Preface

This book developed out of my year-long course on asymptotic theory at Purdue University. To some extent, the topics coincide with what I cover in that course. There are already a number of well-known books on asymptotics. This book is quite different. It covers more topics in one source than are available in any other single book on asymptotic theory. Numerous topics covered in this book are available in the literature in a scattered manner, and they are brought together under one umbrella in this book. Asymptotic theory is a central unifying theme in probability and statistics. My main goal in writing this book is to give its readers a feel for the incredible scope and reach of asymptotics. I have tried to write this book in a way that is accessible and to make the reader appreciate the beauty of theory and the insights that only theory can provide.

Essentially every theorem in the book comes with at least one reference, preceding or following the statement of the theorem. In addition, I have provided a separate theorem-by-theorem reference as an entry on its own in the front of the book to make it extremely convenient for the reader to find a proof that was not provided in the text. Also particularly worth mentioning is a collection of nearly 300 practically useful inequalities that I have collected together from numerous sources. This is appended at the very end of the book. Almost every inequality in this collection comes with at least one reference. I have often preferred to cite a book rather than an original publication for these inequalities, particularly if the book contained many of the inequalities that I present. I also emphasize in this book conceptual discussion of issues, working out many examples and providing a good collection of unusual exercises. Another feature of this book is the guidance to the literature for someone who wishes to dig deeper into the topic of a particular chapter. I have tried to make the chapter-by-chapter bibliographies both modern and representative. The book has 574 exercises and 293 worked-out examples. I have marked the more nonroutine exercises with an asterisk.

I hope that this book is useful as a graduate text, for independent reading, and as a general and nearly encyclopedic research reference on asymptotic theory. It should be possible to design graduate-level courses using this book with emphasis on parametric methods or nonparametric methods, on classic topics or more current topics, on frequentist topics or Bayesian topics, or even on probability theory. For the benefit of instructors, I have provided recommended chapters for ten different one-semester courses, with emphasis on different themes. I hope that this provides some useful guidance toward designing courses based on this book.

Because the book covers a very broad range of topics, I do not have a uniform answer for what background I assume for a reader of this book. For most chapters, a knowledge of advanced calculus and linear algebra is enough to enable the reader to follow the material. However, some chapters require some use of measure theory and advanced analysis and some exposure to stochastic processes. One course on statistical theory at the level of Bickel and Doksum (cited in Chapter 3 of this volume) or Casella and Berger (1990) and one on probability at the level of Hoel, Port, and Stone (1971) or Durrett (1994) are certainly needed to follow the discussion in this book. Chapter 1 is essentially a review of somewhat more advanced probability should one need it. The more advanced chapters in this book can be much better appreciated if one has had courses on the two books of Erich Lehmann (Lehmann and Casella (cited in Chapter 16), Lehmann and Romano (cited in Chapter 24)) and a course based on Breiman (1992), Durrett (2004) or Billingsley (see Chapter 1).

My greatest thanks are due to Peter Hall for being an inspirational and caring advisor, reader, and intellectual filter over the last several years as I was writing drafts of this book. Peter has deeply influenced my understanding, appreciation, and taste for probability and statistics, and I have felt grateful that I have had access to him at all times and with unlimited patience. I have received much more from Peter than I could wish or expect. I could not have written this book without Peter's exemplary warmth and mentorship. However, all mistakes and ambiguities in the book are exclusively my responsibility. I would love to know of all serious mistakes that a reader finds in this book, and there must be mistakes in a book of this length.

I also want to express my very special thanks to John Marden and Larry Wasserman for repeatedly offering their friendly and thoughtful suggestions on various decisions I had to make on this book. I want to mention the generous help and support from Erich Lehmann, Peter Bickel, Rabi Bhattacharya, and Jon Wellner on specific chapters in the book. Numerous colleagues, and in particular C. R. Rao, Arup Bose, Persi Diaconis, Joe Eaton, Jianqing Fan, Iain Johnstone, T. Krishnan, Bruce Lindsay, Wei-Liem Loh, Peter McCullagh, Dimitris Politis, B. V. Rao, Bob Serfling, J. Sethuraman, Kesar Singh,

and Michael Woodroofe, made helpful comments on parts of earlier drafts of the book. Chun Han, Martina Muehlbach, and Surya Tokdar helped me graciously with putting together TeX files of the chapers. John Kimmel and Jeffrey Taub at Springer were extremely helpful and professional, and I enjoyed working with them very much. I will work with John and Jeff at any time with pleasure. Several anonymous referees did unbelievably helpful and constructive readings of many parts of the book. The Springer series editors gave me gracious input whenever needed. The copyeditor Hal Henglein and the typesetters – Integra India of Springer did a magnificent job. I am immensely thankful to all of them. I am also thankful to Purdue University for providing me with computing and secretarial assistance. Doug and Cheryl Crabill, in particular, assisted me numerous times with a smile.

I was an impressionable fifteen-year-old when I entered the Indian Statistical Institute (ISI) as a first-year student. I had heard that statisticians do boring calculations with large numbers using clumsy calculating machines. Dev Basu entered the lecture room on my first day at the ISI and instantly changed my perception of statistics. No one I met could explain so effortlessly the study of randomness and how to use what we learn about it to make useful conclusions. There was not one person at the ISI who didn't regard him as an incomparable role model, a personification of scholarship, and an angelic personality. I am fortunate that I had him as my foremost teacher. I am grateful to C. R. Rao for the golden days of the ISI and for making all of us feel that even as students we were equals in his eyes.

At a personal level, I am profoundly grateful to Jenifer Brown for the uniqueness and constancy of her treasured support, counsel, well wishes, and endearing camaraderie for many years, all of which have enriched me at my most difficult times and have helped me become a better human being. I will always remain much indebted to Jenifer for the positive, encouraging, and crystallizing influence she has been at all times. I have considered Jenifer to be an impeccable role model.

I am also thankful to Supriyo and Anuradha Datta, Julie Marshburn, Teena Seele, Gail Hytner, Norma Lucas, Deb Green, Tanya Winder, Hira Koul, Rajeeva Karandikar, Wei-Liem Loh, Dimitris Politis, and Larry Shepp for their loyalty, friendship and warmth. Jim and Ann Berger, Herman Rubin, B.V. Rao, T. Krishnan, Larry Brown, Len Haff, Jianqing Fan, and Bill Strawderman have mentored, supported and cared for me for more than a decade. I appreciate all of them. But most of all, I appreciate the love and warmth of my family. I dedicate this book to the cherished memories of my father, and to my mother on her eightieth birthday.

Anirban DasGupta
Purdue University, West Lafayette, IN

Chapter 12
Invariance Principles

The previous chapters discuss the asymptotic behavior of the sequence of partial sums $S_n = \sum_{i=1}^{n} X_i, n \geq 1$, for an iid sequence $X_1, X_2, \cdots$, under suitable moment conditions. In particular, we have described limit distributions and laws of large numbers for centered and normalized versions of S_n. The sequence of partial sums is clearly a natural thing to consider, given that sample means are so natural in statistics and probability. The central limit theorem says that as long as some moment conditions are satisfied, at any particular large value of n, S_n acts like a normally distributed random variable. In other words, the population from which the X_i came does not matter. The delta theorem says that we can do even better. We can even identify the limit distributions of functions, $h(S_n)$, and this is nice because there are problems in which the right statistic is not S_n itself but some suitable function $h(S_n)$.

Now, the sequence S_n is obviously a discrete-time stochastic process. We can think of a continuous-time stochastic process suitably devised, say $S_n(t)$, on the interval $[0, 1]$ such that, for any n, members of the discrete sequence $\frac{S_1}{\sqrt{n}}, \frac{S_2}{\sqrt{n}}, \cdots, \frac{S_n}{\sqrt{n}}$, are the values of that continuous-time process at the discrete times $t = \frac{1}{n}, \frac{2}{n}, \cdots, \frac{n}{n} = 1$. One can ask, what is the asymptotic behavior of this sequence of continuous-time stochastic processes? And just as in the case of the discrete-time stochastic process S_n, one can look at functionals $h(S_n(t))$ of these continuous-time stochastic processes. There are numerous problems in which it is precisely some suitable functional $h(S_n(t))$ that is the appropriate sequence of statistics. The problems in which a functional of such a type is the appropriate statistic arise in estimation, testing, model selection, goodness of fit, regression, and many other common statistical contexts. The invariance principle says the remarkable thing that, once again, under limited moment conditions, the continuous-time process $S_n(t)$ will act like a suitable continuous-time Gaussian process, say $W(t)$, and any nice enough functional $h(S_n(t))$ will act like the same functional of the limiting Gaussian process $W(t)$. The original F from which the data X_i

A. DasGupta, *Asymptotic Theory of Statistics and Probability*,

came is once again not going to matter. The form of the result is always the same. So, if we can identify that limiting Gaussian process $W(t)$, and if we know how to deal with the distribution of $h(W(t))$, then we have obtained the asymptotic behavior of our statistics $h(S_n(t))$ all at one stroke. It is a profoundly useful fact in the asymptotic theory of probability that all of this is indeed a reality. This chapter deals with such invariance principles and their concrete applications in important problems.

We recommend Billingsley (1968), Hall and Heyde (1980), and Csörgo and Révész (1981) for detailed and technical treatments, Erdös and Kac (1946), Donsker (1951), Komlós, Major, and Tusnady (1975, 1976), Major (1978), Whitt (1980), and Csörgo (1984) for invariance principles for the partial sum process, Mandrekar and Rao (1989) for more general symmetric statistics, Csörgo (1984), Dudley (1984), Shorack and Wellner (1986), Wellner (1992), Csörgo and Horváth (1993), and Giné (1996) for comprehensive treatments of empirical processes and their invariance principles, Heyde (1981), Pyke (1984), and Csörgo (2002) for lucid reviews, Philipp (1979), Dudley and Philipp (1983), Révész (1976), Einmahl (1987), and Massart (1989) for the multidimensional case, and Billingsley (1956), Jain, Jogdeo, and Stout (1975), McLeish (1974, 1975), Philip and Stout (1975), Hall (1977), Sen (1978), and Merlevéde, Peligrad, and Utev (2006) for treatments and reviews of various dependent cases. Other references are given within the specific sections.

12.1 Motivating Examples

Although we only talked about a continuous-time process $S_n(t)$ that suitably interpolates the partial sums $S_1, S_2, \cdots$, another continuous-time process of immense practical utility is the so-called *empirical process*. The empirical process $F_n(t)$ counts the proportion of sample observations among the first n that are less than or equal to a given t. We will discuss it in a little more detail shortly. But first we give a collection of examples of functionals $h(S_n(t))$ or $h(F_n(t))$ that arise naturally as test statistics in important testing problems or in the theory of probability. Their exact finite sample distributions being clumsy or even impossible to write, it becomes necessary to consider their asymptotic behavior. And, here is where an invariance principle of some appropriate type comes into play and settles the asymptotics in an elegant and crisp way.

Here are a small number of examples of such functionals that arise in statistics and probability.

Example 1 $D_n = \sup_{-\infty<t<\infty} |F_n(t) - F_0(t)|$, where F_0 is a given CDF on $\mathcal{R}$.
Example 2 $C_n = \int_{-\infty}^{\infty} (F_n(t) - F_0(t))^2 dF_0(t)$.
Example 3 $M_n = \sup_{0\le t\le 1} S_n(t)$.
Example 4 $\Pi_n = \frac{1}{n}\#\{k : S_k > 0\} = \lambda(t : S_n(t) > 0)$, where λ denotes Lebesgue measure (restricted to $[0, 1]$).

D_n and C_n arise as common test statistics in goodness-of-fit problems; we will study them in greater detail in Chapter 26. M_n and Π_n arise in the theory of random walks as the maximum fortune of a fixed player up to time n and as the proportion of times that she has been ahead. There are numerous such examples of statistics that can be regarded as functionals of either a partial sum process or an empirical process, and typically they satisfy some appropriate continuity property from which point an invariance principle takes over and settles the asymptotics.

12.2 Two Relevant Gaussian Processes

We remarked earlier that $S_n(t)$ will asymptotically act like a suitable Gaussian process. So does the empirical process $F_n(t)$. These two limiting Gaussian processes are closely related and happen to be the Brownian motion and the Brownian bridge, also known as the Wiener process and the tied-down Wiener process. For purposes of completeness, we give the definition and mention some fundamental properties of these two processes.

Definition 12.1 A stochastic process $W(t)$ defined on a probability space $(\Omega, \mathcal{A}, P)$, $t \in [0, \infty)$ is called a Wiener process or the Brownian motion starting at zero if:

(i) $W(0) = 0$ with probability 1;
(ii) for $0 \le s < t < \infty$, $W(t) - W(s) \sim N(0, t - s)$;
(iii) given $0 \le t_0 < t_1 < \cdots < t_k < \infty$, the random variables $W(t_{j+1}) - W(t_j)$, $0 \le j \le k - 1$ are mutually independent; and
(iv) the sample paths of $W(.)$ are almost all continuous (i.e., except for a set of sample points of probability 0), as a function of t, $W(t, \omega)$ is a continuous function.

Definition 12.2 Let $W(t)$ be a Wiener process on $[0, 1]$. The process $B(t) = W(t) - tW(1)$ is called a Brownian bridge on $[0, 1]$.

The proofs of the invariance principle theorems exploit key properties of the Brownian motion and the Brownian bridge. The properties that are the most useful toward this are sample path properties and some distributional results for relevant functionals of these two processes. The most fundamental properties are given in the next theorem. These are also of paramount importance and major historical value in their own right. There is no single source where all of these properties are available or proved. Most of them can be seen in Durrett (1996) and Csörgo (2002).

Theorem 12.1 (Important Properties of the Brownian Motion) Let $W(t)$ *and* $B(t)$ denote, respectively, a Brownian motion on $[0, \infty)$ starting at zero and a Brownian bridge on $[0, 1]$. Then,

(a) $\text{cov}(W(s), W(t)) = \min(s, t); \text{cov}(B(s), B(t)) = \min(s, t) - st$.

(b) **Karhunen-Loeve Expansion** If $Z_1, Z_2, \cdots$ is an infinite sequence of iid $N(0, 1)$ random variables, then $W(t)$ defined as $W(t) = \sqrt{2}\sum_{m=1}^{\infty} \frac{\sin([m-\frac{1}{2}]\pi t)}{[m-\frac{1}{2}]\pi} Z_m$ is a Brownian motion starting at zero and $B(t)$ defined as $B(t) = \sqrt{2}\sum_{m=1}^{\infty} \frac{\sin(m\pi t)}{m\pi} Z_m$ is a Brownian bridge on $[0, 1]$.

(c) **Scale and Time Change** $\frac{1}{\sqrt{c}} W(ct), c > 0, tW(\frac{1}{t})$ are also each Brownian motions on $[0, \infty)$.

(d) $W(t)$ is a Markov process and $W(t), W^2(t) - t, e^{\theta W(t) - \frac{\theta^2 t}{2}}, \theta \in \mathcal{R}$, are each martingales.

(e) **Unbounded Variations** On any nondegenerate finite interval, $W(t)$ is almost surely of unbounded total variation.

(f) **Rough Paths** Almost surely, sample paths of $W(t)$ are nowhere differentiable, but the paths are Holder continuous of order α for all $\alpha < \frac{1}{2}$.

(g) Almost surely, there does not exist any t_0 such that t_0 is a *point of increase* of $W(t)$ in the usual sense of analysis.

(h) **Behavior Near Zero** Almost surely, on any interval $(0, t_0), t_0 > 0$, $W(t)$ has infinitely many zeros.

(i) **Zero Set and Cantor Property** The set of all zeros of $W(t)$ is almost surely a closed, uncountable set of Lebesgue measure zero without any isolated points.

(j) **Strong Markov Property** If τ is a *stopping time* (w.r.t. the $W(t)$ process), then $W(t + \tau) - W(\tau)$ is also a Brownian motion on $[0, \infty)$.

(k) As $t \to \infty$, $\frac{W(t)}{t} \to 0$ with probability 1.

(l) **LIL** Almost surely, $\limsup_{t\downarrow 0} \frac{W(t)}{\sqrt{2t\log\log(1/t)}} = 1$, and $\liminf_{t\downarrow 0} \frac{W(t)}{\sqrt{2t\log\log(1/t)}} = -1$.

(m) Almost surely, $\limsup_{t\to\infty} \frac{W(t)}{\sqrt{2t\log\log t}} = 1$, and $\liminf_{t\to\infty} \frac{W(t)}{\sqrt{2t\log\log t}} = -1$.

(n) **Order of Increments** Almost surely, for any given $c > 0$,

$$\lim_{T\to\infty} \sup_{0\le t\le T-c\log T} \frac{|W(t+c\log T) - W(t)|}{c\log T} = \sqrt{\frac{2}{c}}$$

and

$$\lim_{T\to\infty} \sup_{t\ge 0, t+1\le T} \frac{|W(t+1) - W(t)|}{\sqrt{2\log T}} = 1.$$

(o) **Domination Near Zero** If $r(t) \in C[0,1]$ (the class of real continuous functions on $[0,1]$) and is such that $\inf r(t) > 0$, $r(t)$ is increasing and $\frac{r(t)}{\sqrt{t}}$ is decreasing in some neighborhood of $t = 0$, then $P(|W(t)| < r(t) \forall t \in [0, t_0]$ for some $t_0) = 1$ iff $\int_0^1 t^{-3/2} r(t) e^{-r^2(t)/(2t)} < \infty$.

(p) **Maxima and Reflection Principle** $P(\sup_{0<s\le t} W(s) \ge x) = 2P(W(t) \ge x)$.

(q) **First Arcsine Law** Let T be the point of maxima of $W(t)$ on $[0,1]$. Then T is almost surely unique, and $P(T \le t) = \frac{2}{\pi}\arcsin(\sqrt{t})$.

(r) **Last Zero and the Second Arcsine Law** Let $L = \sup\{t \in [0,1] : W(t) = 0\}$. Then $P(L \le t) = \frac{2}{\pi}\arcsin(\sqrt{t})$.

(s) **Reflected Brownian Motion** Let $X(t) = \sup_{0\le s\le t} |W(s)|$. Then $P(X(t) \le x) = 2\Phi(\frac{x}{\sqrt{t}}) - 1$, $x > 0$.

(t) **Loops and Self-Crossings** Given $d \ge 2$, let $W_1(t), \cdots, W_d(t)$ be independent Brownian motions starting at zero, and let $W^d(t) = (W_1(t), \cdots, W_d(t))$, called a *d-dimensional Brownian motion*. Then, for $d = 2$, for any given finite $k \ge 2$ and any nondegenerate time interval, almost surely there exist times $t_1, \cdots, t_k$ such that $W^d(t_1) = \cdots = W^d(t_k)$; for $d = 3$, given any nondegenerate time interval, almost surely there exist times t_1, t_2 such that $W^d(t_1) = W^d(t_2)$ (called double points or self-crossings); for $d > 3$, $W^d(t)$ has, almost surely, no double points.

(u) **Exit Time from Spheres** Let W^d be a d-dimensional Brownian motion, B the d-dimensional open unit sphere centered at the origin, and $\tau = \inf\{t > 0 : W^d(t) \notin B\}$. For a bounded function $f : \mathcal{R}^d \to \mathcal{R}$,

$E[f(W(\tau))] = \int_{\partial B} \frac{1}{||x||^d} f(x) dS(x)$, where S is the normalized surface measure on the boundary of B.

(v) **Recurrence and Transience** Let W^d be a d-dimensional Brownian motion. W^d is recurrent for $d = 1, 2$ and transient for $d \geq 3$.

12.3 The Erdös-Kac Invariance Principle

Although the invariance principle for partial sums of iid random variables is usually credited to Donsker (Donsker (1951)), Erdös and Kac (1946) contained the basic idea behind the invariance principle and also worked out the asymptotic distribution of a number of key and interesting functionals $h(S_n(t))$. We provide a glimpse into the Erdös-Kac results in this section. Erdös and Kac describe their method of proof as follows:

"The proofs of all these theorems follow the same pattern. It is first proved that the limiting distribution exists and is independent of the distribution of the X_i's; then the distribution of the X_i's is chosen conveniently so that the limiting distribution can be calculated explicitly. This simple principle has, to the best of our knowledge, never been used before."

Theorem 12.2 Let $X_1, X_2, \cdots$ be an infinite iid sequence of zero-mean random variables such that $\frac{S_n}{\sqrt{n}}$ admits the central limit theorem. Then,

$$\lim_{n\to\infty} P(n^{-1/2} \max_{1\leq k\leq n} S_k \leq x) = G_1(x), x \geq 0,$$

$$\lim_{n\to\infty} P(n^{-1/2} \max_{1\leq k\leq n} |S_k| \leq x) = G_2(x), x \geq 0,$$

$$\lim_{n\to\infty} P\left(n^{-2} \sum_{k=1}^{n} S_k^2 \leq x\right) = G_3(x), x \geq 0,$$

$$\lim_{n\to\infty} P\left(n^{-3/2} \sum_{k=1}^{n} |S_k| \leq x\right) = G_4(x), x \geq 0,$$

$$\lim_{n\to\infty} P\left(\frac{1}{n} \sum_{k=1}^{n} I_{S_k>0} \leq x\right) = \frac{2}{\pi} \arcsin(\sqrt{x}), 0 \leq x \leq 1,$$

where

$$G_1(x) = 2\Phi(x) - 1,$$
$$G_2(x) = \frac{4}{\pi}\sum_{m=0}^{\infty}\frac{(-1)^m}{2m+1}e^{-(2m+1)^2\pi^2/(8x^2)},$$

and formulas for G_3 and the Laplace transform of G_4 are provided in Erdös and Kac (1946) (they involve complicated integral representations).

Remark. It is quite interesting that for existence of a limiting distribution of the sum of squares of the partial sums, a fourth moment of the X_i's is not at all necessary, and in fact all that is needed is that the distribution F from which the X_i arise be in the domain of attraction of the normal. In particular, existence of a variance is already enough. We will see this phenomenon reemerge in the next section.

12.4 Invariance Principles, Donsker's Theorem, and the KMT Construction

Donsker (1951) provided the full generalization of the Erdös-Kac technique by providing explicit embeddings of the discrete sequence $\frac{S_k}{\sqrt{n}}, k = 1, 2, \cdots, n$ into a continuous-time stochastic process $S_n(t)$ and by establishing the limiting distribution of a general continuous functional $h(S_n(t))$. In order to achieve this, it is necessary to use a continuous mapping theorem for metric spaces, as consideration of Euclidean spaces is no longer enough. It is also useful to exploit a property of the Brownian motion known as the *Skorohod embedding theorem*. We first describe this necessary background material.

Define $C[0, 1]$ the class of all continuous real-valued functions on $[0, 1]$ and $D[0, 1]$ the class of all real-valued functions on $[0, 1]$ that are right continuous and have a left limit at every point in $[0, 1]$.

Given two functions f, g in either $C[0,1]$ or $D[0,1]$, let $\rho(f,g) = \sup_{0\le t\le 1}|f(t) - g(t)|$ denote the supremum distance between f and g. We will refer to ρ as the *uniform metric*. Both $C[0, 1]$ and $D[0, 1]$ are (complete) metric spaces with respect to the uniform metric ρ. Two common embeddings of the discrete sequence $\frac{S_k}{\sqrt{n}}, k = 1, 2, \cdots, n$ into a continuous-time process are the following:

$$S_{n,1}(t) = \frac{1}{\sqrt{n}}[S_{[nt]} + \{nt\}X_{[nt]+1}]$$

and

$$S_{n,2}(t) = \frac{1}{\sqrt{n}} S_{[nt]},$$

$0 \le t \le 1$. Here, [.] denotes the integer part and {.} the fractional part of a positive real.

The first one simply continuously interpolates between the values $\frac{S_k}{\sqrt{n}}$ by drawing straight lines, but the second one is only right continuous, with jumps at the points $t = \frac{k}{n}$, $k = 1, 2, \cdots, n$. For certain specific applications, the second embedding is more useful. It is because of these jump discontinuities that Donsker needed to consider weak convergence in $D[0, 1]$. It did lead to some additional technical complexities.

The main idea from this point on is not difficult. One can produce a version of $S_n(t)$, say $\widetilde{S}_n(t)$, such that $\widetilde{S}_n(t)$ is *close to* a sequence of Wiener processes $W_n(t)$. Since $\widetilde{S}_n(t) \approx W_n(t)$, if $h(.)$ is a continuous functional with respect to the uniform metric, then one can expect that $h(\widetilde{S}_n(t)) \approx h(W_n(t)) = h(W(t))$ in distribution. $\widetilde{S}_n(t)$ being a version of $S_n(t)$, $h(S_n(t)) = h(\widetilde{S}_n(t))$ in distribution, and so $h(S_n(t))$ should be close to the fixed Brownian functional $h(W(t))$ in distribution, which is the question we wanted to answer.

The results leading to Donsker's theorem are presented below; we recommend Csörgo (2003) for further details on Theorems 12.3–12.5 below.

Theorem 12.3 (Skorohod Embedding) Given $n \ge 1$, iid random variables $X_1, \cdots, X_n$, $E(X_1) = 0$, $\mathrm{Var}(X_1) = 1$, there exists a common probability space on which one can define a Wiener process $W(t)$ starting at zero and a triangular array of nonnegative random variables $\{\tau_{1,n}, \cdots, \tau_{n,n}\}$, iid under each given n such that

$$\text{(a) } \tau_{n,n} \stackrel{\mathcal{L}}{=} \tau_{1,1},$$

$$\text{(b) } E(\tau_{1,1}) = 1,$$

$$\text{(c) } \left\{ \frac{S_1}{\sqrt{n}}, \cdots, \frac{S_n}{\sqrt{n}} \right\} \stackrel{\mathcal{L}}{=} \left\{ W\left(\frac{\tau_{1,n}}{n}\right), \cdots, W\left(\frac{\tau_{1,n} + \cdots + \tau_{n,n}}{n}\right) \right\}.$$

Remark. Much more general versions of the Skorohod embedding theorem are known. See, for example, Oblój (2004). The version above suffices for the following *weak invariance principle for partial sum processes*.

Theorem 12.4 Let $S_n(t) = S_{n,1}(t)$ or $S_{n,2}(t)$ as defined above. Then there exists a common probability space on which one can define a Wiener process $W(t)$ starting at zero and a sequence of processes $\{\widetilde{S}_n(t)\}$, $n \geq 1$, such that

(a) for each n, $S_n(t)$ and $\widetilde{S}_n(t)$ are identically distributed as processes;

(b) $\sup_{0\leq t\leq 1} |\widetilde{S}_n(t) - W(t)| \stackrel{P}{\Rightarrow} 0$.

This leads to the famous Donsker theorem. We state a version that is slightly less general than the original result in order to avoid discussion of the so-called *Wiener measure*.

Theorem 12.5 (Donsker) Let h be a continuous functional with respect to the uniform metric ρ on $C[0, 1]$ or $D[0, 1]$ and let $S_n(t)$ be defined as either $S_{n,1}(t)$ or $S_{n,2}(t)$. Then $h(S_n(t)) \stackrel{\mathcal{L}}{\Rightarrow} h(W(t))$ as $n \to \infty$.

Example 12.1 The five examples worked out by Erdös and Kac now follow from Donsker's theorem by considering the following functionals, each of which is continuous (with the exception of h_5, even which is continuous at *almost all* $f \in C[0, 1]$) with respect to the uniform metric on $C[0, 1]$: $h_1(f) = \sup_{0\leq t\leq 1} f(t); h_2(f) = \sup_{0\leq t\leq 1} |f(t)|; h_3(f) = \int_0^1 f^2(t)dt; h_4(f) = \int_0^1 |f(t)|dt$; $h_5(f) = \lambda\{t \in [0, 1] : f(t) > 0\}$, where λ denotes Lebesgue measure. Note that the formulas for the CDF of the limiting distribution are always a separate calculation and do not follow from Donsker's theorem.

Example 12.2 Consider the functional $h(f) = \int_0^1 f^m(t)dt$, where $m \geq 1$ is an integer. Because $[0, 1]$ is a compact interval, it is easy to verify that h is a continuous functional on $C[0, 1]$ with respect to the uniform metric. Indeed, it follows simply from the algebraic identity $|x^m - y^m| = |x - y||x^{m-1} + x^{m-2}y + \cdots + y^{m-1}|$. On the other hand, by direct integration of the polygonal curve $S_{n,1}(t)$, it follows from Donsker's theorem that $n^{-1-m/2} \sum_{k=1}^n S_k^m \stackrel{\mathcal{L}}{\Rightarrow} \int_0^1 W^m(t)dt$. At first glance, it seems surprising that a nondegenerate limit distribution for partial sums of S_k^m can exist with *only* two moments (and even that is not necessary).

Other examples and classic theory on distributions of functionals of $W(t)$ can be seen in Cameron and Martin (1945), Kac (1951), Durrett (1996), and Fitzsimmons and Pitman (1999).

Contrary to the weak invariance principle described above, there are also *strong invariance principles*, which, roughly speaking, say that the partial

sum process is close to a Brownian motion with probability 1. However, the exact statements need careful description, and the best available results need a fair amount of extra assumptions, as well as sophisticated methods of proof. Furthermore, somewhat paradoxically, the strong invariance principles may not lead to the desired weak convergence results unless enough assumptions have been made so that a good enough almost sure bound on the deviation between the two processes can be established. The first strong invariance principle for partial sums was obtained in Strassen (1964). Since then, a lot of literature has developed, including for the multidimensional case. Good sources for information are Strassen (1967), Komlós, Major, and Tusnady (1976), Major (1978), and Einmahl (1987). We present two results on strong approximations of partial sums below. The results may be seen in Csörgo (1984) and Heyde (1981).

