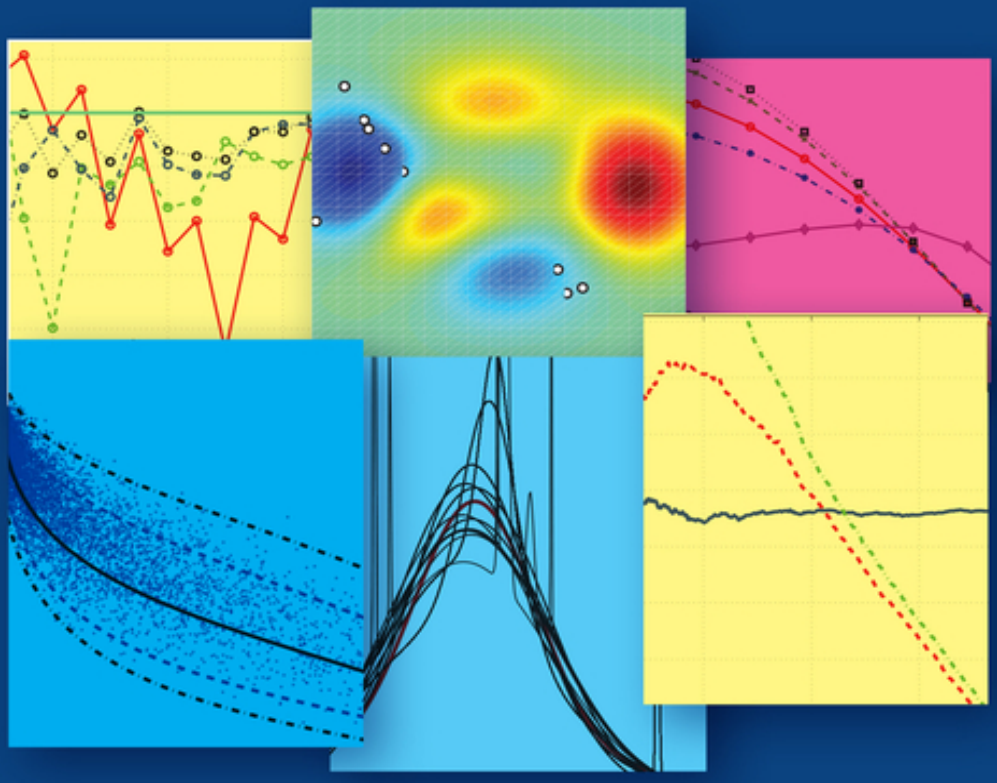
WILEY SERIES IN PROBABILITY AND STATISTICS

Fundamental Statistical Inference

A Computational Approach

Marc S. Paolella



WILEY

Fundamental Statistical Inference

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by *Walter A. Shewhart and Samuel S. Wilks*

Editors: *David J. Balding, Noel A. C. Cressie, Garrett M. Fitzmaurice, Geof H. Givens, Harvey Goldstein, Geert Molenberghs, David W. Scott, Adrian F. M. Smith, Ruey S. Tsay*

Editors Emeriti: *J. Stuart Hunter, Iain M. Johnstone, Joseph B. Kadane, Jozef L. Teugels*

The *Wiley Series in Probability and Statistics* is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches. This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

A complete list of titles in this series can be found at http://www.wiley.com/go/wsps

Fundamental Statistical Inference

A Computational Approach

Marc S. Paolella

Department of Banking and Finance
University of Zurich
Switzerland

WILEY

This edition first published 2018

Registered Offices
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Office
9600 Garsington Road, Oxford, OX4 2DQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloging-in-Publication Data applied for

Hardback ISBN: 9781119417866

Cover design by Wiley
Cover images: Courtesy of Marc S. Paolella

Set in 10/12pt TimesLTStd by SPi Global, Chennai, India

10 9 8 7 6 5 4 3 2 1

Contents

Preface

> Young people today love luxury. They have bad manners, despise authority, have no respect for older people, and chatter when they should be working.
>
> *(Socrates, 470–399 BC)*

This book on statistical inference can be viewed as a continuation of the author's previous two books on probability theory (Paolella, 2006, 2007), hereafter referred to as Books I and II. Of those two, Book I (or any book at a comparable level) is more relevant, in establishing the basics of random variables and distributions as required to understand statistical methodology. Occasional use of material from Book II is made, though most of that required material is reviewed in the appendix herein in order to keep this volume as self-contained as possible. References to those books will be abbreviated as I and II, respectively. For example, Figure 5.1 in (Chapter 5 of) Paolella (2006) is referred to as Figure I.5.1; and similarly for equation references, where (I.5.22) and (II.4.3) refer to equations (5.22) and (4.3) in Paolella (2006) and Paolella (2007) respectively (and both are the Cauchy–Schwarz inequality).

Further prerequisites are the same as those for Book I, namely a solid command of basic undergraduate calculus and matrix algebra, and occasionally very rudimentary concepts from complex analysis, as required for working with characteristic functions. As with Books I and II, a solutions manual to the exercises is available.

Certainly, no measure theory is required, nor any previous exposure to statistical inference, though it would be useful to have had an introductory course in statistics or data analysis. The book is aimed at beginning master's students in statistics, though it is written to be fully accessible to master's students in the social sciences. In particular, I have in mind students in economics and finance, as I provide introductory coverage of some nonstandard topics, notably Chapter 9 on heavy-tailed distributions and tail estimation, and detailed coverage of the mixed normal distribution in Chapter 5.

Naturally, the book can be also used for undergraduates in a mathematics program. For the intended audience of master's students in statistics or the social sciences, the instructor is welcome to skip material that uses concepts from convergence and limit theorems if the target audience is not ready for such mathematics. This is one of the points of this book: such material is included so that, for example, accessible, detailed proofs of the Glivenko–Cantelli theorem and the limiting distribution of the maximum likelihood estimator can be demonstrated at a reasonably rigorous level. The vast majority of the book only requires simple algebra and basic calculus.

