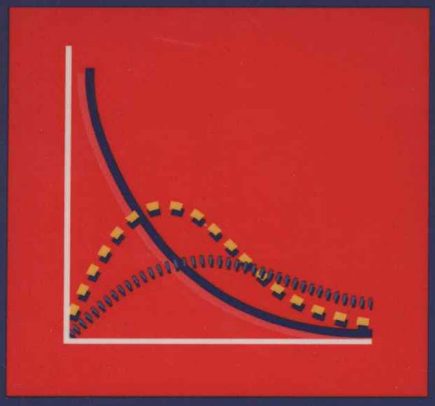
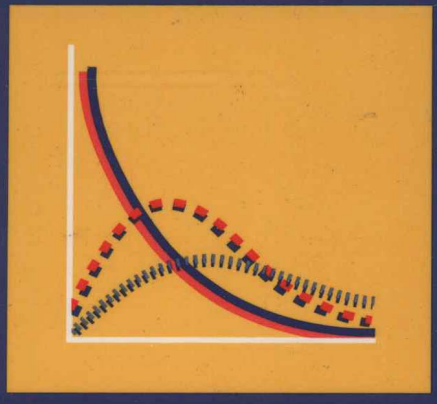
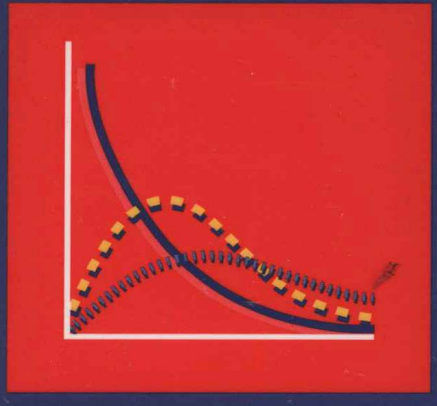
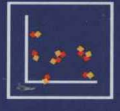






PROBABILITY and STATISTICS

KEVIN J. HASTINGS



ADDISON-WESLEY
ADVANCED SERIES
IN STATISTICS

PROBABILITY AND STATISTICS

KEVIN J. HASTINGS
Knox College

ADDISON-WESLEY

An imprint of Addison Wesley Longman, Inc.

Reading, Massachusetts • Menlo Park, California • New York • Harlow, England
Don Mills, Ontario • Sydney • Mexico City • Madrid • Amsterdam

Senior Editor: Julia Berrisford
Associate Editor: Jennifer Albanese
Production Services: Diane Freed
Production Supervisor: Peggy McMahon
Senior Manufacturing Manager: Roy Logan
Associate Marketing Manager: Benjamin Rivera
Cover Design Supervisor: Meredith Nightingale
Cover Designer: Darci Meehall
Compositor: Intercontinental Photocomposition, Ltd.
Technical Art Illustration: George Nichols

Library of Congress Cataloging-in-Publication Data

Hastings, Kevin J.
Probability and statistics / by Kevin J. Hastings. -- 1st ed.
p. cm.
Includes bibliographical references (p -) and index.
ISBN 0-201-59278-9
1. Probabilities. 2. Mathematical statistics. I. Title.
QA273.H375 1997 96-11393
519.2–dc20 CIP

2 3 4 5 6 7 8 9 10 MA 999897

PREFACE

It is not an easy decision to write a textbook. You can be motivated by the negative point of view that there is not a book out there on your subject that is worth using. But this is not true in the case of probability and statistics. Or you can believe, enthralled by your own ego, that you have the definitive approach to which the world simply must be exposed. I hope that I had no such motivations. For me the reason to write this book had a lot to do with the joy of writing itself, and with the challenge. I did go so far as to think that I had several productive new ideas of dealing with an established topic that might interest and enlighten people and could make a worthwhile addition to the great library of probability and statistics.