Theorem 12.6 (Strassen) Given iid random variables $X_1, X_2, \cdots$ with $E(X_1) = 0$, $\text{Var}(X_1) = 1$, there exists a common probability space on which one can define a Wiener process $W(t)$ and iid random variables $Y_1, Y_2, \cdots$ such that

$$\text{(a)} \quad \left\{ S_n = \sum_{i=1}^{n} X_i, n \geq 1 \right\} \stackrel{\mathcal{L}}{=} \left\{ \widetilde{S}_n = \sum_{i=1}^{n} Y_i, n \geq 1 \right\},$$

$$\text{(b)} \quad \sup_{0 \leq t \leq 1} \frac{|\tilde{S}_{[nt]} - W(nt)|}{\sqrt{n \log \log n}} \to 0,$$

almost surely, as $n \to \infty$.

Remark. The $\sqrt{n \log \log n}$ bound cannot be improved in general without further assumptions on the distribution of the X_i's. The next theorem says that if we assume finiteness of more than two moments of the X_i's, or even better the moment-generating function itself, then the error can be made sufficiently small to allow the weak convergence result to be derived directly from the strong invariance principle itself. The improved rate under the existence of the mgf is the famous *KMT construction* due to Komlós, Major, and Tusnady (1976).

Theorem 12.7 Given iid random variables $X_1, X_2, \cdots$ with $E(X_1) = 0$, $\text{Var}(X_1) = 1$, $E(|X_1|^{2+\delta}) < \infty$, for some $\delta > 0$, Theorem 12.6 holds with $n^{1/(2+\delta)}$ in place of $\sqrt{n \log \log n}$. If $E(X_1^4) < \infty$, the result holds with $(n \log \log n)^{1/4} \sqrt{\log n}$ in place of $\sqrt{n \log \log n}$. If the mgf $E(e^{tX_1}) < \infty$ in

some neighborhood of zero, then the $O(\log n)$ rate holds, almost surely, in place of $o(\sqrt{n \log\log n})$.

12.5 Invariance Principle for Empirical Processes

Of our four motivating examples, the statistics D_n and C_n are not functionals of a partial sum process; they are functionals of the empirical process. If weak and strong invariance principles akin to the case of partial sum processes were available for the empirical process, clearly that would be tremendously helpful in settling the questions of useful asymptotics of D_n, C_n, and more generally *nice* functionals of the empirical process. It turns out that there are indeed such invariance principles for the empirical process. Although new and significant developments are still taking place, a large part of this literature is classic, dating back to at least Kolmogorov (1933) and Smirnov (1944). In this section, we provide a brief description of some major results on this and give applications. The topic is also discussed with applications in the context of goodness-of-fit tests in Chapter 26.

To describe the weak and strong approximations of the empirical process, we first need some notation and definitions. Given a sequence of iid $U[0,1]$ random variables $U_1, U_2, \cdots$, we define the *uniform empirical process* as $G_n(t) = \frac{1}{n}\sum_{i=1}^{n} I_{U_i \le t}$ and the normalized uniform empirical process $\alpha_n(t) = \sqrt{n}(G_n(t) - t), n \ge 1, 0 \le t \le 1$. For an iid sequence $X_1, X_2, \cdots$ distributed as a general F, the empirical process is defined as $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \le x}$. The normalized empirical process is $\beta_n(x) = \sqrt{n}(F_n(x) - F(x)), -\infty < x < \infty$.

The weak invariance principle is very similar to that for partial sum processes and is given below; see Dudley (1984) for further details.

Theorem 12.8 Given a sequence of iid random variables $X_1, X_2, \cdots \sim F$ and h a continuous functional on $D(-\infty, \infty)$ with respect to the uniform metric, $h(\beta_n(.)) \stackrel{\mathcal{L}}{\Rightarrow} h(B_F(.))$, where $B_F(.)$ is a centered Gaussian process with covariance kernel

$$\mathrm{cov}(B_F(x), B_F(y)) = F(x \wedge y) - F(x)F(y).$$

An important consequence of this result is the asymptotic distribution of D_n, an extremely common statistic in the goodness-of-fit literature, which we will revisit in Chapter 26.

Example 12.3 Since the functional $h(f) = \sup_x |f(x)|$ satisfies the continuity assumption under the uniform metric, it follows that $D_n = \sup_x |\beta_n(x)| \stackrel{\mathcal{L}}{\Rightarrow} \sup_x |B_F(x)|$, which is equal to the supremum of the absolute value of a Brownian bridge (on $[0, 1]$) in distribution. This distribution was calculated in closed form by Kolmogorov (1933) and is a classic result. Kolmogorov found the CDF of the supremum of the absolute value of a Brownian bridge to be $H(z) = 1 - \sum_{k=-\infty}^{\infty} (-1)^{k-1} e^{-2k^2 z^2}$, from which the quantiles of the limiting distribution of D_n can be (numerically) computed.

Similar to partial sum processes, there are strong invariance principles for empirical processes as well. Some of the first ideas and results were due to Brillinger (1969) and Kiefer (1972). Given a sequence of independent Brownian bridges $B_n(t), n \geq 1, 0 \leq t \leq 1$, a *Kiefer process* (with two time parameters, namely n and t) is defined as $K(n, t) = \sum_{i=1}^{n} B_i(t)$. The better-known strong approximations for the empirical process are not the initial ones; we present one of the modern strong approximations below. See Komlós, Major, and Tusnady (1975) for the method of proof, which is different from the original Skorohod embedding technique.

Theorem 12.9 Given iid random variables $X_1, X_2, \cdots \sim F$, let $\beta_n(x) = \sqrt{n}(F_n(x) - F(x)) \stackrel{\mathcal{L}}{=} \alpha_n(F(x))$. Then, there exists a probability space on which one can define a Kiefer process $K(n, F(x))$, Brownian bridges $B_n(t), n \geq 1$, and $\tilde{\beta}_n(x), n \geq 1$, such that

(a) $\{\tilde{\beta}_n(x), n \geq 1\} \stackrel{\mathcal{L}}{=} \{\beta_n(x), n \geq 1\}$;

(b) $\sup_{-\infty<x<\infty} |\tilde{\beta}_n(x) - n^{-1/2} K(n, F(x))| = O(n^{-1/2}(\log n)^2)$ almost surely;

(c) for suitable constants C_1, C_2, λ,

$$P(\sup_{1\leq k\leq n, -\infty<x<\infty} |\sqrt{k}\tilde{\beta}_k(x) - K(k, F(x))| > C_1(\log n)^2 + z \log n) \leq C_2 e^{-\lambda z}$$

for any z and any n;

(d) for suitable constants C_1, C_2, λ,

$$P(\sup_{-\infty<x<\infty} |\tilde{\beta}_n(x) - B_n(F(x))| > n^{-1/2}(C_1 \log n + z)) \leq C_2 e^{-\lambda z}$$

for any z and any n;

(e) $\sup_{-\infty<x<\infty} |\tilde{\beta}_n(x) - B_n(F(x))| = O(n^{-1/2} \log n)$ almost surely.

Remark. No additional improvement in the rate in part (e) is possible. Multidimensional versions are available in many places; a recent reference is Massart (1989).

We now give an application of the result above.

Example 12.4 We have noted that the general (normalized) empirical process has the property that $\sup_{-\infty<x<\infty} |\beta_n(x)|$ converges in law to a distribution with CDF $H(z) = 1 - \sum_{k=-\infty}^{\infty}(-1)^{k-1}e^{-2k^2z^2}$ as $n \to \infty$. The key point is that this is the CDF of the absolute value of a Brownian bridge. We can combine this fact with part (c) of Theorem 12.9 to produce guaranteed coverage confidence bands for the CDF $F(.)$ at any given n. The form of this *nonparametric confidence band* is $F_n(x) \pm \frac{\delta}{\sqrt{n}}$, where δ is to be chosen appropriately. If $r(n) = (C_1 + \epsilon\lambda^{-1})n^{-1/2}\log n$, then we need $H(\delta - r(n)) - C_2 n^{-\epsilon}$ to be greater than or equal to the nominal coverage $1 - \alpha$. To execute this, we need values for the constants C_1, C_2, λ; they may be taken to be 100, 10, and $\frac{1}{50}$, respectively (see Csörgo and Hall (1984)). However, these values are not sharp enough to produce useful (or even nontrivial) confidence bands at moderate values of n. But the coverage property is exact; i.e., there is no need to say that the coverage is approximately $1 - \alpha$ for large n.

12.6 Extensions of Donsker's Principle and Vapnik-Chervonenkis Classes

As we have seen, an important consequence of the weak invariance principle is the derivation of the limiting distribution of $D_n = \sqrt{n}\sup_{-\infty<x<\infty} |F_n(x) - F(x)|$ for a continuous CDF F and the empirical CDF $F_n(x)$. If we let $\mathcal{F} = \{I_{-\infty,x} : x \in \mathcal{R}\}$, then the Kolmogorov-Smirnov result says that $\sqrt{n}\sup_{f\in\mathcal{F}}(E_{P_n} f - E_P f) \stackrel{\mathcal{L}}{\Rightarrow} \sup_{0\le t\le 1} |B(t)|$, P_n, P being the probability measures corresponding to F_n, F, and $B(t)$ being a Brownian bridge. Extensions of this involve studying the asymptotic behavior of $\sup_{f\in\mathcal{F}}(E_{P_n} f - E_P f)$ for much more general classes of functions $\mathcal{F}$ and the range space of the random variables X_i; they need not be $\mathcal{R}$, or $\mathcal{R}^d$ for some finite d. Examples of asymptotic behavior include derivation of laws of large numbers and central limit theorems.

There are numerous applications of these extensions. To give just one motivating example, suppose $X_1, X_2, \cdots, X_n$ are d-dimensional iid random vectors from some P and we want to test the null hypothesis that $P = P_0$ (specified). Then, a natural statistic to assess the truth of the hypothesis is $T_n = \sup_{C\in\mathcal{C}} |P_n(C) - P_0(C)|$ for a suitable class of sets $\mathcal{C}$. Now, if $\mathcal{C}$ is too rich (for example, if it is the class of all measurable sets), then clearly there cannot be any meaningful asymptotics if P_0 is absolutely continuous. On the other hand, if $\mathcal{C}$ is too small, then the statistic cannot be good enough for detecting departures from the null hypothesis. So these extensions study the question of what kinds of families $\mathcal{C}$ or function classes $\mathcal{F}$ allow meaningful

asymptotics and also result in good and commonsense tests or estimators. In some sense, the topic is a study in the art of the possible.

The technical tools required for such generalizations are extremely sophisticated and have led to striking new discoveries and mathematical advances in the theory of empirical processes. Along with these advances have come numerous new and useful statistical and probabilistic applications. The literature is huge; we strongly recommend Wellner (1992), Giné (1996), Pollard (1989), and Giné and Zinn (1984) for comprehensive reviews and sources for major theorems and additional references; specific references to some results are given later. We limit ourselves to a short description of a few key results and tools.

12.7 Glivenko-Cantelli Theorem for VC Classes

We first discuss plausibility of strong laws more general than the well-known Glivenko-Cantelli theorem, which asserts that in the one-dimensional iid case $\sup_x |F_n(x) - F(x)| \stackrel{\text{a.s.}}{\to} 0$. We need a concept of combinatorial richness of a class of sets $\mathcal{C}$ that will allow us to make statements like $\sup_{C\in\mathcal{C}} |P_n(C) - P(C)| \stackrel{\text{a.s.}}{\to} 0$. A class of sets for which this property holds is called a *Glivenko-Cantelli class*. A useful such concept is the *Vapnik-Chervonenkis dimension* of a class of sets. Meaningful asymptotics will exist for classes of sets that have a finite Vapnik-Chervonenkis dimension. It is therefore critical to know what it means and what are good examples of classes of sets with a finite Vapnik-Chervonenkis dimension. A basic treatment of this is given next.

Definition 12.3 Let $A \subset \mathcal{S}$ be a fixed set and $\mathcal{C}$ a class of subsets of $\mathcal{S}$. A is said to be *shattered by* $\mathcal{C}$ if every subset U of A is the intersection of A with some member C of $\mathcal{C}$ (i.e., $\{A \cap C : C \in \mathcal{C}\} = P(A)$, where $P(A)$ denotes the power set of A).

Sometimes the phenomenon is colloquially described as *every subset of A is picked up by some member of $\mathcal{C}$*.

Definition 12.4 The Vapnik-Chervonenkis (VC) dimension of $\mathcal{C}$ is the size of the largest set A that can be shattered by $\mathcal{C}$.

Although this is already fine as a definition, a more formal definition is given by using the concept of *shattering coefficients*.

Definition 12.5 For $n \geq 1$, the nth shattering coefficient of $\mathcal{C}$ is defined to be

$$S(n, \mathcal{C}) = \max_{x_1,x_2,\cdots,x_n\in\mathcal{S}} \text{Card}\{\{x_1, x_2, \cdots, x_n\} \cap C : C \in \mathcal{C}\}.$$

That is, $S(n, \mathcal{C})$ is the largest possible number of subsets of some (wisely chosen) set of n points that can be formed by intersecting the set with members of $\mathcal{C}$. Clearly, for any n, $S(n, \mathcal{C}) \leq 2^n$.

Here is an algebraic definition of the VC dimension of a class of sets.

Definition 12.6 The VC dimension of $\mathcal{C}$ equals $VC(\mathcal{C}) = \min\{n : S(n, \mathcal{C}) < 2^n\} - 1 = \max\{n : S(n, \mathcal{C}) = 2^n\}$.

Definition 12.7 $\mathcal{C}$ is called a Vapnik-Chervonenkis (VC) class if $VC(\mathcal{C}) < \infty$.

The following remarkable result is known as *Sauer's lemma* (Sauer (1972)).

Proposition 12.1 *Either* $S(n, \mathcal{C}) = 2^n \forall n$ *or* $\forall n$, $S(n, \mathcal{C}) \leq \sum_{i=0}^{\mathrm{VC}(\mathcal{C})} \binom{n}{i}$.

Remark. Sauer's lemma says that either a class of sets has infinite VC dimension or its shattering coefficients grow polynomially. A few other important and useful properties of the shattering coefficients are listed below; most of them are derived easily. These properties are useful for generating new classes of VC sets from known ones by using various Boolean operations.

Theorem 12.10 The shattering coefficients $S(n, \mathcal{C})$ of a class of sets $\mathcal{C}$ satisfy

(a) $S(m, \mathcal{C}) < 2^m$ for some $m \Rightarrow S(n, \mathcal{C}) < 2^n \, \forall n > m$;

(b) $S(n, \mathcal{C}) \leq (n+1)^{VC(\mathcal{C})} \forall n \geq 1$;

(c) $S(n, \mathcal{C}^c) = S(n, \mathcal{C})$, where $\mathcal{C}^c$ is the class of complements of members of $\mathcal{C}$;

(d) $S(n, \mathcal{B} \cap \mathcal{C}) \leq S(n, \mathcal{B}) S(n, \mathcal{C})$, where the $\cap$ notation means the class of sets formed by intersecting members of $\mathcal{B}$ with those of $\mathcal{C}$;

(e) $S(n, \mathcal{B} \otimes \mathcal{C}) \leq S(n, \mathcal{B}) S(n, \mathcal{C})$, where the $\otimes$ notation means the class of sets formed by taking Cartesian products of members of $\mathcal{B}$ and those of $\mathcal{C}$;

(f) $S(m+n, \mathcal{C}) \leq S(m, \mathcal{C}) S(n, \mathcal{C})$.

See Vapnik and Chervonenkis (1971) and Sauer (1972) for many of the parts in this theorem. Now we give a few quick examples.

Example 12.5 Let $\mathcal{C}$ be the class of all left unbounded closed intervals on the real line; i.e., $\mathcal{C} = \{(-\infty, x] : x \in \mathcal{R}\}$. To illustrate the general formula, suppose $n = 2$. What is $S(n, \mathcal{C})$? Clearly, if we pick up the larger one among

x_1, x_2, we will pick up the smaller one, too. Or we may pick up none of them or just the smaller one. So we can pick up three distinct subsets from the power set of $\{x_1, x_2\}$. The same argument shows that the general formula for the shattering coefficients is $S(n, \mathcal{C}) = n + 1$. Consequently, this is a VC class with VC dimension one.

Example 12.6 Although topologically there are just as many left unbounded intervals on the real line as there are arbitrary intervals, in the VC index they act differently. This is interesting. Thus, let $\mathcal{C} = \{(a, b) : a \leq b \in \mathcal{R}\}$. Then it is easy to establish the formula $S(n, \mathcal{C}) = 1 + \binom{n+1}{2}$. For $n = 2$, this is equal to 4, which is also 2^2. Consequently, this is a VC class with VC dimension two.

Example 12.7 The previous example says that, on $\mathcal{R}$, the class of all convex sets is a VC class. However, this is far from being true, even in two dimensions. Indeed, if we let $\mathcal{C}$ be just the class of convex polygons in the plane, it is clear geometrically that for any n, $\mathcal{C}$ can shatter n points. So, convex polygons in $\mathcal{R}^2$ have an infinite VC dimension.

More examples of exact values of VC dimensions are given in the chapter exercises. For actual applications of these ideas to concrete extensions of Donsker's principles, it is extremely useful to know what other natural classes of sets in various spaces are VC classes. The various parts of the following result are available in Vapnik and Chervonenkis (1971) and Dudley (1978, 1979).

Theorem 12.11 Each of the following classes of sets is a VC class:

(a) southwest quadrants of $\mathcal{R}^d$ (i.e., the class of all sets of the form $\prod_{i=1}^{d} (-\infty, x_i]$);

(b) closed half-spaces of $\mathcal{R}^d$;

(c) closed balls of $\mathcal{R}^d$;

(d) closed rectangles of $\mathcal{R}^d$;

(e) $\mathcal{C} = \{\{x \in \mathcal{R}^d : g(x) \geq 0\} : g \in G\}$, where G is a finite-dimensional vector space of real-valued functions defined on $\mathcal{R}^d$.

We can now state a general version of the familiar Glivenko-Cantelli theorem. However, to appreciate the probabilistic utility of the combinatorial concept of shattering coefficients, it is useful to see also a famous theorem of Vapnik and Chervonenkis (1971) on Euclidean spaces, which we also provide.

Theorem 12.12 Let $X_1, X_2, \cdots \overset{\text{iid}}{\sim} P$, a probability measure on $\mathcal{R}^d$ for some finite d. Given any class of (measurable) sets $\mathcal{C}$, for $n \geq 1, \epsilon > 0$, $P(\sup_{\mathcal{C}} |P_n(C) - P(C)| > \epsilon) \leq 8E[\text{Card}\{\{X_1, X_2, \cdots, X_n\} \cap C : C \in \mathcal{C}\}]e^{-n\epsilon^2/32} \leq 8S(n, \mathcal{C})e^{-n\epsilon^2/32}$.

Remark. This theorem implies that for classes of sets that are of the right complexity as measured by the VC dimension, the empirical measure converges to the true measure at an essentially exponential rate. This is a sophisticated generalization of the one-dimensional DKW inequality. The improved bound of the theorem is harder to implement because it involves computation of a hard expectation with respect to a sample of n observations from the underlying P. It would usually not be possible to find this expectation, although simulating the quantity $\text{Card}\{\{X_1, X_2, \cdots, X_n\} \cap C : C \in \mathcal{C}\}$ is an interesting exercise.

The general theorem is given next; see Giné (1996).

Theorem 12.13 Let P be a probability measure on a general measurable space $\mathcal{S}$ and let $X_1, X_2, \cdots \overset{\text{iid}}{\sim} P$. Let P_n denote the sequence of empirical measures and let $\mathcal{C}$ be a VC class of sets in $\mathcal{S}$. Then, under suitable measurability conditions, $\sup_{C \in \mathcal{C}} |P_n(C) - P(C)| \overset{\text{a.s.}}{\to} 0$ as $n \to \infty$.

12.8 CLTs for Empirical Measures and Applications

This result gives us hope for establishing CLTs for suitably normalized versions of $\sup_{C \in \mathcal{C}} |P_n(C) - P(C)|$ in general spaces and with general VC classes of sets. It is useful to think of this as an analog of the one-dimensional Kolmogorov-Smirnov statistic for real-valued random variables, namely $\sup_x |F_n(x) - F(x)|$. Invariance principles allowed us to conclude that the limiting distribution is related to a Brownian bridge with real numbers in $[0, 1]$ as the time parameter. Now, however, the setup is much more abstract. The space is not a Euclidean space, and the time parameter is a set or a function. So the formulation and description of the appropriate CLTs is more involved, and although suitable Gaussian processes will still emerge as the relevant processes that determine the asymptotics, they are not Brownian bridges, and they even depend on the underlying P from which we are sampling. Some of the most profound advances in the theory of statistics and probability in the twentieth century took place around this problem, resulting along the way in deep mathematical developments and completely new tools. A short description of this is provided next.

12.8.1 Notation and Formulation

First, we give some notation and definitions. The notation $(P_n - P)(f)$ would mean $\int f\,dP_n - \int f\,dP$. Here, f is supposed to belong to some suitable class of functions $\mathcal{F}$. For example, $\mathcal{F}$ could be the class of indicator functions of the members C of a class of sets $\mathcal{C}$. In that case, $(P_n - P)(f)$ would simply mean $P_n(C) - P(C)$; we have just talked about strong laws for their suprema as C varies over $\mathcal{C}$. That is a uniformity result. Likewise, we will now need certain uniformity assumptions on the class of functions $\mathcal{F}$. We assume that

$$\text{(a) } \sup_{f\in\mathcal{F}} |f(s)| := F(s) < \infty \,\forall s \in \mathcal{S}$$

(measurability of F is clearly not obvious but is being ignored here) and

$$\text{(b) } F \in L_2(P).$$

In the case of real-valued random variables and for the problem of convergence of the process $F_n(x) - F(x)$, the corresponding functions, as we noted before, are indicator functions of $(-\infty, x)$, which are uniformly bounded functions. Now the time parameter has become a function itself, and we will need to talk about uniformly bounded functionals of functions; we will use the notation

$$l_\infty(\mathcal{F}) = \{h : \mathcal{F} \to \mathcal{R} : \sup_{f\in\mathcal{F}} |h(f)| < \infty\}.$$

Furthermore, we will refer to $\sup_{f\in\mathcal{F}} |h(f)|$ as the uniform norm and denote it as $||h||_{\infty,\mathcal{F}}$.

The two other notions we need to define are those of convergence of the process $(P_n - P)(f)$ (on normalization) and of a limiting Gaussian process that will play the role of a Brownian bridge in these general circumstances.

The Gaussian process, which we will denote as $B_P(f)$, will continue to have continuous sample paths, as was the case for the ordinary Brownian bridge, but now with the time parameter being a function and continuity being with respect to $\rho_P(f,g) = \sqrt{E_P(f(X) - g(X))^2}$. B_P has mean zero, and the covariance kernel $\text{cov}(B_P(f), B_P(g)) = P(fg) - P(f)P(g) := E_P(f(X)g(X)) - E_P(f(X))\,E_P(g(X))$. Note that due to our assumption that $F \in L_2(P)$, the covariance kernel is well defined. Trajectories of our Gaussian process B_P are therefore members of $l_\infty(\mathcal{F})$, also (uniformly) continuous with respect to the norm ρ_P we have defined above.

Finally, as in the Portmanteau theorem in Chapter 1, convergence of the process $\sqrt{n}(P_n - P)(f)$ to $B_P(f)$ would mean that expectation of any functional H of $\sqrt{n}(P_n - P)(f)$ will converge to the expectation of $H(B_P(f))$,

H being a bounded and continuous functional defined on $l_\infty(\mathcal{F})$ and taking values in $\mathcal{R}$. We remind the reader that continuity on $l_\infty(\mathcal{F})$ is with respect to the uniform norm we have already defined there. A class of functions $\mathcal{F}$ for which this central limit property holds is called a *P-Donsker class*; if the property holds for *every* probability measure P on $\mathcal{S}$, it is called a *universal Donsker class*.

12.8.2 Entropy Bounds and Specific CLTs

We can now state what sorts of assumptions on our class of functions $\mathcal{F}$ will ensure that convergence occurs (i.e., a CLT holds) and what are some good applications of such CLTs. There are multiple sets of assumptions on the class of functions that ensure a CLT. Here we describe only two, one of which relates to the concept of VC classes and the second related to *metric entropy* and *packing numbers*. Since we are already familiar with the concept of VC classes, we first state a CLT based on a VC assumption of a suitable class of sets.

Definition 12.8 A family $\mathcal{F}$ of functions f on a (measurable) space $\mathcal{S}$ is called a *VC-subgraph* if the class of subgraphs of $f \in \mathcal{F}$ is a VC class of sets, where the subgraph of f is defined to be $C_f = \{(x, y), x \in \mathcal{S}, y \in \mathcal{R} : 0 \le y \le f(x) \text{ or } f(x) \le y \le 0\}$.

Theorem 12.14 Given $X_1, X_2, \cdots \overset{\text{iid}}{\sim} P$, a probability measure on a measurable space $\mathcal{S}$, and a family of functions $\mathcal{F}$ on $\mathcal{S}$ such that $F(s) := \sup_{f \in \mathcal{F}} |f(s)| \in L_2(P)$, $\sqrt{n}(P_n - P)(f) \overset{\mathcal{L}}{\Rightarrow} B_P(f)$ if $\mathcal{F}$ is a VC-subgraph family of functions.

An important application of this theorem is the following result.

Corollary 12.1 Under the other assumptions made in Theorem 12.14, $\sqrt{n}(P_n - P)(f) \overset{\mathcal{L}}{\Rightarrow} B_P(f)$ if $\mathcal{F}$ is a finite-dimensional space of functions or if $\mathcal{F} = \{I_C : C \in \mathcal{C}\}$, where $\mathcal{C}$ is any VC class of sets.

Theorem 12.14 beautifully connects the scope of a Glivenko-Cantelli theorem to that of a CLT via the same VC concept, modulo the extra qualification that $F \in L_2(P)$. One can see more about this key theorem in Alexander (1984, 1987) and Giné (1996).

A pretty and useful statistical application of the result above is the following example on extension (due to Beran and Millar (1986)) of the familiar Kolmogorov-Smirnov test for goodness of fit to general spaces.

Example 12.8 Let $X_1, X_2, \cdots$ be iid observations from P on some space $\mathcal{S}$, and consider testing the null hypothesis $H_0 : P = P_0$ (specified). The natural Kolmogorov-Smirnov type test statistic for this problem is $T_n = \sqrt{n}\sup_{C\in\mathcal{C}} |P_n(C) - P_0(C)|$ for a judiciously chosen family of (measurable) sets $\mathcal{C}$. Theorem 12.14 implies that T_n converges under the null in distribution to the supremum of the absolute value of the Gaussian process $|B_{P_0}(f)|$, the sup being taken over all $f = I_C, C \in \mathcal{C}$, a VC class of subsets of $\mathcal{S}$. In principle, therefore, the null hypothesis can be tested by using this Kolmogorov-Smirnov type statistic. Note, however, that the limiting Gaussian process depends on P_0. Evaluation of the critical points of the limiting distribution of T_n under the null needs more work; see Giné (1996) for more discussion and references on this computational issue.

The second CLT we will present requires the concepts of metric entropy and bracketing numbers, which we introduce next.

Definition 12.9 Let $\mathcal{F}^*$ be a space of real-valued functions defined on some space $\mathcal{S}$, and suppose $\mathcal{F}^*$ is equipped with a norm $||.||$. Let $\mathcal{F}$ be a specific subcollection of $\mathcal{F}^*$. The *covering number* of $\mathcal{F}$ is defined to be the smallest number of balls $B(g, \epsilon) = \{h : ||h - g|| < \epsilon\}$ needed to cover $\mathcal{F}$, where $\epsilon > 0$ is arbitrary but fixed, $g \in \mathcal{F}^*$, and $||g|| < \infty$.

The covering number of $\mathcal{F}$ is denoted as $N(\epsilon, \mathcal{F}, ||.||)$. $\log N(\epsilon, \mathcal{F}, ||.||)$ is called the *entropy without bracketing* of $\mathcal{F}$.

Definition 12.10 In the same setup as in the previous definition, a *bracket* is the set of functions sandwiched between two given functions l, u (i.e., a bracket is the set $\{f : l(s) \le f(s) \le u(s) \forall s \in \mathcal{S}\}$). It is denoted as $[l, u]$.

Definition 12.11 The *bracketing number* of $\mathcal{F}$ is defined to be the smallest number of brackets $[l, u]$ needed to cover $\mathcal{F}$ under the restriction $||l - u|| < \epsilon\, with\, \epsilon > 0$ an arbitrary but fixed number.

The bracketing nnumber of $\mathcal{F}$ is denoted as $N_{[]}(\epsilon, \mathcal{F}, ||.||)$. $\log N_{[]}(\epsilon, \mathcal{F}, ||.||)$ is called the *entropy with bracketing* of $\mathcal{F}$.

