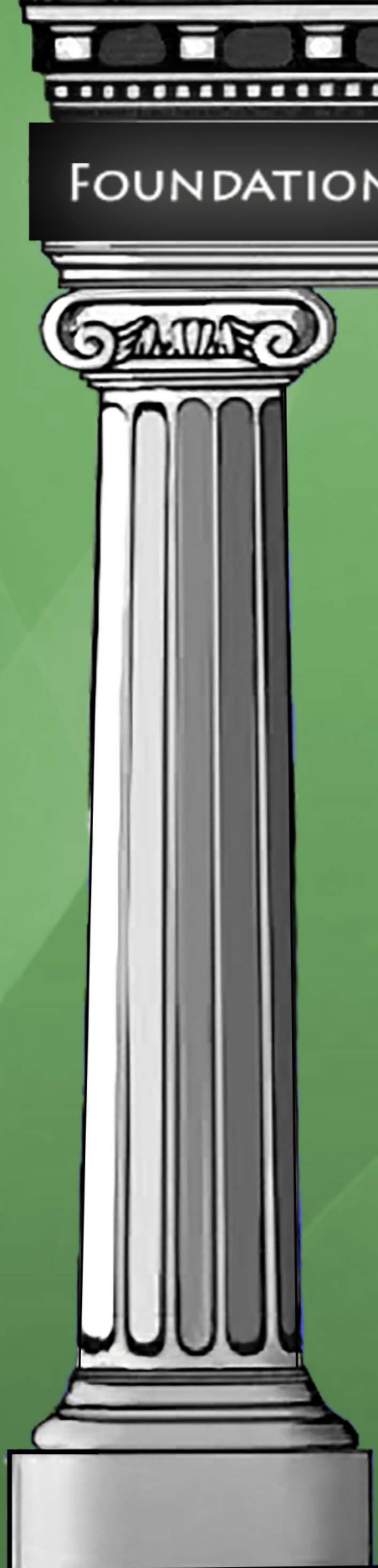
FOUNDATIONS OF QUANTITATIVE FINANCE



Book 4:

Distribution Functions and Expectations

ROBERT R. REITANO

Foundations of Quantitative Finance: 4. Distribution Functions and Expectations

Robert R. Reitano

Brandeis International Business School

Waltham, MA 02454

September, 2017

Contents

Preface

The idea for a reference book on the mathematical foundations of quantitative finance has been with me throughout my career in this field. But the urge to begin writing it didn't materialize until shortly after completing my first book, Introduction to Quantitative Finance: A Math Tool Kit, in 2010. The one goal I had for this reference book was that it would be complete and detailed in the development of the many materials one finds referenced in the various areas of quantitative finance. The one constraint I realized from the beginning was that I could not accomplish this goal, plus write a complete survey of the quantitative finance applications of these materials, in the 700 or so pages that I budgeted for myself for my first book. Little did I know at the time that this project would require a multiple of this initial page count budget even without detailed finance applications.

I was never concerned about the omission of the details on applications to quantitative finance because there are already a great many books in this area that develop these applications very well. The one shortcoming I perceived many such books to have is that they are written at a level of mathematical sophistication that requires a reader to have significant formal training in mathematics, as well as the time and energy to fill in omitted details. While such a task would provide a challenging and perhaps welcome exercise for more advanced graduate students in this field, it is likely to be less welcome to many other students and practitioners. It is also the case that quantitative finance has grown to utilize advanced mathematical theories from a number of fields. While there are also a great many very good references on these subjects, most are again written at a level that does not in my experience characterize the backgrounds of most students and practitioners of quantitative finance.

So over the past several years I have been drafting this reference book, accumulating the mathematical theories I have encountered in my work in this field, and then attempting to integrate them into a coherent collection of books that develops the necessary ideas in some detail. My target readers would be quantitatively literate to the extent of familiarity, indeed comfort, with the materials and formal developments in my first book, and sufficiently motivated to identify and then navigate the details of the materials they were attempting to master. Unfortunately, adding these details supports learning but also increases the lengths of the various developments. But this book was never intended to provide a "cover-to-cover" reading challenge, but rather to be a reference book in which one could find detailed foundational materials in a variety of areas that support further studies in quantitative finance.

Over these past years, one volume turned into two, which then became a work not likely publishable in the traditional channels given its unforgiving size and likely limited target audience. So I have instead decided to self-publish this work, converting the original chapters into stand-alone books, of which there are now nine. My goal is to finalize each book over the coming year or two.

I hope these books serve you well.

I am grateful for the support of my family: Lisa, Michael, David, and Jeffrey, as well as the support of friends and colleagues at Brandeis International Business School.

Robert R. Reitano

Brandeis International Business School

to Dorothy and Domenic

Introduction

This is the fourth book in a series of several that will be self-published under the collective title of *Foundations of Quantitative Finance*. Each book in the series is intended to build from the materials in earlier books, with the first several alternating between books with a more foundational mathematical perspective, which was the case with the first and third book, and books which develop probability theory and some quantitative applications to finance, the focus of the second and now this fourth book. But while providing many of the foundational theories underlying quantitative finance, this series of books does not provide a detailed development of these financial applications. Instead this series is intended to be used as a reference work for researchers and practitioners of quantitative finance who already have other sources for these detailed financial applications but find that such sources are written at a level which assume significant mathematical expertise, which if not possessed can be difficult to supplement.

Because the goal of many books in quantitative finance is to develop financial applications from an advanced point of view, it is often the case that advanced foundational materials from mathematics and probability theory are introduced and summarized but without a complete and formal development that would take the respective authors too far astray from their intended objectives. And while there are a great many excellent books on mathematics and probability theory, a number of which are cited in the references, such books typically develop materials with a eye to comprehensiveness in the subject matter, and not with an eye toward efficiently curating and developing the theory needed for applications in quantitative finance.

Thus the goal of this series is to introduce and develop in some detail a number of the foundational theories underlying quantitative finance, with topics curated from a vast mathematical and probability literature for the

express purpose of supporting applications in quantitative finance. In addition, the development of these topics will be found to be at a much greater level of detail than in most advanced quantitative finance books, and certainly in far more detail that most advanced mathematics texts.

The title of this fourth book, *Distribution Functions and Expectations,* is at once relatively generic yet at the same time succinctly identifies the overriding themes of this volume with Distribution Functions the title of chapter 1. Using the formidable tools of the Lebesgue integration and differentiation theory of book 3, the first section of chapter 1 develops a characterization of distribution functions on $\mathbb{R}$. The result is that every such distribution function is the sum of three component functions: a discrete component analogous to the discrete distribution functions of elementary probability theory; an absolutely continuous component that generalizes somewhat the distribution functions of the continuous theory; and finally a singular component. The chapter then introduces many common and familiar distribution functions from applications, and begins the derivation of distribution functions of transformed random variables. This theory is applied to derive the distribution functions of sums and ratios of random variables, with applications noted.

Chapter 2 is entitled Order Statistics, a topic which plays an important role in extreme value theory and elsewhere. The goal of this chapter is to derive the various distribution and density functions of the individual order statistics, as well as the joint, marginal and conditional distributions. The important Rényi representation theory for exponential order statistics is also developed.

Turning to the second theme of this book, Expectations of Random Variables 1 is the title of chapter 3, introducing the expectations operator in terms of the Riemann-Stieltjes integrals of book 3 and developing the familiar representations in the case of discrete and continuous distribution functions. Also addressed is the subtle ambiguity of such definitions as well as the framework for resolving such questions utilizing the more advanced integration theory of book 5 which will be applied in book 6. The chapter then introduces the various moments and the moment generating function, develops properties, provides examples using the distributions of chapter 1, and derives a number of important inequalities related to moments. The problem of uniqueness of moments and moment generating functions is then studied, as is the relationship between weak convergence of distributions and convergence of moments or moment generating functions.

In chapter 4, Simulating Samples of RVs - Examples, we return to the question of generating samples of random variables. With the theory ad-

dressed in chapter 4 of book 2, this chapter focuses on applications of this theory to the distribution functions exemplified. Both random and ordered samples are addressed, and the Rényi representation theory is seen to be an efficient tool for generating partial subsets of order statistics.

Chapter 5 is called Limit Theorems and begins with a review of the theoretical framework for the probability spaces introduced in book 2, on which collections of independent and sometimes identically distributed random variables are defined, and in which probability statements related to limits of these samples are meant to be measured. The chapter's many results are grouped by type into the three categories: Weak Convergence of Distributions; Laws of Large Numbers; and Convergence of Empirical Distribution Functions. Many of the most important and applicable limit theorems will be found here, including often overlooked results on quantiles of order statistics. The most general versions of the central limit theorems are deferred to book 6.

The final chapter 7, Estimating Tail Events, picks up where chapter 9 of book 2 left off. As in the book 2 development, this chapter is again split between large deviation theory, with the goal here being the Cramér-Chernoff Theorem, and extreme value theory. This latter investigation focuses on the development of the Hill estimator for the tail index as well as the Pickands–Balkema–de Haan theorem.

Chapter 1

Distribution Functions

1.1 A Characterization of Distribution Functions on $\mathbb{R}$

In this section we utilize some of the results from chapter 3 of book 3 to derive a characterization of the distribution function of a random variable. To state the proposition which summarizes the result we need some terminology. The following definition introduces **saltus functions,** which are generalized step functions with potentially countably many steps defined by a collection $\{x_n\}$. The general definition also allows two jumps at each domain point x_n, one of size u_n reflecting the discontinuity from the left, and one of size v_n, reflecting the discontinuity from the right. For the application in hand only u_n will be needed below since distribution functions are right continuous. Also, while in the general definition both u_n and v_n can be positive or negative, in the distribution function application it will always be the case that $u_n > 0$ and $v_n = 0$.

Definition 1.1 (Saltus function) *Given* $\{x_n\}_{n=1}^{\infty} \subset \mathbb{R}$ *and real sequences* $\{u_n\}_{n=1}^{\infty}$, $\{v_n\}_{n=1}^{\infty}$ *which are absolutely convergent:*

$$\sum_{n=1}^{\infty} |u_n| < \infty, \ \sum_{n=1}^{\infty} |v_n| < \infty,$$

a ***saltus function*** $f(x)$ *is defined as:*

$$f(x) = \sum_{n=1}^{\infty} f_n(x)$$

where

$$f_n(x) = \begin{cases} 0, & x < x_n, \\ u_n, & x = x_n, \\ u_n + v_n & x > x_n. \end{cases}$$

In other words,

$$f(x) = \sum\nolimits_{x_n \leq x} u_n + \sum\nolimits_{x_n < x} v_n.$$

In addition to saltus functions, recall the following notions from book 3:

Definition 1.2 *A function $f(x)$ is **singular** on the interval $[a, b]$ if $f(x)$ is continuous, monotonically increasing with $f(b) > f(a)$, and $f'(x) = 0$ almost everywhere.*

*A function $f(x)$ defined on $[a, b]$ is **absolutely continuous** if for any $\epsilon > 0$ there is a δ so that*

$$\sum\nolimits_{i=1}^{n} \left| f(x_i) - f(x_i') \right| < \epsilon$$

for any finite collection of disjoint subintervals, $\{(x_i', x_i)\} \subset [a, b]$, with

$$\sum\nolimits_{i=1}^{n} \left| x_i - x_i' \right| < \delta.$$

The relevant facts from book 3 on such functions are:

- **Singular functions**: Indeed, that these exist. An example is the **Cantor function**, named for **Georg Cantor** (1845 – 1918) and given in definition 3.49 of book 3.

- **Absolutely continuous functions:** As summarized in proposition 3.61 of book 3, $f(x)$ is absolutely continuous on $[a, b]$ if and only if $f(x)$ equals the Lebesgue integral of its derivative on this interval:

  $$f(x) = f(a) + (\mathcal{L}) \int_a^x f'(y) dy,$$

  which is version I of the fundamental theorem of calculus. Implicit in this result is that $f'(x)$ exists almost everywhere and is Lebesgue integrable.

1.1 A CHARACTERIZATION OF DISTRIBUTION FUNCTIONS ON $\mathbb{R}3$

Proposition 1.3 *Let X be a random variable on a probability space $(\mathcal{S}, \mathcal{E}, \lambda)$ with distribution function $F(x) \equiv \lambda\left[X^{-1}(-\infty, x]\right]$. Then $F(x)$ is differentiable almost everywhere and:*

$$F(x) = F_{SLT}(x) + F_{AC}(x) + F_{SN}(x), \tag{1.1}$$

where:

- *$F_{SLT}(x)$ is a saltus function.*
- *$F_{AC}(x)$ is absolutely continuous, and thus the density function $f_{AC}(y) \equiv F'_{AC}(y)$ exists almost everywhere, is measurable, and*

$$F_{AC}(x) = (\mathcal{L}) \int_{-\infty}^{x} f_{AC}(y)dy.$$

- *$F_{SN}(x)$ is identically 0, or a singular function.*

Proof. *By proposition 3.60 of book 1 and subsequent remarks, distribution functions are increasing and thus differentiable almost everywhere by proposition 3.15 of book 3. Such functions are also continuous from the right and have left limits, and have at most countably many points of discontinuity which we denote $\{x_n\}_{n=1}^{\infty}$. At such points define $u_n = F(x_n) - F(x_n^-)$ where $F(x^-) \equiv \lim_{y \to x-} F(y)$, and hence from 3.26 of book 1: $u_n = \lambda(X^{-1}(x_n)) > 0$. Define:*

$$F_{SLT}(x) \equiv \sum\nolimits_{x_n \leq x} u_n,$$

and note that $F_{SLT}(x)$ is increasing by definition, and right continuous. To prove right continuity let x and $\epsilon > 0$ be given. Then for $x \leq y$,

$$F_{SLT}(y) - F_{SLT}(x) = \sum\nolimits_{x < x_n \leq y} u_n,$$

and this summation is finite or countable. In the former case there is δ so that for $y \leq x + \delta$ this summation is zero and is then bounded by ϵ. If countable, then since $\sum u_n \leq 1$ the summation is convergent and thus can be made arbitrarily small by eliminating finitely many terms, or equivalently, reducing y.

Next, $G(x) \equiv F(x) - F_{SLT}(x)$ is an increasing function. If $y > x$ then $F(y^-) \geq F(x^-)$ and:

$$\begin{aligned} F_{SLT}(y) - F_{SLT}(x) &\equiv \left[F(y) - F(y^-)\right] - \left[F(x) - F(x^-)\right] \\ &\leq F(y) - F(x), \end{aligned}$$

which obtains $G(y) - G(x) \geq 0$. *Further,* $G(x)$ *is continuous since if* $y > x$:

$$G(y) - G(x) = F(y) - F(x) - \sum\nolimits_{x<x_n\leq y} u_n,$$

and the result follows from right continuity of $F(x)$ *and that* $\sum u_n \to 0$ *as* $y \to x$ *as noted above. If* $y < x$ *then*

$$G(x) - G(y) = F(x) - F(y) - \sum\nolimits_{y<x_n\leq x} u_n,$$

and the result will follow as an exercise by consideration of the two cases where x *is a continuity point, or a left discontinuity point, of* $F(x)$.

Hence by propositions 3.15 and 3.16 and remark 3.17 of book 3, $G(x)$ *is differentiable almost everywhere,* $G'(x) \geq 0$, *and is Lebesgue integrable. Defining*

$$F_{AC}(x) \equiv (\mathcal{L}) \int_{-\infty}^{x} G'(y)dy,$$

then $F_{AC}(x)$ *is increasing and by proposition 3.16 of book 3, for every interval* $[a, b]$:

$$(\mathcal{L}) \int_{a}^{b} G'(y)dy \leq G(b) - G(a).$$

By proposition 3.61 of book 3, $F_{AC}(x)$ *is absolutely continuous and equal almost everywhere to the Lebesgue integral of its derivative:*

$$F_{AC}(x) = (\mathcal{L}) \int_{-\infty}^{x} F'_{AC}(y)dy,$$

noting that $F(x) \to 0$ *as* $x \to -\infty$ *from book 1.*

Finally consider $H(x) \equiv G(x) - F_{AC}(x)$. *Then as a difference of continuous increasing functions* $H(x)$ *is continuous and of bounded variation, but also by construction* $H'(x) = 0$ *almost everywhere. In addition, for every interval* $[a, b]$:

$$F_{AC}(b) - F_{AC}(a) = \int_{a}^{b} F'_{AC}(y)dy = \int_{a}^{b} G'(y)dy \leq G(b) - G(a),$$

and so $H(a) \leq H(b)$ *and* $H(x)$ *is increasing. If* $H(x)$ *is constant it must be identically* 0 *since* $H(x) \to 0$ *as* $x \to -\infty$. *Otherwise, there exists an interval* $[a, b]$ *for which* $H(a) < H(b)$ *and by definition,* $H(x)$ *is a singular function. In either case we define* $F_{SN}(x) \equiv H(x)$. ■

Remark 1.4 *Note that it is possible to have either or both $F_{SLT}(x) \equiv 0$ or $F_{AC}(x) \equiv 0$ in the above proposition. The only reason to highlight the fact that we may have $F_{SN}(x) \equiv 0$ is that by definition, a singular function must have a positive increase somewhere, whereas there is no definitional reason that saltus functions and absolutely continuous functions cannot be identically zero.*

1.2 Examples of Distribution Functions on $\mathbb{R}$

The above section provides a framework for thinking about the distribution functions of a random variable. In this section we exemplify a variety of popular examples of such functions within this framework.

1.2.1 Discrete Distribution Functions

Discrete probability theory studies random variables for which $F_{AC}(x) = F_{SN}(x) = 0$, and so

$$F(x) = F_{SLT}(x), \tag{1.2}$$

and hence for which $\sum_{n=1}^{\infty} u_n = 1$. Since $u_n = \mu(X^{-1}(x_n)) > 0$ as noted above, it is conventional to define the **probability density function (p.d.f.)** associated with a discrete random variable by

$$f(x) = \begin{cases} u_n, & x = x_n, \\ 0, & \text{otherwise}, \end{cases}$$

and hence the **distribution function (d.f.)** is given by:

$$F(x) = \sum\nolimits_{x_n \leq x} f(x_n). \tag{1.3}$$

In most cases, it is the probability density functions that are explicitly defined in a given application. Frequently encountered examples of discrete distribution functions are the discrete rectangular, binomial, geometric, negative binomial, and Poisson distribution functions.

Example 1.5 **1**. ***Discrete Rectangular Distribution:*** *The defining collection $\{x_n\}_{n=1}^{\infty}$ for this distribution is finite and can otherwise be arbitrary. However this collection is conventionally taken as $\{j/n\}_{j=1}^{n}$ or $\{(j-1)/n\}_{j=1}^{n}$, so in either case $\{x_j\}_{j=1}^{n} \subset [0,1]$ and the discrete*

rectangular random variable is modelled by $X^R : \mathcal{S} \longrightarrow [0,1]$. *For given* n, *the probability density function of the* ***discrete rectangular distribution,*** *also called the* ***discrete uniform distribution,*** *is defined in the first case on* $\left\{\frac{j}{n}\right\}_{j=1}^{n}$ *by:*

$$f_R(j/n) = 1/n,\ j = 1, 2, .., n, \tag{1.4}$$

and analogously in the second case. In effect, this random variable splits $\mathcal{S}$ *into* n *sets of equal measure:*

$$\lambda(S_j) = 1/n,$$

where $\cup S_j = \mathcal{S}$ *and* $S_j \equiv X^{-1}(j/n)$ *in the first case, and analogously in the second.*

By rescaling this distribution can be supported on any interval $[a, b]$, *defining* $Y^R = (b-a)X^R + a$.

2. ***Binomial Distribution****: For a given* p, $0 < p < 1$, *the* ***standard binomial*** *random variable is defined:* $X_1^B : \mathcal{S} \longrightarrow \{0, 1\}$, *where the associated p.d.f. is defined:* $f(1) = p$, $f(0) = p' \equiv 1 - p$. *This is often expressed:*

$$X_1^B = \begin{cases} 1, & \Pr = p, \\ 0, & \Pr = p', \end{cases}$$

or to emphasize the associated p.d.f.:

$$f_{B_1}(j) = \begin{cases} p, & j = 1, \\ p', & j = 0. \end{cases} \tag{1.5}$$

A simple application for this random variable is as a model for a single coin flip. So $\mathcal{S} = \{H, T\}$, *a probability measure is defined on* $\mathcal{S}$ *by:* $\lambda(H) = p$, *and* $\lambda(T) = p'$, *and the random variable defined by* $X_1^B(H) \equiv 1$ *and* $X_1^B(T) \equiv 0$. *This random variable is sometimes referred to as a* ***Bernoulli trial****, and the associated d.f. as the* ***Bernoulli distribution*** *after* ***Jakob Bernoulli*** *(1654 – 1705).*

This standard formulation is then transformed to a ***shifted standard binomial*** *random variable:* $Y_1^B = b + (a-b)X_1^B$, *which is defined:*

$$Y_1^B = \begin{cases} a, & \Pr = p, \\ b, & \Pr = p', \end{cases}$$

where the example of $b = -a$ is common in discrete stock price modelling for example.

Similarly, this model can be extended to accommodate sample spaces of n-coin flips, producing the ***general binomial*** *random variable with two parameters, p and $n \in \mathbb{N}$. That is, $\mathcal{S} = \{(F_1F_2...F_n) \mid$ each $F_j = H$ or $T\}$, and X_n^B is defined as the "head counting" random variable:*

$$X_n^B(F_1F_2...F_n) = \sum\nolimits_{j=1}^{n} X_1^B(F_j).$$

It is apparent that X_n^B assumes values: $0, 1, 2, ..., n$, and that using a standard combinatorial analysis the associated probabilities are given for $j = 0, 1, .., n$ by:

$$X_n^B = \Big\{ \ j, \ \Pr = \tbinom{n}{j} p^j (1-p)^{n-j},$$

or to emphasize the associated p.d.f.:

$$f_{B_n}(j) = \binom{n}{j} p^j (1-p)^{n-j}, \quad j = 0, 1, 2, ..., n. \tag{1.6}$$

Recall that $\binom{n}{j}$ denotes the ***binomial coefficient*** *defined by:*

$$\binom{n}{j} = \frac{n!}{(n-j)!j!}, \tag{1.7}$$

where by convention, $0! = 1$. This expression is sometimes denoted ${}_nC_j$ and read, "n choose j." The name "binomial coefficient" follows from the expansion of a binomial, $a + b$, raised to the power n, producing the ***binomial theorem****:*

$$(a+b)^n = \sum\nolimits_{m=0}^{n} \binom{n}{m} a^m b^{n-m}. \tag{1.8}$$

3. ***Geometric Distribution:*** *For a given p, $0 < p < 1$, the* ***geometric distribution*** *is defined on the nonnegative integers and its p.d.f. is given by:*

$$f_G(j) = p(1-p)^j, \ j = 0, 1, 2, .. \tag{1.9}$$

and thus by summation:

$$F_G(j) = 1 - (1-p)^{j+1}, \ j = 0, 1, 2, .. \tag{1.10}$$

This distribution is related to the standard binomial distribution in a natural way. The underlying sample space can be envisioned as the collection of all coin-flip sequences which terminate on the first H. So:

$$\mathcal{S} = \{H, TH, TTH, TTTH, ...\},$$

and the random variable X^G is defined as the number of flips before the first H. Consequently, $f_G(j)$ above is the probability in $\mathcal{S}$ of the sequence of j-Ts and then 1-H. That is, the probability that the first H occurs after j-Ts. Of course, $f_G(j)$ is indeed a p.d.f. in that $\sum_{j=0}^{\infty} p(1-p)^j = 1$ as is verified noting that $\sum_{j=0}^{\infty}(1-p)^j$ is a geometric summation.

The geometric distribution is sometimes parametrized as:

$$f_{G'}(j) = p(1-p)^{j-1},\ j = 1, 2, ..,$$

and then represents the probability of the first head in a coin flip sequence appearing on flip j. These representations are conceptually equivalent, but mathematically distinct due to the shift in domain.

One way of generalizing the geometric distribution is to allow the probability of a head to vary with the sequential number of the coin flip. This is the basic model in all financial calculations relating to payments ***contingent on death or survival****, as well as to various other vitality-based outcomes. Specifically, if*

$$\Pr[H \mid kth\ flip] = p_k,$$

then with a simplifying change in notation to exclude the case $j = 0$, a ***generalized geometric distribution*** *can be defined by:*

$$f_{GG}(j) = p_j \prod_{k=1}^{j-1}(1-p_k), \quad j = 1, 2, 3, ..., \tag{1.11}$$

where $f_{GG}(j)$ is the probability of the first head appearing on flip j. By convention, $\prod_{k=1}^{0}(1-p_k) \equiv 1$ when $j = 1$. Of course, if $p_k = p > 0$ for all k, then $f_G(j)$ is a p.d.f. as noted above. With non-constant probabilities this conclusion is also true with a small restriction. Specifically, if $0 < a \leq p_k \leq b < 1$ for all j, then the summation is finite since $f_{GG}(j) < b(1-a)^{j-1}$ and thus $\sum_{j=1}^{\infty} f_{GG}(j)(j) < b/a$ by a geometric series summation. The details are left to the interested reader, or see Reitano pp. 314 – 319.

4. ***Negative Binomial Distribution:*** *The name of this distribution calls out yet another connection to the binomial distribution, and here we generalize the idea behind the geometric distribution. There, $f_G(j)$ was defined as the probability of j-Ts before the first H. The negative binomial, $f_{NB}(j)$ introduces another parameter, k, and is defined as the probability of j-Ts before the kth-H. So when $k = 1$, the negative binomial is the same as the geometric. The p.d.f. is then defined with parameters p and $k \in \mathbb{N}$ as follows:*

$$f_{NB}(j) = \binom{j+k-1}{k-1} p^k (1-p)^j, \quad j = 0, 1, 2, .. \tag{1.12}$$

This formula can be derived analogously to that for the geometric by considering in the sample space of all coin flip sequences, those which are terminated on the occurrence of the kth-H. The probability of any such sequence with j-Ts and k-Hs is then $p^k(1-p)^j$. Next we must determine the number of such sequences in the sample space. First off, since every such sequence terminates with an H, there are only the first $j + k - 1$ positions that need to be addressed. Each such sequence is then determined by the placement of the first $(k-1)$-Hs, and so the total count of these sequences is $\binom{j+k-1}{k-1}$. Multiplying the probability and the count, we have 1.12.

5. ***Poisson Distribution:*** *The Poisson distribution is named for* ***Siméon-Denis Poisson*** *(1781 – 1840) who discovered this p.d.f. and studied its properties. This distribution is characterized by a single parameter $\lambda > 0$, and is defined on the nonnegative integers by:*

$$f_P(j) = e^{-\lambda}\lambda^j/j!, \; j = 0, 1, 2, ... \tag{1.13}$$

That $\sum_{j=0}^{\infty} f_P(j) = 1$ is an application of the Taylor series expansion for e^λ :

$$e^\lambda = \sum_{j=0}^{\infty} \lambda^j/j!$$

One important application of the Poisson distribution is provided by the ***Poisson Limit theorem****, discussed in chapter 5 on Limit Theorems. This result states that the Poisson distribution provides a good approximation to the binomial distribution when the binomial parameter p is "small" and n is "large." See proposition 5.5. In addition, the binomial probabilities in 1.6 can be approximated by the Poisson*

probabilities above, in that with p "small" and n "large," and $\lambda = np$:

$$\binom{n}{j} p^j (1-p)^{n-j} \simeq e^{-np}(np)^j/j! \tag{1.14}$$

Another important property of the Poisson distribution is that it is the unique p.d.f. which characterizes "arrivals" during a given period of time under reasonable and frequently encountered assumptions. For example, the model might be one of automobile arrivals at a stop light; or telephone calls to a switchboard; or internet searches to a server; or radio-active particles to a Geiger counter; or insurance claims of any type (injuries, deaths, automobile accidents, etc.) from a large group of policyholders; or defaults from a large portfolio of loans and bonds; etc. For this result and the precise meaning of the approximation in 1.14, including the meaning of p "small" and n "large," see also proposition 7.51 in Reitano and the subsequent discussion.

1.2.2 Absolutely Continuous Distribution Functions

At the other end of the spectrum from discrete probability theory is what is sometimes called **continuous probability theory,** for which $F_{SLT}(x) = F_{SN}(x) = 0$ and so

$$F(x) \equiv F_{AC}(x), \tag{1.15}$$

is absolutely continuous. Hence by proposition 3.61 of book 3, $F'(x)$ exists almost everywhere, and

$$F(x) = (\mathcal{L}) \int_{-\infty}^{x} F'(y)dy.$$

Then $f(x) \equiv F'(x)$ is a **probability density function (p.d.f.) associated with the distribution function** F, which is defined in general as any measurable function f for which:

$$F(x) = (\mathcal{L}) \int_{-\infty}^{x} f(y)dy.$$

In the general case density functions are not unique in that any $f(x)$ with $f(x) = F'(x)$ a.e. is also a density function associated with F. That is because if $f(x)$ is a density function associated with F, all that can be said is that:

$$F(x) = \int_{-\infty}^{x} f(y)dy, \qquad F'(x) = f(x) \text{ a.e.} \tag{1.16}$$

For such absolutely continuous distribution functions, it then also follows that for any a, b :

$$\mu\{X^{-1}(a,b]\} \equiv F(b) - F(a) = \int_a^b f(y)dy.$$

Hence $\mu\{X^{-1}[a]\} = 0$ for all a and so $\mu\{X^{-1}(a,b]\}$ is the same for closed, open or semi-open intervals.

In many applications $F_{AC}(x)$ is in fact modelled to be **continuously differentiable** so $f(x) \equiv F'(x)$ is continuous and uniquely defined as the only continuous density function associated with F . A random variable $X : \mathcal{S} \to \mathbb{R}$ with such a distribution function is said to have a **continuous density function.** One then has from the standard version of the fundamental theorem of calculus that for all x :

$$F(x) = \int_{-\infty}^{x} f(y)dy, \qquad F'(x) = f(x). \tag{1.17}$$

There are many distributions with continuous density functions used in finance and other applications, but the most common are the uniform, exponential, gamma, beta, normal and lognormal. The Cauchy distribution will be introduced in the section Examples of Moments and M.G.F.s. In addition, other such examples are found with the **extreme value distributions** discussed in chapter 9 of book 2 as well as chapter 6 below.

It will be noted that we do not require a continuous density function to be continuous on $\mathbb{R}$, but only on its domain of definition.

Example 1.6 1. ***Continuous Uniform Distribution****: Perhaps the simplest continuous probability density that can be imagined is one which assumes the same value on every sample point. The domain of this distribution is arbitrary, though of necessity bounded, and is conventionally denoted as the interval* $[a, b]$*. The p.d.f. of the* ***continuous uniform distribution,*** *sometimes called the* ***continuous rectangular distribution*** *is defined on* $[a, b]$ *by the density function:*

$$f_U(x) = 1/(b-a), \quad x \in [a,b], \tag{1.18}$$

and $f_U(x) = 0$ *otherwise. This distribution is called "uniform" because if* $a \leq s < t \leq b$ *and* $X_U : \mathcal{S} \to \mathbb{R}$ *denotes the underlying random variable, then*

$$F_U(t) - F_U(s) = \lambda\{X_U^{-1}[s,t]\} = (t-s)/(b-a).$$

This probability is translation invariant within $[a, b]$, thus justifying the uniformity label.

This p.d.f. is important to a large degree because of the book 2 results in propositions 4.5 and 4.8 which state that for any random variable, X, the distribution function of the random variable $Y \equiv F(X)$ is uniformly distributed on $[0, 1]$ if and only if F is continuous. Of course, the random variable Y is the composition of $X : \mathcal{S} \to \mathbb{R}$ with $F : \mathbb{R} \to [0, 1]$. More generally, in all cases if Y is a uniformly distributed random variable on $(0, 1)$, the random variable $F^*(Y)$ has distribution function F, where F^* denotes the left continuous inverse function. This result can be used to generate independent random variates with distribution function $F(x)$, the theory of which was developed in chapter 4 of book 2, and the applications will be seen in chapter 4 below.

To summarize these results, recall the definition of left continuous inverse which is applicable to any increasing function:

Definition 1.7 *Let $F(x)$ be an increasing function, $F : \mathbb{R} \to \mathbb{R}$. The* ***"left-continuous" inverse of*** *F, denoted F^*, is defined by:*

$$F^*(y) = \inf\{x | y \leq F(x)\}. \tag{1.19}$$

By convention, if $\{x|F(x) \geq y\} = \emptyset$ then we define $F^*(y) = \infty$. Similarly, if $\{x|F(x) \geq y\} = \mathbb{R}$, then by 1.19, $F^*(y) = -\infty$.

The summary results are then:

Proposition: *Let $(\mathcal{S},\mathcal{E},\lambda)$ be given, and $X : (\mathcal{S}, \mathcal{E}, \lambda) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}), m_L)$ a random variable with distribution function $F(x)$, left continuous inverse $F^*(y)$, and m_L Lebesgue measure. Then:*

- *$F(X) : (\mathcal{S}, \mathcal{E}, \lambda) \to ((0, 1), \mathcal{B}((0, 1), m_L)$ defined by $F(X)(s) = F(X(s))$ is a random variable on $\mathcal{S}$, with distribution function $F_{F(X)}(y)$, satisfying:*

  $$F_{F(X)}(y) \leq y.$$

  Further, $F_{F(X)}(y) = y$ if and only if F is continuous.

- *$F^* : ((0, 1), \mathcal{B}(0, 1), m_L) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}), m_L)$ is a random variable on $(0, 1)$, which has distribution function $F(x)$.*

- *If $\{Y_j\}_{j=1}^n$ are independent, continuous uniformly distributed random variables, then $\{X_j\}_{j=1}^n \equiv \{F^*(Y_j)\}_{j=1}^n$ are independent random variables with distribution function $F(x)$.*
- *If $F(x)$ is continuous and $\{X_j\}_{j=1}^n$ are independent random variables with distribution function $F(x)$, then $\{Y_j\}_{j=1}^n \equiv \{F(X_j)\}_{j=1}^n$ are independent, continuous uniformly distributed random variables.*

The significance of these results is that one can convert a uniformly distributed sample $\{Y_j\}$ from $[0,1]$, into a random sample of the $\{X_j\}$ random variables by defining:

$$X_j = F^*(Y_j),$$

with $F^*(Y_j)$ defined as in 1.19. And note that despite being defined as an infimum, the value of $F^*(Y_j)$ produced by this definition is truly in the domain of the random variable X because $F(x)$ is right continuous.

2. ***Exponential and Gamma Distributions:*** *The **exponential density function** is defined with parameter $\lambda > 0$:*

$$f_E(x) = \lambda e^{-\lambda x}, \quad x \geq 0, \tag{1.20}$$

and $f_E(x) = 0$ for $x < 0$. The associated distribution function is calculated to be:

$$F_E(x) = 1 - \exp(-\lambda x), \quad x \geq 0, \tag{1.21}$$

and $F_E(x) = 0$ for $x < 0$.

*The **gamma distribution** is a two parameter generalization of the exponential, with $\alpha > 0, \lambda > 0$, and density function $f_G(x)$ defined by*

$$f_\Gamma(x) = \lambda^\alpha x^{\alpha-1} e^{-\lambda x}/\Gamma(\alpha), \quad x \geq 0, \tag{1.22}$$

*and $f_\Gamma(x) = 0$ for $x < 0$. When $0 < \alpha < 1$, $f_\Gamma(x)$ is unbounded at $x = 0$ but is integrable. The **gamma function,** $\Gamma(\alpha)$, is defined by*

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx, \tag{1.23}$$

and thus $\int_0^\infty f_\Gamma(x)dx = 1$ with a change of variable. An integration by parts shows that for $\alpha > 0$ the gamma function satisfies

$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha), \tag{1.24}$$

and since $\Gamma(1) = 1$ *this function generalizes the factorial function in that for any integer* n,

$$\Gamma(n) = (n-1)! \tag{1.25}$$

When the parameter $\alpha = k$, *a positive integer, then by 1.25 the probability density function becomes*

$$f_\Gamma(x) = \lambda^k x^{k-1} e^{-\lambda x}/(k-1)!, \quad x \geq 0.$$

The associated distribution function is:

$$F_\Gamma(x) = e^{-\lambda x} \sum_{j=k}^{\infty} (\lambda x)^j / j! \tag{1.26}$$

which can be verified by differentiation, that $F'_\Gamma(x) = f_\Gamma(x)$ *since* $F_\Gamma(0) = 0$.

The exponential and gamma distributions are important in the generation of ***ordered random samples.*** *See chapter 4 below.*

Remark 1.8 *When a random variable* Y *has a gamma distribution with* $\lambda = 1/2$ *and* $\alpha = n/2$, *it is said to have a* ***Chi-squared distribution with*** n ***degrees of freedom,*** *which is sometimes denoted* $\chi^2_{n\ d.f.}$. *See example 1.14.*

3. ***Beta Distribution****: The* ***beta distribution*** *contains two shape parameters,* $v > 0, w > 0$, *and is defined on the interval* $[0, 1]$ *by the density function:*

$$f_\beta(x) = x^{v-1}(1-x)^{w-1}/B(v, w). \tag{1.27}$$

Here the ***beta function*** $B(v, w)$ *is defined by a definite integral which in general requires numerical evaluation:*

$$B(v, w) = \int_0^1 y^{v-1}(1-y)^{w-1} dy. \tag{1.28}$$

By definition, therefore, $\int_0^1 f_\beta(x)dx = 1$.

If v *or* w *or both parameters are less than* 1, *the beta density is unbounded at* $x = 0$ *or* $x = 1$ *or both, but this integral converges as an improper integral discussed in book* 3 *because the exponents of both the* x *and* $1 - x$ *terms exceed* -1. *If both parameters are greater than* 1 *this density function is* 0 *at the interval endpoints, and has a unique*

maximum at $x = (v-1)/(v+w-2)$. *When both parameters equal* 1, *this distribution reduces to the continuous uniform distribution on* $[0,1]$.

The beta function is closely related to the gamma function above. Substituting $x = y/(1-y)$ *into the integral in 1.28:*

$$B(v,w) = \int_0^\infty \frac{x^{v-1}}{(1+x)^{v+w}} dx.$$

Letting $\alpha = v + w$ *in 1.23 obtains by substitution* $x \to (1+x)y$*:*

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx, \tag{1.29}$$

$$1/(1+x)^{v+w} = \int_0^\infty y^{v+w-1} e^{-(1+x)y} dy/\Gamma(v+w),$$

and so

$$B(v,w) = \int_0^\infty y^{v+w-1} e^{-y} \left(\int_0^\infty e^{-xy} x^{v-1} dx \right) dy/\Gamma(v+w).$$

A final substitution $z = xy$ *and two applications of 1.23 produces the identity:*

$$B(v,w) = \frac{\Gamma(v)\Gamma(w)}{\Gamma(v+w)}. \tag{1.30}$$

Thus for integer n *and* m, *it follows from 1.25 that*

$$B(n,m) = \frac{(n-1)!\,(m-1)!}{(n+m-1)!}. \tag{1.31}$$

4. ***Normal Distribution****: The* ***normal distribution*** *is defined on* $(-\infty, \infty)$, *depends on a location parameter* $\mu \in \mathbb{R}$ *and a scale parameter* $\sigma > 0$, *and is defined by the probability density function:*

$$f_N(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-(x-\mu)^2 / 2\sigma^2 \right), \tag{1.32}$$

where $\exp A \equiv e^A$ *to simplify notation. In many books the parametrization is defined in terms of* μ *and* σ^2, *where in 1.32* σ *is taken as the positive square root. When* $\mu = 0$ *and* $\sigma = 1$ *this is known as the* ***standard normal distribution,*** *or* ***unit normal distribution,*** *and denoted* $\phi(x)$:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2 \right). \tag{1.33}$$

A change of variables in the associated integral shows that if a random variable X is normally distributed with parameters μ and σ^2, then $(X-\mu)/\sigma$ has a standard normal distribution. Conversely, if X is standard normal, then $\sigma X+\mu$ is normally distributed with parameters μ and σ^2.

Perhaps the greatest significance of the normal distribution is that it can be used as an approximating distribution to the distribution of sums and averages of a random sample of "scaled" random variables under relatively mild assumptions. When applied to approximate the binomial distribution, this result is called the ***De Moivre-Laplace Theorem****, named for* ***Abraham de Moivre*** *(1667 – 1754) who demonstrated the special case of $p = 1/2$, and* ***Pierre-Simon Laplace*** *(1749 – 1827) who many years later generalized to all p, $0 < p < 1$. In the most general case this result is known as the the* ***Central Limit Theorem****. See chapter 5 below on Limit Theorems.*

5. ***Lognormal Distribution****: The* ***lognormal distribution*** *is defined on $[0,\infty)$, depends on a location parameter $\mu \in \mathbb{R}$ and a shape parameter, $\sigma > 0$, and unsurprisingly is intimately related to the normal distribution above. However, to some the name "lognormal" appears to be opposite of the relationship that exists. Stated one way, a random variable X is lognormal with parameters (μ, σ^2) if $X = e^Z$ where Z is normal with the same parameters. So X can be understood as an exponentiated normal. Stated another way, a random variable X is lognormal with parameters (μ, σ^2) if $\ln X$ is normal with the same parameters. The name comes from the second statement, in that the log of a lognormal variate is a normal variate.*

   The probability density function of the lognormal is defined as follows, again using $\exp A \equiv e^A$ to simplify notation:

$$f_L(x) = \frac{1}{\sigma x\sqrt{2\pi}} \exp\left(-\left(\ln x - \mu\right)^2 / 2\sigma^2\right). \tag{1.34}$$

Remark 1.9 *By change of variable, $\int_0^\infty f_L(x)dx = 1$ if and only if $\int_{-\infty}^\infty f_N(x)dx = 1$. While there are no elementary proofs of the latter identity, there is clever proof used by virtually everyone. It involves embedding the integral $\int_{-\infty}^\infty f_N(x)dx$ into a 2-dimensional integral and applying a change of variables which allows direct calculation. This derivation also requires the ability to move back and forth between 2-dimensional and iterated integrals, another highly nontrivial result called Fubini's theorem. The technical results needed for this*

derivation will be developed in book 5 and applied in book 6 to the result at hand.

1.2.3 Mixed Distribution Functions

So-called **mixed distribution functions,** for which $F_{SLT}(x) = 0$, are commonly encountered in finance and defined:

$$F(x) \equiv F_{SN}(x) + F_{AC}(x), \tag{1.35}$$

where again it is assumed that $F'_{AC}(x) = f(x)$ is continuous. In other words, such a distribution function represents a random variable with both continuously distributed and discrete parts. For example, in chapter 1 of book 2 was exemplified a loan portfolio and of particular interest was the modeling of default losses. For a portfolio of n loans, if X denotes the number of defaults then $X : \mathcal{S} \to \{0, 1, 2, ..., n\}$ is discrete, while if $Y : \mathcal{S} \to \mathbb{R}$ denotes the dollars of loss, then typically the model for Y will be "mixed." This is because with fixed default probability p, say, the probability of no defaults,

$$\Pr\{Y = 0\} = \Pr\{X = 0\} = (1 - p)^n,$$

and this will typically be quite large compared to $\Pr\{0 < Y \leq \epsilon\}$, which is the probability of one or several defaults but very low loss amounts. So at the minimum, it is expected that $F(y)$ will have a discrete part at $y = 0$ and then a continuous or mixed distribution for $y > 0$ depending on how losses are modeled when defaults occur.

