Preface

> Writing a book is an adventure. To begin with, it is a toy and an amusement; then it becomes a mistress, and then it becomes a master, and then a tyrant. The last phase is that just as you are about to be reconciled to your servitude, you kill the monster, and fling him out to the public.
>
> (Sir Winston Churchill)
>
> *Reproduced by permission of Little Brown*

Like many branches of science, probability and statistics can be effectively taught and appreciated at many mathematical levels. This book is the first of two on probability, and designed to accompany a course aimed at students who possess a basic command of freshman calculus and some linear algebra, but with no previous exposure to probability. It follows the more or less traditional ordering of subjects at this level, though with greater scope and often more depth, and stops with multivariate transformations. The second book, referred to as Volume II, continues with more advanced topics, 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.

A subsequent book will deal with the subject of statistical modeling and extraction of information from data. In principle, such prerequisites render the books appropriate for upper division undergraduates in a variety of disciplines, though the *amount* of material and, in some topics covered, the depth and use of some elements from advanced calculus, make the books especially suited for students with more focus and mathematical maturity, such as undergraduate math majors, or beginning graduate students in statistics, finance or economics, and other fields which use the tools of probability and statistical methodology.

Motivation and approach

The books grew (and grew. . .) from a set of notes I began writing in 1996 to accompany a two-semester sequence I started teaching in the statistics and econometrics faculty in

Kiel University in Germany.[1] The group of students consisted of upper-division undergraduate and beginning graduate students from quantitative business (50 %), economics (40 %) and mathematics (10 %). These three disparate groups, with their different motivations and goals, complicated the choice of which topics to cover and textbook to use, and I eventually opted to concentrate strictly on my own notes, pursuing a flexible agenda which could satisfy their varying interests and abilities.

The two overriding goals were (i) to emphasize the practical side of matters by presenting a large variety of examples and discussing computational matters, and (ii) to go beyond the limited set of topics and standard, ‘safe’ examples traditionally taught at this level (more on this below). From my experience, this results in a much more enjoyable and unified treatment of the subject matter (theory, application and computation) and, more importantly, highly successful students with an above average competence, understanding, breadth and appreciation for the usefulness of our discipline.

It must be mentioned, however, that while these two goals certainly complement each other, they gave rise to a substantially larger project than similarly positioned books. It is thus essential to understand that *not everything in the texts is supposed to be (or could be) covered in the classroom.* The teacher has the choice of what to cover in class and what to assign as outside reading, and the student has the choice of how much of the remaining material he or she wishes to read and study. In my opinion, a textbook should be a considerably larger superset of material covered in class, bursting with varied, challenging examples, full derivations, graphics, programming techniques, references to old and new literature, and the like, and not a dry protocol of a lecture series, or a bare-bones outline of the major formulae and their proofs.

With respect to computation, I have chosen Matlab, which is a matrix-based prototyping language requiring very little initial investment to be productive. Its easy structure, powerful graphics interface and enormous computational library mean that we can concentrate on the problem we wish to solve, and avoid getting sidetracked by the intricacies and subtleties of the software used to do it. All these benefits also apply to S-Plus® and R, which certainly could be used in place of Matlab if the instructor and/or student wishes. With that in mind, I have attempted to make the programming code and data structures as simple and universal as possible, so that a line-by-line translation of the numerous programs throughout would often be straightforward.

Numeric methods are used throughout the book and are intentionally woven tightly together with the theory. Overall, students enjoy this approach, as it allows the equations and concepts to ‘come to life’. While computational and programming exercises are valuable in themselves, experience suggests that most students are at least as willing to play around on the computer as they are on paper, with the end result that more time is spent actively thinking about the underlying problem and the associated theory, and so increasing the learning effect.

Of course, because of its reliance on modern hardware and software, this approach does not enjoy a long tradition but is quickly gaining in acceptance, as measured by the number of probability and statistics textbooks which either incorporate computational aspects throughout (e.g. Kao, 1996; Childers, 1997; Rose and Smith, 2002; Ruppert,

[1] For a bit of trivia, Kiel’s mathematics department employed William Feller until 1933, as well as Otto Toeplitz from 1913 to 1928. In 1885, Max Planck became a professor of theoretical physics in Kiel, which is also where he was born.