Clearly, the smaller the radius of the balls or the width of the brackets, the greater the number of balls or brackets necessary to cover the function class $\mathcal{F}$. The important thing is to pin down qualitatively the rate at which the entropy (with or without bracketing) is going to ∞ for a given $\mathcal{F}$. It turns out, as we shall see, that for many interesting and useful classes of functions $\mathcal{F}$, this rate would be of the order of $(-\log \epsilon)$, and this will, by virtue of some theorems to be given below, ensure that the class $\mathcal{F}$ is P-Donsker.

Theorem 12.15 Assume that $F \in L_2(P)$. Then, $\mathcal{F}$ is P-Donsker if either

$$\int_0^\infty \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, ||.|| = L_2(P))} d\epsilon < \infty$$

or

$$\int_0^\infty \sup_Q (\sqrt{\log N(\epsilon ||F||_{2,Q}, \mathcal{F}, ||.|| = L_2(Q))}) d\epsilon < \infty,$$

where Q denotes a general probability measure on $\mathcal{S}$.

We have previously seen that if $\mathcal{F}$ is a VC-subgraph, then it is P-Donsker. It turns out that this result follows from Theorem 12.15 on the integrability of $\sup_Q \sqrt{\log N}$. What one needs is the following upper bound on the entropy without bracketing of a VC-subgraph class. See van der Vaart and Wellner (1996) for its proof.

Proposition 12.2 *Given a VC-subgraph class $\mathcal{F}$, for any probability measure Q and any $r \geq 1$, for all $0 < \epsilon < 1$, $N(\epsilon ||F||_{r,Q}, \mathcal{F}, ||.|| = L_r(Q)) \leq C(\frac{1}{\epsilon})^{r\,\mathrm{VC}(\mathcal{C})}$, where the constant C depends only on* $\mathrm{VC}(\mathcal{C})$, *$\mathcal{C}$ being the subgraph class of $\mathcal{F}$.*

Here are some additional good applications of the entropy results.

Example 12.9 As mentioned above, the key to the applicability of the entropy theorems is a good upper bound on the rate of growth of the entropy numbers of the class. Such bounds have been worked out for many intuitively interesting classes. The bounds are sometimes sharp in the sense that lower bounds can also be obtained that grow at the same rate as the upper bounds. In nearly every case mentioned in this example, the derivation of the upper bound is completely nontrivial. A very good reference is van der Vaart and Wellner (1996), particularly Chapter 2.7 there.

Uniformly Bounded Monotone Functions on $\mathcal{R}$
For this function class $\mathcal{F}$, $\log N_{[]}(\epsilon, \mathcal{F}, ||.|| = L_2(P)) \leq \frac{K}{\epsilon}$, where K is a universal constant independent of P, and so this class is in fact universal P-Donsker.

Uniformly Bounded Lipschitz Functions on Bounded Intervals in $\mathcal{R}$
Let $\mathcal{F}$ be the class of real-valued functions on a bounded interval $\mathcal{I}$ in $\mathcal{R}$ that are uniformly bounded by a universal constant and are uniformly Lipschitz of some order $\alpha > \frac{1}{2}$ (i.e., $|f(x) - f(y)| \leq M|x - y|^\alpha$) uniformly in x, y and for some finite universal constant M. For this class,

$\log N_{[]}(\epsilon, \mathcal{F}, ||.|| = L_2(P)) \leq K(\frac{1}{\epsilon})^{1/\alpha}$, where K depends only on the length of $\mathcal{I}$, M, and α, and so this class is also universal P-Donsker.

Compact Convex Subsets of a Fixed Compact Set in $\mathcal{R}^{\mathbf{d}}$
Suppose S is a compact set in $\mathcal{R}^d$ for some finite d, and let $\mathcal{C}$ be the class of all compact convex subsets of S. For any absolutely continuous P, this class satisfies $\log N_{[]}(\epsilon, \mathcal{C}, ||.|| = L_2(P)) \leq K(\frac{1}{\epsilon})^{d-1}$, where K depends on S, P, and d. Here it is meant that the function class is the set of indicators of the members of $\mathcal{C}$. Thus, for $d = 2$, $\mathcal{F}$ is P-Donsker for any absolutely continuous P.

A common implication of all of these applications of the entropy theorems is that, in the corresponding setups, asymptotic goodness-of-fit tests can be constructed by using these function classes.

12.9 Dependent Sequences: Martingales, Mixing, and Short-Range Dependence

Central limit theorems for certain types of dependent sequences were described in Chapters 9 and 10. The progression to a weak and then strong invariance principles for most of those processes was achieved by the mid-1970s; some key references are Billingsley (1956), Philip and Stout (1975), Hall (1977), and Andrews and Pollard (1994). McLeish (1975) unified much of the work by introducing what he called *mixingales*. Depending on the nature of the dependence, the norming constant for an invariance principle for the partial sum process can be $\sqrt{n}$, or something more complicated involving suitable moments of $|S_n|$. Merlevéde, Peligrad, and Utev (2006) have provided a modern review, including the latest technical innovations. We provide a brief treatment of a few classic results in this section.

Theorem 12.16 Let $\{X_k\}_{k\geq 0}$ be a stationary process with mean zero and finite variance. Assume the condition

$$\sum_{n=1}^{\infty} \frac{[E(S_n^2|X_0)]^{1/2}}{n^{3/2}} < \infty.$$

Let $S_n(t) = S_{n,1}(t), 0 \leq t \leq 1$. Then there exists a common probability space on which one can define a Wiener process $W(t)$ starting at zero and a sequence of processes $\{\hat{S}_n(t)\}, n \geq 1$, such that

(a) $\{\hat{S}_n(t), n \geq 1\} \stackrel{\mathcal{L}}{=} \{S_n(t), n \geq 1\}$;

(b) $\hat{S}_n(t) \stackrel{\mathcal{L}}{\Rightarrow} \eta W(t)$, where η is a nonnegative random variable (on the same probability space) with $E(\eta^2) = \lim_{n\to\infty} \frac{E(S_n^2)}{n}$.

Remark. The random variable η can in fact be explicitly characterized; see Merlevéde, Peligrad, and Utev (2006). If $\{X_k\}_{k\geq 0}$ is also *ergodic*, then η is a trivial random variable and $\eta^2 = E(X_0^2) + 2\sum_{k>0} E(X_k X_0)$.

Theorem 12.16 already applies to a broad variety of time series used in practice, although generalizations with norming different from $\sqrt{n}$ in the definition of $S_n(t)$ are available. To give specific applications of this theorem, we need some definitions. In particular, we need the definition of the concept of a *mixing sequence*, introduced in Rosenblatt (1956).

Definition 12.12 Given a process $\{X_k\}_{k\geq 0}$ (not necessarily stationary), let $\alpha(j,n) = \sup\{|P(A\cap B) - P(A)P(B)|, A \in \sigma(X_k, k \leq j), B \in \sigma(X_k, k \geq j+n)\}$. $\{X_k\}_{k\geq 0}$ is called strongly mixing if $\alpha(n) := \sup_j \alpha(j,n) \to 0$ as $n \to \infty$.

Definition 12.13 Given a process $\{X_k\}_{k\geq 0}$ (not necessarily stationary), let $\rho(j,n) = \sup\{\rho(f,g) : f \in L_2(\sigma(X_k, k \leq j)), g \in L_2(\sigma(X_k, k \geq j+n))\}$, where $\rho(f,g)$ denotes the correlation between f and g. $\{X_k\}_{k\geq 0}$ is called ρ-mixing if $\rho(n) := \sup_j \rho(j,n) \to 0$ as $n \to \infty$.

Definition 12.14 $\{X_k\}_{k\geq 0}$ is called a short-range dependent one-sided linear process if $X_k = \sum_{i\geq 0} a_i Z_{k-i}$, where $\{Z_k\}_{-\infty<k<\infty}$ is a martingale difference sequence and $\sum_{i=0}^{\infty} |a_i| < \infty$.

The general inequality $4\alpha(j,n) \leq \rho(j,n)$ is true, so that ρ-mixing implies strong mixing. For stationary Gaussian processes, the two concepts are equivalent. Many other concepts of mixing-type asymptotic independence are known. Among them is the concept of ϕ-mixing, which says that $P(B|A)$ and $P(B)$ should be uniformly close together in the sense of a small difference when A and B are events separated by a large lag. We will not discuss the other types of mixing here. A good general reference is Doukhan (1994). More specifically, for invariance principles for mixing empirical processes and the rates of convergence, two good references are Yu (1994) and Arcones and Yu (1994). Here is a result that gives good applications of Theorem 12.16 above; the assumptions of zero mean and finite variance are implicit below. We have not stated parts of Theorem 12.17 below under the best currently known conditions in order to avoid possibly hard-to-verify assumptions. Merlevéde, Peligrad, and Utev (2006) can be consulted for the best conditions or references to the best conditions.

Theorem 12.17 (a) If $\{X_k\}$ is stationary and strongly mixing satisfying

$$E(X_0^4) < \infty; \sum_{n=1}^{\infty} [\alpha(2^n)]^{1/4} < \infty; \mathrm{Var}(S_n) \to \infty; E(S_n)^4 = O(\mathrm{Var}(S_n))^2,$$

then Theorem 12.16 holds.

(b) If $\{X_k\}$ is stationary and ρ-mixing satisfying $\sum_{n=1}^{\infty} \rho(2^n) < \infty$, then Theorem 12.16 holds.

(c) If $\{X_k\}$ is a stationary and short-range dependent one-sided linear process, then Theorem 12.16 holds.

Example 12.10 Numerous examples of common processes with various mixing properties are known; a very recent reference is Bradley (2005). For example, obviously, a stationary m-dependent process for any finite m is strongly mixing and also ρ-mixing.

The mixing properties of Markov chains are intriguing. For example, the remarkable fact that if $\alpha(n) < \frac{1}{4}$ for some $n \geq 1$ then a strictly stationary ergodic aperiodic chain with state space $\mathcal{R}$ is strongly mixing is attributed in Bradley (2005) to be implicitly proved in Rosenblatt (1972). Likewise, if $\rho(n) < 1$ for some $n \geq 1$, then even without stationarity, the chain is ρ-mixing, and $\rho(n) \to 0$ at an exponential rate. If the state space is countable, then the conditions on the chain can be relaxed; see, again, Bradley (2005). A stationary autoregressive process of some finite order p is strongly mixing under quite mild conditions. If the roots of the characteristic polynomial of the process are all inside the unit circle and the errors are iid with a finite mean and have a density, then the process is strongly mixing. See Chanda (1974), Withers (1981), and Athreya and Pantula (1986); useful information is also given in Koul (1977) and Gastwirth and Rubin (1975).

Central limit theorems for martingales were described in Chapter 9. We commented there that martingale central limit theorems can be obtained with random or nonrandom norming; which one holds depends on exactly what assumptions one makes. Likewise, invariance principles for martingales (or, more generally, martingale arrays) have been obtained under various sets of conditions. One possibility is to assume a kind of Lindeberg condition; alternatively, one can assume suitable growth conditions in terms of L_p norms; see Brown (1971), McLeish (1974), and Hall (1977). Here we report presumably the first invariance principle obtained for martingale sequences by assuming a Lindeberg type condition; see Brown (1971). These are not the weakest conditions under which an invariance principle obtains; the other

references above give more general theorems under somewhat weaker conditions.

We follow the notation in Brown (1971). Let $\{S_n, \mathcal{F}_n\}_{n\geq 1}$ be a martingale and let $X_n = S_n - S_{n-1}$, $n \geq 1$, with $S_0 = 0$. Let $E_{j-1}(.)$ denote conditional expectation given $\mathcal{F}_{j-1}$. Define

$$\sigma_n^2 = E_{n-1}(X_n^2); V_n^2 = \sum_{j=1}^{n} \sigma_j^2; s_n^2 = E(S_n^2)(= E(V_n^2)).$$

Consider the piecewise linear function $\xi_n(t)$ on $[0, 1]$ that joins the discrete set of points $(s_k^2/s_n^2, S_k/s_n)$, $0 \leq k \leq n$. Thus,

$$\xi_n(t) = s_n^{-1}[S_k + X_{k+1}(ts_n^2 - s_k^2)/(s_{k+1}^2 - s_k^2)], s_k^2 \leq ts_n^2 \leq s_{k+1}^2,$$

$0 \leq k \leq n$. The invariance principle addresses the question of weak convergence of the process $\xi_n(t)$ as an element of $C[0, 1]$.

Assume the following:

$$(1) \qquad V_n^2/s_n^2 \stackrel{P}{\Rightarrow} 1,$$

$$(2) \quad s_n^{-2} \sum_{j=1}^{n} E[X_j^2 I_{|X_j|\geq \epsilon s_n}] = o_p(1).$$

Condition (2) is the *Lindeberg condition for Martingales*. The following result is proved in Brown (1971).

Theorem 12.18 Under conditions (1) and (2) stated above, the assertion of Donsker's theorem holds.

12.10 Weighted Empirical Processes and Approximations

Recently, there has been a growth in interest in studying weighted empirical processes due to their use in modern multiple testing problems. We will see the actual uses in Chapter 34. It turns out that asymptotic behavior of weighted empirical processes is surprisingly subtle. The subtlety comes from mutual interaction of the tail of a Wiener process and that of the weighting function. For example, since the normalized uniform empirical process $\alpha_n(t) = \sqrt{n}(G_n(t) - t)$ behaves like a Brownian bridge asymptotically, one might hope that for a strictly positive function $\delta(t)$, $\frac{\alpha_n(t)}{\delta(t)}$ may behave asymptotically like the weighted Brownian bridge $\frac{B(t)}{\delta(t)}$ on $[0, 1]$. This is not true

for some natural choices of the weighting function $\delta(t)$; even more, it fails because of the tail behaviors. If we truncated the interval $[0, 1]$ at suitable rates, the intuition would in fact work. The weighting functions $\delta(t)$ that do admit an invariance principle have no simple descriptions. Characterizing them requires using deep tail properties of the Wiener sample paths. Because of the extremely high current interest in weighted empirical processes in the multiple testing literature, we give a short description of the asymptotics and invariance principles for them. We consider the case of the uniform empirical process, as the general case can be reduced to it by a time change. We recommend Csörgo and Horváth (1993) for the topic of this section.

First we need some notation. The operators that we define below are related to what are called *lower and upper functions* of Wiener processes; see Section 12.2 in this chapter for some information on what the link is.

Let

$$\mathcal{F}_{0,1} = \{\delta : \inf_{\epsilon<t<1-\epsilon}\delta(t) > 0 \quad \forall \quad 0 < \epsilon < \frac{1}{2}, \delta \uparrow \text{near}\, 0, \delta \downarrow \text{near}\, 1\}$$

$$I(c, \delta) = \int_0^1 \frac{1}{t(1-t)} e^{-c\delta^2(t)/[t(1-t)]} dt$$

$$E(c, \delta) = \int_0^1 \frac{\delta(t)}{[t(1-t)]^{3/2}} e^{-c\delta^2(t)/[t(1-t)]} dt.$$

Whether $\frac{\alpha_n(t)}{\delta(t)}$ is uniformly asymptotically close to a correspondingly weighted Brownian bridge is determined by the two operators I, E. Here is a complete characterization; see Csörgo (2002) for Theorems 12.19–12.21.

Theorem 12.19 Let $\delta \in \mathcal{F}_{0,1}$ and let $\alpha_n(t)$ be the normalized uniform empirical process $\alpha_n(t) = \sqrt{n}(G_n(t) - t), t \in [0, 1]$. Then, there exists a sequence of Brownian bridges $B_n(t)$ (on the same space) such that $\sup_{t\in(0,1)} \frac{|\alpha_n(t)-B_n(t)|}{\delta(t)} = o_p(1)$ if and only if

$$I(c, \delta) < \infty \, \forall \, c > 0$$

or

$$E(c, \delta) < \infty \, \forall \, c > 0, \lim_{t\to 0} \delta(t)/\sqrt{t} = \lim_{t\to 1} \delta(t)/\sqrt{1-t} = \infty.$$

Example 12.11 A natural choice for the weighting function is $\delta(t) = \sqrt{t(1-t)}$; after all, the pointwise variance of $\alpha_n(t)$ is $t(1-t)$. Clearly $\sqrt{t(1-t)} \in \mathcal{F}_{0,1}$. But, as is obvious, for no $c > 0$, $I(c, \delta)$ is finite. Thus, $\frac{\alpha_n(t)}{\sqrt{t(1-t)}}$ cannot be uniformly asymptotically approximated in probability by any sequence of the correspondingly weighted Brownian bridges over the whole unit interval. The tails create the problem. In fact, for any sequence of Brownian bridges whatsoever,

$$P\left(\sup_{t\in(0,1)} \frac{|\alpha_n(t) - B_n(t)|}{\sqrt{t(1-t)}} = \infty\right) = 1.$$

If we truncate the tails, we can control the maximum weighted deviation. The next result makes it precise.

Theorem 12.20 For any $u > 0$ and any $0 < \nu \le \frac{1}{2}$, there exist Brownian bridges $B_n(t)$ such that

$$n^{1/2-\nu} \sup_{t\in[\frac{u}{n}, 1-\frac{u}{n}]} \frac{|\alpha_n(t) - B_n(t)|}{[t(1-t)]^\nu} = O_p(1),$$

while, for $0 < \nu < \frac{1}{2}$,

$$n^{1/2-\nu} \sup_{t\in(0,1)} \frac{|\alpha_n(t) - B_n(t)|}{[t(1-t)]^\nu} = O_p(1).$$

Furthermore, the bounds cannot be made $o_p(1)$.

Note that we should not expect convergence in law of $\frac{\alpha_n(t)}{\sqrt{t(1-t)}}$ to $\frac{B(t)}{\sqrt{t(1-t)}}$ because of the failure to have a uniform $o_p(1)$ bound on the maximum deviation, as we just saw. The question arises as to when we can ensure a weak convergence result. The answer is essentially already contained in Theorem 12.19.

Theorem 12.21 Under the assumptions of Theorem 12.19,

$$\sup_{t\in(0,1)} \frac{|\alpha_n(t)|}{\delta(t)} \stackrel{\mathcal{L}}{\Rightarrow} \sup_{t\in(0,1)} \frac{|B(t)|}{\delta(t)},$$

where $B(t)$ is a Brownian bridge.

Remark. An important relaxation in Theorem 12.21 is that the "for all c" requirement can be relaxed to "for some c"; sometimes this can be useful.

Although Theorem 12.21 specifies the limiting distribution of $\sup_{t\in(0,1)} \frac{|\alpha_n(t)|}{\delta(t)}$ for appropriate $\delta(t)$, the following inequality of Birnbaum and Marshall (1961) is useful because it is explicit and holds for all n.

Proposition 12.3 *Suppose $\delta(t)$ is right continuous and belongs to $\mathcal{F}_{0,1}$, and $M = \int_0^1 \delta^{-2}(t)dt < \infty$. Then, $\forall\, n, x$, $P(\sup_{t\in(0,1)} \frac{|\alpha_n(t)|}{\delta(t)} > x) \leq \frac{M}{x^2}$.*

Example 12.12 The results imply that a symmetric function $\delta(t)$ is not amenable to an invariance principle for the weighted uniform empirical process if $\delta(t) \sim \sqrt{t}$ near zero. But, for example, $\delta(t) = [t(1-t)]^{1/3}$works, and $\sup_{t\in(0,1)} \frac{\alpha_n(t)}{[t(1-t)]^{1/3}}$ will converge weakly to $\sup_{t\in(0,1)} \frac{B(t)}{[t(1-t)]^{1/3}}$. Explicit evaluation of the distribution of this latter functional (or similar ones corresponding to other $\delta(t)$) usually would not be possible. However, bounds on their CDFs often would be possible.

12.11 Exercises

Exercise 12.1 For $n = 25$ and 50, approximate the probability $P(\max_{1\leq k\leq n} |S_k| > 2\sqrt{n})$ when the sample observations are iid $U[-1, 1]$. Does the $U[-1, 1]$ assumption have any role in your approximation?

Exercise 12.2 * Plot the density function of G_2, the limiting CDF of the normalized maximum of absolute partial sums.

Exercise 12.3 For $n = 25$ and 50, approximate the probability that at least 60% of the time, a simple symmetric random walk remains over the axis.

Exercise 12.4 Give examples of three functions that are members of $D[0, 1]$ but not of $C[0, 1]$.

Exercise 12.5 * Prove that each of the functionals h_i, $i = 1, 2, 3, 4$ are continuous on $C[0, 1]$ with respect to the uniform metric.

Exercise 12.6 * Why is the functional h_5 not everywhere continuous with respect to the uniform metric?

Exercise 12.7 * Approximately simulate 20 paths of a Brownian motion by using its Karhunen-Loeve expansion (suitably truncated).

Exercise 12.8 Formally prove that the CLT for partial sums follows from the strong invariance principle available if four moments are assumed to be finite.

Exercise 12.9 * Compute and plot a 95% nonparametric confidence band for the CDF based on the KMT theorem for $n = 100, 500$ when the data are simulated from $U[0, 1]$, $N[0, 1]$.

Exercise 12.10 * Find the VC dimension of the following classes of sets :

(a) southwest quadrants of $\mathcal{R}^d$;
(b) closed half-spaces of $\mathcal{R}^d$;
(c) closed balls of $\mathcal{R}^d$;
(d) closed rectangles of $\mathcal{R}^d$.

Exercise 12.11 Give examples of three nontrivial classes of sets in $\mathcal{R}^d$ that are not VC classes.

Exercise 12.12 * Design a test for testing that sample observations in $\mathcal{R}^2$ are iid from a uniform distribution in the unit square by using suitable VC classes and applying the CLT for empirical measures.

Exercise 12.13 * Find the VC dimension of all polygons in the plane with four vertices.

Exercise 12.14 * Is the VC dimension of the class of all ellipsoids of $\mathcal{R}^d$ the same as that of the class of all closed balls of $\mathcal{R}^d$?

Exercise 12.15 * Consider the class of *Box-Cox transformations* $\mathcal{F} = \{\frac{x^\lambda - 1}{\lambda}, x > 0, \lambda \neq 0\}$. Show that $\mathcal{F}$ is a VC-subgraph class (see p. 153 in van der Vaart and Wellner (1996) for hints).

Exercise 12.16 Give an example of a stationary Gaussian process for which the condition $\sum \rho(2^n) < \infty$ holds and a few examples where the condition does not hold.

Exercise 12.17 * Prove the general inequality $4\alpha(j, n) \leq \rho(j, n)$.

Exercise 12.18 * Define $\psi(j, n) = \sup\{|\frac{P(A \cap B)}{P(A)P(B)} - 1| : A \in \sigma(X_k, 0 \leq k \leq j), B \in \sigma(X_k, k \geq j + n)\}$, and call $\{X_k\}_{k \geq 0}$ ψ-mixing if $\sup_j \psi(j, n) \to 0$ as $n \to \infty$. Show that

(a) stationary m-dependence implies ψ-mixing;

(b) ψ-mixing implies ρ-mixing.
Note: Part (b) is hard.

Exercise 12.19 Suppose $X_k = \sum_{i=0}^{\infty} \frac{1}{(i+1)^2} \epsilon_{k-i}$, where ϵ_j are iid $U[-1, 1]$. Is $\{X_k\}_{k \geq 0}$ a short-range dependent process? Is X_k summable with probability 1?

Exercise 12.20 * Derive a martingale central limit theorem with nonrandom norming by using Theorem 12.18.

Exercise 12.21 * Prove that $P\left(\sup_{t \in (0,1)} \frac{|\alpha_n(t)|}{\sqrt{t(1-t)}} = \infty\right) = 1 \, \forall \, n \geq 1$. Hint: Look at t near zero and consider the law of the iterated logarithm.

Exercise 12.22 Give examples of functions $\delta(t)$ that satisfy the assumptions of the Birnbaum-Marshall inequality in Section 12.10.

Exercise 12.23 * Approximate the probability $P\left(\sup_{t \in (0,1)} \frac{|\alpha_n(t)|}{[t(1-t)]^{1/3}} > x\right)$ by using the weak convergence result of Theorem 12.21.

References

Alexander, K. (1984). Probability inequalities for empirical processes and a law of the iterated logarithm, Ann. Prob., 12, 1041–1067.

Alexander, K. (1987). The central limit theorem for empirical processes on Vapnik-Chervonenkis classes, Ann. Prob., 15, 178–203.

Andrews, D. and Pollard, D. (1994). An introduction to functional central limit theorems for dependent stochastic processes, Int. Stat. Rev., 62, 119–132.

Arcones, M. and Yu, B. (1994). Central limit theorems for empirical and U-processes of stationary mixing sequences, J. Theor. Prob., 1, 47–71.

Athreya, K. and Pantula, S. (1986). Mixing properties of Harris chains and autoregressive processes, J. Appl. Prob., 23, 880–892.

Beran, R. and Millar, P. (1986). Confidence sets for a multinomial distribution, Ann. Stat., 14, 431–443.

Billingsley, P. (1956). The invariance principle for dependent random variables, Trans. Am. Math. Soc., 83(1), 250–268.

Billingsley, P. (1968). *Convergence of Probability Measures*, John Wiley, New York.

Birnbaum, Z. and Marshall, A. (1961). Some multivariate Chebyshev inequalities with extensions to continuous parameter processes, Ann. Math. Stat., 32, 687–703.

Bradley, R. (2005). Basic properties of strong mixing conditions: a survey and some open problems, Prob. Surv., 2, 107–144.

Brillinger, D. (1969). An asymptotic representation of the sample df, Bull. Am. Math. Soc., 75, 545–547.

Brown, B. (1971). Martingale central limit theorems, Ann. Math. Stat., 42, 59–66.
Cameron, R. and Martin, W. (1945). Evaluation of various Wiener integrals by use of certain Sturm-Liouville differential equations, Bull. Am. Math. Soc., 51, 73–90.
Chanda, K. (1974). Strong mixing properties of linear stochastic processes, J. Appl. Prob., 11, 401–408.
Csörgo, M. and Révész, P. (1981). *Strong Approximations in Probability and Statistics*, Academic Press, New York.
Csörgo, M. and Horváth, L. (1993). *Weighted Approximations in Probability and Statistics*, John Wiley, New York.
Csörgo, M. (1984). Invariance principle for empirical processes, in *Handbook of Statistics*, P. K. Sen and P. R. Krishraiah (eds.), Vol. 4, North-Holland, Amsterdam, 431–462.
Csörgo, M. (2002). A glimpse of the impact of Paul Erdös on probability and statistics, Can. J. Stat., 30(4), 493–556.
Csörgo, S. and Hall, P. (1984). The KMT approximations and their applications, Aust. J. Stat., 26(2), 189–218.
Donsker, M. (1951). An invariance principle for certain probability limit theorems, Mem. Am. Math. Soc., 6.
Doukhan, P. (1994). *Mixing: Properties and Examples*, Lecture Notes in Statistics, Vol. 85, Springer, New York.
Dudley, R. (1978). Central limit theorems for empirical measures, Ann. Prob., 6, 899–929.
Dudley, R. (1979). Central limit theorems for empirical measures, Ann. Prob., 7(5), 909–911.
Dudley, R. and Philipp, W. (1983). Invariance principles for sums of Banach space valued random elements and empirical processes, Z. Wahr. Verw. Geb., 62, 509–552.
Dudley, R. (1984). *A Course on Empirical Processes*, Lecture Notes in Mathematics, Springer, Berlin.
Durrett, R. (1996). *Probability: Theory and Examples*, 2nd ed., Duxbury Press, Belmont, CA.
Einmahl, U. (1987). Strong invariance principles for partial sums of independent random vectors, Ann. Prob., 15(4), 1419–1440.
Erdös, P. and Kac, M. (1946). On certain limit theorems of the theory of Probability, Bull. Am. Math. Soc., 52, 292–302.
Fitzsimmons, P. and Pitman, J. (1999). Kac's moment formula and the Feynman-Kac formula for additive functionals of a Markov process, Stoch. Proc. Appl., 79(1), 117–134.
Gastwirth, J. and Rubin, H. (1975). The asymptotic distribution theory of the empiric cdf for mixing stochastic processes, Ann. Stat., 3, 809–824.
Giné, E. (1996). Empirical processes and applications: an overview, Bernoulli, 2(1), 1–28.
Giné, E. and Zinn, J. (1984). Some limit theorems for empirical processes, with discussion, Ann. Prob., 12(4), 929–998.
Hall, P. (1977). Martingale invariance principles, Ann. Prob., 5(6), 875–887.
Hall, P. and Heyde, C. (1980). *Martingale Limit Theory and Its Applications*, Academic Press, New York.
Heyde, C. (1981). Invariance principles in statistics, Int. Stat. Rev., 49(2), 143–152.
Jain, N., Jogdeo, K., and Stout, W. (1975). Upper and lower functions for martingales and mixing processes, Ann. Prob., 3, 119–145.

