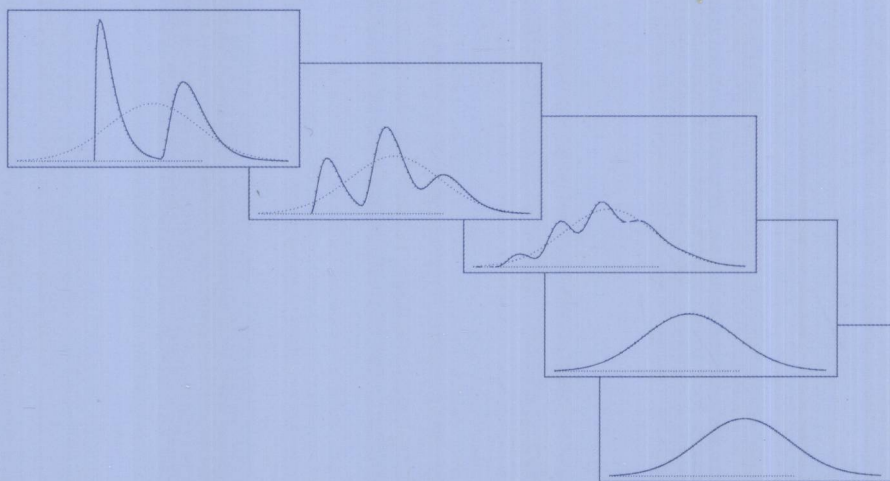


# A Modern Introduction to Probability and Statistics

Understanding Why  
and How



F.M. Dekking  
C. Kraaikamp  
H.P. Lopuhaä  
L.E. Meester



Springer

021  
M689

F.M. Dekking C. Kraaikamp  
H.P. Lopuhaä L.E. Meester

# A Modern Introduction to Probability and Statistics

Understanding Why and How

With 120 Figures



E200501456



Springer

Frederik Michel Dekking  
Cornelis Kraaikamp  
Hendrik Paul Lopuszanski  
Ludolf Erwin Meester  
Delft Institute of Applied Mathematics  
Delft University of Technology  
Mekelweg 4  
2628 CD Delft  
The Netherlands

Whilst we have made considerable efforts to contact all holders of copyright material contained in this book, we may have failed to locate some of them. Should holders wish to contact the Publisher, we will be happy to come to some arrangement with them.

British Library Cataloguing in Publication Data  
A modern introduction to probability and statistics. —  
(Springer texts in statistics)  
1. Probabilities 2. Mathematical statistics  
I. Dekking, F. M.  
519.2  
ISBN 1852338962

Library of Congress Cataloging-in-Publication Data  
A modern introduction to probability and statistics : understanding why and how / F.M. Dekking ... [et al.].  
p. cm. — (Springer texts in statistics)  
Includes bibliographical references and index.  
ISBN 1-85233-896-2  
1. Probabilities—Textbooks. 2. Mathematical statistics—Textbooks. I. Dekking, F.M. II. Series.  
QA273.M645 2005  
519.2—dc22  
2004057700

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms of licences issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms should be sent to the publishers.

ISBN-10: 1-85233-896-2  
ISBN-13: 978-1-85233-896-1

Springer Science+Business Media  
springeronline.com

© Springer-Verlag London Limited 2005

The use of registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant laws and regulations and therefore free for general use.

The publisher makes no representation, express or implied, with regard to the accuracy of the information contained in this book and cannot accept any legal responsibility or liability for any errors or omissions that may be made.

Printed in the United States of America  
12/3830/543210 Printed on acid-free paper SPIN 10943403

