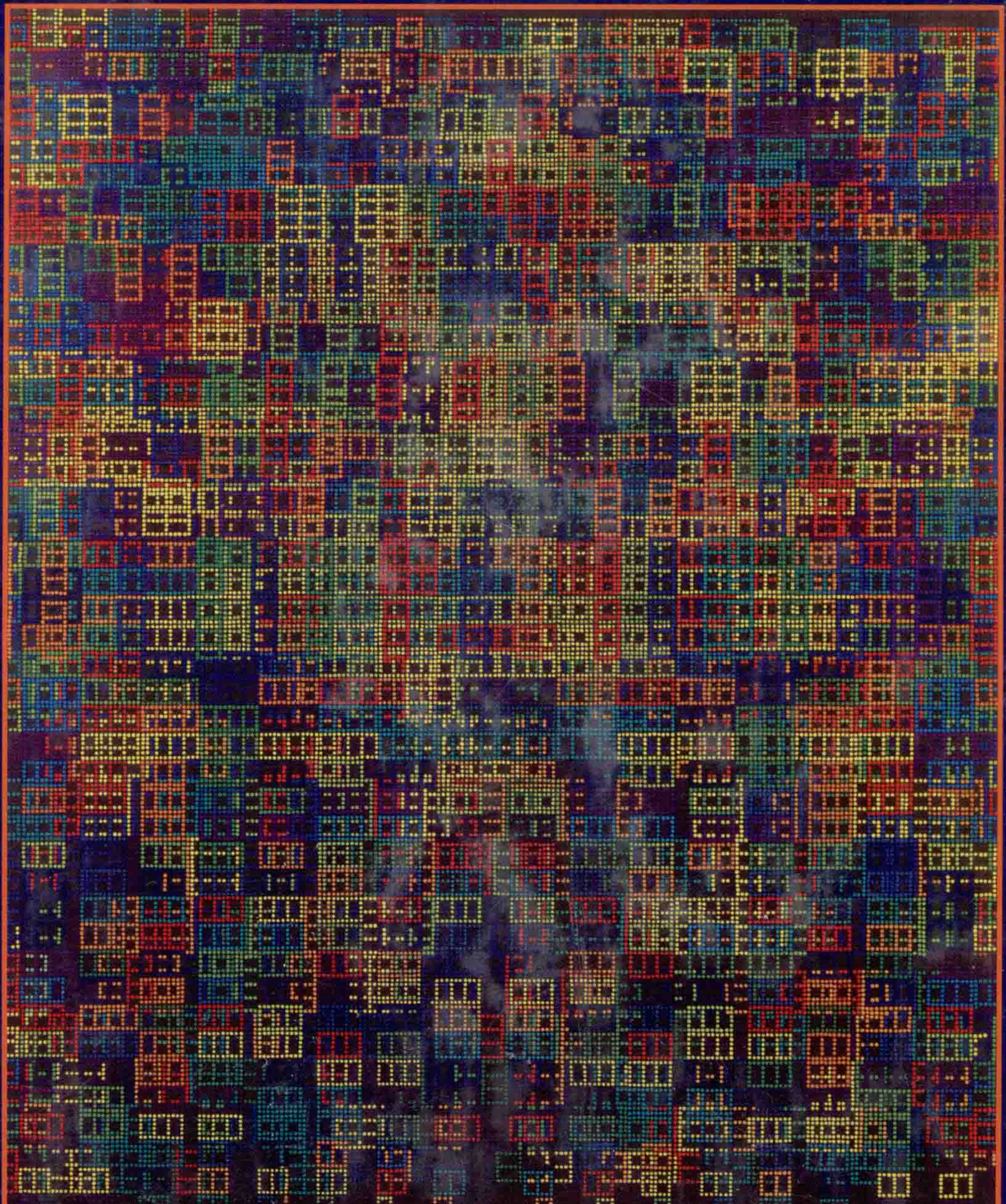
PROBABILITY AND STATISTICS



THE SCIENCE OF UNCERTAINTY

Second Edition

Michael J. Evans   Jeffrey S. Rosenthal

Probability and Statistics

The Science of Uncertainty

Second Edition

Michael J. Evans and Jeffrey S. Rosenthal

University of Toronto

W. H. Freeman and Company

New York

Senior Publisher: Craig Bleyer
Publisher: Ruth Baruth
Executive Marketing Manager: Jennifer Somerville
Developmental Editor: Anne Scanlan-Rohrer
Supplements Editor: Jodi Isman
Associate Editor: Katrina Wilhelm
Photo Editor: Ted Szczepanski
Cover Image: Xylor Jane
Cover Designer: Victoria Tomaselli
Project Editor: Vivien Weiss
Illustrations and Composition: Michael J. Evans, Jeffrey S. Rosenthal
Director of Production: Ellen Cash
Printing and Binding: RR Donnelley

Library of Congress Control Number: 2009936557

Printed in the United States of America

First printing 2009

W. H. Freeman and Company
41 Madison Avenue
New York, NY 10010
Houndmills, Basingstoke RG21 6XS, England

www.whfreeman.com

Preface

This book is an introductory text on probability and statistics, targeting students who have studied one year of calculus at the university level and are seeking an introduction to probability and statistics with mathematical content. Where possible, we provide mathematical details, and it is expected that students are seeking to gain some mastery over these, as well as to learn how to conduct data analyses. All the usual methodologies covered in a typical introductory course are introduced, as well as some of the theory that serves as their justification.

The text can be used with or without a statistical computer package. It is our opinion that students should see the importance of various computational techniques in applications, and the book attempts to do this. Accordingly, we feel that computational aspects of the subject, such as Monte Carlo, should be covered, even if a statistical package is not used. Almost any statistical package is suitable. A **Computations** appendix provides an introduction to the R language. This covers all aspects of the language needed to do the computations in the text. Furthermore, we have provided the R code for any of the more complicated computations. Students can use these examples as templates for problems that involve such computations, e.g., using Gibbs sampling. Also, we have provided, in a separate section of this appendix, Minitab code for those computations that are slightly involved, e.g., Gibbs sampling. No programming experience is required of students to do the problems.