The most important audience for a textbook is the students, and so I have at all times attempted to write with the student reader in mind. This does not mean talking down to the reader; quite the opposite, a proper style acknowledges the reader's intelligence and willingness to work. It does mean writing in a lively style, involving the reader in the process, avoiding obscurity and ambiguity, and not fearing the occasional use of humor. I have a little sermon to preach here on the subject of mathematical writing style. An early reviewer complained rather bitterly about what he or she considered my excessive use of the pronoun *we*, as in "we can now see that ...," but I disagree. This style makes us, the reader and I who are studying the objects of probability and statistics together, the subjects of sentences. The indirect passive style makes the objects of study into subjects and pushes the reader into the background, as in "Random variables are said to be independent if ..." This style is traditional in mathematics, but it is distinctly dry and uninvolving. It also makes sentences weak and unnatural through the use of passive verbs like *are*. Since I am at the mercy of my own imperfections and years of countertraining, I have not nearly attained my ideal of avoiding the passive voice completely throughout this long book. But I hope that the more active style that I have attempted, and other aspects of the book, will appeal to both students and instructors.

Let's turn now to issues of content and pedagogy. When I first started to think about this project, I had several main ideas in mind, the totality of which would make this book a little different from others in the area: (1) a full exploitation of matrix algebra and vector calculus; (2) an interactive, problem-solving approach; (3) a fuller coverage of graphical methods in data analysis; (4) computer laboratory investigations to augment the usual pencil-and-paper exercises and straightforward data analysis exercises using statistical packages; and (5) suggestions for student research projects. I want to elaborate on these features one by one.

It has always seemed to me that instructors do themselves and their stu-

dents a disservice by attempting to package courses to be as independent of one another and prerequisite-free as possible. This gives students an unrealistic view of subject areas as being separate, rather than interdependent, and it gives them no incentive to carry knowledge, concepts, and techniques from one course to another. In our particular case, there are many ways to use matrix algebra and multivariate calculus to simplify, clarify, and extend results in probability and statistics, but most books do not fully exploit them, fearing perhaps that to do so would restrict the audience too much. I have used matrix analysis and multivariable calculus freely in this book, and so I assume that the student has had previous courses in these subjects. When it seemed appropriate, I reviewed the results I needed in context. I have included a short appendix (Appendix E) that summarizes briefly some of the most useful results in matrix algebra.

The second main theme of this book is harder to define. My philosophy is that the students need to be active participants in their education, not passive recipients. I tried every device I could think of to encourage this, including the stylistic one already discussed, and the springboard appendix to independent study. For example, I included numerous data sets, both in examples and exercises, which present the problem solver with a mass of numbers and an ill-stated question. The challenge is to make sense of the question, to get a feeling for what the numbers are saying, to construct a statistical procedure, and interpret the results in terms of the original problem. Every section includes several self-check questions for the reader to answer, as a device to monitor understanding. I left some of the theoretical material to the exercises, instead of laying out everything in a nice, neat package. I also let the students discover some of the cases of hypothesis tests and confidence intervals by solving problems instead of plodding mulelike through textual exposition of all of them. And probably the most important example of the implementation of the philosophy is the problem sets, which contain quite a few exercises that my own students have found challenging and thought-provoking.

The third main idea involves graphics. Statistics does not begin and end with analytical methods. Statistical consultants are coming around to the view that one of their most important roles is to present information to their clients that is understandable and convincing, and that using graphics is a good way to accomplish this. Moreover, it is now generally understood that the statistician must be concerned not only with graphical *methods*, but also with *principles* of good graph construction. Probability and statistics books have been successful at illustrating the methods, and this text does not attempt to break much new ground in that area. But the construction and design ideas are understressed in other books (perhaps because they don't seem mathematical enough), and I have given these more than the usual amount of attention. Appropriate graphs are regularly included in the more analytical chapters to solidify intuition.

