Reviews of Books and Teaching Materials

Fundamental Probability: A Computational Approach.

Marc S. Paolella. Chichester, West Sussex: Wiley, 2006, xxii+466 pp., $130.00 (H), ISBN: 0-470-02594-8.

Fundamental Probability: A Computational Approach by Marc S. Paolella is the first of what is to be a three-volume set on probability and statistics. There is much to say about the first volume. I recently obtained the second volume, and am looking forward to reading through it. If Volume II is as well done as Volume I, then I will surely purchase Volume III when it becomes available. Volume I is a solid, 466-page coverage of combinatorics, probability spaces, counting, conditioning, and discrete and continuous random variables. Paolella also includes a detailed, self-contained appendix of calculus tools used throughout the book, appendices with tables which collect notational, distributional, and other information in a concise manner.

Fundamental Probability is primarily a text. It is designed for students with a basic command of freshman calculus and some linear algebra, but with no previous exposure to probability. For material presented at this academic level, many of the topics are standard—the Bernoulli and binomial distributions, sums of random variables, the probability integral transformation, and t distributions; and some are not standard—utility functions, stochastic dominance, and difference equations. The sheer number of topics can be overwhelming, especially if this volume is being considered as a text for an introductory course in probability. However, as Paolella emphasizes in the Preface, "not everything in the text is supposed to be (or could be) covered in the classroom." Paraphrasing, Paolella's two overriding goals in writing the book are (1) to emphasize the practical side of probability by presenting a wide variety of examples and (2) to go beyond the limited set of topics in standard, "safe" examples traditionally taught at this level. He accomplishes these overriding goals masterfully.

Examples cover a broad range with topics in areas such as legal issues and litigation support, political science, biology and medicine, marketing, risk assessment, and reliability. The examples illustrate previously discussed theory, enrich the theory with special detailed cases, and introduce new, but specialized, material or applications. As previously mentioned, Paolella admits that the amount of material in *Fundamental Probability* is too much to cover in lecture. On the other hand, the amount of material is not so much that students cannot read material that is not specifically covered in lecture. Paolella's approach is rich, interesting, informative, and (dare I say it?) at times, entertaining. Students will be engaged by the practical applications of a subject that can seem—at least to them—simply a mental exercise, or an intermediate step to get to what they consider the truly useful tools. As a result of Paolella's hard work and skill in integrating so many varied applications, students may even change their minds with regard to their narrow perceptions about probability.

For computation, Paolella uses Matlab, a high-performance language for technical computing that integrates computation, visualization, and programming in an easy-to-use environment. The Matlab programming language is matrix-based and uses familiar mathematical notation; thus, it requires very little initial investment for students to be productive. All in all, Matlab is a good choice. The programming code and data structures provided by Paolella are as universal as possible, which makes the translation into S-Plus, or the freely available R, an easy task.

The text is divided into three parts. Part I covers basic probability across three chapters spanning 102 pages. Chapter 1 covers the basics of combinatorics, and begins with the most fundamental idea in counting, $n!$. The chapter quickly advances to combinatoric identities and proof by induction, the binomial and multinomial theorems, and the complete and incomplete gamma and beta functions evaluated for integer arguments. In this chapter, the application of Matlab is a function that computes values of $\binom{n}{k}$ for possible vector values of n and k. Included in the section on the gamma function is Stirling's approximation. Examples in this chapter are usually more mechanical in nature, since the main gist is illustrating manipulation of the various mathematical quantities. Also of importance in this chapter is the illustration of the usefulness of the mathematical quantities with respect to problems in probability, and proving various important results associated with them (like Vandermonde's theorem). The details in the examples are sufficient. There are examples for illustrating practical application, although these examples are, for the most part, trite. However, some examples do contain entertaining historical trivia, as do the quotes and associated footnotes that open each section throughout the book. The chapter ends with a set of 19 exercises of varying levels of difficulty, rating from zero to three $\star$'s. One exercise requires Matlab.