In this book, I stick to the independent, identically distributed (i.i.d.) setting, using it as a platform for introducing the major concepts arising in statistics without the additional overhead and complexities associated with, say, (generalized) linear models, survival analysis, copula methods, and time series. This also allows for more in-depth coverage of important topics such as bootstrap techniques, nonparametric inference via the empirical c.d.f., numerical optimization, discrete mixture models, bias-adjusted estimators, tail estimation (as a nice segue into the study of extreme value theory), and the method of indirect inference. A future project, referred to as Book IV, builds on the framework in the present volume and is dedicated to the linear model (regression and ANOVA) and, primarily, time series analysis (univariate ARMAX models), GARCH, and multivariate distributions for modeling and predicting financial asset returns.

Before discussing the contents of this volume, it is important to mention that, similar to Books I and II, the overriding goals are:

(i) to emphasize the practical side of matters by addressing computation issues;
(ii) to motivate students to actively engage in the material by replicating and extending reported results, and to read the literature on topics of their interest;
(iii) to go beyond the standard topics and examples traditionally taught at this level, albeit still within the i.i.d. framework; and
(iv) to set the stage for students intending to pursue further courses in statistical/econometric inference (and quantitative risk management), as well as those embarking on careers as modern data analysts and applied quantitative researchers.

Regarding point (i), I explain to students that computer programming skills are necessary, but far from sufficient, to be successful in applied research. In an occasional lecture dedicated to programming issues, I emphasize (not sarcastically – I do not test computer skills) that it is fully optional, and those students who are truly mathematically talented can skip it, explaining that they will always have programmers in their team (in industry) or PhD students and co-authors (in academics) as resources to do the computer grunt work implementing their theoretical constructs. Oddly, nobody leaves the room.

With respect to point (ii), the reader will notice that some chapters have few (or no) exercises (some have many). This is because I believe the nature of the material presented is such that it offers the student a judicious platform for self experimentation, particularly with respect to numerical implementation. Some of the material could have been packaged as exercises (and much is), though I prefer to illustrate important concepts, distributions, and methods in a detailed way, along with code and graphics, instead of banishing it to the exercises (or, far worse, littering the exercises with trite, useless algebraic manipulations devoid of genuine application) and instead encourage the student to replicate, complement,

and extend the presented material. The reader will no doubt tire at my occasional explicit suggestions for this ("The reader is encouraged … "). One of my role model authors is Hamilton (1994), whose book has *no* exercises, is twice the size of this book, and has been praised as an outstanding presentation of time series. Hamilton clearly intended to *teach* the material in a straightforward, clear way, with highly detailed and accessible derivations. I aspire to a similar approach, as well as adding numeric illustrations and Matlab code.[1]

Regarding point (iii), besides the obvious benefit of giving students a more modern viewpoint on methods and applications in statistics, having a large variety of such is useful for students (and instructors) looking for interesting, relevant topics for master's theses. An example of a nonstandard topic of interest is in Chapter 5, giving a detailed discussion on the problems associated with, and solutions to, estimating the (univariate) discrete mixed normal, via a variety of non-m.l.e. methods (empirical m.g.f., c.f., quantile-based methods, etc.), and the use of the EM algorithm with shrinkage, with its immediate extension to the multivariate case. For the latter, I refer to recent work of mine using the minimum covariance determinant (MCD) for parameter estimation, this also serving as an example of (i) what can be done when, here, the multivariate normal mixture is surely misspecified, and (ii) use of a most likely inconsistent estimator (which outperforms the m.l.e. in terms of density forecasting and portfolio allocation for financial returns data).

Particularly with the less common topics developed in Part III of this book, the result is, like Books I and II, a substantially larger project than some similarly positioned books. It is thus essential to understand that *not everything in the text is supposed to be (or could be) covered in the classroom*, at least not in one semester. In my opinion, students (even in mathematics departments, but particularly those in the social sciences) benefit from having clearly laid out explanations, detailed proofs, illustrative examples, a variety of approaches, introductions to modern techniques, and discussions of important, possibly controversial topics (e.g., the irrelevance of consistent estimators in light of the notion that, in realistic settings, the model is wrong anyway, and changing through time or space; and the arguable superfluousness, if not danger, of the typical hypothesis testing framework), as well as topics that could initially be skipped in a first course, but returned to later or assigned as outside reading, depending on the interests and abilities of the students.

I wish to emphasize that this book is for *teaching*, as (obviously) opposed to being a research monograph, or (less obviously) a dry regurgitation of traditional concepts and examples. An anonymous reviewer of Book I, when I initially submitted it to the publisher Wiley, remarked "it's too much material: It seems the author has written a brain dump." While I like to think I have much more in my head than what was written in that book, he (his gender was indeed disclosed to me) apparently believes that students (let alone instructors) are incapable of assessing what material is core, and what can be deemed "extra," or suitable for reading after the main concepts are mastered. It is trivial to just skip some material, whereas not having it at all results in an admittedly shorter book (who cares, besides arguably the publisher?) that accomplishes far less, and might even give the student a false sense of understanding and competence (which will be painfully revealed in a quant job interview). Fortunately, not everyone agrees with him: Besides heart-warming student feedback over the years on Book I (from master's students) and Book II (from doctoral

[1] While I am at it, Severini (2005) is another book I consider exemplary for teaching at the graduate level, as it is highly detailed and accessible, covers a range of important topics, and is at the same mathematical level as, and has some overlap with, my Book II. Though beware of the typos (which go far beyond his current errata sheet)!

students), I cherish the detailed, insightful, and highly positive reviews of Books I and II, by Harvill (2008, 2009). (I still need to send her flowers.)

The choice of precisely what material to cover (and what not to) is crucial. My decision is to blend "old" and "new," helping to emphasize that the subject has important roots going back over a century, and continues to develop unabated. (The reader will quickly see my adoration of Karl Pearson and Ronald Fisher, the founders of our subject; both fascinating, albeit complicated personalities, polymaths, and, at times, adversaries.)

Chapter 1 starts modestly with basic concepts of point estimation, and includes my diatribe on the unnecessary obsession with consistent estimators *in some contexts*. The same chapter then progresses to a very basic development of the single and double bootstrap for computing confidence intervals. If one were to imagine that the field of statistics somehow did not exist, I argue that a student versed in basic probability theory and with access to, and skills with, modern computing power would immediately discover on his/her own the (percentile, single, parametric) bootstrap as a natural way of determining a confidence interval. As such, it is presented before the usual asymptotic Wald intervals and analytic methods. The latter *are* important, as conceptual entities, and work well when applicable, but their relevance to the tasks and goals faced by the new generation of students dealing with modern, sophisticated models and/or big data applications is difficult to motivate.