The fourth of the distinctive qualities that I have given to the book is the set of computer investigations. I have included ten of them, in Appendix C, and have ideas for quite a few more. You the instructor may already have some of your own, or your imagination may suggest others by looking at

the ones I have created. Now that the technology is widely available, it is possible to look at questions that we could not easily look at before. Parameter sensitivity, robustness against outliers, empirical power comparisons, and asymptotic properties are areas in which technology can assist understanding, enable the user to branch off in new directions, or go beyond the barrier of computational intensiveness. Simulation is a necessary means to study some of these problems, but I have found in my teaching experience that more can be said for simulation: Doing simulations helps the student better understand the fundamental concepts of randomness, random variables, and distributions. Hence simulation is not just a means but an end. In general, when used appropriately, technology is an indispensable means of enhancing education. These lab projects assume some familiarity with *Mathematica*, but they also give detailed assistance on the necessary commands.

And the fifth main idea is a lengthy appendix (Appendix D) to suggest and support projects for student investigation. Many books of this genre use the theory of probability mostly to set up statistical inference and do not take the opportunity to illustrate the beautiful applications of probability itself. I have chosen several applications (Markov chains, reliability, portfolios, and queueing) that I like and have worked on in the past, which the instructor could assign as small research projects culminating in an oral presentation and a written report. With a reluctant sigh, I decided to group these applications in Appendix D rather than integrating them, mainly because the book was getting rather long. There are suggestions for research in other areas of statistics such as Bayesian statistics and time series analysis, with references.

Apart from the differences discussed here, the organization and content of the book are fairly standard, and the instructor should not have much trouble adapting this text to an existing course. Here are some other less consequential ways in which this book differs from some others. There is a bit less coverage here than usual on the theoretical properties of point estimators and hypothesis tests. I preferred to devote more space to matrix methods and graphical data analysis and to extend discussion of the design (as opposed to just the analysis) of experiments. Nonparametric methods are integrated so as to present them along with the problem for which they are meant, rather than marginally.

Following is a quick run-through of the ordering of topics and other items worthy of note when preparing to use this book for a course. One important general observation is that I frequently use the device of giving sneak previews of later topics, and returning to earlier topics. I think that the reinforcement resulting from multiple exposure helps students to learn better and retain longer. A drawback of this construction, which is more like a network than a line, is that it is dangerous to skip things because there is so much interdependence. Here are some specifics about Chapters 1–6 on Probability.

Chapter 1 is meant to throw the reader into the deep water of the basic concepts of probability quickly: sample spaces, random variables, distributions, conditioning, and independence. The Law of Total Probability is highlighted in the text and in the exercises, and a rather general definition of the

concept of distribution is used. Especially with regard to random variables and their distributions, the strategy is to expose students to these concepts and let them be puzzled for a while as they do exercises. Then they will come back to them in other contexts later in Chapters 2, 3, and 4 on discrete distributions, continuous distributions, and independence, respectively. For instance, my students have found that they have had trouble early with the idea of a continuous probability distribution, but by the end of Section 3.2 something had clicked, and they came away with a deep understanding that they put to use later in the book.

Chapter 2 brings the student to the point of familiarity with basic combinatorial probability (focusing on sampling), discrete distributions and applications related to binomial experiments, phenomena modeled by the Poisson distribution, and discrete expectation. The treatment is fairly standard here, with the possible exception of the extra thoroughness of the material on Poisson processes.

The third chapter explores continuous probability distributions and expectation, highlighting the gamma family and the normal distribution. There are vector expectation results given in the second section which are used occasionally later. The multivariate normal distribution is also introduced there, which becomes a recurring theme as the exposition progresses. We return to it in Sections 4.4 and 5.4 as we learn more about conditional distributions and transformations.

Chapter 4 is a straightforward discussion of conditional distributions and independence of random variables, including a couple of nice examples of the application of the Law of Total Probability for expectation. The third section on covariance and correlation includes a discussion of covariance matrices, and a result about the covariance of a constant matrix times a random vector, an unusual topic for books at this level.

The material in Chapter 5 on transformations is fairly mainstream, including the c.d.f. technique with an emphasis on the simulation theorem, multivariate change of variables, moment-generating functions, and applications to transformations of normal random variables. But there is consistent attention given to multivariate results and methods, especially as regards quadratic forms of normal vectors. These ideas are used heavily later in Chapters 10 and 11 on regression and the analysis of variance.