For example, if we assume that on default the loss given default is uniformly distributed on $[0, L_j]$ where L_j denotes the loan amount on the jth bond, then one expects a continuous distribution on $y > 0$. This is because if $l > 0$ denotes total losses, then $F(y)$ will be continuous from the left at l, and hence continuous at l since all distribution functions are continuous from the right. This left continuity follows from the observation, which was formalized in the so-called **law of total probability** in proposition 1.35 of book 2, that if L denotes total losses and $\epsilon < l$,

$$\Pr\{l - \epsilon < L \leq l\} = \sum\nolimits_{j=1}^{n} \Pr\{l - \epsilon < L \leq l \mid j \text{ defaults}\} \Pr\{j \text{ defaults}\}. \tag{1.36}$$

Given the assumption of uniformly distributed losses, $\Pr\{l - \epsilon < L \leq l \mid j \text{ defaults}\}$ is defined in terms of a fixed sum of j continuously distributed

random variables, and so it is continuously distributed as motivated by exercise 1.10 below and provable with the methods of book 6. So as $\epsilon \to 0$,

$$\Pr\{l - \epsilon < L \leq l \mid j \text{ defaults}\} \to 0,$$

and hence

$$\Pr\{l - \epsilon < L \leq l\} \to 0.$$

Since

$$\Pr\{l - \epsilon < L \leq l\} = F(l) - F(l - \epsilon),$$

this demonstrates left continuity and hence continuity of $F(y)$ for $y > 0$.

Mixed distributions similarly arise in other financial contexts related to various insurance claim amount distributions, since these are modeled much like loan default losses.

Exercise 1.10 *Show that if X and Y are continuously distributed random variables, then the distribution function for $Z \equiv X + Y$ is given:*

$$F_Z(z) = \int f_X(x) F_Y(z - x) dx. \qquad (*)$$

Hint: If $\{x_j\}$ is a partition of the x-axis, and $\widehat{x}_j \in [x_j, x_{j+1}]$, then:

$$F_Z(z) \approx \sum [F_X(x_{j+1}) - F_X(x_j)] F_Y(z - \widehat{x}_j).$$

Use the mean value theorem for integrals to derive $F_X(x_{j+1}) - F_X(x_j) = f_X(\tilde{x}_j)\Delta x$ for some $\tilde{x}_j \in [x_j, x_{j+1}]$. Choosing $\widehat{x}_j = \tilde{x}_j$ and letting the partition mesh size $\mu \to 0$, $(*)$ follows as a Riemann integral. See also the section Distribution Functions of Sums and Ratios of Random Variables below.

This result is enough to derive that the distribution function for the sum of loan losses given two defaults is continuously distributed, for example.

Remark 1.11 *The result in $(*)$ implies that Z is continuously distributed and:*

$$f_Z(z) = \int f_X(x) f_Y(z - x) dx,$$

but to derive this requires a differentiation "under the integral." Alternatively, one can integrate $f_Z(z)$ to yield the formula above for $F_Z(z)$, but this then requires a manipulation of iterated integrals, which is Fubini's theorem of book 5.

1.2.4 Other Distribution Functions

While discrete, continuously distributed, and mixed distributions are most common in finance in practice, it is important to be mindful of the above result that in general, distribution functions can be more general in two ways:

1. Although the $F_{AC}(x)$ component is typically modelled as a continuously differentiable function, so that $f_{AC}(x) \equiv F'_{AC}(x)$ is assumed continuous and hence is unique among continuous densities, the theory only assures that $F'_{AC}(x)$ exists almost everywhere, is measurable and Lebesgue integrable, and can be used to recover $F_{AC}(x)$ by the fundamental theorem, version 1. Any assumption that implies additional properties of $F'_{AC}(x)$ is, by definition, an assumption.

2. In general, distribution functions can have singular components, $F_{SN}(x)$, which can only be ignored when they are assumed to not exist.

1.3 Distribution Functions of Transformed Random Variables

Given a random variable X on a probability space $(\mathcal{S},\mathcal{E},\lambda)$ with distribution function $F(x)$, we are sometimes interested in evaluating the distribution function of $Y \equiv g(X)$ for Borel measurable function $g : \mathbb{R} \to \mathbb{R}$. By proposition 3.33 of book 1, for such g the composite functions $g(X)$ is a λ-measurable function on $\mathcal{S}$ and is hence a random variable. By definition, $F_Y(y) \equiv \lambda\left[Y^{-1}(-\infty, y]\right]$ and so:

$$F_Y(y) = \lambda\left[X^{-1}\left(g^{-1}(-\infty, y]\right)\right], \tag{1.37}$$

where $g^{-1}(-\infty, y] \equiv \{z|g(z) \leq y\}$.

In two simple cases, g^{-1} is easy to work with.

- **With g an increasing function**: Then $g^{-1}(-\infty, y] = (-\infty, g^{-1}(y)]$ and thus:

  $$F_Y(y) = \lambda\left[X^{-1}(-\infty, g^{-1}(y)]\right],$$

  or equivalently:

  $$F_{g(X)}(y) = F_X(g^{-1}(y)). \tag{1.38}$$

  When F is absolutely continuous and has an associated density function f, and g is continuously differentiable with $g'(x) \neq 0$ for all x,

then by differentiation F_Y will have an associated continuous density function given by:

$$f_{g(X)}(y) = f_X(g^{-1}(y))/g'(g^{-1}(y)). \tag{1.39}$$

- **With g is a decreasing function**: Because now $g^{-1}(y, \infty) = (-\infty, g^{-1}(y))$, using complementarity:

$$F_Y(y) = 1 - \lambda\left[X^{-1}(g^{-1}(y, \infty))\right],$$

and thus:

$$F_{g(X)}(y) = 1 - F_X(g^{-1}(y)^-), \tag{1.40}$$

where $F(x^-)$ denotes the left limit of F at x.

For absolutely continuous distribution functions which are thus continuous, $F(x^-) = F(x)$. So if F has an associated density function f, and g is continuously differentiable with $g'(x) \neq 0$ for all x, then by differentiation F_Y has an associated continuous density function given by:

$$f_{g(X)}(y) = -f_X(g^{-1}(y))/g'(g^{-1}(y)). \tag{1.41}$$

Remark 1.12 *A formula that is often given for absolutely continuous distribution functions, F, and monotonic g with $g^{-1}(y)$ continuously differentiable, is:*

$$f_{g(X)}(y) = f_X(g^{-1}(y))\left|\frac{dg^{-1}(y)}{dy}\right|. \tag{1.42}$$

If $g'(x) \neq 0$, then $g^{-1}(y)$ is differentiable with $\frac{dg^{-1}(y)}{dy} = 1/g'(g^{-1}(y))$ as above. Comparing with 1.39 and 1.41 obtains that the $\pm$ in the respective formulas simply yields $\left|\frac{dg^{-1}(y)}{dy}\right|$.

Example 1.13 *1. If X is normally distributed as in 1.32 and $Y = e^X$, then by 1.39 we find that*

$$f_{g(X)}(y) = \frac{1}{\sigma y\sqrt{2\pi}} \exp\left(-(\ln y - \mu)^2/2\sigma^2\right),$$

which is the density function of the lognormal distribution in 1.34.

2. Similarly, one shows that if X is lognormally distributed as in 1.34, then $Y = \ln X$ is normally distributed as in 1.32.

The above approach can sometimes be adapted in situations where g is not strictly monotonic.

Example 1.14 ***Normal Squared is Gamma:*** *If X is standard normal, consider $Y = X^2$. Here g is not increasing on the range of X, but we can explicitly evaluate F_Y by 1.37. Because $\lambda\left[X^{-1}(x)\right] = 0$ for all x we can be casual about interval endpoints:*

$$\begin{aligned} F_Y(y) &= \lambda\left[X^{-1}([-\sqrt{y}, \sqrt{y}])\right] \\ &= \lambda\left[X^{-1}((-\infty, \sqrt{y}])\right] - \lambda\left[X^{-1}((-\infty, -\sqrt{y}])\right] \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}). \end{aligned}$$

The associated density function can then be calculated, and using the symmetry of f_X we derive that for $y > 0$:

$$\begin{aligned} f_Y(y) &= f_X(\sqrt{y})/\sqrt{y} \\ &= \frac{1}{\sqrt{2\pi}} y^{-1/2} \exp(-y/2). \end{aligned}$$

Comparing this to 1.22, we see that Y has a gamma density with $\lambda = \alpha = 1/2$ if

$$\Gamma(1/2) = \sqrt{\pi}. \tag{1.43}$$

This is indeed verified by substitution into the Riemann integral that defines $\Gamma(t)$, and recalling that the standard normal density integrates to 1. Specifically, letting $y = \sqrt{2x}$:

$$\begin{aligned} \Gamma(1/2) &= \int_0^\infty x^{-1/2} e^{-x} dx \\ &= \sqrt{2} \int_0^\infty e^{-y^2/2} dy \\ &= \sqrt{\pi}. \end{aligned}$$

Notation 1.15 *As noted in remark 1.8, when a random variable Y has a gamma distribution with $\lambda = 1/2$ and $\alpha = n/2$, it is said to have a* ***Chi-squared distribution with*** *n* ***degrees of freedom,*** *which is sometimes denoted $\chi^2_{n\ d.f.}$. Thus if X is standard normal, then X^2 is $\chi^2_{1\ d.f.}$. In general, the density function of the Chi-squared distribution with n degrees of freedom is given by 1.22 with $\lambda = 1/2$ and $\alpha = n/2$, defined on $x \geq 0$ as:*

$$f_{\chi^2_{n\ d.f.}}(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}. \tag{1.44}$$

We will see in example 3.59 that for n independent gammas $\{Y_i\}_{i=1}^n$ *with common parameter* λ *and parameters* $\{\alpha_i\}_{i=1}^n$*, that the sum* $\sum_{i=1}^n Y_i$ *is gamma with parameters* λ *and* $\sum_{i=1}^n \alpha_i$*. In example 1.18 we derive this result for the special case when the* Y_i *are exponential and hence* $\alpha_i = 1$*. The example 3.59 result then implies that if* $\{X_i\}_{i=1}^n$ *are independent standard normals, then* $\sum_{i=1}^n X_i^2$ *is* $\chi^2_{n\ d.f.}$ *because* $\{X_i^2\}_{i=1}^n$ *are independent by example 3.58 of book 2.*

Jumping ahead a bit to chapter 3, this result provides the theoretical basis for estimating the unknown "variance" σ^2 *of a normal distribution with known "mean"* μ_0*, based on a sample:* $\{X_i\}_{i=1}^n$*. Then in this case:*

$$\sum_{i=1}^n \left(\frac{X_i - \mu_0}{\sigma}\right)^2$$

is $\chi^2_{n\ d.f.}$*, and so confidence intervals for this chi-squared distribution can be translated to confidence intervals for* σ^2*.*

Remark 1.16 *In this section we introduced special cases of the distribution function of a transformed random variable. See book 6 for the generalization of these results and to transformed random vectors using the integration theory of book 5.*

1.4 Distribution Functions of Sums and Ratios of RVs

Given distribution functions of certain random variables, it is often of interest to determine the distribution function of the sum of these variables, or of certain ratios. We develop some of these ideas in this section.

1.4.1 Sums of Independent Random Variables

In this section we introduce a calculation that will be formalized in book 6 using the integration tools of book 5. This calculation is the **convolution of functions**, and was motivated in exercise 1.10. The question of interest is, given random variables X and Y on a probability space $(\mathcal{S},\mathcal{E},\lambda)$ with respective distribution functions F_X and F_Y, what is the distribution function of the random variable $Z \equiv X + Y$? If X and Y have density functions, what is the density function of $X + Y$? In this section we will assume that X and Y have distribution functions which are either both

saltus functions, or both absolutely continuous with continuous densities, and generalize the theory in various ways in book 6. As above:

$$\text{Saltus:} \quad F(x) = \sum\nolimits_{y \le x} f(y);$$
$$\text{Absolutely Continuous:} \quad F(x) = \int_{-\infty}^{x} f(y)dy.$$

Distribution Functions of Sums

Let $F(x,y)$ be the joint distribution function of (X,Y), defined as in 3.11 of book 2 by:

$$F(x,y) = \lambda\left[\{s|X(s) \le x\} \bigcap \{s|Y(s) \le y\}\right],$$

and $f(x,y)$ a measurable joint density function so that the joint distribution function is given by:

$$F(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f(u,v)dudv.$$

Defining $Z = X + Y$:

$$\begin{aligned} F_Z(z) &\equiv \lambda\left[\{s|X(s) + Y(s) \le z\}\right] \\ &= \lambda\left[(X,Y)^{-1}\left[A_z\right]\right] \end{aligned}$$

where

$$\begin{aligned} A_z &\equiv \{(x,y)|x + y \le z\} \subset \mathbb{R}^2, \\ (X,Y)^{-1}\left[A_z\right] &= \{s|\,(X(s),Y(s)) \in A_z\}. \end{aligned}$$

We expect that in this case, and this will be assumed for now and formalized in book 6, that:

$$F_Z(z) = \iint_{A_z} f(x,y)dxdy.$$

This follows from the more general result that if $A \in \mathcal{B}(\mathbb{R}^2)$, then

$$\lambda\left[(X,Y)^{-1}(A)\right] = \iint_{A} f(x,y)dxdy. \tag{1.45}$$

In the discrete case the book 6 theory is not needed. First:

$$F(x,y) = \sum\nolimits_{u \le x}\sum\nolimits_{v \le y} f(u,v),$$

and it follows from absolute convergence of the summation that we can rearrange terms to produce:

$$F_Z(z) = \sum\nolimits_{(x,y)\in A_z} f(x,y).$$

In either case, if X and Y are independent, so that $F(x,y) = F_X(x)F_Y(y)$ by proposition 3.53 of book 2, the calculations needed for $F_Z(z)$ are significantly simplified. Indeed, since A_z can be parametrized as:

$$A_z \equiv \{(x,y)|y \in \mathbb{R},\ x \leq z - y\},$$

and we have in the case of independent random variables that:

$$F_Z(z) = \int_{-\infty}^{\infty}\int_{-\infty}^{z-y} f_X(x)f_Y(y)dxdy.$$

Assuming that integrals can be calculated iteratively (Fubini's theorem, book 5) and completing the dx-integral:

$$F_Z(z) = \int_{-\infty}^{\infty} F_X(z-y)f_Y(y)dy. \tag{1.46}$$

Analogously, in the discrete case:

$$F_Z(z) = \sum\nolimits_y F_X(z-y)f_Y(y). \tag{1.47}$$

By parametrizing

$$A_z \equiv \{(x,y)|x \in \mathbb{R},\ y \leq z - x\},$$

the roles of x and y are reversed in these formulas:

$$\begin{aligned} F_Z(z) &= \int_{-\infty}^{\infty} F_Y(z-x)f_X(x)dx, \\ F_Z(z) &= \sum\nolimits_x F_Y(z-x)f_X(x). \end{aligned}$$

Notation 1.17 *In the terminology and notation of book 5, the distribution function $F_Z(z)$ expressed as in 1.46 or 1.47 equals the* ***convolution of*** F_X ***and*** f_Y*, or equivalently by the above remark, the* ***convolution of*** f_X ***and*** F_Y*. Notationally:*

$$F_Z(z) = F_X * f_Y(z) = f_X * F_Y(z).$$

Convolutions are commutative, in the sense that $F_X * f_Y(z) = f_Y * F_X(z)$. *For example:*

$$\int_{-\infty}^{\infty} F_X(z-y) f_Y(y) dy = \int_{-\infty}^{\infty} F_X(y) f_Y(z-y) dy,$$

and similarly for 1.47, as a change of variables in the summation or integral verifies. This change of variables in the integral is to be formally justified in book 5 when such integrals are Lebesgue, while for summations this result requires no new theory.

Convolution of 3 *or more functions is then defined iteratively, by*

$$f * g * h(z) \equiv (f * g) * h(z) = f * (g * h)(z).$$

This definition implies the result that convolution is also associative, and this will be proved with the aid of Fubini's theorem in book 5.

Example 1.18 1. ***Sums of Exponentials are Gamma:*** *Let* X *and* Y *be independent and have exponential distributions as in 1.20 and with the same parameter* λ, *so* $f_E(x) = \lambda e^{-\lambda x}$ *and* $F_E(x) = 1 - e^{-\lambda x}$ *for* $x \geq 0$. *Then by 1.46, noting the range of integration in* y :

$$\begin{aligned} F_Z(z) &= \int_0^z \left(1 - e^{-\lambda(z-y)}\right) \lambda e^{-\lambda y} dy \\ &= 1 - e^{-\lambda z}(1 + \lambda z). \end{aligned}$$

This formula is satisfied for $z \geq 0$, *and hence by differentiation,*

$$f_Z(z) = \lambda^2 z e^{-\lambda z}$$

for $z \geq 0$. *By 1.25,* f_Z *is the gamma density function in 1.22 with parameters* λ *and* $\alpha = 2$.

Exercise 1.19 *Prove by induction that the sum of* k *independent exponentials with the same parameter* λ *has a gamma distribution with parameters* λ *and* $\alpha = k$. *This result will be further generalized in example 3.59 to the statement that sums of independent gamma random variables are gamma as long as they have a common* λ, *and then the resultant* α *parameter satisfies* $\alpha = \sum_i \alpha_i$.

2. ***Sums of Exponentials and the Poisson:*** *The fact from* 1, *that the sum of* k *independent exponentials with common parameter* λ *produces*

a Gamma with parameters λ and $\alpha = k$, motivates an interesting connection between sums of such exponentials and the Poisson distribution. Given such independent exponentials $\{X_j\}_{j=1}^{\infty}$, let $S_n = \sum_{j=1}^{n} X_j$ and define a new random variable N by:

$$N \equiv \max\{n | S_n \leq 1\}.$$

Then $N = n$ if and only if $S_n \leq 1 < S_{n+1}$, and hence $N \geq n$ if and only if $S_n \leq 1$. Recall the distribution function of S_n is given in 1.26, and since $\Pr[S_n \leq 1] = F_{S_n}(1)$, we obtain:

$$\Pr[N \geq n] = e^{-\lambda} \sum_{j=n}^{\infty} (\lambda)^j / j!.$$

Since $\Pr[N \geq n] = 1 - F(n-1)$, and $f(n) = F(n) - F(n-1)$:

$$\Pr[N = n] = e^{-\lambda} \lambda^n / n!.$$

In other words, N has a Poisson distribution with parameter λ.

Remark 1.20 *Generalizing the above, define $N_0 = 0$ and for $t > 0$:*

$$N_t \equiv \max\{n | S_n \leq t\}.$$

*It can then be shown analogously that for $t > 0$ that N_t has a Poisson distribution with parameter λt. Thus by generating independent exponentials, $\{X_j\}_{j=1}^{\infty}$, one can create a "stochastic process" N_t, $t \geq 0$, such that each N_t has a Poisson distribution with parameter λt. This is called a **Poisson process** and has other important properties. See section 23 of Billingsley for details.*

3. ***Sums of Binomials are Binomial:*** *Let X and Y be independent and have binomial distributions as in 1.6 with common parameter p, but with respective parameters n and m. Then by 1.47, again noting the range of summation in y :*

$$\begin{aligned} F_Z(z) &= \sum_{j=0}^{z} F_X(z-j) \binom{m}{j} p^j (1-p)^{m-j} \\ &= \sum_{j=0}^{z} \left[\sum_{k=0}^{z-j} \binom{n}{k} p^k (1-p)^{n-k} \right] \binom{m}{j} p^j (1-p)^{m-j} \\ &= \sum_{j=0}^{z} \sum_{k=0}^{z-j} \binom{n}{k} \binom{m}{j} p^{j+k} (1-p)^{n+m-(j+k)}. \end{aligned}$$

This double summation can now be rearranged. Define $i = j + k$, then $k = i - j$ and the k-sum becomes an i-sum from j to z. Reversing the resulting double summations:

$$\sum_{j=0}^{z}\sum_{k=0}^{z-j} = \sum_{j=0}^{z}\sum_{i=j}^{z} = \sum_{i=0}^{z}\sum_{j=0}^{i},$$

obtains:

$$F_Z(z) = \sum_{i=0}^{z}\sum_{j=0}^{i}\binom{n}{i-j}\binom{m}{j}p^i(1-p)^{n+m-i}.$$

Finally,

$$\sum_{j=0}^{i}\binom{n}{i-j}\binom{m}{j} = \binom{m+n}{i}$$

as can be verified as an exercise by evaluating and comparing the coefficients of x^i in the identity:

$$(1+x)^{n+m} = (1+x)^n(1+x)^m,$$

and so:

$$F_Z(z) = \sum_{i=0}^{z}\binom{m+n}{i}p^i(1-p)^{n+m-i}.$$

In other words, X has a binomial distribution with parameters p and $n + m$.

Exercise 1.21 *Prove by induction that the sum of N such independent binomials has a binomial distribution with parameters p and $n = \sum_{i=1}^{N} n_i$.*

Density Functions of Sums

Continuing with a little informality that will be rectified in book 6, we can identify the density functions for $Z = X + Y$ in the above special cases of absolutely continuous or discrete distribution functions. When F_X and F_Y are absolutely continuous with continuous densities, then

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y)f_Y(y)dy. \tag{1.48}$$

To see this, we only need to verify that $f_Z(z)$ gives rise to $F_Z(z)$ in the usual way. Assuming we can change the order of the integrals and

assuming for now that all densities are continuous and integrals are Riemann, we then have:

$$\begin{aligned}\int_{-\infty}^{z} f_Z(x)dx &= \int_{-\infty}^{z}\int_{-\infty}^{\infty} f_X(x-y)f_Y(y)dydx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{z} f_X(x-y)dxf_Y(y)dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{z-y} f_X(x)dxf_Y(y)dy \\ &= \int_{-\infty}^{\infty} F_X(z-y)f_Y(y)dy \\ &= F_Z(z).\end{aligned}$$

For discrete distribution functions:

$$f_Z(z) = \sum\nolimits_y f_X(z-y)f_Y(y). \tag{1.49}$$

To verify requires the reordering of summations and change of variables, but in this context no new theory is needed and it can be verified that:

$$\sum\nolimits_{x\le z} f_Z(x) = F_Z(z).$$

In summary, using the notation of convolutions,

$$f_Z(z) = f_X * f_Y(z),$$

in both cases.

1.4.2 Ratios of Random Variables

In this section we alter the question of the prior section somewhat. Given random variables X and Y on a probability space $(\mathcal{S},\mathcal{E},\lambda)$ with respective distribution functions F_X and F_Y and where Y has range $(0,\infty)$, what is the distribution function of the random variable $Z \equiv X/Y$?

Independent Random Variables

Continuing the informality of the previous section to be justified in book 6, let $F(x,y)$ be the joint distribution function of (X,Y) and $f(x,y)$ the joint

density function which is assumed continuous for simplicity. Then defining $Z = X/Y$:

$$\begin{aligned} F_Z(z) &\equiv \lambda\left[\{s|X(s)/Y(s) \leq z\}\right] \\ &= \lambda\left[(X,Y)^{-1}\left[B_z\right]\right], \end{aligned}$$

where

$$\begin{aligned} B_z &\equiv \{(x,y)|x/y \leq z\} \subset \mathbb{R}^2, \\ (X,Y)^{-1}\left[B_z\right] &= \{s|\left(X(s),Y(s)\right) \in B_z\}. \end{aligned}$$

As another application of 1.45 it is expected that:

$$F_Z(z) = \iint_{B_z} f(x,y)dxdy,$$

while in the discrete case with no new theory:

$$F_Z(z) = \sum\nolimits_{(x,y)\in B_z} f(x,y).$$

If X and Y are independent random variables, so that $F(x,y) = F_X(x)F_Y(y)$, then since B_z can be parametrized as:

$$B_z \equiv \{(x,y)|y \in (0,\infty),\ x \leq zy\},$$

it follows that in the case of independent random variables:

$$F_Z(z) = \int_0^\infty \int_{-\infty}^{zy} f_X(x)f_Y(y)dxdy,$$

and so

$$F_Z(z) = \int_0^\infty F_X(zy)f_Y(y)dy. \tag{1.50}$$

Analogously in the discrete case:

$$F_Z(z) = \sum\nolimits_y F_X(zy)f_Y(y). \tag{1.51}$$

When X and Y have density functions the density function for Z can also be derived. In the absolutely continuous case:

$$f_Z(z) = \int_0^\infty y f_X(zy)f_Y(y)dy, \tag{1.52}$$

and in the discrete case:

$$f_Z(z) = \sum\nolimits_y y f_X(zy) f_Y(y). \tag{1.53}$$

We formalize these manipulations of integrals in book 6 using the tools of book 5, but for discrete densities and summations, or, continuous density functions and Riemann integrals, the manipulations will be familiar. For example, assuming we can reorder integrals, a simple change of variables produces the result that $F_Z(z)$ in 1.50 is the distribution function associated with the density in 1.52:

$$\begin{aligned}\int_0^z f_Z(w)dw &= \int_0^z \int_0^\infty y f_X(wy) f_Y(y) dy dw \\ &= \int_0^\infty \int_0^z y f_X(wy) dw f_Y(y) dy \\ &= \int_0^\infty F_X(zy) f_Y(y) dy.\end{aligned}$$

Example 1.22 *1. **Ratio of Gammas and the F-distribution:** Let X and Y be independent gamma random variables with common λ parameter and respective parameters of α_1 and α_2. Then if $Z = X/Y$, we have from substituting 1.22 into 1.52 that for $z > 0$:*

$$\begin{aligned}f_Z(z) &= \frac{\lambda^{\alpha_1+\alpha_2} z^{\alpha_1-1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^\infty y^{\alpha_1+\alpha_2-1} e^{-\lambda(1+z)y} dy \\ &= \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \frac{z^{\alpha_1-1}}{(1+z)^{\alpha_1+\alpha_2}}.\end{aligned}$$

Comparing to 1.27 and recalling the identity in 1.30, the density function of Z is very closely related to the beta distribution with parameters $v = \alpha_1$ and $w = \alpha_2$. Specifically, let $F_B(w)$ denote this beta distribution function. Substituting $y = \frac{x}{1+x}$ obtains:

$$\begin{aligned}F_Z(z) &= \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^z \frac{x^{\alpha_1-1} dx}{(1+x)^{\alpha_1+\alpha_2}} \\ &= \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^{\frac{z}{1+z}} y^{\alpha_1-1}(1-y)^{\alpha_2-1} dy \\ &= F_B\left(\frac{z}{1+z}\right).\end{aligned}$$

Remark 1.23 *In the special case where* $\lambda = 1/2$, $\alpha_1 = n/2$ *and* $\alpha_2 = m/2$, *these* X *and* Y *gamma variates are called* ***Chi-squared variables with respective degrees of freedom of*** n ***and*** m *as noted in remark 1.8 above. In this special case, the random variable*

$$F \equiv \frac{m}{n} Z = \frac{mX}{nY},$$

is said to have an F***-distribution with*** n ***and*** m ***degrees of freedom****, and sometimes a* ***Snedecor's*** F ***distribution*** *or a* ***Fisher-Snedecor distribution****, and named for* ***R. A. Fisher*** *(1890 – 1962) and* ***George W. Snedecor*** *(1881 – 1974).*

The distribution function for the F *variate satisfies:*

$$f_F(x) = f_Z(\frac{n}{m}x),$$

and so from the above calculation the density function of an F*-variate with* n *and* m *degrees of freedom is:*

$$f_{F_{n,m}}(x) = \frac{\Gamma((n+m)/2)}{\Gamma(n/2)\Gamma(m/2)} \frac{(nx/m)^{n/2-1}}{(1+nx/m)^{(n+m)/2}}. \tag{1.54}$$

As will be seen below in the section Examples of Moments and M.G.F.s, n *is the mean of* X *and* m *the mean of* Y, *so the scalings in the definition of* $F \equiv \frac{X/n}{Y/m}$ *produce a ratio of Chi-squared variates which have been normalized to each have a mean of* 1. *Because sample variances of normal variates with known means have been shown to have Chi-squared distributions in example 1.14 above, the* F*-distribution can be used to establish confidence intervals for the ratio of the sample variances, and hence test the hypothesis that the samples have the same variance.*

2. ***Ratio of Normal and "Chi" and the Student T distribution:*** *Another important example of a variate defined as the ratio of independent variates is developed as follows. If* X *is standard normal, and* Y *is Chi-squared with* n *degrees of freedom, then the distribution function of*

$$T \equiv X/\sqrt{Y/n} = X\sqrt{n}/\sqrt{Y},$$

is useful as will be seen momentarily. As noted in the above remark, Y/n *is Chi-squared but normalized to have a mean of* 1, *and*

the square root of this variate is the facetiously labelled "chi" variate in the title. To find the density function of T we apply 1.52, writing $T = X\sqrt{n}/\sqrt{Y}$, as the ratio of a normal variate with $\mu = 0$ and $\sigma^2 = n$, and the square root of a Chi-squared variate.

The density of the normal is found in 1.32, while for the latter we apply 1.39 which states that if f_Y denotes the distribution function of the Chi-squared with n degrees of freedom, then the density function of $\sqrt{Y}$ is given by:

$$f_{\sqrt{Y}}(y) = 2yf_Y(y^2).$$

With a little algebra the density function for T is derived:

$$\begin{aligned} f_T(t) &= \int_0^\infty y f_X(ty) f_{\sqrt{Y}}(y) dy \\ &= \frac{1}{\sqrt{\pi n} 2^{(n-1)/2} \Gamma(n/2)} \int_0^\infty y^n \exp\left(-\frac{1}{2}\left(1 + \frac{t^2}{n}\right) y^2\right) dy. \end{aligned}$$

The substitution $s = \frac{1}{2}\left(1 + \frac{t^2}{n}\right) y^2$ produces:

$$f_T(t) = \frac{\Gamma((n+1)/2)}{\sqrt{\pi n}\Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}. \tag{1.55}$$

Remark 1.24 *The distribution function of T defined above is known as the* ***Student T distribution*** *or* ***Student's T distribution*** *with n* ***degrees of freedom****. While the above derivation and many applications call for n to be an integer, this distribution is defined more generally with $\nu > 0$ degrees of freedom, by:*

$$f_T(t) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\pi \nu}\Gamma(\nu/2)} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}. \tag{1.56}$$

It is named for ***William Sealy Gosset*** *(1876 – 1937) who published under the pen name of Student.*

An important application of this distribution is for the determination of confidence intervals for the mean μ of a normal distribution based on a given sample, where the variance parameter σ^2 is unknown. In this application, given a sample $\{Z_i\}_{i=1}^n$ from a normal distribution with unknown parameters (μ, σ^2), define the "sample mean" $\bar{Z} = \sum_{i=1}^n Z_i/n$, and the "sample variance" $s^2 = \sum_{i=1}^n \left(Z_i - \bar{Z}\right)^2/n$. Now let

$$X = \frac{\overline{Z} - \mu}{\sigma/\sqrt{n}}, \qquad Y^2 = \frac{ns^2}{\sigma^2}.$$

As we will see in the below section on ***expectations****, X is standard normal and we now motivate the fact that Y^2 is chi-squared with $n-1$ degrees of freedom. To this end, note that from*

$$\begin{aligned} ns^2 &= \sum_{i=1}^{n} \left[(Z_i - \mu) - (\overline{Z} - \mu) \right]^2 \\ &= \sum_{i=1}^{n} (Z_i - \mu)^2 - n \left(\overline{Z} - \mu \right)^2, \end{aligned}$$

obtains

$$Y^2 + \left(\frac{\bar{Z} - \mu}{\sigma/\sqrt{n}} \right)^2 = \sum_{i=1}^{n} \left(\frac{Z_i - \mu}{\sigma} \right)^2.$$

From the discussion in notation 1.15 on the Chi-squared distribution we conclude that the summation on the right is Chi-squared with n degrees of freedom, and similarly, that $X = \frac{\overline{Z}-\mu}{\sigma/\sqrt{n}}$ is standard normal assures that $\left(\frac{\bar{Z}-\mu}{\sigma/\sqrt{n}} \right)^2$ is Chi-squared with 1 degree of freedom. The independence of s^2 and $\bar{Z}$ for normal distributions will be proved in book 6 in the section on the Multivariate Normal Distribution, and this implies the independence of Y^2 and $\left(\frac{\bar{Z}-\mu}{\sigma/\sqrt{n}} \right)^2$ by proposition 3.56 of book 2. The results of section 3.2.7 below then demonstrate that Y^2 is Chi-squared with $n-1$ degrees of freedom. See example 3.59 for details.

Hence, $T = \frac{X\sqrt{n-1}}{\sqrt{Y}}$ is Student T with $n-1$ degrees of freedom, and a calculation shows that:

$$T = \frac{\left(\bar{Z} - \mu \right) / \sqrt{n-1}}{s}.$$

Consequently, confidence intervals for such T can then be translated to confidence intervals for μ. See the discussion following proposition 5.52 for an example of this application.

3. *In book 6, once the mathematics of this section is formalized, it will be shown that the Cauchy distribution defined below in 3.64 is the distribution function of the ratio of two independent standard normal variates.*

A Special Case without Independence

The following result will be applied in chapter 4, Simulating Samples of Random Variables - Examples. Because this special case involves the ratio

of random variables which are not independent, we address the solution from "first principles." In the process we again assume that a Riemann integral in $\mathbb{R}^2$ can be evaluated iteratively, and the order of integration reversed. As a Riemann integral this is justified in calculus based on the absolute integrability of the integrand, but more generally requires Fubini's theorem of book 5.

Proposition 1.25 *Let X and Y be independent Gamma random variables with parameters α_1, λ, and α_2, λ, respectively. Define the random variable $Z = \frac{X}{X+Y}$. Then Z is supported on the interval $[0,1]$, is independent of the parameter λ, and has density function given by:*

$$f(z) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} z^{\alpha_1 - 1} (1-z)^{\alpha_2 - 1}. \tag{1.57}$$

Thus by 1.27 and 1.30, Z is a Beta random variable with parameters $v = \alpha_1$ and $w = \alpha_2$.

Proof. *First note that since X and Y are supported on $[0, \infty)$, Z is potentially undefined when $X = Y = 0$, but as this event has probability 0 it causes no problem. In detail, by independence and 1.21:*

$$\Pr[X \leq \epsilon_1, Y \leq \epsilon_2] = [1 - \exp(-\lambda \epsilon_1)][1 - \exp(-\lambda \epsilon_2)] \to 0,$$

as $\epsilon_j \to 0$. Thus we can redefine X and Y on $(0, \infty)$ without changing their distributions or the calculations below.

Because of the independence assumption, the distribution function of Z is given by

$$F(z) = \iint_{x/(x+y) \leq z} f_X(x) f_Y(y) dx dy.$$

Now $x/(x+y) \leq z$ if and only if $x \leq zy/(1-z)$, so iterating this double integral, substituting $x = uy/(1-u)$, and reversing the integrals obtains:

$$\begin{aligned} F(z) &= \int_0^\infty \int_0^{zy/(1-z)} f_X(x) f_Y(y) dx dy \\ &= \int_0^\infty \left[\int_0^z f_X\left(\frac{uy}{1-u}\right) \frac{y}{(1-u)^2} du \right] f_Y(y) dy \\ &= \int_0^z \left[\frac{1}{(1-u)^2} \int_0^\infty y f_X\left(\frac{uy}{1-u}\right) f_Y(y) dy \right] du. \end{aligned}$$

By the fundamental theorem of calculus version II, which is proposition 3.2 of book 3:

$$f(z) = \frac{1}{(1-z)^2} \int_0^\infty y f_X\left(\frac{zy}{1-z}\right) f_Y(y) dy,$$

which can be now evaluated by a direct substitution of the Gamma density functions. Specifically, by 1.22:

$$f(z) = \frac{\lambda^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)\,(1-z)^2} \int_0^\infty \left(\frac{zy}{1-z}\right)^{\alpha_1-1} y^{\alpha_2} \exp\left[-\lambda\left(\frac{y}{1-z}\right)\right] dy,$$

and a substitution of $w = \frac{\lambda y}{1-z}$ *into this Riemann integral produces:*

$$f(z) = \frac{z^{\alpha_1-1}\,(1-z)^{\alpha_2-1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^\infty w^{\alpha_1+\alpha_2-1} \exp\left[-w\right] dw.$$

The result in 1.57 now follows by the definition of $\Gamma(\alpha_1+\alpha_2)$ *in 1.23.* ■

Chapter 2

Order Statistics

Given any sample of independent, identically distributed variates, $\{X_j\}_{j=1}^M$, these can be re-ordered into $\{X_{(k)}\}_{k=1}^M$ where $X_{(k)} \leq X_{(k+1)}$. Each $X_{(k)}$ is then called a *kth* **order statistic** when M is apparent from the context, or the *kth* **order statistic from a sample of** M, otherwise**.** To further emphasize M many authors use notation such as:

$$X_{(k,M)} \equiv X_{(k)}.$$

Notation 2.1 *It is important to note that, perhaps ironically, there is no universal notational convention for the* ***order of order statistics.*** *In other words, in some references order statistics are ordered in the natural numerical order, so* $X_{(k)} \leq X_{(k+1)}$ *as above. However, it is not uncommon to see order statistics denoted so that* $X_{(1)}$ *is the largest, and hence* $X_{(k+1)} \leq X_{(k)}$.

In this section we derive the distribution function of *kth* **order statistics** and related functions, and introduce the **Rényi representation theorem** for exponential order statistics which will be applied in the section below, Extreme Value Theory 2.

2.1 Distribution Functions for kth Order Statistics

Let X be a random variable defined on $(\mathcal{S}, \mathcal{E}, \lambda)$ with distribution function F, and $\{X_j\}_{j=1}^M$ a given random sample (see chapter 4 of book 2 to formalize this notion). Then the distribution function of the kth order statistic $X_{(k)}$, is relatively straightforward to derive. Indeed if $X_{(k)} \leq x$, then **at least** k of the variates satisfy this constraint and at most $n - k$

variates exceed x. Denoting this distribution function by $F_{(k)}(x)$, we have the following:

Proposition 2.2 *Given independent random variables with common distribution function F, the distribution function of the kth order statistic $F_{(k)}(x)$ is given by:*

$$F_{(k)}(x) = \sum_{j=k}^{M} \binom{M}{j} F^j(x) \left(1 - F(x)\right)^{M-j}, \tag{2.1}$$

and is defined on the same domain as is $F(x)$.

Proof. *For a given ordering of independent variates, $(X_1, ..., X_M)$ and j components specified, the probability that exactly these j variates are less than or equal to x and the remaining $M - j$ variates greater than x is $F^j(x)\left(1 - F(x)\right)^{M-j}$. There are $\binom{M}{j}$ such specifications possible, so by independence, $\binom{M}{j} F^j(x)\left(1 - F(x)\right)^{M-j}$ is the probability that exactly j variates are less than or equal to x. As noted above, $X_{(k)} \leq x$ is the event $j \geq k$, and addition of probabilities is justified since these events are disjoint.* ■

Example 2.3 *1.* ***Extreme Value Distributions***

When $k = M$, $F_{(M)}(x) = F^M(x)$ is the distribution function introduced and characterized in the book 2 section, Extreme Value Theory 1, representing the distribution function of $\max\{X_j\}$. This study is continued below in Extreme Value Theory 2.

2. ***Uniform Continuous Distribution on*** $[0, 1]$

If $F_U(x) = x$ on $[0, 1]$, the distribution function of the kth order statistic is given by:

$$F_{(k)}(x) = \sum_{j=k}^{M} \binom{M}{j} x^j \left(1 - x\right)^{M-j}.$$

In particular, the distribution functions for the smallest and largest uniform variate are respectively given on $[0, 1]$ by:

$$F_{(1)}(x) = 1 - \left(1 - x\right)^M; \qquad F_{(M)}(x) = x^M.$$

3. ***Exponential Distribution, Parameter*** λ

If $F_\Gamma(x) = 1 - e^{-\lambda x}$ on $[0, \infty)$, the distribution function of the kth order statistic is given by:

$$F_{(k)}(x) = \sum_{j=k}^{M} \binom{M}{j} \left(1 - e^{-\lambda x}\right)^j \left(e^{-\lambda x}\right)^{M-j}.$$

In particular, the distribution functions for the smallest and largest exponential variates are given on $[0, \infty)$ *by:*

$$F_{(1)}(x) = 1 - e^{-\lambda M x}; \qquad F_{(M)}(x) = \left(1 - e^{-\lambda x}\right)^M.$$

Consequently, the 1*st order statistic* $X_{(1)}$ *of an exponential with parameter* λ *is exponentially distributed with parameter* λM*. This observation will be expanded upon in the section below on the* ***Rényi representation theorem on order statistics.***

2.2 Density Functions for kth Order Statistics

If $F(x)$ is absolutely continuous then by version I of the fundamental theorem of calculus in proposition 3.61 of book 3, $F(x)$ has an associated measurable density function $f(x)$ with $f(x) = F'(x)$ almost everywhere. If $F(x)$ is continuously differentiable, then $f(x) = F'(x)$ for all x by the proposition 3.3 of book 3. In either case, this assumption on $F(x)$ yields the same assumption on the distribution function of $F_{(k)}(x)$ in 2.1. To simplify the next statement, we assume $F(x)$ is continuously differentiable. The absolutely continuous result is derived by qualifying the density function as being valid almost everywhere.

Proposition 2.4 *If* $F(x)$ *is continuously differentiable with density* $f(x)$*, then the density function of the kth order statistic is given by:*

$$f_{(k)}(x) = c_{(k)} \left[F(x)\right]^{k-1} \left(1 - F(x)\right)^{M-k} f(x), \tag{2.2}$$

where $c_{(k)}$ *is alternately expressed:*

$$c_{(k)} = \frac{M!}{(k-1)!\,(M-k)!} = M\binom{M-1}{k-1} = k\binom{M}{k} = \frac{\Gamma(M+1)}{\Gamma(k)\Gamma(M-k+1)}. \tag{2.3}$$

Proof. *Differentiating* $F_{(k)}(x)$ *in 2.1 yields:*

$$F'_{(k)}(x) = \sum_{j=k}^{M} \binom{M}{j} \left[j \left[F(x)\right]^{j-1} \left(1 - F(x)\right)^{M-j} - (M-j) \left[F(x)\right]^{j} \left(1 - F(x)\right)^{M-j-1} \right] f(x).$$

This summation "telescopes" by noting that:

$$j\binom{M}{j} = M\binom{M-1}{j-1}, \qquad (M-j)\binom{M}{j} = M\binom{M-1}{j},$$

and this then produces 2.2. The equivalent formulations for the constant $c_{(k)}$ *can be verified as an exercise.* ■

One application of 2.2 is the following:

Example 2.5 *Let Y be continuous and uniformly distributed on $[0,1]$. Then the density function of the kth order statistic of a sample of n, $f_{(k)}(y)$, is a Beta density with parameters $v = k$ and $w = n - k + 1$:*

$$f_{(k)}(y) = \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)} y^{k-1} (1-y)^{n-k}. \tag{2.4}$$

This follows from 2.2 since $F(y) = y$ here, recalling 1.27 and the last expression in 2.3.