2004), or are fully dedicated to such issues (e.g. Venables and Ripley, 2002; Dalgaard, 2002; Givens and Hoeting, 2005).

Ready access to complex, numerically intensive algorithms and extraordinarily fast hardware should not be seen as causing a 'break' or 'split' with the classical methodological aspects of our subject, but viewed rather as a driving factor which gradually and continually instigates a progression of ideas and methods of solving problems to which we need to constantly adjust. As a simple example, the admirable and difficult work invested into the Behrens–Fisher problem resulted in several relatively simple closed-form approximations for the distribution of the test statistic which could actually be calculated for a given data sample using the modest computational means available at the time. Now, with more computing power available, bootstrap methodology and MCMC techniques could readily be applied. Similar examples abound in which use of computationally demanding methods can be straightforwardly used for handling problems which are simply not amenable to algebraic (i.e. easily computed) solutions.

This is not to say that certain classical concepts (e.g. central limit theorem-based approximations, sufficient statistics, UMVU estimators, pivotal quantities, UMP tests, etc.) are not worth teaching; on the contrary, they are very important, as they provide the logical foundation upon which many of the more sophisticated methods are based. However, I believe these results need to be augmented *in the same course* with (i) approaches for solving problems which make use of modern theoretical knowledge and, when necessary, modern computing power, and (ii) genuinely useful examples of their use, such as in conjunction with the linear model and basic time series analysis, as opposed to overly simplistic, or even silly, i.i.d. settings because they happen to admit tractable algebraic solutions.

My experience as both a student and teacher suggests that the solid majority of students get interested in our subject from the desire to answer questions of a more applied nature. That does not imply that we should shy away from presenting mathematical results, or that we should not try to entice some of them to get involved with more fundamental research, but we should respect and work with their inherent curiosity and strike a balance between application and theory at this level. Indeed, if the climax of a two-course sequence consists of multivariate transformations, rudimentary uses of moment generating functions, and the 'trinity' of Cramer–Rao, Lehmann–Scheffé and Neyman–Pearson, then the student will leave the course not only with an extremely limited toolbox for answering questions with the slightest hint of realism, but also with a very warped and anachronistic view of what our subject has to offer society at large.

This also makes sense even for people majoring in mathematics, given that most of them will enter private industry (e.g. financial institutions, insurance companies, etc.) and will most likely find themselves making use of simulation techniques and approximate inversion formulae instead of, say, minimal sufficiency theorems associated with exponential families.

Examples

The examples throughout the text serve the purpose of (i) illustrating previously discussed theory with more concrete 'objects' (actual numbers, a distribution, a proof, etc.), and/or (ii) enriching the theory by detailing a special case, and/or (iii) introducing new, but specialized, material or applications. There is quite an abundance of examples throughout the text, and not all of them need to be studied in detail in a first passing, or at all. To help guide the reader, symbols are used to designate the importance of the example, with ⊖, ⊚ and ⊛ indicating low, medium and high importance, respectively, where importance refers to '*relevance to the bigger picture*'.

I understand the bigger picture to be: establishing a foundation for successfully pursuing a course in statistical inference (or related disciplines such as biostatistics, actuarial and loss models, etc.), or a course in stochastic processes, or a deeper mathematical treatment of probability theory. In this case, most examples are not highly relevant for the bigger picture. However, an example sporting only the low-importance symbol ⊖ could actually be very helpful, or even critical, for understanding the material just presented, or might introduce new, useful material, or detail an interesting application, or could be instructive in terms of Matlab programming. Thus, the low symbol ⊖ should certainly not be interpreted as an indication that the example should be skipped! Of course, the high importance symbol ⊛ indicates mandatory reading and understanding.

