqrt{2}, \qquad \lambda_2 = -\sqrt{2}.$$

M also induces a map on the torus S, given by the formula

$$f((x, y)) = (u \text{ (mod } 1), v \text{ (mod } 1))$$

where $M(x, y)^T = (u, v)$. Clearly, the derivative of f at every point $(x, y) \in S$ is the matrix M itself. Hence, the map is hyperbolic.

The Central Limit Theorem behavior (9) for hyperbolic maps can be proved for *Lipschitz continuous* test functions $\phi : S \to \mathbf{R}$, that is, for maps for which the differential ratio $(f(x) - f(y))/|x - y|$ remains bounded over the state space, but the proof is difficult and well beyond the scope of this text (see the Bibliographical Notes for further references). However, the constants appearing in its formulation (8, 9) easily can be justified analytically.

To begin with, the constant A is the common average of all the iterations f^i of f since, in view of the integral's invariance with respect to the invariant measure,

$$A = \int_S \phi(y)\, \mu(dy) = \int_S \phi(f^i(y))\, \mu(dy).$$

To justify the formula (8) for the variance v^2, at least asymptotically, one needs to observe that

$$v_n^2 = \int_S \Big(\phi(y) + \phi(f(y)) + \ldots + \phi(f^{n-1}(y)) - nA \Big)^2 \mu(dy)$$

$$= \sum_{k=0}^{n-1} \int_S \Big(\phi(f^k(y)) - A \Big)^2 \mu(dy)$$

$$+\sum_{0\le l,k\le n-1,k\ne l}\int_S\Big(\phi(f^k(y))-A\Big)\Big(\phi(f^l(y))-A\Big)\mu(dy).$$

Separating the diagonal and off-diagonal terms in the square, and taking advantage of the invariance of the integral, we see that the above expression

$$=n\int_S\Big(\phi(y)-A\Big)^2\mu(dy)$$

$$+2\sum_{0\le l,k\le n-1,l-k=1}\int_S\Big(\phi(y)-A\Big)\Big(\phi(f(y))-A\Big)\mu(dy)$$

$$+2\sum_{0\le l,k\le n-1,l-k=2}\int_S\Big(\phi(y)-A\Big)\Big(\phi(f^2(y))-A\Big)\mu(dy)$$

....

$$+2\sum_{0\le l,k\le n-1,l-k=n-1}\int_S\Big(\phi(y)-A\Big)\Big(\phi(f^{n-1}(y))-A\Big)\mu(dy)$$

$$=n\int_S\Big(\phi(y)-A\Big)^2\mu(dy)$$

$$+2\sum_{k=1}^{n-1}(n-k)\int_S\Big(\phi(y)-A\Big)\Big(\phi(f^k(y))-A\Big)\mu(dy).$$

If the series representing v^2 in the formulation of the Central Limit Theorem converges, then it is reasonable to assume that

$$v_n\sim v\sqrt{n}.$$

Example 6.4.2

Consider the automorphism of the torus considered in Example 6.4.2. The projection $\phi((x,y))=x$ is a Hölder continuous function. Since the Lebesgue measure is the invariant measure, the mean $A_\infty(\phi)$ is easily computable to be

$$A = \int_0^1 \int_0^1 y dx dy = 1/2.$$

The assertion of the above Central Limit Theorem holds true also in some other cases which are not necessarily hyperbolic. One such class consists of piecewise invertible maps on an interval which are twice differentiable, with the derivative larger than 1, and which admit an absolutely continuous invariant measure μ. Such maps are called Lasota-Yorke maps. The tent map is an example of such a map.

Another class of maps for which the Central Limit Theorem holds true is a class that admits an absolutely continuous invariant measure μ and which is related to so called expanding maps. A map f is called *expanding* if it maps open sets into open sets and if it expands distances, that is, if there exists an expanding constant $L > 1$ such that, for any sufficiently close pair $x, y \in S$, we have

$$\text{dist } (f(x), f(y)) \geq L \text{ dist } (x, y).$$

Example 6.4.3
The map $z \mapsto z^2 + 5$ in the complex plane **C** has an expanding repeller J. The map restricted to the repeller is expanding and satisfies the Central Limit Theorem.

There is a lot of research activity going on in this area. For example, as of this writing, the latest word on the familiar logistic system $f_a(x) = ax(1 - x)$, $0 < a \leq 4$, is that the mapping f_a has an attracting cycle, and thus is hyperbolic, for an open and dense set of parameters a, see the *Annals of Mathematics* paper by Jacek Graczyk and Grzegorz Świątek cited in the Bibliographical Notes.

6.5 Attractors, fractals, and entropy

Even simple dynamical systems $f : S \to S$ can have a very "strange" behavior. In particular, *attractors* of dynamical systems, loosely defined here as a set of states where the orbits $\{f^n(x)\}$ accumulate for large times n, can have fractal structure. This phenomenon had been known in mathematics for more than 50 years, but its study from the viewpoint of physical, and other science and technology applications, is much more recent.

Example 6.5.1 Cantor Set as an "Attractor". Consider a map of the unit interval pictured in Fig. 6.5.1.

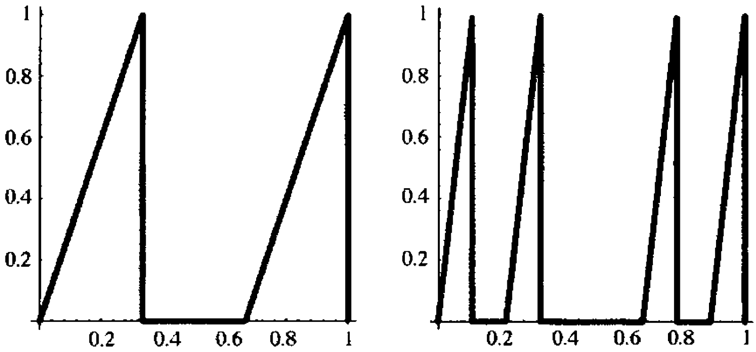


FIGURE 6.5.1

The graph of the map $f : [0, 1] \to [0, 1]$ *defined by formula (1) and its first iterate* $f^2(x) = f(f(x))$. *The sets where the iterates do not vanish form an approximation to the Cantor set.*

and defined by the formula

$$f(x) = \begin{cases} 3x & \text{for } 0 \le x \le 1/3; \\ 0 & \text{for } 1/3 < x < 2/3; \\ -2 + 3x & \text{for } 2/3 \le x \le 1. \end{cases} \tag{1}$$

Its first iterate $f^2(x) = f(f(x))$ is also shown in Fig. 6.5.1. It is clear that the sets of states $C_n, n = 1, 2, \dots,$ where the iterations $f^n(x), n = 1, 2, \dots,$ do not vanish, form approximations to the fractal Cantor set C discussed in Section 2.7.

In general, an *attractor* of $f : S \to S$ is an invariant subset $T \subset S$ (i.e., $f(T) \subset T$) such that there exists an open set $O \subset S$ satisfying the following three conditions:

1. $T \subset O$
2. $f(O) \subset O$
3. $\bigcap_n f^n(O) = T$

Similarly, a *repeller* set $T \subset S$ is defined by the above condition 1, and conditions

2'. $f^{-1}(O) \subset O$

3'. $\bigcap_n f^{-n}(O) = T$

If the system has a repeller (or an attractor) T, and if it is sensitive to initial conditions (that is, has a positive Liapunov exponent) in the neighborhood of T, then the repeller (attractor) is called *strange.*

The presence of a strange attractor in a dynamical system can be viewed as evidence of self-organization of the system at the post–transient stage, i.e., on the attractor T; on the latter the system has a lower number of degrees of freedom than the whole system. It is thus attracted to a lower dimensional phase space, and the dimension of this reduced phase space represents the number of active degrees of freedom in the self-organized system. In particular, it is of interest to determine this dimension.

Example 6.5.2 Attractor of the Hénon map. Consider the map $f : \mathbf{R}^2 \to \mathbf{R}^2$ defined by the formula

$$f((x, y)) = \Big(1 - 1.4x^2 + y,\ 0.3x\Big). \tag{2}$$

Fig. 6.5.2(i) shows the orbit of 100,000 iterates of the Henon map with the starting point at $(0, 0)$, while Fig. 6.5.2(ii) shows the same number of iterates but starting at $(-1, 0)$.

A more extensive experimentation would show that the orbits starting at arbitrary points will either converge to the structure shown on Fig. 6.5.2(i) or would diverge to infinity. The exact mathematical nature of this strange attractor is still unknown. However, we will take an experimental and statistical approach to this problem and will determine the correlation dimension (see Section 2.7) of the attractor (which for strange attractors and repellers is the same as the Hausdorff dimension).

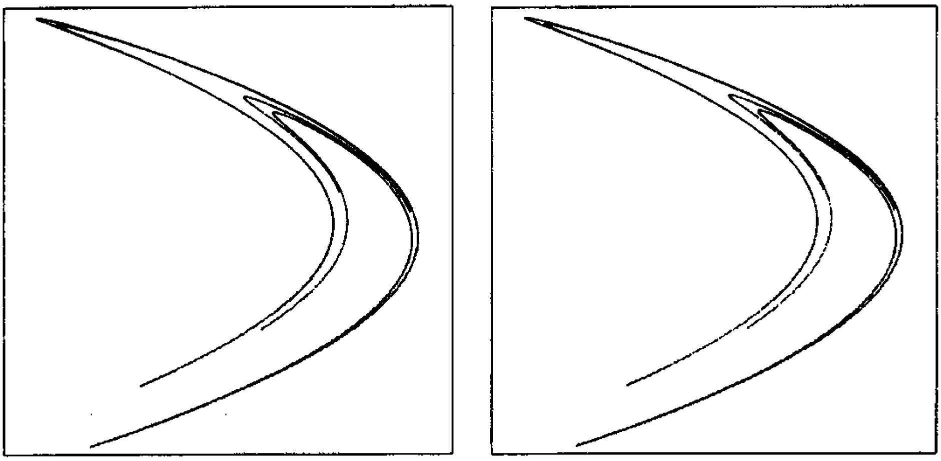


FIGURE 6.5.2
(i) Orbit of 100,000 iterates of the Henon map with the starting point at $(0, 0)$, *(ii) shows the same number of iterates but starting at* $(-1, 0)$.

Mathematica Experiment 1. Correlation Dimension of the Hénon Map. For a

(Gibbs) measure μ, the correlation integrals are defined as follows:

$$C(\epsilon) = \int \mu(B(x,\epsilon))\mu(dx),$$

where ϵ is any positive number and where $B(x,\epsilon)$ denotes the ball of radius ϵ and center at x. The quantity $\mu(B(x,\epsilon))$ is the measure of all points y which are at a distance less than ϵ from x. The *Grassberger-Procaccia correlation dimension* of μ is defined by the formula

$$d_{cor} = \lim_{\epsilon\to 0} \frac{\log C(\epsilon)}{\log \epsilon},$$

whenever this limit exists. Heuristically, it means that $C(\cdot)$ is a function of the form

$$C(\epsilon) = K\epsilon^{d_{cor}} + \text{lower order terms},$$

so that, to find the correlation dimension it seems reasonable to estimate the slope of the regression line in log-log coordinates. Details of this procedure have been explained in Section 2.7.

Now, consider a finite segment

$$T = \{x_0 = x,\ x_1 = f(x),\ x_2 = f^2(x), \ldots, x_{n-1} = f^{n-1}(x)\}$$

of an orbit of the dynamical system $f : S \to S$. For a finite set $\{\epsilon_1, \epsilon_2, \ldots, \epsilon_m\}$ of (different) radii, we will find the best linear fit for the data set

$$\Big\{(\ln\epsilon_1, \ln C^n(\epsilon_1)),\ (\ln\epsilon_2, \ln C^n(\epsilon_2)), \ldots,\ (\ln\epsilon_m, \log C^n(\epsilon_m))\Big\}.$$

The estimated slope

$$\hat{d} = \frac{\sum_{i=1}^m \log\epsilon_i \log C^n(\epsilon_i) - m\bar{\epsilon}\overline{\log C^n(\epsilon)}}{\sum_{i=1}^m r_i^2 - m\bar{\epsilon}^2},$$

where

$$\bar{\epsilon} = \frac{1}{m}\sum_{i=1}^m \epsilon_i,$$

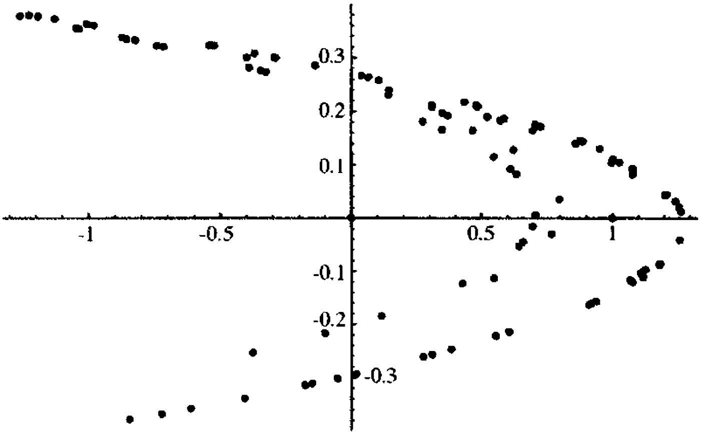
$$C^n(\epsilon, T) = C^n(\epsilon) = \frac{\#\{(x_i, x_j) : |x_i - x_j| < \epsilon, 0 \le i, j \le n-1\}}{n^2}$$

and

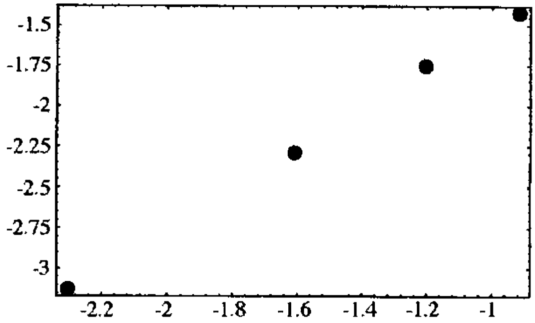
$$\overline{\log C^n(\epsilon)} = \frac{1}{m}\sum_{i=1}^{m}\log C^n(\epsilon_i),$$

will serve as an estimator of the correlation dimension d_{cor}. The experiment is based on 100 iterations, a rather low number. However, the result $\hat{d} = 1.23893$ is quite satisfactory compared to the true value, which is not known exactly, but lies around 1.25.

```
In[1]:= <<Statistics`LinearRegression`
In[2]:= f[{x_,y_}]:= {1-1.4 x^2 + y, .3 x}
In[3]:= hen=NestList[f,{0,0},100]
In[4]:= ListPlot[hen, PlotStyle->PointSize[0.01]]
Out[4]= -Graphics-
```



```
In[4]:= c[r_]:= (1/100.)^2 Sum[ Sum[ If[ Sum[
          (hen[[i]][[k]]-hen[[j]][[k]])^2, {k,1,2}]<r^2,1,0],
          {j,1,100}],{i,1,100}]
In[5]:= reg= Table[{Log[i/10.], Log[c[i/10.]]}, {i,1,4}];
In[5]:= ListPlot[reg]
Out[5]= -Graphics-
```



```
In[6]:= Fit[reg, {1,x},x]
Out[6]= -0.279958 + 1.23893 x
```

Example 6.5.3 The Ikeda Map.

The Ikeda map is derived from a model of the plane wave interaction field in the optical ring laser. The defining map $f : \mathbf{R}^2 \mapsto \mathbf{R}^2$ can be represented by the formula:

$$f((x, y)) = \Big(0.97 + 0.9(x\cos\tau - y\sin\tau),\ 0.9(x\sin\tau + y\cos\tau)\Big), \qquad (3)$$

where

$$\tau = 0.4 - \frac{6.0}{1.0 + x^2 + y^2}. \qquad (4)$$

It has a strange attractor whose fractal dimension is ≈ 1.7 (see Fig. 6.5.3).

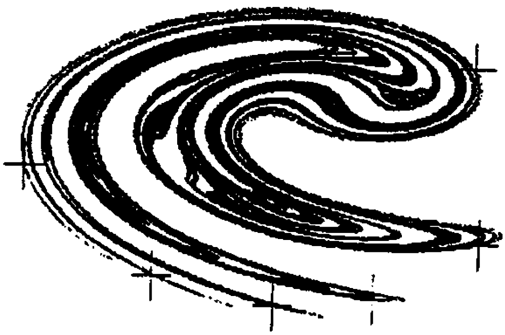


FIGURE 6.5.3
The chaotic strange attractor of the Ikeda map (3-4).

Thus far, we have viewed the chaotic behavior of the system, with its sensitive (exponential) dependence on initial conditions, through the existence of positive

Liapunov exponents and strange attractors. However, the sensitive dependence can be seen as a way to create information, and the way to measure the rate of the information contents creation was via the *entropy* (see Section 2.8). Intuitively speaking, the phenomenon is clear. If we work with fixed resolution of our data, the sensitive dependence leads to the situation where indistinguishable states (differing only in insignificant digits) lead to distinguishable states (differing in significant digits). In other words, with each new step the system creates new information.

Example 6.5.4 Entropy of the Shift Map for Binary (and Other) Strings.
Consider again the shift map σ on the state space of all binary strings or, equivalently, the $2x$ (mod 1) map, if binary strings are viewed as representations of real numbers in the unit interval. The invariant measure of the system corresponds to the length (Lebesgue) measure in the unit interval interpretation, or to the underlying probability measure for any sequence of independent identically distributed symmetric Bernoulli random variables. For the single Bernoulli random variable, the entropy (as introduced in Section 2.8) is

$$H(1/2, 1/2) = -\frac{1}{2}\ln\frac{1}{2} - \frac{1}{2}\ln\frac{1}{2} = \ln 2. \tag{5}$$

Thus, for a string of length k, where 2^k possible states are possible, each taken with probability 2^{-k}, the total entropy is (entropy is additive for independent random variables)

$$H(2^{-k}, \dots, 2^{-k}) = -2^k\frac{1}{2^k}\ln\frac{1}{2^k} = k\ln 2. \tag{6}$$

So, the rate of entropy creation per unit digit (time step) is constant and equal to

$$h = \frac{1}{k}H(2^{-k}, \dots, 2^{-k}) = \ln 2. \tag{7}$$

This number is called the *Kolmogorov-Sinai entropy* of the shift dynamical system.

For a general dynamical system $f : S \to S$ admitting an invariant normalized measure μ the approach has to be more subtle and one cannot expect a fixed rate of entropy creation per unit time-step. The Kolmogorov-Sinai approach is as follows:

(i) Consider a partition $\mathcal{P}$ of the state space S into a union of pairwise disjoint sets $P_1, \dots, P_m$, i.e., $P_i \cap P_j = \emptyset$ and

$$S = P_1 \cup \dots \cup P_m,$$

and define its entropy (information content) by the formula

$$H(\mathcal{P}) = -\sum_{i=1}^{m} \mu(P_i) \ln \mu(P_i). \tag{8}$$

For the shift dynamical system of Example 6.5.4, one can take as the initial partition of the unit interval

$$[0, 1) = [0, 1/2) \cup [1/2, 1),$$

so that (with the Lebesgue measure as the invariant measure) its entropy is ln 2 as given by formula (5).

(ii) After the first time step, the map $f : S \mapsto S$ generates a new partition of S

$$f^{-1}(\mathcal{P}) = (f^{-1}(P_1), \dots, f^{-1}(P_m)), \tag{9}$$

which, in the case of the shift system [see the graph of function $f(x) = 2x \pmod 1$] consists of two sets

$$f^{-1}([0, 1/2)) = [1/4, 1/2) \cup [3/4, 1),$$

and

$$f^{-1}([1/2, 1)) = [0, 1/4) \cup [1/2, 3/4).$$

So, the finer, cumulative partition after the first time-step, generated by both the initial partition, and partition after the first time-step, is built of all the sets of the form

$$P_i \cap f^{-1}(P_j), \qquad i, j = 1, \dots, m.$$

We will denote it by $\mathcal{P}^{(1)}$. In the case of the shift system, it consists of four intervals

$$[0, 1/4),\ [1/4, 1/2),\ [1/2, 3/4),\ [3/4, 1).$$

Its entropy, defined by the general formula (8), is, in the particular case of the shift system, equal to

$$H(\mathcal{P}^{(1)}) = 2 \ln 2.$$

(iii) In general, after n time-steps, the map $f : S \mapsto S$ generates a n-th order partition of S

$$f^{-n}(\mathcal{P}) = (f^{-n}(P_1), \dots, f^{-n}(P_m)), \tag{10}$$

and the cumulative (refined) partition of the n-th order

$$\mathcal{P}^{(n)} = P_{i_1} \cap f^{-1}(P_{i_1}) \cap \dots \cap f^{-n}(P_{i_n}), \qquad i_1, \dots, i_n = 1, \dots, m,$$

which, in the case of the shift system, divides the unit interval into 2^{n+1} equal dyadic subintervals. The corresponding entropy is $H(\mathcal{P}^{(n)})$, which is just $n \ln 2$ for the shift system.

(iv) The entropy of the partition $\mathcal{P}$ (the asymptotic rate of creation of information per unit time-step) is defined by the formula

$$h(\mathcal{P}) = \lim_{n\to\infty} \frac{H(\mathcal{P}^{(n)})}{n}. \tag{11}$$

This limit always exists. The Kolmogorov entropy $h(\mu)$ of the system is obtained by taking the largest of possible partition entropies or, more precisely,

$$h(\mu) = \sup h(\mathcal{P}),$$

where the supremum is taken over all partitions $\mathcal{P}$. For a concrete dynamical system, it is often very difficult to prove that there exists a partition for which the supremum is attained. For the Bernoulli shift, this maximizing partition happens to be $[0, 1/2) \cup [1/2)$, so that $h(\mu) = \ln 2$.

A Mathematical Aside: Entropy vs. Fractal Dimension vs. Liapunov Exponents. As could be guessed, all three quantities (entropy, Liapunov exponents, and fractal dimension) are interconnected, although at this points only incomplete information is available.

For example, it is known that if $f : \mathbf{R}^n \mapsto \mathbf{R}^n$ has derivatives of all orders which satisfy the Hölder condition, and the ergodic measure μ for f has a density with respect to the Lebesgue measure, then

$$h(\mu) = \sum_{\lambda_i>0} \lambda_i,$$

that is, the Kolmogorov-Sinai entropy is simply the sum of positive Liapunov exponents which individually determine the rates of exponential growth of the orbits (in, perhaps, different directions).

On the other hand, the *information dimension* of the ergodic measure μ for f acting on (some smooth surface of) $\mathbf{R}^n$, which is defined as

$$\dim_H(\mu) = \inf\{\dim_H T : \mu(T) = 1\}, \tag{12}$$

can be also calculated from the formula

$$\lim_{\epsilon\to 0} \frac{\ln \mu(B_x(\epsilon))}{\ln \epsilon} = \alpha, \tag{13}$$

where $B_x(\epsilon)$ is the ball of radius ϵ centered at x, if α is independent of x. In that case, the mass $\mu(B_x(\epsilon))$ is scaling like ϵ^α, independently of x.[7]

Again, if $f : \mathbf{R}^n \to \mathbf{R}^n$ has derivatives of all orders which satisfy the Hölder condition and the ergodic measure μ for f has a density with respect to the Lebesgue measure, then

$$\dim_H(\mu) = k + \frac{\lambda_1 + \ldots + \lambda_k}{|\lambda_{k+1}|},$$

where

$$k = \max\{i : \lambda_1 + \ldots + \lambda_i > 0\}.$$

As before, here $\lambda_1, \ldots, \lambda_r$ are here the Liapunov exponents of μ.

In the case where a twice smoothly differentiable f acts on a compact surface S and μ is an ergodic measure with Liapunov exponents $\lambda_1 > 0 > \lambda_2$, then the information dimension, entropy, and the Liapunov exponents are related by the *Young formula*:

$$\dim_H(\mu) = h(\mu)\left(\frac{1}{\lambda_1} + \frac{1}{|\lambda_2|}\right).$$

Example 6.5.5 Entropy, Liapunov Exponent and Fractal Dimension for Asymmetric Bernoulli Systems.[8]

Let $S = [0, 1]$, $x \in S$ be considered as a binary sequence, and let $S_p \subset S$ be the subset of the binary strings $x = (x_1, x_2, \ldots)$ for which the relative frequency of 1s is equal to p, $0 < p < 1$. By the equipartition theorem we know that the Hausdorff dimension

$$\dim_H S_{1/2} = 1.$$

For other ps, one can prove that

$$\dim_H S_p = \frac{1}{\ln 2}\Big[-p \ln p - (1-p)\ln(1-p)\Big]. \tag{14}$$

Consider the shift map σ on S_p [that is, $\sigma(x) = 2x \pmod 1$)], and let μ be the invariant normalized measure on S satisfying

$$\mu_p\{x : x_1 = 1\} = p,$$

which, for $p = 1/2$, is the usual Lebesgue measure. The Liapunov exponent

$$\lambda = \ln\frac{d\sigma}{dx} = \ln 2,$$

[7] If $\alpha = \alpha(x)$ depends on x, then one often talks about the *multifractal* structure of the attractor.

[8] Adapted from Ruelle (1989).

and the Kolmogorov-Sinai entropy

$$h(\mu_p) = -p \ln p - (1-p)\ln(1-p).$$

Then, the formula (14) implies that there exists a set $T \subset S$ with $\mu_p(T) = 1$ and the Hausdorff dimension

$$\dim_H(T) = \frac{h(\mu_p)}{\lambda}.$$

The proofs and further literature on this subject can be found in the Bibliographical Notes.

To see the relationship between various concepts of dimension and entropy introduced above, notice that the quantity $C(\epsilon)$ in the above definition of the correlation dimension can be rewritten in the form

$$C(\epsilon) = \lim_{N\to\infty} \frac{1}{N^2}\sum_{i=1}^{N(\epsilon)} k_i^2 = \sum_{i=1}^{N(\epsilon)} \lim_{N\to\infty} \frac{k_i^2}{N^2} = \sum_{i=1}^{N(\epsilon)} f_i^2,$$

where k_i is the number of points in the ith volume element that are within ϵ of each other, $N(\epsilon)$ is the number of ϵ-volume elements in the coverage of S, and f_i is the (limit) relative frequency that an observation falls in the ith volume element. Then, the formula for the correlation dimension can be rephrased

$$d_{cor}(S) = \lim_{\epsilon\to 0} \frac{\ln \sum_{i=1}^{N(\epsilon)} f_i^2}{\ln \epsilon},$$

where the numerator

$$\ln \sum_{i=1}^{N(\epsilon)} f_i^2$$

is known as the *Renyi entropy of order 2*.

If, instead of the Renyi entropy, one uses the ordinary entropy

$$H(\epsilon) = -\sum_{i=1}^{N(\epsilon)} f_i \ln f_i$$

of the above relative frequency distribution, then one arrives at the definition of the *information dimension*

$$d_{inf}(S) = \lim_{\epsilon\to 0} \frac{\ln H(\epsilon)}{\ln(1/\epsilon)}. \tag{4}$$

It is known that

$$d_{cor} \leq d_{inf} \leq d_{cap}.$$

A Mathematical Aside: Entropy vs. Kolmogorov Complexity. There is a remarkable theorem due to the Russian mathematician A.A. Brudno (see Bibliographical Notes) which shows the connection between Kolmogorov complexity and entropy. It turns out that if the infinite string x is the code of an orbit of a dynamical system with respect to a generating partition $\mathcal{P}$, then the Kolmogorov complexity $K(x)$ is equal to the entropy for almost every x.

6.6 Experiments, exercises, and projects

1. For each of the following maps f, and each given point x, determine $f(x)$, $f^2(x)$, and $f^3(x)$.

   (a) $f(x) = \frac{x}{3}(1-x)$; $x = .230835$; $x = .002546096$; $x = .827993$;

   (b) f = the Bernoulli shift map ; $x = 0010111010111101...$;

   (c) $f(x) = \pi \sin x$; $x = 1.3843302$; $x = -.983775$.

2. A particle moves at a constant speed $V = 3\frac{\text{ft}}{\text{sec}}$ in the 3-dimensional space parallel to the x–axis. Determine the map f describing the position of the particle after 1 second.

3. A particle moves at a constant speed $V = 1.27\frac{\text{ft}}{\text{sec}}$ in 3-dimensional space along a parabola $y = x^2 + c$; $z = d$, where c and d are constants. Determine the map f describing the position of the particle after 1 second.

4. For the tent map T_a determine the values a for which 0 is a periodic point with the principal period 3, 4, and 5.

5. Conduct a study of the fluctuations for the logistic model $f(x) = 4x(1-x)$, analogous to the experiment of Section 6.4 for the tent map. Use *Mathematica* package `UVW'DynSyst'`.

6. Conduct a study of the frequency and fluctuations' stability for the map of the unit interval given by the formula $f(x) = 1/(1+x)$. Iterations of this map are *continued fractions*. Find the invariant measure for this map.

7. Show that the rotation map of the unit circle displays no sensitive dependence on the initial conditions.

8. Verify analytically the form of the invariant density given for the tent map in Section 6.3.

9. Verify analytically the form of the invariant density given for the logistic map in Section 6.3.

10. Consider the logistic function $f(x) = ax(1-x)$.

(a) Compute 100 iterates of this function (start with $x_0 = 1/2$, for example) for the following values of a: 0.5, 0.75; 1, 1.5, 2.0, 3.0, 3.25, 3.5, 3.55, 3.83, 4.

(b) Graph each orbit $f^n(x_0)$, $n = 0, \dots, 100$ as a function of n.

(c) In the case $a = 4$, graph orbits of the following starting points: 0.1, 0.25, 0.3, 0.51, 0.51, 0.749.

(d) Produce the orbit diagram for the above system.

11. Check via simulation the validity of the equipartition theorem for irrational rotations of the unit circle. Use three different sets and find relative frequencies of visits to them, for 10 different (randomly chosen) starting points. Use 1000 iterations and take advantage of *Mathematica* "infinite precision" capabilities.

12. Illustrate the Central Limit Theorem for the tent map using randomized starting points rather than the uniformly spaced starting points.

13. Approximate the function $C(r)$ (needed in the correlation dimension calculation) for the map

$$f((x, y)) = (1 - 1.4x^2 + y, .3x).$$

Apply the linear regression procedure from Chapter 1 to the function $\log C(r)$ in order to detect linearity in variable $\log r$. Then use the `CorrelationDimension` command of `UVW'DynSyst'` package to confirm your results.

14. Find experimentally an approximation for the invariant density for the tent map using starting points different than the one used in Section 6.3.

15. Use *Mathematica* to produce Fig. 6.5.1.

16. Study experimentally the equipartition property of the Champernowne numbers in base 2 and 10.

17. Find the correlation dimensional of data `DROPS` from the `UVW Web Site` (also, see Example 1.5.2).

18. Repeat Experiment 6.4.1 for other test functions.

19. *Mathematica Project.* Design an experiment reproducing Fig. 6.2.10.
A number of other experiments and examples are included in the *Mathematica* `UVW'DynSyst'` and `UVW'RandomWalk'` packages included in Appendix E.

20. *Mathematica Project.* Use the mapping $[0,1] \times [0,1] \ni (x, y) \mapsto (x, y+\sqrt{2}+x(\text{mod } 1)) \in [0,1] \times [0,1]$ to illustrate the ergodic behavior. Experiment with other, more complex, mappings of the unit square (both linear and nonlinear) to see if similar effects are encountered.

21. A Non-Mathematica Project. Study the toss of a coin of radius R as a physical dynamical system. Assume that the evolution of the system is described by the Newton dynamics $y''(t) = -g$, $\theta''(t) = 0$, with the initial conditions $y(0) = R$, $y'(0) = V > 0$, $\theta(0) = 0$, $\theta'(0) = \Omega > 0$, where $y(t)$ denotes the height of the coin's center over the soft (no bounces!) and flat landing surface S, and $\theta(t)$ is the angle between the normal to the landing surface and normal to the heads side of the coin. Let us denote by t_0 the time when the edge of the coin touches S.

(a) Solve the above Newton evolution equations.

(b) Find a condition on $\theta(t_0)$ equivalent to the coin landing on the heads side.

(c) Find the region H in the 2-D phase space of the initial conditions $(V/g, \Omega)$ corresponding to the coin landing on the heads side. Compare it to the complementary region. Draw conclusions in terms of the sensitivity of the system to the initial conditions. Experiment (in real, not virtual worlds) to get an idea of what the range of realistic values of V and Ω might be.

(d) Show that, if the initial conditions V and Ω are random variables of the form $V = W + v$ and $\Omega = Z + v/R$, where the joint density $f(w, z)$ of (W, Z) is strictly positive, then $P(H) \to 1/2$, as $v \to \infty$. Interpret this result.

6 7 Bibliographical notes

A nice and concise (about 100 pages) introduction to the subject of dynamical systems is

[1] D. Ruelle, *Chaotic Evolution and Strange Attractors*, Cambridge University Press, Cambridge, 1989,

and

[2] N.B. Tufillaro, T. Abbott, and J. Reilly, *An Experimental Approach to Nonlinear Dynamics and Chaos*, Addison-Wesley, Reading, MA, 1992

is very much in the general spirit of our book. An easily accessible text is

[3] R.L. Devaney, *An Introduction to Chaotic Dynamical Systems*, Benjamin/Cummings, Menlo Park, CA,1986

with technically more advanced

[4] A.F. Beardon, *Iteration of Rational Maps*, Springer-Verlag, New York, 1991

[5] R.L. Devaney and L. Keen, Eds., *Chaos and Fractals*, American Mathematical Society, Providence, RI, 1989

[6] D.A. Lasota and M.C. Mackey, *Chaos, Fractals, Noise. Stochastic Aspects of Dynamics*, Springer-Verlag, New York, 1994.

A popular exposition with excellent illustrations can be found in

[7] H.O. Peitgen and D. Saupe, Eds., *The Science of Fractal Images*, Springer-Verlag , New York, 1988,

Two classics of dynamical systems are

[8] H. Poincaré, *Le mèthodes nouvelles de la mécanique céleste, Vol 1-3,* Gauthier-Villars, Paris, 1899,

[9] S. Ulam and J. von Neumann, On the combination of stochastic and deterministic processes, *Bull. Am. Math. Soc.*, 53(1947), 1120.

The former has been translated into English and published by Dover. The latter was the first study of the logistic function from the view point of chaotic behavior. The latest on the same subject is

[10] J.Graczyk and G. Świątek, Generic hyperbolicity in the logistic family, *Ann. Math.* 146(1997), 1-52.

The monograph

[11] S. Wiggins, *Chaotic Transport in Dynamical Systems*, Springer-Verlag, New York, 1992,

puts emphasis on transport and mixing phenomena in fluid mechanics. Two full fledged mathematical monographs in the area of ergodic theory are

[12] I. P. Cornfeld, S. V. Fomin, and Ya., G. Sinai, *Ergodic Theory*, Springer-Verlag, New York, 1982,

[13] U. Krengel, *Ergodic Theorems*, Walter de Gruyter, Berlin, 1985.

The original results connecting the Kolmogorov complexity and entropy of strings generated as orbits of dynamical systems were contained in

[14] A.A. Brudno, Entropy and the complexity of the trajectories of a dynamical system, *Trans. Moscow Math. Soc.*, 2(1983), 127-151. Springer-Verlag, New York, 1992.

They were later complemented by

[15] H.S. White, Algorithmic complexity of points in dynamical systems, *Ergodic Theory, Dynamical Systems*, 13(1993), 807-830.

Those interested in practical analysis of observed chaotic data should consult two recent volumes:

[16] C.D. Cutler and D.T. Kaplan, Eds., *Nonlinear Dynamics and Time Series*, Amer. Math. Soc., Providence, RI, 1997,

[17] H.D.I. Abarbanel, *Analysis of Observed Chaotic Data*, Springer-Verlag, New York, 1996.

The first book is more statistically and mathematically oriented, while in the second an emphasis is on the physical picture.

The physical coin toss experiments (see Project 6.6.21) were analyzed in numerous physics and mathematics publications, see, e.g.,

[18] J. Keller, The probability of heads, *American Math. Monthly* 93 (1986), 191-196.

Part III

MODEL SPECIFICATION-DESIGN OF EXPERIMENTS

Chapter 7

General Principles of Statistical Analysis

The exploration of experimental data and the reliability of the statistical inference based on these data depend heavily on the selection of the mathematical model and on the design of the data collection method.

Many of the principles of modern statistical analysis were formulated by R.A. Fisher. Such an analysis can usually be divided into the following steps, with some feedback among them:

Design of experiments and planning of investigation;
Specification of the model;

- Determining the method of statistical inference.