We have organized the exercises in the book into groups, as an aid to users. **Exercises** are suitable for all students and offer practice in applying the concepts discussed in a particular section. **Problems** require greater understanding, and a student can expect to spend more thinking time on these. If a problem is marked (MV), then it will require some facility with multivariable calculus beyond the first calculus course, although these problems are not necessarily hard. **Challenges** are problems that most students will find difficult; these are only for students who have no trouble with the **Exercises** and the **Problems**. There are also **Computer Exercises** and **Computer Problems**, where it is expected that students will make use of a statistical package in deriving solutions.

We have included a number of **Discussion Topics** designed to promote critical thinking in students. Throughout the book, we try to point students beyond the mastery of technicalities to think of the subject in a larger frame of reference. It is important that students acquire a sound mathematical foundation in the basic techniques of probability and statistics, which we believe this book will help students accomplish. Ultimately, however, these subjects are applied in real-world contexts, so it is equally important that students understand how to go about their application and understand what issues arise. Often, there are no right answers to **Discussion Topics**; their purpose is to get a student thinking about the subject matter. If these were to be used for evaluation, then they would be answered in essay format and graded on the maturity the student showed with respect to the issues involved. **Discussion Topics** are probably most suitable for smaller classes, but these will also benefit students who simply read them over and contemplate their relevance.

Some sections of the book are labelled **Advanced**. This material is aimed at students who are more mathematically mature (for example, they are taking, or have taken, a second course in calculus). All the **Advanced** material can be skipped, with no loss of continuity, by an instructor who wishes to do so. In particular, the final chapter of the text is labelled **Advanced** and would only be taught in a high-level introductory course aimed at specialists. Also, many proofs appear in the final section of many chapters, labelled **Further Proofs (Advanced)**. An instructor can choose which (if any) of these proofs they wish to present to their students.