Problems

Some comments on the problem sets provided in the texts are in order. First, I generally avoid overly simplistic exercises of a 'plug and chug' nature, preferring instead to encourage students to test their initial understanding of the material by trying to reproduce some of the more basic examples in the text on their own. Once such a basis is established, the problems can be fruitfully attempted. They range in difficulty, and are marked accordingly (with respect to my subjective beliefs), from rather easy (no stars), to particularly challenging and/or time consuming (three stars).

Second, I try to acknowledge the enormous demands on students' time and the frustration which often arises when attempting to solve homework problems, and thus give them access to the exercise solutions. Detailed answers to all of the problems are provided in a document available from the book's web page `www.wiley.com/go/fundamental`. This allows for self study and self correction and, in my experience, actually *increases* students' motivation and time spent working problems because they can, at any time, check their performance.

Third, I advocate encouraging students to solve problems in various ways, in accordance with their professional goals, individual creativity, mathematical expertise, and time budget. For example, a problem which asks for a proof could, on occasion, be 'numerical verified'. Of course, any problem which begins with 'Using one million Cauchy observations...' should be a clear signal that a computer is required! Less trivially, a problem like 'Compare the bias of the three estimators...' can be attacked algebraically or computationally, or both. Use of symbolic computing packages such as Mathematica, Maple or MuPAD straddles the two fields in some senses, and is also strongly encouraged.

Appendices

Appendix A serves to refresh some essential mathematics from analysis used throughout the text. My suggestion is to have the student skim through its entirety to become familiar with its contents, concentrating on those sections where gaps in the student's knowledge become apparent. There was much vacillation regarding its inclusion because of its size, but experience has shown that most students do not readily exhume their previous math books (let alone purchase different, or more advanced ones) and work through them, but they *will* read a shortened, streamlined version which explicitly ties into the rest of the book, and which is physically attached to the main text.

Appendix B consists of various notation tables, including abbreviations, Latin expressions, Greek letters, a list of special functions, math and distributional notation and moments of random variables.

Appendix C collects tables of distributions, including their naming conventions, p.d.f., c.d.f., parameter space, moments and moment generating function (though the latter is first introduced at the beginning of Volume II). Volume II also includes further tables, with generating function notation, inversion formulae and a table illustrating various relationships among the distributions. What you will not find anywhere in these books are tables of standard normal c.d.f. values, chi-square, t and F cut-off values, etc., as they are no longer necessary with modern computing facilities and no statistical analysis or write-up is done without a computer.

Topics covered in Volume I

Instead of writing here a long summary, I opt for a very short one, and invite the reader and instructor to gloss over the table of contents, which – by construction – is the book's outline. Of course, many topics are standard; some are not, and their respective sections could be briefly skimmed. Some highlights include:

- a detailed, self-contained chapter (Appendix A) on the calculus tools used throughout the book, with many proofs (e.g. fundamental theorem of calculus, univariate and multivariate Taylor series expansions, Leibniz' rule, etc.) and examples (convolution of Cauchy r.v.s, resolution of the normal density's integration constant both with and without the use of polar coordinates, etc.);

- numerous tables (Appendices B and C) which collect notational, distributional and other information in a concise, coherent way;

- a strong use of combinatorics and its use in answering probabilistic questions via De Moivre–Jordan's theorem;

- a detailed treatment of the problem of the points and its historical significance, as well as some references to the historical development of probability;

- full coverage of all four 'basic sampling schemes', e.g. derivation of moments for sampling until r successes without replacement (inverse hypergeometric);

- extension of each of the four 'basic sampling schemes' to the multivariate case, with emphasis on random number generation and simulation;

- ample discussion of occupancy distributions, calculation of the p.m.f., c.d.f. and moments using a variety of methods (combinatoric and, in Volume II, characteristic function inversion and saddlepoint approximation);

- a thorough discussion of basic runs distributions, with full derivation of all common results (and a peek at more recent results);

- use of exchangeability for decomposing random variables such as occupancy, hypergeometric and inverse hypergeometric;

- introduction of the common univariate continuous distributions in such a way as to teach the student how to 'approach' a new distribution (recognizing the roles of parameters, major applications, tail behavior, etc.);

- introduction of some basic concepts used in finance, such as utility functions, stochastic dominance, geometric expectation, continuous returns, value at risk and expected shortfall.