This chapter will discuss these steps in some detail.

7.1 Design of experiments and planning of investigation

Standard statistical analysis of observations from experiments is applicable only if the latter are conducted *independently* of each other, and are *repeatable*. Thus, the issue of random sample selection becomes paramount. If the population to be studied is uniform and finite, pseudo-random numbers are used to select a sample. All other cases have to be reduced to this case by splitting the whole population into uniform subpopulations. For example, if fifty trucks filled with iron ore are to be tested for the quality of the delivered mineral, then one way to proceed would be to select a smaller number of trucks randomly, then divide loads in the selected trucks in a systematic fashion into N portions, select a small random sample of size $n \ll N$ of the portions, and test only the mineral contained in the selected portions. It is obvious that the process of taking random samples is often tricky, and by no means simple. Another well known example comes from studies of consumers' behavior, when the general population has to be split into uniform subpopulations according to, e.g., age, income, gender, etc.

M. Denker, W. Woyczynski, *Introductory Statistics and Random Phenomena*,
Modern Birkhäuser Classics, DOI 10.1007/978-3-319-66152-0_7

Example 7.1.1 Quality Control.
The usual quality control methods for mass produced items rely on selection of a limited number of items to be tested. The selection process must be random to ensure that no systematic factors affect the quality of chosen items.

Example 7.1.2 Randomized Experiments.
In experimental situations, the (experimental) units may have to be randomized in the following sense. When examining the yield of corn under different watering and fertilizer conditions, one has to take into account the possibility that other factors are also influencing the yield. If the type of soil affects the yield, the design of the experiment has to eliminate this influence by randomization over the types of soil; that is, by distributing experimental units equally over different types of soil.

The above two situations are only very simple examples of what we mean by *planning an experiment*. In practice it may be very difficult to ensure the randomness of the experiment. Also, there are no generally accepted rules to ensure randomness. Nevertheless, there are certain minimal requirements to be satisfied. An obvious demand is that the choice of the experimental unit must depend neither on its properties nor on the investigator's own preferences and biases. Tests of randomness were discussed in previous chapters; their detailed exposition is beyond the scope of this book.

In most of the cases (excepting some special statistical procedures such as sequential analysis) the number of observations must be determined in advance. This number determines the precision of the estimators used and, thus, the reliability of the confidence bounds and power in testing procedures. These terms, used colloquially here, will acquire a technical meaning later on. They were discussed in a preliminary fashion in Example 3.7.1.

In principle, the number of observations affects the precision through the variance of the *statistic* (a function of observations) used. It is important to notice that in populations with more complex structure, this variance may not be uniform throughout the whole population and may vary from one subpopulation to another. For example, in an automobile assembly plant, various subassembly units may have different statistical properties from the view point of quality control. In such cases, sampling should consider this structure and divide the population into a (finite) number of homogeneous subpopulations, drawing a random sample from each subpopulation, instead of a single sample from the total population. The number of samples in a subpopulation has to be chosen in an optimal fashion to minimize the errors. These problems are simple examples of issues addressed in the theory of experimental designs and, more specifically, in the context of analysis of variance (ANOVA).

In practice, it is sometimes possible to rely on previous experience or "intelligent guessing" to find an (almost) optimal design for the experiments. It is often helpful to have some idea as to what are the approximate values of parameters of interest. Computer simulations often come in handy in gaining approximate prior experience. In Chapters 1, 2, and 4, we discussed the problem of pseudo-random number generation and demonstrated how these numbers can be used to simulate populations with prescribed probability distributions. Performing statistical analysis on such data can give some insight into the strengths and weaknesses of the statistical method. Whenever a new practical situation arises such an approach is advisable.

The design of an experiment, such as the decision on the sample size, can often be affected by the selection of a model; a feedback between these two steps is needed.

Mathematica Experiment 1. Random Sample Selection. We conclude this section performing a random selection of n different objects from a population of size N. The selection can be done with replacement or without replacement; the latter alternative was selected below.

```
In[1]:= objects = Table[N[5 Sin[x]+3 Cos[x]], {x,1,40}]
Out[1]= {5.82826, 3.29805, . . . ,1.72475}
In[2]:= f[n_]:= Random[Integer, {1,41-n}]
In[3]:= t1=Table[f[n], {n,1,11}]
Out[3]= {28, 30, 28, 35, 6, 27, 30, 16, 26, 4, 8}
In[4]:= t2=Table[t1[[i]]+ Sum[If[t1[[j]]>t1[[i]],0,1],
         {j,1,i-1}],{i,1,11}]
Out[4]= {28, 31, 29, 38, 6, 28, 35, 17, 28, 4, 10}
In[5]:= choice=Table[objects[[t2[[i]]]],{i,1,11}]
Out[5]= {-1.53329, 0.724039, -5.56234, 4.34706, 1.48343,
         -1.53329, -4.85199, -5.63248, -1.53329, -5.74494,
         -5.23732}
```

7.2 Model selection

The next step is to select a mathematical model for observations to be made. The models are most often statistical (probabilistic) in nature, but as we have seen in Chapters 4 through 6, other models are sometimes more desirable, especially when one deals with chaotic dynamical systems.

In the simplest cases, when the observations come from independently performed experiments, one only needs to specify the distribution function describing the probabilistic features of parameters of interest in the experiment. In the success/failure types of experiments, such a parameter can be, for example, the probability of success in a single trial; it can be the diameter of rivet heads manufactured in a fastener plant, the starch content in potatoes grown on a farm, or the specific pollutant concentration in a river.

Of course, there are some standard models for such probability distributions. If the parameter is constant and randomness occurs only as a result of errors in measurement, inaccuracy of the equipment, etc., then the normal distribution with density

$$f(x;\theta,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\theta)^2}{2\sigma^2}\right]$$

is often adequate as a statistical model. As a matter of fact, it was the early 19th-century analysis of this situation by Carl Friedrich Gauss of Göttingen University that marked the beginning of modern statistical theory. More generally, if the measured parameter fluctuates due to many small influences, then the normal distribution can be reasonably assumed in view of the Central Limit Theorem which asserts that the sum of many small independent random quantities is asymptotically normal (see Chapter 5). A bit of caution is in order here, since the Central Limit Theorem is proved under certain (admittedly mild) assumptions which, however, in some practical cases (such as the analysis of rare events and heavy tail distributions) need not be fulfilled.

Another type of continuous model is related to the so-called *extreme value statistics*. In this case the experimental unit is a system of many components, and one is interested in preventing the failure of the system, which occurs if one of the components fails. If the random quantities X_i denote the time until failure of the i-th component, the failure time of the whole system is the minimum value of all the random quantities X_i. This minimum is a random quantity with the specific probability distributions discussed in Section 3.8. In the special case of constant failure rate, the above distribution is exponential with the density

$$f(x)=\frac{1}{\lambda}\exp\left[-\frac{x}{\lambda}\right],\qquad x\geq 0.$$

The associated number N of components failing during the unit time interval has the discrete Poisson probability distribution

$$\mathrm{P}(N=k)=\frac{\lambda^k}{k!}\exp[-\lambda].$$

Clearly, failures of many components are rare events because the inverse of the factorial functions decays very rapidly.

In some experiments the outcomes are binary. For example, a yes/no decision is being made on the basis of each observation. Such situations occur, for example, in simple models of quality control where an item is only judged as functioning or not functioning. We discussed such models at length in Chapter 3.

However, in many cases, the statistical model of the experiment is unknown or only partially known. Then the standard statistical models of the types described above cannot be utilized. In such situations, the statistician first tries to draw some conclusions from graphical descriptions of the data; it can be called the graphical analysis. For example, if we are considering random observations Y_i depending on some known "independent" variable X_i, $(i = 1, .., n)$, the model may be specified by a linear dependence relation

$$Y_i = aX_i + b + \epsilon_i,$$

where ϵ_i are random fluctuation effects superposed on the linear relationship. Whether such a statistical model is suitable or not is, in practical application, determined by inspection of the *scatterplot*, that is the plot of Y_is against X_is. This simple form of *linear regression* was considered in Section 2.7. A nonlinear regression (e.g., polynomial) and time series models are other, more complex, possibilities that, however, go beyond the scope of this book.

In many applications of interest the collected data is not numerical and, as a result, there is no natural random quantity attached to the outcome of the experiment. Often, only the relative rank (or category) of observations is of importance. Thus, for example, one may have observations $x_1, \ldots, x_n$ where one is only interested in the relative *rank* r_j of x_j among all $x_1, \ldots, x_n$. Formally, $r_j = 1$ if and only if x_j is the smallest observation, and $r_j = k$ if and only if there are exactly $k - 1$ observations smaller and $n - k$ observations larger than x_j. It is clear that for a sample $x_1, \ldots, x_n$ of observations taken from independent identically distributed observations $X_1, \ldots, X_n$, the ordering of the observations is not relevant in the sense that the joint distribution of the random vector $(X_1, \ldots, X_n)$ does not change under coordinate permutation. The latter property of the random vector is often called *exchangeability*. Since sample $x_1, \ldots, x_n$ can be ordered in $n!$ ways, the probability of a given *rank vector* $(r_1, \ldots, r_n)$ equals $1/n!$ (one has to be a little bit cautious here and make sure that all the x_is are different). Thus, transforming the data yields new observables with the known distribution. When a parametric model cannot be specified, ranking procedures provide an important tool in nonparametric statistical inference.

The first step in specifying the statistical model must include determination of whether the data are *numerical* or *categorical*. This classification was explained in Sections 2.1 and 2.2.

For numerical data, statistical analysis depends on their exact values. Numerical data can be of two types: *interval* and*proportional*. Population size of a bee colony is an example of the former, and the Dollar/Yen exchange is an example of the latter.

The categorical data types are further subdivided into *nominal data* and *ordinal data*. Nominal data are fictitious numbers attached to certain characteristics of sample points. For example, the individual's gender can be encoded by numbers 0 and 1, or by -1 and $+1$. This assignment may make the data recording more convenient, but it really has no relation to the observed phenomenon. Automobile license plates can also serve as a convenient label in the population of all cars, although *per se* they do not reflect any physical characteristics of particular cars.

Ordinal data, in addition to categorization of sample points, give extra information through their ordering. For example, the examination grades A, A^-, B^+, ..., D^-, F, can also be recorded as 4.00, 3.66, 3.33, ..., 0.66, 0.00, in which case the relation $1.66 < 3.33$ tells us that the student whose grade was recorded as 3.33 performed better than one whose record shows 1.66. Data of this type are also often collected when a new drug is being tested. If the aim of the new drug is to relieve pain, then the degree of pain relief provided is not objectively measurable and has to be judged on the basis of patients' subjective reporting and physicians' observations. The only practical solution may be to judge whether patient A is better off than patient B. Social status, athletic ranking, and grades of merchandise are other examples of ordinal data.

Also, note that it is possible to measure numerical data but to use them as ordinal data. The converse is, of course, not possible. Usually, in case of numerical data, a parametric approach is desirable and the possible class of distributions has to be determined by one of the approaches sketched above. In the case of ordinal data, nonparametric (or semiparametric) models are often preferable.

In general, if the model cannot be determined *a priori*, the graphical methods can be used. In Chapter 2 we have already seen that the Kolmogorov-Smirnov statistic can be used to provide an estimate (including the confidence bounds) for an unknown distribution. Other methods discussed in this book include box-and-whiskers plots, Q-Q plots, and the chi-square goodness-of-fit test. Thus, a preliminary experiment to determine approximate distribution may be warranted. It would consist in plotting the empirical distribution function and making a reasonable guess (based on basic characteristics of distributions such as skewness, symmetry, mean value, etc.) as to which of the known theoretical distributions provides the best fit.

In conclusion, it is fair to say that model selection is a delicate problem and that it is best solved in a cooperative effort of a statistician, a probabilist, and an experimentalist.

7.3 Determining the method of statistical inference

Once the design of the experiment and the model are selected, one has to turn to the problem of determining the most appropriate method of statistical analysis of collected data. In this section we will concentrate on the problem of estimating parameters in the model family of distributions.

Some parameters possess useful additional properties. For example, the mean (expectation) behaves nicely under translations and dilations and is thus a good *location parameter*. Indeed, if $E(X) = \mu$, then

$$E(X + \nu) = \mu + \nu, \quad \text{and} \quad E(\alpha X) = \alpha\mu.$$

On the other hand, the variance $\sigma^2(X)$ scales nonlinearly, since

$$\mathrm{Var}\,(\alpha X) = \alpha^2 \mathrm{Var}\,(X).$$

Thus, the standard deviation $\sigma(X)$ plays the role of a *scaling parameter*. In general, recall that if the random variable X has the cumulative d.f. $F(x)$, and ν and α are constants, then the random variable $X + \nu$ has the cumulative d.f. $F(x - \nu)$, and the random variable αX has the cumulative d.f. $F(x/\alpha)$. In the most common situations, where the normal or binomial models arise, the statistical inference is done for these parameters. If the population is not homogeneous, then the situation may become more complicated as each subpopulation could have different location and variability (scaling) parameters.

In this section we will discuss three elementary methods of parameter estimation based on different principles. The first, the *maximum likelihood estimation*, relies on the notion that the true parameters are most likely (probable) to be reflected in a given collected sample. The second method, the *least squares method*, minimizes the mean square error of the estimate. The third, the *method of moments*, relies on the interpretation of the Law of Large Numbers.

7.3.1 Maximum likelihood estimator (MLE)

Let $x_1, \ldots, x_m$ be a sample of size m taken from a probability distribution with a density function $f(x; \theta)$, depending on an unknown parameter θ. We assume that the mapping $\theta \mapsto f(x, \theta)$ is smooth and has a unique maximum for each given x.

The maximum likelihood estimation procedure provides an estimator

$$\hat{\theta} = \hat{\theta}(x_1, \ldots, x_m) \tag{1}$$

for the parameter θ which maximizes (over θ) the *likelihood function*

$$\theta \longmapsto L(x_1, \ldots, x_m; \theta) = f(x_1; \theta) \cdot \ldots \cdot f(x_m; \theta). \tag{2}$$

This is the joint probability density of m independent random variables, each with density $f(x; \theta)$ evaluated at $x_1, \ldots, x_m$, with only one unknown parameter θ. If the density f is smooth and has a unique maximum, then it is most convenient to find this maximum via differentiation (with respect to θ) of the logarithm

$$\log L = \log f(x_1; \theta) + \log f(x_2; \theta) + \ldots + \log f(x_m; \theta) \tag{3}$$

of the likelihood function. The logarithm is a strictly increasing function, so the functions L and $\log L$ attain their maximum at the same point.

Thus, we find the MLE $\hat{\theta}$ by solving for θ equation

$$\frac{\partial \log L}{\partial \theta} = \frac{f_\theta(x_1;\theta)}{f(x_1;\theta)} + \frac{f_\theta(x_2;\theta)}{f(x_2;\theta)} + \ldots + \frac{f_\theta(x_m;\theta)}{f(x_m;\theta)} = 0, \tag{4}$$

where $f_\theta = \partial f/\partial\theta$.

Example 7.3.1 Gaussian Family.
Let

$$f(x;\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\theta)^2}{2\sigma^2}\right] \tag{5}$$

be the family of normal distributions with unknown parameter θ. Parameter σ^2 may be known or unknown; in this example it does not matter. We want to determine the MLE $\hat{\theta}$ for θ based on a random sample $x_1, \ldots, x_m$ of size m. The likelihood function

$$L(x_1,\ldots,x_m;\theta) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^m \exp\left[-\frac{(x_1-\theta)^2 + \ldots + (x_m-\theta)^2}{2\sigma^2}\right] \tag{6}$$

and

$$\log L(x_1,\ldots,x_m;\theta) = -\frac{m}{2}\log(2\pi\sigma^2) - \frac{(x_1-\theta)^2 + \ldots + (x_m-\theta)^2}{2\sigma^2}.$$

The derivative with respect to θ is

$$\frac{\partial \log L(x_1,\ldots x_m;\theta)}{\partial\theta} = \frac{x_1 + x_2 + \ldots + x_m - m\theta}{\sigma^2}.$$

The equation

$$\frac{\partial \log L}{\partial \theta} = 0$$

has the unique solution

$$\hat{\theta} = \hat{\theta}(x_1,\ldots,x_m) = \frac{x_1 + \ldots + x_m}{m} = \bar{\boldsymbol{x}}. \tag{7}$$

So, the maximum likelihood estimator for parameter θ in the normal distribution happens to be the sample mean.

Two observations are in order regarding the MLE in the above example. First, note that if we consider

$$\hat{\theta} = \hat{\theta}(X_1, \ldots, X_m) = \frac{X_1 + \ldots + X_m}{m}$$

as a random quantity depending on the normally distributed ensemble $(X_1, \ldots, X_m)$ of all sample points $(x_1, \ldots, x_m)$, then the expectation

$$E(\hat{\theta}) = E(\hat{\theta}(X_1, \ldots, X_m)) = E\left(\frac{X_1 + \ldots + X_m}{m}\right) = \theta. \tag{8}$$

In other words, the expected value of the estimator is equal to the estimated parameter—obviously a desirable property. In such cases the estimator is called *unbiased*. Second, in view of the Law of Large Numbers

$$\hat{\theta}(X_1, \ldots, X_m) = \frac{X_1 + \ldots + X_m}{m} \longrightarrow E(X_1) = \theta \tag{9}$$

as the sample size m goes to ∞. Thus, as the sample size increases, the error of the estimator becomes asymptotically negligible. Estimators enjoying this property are called *consistent*.

Example 7.3.1 Continued. Gaussian Family.
We shall find the MLE for σ^2, based on the knowledge that $\theta = \hat{\theta} = \bar{x}$. The likelihood function becomes

$$\log L = -\frac{m}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{m} (x_i - \bar{x})^2.$$

Differentiating with respect to σ^2 (set $\sigma^2 = y$, and take the derivative with respect to y) gives

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{m}{2\sigma^2} + \frac{1}{\sigma^4} \sum_{i=1}^{m} (x_i - \bar{x})^2.$$

Equating this expression 0 we obtain

$$\hat{\sigma}^2(x_1, \ldots, x_m) = \frac{1}{m} \sum_{i=1}^{m} (x_i - \bar{x})^2.$$

This is not an unbiased estimator because

$$E(\hat{\sigma}^2(X_1, \ldots, X_m)) = \frac{m-1}{m}\sigma^2.$$

It follows that

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

is an unbiased estimator. Both estimators of σ^2 are, however, consistent.

Remark 7.3.1 The method of partial derivatives does not always lead to the correct result, or even to any result. Here is an illuminating example. Consider a family of probability d.f.s defined by the formula

$$f(x;\theta) = \begin{cases} 1/\theta, & \text{if } 0 \le x \le \theta; \\ 0, & \text{otherwise.} \end{cases}$$

Based on the sample $x_1, \ldots, x_m$, the likelihood function is

$$L(x_1, \ldots, x_m; \theta) = \begin{cases} 1/\theta^m, & \text{if } 0 \le \max_{1\le i\le m} x_i \le \theta; \\ 0, & \text{otherwise.} \end{cases}$$

Use *Mathematica* to graph this function for different data sets, and verify that L attains maximum for

$$\hat{\theta}(x_1, \ldots, x_m) = \max_{1\le i\le m} x_i.$$

7.3.2 Least squares estimator (LSE)

In this subsection we will consider a *linear model* in which the observed quantity y is assumed to be an unknown linear function

$$y = \beta_1 x_1 + \ldots + \beta_n x_n \tag{10}$$

of the vector $\boldsymbol{x} = (x_1, \ldots, x_n)$. Statistical inference for $\beta_1, \ldots, \beta_n$, is to be based on $m > n$ observations $y_1, \ldots, y_m$ of the quantity y taken at m levels $(x_{i1}, \ldots, x_{in})$, $i = 1, 2, \ldots, m$, of the independent variable vector $\boldsymbol{x}$. Assuming the presence of a random additive noise $\epsilon_1, \ldots, \epsilon_m$, in the system, independent for different observations, the statistical model equations for observations y_i can be written in the form

$$
\begin{aligned}
y_1 &= \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_n x_{1n} + \epsilon_1 \\
y_2 &= \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_n x_{2n} + \epsilon_2 \\
\ldots\ldots &= \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
y_m &= \beta_1 x_{m1} + \beta_2 x_{m2} + \ldots + \beta_n x_{mn} + \epsilon_m .
\end{aligned}
$$

Traditionally, the matrix $\mathcal{X} = \{x_{ij}\}$ is called the *design matrix* and $(\epsilon_1, \ldots, \epsilon_m)$, the random *error vector*. The model is a generalization of the simple regression model considered in Section 2.7.

The least squares estimator $\tilde{\beta}$ of $\beta = (\beta_1, \beta_2, \ldots, \beta_n)$ minimizes the quadratic error function

$$\epsilon^2(\beta) = \sum_{i=1}^{m} \epsilon_i^2 = \sum_{i=1}^{m} \Bigl(y_i - \sum_{j=1}^{n} \beta_j x_{ij}\Bigr)^2, \qquad (11)$$

which also happens to be the square of the m-dimensional Euclidean distance between the vector $(y_1, \ldots, y_m)$ of *observed responses* and the vector $(\beta_1 x_{11} + \ldots + \beta_n x_{1n}, \ldots, \beta_1 x_{m1} + \ldots + \beta_n x_{mn})$ of predicted responses.

Example 7.3.2 Measuring a Constant.

Suppose that m measurements of a single unknown quantity β are made in the presence of some errors. The corresponding linear model is

$$y_i = \beta + \epsilon_i, \qquad i = 1, \ldots, m.$$

In this case, $n = 1$, and the design matrix is

$$\mathcal{X} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

The LSE method requires that we minimize the expression

$$\epsilon^2(\beta) = \sum_{i=1}^{m} (y_i - \beta)^2.$$

Taking the derivative and equating it to zero we get

$$\sum_{i=1}^{m}(y_i - \beta) = 0.$$

Hence, in this case, the LSE

$$\hat{\beta} = \frac{y_1 + \ldots + y_m}{m} = \bar{y}$$

is again the sample mean.

Example 7.3.3 Optimizing the Manufacturing of Chips.

A chip manufacturer can choose between two manufacturing processes (say, I and II). He wants to organize his production lines to maximize his profits defined as the retail value minus production costs and minus distribution costs. Two samples of m chips each are manufactured using processes I and II, respectively. If a chip produced via process I (resp., II) passes the quality control, its profit is β_1 (resp., β_2). If the chip does not pass the quality standard, then its profit is β_3 (resp., β_4). Denoting by $m_1 \leq m$ the number of chips in the batch manufactured by process I that meet the quality standard, and by $m_2 \leq m$ the corresponding number for process II, the appropriate linear model is

$$y_i = \begin{cases} \beta_1 + \epsilon_i & \text{for } i = 1, \ldots, m_1 \\ \beta_2 + \epsilon_i & \text{for } i = m_1 + 1, \ldots, m_1 + m_2 \\ \beta_3 + \epsilon_i & \text{for } i = m_1 + m_2 + 1, \ldots, m + m_2 \\ \beta_4 + \epsilon_i & \text{for } i = m + m_2 + 1, \ldots, 2m. \end{cases}$$

The design matrix $\mathcal{X}$ becomes

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

and the LSE for the vector $(\beta_1, \beta_2, \beta_3, \beta_4)$ is found by minimizing the expression

$$\sum_{i=1}^{m_1}(y_i-\beta_1)^2+\sum_{i=m_1+1}^{m_1+m_2}(y_i-\beta_2)^2+\sum_{i=m_1+m_2+1}^{m+m_2}(y_i-\beta_3)^2+\sum_{i=m+m_2+1}^{2m}(y_i-\beta_4)^2.$$

Taking the partial derivatives with respect to $\beta_1, \beta_2, \beta_3, \beta_4$, and equating them to zero, we obtain the following LSE for these parameters:

$$\hat{\beta}_1=\frac{1}{m_1}\sum_{i=1}^{m_1}y_i,$$

$$\hat{\beta}_2=\frac{1}{m_2}\sum_{i=m_1+1}^{m_1+m_2}y_i$$

$$\hat{\beta}_3=\frac{1}{m-m_1}\sum_{i=m_1+m_2+1}^{m+m_2}y_i,$$

$$\hat{\beta}_4=\frac{1}{m-m_2}\sum_{i=m+m_2+1}^{2m}y_i.$$

Example 7.3.4 Regression Lines and Curves.
This example has been studied in Section 2.7, but it is worthwhile to recall it in the present context of general linear models. Suppose that $y_1, \ldots, y_m$ are the observed responses, and $x_1, \ldots, x_m$ are the corresponding values of the independent variable and, after a glance at the scatter plot of the paired data (x_i, y_i), we suspect that there is a linear relationship between an independent variable x and the observed response y. The appropriate linear model is

$$y_i=\beta_0+\beta_1x_i+\epsilon_i, \qquad i=1,\ldots,m, \tag{12}$$

which can be compactly written in the form

$$\boldsymbol{y}=\mathcal{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}, \tag{13}$$

where

$$\mathcal{X}=\begin{pmatrix}1 & x_1\\ \vdots & \vdots\\ 1 & x_m\end{pmatrix} \tag{14}$$

is the design matrix, $\beta = (\beta_0, \beta_1)^T$ is the vector of regression coefficients, and $\epsilon = (\epsilon_1, \ldots, \epsilon_m)^T$ is the error (residuals) vector. To find the LSEs $\hat{\beta}_0$ and $\hat{\beta}_1$, we need to minimize the function

$$\epsilon^2(\beta_0, \beta_1) = \sum_{i=1}^{m} \epsilon_i^2 = \sum_{i=1}^{m} (y_i - \beta_0 - \beta_1 x_i)^2.$$

The partial differentiation gives

$$\frac{\partial \epsilon^2}{\partial \beta_0} = -2 \sum_{i=0}^{m} y_i - \beta_0 - \beta_1 x_i,$$

and

$$\frac{\partial \epsilon^2}{\partial \beta_1} = -2 \sum_{i=1}^{m} x_i (y_i - \beta_0 - \beta_1 x_i).$$

Denoting by $\bar{x}, \bar{y}, \overline{x^2}$, and $\overline{xy}$ the means of samples $(x_1, \ldots, x_m)$, $(y_1, \ldots, y_m)$, $(x_1^2, \ldots, x_m^2)$, and $(x_1 y_1, \ldots, x_m y_m)$, respectively, the equations

$$\frac{\partial \epsilon^2}{\partial \beta_0} = 0, \qquad \frac{\partial \epsilon^2}{\partial \beta_1} = 0, \tag{15}$$

can be rewritten in a more transparent form:

$$\beta_0 + \beta_1 \bar{x} = \bar{y}, \qquad \beta_0 \bar{x} + \beta_1 \overline{x^2} = \overline{xy}. \tag{16}$$

Thus, we finally obtain the LSEs

$$\hat{\beta}_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{17}$$

As we observed in Chapter 2, the first order polynomial regression is only the simplest example of the general polynomial model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_d x^d + \epsilon \tag{18}$$

of degree d, with the design matrix

$$\begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^d \\ 1 & x_2 & x_2^2 & \dots & x_2^d \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_m & x_m^2 & \dots & x_m^d \end{pmatrix}. \tag{19}$$

To carry out statistical inference on the linear regression model, one needs to propose and test a statistical model for the residuals $\epsilon_1, \dots, \epsilon_m$, which are random quantities. One such approach will be discussed in Chapter 8.

7.3.3 Method of moments (MM)

The method of moments is based on the fact, for certain cumulative d.f.s $F_X(x)$, the sequence

$$\mu_1 = EX, \quad \mu_2 = EX^2, \quad \mu_3 = EX^3, \quad \dots, \tag{20}$$

of moments of the random quantity X, completely determines the distribution $F_X(x)$ itself. This is the case, for example, when the moment generating function $\phi(u) = Ee^{uX}$ is well defined and analytic in the neighborhood of 0, since then one can expand it in the power series

$$\phi(u) = \sum_{k=0}^{\infty} \frac{\phi^{(k)}(0)}{k!} u^k, \qquad \phi^{(k)}(0) = EX^k, \tag{21}$$

with coefficients determined by the moment sequence [see (5.4.14)]. Actually many families of distributions are determined just by a couple of moments (e.g., Gaussian, exponential, etc.).

By the Law of Large Numbers, the average of independent identically distributed random observations $X, X_1, \dots, X_n$, approaches, as n increases to ∞, the expected value $\mu_1 = EX$ of the model distribution (if it exists). Thus the sample mean, that is, the first sample moment (see Section 2.3)

$$m_1(\boldsymbol{x}) = \bar{\boldsymbol{x}} = \frac{x_1 + x_2 + .. + x_n}{n} \tag{22}$$

may serve as an estimator for $\mu_1 = EX$, which is automatically consistent and unbiased.

More generally, we can estimate the k-th moment μ_k of the model cumulative d.f. $F_X(x)$ (again, if it exists) by the sample k-th moment

$$m_k(\boldsymbol{x}) = \overline{\boldsymbol{x}^k} = \frac{x_1^k + x_2^k + \ldots + x_n^k}{n}. \tag{23}$$

Again, by the Law of Large Numbers applied to the sequence $X_1^k, X_2^k, \ldots,$ of independent and identically distributed random variables, the random quantity $m_k(X)$ converges to EX^k. In particular, for $k = 2$, the above formula gives an unbiased and consistent estimator for the second moment EX^2.

Since the variance

$$\text{Var}\,(X) = \sigma^2(X) = E(X^2) - (EX)^2,$$

one could suggest

$$\hat{\sigma}^2(\boldsymbol{x}) = \frac{x_1^2 + .. + x_n^2}{n} - \left(\frac{x_1 + .. + x_n}{n}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{24}$$

as a good estimator for the variance of X. However, this estimator is biased because an easy calculation shows that

$$E(\hat{\sigma}^2(X_1, \ldots, X_n)) = \frac{n-1}{n}\sigma^2(X) \neq \sigma^2(X), \tag{25}$$

although, as $n \to \infty$, the discrepancy between $E\hat{\sigma}^2$ and σ^2 disappears (see Subsection 7.3.2). To remedy this difficulty, one usually considers the estimator

$$s^2(\boldsymbol{x}) = \frac{n}{n-1}\hat{\sigma}^2(\boldsymbol{x}) = \frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - \frac{n}{n-1}\bar{x}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \tag{26}$$

which, obviously, is an unbiased estimator of the variance; that is

$$E\Big(s^2(X_1, \ldots, X_n)\Big) = \sigma^2(X). \tag{27}$$

In statistics, the primary problem is usually not the moment estimation *per se* but estimation of parameters of an unknown distribution. Often these parameters are functions of moments. Thus, we are given a family $f(x; \theta)$ of density functions with an unknown parameter $\theta \in \mathbf{R}^d$, where $d \geq 1$ is an integer. For example, if $f(\,.\,; \theta)$ is the normal family, then $\theta = (\mu, \sigma^2) \in \mathbf{R}^2$. To be more precise, assume that the first d moments exist for every distribution in the family, and that the k-th moment $\mu_k = \lambda_k(\theta)$ is a function of θ. Then, the method of moments (MM) estimator of θ is a vector $\hat{\theta}$ satisfying the condition

$$\overline{x^k} = \lambda_k(\hat{\theta}),$$

for every $k = 1, \ldots, d$. Of course, this estimator need not exist, nor need it be unique. But in many important cases these equations have a unique solution.

In order to see how to apply the MM estimators, let us return once more to the estimation of parameter σ^2 for the normal family with the vector parameter $\theta = (\mu, \sigma^2)$. The first and second moments are

$$\mu_1 = \int x f(x; \theta) dx = \mu,$$

and

$$\mu_2 = \int x^2 f(x; \theta) dx = \sigma^2 + \mu^2.$$

Hence, the MM estimator $\hat{\sigma}^2$ is the solution of the following two equations:

$$\bar{x} = \mu$$

and

$$\overline{x^2} = \sigma^2 + \mu^2.$$

As another example, consider the family of Bernoulli distributions given by $P(1; \theta) = \theta$, and $P(-1; \theta) = 1 - \theta$. Then,

$$\mu_1 = 1 \cdot P(1; \theta) + (-1) \cdot P(-1; \theta) = \theta - (1 - \theta) = 2\theta - 1.$$

Thus, the estimator for θ (the MM estimator of θ) is

$$\hat{\theta} = \frac{1}{2}(\bar{x} + 1).$$

7.3.4 Concluding remarks

The above three types of estimators—MLE, LSE, and MM—are only the most prominent and most frequently used estimators. They are usually followed by calculation of confidence intervals, or by test procedures and under a variety of models, including the normal models, regression model, ANOVA, the binomial model (appropriate when we want to estimate an unknown probability of success or failure), or the Poisson model which is often used to analyze rare (bad luck) events. More sophisticated estimators, such as median, rank, etc. will not be considered in this book (see the Bibliographical Notes).

Confidence intervals give error bounds for the estimators. A typical statement would be that the true value of the parameter θ is within distance δ of the estimator $\tilde{\theta}$, that is

$$|\tilde{\theta} - \theta| \leq \delta,$$

with a certain probability α, $0 < \alpha < 1$. The latter is called the confidence level. Finding such confidence intervals is preferable if one wants to obtain a precise measurement of an unknown numerical quantity (for example, the starch content of a stock of potatoes used in a manufacturing plant).

Testing procedures are usually more sophisticated and result in a decision whether or not to reject a certain hypothesis. In general, natural sciences progress by suggesting theories purporting to explain various natural phenomena, and then testing those theories against experiments. As long as no contradiction appears, the theory is not rejected. So, new theories abound, and a lot of effort is directed at testing various hypothesis with an eye towards rejecting them. This is also the basic attitude of statisticians. We will formalize these concepts in Chapters 8 and 9.

7.4 Estimation of fractal dimension

As a characteristic of data sets, the fractal dimension was already discussed in Sections 2.8 and 6.5 in connection with data compression techniques and attractors in nonlinear dynamical systems. In this section, we will develop a method of estimation of correlation integrals (and, thus, the correlation dimension) when the sample is taken from a sequence of independent, identically distributed d-dimensional random vectors. This methodology is also widely applicable to chaotic data, where the independence assumption is not satisfied.

Let us consider a probability distribution μ on the d-dimensional Euclidean space $\mathbf{R}^d$. By definition, the support of μ is the smallest closed set in $\mathbf{R}^d$ of full μ measure 1. The support of μ may be a proper subset of $\mathbf{R}^d$, and it can have fractal dimension.

As a preliminary example consider the normal probability distribution μ_2 on $\mathbf{R}^2$, with density

$$f(x, y) = \frac{1}{2\pi} e^{-(x^2+y^2)/2}, \tag{1}$$

and the standard normal distribution μ_1 on $\mathbf{R} \subset \mathbf{R}^2$ which, considered as a measure on $\mathbf{R}^2$, acts on test functions $\phi(x, y)$ via the formula

$$\int_{\mathbf{R}^2} \phi(x, y)\mu_1(dx, dy) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \phi(x, 0)e^{-x^2/2}dx. \tag{2}$$

Observe that for a square $A = [-a, a] \times [-a, a] \subset \mathbf{R}^2$, we have that $\mu_2(A) = \mu_1(A)^2$. Thus, if $a \to 0$, then the measure $\mu_2(A)$ converges to 0 at a rate equal to the square of the rate of convergence for $\mu_1(A)$. This is directly related to the fact that the dimension of the support of μ_2 is 2, whereas the dimension of the support of μ_1 is 1.

The concept of correlation integral is based on a similar idea of measuring such differences in asymptotics. Instead of the square A, we will take the integral over a ball of fixed radius and, then, average it over all possible balls to avoid the artificial centering at 0. More precisely, the correlation integral of a probability measure μ on $\mathbf{R}^d$ is the function

$$C_\mu : (0, \epsilon_0) \to [0, \infty) \tag{3}$$

defined by

$$C_\mu(\epsilon) = \int_{\mathbf{R}^d} \int_{\mathbf{R}^d} \mathbf{1}_{\{|x-y|<\epsilon\}}\, \mu(dx)\, \mu(dy) = \int_{\mathbf{R}^d} \mu(B(x, \epsilon))\, \mu(dx), \tag{4}$$

where $\mathbf{1}_{\{|x-y|<\epsilon\}} = 1$, if $|x - y| < \epsilon$, and $= 0$, otherwise, and where $\mu(B(x, \epsilon))$ denotes the measure of the ball $B(x, \epsilon)$ with radius ϵ and center $x \in \mathbf{R}^d$.