2.3 Joint Distribution of all Order Statistics

We next derive the joint distribution function, $F_{(1,...,M)}(x_1, ..., x_M)$ of all order statistics. By definition, for $x_1 \leq x_2 \leq \cdots \leq x_M$:

$$F_{(1,...,M)}(x_1, ..., x_M) = \Pr\{X_{(j)} \leq x_j \text{ for all } j = 1, ..., M\}.$$

Recall the following:

Definition 2.6 *Given an ordered set, $A \equiv (x_1, x_2, ..., x_M)$ where M can be infinite, a **permutation** $\pi : A \to A$ is a one-to-one and onto mapping:*

$$(x_1, x_2, ..., x_M) \to (\pi(x_1), \pi(x_2), ..., \pi(x_M)). \tag{2.5}$$

Exercise 2.7 *Show that if M is finite then there are $M!$ possible permutations including the identity permutation, $\pi(x_j) = x_j$, whereas if $M = \infty$, then there are uncountably many. Hint for $M = \infty$: Given a binary expansion for $b \in [0,1)$, $b_{(2)} = b_1 b_2, ...$ with each $b_j \in \{0,1\}$, define a permutation π_b so that for each j :*

- *If $b_j = 0 : \pi_b(2x_j) = 2x_j$, $\pi_b(2x_j + 1) = 2x_j + 1$.*
- *If $b_j = 1 : \pi_b(2x_j) = 2x_j + 1$, $\pi_b(2x_j + 1) = 2x_j$.*

Given a permutation $\pi : (1, 2, ..., M) \to (\pi(1), ..., \pi(M))$, define $D_\pi \subset \mathbb{R}^M$ by:

$$D_\pi = \{(x_1, ..., x_M) | x_{\pi(1)} \leq x_{\pi(2)} \leq \cdots \leq x_{\pi(M)}\},$$

and note that D_π is closed and Lebesgue measurable. This follows because D_π is the intersection of $M-1$ closed sets: $\{x_{\pi(j)} \leq x_{\pi(j+1)}\}$, and closed sets

are measurable. Also, the sets $\{D_\pi\}$ are "nearly" disjoint in the following sense. Defining the **interior** of D_π:

$$D_\pi^o = \{((x_1, ..., x_M) | x_{\pi(1)} < x_{\pi(2)} < \cdots < x_{\pi(M)}\},$$

then given any two permutations, $\pi_1 \neq \pi_2$:

$$D_{\pi_1}^o \cap D_{\pi_2}^o = \emptyset.$$

In addition, $D_{\pi_1} \cap D_{\pi_2}$ is at most an $(M-1)$-dimensional subset of $\mathbb{R}^M$ since if $x \in D_{\pi_1} \cap D_{\pi_2}$ then for some $j \neq k$ both $x_j \leq x_k$ and $x_k \leq x_j$ and thus $x_j = x_k$. As the intersection of closed and measurable sets, $D_{\pi_1} \cap D_{\pi_2}$ is closed and measurable, as is any such intersection of these sets.

Exercise 2.8 *Verify that any such intersection set of* $\{D_\pi\}$ *has Lebesgue measure* 0, *and:*

$$\bigcup\nolimits_\pi D_\pi = \mathbb{R}^M,$$

where this union is over all $M!$ *permutations.*

Before developing the general result, an example is needed.

Example 2.9 *Let* $M = 3$. *If* $\pi : (1,2,3) \to (1,2,3)$, *the integral over* D_π *can be expressed as iterated integrals as has been noted above (Fubini, book 5):*

$$\iiint_{D_\pi} dy = \int_{\mathbb{R}} \int_{y_2 \leq y_3} \int_{y_1 \leq y_2} dy_1 dy_2 dy_3.$$

Assume that X *is defined on* $(\mathcal{S}, \mathcal{E}, \lambda)$ *and has a continuous density for simplicity. Also, recall that an expression such as* $\Pr[(X_1, X_2, X_3) \in D_\pi]$ *is defined as* $\lambda\left[(X_1, X_2, X_3)^{-1} D_\pi\right]$. *Then* $f(y_1, y_2, y_3) = f(y_1) f(y_2) f(y_3)$ *by independence of the random variables, and an integration by parts yields:*

$$\begin{aligned}
\Pr[(X_1, X_2, X_3) \in D_\pi] &= \int_{\mathbb{R}} \int_{y_2 \leq y_3} \int_{y_1 \leq y_2} f(y_1) f(y_2) f(y_3) dy_1 dy_2 dy_3 \\
&= \int_{\mathbb{R}} \int_{y_2 \leq y_3} F(y_2) f(y_2) f(y_3) dy_2 dy_3 \\
&= \frac{1}{2} \int_{\mathbb{R}} F^2(y_3) f(y_3) dy_3 \\
&= 1/3!
\end{aligned}$$

By the same calculation,

$$\Pr[(X_1, X_2, X_3) \in D_\pi] = \Pr[(X_1, X_2, X_3) \in D_\pi^o],$$

since the boundaries of D_π have measure zero.

In M dimensions, if X has a continuous density the same calculations derive that for any π,

$$\Pr\left[(X_1, ..., X_M) \in D_\pi\right] = \Pr\left[(X_1, ..., X_M) \in D_\pi^o\right] = 1/M!,$$

and hence since this result is the same for any π :

$$\sum\nolimits_\pi \Pr\left[(X_1, ..., X_M) \in D_\pi\right] = 1. \tag{2.6}$$

Note that continuity of densities produces the simple result in 2.6 because this assumption assures that for any permutation,

$$\Pr\left[(X_1, ..., X_M) \in D_\pi\right] = \Pr\left[(X_1, ..., X_M) \in D_\pi^o\right],$$

or equivalently,

$$\Pr\left[(X_1, ..., X_M) \in (D_\pi - D_\pi^o)\right] = 0.$$

That continuity is not a superfluous assumption is easy to exemplify.

Example 2.10 *If $X : \mathcal{S} \to \{0, 1\}$ is binomial, with $\lambda[X^{-1}(1)] = p$ with $0 < p < 1$, then with $M = 2$ there are only 2 permutations and for either:*

$$\lambda\left[(X_1, X_2)^{-1} D_\pi\right] = 1 - p(1 - p).$$

This follows because if $\pi_1 : (1, 2) \to (1, 2)$ then $(X_1, X_2)^{-1} D_{\pi_1} = \{X_1 \leq X_2\}$ and:

$$D_{\pi_1} = \{(0, 0), (0, 1), (1, 1)\},$$

while if $\pi_2 : (1, 2) \to (2, 1)$ then $(X_1, X_2)^{-1} D_{\pi_2} = \{X_2 \leq X_1\}$ and:

$$D_{\pi_2} = \{(0, 0), (1, 0), (1, 1)\}.$$

Thus:

$$\sum\nolimits_\pi \lambda\left[(X_1, X_2)^{-1} D_\pi\right] = 2\left[1 - p(1 - p)\right] > 1.$$

This sum always exceeds 1 by exactly $p^2 + (1 - p)^2$, and this is because for $A \equiv D_{\pi_1} \cap D_{\pi_2} = \{(0, 0), (1, 1)\}$:

$$\lambda\left[(X_1, X_2)^{-1} A\right] = p^2 + (1 - p)^2 \neq 0.$$

Put another way, $\lambda\left[(X_1, X_2)^{-1} D_\pi\right] \neq \lambda\left[(X_1, X_2)^{-1} D_\pi^o\right]$.

With the above warm-up, we now state the main result.

Proposition 2.11 *Given independent random variables with common distribution function $F(x)$ and a continuous density function, the joint distribution function of all order statistics is given for $x_1 \leq x_2 \leq \cdots \leq x_M$ by:*

$$F_{(1,...,M)}(x_1, ..., x_M) = M!F(x_1) \prod_{j=2}^{M} [F(x_j) - F(x_{j-1})]. \tag{2.7}$$

Proof. *Given $x \equiv (x_1, ..., x_M)$ with $x_1 \leq x_2 \leq \cdots \leq x_M$, define the right semi-closed infinite rectangle $R_x = \{(y_1, ..., y_M) | y_j \leq x_j \text{ for all } j\}$. Then since*

$$\begin{aligned} F_{(1,...,M)}(x_1, ..., x_M) &\equiv \Pr\{X_{(j)} \leq x_j,\ j = 1, ..., M\} \\ &= \bigcup\nolimits_{\pi} \Pr\{X_{\pi(j)} \leq x_j,\ j = 1, ..., M\}, \end{aligned}$$

it follows that:

$$\begin{aligned} F_{(1,...,M)}(x_1, ..., x_M) &= \lambda \left[(X_1, ..., X_M)^{-1} \left(\bigcup\nolimits_{\pi} [D_\pi \cap R_x] \right) \right] \\ &= \lambda \left[\bigcup\nolimits_{\pi} (X_1, ..., X_M)^{-1} [D_\pi \cap R_x] \right] \\ &= \lambda \left[\bigcup\nolimits_{\pi} (X_1, ..., X_M)^{-1} [D_\pi^o \cap R_x] \right]. \end{aligned}$$

Note that the third equation follows from continuity of the density function and thus as above, $\lambda \left[\bigcup_\pi (X_1, ..., X_M)^{-1} [D_\pi - D_\pi^o] \right] = 0$.

Now since $\{(X_1, ..., X_M)^{-1} [D_\pi^o \cap R_x]\}_\pi$ are disjoint sets, finite additivity of λ and independence obtain that:

$$\begin{aligned} F_{(1,...,M)}(x_1, ..., x_M) &= \sum\nolimits_{\pi} \lambda \left[(X_1, ..., X_M)^{-1} [D_\pi^o \cap R_x] \right] \\ &= \sum\nolimits_{\pi} \lambda \left[(X_1, ..., X_M)^{-1} [D_\pi \cap R_x] \right] \\ &= \sum\nolimits_{\pi} \lambda \{X_{\pi(1)}(s) \leq x_1\ \} \prod_{j=2}^{M} \lambda \{x_{j-1} \leq X_{\pi(j)}(s) \leq x_j\ \} \\ &= \sum\nolimits_{\pi} F_{\pi(1)}(x_1) \left[F_{\pi(2)}(x_2) - F_{\pi(1)}(x_1) \right] ... \left[F_{\pi(M)}(x_M) - F_{\pi(M-1)}(x_{M-1}) \right]. \end{aligned}$$

But since $X_j : (\mathcal{S}, \mathcal{E}, \mu) \to \mathbb{R}$ are identically distributed, $F_{\pi(j)}(x_j) = F(x_j)$ for any π and 2.7 follows as the summation of $M!$ identical expressions. ■

Example 2.12 *When the density function is not continuous, this result is not valid. For X binomial as in example 1.5 above, a calculation shows that $F_{(1,2)}(0, 1) = 1 - p^2$ while $2F(0)[F(1) - F(0)] = 2p(1 - p)$.*

2.4 Multivariate Density Functions

In the next few sections we will derive density functions associated with various multivariate distribution functions. To be completely rigorous requires the general integration theory of book 5 which will be applied as below in book 6. But intuitively as has been illustrated in sections 1.3 and 1.4 above, the notion of a multivariate density function is identical with that in the one variable context. If $F(x)$ is the joint distribution function of $X \equiv (X_1, X_2, ..., X_n)$ defined on $(\mathcal{S}, \mathcal{E}, \lambda)$ with range in $\mathbb{R}^n$, we say that $f(x)$ is an associated joint density function if with $x \equiv (x_1, x_2, ..., x_n)$:

$$F(x) = \int_{y \leq x} f(y)dy,$$

where the domain of integration, $R_x \equiv \{y \leq x\}$ is shorthand for $\{y_j \leq x_j$ for all $j\}$. This of course is consistent with definition 3.28 of book 2:

$$\begin{aligned} F(x_1, x_2, ..., x_n) &\equiv \lambda \left[\bigcap_{j=1}^{n} X_j^{-1}(-\infty, x_j] \right] \\ &= \lambda \left[X^{-1} R_x \right]. \end{aligned}$$

Just as for the one variable case with singular distribution functions the example, joint density functions need not exist for a given joint distribution function.

When $f(x)$ exists this integral will in general be defined as a Lebesgue integral in $\mathbb{R}^n$, but in many applications $f(x)$ will be continuous and this will be an ordinary Riemann integral. As is the case for the Lebesgue integrals of book 3, such $f(x)$ is in general not uniquely defined since if $f(x) = g(x)$ outside a set of Lebesgue measure 0, then the Lebesgue integrals of $f(x)$ and $g(x)$ agree and thus $g(x)$ is also a density function for $F(x)$. The same is true in the Riemann context as long as $g(x)$ is also Riemann integrable. That said, in cases where there is a continuous density function, this version is often regarded as if unique.

With the aid of Fubini's theorem such integrals in $\mathbb{R}^n$ can be "iterated" as has been applied above. That is:

$$F(x) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f(y_1, y_2, ..., y_n) dy_1 dy_2 ... dy_n,$$

where this notation implies that these integrals can be evaluated one at a time, in this or in any given order. This will be formalized in book 5, so

here for expedience we simply assume these manipulations are valid for the results below. Any such $f(y_1, y_2, ..., y_n)$ is nonnegative and thus absolutely integrable over $\mathbb{R}^n$, and when also continuous, an assumption often true in applications, version II of the fundamental theorem of calculus of proposition 3.2 of book 3 applies one variable at a time to derive that:

$$f(x_1, x_2, ..., x_n) = \frac{\partial^n F}{\partial x_1 \partial x_2 \cdots \partial x_n}. \tag{2.8}$$

2.4.1 Joint Density of all Order Statistics

The **joint density function** of all order statistics associated with the joint distribution function, $F_{(1,...,M)}(x_1, ..., x_M)$, can be derived from 2.8 assuming continuity of the density function $f(x)$, the same assumption underlying 2.7. However, the relationship between the integral of the density function and the distribution function noted above must be modified for order statistics because $x_j \leq x_{j+1}$. Specifically, for $x_1 \leq x_2 \leq \cdots \leq x_M$:

$$F_{(1,...,M)}(x_1, ..., x_M) = \int_{x_{M-1}}^{x_M} \cdots \int_{x_1}^{x_2} \int_{-\infty}^{x_1} f_{(1,...,M)}(y_1, y_2, ..., y_M) dy_1 dy_2 ... dy_M. \tag{2.9}$$

This formula will significantly complicate the relationship between $f_{(1,...,M)}(x_1, ..., x_M)$ and derivatives of $F_{(1,...,M)}(x_1, ..., x_M)$ compared with the result in 2.8. Fortunately, for the current application the hard work has already been done.

Proposition 2.13 *Given independent random variables with common distribution function $F(x)$ and with continuous density function $f(x)$, the continuous joint density function of all order statistics is given for $x_1 \leq x_2 \leq \cdots \leq x_M$ by:*

$$f_{(1,...,M)}(x_1, ..., x_M) = M! f(x_1) f(x_2) ... f(x_M). \tag{2.10}$$

Proof. *From 2.9, the joint distribution function in 2.7 can be expressed for $x_1 \leq x_2 \leq \cdots \leq x_M$:*

$$M! F(x_1) \prod_{j=2}^{M} [F(x_j) - F(x_{j-1})] = \int_{x_{M-1}}^{x_M} \cdots \int_{x_1}^{x_2} \int_{-\infty}^{x_1} f_{(1,...,M)}(y_1, y_2, ..., y_M) dy_1 dy_2 ... dy_M.$$

Because for $j \geq 2$:

$$\int_{x_{j-1}}^{x_j} f(y_j) dy_j = F(x_j) - F(x_{j-1}),$$

and similarly for $j = 1$ and $x_{j-1} = -\infty$, the density function $f_{(1,...,M)}(x_1, ..., x_M)$ in 2.10 satisfies this identity and is thus the continuous version of this density function. ■

An application of 2.10 will be given below in section 2.5, The Rényi Representation Theorem.

2.4.2 Marginal Densities of Order Statistics

Recall definition 3.34 of book 2 which defined the **marginal distribution functions** given the joint distribution function $F(x_1, ..., x_n)$.

Definition 2.14 *Given $F(x_1, x_2, ..., x_n)$ and $I = \{i_1, ..., i_m\} \subset \{1, 2, ..., n\}$, let $x_I = (x_{i_1}, x_{i_2}, ..., x_{i_m})$, and $x_J = (x_{j_1}, x_{j_2}, ..., x_{j_{n-m}})$ for $j_k \in J \equiv \tilde{I}$. Then the **marginal distribution function** $F_I(x_{i_1}, x_{i_2}, ..., x_{i_m})$ is defined on $\mathbb{R}^m$ by:*

$$F_I(x_{i_1}, x_{i_2}, ..., x_{i_m}) \equiv \lim_{x_J \to \infty} F(x_1, x_2, ..., x_n). \tag{2.11}$$

Thus given $F(x_1, x_2, ..., x_n)$, there are $2^n - 2$ **proper** marginal distribution functions defined by the $2^n - 2$ proper subsets of $\{1, 2, ..., n\}$.

For a general distribution function defined on $\mathbb{R}^n$ with a density $f(y_1, y_2, ..., y_n) \equiv f(y_I, y_J)$, 2.11 implies that with apparent notation:

$$F_I(x_{i_1}, x_{i_2}, ..., x_{i_m}) \equiv \int_{-\infty}^{x_I} \int_{-\infty}^{\infty} f(y_I, y_J) dy_J dy_I,$$

and thus

$$F_I(x_{i_1}, x_{i_2}, ..., x_{i_m}) = \int_{-\infty}^{x_I} f_I(y_I) dy_I.$$

where the marginal density function $f_I(x_I)$ defined:

$$f_I(x_I) = \int_{-\infty}^{\infty} f(x_I, y_J) dy_J. \tag{2.12}$$

In the current application to $F_{(1,...,M)}(x_1, ..., x_M)$, we must take more care with this integration since the variables are ordered. For example, with $x_J = x_1$, it would make no sense to let $x_1 \to \infty$ in the distribution function since $x_1 \leq x_2 \leq \cdots \leq x_M$ and thus $x_1 \to \infty$ here really means $x_1 \to x_2$. Thus integrating the x_1 variate of the density function over $(-\infty, \infty)$ must be interpreted as the integral over $(-\infty, x_2]$. In general, for the definition of marginal distribution function we must interpret "∞" as

the upper boundary point of the domain of the variable, and this differs depending on which indexes are in the x_J vector. Similar comments apply to the lower limit of integration, that $-\infty$ must also be interpreted in terms of the lower boundary point of the domain of the given variable. The resulting calculations can become tedious and so we provide the result when $I = (i, j)$ with $1 \leq i < j \leq M$, deriving the marginal density functions $f_{(i,j)}(x_i, x_j) \equiv f_I(x_i, x_j)$.

Within the proof we will also develop the marginal distribution functions $f_{(1,...,j)}(x_1, ..., x_j)$ and $f_{(i,...,j)}(x_i, ..., x_j)$, and these are summarized as a corollary.

Proposition 2.15 *Given the joint distribution function $F_{(1,...,M)}(x_1, ..., x_M)$ in 2.7 and $I = (i, j)$ with $1 \leq i < j \leq M$, the marginal density function $f_{(i,j)}(x_i, x_j)$ is given for $x_i \leq x_j$ by:*

$$f_{(i,j)}(x_i, x_j) = \frac{M!}{(M-j)!(j-i-1)!(i-1)!} f(x_i) f(x_j) \left[1 - F(x_j)\right]^{M-j} \left[F(x_j) - F(x_i)\right]^{j-i-1} F^{i-1}(x_i). \tag{2.13}$$

Proof. *Given the above remarks, $f_{(i,j)}(x_i, x_j)$ will be derived from the joint density function in 2.10 by integrating in three steps as follows:*

- *For variates x_k with $k > j$, the integral is from x_{k-1} to ∞,*
- *For variates x_k with $k < i$, the integral is from $-\infty$ to x_{k+1},*
- *For variates x_k with $i < k < j$, the integral is from x_{k-1} to x_j.*

Using the first step we calculate the marginal density $f_{(1,...,j)}(x_1, ..., x_j)$ from $f_{(1,...,M)}(x_1, ..., x_M)$, first dividing by $M! f(x_1) f(x_2) ... f(x_j)$ to simplify notation:

$$\begin{aligned}
& f_{(1,...,j)}(x_1, ..., x_j) / \left[M! f(x_1) f(x_2) ... f(x_j)\right] \\
= & \int_{x_j}^{\infty} \left[\cdots \int_{x_{M-2}}^{\infty} \left[\int_{x_{M-1}}^{\infty} f(y_M) dx_M \right] f(y_{M-1}) dy_{M-1} \cdots \right] f(y_{j+1}) dy_{j+1} \\
= & \int_{x_j}^{\infty} \left[\cdots \int_{x_{M-1}}^{\infty} \left[1 - F(y_{M-1})\right] f(y_{M-1}) dy_{M-1} \cdots \right] f(y_{j+1}) dy_{j+1} \\
& \vdots \\
= & \frac{1}{(M-j)!} \left[1 - F(x_j)\right]^{M-j},
\end{aligned}$$

and so

$$f_{(1,...,j)}(x_1,...,x_j) = \frac{M!}{(M-j)!} f(x_1)f(x_2)...f(x_j)\left[1-F(x_j)\right]^{M-j}$$

Next, $f_{(i,...,j)}(x_i,...,x_j)$ *is derived from* $f_{(1,...,j)}(x_1,...,x_j)$ *in the second step, first dividing by* $\frac{M!}{(M-j)!} f(x_i)f(x_{i+1})...f(x_j)\left[1-F(x_j)\right]^{M-j}$ *to simplify notation:*

$$\begin{aligned}
& f_{(i,...,j)}(x_i,...,x_j) \Big/ \left[\frac{M!}{(M-j)!} f(x_i)f(x_{i+1})...f(x_j)\left[1-F(x_j)\right]^{M-j}\right] \\
= & \int_{-\infty}^{x_i}\left[\cdots\int_{-\infty}^{x_3}\left[\int_{-\infty}^{x_2} f(y_1)dy_1\right] f(y_2)dy_2\cdots\right] f(y_{i-1})dy_{i-1} \\
= & \int_{-\infty}^{x_i}\left[\cdots\int_{-\infty}^{x_3} F(y_2)f(y_2)dy_2\cdots\right] f(y_{i-1})dy_{i-1} \\
& \vdots \\
= & \frac{1}{(i-1)!}F^{i-1}(x_i),
\end{aligned}$$

and so

$$f_{(i,...,j)}(x_i,...,x_j) = \frac{M!}{(M-j)!(i-1)!} f(x_i)f(x_{i+1})...f(x_j)\left[1-F(x_j)\right]^{M-j} F^{i-1}(x_i).$$

The final step requires the following integral of $f(x_{i+1})...f(x_{j-1})$*:*

$$\begin{aligned}
& \int_{x_{i+1}}^{x_j}\left[\cdots\int_{x_{j-3}}^{x_j}\left[\int_{x_{j-2}}^{x_j} f(y_{j-1})dy_{j-1}\right] f(y_{j-2})dy_{j-2}\cdots\right] f(y_{i+1})dy_{i+1} \\
= & \int_{x_i}^{x_j}\left[\cdots\int_{x_{j-3}}^{x_j}\left[F(y_j)-F(y_{j-2})\right] f(y_{j-2})dy_{j-2}\cdots\right] f(y_{i+1})dy_{i+1} \\
& \vdots \\
= & \frac{[F(x_j)-F(x_i)]^{j-i-1}}{(j-i-1)!}.
\end{aligned}$$

Combining obtains 2.13. ■

Corollary 2.16 *Given the joint distribution function* $F_{(1,...,M)}(x_1,...,x_M)$ *in 2.7 and* $1 \le i < j \le M$*, the marginal density functions* $f_{(1,...,j)}(x_1,...,x_j)$ *and* $f_{(i,...,j)}(x_i,...,x_j)$ *are given for* $x_1 \le x_2 \le \cdots x_i \le \cdots \le x_j$ *as follows:*

$$f_{(1,...,j)}(x_1,...,x_j) = \frac{M!}{(M-j)!} f(x_1)f(x_2)...f(x_j)\left[1-F(x_j)\right]^{M-j}, \quad (2.14)$$

$$f_{(i,...,j)}(x_i,...,x_j) = \frac{M!}{(M-j)!(i-1)!} f(x_i)f(x_{i+1})...f(x_j)\left[1-F(x_j)\right]^{M-j} F^{i-1}(x_i). \tag{2.15}$$

Remark 2.17 *Note that the corresponding* ***marginal distribution functions*** *can be defined from the marginal density functions with the usual variate restrictions. For example, $F_{(i,j)}(x_i,x_j)$ for $x_i \leq x_j$ is defined:*

$$F_{(i,j)}(x_i,x_j) = \int_{-\infty}^{x_i}\int_{y_i}^{x_j} f_{(i,j)}(y_i,y_j)dy_j dy_i. \tag{2.16}$$

2.4.3 Conditional Densities of Order Statistics

Recall definition 3.39 from book 2 which defined the **conditional distribution functions** given the joint distribution function $F(x_1,...,x_n)$.

Definition 2.18 *Let $X \equiv (X_1,X_2,...,X_n)$ be a random vector defined on $(\mathcal{S},\mathcal{E},\lambda)$ and $J \equiv \{j_1,...,j_m\} \subset \{1,2,...,n\}$ and $X_J \equiv (X_{j_1},X_{j_2},...,X_{j_m})$. Given a Borel set $B \in \mathcal{B}(\mathbb{R}^m)$ with $\lambda\left[X_J^{-1}(B)\right] \neq 0$, define the conditional distribution function of X given $X_J \in B$, denoted $F(x_1,x_2,...,x_n|X_J \in B)$ by:*

$$F(x_1,x_2,...,x_n|X_J \in B) \equiv \lambda\left[X^{-1}\left(\prod\nolimits_{i=1}^{n}(-\infty,x_i]\right) \mid X_J^{-1}(B)\right],$$

and so by 1.27 (book 2):

$$F(x_1,x_2,...,x_n|X_J \in B) = \lambda\left[X^{-1}\left(\prod\nolimits_{i=1}^{n}(-\infty,x_i]\right) \cap X_J^{-1}(B)\right] / \lambda\left[X_J^{-1}(B)\right]. \tag{2.17}$$

Example 2.19 *A very important example for extreme value theory is the 1-dimensional case and $B = \{X > t\}$. Then $F(x|B) = 0$ for $x \leq t$ by 2.17 since then $X^{-1}((-\infty,x]) \cap X^{-1}((-\infty,t]) = \emptyset$, while for $x \geq t$:*

$$F(x|B) = \frac{F(x)-F(t)}{1-F(t)}.$$

For the next result we recall a common notion from elementary probability theory – that of a conditional distribution function where the conditional set B is replaced by a single point, $B \equiv y_0$ – here applied to a bivariate distribution function $F(x,y)$. Of course in many applications of interest it will be the case that $\lambda\left[Y^{-1}(y_0)\right] = 0$, for example when the the marginal distribution function $F_Y(y)$ is continuous. However, the intuition is compelling

that given the distribution function $F(x, y)$ and $Y = y_0$, there must be a definable distribution function of x:

$$F(x|y_0) \equiv F(x, y|y = y_0),$$

the values of which will depend on y_0. We will return to a very general model for this notion and related ideas in book 6 in the study of **conditional probability measures** and **conditional expectations,** but for the current application it is enough to recall example 3.42 of book 2.

In that example was developed an approach to defining the distribution function $F(x|y_0)$ from the bivariate joint distribution function $F(x, y)$ which is assumed to have a continuous joint density function, an assumption made throughout this section. Dropping the subscript on y, the book 2 derivation defined $F(x|y)$ as the limit:

$$F(x|y) \equiv \lim_{\Delta y \to 0} F(x, y|Y \in [y, y + \Delta y]).$$

Assuming that $\frac{\partial F(y)}{\partial y} \equiv f(y) \neq 0$, where $F(y)$, $f(y)$ are the associated marginal distribution and density functions, the result derived was:

$$F(x|y) = \frac{\partial F(x, y)}{\partial y} \Big/ \frac{\partial F(y)}{\partial y}, \qquad f(x|y) = f(x, y)/f(y).$$

The goal of this section is to apply this example's derivation to a conditional density function of interest, $f_{(i+1|i)}(x_{i+1}|x_i)$, the density function associated with the conditional distribution function $F_{(i+1|i)}(x_{i+1}|x_i)$ of $X_{(i+1)}$ given $X_{(i)}$. The same analysis can be applied to $F_{(j|i)}(x_j|x_i)$ for $j > i$ and is left as an exercise.

Proposition 2.20 *Given the joint distribution function $F_{(1,...,M)}(x_1, ..., x_M)$ in 2.7 and i with $1 \leq i < M$, the conditional distribution function $F_{(i+1|i)}(x_{i+1}|x_i)$ is:*

$$F_{(i+1|i)}(x_{i+1}|x_i) = 1 - \left(\frac{1 - F(x_{i+1})}{1 - F(x_i)}\right)^{M-i}, \tag{2.18}$$

and conditional density function $f_{(i+1|i)}(x_{i+1}|x_i)$ is:

$$f_{(i+1|i)}(x_{i+1}|x_i) = (M - i)f(x_{i+1})\frac{(1 - F(x_{i+1}))^{M-i-1}}{(1 - F(x_i))^{M-i}}. \tag{2.19}$$

Proof. *Formulaically, the marginal distribution function $F_{(i,i+1)}(x_i, x_{i+1})$ can be expressed as in 2.16 using the marginal density function $f_{(i,i+1)}(x_i, x_{i+1})$*

given in 2.13 for $x_i \leq x_{i+1}$:

$$F_{(i,i+1)}(x_i, x_{i+1}) = \int_{-\infty}^{x_i} \int_{x}^{x_{i+1}} f_{(i,i+1)}(x, y) dy dx.$$

As in the derivation in example 3.42 of book 2, using the fundamental theorem of calculus obtains:

$$\begin{aligned} F_{(i+1|i)}(x_{i+1}|x_i) &= \frac{\partial F_{(i,i+1)}(x_i, x_{i+1})}{\partial x_i} \Big/ \frac{\partial F_{(i,i+1)}(x_i, \infty)}{\partial x_i} \\ &= \int_{x_i}^{x_{i+1}} f_{(i,i+1)}(x_i, y) dy \Big/ \int_{x_i}^{\infty} f_{(i,i+1)}(x_i, y) dy \\ &= \int_{x_i}^{x_{i+1}} f(y) \left[1 - F(y)\right]^{M-i-1} dy \Big/ \int_{x_i}^{\infty} f(y) \left[1 - F(y)\right]^{M-i-1} dy \ , \end{aligned}$$

since the factorial constants, as well as the common factor of $f(x_i)F^{i-1}(x_i)$, *cancel from numerator and denominator. These integrals can be evaluated to produce 2.18.*

Since x_i *is implicitly fixed:*

$$F_{(i+1|i)}(x_{i+1}|x_i) = \int_{x_i}^{x_{i+1}} f_{(i+1|i)}(y|x_i) dy,$$

the associated density function in 2.19 can be obtained by differentiation of the above result,

$$f_{(i+1|i)}(x_{i+1}|x_i) = \frac{\partial}{\partial x_{i+1}} F_{(i+1|i)}(x_{i+1}|x_i),$$

or directly with $f_{(i+1|i)}(x_{i+1}|x_i) = f_{(i,i+1)}(x_i, x_{i+1}) / f_{(i)}(x_i)$, *and using 2.13 and 2.2.* ■

Not at all surprising is the fact that in general, $F_{(i+1|i)}(x_{i+1}|x_i)$ depends on the value of x_i. That is, the distribution function of $X_{(i+1)}$ depends on the value of $X_{(i)}$ and this is logically expected because it must be the case that $X_{(i+1)} \geq X_{(i)}$. More on this in the section below on the **Rényi representation theorem on order statistics,** but in the meantime we look at an example.

Example 2.21 *Assuming that* F *is the exponential distribution with parameter* λ, *then 2.18 becomes:*

$$F^E_{(i+1|i)}(x_{i+1}|x_i) = 1 - e^{-\lambda(M-i)(x_{i+1}-x_i)}.$$

Thus while the distribution function of $X_{(i+1)}$ depends on the value of $X_{(i)}$, the distribution function of the difference, $X_{(i+1)} - X_{(i)}$ does not. This formula states that whatever the value of $X_{(i)}$, the value of $X_{(i+1)}$ is given by:

$$X_{(i+1)} = X_{(i)} + Y_i,$$

where Y_i is exponentially distributed with parameter $\lambda(M-i)$. And this is true for all i.

*The remarkable insight in the development of the Rényi representation theorem below is that $\{Y_i\}_{i=0}^{M-1}$ so defined are **independent exponentials**.*

2.5 The Rényi Representation Theorem

As will be seen below in the section Simulating Samples of Random Variables - Examples, the standard theory for generating random samples provides a framework for generating ordered samples of a given random variable. However this approach is potentially costly in computer time to generate the necessary uniformly distributed Y-samples, and even more costly to evaluate the necessary inversions of the given distribution function F. An alternative approach using **ordered exponential variables** is developed in this section and is based on the **Rényi representation theorem on order statistics,** named for **Alfréd Rényi** (1921 – 1970).

The Rényi representation theorem derives the surprising conclusion that if F is an exponential distribution and $\{X_{(k)}\}_{k=1}^{M}$ its *kth* **order statistics,** so $X_{(k)} \leq X_{(k+1)}$, then $\{X_{(k+1)} - X_{(k)}\}_{k=0}^{M-1}$ are **independent,** exponentially distributed random variables. Here and below we define $X_{(0)} = 0$ to simplify notation. As it turns out, the independence of $\{X_{(k+1)} - X_{(k)}\}$ is unique to the exponential distribution and reflects the following insight.

Recall the definition of conditional probability of 1.27 of book 2, which if applied to the event $\Pr\{X \leq x + y | X > x\}$ for exponential X yields:

$$\begin{aligned} \Pr\{X &\leq x + y | X > x\} \\ &= \Pr\{x < X \leq x + y\} / \Pr\{X > x\} \\ &= \left[F^E(x+y) - F^E(x)\right] / \left[1 - F^E(x)\right]. \end{aligned}$$

With $F^E(x) = 1 - e^{-\lambda x}$, a calculation obtains:

$$\left[F^E(x+y) - F^E(x)\right] / \left[1 - F^E(x)\right] = F^E(y). \qquad ((*))$$

In other words, letting x denote the value of $X_{(k)}$, this calculation states that the distribution function of the excess of $X_{(k+1)}$ over $X_{(k)}$, here denoted y, is independent of $X_{(k)}$.

On an intuitive level it is clear that such a statement could not possibly be true for many distribution functions. Indeed, if $F^U(x) = x$ is the uniform distribution function, then a calculation produces:

$$\left[F^U(x+y) - F^U(x)\right] / \left[1 - F^U(x)\right] = \left[\min\left(x+y, 1\right) - x\right] / \left[1 - x\right],$$

and the distribution of allowable y values is compressed into $[0, 1-x]$.

More formally, it is known that the exponential distribution function is unique in this respect. First, there is a famous result proved by **Augustin-Louis Cauchy** (1789 – 1857) that if f is a continuous function on $\mathbb{R}$ that satisfies **Cauchy's functional equation:**

$$f(x+y) = f(x) + f(y),$$

then there is a constant c so that $f(x) = cx$. Defining $f(x) = \ln\left[1 - F(x)\right]$, then Cauchy's functional equation is equivalent to that in $(*)$:

$$F(x+y) - F(x) = F(y)\left(1 - F(x)\right),$$

and the conclusion is then $F(x) = 1 - e^{cx}$. This then proves that only the exponential distribution satisfies $(*)$.

On the other hand, it is in theory possible for a given distribution function $F(x)$ that x and y are again independent, but with:

$$\left[F(x+y) - F(x)\right] / \left[1 - F(x)\right] = F_1(y), \qquad ((**))$$

where $F_1(y)$ is a different distribution function. This would then not contradict Cauchy's result, but would provide another example for which the excess variable, y, was independent of x.

Exercise 2.22 *Show that if $(**)$ was valid and $F(x)$ is a differentiable distribution function, then $F(x)$ is again the exponential distribution and thus $F_1 = F$. Hint: By $(**)$ since independent of x, the x-derivative of $\left[F(x+y) - F(x)\right] / \left[1 - F(x)\right]$ is zero, and this obtains that $F'(x)/[1-F(x)]$ is constant. But this expression is $-\frac{d}{dx}\ln[1 - F(x)]$.*

We now prove Rényi's result and show that $\{X_{(k+1)} - X_{(k)}\}_{k=0}^{M-1}$ are independent exponential random variables with respective parameters $\{\lambda(M-k)\}_{k=0}^{M-1}$.

Proposition 2.23 (Rényi Representation Theorem) *Let $\{X_k\}_{k=1}^M$ denote independent random variables from an exponential distribution with parameter λ, and $\{X_{(k)}\}_{k=1}^M$ the associated ordered random variables. Then with $X_{(0)} \equiv 0$, $\{X_{(k+1)} - X_{(k)}\}_{k=0}^{M-1}$ are independent, exponentially distributed random variables with respective parameters $\{\lambda(M-k)\}_{k=0}^{M-1}$.*

Proof. *For this proof we will evaluate another integral of the joint density function, $f_{(1,...,M)}(x_1, ..., x_M)$ in 2.10. The goal is to show that the joint distribution function $G(a_1, a_2, ..., a_M)$ of $(X_{(1)}, X_{(2)} - X_{(1)}, ..., X_{(M)} - X_{(M-1)})$ satisfies:*

$$G(a_1, a_2, ..., a_M) = \prod_{k=0}^{M-1} G_k(a_{k+1}),$$

where $G_k(x)$ is the distribution function of an exponential with parameter $\lambda(M-k)$. This then proves that $\{X_{(k+1)} - X_{(k)}\}_{k=0}^{M-1}$ are independent random variables by proposition 3.53 of book 2.

Consider the first step in the integration, recalling the constraint that $x_k \geq x_{k-1}$ for all k :

$$\begin{aligned}
& G(a_1, a_2, ..., a_M) \\
\equiv\ & \Pr\left[X_{(M)} - X_{(M-1)} \leq a_M, ..., X_{(2)} - X_{(1)} \leq a_2,\ X_{(1)} \leq a_1\right] \\
=\ & M! \int_{-\infty}^{a_1} \cdots \int_{x_{M-2}}^{x_{M-2}+a_{M-1}} \left(\int_{x_{M-1}}^{x_{M-1}+a_M} f(x_M) dx_M \right) f(x_{M-1}) dx_{M-1} \cdots f(x_1) dx_1 \\
=\ & M! \int_{-\infty}^{a_1} .. \int_{x_{M-2}}^{x_{M-2}+a_{M-1}} \left[F(x_{M-1} + a_M) - F(x_{M-1})\right] f(x_{M-1}) dx_{M-1} .. f(x_1) dx_1.
\end{aligned}$$

Because F is exponential with parameter λ :

$$\begin{aligned}
\left[F(x_{M-1} + a_M) - F(x_{M-1})\right] f(x_{M-1}) &= \left[e^{-\lambda x_{M-1}} - e^{-\lambda(x_{M-1}+a_M)}\right] \lambda e^{-\lambda x_{M-1}} \\
&= \left[1 - e^{-\lambda a_M}\right] \lambda e^{-2\lambda x_{M-1}} \\
&= \frac{1}{2}\left[1 - e^{-\lambda a_M}\right] f_2(x_{M-1}),
\end{aligned}$$

where $f_2(x_{M-1})$ is the exponential density with parameter 2λ. Repeating this calculation, letting $f_k(x_{M-1})$ denote the exponential density with parameter

$k\lambda$:

$$
\begin{aligned}
& G(a_1, a_2, ..., a_M) \\
= & \frac{M!}{2}\left[1 - e^{-\lambda a_M}\right] \int_{-\infty}^{a_1} \left[\cdots \left[\int_{x_{M-2}}^{x_{M-2}+a_{M-1}} f_2(x_{M-1}) dx_{M-1}\right] \cdots\right] f(x_1) dx_1 \\
= & \frac{M!}{2}\left[1 - e^{-\lambda a_M}\right] \int_{-\infty}^{a_1} \left[\cdots \left[\frac{1}{3}\left[1 - e^{-2\lambda a_{M-1}}\right] f_3(x_{M-2})\right] \cdots\right] f(x_1) dx_1 \\
= & \frac{M!}{3!}\left[1 - e^{-\lambda a_M}\right]\left[1 - e^{-2\lambda a_{M-1}}\right] \times \\
& \int_{-\infty}^{a_1} \left[\cdots \int_{-\infty}^{x_{M-3}+a_{M-2}} f_3(x_{M-2}) dx_{M-2} \cdots\right] f(x_1) dx_1 \\
& \vdots \\
= & \frac{M!}{(M-1)!} \prod_{k=0}^{M-2}\left[1 - e^{-(M-k)\lambda a_{k+1}}\right] \int_{-\infty}^{a_1} f_{M-1}(x_1) dx_1 \\
= & \prod_{k=0}^{M-1}\left[1 - e^{-(M-k)\lambda a_{k+1}}\right].
\end{aligned}
$$

■

The essence of this representation theorem is that each of the *kth* order statistics, or any sequential grouping of *kth* order statistics, can be generated **sequentially** as a sum of independent exponential random variables. This is in contrast to the definitional procedure whereby the entire collection $\{X_j\}_{j=1}^{M}$ must be generated, then reordered to $\{X_{(j)}\}_{j=1}^{M}$ to identify each variate. To generate a larger collection requires more variates, $\{X_j\}_{j=M+1}^{M'}$, and then a complete reordering of the entire collection $\{X_j\}_{j=1}^{M'}$.

With Rényi's result we can directly generate $X_{(1)}$, then $X_{(2)}$, and so forth. For example, $X_{(1)}$ is exponential with parameter $M\lambda$, while $X_{(2)}$ is a sum of $X_{(1)}$ and an independent exponential variate with parameter $(M-1)\lambda$, and so forth. The following corollary simplifies this insight further, to allow the use of identically distributed exponentials with $\lambda = 1$, which are also called "standard" exponentials.

Exercise 2.24 *Prove that if E is an exponential variable with parameter $\lambda = 1$, then E/α is an exponential variable with parameter $\lambda = \alpha$.*

Corollary 2.25 (Rényi Representation Theorem) *Let $\{X_k\}_{k=1}^{M}$ denote independent random variables from an exponential distribution with parameter λ, and $\{X_{(k)}\}_{k=1}^{M}$ the associated ordered random variables. Then in*

distribution,

$$X_{(k)} = \sum_{j=1}^{k} \frac{E_j}{\lambda(M-j+1)}, \tag{2.20}$$

where $\{E_j\}_{j=1}^{M}$ *are independent standard exponential random variables,* $\lambda = 1$.