As such, we feel that the material in the text is presented in a flexible way that allows the instructor to find an appropriate level for the students they are teaching. A **Mathematical Background** appendix reviews some mathematical concepts, from a first course in calculus, in case students could use a refresher, as well as brief introductions to partial derivatives, double integrals, etc.

Chapter 1 introduces the probability model and provides motivation for the study of probability. The basic properties of a probability measure are developed.

Chapter 2 deals with discrete, continuous, joint distributions, and the effects of a change of variable. It also introduces the topic of simulating from a probability distribution. The multivariate change of variable is developed in an Advanced section.

Chapter 3 introduces expectation. The probability-generating function is discussed, as are the moments and the moment-generating function of a random variable. This chapter develops some of the major inequalities used in probability. A section on characteristic functions is included as an Advanced topic.

Chapter 4 deals with sampling distributions and limits. Convergence in probability, convergence with probability 1, the weak and strong laws of large numbers, convergence in distribution, and the central limit theorem are all introduced, along with various applications such as Monte Carlo. The normal distribution theory, necessary for many statistical applications, is also dealt with here.

As mentioned, Chapters 1 through 4 include material on Monte Carlo techniques. Simulation is a key aspect of the application of probability theory, and it is our view that its teaching should be integrated with the theory right from the start. This reveals the power of probability to solve real-world problems and helps convince students that it is far more than just an interesting mathematical theory. No practitioner divorces himself from the theory when using the computer for computations or vice versa. We believe this is a more modern way of teaching the subject. This material can be skipped, however, if an instructor believes otherwise or feels there is not enough time to cover it effectively.

Chapter 5 is an introduction to statistical inference. For the most part, this is concerned with laying the groundwork for the development of more formal methodology in later chapters. So practical issues — such as proper data collection, presenting data via graphical techniques, and informal inference methods like descriptive statistics — are discussed here.

Chapter 6 deals with many of the standard methods of inference for one-sample problems. The theoretical justification for these methods is developed primarily through the likelihood function, but the treatment is still fairly informal. Basic methods of inference, such as the standard error of an estimate, confidence intervals, and P-values, are introduced. There is also a section devoted to distribution-free (nonparametric) methods like the bootstrap.

Chapter 7 involves many of the same problems discussed in Chapter 6, but now from a Bayesian perspective. The point of view adopted here is not that Bayesian methods are better or, for that matter, worse than those of Chapter 6. Rather, we take the view that Bayesian methods arise naturally when the statistician adds another ingredient — the prior — to the model. The appropriateness of this, or the sampling model for the data, is resolved through the model-checking methods of Chapter 9. It is not our intention to have students adopt a particular philosophy. Rather, the text introduces students to a broad spectrum of statistical thinking.

Subsequent chapters deal with both frequentist and Bayesian approaches to the various problems discussed. The Bayesian material is in clearly labelled sections and can be skipped with no loss of continuity, if so desired. It has become apparent in recent years, however, that Bayesian methodology is widely used in applications. As such, we feel that it is important for students to be exposed to this, as well as to the frequentist approaches, early in their statistical education.

Chapter 8 deals with the traditional optimality justifications offered for some statistical inferences. In particular, some aspects of optimal unbiased estimation and the Neyman–Pearson theorem are discussed. There is also a brief introduction to decision theory. This chapter is more formal and mathematical than Chapters 5, 6, and 7, and it can be skipped, with no loss of continuity, if an instructor wants to emphasize methods and applications.

Chapter 9 is on model checking. We placed model checking in a separate chapter to emphasize its importance in applications. In practice, model checking is the way statisticians justify the choices they make in selecting the ingredients of a statistical problem. While these choices are inherently subjective, the methods of this chapter provide checks to make sure that the choices made are sensible in light of the objective observed data.