Chapter 2 covers counting techniques, and develops the idea of a probability space and some of its most useful properties. Beginning with sampling with and without replacement, ordered and unordered events, the counting formulas associated with the four combinations are developed in a traditional way. Paolella also notes that unordered events coupled with drawing with replacement can be alternatively considered as an occupancy problem and relates occupancy problems to problems in physics.

The chapter then moves on to the development of probability spaces. Keeping in mind the academic level of his target audience, Paolella begins developing the abstract ideas of σ-field, Borel field, and probability measure by using a simple example to which his readers can relate—drawing a card from a standard deck of 52. Examples in this chapter are funny, and many still remain trite. But here is also the first occurrence of an example "reaching beyond" the material presented in the chapter through the introduction of the idea of difference equations having solutions that are computable but that do not occur in a closed form and long-run expectation. Several Matlab functions are used to illustrate long-run behaviors via simulation. He also provides a thoughtful discussion on the complexities of generating random numbers, sketching an outline that explains why the values, which are considered "random" from a computer are completely deterministic, and should therefore be considered "pseudo-random." Several easily accessible, authoritative references (e.g., Ripley 1987 among others) are provided for the more interested reader. The last sections of Chapter 2 present basic and advanced properties of probability measures, Venn diagrams, and some limiting behaviors. Accordingly, results such as Boole's inequality, Bonferroni's inequality, Poincaré's theorem, and the like are presented, and some are proved. The chapter ends with a nice combination of 25 applied, computational, and theoretical exercises.

Chapter 3 presents the ideas of a symmetric probability space and conditional probability. The beginning of the discussion on a symmetric probability space is easy to follow, and is quickly stretched with many thought-provoking examples that illustrate a number of ways symmetric probability spaces are applicable. Some of these examples also include simulation as a visualization tool. The extensive section on conditional probability includes the typical topics (law of total probability and Bayes' rule). Some of the examples serve to simply illustrate the utility of these results. But many other examples are relevant, interesting, and clever. Included are the "Monty Hall Puzzle," (along with its history), "Simpson's Paradox," "The Problem of Points," and the "Gambler's Ruin Problem." As an added bonus, for some examples Paolella provides more than one solution, thereby guiding students in reasoning through the various dimensions of the problem. The chapter ends with 21 applied or theoretical exercises of varying levels of difficulty.

Part II is three chapters long. The first chapter (Chapter 4) discusses univariate discrete random variables. Chapter 5 contains a discussion of multivariate random variables. The last chapter covers sums of random variables. Chapter 4 begins by restating the general probability space $\{\Omega, \mathcal{A}, \Pr(\cdot)\}$, and defining a univariate random variable X relative to the collection of measurable events $\mathcal{A}$ as a function with domain Ω and range the real number line. This definition is immediately followed by a sound intuitive explanation. This illustrates again Paolella's continual gentle stressing of the complicated underlying structure without overwhelming technical detail, thus refusing to allow it to be ignored.

The obligatory treatment of the traditional topics—Bernoulli and binomial, hypergeometric, geometric, and negative binomial—is thorough and predictable, developing the distributions from sampling and counting schemes. Unlike most texts, *Fundamental Probability* also includes a nice development of the inverse hypergeometric distribution, Naor's distribution, and lattice distributions. There is a brief introduction of a few basic continuous distributions (standard uniform, exponential, standard normal) to illustrate the difference between discrete and continuous. The uniform and exponential distributions are more thoroughly discussed in Part III. It is also in Chapter 4 that Paolella addresses the age-old controversy of continuity in application. I have heard statisticians and mathematicians vehemently argue two contrasting points of view: (1) that a continuous model approximates a discrete reality, and (2) that a discrete mea-

©2008 American Statistical Association DOI: 10.1198/000313008X297070

surement is an approximation of a continuous reality. Which side Paolella takes in this argument is most easily seen in a quote from page 118 in Chapter 4:

> . . . any phenomenon (that being an observable fact or event) can only be measured with finite precision so that, if X is a standard uniform random variable, then it can, realistically speaking, take on only a finite number of different values, . . . Thus, ultimately, X is a discrete uniform random variable . . . but, *for practical purposes*, when the number of values in the genuine support of X is large, we envisage f_X to be the idealized, mathematical limited case.