Chapter 2 spends more time than usual on the empirical c.d.f., and shows, among other things, two simple, instructive proofs of the Glivenko–Cantelli theorem, as opposed to not mentioning it at all, or, perhaps worse, the dreaded "it can be shown . . . ." Besides being a fundamental result of enormous importance, this serves as a primer for students interested in point processes. The chapter also introduces the major concepts associated with hypothesis testing and p-values, within the context of distribution testing. I argue in the chapter that this is a very good platform for use of hypothesis testing, and then provide yet another diatribe about why I shy away from presenting the standard material on the subject when applied to parameters of a model.

The rest of Part I consists of three related chapters on parameter estimation. The five chapters of Part I are what I consider to be the core of fundamental statistical inference, and are best read in the order presented, though Chapter 4 can be studied independently of other chapters and possibly assigned as outside reading.

The cornerstone Chapter 3 introduces likelihood, and contains many standard examples, but also some nonstandard material, such as the MCD method to emphasize the relevance of robust statistics and the pernicious issue of masking. Chapter 4 is about numerical optimization, motivating the development of multivariate Hessian-based techniques via repeated application of simple, univariate methods that every student understands, such as bisection. This chapter also includes discussions, with Matlab code, for genetic algorithms and why they are of such importance in many applications.

Chapter 5 is rather unique, using the mixed normal distribution (itself of great relevance, notably now in machine learning) as a platform for showing numerous other methods of point estimation that can outperform the m.l.e. in smaller samples, serve as starting values for computing the m.l.e., or be used when the likelihood is not accessible. Chapter 5 also introduces the use of shrinkage as a penalty factor in the likelihood, and the EM algorithm in the context of the discrete mixed normal distribution.

The chapters of Part II are written to be more or less orthogonal. The instructor (or student working independently) can choose among them, based on his/her interests. The lengthy Chapter 6, on Q-Q plots and distribution testing, builds on the material in

Chapter 2. It emphasizes the distinction between one-at-a-time and simultaneous intervals, and presents various tests for composite normality, including a test of mine, conveniently abbreviated MSP: it is not the most powerful test against all alternatives (no such test yet exists), but its *development* illustrates numerous important concepts – and that is the point of the book.

Chapters 7 and 8 (and Section 3.2 on the univariate and multivariate Cramér–Rao lower bound) are the most "classic," on well-worn results for point and interval estimation, respectively, though Chapter 7 contains some more modern techniques for bias reduction and new classes of estimators. As most of this is standard textbook material at this level, the goal was to develop it in the clearest way possible, with accessible, detailed (sometimes multiple) proofs, and a large variety of examples and end-of-chapter algebraic exercises. There are now several excellent advanced books on mathematical statistics: Schervish (1995), Lehmann and Casella (1998), Shao (2003), and Robert (2007) come to mind, and it is pointless to compete with them, nor is it the goal of this book to do so.

The two chapters of Part III are more associated with financial econometrics and quantitative risk management, though I believe the material should be of interest to a general statistics audience. Chapter 9 covers much ground. It introduces the basics of tail estimation, with a simple derivation of the Hill estimator, discussion of its problems (along with customary Hill horror plots), and enough of a literature review for the interested student to pursue. Also in this chapter (and in Section A.16), the (univariate, asymmetric) stable Paretian distribution receives much attention: I dispel myths about its inapplicability or difficulty in estimation, and discuss several methods for the latter, as well as including recent work on *testing* the stability assumption.

The relatively short Chapter 10 introduces the concept and methodology of indirect inference, a topic rarely presented at this level but of fundamental importance in a variety of challenging contexts. One of the examples used for its demonstration involves the randomized response technique for dealing with awkward questions in surveys (this being notably a topic squarely within statistics, as opposed to econometrics). This elegant solution for obtaining point estimators appears to be new.

The appendix is primarily a review of important and useful facts from probability theory, condensed from Books I and II (where more detail can obviously be found), with its equations being referenced throughout, thus helping to keep this book as self-contained as possible. It also includes a large section of exercises, many of which are not in Books I or II and some of which are challenging, enabling the student to refresh, extend, and self-assess his/her abilities, and/or enabling the instructor to give an initial exam to determine if the student has the requisite knowledge. All the solutions are provided at the end of the appendix.

This appendix also includes some new material not found in Books I and II, such as (i) more results, with proofs, on convergence in distribution (as required for proving the asymptotic properties of the m.l.e.); (ii) a detailed section on expected shortfall (ES), including Stein's lemma, as required for illustrating the shrinkage estimator in Section 5.4; (iii) additional Matlab programs (not in Book II) for the p.d.f., c.d.f., quantiles and ES of the asymmetric stable; and (iv) among the exercises, some potentially useful ones, such as saddlepoint approximations and characteristic function inversion for computing the distribution and ES of a convolution of independent skew-normal random variables.

Numerous topics of relevance were omitted (and some notes deleted – which would delight my "brain dump" accuser), such as ancillarity, hierarchical models, rank and permutation tests, and, most notably, Bayesian methodology. For the latter, there are now many good textbooks on the topic, in both pure statistics and also econometrics, and the last thing I want is that the reader ignore the Bayesian approach. I think a solid grounding in basic principles, likelihood-based inference, and a strong command of computing serve as an excellent background for pursuing dedicated works on Bayesian methodology. Section 5.1.6 does introduce the idea of quasi-Bayesian estimation and its connection to shrinkage estimation, and illustrates (without needing to break the proverbial full Bayesian egg)[2] the effectiveness and importance of these methods.

With respect to computing, I chose (no doubt to the annoyance of some) Matlab as the vehicle for prototyping, though I strongly encourage readers versed in R to continue using R, or Python, or even to learn the relatively new and highly promising language Julia. Unlike with the Matlab codes in Book I, I do not (so far) provide R translations, though every attempt was made to use the most basic coding and data structures possible, so that translations should be straightforward, and also occasionally separating the very-specific-to-Matlab commands, such as for graphics.