In a short sixth chapter on large sample theory, the Weak and Strong Laws of Large Numbers and the Central Limit Theorem are discussed. I have carefully tried to clarify the distinction between weak and almost everywhere convergence. The weak law depends on Chebyshev's inequality. The Central Limit Theorem is treated in the usual way. Here, and earlier in the section of Chapter 5 on normal transformations, a sneak preview of statistical analysis using the sample mean is given.

As for Chapters 7–12 on Statistics, Chapter 7 leads off the study with a treatment of random sampling and the elementary summary statistics. We then proceed to graphical methods, covering the usual types of graphs, including histograms, dot and box plots, and normal scores plots. But the chapter ends with a rather unique discussion of good graph-making techniques, done

from a problem-solving point of view: If a graph is to tell a story, how should it be constructed, and how should the fine details be implemented so as to enhance, or at least not to obscure, the story to be told?

Chapter 8 deals with parameter estimation, highlighting maximum likelihood and the notions of unbiasedness and precision of estimators. Sufficiency is included, as is a short treatment of confidence intervals, which should dovetail well with the idea of precision of an estimator.

In the ninth chapter I have tried to cover as efficiently as possible the classical normal theory tests for means, variances, and proportions. But the chapter takes a conceptual approach, rather than a smorgasbord approach. The reader is encouraged to go at data analysis problems by asking questions: What am I trying to find out about a certain population? and What commonsense summary measures would be appropriate to look at? The first section introduces the basic terms and ideas and includes a discussion of both the large sample test for the proportion and the sign test. This is an attempt to get the reader to think of a hypothesis test as a problem to be solved creatively, rather than a standard technique to be applied. The rest of the chapter passes through one- and two-sample location tests, including integrated subsections on the Wilcoxon and Mann–Whitney tests, and then the chapter proceeds to dispersion tests. In Section 9.3 where dispersion tests are covered there is also some material on Mood's nonparametric test. Beyond just giving the nonparametric procedures their fair due, I think that the integration of classical normal theory with some distribution-free theory stresses to the student the fact that assumptions are important and that testing hypotheses is a creative, problem-solving enterprise. The chapter ends with a brief introduction to the likelihood ratio testing criterion.

Chapter 10 on regression and correlation takes more than the customary advantage of multivariate theory from the first part of the book. It considers the regression model as a matrix linear equation in the unknown parameters, and in so doing previews the following chapter. The first two sections set up the problem and cover least squares parameter estimation in roughly the usual way. Section 10.3 contains very powerful theorems, derived with matrix techniques, on the distribution of estimators and sums of squares. With these in hand, the student is ready to apply and understand the statistical inference procedures. There is some discussion of residuals and diagnostic checking in the fourth section of this chapter, using an example-oriented approach. Correlation analysis completes the chapter.

On a recent sabbatical I became rather interested in the area of experimental design, especially in the fact that despite the appearance of being a disjointed amalgamation of problems and techniques, it is thematically unified by the basic structure of the linear model. So, I decided to stress this idea in Chapter 11, and as well, to include a longer discussion of data gathering in designed experiments than is normally done (Section 11.1). We study one-factor problems, including random effects models and blocking, and two-factor problems including random effects. The topic of Latin Square designs is left as one of the research project possibilities in Appendix D. There is

heavy usage throughout the chapter of the results on quadratic forms from Chapter 5.

Chi-square goodness-of-fit testing occupies most of our attention in Chapter 12. We begin by looking at the simple multinomial fit problem with all category probabilities known, then move on to testing the fit of distributions, and problems in which parameters are estimated. The Kolmogorov–Smirnov procedure is also discussed here. The chapter closes with a rather standard presentation of independence testing on contingency tables using the classical chi-square statistic. The nonparametric runs test for problems involving trends is also covered.

The supporting material in the appendices includes statistical tables, a short guide to the *Minitab* statistical system, computer exercises with hints on using *Mathematica*, research project ideas, and a linear algebra review. Answers to many of the exercises follow. A complete solutions manual is available.