Kac, M. (1951). On some connections between probability theory and differential and integral equations, in Proceedings of the Second Berkeley Symposium, J. Neyman (ed.), University of California Press, Berkeley, 189–215.
Kiefer, J. (1972). Skorohod embedding of multivariate rvs and the sample df, Z. Wahr. Verw. Geb., 24, 1–35.
Kolmogorov, A. (1933). Izv. Akad. Nauk SSSR, 7, 363–372 (in German).
Komlós, J., Major, P., and Tusnady, G. (1975). An approximation of partial sums of independent rvs and the sample df: I, Z. Wahr. Verw. Geb., 32, 111–131.
Komlós, J., Major, P., and Tusnady, G. (1976). An approximation of partial sums of independent rvs and the sample df: II, Z. Wahr. Verw. Geb., 34, 33–58.
Koul, H. (1977). Behavior of robust estimators in the regression model with dependent errors, Ann. Stat., 5, 681–699.
Major, P. (1978). On the invariance principle for sums of iid random variables, J. Multivar. Anal., 8, 487–517.
Mandrekar, V. and Rao, B.V. (1989). On a limit theorem and invariance principle for symmetric statistics, Prob. Math. Stat., 10, 271–276.
Massart, P. (1989). Strong approximation for multivariate empirical and related processes, via KMT construction, Ann. Prob., 17(1), 266–291.
McLeish, D. (1974). Dependent central limit theorems and invariance principles, Ann. Prob., 2, 620–628.
McLeish, D. (1975). Invariance principles for dependent variables, Z. Wahr. Verw. Geb., 3, 165–178.
Merlevéde, F., Peligrad, M., and Utev, S. (2006). Recent advances in invariance principles for stationary sequences, Prob. Surv., 3, 1–36.
Oblój, J. (2004). The Skorohod embedding problem and its offspring, Prob. Surv., 1, 321–390.
Philipp, W. (1979). Almost sure invariance principles for sums of *B*-valued random variables, in *Problems in Banach Spaces*, A. Beck (ed.), Vol. II, Lecture Notes in Mathematics, Vol. 709, Springer, Berlin, 171–193.
Philipp, W. and Stout, W. (1975). Almost sure invariance principles for partial sums of weakly dependent random variables, Mem. Am. Math. Soc., 2, 161.
Pollard, D. (1989). Asymptotics via empirical processes, Stat. Sci., 4, 341–366.
Pyke, R. (1984). Asymptotic results for empirical and partial sum processes: a review, Can. J. Stat., 12, 241–264.
Révész, P. (1976). On strong approximation of the multidimensional empirical process, Ann. Prob., 4, 729–743.
Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition, Proc. Natl. Acad. Sci. USA, 42, 43–47.
Rosenblatt, M. (1972). Uniform ergodicity and strong mixing, Z. Wahr. Verw. Geb., 24, 79–84.
Sauer, N. (1972). On the density of families of sets, J. Comb. Theory Ser. A, 13, 145–147.
Sen, P.K. (1978). An invariance principle for linear combinations of order statistics, Z. Wahr. Verw. Geb., 42(4), 327–340.
Shorack, G. and Wellner, J. (1986). *Empirical Processes with Applications to Statistics*, John Wiley, New York.
Smirnov, N. (1944). Approximate laws of distribution of random variables from empirical data, Usp. Mat. Nauk., 10, 179–206.

Strassen, V. (1964). An invariance principle for the law of the iterated logarithm, Z. Wahr. Verw. Geb., 3, 211–226.

Strassen, V. (1967). Almost sure behavior of sums of independent random variables and martingales, in *Proceedings of the Fifth Berkeley Symposium*, L. Le Cam and J. Neyman (eds.), Vol. 1 University of California Press, Berkeley, 315–343.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes*, Springer-Verlag, New York.

Vapnik, V. and Chervonenkis, A. (1971). On the uniform convergence of relative frequencies of events to their probabilities, Theory Prob. Appl., 16, 264–280.

Wellner, J. (1992). Empirical processes in action: a review, Int. Stat. Rev., 60(3), 247–269.

Whithers, C. (1981). Conditions for linear processes to be strong mixing, Z. Wahr. Verw. Geb., 57, 477–480.

Whitt, W. (1980). Some useful functions for functional limit theorems, Math. Oper. Res., 5, 67–85.

Yu, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences, Ann. Prob., 22, 94–116.

Chapter 2
Metrics, Information Theory, Convergence, and Poisson Approximations

Sometimes it is technically convenient to prove a certain type of convergence by proving that, for some suitable metric d on the set of CDFs, $d(F_n, F) \to 0$ instead of proving the required convergence directly from the definition. Here F_n, F are CDFs on some space, say the real line. Metrics are also useful as statistical tools to assess errors in distribution estimation and to study convergence properties in such statistical problems. The metric, of course, will depend on the type of convergence desired.

The central limit theorem justifiably occupies a prominent place in all of statistics and probability theory. Fourier methods are most commonly used to prove the central limit theorem. This is technically efficient but fails to supply any intuition as to *why* the result should be true. It is interesting that proofs of the central limit theorem have been obtained that avoid Fourier methods and use instead much more intuitive information-theoretic methods. These proofs use convergence of entropies and Fisher information in order to conclude convergence in law to normality. It was then realized that such information-theoretic methods are useful also to establish convergence to Poisson limits in suitable paradigms; for example, convergence of appropriate Bernoulli sums to a Poisson limit. In any case, Poisson approximations are extremely useful in numerous complicated problems in both probability theory and statistics. In this chapter, we give an introduction to the use of metrics and information-theoretic tools for establishing convergences and also give an introduction to Poisson approximations.

Good references on metrics on distributions are Dudley (1989), Rachev (1991), and Reiss (1989). The role of information theory in establishing central limit theorems can be seen, among many references, in Linnik (1959), Brown (1982), and Barron (1986). Poisson approximations have a long history. There are first-generation methods and then there are the modern methods, often called the *Stein-Chen methods*. The literature is huge. A few references are LeCam (1960), Sevas'tyanov (1972), Stein (1972, 1986),

A. DasGupta, *Asymptotic Theory of Statistics and Probability*,

Chen (1975), and Barbour, Holst and Janson (1992). Two other references where interesting applications are given in an easily readable style are Arratia, Goldstein, and Gordon (1990) and Diaconis and Holmes (2004).

2.1 Some Common Metrics and Their Usefulness

There are numerous metrics and distances on probability distributions on Euclidean spaces. The choice depends on the exact purpose and on technical feasibility. We mention a few important ones only and give some information about their interrelationships, primarily in the form of inequalities. The inequalities are good to know in any case.

(i) Metric for convergence in probability
$d_E(X, Y) = E\left(\frac{|X-Y|}{1+|X-Y|}\right)$. This extends to the multidimensional case in the obvious way by using the Euclidean norm $||X - Y||$.
(ii) Kolmogorov metric
$d_K(F, G) = \sup_x |F(x) - G(x)|$. This definition includes the multidimensional case.
(iii) Lévy metric
$d_L(F, G) = \inf\{\epsilon > 0 : F(x-\epsilon) - \epsilon \le G(x) \le F(x+\epsilon) + \epsilon \quad \forall x\}$.
(iv) Total variation metric
$d_{TV}(P, Q) = \sup_{\text{Borel } A} |P(A) - Q(A)|$. This also includes the multidimensional case. If P, Q are both absolutely continuous with respect to some measure μ, then $d_{\text{TV}}(P, Q) = \frac{1}{2}\int |f(x) - g(x)| d\mu(x)$, where f is the density of P with respect to μ and g is the density of Q with respect to μ.
(v) Kullback-Leibler distance
$K(P, Q) = -\int(\log \frac{q}{p})dP = -\int(\log \frac{q}{p})p d\mu$, where $p = \frac{dP}{d\mu}$ and $q = \frac{dQ}{d\mu}$ for some μ. Again, the multidimensional case is included. Note that K is not symmetric in its arguments P, Q.
(vi) Hellinger distance
$H(P, Q) = \left[\int(\sqrt{p} - \sqrt{q})^2 d\mu\right]^{1/2}$, where again $p = \frac{dP}{d\mu}$ and $q = \frac{dQ}{d\mu}$ for some μ, and the multidimensional case is included.

Theorem 2.1

(i) $X_n \xrightarrow{\mathcal{P}} X$ iff $d_E(X_n, X) \to 0$.
(ii) $X_n \xrightarrow{\mathcal{L}} X$ iff $d_L(F_n, F) \to 0$, where $X_n \sim F_n$ and $X \sim F$.

(iii) $X_n \xrightarrow{\mathcal{L}} X$ if $d_K(F_n, F) \to 0$, the reverse being true only under additional conditions.
(iv) If $X \sim F$, where F is continuous and $X_n \sim F_n$, then $X_n \xrightarrow{\mathcal{L}} X$ iff $d_K(F_n, F) \to 0$ (Polyá's theorem).
(v) $X_n \xrightarrow{\mathcal{L}} X$ if $d_{TV}(P_n, P) \to 0$, where $X_n \sim P_n$, $X \sim P$ (the converse is not necessarily true).
(vi) $H(P, Q) \leq \sqrt{K(P, Q)}$.
(vii) $H(P, Q) \geq d_{\mathrm{TV}}(P, Q)$.
(viii) $H(P, Q)/\sqrt{2} \leq \sqrt{d_{\mathrm{TV}}(P, Q)}$.

Proofs of parts of Theorem 2.1 are available in Reiss (1989).

Corollary 2.1 (a) The total variation distance and the Hellinger distance are equivalent in the sense $d_{TV}(P_n, P) \to 0 \Leftrightarrow H(P_n, P) \to 0$.

(b) If P_n, P are all absolutely continuous with unimodal densities, and if P_n converges to P in law, then $H(P_n, P) \to 0$.

(c) Convergence in Kullback-Leibler distance implies convergence in total variation and hence convergence in law.

Note that the proof of part (b) also uses Ibragimov's theorem stated below.

Remark. The Kullback-Leibler distance is very popular in statistics. Specifically, it is frequently used in problems of model selection, testing for goodness of fit, Bayesian modeling and Bayesian asymptotics, and in certain estimation methods known as minimum distance estimation. The Kolmogorov distance is one of the easier ones computationally and has been used in many problems, too, and notably so in the literature on robustness and Bayesian robustness. The Hellinger distance is a popular one in problems of density estimation and in time series problems. The Lévy metric is technically hard to work with but metrizes weak convergence, a very useful property. It, too, has been used in the robustness literature, but it is more common in probability theory. Convergence in total variation is extremely strong, and many statisticians seem to consider it unimportant. But it has a direct connection to $\mathcal{L}_1$ distance, which is intuitive. It has a transformation invariance property and, when it holds, convergence in total variation is extremely comforting.

Notice the last two parts in Theorem 2.1. We have inequalities in *both directions* relating the total variation distance to the Hellinger distance. Since computation of the total variation distance is usually difficult, Hellinger distances are useful in establishing useful bounds on total variation.

2.2 Convergence in Total Variation and Further Useful Formulas

Next, we state three important results on when convergence in total variation can be asserted; see Reiss (1989) for all three theorems and also almost any text on probability for a proof of Scheffé's theorem.

Theorem 2.2 (Scheffé) Let $f_n, n \geq 0$ be a sequence of densities with respect to some measure μ. If $f_n \to f_0$ a.e. (μ), then $d_{\mathrm{TV}}(f_n, f_0) \to 0$.

Remark. Certain converses to Scheffé's theorem are available, and the most recent results are due to Sweeting (1986) and Boos (1985). As we remarked before, convergence in total variation is very strong, and even for the simplest weak convergence problems, convergence in total variation should not be expected without some additional structure. The following theorem exemplifies what kind of structure may be necessary. This is a general theorem (i.e., no assumptions are made on the structural forms of the statistics). In the Theorem 2.4 below, convergence in total variation is considered for sample means of iid random variables (i.e., there is a restriction on the structural form of the underlying statistics). It is not surprising that this theorem needs fewer conditions than Theorem 2.3 to assert convergence in total variation.

Theorem 2.3 (Ibragimov) Suppose P_0 and (for large n) P_n are unimodal, with densities $f_0 = \frac{dP_0}{d\lambda}$ and $f_n = \frac{dP_n}{d\lambda}$, where λ denotes Lebesgue measure. Then $P_n \xrightarrow{\mathcal{L}} P_0$ iff $d_{\mathrm{TV}}(P_n, P_0) \to 0$.

Definition 2.1 A random variable X is said to have a lattice distribution if it is supported on a set of the form $\{a + nh : n \in \mathcal{Z}\}$, where a is a fixed real, h a fixed positive real, and $\mathcal{Z}$ the set of integers.

Theorem 2.4 Suppose $X_1, \ldots, X_n$ are iid nonlattice random variables with a finite variance and characteristic function $\psi(t)$. If, for some $p \geq 1$, $\psi \in L^p(\lambda)$, where λ denotes Lebesgue measure, then $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$ converges to $N(0, 1)$ in total variation.

Example 2.1 Suppose X_n is a sequence of random variables on $[0, 1]$ with density $f_n(x) = 1 + \cos(2\pi nx)$. Then, $X_n \overset{\mathcal{L}}{\Rightarrow} U[0, 1]$ by a direct verification of the definition using CDFs. However, note that the densities f_n *do not* converge to the uniform density 1 as $n \to \infty$. The limit distribution P_0 is unimodal, but the distribution P_n of X_n is not unimodal. The example

shows that the condition in Ibragimov's theorem above that the P_n need to be unimodal as well cannot be relaxed.

Example 2.2 Suppose $X_1, X_2, \ldots$ are iid $\chi^2(2)$ with density $\frac{1}{2}e^{-x/2}$. The characteristic function of X_1 is $\psi(t) = \frac{1}{1-2it}$, which is in $L^p(\lambda)$ for any $p > 1$. Hence, by Theorem 2.4, $\frac{\sqrt{n}(\overline{X}-2)}{2}$ converges in total variation to $N(0,1)$. We now verify that in fact the density of $Z_n = \frac{\sqrt{n}(\overline{X}-2)}{2}$ converges pointwise to the $N(0,1)$ density, which by Scheffé's theorem will also imply convergence in total variation. The pointwise convergence of the density is an interesting calculation.

Since $S_n = \sum_{i=1}^n X_i$ has the $\chi^2(2n)$ distribution with density $\frac{e^{-x/2}x^{n-1}}{2^n\Gamma(n)}$, Z_n has density $f_n(z) = \frac{e^{-(z\sqrt{n}+n)}(1+\frac{z}{\sqrt{n}})^{n-1}n^{n-\frac{1}{2}}}{\Gamma(n)}$. Hence, $\log f_n(z) = -z\sqrt{n} - n + (n-1)(\frac{z}{\sqrt{n}} - \frac{z^2}{2n} + O(n^{-3/2})) + (n-\frac{1}{2})\log n - \log\Gamma(n) = -z\sqrt{n} - n + (n-1)(\frac{z}{\sqrt{n}} - \frac{z^2}{2n} + O(n^{-3/2})) + (n-\frac{1}{2})\log n - (n\log n - n - \frac{1}{2}\log n + \log\sqrt{2\pi} + O(n^{-1}))$ on using Stirling's approximation for $\log\Gamma(n)$.

On canceling terms, this gives $\log f_n(z) = -\frac{z}{\sqrt{n}} - \log\sqrt{2\pi} - \frac{(n-1)z^2}{2n} + O(n^{-1/2})$, implying that $\log f_n(z) \to -\log\sqrt{2\pi} - \frac{z^2}{2}$, and hence $f_n(z) \to \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$, establishing the pointwise density convergence.

Example 2.3 The Hellinger and the Kullback-Leibler distances are generally easier to calculate than the total variation distance. The normal case itself is a good example. For instance, the Kullback-Leibler distance $K(N_p(\mu, \mathbf{I}), N_p(0, \mathbf{I})) = \frac{1}{2}||\mu||^2$.

Many bounds on the total variation distance between two multivariate normal distributions are known; we mention a few below that are relatively neat.

$$d_{\mathrm{TV}}(N_p(\mu_1, \mathbf{I}), N_p(\mu_2, \mathbf{I})) \le \frac{1}{\sqrt{2}}||\mu_1 - \mu_2||,$$

$$d_{\mathrm{TV}}(N_p(0, \Sigma), N_p(0, \mathbf{I})) \le \min \begin{cases} \frac{1}{\sqrt{2}}\left(\sum_{i=1}^p(\sigma_i^2 - 1) - \log|\Sigma|\right)^{\frac{1}{2}} \\ p2^{p+1}||\Sigma - \mathbf{I}||_2 \end{cases},$$

where $||A||_2$ denotes the usual Euclidean matrix norm $(\sum_i \sum_j a_{ij}^2)^{1/2}$. These and other bounds can be seen in Reiss (1989).

Example 2.4 Suppose $X_n \sim N(\mu_n, \sigma_n^2)$ and $X_0 \sim N(\mu, \sigma^2)$. Then X_n converges to X_0 in total variation if and only if $\mu_n \to \mu$ and $\sigma_n^2 \to \sigma^2$. This can be proved directly by calculation.

Remark. There is some interest in finding projections in total variation of a fixed distribution to a given class of distributions. This is a good problem but usually very hard, and even in simple one-dimensional cases, the projection can only be found by numerical means. Here is an example; the exercises at the end of the chapter offer some more cases.

Example 2.5 If $X_n \sim \text{Bin}(n, p_n)$ and $np_n \to \lambda, 0 < \lambda < \infty$, then X_n converges in law to the Poi(λ) distribution. In practice, this result is used to approximate a Bin(n, p) distribution for large n and small p by a Poisson distribution with mean np. One can ask what is the best Poisson approximation for a given Bin(n, p) distribution (e.g., what is the total variation projection of a given Bin(n, p) distribution onto the class of all Poisson distributions). An explicit description would not be possible. However, the total variation projection can be numerically computed.

For instance, if $n = 50$, $p = .01$, then the total variation projection is the Poisson distribution with mean .5025. If $n = 100$, $p = .05$, then the total variation projection is the Poisson distribution with mean 5.015. The best Poisson approximation seems to have a mean slightly off from np. In fact, if the total variation projection has mean λ_n, then $|\lambda_n - \lambda| \to 0$. We will come back to Poisson approximations to binomials later in this chapter.

2.3 Information-Theoretic Distances, de Bruijn's Identity, and Relations to Convergence

Entropy and Fisher information are two principal information-theoretic quantities. Statisticians, by means of well-known connections to inference such as the Cramér-Rao inequality and maximum likelihood estimates, are very familiar with the Fisher information. Probabilists, on the other hand, are very familiar with entropy. We first define them formally.

Definition 2.2 Let f be a density in $\mathcal{R}^d$. The entropy of f, or synonymously of a random variable $X \sim f$, is $H(X) = -\int f(x) \log f(x) dx = -E_f[\log f(X)]$.

For integer-valued variables, the definition is similar.

Definition 2.3 Let X be integer valued with $P(X = j) = p_j$. Then, the entropy of X is $H(X) = -\sum_j p(j) \log p(j)$.

Fisher information is defined only for smooth densities. Here is the definition.

Definition 2.4 Let f be a density in $\mathcal{R}^d$. Suppose f has one partial derivative with respect to each coordinate everywhere in its support $\{x : f(x) > 0\}$. The Fisher information of f, or synonymously of a random variable $X \sim f$, is $I(X) = \int_{x: f(x)>0} \frac{||\nabla f(x)||^2}{f(x)} dx = E_f[|| \nabla \log f(X)||^2]$, where $\nabla(.)$ denotes the gradient vector.

Remark. The function $\nabla \log f(x)$ is called *the score function* of f.

Entropy and Fisher information each satisfy certain suitable subadditivity properties. We record their most basic properties below. Johnson (2004) can be consulted for proofs of the theorems in this section apart from the specific references given for particular theorems below.

Theorem 2.5 (a) For jointly distributed random variables X, Y, $H(X, Y) \leq H(X) + H(Y)$ with equality iff X, Y are independent:

(b) For any $\sigma > 0$, $H(\mu + \sigma X) = \log \sigma + H(X)$.

(c) For independent random variables X, Y, $H(X+Y) \geq \max\{H(X), H(Y)\}$.

(d) For jointly distributed random variables X, Y, $I(X, Y) \geq \max\{I(X), I(Y)\}$.

(e) For any σ, $I(\mu + \sigma X) = \frac{I(X)}{\sigma^2}$.

(f) For independent random variables X, Y, $I(X + Y) \leq \alpha^2 I(X) + (1-\alpha)^2 I(Y) \forall 0 \leq \alpha \leq 1$ with equality iff X, Y are each normal.

(g) For independent random variables X, Y, $I(X + Y) \leq \left(\frac{1}{I(X)} + \frac{1}{I(Y)}\right)^{-1}$ with equality iff X, Y are each normal.

Example 2.6 For some common distributions, we give expressions for the entropy and Fisher information when available.

Distribution	$H(X)$	$I(X)$
Exponential(1)	1	1
$N(0, 1)$	$\frac{1}{2}\log(2\pi) + \frac{1}{2}$	1
Gamma$(\alpha, 1)$	$\alpha + \log\Gamma(\alpha) + (\alpha - 1)\psi(\alpha)$	$\frac{1}{\alpha-2}(\alpha > 2)$
$C(0, 1)$	—	$\frac{1}{2}$
$N_d(0, \Sigma)$	$\frac{d}{2}\log(2\pi) + \log\lvert\Sigma\rvert + \frac{d}{2}$	$\mathrm{tr}\Sigma^{-1}$

Remark. In the table above, ψ is the di-Gamma function (i.e., the derivative of $\log\Gamma$).

Entropy and Fisher information, interestingly, are connected to each other. They are connected by a link to the normal distribution and also through an

algebraic relation known as *de Bruijn's identity*. We mention the link through the normal distribution first.

Theorem 2.6 Among all densities with mean 0 and variance $\sigma^2 < \infty$, the entropy is maximized by the $N(0, \sigma^2)$ density. On the other hand, among all densities with mean 0 and variance $\sigma^2 < \infty$ such that the Fisher information is defined, Fisher information is minimized by the $N(0, \sigma^2)$ density.

Remark. The theorem says that normal distributions are extremals in two optimization problems with a variance constraint, namely the maximum entropy and the minimum Fisher information problems. Actually, although we state the theorem for $N(0, \sigma^2)$, the mean is irrelevant. This theorem establishes an indirect connection between H and I inherited from a connection of each to normal distributions.

We can use H and I to define distances between two different distributions. These are defined as follows.

Definition 2.5 Let $X \sim f, Y \sim g$, and assume that $g(x) = 0 \Rightarrow f(x) = 0$. The *entropy divergence or differential entropy between* f *and* g is defined as

$$D(f||g) = \int f(x) \log\left(\frac{f(x)}{g(x)}\right).$$

The *Fisher information distance between* f *and* g is defined as

$$I(f||g) = I(X||Y) = \int [|| \bigtriangledown \log f - \bigtriangledown \log(g)||^2] f(x) dx.$$

Using the normal distribution as a benchmark, we can define a standardized Fisher information as follows.

Definition 2.6 Let $X \sim f$ have finite variance σ^2. The *standardized Fisher information of* f is defined as $I_s(f) = I_s(X) = \sigma^2 I(f||N(0, \sigma^2))$.

The advantage of the standardization is that $I_s(f)$ can be zero only when f itself is a normal density. Similarly, the entropy divergence of a density f with a normal density can be zero only if f is that same normal density.

We state the elegant algebraic connection between entropy divergence and standardized Fisher information next.

Theorem 2.7 (**De Bruijn's Identity**) Let $X \sim f$ have variance 1. Let Z be a standard normal variable independent of X. For $t > 0$, let f_t denote the density of $X + \sqrt{t}Z$. Then, $I(f_t) = 2\frac{d}{dt}[H(f_t)]$.

Remark. De Bruijn's identity (which extends to higher dimensions) is a consequence of the heat equation of partial differential equations; see Johnson (2004). A large number of such $\frac{d}{dt}$ identities of use in statistics (although not de Bruijn's identity itself) are proved in Brown et al. (2006). That such a neat algebraic identity links entropy with Fisher information is a pleasant surprise.

We now describe how convergence in entropy divergence is a very strong form of convergence.

Theorem 2.8 Let f_n, f be densities in $\mathcal{R}^d$. Suppose $D(f_n||f) \to 0$. Then f_n converges to f in total variation; in particular, convergence in distribution follows.

This theorem has completely general densities f_n, f. In statistics, often one is interested in densities of normalized convolutions. Calculating their entropies or entropy distances from the density of the limiting $N(0, 1)$ distribution could be hard because convolution densities are difficult to write. In a remarkable result, Barron (1986) proved the following.

Theorem 2.9 Let $X_1, X_2, \cdots$ be iid zero-mean, unit-variance random variables and let f_n denote the density (assuming it exists) of $\sqrt{n}\bar{X}$. If, for some m, $D(f_m||N(0, 1)) < \infty$, then $D(f_n||N(0, 1)) \to 0$.

Analogously, one can use Fisher information in order to establish weak convergence. The intuition is that if the Fisher information of $\sqrt{n}\bar{X}$ is converging to 1, which is the Fisher information of the $N(0, 1)$ distribution, then by virtue of the unique Fisher information minimizing property of the $N(0, 1)$ subject to a fixed variance of 1 (stated above), it ought to be the case that $\sqrt{n}\bar{X}$ is converging to $N(0, 1)$ in distribution. The intuition is pushed to a proof in Brown (1982), as stated below.

Theorem 2.10 Let $X_1, X_2, \cdots$ be iid zero-mean, unit-variance random variables and $Z_1, Z_2, \cdots$ be an iid $N(0, 1)$ sequence independent of the $\{X_i\}$. Let $v > 0$ and $Y_n(v) = \sqrt{n}\bar{X} + \sqrt{v}Z_n$. Then, for any v, $I_s(Y_n(v)) \to 0$ and hence $\sqrt{n}\bar{X} \stackrel{\mathcal{L}}{\Rightarrow} N(0, 1)$.

Remark. It had been suspected for a long time that there should be such a proof of the central limit theorem by using Fisher information. It was later found that Brown's technique was so powerful that it extended to central limit theorems for many kinds of non-iid variables. These results amounted to a triumph of information theory tools and provided much more intuitive proofs of the central limit results than proofs based on Fourier methods.

An interesting question to ask is what can be said about the rates of convergence of the entropy divergence and the standardized Fisher information

in the canonical CLT situation (i.e., for $\sqrt{n}\bar{X}$ when X_i are iid with mean 0 and variance 1). This is a difficult question. In general, one can hope for convergence at the rate of $\frac{1}{n}$. The following is true.

Theorem 2.11 Let $X_1, X_2, \cdots$ be iid zero-mean, unit-variance random variables. Then, each of $D(\sqrt{n}\bar{X}||N(0,1))$ and $I_s(\sqrt{n}\bar{X})$ is $O(\frac{1}{n})$.

Remark. This is quite a bit weaker than the best results that are now known. In fact, one can get bounds valid for *all* n, although they involve constants that usually cannot be computed. Johnson and Barron (2003) may be consulted to see the details.

2.4 Poisson Approximations

Exercise 1.5 in Chapter 1 asks to show that the sequence of $\text{Bin}(n, \frac{1}{n})$ distributions converges in law to the Poisson distribution with mean 1. The $\text{Bin}(n, \frac{1}{n})$ is a sum of n independent Bernoullis but with a success probability that is small and also depends on n. The $\text{Bin}(n, \frac{1}{n})$ is a count of the total number of occurrences among n independent rare events. It turns out that convergence to a Poisson distribution can occur even if the individual success probabilities are small but not the same, and even if the Bernoulli variables are not independent. Indeed, approximations by Poisson distributions are extremely useful and accurate in many problems. The problems arise in diverse areas. Poisson approximation is a huge area, with an enormous body of literature, and there are many book-length treatments. We provide here a glimpse into the area with some examples.

Definition 2.7 Let p, q be two mass functions on the integers. The total variation distance between p and q is defined as $d_{\text{TV}}(p,q) = \sup_{A \subseteq \mathcal{Z}} |P_p(X \in A) - P_q(X \in A)|$, which equals $\frac{1}{2}\sum_j |p(j) - q(j)|$.

A simple and classic result is the following.

Theorem 2.12 (LeCam (1960)) (a) $d_{\text{TV}}(\text{Bin}(n, \frac{\lambda}{n}), \text{Poi}(\lambda)) \le \frac{8\lambda}{n}$. (b) For $n \ge 1$, let $\{X_{in}\}_{i=1}^n$ be a triangular array of independent $\text{Ber}(p_{in})$ variables. Let $S_n = \sum_{i=1}^n X_{in}$ and $\lambda_n = \sum_{i=1}^n p_{in}$. Then, $d_{\text{TV}}(S_n, \text{Poi}(\lambda_n)) \le \frac{8}{\lambda_n}\sum_{i=1}^n p_{in}^2$, if $\max\{p_{in}, 1 \le i \le n\} \le \frac{1}{4}$.

A neat corollary of LeCam's theorem is the following.

Corollary 2.2 If X_{in} is a triangular array of independent $\text{Ber}(p_{in})$ variables such that $\max\{p_{in}, 1 \le i \le n\} \to 0$, and $\lambda_n = \sum_{i=1}^n p_{in} \to \lambda, 0 < \lambda < \infty$, then $d_{\text{TV}}(S_n, \text{Poi}(\lambda)) \to 0$ and hence $S_n \stackrel{\mathcal{L}}{\Rightarrow} \text{Poi}(\lambda)$.

The Poisson distribution has the property of having equal mean and variance, so intuition would suggest that if a sum of independent Bernoulli variables had, asymptotically, an equal mean and variance, then it should converge to a Poisson distribution. That, too, is true.

Corollary 2.3 If X_{in} is a triangular array of independent Ber(p_{in}) variables such that $\sum_{i=1}^{n} p_{in}$ and $\sum_{i=1}^{n} p_{in}(1 - p_{in})$ each converge to λ, $0 < \lambda < \infty$, then $S_n \stackrel{\mathcal{L}}{\Rightarrow} \text{Poi}(\lambda)$.