No single book will ever cover every topic or aspect the author would like. As a complement to this book, I recommend students concurrently read some sections of Pawitan (2001) (with an updated and paperback version now available), Davison (2003), and Casella and Berger (2002), three books that I hold as exemplary; they cover additional topics I have omitted, and, in the case of the former two, contain far more examples with real data.

I recall a review of a book in financial econometrics (which I had best not name). Paraphrasing, the reviewer stated that academic books tend to have one of two purposes: (i) to teach the material; or (ii) to impress the reader and, particularly, colleagues with the authors' knowledge. The reviewer then went on to say how the book accomplished neither. My hope is that the reader and instructor understand my goal to be the former, with little regard for the latter: As emphasized above, the book contains much material, computer codes, and touches upon some recent developments. When proofs are shown, they are simple and detailed. I wrote the book for motivated students who want straightforward explanations, clear demonstrations, and discussions of more modern topics, particularly in a non-Gaussian setting. My guiding principle was to write the book that I would have killed for as a graduate student.

Some acknowledgments are in order. I owe an enormous amount of gratitude to the excellent scientists and instructors I worked with during and after my graduate studies. Alphabetically, these include professors Peter Brockwell, Ronald Butler, Richard Davis, Hariharan (Hari) Iyer, Stefan Mittnik, and Svetlozar (Zari) Rachev. All of these individuals also have textbooks that I highly recommend, and some of which will be mentioned in the preface to book IV. As the years go by, the proverbial circle starts to close, and I have my own doctoral students, all of whom have contributed in various ways to my book projects. Notable mention goes to Simon Broda, Pawel Polak (both of whom are now professors themselves) and (current PhD students) Marco Gambacciani and Patrick Walker, who, along with professors Kai Carstensen, Walter Farkas, Markus Haas, Alexander McNeil, Nuttanan (Nate) Wichitaksorn, and Michael Wolf, have read parts of this manuscript (and

[2] This refers to the oft-quoted statement in Savage (1961, p. 578) that Fisher's fiducial inferential method is "a bold attempt to make the Bayesian omelet without breaking the Bayesian eggs".

book IV) and helped tease out mistakes and improve the presentation. Finally, I am indebted to my copy editor Richard Leigh from Wiley, who read every line of the book, checked every graphic and bibliography reference, and made uncountable corrections and suggestions to the scientific English presentation, as well as (embarrassingly) caught a few math mistakes. I have obviously suggested to the editor to have him work on my book IV (and double his salary).

My gratitude to these individuals cannot be overstated.

Part I

Essential Concepts in Statistics

1

Introducing Point and Interval Estimation

> The discussions of theoretical statistics may be regarded as alternating between problems of estimation and problems of distribution. In the first place a method of calculating one of the population parameters is devised from common-sense considerations: we next require to know its probable error, and therefore an approximate solution of the distribution, in samples, of the statistics calculated.
>
> *(R. A. Fisher, 1922, reproduced in Kotz and Johnson, 1992)*

This chapter and the next two introduce the primary tools and concepts underlying most all problems in statistical inference. We restrict ourselves herein to the independent, identically distributed (i.i.d.) framework, in order to emphasize the fundamental concepts without the need for addressing the additional issues and complexities associated with the workhorse models of statistics, such as linear models, analysis of variance, design of experiments, and time series. The overriding goal is to extract relevant information from the available sample in order to learn about the underlying population from which it was drawn.

We begin with the basic definitions associated with point estimation, and introduce the maximum likelihood estimator (m.l.e.). We will have more to say about point estimation and m.l.e.s in Chapters 3 and 5. The remainder of the chapter is dedicated to individual parameter confidence intervals (c.i.s), restricting attention to the intuitive use of computer-intensive methods for their construction, as they are generally applicable and, for more complex problems, often the only available choice. In particular, a natural progression is made from simulation to the parametric bootstrap, to the nonparametric bootstrap, to the double nonparametric bootstrap, and finally to the double bootstrap with analytic inner loop, the latter using techniques from Chapter 8.

Fundamental Statistical Inference: A Computational Approach, First Edition. Marc S. Paolella.

1.1 POINT ESTIMATION

To introduce the notion of parameter estimation from a sample of data, we make use of two simple models, the Bernoulli and geometric.

1.1.1 Bernoulli Model

Consider an idealized experiment that consists of randomly drawing a marble from an urn containing R red and W white marbles; its color is noted and it is then placed back into the urn. This is repeated n times, whereby n is a known, finite constant, *but R and W are unknown*. This corresponds to a sequence of Bernoulli trials with unknown probability $p = R/(R+W)$ or $p = W/(R+W)$, depending on what one wants to consider a "success." Assuming the former, let X_i, $i = 1, \ldots, n$, denote the outcomes of the experiment, with $X_i \overset{\text{i.i.d.}}{\sim} \text{Bern}(p)$, each with support $\mathcal{S} = \{0, 1\}$. The ultimate goal is to determine the value of p. If n is finite (as reality often dictates), this will be an impossible task. Instead, we content ourselves with attempting to infer as much information as possible about the value of p.

As a starting point, in line with Fisher's "common-sense considerations" in the opening quote, it seems reasonable to examine the proportion of successes. That is, we would compute s/n, where s is the observed number of successes. The value s/n is referred to as a **point estimate** of p and denoted $\hat{p}$, pronounced "p hat." Sometimes it is advantageous to write $\hat{p}_n$, where the subscript indicates the sample size. From the way in which the experiment is defined, it should be clear that s is a realization from the binomial random variable (r.v.) $S = \sum_{i=1}^{n} X_i \sim \text{Bin}(n, p)$. To emphasize this, we also write $\hat{p} = S/n$ and call this a **point estimator** of p, the distinction being that a point estimator is a random variable, while a point estimate is a realization of this random variable resulting from the outcome of a particular experiment.