# *Springer Texts in Statistics*

*Advisors:*

George Casella   Stephen Fienberg   Ingram Olkin

## Springer Texts in Statistics

---

- Alfred*: Elements of Statistics for the Life and Social Sciences  
*Berger*: Introduction to Probability and Stochastic Processes, Second Edition  
*Bilodeau and Brenner*: Theory of Multivariate Statistics  
*Blom*: Probability and Statistics: Theory and Applications  
*Brockwell and Davis*: An Introduction to Time Series and Forecasting  
*Carmona*: Statistical Analysis of Financial Data in S-Plus  
*Chow and Teicher*: Probability Theory:  
Independence, Interchangeability,  
Martingales, Third Edition  
*Christensen*: Advanced Linear Modeling: Multivariate, Times Series, and  
Spatial Data; Nonparametric Regression and Response Surface  
Maximization, Second Edition  
*Christensen*: Log-Linear Models and Logistic Regression, Second Edition  
*Christensen*: Plane Answers to Complex Questions: The Theory of Linear  
Models, Second Edition  
*Creighton*: A First Course in Probability Models and Statistical Inference  
*Davis*: Statistical Methods for the Analysis of Repeated Measurements  
*Dean and Voss*: Design and Analysis of Experiments  
*Dekking et al*: A Modern Introduction to Probability and Statistics:  
Understanding Why and How  
*du Toit, Steyn, and Stumpf*: Graphical Exploratory Data Analysis  
*Durrett*: Essential of Stochastic Processes  
*Edwards*: Introduction to Graphical Modeling, Second Edition  
*Everitt*: An R and S-PLUS® Companion to Multivariate Analysis  
*Finkelstein and Levin*: Statistics for Lawyers  
*Flury*: A First Course in Multivariate Statistics  
*Heiberger and Holland*: Statistical Analysis and Data Display:  
An Intermediate Course with Examples in S-PLUS, R, and SAS  
*Jobson*: Applied Multivariate Data Analysis, Volume I: Regression and  
Experimental Design  
*Jobson*: Applied Multivariate Data Analysis, Volume II: Categorical and  
Multivariate Methods  
*Kalbfleisch*: Probability and Statistical Inference, Volume I: Probability,  
Second Edition  
*Kalbfleisch*: Probability and Statistical Inference, Volume II: Statistical  
Interference, Second Edition  
*Karr*: Probability  
*Keyfitz*: Applied Mathematical Demography, Second Edition  
*Kiefer*: Introduction to Statistical Inference  
*Kokoska and Nevison*: Statistical Tables and Formulae

(continued after index)

---

## Preface

Probability and statistics are fascinating subjects on the interface between mathematics and applied sciences that help us understand and solve practical problems. We believe that you, by learning how stochastic methods come about and why they work, will be able to understand the meaning of statistical statements as well as judge the quality of their content, when facing such problems on your own. Our philosophy is one of *how* and *why*: instead of just presenting stochastic methods as cookbook recipes, we prefer to explain the principles behind them.

In this book you will find the basics of probability theory and statistics. In addition, there are several topics that go somewhat beyond the basics but that ought to be present in an introductory course: simulation, the Poisson process, the law of large numbers, and the central limit theorem. Computers have brought many changes in statistics. In particular, the bootstrap has earned its place. It provides the possibility to derive confidence intervals and perform tests of hypotheses where traditional (normal approximation or large sample) methods are inappropriate. It is a modern useful tool one should learn about, we believe.

Examples and datasets in this book are mostly from real-life situations, at least that is what we looked for in illustrations of the material. Anybody who has inspected datasets with the purpose of using them as elementary examples knows that this is hard: on the one hand, you do not want to boldly state assumptions that are clearly not satisfied; on the other hand, long explanations concerning side issues distract from the main points. We hope that we found a good middle way.

A first course in calculus is needed as a prerequisite for this book. In addition to high-school algebra, some infinite series are used (exponential, geometric). Integration and differentiation are the most important skills, mainly concerning one variable (the exceptions, two dimensional integrals, are encountered in Chapters 9–11). Although the mathematics is kept to a minimum, we strived

to be mathematically correct throughout the book. With respect to probability and statistics the book is self-contained.

The book is aimed at undergraduate engineering students, and students from more business-oriented studies (who may gloss over some of the more mathematically oriented parts). At our own university we also use it for students in applied mathematics (where we put a little more emphasis on the math and add topics like combinatorics, conditional expectations, and generating functions). It is designed for a one-semester course: on average two hours in class per chapter, the first for a lecture, the second doing exercises. The material is also well-suited for self-study, as we know from experience.

We have divided attention about evenly between probability and statistics. The very first chapter is a sampler with differently flavored introductory examples, ranging from scientific success stories to a controversial puzzle. Topics that follow are elementary probability theory, simulation, joint distributions, the law of large numbers, the central limit theorem, statistical modeling (informal: why and how we can draw inference from data), data analysis, the bootstrap, estimation, simple linear regression, confidence intervals, and hypothesis testing. Instead of a few chapters with a long list of discrete and continuous distributions, with an enumeration of the important attributes of each, we introduce a few distributions when presenting the concepts and the others where they arise (more) naturally. A list of distributions and their characteristics is found in Appendix A.