It is a fact that, in many applications, although the variable can be represented as a sum of Bernoulli variables, they are not independent. The question arises if a Poisson limit can still be proved. The question is rather old. Techniques that we call *first-generation techniques*, using combinatorial methods, are successful in some interesting problems. These methods typically use generating functions or sharp Bonferroni inequalities. Two very good references for looking at those techniques are Kolchin, Sevas'tyanov, and Chistyakov (1978) and Galambos and Simonelli (1996). Here is perhaps the most basic result of that type.

Theorem 2.13 For $N \geq 1$, let $X_{in}, i = 1, 2, \cdots, n = n(N)$ be a triangular array of Bernoulli random variables, and let $A_i = A_{in}$ denote the event where $X_{in} = 1$. For a given k, let $M_k = M_{kn}$ be the kth binomial moment of S_n; i.e., $M_k = \sum_{j=k}^{n} \binom{j}{k} P(S_n = j)$. If there exists $0 < \lambda < \infty$ such that, for every fixed k, $M_k \to \frac{\lambda^k}{k!}$ as $N \to \infty$, then $S_n \stackrel{\mathcal{L}}{\Rightarrow} \text{Poi}(\lambda)$.

Remark. In some problems, typically of a combinatorial nature, careful counting lets one apply this basic theorem and establish convergence to a Poisson distribution.

Example 2.7 (**The Matching Problem**) Cards are drawn one at a time from a well-shuffled deck containing N cards, and a match occurs if the card bearing a number, say j, is drawn at precisely the jth draw from the deck. Let S_N be the total number of matches. Theorem 2.13 can be used in this example. The binomial moment M_k can be shown to be $M_k = \binom{N}{k} \frac{1}{N(N-1)\ldots(N-k+1)}$, and from here, by Stirling's approximation, for every fixed k, $M_k \to \frac{1}{k!}$, establishing that the total number of matches converges to a Poisson distribution with mean 1 as the deck size $N \to \infty$. Note that the mean value of S_N is exactly 1 for any N. Convergence to a Poisson distribution is extremely fast in this problem; even for $N = 5$, the Poisson approximation is quite good. For $N = 10$, it is almost exact!

For information, we note the following superexponential bound on the error of the Poisson approximation in this problem; this is proved in DasGupta (1999).

Theorem 2.14 $d_{\mathrm{TV}}(S_N, \mathrm{Poi}(1)) \le \frac{2^N}{(N+1)!}\ \forall N$.

Example 2.8 (**The Committee Problem**) From n people, $N = N(n)$ committees are formed, each committee of a fixed size m. We let $N, n \to \infty$, holding m fixed. The Bernoulli variable X_{in} is the indicator of the event that the ith person is not included in any committee. Under the usual assumptions of independence and also the assumption of random selection, the binomial moment M_k can be shown to be $M_k = \binom{n}{k}[\frac{\binom{n-k}{m}}{\binom{n}{m}}]^N$.

Stirling's approximation shows that $M_k \sim \frac{n^k}{k!}e^{-kN(\frac{m}{n}+O(n^{-2}))}$ as $n \to \infty$. One now sees on inspection that if N, n are related as $N = \frac{n\log n}{m} - n\log\lambda + o(n^{-1})$ for some $0 < \lambda < \infty$, then $M_k \to \frac{\lambda^{km}}{k!}$ and so, from the basic convergence theorem above, the number of people who are left out of *all* committees converges to $\mathrm{Poi}(\lambda^m)$.

Example 2.9 (**The Birthday Problem**) This is one of the most colorful examples in probability theory. Suppose each person in a group of n people has, mutually independently, a probability $\frac{1}{N}$ of being born on any given day of a year with N calendar days. Let S_n be the total number of pairs of people (i, j) such that they have the same birthday. $P(S_n > 0)$ is the probability that there is at least one pair of people in the group who share the same birthday. It turns out that if n, N are related as $n^2 = 2N\lambda + o(N)$, for some $0 < \lambda < \infty$, then $S_n \stackrel{\mathcal{L}}{\Rightarrow} \mathrm{Poi}(\lambda)$. For example, if $N = 365, n = 30$, then S_n is roughly Poisson with mean 1.233.

A review of the birthday and matching problems is given in DasGupta (2005). Many of the references given at the beginning of this chapter also discuss Poisson approximation in these problems.

We earlier described the binomial moment method as a first-generation method for establishing Poisson convergence. The modern method, which has been fantastically successful in hard problems, is known as the *Stein-Chen method*. It has a very interesting history. In 1972, Stein gave a novel method of obtaining error bounds in the central limit theorem. Stein (1972) gave a technique that allowed him to have dependent summands and also allowed him to use non-Fourier methods, which are the classical methods in that problem. We go into those results, generally called Berry-Esseen bounds, later in the book (see Chapter 11). Stein's method was based on a very simple identity, now universally known as Stein's iden-

tity (published later in Stein (1981)), which says that if $Z \sim N(0,1)$, then for *nice* functions f, $E[Zf(Z)] = E[f'(Z)]$. It was later found that if Stein's identity holds for *many* nice functions, then the underlying variable Z *must* be $N(0,1)$. So, the intuition is that if for some random variable $Z = Z_n$, $E[Zf(Z) - f'(Z)] \approx 0$, then Z should be close to $N(0,1)$ in distribution. In a manner that many still find mysterious, Stein reduced this to a comparison of the mean of a suitable function h, related to f by a differential equation, under the true distribution of Z and the $N(0,1)$ distribution. From here, he was able to obtain non-Fourier bounds on errors in the CLT for dependent random variables. A Stein type identity was later found for the Poisson case in the decision theory literature; see Hwang (1982). Stein's method for the normal case was successfully adapted to the Poisson case in Chen (1975). The Stein-Chen method is now regarded as the principal tool in establishing Poisson limits for sums of dependent Bernoulli variables. Roughly speaking, the dependence should be weak, and for any single Bernoulli variable, the number of other Bernoulli variables with which it shares a dependence relation should not be very large. The Stein-Chen method has undergone a lot of evolution with increasing sophistication since Chen (1975). The references given in the first section of this chapter contain a wealth of techniques, results, and, most of all, numerous new applications. Specifically, we recommend Arratia, Goldstein, and Gordon (1990), Barbour, Holst and Janson (1992), Dembo and Rinott (1996), and the recent monograph by Diaconis and Holmes (2004). See Barbour, Chen, and Loh (1992) for use of the Stein-Chen technique for compound Poisson approximations.

2.5 Exercises

Exercise 2.1 * Let $X \sim F$ with density $\frac{1}{\pi(1+x^2)}$, $-\infty < x < \infty$. Find the total variation projection of F onto the family of all normal distributions.

Exercise 2.2 For each of the following cases, evaluate the indicated distances.

(i) $d_{\mathrm{TV}}(P, Q)$ when $P = \mathrm{Bin}(20, .05)$ and $Q = \mathrm{Poisson}(1)$.
(ii) $d_K(F, G)$ when $F = N(0, \sigma^2)$ and $G = \mathrm{Cauchy}(0, \tau^2)$.
(iii) $H(P, Q)$ when $P = N(\mu, \sigma^2)$ and $Q = N(\nu, \tau^2)$.

Exercise 2.3 * Write an expansion in powers of ϵ for $d_{\mathrm{TV}}(P, Q)$ when $P = N(0,1)$ and $Q = N(\epsilon, 1)$.

Exercise 2.4 Calculate and plot (as a function of μ) $H(P, Q)$ and $d_{\mathrm{TV}}(P, Q)$ when $P = N(0, 1)$ and $Q = N(\mu, 1)$.

Exercise 2.5 * Suppose $P_n = \mathrm{Bin}(n, p_n)$ and $P = \mathrm{Poi}(\lambda)$. Give a sufficient condition for $d_{\mathrm{TV}}(P_n, P) \to 0$. Can you give a nontrivial necessary condition?

Exercise 2.6 Show that if $X \sim P$, $Y \sim Q$, then $d_{\mathrm{TV}}(P, Q) \leq P(X \neq Y)$.

Exercise 2.7 Suppose $X_i \overset{\text{indep.}}{\sim} P_i$, $Y_i \overset{\text{indep.}}{\sim} Q_i$. Then $d_{\mathrm{TV}}(P_1 * P_2 * \cdots * P_n, Q_1 * Q_2 * \cdots * Q_n) \leq \sum_{i=1}^{n} d_{\mathrm{TV}}(P_i, Q_i)$, where $*$ denotes convolution.

Exercise 2.8 Suppose X_n is a Poisson variable with mean $\frac{n}{n+1}$ and X is Poisson with mean 1.

(a) Show that the total variation distance between the distributions of X_n and X converges to zero.

(b) * (Harder) Find the rate of convergence to zero in part (a).

Exercise 2.9 * Let $P = N(0, 1)$ and $Q = N(\mu, \sigma^2)$. Plot the set $S = \{(\mu, \sigma) : d_{\mathrm{TV}}(P, Q) \leq \epsilon\}$ for some selected values of ϵ.

Exercise 2.10 Suppose $X_1, X_2, \ldots$ are iid Exp(1). Does $\sqrt{n}(\bar{X} - 1)$ converge to standard normal in total variation?

Exercise 2.11 If X_i are iid, show that $\bar{X}_n \xrightarrow{\mathcal{P}} 0$ iff $E\left(\frac{\bar{X}_n^2}{1+X_n^2}\right) \to 0$.

Exercise 2.12 * Let $X \sim U[-1, 1]$. Find the total variation projection of X onto the class of all normal distributions.

Exercise 2.13 * Consider the family of densities with mean equal to a specified μ. Find the density in this family that maximizes the entropy.

Exercise 2.14 * (**Projection in Entropy Distance**) Suppose X has a density with mean μ and variance σ^2. Show that the projection of X onto the class of all normal distributions has the same mean and variance as X.

Exercise 2.15 * (**Projection in Entropy Distance Continued**) Suppose X is an integer-valued random variable with mean μ. Show that the projection of X onto the class of all Poisson distributions has the same mean as X.

Exercise 2.16 * First write the exact formula for the entropy of a Poisson distribution, and then prove that the entropy grows at the rate of $\log \lambda$ as the mean $\lambda \to \infty$.

Exercise 2.17 What can you say about the existence of entropy and Fisher information for Beta densities? What about the double exponential density?

Exercise 2.18 Prove that the standardized Fisher information of a Gamma$(\alpha, 1)$ density converges to zero at the rate $\frac{1}{\alpha}$, α being the shape parameter.

Exercise 2.19 * Consider the Le Cam bound $d_{\mathrm{TV}}(\mathrm{Bin}(n, p), \mathrm{Poi}(np)) \leq 8p$. Compute the ratio $\frac{d_{\mathrm{TV}}(\mathrm{Bin}(n,p),\mathrm{Poi}(np))}{p}$ for a grid of (n, p) pairs and investigate the best constant in Le Cam's inequality.

Exercise 2.20 * For $N = 5, 10, 20, 30$, compute the distribution of the total number of matches in the matching problem, and verify that the distribution in each case is unimodal.

Exercise 2.21 Give an example of a sequence of binomial distributions that converge neither to a normal (on centering and norming) nor to a Poisson distribution.

References

Arratia, R., Goldstein, L., and Gordon, L. (1990). Poisson approximation and the Chen-Stein method, Stat. Sci., 5(4), 403–434.

Barbour, A., Chen, L., and Loh, W-L. (1992). Compound Poisson approximation for non-negative random variables via Stein's method, Ann. Prob., 20, 1843–1866.

Barbour, A., Holst, L., and Janson, S. (1992). *Poisson Approximation*, Clarendon Press, New York.

Barron, A. (1986). Entropy and the central limit theorem, Ann. Prob., 14(1), 336–342.

Boos, D. (1985). A converse to Scheffe's theorem, Ann. Stat., 1, 423–427.

Brown, L. (1982). A proof of the central limit theorem motivated by the Cramér-Rao inequality, G. Kallianpur, P.R. Krishnaiah, and J.K. Ghosh *Statistics and Probability, Essays in Honor of C.R. Rao*, North-Holland, Amsterdam, 141–148.

Brown, L., DasGupta, A., Haff, L.R., and Strawderman, W.E. (2006). The heat equation and Stein's identity: Connections, applications, J. Stat. Planning Infer, Special Issue in Memory of Shanti Gupta, 136, 2254–2278.

Chen, L.H.Y. (1975). Poisson approximation for dependent trials, Ann. Prob., 3, 534–545.

DasGupta, A. (1999). The matching problem and the Poisson approximation, Technical Report, Purdue University.

DasGupta, A. (2005). The matching, birthday, and the strong birthday problems: A contemporary review, J. Stat. Planning Infer, Special Issue in Honor of Herman Chernoff, 130, 377–389.

Dembo, A. and Rinott, Y. (1996). Some examples of normal approximations by Stein's method, in *Random Discrete Structures*, D. Aldous and Pemantle R. IMA Volumes in Mathematics and Its Applications, Vol. 76, Springer, New York, 25–44.

Diaconis, P. and Holmes, S. (2004). *Stein's Method: Expository Lectures and Applications*, IMS Lecture Notes Monograph Series, vol. 46, Institute of Mathematical Statistics, Beachwood, OH.

Dudley, R. (1989). *Real Analysis and Probability*, Wadsworth, Pacific Grove, CA.

Galambos, J. and Simonelli, I. (1996). *Bonferroni-Type Inequalities with Applications*, Springer, New York.

Hwang, J.T. (1982). Improving upon standard estimators in discrete exponential families with applications to Poisson and negative binomial cases, Ann. Stat., 10(3), 857–867.

Johnson, O. (2004). *Information Theory and the Central Limit Theorem*, Imperial College Press, Yale University London.

Johnson, O. and Barron, A. (2003). Fisher information inequalities and the central limit theorem, Technical Report.

Kolchin, V., Sevas'tyanov, B., and Chistyakov, V. (1978). *Random Allocations*, V.H. Winston & Sons, Washington, distributed by Halsted Press, New York.

LeCam, L. (1960). An approximation theorem for the Poisson binomial distribution, Pac. J. Math., 10, 1181–1197.

Linnik, Y. (1959). An information theoretic proof of the central limit theorem, Theory Prob. Appl., 4, 288–299.

Rachev, S. (1991). *Probability Metrics and the Stability of Stochastic Models*, John Wiley, Chichester.

Reiss, R. (1989). *Approximate Distributions of Order Statistics, with Applications to Nonparametric Statistics*, Springer-Verlag, New York.

Sevas'tyanov, B.A. (1972). A limiting Poisson law in a scheme of sums of dependent random variables, Teor. Veroyatni. Primen., 17, 733–738.

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables, L. Le Cam, J. Neyman, and E. Scott in *Proceedings of the Sixth Berkeley Symposium*, Vol. 2, University of California Press, Berkeley, 583–602.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution, Ann. Stat., 9, 1135–1151.

Stein, C. (1986). *Approximate Computations of Expectations*, Institute of Mathematical Statistics, Hayward, CA.

Sweeting, T. (1986). On a converse to Scheffe's theorem, Ann. Stat., 3, 1252–1256.

Chapter 29
The Bootstrap

The bootstrap is a resampling mechanism designed to provide information about the sampling distribution of a functional $T(X_1, X_2, ..., X_n, F)$, where $X_1, X_2, ..., X_n$ are sample observations and F is the CDF from which $X_1, X_2, ..., X_n$ are independent observations. The bootstrap is not limited to the iid situation. It has been studied for various kinds of dependent data and complex situations. In fact, this versatile nature of the bootstrap is the principal reason for its popularity. There are numerous texts and reviews of bootstrap theory and methodology at various technical levels. We recommend Efron and Tibshirani (1993) and Davison and Hinkley (1997) for applications-oriented broad expositions and Hall (1992) and Shao and Tu (1995) for detailed theoretical development. Modern reviews include Hall (2003), Beran (2003), Bickel (2003), and Efron (2003). Bose and Politis (1992) is a well-written nontechnical account, and Lahiri (2003) is a rigorous treatment of the bootstrap for various kinds of dependent data.

Suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F$ and $T(X_1, X_2, ..., X_n, F)$ is a functional; e.g., $T(X_1, X_2, ..., X_n, F) = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$, where $\mu = E_F(X_1)$ and $\sigma^2 = \text{Var}_F(X_1)$. In statistical problems, we frequently need to know something about the sampling distribution of T; e.g., $P_F(T(X_1, X_2, ..., X_n, F) \le t)$. If we had replicated samples from the population, resulting in a series of values for the statistic T, then we could form estimates of $P_F(T \le t)$ by counting how many of the T_i's are $\le t$. But statistical sampling is not done that way. We do not usually obtain replicated samples; we obtain just one set of data of some size n. However, let us think for a moment of a finite population. A large sample from a finite population should be well representative of the full population itself, so replicated samples (with replacement) from the original sample, which would just be an iid sample from the empirical CDF F_n, could be regarded as proxies for replicated samples from the population itself, provided n is large. Suppose that for some number B we draw B resamples of size n from the original sample. Denoting the resamples from the original

A. DasGupta, *Asymptotic Theory of Statistics and Probability*,

sample as $(X_{11}^*, X_{12}^*, ..., X_{1n}^*), (X_{21}^*, X_{22}^*, ..., X_{2n}^*), ..., (X_{B1}^*, X_{B2}^*, ..., X_{Bn}^*)$, with corresponding values $T_1^*, T_2^*, ..., T_B^*$ for the functional T, one can use simple frequency-based estimates such as $\frac{\#\{j:T_j^* \le t\}}{B}$ to estimate $P_F(T \le t)$. This is the basic idea of the bootstrap. Over time, the bootstrap has found its use in estimating other quantities, e.g., $\mathrm{Var}_F(T)$ or quantiles of T. The bootstrap is thus an omnibus mechanism for approximating sampling distributions or functionals of sampling distributions of statistics. Since frequentist inference is mostly about sampling distributions of suitable statistics, the bootstrap is viewed as an immensely useful and versatile tool, further popularized by its automatic nature. However, it is also frequently used in situations where it should not be used. In this chapter, we give a broad methodological introduction to various types of bootstraps, explain their theoretical underpinnings, discuss their successes and limitations, and try them out in some trial cases.

29.1 Bootstrap Distribution and the Meaning of Consistency

The formal definition of the bootstrap distribution of a functional is the following.

Definition 29.1 Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F$ and $T(X_1, X_2, ..., X_n, F)$ be a given functional. The *ordinary bootstrap distribution* of T is defined as

$$H_{\text{Boot}}(x) = P_{F_n}(T(X_1^*, ..., X_n^*, F_n) \le x),$$

where $(X_1^*, ..., X_n^*)$ is an iid sample of size n from the empirical CDF F_n.

It is common to use the notation P_* to denote probabilities under the bootstrap distribution.

Remark. $P_{F_n}(\cdot)$ corresponds to probability statements corresponding to all the n^n possible resamples with replacement from the original sample $(X_1, \ldots, X_n)$. Since recalculating T from all n^n resamples is basically impossible unless n is very small, one uses a smaller number of B resamples and recalculates T only B times. Thus $H_{\text{Boot}}(x)$ itself is estimated by a Monte Carlo, known as the *bootstrap Monte Carlo*, so the final estimate for $P_F(T(X_1, X_2, ..., X_n, F_n) \le x)$ absorbs errors from two sources: (i) pretending $(X_{i1}^*, X_{i2}^*, ..., X_{in}^*)$ to be bona fide resamples from F; (ii) estimating the true $H_{\text{Boot}}(x)$ by a Monte Carlo. By choosing B adequately large, the Monte Carlo error is generally ignored. The choice of B that would let one ignore the Monte Carlo error is a hard mathematical problem; Hall (1986, 1989a) are two key references. It is customary to choose $B \approx 300$ for variance

estimation and a somewhat larger value for estimating quantiles. It is hard to give any general reliable prescriptions on B.

It is important to note that the resampled data need not necessarily be obtained from the empirical CDF F_n. Indeed, it is a natural question whether resampling from a smoothed nonparametric distribution estimator can result in better performance. Examples of such smoothed distribution estimators are integrated kernel density estimates. It turns out that, in some problems, smoothing does lead to greater accuracy, typically in the second order. See Silverman and Young (1987) and Hall, DiCiccio, and Romano (1989) for practical questions and theoretical analysis of the benefits of using a smoothed bootstrap. Meanwhile, bootstrapping from F_n is often called the *naive or orthodox bootstrap*, and we will sometimes use this terminology.

Remark. At first glance, the idea appears to be a bit too simple to actually work. But one has to have a definition for what one means by the bootstrap working in a given situation. It depends on what one wants the bootstrap to do. For estimating the CDF of a statistic, one should want $H_{\text{Boot}}(x)$ to be numerically close to the true CDF $H_n(x)$ of T. This would require consideration of metrics on CDFs. For a general metric ρ, the definition of "the bootstrap working" is the following.

Definition 29.2 Let F and G be two CDFs on a sample space $\mathcal{X}$. Let $\rho(F, G)$ be a metric on the space of CDFs on $\mathcal{X}$. For $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F$, and a given functional $T(X_1, X_2, ..., X_n, F)$, let

$$H_n(x) = P_F(T(X_1, X_2, ..., X_n, F) \le x),$$
$$H_{\text{Boot}}(x) = P_*(T(X_1^*, X_2^*, ..., X_n^*, F_n) \le x).$$

We say that the bootstrap is weakly consistent under ρ for T if $\rho(H_n, H_{\text{Boot}}) \overset{P}{\Rightarrow} 0$ as $n \to \infty$. We say that the bootstrap is strongly consistent under ρ for T if $\rho(H_n, H_{\text{Boot}}) \overset{\text{a.s.}}{\Rightarrow} 0$.

Remark. Note that the need for mentioning convergence to zero in probability or a.s. in this definition is due to the fact that the bootstrap distribution H_{Boot} is a random CDF. That H_{Boot} is a random CDF has nothing to do with bootstrap Monte Carlo; it is a random CDF because as a function it depends on the original sample $(X_1, X_2, ..., X_n)$. Thus, the bootstrap uses a random CDF to approximate a deterministic but unknown CDF, namely the true CDF H_n of the functional T.

Example 29.1 How does one apply the bootstrap in practice? Suppose, for example, $T(X_1, \ldots, X_n, F) = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$. In the orthodox bootstrap scheme,

we take iid samples from F_n. The mean and the variance of the empirical distribution F_n are $\bar{X}$ and $s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ (note the n rather than $n-1$ in the denominator). The bootstrap is a device for estimating $P_F(\frac{\sqrt{n}(\bar{X}-\mu(F))}{\sigma} \le x)$ by $P_{F_n}(\frac{\sqrt{n}(\bar{X}_n^*-\bar{X})}{s} \le x)$. We will further approximate $P_{F_n}(\frac{\sqrt{n}(\bar{X}_n^*-\bar{X})}{s} \le x)$ by resampling only B times from the original sample set $\{X_1, \ldots, X_n\}$. In other words, finally we will report as our estimate for $P_F(\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \le x)$ the number $\#\{j : \frac{\sqrt{n}(\bar{X}_{n,j}^*-\bar{X})}{s} \le x\}/B$.

29.2 Consistency in the Kolmogorov and Wasserstein Metrics

We start with the case of the sample mean of iid random variables. If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$ and if $\text{Var}_F(X_i) < \infty$, then $\sqrt{n}(\bar{X} - \mu)$ has a limiting normal distribution by the CLT. So a probability such as $P_F(\sqrt{n}(\bar{X}-\mu) \le x)$ could be approximated for example by $\Phi(\frac{x}{s})$, where s is the sample standard deviation. An interesting property of the bootstrap approximation is that, even when the CLT approximation $\Phi(\frac{x}{s})$ is available, the bootstrap approximation may be more accurate. We will later describe theoretical results in this regard. But first we present two consistency results corresponding to the following two specific metrics that have earned a special status in this literature:

(i) Kolmogorov metric

$$K(F, G) = \sup_{-\infty<x<\infty} |F(x) - G(x)|;$$

(ii) Mallows-Wasserstein metric

$$\ell_2(F, G) = \inf_{\Gamma_{2,F,G}} (E|Y - X|^2)^{\frac{1}{2}},$$

where $X \sim F$, $Y \sim G$, and $\Gamma_{2,F,G}$ is the class of all joint distributions of (X, Y) with marginals F and G, each with a finite second moment.

ℓ_2 is a special case of the more general metric

$$\ell_p(F, G) = \inf_{\Gamma_{p,F,G}} (E|Y - X|^p)^{\frac{1}{p}},$$

with the infimum being taken over the class of joint distributions with marginals as F, G, and the pth moment of F, G being finite.

Of these, the Kolmogorov metric is universally regarded as a natural one. But how about ℓ_2? ℓ_2 is a natural metric for many statistical problems because of its interesting property that $\ell_2(F_n, F) \to 0$ iff $F_n \stackrel{\mathcal{L}}{\Rightarrow} F$ and $E_{F_n}(X^i) \to E_F(X^i)$ for $i = 1, 2$. Since one might want to use the bootstrap primarily for estimating the CDF, mean, and variance of a statistic, consistency in ℓ_2 is just the right result for that purpose.

Theorem 29.1 Suppose $X_1, X_2, \ldots, X_n \stackrel{\text{iid}}{\sim} F$ and that $E_F(X_1^2) < \infty$. Let $T(X_1, \ldots, X_n, F) = \sqrt{n}(\bar{X} - \mu)$. Then $K(H_n, H_{\text{Boot}})$ and $\ell_2(H_n, H_{\text{Boot}}) \stackrel{\text{a.s.}}{\longrightarrow} 0$ as $n \to \infty$.

Remark. Strong consistency in K is proved in Singh (1981), and that for ℓ_2 is proved in Bickel and Freedman (1981). Notice that $E_F(X_1^2) < \infty$ guarantees that $\sqrt{n}(\bar{X} - \mu)$ admits a CLT. And Theorem 29.1 says that the bootstrap is strongly consistent (w.r.t. K and ℓ_2) under that assumption. This is in fact a very good rule of thumb: if a functional $T(X_1, X_2, ..., X_n, F)$ admits a CLT, then the bootstrap would be at least weakly consistent for T. Strong consistency might require a little more assumption.

We sketch a proof of the strong consistency in K. The proof requires use of the Berry-Esseen inequality, Polya's theorem (see Chapter 1 or Chapter 2), and a strong law known as the Zygmund-Marcinkiewicz strong law, which we state below.

Lemma 29.1 (**Zygmund-Marcinkiewicz SLLN**) Let $Y_1, Y_2, \ldots$ be iid random variables with CDF F and suppose, for some $0 < \delta < 1$, $E_F|Y_1|^\delta < \infty$. Then $n^{-1/\delta} \sum_{i=1}^n Y_i \stackrel{\text{a.s.}}{\Rightarrow} 0$.

We are now ready to sketch the proof of strong consistency of H_{Boot} under K. Using the definition of K, we can write $K(H_n, H_{\text{Boot}}) = \sup_x \left| P_F \{T_n \le x\} - P_* \left\{T_n^* \le x\right\} \right|$

$$
\begin{aligned}
&= \sup_x \left| P_F \left\{ \frac{T_n}{\sigma} \le \frac{x}{\sigma} \right\} - P_* \left\{ \frac{T_n^*}{s} \le \frac{x}{s} \right\} \right| \\
&= \sup_x \left| P_F \left\{ \frac{T_n}{\sigma} \le \frac{x}{\sigma} \right\} - \Phi\left(\frac{x}{\sigma}\right) + \Phi\left(\frac{x}{\sigma}\right) - \Phi\left(\frac{x}{s}\right) + \Phi\left(\frac{x}{s}\right) \right. \\
&\quad \left. - P_* \left\{ \frac{T_n^*}{s} \le \frac{x}{s} \right\} \right| \\
&\quad \le \sup_x \left| P_F \left\{ \frac{T_n}{\sigma} \le \frac{x}{\sigma} \right\} - \Phi\left(\frac{x}{\sigma}\right) \right| + \sup_x \left| \Phi\left(\frac{x}{\sigma}\right) - \Phi\left(\frac{x}{s}\right) \right| \\
&\quad + \sup_x \left| \Phi\left(\frac{x}{s}\right) - P_* \left\{ \frac{T_n^*}{s} \le \frac{x}{s} \right\} \right| \\
&= A_n + B_n + C_n, \quad \text{say.}
\end{aligned}
$$

That $A_n \to 0$ is a direct consequence of Polya's theorem. Also, s^2 converges almost surely to σ^2 and so, by the continuous mapping theorem, s converges almost surely to σ. Then $B_n \Rightarrow 0$ almost surely by the fact that $\Phi(\cdot)$ is a uniformly continuous function. Finally, we can apply the Berry-Esseen theorem to show that C_n goes to zero:

$$\begin{aligned} C_n &\le \frac{4}{5\sqrt{n}} \cdot \frac{E_{F_n}|X_1^* - \overline{X}_n|^3}{[\mathrm{var}_{F_n}(X_1^*)]^{3/2}} = \frac{4}{5\sqrt{n}} \cdot \frac{\sum_{i=1}^n |X_i - \overline{X}_n|^3}{ns^3} \\ &\le \frac{4}{5n^{3/2}s^3} \cdot 2^3 \left[\sum_{i=1}^n |X_i - \mu|^3 + n|\mu - \overline{X}_n|^3 \right] \\ &= \frac{M}{s^3} \left[\frac{1}{n^{3/2}} \sum_{i=1}^n |X_i - \mu|^3 + \frac{|\overline{X}_n - \mu|^3}{\sqrt{n}} \right], \end{aligned}$$

where $M = \frac{32}{5}$.