Now, let $X_1, X_2, \ldots, X_n$ be a finite sequence of independent d-dimensional random vectors with distribution μ. Then

$$E(\mathbf{1}_{\{|X_i-X_k|<\epsilon\}}) = E(\mathbf{1}_{\{|X_1-X_2|<\epsilon\}}) = \int_{\mathbf{R}^d} \mu(B(x, \epsilon))\, \mu(dx) = C_\mu(\epsilon), \tag{5}$$

for all $1 \le i \ne k \le n$. Therefore, because there are exactly $n(n-1)$ pairs $i \ne k$,

$$E\left(\frac{1}{n(n-1)} \sum_{i,k=1; i\ne k}^{n} \mathbf{1}_{\{|X_i-X_k|<\epsilon\}}\right) = C_\mu(\epsilon). \tag{6}$$

Finally, given a random sample $x_1, x_2, \ldots, x_n$, from the distribution μ, we can use an idea influenced by the method of moments of Section 7.3. First, in view of the Law of Large Numbers, $E(\mathbf{1}_{\{|X_1-x|<\epsilon\}}) = \mu(B(x, \epsilon))$ can be estimated by $(1/n)\sum_{i=1}^{n} \mathbf{1}_{|x-x_i|<\epsilon}$. Taking another average, we obtain

$$\hat{C}_\mu(\epsilon) = \frac{1}{n^2} \sum_{i,k=1}^{n} \mathbf{1}_{\{|x_i-x_k|<\epsilon\}} \tag{7}$$

as a consistent estimator for $C_\mu(\epsilon)$. Like the MM variance estimator, it is not unbiased; the unbiased estimator is obtained by deleting the diagonal sum. Indeed,

$$\tilde{C}_\mu(\epsilon) = \frac{1}{n(n-1)} \sum_{i,k=1;i\neq k}^{n} \mathbf{1}_{\{|x_i - x_k| < \epsilon\}} \tag{8}$$

is an unbiased and consistent estimator of $C_\mu(\epsilon)$.

The Grassberger-Procaccia correlation dimension d_μ for μ uses the correlation integral $C_\mu(\epsilon)$. It is defined by the formula

$$d_\mu = \lim_{\epsilon\to 0} \frac{\log C(\epsilon)}{\log \epsilon}, \tag{9}$$

whenever this limit exists. Intuitively, it means that $C(\cdot)$ is roughly a function of the form

$$C(\epsilon) = K\epsilon^{d_\mu} + \text{lower order terms}. \tag{10}$$

Therefore, as suggested in Chapters 2 and 6, for a finite set of (different) radii $\{\epsilon_1, \epsilon_2, \ldots, \epsilon_m\}$, the correlation dimension d_μ is estimated by the slope of the regression line for the data set

$$\Big\{(\log \epsilon_1, \log \hat{C}_\mu(\epsilon_1)), (\log \epsilon_2, \log \hat{C}_\mu(\epsilon_2)), \ldots, (\log \epsilon_m, \log \hat{C}_\mu(\epsilon_m))\Big\}.$$

As a result,

$$\hat{d} = \frac{\sum_{i=1}^m \log \epsilon_i \log \hat{C}_\mu(\epsilon_i) - m\bar{\epsilon}\overline{\log \hat{C}_\mu(\epsilon)}}{\sum_{i=1}^m r_i^2 - m\bar{\epsilon}^2}, \tag{11}$$

where

$$\bar{\epsilon} = \frac{1}{m}\sum_{i=1}^m \epsilon_i,$$

and

$$\log \hat{C}_\mu(\epsilon) = \frac{1}{m}\sum_{i=1}^m \log \hat{C}_\mu(\epsilon_i).$$

Mathematica Experiment 1. Correlation Dimension of the Cantor Set. We will use the above procedure to estimate the correlation integral and dimension of the Cantor set. Recall (see Section 2.8) that the "middle-third-removed" Cantor set consists of all real numbers in the interval [0, 1] which have a triadic representation $\sum x_k 3^{-k}$, where each x_k is either 0 or 2. A natural measure μ on the Cantor set C is obtained by transporting the Lebesgue measure from [0, 1], using the map

$$[0, 1] \ni x = \sum x_k 2^{-k} \longmapsto y = y(x) = \sum x_k 3^{-k} \in C.$$

Technically speaking, our goal is to estimate the correlation integrals and dimension of μ.

The first task is to produce a realization of a sequence of independent random variables with the probability distribution μ on the Cantor set. For this measure, sequences of 0s and 2s are like sequences of 0s and 1s for the symmetric Bernoulli measure (= Lebesgue measure on [0, 1]). Hence, it suffices to produce a string of 0s and 1s as a realization of the symmetric Bernoulli sequence and multiply each term by 2. Such a string, $x_1, x_2, \ldots, x_l$, represents a random number

$$x = \sum_{i=1}^{l} x_i 3^{-i}$$

in the Cantor set, with the probability distribution approximately equal to μ.

```
In[1]:= <<Statistics‘LinearRegression‘
In[2]:= ran= Table[ N[Sum[ 2*3^(-i)*Table[Random[Integer],
          {6}][[i]], {i,1,6}]], {80}];
In[3]:= c[r_]:= (1/80.)^2 Sum[ Sum[ If[ Abs[ran[[i]]-
              ran[[j]]]<r, 1,0], {j,1,80}],{i,1,80}]
In[4]:= reg= Table[{Log[1.8^(-i-2)], Log[c[1.8^(-i-2)]]},
              {i,1,5}]
Out[5]= {{-1.76336, -1.19671}, {-2.35115, -1.46696},
        {-2.93893, -1.83846},  {-3.52672, -2.14398},
        {-4.11451, -2.54148}}
In[6]:= Fit[reg, {1,x},x]
Out[6]= -0.154239 + 0.572751 x
In[7]:= N[Log[2]/Log[3]]
Out[7]= 0.63093
```

So, in a single run of a relatively moderate length, we obtained 0.572751 as an estimate of the correlation dimension of the Cantor set, not a bad approximation to the true theoretical value of $\log 2/\log 3 \approx 0.63093$. Of course, since we operate with random samples, each time you run the above program, the result is going to be slightly different. When we run it for the second time, we obtained the estimate 0.578216.

7.5 Practical side of data collection and analysis

We shall conclude this chapter with a few comments on the practical side of data collection and analysis. First of all, one has to remember that statistical methods are well suited only for studies of random phenomena. This randomness may be systemic (as in numerous natural and chaotic phenomena), or it may come from inaccurate and noisy measurements, incomplete information, or other similar sources.

When collecting data in a survey and/or a series of experiments, one has to ensure that this randomness is not affected by systematic errors which can have various causes. Quite often the experimenter himself is responsible for faulty measurements, not to mention cases of ethical lapses when data sets are just made up or tampered with. But an instrument can also add some systematic measurement errors due to faulty calibration, positioning, fixed but unknown external fields, etc. Also, the process of instrument reading by different people may result in a systematic error. Often, the procedure for taking a random sample is not sound. Testing the general population's habits by polling a random sample of customers at an upscale shopping center may be a wrong approach as this is already a self-selected population.

The most important procedural point in data collection is to make sure that the complete original record for each experiment (the raw data) is securely stored and preserved. In order to make this *documentation* available for computer analysis, the data should be collected in computerized files, either immediately after the experiment is concluded, or as soon as possible. There are many statistical software packages that facilitate data collection. Often, the experiments are costly, and one wants data files readily available for analysis.

Second, one has to keep in mind that the statistical analysis is a scientific report, and such reports have to be *reproducible* as a scientific investigation. Data analysis often begins with the back-of-the-envelope stem and leaf diagrams (not only in the one sample case) to obtain some indication whether or not collected data approximately satisfy model assumptions. This initial interpretation of data requires a lot of experience and intuition, as there are only a few basic rules to be followed: checking for the presence of *outliers*—that is values which, with very high probability, seem to be outside the reasonable range; inspecting if the empirical distribution seems to have the correct range, symmetry properties and location parameters; constructing histograms is usually helpful in such situations.

The next step in practical statistical analysis is a test of the hypothesis or construction of the confidence intervals. These are done using one of the numerous commercial statistical software such as S^+ or SAS but it is not our aim in this book to train students in their use. For many simpler situations, *Mathematica* `Statistics` packages will do a creditable job. One has to remember though that the procedures vary quite a bit from one software package to another, and changing your computer tools may require a big time investment and a steep learning curve.

Computer analysis provides the statistician with a basis for his conclusions: the hypothesis is rejected or not rejected, the confidence interval is established. Sometimes, however, statistical results are used to draw conclusions that are too optimistic or not really justified. To avoid such mistakes, besides using correct formal statistical tools, the experimenter has to be guided by an in-depth knowledge of his own field. In a study of 149 articles which appeared in 10 respected medical journals (see Schoor and Karsten, Statistical evaluation of medical journal manuscripts, *J. Amer. Med. Asso.* 195 (1966), 1123–1128), 12% were found to contain conclusions that were not justified! Remember Huff's book *How to Lie with Statistics* quoted in Chapter 1.

Finally, one has to guard against creating too high expectations for the results of statistical analysis. Statistical methods cannot prove anything in the classical, deterministic sense. They will not establish rigid cause-and-effect relationships in the sense that is possible, for example, in the classical Newtonian mechanics. The rigorous statements are obtained only for probabilities, but in the complex and chaotic world we live in, this is often the best one can do.

7.6 Experiments, exercises, and projects

1. Find equations for the MLE of the parameter β in the Gamma distribution. Solve them for the simple exponential distribution with parameter λ. Show that the MLE for λ is unbiased and consistent.

2. Find the MLE $\hat{N}$ for the parameter N in the discrete uniform distribution over integers $1, 2, \dots, N$, based on a sample of size n.

3. Find the MLE for the parameter p of the Bernoulli distribution. Show that it is unbiased and consistent.

4. Find the MLE for the parameter θ of the family of densities

$$f(x;\theta) = \begin{cases} e^{-x+\theta}, & \text{for } x \geq \theta; \\ 0, & \text{otherwise.} \end{cases}$$

5. Find the MM estimators for parameters a and b in the uniform distribution on the interval $[a, b]$.

6. Find the MLE for the parameter λ in the Poisson distribution, based on a random sample of size n. Show that it is unbiased and consistent.

7. Find the MLE for the parameter p in the geometric distribution, based on a random sample of size n.

8. Find the MM estimator for the parameter of the Bernoulli, Poisson and exponential distribution.

9. Find the MM estimator for the parameter θ in the family of distributions from Exercise 4.

10. Draw a random sample of size 35 from the data set `rivet` on the `UVW Web Site`. Do it first with replacements, and then without replacements.

11. Let random variables $X_1, \ldots, X_n$ be independent with an identical symmetric probability distribution. Show that for the unbiased estimators $m_1(\boldsymbol{x}) = \bar{\boldsymbol{x}}$ and $s^2(\boldsymbol{x})$ for the mean and variance, respectively, $\operatorname{Cov}(m_1(X_1, \ldots, X_n), s^2(X_1, \ldots, X_n)) = 0$.

12. Find the likelihood and the log likelihood functions for the gamma distribution with parameters α and β. Then find the equations that define the MLEs for α and β. Can you solve them explicitly? Show that the MLE for $\mu = \alpha\beta$ is $\bar{\boldsymbol{x}}$.

13. Fit a regression model $Y = \beta x + \epsilon$ (i.e., the true regression line passes through the origin) using the LSE to find the estimate for β.

14. Let $t_1, \ldots, t_n$ be a random sample from a Weibull distribution (see Exercise 5.8.26). Find equations for the maximum likelihood estimators $\hat{\alpha}$ and $\hat{\beta}$. Do not try to solve them explicitly though; explain why not.

15. Estimate the correlation dimension of the Cantor set C using 200 randomly selected points.

16. Estimate the dimension of the set $\{(x, y) \in \mathbf{R}^2 : x \in C, 0 \leq y \leq 1\}$. Try to calculate this dimension analytically.

17. Estimate the dimension of the set $\{(x, y) \in \mathbf{R}^2 : x, y \in C\}$. Try to calculate this dimension analytically.

18. Estimate the dimension of the "second-and-fourth-fifth-removed" Cantor-type set. It is obtained by a procedure similar to the "middle-third-removed" procedure that produces the standard Cantor set C, except that the partition at each step consists of five equal subintervals, of which the second and the fourth are removed. Try to calculate this dimension analytically.

7.7 Bibliographical notes

There is a huge literature on statistical inference from, e.g., very practical oriented

[1] S.B. Vardeman, *Statistics for Engineering Problem Solving*, PWS Publishing, Boston, 1994,

[2] D.C. Montgomery and G.C. Runger, *Applied Statistics and Probability for Engineers*, John Wiley, Inc., New York, 1994.

to very theoretical

[3] M.J. Schervish, *Theory of Statistics*, Springer-Verlag, New York, 1995,

with a plethora of more specialized monographs such as

[4] S.R. Searle, *Linear Models*, J. Wiley, 1971,

[5] E.H. Lehman, *Theory of Point Estimation*, John Wiley, New York, 1981,

[6] E.H. Lehman, *Testing Statistical Hypotheses*, John Wiley, New York, 1959,

[7] M. Hollander and G.A. Wolfe, *Nonparametric Statistical Analysis*, John Wiley, New York, 1973.

Additional statistical titles will be quoted in Chapters 8 and 9.

Chapter 8

Statistical Inference for Normal Populations

In this chapter the general assumption is that the statistical model is *normal*. We begin by discussing the general issue of parametric inference and then quickly move to construction of confidence intervals for one-sample models and the related hypothesis testing issues. A few remarks on the two-sample model follow and the chapter concludes with the regression analysis for the normal model and a goodness-of-fit test.

8.1 Introduction; parametric inference

In the simplest *one-sample experimental design*, the model is completely described by the 1-dimensional normal distribution with the density

$$f(x;\mu,\sigma)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \tag{1}$$

where μ is a real number and $\sigma^2>0$. The sample $x_1,\dots,x_n$, comes from a finite number of independently performed experiments represented by a sequence $X_1,\dots,X_n$ of independent random normal quantities, each with the probability density (1). The statistical decisions to be made in this model are about two parameters, μ and σ^2.

A more involved, *two-sample design* calls for sampling from two independent sequences of independent random normal quantities $X_1,\dots,X_n$, and $Y_1,\dots,Y_m$, each sample with its own sample size, n and m, respectively, and possibly different parameters μ_1 and σ_1^2, and, respectively, μ_2 and σ_2^2. The corresponding densities

M. Denker, W. Woyczynski, *Introductory Statistics and Random Phenomena*,
Modern Birkhäuser Classics, DOI 10.1007/978-3-319-66152-0_8

are

$$f_1(x) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left[-\frac{(x-\mu_1)^2}{2\sigma_1^2} \right], \tag{2}$$

and

$$f_2(y) = \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left[-\frac{(y-\mu_2)^2}{2\sigma_2^2} \right]. \tag{3}$$

In the two-sample model there are four parameters to be estimated. Each sample comes from a finite set of independent normally distributed experiments.

However, there is another two-sample model, called the *paired two sample model*, where $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are two samples such that the paired variables $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent. It follows that in each of the subsamples, $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, random variables are independent. In this model it is assumed that the differences

$$D_i = X_i - Y_i \tag{4}$$

are normal with their probability d.f. given by (1). Such models (and we shall investigate one of them in Example 8.1.1) are treated as one-sample models though the original data set has been obtained from a two-sample design.

In a similar fashion one can consider multi-sample designs, normal or not normal. These designs can be further categorized, and we shall deal with these problems in the chapter on ANOVA, the *analysis of variance*. The simplest ANOVA design will be, for example, the model with k independent samples, all with the same variance but, possibly, different means. That is, we will be given a family $X_{ij}, \quad i = 1, \ldots, k, \quad j = 1, \ldots, n_i$, of independent random variables, which, for each i have the identical normal probability d.f.

$$f_i(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu_i)^2}{\sigma^2} \right]. \tag{5}$$

The final normal design to be considered in this chapter is the *regression model* already introduced in Chapters 2 and 8, and specified by equations

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \ldots, n,$$

where β_0 and β_1 are real parameters and where ϵ_i, $i = 1, \ldots, n$, is a sequence of independent normal random variables with zero mean and (mostly unknown) variance $\sigma^2 > 0$. In this design, statistical inference can be made on three parameters β_0, β_1, and $\sigma^2 > 0$.

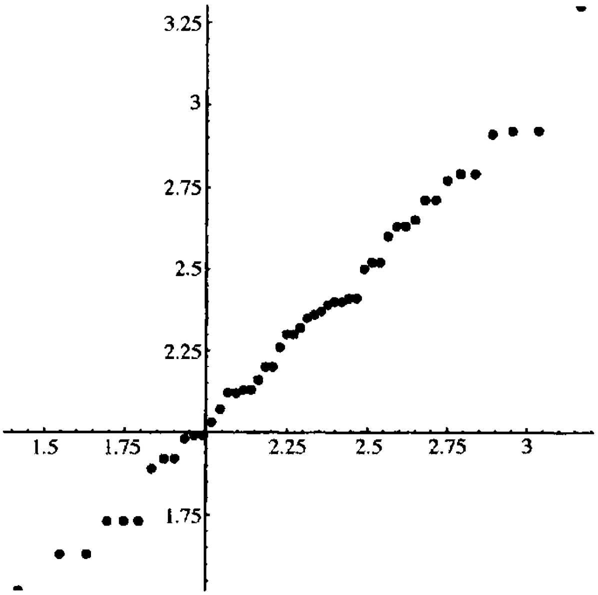
In applications it is most important that the correct model be selected. We will illustrate the model selection process on a number of examples containing concrete experimental data.

Mathematica Experiment 1. Cotton Threads. In order to examine the quality of linen manufactured at a plant, the *quality control* staff has taken a sample of 50 cotton threads and tested their breaking strengths (in kg). The observations, already ordered according to their magnitude, are given in Table 2.5.1. The data represent a one-sample design and the inference to be drawn is about the population mean. Given that the thread's strength depends on the strengths of a large number of independent fibers, the plausible assumption is that the normal model is appropriate. Of course, in the second approximation one could argue that the fibers are not really acting independently of each other given the friction between them. Then a more sophisticated model would be in order.

As the first step in deciding whether or not the data are normal, and what its parameters could be, one usually produces the location, dispersion and shape reports, and a Q-Q plot of the data vs. the normal quantiles with the same mean and standard deviation. We shall use the `COTTON` data from the `UVW Web Site`.

```
In[1]:= <<Statistics'ContinuousDistributions'
In[2]:= <<Statistics'DescriptiveStatistics'
In[3]:= cotton={ 1.10, 2.32,1.52,2.35,1.63,2.36,1.69,2.37,
          1.73, 2.39, 1.73,2.40,1.78,2.40,1.89,2.41,1.92,
          2.47,1.95,2.50, 1.98,2.52,1.99,2.55, 2.02,2.60,
          2.03,2.63,2.07,2.64,2.12, 2.65,2.12,2.71,2.13,
          2.71,2.15,2.77,2.16,2.79,2.20,2.86,
          2.23,2.91,2.26,2.92,2.30,3.02, 2.31,3.30};
In[4]:= LocationReport[cotton]
Out[4]= {Mean -> 2.2912, HarmonicMean -> 2.203,
         Median -> 2.315}
In[5]:= DispersionReport[cotton]
Out[5]= {Variance -> 0.180203, StandardDeviation -> 0.424503,
         SampleRange -> 2.2,  MeanDeviation -> 0.331904,
         MedianDeviation -> 0.29,  QuartileDeviation -> 0.29}
In[6]:= ShapeReport[cotton]
Out[6]= {Skewness -> -0.196803, QuartileSkewness -> -0.0172414,
            KurtosisExcess -> 0.0875748}
In[7]:= Length[cotton]
Out[4]= 50
In[8]:= m=Mean[cotton]
Out[5]= 2.2912
In[9]:= s=StandardDeviation[cotton]
Out[9]= 0.424503
In[10]:= t1=Table[{Quantile[NormalDistribution[m,s],k*0.02],
                              Quantile[cotton,k*0.02]},{k,1,49}]
```

```
Out[10]= {{1.41938, 1.52}, {1.54803, 1.63}, {1.63119, 1.63},
          . . . ,
          {2.95121, 2.92}, {3.03437,2.92}, {3.16302, 3.3}}
In[11]:= ListPlot[t1, AspectRatio->1,
          PlotStyle->PointSize[0.015]]
Out[11]= -Graphics-
```



The mean and median are almost the same, indicating lack of asymmetry in the distribution. The `MeanDeviation` is the *mean absolute deviation*:

$$\frac{1}{n}\sum_i |x_i - \bar{x}| = 0.331904.$$

It is a dispersion parameter which is less sensitive to extreme outliers than the standard deviation. The `MedianDeviation` stands for the *median absolute deviation*

$$\text{med}\left((x_1 - \text{med}(x), \ldots, (x_n - \text{med}(x)\right) = 0.29.$$

The *skeweness*, which is calculated as the third central moment nondimensionalized by dividing it by the cube of the standard deviation

$$\frac{\sum_i (x_i - \bar{x})^3}{\sigma^3(x)} = -0.196803$$

is here quite small, further reinforcing the conclusion that the data are quite symmetric. The fact that it is negative indicates that the underlying probability distribution has a longer left-sided tail.

The `KurtosisExcess` is the *kurtosis coefficient*

$$\frac{\sum_i (x_i - \bar{\boldsymbol{x}})^4}{\sigma^4(\boldsymbol{x})}$$

shifted by -3 so that it is zero for the normal distribution. It is positive for distributions with prominent peaks and heavy tails, and negative for distributions with prominent flanks (relative to the normal distribution). In our case it is very small and equal to 0.0875748, an indication that the normality hypothesis has to be taken seriously. The linearity of the Q-Q plot clearly indicates that the normality assumption is warranted here.

A more formal assessment of the normality of the data can be made on the basis of *nonparametric inference* relying on the *goodness-of-fit* test which applies the *Kolmogorov-Smirnov Theorem* discussed in Section 3.9. Recall that the Kolmogorov-Smirnov Theorem states that if $x_1, \ldots, x_n$, is a random sample from any continuous cumulative d.f. $F(x)$, with order statistics $X_{(1)}, \ldots, X_{(n)}$, and

$$\hat{F}_n(x) = \frac{i}{n}, \qquad \text{for } X_{(i)} \le x < X_{(i+1)},\ i = 0, 1, 2, \ldots, n, \tag{6}$$

is the sample (empirical) cumulative distribution function, then the statistic

$$D_n = \sup_x |\hat{F}_n(x) - F(x)| \tag{7}$$

is independent of the distribution $F(x)$ (in other words, it is *distribution free*), and for every $z \ge 0$,

$$\lim_{n\to\infty} \Pr(D_n n^{1/2} \le z) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}\exp[-2i^2z^2]. \tag{8}$$

In Section 3.9 we have shown how, with the help from *Mathematica*, this result can be used to find d_α such that, for large sample size n,

$$\Pr\{\sqrt{n}D_n \le d_\alpha\} \approx 1 - \alpha. \tag{9}$$

See Section 8.6, for another goodness-of-fit test based on the chi-square distribution.

Mathematica Experiment 2. Rivets. The following table gives the frequency distribution of head diameters for a sample of $n = 500$ rivets. Observe that here—as is often the case—the raw data are no longer available—only a class distribution (sometimes called a grouped distribution) has been presented.

Table 8.1.1 Distribution of head diameters of 500 rivets (in mm)
(Class length: 0.05 mm)

Class midpoint t	Number of rivets ϕ
13.07	1
13.12	4
13.17	4
13.22	18
13.27	38
13.32	56
13.37	69
13.42	96
13.47	72
13.52	68
13.57	41
13.62	18
13.67	12
13.72	2
13.77	1

Since we can assume that the sample was taken from a homogeneous population, and the fluctuations resulted from purely random errors of measurements and manufacturing, the normal model seems to be appropriate here. Note that since we no longer have the list of the original data, but just the grouped data frequencies, only an approximate calculation of the sample mean and variance are possible. Thus, if ϕ_j denotes the number of rivets in class j with the class midpoint t_j, and $n = \sum \phi_j$, the approximate sample mean is the weighted mean, that is

$$\bar{x} \approx \frac{\sum \phi_j t_j}{n} \tag{10}$$

The sample standard deviation also can only be approximated by the formula

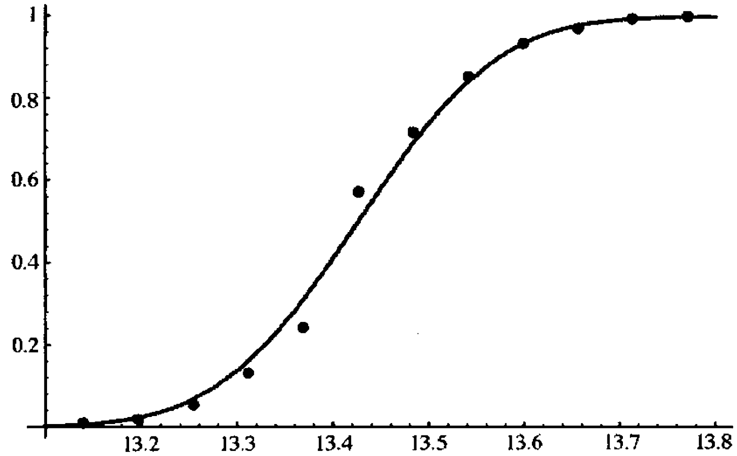
$$s^2 \approx \frac{1}{n-1}\sum_j \phi_j (t_j - \overline{x})^2, \tag{11}$$

and, similarly, for the empirical distribution we obtain the expression

$$\hat{F}_n(t) \approx \frac{1}{n}\sum_j \phi_j H(t - t_j), \tag{12}$$

where $H(t)$ is the Heaviside unit jump function.

```
In[1]:= <<Statistics‘NormalDistribution‘
In[2]:= t={13.07, 13.12, 13.17, 13.22, 13.27, 13.32, 13.37, 13.42,
          13.47, 13.52, 13.57, 13.62, 13.67, 13.72, 13.77};
In[3]:= Length[t]
Out[3]= 15
In[4]:= phi={1, 4, 4, 18, 38, 56, 69, 96, 72, 68, 41, 18, 12, 2, 1};
In[5]:= barx=(1/500)Sum[phi[[j]]*t[[j]],{j,1,15}]
Out[5]= 13.4264
In[6]:= s=Sqrt[ (1/499)Sum[phi[[j]]*(t[[j]]-barx)^2,{j,1,15}]]
Out[6]= 0.
In[7]:= H[x_]:=If[x<0,0,1]
In[8]:= F[x_]:=(1/500) Sum[ phi[[j]] * H[ x-t[[j]] ],{j,1,15}]
In[9]:= tab= Table[{barx-3s+0.5*k*s,F[barx-3s+0.5*k*s]},{k,12}];
In[10]:= p1=ListPlot[ tab,PlotStyle->PointSize[0.015]];
In[11]:= p2=Plot[CDF[NormalDistribution[barx,s],x],{x,13.1,13.8}];
In[12]:= Show[p1,p2]
Out[12]= -Graphics-
```



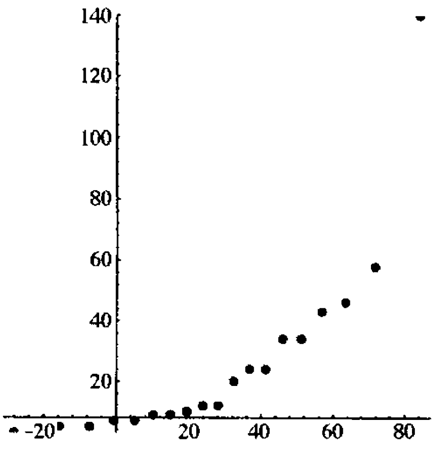
Mathematica Experiment 3. Pesticide Effectiveness. In an experiment testing the effectiveness of a pesticide, two populations of flies were exposed to nerve gas for 30 and 60 seconds, respectively. The quantity measured was the time elapsed from the instant the fly was exposed to the pesticide to the moment when it could no longer stand up. The design of the experiment called for assigning at random (using the pseudorandom number generator) a total of 31 flies to one of the two exposures. Thus, the model was a two-sample design. A *Mathematica* session analyzing the data follows.

```
In[1]:= <<Statistics'ContinuousDistributions'
In[2]:= <<Statistics'DescriptiveStatistics'
In[3]:= t30= {3., 5., 5., 7., 8.99, 8.99, 10., 12., 20., 24., 24.,
                      34., 43.1, 46., 57.9, 140.};
In[4]:= t60={2., 5., 5., 7., 8., 8.99, 14., 18., 24., 26., 26., 34.,
                      37., 42., 89.9};
In[5]:= {Length[ t30],Length[ t60]}
Out[5]= {16, 15}
In[6]:= {ShapeReport[t30], ShapeReport[t60]}
Out[6]= {{Skewness -> 2.12991, QuartileSkewness -> 0.474753,
                                    KurtosisExcess -> 4.28096},
          {Skewness -> 1.63001, QuartileSkewness -> 0.133238,
                                    KurtosisExcess -> 2.44}}
In[7]:=  m={Mean[ t30], Mean[ t60]}
Out[7]= {28.0641, 23.1278}
In[8]:= s={StandardDeviation[ t30], StandardDeviation[ t60]}
Out[8]= {34.1968, 22.4419}
In[9]:= q1=Table[{Quantile[NormalDistribution[m[[1]],s[[1]]],k*0.05],
                 Quantile[t30,k*0.05]},{k,1,19}];
In[10]:= ListPlot[q1, PlotRange->{3,140}, AspectRatio->1]
```

Table 8.1.2 Times of reaction to pesticide of 31 (16+15) flies

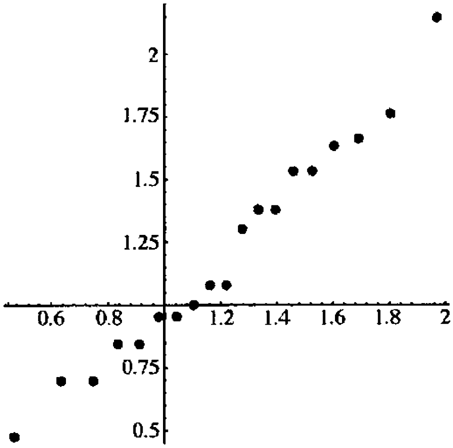
Exposure time: 30 seconds	Exposure time: 60 seconds
Logarithm of reaction time	Logarithm of reaction time
0.477	0.301
0.699	0.699
0.699	0.699
0.845	0.845
0.954	0.903
0.954	0.954
1.000	1.146
1.079	1.255
1.302	1.380
1.380	1.415
1.380	1.415
1.532	1.532
1.634	1.568
1.663	1.623
1.763	1.954
2.146	—

```
Out[10]= -Graphics-
```



The session using the original data shows a skewed to the left underlying distribution and a very nonlinear Q-Q plot. So, the normality hypothesis has to be rejected. However, the Q-Q plot suggests that the logarithmic transformation of the data could produce an approximately normal distribution of values.

```
In[11]:= logt30=  N[Table [Log[t30[[i]], 10], {i,1,16}], 4]
Out[11]= {0.477, 0.699, 0.699, 0.845, 0.954, 0.954, 1., 1.079,
           1.302, 1.38, 1.38,  1.532, 1.634, 1.663, 1.763, 2.146}
In[12]:= ShapeReport[logt30]
Out[12]= {Skewness -> 0.238372, QuartileSkewness -> 0.1485,
                                KurtosisExcess -> -0.997441}
In[13]:= logm= Mean[logt30]
Out[13]= 1.21919
In[14]:= logs= StandardDeviation[logt30]
Out[14]= 0.0456112
In[15]: q2= Table[{Quantile[NormalDistribution[logm,logs], k*0.05],
                     Quantile[logt30,k*0.05]},{k,1,19}];
In[16]:= ListPlot[q2, PLotRange->{0.45,2.2},AspectRatio->1]
Out[16]= -Graphics-
```



A similar analysis of the data t60 is left as an Exercise. The logarithmic transformation of the data (also shown in Table 8.1.1) dramatically decreased skewness and kurtosis excess. The Q-Q plot is approximately linear. Now the transformed data can be analyzed under a two-sample normal model to be discussed in Section 8.4. A random quantity X such that log X has the normal distribution is said to have a *log-normal distribution*. Such distributions appear in many areas of engineering and sciences.

Example 8.1.1 Tire Wear.

A car company wants to compare the durability of two types, A and B, of tires to be mounted on the company's cars as the original equipment. The experimental design calls for nine cars being selected at random from the production lines and mounted with tires A, and then another nine cars being selected at random and mounted with tires B. Each car is driven for 20,000 miles under normal conditions and the wear is quantified by measuring at the end of the experiment the remaining thread depth in millimeters. Table 8.1.3 gives the results of the experiment. Clearly, a two-sample design with equal variances is called for here.

Table 8.1.3 Remaining thread depth (in mm) after 20,000 miles

Type A tire		Type B tire	
Car no.	Wear	Car no.	Wear
1	12.0	1	10.2
2	11.4	2	11.3
3	12.2	3	12.4
4	11.3	4	10.7
5	11.7	5	11.2
6	12.1	6	12.0
7	12.3	7	12.2
8	11.2	8	11.1
9	12.2	9	11.7

An argument can be made that a random selection of the two groups of cars introduced too much randomness in the design, making the comparison more difficult. The experimental design can be adjusted to answer this criticism by, for example, mounting two types of tires on the same car, say, type A on the left-hand side, type B on the right-hand side (or vice versa). In this case we could consider a one–sample design for the differences D_i of observations provided in Table 8.1.4.

Table 8.1.4 One-sample design for measuring tire wear

Car	Wear A	Wear B	$D_i = A_i - B_i$
1	12.0	10.2	1.8
2	11.4	11.3	0.1
3	12.2	12.4	-0.2
4	11.3	10.7	0.6
5	11.7	11.2	0.5
6	12.1	12.0	0.1
7	12.3	12.2	0.1
8	11.2	11.1	0.1
9	12.2	11.7	0.5

If the random quantities A and B are normal and can be assumed independent, then the difference $D = A - B$ is also a normal random quantity with mean $\mu = \mu_1 - \mu_2$ and unknown variance.

Example 8.1.2

The same car manufacturing company is interested in durability of three types of tires, A, B, and C, under three different types of weather conditions: snow, rain, and dry (more factors such as driving habits, traffic type, or road conditions could also be introduced here). This leads to a multi-sample model given by a *two-way classification design*. For each of the combinations of tire type—A,B, and C—and weather condition—snow, rain, and dry—we make one observation X_{ij} and arrange it in the matrix:

	A	B	C
snow	X_{11}	X_{12}	X_{13}
rain	X_{21}	X_{22}	X_{23}
dry	X_{31}	X_{32}	X_{33}

This is the design matrix for our experiment. Of course, the measurement of each of X_{ij} can be independently repeated a finite number n_{ij} of times. One says then that each cell ij contains n_{ij} independent observations. This would give more precise statistical information.

Mathematica Experiment 6. Lumberjack's Hypothesis. A forester needs to estimate the volume of lumber in a stand of trees. The conventional wisdom among lumberjacks is that the volume V of usable lumber in a tree is equal to 100 plus two thirds of the square C^2 of the circumference C of the tree trunk measured 24 inches above the ground. The measurements of both, the circumferences and the volumes, from a random sample of size 25 are given in Table 8.1.5.