With the exception of the first one, chapters in this book consist of three main parts. First, about four sections discussing new material, interspersed with a handful of so-called Quick exercises. Working these—two-or-three-minute—exercises should help to master the material and provide a break from reading to do something more active. On about two dozen occasions you will find indented paragraphs labeled *Remark*, where we felt the need to discuss more mathematical details or background material. These remarks can be skipped without loss of continuity; in most cases they require a bit more mathematical maturity. Whenever persons are introduced in examples we have determined their sex by looking at the chapter number and applying the rule “He is odd, she is even.” Solutions to the quick exercises are found in the second to last section of each chapter.

The last section of each chapter is devoted to exercises, on average thirteen per chapter. For about half of the exercises, answers are given in Appendix C, and for half of these, full solutions in Appendix D. Exercises with both a short answer and a full solution are marked with  $\boxplus$  and those with only a short answer are marked with  $\square$  (when more appropriate, for example, in “Show that ...” exercises, the short answer provides a hint to the key step). Typically, the section starts with some easy exercises and the order of the material in the chapter is more or less respected. More challenging exercises are found at the end.

## Springer Texts in Statistics (continued from page II)

---

- Kokoska and Nevison*: Statistical Tables and Formulae  
*Kulkarni*: Modeling, Analysis, Design, and Control of Stochastic Systems  
*Lange*: Applied Probability  
*Lange*: Optimization  
*Lehmann*: Elements of Large-Sample Theory  
*Lehmann and Romano*: Testing Statistical Hypotheses, Third Edition  
*Lehmann and Casella*: Theory of Point Estimation, Second Edition  
*Lindman*: Analysis of Variance in Experimental Design  
*Lindsey*: Applying Generalized Linear Models  
*Madansky*: Prescriptions for Working Statisticians  
*McPherson*: Applying and Interpreting Statistics: A Comprehensive Guide, Second Edition  
*Mueller*: Basic Principles of Structural Equation Modeling: An Introduction to LISREL and EQS  
*Nguyen and Rogers*: Fundamentals of Mathematical Statistics, Volume I: Probability for Statistics  
*Nguyen and Rogers*: Fundamentals of Mathematical Statistics, Volume II: Statistical Inference  
*Noether*: Introduction to Statistics: The Nonparametric Way  
*Nolan and Speed*: Stat Labs: Mathematical Statistics Through Applications  
*Peters*: Counting for Something: Statistical Principles and Personalities  
*Pfeiffer*: Probability for Applications  
*Pitman*: Probability  
*Rawlings, Pantula, and Dickey*: Applied Regression Analysis  
*Robert*: The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation, Second Edition  
*Robert and Casella*: Monte Carlo Statistical Methods, Second Edition  
*Rose and Smith*: Mathematical Statistics with *Mathematica*  
*Ruppert*: Statistics and Finance: An Introduction  
*Santner and Duffy*: The Statistical Analysis of Discrete Data  
*Saville and Wood*: Statistical Methods: The Geometric Approach  
*Sen and Srivastava*: Regressions Analysis: Theory, Methods, and Applications  
*Shao*: Mathematical Statistics, Second Edition  
*Shorack*: Probability for Statisticians  
*Shumway and Stoffe*: Time Series Analysis and Its Applications  
*Simonoff*: Analyzing Categorical Data  
*Terrell*: Mathematical Statistics: A Unified Introduction  
*Timm*: Applied Multivariate Analysis  
*Toutenburg*: Statistical Analysis of Designed Experiments, Second Edition  
*Wasserman*: All of Statistics: A Concise Course in Statistical Inference  
*Whittle*: Probability via Expectation, Fourth Edition  
*Zacks*: Introduction to Reliability Analysis: Probability Models and Statistical Methods