Chapter 10 is concerned with the statistical analysis of relationships among variables. This includes material on simple linear and multiple regression, ANOVA, the design of experiments, and contingency tables. The emphasis in this chapter is on applications.

Chapter 11 is concerned with stochastic processes. In particular, Markov chains and Markov chain Monte Carlo are covered in this chapter, as are Brownian motion and its relevance to finance. Fairly sophisticated topics are introduced, but the treatment is entirely elementary. Chapter 11 depends only on the material in Chapters 1 through 4.

A one-semester course on probability would cover Chapters 1–4 and perhaps some of Chapter 11. A one-semester, follow-up course on statistics would cover Chapters 5–7 and 9–10. Chapter 8 is not necessary, but some parts, such as the theory of unbiased estimation and optimal testing, are suitable for a more theoretical course.

A basic two-semester course in probability and statistics would cover Chapters 1–6 and 9–10. Such a course covers all the traditional topics, including basic probability theory, basic statistical inference concepts, and the usual introductory applied statistics topics. To cover the entire book would take three semesters, which could be organized in a variety of ways.

The Advanced sections can be skipped or included, depending on the level of the students, with no loss of continuity. A similar approach applies to Chapters 7, 8, and 11.

Students who have already taken an introductory, noncalculus-based, applied statistics course will also benefit from a course based on this text. While similar topics are covered, they are presented with more depth and rigor here. For example, *Introduction to the Practice of Statistics*, 6th ed., by D. Moore and G. McCabe (W. H. Freeman, 2009) is an excellent text, and we believe that our book would serve as a strong basis for a follow-up course.

There is an **Instructor's Solutions Manual** available from the publisher.

The second edition contains many more basic exercises than the first edition. Also, we have rewritten a number of sections, with the aim of making the material clearer to students. One goal in our rewriting was to subdivide the material into smaller, more digestible components so that key ideas stand out more boldly. There has been a complete typographical redesign that we feel aids in this as well. In the appendices, we have added material on the statistical package R as well as answers for the odd-numbered exercises that students can use to check their understanding.

Many thanks to the reviewers and users for their comments: Abbas Alhakim (Clarkson University), Michelle Baillargeon (McMaster University), Arne C. Bathke (University of Kentucky), Lisa A. Bloomer (Middle Tennessee State University), Christopher Brown (California Lutheran University), Jem N. Corcoran (University of Colorado), Guang Cheng (Purdue University), Yi Cheng (Indiana University South Bend), Eugene Demidenko (Dartmouth College), Robert P. Dobrow (Carleton College), John Ferdinands (Calvin College), Soledad A. Fernandez (The Ohio State University), Paramjit Gill (University of British Columbia Okanagan), Marvin Glover (Milligan College), Ellen Gundlach (Purdue University), Paul Gustafson (University of British Columbia), Jan Hannig (Colorado State University), Solomon W. Harrar (The University of Montana), Susan Herring (Sonoma State University), George F. Hilton (Pacific Union College), Chun Jin (Central Connecticut State University), Paul Joyce (University of Idaho), Hubert Lilliefors (George Washington University), Andy R. Magid (University of Oklahoma), Phil McDonnough (University of Toronto), Julia Morton (Nipissing University), Jean D. Opsomer (Colorado State University), Randall H. Rieger (West Chester University), Robert L. Schaefer (Miami University), Osnat Stramer (University of Iowa), Tim B. Swartz (Simon Fraser University), Glen Takahara (Queen's University), Robert D. Thompson (Hunter College), David C. Vaughan (Wilfrid Laurier University), Joseph J. Walker (Georgia State University), Chad Westerland (University of Arizona), Dongfeng Wu (Mississippi State University), Yuehua Wu (York University), Nicholas Zaino (University of Rochester). In particular, Professor Chris Andrews (State University of New York) provided many corrections to the first edition.