Proof. *By definition:*

$$X_{(k)} = \sum_{j=1}^{k} \left(X_{(j)} - X_{(j-1)}\right).$$

The result then follows from the prior proposition, and exercise 2.24. ■

The final result requires the results of the next section, Expectations of Random Variables 1, but is included here for completeness. If unfamiliar with these notions the reader should read ahead and come back to this result.

Corollary 2.26 (Rényi Representation Theorem) *Let* $\{X_k\}_{k=1}^{M}$ *denote independent random variables from an exponential distribution with parameter* λ, *and* $\{X_{(k)}\}_{k=1}^{M}$ *the associated ordered random variables. Then denoting by* $\mu_{(k)}$ *and* $\sigma^2_{(k)}$ *the mean and variance of* $X_{(k)}$*:*

$$\mu_{(k)} = \frac{1}{\lambda}\sum_{j=0}^{k-1}\frac{1}{M-j} = \frac{1}{\lambda}\sum_{j=M-k+1}^{M}\frac{1}{j}, \tag{2.21}$$

$$\sigma^2_{(k)} = \frac{1}{\lambda^2}\sum_{j=0}^{k-1}\frac{1}{(M-j)^2} = \frac{1}{\lambda^2}\sum_{j=M-k+1}^{M}\frac{1}{j^2}. \tag{2.22}$$

Further, with $M_{(k)}(t)$ *denoting the moment generating function of* $X_{(k)}$*:*

$$M_{(k)}(t) = \prod_{j=0}^{k-1}\left(1 - \frac{t}{\lambda\,(M-j)}\right)^{-1}, \qquad |t| < \lambda\,(M-k+1)\,. \tag{2.23}$$

Proof. *By the prior proposition* $X_{(k)}$ *is the sum of* k *independent exponentials with parameters* $\lambda\,(M-j)$ *for* $j = 0$ *to* $k-1$. *Thus these results follow from the section Moments of Sums of RVs, using 3.58 and 3.59 below.* ■

Remark 2.27 **1**. *Because* $\sum_{j=1}^{N}\frac{1}{j} \approx \ln N$ *as* $N \to \infty$, *we have that for* M *large:*

$$\mu_{(M)} \approx \frac{1}{\lambda}\ln M.$$

More generally, by comparing the series to the integral of $1/x$*:*

$$\ln N + 1/N < \sum_{j=1}^{N}\frac{1}{j} < \ln N + 1.$$

Thus:

$$\frac{1}{\lambda}\left[\ln M + 1/M\right] < \mu_{(M)} < \frac{1}{\lambda}\left[\ln M + 1\right],$$

and for $k < M$:

$$\frac{1}{\lambda}\left[\ln\left(\frac{M}{M-k}\right) - \frac{M-1}{M}\right] < \mu_{(k)} < \frac{1}{\lambda}\left[\ln\left(\frac{M}{M-k}\right) + \frac{M-k-1}{M-k}\right].$$

2. *The power of the* ***Rényi representation theorem*** *can be appreciated by attempting to derive the formulas in 2.21, 2.22 and 2.23 directly from the density function of* $X_{(k)}$ *in 2.2 and the formulas of the next section.*

3. *This representation also plays an important role in generating random samples of ordered exponentials variates in the section, Simulating Samples of Random Variables - Examples, as well as in the continued development in Extreme Value Theory II.*

Exercise 2.28 *Derive bounds for* $\sigma^2_{(k)}$ *as was done for* $\mu_{(k)}$ *by estimating bounds for* $\sum_{j=1}^{N} 1/j^2$.

Chapter 3

Expectations of Random Variables 1

In this section we begin the study of "expectations" of random variables and introduce some of their important properties. These properties can be fully realized with the current state of our theoretical development in the special cases of absolutely continuous and discrete distribution functions. But as will be reviewed below, even this development requires a small "leap of faith" regarding the fundamental definitions. This definitional ambiguity, as well as generalizations to general distribution functions can only be fully resolved with the more advanced integration theory of book 5.

3.1 General Definitions

We begin by introducing the definition of "expectation" of a random variable in the general case, and also in the special case of absolutely continuous and discrete distribution functions which will undoubtedly look familiar. It is then both necessary and important to identify an inherent ambiguity in this definition, and review the forthcoming mathematical tools of book 5 that will ultimately be used to put this definition on a solid, unambiguous footing in book 6.

For the following definition recall the development in chapter 4 of book 3 of the Riemann-Stieltjes integral. The next section discusses inherent but temporary ambiguities in this definition.

Definition 3.1 (Expectation) *Let $X : \mathcal{S} \to \mathbb{R}$ be a random variable defined on a probability space $(\mathcal{S},\mathcal{E},\lambda)$, and $F(x)$ the associated distribution*

function. If $g(x)$ is a Borel measurable function, the ***expectation of*** $g(X)$, *denoted $E([g(X)]$, is defined by the* ***Riemann-Stieltjes integral****:*

$$E[g(X)] = \int g(x)dF, \tag{3.1}$$

when

$$\int |g(x)|\, dF < \infty. \tag{3.2}$$

Remark 3.2 *1. On the matter of existence, since $F(x)$ is increasing and bounded, proposition 4.24 of book 3 assures that this integral exists for $g(x)$ continuous and bounded, or, of bounded variation as long is any discontinuity points of $g(x)$ are then continuity points of $F(x)$. The existence theory applies to $|g(x)|$ by proposition 4.26. Of course boundedness of $g(x)$ is a big restriction, but we also know that at least in the special cases of integrators addressed in proposition 4.30 of book 3, that this integral will also exist for unbounded integrands under conditions discussed below.*

2. Note that the above definition is equally applicable when $X : \mathcal{S} \to \mathbb{R}^n$ is a random vector defined on a probability space $(\mathcal{S},\mathcal{E},\lambda)$, $F(x)$ the associated joint distribution function, and $g : \mathbb{R}^n \to \mathbb{R}$ a Borel measurable function using the theory and results from section 4.2 of book 3.

By 1.1 any such distribution function can be decomposed:

$$F(x) = F_{SLT}(x) + F_{AC}(x) + F_{SN}(x).$$

At least in the case of $g(x)$ continuous and bounded, the book 3 existence theory of proposition 4.24 applies to each of the Riemann-Stieltjes integrals defined with respect to the three component functions, and then proposition 4.26 assures that for such $g(x)$:

$$\int g(x)dF = \int g(x)dF_{SLT} + \int g(x)dF_{AC} + \int g(x)dF_{SN}.$$

Further,

$$F_{SLT}(x) \equiv \sum\nolimits_{x_n \le x} f_{SLT}(x_n),$$

with $f_{SLT}(x_n) \equiv F(x_n) - F(x_n^-)$ and defined on the at most countably many discontinuities $\{x_n\}$ of $F(x)$, while $F_{AC}(x)$ is absolutely continuous and thus

$f_{AC}(x) \equiv F'_{AC}(x)$ exists almost everywhere, is measurable, and:

$$F_{AC}(x) = (\mathcal{L}) \int_{-\infty}^{x} f_{AC}(y)dy.$$

In the special case when $F'_{AC}(x)$ exists everywhere and is continuous, the integral representation for $F_{AC}(x)$ is valid as a Riemann integral. For these special component functions in the decomposition of $F(x)$, the Riemann-Stieltjes integral can be recast by proposition 4.30 of book 3.

Definition 3.3 (Expectation - Special Cases) *Let $X : \mathcal{S} \to \mathbb{R}$ be a random variable defined on a probability space $(\mathcal{S},\mathcal{E},\lambda)$ and $F(x)$ the associated distribution function assumed to be of the form*

$$F(x) = F_{SLT}(x) + F_{AC}(x),$$

*where $f_{AC}(x) \equiv F'_{AC}(x)$ is continuous, and the collection $\{x_n\}$ which define $f_{SLT}(x)$ have no accumulation points. If $g(x)$ is a continuous function, the **expectation of** $g(X)$, denoted $E([g(X)]$, is defined by:*

$$E[g(X)] = \sum\nolimits_n g(x_n) f_{SLT}(x_n) + (\mathcal{R}) \int_{-\infty}^{\infty} g(x) f_{AC}(x) dx, \tag{3.3}$$

when

$$\sum\nolimits_n |g(x_n)| \, f_{SLT}(x_n) + (\mathcal{R}) \int_{-\infty}^{\infty} |g(x)| \, f_{AC}(x) dx < \infty. \tag{3.4}$$

In many applications using the distribution functions such as those defined above, only one of $F_{SLT}(x)$ or $F_{AC}(x)$ will be present and thus $E[g(X)]$ will be defined in terms of only one of the components in 3.3.

The following result is an immediate application of proposition 4.26 of book 3.

Proposition 3.4 *If $X : \mathcal{S} \to \mathbb{R}$ is a random variable defined on a probability space $(\mathcal{S},\mathcal{E},\lambda)$ with associated distribution function $F(x)$, and g and h Borel measurable functions for which $E[g(x)]$ and $E[h(x)]$ are well defined in the sense of 3.4, then for any real constants a, b, we have that $E[ag(x) + bh(x)]$ exists and:*

$$E[ag(x) + bh(x)] = aE[g(x)] + bE[h(x)]. \tag{3.5}$$

Proof. *Once we prove that $|ag(x) + bh(x)|$ is integrable and thus $E[ag(x) + bh(x)]$ exists, the equality in 3.5 follows from proposition 4.26 of book 3. For existence we have by the triangle inequality that:*

$$|ag(x) + bh(x)| \leq |a| \, |g(x)| + |b| \, |h(x)| \, ,$$

and hence linearity of the integral and integrability of $|g(x)|$ *and* $|h(x)|$ *implies the result.* ■

3.1.1 Is Expectation Well Defined?

Undoubtedly the definition of $E[g(x)]$ in the special cases of definition 3.3 is quite familiar to students of probability theory, even if the more unified Riemann-Stieltjes approach of definition 3.1 is perhaps new. However, a closer examination of the above definitions reveals a definitional ambiguity, and potential concern that $E[g(X)]$ may not be well defined. This ambiguity is sometimes ignored by authors of introductory texts because as we will see below, the resolution involves advanced notions that would likely be outside such texts' mathematical tool kits. The ambiguity identified below is equally applicable whether X is a random variable or random vector, but equally resolvable with the integration theory of book 5. In this book we focus on results for random variables and defer the more general discussion to book 6.

Here is the problem. Note that if X is a random variable on $\mathcal{S}$ then so too is $Y \equiv g \circ X$ as the composition of measurable $X : \mathcal{S} \to \mathbb{R}$ and Borel measurable $g : \mathbb{R} \to \mathbb{R}$. Let $F_Y(y)$ denote the distribution function of Y, so

$$F_Y(y) = \lambda[(g \circ X)^{-1}(-\infty, y]] = \lambda[X^{-1}[g^{-1}(-\infty, y]]].$$

Of course, $F_Y(y)$ is well defined because $g^{-1}(-\infty, y] \in \mathcal{B}(\mathbb{R})$ so $X^{-1}[g^{-1}(-\infty, y]] \in \mathcal{E}$. Also, $F_Y(y)$ is an increasing, right continuous function, as was $F_X(x)$.

We now have two approaches to the definition of expectation of $Y \equiv g(X)$:

1. As a function of X :

$$E[g(X)] = \int g(x) dF_X.$$

2. As the random variable Y :

$$E[Y] = \int y dF_Y.$$

Of course the outstanding question is, must it be the case that both integrals exist or don't exist together, and furthermore, when they both exist, must

$$\int g(x) dF_X = \int y dF_Y \ ? \tag{3.6}$$

For the special cases of definition 3.3, we can derive insights to the validity of 3.6.

- $F_X(x) = F^X_{SLT}(x)$ **a Saltus Function, and** $g(x)$ **monotonic.**

  Assume that $F_X(x)$ is defined by $\{x_n\}$ and probabilities $\{f_X(x_n)\}$ and define $y_n \equiv g(x_n)$. If $g(x)$ is **increasing** then $\{y_n\}$ is an increasing sequence and so:

  $$F_Y(y_n) \equiv \lambda[X^{-1}[g^{-1}(-\infty, g(x_n)]]] = \lambda[X^{-1}[(-\infty, x_n]]] = F_X(x_n).$$

  Thus

  $$f_Y(y_n) \equiv F_Y(y_n) - F_Y(y_{n-1}) = f_X(x_n),$$

  When $g(x)$ is **decreasing** then $\{y_n\}$ is a decreasing sequence and so:

  $$1 - F_Y(y_n) \equiv \lambda[X^{-1}[g^{-1}(g(x_n), \infty)]] = \lambda[X^{-1}[(-\infty, x_n)]] = F_X(x_{n-1}).$$

  But now

  $$f_X(x_n) = F_X(x_n) - F_X(x_{n-1}) = F_Y(y_n) - F_Y(y_{n+1}) = f_Y(y_n).$$

  In either case by 3.3:

  $$\int g(x) dF_X = \sum\nolimits_n g(x_n) f_X(x_n) = \sum\nolimits_n y_n f_Y(y_n) = \int y dF_Y.$$

- $F_X(x) = F^X_{AC}(x)$ **a Continuously Differentiable Function and so** $f_{AC}(x)$ **is continuous, and** $g(x)$ **monotonic and continuously differentiable with** $g'(x) \neq 0$ **for all** x**.**

  Both cases can be accommodated with the aid of 1.42, that:

  $$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right|.$$

  Thus as Riemann integrals with substitution $x = g^{-1}(y)$, noting that for decreasing $g(x)$ the use of $\left|\frac{dg^{-1}(y)}{dy}\right|$ eliminates the need to reverse the limits of integration:

  $$\begin{aligned} E[g(X)] &= \int g(x) f_X(x) dx \\ &= \int y f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right| dy \\ &= \int y f_Y(y) dy = E[Y]. \end{aligned}$$

Exercise 3.5 *Show that with $F_X(x)$ a given general distribution function, and $g(x)$ monotonically increasing and continuously differentiable with $g'(x) \neq 0$ for all x, then 3.6 is satisfied with $Y \equiv g(X)$. Hint: From 1.38 $F_Y(y) = F_X(g^{-1}(y))$. Investigate the Riemann-Stieltjes summations for $\int y dF_Y$, approximating ΔF_Y with ΔF_X and noting that if $\{y_j\}$ is a partition for the dF_Y-integral, then by continuity of $g^{-1}(y)$ it follows that $\{g^{-1}(y_j)\}$ is a partition for the dF_X-integral, and mesh sizes go to zero together. To simplify first investigate $\int y dF_Y$ defined over $[a, b]$ rather than $\mathbb{R}$, and determine the limits of integration of the dF_X integral, then generalize.*

3.1.2 Formal Resolution of Well-Definedness

While the well-definedness question has been partially answered by the results of the prior section, it is clear that we are still a long way from a statement for general $F_X(x)$ and general Borel measurable $g(x)$, the minimal requirement needed to ensure that $Y = g(X)$ is a random variable on $(\mathcal{S},\mathcal{E},\lambda)$. The resolution will be formalized in book 6 using the following results that will be developed in book 5.

1. **General Definition:**

   If $X : \mathcal{S} \to \mathbb{R}$ is a random variable defined on a probability space $(\mathcal{S},\mathcal{E},\lambda)$ we will formally define:

   $$E[X] = \int_{\mathcal{S}} X(s) d\lambda(s), \tag{3.7}$$

   when

   $$\int_{\mathcal{S}} |X(s)| \, d\lambda(s) < \infty. \tag{3.8}$$

   This definition requires the development of an integration theory on the measure space $(\mathcal{S},\mathcal{E},\lambda)$ which will be addressed in the second chapter of book 5. But once done, the above ambiguities disappear. Defining the new random variable $Y \equiv g(X)$ with Borel measurable $g(x)$:

   $$\begin{aligned} E[Y] &\equiv \int_{\mathcal{S}} Y(s) d\lambda(s) \\ &\equiv \int_{\mathcal{S}} g(X(s)) d\lambda(s) \equiv E[g(X)]. \end{aligned}$$

   Because the general definition does not even mention the distribution function of the variable we are integrating, it matters not whether we consider Y as the random variable, or X as the random variable which is then composed with $g(x)$. On $\mathcal{S}$, $Y(s) \equiv g(X(s))$.

2. Change of Variables I - Transformation from $\mathcal{S}$ to $\mathbb{R}$:

While this approach circumvents the apparent definitional problem, it raises the question of how in any given application one actually evaluates such an integral on $\mathcal{S}$. If $F(x)$ is the distribution function associated with X, and μ_{F_X} the associated Borel measure on $\mathbb{R}$ as developed in chapter 5 of book 1, we will show that for any measurable function g,

$$\int_{\mathcal{S}} g(X(s))d\lambda(s) = \int_{\mathbb{R}} g(x)d\mu_{F_X}, \tag{3.9}$$

where the integral on the right is a **Lebesgue-Stieltjes integral.** This integral was briefly discussed in chapter 4 of book 3, and is named for **Henri Lebesgue** (1875 – 1941) and **Thomas Stieltjes** (1856 – 1894). That such a Lebesgue-Stieltjes integral is well defined will follow from the general theory of integration applied to the Borel measure space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu_{F_X})$. This integral will be well defined for any distribution function $F(x)$.

Applying 3.9 to the Y-integral obtains:

$$\int_{\mathcal{S}} Y(s)d\lambda(s) = \int_{\mathbb{R}} yd\mu_{F_Y},$$

and thus as integrals on $\mathbb{R}$ this implies a change of variables result that for $y = g(x)$;

$$\int_{\mathbb{R}} g(x)d\mu_{F_X} = \int_{\mathbb{R}} yd\mu_{F_Y}.$$

When $g(x)$ is continuous it will turn out that Lebesgue-Stieltjes and Riemann-Stieltjes integrals agree, and thus for example:

$$\int_{\mathbb{R}} g(x)d\mu_F = \int_{\mathbb{R}} g(x)dF,$$

with the integral on the right defined as a Riemann-Stieltjes integral of book 3.

3. Change of Variables II - Transformation from $d\mu_F$ to dx for Special Distributions:

In this last step we derive the above formulas for $E[g(x)]$ in the special case where $F(x)$ has no singular component.

(**a**) In the special case where $F(x) = F_{AC}(x)$ is absolutely continuous, then with $f_{AC}(x) \equiv F'_{AC}(x)$ and Lebesgue measurable $g(x)$,

$$\int_{\mathbb{R}} g(x)d\mu_F = (\mathcal{L}) \int_{\mathbb{R}} g(x)f_{AC}(x)dx,$$

where the integral on the right is defined as a Lebesgue integral. Then from proposition 2.64 of book 3 it will follow that if $f_{AC}(x)$ and $g(x)$ are continuous, then the integral on the right can also be defined as a Riemann integral, which is the usual set-up in "continuous" probability theory. In this case the integral on the left is the **Riemann-Stieltjes integral** $\int_{\mathbb{R}} g(x)dF$ as noted above, and 3.10 is proposition 4.30 of book 3:

$$\int_{\mathbb{R}} g(x)dF = (\mathcal{R}) \int_{\mathbb{R}} g(x)f_{AC}(x)dx. \tag{3.10}$$

(**b**) In the special case where $F(x) = F_{SLT}(x)$ is discrete with discontinuity set $\{x_n\}$, then with $f_{SLT}(x_n) \equiv F(x_n) - F(x_n^-)$ and continuous $g(x)$:

$$\int_{\mathbb{R}} g(x)d\mu_F = \int_{\mathbb{R}} g(x)dF = \sum\nolimits_n g(x_n)f_{SLT}(x_n), \tag{3.11}$$

and 3.11 is proposition 4.30 of book 3.

(**c**) Consequently, when $F(x) = F_{SLT}(x) + F_{AC}(x)$, the formula in 3.3 is produced by proposition 4.26 of book 3.

Remark 3.6 *The above program of study will also apply when $X : \mathcal{S} \to \mathbb{R}^n$ is a random vector defined on a probability space $(\mathcal{S},\mathcal{E},\lambda)$, $F(x)$ is the associated joint distribution function, and $g : \mathbb{R}^n \to \mathbb{R}$ a Borel measurable function. In the section Moments of Sums of RVs below, we will review the multivariate version of steps 2 and 3..*

3.2 Moments of Distributions

As may be recalled from past experiences in probability theory, there are a number of special expectation values which are commonly defined relative to the functions $g(x) = x^n$ and $g(x) = (x - a)^n$ for positive integer n, and for a special value of a. It is common to then refer to these expectations as **the moments of the distribution,** and sometimes, **the moments of**

the random variable. Since such functions are not bounded there is in general no applicable existence theory from book 3 and thus it is important to note that such moments need not exist. We will present these definitions in the general form of the above Riemann-Stieltjes integral in 3.1 to simplify notation, and note that in the special but common cases of continuously differentiable and/or discrete distribution functions, these definitions transform to the familiar results involving Riemann integrals and/or summations as in 3.3.

3.2.1 Common Types of Moments

There are three types of moments commonly defined. The first two are more common and with well established notational conventions, while the third is more specialized and used primarily for moment inequalities.

1. Moments About the Origin

Sometimes referred to as the **raw moments** or simply the **moments,** these are the expectations defined relative to the function $g(x) = x^n$.

Definition 3.7 *If $X : \mathcal{S} \to \mathbb{R}$ is a random variable defined on a probability space $(\mathcal{S},\mathcal{E},\lambda)$ with associated distribution function $F(x)$, the nth **moment,** denoted μ'_n, is defined by:*

$$\mu'_n \equiv \int_{\mathbb{R}} x^n dF, \tag{3.12}$$

when 3.2 is satisfied, and undefined otherwise.

*When $n = 1$, μ'_1 is called the **mean of the distribution** F and denoted by μ :*

$$\mu \equiv \mu'_1. \tag{3.13}$$

2. Central Moments

The **central moments** are defined with $g(x) = (x-\mu)^n$, where μ denotes the mean of the distribution.

Definition 3.8 *If $X : \mathcal{S} \to \mathbb{R}$ is a random variable defined on a probability space $(\mathcal{S},\mathcal{E},\lambda)$ with associated distribution function $F(x)$, the nth **central moment,** denoted μ_n, is defined by:*

$$\mu_n \equiv \int_{\mathbb{R}} (x - \mu)^n dF, \tag{3.14}$$

when 3.2 is satisfied, and undefined otherwise.

When $n = 2$, μ_2 is called the ***variance of the distribution*** *F, and denoted by σ^2,*

$$\sigma^2 \equiv \mu_2, \tag{3.15}$$

and $\sigma \equiv \sqrt{\mu_2}$, the positive square root, is called the ***standard deviation of the distribution*** *F.*

3. Absolute Moments

There are both **absolute moments** and **absolute central moments** defined respectively in terms of $g(x) = |x|^n$ and $g(x) = |x - \mu|^n$. Of course, the absolute value is redundant when n is an even integer. By definition, these moments exist whenever the associated moments and central moments exist due to the constraint in 3.2. There is no standard notation for these moments, but $\mu'_{|n|}$ and $\mu_{|n|}$ seem self-explanatory and will be used in this text.

3.2.2 Moment Generating Function

In contrast to the above moment definitions which produce numerical values, the moment generating function is defined as an expectation which produces a function. This is accomplished by choosing $g(x)$ to have the form, $g(x) = e^{tx}$, and thus the moment generating function is parametrized in t.

Definition 3.9 *If $X : \mathcal{S} \to \mathbb{R}$ is a random variable defined on a probability space $(\mathcal{S},\mathcal{E},\lambda)$ with associated distribution function $F(x)$, the* ***moment generating function of*** *X, denoted $M_X(t)$, is defined by:*

$$M_X(t) \equiv \int_{\mathbb{R}} e^{tx} dF(x), \tag{3.16}$$

for any t for which the integral is finite.

Notation 3.10 *It is common is probability theory to abbreviate moment generating function by m.g.f. in the same way as p.d.f. and d.f. are used as abbreviations for the probability density function and distribution function.*

It is also common to denote $M_X(t)$ by $M(t)$ when the random variable is obvious from the context, or by $M_F(t)$ when one wants to highlight the distribution function.

Remark 3.11 *When written as a Lebesgue-Stieltjes integral as in 3.9, the integral in 3.16 is related to the* ***bilateral*** *or* ***two-sided Laplace transform of the induced Borel measure*** *μ_F, and named for* ***Pierre-Simon***

***Laplace** (1749 – 1827). However, it is common to use the exponential e^{-tx} in this definition and then define this function on complex $t = a + ib$. Thus the moment generating function is the two-sided Laplace transform of the Borel measure μ_F restricted to the real numbers, and with reversed orientation.*

*Written as a **Riemann-Stieltjes integral** with F continuously differentiable and thus with a continuous density function f, 3.16 can be expressed as in 3.3 with the familiar formula:*

$$M_X(t) \equiv (\mathcal{R}) \int_{\mathbb{R}} e^{tx} f(x) dx. \tag{3.17}$$

When F is a discrete distribution function with discontinuities on $\{x_i\}_{i=1}^{\infty}$, then in the most common case where these points have no accumulation point 3.16 can again be expressed as in 3.3 with the familiar formula:

$$M_X(t) \equiv \sum\nolimits_{i=1}^{\infty} e^{tx_i} f(x_i). \tag{3.18}$$

By splitting the integral in 3.16 as follows:

$$M_X(t) = \int_{-\infty}^{0} e^{tx} dF(x) + \int_{0}^{\infty} e^{tx} dF(x),$$

it is apparent that the first integral exists for all $t \geq 0$ since $\int_{-\infty}^{0} dF(x) = F(0)$ and $0 < e^{tx} < 1$, and similarly the second integral exists for all $t \leq 0$. But for the general case, it is not at all obvious that $M_X(t)$ exists for some interval about the origin, $(-t_0, t_0)$, which is essential for important results in the next section. While $M_X(0)$ exists for any $F(x)$:

$$M_X(0) = \int_{\mathbb{R}} dF(x) \ = 1,$$

it is possible that $M_X(t)$ exists only for $t = 0$.

Exercise 3.12 *Using 3.17 and the continuous density function for the log-normal distribution in 1.34 defined on $[0, \infty)$, show that $M_{LN}(t)$ exists only for $t = 0$. See also remark 3.30*

Proposition 3.13 *Recalling the above splitting of the integral defining $M_X(t)$ into negative and positive domains of integration, if the first integral exists for some $t_0' < 0$ and the second exists for some $t_0'' > 0$, then $M_X(t)$ is well defined on $(-t_0, t_0)$ for $t_0 = \min[|t_0'|, t_0'']$.*

Proof. *Note that if* $\int_{-\infty}^{0} e^{t_0' x} dF(x) < \infty$, *then for* $t_0' < t \leq 0$ *it follows from* $e^{tx} \leq e^{t_0' x}$ *for* $x \leq 0$ *and proposition 4.26 of book 3 that* $\int_{-\infty}^{0} e^{tx} dF(x) < \infty$. *And as noted above, the second integral automatically exists for all* $t \leq 0$, *and so* $M_X(t)$ *exists on* $(t_0', 0]$. *Similarly, if the second integral exists for some* $t_0'' > 0$ *it follows that* $M_X(t)$ *exists on* $[0, t_0'')$, *and the result follows.* ■

Finally, we note a simple but useful result which is an application of 3.5.

Proposition 3.14 *If* $M_X(t)$ *exists for* $t \in (-t_0, t_0)$, *then with* $Y \equiv aX + b$ *for* $a, b \in \mathbb{R}$, $M_Y(t)$ *exists for* $t \in (-t_0/|a|, t_0/|a|)$, *and:*

$$M_Y(t) = e^{bt} M_X(at). \tag{3.19}$$

Proof. *By definition and 3.5:*

$$M_Y(t) = E\left[e^{(aX+b)t}\right] = e^{bt} E\left[e^{Xat}\right],$$

and the result follows. ■

3.2.3 Moments of Sums of RVs

Theory

If $X_i : \mathcal{S} \to \mathbb{R}$ are random variables defined on a probability space ($\mathcal{S}$,$\mathcal{E}$,λ) for $i = 1, 2, ..., n$, we are interested in the moments of $\sum_{i=1}^{n} X_i$. More generally, if $g : \mathbb{R}^n \to \mathbb{R}$ is a Borel measurable function, we are interested in the expectation of $g(X_1, X_2, ..., X_n)$. For such general situations we again require the tools of book 5 and specifically, the multivariate version of the development of $E[g(X)]$ above. We summarize here to provide an intuitive context for the calculations below.

1. **Definition:**

   There is no change to the general definition given above. On the probability space ($\mathcal{S}$,$\mathcal{E}$, λ), if the **random vector** $X : \mathcal{S} \longrightarrow \mathbb{R}^n$ is defined by:

   $$X(s) = (X_1(s), X_2(s), ..., X_n(s)),$$

   then for Borel measurable $g : \mathbb{R}^n \to \mathbb{R}$ it follows that $g(X) : \mathcal{S} \longrightarrow \mathbb{R}$ is a random variable. Thus by 3.3:

   $$E[g(X)] \equiv \int_{\mathcal{S}} g(X(s)) d\lambda$$

when

$$\int_{\mathcal{S}} |g(X(s))|\, d\lambda < \infty,$$

and $E[g(X)]$ is undefined when this constraint is not satisfied.

Remark 3.15 *In the special case when $g(X)$ is a summation:*

$$g(X) = \sum\nolimits_{i=1}^{n} X_i,$$

the linearity property of integrals on $\mathcal{S}$ will obtain that for ***all random variables****:*

$$E\left[\sum\nolimits_{i=1}^{n} X_i\right] = \sum\nolimits_{i=1}^{n} E\left[X_i\right]. \tag{3.20}$$

2. Change of Variables I - Transformation from $\mathcal{S}$ to $\mathbb{R}^n$:

In order to evaluate such expectations it is necessary to move this integral to $\mathbb{R}^n$. Using the same mathematical result as in the one dimensional case, it will turn out that with $F(x_1, x_2, ..., x_n)$ defined as the joint distribution function of $(X_1, X_2, ..., X_n)$, that:

$$\int_{\mathcal{S}} g(X(s))d\lambda = \int_{\mathbb{R}^n} g(x_1, x_2, ..., x_n)d\mu_F^n, \tag{3.21}$$

where μ_F^n is the Borel measure on $\mathbb{R}^n$ induced by F. The integral over $\mathbb{R}^n$ is again a **Lebesgue-Stieltjes integral** as was the case for the one dimensional result.

Recall from chapter 8 of book 1 that the measure μ_F^n is defined on rectangles of the form $\prod_{j=1}^{n}(-\infty, x_j]$ by:

$$\mu_F^n\left[\prod\nolimits_{j=1}^{n}(-\infty, x_j]\right] \equiv F(x_1, x_2, ..., x_n),$$

while the measure of a bounded measurable rectangle is given as in 8.7 of book 1:

$$\mu_F\left[\prod\nolimits_{i=1}^{n}(a_i, b_i]\right] = \sum\nolimits_{x} sgn(x)F(x). \tag{3.22}$$

Each $x = (x_1, ..., x_n)$ in the summation is one of the 2^n vertices of $\prod_{i=1}^{n}(a_i, b_i]$, so each $x_i = a_i$ or $x_i = b_i$, and $sgn(x)$ is defined as -1 if the number of a_i-components of x is odd and $+1$ otherwise. This basic set function then extends to a product measure μ_F^n on the Borel sigma algebra $\mathcal{B}(\mathbb{R}^n)$ as well as to a complete sigma algebra denoted $\mathcal{M}_F(\mathbb{R}^n)$.

3. **Change of Variables II - Transformation from $d\mu_F^n$ to $\prod_{j=1}^n dx_i$ for Special Distributions:**

Similar to 3.10, in the special case where there exists a measurable multivariate density function $f(x_1, x_2, ..., x_n)$ so that

$$F(x_1, x_2, ..., x_n) = (\mathcal{L}) \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f(y_1, y_2, ..., y_n) dy,$$

where dy denotes the Lebesgue product measure on $\mathbb{R}^n$, then

$$\int_{\mathbb{R}^n} g(x_1, x_2, ..., x_n) d\mu_F^n = (\mathcal{L}) \int_{\mathbb{R}^n} g(x_1, x_2, ..., x_n) f(x_1, x_2, ..., x_n) dx. \tag{3.23}$$

A similar transformation is valid when $F(x)$ is a discrete multivariate distribution function, replacing the Lebesgue integral with a summation.

When $f(x)$ and $g(x)$ are continuous, the integral on the right in 3.23 is definable as a Riemann integral, and thus by proposition 4.75 of book 3:

$$\int_{\mathbb{R}^n} g(x_1, x_2, ..., x_n) d\mu_F^n = \int_{\mathbb{R}^n} g(x_1, x_2, ..., x_n) dF,$$

defined as a **Riemann-Stieltjes integral.**

4. **Iterated Integrals and Independent Random Variables:**

There is one additional step needed to evaluate multivariate integrals. The Lebesgue integral in 3.23 can be expressed as an "iterated" integral:

$$\begin{aligned} &\int_{\mathbb{R}^n} g(x_1, x_2, ..., x_n) f(x_1, x_2, ..., x_n) dx \qquad (3.24) \\ = &\int_{\mathbb{R}} \cdots \int_{\mathbb{R}} g(x_1, x_2, ..., x_n) f(x_1, x_2, ..., x_n) dx_1 dx_2 ... dx_n. \end{aligned}$$

This important result from book 5 is **Fubini's theorem, and** named for **Guido Fubini** (1879 – 1943). It states that given the constraint in 3.4, this integral over $\mathbb{R}^n$ can be evaluated in an iterated fashion, one variable at a time, and in any order. This is especially useful in the case of **independent random variables**.

Recall from book 2 that the distribution function of independent $\{X_i\}_{i=1}^n$ is given in proposition 3.53 by:

$$F(x_1, x_2, ..., x_n) = \prod_{j=1}^n F_j(x_j),$$

and in the case of discrete or absolutely continuous distribution functions, this identity extends to an identity in probability density functions:

$$f(x_1, x_2, ..., x_n) = \prod_{j=1}^{n} f(x_j).$$

Thus given $g(x_1, x_2, ..., x_n)$ and continuous densities:

$$E\left[g(X_1, X_2, ..., X_n)\right] = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} g(x_1, x_2, ..., x_n) \prod_{j=1}^{n} f(x_j) dx_1 dx_2 ... dx_n. \tag{3.25}$$

Applications

For independent random variables the evaluation of moments is now relatively simple in theory using 3.20 or more generally 3.25, though some calculations may be complicated from a combinatorial point of view. We provide more detail on the application of this result to the calculation of the mean of a sum and the variance of an independent sum. In the case of variance, we also derive the general formula without the assumption of independence.

Notation 3.16 *In an attempt at notational clarity we denote by $E^{(n)}$ the expectation of a multivariate function relative to the joint distribution function as defined in 3.24 or its discrete counterpart, and by E an expectation of any one of the individual random variables relative to its marginal distribution function as in 3.10 or its discrete counterpart. Many books avoid this detail.*

1. **Mean:**

   Let $X = \sum_{i=1}^{n} X_i$ where $\{X_i\}$ are random variables defined on $(\mathcal{S},\mathcal{E},\lambda)$ for $i = 1, 2, ..., n$. Using 3.20 obtains that $E^{(n)}[X] = \sum_{i=1}^{n} E^{(n)}[X_i]$. In the case of independent variates, from 3.25 and $\int_{\mathbb{R}} f(x_j)dx_j = 1$ obtains:

   $$\begin{aligned} E^{(n)}[X_i] &= \int_{\mathbb{R}} x_i f(x_i) dx_i \prod_{j \neq i} \int_{\mathbb{R}} f(x_j) dx_j \\ &= \mu_i, \end{aligned}$$

   where $\mu_i = E[X_i]$ denotes the mean of X_i as a random variable on $\mathcal{S}$. Hence, for **independent random variables:**

   $$E^{(n)}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E\left[X_i\right] = \sum_{i=1}^{n} \mu_i. \tag{3.26}$$

This same formula applies for variates which are **not independent**, but requires the notion of a marginal density function defined above. Applying 3.24:

$$\begin{aligned} E^{(n)}[X_i] &= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} x_i f(x_1, x_2, ..., x_n) dx_1 dx_2 ... dx_n \\ &= \int_{\mathbb{R}} x_i f(x_i) dx_i \end{aligned}$$

as before. This calculation follows because the density for X_i is the *ith* **marginal density function of** $f(x_1, x_2, ..., x_n)$ as in 2.12:

$$f(x_i) = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f(x_1, x_2, ..., x_n) \prod_{j \neq i} dx_j.$$

2. **mth Moment:**

Using the so-called **multinomial theorem**, which is to be proved in exercise 3.17,

$$\left(\sum_{i=1}^n X_i\right)^m = \sum_{m_1, m_2, .. m_n} \frac{m!}{m_1! m_2! ... m_n!} X_1^{m_1} X_2^{m_2} ... X_n^{m_n}, \quad (3.27)$$

where this summation is over all distinct n-tuples $(m_1, m_2, ..m_n)$ with $m_j \geq 0$ and $\sum_{j=1}^n m_j = m$. Hence, using the same approach as for the mean, it follows that for independent variates,

$$E^{(n)}\left[\left(\sum_{i=1}^n X_i\right)^m\right] = \sum_{m_1, m_2, .. m_n} \frac{m!}{m_1! m_2! ... m_n!} \mu_{m_1}^{(1)\prime} \mu_{m_2}^{(2)\prime} \cdots \mu_{m_n}^{(n)\prime}, \quad (3.28)$$

where $\mu_{m_i}^{(i)\prime}$ is the $m_i th$ moment of X_i, with $\mu_0^{(i)\prime} = 1$.

Exercise 3.17 *Prove the formula in 3.27 using induction on m.*

3. **mth Central Moment:**

Exactly as for the *mth* moments:

$$E^{(n)}\left[\left(\sum_{i=1}^n (X_i - \mu_i)\right)^m\right] = \sum_{m_1, m_2, .. m_n} \frac{m!}{m_1! m_2! ... m_n!} \mu_{m_1}^{(1)} \mu_{m_2}^{(2)} \cdots \mu_{m_n}^{(n)}, \quad (3.29)$$

where this summation is over all distinct n-tuples $(m_1, m_2, ..m_n)$ with $m_j \geq 0$ and $\sum_{j=1}^n m_j = m$. Here $\mu_{m_i}^{(i)}$ is the $m_i th$ central moment of X_i,with $\mu_0^{(i)} = 1$ and $\mu_1^{(i)} = 0$.

4. Variance for Independent Variates:

With $X = \sum_{i=1}^{n} X_i$, it follows that $X - E^{(n)}[X] = \sum_{i=1}^{n} (X_i - \mu_i)$ and so:

$$\left(X - E^{(n)}[X]\right)^2 = \sum\nolimits_{i=1}^{n} (X_i - \mu_i)^2 + 2 \sum\nolimits_{j<i} \left(X_j - \mu_j\right)(X_i - \mu_i).$$

By 3.25 applied to independent random variables, the variance of a sum of **independent random variables** is given:

$$Var\left[\sum\nolimits_{i=1}^{n} X_i\right] = \sum\nolimits_{i=1}^{n} \sigma_i^2. \tag{3.30}$$

This follows by iterating integrals because for $i < j$, after integrating the densities $f(x_k)$ to 1 for $k \neq i, j$:

$$\begin{aligned} &\int \left(X_j - \mu_j\right)(X_i - \mu_i) f(x_i) f(x_j) dx_i dx_j \\ = &\int \left(X_j - \mu_j\right) f(x_j) dx_j \int (X_i - \mu_i) f(x_i) dx_i = 0, \end{aligned}$$

while

$$\int (X_i - \mu_i)^2 f(x_i) dx_i = \sigma_i^2,$$

the variance of X_i.

Notation 3.18 *It is common in probability to use Var to denote the variance of a complicated expression. In this application, one will also see σ^2.*

5. Variance for Dependent Variates:

Note that when $\{X_i\}$ are not independent, we obtain as above:

$$Var\left[\sum\nolimits_{i=1}^{n} X_i\right] = \sum\nolimits_{i=1}^{n} \sigma_i^2 + 2 \sum\nolimits_{j<i} E^{(n)}\left[\left(X_j - \mu_j\right)(X_i - \mu_i)\right],$$

where by 3.24:

$$E^{(n)}\left[\left(X_j - \mu_j\right)(X_i - \mu_i)\right] = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \left(X_j - \mu_j\right)(X_i - \mu_i) f(x_1, x_2, ..., x_n) dx_1 dx_2 ... dx_n.$$

This again involves the marginal density function of 2.12:

$$f(x_i, x_j) = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f(x_1, x_2, ..., x_n) \prod\nolimits_{k \neq i,j} dx_k,$$

and so:

$$E^{(n)}\left[\left(X_j-\mu_j\right)\left(X_i-\mu_i\right)\right] = \int_{\mathbb{R}}\int_{\mathbb{R}}\left(X_j-\mu_j\right)\left(X_i-\mu_i\right)f(x_i,x_j)dx_idx_j$$
$$= E^{(2)}\left[\left(X_j-\mu_j\right)\left(X_i-\mu_i\right)\right].$$

Definition 3.19 *Given random variables X, Y on a probability space $(\mathcal{S},\mathcal{E},\lambda)$, the* ***covariance of*** *X* ***and*** *Y, denoted $cov(X,Y)$, is defined by:*

$$cov(X,Y) \equiv E^{(2)}[(X-\mu_X)(Y-\mu_Y)] = E^{(2)}[XY]-\mu_X\mu_Y, \quad (3.31)$$

and the ***correlation between*** *X* ***and*** *Y, denoted $corr(X,Y)$ and often $\rho(X,Y)$ or ρ_{XY}, is defined by:*

$$corr(X,Y) \equiv \frac{cov(X,Y)}{\sigma_X\sigma_Y}. \quad (3.32)$$

Hence, the general formula for the variance of a summation can be expressed:

$$Var\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \sigma_i^2 + 2\sum_{j<i} cov(X_i,X_j), \quad (3.33)$$

or equivalently:

$$Var\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \sigma_i^2 + 2\sum_{j<i} \rho_{ij}\sigma_i\sigma_j. \quad (3.34)$$

6. Moment Generating Function:

By the definition in 3.16 applied to a sum of **independent variates** $X=\sum_{i=1}^n X_i$**:**

$$M_X(t) = E^{(n)}\left[\exp\left(\sum_{i=1}^n tX_i\right)\right]$$
$$= \prod_{i=1}^n \int \exp\left(tx_i\right) f\left(x_i\right) dx_i,$$

and so:

$$M_X(t) = \prod_{i=1}^n M_{X_i}(t). \quad (3.35)$$

In other words, if $M_{X_i}(t)$ exists for $t \in (-t_0^{(i)}, t_0^{(i)})$, then $M_X(t)$ exists on $(-t_0,t_0) \equiv \bigcap_{i=1}^n (-t_0^{(i)}, t_0^{(i)})$ and on this interval 3.28 is satisfied.