A note to the student (and instructor)

Mathematics

The reader will quickly discover that considerable portions of this text are spent deriving mathematically stated results and illustrating applications and examples of such which involve varying degrees of algebraic manipulation. It should thus come as no surprise that a solid command of 'first-year calculus' (univariate differentiation, integration and Taylor series, partial differentiation and iterated integrals) is critical for this course. Also necessary is the ability to work with the gamma and beta functions (detailed from scratch in Section 1.5) and a good working knowledge of basic linear/matrix algebra for Volume II (in which an overview is provided). With these tools, the reader should be able to comfortably follow about 95 % of the material.

The remaining 5 % uses bits of advanced calculus, e.g. basic set theory, limits, Leibniz' rule, exchange of derivative and integral, along with some topics from multivariate calculus, including polar coordinates, multivariate change of variables, gradients, Hessian matrices and multivariate Taylor series. Appendix A provides a streamlined introduction to these topics. Rudiments of complex analysis are required in Volume II, and an overview is provided there.

Thus, with the help of Appendix A to fill in some gaps, the formal prerequisites are very modest (one year of calculus and linear algebra), though a previous exposure to real analysis or advanced calculus is undoubtedly advantageous for two reasons. First, and obviously, the reader will already be familiar with some of the mathematical tools used here. Secondly, and arguably more important, the student is in a much better position to *separate mathematics from the probabilistic concepts and distributional results introduced in the text*. This is very important, as it allows you to avoid getting bogged down in the particulars of the mathematical details and instead you can keep an eye on the concepts associated with the study of probability and the important distributions which are essential for statistical analysis.

The ability to recognize this distinction is quite simple to obtain, once you become conscious of it, but some students ultimately experience difficulty in the course because they do not do so. For example, the probability mass function of a random variable may be easily expressed recursively, i.e. as a difference equation, but whose solution is not readily known. Such a difference equation could have also arisen in, say, oceanography, or any of dozens of scientific fields, and the method of mathematical solution has nothing to do per se with the initial probabilistic question posed. This was another motivation for including the math review in Appendix A: by reading it first, the student explicitly sees precisely those math tools which are used throughout the book – but in a 'pure math' context unburdened by the new concepts from probability. Then, when a messy series of set operations, or a complicated integral, or a Taylor series expression, arises in the main chapters, it can be decoupled from the context in which it arose.

Having now said that, a big caveat is in order: often, a 'mathematical' problem might be more easily solved in the context of the subject matter, so that the two aspects of the problem are actually wedded together. Nevertheless, during a first exposure to the subject, it is helpful to recognize the border between the new concepts in probability and statistics, and the mathematical tools used to assist their study. As an analogy to a similar situation, it seems natural and wise to separate the derivation of a mathematical solution, and the computer language, algorithmic techniques and data structures used to implement its numeric calculation. At a deeper level, however, the form of the solution might be influenced or even dictated by knowledge of the capabilities and limitations of a possible computer implementation.

Motivation and learning habits

It seems safe to say that there are some universally good ways of learning any kind of subject matter, and some bad ways. I strongly advocate an 'active approach', which means grabbing the bull by the horns, jumping right into the cold water (with the bull), and getting your hands dirty. Concretely, this means having pencil, paper and a computer at hand, and picking to pieces any equation, program, derivation or piece of logic which is not fully clear. You should augment my explanations with more detail, and replace my solutions with better ones. Work lots of problems, and for those which you find difficult, read and master the solutions provided, then close the book and do it yourself. Then do it again tomorrow. The same goes for the Matlab programs.