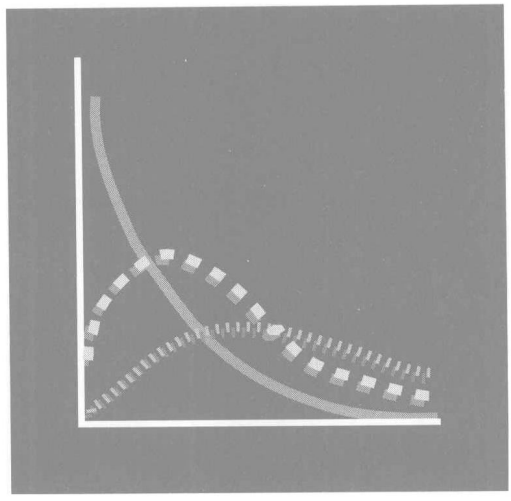
I would like to thank the people at Addison-Wesley, particularly Julia Berrisford, for their help and their understanding when I was just a tad past our agreed upon deadlines. Much of whatever credit the book earns is really due to the great teachers and colleagues of probability and statistics that I have known and worked with: Rothwell Stephens at Knox College, Stan Pliska and my dissertation adviser, Erhan Cinlar, at Northwestern, Marcel Neuts and others at the University of Delaware, and Russ Lenth at the University of Iowa. The many students who I have had, including those whose projects are mentioned in the text and those who helped me with the solution manual, may have made as great a contribution to my education as I have to theirs, and for that I thank them. I am very grateful to the reviewers of the manuscript: Bernice Auslander, University of Massachusetts, Boston; James Conklin, Ithaca College; Joseph Glaz, University of Connecticut; H. Allan Knappenberger, Wayne State University; Karen Messer, California State University, Fullerton; Don Ridgeway, North Carolina State University; Lyndon Weberg, University of Wisconsin, River Falls; and Douglas Wolfe, The Ohio State University. Most importantly, I thank my wife, Gay Lynn, whose love and support have helped carry me through the second half of this project and who also helped with a good deal of the technical word processing. And to my precious baby, Emily Marie, who gives my life so much joy, thanks for going to sleep at 8:00 so that Mommy and Daddy could work.

I sincerely hope that this text will be a positive contribution to the field. In some ways it is on the demanding side, but not unreasonably so if the student has honestly attempted to commit calculus and linear algebra to long-term memory. More than anything, I hope that the approach that I decided to take will spark the same kind of excitement about probability and statistics that I felt when I was first learning. The subjects deserve no less.

KJH
Galesburg, IL

CONTENTS

CHAPTER 1



SAMPLE SPACES AND RANDOM VARIABLES

1.1 Introduction and Examples

Our investigation of probability and statistics must begin with an understanding of the word *random*. The Second College Edition of *Webster's New World Dictionary* makes several attempts at a definition, including: "without careful choice," "haphazardly," and the rather contradictory pair "not uniform" and "pertaining to a set of items, every member of which has an equal chance of occurring." The last definition comes closest to what we need, although it is far too restrictive. The subjects of probability and statistics exist because of a need to understand events whose occurrence we cannot predict with certainty. However, we may know how likely these events are. So for us, the word *random* will pertain to an experiment or phenomenon whose result (or *outcome*, in the language of probability) remains uncertain until the experiment is performed or the phenomenon is observed. The most familiar and elementary random phenomenon is probably the flip of a coin. It is unknown

prior to the flip which of the two possible outcomes, "head" or "tail," will occur, but the outcome is observable after the flip. Note that we needn't assume that each possible outcome of a random experiment is as likely as each other outcome, just as we needn't assume that our coin is fair.

Here are several examples of random phenomena and their possible outcomes.