The design clearly calls for a linear regression model in variable C^2. If V_i denotes the lumber volume of the i-th tree (selected at random) then the working hypothesis is that $V_i = \beta_0 + \beta_1 C_i^2 + \epsilon_i$, where C_i^2 denotes the square of the circumference of the i-th tree. In this example, the first guess is that the residual errors ϵ_i are random and normally distributed, with expectation 0 and common variance $\sigma^2 > 0$. Thus, the initial job here is to find the regression line

$$EV_i = \beta_0 + \beta_1 C_i^2.$$

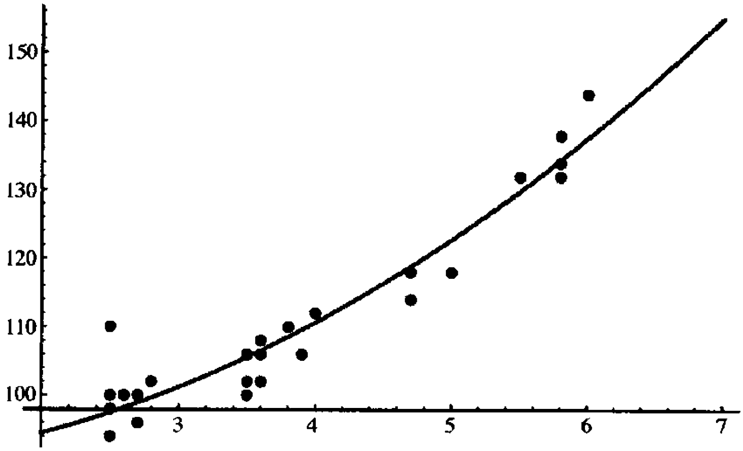
Here is a simple fit via *Mathematica*.

```
In[1]:= <<Statistics'LinearRegression'
In[2]:= circum={4.7, 2.7, 3.5, 2.5, 4., 3.6, 2.7, 5.5, 5.,
```

Table 8.1.5 Circumferences C of tree trunks and lumber volumes V

No. of tree	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
C (in ft)	4.7	2.7	3.5	2.5	4.0	3.6	2.7	5.5	5.0	2.6	2.5	6.0	3.8	5.8	3.9	2.8	5.8	3.5	3.6	2.5	3.5	4.7	3.6	2.5	5.8
C^2 (in ft^2)	22.09	7.29	12.25	6.25	16.0	12.96	7.29	30.25	25.0	6.76	6.25	36.0	14.44	33.64	15.21	7.84	33.64	12.25	12.96	6.25	12.25	22.09	12.96	6.25	33.64
V (in ft^3)	114	96	100	110	112	108	100	132	118	100	94	144	110	134	106	102	138	102	102	100	106	118	106	98	132

```
         2.6, 2.5, 6., 3.8, 5.8, 3.9, 2.8, 5.8, 3.5, 3.6, 2.5,
         3.5, 4.7, 3.6, 2.5, 5.8};
In[3]:= volume={114, 96, 100, 110, 112, 108, 100, 132, 118, 100,
          94, 144, 110, 134, 106, 102,  138, 102, 102, 100, 106,
         118, 106, 98, 132};
In[4]:= data=Table[{circum[[i]], volume[[i]]}, {i,25}]
Out[4]= {{4.7, 114}, {2.7, 96}, {3.5, 100}, {2.5, 110},
       {4., 112}, {3.6, 108}, {2.7, 100}, {5.5, 132}, {5., 118},
       {2.6, 100}, {2.5, 94}, {6., 144}, {3.8, 110}, {5.8, 134},
       {3.9, 106}, {2.8, 102}, {5.8, 138}, {3.5,102}, {3.6, 102},
       {2.5, 100}, {3.5, 106}, {4.7, 118}, {3.6, 106}, {2.5, 98},
       {5.8, 132}}
In[5]:= Fit[data,{1,x^2}, x]
Out[5]= 89.0685  + 1.34841*x^2
In[6]:= f[x_]:=89.0685  + 1.34841*x^2;
In[7]:= p1= ListPlot[data];
In[8]:= p2= Plot[f[x], {x, 2, 7}];
In[9]:= Show[p1, p2]
Out[9]= -Graphics-
```



So the best fitting curve comes close to the lumberjacks' lore. The issue for the statistician, which we will address in Section 8.5, is whether this hypothesis is justifiable and whether the normality assumption on the residuals is valid.

8.2 Confidence intervals for one-sample model

In this section $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is a random sample of size n, from the normal population with the density

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \tag{1}$$

where the mean $\mu \in \mathbf{R}$, and variance $\sigma^2 > 0$. Our goal here is to do statistical inference for the parameters μ and σ^2 when one (or both of them) is unknown. We shall do it first by constructing the *confidence intervals* for them. In the simple special case of Bernoulli data, we have introduced this simple but effective estimation method in Section 3.7.

Consider the sample mean

$$\hat{\mu} = \bar{\boldsymbol{x}}, \tag{2}$$

which happens to be the maximum likelihood estimator (MLE, see Chapter 7) for the true value of the unknown parameter μ. It is clear that with probability 1, the numerical value of the estimate $\hat{\mu}$ is different from μ, but we would certainly hope that they are close to each other. The statistical and *coupled* questions are: "How close?", and "With what probability?".

More exactly, having selected a precision level $\epsilon > 0$, we would like to repeat the above estimation procedure independently numerous times and find the relative frequency of the event "the estimate $\hat{\mu} = \bar{x}$ is within ϵ of the true value μ of the mean", or, in other words,

$$\mu \in [\bar{x} - \epsilon, \bar{x} + \epsilon]. \tag{3}$$

Note that the above interval is random; that is, it depends on the realization of the above estimation procedure.

If, more formally, we interpret the sample $x = (x_1, x_2, \dots, x_n)$ as a realization of independent random quantities $(X_1, X_2, \dots, X_n) = \boldsymbol{X}$, each with density (1), then the question about the relative frequency of the event (3) becomes, in view of the Law of Large Numbers, a more easily answered question about the probability

$$\gamma(\mu, \epsilon, n) = \Pr\{\bar{X} - \epsilon \leq \mu \leq \bar{X} + \epsilon\} = \Pr\{\mu - \epsilon \leq \bar{X} \leq \mu + \epsilon\}, \tag{4}$$

where

$$\bar{X} = \frac{X_1 + \dots + X_n}{n}. \tag{5}$$

The probability $\gamma(\mu, \epsilon)$ depends, *a priori*, on the unknown parameter μ, the sample size n, and the precision level ϵ.

In the case when the probability $\gamma(\mu, \epsilon, n)$ does not depend on the true value of parameter μ, i.e., $\gamma(\mu, \epsilon, n) = \gamma(\epsilon, n)$ then the random interval

$$[\bar{X} - \epsilon, \bar{X} + \epsilon] \tag{6}$$

is called an $\gamma \times 100$ percent confidence interval for μ. The probability $\gamma = \gamma(\epsilon, n)$ is also called the *confidence level*. Traditionally, one selects a small positive α, and one talks about the $(1 - \alpha)$-confidence level. This notation is justified by the hypothesis testing notation in Section 8.4.

In more generality, we will record the following definition:

Definition 8.2.1 Confidence Intervals.

Let $X_1, \dots, X_n$ be a random sample, i.e., independent, identically distributed random quantities, from the unknown normal distribution with the density (1) with parameters μ and σ^2. A $(1 - \alpha) \times 100\%$ confidence interval for μ (respectively, σ^2) is defined by two estimators

$$L_\mu = L_\mu(X_1, \dots, X_n), \qquad \text{and} \qquad U_\mu = U_\mu(X_1, \dots, X_n), \tag{7}$$

such that, regardless of the true value of μ,

$$\Pr\{L_\mu \leq \mu \leq U_\mu\} = 1 - \alpha \tag{8}$$

(respectively, by L_{σ^2} and U_{σ^2} such that, $\Pr\{L_{\sigma^2} \leq \sigma^2 \leq U_{\sigma^2}\} = 1 - \alpha$*).*

Note that in the case of the MLE estimator $\hat{\mu} = \bar{X}$ for μ

$$L_\mu(X_1, \ldots, X_n) = \bar{X} - \epsilon, \quad \text{and} \quad U_\mu(X_1, \ldots, X_n) = \bar{X} + \epsilon. \tag{9}$$

In practice, the problems are:

- Find the size ϵ of the confidence interval given the sample size n and the confidence level $1 - \alpha$.

- Find the sample size n that would guarantee the desired confidence level $1 - \alpha$, for a given size ϵ of the confidence interval.

The latter is, of course, one of the first questions in experimental design.

Example 8.2.1 Confidence Intervals for Unknown μ, with Known σ^2.
If σ^2 is known, then the random quantity

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \tag{10}$$

is normal, with mean zero and variance 1. Hence,

$$\Pr\{-a \leq Z \leq a\} = 1 - 2\Big(1 - \Phi(a)\Big) = 2\Phi(a) - 1 \tag{11}$$

where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution.

Thus, given the confidence level α, and the tail quantile function $z_\alpha = \Phi^{-1}(1 - \alpha)$ of the standard normal distribution, we find that for

$$a = z_{\alpha/2}, \tag{12}$$

see Fig. 8.2.1, the probability from (11) is at the desired confidence level $1 - \alpha$, i.e.,

$$2\Phi(a) - 1 = 1 - \alpha. \tag{13}$$

In other words, putting (10), (11), and (13) together,

$$\Pr\left\{\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha.$$

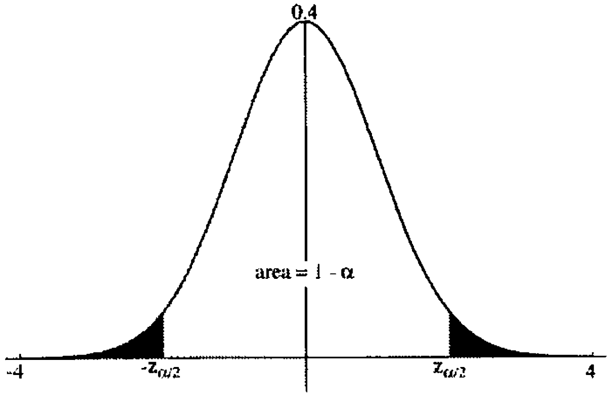


FIGURE 8.2.1
Tail quantiles of the $N(0, 1)$ *distribution. The tail quantile* $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$, *where* $\Phi^{-1}(\alpha)$ *is the quantile function of the standard normal distribution. The area under the density, between* $-z_{\alpha/2}$ *and* $z_{\alpha/2}$, *is* $1 - \alpha$.

Alternatively, we can say that with

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

the interval $[\bar{X} - \epsilon, \bar{X} + \epsilon]$ is the $(1 - \alpha) \times 100\%$ confidence interval. The *length of the confidence interval* is $2\epsilon = 2z_{\alpha/2}\sigma/\sqrt{n}$.

So, for example, if the desired confidence level $1 - \alpha = 0.95$, then the corresponding $a = 1.96$ in (12), and the random interval

$$\left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right] \tag{14}$$

is the 95% confidence interval for μ. In particular, if the known variance $\sigma = 2$ and the sample size $n = 10,000$, then, with probability 0.95, the true mean μ will be inside the interval $[\bar{X} - 0.0392, \bar{X} + 0.0392]$.

On the other hand, assuming the same known variance $\sigma = 2$, if we demand that the size of the confidence interval be $2\epsilon = 0.1$, and the confidence level be $1 - \alpha = .95$, then the sample size n has to be at least such that

$$2 \times 1.96\frac{2}{\sqrt{n}} = 0.1$$

so that the sufficient condition for the sample size n is that $n \geq (78.4)^2 \approx 6147$.

Mathematica Experiment 1. Cotton Threads. The Statistics 'ConfidenceIntervals' package implements the above algorithm. The commands are self-explanatory. We will apply it, at different confidence levels, to find the confidence intervals for the mean of data Cotton from Section 8.1, assuming that the true variance σ^2 is known and equal to 0.18.

```
In[1]:= <<Statistics'ConfidenceIntervals'
In[2]:= <<Statistics'ContinuousDistributions'
In[3]:= cotton={ 1.10, 2.32,1.52,2.35, . . . , 3.02, 2.31,3.30}
In[4]:= MeanCI[cotton, KnownVariance->0.18,
         ConfidenceLevel->0.9]
Out[4]= {2.19251, 2.38989}
In[5]:= MeanCI[cotton, KnownVariance->0.18,
         ConfidenceLevel->0.95]
Out[5]= {2.1736, 2.4088}
In[6]:= MeanCI[cotton, KnownVariance->0.18,
         ConfidenceLevel->0.99]
Out[6]= {2.13665, 2.44575}
```

Of course, the length of the confidence interval increases as we increase the desired confidence level. Next, we calculate the sample size n sufficient to guarantee the confidence levels $1 - \alpha$ with a prescribed confidence interval size 2ϵ. The command Ceiling[x] calculates the smallest integer greater than or equal to x.

```
In[7]:= Q[clevel_]:= Quantile[NormalDistribution[0,1], clevel]
In[8]:= samplesize[ clevel_, sigma_, epsilon_]:=
         Ceiling[(sigma*Q[(1+clevel)/2]/epsilon)^2]
In[9]:= {samplesize[ 0.95, 0.4, 0.2] ,samplesize[ 0.95, 0.4, 0.1],
          samplesize[ 0.95, 0.4, 0.05]}
Out[9]= {16, 62, 246}
In[10]:= {samplesize[ 0.90, 0.4, 0.1],samplesize[ 0.95, 0.4, 0.1],
          samplesize[ 0.99, 0.4, 0.1]}
Out[10]= {44, 62, 107}
```

Obviously, given the confidence level, the necessary sample size grows as the desired confidence interval's size decreases. On the other hand, for a fixed size of the confidence interval, the necessary sample size increases as the desired confidence level increases.

Construction of confidence intervals for the mean μ with known variance σ^2 is relatively easy because the probability distribution of the standardized sample mean $\bar{X}$ remains normal. Other cases are not so simple and more complicated statistics L and U have to be considered. If the variance is unknown, it has to be replaced by the sample variance which is a random quantity. Hence, the size of the confidence interval becomes random as well.

Example 8.2.2 Confidence Intervals for μ with Unknown σ^2.
In the case of unknown μ and σ^2, simultaneously estimated by the unbiased estimators

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n}, \quad \text{and} \quad S^2(X) = \frac{(X_1 - \bar{X})^2 + \ldots + (X_n - \bar{X})^2}{n-1}, \tag{15}$$

the critical observation is that the condition

$$\mu \in [\bar{X} - \epsilon, \bar{X} + \epsilon] \tag{16}$$

can be rewritten as an equivalent condition

$$\frac{-\sqrt{n}\epsilon}{S} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq \frac{\sqrt{n}\epsilon}{S}, \tag{17}$$

and that the random variable

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \tag{18}$$

has the standard Student t-distribution with $n - 1$ degrees of freedom, which is *independent* of μ and σ^2! The values of its upper tail quantile function $t_\alpha(n - 1) = Q(1 - \alpha; n - 1)$, where $Q(\beta; n - 1)$ is the quantile function of the Student t-distribution with $n - 1$ degrees of freedom, are provided in the table in Appendix F. They are also available in *Mathematica* `Statistics 'ContinuousDistributions'` package. Hence, if $\bar{X}$ and $S^2 = S^2(X)$ are the sample mean and unbiased sample variance, respectively, used as the estimators for μ and σ^2, then the $(1 - \alpha) \times 100\%$ confidence interval for μ is

$$\left[\bar{X} - t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}, \ \bar{X} + t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right]. \tag{19}$$

Notice that in this case, finding the sample size n sufficient to guarantee the desired confidence level α and the desired confidence interval size 2ϵ is not that simple as n enters in the interval construction (19) not only through $\sqrt{n}$ but also through the tail quantile $t_{\alpha/2}(n - 1)$; explicit analytic solution is not feasible here.

In general, the Student t-distributions are flatter, with broader flanks and lower peaks than the N(0,1) distribution, see Example 3.8.8. This bigger dispersion is easy to explain: the lack of information about the variance causes more variability. In the *Mathematica* Experiment 3.8.7, the Student t-density with 2 degrees of freedom has the lowest peak and the one with 35 degrees of freedom, has the highest. But as the number of degrees of freedom increases, the Student t-distribution converges to the N(0,1) distribution; for $n = 25$ the difference is already very small. Again, it is easy to understand. For large n, the sample variance $S^2(\boldsymbol{X})$ well approximates the true variance σ^2; it is a consistent estimator.

Mathematica Experiment 2. Cotton Threads. The command `MeanCI [data, ConfidenceLevel->clevel]` automatically calculates the confidence intervals for the parameter μ with unknown variance, using the Student t-distribution, at a given confidence level `clevel`. We will apply it to the `cotton` data used above. Not surprisingly, the confidence intervals are a little wider than in the `Mathematica` Experiment 1.

```
In[1]:= <<Statistics'ConfidenceIntervals'
In[2]:= cotton={ 1.10, 2.32,1.52, . . . ,3.02, 2.31,3.30}
In[3]:= MeanCI[cotton, ConfidenceLevel->0.9]
Out[3]= {2.19055, 2.39185}
In[4]:= MeanCI[cotton, ConfidenceLevel->0.95]
Out[4]= {2.17056, 2.41184}
In[5]:= MeanCI[cotton, ConfidenceLevel->0.99]
Out[5]= {2.13031, 2.45209}
```

Example 8.2.3 Confidence Intervals for σ^2.

In this estimation problem we will utilize the fact that if $X_1, \ldots, X_n$, are independent normal random quantities with identical densities $f(x; \mu, \sigma^2)$, then the random quantity

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 \tag{20}$$

has the chi-square $(\chi^2(n))$ distribution with n degrees of freedom and the random quantity

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2 \tag{21}$$

has the $\chi^2(n-1)$ distribution with $n-1$ degrees of freedom, see Examples 3.8.7 and 5.5.4, where you can also find the plots of selected chi-square densities. Consequently, if μ is known, and $\chi^2_\alpha(n) = Q(1-\alpha; n)$ denotes the chi-square tail quantile function (which can be calculated via *Mathematica*; selected values of the chi-square quantile function $Q(\alpha; n)$ are given in Appendix F), then we have that the probabilities

$$\Pr\left\{\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 \le \chi^2_{1-\alpha/2}(n)\right\} \tag{22}$$

and

$$\Pr\left\{\chi^2_{\alpha/2}(n) \le \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2\right\} \tag{23}$$

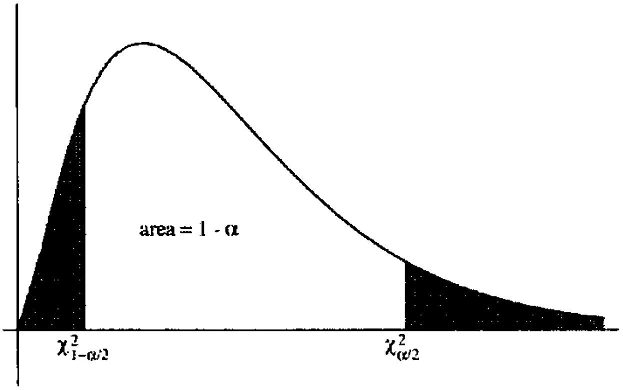


FIGURE 8.2.2
Selection of the confidence interval for σ^2 using symmetric tail quantiles $\chi^2_\alpha(n)$ of the chi-square distribution.

are both equal to $\alpha/2$ (see Fig. 8.2.2), so that

$$\Pr\left\{\frac{1}{\chi^2_{\alpha/2}(n)}\sum_{i=1}^{n}(X_i-\mu)^2 \le \sigma^2 \le \frac{1}{\chi^2_{1-\alpha/2}(n)}\sum_{i=1}^{n}(X_i-\mu)^2\right\} = 1-\alpha. \quad (24)$$

In other words,

$$\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n)}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n)}\right] \quad (25)$$

is a $(1-\alpha)\times 100\%$ confidence interval for σ^2.

If μ is not known, replacing μ by $\bar{X}$, and the tail quantiles of the $\chi^2(n)$ distribution by those of the $\chi^2(n-1)$ distribution gives the $(1-\alpha)$ 100% confidence interval of the form

$$\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right] \quad (26)$$

The package `Statistics'ConfidenceIntervals'` automatically constructs the confidence intervals for σ^2 with μ unknown, via the command `VarianceCI[data, ConfidenceLevel-> clevel]`.

The above formulas for confidence intervals are summarized in Table 8.2.1. Recall that

$$\frac{\bar{X} = X_1 + \ldots + X_n}{n} \quad (27)$$

denotes the sample mean, and that

$$S^2 = \frac{(X_1 - \bar{X})^2 + \ldots + (X_n - \bar{X})^2}{n - 1} \tag{28}$$

is the unbiased sample variance.

Table 8.2.1 Confidence intervals for μ and σ^2

Parameter estimated	Other parameter	$(1-\alpha)$-confidence interval
μ	σ^2 known	$\bar{X} - \frac{\sigma z_{\alpha/2}}{\sqrt{n}} \le \mu \le \bar{X} + \frac{\sigma z_{\alpha/2}}{\sqrt{n}}$
	σ^2 unknown	$\bar{X} - \frac{S t_{\alpha/2}(n-1)}{\sqrt{n}} \le \mu \le \bar{X} + \frac{S t_{\alpha/2}(n-1)}{\sqrt{n}}$
σ^2	μ known	$\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{\alpha/2}(n)} \le \sigma^2 \le \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{1-\alpha/2}(n)}$
	μ unknown	$\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}$

8.3 From confidence intervals to hypothesis testing

Let $X_1, \ldots, X_n$ be a random sample from a distribution with an unknown parameter θ. Suppose that we have found an $(1 - \alpha)$-confidence interval I for the parameter θ. In other words, with probability $1 - \alpha$, the true value of the parameter is contained in the random region I which depends on the particular sample. Statistical *hypothesis testing* is the process that leads to a decision on the truth/falsity of the hypothesis that the parameter θ is located in a given, fixed, and nonrandom (i.e., independent of the sample) region H. Armed with information about the confidence intervals we can now follow one of the several alternatives:

1. If $I \cap H = \emptyset$, i.e., H has an empty intersection with I, then it is *unlikely* that the true value of parameter θ lies in H since it is likely that θ is in I. In this case, the decision should *reject the hypothesis* H, that is proclaim that the true value of θ is not in H but in its complement.
2. If $I \subset H$, i.e., I is contained in H, then it is *likely* that the true value of parameter θ lies in H since it is likely that θ is in I. In this case, the decision should be *not to reject the hypothesis* H.
3. If both $I \cap H \neq \emptyset$ and $I^c \cap H \neq \emptyset$, i.e., H has a nonempty intersection with I and with its complement I^c, the situation is obviously ambiguous. The accepted scientific method calls for keeping a hypothesis as a working hypothesis, i.e., not rejecting it as long as it is not disproved.

The two possible decisions are thus:

I. Reject the hypothesis H if $I \cap H = \emptyset$.

II. Do not reject the hypothesis H if $I \cap H \neq \emptyset$.

Obviously, either of decisions I and II can be erroneous in view of randomness of the confidence interval I, but the error probabilities can be estimated.

Type I Error: It occurs when the true hypothesis H is rejected. The upper bound for the probability of this error is

$$\Pr\{\theta \in H\} \leq \Pr\{\theta \in I^c\} = \alpha. \tag{1}$$

Type II error: It occurs when the false hypothesis H is not rejected. However, if the confidence level $1-\alpha$ is high, that is, α is small, the probability of type II error cannot be made small since

$$\Pr\{H \text{ not rejected}\} = \Pr\{I \cap H \neq \emptyset\}$$

can be as large as $1-\alpha$ if the true value of parameter θ is in $I \cap H^c$.

The above discussion exemplifies the general hypothesis testing problem where, on the basis of a finite sample $x_1, \ldots, x_n$, from an unknown distribution F, the decision has to be made on whether or not F belongs to a specified family H of distributions. In *parametric problems* such a family is described by certain parameters belonging to a preselected range. For example, the normal family is parametrized by two parameters, the mean $-\infty < \mu < \infty$, and the variance $\sigma^2 > 0$, and we may want, for a given μ_0, to test the hypothesis $H : \mu \geq \mu_0$ vs. the alternative $H_1 : \mu < \mu_0$. The Bernoulli distributions are parametrized by the single parameter $0 < p < 1$, and we may want to test the hypothesis $H_0 : p \geq 1/2$ vs. its alternative $H_1 : p < 1/2$.

Let us denote by C the set of all points $(x_1, \ldots, x_n)$ in $\mathbf{R}^n$ for which the associated confidence interval I based on $x_1, \ldots, x_n$, has an empty intersection with H ($I \cap H = \emptyset$), i.e., given random sample $x_1, \ldots, x_n$, the hypothesis H is rejected. The set C is called the *critical region* of the test based on the confidence interval I. In general, we shall introduce the following definition:

Definition 8.3.1 Critical Region.
A test procedure for hypothesis H specifies a region C in the n-dimensional space $\mathbf{R}^n$, called the critical or rejection region. If the sample vector $(x_1, \ldots, x_n)$ is located in the critical region C, then the hypothesis H is rejected.

Of course, as the above discussion indicates, decisions based on the critical region C may be wrong for one of two reasons:

Type I Error. The hypothesis H is true but $(x_1, \ldots, x_n) \in C$ and H is rejected;

Type II Error. The hypothesis H is false but $(x_1, \ldots, x_n) \notin C$ and H is not rejected.

The probability of the type I error is bounded from above by the number α which is the maximum, taken over all F from the family H, of the probabilities that the random vector $(X_1, \ldots, X_n)$, with independent components X_i with identical cumulative d.f. $F(x)$, has values in the critical region C:

$$\alpha = \max_{F \in H} \Pr\left\{(X_1, \ldots, X_n) \in C\right\} = \max_{F \in H} \int \ldots \int_C dF(x_1) \ldots dF(x_n) \qquad (2)$$

$$= \max_{\theta \in H} \int \ldots \int_C f(x_1; \theta) \cdot \ldots \cdot f(x_n; \theta)\, dx_1 \ldots dx_n,$$

in case the density exists. The number α is called the *significance level* of the test based on the critical region C. Clearly, it is desirable that the critical region C be selected so that the significance level α is as small as possible.

Calculation of the probability of the type II error, that is the probability that we do not reject the false hypothesis H, is usually very difficult, although a bound can be found by minimizing the type II error of the test

$$\beta(F) = \Pr\{(X_1, \ldots, X_n) \notin C\} \qquad (3)$$

over F outside H.

It is also quite clear that if the critical region C is shrunk, then the corresponding significance level α decreases. However, this decrease of the probability of the type I error also leads to the increase of the probability β of the type II error. Thus, optimization of the hypothesis testing procedure by selection of the critical region C so that the probabilities of the type I and type II errors are minimized simultaneously is impossible. The significance level α of the smallest (in a certain class) critical region C which causes rejection of a given random sample $x_1, \ldots, x_n$, is called the *P-value* of the random sample (for this class). Its precise meaning will be explained below.

In the parametric case, the common approach, which combines the study of probabilities of type I and type II errors, is to consider the *power function* of the test

$$\pi(\theta, C) = \Pr\{H \text{ is rejected while the true value is } \theta\}. \tag{4}$$

Thus $\pi(\theta, C)$ is the probability of type I error for $\theta \in H$, and $p(\theta) = 1 - \pi(\theta, C)$ is the probability of type II error for $\theta \notin H$.

The α-significance level test procedure based on the critical region C is said to be *optimal at a given significance level* α if, for any other α-significance level critical region C_1, and any $\theta \notin H$,

$$\pi(\theta, C_1) \leq \pi(\theta, C). \tag{5}$$

In other words, for the optimal critical region at a given significance level α, the probability of the type II error is the smallest possible.

Example 8.3.1 Hypothesis Testing for the Mean of the Normal Distribution.

We shall consider the *simple hypothesis* H_0 that says that the sample $x_1, \ldots, x_n$ is a random sample from the $N(\mu_0, \sigma^2)$ distribution with known σ^2, and test it against the *simple alternative hypothesis* H_1 that says that the sample comes from the $N(\mu_1, \sigma^2)$ distribution with, say, $\mu_1 > \mu_0$. Assuming tacitly the normality of the population, this is often written as the testing problem for the *null hypothesis* $H_0 : \mu = \mu_0$ vs. the *alternative hypothesis* $H_1 : \mu = \mu_1$. In this case, the whole parameter space under consideration consists of just two points, μ_0 and μ_1.

We will base our test on the sample mean statistics $\bar{x}$. Note that under H_0 the distribution of the random quantity $\bar{X}$ is $N(\mu_0, \sigma/\sqrt{n})$, and that under H_1 the distribution of $\bar{X}$ is $N(\mu_1, \sigma/\sqrt{n})$, see Fig. 8.3.1.

We will consider the critical regions of the form

$$C_c = \left\{(x_1, \ldots, x_n) : \bar{x} = \frac{x_1 + \ldots + x_n}{n} \geq c\right\} \subset \mathbf{R}^n \tag{6}$$

In other words, we will reject $H_0 : \mu = \mu_0$ if the sample mean $\bar{x}$ exceeds a certain threshold value c. Then, assuming that $H_0 : \mu = \mu_0$ is true, the probability of the type I error is

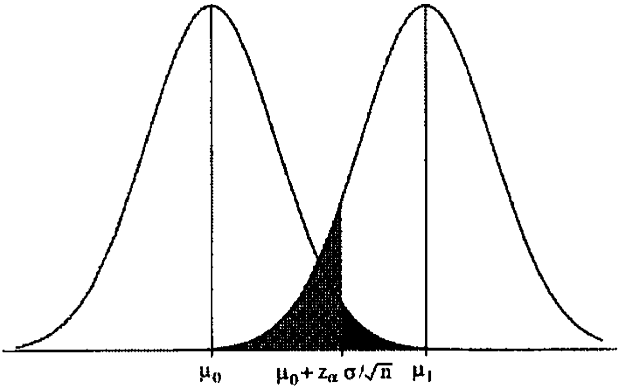


FIGURE 8.3.1
Probability d.f.s of $\bar{X}$ under $H_0 : \mu = \mu_0$ (left curve), and under $H_1 : \mu = \mu_1$ (right curve). At the significance level α, the rejection region, expressed in terms of the statistic $\bar{x}$, is to the right of the point $\mu_0 + z_\alpha\sigma/\sqrt{n}$. The shaded region on the right has the area $= \alpha$, probability of type I error, and the shaded region on the left has the area $= \beta$, probability of type II error.

$$\begin{aligned}\alpha = \pi(\mu_0, C) &= \Pr\left\{\bar{X} \geq c\right\} \qquad (7)\\ &= \Pr\left\{\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \geq \frac{c-\mu_0}{\sigma/\sqrt{n}}\right\}\\ &= \Pr\left\{Z \geq \frac{c-\mu_0}{\sigma/\sqrt{n}}\right\}\\ &= 1 - \Phi\left(\frac{c-\mu_0}{\sigma/\sqrt{n}}\right),\end{aligned}$$

where Z is the N(0,1) random variable and $\Phi(z)$ is its cumulative d.f. So, given the significance level α, it follows immediately from (7) that the corresponding rejection region is $C_{c(\alpha)}$, see (6), with

$$c(\alpha) = \mu_0 + z_\alpha\frac{\sigma}{\sqrt{n}}, \qquad (8)$$

where z_α is the upper tail α-quantile of $N(0, 1)$, i.e., $1 - \Phi(z_\alpha) = \alpha$.

Similarly, the probability of type II error

$$\beta = 1 - \pi(\mu_1, C) = \Pr\left\{Z < \frac{c-\mu_1}{\sigma/\sqrt{n}}\right\} = \Phi\left(\frac{c-\mu_1}{\sigma/\sqrt{n}}\right). \tag{9}$$

So, for a given significance level α, taking into account (8),

$$\beta = \beta(\alpha) = \Phi\left(\frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + z_\alpha\right). \tag{10}$$

The error probabilities α and β, and the relationship between them, depends of course on the sample size n, the location of the threshold point c, and σ. For a given sample $\bar{x} = (x_1, \dots, x_n)$, the corresponding P-value is

$$p = p(x_1, ..., x_n) = \begin{cases} 1 - \Phi((\bar{\boldsymbol{x}} - \mu_0)\sqrt{n}/\sigma) & \text{if } \bar{\boldsymbol{x}} > \mu_0 \\ \Phi((\bar{\boldsymbol{x}} - \mu_0)\sqrt{n}/\sigma) & \text{if } \bar{\boldsymbol{x}} < \mu_0 \end{cases} \tag{11}$$

If, say, $\mu_0 = 0$, $\mu_1 = 1$, $\sigma = 1, n = 16$, then for $c = 1/2$, that is for the critical region $C_{1/2} = \{\bar{x} \geq 1/2\}$

$$\alpha = 1 - \Phi(0.5 \cdot 4) = 0.0227, \quad \beta = \Phi(-0.5 \cdot 4) = 0.0227.$$

However, if $c = 1/10$, that is, the critical region $C = \{\bar{x} \geq 1/10\}$,

$$\alpha = 1 - \Phi(0.1 \cdot 4) = 0.3446, \quad \beta = \Phi(-0.9 \cdot 4) = 0.0002.$$

The power of the test is defined by the type-two error via

$$\pi(\mu_1) = 1 - \Phi((c - \mu_1)\sqrt{n}/\sigma). \tag{12}$$

If the alternative H_1 is no longer simple, then the power function is defined by the formula

$$\pi(\mu) = \Pr\{H_0 \text{ is rejected while the true value is } \mu\} = 1 - \Phi\left(\frac{c-\mu}{\sigma/\sqrt{n}}\right).$$

This formula implies that the power tends to zero (resp., one) if μ tends to $-\infty$ (resp., $+\infty$), and approaches α as μ approaches μ_0.

On the other hand, for a given significance level α, the rejection region is

$$C_c = \{\bar{\boldsymbol{x}} \geq c\}, \quad \text{with } c = \mu_0 + z_\alpha \sigma/\sqrt{n}. \tag{13}$$

The rejection regions for other hypotheses about the mean with the variance σ^2 unknown, and various hypotheses about the variance, are constructed using the Student t-distribution and the chi-square distribution, respectively, just the way we did it for confidence intervals in Section 8.2. All of these rejection regions are summarized in Table 8.3.1.

Mathematica Experiment 1. Testing Hypotheses About the Strength of Cotton Yarn. Hypothesis testing is implemented in the *Mathematica* package `Statistics'HypothesisTests'`. Recall that in the *Mathematica Experiment 8.1.1* we obtained the sample mean for `cotton` data $\bar{x} = 2.2912$ kg, and the sample variance as $s^2 = 0.1802$. We will test the *one-sided hypothesis* $H_0 : \mu_0 = 2.4$ vs. the alternative $H_1 : \mu < 2.4$, as well as the *two-sided hypothesis* $H_0 : \mu_0 = 2.4$ vs. the alternative $H_1 : \mu \neq 2.4$. Both the case of known variance (we take then $\sigma^2 = 0.15$) and the unknown variance are considered. Then we will also experiment with testing the hypothesis about the variance, $H_0 : \sigma^2 = 0.15$ vs. $H_1 : \sigma^2 > 0.15$ The role of P-values, which is the probability of the sample estimate being as extreme as it is, given that the hypothesized population parameter is true, is clearly indicated throughout this experiment.

```
In[1]:= <<Statistics'HypothesisTests'
In[2]:= cotton={ 1.10, . . . ,3.30}
Out[2]= {1.10, . . . ,3.30}
In[3]:= MeanTest[cotton, 2.4]
Out[3]= OneSidedPValue -> 0.0380336
In[4]:= MeanTest[cotton, 2.4, SignificanceLevel -> 0.05]
Out[4]= {OneSidedPValue -> 0.0380336,
          Reject null hypothesis at significance level -> 0.05}


In[5]:= MeanTest[cotton, 2.2, SignificanceLevel -> 0.01]
Out[5]= {OneSidedPValue -> 0.0675763,
            Accept null hypothesis at significance level -> 0.01}
In[6]:= MeanTest[cotton, 2.4, SignificanceLevel -> 0.05,
                                                 TwoSided->True]
Out[6]= {TwoSidedPValue -> 0.0760672,
           Accept null hypothesis at significance level -> 0.05}
In[7]:= MeanTest[cotton, 2.4, SignificanceLevel -> 0.01,
                                                 TwoSided->True]
Out[7]= {TwoSidedPValue -> 0.0760672,
           Accept null hypothesis at significance level -> 0.01}
In[8]:= MeanTest[cotton, 2.4, SignificanceLevel -> 0.01,
        TwoSided->True, FullReport->True,  KnownVariance -> 0.15]
Out[8]= {FullReport -> Mean  TestStat, NormalDistribution,
                       2.2912  -1.98641
         TwoSidedPValue -> 0.0469881,
         Accept null hypothesis at significance level -> 0.01}
```

Table 8.3.1 Rejection regions at significance level α for hypotheses about normal samples of size n

Hypothesis	Other parameter	Rejection region
$H_0 : \mu \geq \mu_0$	σ^2 known	$\bar{x} \leq \mu_0 - z_\alpha \sigma/\sqrt{n}$
	σ^2 unknown	$\bar{x} \leq \mu_0 - t_\alpha(n-1)s/\sqrt{n}$
$H_0 : \mu \leq \mu_0$	σ^2 known	$\bar{x} \geq \mu_0 + z_\alpha \sigma/\sqrt{n}$
	σ^2 unknown	$\bar{x} \geq \mu_0 + t_\alpha(n-1)s/\sqrt{n}$
$H_0 : \mu = \mu_0$	σ^2 known	$\lvert\bar{x} - \mu_0\rvert \geq z_{\alpha/2}\sigma/\sqrt{n}$
	σ^2 unknown	$\lvert\bar{x} - \mu_0\rvert \geq t_{\alpha/2}(n-1)s/\sqrt{n}$
$H_0 : \sigma^2 \geq \sigma_0^2$	μ known	$\sum(x_i - \mu)^2 \leq \chi^2_{1-\alpha}(n)\sigma_0^2$
	μ unknown	$(n-1)s^2 \leq \chi^2_{1-\alpha}(n-1)\sigma_0^2$
$H_0 : \sigma^2 \leq \sigma_0^2$	μ known	$\sum(x_i - \mu)^2 \geq \chi^2_\alpha(n)\sigma_0^2$
	μ unknown	$(n-1)s^2 \geq \chi^2_\alpha(n-1)\sigma_0^2$
$H_0 : \sigma^2 = \sigma_0^2$	μ known	$\frac{\sum(x_i-\mu)^2}{\sigma_0^2} \notin [\chi^2_{1-\alpha/2}(n), \chi^2_{\alpha/2}(n)]$
	μ unknown	$\frac{(n-1)s^2}{\sigma_0^2} \notin [\chi^2_{1-\alpha/2}(n-1), \chi^2_{\alpha/2}(n-1)]$

In Table 8.3.1, z_α is the upper α-tail quantile of the standard normal distribution, $t_\alpha(n)$—upper α-tail quantile of the Student t-distribution with n degrees of freedom, and $\chi_\alpha(n)$—upper α-tail quantile of the χ^2-distribution with n degrees of freedom.

```
In[9]:= VarianceTest[cotton, 0.15, SignificanceLevel->0.05,
                        TwoSided->True, FullReport->True]
Out[9]= {FullReport ->  Variance   TestStat   DF,
                        0.180203   60.0675    49
        ChiSquare Distribution, OneSidedPValue -> 0.133576,
        Accept null hypothesis at significance level -> 0.05}
```

8.4 Statistical inference for two–sample normal models

In this section we consider the issues related to parameter estimation and hypothesis testing for the two-sample normal model, where $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_m)$ are two, independent of each other, random vectors, each with independent, identically distributed components with distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively. As in Section 8.2, for μ_1 and μ_2, respectively, we obtain unbiased estimators

$$\bar{X} = \frac{X_1 + .. + X_n}{n}, \qquad \bar{Y} = \frac{Y_1 + .. + Y_m}{m}. \tag{1}$$

Note that $\bar{X} \sim N(\mu_1, \sigma_1^2/n)$, and $\bar{Y} \sim N(\mu_2, \sigma_2^2/m)$.