Learn the material with the intention of making it your own, so that you could even teach it with competence and confidence to others. Construct your own questions and write up your own solutions (good opportunity to learn LaTeX), challenge your instructor with good questions (that you have given your own thought to), borrow or buy other textbooks at a lower, similar or higher level, think about applications in fields of your interest, and build learning groups with other motivated students. Now is the time to challenge yourself and learn as much as you can, because once you leave university, the realisms of life (such as family and employment concerns) will confront you and take up most of your time, and learning new things becomes difficult.

Some of the ingredients for success can be put into an acronym: get TIMED! That is, you will need Time, Interest, Mathematical maturity, Energy and Discipline. A deficit in one can often be accommodated by an increase in one or more of the others, though all are required to some extent. If one of these is failing, figure out why, and how to solve it.

A note to the instructor (and student)

Book usage and course content

As mentioned above, the book is ideal for beginning graduate students in the social or mathematical sciences, but is also suited (and designed) for use with upper division undergraduates, provided that the more difficult material (about a quarter of the book) is omitted. Let me first discuss how I use the book in a graduate program, and then in an undergraduate setting.

For teaching graduate students, I begin by quickly covering the math appendix (this has positive externalities because they need the material for other courses anyway), spending about 3 hours (two or three lectures) in total. I just emphasize the major results, and give the students the chance to work on a few basic examples during class (see my comments on teaching style below) which I often take from the easier homework exercises in real analysis textbooks. From the main part of the textbook, I spend disproportionately more time with the first chapter (combinatorics), because (i) students often have problems with this material, (ii) some of it is critical for subsequent chapters, and (iii) it covers mathematics topics which the students can profit from in other courses, e.g. proof by induction, the binomial and multinomial theorems, and the gamma and beta functions. Most of the material in the remaining chapters gets 'covered', but not necessarily in class. In fact, much of it is necessarily read and learned outside of class, and some material can just be skimmed over, e.g. the derivation of the runs distributions and the inverse hypergeometric moments.

For teaching junior or senior undergraduate courses in probability, I also start with the math appendix, but only cover the 'easier' material, i.e. differentiation, the basics of Riemann integration, and the fundamental theorem of calculus, skipping much of the sections on series and multivariate calculus. I also spend a significant amount of time with Chapter 1, for the reasons mentioned above. For the remaining chapters, I skip most of the more 'exotic' material. For example, in Chapter 2, I omit a discussion of occupancy problems, the more advanced set-theoretic material (σ-fields, etc.), and much of the material in the section on advanced properties (e.g. Poincaré's theorem for $n > 3$, the De Moivre–Jordan theorem, and Bonferroni's general inequality), as well as the proof that $\lim_{i\to\infty} \Pr(A_i) = \Pr(\lim_{i\to\infty} A_i)$.

Teaching style

All lecturers have different styles – some like to 'talk and chalk', i.e. write everything on the blackboard, others use overheads or beamer presentations, etc. I personally prefer the use of overheads for showing the main results, which can then be openly discussed (as opposed to having $n + 1$ people quietly acting like human copy machines), and using the blackboard for excursions or 'live' derivations. As many instructors have learned, students can write, or think, but not both at the same time, so my strategy was to prepare a book which contains almost everything I would possibly write down – so the students don't have to. Of course, if a student is not writing, it does not imply that he or she is thinking, and so to help stimulate discussion and liveliness, I spend about a quarter of the lecture time allowing them to work some problems on their own (or in teams) which we subsequently discuss (or I choose a team to present their findings). What I just described accounts for about two thirds of the lecture; the remaining time takes place in the computer room, letting the students actively work with each other and the PC.

I must now repeat what I said earlier: *not everything in the texts is supposed to be (or could be) covered in the classroom.* Parts should be assigned for reading outside of class, while some of the material can of course be skipped. In my opinion, students at this level should be encouraged to start the transition of learning from lectures, to learning from books and research articles. For high school students and beginning undergraduates, the lecture is the main arena for knowledge transfer, and the book is often used only as a secondary resource, if at all. At the other end of the spectrum, doctoral research students, and professionals in and outside academia, attain most of their new knowledge on their own, by reading (and implementing). I 'assist' this transition by taking some exam questions from material which the students should read on their own.