---

# Contents

<b>1</b>	<b>Why probability and statistics?</b>	<b>1</b>
1.1	Biometry: iris recognition	1
1.2	Killer football	3
1.3	Cars and goats: the Monty Hall dilemma	4
1.4	The space shuttle <i>Challenger</i>	5
1.5	Statistics versus intelligence agencies	7
1.6	The speed of light	9
<b>2</b>	<b>Outcomes, events, and probability</b>	<b>13</b>
2.1	Sample spaces	13
2.2	Events	14
2.3	Probability	16
2.4	Products of sample spaces	18
2.5	An infinite sample space	19
2.6	Solutions to the quick exercises	21
2.7	Exercises	21
<b>3</b>	<b>Conditional probability and independence</b>	<b>25</b>
3.1	Conditional probability	25
3.2	The multiplication rule	27
3.3	The law of total probability and Bayes' rule	30
3.4	Independence	32
3.5	Solutions to the quick exercises	35
3.6	Exercises	37

<b>4</b>	<b>Discrete random variables</b>	41
4.1	Random variables	41
4.2	The probability distribution of a discrete random variable	43
4.3	The Bernoulli and binomial distributions	45
4.4	The geometric distribution	48
4.5	Solutions to the quick exercises	50
4.6	Exercises	51
<b>5</b>	<b>Continuous random variables</b>	57
5.1	Probability density functions	57
5.2	The uniform distribution	60
5.3	The exponential distribution	61
5.4	The Pareto distribution	63
5.5	The normal distribution	64
5.6	Quantiles	65
5.7	Solutions to the quick exercises	67
5.8	Exercises	68
<b>6</b>	<b>Simulation</b>	71
6.1	What is simulation?	71
6.2	Generating realizations of random variables	72
6.3	Comparing two jury rules	75
6.4	The single-server queue	80
6.5	Solutions to the quick exercises	84
6.6	Exercises	85
<b>7</b>	<b>Expectation and variance</b>	89
7.1	Expected values	89
7.2	Three examples	93
7.3	The change-of-variable formula	94
7.4	Variance	96
7.5	Solutions to the quick exercises	99
7.6	Exercises	99
<b>8</b>	<b>Computations with random variables</b>	103
8.1	Transforming discrete random variables	103
8.2	Transforming continuous random variables	104
8.3	Jensen's inequality	106

8.4	Extremes .....	108
8.5	Solutions to the quick exercises .....	110
8.6	Exercises .....	111
<b>9</b>	<b>Joint distributions and independence .....</b>	<b>115</b>
9.1	Joint distributions of discrete random variables .....	115
9.2	Joint distributions of continuous random variables .....	118
9.3	More than two random variables .....	122
9.4	Independent random variables .....	124
9.5	Propagation of independence .....	125
9.6	Solutions to the quick exercises .....	126
9.7	Exercises .....	127
<b>10</b>	<b>Covariance and correlation .....</b>	<b>135</b>
10.1	Expectation and joint distributions .....	135
10.2	Covariance .....	138
10.3	The correlation coefficient .....	141
10.4	Solutions to the quick exercises .....	143
10.5	Exercises .....	144
<b>11</b>	<b>More computations with more random variables .....</b>	<b>151</b>
11.1	Sums of discrete random variables .....	151
11.2	Sums of continuous random variables .....	154
11.3	Product and quotient of two random variables .....	159
11.4	Solutions to the quick exercises .....	162
11.5	Exercises .....	163
<b>12</b>	<b>The Poisson process .....</b>	<b>167</b>
12.1	Random points .....	167
12.2	Taking a closer look at random arrivals .....	168
12.3	The one-dimensional Poisson process .....	171
12.4	Higher-dimensional Poisson processes .....	173
12.5	Solutions to the quick exercises .....	176
12.6	Exercises .....	176
<b>13</b>	<b>The law of large numbers .....</b>	<b>181</b>
13.1	Averages vary less .....	181
13.2	Chebyshev's inequality .....	183