The estimators for variances σ_1^2 and σ_2^2, are, respectively,

$$S_1^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu_1)^2, \qquad S_2^{*2} = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \mu_2)^2, \tag{2}$$

when μ_1 and μ_2 are known, and

$$S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \qquad S_2^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y})^2, \tag{3}$$

when μ_1 and μ_2 are unknown. Finally, if $\sigma_1^2 = \sigma_2^2$,

$$S^2 = \frac{1}{n+m-2}\left(\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{j=1}^{m}(Y_j - \bar{Y})^2\right) \tag{4}$$

is the total variance estimator for this two-sample model.

One can prove that

> $nS_1^{*^2}/\sigma_1^2$ has the χ^2 distribution with n degrees of freedom;
>
> $(n-1)S_1^2/\sigma_1^2$ has the χ^2 distribution with $n-1$ degrees of freedom;
>
> $mS_2^{*^2}/\sigma_2^2$ has the χ^2 distribution with m degrees of freedom;
>
> $(m-1)S_2^2/\sigma_2^2$ has the χ^2 distribution with $m-1$ degrees of freedom;
>
> $(m+n-2)S^2/\sigma^2$ has the χ^2 distribution with $n+m-2$ degrees of freedom, if $\sigma_1^2=\sigma_2^2$.

Now, as for the one-sample model discussed in Section 8.2, we can construct confidence intervals for the difference of two means $\mu_1-\mu_2$ and for the ratio of two variances σ_2^2/σ_1^2, and the corresponding rejection regions for tests of equality of two population means and variances, and other hypotheses.

In the case of known variances σ_1^2 and σ_2^2, the $(1-\alpha)$- confidence interval for the difference $\mu_1-\mu_2$ is

$$(\bar{X}-\bar{Y})-z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}\le\mu_1-\mu_2\le(\bar{X}-\bar{Y})+z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}. \quad (5)$$

In the case of unknown variances σ_1^2 and σ_2^2, satisfying condition $\sigma_1^2=\sigma_2^2$, the $(1-\alpha)$-confidence interval for $\mu_1-\mu_2$ has the endpoints

$$(\bar{X}-\bar{Y})\pm t_{\alpha/2}(n+m-2)S\sqrt{\frac{1}{n}+\frac{1}{m}}. \quad (6)$$

In the case of unknown μ_1 and μ_2, the $(1-\alpha)$-confidence interval for the ratio σ_2^2/σ_1^2 is

$$f_{1-\alpha/2}(n-1,m-1)\frac{S_2^2}{S_1^2}\le\frac{\sigma_2^2}{\sigma_1^2}\le f_{\alpha/2}(n-1,m-1)\frac{S_2^2}{S_1^2}, \quad (7)$$

where $f_\alpha(n,m)$ is the upper tail α-quantile of the F-distribution with (n,m) degrees of freedom. Recall that the F-distribution with (n,m) degrees of freedom is defined as the distribution of the ratio

$$F=\frac{X/n}{Y/m} \quad (8)$$

of independent random variables X and Y with χ^2 distributions with, respectively, n and m degrees of freedom.

The related rejection regions for most common hypotheses about the two-sample normal model are summarized in Table 8.4.1.

In order to apply these procedures, use the following checklist:

1. Is the problem a true two-sample problem, that is, are the two samples independent? If not, take the differences of the paired observation $d_i = x_i - y_i$ and proceed for the d_i as in the one-sample model. If the answer is 'yes', move on to the next step.
2. Determine whether the problem is about confidence intervals or hypothesis testing (two- or one-sided tests?). Determine for which parameter is the statistical inference to be done.
3. Determine whether the other parameter is known or not. If σ_1, σ_2 are unknown, is there evidence to support that $\sigma_1^2 = \sigma_2^2$?
4. Apply the appropriate *Mathematica* package.

Mathematica Experiment 1. Testing Pesticide Effectiveness. Armed with more sophisticated estimation and hypothesis tools, we now return (see *Mathematica* Experiment 8.1.3) to the analysis of effectiveness of pesticide applied to two fly populations for $t = 30$ and $t = 60$ seconds, respectively. The original data t30 ($n = 16$) and t60 ($m = 15$) turned out to be not normal, but their logarithms logt30 and logt60 were approximately normal, with sample means $\bar{x} = 1.21919$ and $\bar{y} = 1.17927$, and sample standard deviations $s_1 = 0.456112$ and $s_2 = 0.439349$. We shall start with finding confidence intervals.

The command MeanDifferenceCI[data1,data2, ConfidenceLevel->c] of the Statistics‘ConfidenceIntervals‘ package gives the c confidence interval for the difference between the population mean of data1 and data2 based on the Student t-distribution. The additional options are KnownVariance-> {var1,var2} which returns confidence interval based on the normal distribution, and EqualVariances->True; note that if equal variances cannot be assumed, and $n = m$, the computed interval is larger, or more conservative, and is based on different statistics. The command VarianceRatioCI[data1,data2, ConfidenceLevel->c] gives the c confidence interval for the ratio of the population variance of data1 to population variance of data2 base on the F-ratio distribution.

```
In[1]:= <<Statistics`ConfidenceIntervals`
In[2]:= <<Statistics`HypothesisTests`
In[3]:= logt30={0.477,0.699,0.699 ,0.845,0.954 ,0.954 ,
        1.000 ,1.079 ,1.302 ,1.380,1.380 ,1.532 ,1.634 ,1.663 ,
        1.763, 2.146} ;
In[4]:= logt60={0.301,0.699, 0.699, 0.845, 0.903, 0.954, 1.146,
        1.255,1.380,1.415, 1.415, 1.532, 1.568, 1.623, 1.954};
In[5]:= logt301={0.477,0.699,0.699 ,0.845,0.954 ,0.954 ,
```

Table 8.4.1 Rejection regions at significance level α for the two-sample normal model

Hypothesis	Other parameter	Rejection region
$H_0 : \mu_2 \geq \mu_1$	σ_1^2, σ_2^2 known	$\frac{\bar{x}-\bar{y}}{\sqrt{\sigma_1^2/n+\sigma_2^2/m}} \geq z_\alpha$
	σ_1^2, σ_2^2 unknown, $\sigma_1^2 = \sigma_2^2$	$\frac{\bar{x}-\bar{y}}{s\sqrt{\frac{1}{n}+\frac{1}{m}}} \geq t_\alpha(m+n-2)$
$H_0 : \mu_2 \leq \mu_1$	σ_1^2, σ_2^2 known	$\frac{\bar{x}-\bar{y}}{\sqrt{\sigma_1^2/n+\sigma_2^2/m}} \leq -z_\alpha$
	σ_1^2, σ_2^2 unknown, $\sigma_1^2 = \sigma_2^2$	$\frac{\bar{x}-\bar{y}}{s\sqrt{\frac{1}{n}+\frac{1}{m}}} \leq -t_\alpha(n+m-2)$
$H_0 : \mu_1 = \mu_2$	σ_1^2, σ_2^2 known	$\frac{\lvert\bar{x}-\bar{y}\rvert}{\sqrt{\sigma_1^2/n+\sigma_2^2/m}} \geq z_{\alpha/2}$
	σ_1^2, σ_2^2 unknown, $\sigma_1^2 = \sigma_2^2$	$\frac{\lvert\bar{x}-\bar{y}\rvert}{s\sqrt{\frac{1}{n}+\frac{1}{m}}} \geq t_{\alpha/2}(n+m-2)$
$H_0 : \sigma_1^2 \leq \sigma_2^2$	μ_1, μ_2 unknown	$\frac{s_1^2}{s_2^2} \geq f_\alpha(n-1, m-1)$
$H_0 : \sigma_1^2 \geq \sigma_2^2$	μ_1, μ_2 unknown	$\frac{s_1^2}{s_2^2} \leq f_{1-\alpha}(n-1, m-1)$
$H_0 : \sigma_1^2 = \sigma_2^2$	μ_1, μ_2 unknown	$\frac{s_1^2}{s_2^2} \notin [f_{1-\alpha/2}(n-1, m-1),$ $f_{\alpha/2}(n-1, m-1)]$

```
        1.000 ,1.079 ,1.302 ,1.380,1.380 ,1.532 ,1.634 ,1.663,
        1.763  }
In[6]:= MeanDifferenceCI[logt30 ,logt60, ConfidenceLevel->0.95 ]
Out[6]= {-0.289055, 0.368897}
In[7]:= MeanDifferenceCI[logt30, logt60, ConfidenceLevel->0.95,
                                   EqualVariances->True]
Out[7]= {-0.289454, 0.369296}
In[8]:= MeanDifferenceCI[logt301, logt60, ConfidenceLevel->0.95]
Out[8]= {-0.335126, 0.291393}
```

```
In[9]:=  MeanDifferenceCI[logt301, logt60, ConfidenceLevel->0.95,
                                          EqualVariances->True]
Out[10]= {-0.33498, 0.291247}
In[11]:= VarianceRatioCI[logt30, logt60, ConfidenceLevel->0.95]
Out[11]= {0.365428, 3.11633}
In[12]:= VarianceRatioCI[logt30, logt60, ConfidenceLevel->0.80]
Out[12]= {0.536325, 2.13971}
```

The command `MeanDifferenceTest[data1, data2, diff, EqualVariances->True]` of the `Statistics'HypothesisTests'` package gives the P-value for the test that the difference in population means is `diff` based on the Student t-distribution. The same command, but without the `EqualVariances->True` option, returns the P-value for the so-called Welch's approximate t-test (with a special formula to calculate the number of degrees of freedom) that the difference in population means if `diff`. This case was not discussed in this section. Other options are as in the previously discussed commands. The null hypotheses tested are $H_0 : \mu_0 - \mu_1 = 0$, concerning the means and, then, $H_0 : \sigma_1^2/\sigma_2^2 = 1$, concerning the variances. In the `VarianceRatioTest` command, the `FullReport->True` option also lists explicitly the numerator's and denominator's numbers of degrees of freedom.

```
In[13]:= MeanDifferenceTest[logt30, logt60, 0,
      SignificanceLevel-> 0.05, TwoSided-> True, FullReport-> True]
Out[13]= {FullReport -> MeanDiff    TestStat   DF,
                         0.0399208   0.248195   28.9751
         StudentTDistribution, TwoSidedPValue -> 0.805734,
         Accept null hypothesis at significance level -> 0.05}
In[14]:= MeanDifferenceTest[logt30, logt60, 0,
         SignificanceLevel -> 0.05, TwoSided -> True,
         EqualVariances -> True, FullReport -> True]
Out[14]= {FullReport ->   MeanDiff    TestStat   DF,
                           0.0399208   0.247886   29
         StudentTDistribution,  TwoSidedPValue -> 0.80597,
         Accept null hypothesis at significance level -> 0.05}
In[15]:= VarianceRatioTest[logt30, logt60, 1,
      SignificanceLevel-> 0.05, TwoSided-> True, FullReport-> True]
Out[15]= {FullReport -> Ratio     TestStat   NumDF   DenDF,
                         1.07776   1.07776    15      14
         FRatio Distribution, TwoSidedPValue -> 0.893489,
         Accept null hypothesis at significance  level -> 0.05}
```

8.5 Regression analysis for the normal model

The linear least-squares fit of paired data

$$(\boldsymbol{x}, \boldsymbol{y})^T = ((x_1, y_1), \dots, (x_n, y_n)) \tag{1}$$

was discussed in a preliminary way in Section 2.7, where we found the following nondimensionalized form of the regression line

$$\frac{y - \bar{\boldsymbol{y}}}{\operatorname{std} \boldsymbol{y}} = \operatorname{corr}(\boldsymbol{x}, \boldsymbol{y}) \frac{x - \bar{\boldsymbol{x}}}{\operatorname{std} \boldsymbol{y}} \tag{2}$$

where

$$\operatorname{std} \boldsymbol{y} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{\boldsymbol{y}})^2}, \qquad \operatorname{std} \boldsymbol{x} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{\boldsymbol{x}})^2}, \tag{3}$$

and

$$\operatorname{corr}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\operatorname{cov}(\boldsymbol{x}, \boldsymbol{y})}{\operatorname{std}(\boldsymbol{x})\operatorname{std}(\boldsymbol{y})}, \quad \operatorname{cov}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{\boldsymbol{x}})(y_i - \bar{\boldsymbol{y}}). \tag{4}$$

We returned to these issues in Section 7.3.2 while discussing the Least Squares Estimators.

In this section we undertake a more sophisticated statistical analysis of this model. More precisely, the basic assumption here is that the *response random variables* $Y_1, \dots, Y_n$, have the representation

$$Y_i = a + bx_i + \epsilon_i, \qquad i = 1, \dots, n, \tag{5}$$

where $\epsilon_1, \dots, \epsilon_n$ are normal, independent, zero-mean random variables with common variance $\sigma^2 > 0$. The quantities a, b are real numbers and $x_1, \dots, x_n$, are 'manipulated' variables, chosen by the experimenter.

More generally, to permit several measurements of the response variable y for the same value of the manipulated variable x, we will consider the model

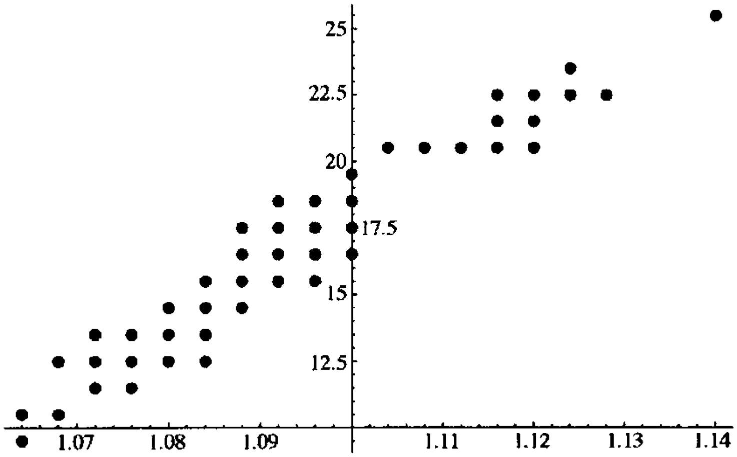
$$Y_{ji} = a + bx_j + \epsilon_{ji}, \qquad j = 1, \dots, n, \quad i = 1, \dots, n_j, \tag{6}$$

where ϵ_{ji} are normal, independent, zero-mean random variables with common variance $\sigma^2 > 0$. Thus, for each value x_j of the manipulated variable, we have n_j values of the response variable Y.

Example 8.5.1 Starch Content of Potatoes.

The starch content y (measured as percent of total weight) in potatoes is not easy to measure directly as it requires a costly chemical analysis; the specific gravity x is a parameter that is much easier to measure. Is there a linear relationship between the two parameters?

A sample of 409 potatoes was taken and Table 8.5.1 shows the number $n_{i,j}$ of times when a specific pair (x_i, y_j) was observed. The raw paired data (x_j, y_{ji}), $j = 1, \ldots, 20$, $i = 1, \ldots, n_j$, is contained in the file `POTATOES` on the `UVW Web Site`. The scatter plot of these data, produced below, is obviously inadequate: the information about frequencies of paired data is lost there.



Let us return to the general statistical analysis for the linear regression model (6). As before, the Least Squares Estimators $\hat{a}$, $\hat{b}$ for the coefficients a and b, are

$$\hat{b} = \frac{\sum_{j=1}^{n}\sum_{i=1}^{n_j}(Y_{ji} - \bar{Y})(x_j - \bar{x})}{\sum_{j=1}^{n}\sum_{i=1}^{n_j}(x_j - \bar{x})^2}, \tag{7}$$

$$\hat{a} = \bar{Y} - \hat{b}\bar{x}, \tag{8}$$

where

$$\bar{x} = \frac{1}{N}\sum_{j=1}^{n} n_j x_j, \qquad \bar{Y} = \frac{1}{N}\sum_{j=1}^{n}\sum_{i=1}^{n_j} Y_{ji}, \tag{9}$$

Table 8.5.1 Starch content and specific gravity of 409 potatoes

Starch content	Specific gravity x (in g/cm^3)					
y (in %)	1.064	1.068	1.072	1.076	1.080	1.084
9.5	1					
10.5	1	1				
11.5			5	1		
12.5		1	5	2	3	1
13.5			1	2	1	9
14.5					6	9
15.5						3

	1.088	1.092	1.096	1.100	1.104	1.108	1.112
14.5	11						
15.5	19	18	6				
16.5	11	30	43	10			
17.5	2	11	33	54			
18.5		2	4	39			
19.5				2			
20.5					1	6	22

	1.116	1.120	1.124	1.128	1.132	1.136	1.140
21.5	11	3					
22.5	1	1	4	1			
23.5			1				
24.5							
25.5							1

and

$$N = \sum_{j=1}^{n} n_j \tag{10}$$

is the total number of measurements of the response variable. Denoting

$$S_{xx} = \sum_{j=1}^{n} n_j (x_j - \bar{x})^2,$$

$$S_{YY} = \sum_{j=1}^{n}\sum_{i=1}^{n_j}(Y_{ji} - \bar{Y})^2,$$

$$S_{xY} = \sum_{j=1}^{n}\sum_{i=1}^{n_j}(Y_{ji} - \bar{Y})(x_j - \bar{x}),$$

the formula (7) for the Least Squares Estimator $\hat{b}$ can be rewritten in the form

$$\hat{b} = \frac{S_{xY}}{S_{xx}}. \tag{11}$$

Given paired data (x_j, y_{ji}), the quantity S_{xx} is called the *corrected sum of squares for x*, S_{yy} is called the *corrected sum of squares for y*, and S_{xy} is called the *corrected sum of cross-products for x and y*. An alternative formula for S_{xy} is

$$S_{xy} = \sum_{j=1}^{n} n_j(\bar{y}_{j\cdot} - \bar{y})(x_j - \bar{x}), \tag{12}$$

with

$$\bar{y}_{j\cdot} = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ji}.$$

Note that in this notation a dot replaces the index over which the averaging has been done.

Under the normality assumptions on ϵ_{ji}s in model (6), for each given x, the S_{xY}, and thus $\hat{b}$ and $\hat{a}$, are random quantities with normal distributions, and S_{YY}, properly normalized, has a chi-square distribution. Using calculations similar to those employed in Sections 8.2-4, one can construct confidence intervals for, and test hypotheses about, parameters $a = E(\hat{b}), b = E(\hat{a})$, the response variable $y = a + bx = E(\hat{a} + \hat{b}x)$ (they are all unbiased estimators), and on future observations $Y^* = a + bx + \epsilon$. The results are summarized in Table 8.5.2 (see also Bibliographical Notes), which is followed by explanations of the notation used and commentaries.

As before, $t_\alpha(N - 2)$ denotes the upper tail α–quantile of the Student t-distribution with $N-2$ degrees of freedom. The unbiased estimator for the variance σ^2,

$$\hat{\sigma}^2 = \frac{SS_E}{N-2} \tag{13}$$

Table 8.5.2 Linear Regression Model $Y_{ji} = a + bx_j + \epsilon_{ji}$

LSE estimator of b	$\hat{b} = S_{xY}/S_{xx}$
LSE estimator of a	$\hat{a} = \bar{Y} - \hat{b}\bar{x}$
LSE estimator of $y = a + bx$	$\hat{y} = \hat{a} + \hat{b}x$
α-confidence interval for b	$\hat{b} - t_{\alpha/2}(N-2)\hat{\sigma}\frac{1}{\sqrt{S_{xx}}} \le b \le \hat{b} + t_{\alpha/2}(N-2)\hat{\sigma}\frac{1}{\sqrt{S_{xx}}}$
α-confidence interval for $y = a + bx$	$\hat{y} - t_{\alpha/2}(N-2)\hat{\sigma}\sqrt{\frac{1}{N} + \frac{(x-\bar{x})^2}{S_{xx}}} \le y \le \hat{y} + t_{\alpha/2}(N-2)\hat{\sigma}\sqrt{\frac{1}{N} + \frac{(x-\bar{x})^2}{S_{xx}}}$
α-prediction interval for $y^* = a + bx + \epsilon$	$\hat{y} - t_{\alpha/2}(N-2)\hat{\sigma}\sqrt{1 + \frac{1}{N} + \frac{(x-\bar{x})^2}{S_{xx}}} \le y^* \le \bar{y} + t_{\alpha/2}(N-2)\hat{\sigma}\sqrt{1 + \frac{1}{N} + \frac{(x-\bar{x})^2}{S_{xx}}}$

depends on the *error sum of squares*

$$SS_E = \sum_{j=1}^{n}\sum_{i=1}^{n_j}(Y_{ji} - (\hat{a} + \hat{b}x_j))^2 = \sum_{j=1}^{n}\sum_{i=1}^{n_j}(Y_{ji} - \hat{y}_j)^2 \tag{14}$$

which measures the fluctuations (variability) of the observations y_{ji} not attributable to the variability of the regression line itself. Indeed, by simple algebra, the total corrected sum of squares

$$S_{YY} = SS_E + SS_R, \tag{15}$$

where

$$SS_R = \sum_{j=1}^{n}\sum_{i=1}^{n_j}(\bar{Y} - (\bar{a} + \bar{b}x_j))^2 = \sum_{j=1}^{n}\sum_{i=1}^{n_j}(\bar{Y} - \hat{y}_j)^2, \tag{16}$$

is called the *regression sum of squares* and measures variability in y_is accounted for by the regression line.

The decomposition (15) of the total variability in response variable Y leads to a simple version of the *analysis of variance* which can be used to test for significance of the regression model. Intuitively speaking, if the regression sum of squares SS_R contributes the lion's share of the total variability S_{YY}, as compared to the error sum of squares SS_E, then we would be inclined to say that the linear regression model is valid. This line of thinking can be quantified if one recognizes that the random variables SS_R/σ^2 and SS_E/σ^2 have chi-square distributions with, respectively, 1 and $(N-2)$ degrees of freedom. Thus, the statistic

$$F_0 = \frac{SS_R/1}{SS_E/(N-2)} \equiv \frac{MS_R}{MS_E} \tag{17}$$

has the $F(1, N-2)$-distribution. We will reject the hypothesis $H_0 : b = 0$ (that the linear regression model is inappropriate) at the significance level α, if $F_0 > f_\alpha(1, N-2)$. The above procedure is usually summarized in the *analysis of variance (ANOVA) table*. Other information usually included in the regression analysis is the *estimated standard errors* for parameters a and b:

$$\text{se}\,(\hat{b}) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}, \qquad \text{and} \qquad \text{se}\,(\hat{a}) = \sqrt{\hat{\sigma}^2\left(\frac{1}{N} + \frac{\bar{x}^2}{S_{xx}}\right)} \tag{18}$$

which reflect the fact that the variances of the unbiased estimators $\hat{b}$ and $\hat{a}$, are

$$\text{Var}\,(\hat{b}) = \frac{\sigma^2}{S_{xx}}, \qquad \text{and} \qquad \text{Var}\,(\hat{a}) = \sigma^2\left(\frac{1}{N} + \frac{\bar{x}^2}{S_{xx}}\right). \tag{19}$$

Another useful parameter is the *coefficient of determination*

$$R^2 = \frac{SS_R}{SS_Y} = 1 - \frac{SS_E}{SS_Y}, \tag{20}$$

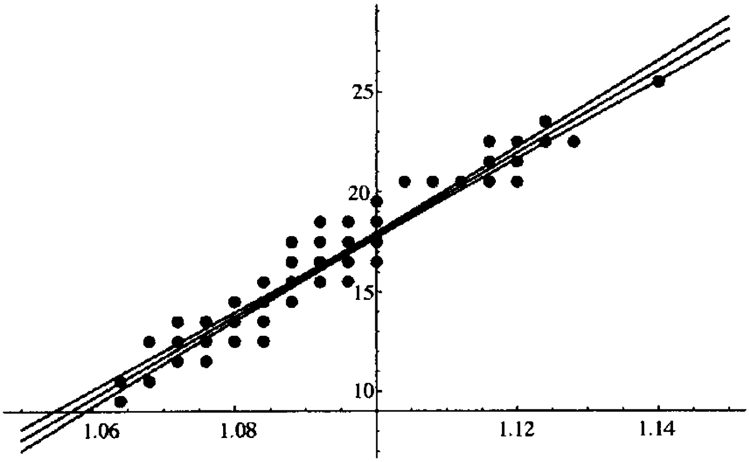
which is the square of the sample correlation coefficient. Obviously, $0 \le R^2 \le 1$, and the closer R^2 is to 1 the bigger percentage of the total variability in the data can be attributed to the regression model.

Mathematica Experiment 1. Starch Content of Potatoes. In this experiment we will do regression analysis on the data from Table 8.5.1 (and file `POTATOES` located on the `UVW Web Site`) using the *Mathematica* package `Statistics'LinearRegression'`. Most of the commands are, by now, self-explanatory.

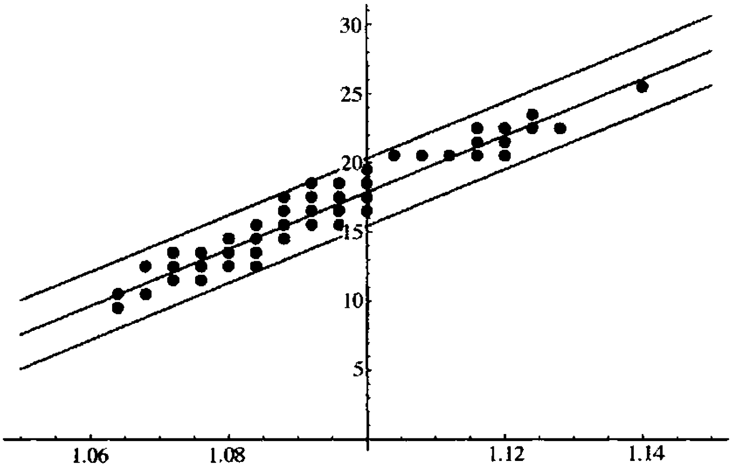
The command Chop replaces the P-values below 10^{-6} with 0. The confidence and prediction intervals are computed at the confidence level 0.999.

```
In[1]:= <<Statistics'ContinuousDistributions'
In[2]:= <<Statistics'LinearRegression'
In[3]:= potatoes={
        {1.064,9.5},{1.064,10.5},
        {1.068,10.5}, {1.068,12.5},
        {1.072, 11.5},{1.072, 11.5},{1.072, 11.5},{1.072, 11.5},
        . . . . . . . . . . . . . . .
        {1.124, 22.5},{1.124, 22.5},{1.124, 22.5},{1.124, 22.5},
        {1.124, 23.5},
        {1.128, 22.5},{1.140, 25.5}
        };
In[4]:= L=Length[potatoes]
Out[4]= 409
In[5]:= PotatoPlot=ListPlot[potatoes, PlotStyle->PointSize[0.015],
        PlotRange->All];
In[6]:=  reg= Regress[potatoes, {1,x},x]; Chop[reg, 10^(-6)]
Out[6]= {ParameterTable -> Estimate    SE         TStat      PValue,
                    1        -208.146   3.67963   -56.5671   0
                    x        205.466    3.35632   61.2177    0
       RSquared -> 0.902036, AdjustedRSquared -> 0.901796,
       EstimatedVariance -> 0.538207,
       ANOVATable ->
            DoF   SoS        MeanSS     FRatio     PValue}
      Model    1   2016.99   2016.99    3747.61   0
      Error 407  219.05    0.538207
      Total 408  2236.04
In[7]:= xList=Part[Transpose[potatoes],1];
In[8]:= xMean=(1/409)Sum[xlist[[i]],{i,1,409}]
Out[8]= 1.09627
In[9]:= Sxx=Sum[(xList[[i]]-xMean)^2, {i,1,L}]
Out[9]= 0.0477773
In[10]:= RegressionLineAndConfidenceBand=
        Plot[
        {-208.146+205.446x,
        -208.146+205.446x+
        Quantile[StudentTDistribution[L-2],.9995]*
        Sqrt[0.538207*((1/L)+(x-xMean)^2/Sxx)],
       -208.146+205.446x-
        Quantile[StudentTDistribution[L-2],.9995]*
        Sqrt[0.538207*((1/L)+(x-xMean)^2/Sxx)]
       },  {x,1.05,1.15}];
```

```
In[11]:= Show[RegressionLineAndConfidenceBand, PotatoPlot]
Out[12]:= -Graphics-
```



```
In[13]:= RegressionLineAndPredictionBand=
        Plot[
        {-208.146+205.446x,
        -208.146+205.446x+
        Quantile[StudentTDistribution[L-2],.9995]*
        Sqrt[0.538207*(1+(1/L)+(x-xMean)^2/Sxx)],
       -208.146+205.446x-
        Quantile[StudentTDistribution[L-2],.9995]*
        Sqrt[0.538207*(1+(1/L)+(x-xMean)^2/Sxx)],
       }, {x,1.05,1.15}];
In[14]:= Show[RegressionLineAndPredictionBand, PotatoPlot]
Out[14]= -Graphics-
```



A more detailed discussion of the analysis of variance techniques can be found in Chapter 9.

8.6 Testing for goodness-of-fit

Testing normality of experimental data is of fundamental importance for implementation of hypothesis testing techniques developed in this chapter. For that purpose, and for testing goodness-of-fit of any data to any particular distribution, we have used the Q-Q plots and the Kolmogorov-Smirnov statistics. In this section we discuss another way to test the hypothesis H_0: the random sample $X = (X_1, \dots, X_n)$ comes from the population with the cumulative d.f. $F(x)$. The procedure is as follows:

(1) The range of the random variable X is divided into a finite number k of disjoint and exhaustive bins $B_1, B_2, \dots, B_k$ and the frequencies (the histogram)

$$\phi_X(B_i) = \sum_{j=1}^{n} 1_{B_i}(X_j) = \#\{j : X_j \in B_i\}, \qquad i = 1, \dots, k, \tag{1}$$

are calculated for all the bins.

(2) Expected probabilities (expected values of the indicator functions of the bins)

$$p_i = \Pr\{X \in B_i\} = \int_{B_i} dF(x) = E1(B_i), \tag{2}$$

are calculated.

(3) The test statistic,

$$X_0^2 = \sum_{i=1}^{k} \frac{(\phi_X(B_i) - np_i)^2}{np_i}, \tag{3}$$

is computed. It has an approximate chi-square distribution with $k - 1$ degrees of freedom if the distribution F is completely specified, and $k - 1 - h$ degrees of freedom if h parameters of F have to be estimated using the random sample $\boldsymbol{X}$.

(4) The hypothesis H_0 is rejected at the significance level α if

$$X_0^2 \geq \chi_\alpha^2(k - 1 - h), \tag{4}$$

where, as before, $\chi_\alpha^2(k - 1 - h)$ is the upper tail α-quantile of the chi-square distribution with $k - 1 - h$ degrees of freedom.

Mathematica Experiment 1. Web Page Hits. A number of hits of the student's Web Page was watched in each 24-hour period over $n = 30$ days. The recorded frequencies were as follows: 0 hits were recorded on 8 days, 1 hit on 11 days, 2 hits on 6 days, 3 hits in 3 days, 4 hits on 1 day, and 6 hits on 1 day. We will test the goodness-of-fit of the Poisson distribution with one ($h = 1$) parameter μ to be estimated from the sample. The $k = 6$ bins were selected as follows:

$$B_1 = \{0\},\ B_2 = \{1\},\ B_3 = \{2\},\ B_4 = \{3\},\ B_5 = \{4\},\ B_6 = \{5, 6, \ldots\}.$$

The corresponding frequencies were

$$\phi(B_1) = 8,\ \phi(B_2) = 11,\ \phi(B_3) = 6,\ \phi(B_4) = 3,\ \phi(B_5) = 1,\ \phi(B_6) = 1.$$

with the estimated mean value

$$\mu = \frac{1}{30}\Big(0 \cdot 8 + 1 \cdot 11 + 2 \cdot 6 + 3 \cdot 3 + 1 \cdot 4 + 1 \cdot 6\Big) = 1.4.$$

Thus, remembering the Poisson distribution $\Pr\{X = m\} = e^{-\mu}\mu^m/m!, m = 0, 1, 2, \ldots$, the respective probabilities p_i from (2) are as follows:

$$p_1 = \Pr\{X = 0\} = 0.247, \quad p_2 = \Pr\{X = 1\} = 0.345,$$

$$p_3 = \Pr\{X = 2\} = 0.241, \quad p_4 = \Pr\{X = 3\} = 0.113,$$

$$p_5 = \Pr\{X = 4\} = 0.040, \quad p_6 = \Pr\{X = 5, 6, \ldots\} = 0.014.$$

With this information the test statistic from (3) is $x_0^2 = 1.2601$ with the number of degrees of freedom $k - 1 - h = 4$. At the $\alpha = .05$ significance level

$$x_0^2 = 1.2601 < \chi^2_{0.05}(4) = 9.488,$$

and the hypothesis that the data come from the Poisson distribution with $\mu = 1.4$ cannot be rejected. Actually, the P-value for our data is very high, a strong indication that the hypothesis is true:

```
In[1]:= Statistics'ContinuousDistributions'
In[2]:= PValue=1-CDF[ChiSquareDistribution[4],1.260]
Out[2]= 0.868125
```

Mathematica Experiment 2. Resistors. The data set RESISTORS was supplied by Jacob K. Matthews, a Chemical Engineering major at Case Western Reserve University. It represents a listing of the resistances (in ohms) of $n = 200$ resistors, which are all rated at 10 kΩ. We will test the goodness-of-fit of the normal distribution $N(\mu, \sigma^2)$, with both ($h = 2$) parameters μ and σ^2 estimated from the sample. The bins B_i were selected to correspond to deciles of the normal distribution.

```
In[1]:= <<Statistics'DataManipulation'
In[2]:= <<Statistics'ContinuousDistributions'
In[3]:= <<Statistics'DescriptiveStatistics'
In[4]:= resistors={9.97910927, 9.833997401, 10.48797923, . . . ,
   10.03217857, 10.19504918, 10.23059564};
In[5]:= Length[resistors]
Out[5]= 200
In[6]:= mu=Mean[resistors]
Out[6]= 9.87989
In[7]:= sigma =StandardDeviation[resistors]
Out[7]= 0.480209
In[8]:= bins=Table[Quantile[NormalDistribution[mu,sigma ],
                            i /10], {i, 9}]
Out[8]= {9.26448, 9.47574, 9.62807, 9.75823, 9.87989,
          10.0015,  10.1317, 10.284, 10.4953}
In[9]:= phi=RangeCounts[resistors,bins]
Out[9]=  {5, 7, 16, 35, 30, 32, 34, 28, 12, 1}
In[10]:= xSquared = Sum[(phi[[i]]-20)^2/20, {i,1,10}]
Out[10]=  78.2
In[11]:= PValue=1-CDF[ChiSquareDistribution[7],78.2]
Out[11]=    3.20539 . 10^(-14)
```

Thus, the hypothesis that the data RESISTOR have the $N(9.88, (0.48)^2)$ distribution has to be rejected at any reasonable significance level.