This is reiterated in the opening sentence of Chapter 7 (on continuous univariate random variables) in the statement, "In many branches of statistics and other fields, continuous distributions offer reasonable and tractable approximations to the underlying process of interest even though virtually all phenomena are, at some level, discrete." Whether or not you agree with Paolella's perspective, is not the point. By including this material that some may call controversial, Paolella opens the door to a thoughtful discussion.

The continuing treatment of discrete (and later) continuous behavior, is a solid one. As Chapter 4 progresses, examples become more interesting and application oriented. The use of Matlab for illustrating mathematical behaviors becomes more prevalent than in past chapters. Computational issues, such as computing $n!$ for large values of n are addressed and solutions provided. Chapter 4 also contains the introduction of transformations, expected value, and higher-order moments. Skewness and kurtosis are included in the discussion as values for quantifying deviations from symmetry. The chapter concludes with a development of Poisson processes. There is a plethora of other topics introduced in Chapter 4. The 13 exercises are varied in application, level of difficulty, and the need for computing.

Chapter 5 contains a development of multivariate analogs of the topics introduced in Chapter 4. Also included are the ideas of exchangeability and of independence. Relative to the other chapters, this one is short. Chapter 6 introduces sums of random variables. A nice discussion of the runs distributions is provided, along with illustrative graphs created from Matlab. Random variable decomposition, especially with respect to the Bernoulli, binomial, negative binomial, hypergeometric, inverse hypergeometric, and Poisson distributions is included. The chapter ends with a discussion of general linear combinations of two random variables. The exercises at the end of both chapters are a little sparse (only six in Chapter 5, and nine in Chapter 6). But the coverage of the topics is engaging and well done.

Part III contains a coverage of continuous random variables in the final three chapters of the book. The organization of Part III is similar to Part II. Chapter 7—the first chapter of Part III—discusses univariate continuous random variables. Chapter 8 discusses joint and conditional random variables. The text concludes in Chapter 9 by presenting multivariate transformations of continuous random variables.

The presentation of continuous distributions begins with a discussion of "most prominent distributions." Paolella provides readers with six practical reasons for why 18 distributions are more prominent in application and literature than others. He follows this with a solid development of the 18 prominent distributions. Through this necessarily lengthy exposition, he includes important concepts like location-scale families, the central limit theorem, Bessel functions, and various computational issues. This section is particularly rich in information and well presented. It also contains a good number of engaging examples from a variety of applications. Some of these examples are not marked as such. Instead they are contained within the development of a particular distribution, making that distribution relevant, as opposed to an exercise in mathematical statistics. That is not to say there are not plenty of examples in the mathematical arena. There are. They are just well mixed with application-oriented discussion. The section on prominent distributions is followed by another section on other popular distributions. Univariate transformations, the probability integral transformation, simulation, kernel density estimation, are also presented in Chapter 7. The chapter ends with a solid set of 24 exercises.

Chapter 8 covers joint and conditional random variables. The first section is a review of basic calculus-oriented methods. Section 2 contains the heart of the discussion in this chapter, covering conditional distributions for both discrete and continuous random variables. The set of graphical illustrations serve to enhance the discussion. The "exchange paradox" (also known as "the wallet game") are included as examples, but for the most part, the examples in this section are more mathematically oriented. The second section also covers conditional moments, independence, and a nice discussion of computing probabilities via conditioning. The chapter ends with 14 exercises.