13.3	The law of large numbers . . . . .	185
13.4	Consequences of the law of large numbers . . . . .	188
13.5	Solutions to the quick exercises . . . . .	191
13.6	Exercises . . . . .	191
<b>14</b>	<b>The central limit theorem . . . . .</b>	<b>195</b>
14.1	Standardizing averages . . . . .	195
14.2	Applications of the central limit theorem . . . . .	199
14.3	Solutions to the quick exercises . . . . .	202
14.4	Exercises . . . . .	203
<b>15</b>	<b>Exploratory data analysis: graphical summaries . . . . .</b>	<b>207</b>
15.1	Example: the Old Faithful data . . . . .	207
15.2	Histograms . . . . .	209
15.3	Kernel density estimates . . . . .	212
15.4	The empirical distribution function . . . . .	219
15.5	Scatterplot . . . . .	221
15.6	Solutions to the quick exercises . . . . .	225
15.7	Exercises . . . . .	226
<b>16</b>	<b>Exploratory data analysis: numerical summaries . . . . .</b>	<b>231</b>
16.1	The center of a dataset . . . . .	231
16.2	The amount of variability of a dataset . . . . .	233
16.3	Empirical quantiles, quartiles, and the IQR . . . . .	234
16.4	The box-and-whisker plot . . . . .	236
16.5	Solutions to the quick exercises . . . . .	238
16.6	Exercises . . . . .	240
<b>17</b>	<b>Basic statistical models . . . . .</b>	<b>245</b>
17.1	Random samples and statistical models . . . . .	245
17.2	Distribution features and sample statistics . . . . .	248
17.3	Estimating features of the “true” distribution . . . . .	253
17.4	The linear regression model . . . . .	256
17.5	Solutions to the quick exercises . . . . .	259
17.6	Exercises . . . . .	259

<b>18</b>	<b>The bootstrap</b>	269
18.1	The bootstrap principle	269
18.2	The empirical bootstrap	272
18.3	The parametric bootstrap	276
18.4	Solutions to the quick exercises	279
18.5	Exercises	280
<b>19</b>	<b>Unbiased estimators</b>	285
19.1	Estimators	285
19.2	Investigating the behavior of an estimator	287
19.3	The sampling distribution and unbiasedness	288
19.4	Unbiased estimators for expectation and variance	292
19.5	Solutions to the quick exercises	294
19.6	Exercises	294
<b>20</b>	<b>Efficiency and mean squared error</b>	299
20.1	Estimating the number of German tanks	299
20.2	Variance of an estimator	302
20.3	Mean squared error	305
20.4	Solutions to the quick exercises	307
20.5	Exercises	307
<b>21</b>	<b>Maximum likelihood</b>	313
21.1	Why a general principle?	313
21.2	The maximum likelihood principle	314
21.3	Likelihood and loglikelihood	316
21.4	Properties of maximum likelihood estimators	321
21.5	Solutions to the quick exercises	322
21.6	Exercises	323
<b>22</b>	<b>The method of least squares</b>	329
22.1	Least squares estimation and regression	329
22.2	Residuals	332
22.3	Relation with maximum likelihood	335
22.4	Solutions to the quick exercises	336
22.5	Exercises	337

<b>23</b>	<b>Confidence intervals for the mean</b>	341
23.1	General principle	341
23.2	Normal data	345
23.3	Bootstrap confidence intervals	350
23.4	Large samples	353
23.5	Solutions to the quick exercises	355
23.6	Exercises	356
<b>24</b>	<b>More on confidence intervals</b>	361
24.1	The probability of success	361
24.2	Is there a general method?	364
24.3	One-sided confidence intervals	366
24.4	Determining the sample size	367
24.5	Solutions to the quick exercises	368
24.6	Exercises	369
<b>25</b>	<b>Testing hypotheses: essentials</b>	373
25.1	Null hypothesis and test statistic	373
25.2	Tail probabilities	376
25.3	Type I and type II errors	377
25.4	Solutions to the quick exercises	379
25.5	Exercises	380
<b>26</b>	<b>Testing hypotheses: elaboration</b>	383
26.1	Significance level	383
26.2	Critical region and critical values	386
26.3	Type II error	390
26.4	Relation with confidence intervals	392
26.5	Solutions to the quick exercises	393
26.6	Exercises	394
<b>27</b>	<b>The <math>t</math>-test</b>	399
27.1	Monitoring the production of ball bearings	399
27.2	The one-sample $t$ -test	401
27.3	The $t$ -test in a regression setting	405
27.4	Solutions to the quick exercises	409
27.5	Exercises	410

<b>28 Comparing two samples</b> .....	415
28.1 Is dry drilling faster than wet drilling? .....	415
28.2 Two samples with equal variances .....	416
28.3 Two samples with unequal variances .....	419
28.4 Large samples .....	422
28.5 Solutions to the quick exercises .....	424
28.6 Exercises .....	424
<b>A Summary of distributions</b> .....	429
<b>B Tables of the normal and <math>t</math>-distributions</b> .....	431
<b>C Answers to selected exercises</b> .....	435
<b>D Full solutions to selected exercises</b> .....	445
<b>References</b> .....	475
<b>List of symbols</b> .....	477
<b>Index</b> .....	479

# Why probability and statistics?

Is everything on this planet determined by randomness? This question is open to philosophical debate. What is certain is that every day thousands and thousands of engineers, scientists, business persons, manufacturers, and others are using tools from probability and statistics.