8.7 Experiments, exercises, and projects

1. Test the assumption of normality for the COTTON data from *Mathematica* Experiment 8.1.1 using the Kolmogorov-Smirnov test and the goodness-of-fit test of Section 8.6. Construct 0.95-confidence intervals for the variance of the population.

2. Test the assumption of normality for the RIVET data from *Mathematica* Experiment 8.1.2 using the Kolmogorov-Smirnov test and the goodness-of-fit test of Section 8.6. Construct 0.95-confidence intervals for the mean and variance of the population. Find the P-values for the sample mean and sample variance.

3. Test the assumption of normality for the logarithms of the pesticide data from *Mathematica* Experiment 8.1.3 using the Kolmogorov-Smirnov test and the goodness of fit test of Section 8.6. Construct 0.95-confidence intervals for the means and variances of the two populations. Find the P-values for the sample means and sample variances.

4. Using the data from Example 8.1.1, test the hypothesis that the tire wear is the same for two types of tires tested. Do it at 0.1, 0.05, and 0.01 significance levels.

5. *Theoretical Project.* On the basis of the Central Limit Theorem, derive the large sample confidence intervals for parameter p in the Bernoulli distribution. Start with the discussion in Section 3.7. Then, analyze and complete the following reasoning:

   Assume that in n independent trials, k successes were observed. By the CLT, the random quantity

   $$Z = \frac{k/n - p}{\sqrt{p(1-p)/n}}$$

   is asymptotically standard normal. Thus, for a large sample size n,

   $$\Pr\left(\frac{|k/n - p|}{\sqrt{p(1-p)/n}} < z_{\alpha/2}\right) \approx 1 - \alpha.$$

   The inequality inside the parentheses can be rewritten in the form

   $$\left(\frac{k}{n}\right)^2 - 2p\frac{k}{n} + p^2 < p\frac{z_{\alpha/2}^2}{n} - p^2\frac{z_{\alpha/2}^2}{n}.$$

This inequality is quadratic in p and can be solved explicitly, showing that p has to be contained in the interval with the end points

$$\frac{1}{1+z_{\alpha/2}^2/n}\left(\frac{k}{n}+\frac{z_{\alpha/2}^2}{2n}\pm z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}+\frac{z_{\alpha/2}^2}{4n^2}}\right), \qquad (1)$$

which give the desired $1-\alpha$ confidence limits. Taking n large, derive the familiar approximate confidence limits

$$\frac{k}{n}\pm z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}}. \qquad (2)$$

Test how good these approximations are using *Mathematica*. For p which is neither close to 0, nor close to 1, a sample size of 25 should be adequate for (1), but one needs at least $n = 100$ to get good results with the cruder (2).

6. A machine produces lids for jars. In order to obtain a proper fit for jars with a diameter of 3 inches, the deviation σ should not exceed .05 inches. A random sample of 12 lids showed the following diameters: 3.10, 3.05, 3.02, 2.98, 2.97, 3.02, 3.03, 2.98, 2.95, 2.92, 3.07, 3.04. Obtain the 95% confidence interval for σ using *Mathematica*.

7. Obtain 0.95-level confidence bands and prediction bands for the regression problem considered in Mathematica Experiment 8.1.6. Test the lumberjack's hypothesis.

8. Test the assumption of normality for the `POTATOES` regression problem from Table 8.5.1. Find the residuals first.

9. Using *Mathematica* devise an experimental method of verifying the statements concerning distributions of the sample variances, contained in bullets in Section 8.4.

10. Using *Mathematica* produce the table of maximal differences between the values of the $N(0, 1)$ density and the Student T-densities, for the number of degrees of freedom equal to $1, \ldots, 20$.

11. Repeat the *Mathematica* Experiment 8.1.3 on data `t60`.

12. Construct confidence intervals for the variance σ^2 of the $N(\mu, \sigma^2)$, based on a random sample of size n and assuming that the mean μ is known.

13. Use *Mathematica* pseudo-random number generator to create "virtual" random data for the regression analysis under the *normal* model. Start by creating tables of random residuals $e_1, \ldots, e_n$, with various normal zero-mean distributions. Then create data sets by preselecting values of

$x_1, \ldots, x_n$, and considering as response variables $y_i = f(x_i) + e_i$ where $f(x)$ are selected from the following list:

$$f(x) = 2 + 3x$$

$$f(x) = 2 - 3x$$

$$f(x) = -2 + 3x$$

$$f(x) = -2 - 3x$$

$$f(x) = 2 + 3x^2$$

$$f(x) = 2 - 3x^2$$

$$f(x) = 3\exp x$$

$$f(x) = 3\ln x \quad (x > 0)$$

Draw the scatterplots and calculated regression curves, together with confidence and prediction bands, at different significance levels. Repeat the experiment for the model $y_{ji} = f(x_j) + e_{ji}$.

14. Test the *Mathematica* pseudo-random number generator using the goodness-of-fit test. Do not stop at checking the uniform distribution of the digits. Test the uniform distribution of pairs, triples, etc. Repeat the experiment for various pseudo-random number generators discussed earlier in this book, including the decimal expansions of numbers π and e .

15. Test the lumberjack's hypothesis from *Mathematica* Experiment 8.1.6, using the full power of Section 8.5. Do not forget to check the normality of residuals.

8.8 Bibliographical notes

Two classics on parameter estimation and hypotheses testing are

[1] E.H. Lehman, *Theory of Point Estimation*, John Wiley & Sons, New York, 1981,

[2] E.H. Lehman, *Testing Statistical Hypotheses*, John Wiley & Sons, New York, 1959.

The multivariate issues are addressed in

[3] T.W. Anderson, *An Introduction to Mutivariate Statistical Analysis*, John Wiley & Sons, New York, 1984.

A medium level, but fairly comprehensive, general text on the subject is

[4] D.C. Montgomery and G.C. Runger, *Applied Statistics and Probability for Engineers*, John Wiley & Sons, New York, 1994.

On the subject of linear regression, we would like to mention

[5] W. Mendenhall and T. Sincich, *A Second Course in Statistics: Regression Analysis*, Prentice Hall, Englewood Cliffs, N.J., 1996,

practice oriented

[6] T. Ryan, *Modern Regression Methods*, John Wiley & Sons, New York, 1997,

and, on a more advanced level,

[7] S.R. Searle, *Linear Models*, John Wiley & Sons, New York, 1970.

Chapter 9

Analysis of Variance

Analysis of Variance (ANOVA) is used to test whether the variability in response data taken for different levels of manipulated variables can be attributed just to random fluctuations, or is caused by the impact of different input levels. Such an approach has been briefly discussed in Section 8.5. A more general case, with several manipulated categorical variables (factors), will be sketched in this chapter. It is one of the basic tools in the design and analysis of experiments.

9.1 Single-factor ANOVA

In this model we consider the model of k independent random vectors

$$\boldsymbol{X}_1 = (X_{11}, \ldots, X_{1n_1})$$

$$\boldsymbol{X}_2 = (X_{21}, \ldots, X_{2n_2})$$

$$\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\boldsymbol{X}_k = (X_{k1}, \ldots, X_{kn_k})$$

of independent, normally distributed random variables $X_{l,j}$ with means μ_l, $l = 1, \ldots, k$, and common variance $\sigma^2 > 0$. The goal is to develop procedures for testing the hypothesis

$$H_0 : \mu_1 = \mu_2 = .. = \mu_k, \tag{1}$$

based on the realization $\boldsymbol{x}_1, \ldots \boldsymbol{x}_k$, of the above random vectors obtained from an experiment.

M. Denker, W. Woyczynski, *Introductory Statistics and Random Phenomena*,
Modern Birkhäuser Classics, DOI 10.1007/978-3-319-66152-0_9

Example 9.1.1 Fertilizer Yields.

The yield of a crop depends on the type of fertilizer that has been applied. To test effects of fertilizers A, B, C, and D, a completely randomized design has been used to ensure that there was no systematic bias (in terms of soil type, drainage, exposure to sun, etc.) in selection of plots for application of different fertilizers. Fertilizer A has been applied on 8 plots, B—on 7, C—on 6, and D—on 9. The results of the experiment are summarized in Table 9.1.1.

Table 9.1.1 Fertilizers' effects on crop yield

Fertilizer	Yield in bushels
A	151, 158, 162, 149, 153, 151, 150, 159
B	142, 143, 142, 145, 147, 150, 148
C	147, 142, 143, 146, 144, 142
D	137, 139, 141, 138, 139, 137, 142, 136, 140

This is a typical *single-factor* (or, one-way) experiment, the factor being the type of fertilizer *treatment* applied to the crop. The null hypothesis $H_0 : \mu_A = \mu_B = \mu_C = \mu_D$ asserts that there is no difference in effects of different fertilizer treatments on the distribution (mean) of crop yields. The alternative hypothesis H_1 is that two or more treatments have (significantly) different effects.

Recall that for the two-sample case ($k = 2$) discussed in Section 8.4, the rejection region at significance level 2α is given by the condition

$$\frac{|\bar{x}_{1\cdot} - \bar{x}_{2\cdot}|}{s\sqrt{1/n_1 + 1/n_2}} \geq t_\alpha(n_1 + n_2 - 2), \tag{2}$$

where

$$s^2 = \frac{1}{n_1 + n_2 - 2}\left(\sum_{j=1}^{n_1}(x_{1j} - \bar{x}_{1\cdot})^2 + \sum_{l=1}^{n_2}(x_{2l} - \bar{x}_{2\cdot})^2\right), \tag{3}$$

$$\bar{x}_{1\cdot} = \frac{1}{n_1}\sum_{j=1}^{n_1} x_{1j}, \quad \text{and} \quad \bar{x}_{2\cdot} = \frac{1}{n_2}\sum_{l=1}^{n_2} x_{2l}, \tag{4}$$

and where $t_\alpha(n_1 + n_2 - 2)$ is the upper tail α-quantile of the Student t-distribution with $n_1 + n_2 - 2$ degrees of freedom. Since, by simple algebra,

$$\sum_{j=1}^{2} n_j(\bar{x}_{j\cdot} - \bar{x}_{\cdot\cdot})^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{x}_{1\cdot} - \bar{x}_{2\cdot})^2, \tag{5}$$

where

$$\bar{x}_{\cdot\cdot} = \frac{1}{n_1 + n_2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} x_{ij}, \tag{6}$$

the rejection region (2) can be rewritten in the form

$$\frac{n_1(\bar{x}_{1\cdot} - \bar{x}_{\cdot\cdot})^2 + n_2(\bar{x}_{2\cdot} - \bar{x}_{\cdot\cdot})^2}{s^2} \geq t_\alpha^2(n_1 + n_2 - 2). \tag{7}$$

Since $t_\alpha^2(n_1 + n_2 - 2)$ also happens to be an upper tail α-quantile for the F-distribution with $(1, n_1 + n_2 - 2)$ degrees of freedom, which is the distribution of

$$\frac{n_1(\bar{X}_{1\cdot} - \bar{X}_{\cdot\cdot})^2 + n_2(\bar{X}_{2\cdot} - \bar{X}_{\cdot\cdot})^2}{S^2}, \tag{8}$$

the following extension of the above test to the k sample design is natural.

Consider the total sum of squared deviations from the grand mean $\bar{X}_{\cdot\cdot}$:

$$TSS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{\cdot\cdot})^2, \tag{9}$$

where $N = n_1 + \ldots + n_k$, and

$$\bar{X}_{\cdot\cdot} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}, \tag{10}$$

is the grand mean taken over all samples. Again, by simple algebra,

$$TSS = SS_T + SS_E, \tag{11}$$

where

$$SS_T = \sum_{i=1}^{k} n_i(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2 \tag{12}$$

is the *treatment sum of squares* and

$$SS_E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i\cdot})^2, \tag{13}$$

is the *error sum of squares*. In other words, the total variability in the pooled samples has been split into the variability due to the effects of treatments and random fluctuations due to errors.

Decomposition (11) is a foundation of the ANOVA method. The null hypothesis (1) will be rejected if the variability SS_T due to treatments is large relative to the variability SS_E due to errors. As always, to construct rejection regions at specific significance levels we need information about the distributions of these two estimators.

It can be demonstrated that the *mean treatment sum of squares*

$$MS_T = \frac{SS_T}{k-1} = \frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2, \tag{14}$$

is an unbiased estimator for σ^2, and SS_{T/σ^2} has a chi-square distribution with $k-1$ degrees of freedom, when the hypothesis H_0 is true. It is biased under H_1, in which case it overestimates σ^2 as $E(MSS_T) > \sigma^2$.

On the other hand, the *mean error sum of squares*

$$MS_E = \frac{SS_E}{N-k} = \frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\cdot})^2, \tag{15}$$

is always an unbiased estimator for σ^2, and SS_{E/σ^2} has a chi-square distribution with $N-k$ degrees of freedom, when the hypothesis H_0 is true. Thus, the statistic

$$F = \frac{MS_T}{MS_E} \tag{16}$$

has an F–distribution with $(k-1, N-k)$ degrees of freedom, and the rejection region for H_0 at the significance level α is

$$\frac{\frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{\boldsymbol{x}}_{i\cdot} - \bar{\boldsymbol{x}}_{\cdot\cdot})^2}{\frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{\boldsymbol{x}}_{i\cdot})^2} \geq f_\alpha(k-1, N-k). \tag{17}$$

The above procedure is summarized in Table 9.1.2.

Mathematica Experiment 1. Fertilizer Yield. In this experiment we will implement the analysis of variance procedures for the fertilizer yield data from Example 9.1.1.

Table 9.1.2 Single-factor ANOVA table

Source of variability	Deg. of freedom	Sums of squares	Mean sums of squares	$F-$ Ratio
Treatment	$k-1$	$SS_T = \sum_{i=1}^{k} n_i(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2$	$MS_T = \frac{SS_T}{k-1}$	$\frac{MS_T}{MS_E}$
Error	$N-k$	$SS_E = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\cdot})^2$	$MS_E = \frac{SS_E}{N-k}$	—
Total error	$N-1$	$TSS = \sum_{i=1}^{k}\sum_{j=1}^{n_j}(X_{ij} - \bar{X}_{\cdot\cdot})^2$	—	—

```
In[1]:= <<Statistics‘ContinuousDistributions‘
In[2]:= <<Statistics‘DescriptiveStatistics‘
In[3]:= X ={{151,158,162,149,153,151,150,159},
           {142,143,142,145,147,150,148},
           {147,142,143,146,144,142},
           {137,139,141,138,139,137,142,136,140}};
In[4]:= k=4;
In[5]:= n =Table[Length[X[[i]]], {i,1,k}]
Out[5]= {8, 7, 6, 9}
In[6]:= CapN=Sum[n[[i]],{i,1,k}]
Out[6]= 30
In[7]:= BX=N[ Table[Mean[X[[i]]] ,{i,1,k} ]]
Out[7]= {154.125, 145.286, 144., 138.778}
In[8]:= BarX=N[(1/CapN)Sum[  n[[i]] BX[[i]], {i,1,k}]]
Out[8]= 145.433
In[9]:= MST=(1/(k-1))Sum[n[[i]](BX[[i]]-BarX)^2, {i,1,k}]
Out[9]= 338.503
In[10]:= MSE=(1/(CapN-k))Sum[ Sum[(X[[i]][[j]]-BX[[i]])^2,
                                  {j,1,n[[i]]}], {i,1,k}]
Out[10]= 10.6869
In[11]:= MST/MSE
Out[11]= 31.6746
In[12]:= f[alpha_]:=Quantile[FRatioDistribution[k-1, CapN-k],
                                         1-alpha]
```

```
In[13]:= {f[0.10], f[0.05],f[0.01]}
Out[13]= {2.30749, 2.97515, 4.63657}
In[14]:= PValue=1-CDF[FRatioDistribution[k-1, CapN-k], MST/MSE]
Out[14]=  7.77406 .10^(-9)
```

So, at any reasonable significance level, the null hypothesis H_0 is rejected; that is, the conclusion is that there is significant evidence that the effects of fertilizers A, B, C, and D on crop yield are not the same. For the one-way ANOVA table, see Section 9.3.

9.2 Two-factor ANOVA

In this section we analyze the experimental design where the outcomes depend on two factors, say A and B, which can have several levels, say,

$$i = 1, 2, \ldots, I,$$

$$j = 1, 2, \ldots, J,$$

respectively. Thus the appropriate normal model, related to this *two-way* classification scheme is a family of normal random variables

$$\boldsymbol{X} = (X_{ijk}), \tag{1}$$

indexed by three indices i, j, and k, with means μ_{ij} depending on factor levels. The variance σ^2 is assumed to be the same for all X_{ijk}s. The index

$$k = 1, 2, \ldots, k_{ij}$$

indexes sample points for fixed levels i, j of the two factors. In general, the sample sizes can depend in the factor levels, but in what follows we will restrict our attention to the case when the sample sizes are uniform, that is

$$k_{ij} = K, \qquad i = 1, \ldots, I,\ j = 1, \ldots, J.$$

The collection of random variables $\{X_{ijk} : k = 1, \ldots, K\}$, corresponding to K repetitions of the experiment with factor A at level i and factor B at level j, is often called the (i, j)-*cell*.

Example 9.2.1 Workers' Productivity.

A company wants to evaluate the productivity of unskilled (U) and skilled (S) workers, who can use either an old (O) or new (N). The situation calls for a two-factor design, where the first factor is the workers' skill (and it can be at either of two levels, U and S) and the machine type (which can also have two levels, O and N). Taking samples of size $K = 5$, where each sample point represents the number of items manufactured in a week, resulted in the two-way Table 9.2.1.

Table 9.2.1 Workers' productivity

	Machines	
Workers	O	N
U	69	95
	69	98
	72	100
	74	96
	75	97
S	81	105
	88	110
	84	107
	87	112
	88	118

The question of interest is whether the workers' skill and machine age have any incremental or *additive* effects on workers' productivity. Note that labeling the cell means just μ_{ij} would not take the full advantage of the data structure. For that reason we will denote by

$$\mu$$

the 'basic' mean productivity of the unskilled worker U working on an old machine O, by

$$\mu + \alpha$$

the mean productivity of the unskilled worker U working on a new machine N, by

$$\mu + \beta$$

the mean productivity of the skilled worker S working on an old machine O, and, finally, by

$$\mu + \alpha + \beta + \gamma$$

the mean productivity of the skilled worker S working on a new machine N.

Here, the parameters α and β measure, but also separate, the additive effects of the skill and the modernization factors on productivity, while the parameter γ measures the interactive effects of the two factors. To verify whether either of the two factors has any effect on the productivity, we will test the hypotheses $H_0 : \alpha = 0$ and $H_0 : \beta = 0$. To verify the interactive effects of the two factors, we will test the hypothesis $H_0 : \gamma = 0$. Table 9.2.2 summarizes the structure of the means in our example.

Table 9.2.2 Means' structure in productivity data

	Machines	
Workers	O	N
U	μ	$\mu + \beta$
S	$\mu + \alpha$	$\mu + \alpha + \beta + \gamma$

The above example is typical for the general case in which the cell means in the experimental design (1) have the following 'additive' structure (decomposition):

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}, \tag{2}$$

where, to guarantee uniqueness, $\sum_i \alpha_i = \sum_i \beta_i = \sum_i \gamma_{il} = \sum_j \gamma_{kj} = 0$, for every $1 \leq k \leq I$ and $1 \leq l \leq J$. Then, we say that α_i represents the effect of the ith level of the first factor, β_j represents the effect of the jth level of the second factor, and γ_{ij} represents the interaction effect of levels i and j of the two factors, with the specific structure of the means, and we consider the linear model

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk},$$

where ϵ_{ijk} are independent $N(0, \sigma^2)$ random variables.

Such an experimental design is visualized in Table 9.2.3. The dots indicate indices over which the averaging is performed. The notation in the table is as follows:

$$\bar{x}_{ij\cdot} = \frac{1}{K}\sum_{k=1}^{K} x_{ijk},$$

$$\bar{x}_{i\cdot\cdot} = \frac{1}{J}\sum_{j=1}^{J} \bar{x}_{ij\cdot}, \qquad \bar{x}_{\cdot j\cdot} = \frac{1}{I}\sum_{i=1}^{I} \bar{x}_{ij\cdot},$$

$$\bar{x}_{\cdots} = \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} x_{ijk}.$$

Table 9.2.3 Data and means representation in two-factor experiments

	Factor B				Row
Factor A	1	2	....	J	means
1	x_{111} x_{112} · x_{11K}	x_{121} · · x_{12K}		x_{1J1} · · x_{1JK}	$\bar{x}_{1\cdot\cdot}$
2					
· · ·					
I	x_{I11} · x_{I1K}			x_{IJ1} · x_{IJK}	$\bar{x}_{I\cdot\cdot}$
Column means	$\bar{x}_{\cdot 1\cdot}$	$\bar{x}_{\cdot 2\cdot}$		$\bar{x}_{\cdot J\cdot}$	$\bar{x}_{\cdots}$

Since the number K of observations in each cell is constant, one can first calculate the mean in each cell and then calculate the row and column means as in Table 9.2.4.

Table 9.2.4 Reduced table of means

	Factor B				Row
Factor A	1	2		J	means
1	$\bar{x}_{11\cdot}$	$\bar{x}_{12\cdot}$		$\bar{x}_{1J\cdot}$	$\bar{x}_{1\cdot\cdot}$
2	$\bar{x}_{21\cdot}$	$\bar{x}_{2J\cdot}$			$\bar{x}_{2\cdot\cdot}$
· ·	· ·	· ·		· ·	· ·
I	$\bar{x}_{I1\cdot}$	$\bar{x}_{I2\cdot}$		$\bar{x}_{IJ}$	$\bar{x}_{I\cdot\cdot}$
Column means	$\bar{x}_{\cdot 1\cdot}$	$\bar{x}_{\cdot 2\cdot}$		$\bar{x}_{\cdot J\cdot}$	$\bar{x}_{\cdots}$

In the above two-factor experimental design, ANOVA will permit us to test the following hypotheses:

$$H_A : \alpha_1 = ... = \alpha_I = 0, \tag{3}$$

which asserts that the level change of factor A has no effect on the outcome of the experiment,

$$H_B : \beta_1 = ... = \beta_J = 0, \tag{4}$$

which makes a similar assertion about the effects of factor B, and

$$H_{AB} : \gamma_{ij} = 0, \quad \text{for all } i = 1, \dots, I, \; j = 1, \dots, J, \tag{5}$$

which asserts that there are no interactions between factors A and B.

The analysis of variance for the two-factor design follows the same pattern as ANOVA in one-factor experiments discussed in Section 9.1. Define the *total sum of squares*

$$TSS = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (X_{ijk} - \bar{X}_{...})^2, \tag{6}$$

measuring the pooled variability of all the samples, the *factor A sum of squares*

$$SS_A = JK \sum_{i=1}^{I} (\bar{X}_{i..} - \bar{X}_{...})^2 \tag{7}$$

representing variability due to factor A, the *factor B sum of squares*

$$SS_B = IK \sum_{j=1}^{J} (\bar{X}_{.j.} - \bar{X}_{...})^2 \tag{8}$$

representing variability due to factor B, the *interaction sum of squares*

$$SS_I = K \sum_{i=1}^{I} \sum_{j=1}^{J} (\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...})^2 \tag{9}$$

representing variability due to the interaction between the two factors, and finally, the *error sum of squares*

$$SS_E = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (X_{ijk} - \bar{X}_{ij.})^2 \tag{10}$$

representing variability due to random errors.

The expectations of all these estimators are given in Table 9.2.5. Note that, again, only the normalized SS_E is always an unbiased estimator for σ^2, while the normalized SS_A (resp., SS_B, and SS_I) is unbiased only under the hypothesis H_A (resp., H_B and H_I).

Table 9.2.5 Expectations of SS estimators in two-factor ANOVA

$E(SS_A)$	$=$	$(I-1)\sigma^2 + JK\sum_{i=1}^{I}\alpha_i^2$
$E(SS_B)$	$=$	$(J-1)\sigma^2 + IK\sum_{j=1}^{J}\beta_j^2$
$E(SS_I)$	$=$	$(I-1)(J-1)\sigma^2 + K\sum_{i=1}^{I}\sum_{j=1}^{J}\gamma_{ij}^2$
$E(SS_E)$	$=$	$IJ(K-1)\sigma^2$

The distributions of the SS estimators, under respective hypotheses, are given in Table 9.2.6.

The fundamental algebraic relation

$$TSS = SS_A + SS_B + SS_I + SS_E. \tag{11}$$

splits the total variability of all the samples into *variance components* representing the variability due to factors A and B separately, to the interaction of factors A and B, and to random errors.

As in one-factor ANOVA, we will reject hypotheses H_A, H_B, and H_{AB}, respectively, whenever the normalized SS estimators (mean sums of squares)

$$MS_A = \frac{SS_A}{I-1}, \tag{12}$$

$$MS_B = \frac{SS_B}{J-I}, \tag{13}$$

and

$$MS_I = \frac{SS_I}{(I-1)(J-1)}, \tag{14}$$

Table 9.2.6 Distributions of SS estimators in two-factor ANOVA

Under H_A,	$\frac{SS_A}{\sigma^2}$	has a χ^2 distribution with $I - 1$ degrees of freedom.
Under H_B,	$\frac{SS_B}{\sigma^2}$	has a χ^2 distribution with $J - 1$ degrees of freedom.
Under H_{AB},	$\frac{SS_I}{\sigma^2}$	has a χ^2 distribution with $(I - 1)(J - 1)$ degrees of freedom.
	$\frac{SS_E}{\sigma^2}$	has a χ^2 distribution with $IJ(K - 1)$ degrees of freedom.

are large relative to the mean error sum of squares

$$MS_E = \frac{SS_E}{IJ(K-1)}, \tag{15}$$

with the boundary of the rejection region determined, for a given significance level α, by the upper tail α-quantile $f_\alpha(\cdot, \cdot)$ of the respective F-distribution.

Thus, we will reject the hypothesis H_A at the significance level α, if

$$\frac{MS_A}{MS_E} \geq f_\alpha(I-1, IJ(K-1)), \tag{16}$$

the hypothesis H_B, if

$$\frac{MS_B}{MS_E} \geq f_\alpha(J-1, IJ(K-1)), \tag{17}$$

and, finally, the interaction hypothesis H_{AB}, if

$$\frac{MS_I}{MS_E} \geq f_\alpha((I-1)(J-1), IJ(K-1)). \tag{18}$$

To be formally correct, the formulas (16-18) should have SS estimates rather than SS estimators (which are random variables) inserted, but we did not want to further complicate our notation. The above discussion is summarized in Table 9.2.7.

Mathematica Experiment 1. Workers' Productivity. In this experiment we will implement the analysis of variance procedures for the workers' productivity data from Example 9.2.1.

```
In[1]:= <<Statistics`ContinuousDistributions`
In[2]:= <<Statistics`DescriptiveStatistics`
In[3]:= X ={{{69,69,72,74,75},{95,98,100,96,97}},
                {{81,88,84,87,88},{105,110,107,112,118}}};
In[4]:= L=2; J=2; K=5;
In[5]:= CellMeans= N[ Table[Mean[X[[i,j]]] , {i, L}, {j, J} ]]
Out[5]= {{71.8, 97.2}, {85.6, 110.4}}
In[6]:= RowMeans= N[Table[(1/J)Sum[CellMeans[[i,j]],
                             {j,1,J}],{i,1,L}]]
Out[6]= {84.5, 98.}
In[7]:= ColumnMeans= N[Table[(1/J)Sum[CellMeans[[i,j]],
                             {i,1,L}],{j,1,J}]]
Out[7]= {78.7, 103.8}
In[8]:= GrandMean=  N[(1/(L*J))Sum[Sum[CellMeans[[i,j]],
                             {j,1,J}],{i,1,L}]]
Out[8]= 91.25
In[9]:= MSA=(1/(L-1))J*K*Sum[ (RowMeans[[i]]-GrandMean)^2,
                             {i,1,L}]
Out[9]= 911.25
In[10]:= MSB=(1/(J-1))L*K*Sum[ (ColumnMeans[[j]]-GrandMean)^2,
                             {j,1,J}]
Out[10]= 3150.05
In[11]:= MSI=(1/(L-1)(J-1))*K* Sum[Sum[
     (CellMeans[[i,j]]-RowMeans[[i]]-ColumnMeans[[j]]+GrandMean)^2,
                            {j,1,J}],{i,1,L}]
Out[11]= 0.45
In[12]:= MSE=(1/(L*J* (K-1)))  Sum[Sum[Sum[
                            (X[[i,j,k]]-GrandMean)^2,
                            {j,1,J}],{i,1,L}],{k,1,K}]
Out[12]= 265.359
In[13]:= FRatios={MSA/MSE, MSB/MSE, MSI/MSE}
Out[13]= {3.43402, 11.8709, 0.00169581}
In[14]:= PValues={
     1-CDF[FRatioDistribution[L-1,L*J*(K-1)],MSA/MSE],
     1-CDF[FRatioDistribution[J-1,L*J*(K-1)],MSB/MSE],
     1-CDF[FRatioDistribution[(L-1)(J-1),L*J*(K-1)],MSI/MSE]
      }
```

```
Out[14]= {0.0824008, 0.00332561, 0.967662}
```

Table 9.2.7 Two-factor ANOVA table

Source of variability	Degrees of freedom	Sums of squares	Mean sums of squares	F–ratio
A	$I-1$	SS_A	$MS_A = \frac{SS_A}{I-1}$	$\frac{MS_A}{MS_E}$
B	$J-1$	SS_B	$MS_B = \frac{SS_B}{J-1}$	$\frac{MS_B}{MS_E}$
Interaction	$(I-1)(J-1)$	SS_I	$MS_I = \frac{SS_I}{(I-1)(J-1)}$	$\frac{MS_I}{MS_E}$
Error	$IJ(K-1)$	SS_E	$MS_E = \frac{SS_E}{IJ(K-1)}$	—
Total error	$IJK-1$	TSS	$MTS = \frac{TSS}{IJK-1}$	—

So, say, at the 0.05 significance level the hypotheses H_A and H_{AB} cannot be rejected, while the hypotheses H_B is rejected. In other words, our data indicate that skill has little impact on worker's productivity while the machine age significantly influences it. Also, there seems to be no significant interaction between these two factors.

To complete the analysis one should check the validity of the model by verifying the normality of residuals and the equality of their variances. For the two-way ANOVA table, see Section 9.3.

9.3 Experiments, exercises, and projects

1. The article "Origin of Precambrian Iron Formations" (*Econ. Geology*, 1964, pp.1025-1057) provided the following data on the total content (in percent) of iron (Fe) for four types of iron formation:
   Carbonate: 20.5, 28.1, 27.8, 27.0, 28.0, 25.2, 25.3, 27.1, 20.5, 31.3;

Slicate: 26.3, 24.0, 26.2, 20.2, 23.7, 34.0, 17.1, 26.8, 23.7, 24.9;
Magnetite: 29.5, 34.0, 27.5, 29.4, 27.9, 26.2, 29.9, 29.5, 30.0, 35.6;
Hematite: 36.5, 44.2, 34.1, 30.3, 31.4, 33.1, 34.1, 32.9, 36.3, 25,5.
Construct an ANOVA table for these data and comment on the results.

2. Four different coatings are being considered for a pipe that can be buried in one of three different types of soil. The two-factor experiment was designed to bury 12 pieces of pipe each coated with one of the four coatings and buried in one of the three types of soil for a fixed time. Afterwards, the depth of the deepest corrosion pit in each piece of pipe was measured (in .0001 in.) and the results were as follows (in order of three types of soil):
Coating 1: 64, 49, 50
Coating 2: 53, 51, 48
Coating 3: 47, 45, 50
Coating 4: 51, 43, 52
Assuming the validity of the additive model, carry out the ANOVA to see whether the amount of corrosion depends on either the type of coating used or the type of soil. Use $\alpha = .05$. Assume that the above numbers represent the means with variance 1. Sample size in each cell $n = 3$.

3. Lifetimes of springs under two different stress levels were measured to be
Stress level I: 225, 171, 198, 189, 189, 135, 162, 135, 117, 162
Stress level II: 216, 162, 153, 216, 225, 216, 306, 225, 243, 189
Find a rejection region at the 5% significance level for the hypothesis $H_0 : \mu_1 = \mu_2$, given that $\sigma_1^2 = \sigma_2^2$, but are unknown. Use the above data to reject or accept H_0.

4. An experiment was conducted to check the effect of C_2F_6 flow rate on the uniformity of the etch on a silicon wafer used in integrated circuit manufacturing. For each of the flow rates shown below, the experiment was replicated six times. The following results were obtained (G.Z. Yin and D.W. Jillie, *Solid State Technology*, May 1987):
Flow rate: 125 Etch uniformity (in percent): 2.7, 4.6, 2.6, 3.0, 3.2, 3.8
Flow rate: 160 Etch uniformity (in percent): 4.9, 4.6, 5.0, 4.2, 3.6, 4.2
Flow rate: 200 Etch uniformity (in percent): 4.6, 3.4, 2.9, 3.5, 4.1, 5.1
Perform the analysis of variance on this data set at significance level 0.05. Draw conclusions. Verify the model assumptions.

5. Two factors, glass type and phosphor type, influence the brightness of a display screen. The response variable measured the current (in microamps) necessary to obtain a specified brightness level. The data, obtained in *Industrial Quality Control* 1956, pp. 5-8, were as follows:

	Phosphor type		
Glass type	1	2	3
1	280	300	290
1	290	310	285
1	285	295	290
2	230	260	220
2	235	240	225
2	240	235	230

Carry out the ANOVA on this data set at several significance levels. Draw conclusions.

6. The `UVW Web Site` contains a *Mathematica* file `anova1` that automatically produces the single-factor ANOVA table. Study the code and use it to analyze the fertilizer yield data from *Mathematica* Experiment 9.1.1.

7. The `UVW Web Site` contains a *Mathematica* file `anova2` that automatically produces the single-factor ANOVA table. Study the code and use it to analyze the workers' productivity data from *Mathematica* Experiment 9.2.1.

8. Show that $t_\alpha^2(n)$ is the upper tail α-quantile of an F-distribution with 1 and n degrees of freedom.

9. Write a *Mathematica* code to verify experimentally the result of the above exercise.

10. Show that the parametrization for the means in the two-factor design,

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

where $\sum_{i=1}^{I} \alpha_i = 0, \sum_{j=1}^{J} \beta_j = 0, \sum_{i=1}^{I} \gamma_{ij} = 0, \sum_{j=1}^{J} \gamma_{ij} = 0$, is unique.

11. 6. The data set `iris1` on the `UVW Web Site` contains the lengths and widths of petals and sepals of three different species of the iris flower.

    (a) Test the hypothesis that there is no difference in the length (resp., width) of the petals among the three different species. Use the significance level 5% (resp., 1%).

    (b) Test the hypothesis that there is no difference in the length (resp., width) of the sepals among the three different species. Use the significance level 2% (resp., 0.5%).

    (c) Consider the two-factor model: Factor A is the species, Factor B has two levels, petals, and sepals. (1) Does the level of factor B have any effect on the length of the petals and sepals? (2) Does the level of factor B have any effect on the width of the petals and sepals? (3)

Does the type of flower have any effect on the length of the petals and sepals? (4) Does the type of flower have any effect on the width of the petals and sepals? (5) Do the two factors have any interaction concerning the length of the petals and sepals? (6) Do the two factors have any interaction concerning the width of the petals and sepals?

9.4 Bibliographical notes

There exists a voluminous literature on the subject of ANOVA. In this chapter we have only scratched the surface. On a practical and elementary level

[1] S.B. Vardeman, *Statistics for Engineering Problem Solving*, PWS Publishing, Boston, 1994.

is a good text. A more complete picture is presented in

[2] D.C. Montgomery and G.C. Runger, *Applied Statistics and Probability Engineers*, John Wiley & Sons, New York, 1994.

A full-fledged, graduate level, theoretical monograph on the ANOVA is

[3] S.R. Searle, G. Casella, and C.E. McCulloch, *Variance Components*, John Wiley & Sons, New York, 1992.