Note that the same notation of adding a "hat" to the parameter of interest is used to denote both estimate and estimator, as this is the common standard. However, the distinction between estimate and estimator is crucial when attempting to assess the properties of $\hat{p}$ (e.g., is it correct on average?) and compare its performance to other possible estimators (e.g., is one point estimator more likely to be correct than the other?). For instance, $\mathbb{E}[s/n] = s/n$, that is, s/n is a post-experiment constant, while $\mathbb{E}[S/n] = (np)/n = p$ can be computed before or after the experiment takes place. In this case, estimator S/n is said to be **(mean) unbiased**.

More formally, let $\hat{\theta}$ be a point estimator of the finite, fixed, unknown parameter $\theta \in \Theta \subset \mathbb{R}$ such that $\mathbb{E}[\hat{\theta}]$ exists. Then:

> The point estimator $\hat{\theta}$ is **(mean) unbiased** (with respect to the set Θ) if its expected value is θ (for all $\theta \in \Theta$); otherwise it is **(mean) biased** with $\text{bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$.

Generally speaking, mean unbiasedness is a desirable property because it implies that we are "correct on average," where the "average" refers to the hypothetical idea of repeating the experiment infinitely often – something that of course does not actually happen in reality. An impressive theoretical framework in mathematical statistics was developed, starting in the 1950s, for the derivation and study of unbiased estimators with minimum variance; see Chapter 7, especially Section 7.2. It is often the case, however, that estimators can be found

that are biased, but, by virtue of having a lower variance, wind up having a lower mean squared error, as seen from (1.2) directly below. This concept is well known, and reflected, for example, in Shao and Tu (1995, p. 67), stating "We need to balance the advantage of unbiasedness against the drawbacks of a large mean squared error." Another type of unbiasedness involves using the median instead of the mean. See Section 7.4.2 for details on median-unbiased estimators.

For the binomial example, the variance of estimator $\hat{p}$ is

$$\mathbb{V}(\hat{p}) = \mathbb{V}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = n\frac{p(1-p)}{n^2} = \frac{p(1-p)}{n}, \tag{1.1}$$

and it clearly goes to zero as the sample size increases. This is also desirable, because, as more samples are collected, the amount of information from which p is to be inferred is growing. This concept is referred to as consistency; recalling the definition of convergence in probability from (A.254) and the weak law of large numbers (A.255), the following definition should seem natural:

> An estimator $\hat{\theta}_n$ based on a sample of n observations is **weakly consistent** (with respect to Θ) if, as $n \to \infty$, $\Pr(|\hat{\theta}_n - \theta| > \epsilon) \to 0$ for any $\epsilon > 0$ (and all $\theta \in \Theta$).

Observe that an estimator can be (mean) unbiased but not consistent: if $X_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu, 1)$, $i = 1, \ldots, n$, then the estimator $\hat{\mu} = X_1$ is unbiased, but it does not converge to μ as the sample size increases.

Another popular measure of the quality of an estimator is its expected squared deviation from the true value, called *mean squared error*, or *m.s.e.*:

> The **mean squared error** of the estimator $\hat{\theta}$ is defined as $\mathbb{E}[(\hat{\theta} - \theta)^2]$.

An important decomposition of the m.s.e. is as follows. With $\Xi = \mathbb{E}[\hat{\theta}]$,

$$\begin{aligned}
\text{m.s.e.}\,(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{E}[(\hat{\theta} - \Xi + \Xi - \theta)^2] \\
&= \mathbb{E}[(\hat{\theta} - \Xi)^2] + \mathbb{E}[(\Xi - \theta)^2] + \text{cross-term, which is zero} \\
&= \mathbb{E}[(\hat{\theta} - \Xi)^2] + (\Xi - \theta)^2 = \mathbb{V}(\hat{\theta}) + [\text{bias}\,(\hat{\theta})]^2.
\end{aligned} \tag{1.2}$$

The reader should quickly verify that the cross-term is indeed zero. Note that, for an unbiased estimator, its m.s.e. and variance are equal. As the estimator $\hat{\theta}$ is a function of the data, it is itself a random variable. With $f = f_{\hat{\theta}}$ the p.d.f. of $\hat{\theta}$, we can write $\Pr(|\hat{\theta} - \theta| > \epsilon)$ for any $\epsilon > 0$ as

$$\int_{|t-\theta|>\epsilon} f(t)\,\mathrm{d}t \le \int_{|t-\theta|>\epsilon} \frac{(t-\theta)^2}{\epsilon^2} f(t)\,\mathrm{d}t \le \int_{-\infty}^{\infty} \frac{(t-\theta)^2}{\epsilon^2} f(t)\,\mathrm{d}t = \frac{\mathbb{E}[(\hat{\theta} - \theta)^2]}{\epsilon^2},$$

so that $\hat{\theta}$ is weakly consistent if m.s.e. $(\hat{\theta}) \to 0$.

The estimator $\hat{p} = S/n$ for the Bernoulli model is rather intuitive and virtually presents itself as being a good estimator of p. It turns out that this $\hat{p}$ coincides with the estimator we obtain when applying a very general and powerful method of obtaining an estimator for an

unknown parameter of a statistical model. We briefly introduce this method now, and will have more to say about it in Section 3.1.

The **likelihood function** $\mathcal{L}(\theta; \mathbf{x})$ is the joint density of a sample $\mathbf{X} = (X_1, \ldots, X_n)$ as a function of the (for now, scalar) parameter θ, for fixed sample values $\mathbf{X} = \mathbf{x}$. That is, $\mathcal{L}(\theta; \mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}; \theta)$, where $f_{\mathbf{X}}$ is the p.m.f. or p.d.f. of $\mathbf{X}$. Let $\ell(\theta; \mathbf{x}) = \log \mathcal{L}(\theta; \mathbf{x})$,[1] and write just $\ell(\theta)$ when the data are clear from the context. Denote the first and second derivatives of $\ell(\theta)$ with respect to θ by $\dot{\ell}(\theta)$ and $\ddot{\ell}(\theta)$, respectively. The **maximum likelihood estimate**, abbreviated m.l.e. and denoted by $\hat{\theta}$ (or, to distinguish it from other estimates, $\hat{\theta}_{\mathrm{ML}}$), is that value of θ that maximizes the likelihood function for a given data set $\mathbf{x}$. The **maximum likelihood estimator** (as opposed to *estimate*) is the function of the X_i, also denoted $\hat{\theta}_{\mathrm{ML}}$, that yields the m.l.e. for an observed data set $\mathbf{x}$.