3.2.4 Properties of Moments and the M.G.F.

Note that for $m \leq n$:

$$(x-a)^m \leq (x-a)^n + 1,$$

and so the existence of an *nth* moment assures the existence of the respective *mth* moments for all $m \leq n$. Hence, one is usually interested in whether moments exist for all n, and if not, determining the largest n for which moments exist.

The following result assures that if one *nth* moment exists, so too do all the others. Of course, the absolute versions of the moments exist by definition of expectation, so we only need to consider μ'_n and μ_n.

Proposition 3.20 *For any n, μ_n exists if and only if μ'_n exists, and:*

$$\mu_n = \sum_{j=0}^{n} (-1)^{n-j} \binom{n}{j} \mu'_j \mu^{n-j}, \qquad \mu'_n = \sum_{j=0}^{n} \binom{n}{j} \mu_j \mu^{n-j}. \tag{3.36}$$

Proof. *Left as an exercise. Hint: Recall the binomial theorem in 1.8.* ■

Corollary 3.21

$$\sigma^2 = \mu'_2 - \mu^2. \tag{3.37}$$

Proof. *This follows from the first identity in 3.36 with $n = 2$.* ■

The next result shows that in order for $M_X(t)$ to exist for some interval about the origin, $(-t_0, t_0)$ with $t_0 > 0$, it is necessary that μ'_n and hence μ_n exist for all n. In other words, the existence of the moment generating function is at least as restrictive on a distribution as is the existence of all finite moments. As will be seen below by example, it is even more restrictive. Specifically, there are distributions for which μ'_n exists for all n but for which $M_X(t)$ exists only for $t = 0$.

Proposition 3.22 *If $M_X(t)$ exists for some interval about the origin, $(-t_0, t_0)$ with $t_0 > 0$, then μ'_n exists for all n.*
Proof. *Choose t with $0 < t < t_0$. Then since $e^{t|x|} \leq e^{tx} + e^{-tx}$, the existence of $M_X(t)$ and $M_X(-t)$ implies that*

$$\int_{\mathbb{R}} e^{t|x|} dF \leq M_X(t) + M_X(-t) < \infty.$$

Given n, note that $e^{t|x|} \geq |x|^n$ for $|x| / \ln |x| \geq n/t$, and since $|x| / \ln |x|$ is increasing and unbounded, choose $\{x_n\}$ with $x_n > 0$ and $|x_n| / \ln |x_n| \geq n/t$. Then:

$$\begin{aligned}\int_{\mathbb{R}} |x|^n \, d\mu_F(x) &= \int_{|x|<x_n} |x|^n \, dF + \int_{|x|\geq x_n} |x|^n \, dF \\ &\leq c\,|x_n|^n + \int_{|x|\geq x_n} e^{t|x|} dF \\ &\leq c\,|x_n|^n + M_X(t) + M_X(-t),\end{aligned}$$

and so μ'_n exists for every n. ■

Remark 3.23 *In the above proof,*

$$\begin{aligned}c &= \int_{|x|<x_n} dF \\ &= F\left(x_n^-\right) - F(-x_n) \leq 1.\end{aligned}$$

The name "moment generating" function for $M_X(t)$ is justified next, in that when $M_X(t)$ exists it not only assures the existence of all moments, but can also be used to generate these moments. To make this proof rigorous, one needs the Lebesgue dominated convergence theorem for Lebesgue-Stieltjes integrals, a result that will be proved in book 5. More specifically, we need the corollary to this result as stated in corollary 2.63 in book 3 in the Lebesgue case, that addresses when the integral of a infinite summation equals the summation of the integrals. Standard results on absolutely convergent double series and the book 3 result are adequate for the special cases of this proposition where dF reflects a discrete or continuously differentiable distribution function. In these cases the Lebesgue-Stieltjes integrals reduce to summations or Lebesgue integrals, respectively. In the latter case these integrals are actually Riemann, but by proposition 2.64 of book 3 they equal their Lebesgue integral counterparts.

For a generalization of 3.39 see the proof of proposition 6.6, which derives a Riemann-Stieltjes formula for $M_X^{(n)}(t)$. See 6.9.

Proposition 3.24 *If $M_X(t)$ exists for some interval about the origin, $(-t_0, t_0)$ with $t_0 > 0$, then*

$$M_X(t) = \sum\nolimits_{n=0}^{\infty} \mu'_n t^n / n!, \tag{3.38}$$

and hence

$$\mu'_n = M_X^{(n)}(0), \tag{3.39}$$

where $M^{(n)}(0)$ denotes the nth derivative of $M_X(t)$ evaluated at $t = 0$.
Proof. *Because $e^{|tx|} \leq e^{tx} + e^{-tx}$, the existence of $M_X(t)$ for $|t| < t_0$ assures that $\int e^{|tx|} dF < \infty$. Recall from calculus that the exponential Taylor series:*

$$e^{tx} \equiv \sum_{n=0}^{\infty} (tx)^n / n! \tag{3.40}$$

is absolutely convergent in x for all t. Thus the following partial sums:

$$\sum_{n=0}^{N} |tx|^n / n! \leq e^{|tx|},$$

are bounded by an integrable function and this then implies by the triangle inequality that for all N :

$$\left| \sum_{n=0}^{N} (tx)^n / n! \right| \leq e^{|tx|}.$$

Thus the corollary to the Lebesgue dominated convergence theorem of book 5 applies to state that for $|t| < t_0$:

$$\begin{aligned} M_X(t) &= \sum_{n=0}^{\infty} (t^n/n!) \int x^n dF \\ &= \sum_{n=0}^{\infty} \mu'_n t^n / n!. \end{aligned}$$

Hence, $M_X(t)$ has a Taylor series expansion for $|t| < t$ and 3.39 follows by a direct calculation. See for example proposition 9.111 in Reitano. ■

For the purpose of calculating only the mean, variance and third central moment of a distribution, the following corollary is often useful.

Corollary 3.25 *If $M_X(t)$ exists for some interval about the origin, $(-t_0, t_0)$ with $t_0 > 0$, then with $g_X(t) \equiv \ln M_X(t)$, the* ***cumulant generating function:*** *:*

$$\mu = g'_X(0), \qquad \sigma^2 = g''_X(0), \qquad \mu_3 = g_X^{(3)}(0). \tag{3.41}$$

Proof. *Left as an exercise.* ■

Remark 3.26 *Note that if $M_X(t)$ exists as a Taylor series for $t \in (-t_0, t_0)$, then $g_X(t) \equiv \ln M_X(t)$ exists as a Taylor series for all such t since $M_X(t) > 0$ by definition. The* ***cumulants*** *of X are then defined in terms of this Taylor series:*

$$\ln M_X(t) = \sum_{n=0}^{\infty} (t^n/n!) \kappa_n,$$

where κ_n denotes the nth ***cumulant of*** *X, and as above, $\kappa_n = g_X^{(n)}(0)$.*

Corollary 3.27 *If $M_X(t)$ exists for some interval about the origin, $(-t_0, t_0)$ with $t_0 > 0$, then $M_X(t)$ is **convex** on on this interval, That is, if $[t_1, t_2] \subset (-t_0, t_0)$, then for any α, $0 \leq \alpha \leq 1$:*

$$M_X(\alpha t_1 + (1-\alpha)t_2) \leq \alpha M_X(t_1) + (1-\alpha)M_X(t_2). \tag{3.42}$$

Proof. *Left as an exercise. Hint: Note that e^{xt} is a convex in t for any x.* ■

Remark 3.28 *While the existence of $M_X(t)$ on the interval $(-t_0, t_0)$ assures the existence of moments of all orders, this result is not reversible as noted above. As will be seen in remark 3.30 below, the lognormal distribution provides an example for which moments of all orders exist, yet $M_X(t)$ is only defined for $t = 0$. The "problem" with this distribution is that μ'_n grows exponentially fast with n, and hence far too fast for the $M_X(t)$ series to converge if $t \neq 0$. See also example 3.53 in the section Uniqueness of Moments and the M.G.F., as well as proposition 3.55 which addresses the question of when the existence of all moments implies the existence of the moment generating function.*

3.2.5 Examples of Moments and M.G.F.s

In this section we present various results related to moments of the discrete and continuous distributions presented in section 1.2. For the various results, which are largely assigned as exercises, we recall the equations 3.10 and 3.11 in the discussion above.

Discrete Distributions

1. **Discrete Rectangular Distribution on** $[0, 1]$:

   Recalling 1.4:

   $$f_R(j/n) = 1/n, \quad j = 1, 2, .., n,$$

   a calculation obtains:

   $$\mu_R = (n+1)/(2n), \qquad \sigma_R^2 = \left(n^2 - 1\right)/\left(12n^2\right), \tag{3.43}$$

   and

   $$M_R(t) = \frac{\exp[(1+1/n)t] - \exp[t/n]}{n(\exp[t/n] - 1)}. \tag{3.44}$$

   These results can be generalized to a discrete rectangular distribution on $[a, b]$,

   $$f_R(j(b-a)/n + a) = 1/n, \quad j = 1, 2, .., n, \tag{3.45}$$

by defining $Y = (b-a)X + a$ and utilizing linearity of the expectations in 3.5 to obtain:

$$\mu_{R_{a,b}} = (b-a)\mu_R + a, \qquad \sigma^2_{R_{a,b}} = (b-a)^2\sigma^2_R, \tag{3.46}$$

and

$$M_{R_{a,b}}(t) = e^{at} M_R([b-a]t). \tag{3.47}$$

2. **Binomial Distribution**:

   From 1.6, let $f_{B_n}(j)$ denote the probability density function of the sum of n independent standard binomials:

$$f_{B_n}(j) = \binom{n}{j} p^j (1-p)^{n-j}, \quad j = 0, 1, .., n.$$

   For the standard binomial random variable, X_1^B, we have from 1.5 and a simple calculation that:

$$E[\left(X_1^B\right)^m] = p, \qquad E[\left(X_1^B - p\right)^m] = p(1-p)^m + (1-p)(-p)^m,$$

   and $M_{X_1^B}(t) = pe^t + (1-p)$. Hence from the section Moments of Sums of RVs:

$$\mu_{B_n} = np, \qquad \sigma^2_{B_n} = np(1-p), \tag{3.48}$$

$$M_{B_n}(t) = \left(1 + p(e^t - 1)\right)^n. \tag{3.49}$$

   Note that while the moment generating function of $f_{B_n}(j)$ has a simple form, the mth moment of B, $\mu'_{B_n,m}$, is a messy polynomial in p of degree m since in 3.28, $\mu'^{(1)}_{m_1}\mu'^{(2)}_{m_2}\cdots\mu'^{(n)}_{m_n} = p^r$ where r equals the number of subscripts with $m_j > 0$. The derivation of $\mu'_{B_n,m}$ is not simplified by using 3.39.

3. **Geometric Distribution:**:

   From 1.9:

$$f_G(j) = p(1-p)^j, \ j = 0, 1, 2, ...$$

   and for this distribution it is easiest to first calculate $M_G(t)$, then evaluate some of the moments using the formula in 3.39. To this end:

$$M_G(t) = p\sum\nolimits_{j=0}^{\infty} \left[(1-p)e^t\right]^j,$$

and as a geometric summation this series is convergent if $(1-p)e^t < 1$, with:

$$M_G(t) = p/\left[1-(1-p)e^t\right]. \tag{3.50}$$

Derivatives can now be evaluated using 3.41 to produce:

$$\mu_G = (1-p)/p, \qquad \sigma_G^2 = (1-p)/p^2. \tag{3.51}$$

4. **Negative Binomial Distribution:**:

   This is another distribution for which it is easiest to evaluate $M_{NB}(t)$ first, and using 1.12:

$$f_{NB}(j) = \binom{j+k-1}{k-1} p^k (1-p)^j, \quad j = 0, 1, 2, ...,$$

   it follows that for $(1-p)e^t < 1$ the resulting series is convergent and:

$$M_{NB}(t) = \left(p/\left[1-(1-p)e^t\right]\right)^k. \tag{3.52}$$

   Calculating two derivatives with 3.41 produces:

$$\mu_{NB} = k(1-p)/p, \qquad \sigma_{NB}^2 = k(1-p)/p^2. \tag{3.53}$$

   Remark 3.29 *Note that by 3.35, $M_{NB}(t)$ is also the moment generating function of the sum of k independent geometric variables, Using the tools of section 3.2.7, we will conclude that the negative binomial is uniquely defined by its moments and moment generating function, and hence conclude that a negative binomial variate equals the sum of k independent geometric variates. See example 3.59.*

5. **Poisson Distribution:**:

   Recalling 1.13:

$$f_P(j) = e^{-\lambda}\lambda^j/j!, \quad j = 0, 1, 2, ...,$$

   the moment generating function is calculated directly as:

$$M_P(t) = \exp[\lambda(e^t - 1)], \tag{3.54}$$

   and mean and variance follow from 3.41:

$$\mu_P = \lambda, \qquad \sigma_P^2 = \lambda. \tag{3.55}$$

Continuous Distributions

1. **Continuous Uniform Distribution**:

   From 1.18:

$$f_U(x) = 1/(b-a), \quad x \in [a,b],$$

   and $f_U(x) = 0$ otherwise, it follows that:

$$\mu_U = (b+a)/2, \qquad \sigma_U^2 = (b-a)^2/12, \tag{3.56}$$

   and

$$M_U(t) = \left(e^{bt} - e^{at}\right) / \left(t\left[b-a\right]\right), \quad t \in \mathbb{R}. \tag{3.57}$$

   Note that $M_U(t)$ is well defined at $t = 0$ despite the apparent singularity as justified using the exponential Taylor series in 3.40.

2. **Exponential Distribution** and **Gamma Distribution**:

   The exponential density is defined with a single scale parameter $\lambda > 0$ in 1.20:

$$f_E(x) = \lambda e^{-\lambda x}, \quad x \geq 0,$$

   and is a special case of the more general gamma density defined with a shape parameter α and scale parameter $\lambda > 0$ in 1.22 by:

$$f_\Gamma(x) = \lambda^\alpha x^{\alpha-1} e^{-\lambda x}/\Gamma(\alpha), \quad x \geq 0,$$

   with the **gamma function,** $\Gamma(\alpha)$, defined by:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx.$$

   Moments of the gamma are derived by integration by parts and 1.24, that for $c > 1$:

$$\Gamma(c) = (c-1)\Gamma(c-1),$$

   which produces $\Gamma(n) = (n-1)!$ for integer n as in 1.25. This is then used to derive:

$$\mu'_{\Gamma,n} = \alpha \left[\prod_{j=1}^{n-1} (\alpha + j)\right] / \lambda^n, \qquad \mu_\Gamma = \alpha/\lambda, \qquad \sigma_\Gamma^2 = \alpha/\lambda^2. \tag{3.58}$$

   The moment generating function can also be calculated:

$$M_\Gamma(t) = (1 - t/\lambda)^{-\alpha}, \qquad t < \lambda. \tag{3.59}$$

These formulas produce results for the **exponential distribution** with parameter λ by setting $\alpha = 1$, and by notation 1.15 provide results for the **Chi-squared distribution with n-degrees of freedom** by setting $\lambda = 1/2$ and $\alpha = n/2$.

3. Beta Distribution

The **beta density** contains two shape parameters, $v > 0, w > 0$, and is defined on the interval $[0, 1]$ by the density function in 1.27:

$$f_\beta(x) = x^{v-1}(1-x)^{w-1}/B(v,w),$$

where the **beta function** $B(v, w)$ is defined by a definite integral as in 1.28:

$$B(v,w) = \int_0^1 y^{v-1}(1-y)^{w-1}dy.$$

For any positive integer n it follows by definition that:

$$E[x^n] = \frac{B(v+n,w)}{B(v,w)}.$$

Also, the beta function $B(v, w)$ satisfies an important identity which is useful in evaluating these moments:

$$B(v+1,w) = \frac{v}{v+w}B(v,w), \tag{3.60}$$

as can be derived using 1.30 and 1.24. Applying the iterative formula in 3.60 and $B(1, 1) = 1$ produces:

$$\begin{aligned} \mu_\beta &= \frac{v}{v+w}, \qquad \mu'_{\beta,n} = \prod_{j=0}^{n-1}\left(\frac{v+j}{v+w+j}\right), \\ \sigma^2_\beta &= \frac{vw}{(v+w)^2(v+w+1)}. \end{aligned} \tag{3.61}$$

Thus $\mu'_{\beta,n} \to 0$ as $n \to \infty$.

Next, note that $M_\beta(t)$ exists for all t since for $x \in [0, 1]$, $e^{xt} \leq \max[1, e^t]$, and so:

$$|M_\beta(t)| \leq \max[1, e^t].$$

Thus $M_\beta(t)$ can be expressed in terms of its moments:

$$M_\beta(t) = 1 + \sum_{n=1}^{\infty}\prod_{i=0}^{n-1}\left(\frac{v+i}{v+w+i}\right)\frac{t^n}{n!}. \tag{3.62}$$

4. Cauchy Distribution

The **Cauchy distribution**, named for **Augustin Louis Cauchy** (1789 – 1857), is of interest as an example of a p.d.f. that has no finite moments. This density function is defined on $\mathbb{R}$ as a function of a location parameter, $x_0 \in \mathbb{R}$, and a scale parameter $\gamma > 0$, by:

$$f_C(x) = \frac{1}{\pi\gamma} \frac{1}{1 + \left([x - x_0] / \gamma \right)^2}, \tag{3.63}$$

while the **standard Cauchy distribution** is parameterized with $x_0 = 0$ and $\gamma = 1$ to

$$f_C(x) = \frac{1}{\pi} \frac{1}{1 + x^2}. \tag{3.64}$$

A substitution into the integral of $\tan z = (x - x_0) / \gamma$ shows this probability function integrates to 1, as it should. While by a symmetry argument one would derive that

$$\frac{1}{\pi} \int_{-\infty}^{\infty} \frac{x dx}{1 + x^2} = 0,$$

it is not the case that $E[X] = 0$ since $E[|X|]$ is unbounded:

$$\frac{1}{\pi} \int_{-N}^{N} \frac{|x| \, dx}{1 + x^2} = \frac{1}{\pi} \int_{0}^{N} \frac{2x dx}{1 + x^2} = \frac{1}{\pi} \ln N,$$

and hence the Cauchy distribution has no finite moments.

5. Normal Distribution

The normal distribution is defined on $(-\infty, \infty)$, depends on a location parameter $\mu \in \mathbb{R}$ and a scale parameter $\sigma > 0$, and is defined in 1.32 by:

$$f_N(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(- \left(x - \mu \right)^2 / \left(2\sigma^2 \right) \right),$$

with the associated unit normal distribution, denoted $\phi(x)$, defined in 1.33 with $\mu = 0$ and $\sigma = 1$:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2 \right).$$

There is no elementary derivation of the fact that $\phi(x)$, and hence $f_N(x)$, integrate to 1. As above, the powerful results of book 5 are required. However, all central moments exist because $\exp\left(-x^2/2 \right) <$

x^{-n} for all n as $x \to \infty$. In addition, all odd central moments are 0 by symmetry, and for even moments an integration by parts shows that

$$\int_{-\infty}^{\infty} x^{2m}\phi(x)dx = (2m-1)\int_{-\infty}^{\infty} x^{2m-2}\phi(x)dx,$$

and this plus mathematical induction obtains that:

$$\int_{-\infty}^{\infty} x^{2m}\phi(x)dy = \frac{(2m)!}{2^m m!}.$$

Hence, justifying the notational convention of parameterizing the normal with μ and σ^2, we have that:

$$\mu'_{N,1} = \mu, \qquad \mu_{N,2} = \sigma^2, \qquad \mu_{N,2m} = \frac{\sigma^{2m}\,(2m)!}{2^m m!}, \qquad \mu_{N,2m+1} = 0. \tag{3.65}$$

Finally, the moment generating function is derived by completing the square in the exponential function, and a substitution to produce:

$$M_N(t) = \exp\left(\mu t + \sigma^2 t^2/2\right), \qquad t \in \mathbb{R}, \tag{3.66}$$

which obtains for the unit normal:

$$M_\Phi(t) = \exp\left(t^2/2\right). \tag{3.67}$$

6. **Lognormal Distribution**:

The **lognormal distribution** is defined on $[0, \infty)$, depends on a location parameter $\mu \in \mathbb{R}$ and a shape parameter $\sigma > 0$, and has probability density function given in 1.34:

$$f_L(x) = \frac{1}{\sigma x\sqrt{2\pi}} \exp\left(-\left(\ln x - \mu\right)^2 / \left(2\sigma^2\right)\right).$$

The substitution $y = (\ln x - \mu)/\sigma$ into the integral of $f_L(x)$ produces the integral of the unit normal $\phi(y)$, and moments of all orders exist for the lognormal and are calculated using the same substitution:

$$\mu'_{L,n} = e^{n\mu} M_\Phi(n\sigma).$$

In other words, the moments of the lognormal can be calculated from the moment generating function of the unit normal. Specifically, using 3.67 obtains:

$$\mu'_{L,n} = e^{n\mu + (n\sigma)^2/2}, \qquad \mu_L = e^{\mu+\sigma^2/2}, \qquad \sigma_L^2 = e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right). \tag{3.68}$$

Remark 3.30 *As noted above in remark 3.28, while the existence of* $M_X(t)$ *on an open interval* $(-t_0, t_0)$ *assures that all moments of the distribution function exists, it is not the case that the existence of all moments assures the existence of* $M_X(t)$ *on an open interval* $(-t_0, t_0)$*. The lognormal distribution provides the classical counterexample, in that while* μ'_{nL} *exists for all* n *by 3.68, the series* $\sum_{n=0}^{\infty} t^n \mu'_n / n!$ *cannot converge for any* $t \neq 0$*, and so* $M_{LN}(t)$ *cannot exist except for* $t = 0$*.*

To see this, recall that the ratio test states that for a positive series, if $\limsup \frac{c_{n+1}}{c_n} < 1$ *then* $\sum_{n=0}^{\infty} c_n$ *converges, while if* $\liminf \frac{c_{n+1}}{c_n} > 1$ *then* $\sum_{n=0}^{\infty} c_n$ *diverges. A calculation shows that for this series and* $t > 0$*:*

$$\frac{c_{n+1}}{c_n} = \frac{e^{\mu+(n+1/2)\sigma^2}}{n+1} t,$$

which is unbounded. When $t < 0$ *this series alternates, and by the alternating series theorem this series converges if and only if* $\left| t^n e^{n\mu+(n\sigma)^2/2}/n! \right| \to 0$ *as* $n \to \infty$*. By Stirling's formula in 3.91 below,* $n!/\left(\sqrt{2\pi} n^{n+1/2} e^{-n}\right) \to 1$ *as* $n \to \infty$ *and thus we can investigate if:*

$$\left| \frac{t^n e^{n\mu+(n\sigma)^2/2}}{\sqrt{2\pi} n^{n+1/2} e^{-n}} \right| \to 0.$$

As the log of this expression is unbounded, it is apparent that so too is this ratio and thus the alternating series does not converge.

In summary, $M_{LN}(t)$ *does not exist except for* $t = 0$*.*

3.2.6 Moments and Inequalities

In this section is developed a number of important inequalities related to moments.

Chebyshev's Inequality

Chebyshev's inequality, sometimes spelled as Chebychev or Tchebysheff, applies to any distribution function that has a mean and variance, and hence it is quite generally applicable. It is named for its discoverer, **Pafnuty Chebyshev** (1821 – 1894), who was a Russian mathematician and hence the many transliterations of his name in English. This inequality can be stated in many ways, and Chebyshev is actually a name now given to a family of inequalities as will be seen below. But it is

often applied as stated in the first proposition when we are interested in an upper bound for the probability that a random variable is far from its mean, where "far" can be defined in absolute units, or units relative to the variance.

Notation 3.31 *For these results, if X is a random variable defined on a probability space $(\mathcal{S},\mathcal{E},\lambda)$ with distribution function F, recall that we use the simplified notation such as $\Pr[|X-\mu| \geq t\sigma]$ as shorthand for:*

$$\begin{aligned}\Pr\left[|X-\mu| \geq t\sigma\right] &\equiv \lambda\left[\{s \in \mathcal{S} | X(s) \leq \mu - t\sigma\} \cup \{s \in \mathcal{S} | X(s) \geq \mu + t\sigma\}\right] \\ &= F(\mu - t\sigma) + 1 - F([\mu + t\sigma]^{-}),\end{aligned}$$

where $F([\mu + t\sigma]^{-})$ denotes the left limit of $F(x)$ as $x \to \mu + t\sigma$.

Proposition 3.32 (Chebyshev's inequality) *If $F(x)$ is a distribution function of a random variable X with mean μ and variance σ^2, then for any real number $t > 0$:*

$$\Pr[|X-\mu| \geq t\sigma] \leq 1/t^2. \tag{3.69}$$

Equivalently, for any $s > 0$:

$$\Pr[|X-\mu| \geq s] \leq \sigma^2/s^2. \tag{3.70}$$

Proof. *By definition,*

$$\begin{aligned}\sigma^2 &\equiv \int_{\mathbb{R}} (x-\mu)^2 dF \\ &\geq \int_{|x-\mu| \geq t\sigma} (x-\mu)^2 dF \\ &\geq (t\sigma)^2 \Pr\left[|X-\mu| \geq t\sigma\right].\end{aligned}$$

The second result is implied by the first with the substitution: $t = s/\sigma$. ■

With the same proof as above, we can extend the above result.

Proposition 3.33 (Generalized Chebyshev inequalities) *If $F(x)$ is a distribution function of a random variable X for which the nth moment exists, then for any real number $t > 0$:*

$$\Pr[|X| \geq t] \leq \mu'_{|n|}/t^n \tag{3.71}$$

$$\Pr[|X-\mu| \geq t] \leq \mu_{|n|}/t^n, \tag{3.72}$$

where $\mu'_{|n|} \equiv E[|X|^n]$ *and* $\mu_{|n|} \equiv E[|X-\mu|^n]$ *denote absolute moments. If the moment generating function exists for all* t*, then*

$$\Pr[X \geq t] \leq M_X(t)e^{-t^2}. \tag{3.73}$$

If the moment generating function exists for all $t' \in (-t_0, t_0)$ *with* $t_0 > 0$*, then*

$$\Pr[|X| \geq t] \leq e^{-t't} \left[M(t') + M(-t')\right]. \tag{3.74}$$

Proof. *Left as exercises. Hint for 3.74: Consider* $t' > 0$ *and* $t' < 0$ *separately.* ■

Corollary 3.34 (Generalized Chebyshev inequalities) *If* $A \in \mathcal{E}$*, then 3.71 generalizes to:*

$$\Pr[\{|X| \geq t\} \cap A] \leq E[|X|^n \chi_A]/t^n, \tag{3.75}$$

where $\chi_A : \mathcal{S} \to \mathbb{R}$ *is the characteristic function defined as* $\chi_A \equiv 1$ *for* $s \in A$ *and is* 0 *otherwise. A similar statement holds for other inequalities.*
Proof. *Left as exercises.* ■

Remark 3.35 **1**. *Note that the statement and proof of the generalized Chebyshev result make sense for real numbers* $n > 0$*, and not just integer values.*

2. *Note also that when* $n = 1$*, the inequality in 3.71 as restated in terms of* $\mu'_{|1|} \equiv E[|X|]$ *is known as* ***Markov's inequality,*** *named for* ***Andrey Markov*** *(1856 – 1922), a student of Chebyshev. This inequality states:*

$$\Pr[|X| \geq t] \leq E[|X|]/t. \tag{3.76}$$

In some texts, all inequalities of the type in 3.71 are called Markov inequalities.

Jensen's Inequality

In corollary 3.27 the notion of convexity was introduced. Here we formalize and expand this definition. Define the secant line of $g(x)$ over an interval $[x, y]$ as the line segment between $(x, g(x))$ and $(y, g(y))$. A convex function is one for which the secant line is always above the graph of the function over this interval, while for concave function the secant line is always below the graph of the function on this interval. More formally:

Definition 3.36 *A function $g(x)$ is **concave** on an interval $I = (a, b)$, which may be infinite, if for any $x, y \in I$:*

$$g(tx + (1-t)y) \geq tg(x) + (1-t)g(y) \quad \text{for } t \in [0,1]. \tag{3.77}$$

*A function $g(x)$ is **convex** on such I if for any $x, y \in I$:*

$$g(tx + (1-t)y) \leq tg(x) + (1-t)g(y) \quad \text{for } t \in [0,1]. \tag{3.78}$$

*When the inequalities are strict for $t \in (0,1)$, such functions are referred to as **strictly concave** and **strictly convex,** respectively.*

An important result from calculus, which we do not prove but can be found in section 9.6 of Reitano and elsewhere, is the following.

Proposition 3.37 *If $g(x)$ is differentiable, then:*

1. *$g(x)$ is concave on an interval if and only if $g'(x)$ is a decreasing function on that interval.*

2. *$g(x)$ is convex on an interval if and only if $g'(x)$ is an increasing function on that interval.*

3. *$g(x)$ is strictly concave iff $g'(x)$ is strictly monotonically decreasing, and strictly convex iff $g'(x)$ is strictly monotonically increasing.*

The next proposition states that the graph of the tangent line dominates the graph of a concave function from above, and supports the graph of a convex function from below. To simplify the proof we assume differentiability. But this result is true without this assumption but where $g'(a)$ is replaced by a one-sided derivative related to the derivates introduced in book 3 in section 3.2.1. If the right upper and lower derivates agree, this common value is called the right derivative, and similarly for the definition of left derivative. It turns out that for concave and convex functions, both of these one-sided derivatives exist at every point, and that they agree except perhaps on a countable collection of points. In other words, a concave or convex function $g(x)$ is differentiable except at most on a countable collection of points. However, we have no further use for this generalization, so we simply assume differentiability.

Proposition 3.38 *If $g(x)$ is differentiable, then for any a:*

1. *If $g(x)$ is concave:*

$$g(x) \leq g(a) + g'(a)(x - a),$$

2. *If $g(x)$ is convex:*

$$g(x) \geq g(a) + g'(a)(x - a).$$

In addition, if $g(x)$ is strictly concave or strictly convex, then these inequalities are strict.

Proof. *By the Mean Value theorem of calculus we have that for any a :*

$$g(x) = g(a) + g'(\theta_x)(x - a),$$

where either $x > \theta_x > a$ or $a < \theta_x < x$. By the above proposition 3.37, if $g(x)$ is concave then $g'(x)$ is a decreasing function and hence, $g'(\theta_x) \leq g(a)$ if $x > a$, and $g'(\theta_x) \geq g(a)$ if $x < a$. In either case, $g'(\theta_x)(x-a) \leq g'(a)(x-a)$. If $g(x)$ is convex, the inequalities reverse. For strictly concave or strictly convex functions, the first derivative inequalities are sharp and so too are the inequalities in the conclusion. ■

We now turn to an important result related to concave and convex functions known as **Jensen's inequality,** and named for its discoverer, **Johan Jensen** (1859 – 1925). We assume differentiability of $g(x)$ to be consistent, but as noted, this is not necessary.

Proposition 3.39 (Jensen's Inequality) *Let $g(x)$ be a differentiable function and X a random variable with range contained in the domain of g, i.e., $Rng(X) \subset Dmn(g)$. Then:*

1. *If $g(x)$ is concave:*

$$E[g(X)] \leq g(E[X]). \tag{3.79}$$

2. *If $g(x)$ is convex:*

$$E[g(X)] \geq g(E[X]). \tag{3.80}$$

If g is strictly concave or strictly convex, the inequalities are strict.

Proof. *Let $a = E[X]$ in the above proposition, then since $E[g'(a)(X - a)] = g'(a)E[(X - a)] = 0$, the result follows.* ■

Kolmogorov's Inequality

Andrey Kolmogorov (1903 – 1987) was responsible for introducing an axiomatic framework for probability theory, and a large number of important results bear his name. Extending Chebyshev's inequality, **Kolmogorov's inequality** addresses a collection of random variables, $\{X_i\}_{i=1}^n$, and provides a probability statement regarding the maximum of the associated partial summations. It is stated for simplicity under the assumption that $E[X_j] = 0$ for all j. However, this is not a true restriction since if we are given $\{Y_j\}_{j=1}^n$ with $E[Y_j] = \mu_j$ we can apply the result to $X_j \equiv Y_j - \mu_j$ with no change in variance since since $Var[X_j] = Var[Y_j]$. Also note that while this result requires that $\{X_j\}_{j=1}^n$ be independent random variables, it does not require that they be "identically distributed," meaning it does not require that they have the same distribution function.

Proposition 3.40 (Kolmogorov's Inequality) *Let $\{X_i\}_{i=1}^n$ be independent random variables with $E[X_j] = 0$ and $Var[X_j] = \sigma_j^2$. Then for $t > 0$:*

$$\Pr\left\{\max_{1\le k\le n}\left|\sum_{j=1}^{k} X_j\right| \ge t\right\} \le \sum_{j=1}^{n} \sigma_j^2/t^2. \tag{3.81}$$

Remark 3.41 *Note that Kolmogorov's inequality is considerably stronger than is Chebyshev's inequality applied to this probability statement. The Chebyshev inequality would state that for any k with $1 \le k \le n$:*

$$\Pr\left\{\left|\sum_{j=1}^{k} X_j\right| \ge t\right\} \le \sum_{j=1}^{k} \sigma_j^2/t^2,$$

since as independent random variables, $Var\left(\sum_{j=1}^{k} X_j\right) = \sum_{j=1}^{k} \sigma_j^2$. Of course for any $k \le n$, $\sum_{j=1}^{k} \sigma_j^2/t^2 \le \sum_{j=1}^{n} \sigma_j^2/t^2$, so at first these inequalities appear similar.

However, Chebyshev's inequality provides probability statements on n separate events, and is silent on the question of the simultaneous occurrence of these n events. Kolmogorov's inequality says that the largest of the n Chebyshev probability bounds is sufficient to bound the probability of the worst case of these n events. Alternatively, Kolmogorov's inequality says that the largest of the n Chebyshev probability bounds is sufficient to bound the probability that all inequalities are satisfied simultaneously.

Proof. *With $S_k \equiv \sum_{j=1}^{k} X_j$, define $A_k = \{s \in \mathcal{S} | \, |S_k| \ge t$ and $|S_j| < t$ for $j < k\}$ with $S_0 \equiv 0$ for A_1. Then $\{A_k\}_{k=1}^n$ are disjoint measurable sets*

and $\sum_{k=1}^{n}\chi_{A_k}(s) \leq 1$ *since if* $|S_n(s)| < t$ *then* $s \notin \bigcup_{k=1}^{n} A_k$. *Recall that* $\chi_{A_k}(s)$ *is the characteristic function of* A_k *and defined as* 1 *for* $s \in A_k$ *and* 0 *otherwise. Thus:*

$$E\left[S_n^2\right] \geq \sum\nolimits_{k=1}^{n} E\left[\chi_{A_k} S_n^2\right].$$

Now $S_n^2 = S_k^2 + 2S_k(S_n - S_k) + (S_n - S_k)^2$, *and so:*

$$\begin{aligned} E\left[S_n^2\right] &\geq \sum\nolimits_{k=1}^{n}\left(E\left[\chi_{A_k} S_k^2\right] + 2E\left[\chi_{A_k} S_k(S_n - S_k)\right] + E\left[\chi_{A_k}(S_n - S_k)^2\right]\right) \\ &\geq \sum\nolimits_{k=1}^{n} E\left[\chi_{A_k} S_k^2\right]. \end{aligned}$$

The last step follows since $E\left[\chi_{A_k}(S_n - S_k)^2\right] \geq 0$, *and* $2E\left[\chi_{A_k} S_k(S_n - S_k)\right] = 0$ *because the random variables* $\chi_{A_k} S_k = \sum_{j=1}^{k}\chi_{A_k} X_j$ *and* $S_n - S_k = \sum_{j=k+1}^{n} X_j$ *are independent by section 3.4.4 of book 2 (see also remark 3.42 below) and* $E[S_n - S_k] = \sum_{j=k+1}^{n} E[X_j] = 0$.

Now $E\left[\chi_{A_k} S_k^2\right] \geq t^2 \Pr\left[A_k\right]$ *by 3.75, and so*

$$\begin{aligned} E\left[S_n^2\right] &\geq t^2 \sum\nolimits_{k=1}^{n} \Pr\left[A_k\right] \\ &= t^2 \Pr\left[\max_{1\leq k\leq n} |S_k| \geq t\right]. \end{aligned}$$

From independence of $\{X_i\}_{i=1}^{n}$, $E\left[S_n^2\right] = \sum_{j=1}^{n}\sigma_j^2$ *by 3.30.* ■

Remark 3.42 *In a little more detail with all references to book 2, independence of* $\{X_i\}_{i=1}^{n}$ *implies independence of* $\{Y_i\}_{i=1}^{n}$ *with* $Y_i \equiv \chi_{A_k} X_i$ *for* $i \leq k$ *and* $Y_i \equiv X_i$ *otherwise, since as noted in remark 3.57,* $\sigma\left(Y_i\right) \subset \sigma\left(X_i\right)$. *The random vectors* $(Y_1, ..., Y_k)$ *and* $(Y_{k+1}, ..., Y_n)$ *are then independent by exercise 3.50. Then using the Borel measurable functions* $g_1(y_1, ..., y_k) \equiv \sum_{j=1}^{k} y_j$ *and* $g_2(y_{k+1}, ..., y_n) \equiv \sum_{j=k+1}^{n} y_j$ *obtains that* $S_1 \equiv g_1(Y_1, ..., Y_k)$ *and* $S_2 \equiv g_2(Y_{k+1}, ..., Y_n)$ *are independent random variables by proposition 3.56, and thus* $F(S_1, S_2) = F(S_1)F(S_2)$ *by proposition 3.53. Recalling notation 3.16 of this book:*

$$E^{(2)}[S_1 S_2] = E[S_1]E[S_2],$$

by an application of Fubini's theorem (book 5) to justify iterated integrals, and finally $E^{(2)}[S_1 S_2] = 0$ *since* $E[S_2] = 0$.

Hölder's and Related Inequalities

Hölder's inequality was derived by **Otto Hölder** (1859 – 1937) in 1884. It generalizes the **Cauchy-Schwarz inequality** which was originally

proved by **Augustin-Louis Cauchy** (1759-1857) in 1821 in the context of the n-dimensional real Euclidean space $\mathbb{R}^n$, and generalized 25 years later to all so-called "inner product spaces" by **Hermann Schwarz** (1843-1921).

To derive Hölder's inequality we require **Young's inequality**, which was derived by **W. H. Young** (1863-1942) in 1912, and who was the father of **L. C. Young** (1905 – 2000) noted in chapter 4 of book 3 in relation to Riemann-Stieltjes integration.

Proposition 3.43 (Young's Inequality) *Given p, q so that $1 < p, q < \infty$, and $\frac{1}{p} + \frac{1}{q} = 1$, then for all $a, b > 0$:*

$$ab \leq a^p/p + b^q/q. \tag{3.82}$$

Proof. *Note that $g(x) = \ln x$ is concave on $(0, \infty)$ by proposition 3.37 since $g'(x) = 1/x$ is a decreasing function. Applying 3.77 with $t = 1/p$:*

$$\begin{aligned} \ln(ab) &\equiv \left(\ln a^p\right)/p + \left(\ln b^q\right)/q \\ &\leq \ln\left(a^p/p + b^q/q\right), \end{aligned}$$

and the result in 3.82 follows by exponentiation. ■

Remark 3.44 *When p, q satisfy $1 < p, q < \infty$ and $\frac{1}{p} + \frac{1}{q} = 1$, they are called **conjugate indexes.** By defining $\frac{1}{\infty} = 0$, the pair $(1, \infty)$ is also called conjugate, and many results on conjugate indexes can be extended to this special case, including **Hölder's inequality** below.*

To set the stage for Hölder's inequality, assume that X, Y are random variables defined on a probability space $(\mathcal{S},\mathcal{E},\lambda)$. Then of course XY as well as $|XY|$ are also random variables and hence their expectations are defined by 3.7 when 3.8 is satisfied. For example, while

$$E[|XY|] \equiv \int_{\mathcal{S}} |X(s)Y(s)|\, d\lambda(s),$$

even the integrability of X and Y does not imply the integrability of XY. The expectation of $|X(s)Y(s)|$ can also be expressed in terms of a Riemann-Stieltjes integral over $\mathbb{R}^2$ reflecting the distribution function $F(x, y)$ and defined as in 3.23, but we do not need this here.

Example 3.45 *With* $(\mathcal{S},\mathcal{E},\lambda) = ([0,1),\mathcal{B}([0,1]),m)$, *the random variables* $X(s) = s^{-a}$ *and* $Y(s) = s^{a-1}$ *are both integrable for* $0 < a < 1$, *but* $XY = s^{-1}$ *is not. Jumping ahead to 3.83, not surprisingly it is also the case that for any such a there is no* $1 < p,q < \infty$ *with* $1/p + 1/q = 1$ *so that both* $E[|X|^p] < \infty$ *and* $E[|Y|^q] < \infty$, *as can be verified as an exercise.*

Proposition 3.46 (Hölder's Inequality) *Given* p,q *with* $1 < p,q < \infty$ *and* $1/p + 1/q = 1$, *assume that* $E[|X|^p]$ *and* $E[|Y|^q]$ *exist. Then* $E[|XY|]$ *exists and:*

$$E[|XY|] \leq (E[|X|^p])^{1/p} (E[|Y|^q])^{1/q}. \tag{3.83}$$

If $p = 1$ *and* $E[|X|]$ *and* $\sup[|Y|]$ *exist, then* $E[|XY|]$ *exists and:*

$$E[|XY|] \leq E[|X|] \sup[|Y|]. \tag{3.84}$$

Proof. *By Young's inequality with* $a = |X| / (E[|X|^p])^{1/p}$ *and* $b = |Y| / (E[|Y|^q])^{1/q}$:

$$\frac{|XY|}{(E[|X|^p])^{1/p} (E[|Y|^q])^{1/q}} \leq \frac{1}{p}\frac{|X|^p}{E[|X|^p]} + \frac{1}{q}\frac{|Y|^q}{E[|Y|^q]}.$$

So the existence of $E[|X|^p]$ *and* $E[|Y|^q]$ *assures the existence of* $E[|XY|]$. *Taking expectations, the right hand side reduces to* 1 *for conjugate indexes, and the inequality follows.*

If $p = 1$ *and* $q = \infty$, *the result follows directly from* $|Y| \leq \sup|Y|$:

$$E[|XY|] \leq \sup|Y| \, E[|X|].$$

■

Remark 3.47 **1**. *Note that when* $p = 1$, *it is not only logical to define* $q = \infty$ *to retain the identity* $1/p + 1/q = 1$, *but as will be seen in book 5 in the study of* L_p*-spaces, as* $q \to \infty$ *it will be seen that when* $\sup[|Y|] < \infty$ *that* $(E[|Y|^q])^{1/q} \to \sup[|Y|]$, *and thus 3.84 generalizes 3.83 in a natural way.*

2. *Hölder's inequality also provides an upper bound for* $|E[XY]|$ *once the triangle inequality in propositions 2.60 and 4.26 of book 3 is proved for general Lebesgue-Stieltjes integrals in book 5. In other words, once it is proved that* $|E[XY]| \leq E[|XY|]$, *then Hölder's inequality implies that:*

$$|E[XY]| \leq (E[|X|^p])^{1/p} (E[|Y|^q])^{1/q}, \tag{3.85}$$

and similarly for the Cauchy-Schwarz inequality below.