Appendix A

Uncertainty Principle in Signal Processing and Quantum Mechanics

Consider a (perhaps complex-valued) signal $f(t)$ such that

$$\int |f(t)|^2 dt = 1. \tag{1}$$

The quantity $|f(t)|^2$ can be thought of as the signal's "mass" density and describes its distribution in time. If the signal $f(t)$ is square integrable but (1) is not satisfied, then one can always normalize it by considering $f(t)/(\int |f(t)|^2 dt)^{1/2}$. In this context, the quantity

$$\int t|f(t)|^2 dt$$

can be interpreted as the location in time of the signal's "center of gravity", or its mean location. For the purposes of this section, and without loss of generality, we will assume that its mean location is at 0 or, in other words, that $\int t|f(t)|^2 dt = 0$. In this case, the quantity

$$\sigma^2[f] = \int t^2|f(t)|^2 dt \tag{2}$$

measures the average square deviation from the mean time location, or the degree of *localization* of the signal around its mean in the time domain.

On the other hand, the Fourier transform

$$\tilde{f}(\omega) = \frac{1}{2\pi}\int f(t)e^{-i\omega t} dt$$

displays no direct information about the signal's time localization, but has explicit information about its frequency localization. The square of its modulus $|\tilde{f}(\omega)|^2$

M. Denker, W. Woyczynski, *Introductory Statistics and Random Phenomena*,
Modern Birkhäuser Classics, DOI 10.1007/978-3-319-66152-0

is the frequency domain counterpart of the time density $|f(t)|^2$. Note that, by Parseval's formula

$$\int |f(t)|^2\,dt = 2\pi \int |\tilde{f}(\omega)|^2\,d\omega,$$

we have

$$\int |\tilde{f}(\omega)|^2 d\omega = \frac{1}{2\pi},$$

so that $2\pi\,|\tilde{f}(\omega)|^2$ can be viewed as the signal's normalized density in the frequency domain. Assume (again, without loss of generality) that the mean frequency

$$2\pi \int \omega|\tilde{f}(\omega)|^2 d\omega = 0.$$

Then the quantity

$$\sigma^2[\tilde{f}] = 2\pi \int \omega^2|\tilde{f}(\omega)|^2 d\omega \tag{3}$$

measures the mean square deviation from the mean frequency location, or the degree of localization of the signal in the frequency domain.

The *uncertainty principle* asserts that there exists a lower bound on the simultaneous localization of the signal in time and frequency domains. More precisely, it states that

$$\sigma^2[f]\sigma^2[\tilde{f}] \geq 1/4, \tag{4}$$

whenever the variances $\sigma^2[f]$ and $\sigma^2[\tilde{f}]$ are well defined. Note the universal constant 1/4.

To see why the uncertainty principle holds true, consider the integral

$$I(x) = \int |xtf(t) + f'(t)|^2 dt \geq 0 \tag{5}$$

where x is a real parameter. Then, since

$$|xtf(t) + f'(t)|^2 = (xtf + f')(xtf^* + (f')^*),$$

where f^* denotes the complex conjugate of f, we get that

$$I(x) = x^2 \int t^2|f|^2 dt + x \int t(f(f')^* + f'f^*)\,dt + \int |f'|^2\,dt. \tag{6}$$

The first integral in (6) is just $\sigma^2[f]$ [by definition (2)]. The second integral is equal to

$$\int t(ff^*)'dt = t|f(t)|^2\Big|_{-\infty}^{\infty} - \int |f|^2 dt = -1,$$

since $t|f|^2$ decays to zero at $\pm\infty$ in view of the assumption $\sigma^2[f] < \infty$. Finally, the third integral is equal to

$$2\pi \int \omega^2 |\tilde{f}(\omega)|^2 d\omega = \sigma^2[\tilde{f}]$$

because of Parseval's formula and the fact that the Fourier transform of f' is $i\omega\tilde{f}(\omega)$. As a result, the integral

$$I(x) = x^2\sigma^2(f) - x + \sigma^2(\tilde{f}). \tag{7}$$

This is a quadratic polynomial in variable x and, in view of (5), it is nonnegative for all values of x. As such, it has a nonpositive discriminant

$$1 - 4\sigma^2(f)\sigma^2(\tilde{f}) \leq 0,$$

which immediately yields the uncertainty principle (4).

Remark A.1 The Heisenberg[1] uncertainty principle in quantum mechanics. The (3-D version of the) above uncertainty principle concerning time-frequency localization has a celebrated interpretation in quantum mechanics, where the principle asserts that the position and the momentum of a particle cannot be simultaneously measured with arbitrary accuracy. Indeed, in quantum mechanics the particle is represented by a complex wave function $f(\boldsymbol{x})$, where $|f(\boldsymbol{x})|^2$ is the probability density of its position in space. The observables are represented by operators A on wave functions; the mean value of the observable is

$$\int (Af)(\boldsymbol{x}) f^*(\boldsymbol{x})\, d\boldsymbol{x}.$$

The *position observable* is represented by a multiplication by variable (vector) $\boldsymbol{x}$ and the *momentum observable* is represented by the operation of differentiation $\partial/\partial\boldsymbol{x}$. However, via the Fourier transform, the latter also becomes an operation of multiplication but by an independent variable (vector) $\boldsymbol{\omega}$ in the frequency domain. Thus, the uncertainty principle (4) gives the universal lower bound for the product

[1] Werner Heisenberg (1901–1976) was a Professor of Physics at the University of Göttingen.

of variances of the probability distributions of the position and of the momentum. In the three-dimensional space, and in the physical units, the lower bound 1/4 in (4) has to be replaced by a different mathematical constant multiplied by a universal physical constant called the Planck constant.

Remark A.2 One can check that the equality in the uncertainty principle (4) obtains only for the Gaussian function $f(t) = \pi^{-1/4}\exp(-t^2/2)$. Thus, the optimal simultaneous time and frequency localization is attained for a Gaussian-shaped signal.

Bibliographical note. Issues related to the uncertainty principle in the context of signals and wavelets are further developed in Chapter 7 of

[1] A.I. Saichev and W.A. Woyczynski, *Distributions in the Physical and Engineering Sciences. Volume 1. Distributional and Fractal Calculus, Integral Transform and Wavelets*, Birkhäuser-Boston, Cambridge, MA, 1997,

from which the material of this appendix has been borrowed. There is, of course, a large literature on the subject of quantum mechanical uncertainty principle. We shall just quote two classics:

[2] R.P. Feynman, R.B. Leighton and M. Sands, *Feynman Lectures on Physics. Volume 3. Quantum Mechanics*, Addison-Wesley, Reading, MA, 1965.

[3] L.D. Landau and E.M. Lifshitz, *Course of Theoretical Physics. Volume 3. Quantum Mechanics: Non-Relativistic Theory*, Pergamon Press, New York, 1977.

Appendix B

Fuzzy Systems and Logic

Fuzzy systems theory provides another, deterministic interpretation of probability and randomness which permits venturing outside the standard two-valued, TRUE-FALSE binary logic to situations with undetermined outcome in which several, or even continuum, of undeterminacy levels are possible; the law of the excluded middle is thus violated. It is spiritually related to the uncertainty principle philosophy, and counts among its predecessors the multivalued logic developed in the early 1930s by the Polish logician Jan Łukasiewicz.

The term *fuzzy set* and the foundations of the theory of fuzzy systems were introduced in 1965 by system scientist Lofti Zadeh. Since then the theory developed a large following in the engineering community and found many practical applications, including the automobile traction control systems. The terms *fuzzy engineer* and *defuzzification* now have technical meaning.

For a nonfuzzy subset A of the universe X the indicator function

$$\mathbf{1}_{\mathbf{A}}(\mathrm{x}) = \begin{cases} 1, & \text{if } x \in A; \\ 0, & \text{if } x \notin A, \end{cases} \tag{1}$$

has just two values, 0 and 1. By contrast, the *fuzzy subset A* is defined by a *membership function*

$$m_A(x) : X \longrightarrow [0, 1], \tag{2}$$

that is, the membership of a point x in a fuzzy set A can be of any degree between 0 and 1. Then, the usual two-valued set algebra is replaced by the continuum valued fuzzy set algebra, where

$$m_{A\cap B}(x) = \min\{m_A(x), m_B(x)\}, \tag{3}$$

$$m_{A\cup B}(x) = \max\{m_A(x), m_B(x)\}, \tag{4}$$

$$m_{A^c}(x) = 1 - m_A(x), \tag{5}$$

M. Denker, W. Woyczynski, *Introductory Statistics and Random Phenomena*,
Modern Birkhäuser Classics, DOI 10.1007/978-3-319-66152-0

These rules create a possibility of nontrivial *overlap* and *underlap* fuzzy sets $A \cap A^c$ and $A \cup A^c$.

A fuzzy system is the set of IF--THEN rules that maps inputs to outputs. Each fuzzy rule defines a *fuzzy Cartesian patch* $A_j \times B_j \subset X \times Y$ with the "if" fuzzy sets $A_j \subset \mathbf{R}^n$ identified by the membership function $a_j : \mathbf{R}^n \to [0, 1]$. An additive system sums the "fired" THEN fuzzy sets B_j' to give

$$B = \sum_{j=1}^{m} B_j' = \sum_{j=1}^{m} a_j(x) B_j, \tag{6}$$

and then computes the output

$$F(x) = \frac{\int_{-\infty}^{\infty} y \sum_{j=1}^{m} b_j'(y)\, dy}{\int_{-\infty}^{\infty} \sum_{j=1}^{m} b_j'(y)\, dy} \tag{7}$$

as the centroid which can be also written in the form

$$F(x) = \frac{\sum_{j=1}^{m} a_j(x) V_j c_j}{\sum_{j=1}^{m} a_j(x) V_j}, \tag{8}$$

where

$$V_j = \int_{-\infty}^{\infty} b_j(y)\, dy > 0, \qquad c_j = \frac{\int_{-\infty}^{\infty} y b_j(y)\, dy}{\int_{-\infty}^{\infty} b_j(y)\, dy}. \tag{9}$$

Provided at least one rule "fires" so that $a_j(x) > 0$ for some j, we have

$$F(x) = \frac{\sum_{j=1}^{m} a_j(x) P_j}{\sum_{j=1}^{m} a_j(x)} \tag{10}$$

if $V_1 = \ldots V_m$ and if the peak P_j of each "then" fuzzy set B_j equals the centroid c_j. This structure of the output $F(x)$ implies that it is the conditional expectation of Y given that the input $X = x$.

The above sketch gives the flavor of the mathematical contents of the fuzzy systems theory.

Bibliographical note. The multivalued logic is discussed in depth in the monograph

[1] H. Rasiowa and R. Sikorski, *The Mathematics of Metamathematics*, PWN Scientific Publishers, Warsaw, 1970.

The original paper

[2] L. Zadeh, Fuzzy Sets, *Information and Control* 8(1965), 338-353.

is still one of the crispest presentations of the foundations. For more recent developments, and textbook-style presentations, we refer to

[3] B. Kosko, *Neural Networks and Fuzzy Systems. A Dynamical Systems Approach to Machine Intelligence*, Prentice-Hall, Englewood Cliffs, NJ, 1992.

[4] L. Zadeh and J. Kacprzyk, Eds. *Fuzzy Logic for the Management of Uncertainty*, John Wiley & Sons, New York, 1992.

[5] W. Pedrycz, *Fuzzy Sets Engineering*, CRC Press, Boca Raton, FL, 1995.

Appendix C

A Critique of Pure Reason

A new psychology says that the mind is not a computer that works by the rules of logic, but a set of tools evolved to help people live pre-industrial lives. [1]

You are a barman and you will lose your license if you serve a drink to an underage drinker. At your bar are four people; you know what two are drinking (one has beer, one has coke) and you know the ages of the other two (one adult, one teenager). Ask the minimum number of questions that will ascertain if you are breaking the law. If your answer is that you need only ask the beer drinker's age and teenager's tipple, then you join the 75% of those asked the question who get it right. Muted congratulations.

Now consider someone laced with cards which have letters on one side and numbers on the other. He wants to check the rule "a card with a D on one side must have a 3 on the other", and he is presented with card D, card F, card 3 and card 7. Which cards must he turn over? More congratulations for the right answer this time, because only 25% of people say D and 7.

The intriguing thing about these two problems is that, to a logician, they are the same. The structure of the card problem, and the answer, are identical to the drinking problem. Why then is one problem easy and the other relatively hard? A small group of psychologists think they know the answer to this meta-question; if they are right, a new theory of the mind will be in order, one which has no such thing as general intelligence within it, and is not dominated by symbolic reasoning skills. The mind is not, they say, a reasoning machine—it is a machine designed for scraping out an existence in a clan of hunter gatherers.

The tests in the first paragraph are called Wason tests, after the psychologist who first tortured people with them. They are the essence of a simple reasoning task—the application of the rule "if p, then q". If the mind were a straightforward reasoning machine, all Wason tests would be equally soluble. In fact their solution depends on the story around them. Psychologists first guessed that it is all about familiarity with the content of the story—a familiar story would be easier for the

[1] Reprinted from *The Economist*, July 4th, 1992. We preserved the British punctuation and spelling.

M. Denker, W. Woyczynski, *Introductory Statistics and Random Phenomena*,
Modern Birkhäuser Classics, DOI 10.1007/978-3-319-66152-0

mind to reduce to soluble *p*s and *q*s. The barman's problem is obvious, the other one is obscure. But experiments have ruled that out. Familiar contexts ("if a person eats hot chilli peppers, then he will drink cold beer") prove difficult, while strange ones ("if a man eats cassava root, then he must have a tattoo on his face") sometimes prove easy.

Leda Cosmides and John Tooby of the University of California at Santa Barbara are among the band that thinks it has an explanation. They argue that the underlying logical structure is of little relevance, and that the familiarity of the context does not matter much either. What matters a lot is the nature of the context. If a rule of the form "he who takes the benefit must pay the cost" is at stake, then solving the problem means spotting cheats. People do this well. The mind is not following abstract reason; it is enforcing a social contract.

To demonstrate this, Dr Cosmides gave students at Stanford University a series of Wason tests. Some were set in a fictitious culture in which rules such as the one that restricts cassava root to tattoo-wearers are laid down by a chief called Big Kiku. Others were simply nonsensical conjunctions of events: "If you eat duiker meat, then you have found an ostrich eggshell". The students proved far better at enforcing Big Kiku's laws than at pursuing arbitrary pieces of if–then logic.

Gerd Gigerenzer, of Salzburg University, and his colleagues have gone one step further. In an ingenious Wason test, they asked two groups of students to turn over cards to test the statement: "If an employee gets a pension, then that employee must have worked for the firm for at least ten years". The statements on the cards were "Gets a pension", "Did not get a pension", "Worked for eight years", "Worked for 12 years". The difference between the two groups was that one was told they were employers, the other that they were employees.

If they were solving the problem in some purely logical way, both groups should get the right answer; the rule is broken only when somebody has worked for less than ten years but gets a pension, so the cards to turn are "Gets a pension" and "Worked for eight years". But if they are looking for cheats, employees will worry about those who worked for 12 years and do not get a pension, even though this is strictly irrelevant to the problem. So it proves. Almost all the employers, whose interests coincide with the right answer, turned the correct cards. The employees, however, apparently more concerned with justice than logic, plumped for "Did not get a pension" and "Worked for 12 years" by a ratio of six to one.

Dr Cosmides and Dr Tooby take all this to mean that the Wason test awakens a specific mental mechanism that keeps the accounts in social contracts and is on the look-out for cheats. Following on from that, they suggest that the brain is a bundle of such job-specific mechanisms, shaped by evolution rather than applying the same general-purpose "reason" to all the problems it encounters.

The inspiration for this notion is the idea that society is based on social exchange of the form "You scratch my back. I'll scratch yours." In animal societies, all apparently altruistic behaviour that is not based on kinship seems to work like this. Baboons help each other in fights and keep a close account of who owes whom favours. A vampire bat that does not regurgitate part of its blood meal with a neighbour who came home hungry forfeits the return favour at a later date.

Some anthropologists are coming to see human societies in much the same light. Kim Hill of the University of Michigan, who studies the Aché in Paraguay, has found that a hunter who returns from the forest laden with meat will give some to his partner and children, some surreptitiously to a woman he wants to have sex with—the trade is explicit—and some to other hunters who might return the favour later. A fatherless Aché family often nearly starves, because nobody has an incentive to share meat with a family that cannot reciprocate. Such a system of debts is well suited to hunters. A hunter may return empty-handed for days on end and then suddenly kill a tapir—far more than he can eat. Better to share it, and thus be owed a debt, than waste it.

Gambling on certainty. Other aspects of rationality are starting to fall by the wayside, too. Dr Gigerenzer, a probability theorist by training, has tried, using similar ideas, and similar experiments to his work on social contracts, to explain the mistakes people make when thinking about probabilities.

Probability theorists have always been split over the question of what probability is. "Bayesians"—named after the originator of their point of view—say it is a measure of subjective certainty about single events: "I'm 90% sure of my horse winning this race." "Frequentists" say it is the long-run frequency of events: "Nine out of my last ten tips were winners." People are quite good at assessing the latter while, to the delight of bookmakers, they are generally hopeless at the former.

One example of this is the psychological paradox known as the "overconfidence effect". Overconfidence tends to be specific rather than general. When asked a general-knowledge question such as "which city is bigger, Bonn or Heidelberg?" people are more likely to think they know the correct answer than actually to know it. But after answering a string of questions, they are good at estimating the number they got right. Psychologists have used such "fallacies" to argue that people are bad at statistics. Dr Gigerenzer thinks, rather, that people are natural frequentists. He has found that merely rephrasing a problem in frequentist rather than Bayesian terms generally increases the number of people who can solve it (see the Section "Think again", below).

Again, a look at primitive life suggests a reason why. The probability of a single event is a meaningless fact in a hunter's world: what can he do about the fact that his chances of killing a tapir today are 3%? But the frequency of past events and past conjunctions is vitally important and always has been: he killed a tapir on three of the last 100 visits to that valley, and five out of 100 visits to this one. A German psychologist, Egon Brunswik, argued as early as the 1960s that human brains are constantly, and unconsciously, assessing such frequencies as guides to future events.

Given this view of man—a natural trader, ever concerned with social debts and an uncertain future—it is little wonder that human minds are interested in detecting cheats, not pursuing pure logic, and in sampling frequencies rather than making risky one-off guesses. Reasoning, in this view, depends on a number of such mental subroutines. Logic is a refinement and codification of their results—a creative and powerful generalisation, but the crowning glory of human intellectual achievement rather than its deep foundation.

Think again. Arguments about Bayesian v. frequentist statistics may sound esoteric, but they touch the real world. Dr Cosmides and Dr Tooby applied Dr Gigerenzer's ideas to a disturbing piece of research done in the late 1970s.

Ward Casscells and his colleagues at Harvard Medical School had stopped 60 doctors in the corridors of a prestigious hospital and asked them the question: "If a test to detect a disease, whose prevalence is 1/1,000, has a false-positive rate of 5%, what is the chance that a person found to have a positive result actually has the disease, assuming you know nothing about the person's symptoms?" Only 11, or 18%, of the doctors knew the answer. Most said 95%, and the average answer was 56%. (The correct answer is one in 51, just under 2%.)

Dr Cosmides and Dr Tooby asked the question of a group of Stanford students and got the same poor success rate. Even when they clarified the meaning of the term "false-positive", which laymen might not be familiar with (though doctors should be), people still got it wrong. Then they rephrased it. Instead of a Bayesian-style question about the chance of a single infection, they asked a more frequentist one: "How many of 1,000 people who tested positive actually had the disease?" They expected the students (who were neither medically nor statistically trained) to do slightly better. They were stunned by the result: three times as many as before got the right answer.

Nor is this just an abstract experiment. One young American recently committed suicide on learning he had tested positive for HIV. The test had a 4% rate of false positives and he believed his chance of carrying the virus was 96%. It was 10%. Beware of Bayesians bearing diagnoses; the mind is a frequentist device.

Appendix D

The Remarkable Bernoulli Family[1]

BERNOULLI. Originally from Antwerp, the family became citizens of Basel in 1622. Coat of arms: In silver, a tri-partite green branch, each having seven (sometimes nine) leaves. The progenitor was Jakob (d.1583), who in 1570 had fled from Antwerp to Frankfurt/Main, thus escaping Count Alba's persecution of heretics. Jakob Bernoulli, a wholesale grocer and tradesman, had 17 children. One line of his descendants stayed in Frankfurt, where it is still flourishing; others settled in Hamburg, Cologne, Breslau (Wrocław), and Basel.

1. **Jakob** (1598–1634) Grandson of progenitor Jakob, druggist and grocer, was made a citizen of Basel in 1622.
2. **Niklaus** (1623–1708) Son of 1. Elected member of the board of the Saffron's Guild; representing the guild on the city's Great Council [legislative] in 1668.
3. **Niklaus** (1662–1716) Son of 2, painter, elected to the Small Council [executive] in 1705.

   The family's reputation for its outstanding impact on science, especially mathematics, was established first by:

4. **Jakob**[2] (1654–1705) Brother of 3. Professor of Mathematics at Basel University. After graduating in theology in 1676, he traveled widely across Switzerland, France, the Netherlands, and England, where he made contact with the most prominent mathematicians. On his return to Basel in 1682, he inaugurated lectures on experimental physics and, in 1687, was appointed to the Chair of Mathematics. Inspired by, yet independently from, Leibniz, he explored the infinitesimal calculus. He published studies on the logarithmic spiral, the loxodromes, infinite fractions, infinite

[1] Adapted, with permission, from The Bernoulli News, June 1994, 15-17.

[2] Author of *Ars Conjectandi*

M. Denker, W. Woyczynski, *Introductory Statistics and Random Phenomena*,
Modern Birkhäuser Classics, DOI 10.1007/978-3-319-66152-0

series (on which occasion the so-called Bernoulli numbers were discovered), etc., as well as studies on the isoperimetric problem—hereby laying the foundation for variational calculus. In 1701 he became a member of the Berlin Academy.

5. **Johann** (1667–1748) Younger brother of 4. Also Professor of Mathematics and even more renowned for his contributions. He had studied medicine as well, qualifying in this discipline at Basel University in 1694. The next year he became Professor of Mathematics and Physics at Groningen. In 1705 he succeeded his brother to the chair at Basel University. Listing all his achievements in mathematics would require a comprehensive survey of the whole higher analysis; suffice it to name his evaluation, using differential calculus, of the limits of quotients whose numerator and denominator both tend to zero, invention of calculus for exponential functions, most integration methods (together with Leibniz), etc. He was elected a corresponding member of the Academies of Paris (1699), Berlin (1701), London (1712), Bologna (1724), and Petersburg (1725). Among his students were his own sons Niklaus, Daniel, and Johann—and also Leonhard Euler.

6. **Niklaus**[3] (1687–1759) Son of 3. Mathematician, lawyer, and philosopher. His favorite field of research was the theory of infinite series. From 1716 to 1719 he was Professor of Mathematics at Padua University; in 1722 Basel University appointed him as Professor of Logic, and as Professor of Codified and Feudal Law in 1731. Member of the Academies of Berlin, London, and Bologna.

7. **Niklaus II** (1695–1726) Son of 5. Mathematician and lawyer. Professor of Law at Bern University in 1723. In 1725 he left (with his brother Daniel) for Petersburg's newly founded Academy, where he suffered an early death on July 26, 1726. His contributions to mathematics lie above all in the field of integral and differential calculus.

8. **Daniel** (1700–1782) Son of 5. Mathematician, physicist, medical doctor, and botanist, studied mathematics with his father and his brother Niklaus from the age of 11, and later studied medicine in Basel, Heidelberg, and Venice. In 1725 he left for Petersburg with his brother Niklaus. Between 1725 and 1757 he was ten times awarded prizes for his mathematical work by the Paris Academy—on one occasion jointly with his father (1734); on another with his youngest brother, Johann. In 1733 he was made Professor of Anatomy and Botany at Basel University. His scientific achievements reach out into the most diverse areas of mathematics, physics, astronomy, etc. Member of the Academies of Petersburg, Berlin, London, and Paris.

[3] Editor of *Ars Conjectandi*. He also applied the Law of Large Numbers to longevity data.

9. **Johann II** (1710–1790) Son of Johann (5.). Mathematician and lawyer. After obtaining his doctorate in law, 1732, he joined his older brother Daniel in Petersburg, but returned with him to Basle the following year. In 1743 he was given the Chair of Rhetoric at Basel University, which he exchanged for a Chair of Mathematics in 1748. Four of his studies were awarded prizes by the Paris Academy.

10. **Johann III** (1744-1807) Eldest son of 9. Astronomer and mathematician. Appointed by the Berlin Academy in 1763, he became Director of the Berlin Observatory in 1767. Member of the Academies of Paris, Petersburg, Rome, London, and Stockholm. Director of the Berlin Academy.

11. **Daniel** (1751–1834) Second son of 9. Doctor of Medicine, Professor of Rhetoric, followed by professorships in physics and medicine; finally, also Major-Domo of the Dean of the Basel Cathedral.

12. **Jakob II** (1759-1789) Youngest son of Johann II (9.). Mathematician and physicist, he served as secretary to the Imperial Ambassador in Venice, from 1779. In 1786 he joined the Petersburg Academy. His research touched on problems of theoretical mechanics.

The descendants of the fourth son of Johann II, Emanuel, live in Venice and Petersburg. From among the descendants of the fifth son, the pharmacist Niklaus, mention should be made of:

13. **Leonhard** (1791–1871) Councillor of the city of Basel.

14. **Niklaus** (1793–1876) President of the Criminal Court.

15. **August** (1839–1921) Son of 13. Ph.D., renowned historian.

16. **August** (1879–) Son of 15. Professor of Physical Chemistry at Basel University.

17. **Hans** (1876–) Grandson of 14. Architect and Professor at ETH Zürich.

Among the descendants of Daniel (11) quite a few have been important figures:

18. **Christoph** (1782–1863) Biologist and technologist; founder and director of the Philotechnic Institute; Professor of Natural History at Basel University.

19. **Carl Christoph** (1861–) Grandson of 18. Ph.D., chief librarian of the Basel Great Council.

20. **Carl Albrecht** (1863–) Licentiate in Theology, novelist and dramatist. His works are listed in the *Schweizerisches Zeitgenossen Lexikon (Swiss Lexicon of Contemporaries).*

21. **Johannes** (1864–1920) Ph.D., Director of the Swiss National Library in Bern, 1895–1908.
22. **Eduard** (1867–) Brother of 21. Professor of History of Music, Zürich University.
23. **Hieronymus** (1745–1829) Great-nephew of 4 and 5. Biologist.
24. **Karl Gustav** (1834–1878) Biologist.
25. **Johann Jakob** (1831–1920) Ph.D., Professor of Archaeology at Basel University.

The Bernoulli family is still thriving; Daniel Bernoulli is currently Professor of Geology at ETH Zürich.

Bibligraphical Note: The above biographical data have been taken, with the help from M.G. Soland of ETH Zürich in preparation of the English version, from

[1] H.Türler et al., Eds., *Historisch-biographisches Lexikon der Schweiz*, 7 vols., Neuchâtel, 1921 - 1934, Vol. 2 (1924).

All Bernoullis active and successful in mathematics and/or any other area of knowledge were given an entry, chronologically listed up to 1922. For detailed information on specific achievements of this famous Swiss family consult

[2] C. C. Gillespie, Ed., *Dictionary of Scientific Biography*, Vol.2, Charles Scribner and Sons, New York 1970.

Appendix E

Uncertain Virtual Worlds
Mathematica Packages

by Bernard YCART

0. Introduction

The word "packages" is not totally appropriate here. The functions are left without input protections. They are not declared as packages in the language, and their denominations are not protected. This may cause some errors. However, it should permit easier 'on line' testing and modifications. The 13 files should be copied from the `UVW Web Site` to a subdirectory `UVW` of the directory `PACKAGES` in the *Mathematica* folder. Then they will be accessible by the command `<<UVW'Package'`. They will work with *Mathematica 2.0* or higher.

The programming style is a compromise between two objectives. The first one is to respect the spirit and style of *Mathematica*, in order to use the language as efficiently as possible. The second one is to make the functions transparent and easy to modify for the user. As an illustration of that compromise, one can compare for instance the function `Distribution` in package `UVW'DiscSamp'`, with the similar `Frequencies` of the standard *Mathematica* package `Statistics'DataManipulation'`. Numerous examples and suggestions for *Mathematica* experiments are included.

All functions of the type

```
RS...[...,n]
```

return, as a list, a Random Sample of size `n`, i.e., a realization of an `n`-tuple of independent identically distributed random variables.

For easy reference, we provide first a complete list of `UVW` packages and commands. The detailed descriptions follow later. The full code is available on the `UVW Web Site`, where it is augmented by various pedagogical programming commentaries.

M. Denker, W. Woyczynski, *Introductory Statistics and Random Phenomena*,
Modern Birkhäuser Classics, DOI 10.1007/978-3-319-66152-0

1. UVW‘Billiard‘ (Billiards with a round obstacle)
   ```
   NextBounce[currentposition, r]
   Trajectory[shootingangle, tmax, r]
   Bundle[listofangles, tmax, r]
   Differences[alpha, deltaalpha, tmax, r]
   ```
2. UVW‘ContSamp‘ (Simulations of continuous distributions)
   ```
   RSContinuousDistribution[density, a, b, n]
   RSIndependent2D[density1, a1, b1, density2, a2 ,b2, n]
   RSNormal2D[sigma1, sigma2, rho, n]
   RSUnitBall[dim, n]
   ```
3. UVW‘DataRep‘ (Data representations)
   ```
   Histogram[listofdata, listofbounds]
   RegularHisto[listofdata, xmin, xmax, nx]
   SamplePlot2D[listof2Ddata]
   Histogram2D[listof2Ddata, xmin, xmax, nx, ymin, ymax, ny]
   LargeNumbers[listofdata]
   CentralLimit[listofdata, mu, sigma, n]
   ```
4. UVW‘DiscSamp‘ (Simulations of discrete distributions)
   ```
   Distribution[listofdata]
   RSPermutation[list, n]
   RSExtract[list, k, n]
   RSDiscreteDistribution[dist, n]
   ```
5. UVW‘DynSyst‘ (Dynamical Systems)
   ```
   Mean[listofdata]
   Variance[listofdata]
   CovarianceFunction[listofdata, n]
   AsymptoticVariance[listofdata, n]
   CorrelationDimension[listofdata, step, nstep]
   TentFunction[a, x]
   LogisticFunction[a, x]
   IterateAPhi[matrixA, functionPhi, vectorX0, n]
   ```
6. UVW‘Fractals‘ (Deterministic and random Von Koch curves)
   ```
   Triangle
   Star
   Island
   Battlement
   Hat[sharpness]
   Wy[angle1, length1, angle2, length2]
   RSWy[n]
   RSTruncs[n]
   TransformSegment[segment, pattern]
   ```

```
IteratePattern[listofsegments, pattern, n]
IterateRandomPattern[listofsegments, patterns, probas, n]
DrawSegments[listofsegments]
```

7. UVW‘Interact‘ (Simulation of spin systems in the plane)

```
Checkerboard
Diagonals
RConfig[p, width, height]
Uniform[lambda, mu]
Ising[Alpha, Beta]
Contact[lambda]
Voter
Cyclic[n, bound]
RepartConfig[config]
Evolution[initialconfig, rates, niter]
DrawConfig[config, opts]
```

8. UVW‘Lorenz‘ (Lorenz's attractor)

```
Lorenz[s, b, r]
LorenzArray[matrix]
```

9. UVW‘PseuGene‘ (Congruential and midsquare generators)

```
CongruGenerator[seed, a, c, m, n]
MidsquareGenerator[seed, n]
CongruentialLoop[seed, a, c, m]
MidsquareLoop[seed]
```

10. UVW‘RandWalk‘ (Random walks and random vector fields)

```
RandomWalk[listof2Dvelocities, deltat]
VectorField[arrayof2Dvelocities]
VectorFieldTrajectory[arrayof2Dvelocities, deltat, tmax]
```

11. UVW‘StoGho‘ (Stochastic Ghost)

```
StoGho[width, mood]
GalleryOfPortraits[matrix]
```

12. UVW‘TimeRep‘ (Queues and time processes)

```
Queue[interarrivals, services]
CumulatedTimes[listoftimes]
Geiger[listoftimes]
```

13. UVW‘ZeroOne‘ (Lists of zeros and ones)

```
RSZeroOne[p, n]
Binary[functionf, listofzeroones]
PlotZeroOne[listofzeroones]
AnimateShift[listofzeroones]
ActualLength[listofzeroones]
```

```
Weight[listofzeroones]
WeightedAlphabeticalOrder[listofzeroones]
Entropy[t]
```

1. UVW'Billiard' — billiards with a round obstacle

`NextBounce[currentposition,r]` computes the next bouncing point of a ball inside a square table with a circular obstacle of radius `r`. The initial conditions are given in the list `currentposition` that has four real coordinates, respectively the abscissa, ordinate, incoming angle of the ball, and the current time. The result is returned as another list of four elements, the abscissa and ordinate of the new bouncing point, the new direction of the ball and the current time incremented by the running time between the two bounces.

`Trajectory[shootingangle,tmax,r]` draws a square table with a centered circular obstacle of radius `r` (default 0.5). Draws inside this support the trajectory of a ball starting at the bottom left corner with initial direction `shootingangle`, up to time `tmax` .

`Bundle[listofangles,tmax,r]` draws a square table with a centered circular obstacle of radius `r`. Draws inside the trajectories of balls starting at the bottom left corner with initial directions read in `listofangles`, up to time `tmax`.

`Differences[angle,dangle,tmax,r]` simulates two trajectories of balls in a square table with a centered circular obstacle with radius `r`. Both trajectories start from the bottom left corner. They are followed up to time `tmax`. The shooting angle of the first trajectory is `angle`, its difference with the second shooting angle is `dangle`. The function represents three consecutive graphics. The first one is the billiard table with the two trajectories . The second one is the evolution of the absolute difference of angles as a function of time. The third one is the norm of the difference of positions as a function of time.

Examples:

Here are some trajectories with a growing obstacle.

```
In[1]:= <<UVW'Billiard'
In[2]:= Trajectory[0.4,100,0.]
In[3]:= Trajectory[0.4,100,0.1]
In[4]:= Trajectory[0.4,100,0.5]
In[5]:= Trajectory[0.4,100,0.8]
In[6]:= Bundle[Range[Pi/12,5*Pi/12,Pi/12],10,0.]
In[7]:= Bundle[Range[Pi/12,5*Pi/12,Pi/12],50,0.]
In[8]:= Bundle[Range[Pi/12,5*Pi/12,Pi/12],50,0.1]
In[9]:= Bundle[Range[Pi/12,5*Pi/12,Pi/12],50,0.5]
In[10]:= Bundle[Range[Pi/12,5*Pi/12,Pi/12],50,0.8]
```

Here is how two trajectories, with close shooting angles, start differing chaotically after several bounces off the circular obstacle.

```
In[1]:= <<UVW'Billiard'
In[2]:= Differences[0.4,0.001,20,0.]
In[3]:= Differences[0.4,0.001,100,0.]
In[4]:= Differences[0.4,0.001,20,0.1]
In[5]:= Differences[0.4,0.001,20,0.5]
In[6]:= Differences[0.4,0.001,100,0.5]
In[7]:= Differences[0.4,0.001,200,0.5]
```

2. UVW'ContSamp' — simulations of continuous distributions

RSContinuousDistribution[functionf,a,b,n] returns a sample of size n for the distribution with density functionf on the interval [a,b].

RSIndependent2D[functionf1,a1,b1,functionf2,a2,b2,n] returns a sample of size n, {{x1,y1},...,{xn,yn}} of a two-dimensional random vector (X, Y), where X and Y are independent, X has density functionf1 on the interval [a1,b1], Y has density functionf2 on the interval [a2,b2].

RSNormal2D[Sigma1,Sigma2,rho,n] returns a sample of size n for the two-dimensional Gaussian vector (X, Y). The means of X and X are zero, their standard deviations are sigma1 and sigma2. Their correlation coefficient is rho.

RSUnitBall[dim,n] returns a sample of size n of vectors uniformly distributed in the unit ball of the space of dimension dim.