The authors would also like to thank many who have assisted in the development of this project. In particular, our colleagues and students at the University of Toronto have been very supportive. Ping Gao, Aysha Hashim, Gun Ho Jang, Hadas Moshonov, and Mahinda Samarakoon helped in many ways. A number of the data sets in Chapter

10 have been used in courses at the University of Toronto for many years and were, we believe, compiled through the work of the late Professor Daniel B. DeLury. Professor David Moore of Purdue University was of assistance in providing several of the tables at the back of the text. Patrick Farace, Anne Scanlan-Rohrer, Chris Spavins, Danielle Swearengin, Brian Tedesco, Vivien Weiss, and Katrina Wilhelm of W. H. Freeman provided much support and encouragement. Our families helped us with their patience and care while we worked at what seemed at times an unending task; many thanks to Rosemary and Heather Evans and Margaret Fulford.

Michael Evans and Jeffrey Rosenthal
Toronto, 2009

Contents

Chapter 1

Probability Models

CHAPTER OUTLINE

This chapter introduces the basic concept of the entire course, namely, probability. We discuss why probability was introduced as a scientific concept and how it has been formalized mathematically in terms of a probability model. Following this we develop some of the basic mathematical results associated with the probability model.

1.1 Probability: A Measure of Uncertainty

Often in life we are confronted by our own ignorance. Whether we are pondering tonight's traffic jam, tomorrow's weather, next week's stock prices, an upcoming election, or where we left our hat, often we do not know an outcome with certainty. Instead, we are forced to guess, to estimate, to hedge our bets.

Probability is the science of uncertainty. It provides precise mathematical rules for understanding and analyzing our own ignorance. It does not tell us tomorrow's weather or next week's stock prices; rather, it gives us a framework for working with our limited knowledge and for making sensible decisions based on what we do and do not know.

To say there is a 40% chance of rain tomorrow is not to know tomorrow's weather. Rather, it is to know what we do not know about tomorrow's weather.

In this text, we will develop a more precise understanding of what it means to say there is a 40% chance of rain tomorrow. We will learn how to work with ideas of randomness, probability, expected value, prediction, estimation, etc., in ways that are sensible and mathematically clear.

There are also other sources of randomness besides uncertainty. For example, computers often use *pseudorandom numbers* to make games fun, simulations accurate, and searches efficient. Also, according to the modern theory of quantum mechanics, the makeup of atomic matter is in some sense *truly* random. All such sources of randomness can be studied using the techniques of this text.

Another way of thinking about probability is in terms of *relative frequency*. For example, to say a coin has a 50% chance of coming up heads can be interpreted as saying that, if we flipped the coin many, many times, then approximately half of the time it would come up heads. This interpretation has some limitations. In many cases (such as tomorrow's weather or next week's stock prices), it is impossible to repeat the experiment many, many times. Furthermore, what precisely does "approximately" mean in this case? However, despite these limitations, the relative frequency interpretation is a useful way to think of probabilities and to develop intuition about them.

Uncertainty has been with us forever, of course, but the mathematical theory of probability originated in the seventeenth century. In 1654, the Paris gambler Le Chevalier de Méré asked Blaise Pascal about certain probabilities that arose in gambling (such as, if a game of chance is interrupted in the middle, what is the probability that each player would have won had the game continued?). Pascal was intrigued and corresponded with the great mathematician and lawyer Pierre de Fermat about these questions. Pascal later wrote the book *Traité du Triangle Arithmetique*, discussing binomial coefficients (Pascal's triangle) and the binomial probability distribution.

At the beginning of the twentieth century, Russians such as Andrei Andreyevich Markov, Andrey Nikolayevich Kolmogorov, and Pafnuty L. Chebychev (and American Norbert Wiener) developed a more formal mathematical theory of probability. In the 1950s, Americans William Feller and Joe Doob wrote important books about the mathematics of probability theory. They popularized the subject in the western world, both as an important area of pure mathematics and as having important applications in physics, chemistry, and later in computer science, economics, and finance.