The final chapter, Chapter 9, presents a nice exposition on multivariate transformations. Not only are the mathematical details carefully spelled out and illustrated, so is their utility. The t and F distributions are developed again (they were first seen in Chapter 7 as "prominent distributions"), but here they are cast as arising from a multivariate transformation of independent standard normal random variables. Section 3 covers further aspects of multivariate transformations, including other important transformations such as the Box–Muller transformation. There are eight exercises at the end of this section. The following appendices include a review of calculus, notational tables, and a table of a summary of the distributions discussed in the text and the properties of those distributions.

Paolella's second volume, *Intermediate Probability: A Computational Approach*, was released late in 2007. The second volume is more suited to beginning graduate students in statistics, and covers more advanced topics in probability, "including (i) a detailed look at sums of random variables via exact and approximate numeric inversion of moment generating and characteristic functions, (ii) a discussion of more advanced distributions including the stable Paretian and generalized hyperbolic, and (iii) a detailed look at noncentral distributions, quadratic forms, and ratios of quadratic forms." The third volume, in which Paolella will "deal with the subject of statistical modeling and extraction of information from data," is in some stage of production. I have already obtained a copy of the second volume for my library. At first glance, it appears to be as well done as Volume I. I fully intend to purchase Volume III when it becomes available. The collection is sure to be a set of remarkable references for students and teachers alike.

Jane L. HARVILL
Baylor University

REFERENCES

Ripley, B. P. (1987), *Stochastic Simulation*, New York: Wiley.

Graphics of Large Datasets: Visualizing a Million.

Antony UNWIN, Martin THEUS, and Heike HOFMANN. New York: Springer, 2006, xiii+271 pp., $84.95 (H), ISBN: 0-387-32906-4.

What would you do if you wanted to look at a scatterplot of a large bivariate continuous dataset, but you had more data points than you had pixels on your computer screen? Or, even if you had enough pixels, how would you discover the important information held hostage by the black mass of overplotting? What if you had either a large multivariate continuous or large multivariate categorical dataset? *Graphics of Large Datasets: Visualizing a Million* addresses these and other similar situations where there is a need to visualize a dataset that is either large in number of cases, large in number of variables, or large in both.

The book is a compilation of the expertise, experience, research, and software of the three main authors (Hofmann, Theus, Unwin), and 10 guest authors (Cook, González-Arévalo, Hernández-Campos, Marron, Miller, Moustafa, Park, Urbanek, Wegman, and Wills). It comprises an introductory chapter and two multichapter parts.

Chapter 1 (Introduction) defines what a large dataset is, and it orients the reader to some of the technical and logistical problems inherent in collecting, storing, retrieving, visualizing, and statistically analyzing large sets. The authors have chosen one million (1M) to represent a large dataset—a definition that is flexible enough so that no matter how "large dataset" has been defined in the graphics, visualization, and data analysis literature, most would agree that 1M of something is large enough to create scale-up problems. Chapter 1 also highlights and provides websites for the many datasets and software, which will be used throughout the book, and it provides the companion website for the book.

Part I, "Basics," is composed of three chapters. Chapter 2 defines, illustrates, and presents visualization issues for some reasonably standard statistical graphics including bar charts, spine plots, mosaic plots, dot plots, box plots, histograms, scatterplots, trellis plots, the Grand Tour, parallel coordinate plots, maps, contour plots, image maps, time series plots, table plots, and missing value plots.

Chapter 3 points out that area-based plots for categorical data such as bar charts, histograms, and mosaic plots may scale up quite nicely with some reasonably simple refinements in construction and presentation. But most graphs for continuous data such as scatterplots and parallel coordinate plots do not fare too well due to overplotting and increased numbers of extreme outliers, unless some type of hierarchical or subset zooming, tonal highlighting, alpha-bending, or colors with alpha-transparency are used.

Chapter 4 describes traditional interactive graphing tools and their refinements for dissecting and displaying large datasets. It also addresses new problems and techniques generated by the larger dataset size, for example, subsetting, aggregating, recoding, transforming, and weighting.