In many cases of interest (including the Bernoulli and geometric examples in this chapter), the m.l.e. satisfies $\dot{\ell}(\hat{\theta}) = 0$ and $\ddot{\ell}(\hat{\theta}) < 0$. For example, with $X_i \overset{\text{i.i.d.}}{\sim} \mathrm{Bern}(\theta)$, $i = 1, \ldots, n$, the likelihood is

$$\mathcal{L}(\theta; \mathbf{x}) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \mathbb{I}_{\{0,1\}}(x_i) = \theta^s (1-\theta)^{n-s} \mathbb{I}_{\{0,1,\ldots,n\}}(s),$$

where $s = \sum_{i=1}^{n} x_i$. Then

$$\dot{\ell}(\theta) = \frac{s}{\theta} - \frac{n-s}{1-\theta} \quad \text{and} \quad \ddot{\ell}(\theta) = -\frac{s}{\theta^2} - \frac{n-s}{(1-\theta)^2},$$

from which it follows (by setting $\dot{\ell}(\hat{\theta}) = 0$ and confirming $\ddot{\ell}(\hat{\theta}) < 0$) that $\hat{\theta}_{\mathrm{ML}} = S/n$ is the m.l.e. It is easy to see that $\hat{\theta}_{\mathrm{ML}}$ is unbiased.

1.1.2 Geometric Model

As in the binomial case, independent draws with replacement are conducted from an urn with R red and W white marbles. However, now the number of trials is not fixed in advance; sampling continues until r red marbles have been drawn. What can be said about $p = R/(R+W)$? Let the r.v. X be the number of necessary trials. From the sampling structure, X follows a negative binomial distribution, $X \sim \mathrm{NBin}(r, p)$, with p.m.f.

$$f_X(x; r, p) = \binom{x-1}{r-1} p^r (1-p)^{x-r} \mathbb{I}_{\{r, r+1, \ldots\}}(x). \tag{1.3}$$

Recall that X can be expressed as the sum of r i.i.d. geometric r.v.s, say $X = \sum_{i=1}^{r} G_i$, where $G_i \overset{\text{i.i.d.}}{\sim} \mathrm{Geo}(p)$, each with support $\{1, 2, \ldots\}$.

This decomposition is important because it allows us to imagine that sampling occurs not necessarily consecutively in time until r successes occur, but rather as r independent (and possibly concurrent) geometric trials using urns with the same red to white ratio, that is, the same p. For example, interest might center on how long it takes a woman to become pregnant using a particular method of assistance (e.g., temperature measurements or hormone treatment). This is worth making an example, as we will refer to it more than once.

[1] Throughout this book, log refers to base e unless otherwise specified.

Example 1.1 (Geometric) *Let* $G_i \overset{\text{i.i.d.}}{\sim} \text{Geo}(\theta)$, $i = 1, \ldots, n$, *with typical p.m.f.*

$$f_G(x;\theta) = \theta(1-\theta)^{x-1}\mathbb{I}_{\{1,2,\ldots\}}(x), \quad \theta \in \Theta = (0,1).$$

Then $\ell(\theta;\mathbf{x})$, *the log-likelihood of the sample* $\mathbf{x} = (x_1, \ldots, x_n)$, *and its first derivative,* $\dot{\ell}(\theta;\mathbf{x})$, *are, with* $s = \sum_{i=1}^n x_i$,

$$\ell(\theta;\mathbf{x}) = n\log(\theta) + \log(1-\theta)\sum_{i=1}^n (x_i - 1), \quad \dot{\ell}(\theta;\mathbf{x}) = \frac{n}{\theta} - \frac{s-n}{1-\theta}.$$

Solving the equation $\dot{\ell}(\theta;\mathbf{x}) = 0$ *and confirming* $\ddot{\ell}(\hat{\theta}) < 0$ *gives* $\hat{\theta}_{ML} = n/S = 1/\bar{G}$. *We will see below and in Section 7.3 that the m.l.e. is not unbiased.*[2] ■

Imagine a study in which each of r couples (independently of each other) attempts to conceive each month until they succeed. In the $r = 1$ case, $X = G_1 \sim \text{Geo}(p)$ and, recalling that $\mathbb{E}[G_1] = 1/p$, an intuitive point estimator of p is $1/G_1$. Interest centers on developing a point estimator for the $r > 1$ case. Of course, in this simple structure, one would just compute the m.l.e. However, we use this easy case to illustrate how one might proceed when simple answers are not immediately available, and some thinking and creativity are required.

Based on the result for $r = 1$, one idea for the $r > 1$ case would be to use the average of the $1/G_i$ values, $r^{-1}\sum_{i=1}^r G_i^{-1}$, which we denote by $\hat{p}_1$. Another candidate is $\hat{p}_2 = 1/\bar{G} = r/\sum_{i=1}^r G_i = r/X$. This happens to be the m.l.e. from Example 1.2. Note that both of these estimators reduce to $1/G_1$ when $r = 1$. We also consider the nonobvious point estimator $\hat{p}_3 = (r-1)/(X-1)$. It will be derived in Section 7.3, and is only useful for $r > 1$.

Instead of algebraically determining the mean and variance of the $\hat{p}_i$, $i = 1, 2, 3$, we will begin our practice of letting the computer do the work. The program in Listing 1.1 computes the three point estimators for a simulated set of G_i; it repeats this `sim` = 10,000 times, and the resulting sample mean and variance of these simulated estimates approximate the true mean and variance.

To illustrate, Figure 1.1 shows the histograms of the simulated point estimators for the case with $p = 0.3$ and $r = 5$. From these, the large upward bias of $\hat{p}_1$ is particularly clear.

```
function [p1vec, p2vec, p3vec]=geometricparameterestimate(p,r,sim)
p1vec = zeros(sim,1); p2vec = p1vec; p3vec = p1vec;
for s=1:sim
  gvec=geornd(p,[r 1]) +1;
  p1 = mean(1./gvec); p2 = 1/mean(gvec); p3 = (r-1) / (sum(gvec)-1);
  p1vec(s) = p1; p2vec(s) = p2; p3vec(s) = p3;
end
bias1 = mean(p1vec)-p, bias2 = mean(p2vec)-p, bias3 = mean(p3vec)-p
var1 = var(p1vec), var2 = var(p2vec), var3 = var(p3vec)
mse1 = var1+bias1^2, mse2 = var2+bias2^2, mse3 = var3+bias3^2
```

Program Listing 1.1: Simulates three point estimators for p in the i.i.d. geometric model. Calling the function with $p = 0.3$ and $r = 5$ corresponds to the true probability of success being 0.3 and using five couples in the experiment.