Since $s \Rightarrow \sigma > 0$ and $\overline{X}_n \Rightarrow \mu$, it is clear that $|\overline{X}_n - \mu|^3/(\sqrt{n}s^3) \Rightarrow 0$ almost surely. As regards the first term, let $Y_i = |X_i - \mu|^3$ and $\delta = 2/3$. Then the $\{Y_i\}$ are iid and

$$E|Y_i|^\delta = E_F|X_i - \mu|^{3\cdot 2/3} = \mathrm{Var}_F(X_1) < \infty.$$

It now follows from the Zygmund-Marcinkiewicz SLLN that

$$\frac{1}{n^{3/2}} \sum_{i=1}^n |X_i - \mu|^3 = n^{-1/\delta} \sum_{i=1}^n Y_i \Rightarrow 0 \text{ a.s.} \quad \text{as } n \to \infty.$$

Thus, $A_n + B_n + C_n \Rightarrow 0$ almost surely, and hence $K(H_n, H_{\text{Boot}}) \Rightarrow 0$.

We now proceed to a proof of convergence under the Wasserstein-Kantorovich-Mallows metric ℓ_2. Recall that convergence in ℓ_2 allows us to conclude more than weak convergence. We start with a sequence of results that enumerate useful properties of the ℓ_2 metric.

These facts (see Bickel and Freedman (1981)) are needed to prove consistency of H_{Boot} in the ℓ_2 metric.

Lemma 29.2 Let $G_n, G \in \Gamma_2$. Then $\ell_2(G_n, G) \to 0$ if and only if

$$G_n \xRightarrow{\mathcal{L}} G \quad \text{and} \quad \lim_{n\to\infty} \int x^k dG_n(x) = \int x^k \, dG(x), \quad k = 1, 2.$$

Lemma 29.3 Let $G, H \in \Gamma_2$, and suppose $Y_1, \ldots, Y_n$ are iid G and $Z_1, \ldots, Z_n$ are iid H. If $G^{(n)}$ is the CDF of $\sqrt{n}(\bar{Y} - \mu_G)$ and $H^{(n)}$ is the CDF of $\sqrt{n}(\bar{Z} - \mu_H)$, then $\ell_2(G^{(n)}, H^{(n)}) \leq \ell_2(G, H), \quad \forall\, n \geq 1.$

Lemma 29.4 (**Glivenko-Cantelli**) Let $X_1, X_2, \ldots, X_n$ be iid F and let F_n be the empirical CDF. Then $F_n(x) \to F(x)$ almost surely, uniformly in x.

Lemma 29.5 Let $X_1, X_2, \ldots, X_n$ be iid F and let F_n be the empirical CDF. Then $\ell_2(F_n, F) \Rightarrow 0$ almost surely.

The proof that $\ell_2(H_n, H_{\text{Boot}})$ converges to zero almost surely follows on simply putting together the lemmas 29.2–29.5. We omit this easy verification.

It is natural to ask if the bootstrap is consistent for $\sqrt{n}(\bar{X} - \mu)$ even when $E_F(X_1^2) = \infty$. If we insist on strong consistency, then the answer is negative. The point is that the sequence of bootstrap distributions is a sequence of random CDFs and so it cannot be expected a priori that it will converge to a fixed CDF. It may very well converge to a random CDF, depending on the particular realization $X_1, X_2, \ldots$. One runs into this problem if $E_F(X_1^2)$ does not exist. We state the result below.

Theorem 29.2 Suppose $X_1, X_2, \ldots$ are iid random variables. There exist $\mu_n(X_1, X_2, ..., X_n)$, an increasing sequence c_n, and a fixed CDF $G(x)$ such that

$$P_*\left(\frac{\sum_{i=1}^{n}(X_i^* - \mu(X_1, \ldots, X_n))}{c_n} \leq x\right) \xrightarrow{\text{a.s.}} G(x)$$

if and only if $E_F(X_1^2) < \infty$, in which case $\frac{c_n}{\sqrt{n}} \longrightarrow 1$.

Remark. The moral of Theorem 29.2 is that the existence of a nonrandom limit itself would be a problem if $E_F(X_1^2) = \infty$. See Athreya (1987), Giné and Zinn (1989), and Hall (1990) for proofs and additional examples.

The consistency of the bootstrap for the sample mean under finite second moments is also true for the multivariate case. We record consistency under the Kolmogorov metric next; see Shao and Tu (1995) for a proof.

Theorem 29.3 Let $\underset{\sim}{X}_1, \cdots, \underset{\sim}{X}_n, \cdots$ be iid F with $\text{cov}_F(\underset{\sim}{X}_1) = \Sigma$, Σ finite. Let $T(\underset{\sim}{X}_1, \underset{\sim}{X}_2, ..., \underset{\sim}{X}_n, F) = \sqrt{n}(\underset{\sim}{\bar{X}} - \underset{\sim}{\mu})$. Then $K(H_{\text{Boot}}, H_n) \xrightarrow{\text{a.s.}} 0$ as $n \to \infty$.

29.3 Delta Theorem for the Bootstrap

We know from the ordinary delta theorem that if T admits a CLT and $g(\cdot)$ is a smooth transformation, then $g(T)$ also admits a CLT. If we were to believe in our rule of thumb, then this would suggest that the bootstrap should be consistent for $g(T)$ if it is already consistent for T. For the case of sample mean vectors, the following result holds; again, see Shao and Tu (1995) for a proof.

Theorem 29.4 Let $\underset{\sim}{X}_1, \underset{\sim}{X}_2, ..., \underset{\sim}{X}_n \overset{\text{iid}}{\sim} F$ and let $\Sigma_{p\times p} = \text{cov}_F(\underset{\sim}{X}_1)$ be finite. Let $T(\underset{\sim}{X}_1, \underset{\sim}{X}_2, ..., \underset{\sim}{X}_n, F) = \sqrt{n}(\underset{\sim}{\bar{X}} - \underset{\sim}{\mu})$ and, for some $m \geq 1$, let $g : \mathbb{R}^p \to \mathbb{R}^m$. If $\nabla g(\cdot)$ exists in a neighborhood of $\underset{\sim}{\mu}$, $\nabla g(\underset{\sim}{\mu}) \neq \underset{\sim}{0}$, and if $\nabla g(\cdot)$ is continuous at $\underset{\sim}{\mu}$, then the bootstrap is strongly consistent w.r.t. K for $\sqrt{n}(g(\underset{\sim}{\bar{X}}) - g(\underset{\sim}{\mu}))$.

Example 29.2 Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F$, and suppose $E_F(X_1^4) < \infty$. Let $\underset{\sim}{Y}_i = \binom{X_i}{X_i^2}$. Then, with $p = 2$, $\underset{\sim}{Y}_1, \underset{\sim}{Y}_2, ..., \underset{\sim}{Y}_n$ are iid p-dimensional vectors with $\text{cov}(\underset{\sim}{Y}_1)$ finite. Note that $\underset{\sim}{\bar{Y}} = \begin{pmatrix} \bar{X} \\ \frac{1}{n}\sum_{i=1}^n X_i^2 \end{pmatrix}$. Consider the transformation $g : \mathbb{R}^2 \to \mathbb{R}^1$ defined as $g(u, v) = v - u^2$. Then $\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - (\bar{X})^2 = g(\underset{\sim}{\bar{Y}})$. If we let $\underset{\sim}{\mu} = E(\underset{\sim}{Y}_1)$, then $g(\underset{\sim}{\mu}) = \sigma^2 = \text{Var}(X_1)$. Since $g(\cdot)$ satisfies the conditions of the Theorem 29.4, it follows that the bootstrap is strongly consistent w.r.t. K for $\sqrt{n}(\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 - \sigma^2)$.

29.4 Second-Order Accuracy of the Bootstrap

One philosophical question about the use of the bootstrap is whether the bootstrap has any advantages at all when a CLT is already available. To be specific, suppose $T(X_1, \ldots, X_n, F) = \sqrt{n}(\bar{X} - \mu)$. If $\sigma^2 = \text{Var}_F(X) < \infty$, then $\sqrt{n}(\bar{X} - \mu) \overset{\mathcal{L}}{\Rightarrow} N(0, \sigma^2)$ and $K(H_{\text{Boot}}, H_n) \overset{\text{a.s.}}{\to} 0$. So two competitive approximations to $P_F(T(X_1, \ldots, X_n, F) \leq x)$ are $\Phi(\frac{x}{\hat{\sigma}})$ and $P_{F_n}(\sqrt{n}(\bar{X}^* - \bar{X}) \leq x)$. It turns out that, for certain types of statistics, the bootstrap approximation is (theoretically) more accurate than the approximation provided by the CLT. Because any normal distribution is symmetric, the CLT cannot capture information about the skewness in the finite sample distribution of T. The bootstrap approximation does so. So the bootstrap succeeds in correcting for skewness, just as an Edgeworth expansion would do. This is called Edgeworth correction by the bootstrap, and the property is called second-order accuracy of the bootstrap. It is important to remember that

second-order accuracy is not automatic; it holds for certain types of T but not for others. It is also important to understand that practical accuracy and theoretical higher-order accuracy can be different things. The following heuristic calculation will illustrate when second-order accuracy can be anticipated. The first result on higher-order accuracy of the bootstrap is due to Singh (1981). In addition to the references we provided in the beginning, Lehmann (1999) gives a very readable treatment of higher-order accuracy of the bootstrap.

Suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F$ and $T(X_1, \ldots, X_n, F) = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$; here $\sigma^2 = \text{Var}_F(X_1) < \infty$. We know that T admits the Edgeworth expansion

$$\begin{aligned} P_F(T \le x) &= \Phi(x) + \frac{p_1(x|F)}{\sqrt{n}}\varphi(x) + \frac{p_2(x|F)}{n}\varphi(x) \\ &\quad + \text{smaller order terms}, \\ P_*(T^* \le x) &= \Phi(x) + \frac{p_1(x|F_n)}{\sqrt{n}}\varphi(x) + \frac{p_2(x|F_n)}{n}\varphi(x) \\ &\quad + \text{smaller order terms}, \\ H_n(x) - H_{\text{Boot}}(x) &= \frac{p_1(x|F) - p_1(x|F_n)}{\sqrt{n}} + \frac{p_2(x|F) - p_2(x|F_n)}{n} \\ &\quad + \text{smaller order terms}. \end{aligned}$$

Recall now that the polynomials p_1, p_2 are given as

$$\begin{aligned} p_1(x|F) &= \frac{\gamma}{6}(1-x^2), \\ p_2(x|F) &= x\left[\frac{\kappa-3}{24}(3-x^2) - \frac{\kappa^2}{72}(x^4 - 10x^2 + 15)\right], \end{aligned}$$

where $\gamma = \frac{E_F(X_1-\mu)^3}{\sigma^3}$ and $\kappa = \frac{E_F(X_1-\mu)^4}{\sigma^4}$. Since $\gamma_{F_n} - \gamma = O_p(\frac{1}{\sqrt{n}})$ and $\kappa_{F_n} - \kappa = O_p(\frac{1}{\sqrt{n}})$, just from the CLT for γ_{F_n} and κ_{F_n} under finiteness of four moments, one obtains $H_n(x) - H_{\text{Boot}}(x) = O_p(\frac{1}{n})$. If we contrast this with the CLT approximation, in general, the error in the CLT is $O(\frac{1}{\sqrt{n}})$, as is known from the Berry-Esseen theorem. The $\frac{1}{\sqrt{n}}$ rate cannot be improved in general even if there are four moments. Thus, by looking at the standardized statistic $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$, we have succeeded in making the bootstrap one order more accurate than the CLT. This is called second-order accuracy of the bootstrap. If one does not standardize, then

$$P_F(\sqrt{n}(\bar{X}-\mu) \le x) = P_F\left(\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \le \frac{x}{\sigma}\right) \to \Phi\left(\frac{x}{\sigma}\right),$$

and the leading term in the bootstrap approximation in this unstandardized case would be $\Phi(\frac{x}{\bar{\sigma}})$. So the bootstrap approximates the true CDF $H_n(x)$ also at the rate $\frac{1}{\sqrt{n}}$; i.e., if one does not standardize, then $H_n(x) - H_{\text{Boot}}(x) = O_p(\frac{1}{\sqrt{n}})$. We have now lost the second-order accuracy. The following second rule of thumb often applies.

Rule of Thumb Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F$ and $T(X_1, \ldots, X_n, F)$ a functional. If $T(X_1, \ldots, X_n, F) \overset{\mathcal{L}}{\Rightarrow} N(0, \tau^2)$, where τ is independent of F, then second-order accuracy is likely. Proving it will depend on the availability of an Edgeworth expansion for T. If τ depends on F (i.e., $\tau = \tau(F)$), then the bootstrap should be just first-order accurate.

Thus, as we will now see, the orthodox bootstrap is second-order accurate for the standardized mean $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$, although from an inferential point of view it is not particularly useful to have an accurate approximation to the distribution of $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$ because σ would usually be unknown, and the accurate approximation could not really be used to construct a confidence interval for μ. Still, the second-order accuracy result is theoretically insightful.

We state a specific result below for the case of standardized and nonstandardized sample means. Let $H_n(x) = P_F(\sqrt{n}(\bar{X} - \mu) \le x)$, $H_{n,0}(x) = P_F(\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \le x)$, $H_{\text{Boot}}(x) = P_*(\sqrt{n}(\bar{X}^* - \bar{X}) \le x)$, $H_{\text{Boot},0}(x) = P_{F_n}(\frac{\sqrt{n}(\bar{X}^*-\bar{X})}{s} \le x)$.

Theorem 29.5 Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F$.

(a) If $E_F|X_1|^3 < \infty$ and F is nonlattice, then $K(H_{n,0}, H_{\text{Boot},0}) = o_p(\frac{1}{\sqrt{n}})$.

(b) If $E_F|X_1|^3 < \infty$ and F is lattice, then $\sqrt{n}K(H_{n,0}, H_{\text{Boot},0}) \overset{P}{\to} c$, $0 < c < \infty$.

Remark. See Lahiri (2003) for a proof. The constant c in the lattice case equals $\frac{h}{\sigma\sqrt{2\pi}}$, where h is the span of the lattice $\{a + kh, k = 0, \pm 1, \pm 2, ...\}$ on which the X_i are supported. Note also that part (a) says that higher-order accuracy for the standardized case obtains with three moments; Hall (1988) showed that finiteness of three absolute moments is in fact necessary and sufficient for higher-order accuracy of the bootstrap in the standardized case. Bose and Babu (1991) investigate the *unconditional probability* that the Kolmogorov distance between H_{Boot} and H_n exceeds a quantity of the order $o(n^{-\frac{1}{2}})$ for a variety of statistics and show that, with various assumptions, this probability goes to zero at a rate faster than $O(n^{-1})$.

Example 29.3 How does the bootstrap compare with the CLT approximation in actual applications? The question can only be answered by case-by-case simulation. The results are mixed in the following numerical table. The X_i are iid Exp(1) in this example and $T = \sqrt{n}(\bar{X} - 1)$ with $n = 20$. For the bootstrap approximation, $B = 250$ was used.

t	$H_n(t)$	CLT approximation	$H_{\text{Boot}}(t)$
-2	0.0098	0.0228	0.0080
-1	0.1563	0.1587	0.1160
0	0.5297	0.5000	0.4840
1	0.8431	0.8413	0.8760
2	0.9667	0.9772	0.9700

29.5 Other Statistics

The ordinary bootstrap that resamples with replacement from the empirical CDF F_n is consistent for many other natural statistics besides the sample mean and even higher-order accurate for some, but under additional conditions. We mention a few such results below; see Shao and Tu (1995) for further details on the theorems in this section.

Theorem 29.6 (Sample Percentiles)

Let $X_1, \ldots, X_n$ be $\overset{\text{iid}}{\sim} F$ and let $0 < p < 1$. Let $\xi_p = F^{-1}(p)$ and suppose F has a positive derivative $f(\xi_p)$ at ξ_p. Let $T_n = T(X_1, \ldots, X_n, F) = \sqrt{n}(F_n^{-1}(p) - \xi_p)$ and $T_n^* = T(X_1^*, \ldots, X_n^*, F_n) = \sqrt{n}(F_n^{*-1}(p) - F_n^{-1}(p))$, where F_n^* is the empirical CDF of $X_1^*, \ldots, X_n^*$. Let $H_n(x) = P_F(T_n \le x)$ and $H_{\text{Boot}}(x) = P_*(T_n^* \le x)$. Then, $K(H_{\text{Boot}}, H_n) = O(n^{-1/4}\sqrt{\log\log n})$ almost surely.

Remark. So again we see that, under certain conditions that ensure the existence of a CLT, the bootstrap is consistent.

Next we consider the class of one-sample U-statistics.

Theorem 29.7 (U-statistics)

Let $U_n = U_n(X_1, \ldots, X_n)$ be a U-statistic with a kernel h of order 2. Let $\theta = E_F(U_n) = E_F[h(X_1, X_2)]$, where $X_1, X_2 \overset{\text{iid}}{\sim} F$. Assume:

(i) $E_F\left(h^2(X_1, X_2)\right) < \infty$.
(ii) $\tau^2 = \text{Var}_F\left(\tilde{h}(X)\right) > 0$, where $\tilde{h}(x) = E_F[h(X_1, X_2)|X_2 = x]$.
(iii) $E_F|h(X_1, X_1)| < \infty$.

Let $T_n = \sqrt{n}(U_n - \theta)$ and $T_n^* = \sqrt{n}(U_n^* - U_n)$, where $U_n^* = U_n(X_1^*, \ldots, X_n^*)$, $H_n(x) = P_F(T_n \le x)$, and $H_{\text{Boot}}(x) = P_*(T_n^* \le x)$. Then $K(H_n, H_{\text{Boot}}) \xrightarrow{\text{a.s}} 0$.

Remark. Under conditions (i) and (ii), $\sqrt{n}(U_n - \theta)$ has a limiting normal distribution. Condition (iii) is a new additional condition and actually cannot be relaxed. Condition (iii) is vacuous if the kernel h is bounded or a function of $|X_1 - X_2|$. Under additional moment conditions on the kernel h, there is also a higher-order accuracy result; see Helmers (1991).

Previously, we observed that the bootstrap is consistent for smooth functions of a sample mean vector. That lets us handle statistics such as the sample variance. Under some more conditions, even higher-order accuracy obtains. Here is a result in that direction.

Theorem 29.8 (Higher-Order Accuracy for Functions of Means)

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$ with $E_F(X_1) = \mu$ and $\text{cov}_F(X_1) = \Sigma_{p \times p}$. Let $g : \mathbb{R}^p \to \mathbb{R}$ be such that $g(\cdot)$ is twice continuously differentiable in some neighborhood of μ and $\nabla g(\mu) \neq 0$. Assume also:

(i) $E_F||X_1 - \mu||^3 < \infty$.
(ii) $\limsup_{||t|| \to \infty} \left| E_F\left(e^{it'X_1}\right) \right| < 1$.

Let $T_n = \frac{\sqrt{n}(g(\bar{X}) - g(\mu))}{\sqrt{(\nabla g(\mu))' \Sigma (\nabla g(\mu))}}$ and $T_n^* = \frac{\sqrt{n}(g(\bar{X}^*) - g(\bar{X}))}{\sqrt{(\nabla g(\bar{X}))' S (\nabla g(\bar{X}))}}$, where $S = S(X_1, \ldots, X_n)$ is the sample variance-covariance matrix. Also let $H_n(x) = P_F(T_n \le x)$ and $H_{\text{Boot}}(x) = P_*(T_n^* \le x)$. Then $\sqrt{n} K(H_n, H_{\text{Boot}}) \xrightarrow{\text{a.s}} 0$.

Finally, let us describe the case of the t-statistic. By our previous rule of thumb, we would expect the bootstrap to be higher-order accurate simply because the t-statistic is already studentized and has an asymptotic variance function independent of the underlying F.

Theorem 29.9 (Higher-Order Accuracy for the t-statistic)

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$. Suppose F is nonlattice and that $E_F(X^6) < \infty$. Let $T_n = \frac{\sqrt{n}(\bar{X} - \mu)}{s}$ and $T_n^* = \frac{\sqrt{n}(\bar{X}^* - \bar{X})}{s^*}$, where s^* is the standard deviation of $X_1^*, \ldots, X_n^*$. Let $H_n(x) = P_F(T_n \le x)$ and $H_{\text{Boot}}(x) = P_*(T_n^* \le x)$. Then $\sqrt{n} K(H_n, H_{\text{Boot}}) \xrightarrow{\text{a.s}} 0$.

29.6 Some Numerical Examples

The bootstrap is used in practice for a variety of purposes. It is used to estimate a CDF, a percentile, or the bias or variance of a statistic T_n. For example, if T_n is an estimate for some parameter θ, and if $E_F(T_n - \theta)$ is the bias of T_n, the bootstrap estimate $E_{F_n}(T_n^* - T_n)$ can be used to estimate the bias. Likewise, variance estimates can be formed by estimating $\mathrm{Var}_F(T_n)$ by $\mathrm{Var}_{F_n}(T_n^*)$. How accurate are the bootstrap-based estimates in reality?

This can only be answered on the basis of case-by-case simulation. Some overall qualitative phenomena have emerged from these simulations. They are:

(a) The bootstrap captures information about skewness that the CLT will miss.

(b) The bootstrap tends to underestimate the variance of a statistic T_n.

Here are a few numerical examples.

Example 29.4 Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Cauchy}(\mu, 1)$. Let M_n be the sample median and $T_n = \sqrt{n}(M_n - \mu)$. If n is odd, say $n = 2k + 1$, then there is an exact variance formula for M_n. Indeed

$$\mathrm{Var}(M_n) = \frac{2n!}{(k!)^2\pi^n} \int_0^{\pi/2} x^k(\pi - x)^k(\cot x)^2 dx;$$

see David (1981). Because of this exact formula, we can easily gauge the accuracy of the bootstrap variance estimate. In this example, $n = 21$ and $B = 200$. For comparison, the CLT-based variance estimate is also used, which is

$$\widehat{\mathrm{Var}(M_n)} = \frac{\pi^2}{4n}.$$

The exact variance, the CLT-based estimate, and the bootstrap estimate for the specific simulation are 0.1367, 0.1175, and 0.0517, respectively. Note the obvious underestimation of variance by the bootstrap. Of course, one cannot be sure if it is the idiosyncrasy of the specific simulation.

A general useful result on consistency of the bootstrap variance estimate for medians under very mild conditions is in Ghosh et al. (1984).

Example 29.5 Suppose $X_1, \ldots, X_n$ are iid Poi(μ), and let T_n be the t-statistic $T_n = \sqrt{n}(\bar{X} - \mu)/s$. In this example, $n = 20$ and $B = 200$, and for the actual data, μ was chosen to be 1. Apart from the bias and the variance of T_n, in this example we also report percentile estimates for T_n. The bootstrap percentile estimates are found by calculating T_n^* for the B resamples and calculating the corresponding percentile value of the B values of T_n^*. The bias and the variance are estimated to be -0.18 and 1.614, respectively. The estimated percentiles are reported in the following table.

α	Estimated 100αPercentile
0.05	-2.45
0.10	-1.73
0.25	-0.76
0.50	-0.17
0.75	0.49
0.90	1.25
0.95	1.58

On observing the $100(1-\alpha)\%$ estimated percentiles, it is clear that there seems to be substantial skewness in the distribution of T. Whether the skewness is truly as serious can be assessed by a large-scale simulation.

Example 29.6 Suppose (X_i, Y_i), $i = 1, 2, \cdots, n$ are iid $BVN(0, 0, 1, 1, \rho)$, and let r be the sample correlation coefficient. Let $T_n = \sqrt{n}(r - \rho)$. We know that $T_n \stackrel{\mathcal{L}}{\Rightarrow} N(0, (1-\rho^2)^2)$; see Chapter 3. Convergence to normality is very slow. There is also an exact formula for the density of r. For $n \geq 4$, the exact density is

$$f(r|\rho) = \frac{2^{n-3}(1-\rho^2)^{(n-1)/2}}{\pi(n-3)!}(1-r^2)^{(n-4)/2}\sum_{k=0}^{\infty}\Gamma\left(\frac{n+k-1}{2}\right)^2\frac{(2\rho r)^k}{k!};$$

see Tong (1990). In the following table, we give simulation averages of the estimated standard deviation of r by using the bootstrap. We used $n = 20$ and $B = 200$. The bootstrap estimate was calculated for 1000 independent simulations, and the table reports the average of the standard deviation estimates over the 1000 simulations.

n	True ρ	True s.d. of r	CLT estimate	Bootstrap estimate
20	0.0	0.230	0.232	0.217
	0.5	0.182	0.175	0.160
	0.9	0.053	0.046	0.046

Again, except when ρ is large, the bootstrap underestimates the variance and the CLT estimate is better.

29.7 Failure of the Bootstrap

In spite of the many consistency theorems in the previous sections, there are instances where the ordinary bootstrap based on sampling with replacement from F_n actually does not work. Typically, these are instances where the functional T_n fails to admit a CLT. Before seeing a few examples, we list a few situations where the ordinary bootstrap fails to estimate the CDF of T_n consistently:

(a) $T_n = \sqrt{n}(\bar{X} - \mu)$ when $\text{Var}_F(X_1) = \infty$.
(b) $T_n = \sqrt{n}(g(\bar{X}) - g(\mu))$ and $\nabla g(\mu) = 0$.
(c) $T_n = \sqrt{n}(g(\bar{X}) - g(\mu))$ and g is not differentiable at μ.
(d) $T_n = \sqrt{n}(F_n^{-1}(p) - F^{-1}(p))$ and $f(F^{-1}(p)) = 0$ or F has unequal right and left derivatives at $F^{-1}(p)$.
(e) The underlying population F_θ is indexed by a parameter θ, and the support of F_θ depends on the value of θ.
(f) The underlying population F_θ is indexed by a parameter θ, and the true value θ_0 belongs to the boundary of the parameter space Θ.

Example 29.7 Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F$ and $\sigma^2 = \text{Var}_F(X) = 1$. Let $g(x) = |x|$ and $T_n = \sqrt{n}(g(\bar{X}) - g(\mu))$. If the true value of μ is 0, then by the CLT for $\bar{X}$ and the continuous mapping theorem, $T_n \overset{\mathcal{L}}{\Rightarrow} |Z|$ with $Z \sim N(0, \sigma^2)$. To show that the bootstrap does not work in this case, we first need to observe a few subsidiary facts.

(a) For almost all sequences $\{X_1, X_2, \cdots\}$, the conditional distribution of $\sqrt{n}(\overline{X}_n^* - \overline{X}_n)$, given $\overline{X}_n$, converges in law to $N(0, \sigma^2)$ by the triangular array CLT (see van der Vaart (1998).
(b) The joint asymptotic distribution of $(\sqrt{n}(\overline{X}_n - \mu), \sqrt{n}(\overline{X}_n^* - \overline{X}_n)) \overset{\mathcal{L}}{\Rightarrow} (Z_1, Z_2)$, where Z_1, Z_2 are iid $N(0, \sigma^2)$.

In fact, a more general version of part (b) is true. Suppose (X_n, Y_n) is a sequence of random vectors such that $X_n \overset{\mathcal{L}}{\Rightarrow} Z \sim H$ (some Z) and $Y_n | X_n \overset{\mathcal{L}}{\Rightarrow} Z$ (the same Z) almost surely. Then $(X_n, Y_n) \overset{\mathcal{L}}{\Rightarrow} (Z_1, Z_2)$, where Z_1, Z_2 are iid $\sim H$.

Therefore, returning to the example, when the true μ is 0,

$$\begin{aligned} T_n^* &= \sqrt{n}(|\overline{X}_n^*| - |\overline{X}_n|) \\ &= |\sqrt{n}(\overline{X}_n^* - \overline{X}_n) + \sqrt{n}\,\overline{X}_n| - |\sqrt{n}\,\overline{X}_n| \\ &\stackrel{\mathcal{L}}{\Rightarrow} |Z_1 + Z_2| - |Z_1|, \end{aligned}$$

where Z_1, Z_2 are iid $N(0, \sigma^2)$. But this is not distributed as the absolute value of $N(0, \sigma^2)$. The sequence of bootstrap CDFs is therefore not consistent when $\mu = 0$.

Example 29.8 Let $X_1, X_2, \ldots, X_n \stackrel{\text{iid}}{\sim} U(0, \theta)$ and let $T_n = n(\theta - X_{(n)})$, $T_n^* = n(X_{(n)} - X_{(n)}^*)$. The ordinary bootstrap will fail in this example in the sense that the conditional distribution of T_n^* given $X_{(n)}$ does not converge to the Exp(θ) a.s. Let us assume $\theta = 1$. Then, for $t \geq 0$,

$$\begin{aligned} P_{F_n}(T_n^* \leq t) &\geq P_{F_n}(T_n^* = 0) \\ &= P_{F_n}(X_{(n)}^* = X_{(n)}) \\ &= 1 - P_{F_n}(X_{(n)}^* < X_{(n)}) \\ &= 1 - \left(\frac{n-1}{n}\right)^n \\ &\stackrel{n\to\infty}{\longrightarrow} 1 - e^{-1}. \end{aligned}$$

For example, take $t = 0.0001$. Then $\lim_n P_{F_n}(T_n^* \leq t) \geq 1 - e^{-1}$, while $\lim_n P_F(T_n \leq t) = 1 - e^{-0.0001} \approx 0$. So $P_{F_n}(T_n^* \leq t) \not\to P_F(T_n \leq t)$.