Examples:

Random samples of standard continuous distributions can be simulated using the standard function Random, together with the distributions defined in the package Statistics'ContinuousDistributions'.

```
In[1]:= Table[Random[],{100}]
In[2]:= Table[Random[Real,{2,4}],{100}]
In[3]:= <<Statistics'ContinuousDistributions'
In[4]:= Table[Random[ExponentialDistribution[1.]],{100}]
In[5]:= Table[Random[NormalDistribution[0.,1.]],{100}]
In[6]:= Table[Random[WeibullDistribution[2.,1.]],{100}]
```

A random sample of a distribution with an arbitrary density can be simulated using RSContinuousDistribution. Notice that the function f only needs to be non-negative over the prescribed interval. RSContinuousDistribution divides it automatically by its integral.

```
In[1]:= <<UVW`ContSamp`
In[2]:= <<UVW`DataRep`
In[3]:= f[x_]:=1+Sin[x]
In[4]:= samp=RSContinuousDistribution[f,0,10,2000];
In[5]:= g1=RegularHisto[samp,0,10,20]
In[6]:= integral=NIntegrate[f[x],{x,0,10}]
In[7]:= fnorm[x_]:=f[x]/integral
In[8]:= g2=Plot[fnorm[x],{x,0,10}]
In[9]:= Show[g1,g2]
```

The function RSIndependent2D returns a two-dimensional sample with independent coordinates. Each coordinate is simulated using RSContinuousDistribution.

```
In[1]:= <<UVW`ContSamp`
In[2]:= <<UVW`DataRep`
In[3]:= f[x_]:=x^2
In[4]:= para2=RSIndependent2D[f,-1,1,f,-1,1,2000];
In[5]:= SamplePlot2D[para2,Frame->True,AspectRatio->1]
In[6]:= Histogram2D[para2,-1,1,8,-1,1,8]
In[7]:= marge1=Transpose[para2][[1]];
In[8]:= RegularHisto[marge1,-1,1,10]
```

The package Statistics`MultinormalDistribution`, delivered with version 3.0 of *Mathematica*, permits all sorts of manipulations with Gaussian vectors, including their simulation by the standard Random function. RSNormal2D returns samples of Gaussian vectors in the plane.

```
In[1]:= <<UVW`ContSamp`
In[2]:= <<UVW`DataRep`
In[3]:= gauss2=RSNormal2D[2,1,-0.8,2000];
In[4]:= SamplePlot2D[gauss2]
In[5]:= Histogram2D[gauss2,-6,6,10,-3,3,10]
In[6]:= combi=gauss2.Transpose[{0.5,1.6}];
In[7]:= RegularHisto[combi,-3,3,20]
In[8]:= CentralLimit[combi,0,1,1]
```

Here is an illustration of the uniform distribution on the unit ball in dimensions 2, 3, and 4. The norm of a random point in the unit ball in dimension n tends to 1 as n tends to infinity.

```
In[1]:= <<UVW`ContSamp`
In[2]:= <<UVW`DataRep`
In[3]:= ball2=RSUnitBall[2,1000];
In[4]:= SamplePlot2D[ball2,AspectRatio->1]
In[5]:= norms=Sqrt[Table[ball2[[i]].ball2[[i]],{i,1000}]];
In[6]:= RegularHisto[norms,0,1,10]
In[7]:= ball3=RSUnitBall[3,1000];
```

```
In[8]:= Show[Graphics3D[Table[Point[ball3[[i]]],{i,1000}]]]
In[9]:= norms=Sqrt[Table[ball3[[i]].ball3[[i]],{i,1000}]];
In[10]:= RegularHisto[norms,0,1,10]
In[11]:= ball4=RSUnitBall[4,1000];
In[12]:= norms=Sqrt[Table[ball4[[i]].ball4[[i]],{i,1000}]];
In[13]:= RegularHisto[norms,0,1,10]
```

3. UVW‘DataRep‘ — data representations

Histogram[listofdata,listofbounds] represents the histogram of the data contained in listofdata. The bounds of the bins are listofbounds.

RegularHisto[listofdata,xmin,xmax,nx] represents the histogram of the data contained in listofdata. There are nx regular classes between xmin and xmax.

SamplePlot2D[listof2Ddata] plots in the plane the points whose coordinates are read in listof2Ddata.

Histogram2D[listof2Ddata,xmin,xmax,nx,ymin,ymax,ny] represents a histogram in three dimensions for the two-dimensional data contained in listof2Ddata. The classes are regular. There are nx classes on the x-axis between xmin and xmax, and ny classes on the y-axis between ymin and ymax.

LargeNumbers[listofdata] plots the partial means of the data contained in listofdata.

CentralLimit[listofdata, mu, sigma, n] takes consecutive groups of n data in listofdata. The sum of each group is centered by n*mu then divided by Sqrt[n]*sigma. The results are represented on a regular histogram with 20 classes.

Examples:

Here is an illustration of the Law of Large Numbers and the Central Limit Theorem applied to the uniform distribution on [0, 1], and to the exponential distribution. The Central Limit Theorem states that the centered and reduced variables associated to the sum of n independent random variables is approximately normally distributed, for n large enough. For the uniform distribution, it is true with a reasonable precision, for n as low as 6. The exponential distribution, being more skewed, requires a much higher value of n.

```
In[1]:= <<UVW'DataRep'
In[2]:= uni=Table[Random[],{2000}];
In[3]:= Histogram[uni,{0.,0.1,0.3,0.6,0.8,1.}]
In[4]:= RegularHisto[uni,0,1,10]
In[5]:= LargeNumbers[uni]
In[6]:= CentralLimit[uni,0.5,Sqrt[1./12],6]
In[7]:= exp=Table[-Log[Random[]],{2000}];
In[8]:= RegularHisto[exp,0,5,10]
In[9]:= LargeNumbers[exp]
In[10]:= CentralLimit[exp,1.,1.,10]
```

The Law of Large Numbers is false if the distribution of the random variables in the independent sequence does not have an expectation. Here is what happens with the Cauchy distribution.

```
In[1]:= <<UVW'DataRep'
In[2]:= <<Statistics'ContinuousDistributions'
In[3]:= samp=Table[Random[CauchyDistribution[0.,1.]],{2000}];
In[4]:= LargeNumbers[samp]
In[5]:= CentralLimit[samp,0,1,10]
```

Here are several random samples in the plane, visualized through `SamplePlot2D` and `Histogram2D`.

```
In[1]:= <<UVW'DataRep'
In[2]:= uni=Table[{Random[],Random[]},{10000}];
In[3]:= SamplePlot2D[uni,Frame->True,AspectRatio->1]
In[4]:= Histogram2D[uni,0,1,8,0,1,8]
In[5]:= tri=Table[{Min[Random[],Random[]],Min[Random[],Random[]]}, {10000}];
In[6]:= SamplePlot2D[tri,Frame->True,AspectRatio->1]
In[7]:= Histogram2D[tri,0,1,8,0,1,8]
In[8]:= x=Transpose[uni][[1]];
In[9]:= y=Transpose[uni][[2]];
In[10]:= uni2=Transpose[{x+y,x-y}];
In[11]:= SamplePlot2D[tri,AspectRatio->1]
In[12]:= tri2=Transpose[{x/(x+y),2*x*y/(x+y)}];
In[13]:= SamplePlot2D[tri2,AspectRatio->1]
In[14]:= Histogram2D[tri2,0,1,8,0,1,8]
```

4. UVW‘DiscSamp‘ — simulations of discrete distributions

Distribution[listofdata] returns a list of pairs. The first element of each pair is the value appearing in listofdata, the second one is the frequency of that value in listofdata.

RSPermutation[list,n] returns a list of n random permutations of list.

RSExtract[list,k,n] returns a list of n lists of length k, extracted at random from list.

RSDiscreteDistribution[dist,n] returns a random sample of size n of the distribution dist, given under the form of a list of non-negative reals. The list dist is first divided by its sum, then interpreted as a probability distribution on the set 1,...,Length[dist].

Examples:

Random samples of classical discrete distributions can be simulated using the standard function Random, together with the distributions defined in the package Statistics‘DiscreteDistributions‘.

```
In[1]:= Table[Random[Integer],{100}]
In[2]:= Table[Random[Integer,{2,4}],{100}]
In[3]:= <<Statistics‘DiscreteDistributions‘
In[4]:= Table[Random[BinomialDistribution[3,0.5]],{100}]
In[5]:= Table[Random[GeometricDistribution[0.5]],{100}]
In[6]:= Table[Random[PoissonDistribution[1.]],{100}]
```

To play Lotto, one has to extract a sample of 6 numbers from the set $\{1, \dots, 49\}$. It can be done with RSPermutation or RSExtract. One can use also the functions RandomPermutation or RandomSubset of the standard package DiscreteMath‘Combinatorica‘. That package contains several other functions that return random discrete objects, such as graphs, tableaus, trees, heaps, etc.

```
In[1]:= <<UVW‘DiscSamp‘
In[2]:= fournine=Range[49];
In[3]:= perm=RSPermutation[fournine,10];
In[4]:= lotto1=Transpose[Take[Transpose[perm],6]];
In[5]:= MatrixForm[%]
In[6]:= lotto2=RSExtract[fournine,6,100];
In[7]:= Distribution[Flatten[lotto2]]
In[8]:= dist=Transpose[%][[2]]]
In[9]:= <<Graphics‘Graphics‘
In[10]:= BarChart[dist]
In[11]:= PieChart[dist]
```

The following example uses RSDiscreteDistribution to illustrate the Law of Large Numbers.

```
In[1]:= <<UVW`DiscSamp`
In[2]:= dist={.2,.3,.5}
In[3]:= sample1=RSDiscreteDistribution[dist,100];
In[4]:= sample2=RSDiscreteDistribution[dist,1000];
In[5]:= sample3=RSDiscreteDistribution[dist,10000];
In[6]:= Distribution[sample1]
In[7]:= Distribution[sample2]
In[8]:= Distribution[sample3]
```

5. UVW‘DynSyst‘ — dynamical systems

`Mean[list]` returns the arithmetic mean of `list`.

`Variance[list]` returns the maximum likelihood estimate of the variance of `list`.

`CovarianceFunction[list,N]` returns in a list the values cov(i) for i ranging from 0 to `N` . The value cov(i) is the covariance of `list` with its i-th shift.

`AsymptoticVariance[list,N]` returns the sum of the values cov(i) for i ranging from 0 to `N`. The value cov(i) is the covariance of list with its i-th shift.

`CorrelationDimension[list,step,nstep]` computes the values of $C(r)$, r being an integer multiple of `step`, up to `nstep` values. Then the values log[$C(r)$] as a function of r are plotted, and the linear regression coefficients are computed. The slope of the regression line and the correlation coefficient are printed.

`TentFunction[a,x]` returns `1-a*Abs[x-1+1/a]`.

`LogisticFunction[a,x]` returns `ax(1-x)`.

`IterateAPhi[matrixA,functionPhi,vectorX0,n]` computes and returns in a list the images by the `functionPhi` of the `vectorX0` and its products by the successive powers of the `matrixA`. The coordinates of these vectors are reduced to their decimal parts.

Examples:

Can successive iterates of the logistic function be taken as random reals in the interval [0, 1]?

```
In[1]:= <<UVW`DynSyst`
In[2]:= <<UVW`DataRep`
In[3]:= f[x_]:=LogisticFunction[4.,x]
In[4]:= samp=NestList[f,0.23,1000];
In[5]:= RegularHisto[samp,0,1,10]
In[6]:= samp2=Partition[samp,2];
In[7]:= SamplePlot2D[samp2]
In[8]:= mu=Mean[samp]
In[9]:= sigma=Sqrt[AsymptoticVariance[samp,10]]
In[10]:= CentralLimit[samp,mu,sigma,10]
```

The function IterateAPhi generates better samples.

```
In[1]:= <<UVW`DynSyst`
In[2]:= <<UVW`DataRepl
In[3]:= mat=1,2,1,1
In[4]:= phi[ell_]:=Apply[Plus,ell]
In[5]:= vec={0.23,0.12}
In[6]:= samp=IterateAPhi[mat,phi,vec,1000];
In[7]:= RegularHisto[samp,0,1,10]
In[8]:= samp2=Partition[samp,2];
In[9]:= SamplePlot2D[samp2]
In[10]:= mu=Mean[samp]
In[11]:= sigma=Sqrt[AsymptoticVariance[samp]]
In[12]:= CentralLimit[samp,mu,sigma,10]
In[13]:= CorrelationDimension[samp,0.05,5]
```

6. UVW‘Fractals‘ — deterministic and random Von Koch curves

All fractal curve approximations are treated by this package as lists of segments, a segment being a list of two points, each of them being a list of two coordinates in the plane. The package contains a few examples of simple lists of segments (patterns). By replacing each of the segments of a list by a pattern, a new list is obtained, which can be used as another pattern, or a new starting point. This package contains the replacement and iteration functions that permit the production of many different curves, and also their representation. We start with examples of basic patterns.

Triangle returns an equilateral triangle as a list of three segments in the plane.

Star returns a star as a list of six segments in the plane.

Island returns a notch and a triangle as a list of eight segments in the plane.

Battlement returns a battlement as a list of eight segments in the plane.

Hat[sharpness] returns a list of four segments in the plane.

Wy[angle1,length1,angle2,length2] returns a Y-shaped list of four segments in the plane. angle1,length1,angle2 and length2 are the parameters of the two branches.

RSWy[n] returns a random sample of size n of outputs of the function Wy.

RSTruncs[n] returns a random sample of vertical segments in the plane.

The following commands replace segments by patterns:

`TransformSegment[segment,pattern]` returns a list of segments obtained from `pattern` by moving it in the plane so as to make its ends coincide with those of the initial `segment`.

`IteratePattern[listofsegments,pattern,n]` applies the command `TransformSegment` to each of the segments in `listofsegments`. Iterates `n` times.

`IterateRandomPattern[listofsegments,patterns,probas,n]` iterates `n` times the following operation: for each of the segments in `listofsegments`, one of the patterns in `patterns` is chosen randomly according to a probability read in `probas`, and the segment is transformed accordingly by `TransformSegment`.

The graphical representation can be obtained by means of the following command:

`DrawSegments[listofsegments]` plots the elements of `listofsegments` in the plane.

Examples:
The celebrated Von Koch curves are constructed by iteratively replacing each segment of a jagged line by a given pattern.

```
In[1]:= <<UVW'Fractals'
In[2]:= h=Hat[Sqrt[3]/2];
In[3]:= DrawSegments[h]
In[4]:= snowflake=IteratePattern[Triangle,h,4];
In[5]:= DrawSegments[snowflake,AspectRatio->1];
In[6]:= snow1=IteratePattern[Triangle,h,1];
In[7]:= snow2=IteratePattern[snow1,h,1];
In[8]:= snow3=IteratePattern[snow2,h,1];
In[9]:= snow4=IteratePattern[snow3,h,1];
In[10]:= g1=DrawSegments[snow1,AspectRatio->1]
In[11]:= g2=DrawSegments[snow2,AspectRatio->1]
In[12]:= g3=DrawSegments[snow3,AspectRatio->1]
In[13]:= g4=DrawSegments[snow4,AspectRatio->1]
In[14]:= picture={{g1,g2},{g3,g4}};
In[15]:= Show[GraphicsArray[picture]]
In[16]:= DrawSegments[Battlement];
In[17]:= IteratePattern[Battlement,Battlement,3];
In[18]:= DrawSegments[%]
In[19]:= y=Wy[Pi/6,1,Pi/6,1];
In[20]:= DrawSegments[y,AspectRatio->1]
In[21]:= IteratePattern[Star,y,4];
In[22]:= DrawSegments[%,AspectRatio->1]
In[23]:= DrawSegments[Island];
In[24]:= IteratePattern[Island,Island,3];
In[25]:= DrawSegments[%]
```

Random fractals are much better models for real life.

```
In[1]:= <<UVW'Fractals'
In[2]:= h=Hat[Sqrt[3]/2];
In[3]:= base={{{0.,0.},{1.,0.}}};
```

```
In[4]:= IterateRandomPatterns[base,{h,Island},{0.5,0.5},3];
In[5]:= DrawSegments[%]
In[6]:= IterateRandomPatterns[base,{h,Island},{0.5,0.5},3];
In[7]:= DrawSegments[%]
In[8]:= weights=Table[0.25,{4}];
In[9]:= madhatter=Table[Hat[2*Random[]],{4}];
In[10]:= IterateRandomPatterns[base,madhatter,weights,4];
In[11]:= DrawSegments[%]
In[12]:= madhatter=Table[Hat[2*Random[]],{4}];
In[13]:= IterateRandomPatterns[base,madhatter,weights,4];
In[14]:= DrawSegments[%]
In[15]:= ry=Wy[(Pi/2)*Random[],Random[],(Pi/2)*Random[],Random[]];
In[16]:= tree=IteratePattern[ry,ry,4];
In[17]:= DrawSegments[tree]
In[18]:= branches=RSWy[4];
In[19]:= forest=IterateRandomPattern[RSTruncs[10],branches,weights,3];
In[20]:= DrawSegments[forest]
```

7. UVW'Interact' — simulation of interacting particle systems in the plane

This package contains functions, examples of configurations, and transition rates needed for simulation of interacting particle systems on the two-dimensional square lattice. The lattice is finite and wrapping around itself, i.e., periodic boundary conditions are assumed. Each site on the lattice can be in one of the two possible states, 0 and 1. The flip rates from one state to the other at each site depend on the state at this site itself as well as on the number of neighboring sites that are in state 1.

We begin with examples of basic configurations. All configurations are two-dimensional arrays of 0s and 1s. The element `config[[x,y]]` is interpreted as the state of site `(x,y)`.

`Checkerboard` returns a 40-by-40 configuration of 0s and 1s arranged in 10 by 10 squares.

`Diagonals` returns a 40-by-40 configuration of 0's and 1's arranged in diagonal stripes.

`RConfig[p,width,height]` returns a `width`-by-`height` array of independent random 0s and 1s, 1 being chosen with probability `p`.

Here are some examples of rates. All lists of rates are returned as a 2-by-5 list of reals. In such a list, `rate[[i,j]]` represents the rate at which the configuration will change at a site in state i (0 or 1) having j (from 0 to 4) neighbors in state 1.

`Uniform[lambda,mu]` are the rates corresponding to the case where the configuration changes from 0 to 1 at rate `lambda` and from 1 to 0 at rate `mu`, independently from the number of neighbors in state 1.

`Ising[Alpha,Beta]` returns the rates corresponding to the symmetric Stochastic Ising Model, admitting a Gibbs measure as a reversible state. `Alpha` is the potential of a site alone, `Beta` is the potential of a pair of neighboring sites.

`Contact[lambda]` returns the rates of the contact process. The transition rate from 1 to 0 (curing) is constant. The transition rate from 0 to 1 (infecting) is proportional to the number of neighbors at 1.

`Voter` returns the rates of the Voter Model, where the transition rate from 0 to 1 is proportional to the number of neighbors in state 1 and vice versa.

The following commands provide different treatments of a configuration:

`Cyclic[n,bound]` returns `bound` if `n` is 0, 1 if `n` is `bound+1`, and `n` in any other case. (Periodic boundary conditions are assumed for all configurations.)"

`RepartConfig[config]` returns a 2-by-5 list of integers. Its element `[[i,j]]` is the number of sites in state `i` (0 or 1) having `j` (from 0 to 4) neighbors in state 1.

`Evolution[initialconfig,rates,niter]` simulates the evolution of a configuration according to the spin system corresponding to `rates`. `niter` iterations are performed. One iteration consists of picking up a site at random and deciding to flip its state or not.

The last command of this package provides for the graphical representation of the results.

`DrawConfig[config,opts]` plots `config` as a rectangular array of black and white squares.

Examples:
Here is the evolution of the voter model on a square grid of 40 × 40 after 5000 and 10000 iterations.

```
In[1]:= <<UVW‘Interact‘
In[2]:= whims=RConfig[0.5,40,40];
In[3]:= g1=DrawConfig[whims]
In[4]:= opinions=Evolution[whims,Voter,5000];
In[5]:= g2=DrawConfig[opinions]
In[6]:= opinions=Evolution[opinions,Voter,5000];
In[7]:= g3=DrawConfig[opinions]
In[8]:= Show[GraphicsArray[{g1,g2,g3}]]
In[9]:= RepartConfig[whims]
In[10]:= RepartConfig[opinions]
```

The contact process is a model of epidemics. If the parameter λ is smaller than its critical value, the population gets healthier and healthier. If it is larger, the epidemics lasts forever.

```
In[1]:=<<UVW`Interact`
In[2]:= outburst=RConfig[0.05,20,20];
In[3]:= DrawConfig[outburst]
In[4]:= epidemy=Evolution[outburst,Contact[2],2000];
In[5]:= DrawConfig[epidemy]
In[6]:= cured=Evolution[epidemy,Contact[0.5],5000];
In[7]:= DrawConfig[cured]
In[8]:= RepartConfig[outburst]
In[9]:= RepartConfig[epidemy]
In[10]:= RepartConfig[cured]
```

Interacting particle systems are also used in image analysis with pixels interpreted as lattice sites that can be in different states. Here is a simple example of an image, first blurred by a random noise, then randomly cleaned up by two spin systems.

```
In[1]:= <<UVW`Interact`
In[2]:= check=Checkerboard;
In[3]:= DrawConfig[check]
In[4]:= noisy=Evolution[check,Uniform[1,1],100]
In[5]:= DrawConfig[noisy]
In[6]:= soaprates={{0,0,0,1,1},{1,1,0,0,0}};
In[7]:= clean1=Evolution[noisy,soaprates,1000];
In[8]:= DrawConfig[clean1]
In[9]:= clean2=Evolution[noisy,Ising[0,1],1000];
In[10]:= DrawConfig[clean2]
In[11]:= RepartConfig[check]
In[12]:= RepartConfig[noisy]
In[13]:= RepartConfig[clean1]
In[14]:= RepartConfig[clean2]
```

8. UVW‘Lorenz‘ — Lorenz attractor

Lorenz[s,r,b] computes and plots an approximate solution of the Lorenz equations with parameters (s,r,b), by the Runge-Kutta method.

LorenzArray[matrix] plots an array of solutions of the Lorenz equations for the values of the parameters contained in matrix.

Examples:

The Lorenz function is rather slow. On a faster computer with a lot of memory it can be coupled with Animate or ShowAnimation.

```
In[1]:= <<UVW`Lorenz`
In[2]:= Lorenz[3,26.5,1]
In[3]:= para={{{3,26.5,1},{3,25,1}},{{4,26.5,1},{4,25,1}}};
In[4]:= LorenzArray[para]
```

Reference: J.C. Culioli, *Introduction à Mathematica, Ellipses*, Paris (1991).

9. UVW‘PseuGene‘ — congruential and midsquare generators

`CongruGenerator[seed,a,c,m,n]` returns a list of the `n` first iterates of the congruential generator `x(n+1) = a x(n) + c modulo m`. `seed` is the first element `x(0)`. If `seed` is an integer, the result will be a list of integers between 0 and `m` . If `seed` is real, all the results will be divided by `m` to return a list of reals between 0 and 1 .

`MidsquareGenerator[seed,n]` returns a list of the first iterates of the midsquare generator, starting with `seed` (a four-digit integer).

`CongruentialLoop[seed,a,c,m]` returns in a list the loop of the congruential generator `x(n+1) = a x(n) + c modulo m`, starting with `x(0) = seed` .

`MidsquareLoop[seed]` returns in a list the loop of the midsquare generator, starting with `seed`.

Examples:
The midsquare generator is not very good.

```
In[1]:= <<UVW‘PseuGene‘
In[2]:= MidsquareGenerator[1245,100]
In[3]:= MidsquareGenerator[1246,100]
In[4]:= MidsquareGenerator[1247,100]
In[5]:= MidsquareLoop[4578]
In[6]:= MidsquareLoop[9854]
```

Some congruential generators can be reasonably good, others very disappointing.

```
In[1]:= <<UVW‘PseuGene‘
In[2]:= samp=CongruGenerator[0.23,181,0,16384,2000];
In[3]:= <<UVW‘DataRep‘
In[4]:= RegularHisto[samp,0,1,10]
In[5]:= LargeNumbers[samp]
In[6]:= CentralLimit[samp,0.5,Sqrt[1./12],6]
In[7]:= samp2=Partition[samp,2];
In[8]:= SamplePlot2D[samp2]
In[9]:= Length[CongruentialLoop[10,181,0,16384]]
In[10]:= Length[CongruentialLoop[1,181,0,16384]]
In[11]:= CongruGenerator[10,181,0,16381,20]
In[12]:= CongruGenerator[1,181,0,16381,20]
```

10. UVW‘RandWalk‘ — Random walks and random vector fields

`RandomWalk[ListofVelocities,Deltat]` represents the trajectory of a point in a square. `ListofVelocities` is a list of two-dimensional vectors, interpreted as consecutive speeds for the point. The point starts from the center with the first speed vector of the list. It changes its speed vector for the next one in the list at each integer multiple of `deltat`.

`VectorField[arrayof2Dvelocities]` represents graphically by segments on a grid the values of a discrete vector field. The `arrayof2Dvelocities` is a list with three levels. The first two correspond to the coordinates `(i,j)` of a point on the grid. The last level corresponds to the two coordinates of the vector attached to point `(i,j)`: `Vx(i,j), Vy(i,j)`. The function represents the grid and the vector attached to each point.

`VectorFieldTrajectory[arrayof2Dvelocities, deltat, tmax]` represents first a vector field on a grid by calling `VectorField[arrayof2Dvelocities]`. Then it draws the trajectory of a point starting at the center of the grid. At each integer multiple of `deltat`, the velocity vector of the point is changed for that of the vector field at the closest point on the grid. The trajectory is followed up to time `tmax`. The boundary conditions are periodic.

Examples:
Here is a representation of the discretized Brownian motion.

```
In[1]:= <<UVW‘RandWalk‘
In[2]:= <<UVW‘ContSamp‘
In[3]:= vel=RSNormal2D[1,1,0,1000];
In[4]:= RandomWalk[vel,0.02];
In[5]:= vel=RSNormal2D[1,1,0,1000];
In[6]:= RandomWalk[vel,0.02];
```

Here are two trajectories, one in a deterministic vector field, the other in a random one.

```
In[1]:= <<UVW‘RandWalk‘
In[2]:= field1=N[Table[{Cos[(i+j)/5Pi],Sin[(i+j)/5Pi]},{i,19},{j,19}];
In[3]:= VectorField[field1]
In[4]:= VectorFieldTrajectory[field1,0.1,10]
In[5]:= field2=Table[{Random[Real,{-1,1}],Random[Real,{-1,1}]},{5},{5}]
In[6]:= VectorFieldTrajectory[field2,0.5,100]
```

11 UVW‘StoGho‘ — stochastic ghost

`StoGho[width, mood]` portraits a gha(us)stly ghost by the name of StoGho who is known to roam around some old European castles. His shape depends on `width` and `mood`, modulo 2.

`GalleryOfPortraits[matrix]` draws an array of portraits of our favorite spook.

Examples:

The body, mouth, eyes, and pupils can easily be reparametrized in order to change the aspect of the ghost. StoGho can be made into a movie star if function `StoGho` is coupled with `Animate` or `ShowAnimation`.

```
In[1]:= <<UVW'StoGho'
In[2]:= StoGho[Random[Real,{0,2}],Random[Real,{0,2}]]
In[3]:= faces=Table[{x,y},{x,0.,1.8,0.6},{y,0.,1.8,0.6}];
In[4]:= GalleryOfPortraits[faces]
```

12. UVW'TimeRep' — queues and other time-dependent random processes

`Queue[interarrivals,services]` represents graphically, as a function of time, the evolution of the number of customers in a queue with one server. The times between consecutive arrivals are read in the first list, the service times in the second one.

`CumulatedTimes[listoftimes]` represents graphically the function of time defined as follows: Starting from 0, it is incremented by one at dates separated by the times read in `listoftimes`.

`Geiger[listoftimes]` plots on a line the dates separated by the durations read in `listoftimes`.

Examples:

Here is an illustration of the Poisson process with different intensities.

```
In[1]:= <<UVW'TimeRep'
In[2]:= <<Statistics'ContinuousDistributions'
In[3]:= times=Table[Random[ExponentialDistribution[1.]],{100}];
In[4]:= Geiger[times]
In[5]:= CumulatedTimes[times]
In[6]:= times=Table[Random[ExponentialDistribution[1.]],{500}];
In[7]:= CumulatedTimes[times]
In[8]:= times=Table[Random[ExponentialDistribution[2.]],{100}];
In[9]:= CumulatedTimes[times]
In[10]:= times=Table[Random[ExponentialDistribution[0.5]],{100}];
In[11]:= CumulatedTimes[times]
```

The so-called M/M/1 single server queue has exponentially distributed interarrival and service times. It may be in equilibrium or saturated according to the values of the mean interarrival and service times.

```
In[1]:= <<UVW'TimeRep'
In[2]:= <<Statistics'ContinuousDistributions'
In[3]:= arr=Table[Random[ExponentialDistribution[1.]],{200}];
```

```
In[4]:= ser=Table[Random[ExponentialDistribution[1.1]],{200}];
In[5]:= Queue[arr,ser]
In[6]:= ser=Table[Random[ExponentialDistribution[0.9]],{200}];
```

Here is the same illustration with the D/M/1 queue (constant interarrival times).

```
In[1]:= <<UVW`TimeRep`
In[2]:= <<Statistics`ContinuousDistributions`
In[3]:= arr=Table[1.,{200}];
In[4]:= ser=Table[Random[ExponentialDistribution[1.1]],{200}];
In[5]:= Queue[arr,ser]
In[6]:= ser=Table[Random[ExponentialDistribution[0.9]],{200}];
```

13. UVW`ZeroOne` — lists of zeros and ones

RSZeroOne[p,n] returns a random sample of n zeros and ones. One is chosen with probability p.

Binary[functionf,listofzeroones] computes the real number in [0,1] which in the binary representation is given by the listofzeroones. Then its image by functionf (acting from [0,1] into **R**) is computed. The fractional part of it is returned in the binary form as a new list of zeroones.

PlotZeroOne[listofzeroones] represents graphically a list of zeros and ones as black and white squares on a grey background.

AnimateShift[listofzeroones] forms a list of zeros and ones three times as long as the initial list, by adding first a list of same length of random digits, then copying the initial list at the end. Then the successive shifts are animated as arrays of black and white squares.

ActualLength[listofzeroones] returns the length of the list obtained when all zeros before the first one are dropped in the listofzeroones.

Weight[listofzeroones] returns the number of ones in the listofzeroones.

WeightedAlphabeticalOrder[listofzeroones] computes the rank of the given list of zeros and ones among lists of same actual length, when they are ranked according to increasing weights and alphabetical order for lists of same weight. That rank is returned in base 2 as another list of zeros and ones.

BinaryNumbers[n] returns a list of all the 2^n lists of binary digits with length n.

Entropy[t] returns -tLog[2,t]-(1-t)Log[2,1-t]"

Examples:

Here is an illustration of the Kolmogorov complexity of a random sequence of zeros and ones.

```
In[1]:= <<UVW'ZeroOne'
In[2]:= samp=RSZeroOne[0.5,100];
In[3]:= len=ActualLength[samp]
In[4]:= wei=Weight[samp]
In[5]:= samp1=WeightedAlphabeticalOrder[samp];
In[6]:= len1=ActualLength[samp1]
In[7]:= wei1=Weight[samp1]
In[8]:= samp2=Binary[Sin,samp];
In[9]:= len2=ActualLength[samp2]
In[10]:= wei2=Weight[samp2]
In[11]:= g=PlotZeroOne[samp]
In[12]:= g1=PlotZeroOne[samp1]
In[13]:= g2=PlotZeroOne[samp2]
In[14]:= Show[GraphicsArray[{g,g1,g2}]]
```

The following session illustrates the shifts of a sequence of zeros and ones. The animation may not work on all platforms.

```
In[1]:= <<UVW'ZeroOne'
In[2]:= letter={0,0,0,0,0, 0,1,1,1,1, 0,0,0,0,1, 0,1,1,1,1,
                 0,0,0,0,0};
In[3]:= PlotZeroOne[letter]
In[4]:= AnimateShift[letter]
```

Appendix F

Tables

The reader of this book is expected to use *Mathematica* (or other similar software, like *Maple*) to obtain numerical values of cumulative distribution functions and quantiles needed to do experiments, exercises and projects. However, when the power is down and/or your hard disk crashed, you still may have to resort to the old fashioned printed tables provided on the following pages. Besides, browsing through printed tables gives a good insight into the structure of various probability distributions.

M. Denker, W. Woyczynski, *Introductory Statistics and Random Phenomena*,
Modern Birkhäuser Classics, DOI 10.1007/978-3-319-66152-0

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9773	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9983	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

$n \backslash \alpha$	0.1000	0.0500	0.0250	0.0100	0.0050	0.0010	0.0005
1	3.078	6.314	12.706	31.821	63.657	318.317	636.61
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.500	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.813	2.228	2.764	3.169	4.144	4.587
11	1.364	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.141
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.584	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.553	2.879	3.610	3.922
19	1.328	1.729	2.093	2.540	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.849
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.320	1.714	2.069	2.500	2.807	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.059	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.312	1.701	2.049	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.311	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

$n\backslash\alpha$	0.9950	0.9900	0.9750	0.9500	0.9000	0.1000	0.0500	0.0250	0.0100	0.0050
1	0.000	0.000	0.001	0.004	0.016	2.706	3.843	5.025	6.637	7.882
2	0.010	0.020	0.051	0.103	0.211	4.605	5.992	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.344	12.937
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.085	16.748
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.440	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.012	18.474	20.276
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.534	20.090	21.954
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.022	21.665	23.587
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.724	26.755
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.041	19.812	22.362	24.735	27.687	29.817
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.600	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.577	32.799
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.407	7.564	8.682	10.085	24.769	27.587	30.190	33.408	35.716
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.843	7.632	8.906	10.117	11.651	27.203	30.143	32.852	36.190	38.580
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.033	8.897	10.283	11.591	13.240	29.615	32.670	35.479	38.930	41.399
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.260	10.195	11.688	13.090	14.848	32.007	35.172	38.075	41.637	44.179
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.558
25	10.519	11.523	13.120	14.611	16.473	34.381	37.652	40.646	44.313	46.925
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.807	12.878	14.573	16.151	18.114	36.741	40.113	43.194	46.962	49.642
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.120	14.256	16.147	17.708	19.768	39.087	42.557	45.772	49.586	52.333
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
31	14.457	15.655	17.538	19.280	21.433	41.422	44.985	48.231	52.190	55.000
32	15.134	16.362	18.291	20.072	22.271	42.585	46.194	49.480	53.486	56.328
33	15.814	17.073	19.046	20.866	23.110	43.745	47.400	50.724	54.774	57.646
34	16.501	17.789	19.806	21.664	23.952	44.903	48.602	51.966	56.061	58.964
35	17.191	18.508	20.569	22.465	24.796	46.059	49.802	53.203	57.340	60.272
36	17.887	19.233	21.336	23.269	25.643	47.212	50.998	54.437	58.619	61.581
37	18.584	19.960	22.105	24.075	26.492	48.363	52.192	55.667	59.891	62.880
38	19.289	20.691	22.878	24.884	27.343	49.513	53.384	56.896	61.162	64.181
39	19.994	21.425	23.654	25.695	28.196	50.660	54.572	58.119	62.426	65.473
40	20.706	22.164	24.433	26.509	29.050	51.805	55.758	59.342	63.691	66.766

$m\backslash n$	1	2	3	4	5	10	15	20	40	120	999
1	161.40	199.50	215.70	224.60	230.20	241.90	245.90	248.00	251.10	253.30	254.30
2	18.51	19.00	19.16	19.25	19.30	19.40	19.43	19.45	19.47	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.79	8.70	8.66	8.59	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	5.96	5.86	5.80	5.72	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.74	4.62	4.56	4.46	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.06	3.94	3.87	3.77	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.64	3.51	3.44	3.34	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.35	3.22	3.15	3.04	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.14	3.01	2.94	2.83	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	2.98	2.85	2.77	2.66	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	2.85	2.72	2.65	2.53	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	2.75	2.62	2.54	2.43	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.67	2.53	2.46	2.34	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.60	2.46	2.39	2.27	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.54	2.40	2.33	2.20	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.49	2.35	2.28	2.15	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.45	2.31	2.23	2.10	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.41	2.27	2.19	2.06	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.38	2.23	2.16	2.03	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.35	2.20	2.12	1.99	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.32	2.18	2.10	1.96	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.30	2.15	2.07	1.94	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.27	2.13	2.05	1.91	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.25	2.11	2.03	1.89	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.24	2.09	2.01	1.87	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.22	2.07	1.99	1.85	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.20	2.06	1.97	1.84	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.19	2.04	1.96	1.82	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.18	2.03	1.94	1.81	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.16	2.01	1.93	1.79	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.08	1.92	1.84	1.69	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	1.99	1.84	1.75	1.59	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	1.91	1.75	1.66	1.50	1.35	1.25
999	3.84	3.00	2.60	2.37	2.21	1.83	1.67	1.57	1.39	1.22	1.00

Index

© Springer International Publishing AG 2017
M. Denker, W. Woyczynski, *Introductory Statistics and Random Phenomena*,
Modern Birkhäuser Classics, DOI 10.1007/978-3-319-66152-0