[2] We use the symbol ■ to denote the end of proofs of theorems, as well as examples and remarks, acknowledging that it is traditionally only used for the former, as popularized by Paul Halmos.

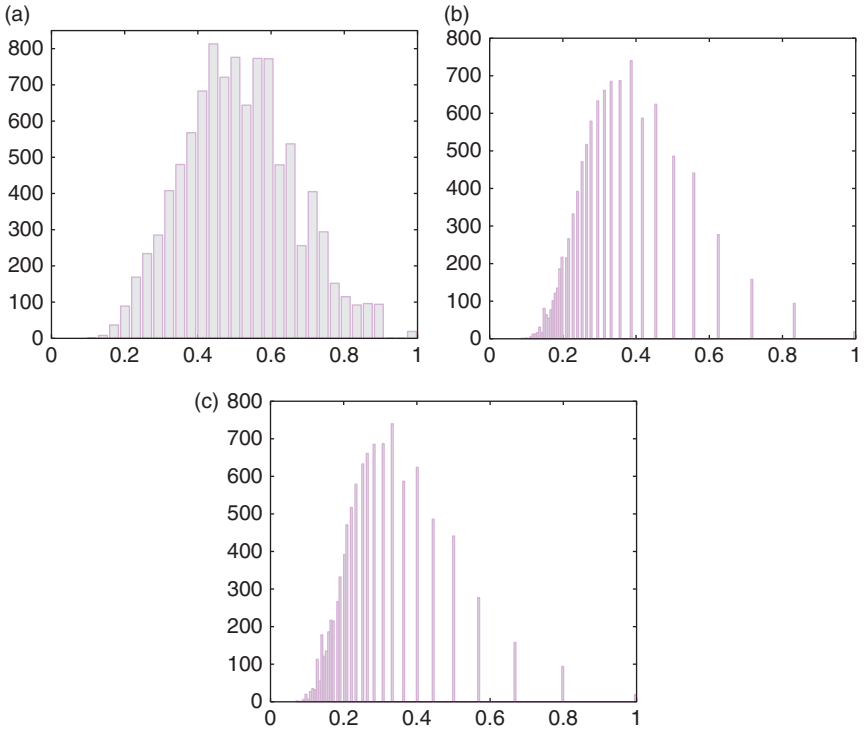


Figure 1.1 *Distribution of point estimators $\hat{p}_1$ (a), $\hat{p}_2$ (b), and $\hat{p}_3$ (c) using output from the program in Listing 1.1 with $p = 0.3$ and $r = 5$, based on simulation with 10,000 replications.*

The discrete nature of $\hat{p}_2$ and $\hat{p}_3$ arises because these two estimators first compute the sum of the observations and then take reciprocals, so that computation of their p.m.f.s is easy. As an example, $\hat{p}_3 = 0.4 \Leftrightarrow X = 11$, which, from (1.3), has probability 0.06, so that approximately 600 of the simulated values depicted in the histogram of $\hat{p}_3$ should be 0.4; there are 624 in the histogram. Similarly, $\hat{p}_3 = 0.8 \Leftrightarrow X = 6$, with probability 0.008505, and 94 in the histogram.

As p increases towards one, the number of points in the supports of $\hat{p}_2$ and $\hat{p}_3$ decreases. This is illustrated in Figure 1.2, showing histograms of $\hat{p}_2$ for $r = 10$ and four different values of p. The code used to make the plots is given in Listing 1.2. Observe how we avoid use of the FOR loop (as was used in Listing 1.1) for generating the 1 million replications, thus providing a significant speed increase. (The use of the `eval` command with concatenated text strings is also demonstrated.)

For the simulation of $\hat{p}_1$, $\hat{p}_2$, and $\hat{p}_3$ from Listing 1.1, with $p = 0.3$ and $r = 5$, the results are shown in the first numeric row of Table 1.1. We see that $\hat{p}_1$ has almost five times the bias of $\hat{p}_2$, while $\hat{p}_2$ has over 100 times the bias of $\hat{p}_3$. The variance of $\hat{p}_1$ is slightly larger than those of $\hat{p}_2$ and $\hat{p}_3$, which are nearly the same. By combining these according to (1.2), it is clear that the m.s.e. will be smallest for $\hat{p}_3$, as also shown in the table. The next row shows the results using a larger sample of 15 couples. While the bias of $\hat{p}_1$ stays the same, those of $\hat{p}_2$ and $\hat{p}_3$ decrease. For all point estimators, the variance decreases.

It turns out that, as the number of couples, r, tends towards infinity, the variance of all the estimators goes to zero, while the bias of $\hat{p}_1$ stays at 0.22 and that of $\hat{p}_2$ goes to zero.