The phenomenon of this example can be generalized essentially to any CDF F with a compact support $[\underline{\omega}(F), \overline{\omega}(F)]$ with some conditions on F, such as existence of a smooth and positive density. This is one of the earliest examples of the failure of the ordinary bootstrap. We will revisit this issue in the next section.

29.8 *m* out of *n* Bootstrap

In the particular problems presented above and several other problems where the ordinary bootstrap fails to be consistent, resampling fewer than n observations from F_n, say m observations, cures the inconsistency problem. This is called the m out of n bootstrap. Typically, consistency will be regained if $m = o(n)$; in some general theorems in this regard, one requires $m^2 = o(n)$ or some similar stronger condition than $m = o(n)$. If the n out of n ordinary

bootstrap is already consistent, then there can still be m out of n schemes with m going to ∞ slower than n that are also consistent, but the m out of n scheme will perform somewhat worse than the n out of n. See Bickel, Göetze, and van Zwet (1997) for an overall review.

We will now present a collection of results that show that the m out of n bootstrap, written as the m/n bootstrap, solves the orthodox bootstrap's inconsistency problem in a number of cases; see Shao and Tu (1995) for proofs and details on all of the theorems in this section.

Theorem 29.10 Let $X_1, X_2, \ldots$ be iid F, where F is a CDF on $\mathbb{R}^d$, $d \geq 1$. Suppose $\mu = E_F(X_1)$ and $\Sigma = \text{cov}_F(X_1)$ exist, and suppose Σ is positive definite. Let $g : \mathbb{R}^d \to \mathbb{R}$ be such that $\nabla g(\mu) = 0$ and the Hessian matrix $\nabla^2 g(\mu)$ is not the zero matrix. Let $T_n = n(g(\bar{X}_n) - g(\mu))$ and $T^*_{m,n} = m(g(\bar{X}^*_m) - g(\bar{X}_n))$ and define $H_n(x) = P_F\{T_n \leq x\}$ and $H_{\text{Boot},m,n}(x) = P_*\{T^*_{m,n} \leq x\}$. Here $\bar{X}^*_m$ denotes the mean of an iid sample of size $m = m(n)$ from F_n, where $m \to \infty$ with n.

(a) If $m = o(n)$, then $K(H_{\text{Boot},m,n}, H_n) \overset{\mathcal{P}}{\Rightarrow} 0$.

(b) If $m = o(\frac{n}{\log\log n})$, then $K(H_{\text{Boot},m,n}, H_n) \overset{\text{a.s.}}{\Rightarrow} 0$.

Theorem 29.11 Let $X_1, X_2, \ldots$ be iid F, where F is a CDF on $\mathbb{R}$. For $0 < p < 1$, let $\xi_p = F^{-1}(p)$. Suppose F has finite and positive left and right derivatives $f(\xi_p+)$, $f(\xi_p-)$ and that $f(\xi_p+) \neq f(\xi_p-)$. Let $T_n = \sqrt{n}(F_n^{-1}(p) - \xi_p)$ and $T^*_{m,n} = \sqrt{m}(F_m^{*-1}(p) - F_n^{-1}(p))$, and define $H_n(x) = P_F\{T_n \leq x\}$and $H_{\text{Boot},m,n}(x) = P_*\{T^*_{m,n} \leq x\}$. Here, $F_m^{*-1}(p)$ denotes the pth quantile of an iid sample of size m from F_n.

(a) If $m = o(n)$, then $K(H_{\text{Boot},m,n}, H_n) \overset{\mathcal{P}}{\Rightarrow} 0$.

(b) If $m = o(\frac{n}{\log\log n})$, then $K(H_{\text{Boot},m,n}, H_n) \overset{a.s.}{\Rightarrow} 0$.

Theorem 29.12 Suppose F is a CDF on $\mathbb{R}$, and let $X_1, X_2, \ldots$ be iid F. Suppose $\theta = \theta(F)$ is such that $F(\theta) = 1$ and $F(x) < 1$ for all $x < \theta$. Suppose, for some $\delta > 0$, $P_F\left\{n^{1/\delta}(\theta - X_{(n)}) > x\right\} \longrightarrow e^{-(x/\theta)^\delta}$, $\forall\, x$. Let $T_n = n^{1/\delta}(\theta - X_{(n)})$ and $T^*_{m,n} = m^{1/\delta}(X_{(n)} - X^*_{(m)})$, and define $H_n(x) = P_F\{T_n \leq x\}$ and $H_{\text{Boot},m,n}(x) = P_*\{T^*_{m,n} \leq x\}$.

(a) If $m = o(n)$, then $K(H_{\text{Boot},m,n}, H_n) \overset{\mathcal{P}}{\Rightarrow} 0$.

(b) If $m = o(\frac{n}{\log\log n})$, then $K(H_{\text{Boot},m,n}, H_n) \overset{a.s.}{\Rightarrow} 0$.

Remark. Clearly an important practical question is the choice of the bootstrap resample size m. This is a difficult question to answer, and no precise prescriptions that have any sort of general optimality are possible. A rule of thumb is to take $m \approx 2\sqrt{n}$.

29.9 Bootstrap Confidence Intervals

The standard method to find a confidence interval for a parameter θ is to find a studentized statistic, sometimes called a pivot, say $T_n = \frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n}$, such that $T_n \stackrel{\mathcal{L}}{\Rightarrow} T$, with T having some known CDF G. An equal-tailed confidence interval for θ, asymptotically correct, is constructed as

$$\hat{\theta}_n - G^{-1}(1-\alpha/2)\hat{\sigma}_n \le \theta \le \hat{\theta}_n - G^{-1}(\alpha/2)\hat{\sigma}_n.$$

This agenda requires the use of a standard deviation estimate $\hat{\sigma}_n$ for the standard deviation of $\hat{\theta}_n$ and the knowledge of the function $G(x)$. Furthermore, in many cases, the limiting CDF G may depend on some unknown parameters, too, that will have to be estimated in turn to construct the confidence interval. The bootstrap methodology offers an omnibus, sometimes easy to implement, and often more accurate method of constructing confidence intervals. Bootstrap confidence intervals and lower and upper one-sided confidence limits of various types have been proposed in great generality. Although, as a matter of methodology, they can be used in an automatic manner, a theoretical evaluation of their performance requires specific structural assumptions. The theoretical evaluation involves an Edgeworth expansion for the relevant statistic and an expansion for their quantiles, called Cornish-Fisher expansions. Necessarily, we are limited to the cases where the underlying statistic admits a known Edgeworth and Cornish-Fisher expansion. The main reference is Hall (1988), but see also Göetze (1989), Hall and Martin (1989), Bickel (1992), Konishi (1991), DiCiccio and Efron (1996), and Lee (1999), of which the article by DiCiccio and Efron is a survey article and Lee (1999) discusses m/n bootstrap confidence intervals. There are also confidence intervals based on more general *subsampling* methods, which work asymptotically under the mildest conditions. These intervals and their extensions to higher dimensions are discussed in Politis, Romano, and Wolf (1999).

Over time, various bootstrap confidence limits have been proposed. Generally, the evolution is from the algebraically simplest to progressively more complicated and computer-intensive formulas for the limits. Many of these limits have, however, now been incorporated into standard statistical software. We present below a selection of these different bootstrap confidence

limits and bounds. Let $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ be a specific estimate of the underlying parameter of interest θ.

(a) *The bootstrap percentile lower bound* (BP). Let $G(x) = G_n(x) = P_F\{\hat{\theta}_n \le x\}$ be the exact distribution and let $\hat{G}(x) = P_*\{\hat{\theta}_n^* \le x\}$ be the bootstrap distribution. The lower $1-\alpha$ bootstrap percentile confidence bound would be $\hat{G}^{-1}(\alpha)$, so the reported interval would be $[\hat{G}^{-1}(\alpha), \infty)$. This was present in Efron (1979) itself, but it is seldom used because it tends to have a significant coverage bias.

(b) *Transformation-based bootstrap percentile confidence bound*. Suppose there is a suitable 1-1 transformation $\varphi = \varphi_n$ of $\hat{\theta}_n$ such that $P_F\{\varphi(\hat{\theta}_n) - \varphi(\theta) \le x\} = \psi(x)$, with ψ being a known continuous, strictly increasing, and symmetric CDF (e.g., the $N(0, 1)$ CDF). Then a transformation-based bootstrap percentile lower confidence bound for θ is $\varphi^{-1}(\hat{\varphi}_n + z_\alpha)$, where $\hat{\varphi}_n = \varphi(\hat{\theta}_n)$ and $z_\alpha = \psi^{-1}(\alpha)$. Transforming may enhance the quality of the confidence bound in some problems. But, on the other hand, it is rare that one can find such a 1-1 transformation with a known ψ.

(c) *Bootstrap-t* (BT). Let $t_n = \frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n}$, where $\hat{\sigma}_n$ is an estimate of the standard error of $\hat{\theta}_n$, and let $t_n^* = \frac{\hat{\theta}_n^* - \hat{\theta}_n}{\hat{\sigma}_n^*}$ be its bootstrap counterpart. As usual, let $H_{\text{Boot}}(x) = P_*\{t_n^* \le x\}$. The bootstrap-$t$ lower bound is $\hat{\theta}_n - H_{\text{Boot}}^{-1}(1-\alpha)\hat{\sigma}_n$, and the two-sided BT confidence limits are $\hat{\theta}_n - H_{\text{Boot}}^{-1}(1-\alpha_1)\hat{\sigma}_n$ and $\hat{\theta}_n - H_{\text{Boot}}^{-1}(\alpha_2)\hat{\sigma}_n$, where $\alpha_1 + \alpha_2 = \alpha$, the nominal confidence level.

(d) *Bias-corrected bootstrap percentile bound* (BC). The derivation of the BC bound involves quite a lot of calculation; see Efron (1981) and Shao and Tu (1995). The BC lower confidence bound is given by $\underline{\theta}_{\text{BC}} = \hat{G}^{-1}[\psi(z_\alpha + 2\psi^{-1}(\hat{G}(\hat{\theta}_n)))]$, where $\hat{G}$ is the bootstrap distribution of $\hat{\theta}_n^*$, ψ is as above, and $z_\alpha = \psi^{-1}(\alpha)$.

(e) *Hybrid bootstrap confidence bound* (BH). Suppose for some deterministic sequence $\{c_n\}$, $c_n(\hat{\theta}_n - \theta) \sim H_n$ and let H_{Boot} be the bootstrap distribution; i.e., the distribution of $c_n(\hat{\theta}_n^* - \hat{\theta}_n)$ under F_n. We know that $P_F\{c_n(\hat{\theta}_n - \theta) \le H_n^{-1}(1-\alpha)\} = 1 - \alpha$.
If we knew H_n, then we could turn this into a $100(1-\alpha)\%$ lower confidence bound, $\theta \ge \hat{\theta}_n - \frac{1}{c_n}H_n^{-1}(1-\alpha)$. But H_n is, in general, not known, so we approximate it by H_{Boot}. That is, the hybrid bootstrap lower confidence bound is defined as $\underline{\theta}_{\text{BH}} = \hat{\theta}_n - \frac{1}{c_n}H_{\text{Boot}}^{-1}(1-\alpha)$.

(f) *Accelerated bias-corrected bootstrap percentile bound* (BC_a). The ordinary bias-corrected bootstrap bound is based on the assumption that we

can find $z_0 = z_0(F, n)$ and ψ (for known ψ) such that

$$P_F\{\hat{\varphi}_n - \varphi + z_0 \le x\} = \psi(x).$$

The accelerated bias-corrected bound comes from the modified assumption that there exists a constant $a = a(F, n)$ such that $P_F\left\{\frac{\hat{\varphi}_n - \varphi}{1+a\varphi} + z_0 \le x\right\} = \psi(x)$. In applications, it is rare that even this modification holds exactly for any given F and n. Manipulation of this probability statement results in a lower bound, $\underline{\theta}_{\mathrm{BC_a}} = \hat{G}^{-1}\left(\psi\left(z_0 + \frac{z_\alpha + z_0}{1-a(z_\alpha - z_0)}\right)\right)$, where $z_\alpha = \psi^{-1}(\alpha)$, a is the acceleration parameter, and $\hat{G}$ is as before. We repeat that, of these, z_0 and a both depend on F and n. They will have to be estimated. Moreover, the CDF ψ will generally have to be replaced by an asymptotic version; e.g., an asymptotic normal CDF of $(\hat{\varphi}_n - \varphi)/(1+a\varphi)$. The exact manner in which z_0 and a depend on F and n is a function of the specific problem. For example, suppose that the problem to begin with is a parametric problem, $F = F_\theta$. In such a case, $z_0 = z_0(\theta, n)$ and $a = a(\theta, n)$. The exact form of $z_0(\theta, n)$ and $a(\theta, n)$ depends on F_θ, $\hat{\theta}_n$, and φ.

Remark. As regards computational simplicity, BP, BT, and BH are the simplest to apply; BC and BC_a are harder to apply and, in addition, are based on assumptions that will rarely hold exactly for finite n. Furthermore, BC_a involves estimation of a very problem-specific acceleration constant a. The bootstrap-t intervals are popular in practice, provided an estimate $\hat{\sigma}_n$ is readily available. The BP method usually suffers from a large bias in coverage and is seldom used.

Remark. If the model is parametric, $F = F_\theta$, and $\hat{\theta}_n$ is the MLE, then one can show the following general and useful formula: $a = z_0 = \frac{1}{6} \times$skewness coefficient of $\dot{\ell}(\theta)$, where $\dot{\ell}(\theta)$ is the score function, $\dot{\ell}(\theta) = \frac{d}{d\theta} \log f(x_1, \ldots, x_n|\theta)$. This expression allows for estimation of a and z_0 by plug-in estimates. Nonparametric estimates of a and z_0 have also been suggested; see Efron (1987) and Loh and Wu (1987).

We now state the theoretical coverage properties of the various one-sided bounds and two-sided intervals.

Definition 29.3 Let $0 < \alpha < 1$ and $I_n = I_n(X_1, \ldots, X_n)$ be a confidence set for the functional $\theta(F^{(n)})$, where $F^{(n)}$ is the joint distribution of $(X_1, \ldots, X_n)$. Then I_n is called kth-order accurate if $P_{F^{(n)}}\left\{I_n \ni \theta(F^{(n)})\right\} = 1 - \alpha + O(n^{-k/2})$.

The theoretical coverage properties below are derived by using Edgeworth expansions as well as Cornish-Fisher expansions for the underlying estimate $\hat{\theta}_n$. If $X_1, X_2, \ldots$ are iid F on $\mathbb{R}^d$, $1 \le d < \infty$, and if $\theta = \varphi(\mu)$, $\hat{\theta} = \varphi(\bar{X})$, for a sufficiently smooth map $\varphi : \mathbb{R}^d \to \mathbb{R}$, then such Edgeworth and Cornish-Fisher expansions are available. In the results below, it is assumed that θ and $\hat{\theta}$ are the images of μ and $\bar{X}$, respectively, under such a smooth mapping φ. See Hall (1988) for the exact details.

Theorem 29.13 The CLT, BP, BH and BC one-sided confidence bounds are first-order accurate. The BT and BC_a one-sided bounds are second-order accurate. The CLT, BP, BH, BT, and BC_a two-sided intervals are all second-order accurate.

Remark. For two-sided intervals, the higher-order accuracy result is expected because the coverage bias for the two tails cancels in the $n^{-1/2}$ term, as can be seen from the Edgeworth expansion. The striking part of the result is that the BT and BC_a can achieve higher-order accuracy even for one-sided bounds.

The second-order accuracy of the BT lower bound is driven by an Edgeworth expansion for H_n and an analogous one for H_{Boot}. One can invert these expansions for the CDFs to get expansions for their quantiles; i.e., to obtain Cornish-Fisher expansions. Under suitable conditions on F, H_n^{-1} and H_{Boot}^{-1} admit expansions of the forms

$$H_n^{-1}(t) = z_t + \frac{q_{11}(z_t, F)}{\sqrt{n}} + \frac{q_{12}(z_t, F)}{n} + o\left(\frac{1}{n}\right)$$

and

$$H_{\mathrm{Boot}}^{-1}(t) = z_t + \frac{q_{11}(z_t, F_n)}{\sqrt{n}} + \frac{q_{12}(z_t, F_n)}{n} + o\left(\frac{1}{n}\right) \text{(a.s.)},$$

where $q_{11}(\cdot, F)$ and $q_{12}(\cdot, F)$ are polynomials with coefficients that depend on the moments of F. The exact polynomials depend on what the statistic $\hat{\theta}_n$ is. For example, if $\hat{\theta}_n = \bar{X}$ and $\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum (X_i - \bar{X})^2}$, then $q_{11}(x, F) = -\frac{\gamma}{6}(1 + 2x^2)$, $q_{12} = x[\frac{x^2+3}{4} - \frac{\kappa(x^2-3)}{12} + \frac{5\gamma^2}{72}(4x^2 - 1)]$, where $\gamma = E_F \frac{(X-\mu)^3}{\sigma^3}$ and $\kappa = E_F \frac{(X-\mu)^4}{\sigma^4} - 3$. For a given t, $0 < t < 1$, on subtraction,

$$\begin{aligned} H_n^{-1}(t) - H_{\mathrm{Boot}}^{-1}(t) = {} & \frac{1}{\sqrt{n}}[q_{11}(z_t, F) - q_{11}(z_t, F_n)] \\ & + \frac{1}{n}[q_{12}(z_t, F) - q_{12}(z_t, F_n)] + o\left(\frac{1}{n}\right) \text{(a.s.)} \end{aligned}$$

$$= \frac{1}{\sqrt{n}} O_p\left(\frac{1}{\sqrt{n}}\right) + \frac{1}{n} O_p\left(\frac{1}{\sqrt{n}}\right) + o\left(\frac{1}{n}\right) \text{(a.s.)}$$
$$= O_p\left(\frac{1}{n}\right).$$

The actual confidence bounds obtained from H_n, H_{Boot} are $\underline{\theta}_{H_n} = \hat{\theta}_n - \hat{\sigma}_n H_n^{-1}(1-\alpha)$ and $\underline{\theta}_{\text{BT}} = \hat{\theta}_n - \hat{\sigma}_n H_{\text{Boot}}^{-1}(1-\alpha)$. On subtraction,

$$|\underline{\theta}_{H_n} - \underline{\theta}_{\text{BT}}| = \hat{\sigma}_n O_p\left(\frac{1}{n}\right) \overset{\text{typically}}{=} O_p(n^{-\frac{3}{2}}).$$

Thus, the bootstrap-t lower bound is approximating the idealized lower bound with third-order accuracy. In addition, it can be shown that $P(\theta \geq \underline{\theta}_{\text{BT}}) = 1 - \alpha + \frac{p(z_\alpha)\varphi(z_\alpha)}{n} + o\left(\frac{1}{n}\right)$, where $p(\cdot)$ is again a polynomial depending on the specific statistic and F. For the case of $\bar{X}$, as an example, $p(x) = \frac{x}{6}(1 + 2x^2)(\kappa - \frac{3}{2}\gamma^2)$. Notice the second-order accuracy in this coverage statement in spite of the fact that the confidence bound is one sided. Again, see Hall (1988) for full details.

29.10 Some Numerical Examples

How accurate are the bootstrap confidence intervals in practice? Only case-by-case numerical investigation can give an answer to that question. We report in the following table results of simulation averages of coverage and length in two problems. The sample size in each case is $n = 20$, in each case $B = 200$, the simulation size is 500, and the nominal coverage $1 - \alpha = .9$.

$\theta(F)$	Type of CI	F					
		$N(0,1)$		$t(5)$		Weibull	
		coverage	length	coverage	length	coverage	length
μ	Regular t	.9	0.76	.91	1.8	.75	2.8
	BP	.91	0.71	.84	1.7	.73	2.6
	BT	.92	0.77	.83	2.7	.83	5.5
σ^2	BP	.79	0.86	.68	1.1	.65	1.3
	BT	.88	1.5	.85	3.2	.83	5.5

From the table, the bootstrap-t interval seems to buy more accuracy (i.e., a smaller bias in coverage) with a larger length than the BP interval. But the BP interval has such a serious bias in coverage that the bootstrap-t may be preferable. To kill the bias, modifications of the BP method have been

suggested, such as the bias-corrected BP and the accelerated bias-corrected BP intervals. Extensive numerical comparisons are reported in Shao and Tu (1995).

29.11 Bootstrap Confidence Intervals for Quantiles

Another interesting problem is the estimation of quantiles of a CDF F on $\mathbb{R}$. We know, for example, that if $X_1, X_2, \ldots$ are iid F, if $0 < p < 1$, and if $f = F'$ exists and is strictly positive at $\xi_p = F^{-1}(p)$, then $\sqrt{n}(F_n^{-1}(p) - \xi_p) \overset{\mathcal{L}}{\Rightarrow} N(0, p(1-p)[f(\xi_p)]^{-2})$. So, a standard CLT-based interval is

$$F_n^{-1}(p) \pm \frac{z_{\alpha/2}}{\sqrt{n}} \cdot \frac{\sqrt{p(1-p)}}{\widehat{f(\xi_p)}},$$

where $\widehat{f(\xi_p)}$ is some estimate of the unknown $f = F'$ at the unknown ξ_p.

For a bootstrap interval, let H_n be the CDF of $\sqrt{n}(F^{-1}(p) - \xi_p)$ and H_{Boot} its bootstrap counterpart. Using the terminology from before, a hybrid bootstrap two-sided confidence interval for ξ_p is

$$\left[F_n^{-1}(p) - H_{\text{Boot}}^{-1}(1 - \tfrac{\alpha}{2})/\sqrt{n},\, F_n^{-1}(p) - H_{\text{Boot}}^{-1}(\tfrac{\alpha}{2})/\sqrt{n}\right].$$

It turns out that this interval is not only asymptotically correct but also comes with a surprising asymptotic accuracy. The main references are Hall, DiCiccio, and Romano (1989) and Falk and Kaufman (1991).

Theorem 29.14 Let $X_1, X_2, \ldots$ be iid and F a CDF on $\mathbb{R}$. For $0 < p < 1$, let $\xi_p = F^{-1}(p)$, and suppose $0 < f(\xi_p) = F'(\xi_p) < \infty$. If I_n is the two-sided hybrid bootstrap interval, then $P_F\{I_n \ni \xi_p\} = 1 - \alpha + O(n^{-1/2})$.

Remark. Actually, the best result available is stronger and says that $P_F\{I_n \ni \xi_p\} = 1 - \alpha + \frac{c(F,\alpha,p)}{\sqrt{n}} + o(n^{-1/2})$, where $c(F, \alpha, p)$ has an explicit but complicated formula. That the bias of the hybrid interval is $O(n^{-1/2})$ is *still* a surprise in view of the fact that the bootstrap distribution of $F_n^{-1}(p)$ is consistent at a very slow rate; see Singh (1981).

29.12 Bootstrap in Regression

Regression models are among the key ones that differ from the iid setup and are also among the most widely used. Bootstrap for regression cannot

be model-free; the particular choice of the bootstrap scheme depends on whether the errors are iid or not. We will only talk about the linear model with deterministic X and iid errors. Additional moment conditions will be necessary depending on the specific problem to which the bootstrap will be applied. The results here are available in Freedman (1981). First let us introduce some notation.

Model: $y_i = \beta' x_i + \epsilon_i$, where β is a $p \times 1$ vector and so is x_i, and ϵ_i are iid with mean 0 and variance $\sigma^2 < \infty$.
X is the $n \times p$ design matrix with ith row equal to x_i'; $H = X(X'X)^{-1}X'$ and $h_i = H_{ii} = x_i'(X'X)^{-1}x_i$.
$\hat{\beta} = \hat{\beta}_{LS} = (X'X)^{-1}X'y$ is the least squares estimate of β, where $y = (y_1, \cdots, y_n)'$ and $(X'X)^{-1}$ is assumed to be nonsingular.

The bootstrap scheme is defined below.

29.13 Residual Bootstrap

Let $e_1, e_2, \cdots, e_n$ denote the residuals obtained from fitting the model (i.e., $e_i = y_i - x_i'\hat{\beta}$); $\overline{e} = 0$ if $x_i = (1, x_{i1}, \cdots, x_{i,p-1})'$ but not otherwise. Define $\tilde{e}_i = e_i - \overline{e}$, and let $e_1^*, \cdots, e_n^*$ be a sample with replacement of size n from $\{\tilde{e}_1, \cdots, \tilde{e}_n\}$. Let $y_i^* = x_i'\hat{\beta} + e_i^*$ and let β^* be the LSE of β computed from (x_i, y_i^*), $i = 1, \cdots, n$. This is the bootstrapped version of $\hat{\beta}$, and the scheme is called the residual bootstrap (RB).

Remark. The more direct approach of resampling the pairs (x_i, y_i) is known as the paired bootstrap and is necessary when the errors are not iid; for example, the case where the errors are still independent but their variances depend on the corresponding covariate values (called the heteroscedastic case). In such a case, the residual bootstrap scheme would not work.

By simple matrix algebra, it can be shown that

$$E_*(\beta^*) = \hat{\beta},$$
$$\mathrm{cov}_*(\beta^*) = \hat{\sigma}^2(X'X)^{-1},$$

where $\hat{\sigma}^2 = (1/n)\sum_{i=1}^n (e_i - \overline{e})^2$. Note that $E(\hat{\sigma}^2) < \sigma^2$. So on average the bootstrap covariance matrix estimate will somewhat underestimate $\mathrm{cov}(\hat{\beta})$. However, $\mathrm{cov}_*(\beta^*)$ is still consistent under some mild conditions. See Shao and Tu (1995) or Freedman (1981) for the following result.

Theorem 29.15 Suppose $|X'X| \to \infty$ and $\max_{1\le i\le n} h_i \to 0$ as $n \to \infty$. Then $[\text{cov}_*(\beta^*)]^{-1}\text{cov}(\hat{\beta}) \Rightarrow I_{p\times p}$ almost surely.

Example 29.9 The only question is, when do the conditions $|X'X| \to \infty$, $\max_{1\le i\le n} h_i \to 0$ hold? As an example, take the basic regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ with one covariate. Then, $|X'X| = n\sum_i (x_i - \bar{x})^2$ and $h_i = (\sum_j x_j^2 - 2x_i \sum_j x_j + nx_i^2)/(n\sum_j (x_j - \bar{x})^2)$.

$$\therefore h_i \le \frac{4n \max_j x_j^2}{n\sum_j (x_j - \bar{x})^2} = \frac{4 \max_j x_j^2}{\sum_j (x_j - \bar{x})^2}.$$

Therefore, for the theorem to apply, it is enough to have $\max|x_j|/\sqrt{\sum(x_j - \bar{x})^2} \to 0$ and $n\sum(x_i - \bar{x})^2 \to \infty$.

29.14 Confidence Intervals

We present some results on bootstrap confidence intervals for a linear combination $\theta = c'\beta_1$, where $\beta' = (\beta_0, \beta_1')$; i.e., there is an intercept term in the model. Correspondingly, $x_i' = (1, t_i')$. The confidence interval for θ or confidence bounds (lower or upper) are going to be in terms of the studentized version of the LSE of θ, namely $\hat{\theta} = c'\hat{\beta}_1$. In fact, $\hat{\beta}_1 = S_{tt}^{-1} S_{ty}$, where $S_{tt} = \sum_i (t_i - \bar{t})(t_i - \bar{t})'$ and $S_{ty} = \sum_i (t_i - \bar{t})(y_i - \bar{y})'$. The bootstrapped version of $\hat{\theta}$ is $\theta^* = c'\beta_1^*$, where $\beta^{*'} = (\beta_0^*, \beta_1^{*'})$ as before. Since the variance of $\hat{\theta}$ is $\sigma^2 c' S_{tt}^{-1} c$, the bootstrapped version of the studentized $\hat{\theta}$ is

$$\theta_s^* = \frac{\theta^* - \hat{\theta}}{\sqrt{\frac{1}{n}\sum_i (y_i - x_i'\beta^*)^2 c' S_{tt}^{-1} c}}.$$

The bootstrap distribution is defined as $H_{\text{Boot}}(x) = P_*(\theta_s^* \le x)$. For given α, let $H_{\text{Boot}}^{-1}(\alpha)$ be the αth quantile of H_{Boot}. We consider the bootstrap-t (BT) confidence bounds and intervals for θ. They are obtained as

$$\underline{\theta}_{\text{BT}}^{(\alpha)} = \hat{\theta} - H_{\text{Boot}}^{-1}(1-\alpha)\sqrt{\hat{\sigma}^2 c' S_{tt}^{-1} c},$$
$$\overline{\theta}_{\text{BT}}^{(\alpha)} = \hat{\theta} - H_{\text{Boot}}^{-1}(\alpha)\sqrt{\hat{\sigma}^2 c' S_{tt}^{-1} c},$$

and the intervals $\theta_{L,\text{BT}} = \underline{\theta}_{\text{BT}}^{(\alpha/2)}$ and $\theta_{U,\text{BT}} = \overline{\theta}_{\text{BT}}^{(\alpha/2)}$.

There are some remarkable results on the accuracy in coverage of the BT one-sided bounds and confidence intervals. We state one key result below.

Theorem 29.16 (a) $P(\theta \geq \underline{\theta}_{\mathrm{BT}}) = (1-\alpha) + O(n^{-3/2})$.
(b) $P(\theta \leq \overline{\theta}_{\mathrm{BT}}) = (1-\alpha) + O(n^{-3/2})$.
(c) $P(\theta_{L,\mathrm{BT}} \leq \theta \leq \theta_{U,\mathrm{BT}}) = (1-\alpha) + O(n^{-2})$.