There are two important results which are now corollaries of Hölder's inequality. The first is the **Cauchy-Schwarz inequality,** which can now be seen as a special case of Hölder's inequality, but this result was proved much earlier as noted above. The second is **Lyapunov's inequality,** named for **Aleksandr Lyapunov** (1857 – 1918).

Corollary 3.48 (Cauchy-Schwarz Inequality) *Assume that* $E[|X|^2]$ *and* $E[|Y|^2]$ *exist. Then* $E[|XY|]$ *exists and:*

$$E[|XY|] \leq \left(E[|X|^2]\right)^{1/2} \left(E[|Y|^2]\right)^{1/2}. \tag{3.86}$$

Proof. *Apply 3.83 with* $p = q = 2$. ■

The Cauchy-Schwarz inequality provides an upper bound for the odd absolute moments of a random variable in terms of its even moments, and also provides a familiar estimate for the first moment of the product of two random variables.

Example 3.49 **1**. *Let* Z *be a random variable defined on a probability space* $(\mathcal{S},\mathcal{E},\lambda)$, *and for integer* k *let* $X = Z^k$ *and* $Y = Z^{k+1}$. *Then:*

$$E[|XY|] = E[|Z|^{2k+1}] \equiv \mu'_{|2k+1|},$$

and so by 3.86:

$$\mu'_{|2k+1|} \leq \left(\mu'_{2k}\mu'_{2k+2}\right)^{1/2},$$

and similarly for central moments defining $X = (Z - \mu_Z)^k$ *and* $Y = (Z - \mu_Z)^{k+1}$*:*

$$\mu_{|2k+1|} \leq \left(\mu_{2k}\mu_{2k+2}\right)^{1/2}.$$

Using 3.85, these inequalities then provide upper bounds for the absolute value of the odd moments μ'_{2k+1} *and* μ_{2k+1}. *Thus:*

$$\left|\mu'_{2k+1}\right| \leq \left(\mu'_{2k}\mu'_{2k+2}\right)^{1/2}, \qquad \left|\mu_{2k+1}\right| \leq \left(\mu_{2k}\mu_{2k+2}\right)^{1/2}. \tag{3.87}$$

2. *Applying the Cauchy-Schwarz inequality to random variables* $X - \mu_X$ *and* $Y - \mu_Y$, *we derive that with transparent notation,*

$$E[|(X - \mu_X)(Y - \mu_Y)|] \leq \sigma_X\sigma_Y,$$

which implies as noted above in 3.85 that:

$$-\sigma_X\sigma_Y \leq E[(X - \mu_X)(Y - \mu_Y)] \leq \sigma_X\sigma_Y. \tag{3.88}$$

Recalling definition 3.19 we obtain:

Corollary 3.50 *Given random variables X, Y, with finite second moments, the covariance $cov(X,Y)$ exists and satisfies:*

$$|cov(X,Y)| \leq \sigma_X \sigma_Y.$$

Hence, the correlation $corr(X,Y)$ is bounded:

$$|corr(X,Y)| \leq 1. \tag{3.89}$$

Finally, we present **Lyapunov's inequality**, which provides a lower bound on the growth rate of moments.

Corollary 3.51 (Lyapunov's Inequality) *Given a random variable X, then for $0 < \alpha < \beta$, assuming all moments exist:*

$$\left(E[|X|^{\alpha}]\right)^{1/\alpha} \leq \left(E[|X|^{\beta}]\right)^{1/\beta}. \tag{3.90}$$

Proof. *Let $p = \beta/\alpha$, then apply Hölder's inequality to $|X|^{\alpha}$ and $Y \equiv 1$.* ■

Example 3.52 *Lyapunov's inequality implies that absolute moments must grow at least geometrically, so that for any $1 \leq m$:*

1. $\left(\mu'_{|m|}\right)^{1/m}$ *and* $\left(\mu_{|m|}\right)^{1/m}$ *are increasing sequences.*
2. *For $m < n$*

$$\left(\mu'_{|m|}\right)^{n/m} \leq \mu'_{|n|}.$$

For example, with $m = 1$:

$$\mu^n \leq \mu'_{|n|}.$$

3.2.7 Uniqueness of Moments and the M.G.F.

It would be interesting to know if distributions were uniquely determined by their moments, $\{\mu'_n\}_{n=1}^{\infty}$, assuming all existed. Since 3.36 provides a unique mapping between $\{\mu'_n\}_{n=1}^{\infty}$ and $\{\mu_n\}_{n=1}^{\infty}$, the answer to this question is independent of which set of moments are used. For example, is the normal distribution the only such distribution for which central moments are given by 3.65 as:

$$\mu_{2m} = \sigma^{2m} (2m)! / (2^m m!), \qquad \mu_{2m+1} = 0?$$

Is the lognormal distribution the only such distribution for which moments are given by 3.68 as: $\mu'_{nL} = e^{n\mu + (n\sigma)^2/2}$?

It turns out that not all moment collections uniquely identify the underlying distribution functions. As will be seen, the moment collection for the normal distribution is unique to the normal, but there are infinitely many distribution functions with moments equal to those of the lognormal distribution. The following example was introduced in 1963 by **C. C. Heyde** (1939 – 2008).

Example 3.53 (Heyde) *Recall the lognormal distribution in 1.34 with* $\mu = 0$ *and* $\sigma = 1$:

$$f_L(x) = \frac{1}{x\sqrt{2\pi}} \exp\left(-\left(\ln x\right)^2/2\right), \qquad x \geq 0,$$

and define for $-1 \leq a \leq 1$:

$$f_a(x) = f_L(x)\left[1 + a\sin(2\pi \ln x)\right].$$

Then $f_0(x) = f_L(x)$ *and* $f_a(x) \geq 0$ *on* $x \geq 0$ *because* $\sin x \geq -1$.

To see that $f_a(x)$ *is a density function and has the same moments as* $f_L(x)$ *we show that for* $n = 0, 1, 2, ...,$

$$I_n \equiv \int_{-\infty}^{\infty} x^n f_L(x)\sin(2\pi \ln x)dx = 0.$$

To this end, the substitution $x = \exp(y + n)$ *produces:*

$$\begin{aligned} I_n &= (2\pi)^{-1/2}\int_{-\infty}^{\infty} \exp(yn + n^2)\exp\left(-\left(y+n\right)^2/2\right)\sin(2\pi(y+n))dy \\ &= (2\pi)^{-1/2}\exp(n^2/2)\int_{-\infty}^{\infty} \exp\left(-y^2/2\right)\sin(2\pi y)dy \\ &= 0. \end{aligned}$$

This follows because the integrand $g(y) = \exp\left(-y^2/2\right)\sin(2\pi y)$ *is absolutely integrable, and is an* ***odd function*** *meaning that* $g(-y) = -g(y)$.

Consequently, for any such a, $f_a(x)$ *is a density function with the same moments as the lognormal* $f_L(x)$.

The next proposition states that given the distribution function $F(x)$, if all moments exist and the power series $\sum_{n=0}^{\infty} \mu_n' t^n/n!$ converges for $t \in (-t_0, t_0)$ with $t_0 > 0$, then $M_F(t)$ exists and by proposition 3.24 is given by this power series. The subscript X is suppressed in this notation since this is really a statement about distribution functions, $F(x)$, and the particular probability space and random variable that gives rise to $F(x)$ is not relevant.

Remark 3.54 *It should be noted that we will not then be able to conclude that $F(x)$ is the **only** distribution function with these moments. That is, it could well be possible that there exists a second distribution function $G(x)$ so that $M_F(t) = M_G(t)$ on $(-t_0, t_0)$ and thus $F(x)$ and $G(x)$ have the same moments, yet $F(x) \neq G(x)$. But see proposition 3.57.*

Proposition 3.55 (Existence of M.G.F.) *Given the distribution function $F(x)$, assume that $\mu'_n \equiv \int_{-\infty}^{\infty} x^n dF$ exists for all n and the power series $\sum_{n=0}^{\infty} \mu'_n t^n/n!$ converges absolutely on $(-t_0, t_0)$ with $t_0 > 0$. Then the moment generating function $M_F(t)$ of $F(x)$ exists on $(-t_0, t_0)$, and is given by this series.*

Proof. *Note that for $t \in (-t_0, t_0)$ the triangle inequality of proposition 4.26 of book 3 obtains:*

$$\left| \int_{\mathbb{R}} \sum\nolimits_{j=0}^{n} (tx)^j / j! dF(x) \right| \leq \sum\nolimits_{j=0}^{n} \int_{\mathbb{R}} \left| (tx)^j / j! \right| dF(x) = \sum\nolimits_{j=0}^{n} |t|^j \mu'_{|j|}/j!,$$

where $\mu'_{|j|}$ denotes the absolute jth moment, $\mu'_{|j|} = E\left[|X|^j\right]$. We show below that this upper limit is bounded for all n for any such t.

Assuming that:

$$\begin{aligned} \left| \int_{\mathbb{R}} e^{tx} dF(x) - \sum\nolimits_{j=0}^{n} \mu'_j t^j/j! \right| &= \left| \int_{\mathbb{R}} e^{tx} dF(x) - \int_{\mathbb{R}} \sum\nolimits_{j=0}^{n} (tx)^j / j! dF(x) \right| \\ &\leq \int_{\mathbb{R}} \left| \sum\nolimits_{j=n+1}^{\infty} (tx)^j / j! \right| dF(x) \\ &\leq \sum\nolimits_{j=n+1}^{\infty} |t|^j \mu'_{|j|}/j!. \end{aligned}$$

For this last inequality, by the triangle inequality and linearity of the integral:

$$\int_{\mathbb{R}} \left| \sum\nolimits_{j=n+1}^{m} (tx)^j / j! \right| dF(x) \leq \sum\nolimits_{j=n+1}^{m} |t|^j \mu'_{|j|}/j! \leq \sum\nolimits_{j=n+1}^{\infty} |t|^j \mu'_{|j|}/j!, \qquad ((*))$$

and this is true for all m, and thus is true for the limit superior of the left hand side. Thus if $\sum_{j=0}^{n} |t|^j \mu'_{|j|}/j!$ is bounded for all n, then the upper bound in $(*)$ can be made arbitrarily small for n large and hence $M_F(t) \equiv \int_{\mathbb{R}} e^{tx} d\mu_F(x)$ exists and is given by this series.

To this end, by example 3.49:

$$\mu'_{|2n+1|} \leq \left(\mu'_{2n} \mu'_{2n+2}\right)^{1/2} \leq \max[\mu'_{2n}, \mu'_{2n+2}] \equiv \mu'_{2n^+}.$$

In other words, the index $2n^+$ *is defined:*

$$2n^+ \equiv \begin{cases} 2n, & \mu'_{2n} \geq \mu'_{2n+2}, \\ 2n+2, & \mu'_{2n} < \mu'_{2n+2}. \end{cases}$$

Now let $s = \lambda t$ *where* $\lambda < 1$*, then:*

$$\frac{\mu'_{|2n+1|}\,|s|^{2n+1}}{(2n+1)!} \leq c_n \frac{\mu'_{2n^+}\,|t|^{2n^+}}{(2n^+)!},$$

where:

$$c_n \equiv \begin{cases} |t|\,\lambda^{2n+1}/(2n+1), & 2n^+ = 2n, \\ (2n+2)\,\lambda^{2n+1}/\,|t|\,, & 2n^+ = 2n+2, \end{cases}$$

and in either case $c_n \to 0$ *as* $n \to \infty$*. Now for each* n*, if* $2n^+ = 2n$ *define* $d_n^- = c_n$ *and* $d_n^+ = 0$*, while if* $2n^+ = 2n+2$ *define* $d_n^- = 0$ *and* $d_n^+ = c_n$*. Then since* $\mu'_{|2n|} = \mu'_{2n}$ *by definition, by splitting the summation into even and odd indexes obtains:*

$$\sum\nolimits_{j=0}^{2n} |s|^j\,\mu'_{|j|}/j! \leq \sum\nolimits_{j=0}^{n} |t|^{2j}\,\mu'_{2j}\left(1 + d_{j+1}^- + d_j^+\right)/\,(2j)!.$$

Because $d_{j+1}^- + d_j^+ \to 0$*, this partial sum in* s *converges as* $n \to \infty$ *for* $|s| < t_0$ *because* $\sum_{j=0}^{\infty} t^j \mu'_j/\,(j)!$ *is absolutely convergent for* $|t| < t_0$ *by assumption.* ■

The following result provides an alternative test for the existence of $M_F(t)$ given $\{\mu'_n\}_{n=1}^{\infty}$, stated in terms of a growth bound on even moments.

Corollary 3.56 (Existence of M.G.F.) *Given the distribution function* $F(x)$*, assume that* $\mu'_n \equiv \int_{-\infty}^{\infty} x^n dF$ *exists for all* n *and:*

$$\limsup\left(\mu'_{2n}\right)^{1/2n}/2n = r < \infty.$$

Then $M_F(t)$ *exists for* $|t| < 1/r$*, and by proposition 3.24 is given by the series in 3.38 on this interval.*

Proof. *The assumed bounded limit superior implies that for all but at most finitely many* n *that:*

$$\mu'_{2n} \leq (2nr)^{2n},$$

and hence by 3.87 for all but at most finitely many n :

$$\begin{aligned}\left|\mu'_{2n+1}\right| &\leq \left(\mu'_{2n}\mu'_{2n+2}\right)^{1/2} \\ &\leq \left(2\left(n+1\right)r\right)^{2n+1}.\end{aligned}$$

Hence in all cases of even or odd m, with at most finitely many exceptions:

$$\left|\mu'_m\right| \leq \left(\left(m+1\right)r\right)^m,$$

and so

$$\begin{aligned}\sum_{m=0}^{\infty}\left|\mu'_m\right|t^m/m! &\leq \sum_{m=0}^{\infty}\left(t\left(m+1\right)r\right)^m/m! \\ &= \sum_{m=0}^{\infty}\left(tr\right)^m\frac{\left(m+1\right)^m}{m!}.\end{aligned}$$

Using the ratio test, it follows that this series is convergent if $|t| < 1/r$, and the result follows from proposition 3.55. ■

The following proposition, to be proved in book 6 using the properties of the **characteristic function of** $F(x)$, provides a key test for when a moment collection uniquely defines a distribution function. Namely and in response to remark 3.54, the absolute convergence of $\sum_{n=0}^{\infty}\mu'_n t^n/n!$ on $(-t_0, t_0)$ with $t_0 > 0$ assures that $F(x)$ is the only distribution function with these moments. Corollary 3.58 below provides another test that uses this result.

Proposition 3.57 (Uniqueness of Moments) *Given the distribution function $F(x)$, assume that $\mu'_n \equiv \int_{-\infty}^{\infty}x^n dF$ exists for all n and that $\sum_{n=0}^{\infty}\mu'_n t^n/n!$ converges absolutely on $(-t_0, t_0)$ with $t_0 > 0$. Then $F(x)$ is the only distribution function with these moments.*

Proof. *See the section on the uniqueness of moments in the book 6 section, The Characteristic Function.* ■

The following corollary provides a converse to the results of propositions 3.22 and 3.24 in section 3.2.4. These stated that if $M_X(t)$ exists on $t \in (-t_0, t_0)$ with $t_0 > 0$, then so too do all moments of X, and for $t \in (-t_0, t_0)$:

$$M_X(t) = \sum\nolimits_{n=0}^{\infty}\mu'_n t^n/n!,$$

and thus $\mu'_n = M_X^{(n)}(0)$.

Corollary 3.58 (Uniqueness of the M.G.F.) *Given the distribution function $F(x)$, assume that $M_F(t) \equiv \int_{-\infty}^{\infty} e^{tx} dF$ exists on $(-t_0, t_0)$ with $t_0 > 0$. Then $F(x)$ is the only distribution function with this moment generating function.*

Proof. *If $F(x)$ has moment generating function $M_F(t)$ which converges on the interval $(-t_0, t_0)$ with $t_0 > 0$, then by proposition 3.24 $F(x)$ has moments of all orders defined by $\mu'_n = M^{(n)}(0)$ and the associated power series $\sum_{n=0}^{\infty} \mu'_n t^n/n!$ converges absolutely on $(-t_0, t_0)$. If $G(x)$ is another distribution function with moment generating function $M_F(t)$ convergent on $(-t_0, t_0)$, then by the same argument $G(x)$ would have the same moments as $F(x)$, contradicting proposition 3.57.*

Hence, $F(x)$ is the only distribution function with this moment generating function. ■

Example 3.59 *In this example we apply the above uniqueness results to resolve statements made earlier in this book.*

1. ***Normal vs. Lognormal:*** *We can now readily verify the statements in the introduction to this section concerning the normal and lognormal distributions.*

   *The **normal distribution** is uniquely defined by the moments, $\mu_{2m} = \sigma^{2m} (2m)!/ (2^m m!)$ and $\mu_{2m+1} = 0$. To see this, note that the moment generating function, $M_N(t)$ exists for all t, and hence the series expression in 3.38 is convergent for all t and so by proposition 3.57 above, this distribution function is uniquely determined by these moments. One can also check that the even moments satisfy the growth bound in corollary 3.56 using **Stirling's formula.** Stirling's formula, also known as **Stirling's approximation,** is named for **James Stirling** (1692 – 1770) and states that as $n \to \infty$:*

   $$n!/ \left(\sqrt{2\pi} n^{n+1/2} e^{-n} \right) \approx e^{1/12n} \to 1. \tag{3.91}$$

   See for example Reitano.

   *On the other hand, the moments of the **lognormal distribution**, $\mu'_n = e^{n\mu + (n\sigma)^2/2}$, do not provide a power series $\sum_{n=0}^{\infty} \mu'_n t^n/n!$ that is convergent on an interval $(-t_0, t_0)$ with $t_0 > 0$ as demonstrated in remark 3.30. Alternatively, looking to the the even moments,*

   $$\left(\mu'_{2n}\right)^{1/2n} /2n = e^{\mu + n\sigma^2}/2n,$$

is unbounded in n. These results are not conclusive in verifying that there are other distributions with these moments, only that the above theory cannot be applied to assure uniqueness. That said, the proof of nonuniqueness is obtained with example 3.53 of ***C. C. Heyde*** *above.*

2. ***Sums of Geometric are Negative Binomial:*** *It was noted in remark 3.29 that for the geometric distribution that when* $(1-p)e^t < 1$ *or equivalently* $t < -\ln(1-p)$*, that* $M_G(t) = p/\left[1-(1-p)e^t\right]$ *by 3.50, while for the negative binomial and the same range of t that* $M_{NB}(t) = \left(p/\left[1-(1-p)e^t\right]\right)^k$ *by 3.52. This latter moment generating function is also recognized to be the moment generating function of a sum of k independent geometric random variables by 3.35. But by corollary 3.58 the negative binomial is the only distribution with this moment generating function, and thus a sum of independent geometric random variables is a negative binomial random variable.*

3. ***Sums of Gamma are Gamma:*** *It was demonstrated in example 1.18 that a sum of independent exponentials are gamma, and then noted in exercise 1.19 that in fact the sum of independent gammas with common* λ *is also gamma. To see this, let* $\{X_i\}_{i=1}^n$ *be independent gammas with parameters* λ *and* $\{\alpha_i\}_{i=1}^n$*. By 3.59* $M_i(t) = (1-t/\lambda)^{-\alpha_i}$ *for* $t < \lambda$*, and thus by 3.35 the moment generating function of* $X \equiv \sum_{i=1}^n X_i$ *is* $M(t) = (1-t/\lambda)^{-\alpha}$ *for* $t < \lambda$ *with* $\alpha \equiv \sum_{i=1}^n \alpha_i$*. Again by corollary 3.58 the gamma is the only distribution with this moment generating function, and thus a sum of independent gamma random variables with common* λ *is a gamma random variable with parameters* λ *and* $\alpha = \sum_{i=1}^n \alpha_i$*.*

4. ***Student T Distribution:*** *Recall the discussion in example 1.22 of the Student T distribution. Simplifying notation, we had concluded that there were random variables A, B, and C with* $A + B = C$*, where A and B were independent and both A and C where Chi-squared with* 1 *and n degrees of freedom, respectively. The claim there was that this assures that B is also Chi-squared, and with* $n-1$ *degrees of freedom. To see this, first note that the moment generating functions of A and C are given in 3.59 with* $\lambda_A = \lambda_C = 1/2$*,* $\alpha_A = 1/2$ *and* $\alpha_C = n/2$*. An application of 3.35 and corollary 3.58 then obtains that B is also Chi-squared with* $n-1$ *degrees of freedom, but this application must be justified by first proving that B does indeed have a moment generating function. This is proved by noting that the moments of A and B sum to the moments of C, and thus the convergence of the moment series*

for A and C assures the convergence of the moment series for B, and thus by proposition 3.55 the moment generating function of B exists.

3.2.8 Weak Convergence and Moment Limits

Assume that a distribution function F is uniquely determined by its moment collection, $\{\mu'_n\}_{n=1}^{\infty}$. Given a sequence of distribution functions $\{F_m\}_{m=1}^{\infty}$ with moment collections $\{\mu'_{m,n}\}_{n=1}^{\infty}$ where $\mu'_{m,n} \to \mu'_n$ for each n, are we then able to conclude that $F_m \Rightarrow F$, that F_m **converges weakly to the distribution function** F? Recalling the definitions in section 8.1 of book 2 this means that $F_m(x) \to F(x)$ for every continuity point of F. Equivalently, the notion $F_m \Rightarrow F$ can be stated other ways:

1. Given random variables $\{X_m\}_{m=1}^{\infty}$ underlying these $\{F_m\}_{m=1}^{\infty}$, $F_m \Rightarrow F$ is equivalent to $X_m \Rightarrow X$, which is to say that X_m **converges in distribution to a random variable** X underlying F.

2. Given the Borel measures induced by these F_m, $\{\mu_{F_m}\}_{m=1}^{\infty}$, $F_m \Rightarrow F$ is equivalent to $\mu_{F_m} \Rightarrow \mu_F$, which is to say that μ_{F_m} **converges weakly to the Borel measure** μ_F induced by F.

The **method of moments** in probability theory is the name sometimes given to the framework within which one can assert the conclusion that $F_m \Rightarrow F$ by demonstrating that that $\mu'_{m,n} \to \mu'_n$ for each n, or conversely. The term "method of moments" is also the name given to the process whereby one estimates the parameters of a given distribution function based on the moments of a sample, by equating the distribution's parametric moment formulas to the numerical values calculated.

Remark 3.60 *In the book 6 section, Weak Convergence and Expectations, the analysis of this section will be reversed and generalized with the aid of the integration theory of book 5. Specifically, we will investigate when $F_m \Rightarrow F$ assures the convergence of expectations, $E[g(X_n)] \to E[g(X)]$, for certain measurable functions g. A more limited result in this direction is found below and also seen in book 5 in the section A Result on Weak Convergence of Measures.*

Any method of moments must certainly require that the distribution function F be uniquely determined by its moment collection, $\{\mu'_n\}_{n=1}^{\infty}$. For example, if $\{F_m\}_{m=1}^{\infty}$ are given and $\mu'_{m,n} \to e^{n\mu+(n\sigma)^2/2}$ for each n, even though these limits are recognizable as the moments of the lognormal distribution, it was shown in example 3.53 that these are also the moments of

an infinite number of other distribution functions. So certainly there can be no statement concerning $F_m \Rightarrow F$.

The main result of this section in proposition 3.73 states that in the case where F is uniquely defined by its moments, and if $\mu'_{m,n} \to \mu'_n$ for each n then $F_m \Rightarrow F$. Corollary 3.73 provides the same conclusion based on convergence of moment generating functions. For these results we will first need a positive result in the opposite direction. Namely, if $F_m \Rightarrow F$ with moment collections $\{\{\mu'_{m,n}\}_{n=1}^{\infty}\}_{m=1}^{\infty}$ and $\{\mu'_n\}_{n=1}^{\infty}$, must it be the case that $\mu'_{m,n} \to \mu'_n$ for each n? Perhaps surprisingly, the answer is in the negative as illustrated in the following examples and discussion.

For this discussion recall the notion of **tightness** of a family of measures as defined in definition 8.16 in book 2.

Definition 3.61 *A sequence of probability measures $\{\mu_n\}$ is said to be **tight** if for any $\epsilon > 0$ there is a finite interval $(a, b]$ so that $\mu_n((a, b]) > 1 - \epsilon$ for all n. A sequence of distribution functions $\{F_n\}$ is said to be **tight** if for any $\epsilon > 0$ there is a finite interval $(a, b]$ so that $F_n(b) - F_n(a) > 1 - \epsilon$ for all n, or equivalently $F_n(b) > 1 - \epsilon$ and $F_n(a) < \epsilon$ for all n.*

Example 3.62 *We provide an example of a sequence of measures which is not tight, then one that is tight, and see that in neither case does weak convergence imply the convergence of moment sequences. Hence, a property stronger than tightness will be needed for a positive conclusion.*

1. $\{F_m\}_{m=1}^{\infty}$ ***is not tight:***

   Define discrete density functions $\{f_m\}$ by $f_m(m) = 1$, $f_m(x) = 0$ for $x \neq m$, and thus $F_m(x) = \chi_{[m,\infty)}(x)$. Clearly, $F_m(x) \to F(x)$ for all x with F defined by $F(x) = 0$, but by definition it is not the case that $F_m \Rightarrow F$ since F is not even a distribution function. That F is not a distribution function is not a surprise from the fact that the associated Borel measures, $\{\mu_{F_m}\}_{m=1}^{\infty}$ are not tight, and the result assigned in exercise 8.21 of book 2. In such a case, there is always a subsequence of measures in the Helly selection process of proposition 8.14 of book 2 which converge to function which is not a distribution function.

   In this case $\mu'_n = 0$ for all n and $\mu'_{m,n} = m^n$ and so it is not that case that $\mu'_{m,n} \to \mu'_n$ for any n.

   ***Conjecture:** It is natural to hypothesize that if $F_m \Rightarrow F$ and F is a distribution function, that we will obtain the positive result that $\mu'_{m,n} \to \mu'_n$ for all n. By proposition 8.18 of book 2, if F is a distribution*

function and $F_m \Rightarrow F$ then $\{F_m\}_{m=1}^{\infty}$ is tight, so we next look at such an example.

2. $\{F_m\}_{m=1}^{\infty}$ ***is tight:***

Define discrete density functions $\{f_m\}$ by $f_m(0) = 1 - 1/m$, $f_m(m) = 1/m$, and $f_m(x) = 0$ for $x \neq 0, m$. Then the associated distribution functions are defined:

$$F_m(x) = \begin{cases} 0, & x < 0, \\ 1 - 1/m, & 0 \leq x < m, \\ 1, & m \leq x. \end{cases}$$

The associated collection of Borel measures $\{\mu_{F_m}\}_{m=1}^{\infty}$ is tight since given $\epsilon > 0$ and $1/m_0 < \epsilon$, then if $a < 0$, $\mu_{F_m}[(a, m_0]] > 1 - \epsilon$ for all m. Not surprisingly, $F_m \Rightarrow F$ with distribution function $F(x) = \chi_{[0,\infty)}(x)$. However, $\mu'_{m,n} = m^{n-1}$ and $\mu'_n = 0$ and so it is not the case that $\mu'_{m,n} \to \mu'_n$ for any n.

Conclusion: *If $F_m \Rightarrow F$ and F is a distribution function, an additional restriction is needed on the sequence beyond tightness to assure convergence of moments.*

Before proceeding we investigate the second example further and more generally, and demonstrate why the desired conclusion is not achieved even with tightness. To this end, assume that $\{F_m\}_{m=1}^{\infty}$ is tight and $F_m \Rightarrow F$ for a distribution function F. Recall that by **Skorokhod's representation theorem** in proposition 8.30 of book 2 that we can then define random variables $\{X_m\}_{m=1}^{\infty}$ and X on the Lebesgue measure space $((0,1),\mathcal{B}(0,1),m_L)$, with respective distribution functions $\{F_m\}_{m=1}^{\infty}$ and F, and for which $X_m \to X$ for all $t \in (0,1)$. We then have as in 3.7 but now as **Lebesgue integrals:**

$$\mu'_{m,n} = \int_0^1 [X_m(t)]^n \, dm_L, \qquad \mu'_n = \int_0^1 [X(t)]^n \, dm_L. \qquad ((*))$$

In the second example an application of the Skorokhod construction yields that $X_m(t) \equiv 0$ for $t \in (0, 1 - 1/m]$ and $X_m(t) \equiv m$ for $t \in (1 - 1/m, 1)$, while $X(t) \equiv 0$. As assured by the Skorokhod representation theorem, since $F_m \Rightarrow F$ it follows that $X_m \to X$ for all $t \in (0,1)$. Not surprisingly, $\mu'_{m,n} = m^{n-1}$ and $\mu'_n = 0$ as before, but now calculated as in $(*)$.

The assumption that $\{\mu_{F_m}\}_{m=1}^{\infty}$ is tight implies that for any $\epsilon > 0$ there is an interval, say $(-N_\epsilon, N_\epsilon]$, such that for all m :

$$\mu_{F_m}\left[(-N_\epsilon, N_\epsilon]\right] \equiv m_L\left[X_m^{-1}(-N_\epsilon, N_\epsilon]\right] \geq 1 - \epsilon.$$

Hence $m_L\left[X_m^{-1}\left[(-\infty, -N_\epsilon] \cup (N_\epsilon, \infty)\right]\right] \leq \epsilon$. We can replace $(-N_\epsilon, N_\epsilon]$ by a slightly larger open interval for notational simplicity, and then have with the same notation:

$$\mu'_{m,n} = \int_{|X_m|<N_\epsilon} [X_m(t)]^n \, dm_L + \int_{|X_m|\geq N_\epsilon} [X_m(t)]^n \, dm_L.$$

With a bit of care in the general case as seen in proposition 3.66 below, we expect that the first integral will convergence as $m \to \infty$ by the bounded convergence theorem of book 3. This is because for any N_ϵ, $m_L\left[|X_m| < N_\epsilon\right] \leq m_L\left[(0,1)\right] = 1$, while by definition $|X_m(t)| \leq N_\epsilon$ on this set for all m, and $X_m \to X$ for all $t \in (0,1)$. Thus the bounded convergence theorem applies and obtains that the first integral will converge:

$$\int_{|X_m|<N_\epsilon} [X_m(t)]^n \, dm_L \to \int_{|X|<N_\epsilon} [X(t)]^n \, dm_L.$$

Hence given tightness the real challenge to achieving the desired result, that $\mu'_{m,n} \to \mu'_n$ for each n, is the convergence of the second integral with unbounded integrands. In other words, the outstanding question is, how can it be assured that:

$$\int_{|X_m|\geq N_\epsilon} [X_m(t)]^n \, dm_L \to \int_{|X_m|\geq N_\epsilon} [X(t)]^n \, dm_L,$$

knowing only that $X_m(t) \to X(t)$ for all t and that $m_L\{t|\,|X_m(t)| \geq N_\epsilon\} \leq \epsilon$?

The answer is that we can not be so assured, and the second example above illustrates why. Namely, knowing that $m_L\{t|\,|X_m(t)| \geq N_\epsilon\} \leq \epsilon$ does not assure that the integrals of $[X_m(t)]^n$ over this set even remain bounded.

Returning to book 2, pointwise convergence assures convergence of the associated Lebesgue integrals in 3 cases:

1. **Bounded convergence theorem:** This is not applicable since $[X_m(t)]^n$ need not be bounded;

2. **Lebesgue's monotone convergence theorem:** Applicable only if we assume $\{X_m(t)\}$ are nonnegative and the convergence $X_m(t) \to X(t)$ is monotonically increasing for all t;

3. **Lebesgue's dominated convergence theorem:** Applicable only if we assume $|X_m(t)| \leq Y(t)$ for Lebesgue integrable Y.

Since it is undesirable to assume that $\{X_m(t)\}$ are nonnegative and monotonically increasing, the desired result can best be achieved by an assumption such as $|X_m(t)| \leq Y(t)$ for some Lebesgue integrable Y. As it turns out there is a somewhat weaker assumption, called **uniform integrability,** which will serve the same purpose. We introduce this definition next in the context of a general probability space even though we only require this notion for Lebesgue integrals. The general integrals in this definition will be developed in book 5.

Definition 3.63 *Given a probability space $(\mathcal{S},\mathcal{E},\lambda)$, a sequence of random variables $\{X_m\}_{m=1}^{\infty}$ is said to be* ***uniformly integrable*** *if:*

$$\lim_{N\to\infty} \sup_m \int_{|X_m|\geq N} |X_m(s)|\, d\lambda(s) = 0. \tag{3.92}$$

Remark 3.64 **1**. *Note that if $N \geq N'$ then for all m :*

$$\int_{|X_m|\geq N} |X_m(s)|\, d\lambda(s) \leq \int_{|X_m|\geq N'} |X_m(s)|\, d\lambda(s),$$

and thus the same is true of the associated suprema. Consequently, the limit $N \to \infty$ in the definition of uniform integrability can be interpreted quite generally, for example with N denoting all real numbers, or integers, or any increasing collection of reals.

2. *Given uniformly integrable $\{X_m\}_{m=1}^{\infty}$, the terminology reflects the fact that if N is large enough to ensure that the above supremum is less that* 1, *then*

$$\int_{\mathcal{S}} |X_m(s)|\, d\lambda(s) \leq N\lambda\left[X_m^{-1}\left[(-N,N)\right]\right] + 1 \leq N + 1,$$

and thus $\{X_m\}_{m=1}^{\infty}$ are all ***integrable****, and these integrals are* ***uniformly bounded****.*

3. *Also note that uniform integrability is assured if $\{X_m\}_{m=1}^{\infty}$ are dominated by integrable Y, meaning that $|X_m(s)| \leq |Y(s)|$ for all m and $\int_{\mathcal{S}} |Y(s)|\, d\lambda(s) < \infty$. And this result is true in a general measure space, not just a probability space which is a finite measure space. But uniform integrability is weaker than the assumption that $|X_m(s)| \leq |Y(s)|$ for such Y, and exercise 3.65 below assigns the development of an example.*

4. *Finally, if* $\{X_m\}_{m=1}^{\infty}$ *are integrable and uniformly bounded, so* $|X_m(s)| \leq c$ *for all* m*, this also implies uniform integrability in a probability space, but not in a general measure space as the example of* $X_m \equiv \chi_{[m.m+1]}(x)$ *defined on* $(\mathbb{R}, \mathcal{B}(\mathbb{R}), m)$ *exemplifies.*

Exercise 3.65 *On the Lebesgue probability space* $((0,1),\mathcal{B}(0,1),m_L)$*, develop an example of uniformly integrable random variables* $\{X_n\}$ *for which there is no integrable* Y *with* $|X_n| \leq |Y|$*. Hint:* $\{X_n\}$ *must be unbounded and uniformly integrable, so for example, make them tightly bounded by* $Y(s) = 1/s$ *which is not integrable.*

We now have the tools needed to make a statement about what $F_m \Rightarrow F$ says about the convergence of moments. This result will be critical for the main proposition on the method of moments, which will be followed by a corollary result on the convergence of moment generating functions.

Proposition 3.66 *Let* $\{F_m\}_{m=1}^{\infty}$ *and* F *be distribution functions with respective moment collections* $\{\{\mu'_{m,n}\}_{n=1}^{\infty}\}_{m=1}^{\infty}$ *and* $\{\mu'_n\}_{n=1}^{\infty}$*. If* $F_m \Rightarrow F$ *and* $\{\mu'_{m,2n}\}_{m=1}^{\infty}$ *is bounded for every* n *then* $\mu'_{m,n} \to \mu'_n$ *for all* n*.*

Proof. *By Skorokhod's representation theorem of proposition 8.30 of book 2 we can define random variables* $\{X_m\}_{m=1}^{\infty}$ *and* X *on the Lebesgue measure space* $((0,1),\mathcal{B}(0,1),m_L)$ *with respective distribution functions* $\{F_m\}_{m=1}^{\infty}$ *and* F*, and for which* $X_m \to X$ *for all* $t \in (0,1)$*. For given* N *define:*

$$X_m^{(N)} = \begin{cases} X_m, & |X_m| < N, \\ 0, & |X_m| \geq N, \end{cases} \qquad X^{(N)} = \begin{cases} X, & |X| < N, \\ 0, & |X| \geq N. \end{cases}$$

Then for each N*,* $\left[X_m^{(N)}\right]^n \to \left[X^{(N)}\right]^n$ *for all* $t \in (0,1)$*, and so by the bounded convergence theorem of proposition 2.37 of book 3:*

$$\int_{(0,1)} \left[X_m^{(N)}(t)\right]^n dm_L \to \int_{(0,1)} \left[X^{(N)}\right]^n dm_L.$$

Also:

$$\int_{(0,1)} [X_m(t)]^n \, dm_L - \int_{(0,1)} \left[X_m^{(N)}(t)\right]^n dm_L = \int_{|X_m(t)| \geq N} [X_m(t)]^n \, dm_L,$$
$$\int_{(0,1)} [X(t)]^n \, dm_L - \int_{(0,1)} \left[X^{(N)}(t)\right]^n dm_L = \int_{|X(t)| \geq N} [X(t)]^n \, dm_L,$$

and from the bounded convergence result we conclude that

$$\begin{aligned}&\limsup_m \left| \int_{(0,1)} [X_m(t)]^n \, dm_L - \int_{(0,1)} [X(t)]^n \, dm_L \right| \\ \leq\ &\sup_m \int_{|X_m(t)|\geq N} |X_m(t)|^n \, dm_L + \int_{|X(t)|\geq N} |X(t)|^n \, dm_L.\end{aligned}$$

The existence of μ'_n assures that the second integral on the right can be made arbitrarily small as $N \to \infty$. To complete the proof, we show that the assumption on boundedness of moments assures that $\{[X_m(t)]^n\}$ are uniformly integrable for any n, and thus the first term on the right converges to zero as $N \to \infty$ by definition.

To this end, the assumed boundedness of $\{\mu'_{m,2n}\}_{m=1}^{\infty}$ implies that

$$\sup_m \int_{(0,1)} |X_m(t)|^{2n} \, dm_L = K_n < \infty.$$

Thus,

$$\begin{aligned}\sup_m \int_{|X_m(t)|\geq N} |X_m(t)|^n \, dm_L \ &\leq\ \sup_m \int_{|X_m(t)|\geq N} |X_m(t)|^{2n} \, dm_L / N^n \\ &\leq\ K_n / N^n,\end{aligned}$$

and so $\{[X_m(t)]^n\}_{m=1}^{\infty}$ are uniformly integrable. ■

Remark 3.67 *Note that for the above proposition, the existence of all moments for F was assumed but could have been part of the conclusion.*

Corollary 3.68 *Let $\{F_m\}_{m=1}^{\infty}$ be distribution functions with moment collections $\{\{\mu'_{m,n}\}_{n=1}^{\infty}\}_{m=1}^{\infty}$. If $F_m \Rightarrow F$ and $\{\mu'_{m,2n}\}_{m=1}^{\infty}$ is bounded for every n then F has moments of all orders, $\{\mu'_n\}_{n=1}^{\infty}$, and $\mu'_{m,n} \to \mu'_n$ for all n.*

Proof. *Once the existence of moments for F is demonstrated, the above proof applies. To this end, let $\{X_m\}_{m=1}^{\infty}$ and X be defined as above. Then since $X_m \to X$ for all $t \in (0,1)$, Fatou's lemma of proposition 2.46 of book 3 obtains:*

$$E\left[|X|^{2n}\right] \leq \liminf E\left[|X_m|^{2n}\right] = \liminf \left[\mu'_{m,2n}\right] < \infty.$$

For odd moments recall example 3.49:

$$E\left[|X_m|^{2n+1}\right] \leq \left(E\left[|X_m|^{2n}\right] E\left[|X_m|^{2n+2}\right]\right)^{1/2}.$$

■

Proposition 3.69 *Let $\{F_m\}_{m=1}^{\infty}$ and F be distribution functions with with respective moment generating functions $\{M_m(t)\}_{m=1}^{\infty}$ and $M(t)$ convergent on a common interval $(-t_0,t_0)$ with $t_0 > 0$. If $F_m \Rightarrow F$ and $\{M_m(t)\}_{m=1}^{\infty}$ is bounded on this interval for each t, then $M_m(t) \to M(t)$ for $t \in (-t_0,t_0)$.*
Proof. *Using the notation of the prior proof:*

$$\int_{(0,1)} \exp\left[tX_m^{(N)}(s)\right] dm_L(s) \to \int_{(0,1)} \exp\left[tX^{(N)}(s)\right] dm_L(s)$$

by the bounded convergence theorem, and hence:

$$\begin{aligned} &\limsup_m \left| \int_{(0,1)} \exp\left[tX_m(s)\right] dm_L(s) - \int_{(0,1)} \exp\left[tX(s)\right] dm_L(s) \right| \\ \leq &\sup_m \int_{|X_m(s)|\geq N} \exp\left[tX_m(s)\right] dm_L(s) + \int_{|X(s)|\geq N} \exp\left[tX(s)\right] dm_L(s). \end{aligned}$$

The second integral converges to 0 for all $t \in (-t_0,t_0)$ by the existence of $M(t)$. To complete the proof we show that the first term will also have limit 0 as $N \to \infty$ by proving that $\{\exp\left[tX_m(s)\right]\}_{m=1}^{\infty}$ are uniformly integrable for $t \in (-t_0,t_0)$.

To this end we use exercise 3.70 below which states that uniform integrability will follow if for some $\epsilon > 0$,

$$\sup_m \int_{(0,1)} \left(\exp\left[tX_m(s)\right]\right)^{1+\epsilon} dm_L(s) = K < \infty.$$

Note however, that

$$\int_{(0,1)} \left(\exp\left[tX_m(s)\right]\right)^{1+\epsilon} dm_L(t) = M_m(t\left[1+\epsilon\right]),$$

and hence is bounded by assumption when $t\left[1+\epsilon\right] \in (-t_0,t_0)$. This completes the proof since for any $t \in (-t_0,t_0)$ there exists $\epsilon > 0$ for which $t\left[1+\epsilon\right] \in (-t_0,t_0)$. ■

Exercise 3.70 *Prove that if for some $\epsilon > 0$:*

$$\sup_m \int_{(0,1)} |Y_m(t)|^{n+\epsilon} dm_L(t) = K < \infty,$$

then $\{[Y_m(t)]^n\}$ are uniformly integrable. Hint: Consider moment inequalities.