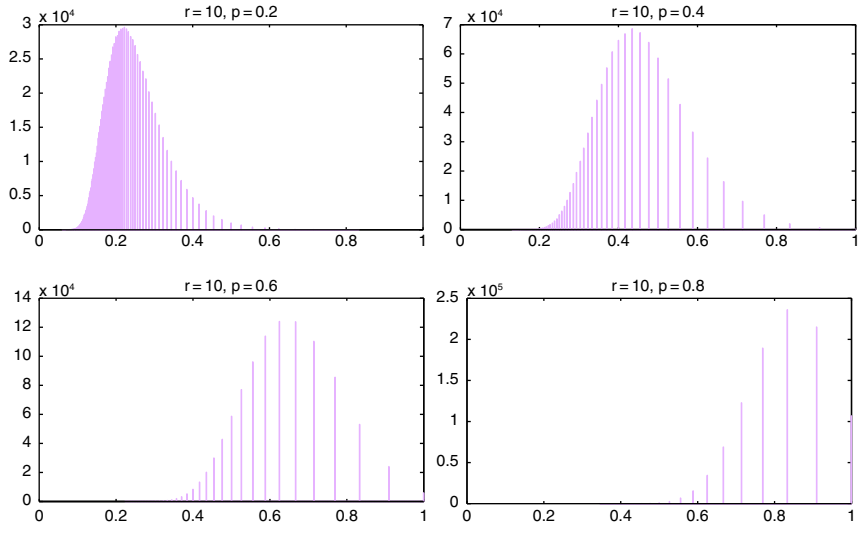


Figure 1.2 *Histogram of point estimator $\hat{p}_2$ for $r = 10$ and four values of p, based on simulation with 1 million replications.*

```
B=1e6; r=10;
for p=0.2:0.2:0.8
  phatvec = 1./mean(geornd(p,[r B])+1); % the MLE
  [histcount, histgrd] =hist(phatvec,1000);
  figure, h1=bar(histgrd,histcount); set(gca,'fontsize',16), xlim([0 1])
  title([ 'r=',int2str(r),', p=',num2str(p)])
  set(h1,'facecolor',[0.94 0.94 0.94],'edgecolor',[0.9 0.7 1])
  eval([ 'print -depsc phatforgeogetsmorediscretep',int2str(10*p)])
end
```

Program Listing 1.2: Generates the graphs in Figure 1.2.

Hence, we say that $\hat{p}_2$ is **asymptotically unbiased**. We will see in Section 7.3 that $\hat{p}_3$ is unbiased – not just asymptotically, but for all $0 < p \leq 1$ and any $r > 1$. This implies that the value 0.0004 in the $\hat{p}_3$ bias column of the table just reflects **sampling error** resulting from using only 10,000 replications in the simulation. In comparison, then, point estimator $\hat{p}_3$ seems to be preferred with respect to all three criteria.

The lower portion of Table 1.1 shows similar results using $p = 0.7$. Again, $\hat{p}_1$ is highly biased, while, comparatively speaking, the bias of $\hat{p}_2$ is much smaller and diminishes with growing sample size r. The bias of $\hat{p}_3$ appears very small and, as already mentioned, is theoretically zero. The interesting thing about this choice of p is that the variance of $\hat{p}_2$ is smaller than that of $\hat{p}_3$. In fact, this reduction in variance causes the m.s.e. of $\hat{p}_2$ to be smaller than that of $\hat{p}_3$ even though the bias of $\hat{p}_3$ is essentially zero. This demonstrates two important points:

(i) An unbiased point estimator need not have the smallest m.s.e.

(ii) The relative properties of point estimators may change with the unknown parameter of interest.

TABLE 1.1 Comparison of three point estimators for the geometric model

		bias			variance			m.s.e.		
p	r	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$
0.3	5	0.22	0.045	0.00040	0.023	0.018	0.017	0.070	0.020	0.017
0.3	15	0.22	0.015	0.00041	0.0077	0.0048	0.0045	0.055	0.0050	0.0045
0.7	5	0.13	0.040	0.00057	0.014	0.027	0.033	0.031	0.028	0.033
0.7	15	0.13	0.013	−0.00078	0.0046	0.0096	0.0102	0.022	0.0098	0.010

Having demonstrated these two facts using just the values in Table 1.1, it would be desirable to graphically depict the m.s.e. of estimators $\hat{p}_2$ and $\hat{p}_3$ as a function of p, for several sample sizes. This is shown in Figure 1.3, from which we see that m.s.e.$(\hat{p}_2) <$ m.s.e.$(\hat{p}_3)$ for (roughly) $p > 0.5$, but as the sample size increases, the difference in m.s.e. of the two estimators becomes negligible.

Facts (i) and (ii) mentioned above complicate the comparison of estimators. Some structure can be put on the problem if we restrict attention to unbiased estimators. Then, minimizing the m.s.e. is the same as minimizing the variance; this gives rise to the following concepts:

> An unbiased estimator, say $\hat{\theta}_{\text{eff}}$, is **efficient** (with respect to Θ) if it has the smallest possible variance of all unbiased estimators (for all $\theta \in \Theta$).
>
> The **efficiency** of an unbiased estimator $\hat{\theta}$ is $\text{Eff}(\hat{\theta}, \theta) = \mathbb{V}(\hat{\theta}_{\text{eff}})/\mathbb{V}(\hat{\theta})$.

We will see later (Chapter 7) that the estimator $\hat{p}_3$ used above is efficient.

In many realistic problems, there may be no unbiased estimators, or no efficient one; and if there is, like $\hat{p}_3$ above, it might not have the smallest m.s.e. over all or parts of Θ. This somewhat diminishes the value of the efficiency concept defined above. All is not lost, however. In many cases of interest, the m.l.e. has the property that, asymptotically, it is (unbiased and) efficient. As such, it serves as a natural benchmark with which to compare competing estimators. We expect that, with increasing sample size, the m.l.e. will eventually be as good as, or better than, all other estimators, with respect to m.s.e., for all $\theta \in \Theta$. This certainly does not imply that the m.l.e. is the best estimator in finite samples, as we see in Figure 1.3 comparing the m.l.e. $\hat{p}_2$ to the efficient estimator $\hat{p}_3$. (Other cases in which the

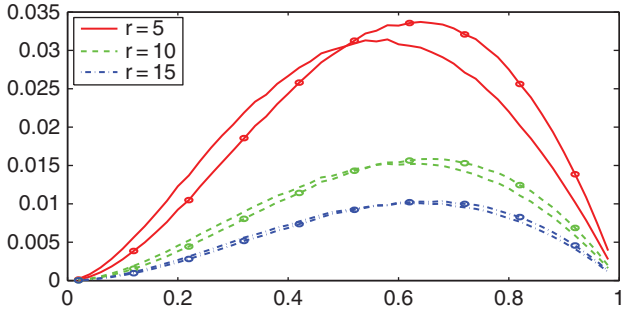


Figure 1.3 *The m.s.e. of estimators $\hat{p}_2$ (lines) and $\hat{p}_3$ (lines with circles) for parameter p in the geometric model, as a function of p, for three sample sizes, obtained by simulation with 100,000 replications.*