1. Will I or won't I be interrupted by a phone call while I am writing today? Until such a phone call comes or I complete my writing for the day, I will not know the answer to this question. Thus the event of being interrupted becomes a random one, with possible outcomes "yes" or "no." It turned out that I was interrupted (an outcome that seems to have high probability).
2. How much will IBM common stock go up today, January 13, 1992, on the New York Stock Exchange? We will not know what will happen until the close of trading, so this situation fits our idea of a random phenomenon. There are many possible outcomes, which are numbers usually reported as multiples of eighths. The observed outcome turned out to be $-5/8$. Past observation of the performance of the stock might give information on the likelihoods of outcomes like $-5/8$.
3. What will be the winning three-digit number today, January 13, 1992, on Michigan's Daily-3 Lottery game? Again, we cannot know until the number is selected by the lottery authorities; after the number is drawn, the outcome is known and it is of the form abc, where a, b, and c are digits from the set of digits 0 through 9. The winning outcome happened to be 517. Theoretically, each three-digit number should be as likely as any other if the mechanism for drawing numbers is fair.
4. What will be the official high temperature in Atlantic City today? Here, outcomes are numerical temperatures. (The observed outcome was 57°F.) At least ideally, if we could measure temperatures with perfect accuracy on a continuous scale, the possible outcomes would form some (uncountably infinite) interval on the real line. One would guess that temperatures nearer to the normal temperature for Atlantic City at this time of year would be more likely than extremely warm or cold temperatures.

?Question 1.1.1 Think of at least three other random phenomena. What are their possible outcomes?

The purpose of this first section is to introduce you to the main concepts of probability, such as random phenomena, very informally, so that you begin the process of acquiring intuition. Then as the next few chapters unfold, we will look at the concepts and their applications more rigorously and in more detail.

Problems involving random phenomena or experiments entail collections of possible *outcomes*. These outcomes are indivisible, meaning that they cannot be broken down into more primitive components. We will refer to the set

of all possible outcomes of the experiment as the *sample space* of the experiment. For instance, the sample space associated with the phone interruption phenomenon cited earlier is {yes, no}. As another example, a lottery game in which four different two-digit numbers are selected includes outcomes such as $\{02, 65, 73, 22\}$, $\{87, 01, 94, 45\}$, and so forth. A formal way to write this sample space is

$$\{\{x_1, x_2, x_3, x_4\} \mid x_i \in \{0, 1, \ldots, 99\}, x_i \neq x_j,\ i, j = 1, 2, 3, 4\}.$$

In Chapter 2 we will learn how to count the number of possible outcomes of an experiment like this.

?Question 1.1.2 Write a formal set-theoretic description of the sample space for the Michigan Daily-3 Lottery game described earlier.

Notice from the examples that outcomes may sometimes be numbers, other times collections or sequences of numbers, and still other times nonnumerical character strings. Because outcomes can take on many different structures depending on the random phenomenon being considered, we will construct a framework for probability in which outcomes are left abstract. The idea of a universal sample space of possible outcomes, however, will remain the consistent unifying theme.

Indivisible outcomes are the building blocks of *events*, which are subsets of the sample space. In reference to some of the examples, the event that IBM rises in value by at least 2/8, the event that 23 is one of the winning Michigan lottery numbers, and the event that the high temperature in Atlantic City is at least 50° all consist of more than one outcome. We will also allow events to contain no outcomes or one outcome.

?Question 1.1.3 For each of your phenomena from Question 1.1.1, give an example of an event.

Probability is a measure of the likelihood of events. Since events are sets of outcomes, it should be enough to define probability for outcomes. The probability associated with an event should then be the total of the probabilities of the outcomes in that event. This last observation is essentially true, although in the kind of idealized experiments mentioned here where the set of outcomes can be uncountably infinite, we need to think more carefully about what the total of outcome probabilities is and to consider whether it even makes sense to give outcomes nonzero probability.

To see the difference between finite (or countable) sample space models and uncountably infinite models, consider this example. A runner in the 100-meter dash will not always complete the run in the same amount of time. One instance may result in a time of 10.3 seconds, another 10.5, another 10.1, and so on. We could propose two probabilistic models.