Remark 3.71 *Once again, the existence of $M(t)$ could have been part of the conclusion.*

Corollary 3.72 *Let $\{F_m\}_{m=1}^{\infty}$ be distribution functions with moment generating functions $\{M_m(t)\}_{m=1}^{\infty}$ convergent on a common interval $(-t_0,t_0)$ with $t_0 > 0$. If $F_m \Rightarrow F$ and $\{M_m(t)\}_{m=1}^{\infty}$ is bounded on this interval for each t, then F has a moment generating function $M(t)$ with $M_m(t) \to M(t)$ for $t \in (-t_0,t_0)$.*
Proof. *As for corollary 3.68, the existence of $M(t)$ for each $t \in (-t_0,t_0)$ follows from Fatou's lemma.* ■

We are now ready for the main results on method of moments. The corollary below will provide the result in the form most commonly used in applications, and that is in the context of the associated moment generating functions.

Proposition 3.73 (Method of Moments) *Assume that a distribution function F is uniquely determined by its moment collection, $\{\mu'_n\}_{n=1}^{\infty}$, for example by proposition 3.57. If $\{F_m\}_{m=1}^{\infty}$ is a sequence of distribution functions with moment collections $\{\{\mu'_{m,n}\}_{n=1}^{\infty}\}_{m=1}^{\infty}$ and $\mu'_{m,n} \to \mu'_n$ for each n, then $F_m \Rightarrow F$.*
Proof. *First note that since $\mu'_{m,2} \to \mu'_2$, the collection $\{\mu'_{m,2}\}_{m=1}^{\infty}$ is bounded, say by K. If X_m is a random variable with distribution F_m, constructed for example using Skorokhod's representation theorem of proposition 8.3 of book 2, then by Chebyshev's inequality in 3.71:*

$$\Pr[|X_m| \geq t] \leq K/t^2,$$

for all m. This implies by definition 3.61 that the collection of distribution functions, $\{F_m\}_{m=1}^{\infty}$, is tight.

By Helly's selection theorem of proposition 8.14 of book 2, there exists a subsequence $\{F_{m_k}\}_{k=1}^{\infty}$, and a right continuous, increasing function, $\widetilde{F}$, so that $F_{m_k}(x) \to \widetilde{F}(x)$ at all continuity points of $\widetilde{F}$. By proposition 8.20 of book 2, $\widetilde{F}$ is a distribution function because $\{F_m\}_{m=1}^{\infty}$ is tight and so $F_{m_k} \Rightarrow \widetilde{F}$. Now since $\mu'_{m,n} \to \mu'_n$ for each n it follows that for this subsequence $\mu'_{m_k,n} \to \mu'_n$ for each n. If it could be proved that μ'_n is the nth moment of $\widetilde{F}$, then by the assumption that F is uniquely determined by its moment collection we could conclude that $\widetilde{F} = F$. Then by corollary 8.22 to Helly's selection theorem, since every such Helly subsequence satisfies $F_{m_k} \Rightarrow F$, it follows that $F_m \Rightarrow F$.

To prove that μ'_n is the nth moment of $\widetilde{F}$, note that the convergence assumption $\mu'_{m,n} \to \mu'_n$ for each n assures that $\{\mu'_{m_k,2n}\}_{m=1}^{\infty}$ is bounded for all n. Thus from corollary 3.68, $F_{m_k} \Rightarrow \widetilde{F}$ obtains that $\widetilde{F}$ has moments $\{\widetilde{\mu}'_n\}_{n=1}^{\infty}$, and $\mu'_{m_k,n} \to \widetilde{\mu}'_n$ for each n. But since $\mu'_{m,n} \to \mu'_n$ for each n we conclude that $\widetilde{\mu}'_n = \mu'_n$. ■

Corollary 3.74 (Method of Moments) *Assume that a distribution function F has moment generating function $M(t)$ which converges on $(-t_0,t_0)$ with $t_0 > 0$. Assume also $\{F_m\}_{m=1}^{\infty}$ is a sequence of distribution functions with respective moment generating functions $\{M_m(t)\}_{m=1}^{\infty}$, convergent on the same interval. If $M_m(t) \to M(t)$ for $t \in (-t_0,t_0)$, then $F_m \Rightarrow F$.*

Proof. *The existence of $M(t)$ on $(-t_0,t_0)$ assures that F is uniquely determined by its moments by proposition 3.57, and also that its moments are derived by $\mu'_n = M^{(n)}(0)$ by proposition 3.24. If X_m is a random variable underlying F_m, again using Skorokhod's representation theorem, then by 3.74, for any $t' \in (-t_0,t_0)$:*

$$\Pr[|X_m| \geq t] \leq e^{-tt'} \left[M_m(t') + M_m(-t')\right] \leq c(t')e^{-tt'},$$

since $M_m(\pm t') \to M(\pm t')$ for all t'. Letting $t \to \infty$ obtains that the collection of distribution functions, $\{F_m\}_{m=1}^{\infty}$, is tight. Hence as in the above proof there is a subsequence, $\{F_{m_k}\}_{k=1}^{\infty}$ and a distribution function $\widetilde{F}$ so that $F_{m_k} \Rightarrow \widetilde{F}$.

The proof that $\widetilde{M}(t)$ exists and $M_{m_k}(t) \to \widetilde{M}(t)$ for $t \in (-t_0,t_0)$, and thus that $\widetilde{M}(t) = M(t)$ follows from corollary 3.72 exactly as for the proof above. Corollary 3.58 on the uniqueness of the moment generating function now assures that $\widetilde{F} = F$ and thus $F_{m_k} \Rightarrow F$. As this conclusion is true for any Helly subsequence we conclude by corollary 8.22 to Helly's selection theorem that $F_m \Rightarrow F$. ■

Remark 3.75 *Many important results will be derived based on corollary 3.74 in the section below on Limit Theorems.*

Chapter 4

Simulating Samples of RVs - Examples

Recall proposition 4.9 of book 2 which proved the following result:

Let $(\mathcal{S},\mathcal{E},\lambda)$ *be given, and* $X : (\mathcal{S},\mathcal{E},\lambda) \to (\mathbb{R},\mathcal{B}(\mathbb{R}),m)$ *a random variable with distribution function* $F(x)$ *and left continuous inverse* $F^*(y)$ *given in 1.19 by:*

$$F^*(y) = \inf\{x|F(x) \geq y\}.$$

If $\{Y_j\}_{j=1}^M$ *are independent, continuous uniformly distributed random variables, then* $\{X_j\}_{j=1}^M \equiv \{F^*(Y_j)\}_{j=1}^M$ *are independent random variables with distribution function* $F(x)$. If F is continuous and $\{X_j\}_{j=1}^M$ *are independent random variables with distribution function* $F(x)$, then $\{Y_j\}_{j=1}^M \equiv \{F(X_j)\}_{j=1}^M$ *are independent, continuous uniformly distributed random variables.*

In the next section we apply this result to develop random samples for a number of the distributions introduced above. Following that, we investigate the generation of **ordered samples.**

We assume throughout these sections that one has access to computer software for generating collections $\{Y_j\}_{j=1}^M$, of independent, continuous uniformly distributed random variables. Fortunately, generating samples of Y is now very easy as most mathematical software has a built-in function which does exactly this. For example, in Microsoft Excel this function is called RAND(), while in MathWorks® MATLAB it is rand, and so forth.

See book 2 for the theoretical framework for random samples, or the introduction to chapter 5 on Limit Theorems for a review of the book 2 results.

4.1 Random Samples

4.1.1 Discrete Distributions

Here of necessity, random variates are calculated directly from the definition of $F^*(y)$. Recall that every distribution function F is right continuous, and in almost all applications a discrete distribution function will be be defined relative to:

1. Finite collection: $\{[a_i, b_i)\}_{i=1}^{N}$, or,

2. One-sided infinite collection:$\{[a_i, b_i)\}_{i=1}^{\infty}$ or $\{[a_i, b_i)\}_{i=-1}^{-\infty}$, or,

3. Doubly infinite collection: $\{[a_i, b_i)\}_{i=-\infty}^{\infty}$,

where in all cases if $a_i = -\infty$ then $[a_i, b_i) \equiv (-\infty, b_i)$. Such intervals partition $\mathbb{R}$ into left semi-closed intervals, $\{[a_i, b_i)\}$ so that $F([a_i, b_i)) = c_i$, and in the most frequently encountered cases:

- Either for some i, $[a_i, b_i) \equiv (-\infty, b_i)$ with $c_i = 0$, or, all $a_i > -\infty$ and $c_i > 0$ for all i;
- Either for some i, $[a_i, b_i) \equiv [a_i, \infty)$ with $c_i = 1$, or all $b_i < \infty$ and $c_i < 1$ for all i.

It is also possible for a discrete distribution function to have no such parametrization, for example by defining:

$$F(x) = \sum\nolimits_{r_n \leq x} 2^{-n},$$

where $\{r_n\}_{n=1}^{\infty} \subset \mathbb{R}$ is an arbitrary ordering of the rationals. But this type of saltus function is rarely encountered in applications, so will be ignored.

1. **Discrete Rectangular Distribution on** $[0, 1]$, parameter n, defined in 1.4:

   Given independent continuous uniform $\{Y_j\}_{j=1}^{M}$, define:

   $$\{X_j^R\}_{j=1}^{M} = \{(\lfloor nY_j \rfloor + 1)/n\}_{j=1}^{M},$$

where $\lfloor x \rfloor$ denotes the **greatest integer function**, also called the **floor function,** defined by:

$$\lfloor x \rfloor \equiv \max\{m \in \mathbb{N} | m \leq x\}.$$

2. **Standard Binomial Distribution**, parameter p, defined in 1.5:

   Given independent continuous uniform $\{Y_j\}_{j=1}^M$, define:

$$\{X_{1,j}^B\}_{j=1}^M = \left\{\chi_{[0,p]}(Y_j)\right\}_{j=1}^M,$$

   where $\chi_{[0,p]}(y)$ is the characteristic function of $[0,p]$, defined to be 1 if $y \in [0,p]$ and 0 otherwise.

3. **General Binomial Distribution**, parameters p, n, defined in 1.6:

   First, $X_{n,j}^B$ can be defined directly in terms of characteristic functions of the intervals defined by $\{F_B(j)\}_{j=0}^n$. To this end, define:

$$a_k \equiv F(k) = \sum\nolimits_{j=0}^k \binom{n}{j} p^j (1-p)^{n-j}, \qquad k = 0, 1, ..., n,$$

   and let $\chi_k(x)$ denote the characteristic function of $[a_{k-1}, a_k)$ where for notational convenience let $a_{-1} \equiv 0$. Then given $\{Y_j\}_{j=1}^M$ define:

$$\{X_{n,j}^B\}_{j=1}^M = \left\{\sum\nolimits_{k=0}^n k\chi_k(Y_j)\right\}_{j=1}^M$$

   Alternatively, using the fact that $X_{n,j}^B$ is the sum of n independent standard binomials, then given independent continuous uniform $\{Y_j\}_{j=1}^{nM}$, define:

$$\{X_{n,j}^B\}_{j=1}^M = \left\{\sum\nolimits_{k=(j-1)n+1}^{jn} \chi_{[0,p]}(Y_k)\right\}_{j=1}^M.$$

4. **Geometric Distribution,** parameter p, defined in 1.9:

   Here we define X_j^G directly in terms of characteristic functions of the intervals defined by $\{F_G(j)\}_{j=0}^\infty$ as noted in 3, but this calculation is simplified by the fact that $F_G^*(y)$ can be explicitly calculated. With $F_G(j)$ defined in 1.10 by $F_G(j) = 1-(1-p)^{j+1}$, a calculation produces:

$$F_G^*(y) = \inf\{j | \ln(1-y)/\ln(1-p) \leq j+1\}.$$

Since the goal is to have $F_G^*[(0,p]] = F_G^*[(0,F_G(0)]] = 0$ and otherwise for $n \geq 0$:

$$F_G^*((F_G(n), F_G(n+1)]) = n+1,$$

we can define F_G^* in terms of the **ceiling function** $\lceil y \rceil$**:**

$$F_G^*(y) = \lceil \ln(1-y)/\ln(1-p) \rceil - 1,$$

where

$$\lceil y \rceil = \min\{m \in \mathbb{N} | m \geq y\}.$$

Hence given independent continuous uniform $\{Y_j\}_{j=1}^M$, define:

$$\begin{aligned} \{X_j^G\}_{j=1}^M &= \{\lceil \ln(1-Y_j)/\ln(1-p) \rceil - 1\}_{j=1}^M \\ &= \{\lceil \ln(Y_j)/\ln(1-p) \rceil - 1\}_{j=1}^M \end{aligned}$$

since Y_j has the same distribution as $1 - Y_j$.

5. **Negative Binomial Distribution,** parameters p, k, defined in 1.12:

   In addition to defining X_j^{NB} directly in terms of characteristic functions of the intervals defined by $\{F_{NB}(j)\}_{j=0}^\infty$, recall example 3.59 which proved that the negative binomial is the sum of k independent geometric random variables. Hence given independent continuous uniform $\{Y_j\}_{j=1}^{kM}$ we can define $\{X_j^{NB}\}_{j=1}^M$ in terms of k-sums of geometric variables as in 3 for the general binomial $X_{n,j}^B$:

$$\{X_j^{NB}\}_{j=1}^M = \left\{ \sum_{k=(j-1)k+1}^{jk} (\lceil \ln Y_j/\ln(1-p) \rceil - 1) \right\}_{j=1}^M.$$

6. **Poisson Distribution,** parameter λ, defined in 1.13:

   In addition to defining X_j^P directly in terms of characteristic functions of the intervals defined by $\{F_P(j)\}_{j=0}^\infty$, we can use the development in part 2 of example 1.18. It was there shown that with X_j^E denoting independent exponential random variables with parameter λ, and

$$N \equiv \max\left\{ n | \sum_{j=1}^n X_j^E \leq 1 \right\},$$

   then $N = X^P$ is a Poisson random variable with parameter λ. As noted below, from independent continuous uniform $\{Y_j\}_{j=1}^M$ such exponentials are generated by:

$$X_j^E = -\ln Y_j/\lambda.$$

Thus independent Poisson random variables can be produced by:

$$\begin{aligned} X^P &= \max\left\{n | \sum_{j=1}^{n} \ln Y_j \geq -\lambda\right\} \\ &= \max\left\{n | \prod_{j=1}^{n} Y_j \geq e^{-\lambda}\right\}. \end{aligned}$$

4.1.2 Continuous Distributions

For many continuous distributions, $F(x)$ is strictly increasing and one can algebraically determine F^* by inverting $F(x) = y$. In other words, $F^* = F^{-1}$ in such cases.

1. **Continuous Uniform Distribution on** $[a, b]$:

   Recalling 1.18, $F(x) = (x - a)/(b - a)$ for $a \leq x \leq b$ and hence given independent continuous uniform $\{Y_j\}_{j=1}^M$:

   $$\left\{X_j^U\right\}_{j=1}^M = \{a + (b - a)Y_j\}_{j=1}^M.$$

2. **Exponential Distribution,** parameter λ, in 1.20:

   Since $F_E(x) = 1 - e^{-\lambda x}$ it follows that $F_E^*(y) = -\ln(1 - y)/\lambda$, and so given independent continuous uniform $\{Y_j\}_{j=1}^M$:

   $$\begin{aligned} \left\{X_j^E\right\}_{j=1}^M &= \{-\ln(1 - Y_j)/\lambda\}_{j=1}^M \\ &= \{-\ln Y_j/\lambda\}_{j=1}^M \end{aligned}$$

   since Y_j has the same distribution as $1 - Y_j$.

3. **Gamma Distribution,** parameters λ, α, in 1.22:

   The Gamma distribution function is not explicitly invertible for general α, but for $\alpha = k \in \mathbb{N}$ a positive integer, exercise 1.19 proves that this Gamma random variable is the sum of k independent exponential random variables with parameter λ. Hence, given independent continuous uniform $\{Y_j\}_{j=1}^{kM}$, we can define $\{X_j^\Gamma\}_{j=1}^M$ in terms of k-sums of exponential variables:

   $$\{X_j^\Gamma\}_{j=1}^M = \left\{-\sum_{k=(j-1)k+1}^{jk} \ln Y_j/\lambda\right\}_{j=1}^M.$$

4. **Beta Distribution**, parameters, v, w, in 1.27:

   The Beta distribution is also not explicitly invertible, but an important connection between the Beta and Gamma distributions makes generating Beta samples feasible when v, w are positive integers. Recalling proposition 1.25, that if X^{Γ_1} and X^{Γ_2} are independent Gamma random variables with parameters v, λ and w, λ, then the random variable $\frac{X^{\Gamma_1}}{X^{\Gamma_1}+X^{\Gamma_2}}$ is Beta with parameters v, w, and this is independent of λ. Hence when $v, w \in \mathbb{N}$ are positive integers, two such Gammas can be generated as above with a total of $v + w$ independent continuous uniform random variables, taking $\lambda = 1$ for simplicity, and so $\{X_j^{\beta}\}_{j=1}^{M}$ can then be generated from $\{Y_j\}_{j=1}^{(v+w)M}$.

5. **Cauchy Distribution,** parameters x_0, γ, in 3.63:

   Because the substitution $(y - x_0)/\gamma = \tan z$ produces:

$$\begin{aligned} F_C(x) &= \frac{1}{\pi\gamma}\int_{-\infty}^{x} \frac{dy}{1 + ((y - x_0)/\gamma)^2} \\ &= \frac{1}{\pi}\int_{-\pi/2}^{\arctan((y-x_0)/\gamma)} dz \\ &= \frac{1}{\pi}\left[\arctan\left((y - x_0)/\gamma\right) + \pi/2\right], \end{aligned}$$

   we can invert $F_C(x)$. Hence, given independent continuous uniform $\{Y_j\}_{j=1}^{M}$, define $\{X_j^C\}_{j=1}^{M}$ by:

$$\{X_j^C\}_{j=1}^{M} = \{x_0 + \gamma \tan\left[\pi\left(Y_j - 1/2\right)\right]\}_{j=1}^{M}.$$

6. **Extreme Value Distributions,** parameter γ, in 6.15 and 6.16:

   Here $G_\gamma^{-1}(y)$ is readily determined when $\gamma \neq 0$ as:

$$G_\gamma^{-1}(y) = \left(\left[\ln\left(1/y\right)\right]^{-\gamma} - 1\right)/\gamma,$$

   while for $\gamma = 0$,

$$G_0^{-1}(y) = \ln\left[1/\ln\left(1/y\right)\right] = -\ln[-\ln y].$$

   Note that as always, the result for G_0^{-1} equals the result for $\lim_{\gamma\to 0} G_\gamma^{-1}$ since this limit is seen to be $f'(0)$ with $f(x) = \left[\ln\left(1/y\right)\right]^{-x}$. Hence, given independent continuous uniform $\{Y_j\}_{j=1}^{M}$, define $\{X_j^{EV}\}_{j=1}^{M}$ by:

$$\{X_j^{EV}\}_{j=1}^{M} = \left\{G_\gamma^{-1}(Y_j)\right\}_{j=1}^{M}.$$

7. **Normal** and **Lognormal Distributions,** parameters μ, σ^2, in 1.32 and 1.34:

   Since $LN = \exp[N]$, the ability to simulate one immediately provides the ability to simulate the other. Unfortunately, the most common approaches to simulation for N, say, is numerical estimation using two approaches:

   - Numerical estimates of $F_N(x)$ for discrete x values, then approximating $F_N^{-1}(y)$ with interpolation,
   - Applying the Central Limit theorem below, which states that for n large, if $\{Y_j\}_{j=1}^M$ are independent and identically distributed, then
     $$\left[\sum_{j=1}^n Y_j - nE[Y_1]\right] \Big/ \sqrt{nVar[Y_1]}$$
     is approximately standard normal. If Y_j is continuous uniform on $[0,1]$, then by 3.56, $E[Y_1] = 1/2$ and $Var[Y_1] = 1/12$. Choosing $n \geq 20$ or so, we conclude that given independent continuous uniform $\{Y_j\}_{j=1}^{Mn}$, we can define $\{X_j^\Phi\}_{j=1}^M$ by:

$$\{X_j^\Phi\}_{j=1}^M = \left\{ \left[\sum_{k=(j-1)n+1}^{jn} Y_j - n/2\right] \Big/ \sqrt{n/12} \right\}_{j=1}^M .$$

4.2 Ordered Random Samples

Because any random sample $\{X_j\}_{j=1}^M$ can be re-ordered into $\{X_{(k)}\}_{k=1}^M$ where $X_{(k)} \leq X_{(k+1)}$, an ordered random sample can always be created by first generating a random M-sample as in the prior section, then reordering. However, there are applications in which we are primarily interested in generating a collection such as $\{X_{(k)_j}\}_{j=1}^N$, a random sample of *kth* **order statistics**, meaning random variates which would each be the *kth* largest of a random M-sample. An example of this from finance is the estimation of so-called **value at risk** at a given percentile where we are interested in financial losses in what would be the **worst** one year event say, with 95% or 99% confidence. In this case we are interested in various M and then generating samples with $k \simeq 0.95M$ or $k \simeq 0.99M$, respectively. While simply generating N such M-samples, reordering, and choosing the appropriate variates is one feasible approach, it is apparent that this is not an efficient approach as it results in discarding most of the generated variates.

There are also situations in which we are interested in properties of the conditional distribution, when $X \geq X_{(k)}$, such as the average value of variates equal to or worse than the identified $X_{(k)}$ variate. For example, in the above finance application the average of such losses is called the **expected shortfall,** or **conditional tail loss.** In this case, we are interested in generating the collections: $\{X_{(k)_j}, X_{(k+1)_j}, ..., X_{(M)_j}\}_{j=1}^{N}$. Again, in theory we can generate full samples and select those of interest, though in practice this is not an attractive option.

4.2.1 Direct Approaches

Given a random variable X with distribution function F, assume that our goal is to generate $\{X_{(k)_j}\}_{j=1}^{N}$, a random sample of N kth **order statistics.** Thus each random variate represents the kth largest of a random M-sample. We then have several possible approaches to achieving this objective:

1. **Complete/Partial Ordered Samples:**

   Beginning with MN uniformly distributed continuous random variables $\{Y_i\}_{i=1}^{MN}$, generate N sets of M-samples of X variates as in the above section:

   $$\left\{\{F^*(Y_i)\}_{i=(j-1)M}^{jM}\right\}_{j=1}^{N},$$

   then reorder each M-sample of X variates, $\{F^*(Y_i)\}_{i=(j-1)M}^{jM}$, to determine $X_{(k)_j}$,

   $$\{F^*(Y_i)\}_{i=(j-1)M}^{jM} \rightarrow \{X_{(k)_j}\}_{k=1}^{M}, \qquad j = 1, ..., N,$$

   resulting in N complete samples of ordered statistics: $\{\{X_{(k)_j}\}_{k=1}^{M}\}_{j=1}^{N}$, from which specific variates or ranges of variates can be selected.

   Of course, it is not necessary to evaluate $F^*(Y_i)$ for every Y_i in the various M-samples unless all order statistics are required. For example, if we require only a sample of $X_{(k)}$ for a fixed k then we only need to determine $F^*\left(Y_{(k)_j}\right)$ where $Y_{(k)_j}$ is the kth order statistic for the jth M-sample of Y variates. This follows since F is increasing and thus the kth order statistic for X is produced by the kth order statistic for uniform variate Y. Even though $\{Y_i\}_{i=1}^{MN}$ are easily generated using a variety of software, the potential shortcoming of this approach is that most of the Y sample points will be discarded.

Summary: Requires MN uniformly distributed continuous random variables, $\{Y_i\}_{i=1}^{MN}$, and KN evaluations of $F^*\left(Y_{(k)_j}\right)$, where $1 \leq K \leq M$ is the number of order statistics desired in each sample, to produce an N-sample of the range of ordered statistics: $\{\{X_{(k_i)_j}\}_{i=1}^{K}\}_{j=1}^{N}$.

2. Direct *kth* Order Statistics 1:

Using example 2.5, recall that the kth order statistic from an M-sample of uniform variates Y has a Beta distribution with $v = k$ and $w = M-k+1$, and this appears promising for the generation of samples of just the one statistic, $\{Y_{(k)_j}\}_{j=1}^{N}$. But based on the prior section, to generate a Beta random variable with $v = k$ and $w = M-k+1$ requires $v + w = M + 1$ uniformly distributed continuous random variables, so in total we require $\{Y_i\}_{i=1}^{(M+1)N}$. This is more than that needed for the complete sample method above despite seeking only a sample of one variate. Each $(M + 1)$-sample of Y variates then produces one such Beta by first producing two independent Gamma random variables. Since the λ parameter in the Gammas is irrelevant for this purpose, we can choose $\lambda = 1$ for simplicity and then set α, the other Gamma parameter, to k and $M - k + 1$, respectively.

Summary: Requires $(M + 1)\, N$ uniformly distributed continuous random variables, $\{Y_i\}_{i=1}^{(M+1)N}$, with each $(M + 1)$-sequence used to produce a Beta as above, and N evaluations of $F^*\left(Y_{(k)_j}\right)$, to produce an N-sample of the kth ordered statistics: $\{X_{(k)_j}\}_{j=1}^{N}$.

3. Direct *kth* Order Statistics 2:

Since $F_{(k)}$, the distribution function for $X_{(k)}$ in 2.1 is known, if it is possible to determine its left continuous inverse $F^*_{(k)}$, then by definition we only need one N-sample $\{Y_j\}_{j=1}^{N}$ which then obtains a sample of $X_{(k)}$:

$$\{X_{(k)_j}\}_{j=1}^{N} = \{F^*_{(k)}(Y_j)\}_{j=1}^{N}.$$

However, even the simplest such distribution function, the uniform distribution F, produced $F_{(k)}$ with the complexity of the Beta distribution, it is unlikely that many examples will be encountered for which one can calculate $F^*_{(k)}$ directly. But in general, to determine $F^*_{(k)}(Y)$ using 2.1 requires two steps, the first of which is almost surely numerical:

(**a**) Solve for Z:

$$Y = \sum_{j=k}^{M} \binom{M}{j} Z^j (1-Z)^{M-j},$$

(**b**) Solve for X :

$$F(X) = Z.$$

Summary: Requires N uniformly distributed continuous random variables $\{Y_i\}_{i=1}^N$, and N evaluations or estimations of $\{Z_i\}_{i=1}^N$ from step a, then N estimates of $\{X_i\}_{i=1}^N$ from part b to produce an N-sample of the kth ordered statistics: $\{X_{(k)_j}\}_{j=1}^N \equiv \{X_i\}_{i=1}^N$.

4.2.2 Using the Rényi Representation

Given a random variable X with distribution function F, the goal of this section to again generate $\{X_{(k)_j}\}_{j=1}^N$, a random sample of kth **order statistics** from random M-samples. While several approaches to achieving this objective were discussed above, in this section we provide a fourth approach based on the **Rényi representation theorem**. While this representation theorem generates ordered **exponential** random variables, it is then a simple matter to convert these to ordered continuous uniform random variables, which then provide for the ordered random variables of interest. This will involve the following result, which is an application of results from book 2.

Proposition 4.1 *For any n, if the collection $\left\{Y_j^U\right\}_{j=1}^n$ are independent, continuous uniform random variables then $\left\{X_j^E\right\}_{j=1}^n \equiv \left\{-\ln\left(1-Y_j^U\right)\right\}_{j=1}^n$ are independent exponential random variables with parameter $\lambda = 1$. Conversely, given independent standard exponentials $\left\{X_j^E\right\}_{j=1}^n$, then $\left\{Y_j^U\right\}_{j=1}^n \equiv \left\{1-\exp\left(-X_j^E\right)\right\}_{j=1}^n$ are independent, continuous uniform random variables.*

Proof. *By proposition 3.22 of book 2, if a distribution function F is continuous and strictly increasing then $F^* = F^{-1}$. Here $F(x) = 1 = e^{-x}$ and so $F^{-1}(y) = -\ln(1-y)$ and this proposition can be stated:*

1. *If $\left\{Y_j^U\right\}_{j=1}^n$ are independent continuous uniform random variables, then $\left\{F^*\left(Y_j^U\right)\right\}_{j=1}^n$ are independent with distribution function F,*

2. *If* $\{X_j\}_{j=1}^n$ *are independent random variables with distribution function* F*, then* $\{F(X_j)\}_{j=1}^n$ *are independent continuous uniform random variables.*

Both distributional results are from proposition 4.5 of book 2, while the independence results are from that book's proposition 4.9 since F *is continuous.* ■

Remark 4.2 *Note that because* Y^U *and* $1-Y^U$ *have the same distribution, given* $\left\{Y_j^U\right\}_{j=1}^n$ *we can in proposition 4.1 optionally generate independent standard exponentials using* $\left\{Y_j^E\right\}_{j=1}^n \equiv \left\{-\ln\left(Y_j^U\right)\right\}_{j=1}^n$*. By the same argument, given* $\left\{Y_j^E\right\}_{j=1}^n$*, we can optionally generate independent uniform variables using* $\left\{Y_j^U\right\}_{j=1}^n \equiv \left\{\exp\left(-Y_j^E\right)\right\}_{j=1}^n$*. But in both cases, the order statistics reverse.*

For example, once we generate exponential kth order statistics using the Rényi representation theorem, $\left\{Y_{(k)}^E\right\}_{k=1}^M$*, then* $\left\{Y_{(k)}^U\right\}_{k=1}^M \equiv \left\{1-\exp\left(-Y_{(k)}^E\right)\right\}_{k=1}^M$ *will be kth order statistics of the uniform distribution, and equivalently by the above remark,* $\left\{Y_{(M-k+1)}^U\right\}_{k=1}^M \equiv \left\{\exp\left(-Y_{(k)}^E\right)\right\}_{k=1}^M$ *will be uniformly distributed kth order statistics, in the reverse order.*

The practical significance of this remark can be seen by returning to the goal of generating $\{X_{(k)_j}\}_{j=1}^N$, a random sample of N *kth* order exponential statistics**.** Adding to the 3 approaches in the above section, we have the following.

4. **Rényi Representation for** *kth* **Order Exponential Statistics**

   By the first corollary to the Rényi representation theorem, each exponential *kth* order statistics, $Y_{(k)}^E$ requires the generation of k standard exponentials and an application of 2.20.

   (**a**) When k is small relative to M, we implement the following N times:

   i. Generate $\{E_j\}_{j=1}^k$ independent standard exponential random variables, $\lambda = 1$, by defining $E_j = -\ln(1-Y_j)$ where $\{Y_j\}_{j=1}^k$ are independent, continuous uniform variates.

   ii. Define $Y_{(k)}^E$ as in 2.20 with $\lambda = 1$.

iii. Define $Y^U_{(k)} = 1 - \exp\left(-Y^E_{(k)}\right)$, producing a kth order statistic of the continuous uniform distribution.

iv. Evaluate $X_{(k)} = F^*\left(Y^U_{(k)}\right)$.

(**b**) When k is large relative to M, we implement the following N times:

i. Generate $\{E_j\}_{j=1}^{M-k+1}$ independent standard exponential random variables, $\lambda = 1$, by defining $E_j = -\ln(1 - Y_j)$ where $\{Y_j\}_{j=1}^{M-k+1}$ are independent continuous uniform variates.

ii. Define $Y^E_{(M-k+1)}$ as in 2.20 with $\lambda = 1$.

iii. Define $Y^U_{(k)} = \exp\left(-Y^E_{(M-k+1)}\right)$, producing a uniformly distributed kth order statistic.

iv. Evaluate $X_{(k)} = F^*\left(Y^U_{(k)}\right)$.

Summary: Requires $N \times \min\{k, M-k+1\}$ uniformly distributed continuous random variables, $\{Y_i\}_{i=1}^{N\min\{k,M-k+1\}}$, and N evaluations of $F^*\left(Y^U_{(k)}\right)$ for an N-sample of kth order statistics.

Chapter 5

Limit Theorems

In this section we study limit theorems grouped by type into three categories:

- Weak Convergence of Distributions
- Laws of Large Numbers
- Convergence of Empirical Distribution Functions

This study will be supplemented in book 6 with more general results related to weak convergence of measures and the central limit theorem.

5.1 Introduction

Throughout this chapter we will repeatedly encounter the notion of a collection of independent random variables, often identically distributed but sometimes not, and often we will want to address questions related to the summation of such variables, which is again declared to be a random variable. This all appears intuitively plausible and yet requires some justification to avoid building the house of emerging theories on a foundation of sand. Among the questions that need to be addressed are:

Q1. If $X : (\mathcal{S}, \mathcal{E}, \lambda) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}), m)$ is a random variable with distribution function $F(x)$, and $\{X_j\}_{j=1}^n$ is a sample of X, meaning independent and identically distributed random variables defined on some probability space, how is this sample constructed and on what probability space is this sample defined? If $\{X_j\}_{j=1}^n$ is a sample of X defined on

the space identified, then an expression like $\sum_{j=1}^{n} X_j$ is again a random variable on this same space. Thus the measure on this space provides meaning to probability statements on such expressions, and underlies the definition of the associated distribution functions.

Q2. If $\{X_j\}_{j=1}^{n}$ is a sample of X defined on the space identified in 1, what does it mean to let $n \to \infty$? Is this increasing collection still defined on the same space so that probability statements regarding this sum, or properties of its distribution function, are all defined relative to a given probability measure?

Q3. More generally, if $X_j : (\mathcal{S}_j, \mathcal{E}_j, \lambda_j) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}), m)$ are random variables with distribution functions $F_j(x)$, is there a common probability space on which all can be defined, and on which these will be independent random variables? If so, then an expression like $\sum_{j=1}^{n} X_j$ is well defined, but in this more general context, Q2 again needs consideration.

In chapter 4 of book 2 the construction sought in Q1 was developed using the infinite dimensional probability space theory of chapter 9 of book 1. While this construction targeted the application to Q1, the book 1 theory applied equally well to the more general construction needed for Q3 and so we summarize the book 2 construction in this more general context. Similarly, while the book 2 construction allowed the construction of a sample space for either finite of infinite samples, we focus on the infinite dimensional model to simultaneously address Q2.

To this end assume the general case notationally and that we are given $\{X_j\}_{j=1}^{\infty}$ defined on $\{(\mathcal{S}_j, \mathcal{E}_j, \lambda_j)\}_{j=1}^{\infty}$ with distribution functions $\{F_j\}_{j=1}^{\infty}$. In the Q1 application we are simply starting with an infinite collection of copies of X and $(\mathcal{S}, \mathcal{E}, \lambda)$, and the indexing just allows a basis for referring to members of these collections. Then by chapter 3 of book 1, each F_j is increasing and right continuous, and by that book's chapter 5 each gives rise to a Borel measure μ_{F_j} defined on $\mathcal{B}(\mathbb{R})$ for which

$$\mu_{F_j}((a,b]) = F_j(b) - F_j(a) \equiv \lambda_j\left(X_j^{-1}((a,b])\right),$$

for all right semi-closed intervals. It then follows by extension that for all $A \in \mathcal{B}(\mathbb{R})$:

$$\mu_{F_j}(A) = \lambda_j\left(X_j^{-1}(A)\right). \tag{5.1}$$

Thus $\mu_{F_j}(\mathbb{R}) = 1$ and $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu_{F_j})$ is a probability space for all j. There is also a complete version of this probability space given in chapter 5 of book

1, $(\mathbb{R}, \mathcal{M}_{\mu_{F_j}}(\mathbb{R}), \mu_{F_j})$ where $\mathcal{B}(\mathbb{R}) \subset \mathcal{M}_{\mu_{F_j}}(\mathbb{R})$, but for the current purpose the Borel space is adequate.

Applying chapter 9 of book 1 to $\{(\mathbb{R}_j, \mathcal{B}(\mathbb{R}_j), \mu_{F_j})\}_{j=1}^{\infty}$, where $\mathbb{R}_j$ and $\mathcal{B}(\mathbb{R}_j)$ are indexed only for notational purposes, we can construct the infinite dimensional probability space, $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}(\mathbb{R}^{\mathbb{N}}), \mu_{\mathbb{N}})$ and complete counterpart $(\mathbb{R}^{\mathbb{N}}, \sigma(\mathbb{R}^{\mathbb{N}}), \mu_{\mathbb{N}})$. For this probability space,

$$\mathbb{R}^{\mathbb{N}} \equiv \{(x_1, x_2, ...) | x_j \in \mathbb{R}_j\},$$

and $\mu_{\mathbb{N}}$ is uniquely defined on the smallest sigma algebra containing the algebra $\mathcal{A}^+$ of **general finite dimensional measurable rectangles** or **general cylinder sets in** $\mathbb{R}^{\mathbb{N}}$. Specifically $H \in \mathcal{A}^+$ if for some positive integer n and n-tuple of positive integers $J = (j(1), j(2), ..., j(n))$:

$$H = \{x \in \mathbb{R}^{\mathbb{N}} | (x_{j(1)}, x_{j(2)}, ...x_{j(n)}) \in A\},$$

where $A \in \mathcal{B}(\mathbb{R}^n)$, the Borel sigma algebra on $\mathbb{R}^n$. Further, $\mu_{\mathbb{N}}$ is defined on $\mathcal{A}^+$ by:

$$\mu_{\mathbb{N}}(H) = \mu_F^{(n)}(A),$$

where $\mu_F^{(n)}$ is the product measure on $\mathbb{R}^n$ induced by $\{\mu_{F_{j(k)}}\}_{k=1}^n$. In the specific case where $A = \prod_{k=1}^n A_{j(k)}$ for $A_{j(k)} \in \mathcal{B}(\mathbb{R}_{j(k)})$, then:

$$\mu_{\mathbb{N}}(H) = \prod_{k=1}^n \mu_{F_{j(k)}}(A_{j(k)}). \tag{5.2}$$

Now define $X_j' : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ by

$$X_j' : (x_1, x_2, ...) = x_j, \tag{5.3}$$

which is the **projection mapping defined** on $\mathbb{R}^{\mathbb{N}}$ to the jth coordinate, and often denoted π_j.

The following result is a modest generalization of proposition 4.4 of book 2, and answers the above questions.

Proposition 5.1 *Let the probability spaces* $\{(\mathcal{S}_j, \mathcal{E}_j, \lambda_j)\}_{j=1}^{\infty}$ *and random variables* $X_j : \mathcal{S}_j \longrightarrow \mathbb{R}$ *with distribution functions* $\{F_j\}_{j=1}^{\infty}$ *be given, and with the notation above, let* $(\mathcal{S}', \mathcal{E}', \mu')$ *denote* $(\mathbb{R}^{\mathbb{N}}, \sigma(\mathbb{R}^{\mathbb{N}}), \mu_{\mathbb{N}})$ *or* $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}(\mathbb{R}^{\mathbb{N}}), \mu_{\mathbb{N}})$. *Then* $\{X_j'\}_{j=1}^{\infty}$ *as defined in 5.3 is a sample of* $\{X_j\}_{j=1}^{\infty}$ *as defined in definition 4.1 of book 2, meaning these are independent random variables defined on* $(\mathcal{S}', \mathcal{E}', \mu')$ *with respective distribution functions* $\{F_j\}_{j=1}^{\infty}$.

Proof. *First, X'_j is measurable and thus a random variable on $\mathbb{R}^{\mathbb{N}}$ since for $A \in \mathcal{B}(\mathbb{R})$, 5.3 yields that $\left(X'_j\right)^{-1}(A) = \{x \in \mathbb{R}^{\mathbb{N}} | x_j \in A\} \in \mathcal{A}^+$, a general cylinder set and thus an element of $\sigma(\mathbb{R}^{\mathbb{N}})$. Further, from 5.2 with $H \equiv \left(X'_j\right)^{-1}(A)$ and 5.1 obtains:*

$$\mu'\left[\left(X'_j\right)^{-1}(A)\right] = \mu_{F_j}(A) \equiv \lambda_j(X_j^{-1}(A)),$$

and thus X'_j and X_j have the same distribution.

Next, let $J = (j(1), j(2), ..., j(n))$ and $A_{j(k)} \in \mathcal{B}(\mathbb{R}_{j(k)})$ be given. Note that by 5.3:

$$\bigcap_{k=1}^{n} \left(X'_{j(k)}\right)^{-1}(A_{j(k)}) = \{x \in \mathbb{R}^{\mathbb{N}} | (x_{j(1)}, x_{j(2)}, ...x_{j(n)}) \in \prod_{k=1}^{n} A_{j(k)}\}.$$

Thus by 5.2 and 5.1:

$$\mu'\left(\bigcap_{k=1}^{n} \left(X'_{j(k)}\right)^{-1}(A_{j(k)})\right) = \prod_{k=1}^{n} \mu_{F_{j(k)}}(A_{j(k)}) \equiv \prod_{k=1}^{n} \mu'\left(\left(X'_{j(k)}\right)^{-1}(A_{j(k)})\right),$$

and thus $\{X'_{j(k)}\}_{k=1}^{n}$ are independent random variables. ■

Notation 5.2 *It is obviously a notational burden to distinguish between $\{X_j\}_{j=1}^{\infty}$ defined on $\{(\mathcal{S}_j, \mathcal{E}_j, \lambda_j)\}_{j=1}^{\infty}$ and independent $\{X'_j\}_{j=1}^{\infty}$ defined on $(\mathcal{S}', \mathcal{E}', \mu')$. So this formality is often suppressed and statements are made about independent, and possibly identically distributed random variables, $\{X_j\}_{j=1}^{\infty}$ defined on some probability space $(\mathcal{S}, \mathcal{E}, \mu)$.*

With the construction of proposition 5.1, the above questions can be answered. If $\{X_j\}_{j=1}^{n}$ are random variables defined on the new space $(\mathcal{S}, \mathcal{E}, \mu)$, then $\sum_{j=1}^{n} X_j$ is a measurable function on $(\mathcal{S}, \mathcal{E}, \mu)$ and thus a random variable by proposition 3.30 of book 1. Thus since all probability statements on such $\sum_{j=1}^{n} X_j$ are made in the same probability space $(\mathcal{S}, \mathcal{E}, \mu)$, these statements remain well defined as $n \to \infty$.

Remark 5.3 *The above framework will not be formally mentioned again in the next two sections on weak convergence of distributions and laws of large numbers. In the final section on convergence of empirical distributions functions, some additional clarifying comments will be required.*

5.2 Weak Convergence of Distributions

In this section, we will develop several important results relating to the weak convergence of certain sequences of distribution functions. The first, the **Poisson Limit theorem**, was stated and proved in proposition 1.11 of book 2 using explicit calculations and the given probability density functions. Here we will prove this result using corollary 3.74 by showing convergence of the associated moment generating functions. This same approach will also be applied to a generalization of the Poisson result, the so-called **weak law of small numbers,** as well as to the **De Moivre-Laplace theorem,** and to a special case of the **Central Limit theorem.** Using other tools we then turn to several results on order statistics.

5.2.1 Poisson Limit Theorem

In book 2 the following result, named for **Siméon-Denis Poisson** (1781 – 1840), was proved using explicit calculations and limiting arguments. Here we implement the method of moments.