These results are derived in Hall (1989).

Remark. It is remarkable that one already gets third-order accuracy for the one-sided confidence bounds and fourth-order accuracy for the two-sided bounds. There seems to be no intuitive explanation for this phenomenon. It just happens that certain terms cancel in the Cornish-Fisher expansions used in the proof for the regression case.

29.15 Distribution Estimates in Regression

The residual bootstrap is also consistent for estimating the distribution of the least squares estimate $\hat{\beta}$ of the full vector β. The metric chosen is the Mallows-Wasserstein metric we used earlier for sample means of iid data. See Freedman (1981) for the result below. We first state the model and the required assumptions below.

Let $y_i = x_i'\beta + \varepsilon_i$, where x_i is the p-vector of covariates for the ith sample unit. Write the design matrix as X_n. We assume that the ε_i's are iid with mean 0 and variance $\sigma^2 < \infty$ and that $\{X_n\}$ is a sequence of nonstochastic matrices. We assume that, for every n ($n > p$), $X_n'X_n$ is positive definite. Let $h_i = x_i'(X'X)^{-1}x_i$ and let $h_{\max} = \max\{h_i\}$. We assume, for the consistency theorem below, that:

(C1) **Stability**: $\frac{1}{n}X_n'X_n \to V$, where V is a $p \times p$ positive definite matrix.
(C2) **Uniform asymptotic negligibility**: $h_{\max} \to 0$.

Under these conditions, we have the following theorem of Freedman (1981) for RB.

Theorem 29.17 Under conditions C1 and C2 above, we have the following:

(a) $\sqrt{n}(\hat{\beta} - \beta) \stackrel{\mathcal{L}}{\Rightarrow} N_p(0, \sigma^2 V^{-1})$.
(b) For almost all $\{\varepsilon_i : i \geq 1\}$, $\sqrt{n}(\beta^* - \hat{\beta}) \stackrel{\mathcal{L}}{\Rightarrow} N_p(0, \sigma^2 V^{-1})$.
(c) $\frac{1}{\sigma}(X_n'X_n)^{1/2}(\hat{\beta} - \beta) \stackrel{\mathcal{L}}{\Rightarrow} N_p(0, I_p)$.
(d) For almost all $\{\varepsilon_i : i \geq 1\}$, $\frac{1}{\hat{\sigma}}(X_n'X_n)^{1/2}(\beta^* - \hat{\beta}) \stackrel{\mathcal{L}}{\Rightarrow} N_p(0, I_p)$.

(e) If H_n and H_{Boot} are the true and bootstrap distributions of $\sqrt{n}(\hat{\beta} - \beta)$ and $\sqrt{n}(\beta^* - \hat{\beta})$, respectively, then for almost all $\{\varepsilon_i : i \geq 1\}$, $\ell_2(H_n, H_{\text{Boot}}) \to 0$.

Remark. This theorem gives a complete picture of the consistency issue for the case of a nonstochastic design matrix and iid errors using the residual bootstrap. If the errors are iid but the design matrices are random, the same results hold as long as the conditions of stability and uniform asymptotic negligibility stated earlier hold with probability 1. See Shao and Tu (1995) for the case of independent but not iid errors (for example, the heteroscedastic case).

29.16 Bootstrap for Dependent Data

The orthodox bootstrap does not work when the sample observations are dependent. This was already pointed out in Singh (1981). It took some time before consistent bootstrap schemes were offered for dependent data. There are consistent schemes that are meant for specific dependence structures (e.g., stationary autoregression of a known order) and also general bootstrap schemes that work for large classes of stationary time series without requiring any particular dependence structure. The model-based schemes are better for the specific models but can completely fall apart if some assumption about the specific model does not hold.

We start with examples of some standard short-range dependence time series models. As opposed to these models, there are some that have a long memory or long-range dependence. The bootstrap runs into problems for long-memory data; see Lahiri (2006).

Standard time series models for short-range dependent processes include:

(a) *Autoregressive processes*. The observations y_t are assumed to satisfy

$$y_t = \mu + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \ldots \theta_p y_{t-p} + \varepsilon_t,$$

where $1 \leq p < \infty$ and the ε_t's are iid white noise with mean 0 and variance $\sigma^2 < \infty$. The $\{y_t\}$ process is stationary if the solutions of the polynomial equation

$$1 + \theta_1 z + \theta_2 z^2 + \ldots + \theta_p z^p = 0$$

lie strictly *outside* the unit circle in the complex plane. This process is called autoregression of order p and is denoted by AR(p).

(b) *Moving average processes*. Given a white noise process $\{\varepsilon_t\}$ with mean 0 and variance $\sigma^2 < \infty$, the observations are assumed to satisfy

$$y_t = \mu + \varepsilon_t - \varphi_1\varepsilon_{t-1} - \varphi_2\varepsilon_{t-2} - \ldots - \varphi_q\varepsilon_{t-q},$$

where $1 \le q < \infty$. The process $\{y_t\}$ is stationary if the roots of

$$1 - \varphi_1 z - \varphi_2 z^2 - \ldots - \varphi_q z^q = 0$$

lie strictly *outside* the unit circle. This process is called a moving average process of order q and is denoted by MA(q).

(c) *Autoregressive moving average processes*. This combines the two previously mentioned models. The observations are assumed to satisfy

$$y_t = \mu + \theta_1 y_{t-1} + \ldots \theta_p y_{t-p} + \varepsilon_t - \varphi_1\varepsilon_{t-1} - \ldots - \varphi_q\varepsilon_{t-q}.$$

The process $\{y_t\}$ is called an autoregressive moving average process of order (p, q) and is denoted by ARMA(p, q).

For all of these processes, the autocorrelation sequence dies off quickly; in particular, if ρ_k is the autocorrelation of lag k, then $\sum_k |\rho_k| < \infty$.

29.17 Consistent Bootstrap for Stationary Autoregression

A version of the residual bootstrap (RB) was offered in Bose (1988) and shown to be consistent and even higher-order accurate for the least squares estimate (LSE) of the vector of regression coefficients in the stationary AR(p) case. For ease of presentation, we assume $\mu = 0$ and $\sigma = 1$. In this case, the LSE of $\theta = (\theta_1, \ldots, \theta_p)'$ is defined as $\hat{\theta} = \arg\min_\theta \sum_{t=1}^n \big[y_t - \sum_{j=1}^p \theta_j y_{t-j}\big]^2$, where $y_{1-p}, \ldots, y_0, y_1, \ldots, y_n$ is the observed data sequence. There is a closed-form expression of $\hat{\theta}$; specifically, $\hat{\theta} = S_{nn}^{-1}\big(\sum_{t=1}^n y_t y_{t-1}, \sum_{t=1}^n y_t y_{t-2}, \ldots, \sum_{t=1}^n y_t y_{t-p}\big)$, where $S_{nn} = ((S_{nn}^{ij}))_{p\times p}$ and $S_{nn}^{ij} = \sum_{t=1}^n y_{t-i} y_{t-j}$. Let $\sigma_k = \mathrm{cov}(y_i, y_{i+k})$ and let

$$\Sigma = \begin{vmatrix} \sigma_0 & \sigma_1 & \ldots \sigma_{p-1} \\ \sigma_1 & \sigma_0 & \ldots \sigma_{p-2} \\ \vdots & & \ddots \\ \sigma_{p-1} & \sigma_{p-2} & \ldots \sigma_0 \end{vmatrix}.$$

Assume Σ is positive definite. It is known that under this condition $\sqrt{n}\Sigma^{-1/2}(\hat{\theta}-\theta) \stackrel{\mathcal{L}}{\Rightarrow} N(0, I)$. So we may expect that with a suitable bootstrap scheme $\sqrt{n}\hat{\Sigma}^{-1/2}$ $(\theta^* - \hat{\theta})$ converges a.s. in law to $N(0, I)$. Here $\hat{\Sigma}$ denotes the sample autocovariance matrix. We now describe the bootstrap scheme given in Bose (1988).

Let $\hat{y}_t = \sum_{j=1}^{p} \hat{\theta}_j y_{t-j}$ and let the residuals be $e_t = y_t - \hat{y}_t$. To obtain the bootstrap data, define $\{y^*_{1-2p}, y^*_{2-2p}, \ldots, y^*_{-p}\} \equiv \{y_{1-p}, y_{2-p}, \ldots . y_0\}$. Obtain bootstrap residuals by taking a random sample with replacement from $\{e_t - \bar{e}\}$. Then obtain the "starred" data by using the equation $y^*_t = \sum_{j=1}^{p} \hat{\theta}_j y^*_{t-j} + e^*_t$. Then θ^* is the LSE obtained by using $\{y^*_t\}$. Bose (1988) proves the following result.

Theorem 29.18 Assume that ε_1 has a density with respect to Lebesgue measure and that $E(\varepsilon_1^8) < \infty$. If $H_n(x) = P\{\sqrt{n}\Sigma^{-1/2}(\hat{\theta} - \theta) \le x\}$ and $H_{\text{Boot}}(x) = P_*\{\sqrt{n}\hat{\Sigma}^{-1/2}(\theta^* - \hat{\theta}) \le x\}$, then $\|H_n - H_{\text{Boot}}\|_\infty = o(n^{-1/2})$, almost surely.

Remark. This was the first result on higher-order accuracy of a suitable form of the bootstrap for dependent data. One possible criticism of the otherwise important result is that it assumes a specific dependence structure and that it assumes the order p is known. More flexible consistent bootstrap schemes involve some form of block resampling, which we describe next.

29.18 Block Bootstrap Methods

The basic idea of the block bootstrap method is that if the underlying series is a stationary process with short-range dependence, then blocks of observations of suitable lengths should be approximately independent and the joint distribution of the variables in different blocks would be (about) the same due to stationarity. So, if we resample blocks of observations rather than observations one at a time, then that should bring us back to the nearly iid situation, a situation in which the bootstrap is known to succeed. The block bootstrap was first suggested in Carlstein (1986) and Künsch (1989). Various block bootstrap schemes are now available. We only present three such schemes, for which the block length is nonrandom. A small problem with some of the blocking schemes is that the "starred" time series is not stationary, although the original series is, by hypothesis, stationary. A version of the block bootstrap that resamples blocks of *random* length allows the "starred" series to be provably stationary. This is called the *stationary bootstrap*, proposed in Politis and Romano (1994), and Politis, Romano, and Wolf (1999). However, later theoretical studies have established that

the auxiliary randomization to determine the block lengths can make the stationary bootstrap less accurate. For this reason, we only discuss three blocking methods with nonrandom block lengths.

(a) *Nonoverlapping block bootstrap (NBB).* In this scheme, one splits the observed series $\{y_1, \ldots, y_n\}$ into nonoverlapping blocks

$$B_1 = \{y_1, \ldots, y_h\}, \ B_2 = \{y_{h+1}, \ldots, y_{2h}\}, \ldots,$$
$$B_m = \{y_{(m-1)h+1}, \ldots, y_{mh}\},$$

where it is assumed that $n = mh$. The common block length is h. One then resamples $B_1^*, B_2^*, \ldots, B_m^*$ at random, with replacement, from $\{B_1, \ldots, B_m\}$. Finally, the B_i^*'s are pasted together to obtain the "starred" series $y_1^*, \ldots, y_n^*$.

(b) *Moving block bootstrap (MBB).* In this scheme, the blocks are

$$B_1 = \{y_1, \ldots, y_h\}, \ B_2 = \{y_2, \ldots, y_{h+1}\}, \ldots, B_N = \{y_{n-h+1}, \ldots, y_n\},$$

where $N = n-h+1$. One then resamples $B_1^*, \ldots, B_m^*$ from $B_1, \ldots, B_N$, where still $n = mh$.

(c) *Circular block bootstrap (CBB).* In this scheme, one periodically extends the observed series as $y_1, y_2, \ldots, y_n, y_1, y_2, \ldots, y_n, \ldots$. Suppose we let z_i be the members of this new series, $i = 1, 2, \ldots$. The blocks are defined as

$$B_1 = \{z_1, \ldots, z_h\}, \ B_2 = \{z_{h+1}, \ldots, z_{2h}\}, \ldots, B_n = \{z_n, \ldots, z_{n+h-1}\}.$$

One then resamples $B_1^*, \ldots, B_m^*$ from $B_1, \ldots, B_n$.

Next we give some theoretical properties of the three block bootstrap methods described above. The results below are due to Lahiri (1999).

Suppose $\{y_i : -\infty < i < \infty\}$ is a d-dimensional stationary process with a finite mean μ and spectral density f. Let $h : \mathbb{R}^d \to \mathbb{R}^1$ be a sufficiently smooth function. Let $\theta = h(\mu)$ and $\hat{\theta}_n = h(\bar{y}_n)$, where $\bar{y}_n$ is the mean of the realized series. We propose to use the block bootstrap schemes to estimate the bias and variance of $\hat{\theta}_n$. Precisely, let $b_n = E(\hat{\theta}_n - \theta)$ be the bias and let $\sigma_n^2 = \mathrm{Var}(\hat{\theta}_n)$ be the variance. We use the block bootstrap-based estimates of b_n and σ_n^2, denoted by $\hat{b}_n$ and $\hat{\sigma}_n^2$, respectively.

Next, let $T_n = \hat{\theta}_n - \theta = h(\bar{y}_n) - h(\mu)$, and let $T_n^* = h(\bar{y}_n^*) - h(E_* \bar{y}_n^*)$. The estimates $\hat{b}_n$ and $\hat{\sigma}_n^2$ are defined as $\hat{b}_n = E_* T_n^*$ and $\hat{\sigma}_n^2 = \mathrm{Var}_*(T_n^*)$. Then the following asymptotic expansions hold; see Lahiri (1999).

Theorem 29.19 Let $h : \mathbb{R}^d \to \mathbb{R}^1$ be a sufficiently smooth function.

(a) For each of the NBB, MBB, and CBB, there exists $c_1 = c_1(f)$ such that

$$E\hat{b}_n = b_n + \frac{c_1}{nh} + o((nh)^{-1}), \quad n \to \infty.$$

(b) For the NBB, there exists $c_2 = c_2(f)$ such that

$$\mathrm{Var}(\hat{b}_n) = \frac{2\pi^2 c_2 h}{n^3} + o(hn^{-3}), \quad n \to \infty,$$

and for the MBB and CBB,

$$\mathrm{Var}(\hat{b}_n) = \frac{4\pi^2 c_2 h}{3n^3} + o(hn^{-3}), \quad n \to \infty.$$

(c) For each of NBB, MBB, and CBB, there exists $c_3 = c_3(f)$ such that $E(\hat{\sigma_n^2}) = \sigma_n^2 + \frac{c_3}{nh} + o((nh)^{-1}), \quad n \to \infty.$

(d) For NBB, there exists $c_4 = c_4(f)$ such that $\mathrm{Var}(\hat{\sigma_n^2}) = \frac{2\pi^2 c_4 h}{n^3} + o(hn^{-3}), \quad n \to \infty$, and for the MBB and CBB, $\mathrm{Var}(\hat{\sigma_n^2}) = \frac{4\pi^2 c_4 h}{3n^3} + o(hn^{-3}), \quad n \to \infty.$

These expansions are used in the next section.

29.19 Optimal Block Length

The asymptotic expansions for the bias and variance of the block bootstrap estimates, given in Theorem 29.19, can be combined to produce MSE-optimal block lengths. For example, for estimating b_n by $\hat{b}_n$, the leading term in the expansion for the MSE is

$$m(h) = \frac{4\pi^2 c_2 h}{3n^3} + \frac{c_1^2}{n^2 h^2}.$$

To minimize $m(\cdot)$, we solve $m'(h) = 0$ to get

$$h_{\mathrm{opt}} = \left(\frac{3c_1^2}{2\pi^2 c_2}\right)^{1/3} n^{1/3}.$$

Similarly, an MSE-optimal block length can be derived for estimating σ_n^2 by $\hat{\sigma}_n^2$. We state the following optimal block-length result of Lahiri (1999) below.

Theorem 29.20 For the MBB and the CBB, the MSE-optimal block length for estimating b_n by $\hat{b}_n$ satisfies

$$h_{\text{opt}} = \left(\frac{3c_1^2}{2\pi^2 c_2}\right)^{1/3} n^{1/3}(1 + o(1)),$$

and the MSE-optimal block length for estimating σ_n^2 by $\hat{\sigma}_n^2$ satisfies

$$h_{\text{opt}} = \left(\frac{3c_3^2}{2\pi^2 c_4}\right)^{1/3} n^{1/3}(1 + o(1)).$$

Remark. Recall that the constants c_i depend on the spectral density f of the process. So, the optimal block lengths cannot be used directly. Plug-in estimates for the c_i may be substituted, or the formulas can be used to try block lengths proportional to $n^{1/3}$ with flexible proportionality constants. There are also other methods in the literature on selection of block lengths; see Hall, Horowitz, and Jing (1995) and Politis and White (2004).

29.20 Exercises

Exercise 29.1 For $n = 10, 20, 50$, take a random sample from an $N(0, 1)$ distribution and bootstrap the sample mean $\bar{X}$ using a bootstrap Monte Carlo size $B = 200$. Construct a histogram and superimpose on it the exact density of $\bar{X}$. Compare the two.

Exercise 29.2 For $n = 5, 25, 50$, take a random sample from an Exp(1) density and bootstrap the sample mean $\bar{X}$ using a bootstrap Monte Carlo size $B = 200$. Construct a histogram and superimpose on it the exact density of $\bar{X}$ and the CLT approximation. Compare the two and discuss if the bootstrap is doing something that the CLT answer does not.

Exercise 29.3 * By using combinatorial coefficient matching cleverly, derive a formula for the number of distinct orthodox bootstrap samples with a general value of n.

Exercise 29.4 * For which, if any, of the sample mean, the sample median, and the sample variance is it possible to explicitly obtain the bootstrap distribution $H_{\text{Boot}}(x)$?

Exercise 29.5 * For $n = 3$, write an expression for the exact Kolmogorov distance between H_n and H_{Boot} when the statistic is $\bar{X}$ and $F = N(0, 1)$.

Exercise 29.6 For $n = 5, 25, 50$, take a random sample from an Exp(1) density and bootstrap the sample mean $\bar{X}$ using a bootstrap Monte Carlo size $B = 200$ using both the canonical bootstrap and the natural parametric bootstrap. Construct the corresponding histograms and superimpose them on the exact density. Is the parametric bootstrap more accurate?

Exercise 29.7 * Prove that under appropriate moment conditions, the bootstrap is consistent for the sample correlation coefficient r between two jointly distributed variables X, Y.

Exercise 29.8 * Give examples of three statistics for which the condition in the rule of thumb on second-order accuracy of the bootstrap does not hold.

Exercise 29.9 * By gradually increasing the value of n, numerically approximate the constant c in the limit theorem for the Kolmogorov distance for the Poisson(1) case (see the text for the definition of c).

Exercise 29.10 * For samples from a uniform distribution, is the bootstrap consistent for the second-largest order statistic? Prove your assertion.

Exercise 29.11 For $n = 5, 25, 50$, take a random sample from an Exp(1) density and compute the bootstrap-t, bootstrap percentile, and the usual t 95% lower confidence bounds on the population mean. Use $B = 300$. Compare them meaningfully.

Exercise 29.12 * Give an example of:

(a) a density such that the bootstrap is not consistent for the median;

(b) a density such that the bootstrap is not consistent for the mean;

(c) a density such that the bootstrap is consistent but not second-order accurate for the mean.

Exercise 29.13 For simulated independent samples from the $U[0, \theta)$ density, let $T_n = n(\theta - X_{(n)})$. For $n = 20, 40, 60$, numerically approximate

$K(H_{\text{Boot},m,n}, H_n)$ with varying choices of m and investigate the choice of an optimal m.

Exercise 29.14 * Suppose (X_i, Y_i) are iid samples from a bivariate normal distribution. Simulate $n = 25$ observations taking $\rho = .5$, and compute:

(a) the usual 95% confidence interval;
(b) the interval based on the variance stabilizing transformation (Fisher's z) (see Chapter 4);
(c) the bootstrap percentile interval;
(d) the bootstrap hybrid percentile interval;
(e) the bootstrap-t interval with $\hat{\sigma}_n$ as the usual estimate;
(f) the accelerated bias-corrected bootstrap interval using φ as Fisher's z, $z_0 = \frac{r}{2\sqrt{n}}$ (the choice coming from theory), and three different values of a near zero.
Discuss your findings.

Exercise 29.15 * In which of the following cases are the results in Hall (1988) not applicable and why?

(a) estimating the 80th percentile of a density on $\mathcal{R}$;
(b) estimating the variance of a Gamma density with known scale and unknown shape parameter;
(c) estimating θ in the $U[0, \theta]$ density;
(d) estimating $P(X > 0)$ in a location-parameter Cauchy density;
(e) estimating the variance of the t-statistic for Weibull data;
(f) estimating a binomial success probability.

Exercise 29.16 Using simulated data, compute a standard CLT-based 95% confidence interval and the hybrid bootstrap interval for the 90th percentile of a (i) standard Cauchy distribution and (ii) a Gamma distribution with scale parameter 1 and shape parameter 3. Compare them and comment. Use $n = 20, 40$.

Exercise 29.17 * Are the centers of the CLT-based interval and the hybrid bootstrap interval for a population quantile always the same? Sometimes the same?

Exercise 29.18 * Simulate a series of length 50 from a stationary $AR(p)$ process with $p = 2$ and then obtain the starred series by using the scheme in Bose (1988).

Exercise 29.19 * For the simulated data in Exercise 29.18, obtain the actual blocks in the NBB and the MBB schemes with $h = 5$. Hence, generate the starred series by pasting the resampled blocks.

Exercise 29.20 For $n = 25$, take a random sample from a bivariate normal distribution with zero means, unit variances, and correlation .6. Implement the residual bootstrap using $B = 150$. Compute a bootstrap estimate of the variance of the LSE of the regression slope parameter. Comment on the accuracy of this estimate.

Exercise 29.21 For $n = 25$, take a random sample from a bivariate normal distri-bution with zero means, unit variances, and correlation .6. Implement the paired bootstrap using $B = 150$. Compute a bootstrap estimate of the variance of the LSE of the regression slope parameter. Compare your results with the preceding exercise.

Exercise 29.22 * Give an example of two design matrices that do not satisfy the conditions C1 and C2 in the text.

Exercise 29.23 * Suppose the values of the covariates are $x_i = \frac{1}{i}$, $i = 1, 2, \cdots, n$ in a simple linear regression setup. Prove or disprove that the residual bootstrap consistently estimates the distribution of the LSE of the slope parameter if the errors are (i) iid $N(0, \sigma^2)$, (ii) iid $t(m, 0, \sigma^2)$, where m denotes the degree of freedom.

Exercise 29.24 * Suppose $\bar{X}_n$ is the sample mean of an iid sample from a CDF F with a finite variance and $\bar{X}_n^*$ is the mean of a bootstrap sample. Consistency of the bootstrap is a statement about the bootstrap distribution, conditional on the observed data. What can you say about the unconditional limit distribution of $\sqrt{n}(\bar{X}_n^* - \mu)$, where μ is the mean of F?

References

Athreya, K. (1987). Bootstrap of the mean in the infinite variance case, Ann. Stat.,15(2), 724–731.

Beran, R. (2003). The impact of the bootstrap on statistical algorithms and theory, Stat. Sci., 18(2), 175–184.

Bickel, P.J. (1992). Theoretical comparison of different bootstrap *t* confidence bounds, in *Exploring the Limits of Bootstrap*, R. LePage and L. Billard (eds.) John Wiley, New York, 65–76.

Bickel, P.J. (2003). Unorthodox bootstraps, Invited paper, J. Korean Stat. Soc., 32(3), 213–224.

Bickel, P.J. and Freedman, D. (1981). Some asymptotic theory for the bootstrap, Ann. Stat., 9(6), 1196–1217.

Bickel, P.J., Göetze, F., and van Zwet, W. (1997). Resampling fewer than *n* observations: gains, losses, and remedies for losses, Stat. Sinica, 1, 1–31.

Bose, A. (1988). Edgeworth correction by bootstrap in autoregressions, Ann. Stat., 16(4), 1709–1722.

Bose, A. and Babu, G. (1991). Accuracy of the bootstrap approximation, Prob. Theory Related Fields, 90(3), 301–316.

Bose, A. and Politis, D. (1992). A review of the bootstrap for dependent samples, in *Stochastic Processes and Statistical Inference*, B.L.S.P Rao and B.R. Bhat, (eds.), New Age, New Delhi.

Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence, Ann. Stat., 14(3), 1171–1179.

David, H.A. (1981). *Order Statistics*, Wiley, New York.

Davison, A.C. and Hinkley, D. (1997). *Bootstrap Methods and Their Application*, Cambridge University Press, Cambridge.

DiCiccio, T. and Efron, B. (1996). Bootstrap confidence intervals, with discussion, Stat. Sci., 11(3), 189–228.

Efron, B. (1979). Bootstrap methods: another look at the Jackknife, Ann. Stat., 7(1), 1–26.

Efron, B. (1981). Nonparametric standard errors and confidence intervals, with discussion, Can. J. Stat., 9(2), 139–172.

Efron, B. (1987). Better bootstrap confidence intervals, with comments, J. Am. Stat. Assoc., 82(397), 171–200.

Efron, B. (2003). Second thoughts on the bootstrap, Stat. Sci., 18(2), 135–140.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*, Chapman and Hall, New York.

Falk, M. and Kaufman, E. (1991). Coverage probabilities of bootstrap confidence intervals for quantiles, Ann. Stat., 19(1), 485–495.

Freedman, D. (1981). Bootstrapping regression models, Ann. Stat., 9(6), 1218–1228.

Ghosh, M., Parr, W., Singh, K., and Babu, G. (1984). A note on bootstrapping the sample median, Ann. Stat., 12, 1130–1135.

Giné, E. and Zinn, J. (1989). Necessary conditions for bootstrap of the mean, Ann. Stat., 17(2), 684–691.

Göetze, F. (1989). Edgeworth expansions in functional limit theorems, Ann. Prob., 17, 1602–1634.

Hall, P. (1986). On the number of bootstrap simulations required to construct a confidence interval, Ann. Stat., 14(4), 1453–1462.

Hall, P. (1988). Rate of convergence in bootstrap approximations, Ann. Prob., 16(4), 1665–1684.

Hall, P. (1989a). On efficient bootstrap simulation, Biometrika, 76(3), 613–617.

Hall, P. (1989b). Unusual properties of bootstrap confidence intervals in regression problems, Prob. Theory Related Fields, 81(2), 247–273.

Hall, P. (1990). Asymptotic properties of the bootstrap for heavy-tailed distributions, Ann. Prob., 18(3), 1342–1360.
Hall, P. (1992). *Bootstrap and Edgeworth Expansion*, Springer-Verlag, New York.
Hall, P. (2003). A short prehistory of the bootstrap, Stat. Sci., 18(2), 158–167.
Hall, P., DiCiccio, T., and Romano, J. (1989). On smoothing and the bootstrap, Ann. Stat., 17(2), 692–704.
Hall, P. and Martin, M.A. (1989). A note on the accuracy of bootstrap percentile method confidence intervals for a quantile, Stat. Prob. Lett., 8(3), 197–200.
Hall, P., Horowitz, J., and Jing, B. (1995). On blocking rules for the bootstrap with dependent data, Biometrika, 82(3), 561–574.
Helmers, R. (1991). On the Edgeworth expansion and bootstrap approximation for a studentized U-statistic, Ann. Stat., 19(1), 470–484.
Konishi, S. (1991). Normalizing transformations and bootstrap confidence intervals, Ann. Stat., 19(4), 2209–2225.
Künsch, H.R. (1989). The Jackknife and the bootstrap for general stationary observations, Ann. Stat., 17(3), 1217–1241.
Lahiri, S.N. (1999). Theoretical comparisons of block bootstrap methods, Ann. Stat., 27(1), 386–404.
Lahiri, S.N. (2003). *Resampling Methods for Dependent Data*, Springer-Verlag, New York.
Lahiri, S.N. (2006). Bootstrap methods, a review, in *Frontiers in Statistics*, J. Fan and H. Koul (eds.), Imperial College Press, London, 231–256.
Lee, S. (1999). On a class of m out of n bootstrap confidence intervals, J.R. Stat. Soc. B, 61(4), 901–911.
Lehmann, E.L. (1999). *Elements of Large Sample Theory*, Springer, New York.
Loh, W. and Wu, C.F.J. (1987). Discussion of "Better bootstrap confidence intervals" by Efron, B., J. Amer. Statist. Assoc., 82, 188–190.
Politis, D. and Romano, J. (1994). The stationary bootstrap, J. Am. Stat. Assoc., 89(428), 1303–1313.
Politis, D., Romano, J., and Wolf, M. (1999). *Subsampling*, Springer, New York.
Politis, D. and White, A. (2004). Automatic block length selection for the dependent bootstrap, Econ. Rev., 23(1), 53–70.
Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*, Springer-Verlag, New York.
Silverman, B. and Young, G. (1987). The bootstrap: to smooth or not to smooth?, Biometrika, 74, 469–479.
Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap, Ann. Stat., 9(6), 1187–1195.
Tong, Y.L. (1990). *The Multivariate Normal Distribution*, Springer, New York.
van der Vaart, A. (1998). *Asymptotic Statistics*, Cambridge University Press, Cambridge.