1.1.1 Why Do We Need Probability Theory?

Probability theory comes up very often in our daily lives. We offer a few examples here.

Suppose you are considering buying a "Lotto 6/49" lottery ticket. In this lottery, you are to pick six distinct integers between 1 and 49. Another six distinct integers between 1 and 49 are then selected at random by the lottery company. If the two sets of six integers are identical, then you win the jackpot.

After mastering Section 1.4, you will know how to calculate that the probability of the two sets matching is equal to one chance in 13,983,816. That is, it is about 14 million times more likely that you will not win the jackpot than that you will. (These are not very good odds!)

Suppose the lottery tickets cost \$1 each. After mastering expected values in Chapter 3, you will know that you should not even *consider* buying a lottery ticket unless the jackpot is more than \$14 million (which it usually is not). Furthermore, if the jackpot is ever more than \$14 million, then likely many other people will buy lottery tickets that week, leading to a larger probability that you will have to *share* the jackpot with other winners even if you do win — so it is probably not in your favor to buy a lottery ticket even then.

Suppose instead that a "friend" offers you a bet. He has three cards, one red on both sides, one black on both sides, and one red on one side and black on the other. He mixes the three cards in a hat, picks one at random, and places it flat on the table with only one side showing. Suppose that one side is red. He then offers to bet his $4 against your $3 that the other side of the card is also red.

At first you might think it sounds like the probability that the other side is also red is 50%; thus, a good bet. However, after mastering conditional probability (Section 1.5), you will know that, conditional on one side being red, the conditional probability that the other side is also red is equal to 2/3. So, by the theory of expected values (Chapter 3), you will know that you should not accept your "friend's" bet.

Finally, suppose he suggests that you flip a coin one thousand times. Your "friend" says that if the coin comes up heads at least six hundred times, then he will pay you $100; otherwise, you have to pay him just $1.

At first you might think that, while 500 heads is the most likely, there is still a *reasonable* chance that 600 heads will appear — at least good enough to justify accepting your friend's $100 to $1 bet. However, after mastering the laws of large numbers (Chapter 4), you will know that as the number of coin flips gets large, it becomes more and more likely that the number of heads is very close to half of the total number of coin flips. In fact, in this case, there is less than one chance in ten billion of getting more than 600 heads! Therefore, you should not accept this bet, either.

As these examples show, a good understanding of probability theory will allow you to correctly assess probabilities in everyday situations, which will in turn allow you to make wiser decisions. It might even save you money!

Probability theory also plays a key role in many important applications of science and technology. For example, the design of a nuclear reactor must be such that the escape of radioactivity into the environment is an extremely rare event. Of course, we would like to say that it is categorically impossible for this to ever happen, but reactors are complicated systems, built up from many interconnected subsystems, each of which we know will fail to function properly at some time. Furthermore, we can never definitely say that a natural event like an earthquake cannot occur that would damage the reactor sufficiently to allow an emission. The best we can do is try to quantify our uncertainty concerning the failures of reactor components or the occurrence of natural events that would lead to such an event. This is where probability enters the picture. Using probability as a tool to deal with the uncertainties, the reactor can be designed to ensure that an unacceptable emission has an extremely small probability — say, once in a billion years — of occurring.

The gambling and nuclear reactor examples deal essentially with the concept of *risk* — the risk of losing money, the risk of being exposed to an injurious level of radioactivity, etc. In fact, we are exposed to risk all the time. When we ride in a car, or take an airplane flight, or even walk down the street, we are exposed to risk. We know that the risk of injury in such circumstances is never zero, yet we still engage in these activities. This is because we intuitively realize that the probability of an accident occurring is extremely low.