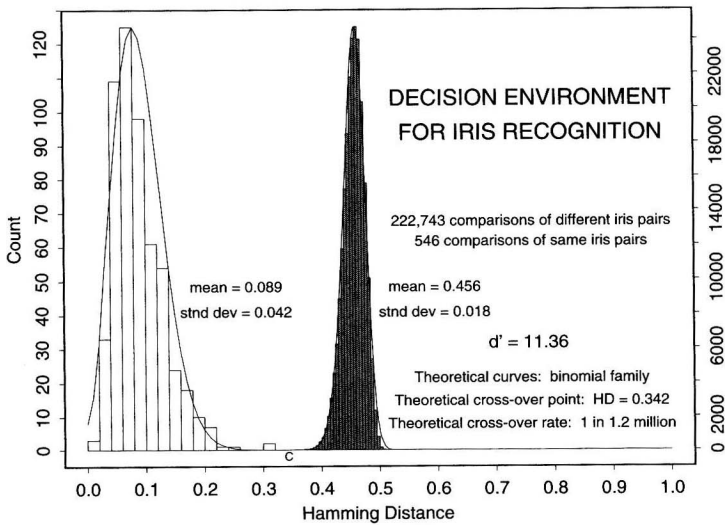
The theory and practice of probability and statistics were developed during the last century and are still actively being refined and extended. In this book we will introduce the basic notions and ideas, and in this first chapter we present a diverse collection of examples where randomness plays a role.

## 1.1 Biometry: iris recognition

Biometry is the art of identifying a person on the basis of his or her personal biological characteristics, such as fingerprints or voice. From recent research it appears that with the human iris one can beat all existing automatic human identification systems. Iris recognition technology is based on the visible qualities of the iris. It converts these—via a video camera—into an “iris code” consisting of just 2048 bits. This is done in such a way that the code is hardly sensitive to the size of the iris or the size of the pupil. However, at different times and different places the iris code of the same person will not be exactly the same. Thus one has to allow for a certain percentage of mismatching bits when identifying a person. In fact, the system allows about 34% mismatches! How can this lead to a reliable identification system? The miracle is that different persons have very different irides. In particular, over a large collection of different irides the code bits take the values 0 and 1 about half of the time. But that is certainly not sufficient: if one bit would determine the other 2047, then we could only distinguish two persons. In other words, single bits may be random, but the correlation between bits is also crucial (we will discuss correlation at length in Chapter 10). John Daugman who has developed the iris recognition technology made comparisons between 222 743 pairs of iris



codes and concluded that of the 2048 bits 266 may be considered as uncorrelated ([6]). He then argues that we may consider an iris code as the result of 266 coin tosses with a fair coin. This implies that if we compare two such codes from different persons, then there is an astronomically small probability that these two differ in less than 34% of the bits—almost all pairs will differ in about 50% of the bits. This is illustrated in Figure 1.1, which originates from [6], and was kindly provided by John Daugman. The iris code data consist of numbers between 0 and 1, each a Hamming distance (the fraction of mismatches) between two iris codes. The data have been summarized in two histograms, that is, two graphs that show the number of counts of Hamming distances falling in a certain interval. We will encounter histograms and other summaries of data in Chapter 15. One sees from the figure that for codes from the same iris (left side) the mismatch fraction is only about 0.09, while for different irides (right side) it is about 0.46.



**Fig. 1.1.** Comparison of same and different iris pairs.

Source: J.Daugman. *Second IMA Conference on Image Processing: Mathematical Methods, Algorithms and Applications*, 2000. © Ellis Horwood Publishing Limited.

You may still wonder how it is possible that irides distinguish people so well. What about twins, for instance? The surprising thing is that although the color of eyes is hereditary, many features of iris patterns seem to be produced by so-called epigenetic events. This means that during embryo development the iris structure develops randomly. In particular, the iris patterns of (monozygotic) twins are as discrepant as those of two arbitrary individuals.