Notation 5.4 *Note that the variate $B_{n,\lambda/n}$ in 5.4 can also be expressed as a summation, $S_n \equiv \sum_{m=1}^{n} X_{1,m}^B$ where for each n, $\{X_{1,m}^B\}_{m=1}^n$ are independent standard binomial variables with $p = \lambda/n$, and so $\mu_{X_1} = \lambda/n$ and $\sigma_{X_1}^2 = \lambda\left(1-\lambda/n\right)/n$. Thus $\mu_{S_n} = \lambda$ and $\sigma_{S_n}^2 = \lambda\left(1-\lambda/n\right)$.*

Proposition 5.5 (Poisson Limit theorem) *Let $F_{B_{n,p}}(j)$ denote the distribution function of the general binomial with parameters n, p associated with the density function in 1.6, and $F_{P_\lambda}(j)$ denote the distribution function of the Poisson with parameter λ associated with the density function in 1.13. For $\lambda = np$ fixed, then as $n \to \infty$:*

$$F_{B_{n,\lambda/n}} \Rightarrow F_{P_\lambda}. \tag{5.4}$$

Proof. *Let $M_{B_{n,\lambda/n}}(t)$ and $M_{P_\lambda}(t)$ denote the respective moment generating functions. It follows from 3.49 that for all t :*

$$M_{B_{n,\lambda/n}}(t) = \left(1 + \frac{\lambda}{n}(e^t - 1)\right)^n,$$

while from 3.54, again for all t:

$$M_{P_\lambda}(t) = \exp[\lambda(e^t - 1)].$$

Now for all $s \in \mathbb{R}$:

$$(1 - s/n)^n \longrightarrow e^{-s} \text{ as } n \to \infty,$$

because:

$$n \ln(1 - s/n) \longrightarrow -s \text{ as } n \to \infty,$$

which in turn follows from the definition of derivative of $f(x) = \ln(1 - sx)$ *at* $x = 0$, *using* $\Delta x = 1/n$.

Hence, $M_{B_{n,\lambda/n}}(t) \to M_{P_\lambda}(t)$ *for all* t, *and by corollary 3.74,* $F_{B_{n,\lambda/n}} \Rightarrow F_{P_\lambda}$. ■

Example 5.6 *The Poisson Limit theorem has immediate applications in finance in any situation in which one is modelling the binomial outcomes of a large group of individuals, with each having the same or similar probabilities of the event being observed. For example, in a relatively large portfolio of bonds with similar credit ratings, or similarly rated bank loans of various types, the event of default is fundamentally binomial with* p *equalling the probability of default over the given period.*

For example, with $n = 200$ *loans with default probability* $p = 0.02$ *in one year, one can model the number of defaults random variable* N *as the sum of* 200 *binomials, or approximately as a Poisson random variable with* $\lambda = np = 4$. *It is then true that in either model, the assumption of independence from loan to loan is reasonable, though the assumed value of* p *is highly dependent on the economic cycle. Thus the decision to increase* p *due to a forthcoming recession say, if justifiable on any loan, is justifiable on all loans since all loans are correlated with this event.*

One can similarly model a variety of insurable events this way. For example, death, disability, hospitalization, etc., can be modeled in the life insurance industry as sums of binomials or approximately Poisson random variables as long as the group being modeled is reasonably homogeneous and individuals have similar values for p. *In the property and casualty industry, various automobile and homeowner insurable events can be modeled with binomials or Poisson random variables when the groups modeled are reasonably homogeneous in terms of claim probabilities.*

We will see next that the Poisson can also be used even in the case of inhomogeneous binomials.

5.2.2 "Weak Law of Small Numbers"

The weak law of small numbers is the tongue-in-cheek name given to various generalizations of Poisson's Limit theorem which weaken the

assumption in 5.4 that for given n, $B_{n,\lambda/n} = \sum_{m=1}^{n} X_m^B$ where $\{X_m^B\}_{m=1}^n$ are independent and **identically distributed** standard binomial variables with $p = \lambda/n$. Versions of the small law assume that for each n, $\{X_{n_m}^B\}_{m=1}^n$ are independent standard binomial variables with $p = p_{n_m}$ and where for each n we have that $\sum_{m=1}^{n} p_{n_m} = \lambda$, or more generally $\sum_{m=1}^{n} p_{n_m} \to \lambda$ as $n \to \infty$. This can even be generalized further to more general "binomials" which assume that $\Pr[X_{n_m}^B = 1] = p_{n_m}$, $\Pr[X_{n_m}^B = 0] = 1 - p_{n_m} - \epsilon_{n_m}$ and $\Pr[X_{n_m}^B \geq 2] = \epsilon_{n_m}$ where $\max_m\{\epsilon_{n_m}\} \to 0$ as $n \to \infty$. In order to ensure that all of the probabilities p_{n_m} become small as $n \to \infty$, and thus this remains a law of "small numbers," it is necessary to also require that $\max_m\{p_{n_m}\} \to 0$ as $n \to \infty$.

We state and prove the version of intermediate generality, but first, a definition.

Definition 5.7 *A collection of random variables, $\{\{X_{n,m}\}_{m=1}^n\}_{n=1}^\infty$ is called a **triangular array** if for each n, the random variables $\{X_{n,m}\}_{m=1}^n$ are independent.*

Remark 5.8 *In many texts triangular arrays are defined to allow $1 \leq m \leq m_n$ where $m_n \to \infty$ as $n \to \infty$. We have no need of the more general notion.*

Proposition 5.9 (Weak Law of Small Numbers) *Let $S_n = \sum_{m=1}^{n} X_{n_m}^B$ where for each n, $\{X_{n_m}^B\}_{m=1}^n$ are independent standard binomial variables with $p = p_{n_m}$, and where both $\sum_{m=1}^{n} p_{n_m} \to \lambda > 0$ and $\max_m\{p_{n_m}\} \to 0$ as $n \to \infty$. If F_{P_λ} denotes the distribution function of the Poisson with parameter λ, then as $n \to \infty$:*

$$F_{S_n} \Rightarrow F_{P_\lambda}. \tag{5.5}$$

Proof. *By 3.35:*

$$M_{S_n}(t) = \prod_{m=1}^{n} \left(1 + p_{n_m}(e^t - 1)\right),$$

and

$$M_{B_{n,\lambda/n}}(t) = \left(1 + \frac{\lambda}{n}(e^t - 1)\right)^n,$$

where $M_{B_{n,\lambda/n}}(t)$ is the moment generating function of the sum of n independent, identically distributed binomials with $p = \lambda/n$. By the Poisson Limit theorem $M_{B_{n,\lambda/n}}(t) \to M_{P_\lambda}(t)$ for all t, and hence 5.5 will be proven if we show that $M_{S_n}(t)/M_{B_{n,\lambda/n}}(t) \to 1$ for all t. We will do this by showing

that $\ln\left[M_{S_n}(t)/M_{B_{n,\lambda/n}}(t)\right] \to 0$ *for all* t *since both* $M(t) > 0$ *and both the logarithmic and exponential functions are continuous.*

To this end, fix t *and for arbitrary* $\epsilon < \max[1,\, e^t - 1]$ *define* N *so that for all* $n \geq N$ *and all* $m \leq n$:

$$\left|p_{n_m}(e^t-1)\right| < \epsilon \text{ and } \left|\frac{\lambda}{n}(e^t-1)\right| < \epsilon.$$

The first bound is possible since $\max_m\{p_{n_m}\} \to 0$ *as* $n \to \infty$, *and the second is apparent. Recall the Taylor series for* $\ln(1+x)$, *that for* $|x| < 1$:

$$\ln(1+x) = \sum\nolimits_{j=1}^{\infty}(-1)^{j+1}x^j/j. \tag{5.6}$$

Thus by the above bounds, both $\ln\left(1+p_{n_m}(e^t-1)\right)$ *and* $\ln\left(1+\frac{\lambda}{n}(e^t-1)\right)$ *can be expanded as absolutely convergent Taylor series. With* $a \equiv e^t - 1$ *and* $A_n \equiv M_{S_n}(t)/M_{B_{n,\lambda/n}}(t)$:

$$\begin{aligned}
\ln A_n &= \sum\nolimits_{m=1}^{n}\sum\nolimits_{j=1}^{\infty}(-1)^{j+1}a^j\left(p_{n_m}^j - (\lambda/n)^j\right)\Big/ j \\
&= \sum_{j=1}^{\infty}(-1)^{j+1}a^j\left[\sum\nolimits_{m=1}^{n}\left(p_{n_m}^j - (\lambda/n)^j\right)\right]\Big/ j \\
&= a\sum\nolimits_{m=1}^{n}(p_{n_m} - \lambda/n) + \sum\nolimits_{j=2}^{\infty}(-1)^{j+1}a^j\left[\sum\nolimits_{m=1}^{n}\left(p_{n_m}^j - (\lambda/n)^j\right)\right]\Big/ j,
\end{aligned}$$

where the interchange in summations is justified by the absolute convergence of these series for $n \geq N$. *Now for* $j \geq 2$, *since* $p_{n_m} < \epsilon/|a|$ *and* $\lambda/n < \epsilon/|a|$,

$$\begin{aligned}
p_{n_m}^j - (\lambda/n)^j &= [p_{n_m} - \lambda/n]\sum\nolimits_{k=0}^{j-1}p_{n_m}^{j-k-1}(\lambda/n)^k \\
&< [p_{n_m} - \lambda/n]\sum\nolimits_{k=0}^{j-1}(\epsilon/|a|)^{j-1} \\
&\leq j\,|p_{n_m} - \lambda/n|\,(\epsilon/|a|)^{j-1}.
\end{aligned}$$

Finally by the triangle inequality:

$$\begin{aligned}
\left|\ln A_n - a\sum\nolimits_{m=1}^{n}(p_{n_m} - \lambda/n)\right| &\leq \sum_{j=2}^{\infty}\frac{|a|^j}{j}\sum\nolimits_{m=1}^{n} j\,|p_{n_m} - \lambda/n|\,(\epsilon/|a|)^{j-1} \\
&= |a|\sum\nolimits_{m=1}^{n}|p_{n_m} - \lambda/n|\sum_{j=2}^{\infty}\epsilon^{j-1} \\
&= |a|\sum\nolimits_{m=1}^{n}|p_{n_m} - \lambda/n|\,\frac{\epsilon}{1-\epsilon} \\
&\leq |a|\left(\sum\nolimits_{m=1}^{n}p_{n_m} + \lambda\right)\frac{\epsilon}{1-\epsilon}.
\end{aligned}$$

Since ϵ was arbitrary and $\sum_{m=1}^{n} p_{n_m} \to \lambda$ by assumption, we conclude that $|\ln A_n| \to 0$. Hence, $M_{S_n}(t) \to M_{P_\lambda}(t)$ for all t, and by corollary 3.74, $F_{S_n} \Rightarrow F_{P_\lambda}$. ■

Example 5.10 *This theorem allows the examples in the above section to be more generally applied. Specifically, a portfolio of bonds with various credit ratings can also be modeled with a Poisson random variable. Similarly for various insurance applications. The requirement that $\max_m\{p_{n_m}\} \to 0$ as $n \to \infty$ provides a constraint, however, that the default or insurance claim probabilities should individually be small.*

For example, given $n = 400$ bonds, 100 with $p_1 = 0.002$, 200 with $p_2 = 0.005$ and 100 with $p_3 = 0.01$, one could reasonable model the number of defaults in this portfolio as a Poisson random variable with $\lambda = 100p_1 + 200p_2 + 100p_3 = 2.2$.

5.2.3 De Moivre-Laplace Theorem

The **De Moivre-Laplace Theorem** is a special case of a very general result discussed below that is known as the **Central Limit Theorem**. The current result addresses another question about the "limiting distribution" of the binomial distribution as $n \to \infty$, and that is the question of probability estimates. If $X_n^B \equiv \sum_{j=1}^{n} X_1^B$ is a binomially distributed random variable with parameters n and p, where X_1^B are independent standard binomial variables, we have from 1.6 that for integers a and b :

$$\Pr[a \le X_n^B \le b] = \sum\nolimits_{j=a'}^{b'} \binom{n}{j} p^j (1-p)^{n-j},$$

where $a' = \max(a, 0)$ and $b' = \min(b, n)$. In this form it makes little sense to attempt to specify what happens to this probability as $n \to \infty$, or indeed identify the distribution function, because the range of the random variable is $[0, n]$ which grows with n. Put another way, we have from 3.48 that:

$$E[X_n^B] = np, \qquad Var[X_n^B] = np(1-p),$$

so in contrast to the Poisson limit theorem, here p is fixed and hence $np \to \infty$ and thus both the mean and variance of X_n^B grow without bound as $n \to \infty$. So in order to investigate quantitatively the limiting probabilities under this distribution as $n \to \infty$, some form of "scaling" is necessary to stabilize the distribution.

The approach used by **Abraham de Moivre** (1667 – 1754) in the special case of $p = \frac{1}{2}$, and many years later generalized to all p, $0 < p < 1$, by **Pierre-Simon Laplace** (1749 – 1827), was to consider what is now called the **normalized random variable,** Y_n^B, defined by:

$$Y_n^B \equiv \frac{X_n^B - E[X_n^B]}{\sqrt{Var[X_n^B]}} = \frac{X_n^B - \mu_B}{\sigma_B}. \tag{5.7}$$

Since $\mu_B = np$ and $\sigma_B = \sqrt{np(1-p)}$ are constants for each n, the random variable Y_n^B has the same binomial probabilities as does X_n^B in the sense that:

$$\Pr\left[Y_n^B = j' \equiv \frac{j - np}{\sqrt{np(1-p)}}\right] = \Pr\left[X_n^B = j\right].$$

The range of Y_n^B :

$$Rng\left[Y_n^B\right] = \left\{ \left. \frac{j - np}{\sqrt{np(1-p)}} \right| 0 \le j \le n \right\},$$

is contained in $[-a\sqrt{n}, \sqrt{n}/a]$ with $a = \sqrt{p/q}$, and thus this may not seem to be a much better result than that for X_n^B with range $[0, n]$. However, a calculation using 3.5 yields that:

$$E[Y_n^B] = 0, \qquad Var[Y_n^B] = 1.$$

In other words, in a certain sense the distribution of Y_n^B "stays put," in contrast the the distribution of X_n^B that "wanders off."

Consequently, with mean and variance both constant and independent of n, the question of investigating and potentially identifying the limiting distribution of Y_n^B as $n \to \infty$ is better defined and its pursuit more compelling. Recalling that the standard normal distribution has mean 0 and variance 1, the following proposition identifies the desired result.

Notation 5.11 *Note that the variate Y_n^B in 5.8 can also be expressed as a summation, $S_n \equiv \sum_{m=1}^n Y_{1,m}^B$ where for each n and independent standard binomials $\{X_{1,m}^B\}_{m=1}^n$:*

$$\{Y_{1,m}^B\}_{m=1}^n \equiv \left\{ \frac{X_{1,m}^B - p}{\sqrt{np(1-p)}} \right\}_{m=1}^n$$

are independent binomial variables with fixed p, and thus $\mu_{Y_1} = 0$ and $\sigma_{Y_1} = 1/\sqrt{n}$. Consequently $\mu_{S_n} = 0$ and $\sigma_{S_n}^2 = 1$ as noted above.

Proposition 5.12 (De Moivre-Laplace theorem) *Let $F_{Y_n^B}(j')$ denote the distribution function of the normalized general binomial with parameters n, p where $0 < p < 1$, associated with the normalized density function in 1.6. In other words, $f_{Y_n^B}(j') \equiv f_{B_n}(j)$ where $j' = \frac{j-np}{\sqrt{np(1-p)}}$ for $0 \leq j \leq n$. With $\Phi(x)$ denoting the distribution function of the standard normal associated with the density function in 1.33, then as $n \to \infty$,*

$$F_{Y_n^B} \Rightarrow \Phi. \tag{5.8}$$

Proof. *Let $M_{B_{n,p}}(t)$ denote the moment generating function of the general binomial, which from 3.49 is given for all t by:*

$$M_{B_{n,p}}(t) = \left(1 + p(e^t - 1)\right)^n.$$

Then from 3.19 and 3.49, the moment generating function for the standardized general binomial is defined for all t as follows, denoting $q \equiv 1 - p$:

$$\begin{aligned} M_{Y_n^B}(t) &= \exp\left[-npt\,/\sqrt{npq}\right] M_{B_{n,p}}(t/\sqrt{npq}) \\ &= \exp\left[-npt\,/\sqrt{npq}\right] \left[1 + p\left(\exp\left[t\,/\sqrt{npq}\right] - 1\right)\right]^n \\ &= \left[q\exp\left(-pt\,/\sqrt{npq}\right) + p\exp\left(qt\,/\sqrt{npq}\right)\right]^n \\ &= \left[q\exp\left(-t\,\sqrt{p/q}\Big/\sqrt{n}\right) + p\exp\left(t\,\sqrt{q/p}\Big/\sqrt{n}\right)\right]^n. \end{aligned}$$

The moment generating function of the standard normal $M_\Phi(t)$ is given in 3.67 and defined for all t by:

$$M_\Phi(t) = \exp\left(t^2/2\right).$$

To prove $M_{Y_n^B}(t) \to M_\Phi(t)$ for all t is equivalent to demonstrating that $\ln M_{Y_n^B}(t) \to t^2/2$ for all t since both $M(t) > 0$ and both the logarithmic and exponential functions are continuous. To this end:

$$\ln M_{Y_n^B}(t) = n \ln\left[q\exp\left[-t\,\sqrt{p/q}\Big/\sqrt{n}\right] + p\left(\exp\left[t\,\sqrt{q/p}\Big/\sqrt{n}\right]\right)\right].$$

Now define the function

$$f(x) = \ln\left[q\exp\left[-t\sqrt{p/q}\sqrt{x}\right] + p\left(\exp\left[t\sqrt{q/p}\sqrt{x}\right]\right)\right],$$

and note that with $\Delta x = 1/n$, that $\lim_{n\to\infty} \ln M_{Y_n^B}(t) = f'(0)$ by definition, assuming that this derivative exists. To investigate existence note that $f(x) = \ln g(x)$, so if $g(x) > 0$ for all x, $f(x)$ is differentiable everywhere $g(x)$

is differentiable. But since $\sqrt{x}$ is not differentiable at $x = 0$, some analysis is needed.

Expanding the exponentials in a Taylor series, we have here that:

$$g(x) \equiv q\left[1 - t\sqrt{p/q}\sqrt{x} + O_1(x)\right] + p\left[1 + t\sqrt{q/p}\sqrt{x} + O_2(x)\right],$$

where $O_j(x)$ denotes the absolutely convergent remainder series in powers of $t\sqrt{x}$ with leading terms of the form ct^2x. A calculation verifies that the $\sqrt{x}$-terms cancel out and that $g(x) = 1 + O(x)$, and so $g(x)$ is differentiable at $x = 0$, and thus so too is $f(x)$. Finally, a calculation of this derivative produces $f'(0) = t^2/2$.

Hence, $M_{Y_n^B}(t) \to M_\Phi(t)$ for all t, and by corollary 3.74, $F_{Y_n^B} \Rightarrow \Phi$. ■

Remark 5.13 (On Approximations) *Because $\Phi(x)$ is continuous everywhere, the De Moivre-Laplace theorem conclusion of $F_{Y_n^B} \Rightarrow \Phi$ implies that*

$$F_{Y_n^B}(y) \to \Phi(y)$$

for all *y. Now with $j' \equiv \frac{j-\mu_B}{\sigma_B}$, where $\mu_B \equiv np$, $\sigma_B \equiv \sqrt{npq}$ and $0 \le j \le n$, we have for any real number w :*

$$\begin{aligned} F_{B_{n,p}}(w) &= \sum\nolimits_{j \le w} \frac{n!}{j!(n-j)!} p^j q^{n-j} \\ &= \sum\nolimits_{j' \le (w-\mu_B)/\sigma_B} \frac{n!}{j!(n-j)!} p^j q^{n-j} \\ &= F_{Y_n^B}\left(\left(w - \mu_B\right)/\sigma_B\right). \end{aligned}$$

The above limiting result then states that as $n \to \infty$:

$$F_{B_{n,p}}(w)/\Phi\left(\left(w - \mu_B\right)/\sigma_B\right) \to 1.$$

Of course we cannot say that $F_{B_{n,p}}(w) \to \Phi\left(\left(w - \mu_B\right)/\sigma_B\right)$ since μ_B and σ_B are functions of n.

In applications where by definition $n << \infty$, this limiting result for the binomial provides an approximation that for n "large,"

$$F_{B_{n,p}}(w) \approx \Phi\left(\left(w - \mu_B\right)/\sigma_B\right),$$

with corresponding approximations for expressions such as $F_{B_{n,p}}(w) - F_{B_{n,p}}(v)$ when $w > v$:

$$F_{B_{n,p}}(w) - F_{B_{n,p}}(v) \approx \Phi\left(\left(w - \mu_B\right)/\sigma_B\right) - \Phi\left(\left(v - \mu_B\right)/\sigma_B\right). \tag{5.9}$$

In practice, "large" is often interpreted as $n \geq 30$ *or* $n \geq 50$ *for most applications.*

In such applications one is typically interested in integer values of w *and/or* v*, and this approximation can not therefore be uniformly useful. For example, if* $w = j$ *an integer and real* $v < j$*, we would conclude that as* $v \to j$ *that:*

$$F_{B_{n,p}}(j) - F_{B_{n,p}}(v) \to F_{B_{n,p}}(j) - F_{B_{n,p}}(j^-) = f_{B_{n,p}}(j),$$

while

$$\Phi((j - \mu_B)/\sigma_B) - \Phi((v - \mu_B)/\sigma_B) \to 0.$$

While not apparent from the above approach to the proof, the problem here can be investigated with a more direct analysis of the binomial probabilities using Stirling's formula as in 3.91 above, and some careful calculations as in Reitano, proposition 8.24. This then reveals that if $j'_n \to y$ *as* $n \to \infty$*, where each* j'_n *is in the range of the random variable* Y_n^B*, then*

$$\lim_{n\to\infty} \sqrt{npq}\, \Pr\{Y_n^B = j'_n\} \to e^{-y^2/2}/\sqrt{2\pi}. \tag{5.10}$$

This produces the approximation:

$$f_{Y_n^B}(j'_n) \approx \exp\left[-(j'_n)^2/2\right]/\sqrt{2\pi npq},$$

which for $0 \leq j \leq n$ *is equivalent to*:

$$f_{B_{n,p}}(j) \approx \exp\left[-\frac{1}{2}\left(\frac{j - \mu_B}{\sigma_B}\right)^2\right] \Big/ \sqrt{2\pi npq}. \tag{5.11}$$

This approximation for $f_{Y_n^B}(j')$ *can be interpreted as a single term in the Riemann summation which approximates the integral of the standard normal density function* $e^{-\frac{1}{2}y^2}/\sqrt{2\pi}$ *using* $\Delta y = 1/\sigma_B$ *where* $\sigma_B = \sqrt{npq}$*. This value of* Δy *is seen to equal* $j'_n(k+1) - j'_n(k)$ *where* $\{j'_n(k)\}_{k=0}^n$ *denote the* $n+1$ *values of this variate. Hence, the discrete probability* $f_{B_{n,p}}(j)$ *in 5.11 is approximated by the standard normal probability of an interval of length* $1/\sigma_B$ *which contains* $(j - \mu_B)/\sigma_B$*. Alternatively, for some* λ *with* $0 < \lambda < 1$*, one wants to approximate* $f_{B_{n,p}}(j)$ *by the integral of the standard normal density over* $[a, b]$ *with* $a = [j - \mu_B - (1 - \lambda)]/\sigma_B$ *and* $b = [j - \mu_B + \lambda]/\sigma_B$*.*

The conventional solution is to make a ***half interval adjustment,*** *or* ***half integer adjustment*** *for the above approximation, producing:*

$$f_{B_{n,p}}(j) \approx \Phi([j + 1/2 - \mu_B]/\sigma_B) - \Phi([j - 1/2 - \mu_B]/\sigma_B), \tag{5.12}$$

from which one obtains:

$$F_{B_{n,p}}(w) - F_{B_{n,p}}(v) \approx \Phi([w + 1/2 - \mu_B]/\sigma_B) - \Phi([v - 1/2 - \mu_B]/\sigma_B). \tag{5.13}$$

5.2.4 The Central Limit Theorem 1

There are many versions of "the" Central Limit theorem, all of which generalize the De Moivre-Laplace theorem in one remarkable way or another. In essence, what every version states and what makes any such version indeed a "Central" Limit theorem is this – under a wide variety of assumptions the distribution function of the sum of n independent random variables, normalized as in 5.7, converges to the standard normal distribution as $n \to \infty$. Remarkably, these random variables need not be identically distributed, just independent, although the need for normalization demands that these random variables have at least 2 moments: means and variances. When not identically distributed there is a requirement that the sequence of variances does not grow too fast so as to preclude latter terms in the random variable series from increasingly dominating the summation, as well as a requirement that they do not converge to 0 so quickly that the average variance converges to 0.

These theorems can be equivalently stated in terms of the **sum** of independent random variables or their **average**. This is because by 3.26 and 3.5:

$$E\left[\sum_{j=1}^{n} X_j/n\right] = E\left[\sum_{j=1}^{n} X_j\right] \Big/ n$$

and from 3.30:

$$Var\left[\frac{1}{n}\sum_{j=1}^{n} X_j\right] = Var\left[\sum_{j=1}^{n} X_j\right] \Big/ n^2.$$

and thus normalized sums equal normalized averages:

$$\frac{\sum_{j=1}^{n} X_j - E\left[\sum_{j=1}^{n} X_j\right]}{\sqrt{Var\left[\sum_{j=1}^{n} X_j\right]}} = \frac{\frac{1}{n}\sum_{j=1}^{n} X_j - E\left[\frac{1}{n}\sum_{j=1}^{n} X_j\right]}{\sqrt{Var\left[\frac{1}{n}\sum_{j=1}^{n} X_j\right]}}. \tag{5.14}$$

So while the ranges of the sum and average of independent random variables are quite different, the associated normalized random variables are identical.

Consequently, Central Limit theorems in general, and the De Moivre-Laplace theorem in particular, apply to the sums of random variables if and only if they apply to the averages of random variables. More generally, if a given version of a Central Limit theorem applies to $\sum_{j=1}^{n} X_j$ for independent $\{X_j\}$, then it applies to $\sum_{j=1}^{n} Y_j$ where $Y_j = aX_j + b$ for constants a and b.

In this section, we provide a proof of a simplified version of the Central limit theorem in the case of independent, identically distributed random

variables which have moments of all orders and a convergent moment generating function. Mechanically, the proof will be quite similar to that of the De Moivre-Laplace theorem, except that we will have to accommodate a more general form of $M_X(t)$.

In book 6, using more powerful tools than the moment generating function, we will present more general versions of this result, with far weaker assumptions.

Proposition 5.14 (Central Limit theorem 1) *Let F_X denote the distribution function of a random variable X with mean and variance denoted μ and σ^2 and moment generating function $M_X(t)$ convergent for $t \in (-t_0, t_0)$ with $t_0 > 0$. Let Y_n denote the normalized random variable associated with the sum or average of n independent values of X, defined as in 5.7:*

$$Y_n = \left[\sum_{j=1}^n X_j - n\mu\right] \Big/ \left[\sqrt{n}\sigma\right] = \left[\frac{1}{n}\sum_{j=1}^n X_j - \mu\right] \Big/ \left[\sigma/\sqrt{n}\right].$$

Then as $n \to \infty$,

$$F_{Y_n} \Rightarrow \Phi, \tag{5.15}$$

where $\Phi(x)$ denotes the distribution function of the standard normal associated with the density function in 1.33.

Proof. *Since $Y_n = \sum_{j=1}^n (X_j - \mu)/\sqrt{n}\sigma$, by 3.35 and 3.19:*

$$M_{Y_n}(t) = \left[\exp\left(-\mu t/\left[\sqrt{n}\sigma\right]\right) M_X\left(t/\left[\sqrt{n}\sigma\right]\right)\right]^n,$$

and by 3.38,

$$M_X\left(t/\left[\sqrt{n}\sigma\right]\right) = \sum_{j=0}^\infty \mu_j' \left(t/\left[\sqrt{n}\sigma\right]\right)^j /j!.$$

Recalling that $\mu_0' = 1$, $\mu_1' = \mu$ and $\mu_2' = \sigma^2 + \mu^2$:

$$M_X\left(t/\left[\sqrt{n}\sigma\right]\right) = 1 + t\mu/\left[\sqrt{n}\sigma\right] + t^2\left(\sigma^2+\mu^2\right)/\left(2n\sigma^2\right) + n^{-\frac{3}{2}}E_1(n),$$

where $E_1(n) = \sum_{j=3}^\infty \mu_j' (t/\sigma)^j n^{\frac{3-j}{2}}/j!$. Now since $M_X(t)$ is by assumption absolutely convergent for $|t| < t_0$, $M_X\left(t/\left[\sqrt{n}\sigma\right]\right)$ and hence $E_1(n)$ are absolutely convergent for $|t| < \sqrt{n}\sigma t_0$, and so for any t, $E_1(n) \to \mu_3' (t/\sigma)^3/6$ as $n \to \infty$.

Similarly, using the Taylor series in 3.40 applied to $\exp\left[-\mu t/\sqrt{n}\sigma\right]$ obtains:

$$\begin{aligned}\exp\left[-\mu t/\left(\sqrt{n}\sigma\right)\right] &= \sum_{j=0}^\infty \left[-\mu t/\sqrt{n}\sigma\right]^j /j! \\ &= 1 - t\mu/\sqrt{n}\sigma + t^2\mu^2/\left(2n\sigma^2\right) + n^{-\frac{3}{2}}E_2(n),\end{aligned}$$

where $E_2(n) = \sum_{j=3}^{\infty} \left[-\mu t/\sigma\right]^j n^{\frac{3-j}{2}}/j!$ *is absolutely convergent for all* t*, and for any* t*,* $E_2(n) \to -\mu^3 \left(t/\sigma\right)^3/6$ *as* $n \to \infty$*. With a bit of algebra:*

$$\exp\left[-\mu t/\sqrt{n}\sigma\right] M_X\left(t/\left[\sqrt{n}\sigma\right]\right) = 1 + t^2/\left(2n\right) + n^{-\frac{3}{2}} E_3(n),$$

where the combined error term $E_3(n)$ *is is absolutely convergent for* $|t| < \sqrt{n}\sigma t_0$*, and* $E_3(n) \to -\mu^3 \mu_3' \left(t/\sigma\right)^6/36$ *as* $n \to \infty$*.*

This expression can now be raised to the nth power, a logarithm taken, and the function $\ln(1+x)$ *expanded in a Taylor series as in 5.6. To simplify, we only keep track of the powers of* n *that are needed for the final limit, meaning only those terms that will not converge to zero as* $n \to \infty$*. This produces:*

$$\begin{aligned}
\ln M_{Y_n}(t) &= n \ln\left[1 + t^2/\left(2n\right) + n^{-\frac{3}{2}} E_3(n)\right] \\
&= n\left[\left(t^2/\left(2n\right) + n^{-\frac{3}{2}} E_3(n)\right) - \frac{1}{2}\left(t^2/\left(2n\right) + n^{-\frac{3}{2}} E_3(n)\right)^2 + O\left[n^{-3}\right]\right] \\
&= t^2/2 + O\left[n^{-1/2}\right].
\end{aligned}$$

To justify the second step in which the power series expansion for $\ln(1+x)$ *is invoked with* $x = t^2/\left(2n\right) + n^{-\frac{3}{2}} E_3(n)$*, we must verify for any* t*, that* $|x| < 1$ *for* n *large. But since* $|x| \leq t^2/\left(2n\right) + n^{-\frac{3}{2}} \left|E_3(n)\right|$ *where* $E_3(n)$ *is continuous and bounded, the conclusion follows.*

Hence for all t*,*

$$\ln M_{Y_n}(t) \to t^2/2$$

as $n \to \infty$*, and equivalently* $M_{Y_n}(t) \to M_\Phi(t)$ *for all* t*. By corollary 3.74,* $F_{Y_n} \Rightarrow \Phi$*.* ■

The assumption in this version of the Central Limit theorem, that $M_X(t)$ exists and is convergent for all $t \in (-t_0, , t_0)$ with $t_0 > 0$, is quite strong since as was proved in proposition 3.22 that this implies that the associated distribution function of X has finite moments of all orders. In fact, in the context of independent and identically distributed random variables, the conclusion in 5.15 is valid under the assumption that X has only two finite moments, a mean and variance. It is also valid more generally for sums of independent random variables which are not identically distributed. But in neither case will the tools of this chapter suffice for the demonstration. The problem is that the moment generating function is a blunt instrument. If it exists on an open interval $(-t_0, , t_0)$ with $t_0 > 0$, then all moments exist

by proposition 3.24. There is no way to adapt this argument in the case of random variables with only finitely many moments.

In book 5, once a general integration theory is developed, we will study the Fourier transform of a measurable function. In the same way that the moment generating function of a random variable is defined relative to the Laplace transform of this measurable function as noted in remark 3.11, in book 6 we will define the **characteristic function** of a random variable in terms of the Fourier transform theory studied in book 5. And unlike the moment generating function which may or may not exist, characteristic functions associated with random variables always exist. Characteristic functions will also be seen to uniquely determine the underlying distribution function of the random variable, so they are useful in proofs the same way that moment generating functions are useful. But more generally, the associated proofs can also be implemented for random variables with only finitely many moments. And finally, the moments of a distribution, to the extent they exist, will be seen to appear in the series expansion of the characteristic function in a familiar way, reminiscent of 3.39.

5.2.5 Smirnov's Limit Theorem on Order Statistics of the Uniform Distribution

In this section we investigate the limiting distribution of the order statistics of independent continuous uniform random variables on $[0,1]$. This result is named for its discoverer, **N. V. (Nikolaĭ Vasil'evich) Smirnov** (1900 – 1966), and will be generalized in book 6 using a generalized version of the central limit theorem introduced above. **Smirnov's Limit Theorem** addresses the limiting distribution of the kth order statistic from a uniform sample, $\{Y_j\}_{j=1}^n$, as $n \to \infty$. It requires that both $k \to \infty$ and $n-k \to \infty$, so for example k cannot be fixed as $n \to \infty$.

Thus this result applies if $k/n \to q \in (0,1)$, the qth quantile of this distribution, since it is then true that both $k \to \infty$ and $n-k \to \infty$ as $n \to \infty$. On the other hand, the requirement that both $k \to \infty$ and $n-k \to \infty$ as $n \to \infty$ also allows that $k/n \to 0, 1$, by choosing $k \approx \sqrt{n}$, or $k = n - \sqrt{n}$, respectively. Hence, this result allows k to converge in relative terms to the tails of the distribution of order statistics. More generally, this theorem allows k/n to not converge to any value as $n \to \infty$, and remains valid as long as both $k \to \infty$ and $n-k \to \infty$.

Proposition 5.15 (Smirnov's Limit Theorem) *Let* $\{Y_{(k)}\}_{k=1}^n$ *be the or-*

der statistics from a continuous uniform distribution on $[0,1]$. *Define*

$$b_n = \frac{k-1}{n-1}, \qquad a_n = \sqrt{\frac{b_n(1-b_n)}{n-1}}, \qquad Y'_{(k)} \equiv \frac{Y_{(k)} - b_n}{a_n},$$

and $F_{(k)}$ *the distribution function of* $Y'_{(k)}$.

Then if $k \to \infty$ *and* $n - k \to \infty$ *as* $n \to \infty$:

$$F_{(k)} \Rightarrow \Phi, \tag{5.16}$$

where $\Phi(x)$ *is the distribution function of the standard normal.*

Remark 5.16 *Recall from example 2.5 that the distribution function for* $Y_{(k)}$ *is Beta with* $v = k$ *and* $w = n - k + 1$. *A calculation using 3.61 will show that* $b_n / E\left[Y_{(k)}\right] \to 1$ *and* $a_n^2 / Var\left[Y_{(k)}\right] \to 1$ *as* $n \to \infty$ *since both* $k \to \infty$ *and* $n - k \to \infty$. *So* $Y'_{(k)}$ *is effectively a* ***normalized random variable*** *as defined above in 5.7. In fact, by 9.18 of proposition 9.16 of book 2, 5.16 assures that* $\widetilde{F}_{(k)} \Rightarrow \Phi$ *if* $\widetilde{F}_{(k)}$ *denotes the distribution function of* $Y'_{(k)}$ *explicitly defined as in 5.7 in terms of the moments of* $Y_{(k)}$. *In corollary 5.18 below we find another application of proposition 9.16 of book 2.*

Proof. *We begin by proving convergence of the associated density functions. By 2.4 the density function of* $Y_{(k)}$ *is defined on* $[0,1]$ *by:*

$$g_{(k)}(y) = \frac{n!}{(k-1)!\,(n-k)!} y^{k-1} (1-y)^{n-k},$$

and thus that of $Y'_{(k)}$ *is defined on* $[-b_n/a_n, (1-b_n)/a_n]$ *by:*

$$\begin{aligned} f_{(k)}(y) &= \frac{n!}{(k-1)!\,(n-k)!} (a_n y + b_n)^{k-1} (1 - a_n y - b_n)^{n-k} \\ &= \frac{n!}{(k-1)!\,(n-k)!} a_n b_n^{k-1} (1-b_n)^{n-k} \left(1 + \frac{a_n y}{b_n}\right)^{k-1} \left(1 - \frac{a_n y}{1-b_n}\right)^{n-k}. \end{aligned}$$

The constant simplifies with Stirling's formula in 3.91:

$$\begin{aligned} & \frac{n!}{(k-1)!\,(n-k)!} a_n b_n^{k-1} (1-b_n)^{n-k} \\ = \; & \frac{n^{n+1/2} e^{-1}}{\sqrt{2\pi}\,(k-1)^{k-1/2}\,(n-k)^{n-k+1/2}} \left(\frac{k-1}{n-1}\right)^{k-1/2} \left(\frac{n-k}{n-1}\right)^{n-k+1/2} \left(\frac{1}{n-1}\right)^{1/2} \\ = \; & \frac{1}{\sqrt{2\pi}} \left(1 + \frac{1}{n-1}\right)^{n+1/2} e^{-1}. \end{aligned}$$

This last expression converges to $\frac{1}{\sqrt{2\pi}}$ *as* $n \to \infty$ *since as in the proof of the Poisson limit theorem,* $\left(1+\frac{1}{n-1}\right)^{n+1/2} \to e$.

The next step is to prove that the expression in y *converges to* $e^{-y^2/2}$. *For this we take logs, apply the Taylor series expansion in 5.6 noting that both* $\frac{a_n y}{b_n}$ *and* $\frac{a_n y}{1-b_n}$ *have absolute value less than* 1, *then explicitly calculate the first two terms to yield:*

$$\begin{aligned}
&\ln\left[\left(1+\frac{a_n y}{b_n}\right)^{k-1}\left(1-\frac{a_n y}{1-b_n}\right)^{n-k}\right] \\
&= (k-1)\left[\sum\nolimits_{j=1}^{\infty}(-1)^{j+1}\left(\frac{a_n y}{b_n}\right)^j \Big/ j\right] + (n-k)\left[-\sum\nolimits_{j=1}^{\infty}\left(\frac{a_n y}{1-b_n}\right)^j \Big/ j\right].
\end{aligned}$$

The coefficient of y^j/j *for* $j \geq 1$ *is:*

$$\begin{aligned}
c_j &\equiv (-1)^{j+1}(k-1)\left(\frac{a_n}{b_n}\right)^j - (n-k)\left(\frac{a_n}{1-b_n}\right)^j \\
&= (-1)^{j+1}(k-1)\left(\frac{n-k}{(n-1)(k-1)}\right)^{j/2} - (n-k)\left(\frac{k-1}{(n-1)(n-k)}\right)^{j/2} \\
&= (-1)^{j+1}(k-1)\left(\frac{-1}{n-1}+\frac{1}{k-1}\right)^{j/2} - (n-k)\left(\frac{-1}{n-1}+\frac{1}{n-k}\right)^{j/2}.
\end{aligned}$$

From this expression it follows that $c_1 = 0$, $c_2 = -1$, *and the coefficient of* y^j/j *converges to zero as* $n \to \infty$ *for* $j \geq 3$ *since* $k \to \infty$ *and* $n-k \to \infty$.

In summary, $f_{(k)}(y) \to \varphi(y) \equiv \frac{1}{\sqrt{2\pi}}e^{-y^2/2}$ *for all* y *as* $n \to \infty$. *The final step is to prove convergence of the associated distribution functions,* $F_{(k)}(y) \to \Phi(y)$ *for all* y. *To this end let* $h_{(k)}(y) \equiv \max\left[\varphi(y) - f_{(k)}(y), 0\right]$ *and note that:*

$$\left|f_{(k)}(y) - \varphi(y)\right| = f_{(k)}(y) - \varphi(y) + 2h_{(k)}(y),$$

and since $f_{(k)}(y)$ *and* $\varphi(y)$ *are densities and integrate to* 1*:*

$$\int \left|f_{(k)}(y) - \varphi(y)\right| dy = 2\int h_{(k)}(y)dy.$$

Now $0 \leq h_{(k)}(y) \leq \varphi(y)$ *and* $h_{(k)}(y) \to 0$ *pointwise, so by Lebesgue's dominated convergence theorem of proposition 2.61 of book 3,* $\int h_{(k)}(y)dy \to 0$*:*

$$\int \left|f_{(k)}(y) - \varphi(y)\right| dy \to 0.$$

This now implies that for any Lebesgue measurable set A:

$$\left|\int_A f_{(k)}(x)dx - \int_A \varphi(x)dx\right| \leq \int_A \left|f_{(k)}(x) - \varphi(x)\right| dx \to 0.$$

In particular for $A = (-\infty, y]$, $G_{(k)}(y) \to \Phi(y)$ *and thus 5.16.* ■

Remark 5.17 *The last paragraph of the above proof is a special case of* ***Scheffé's Theorem****, named for a 1947 result of* ***Henry Scheffé*** *(1907 – 1977). This theorem states that pointwise convergence of density functions assures pointwise convergence of distributions, and thus weak convergence of distributions. It is generally true for density functions defined on arbitrary measure spaces and will be proved in book 6 using the more general version of Lebesgue's dominated convergence theorem of book 5. Otherwise, the general proof is identical with that above.*

In the special case where $k_n/n \to q$, the *qth* quantile of the uniform distribution for $0 < q < 1$, Smirnov's limit theorem can be stated in a simpler way and with a proof that is an application of proposition 9.16 of book 2.

Corollary 5.18 (Smirnov's Limit Theorem) *Let* $\{Y_{(k)}\}_{k=1}^n$ *be the order statistics from a continuous uniform distribution on* $[0, 1]$, *and* k_n *a sequence so that* $k_n/n \to q$ *for* $0 < q < 1$. *Define:*

$$Y'_{(k_n)} \equiv \frac{Y_{(k_n)} - q}{\sqrt{q(1-q)/n}},$$

and let $F_{(k_n)}$ *denote the distribution function of* $Y'_{(k_n)}$.

Then as $n \to \infty$,

$$F_{(k_n)} \Rightarrow \Phi, \tag{5.17}$$

where $\Phi(x)$ *is the distribution function of the standard normal.*

Proof. *Since* $k_n/n \to q$ *implies that both* $k_n \to \infty$ *and* $n - k_n \to \infty$ *as* $n \to \infty$, *the above proposition assures the convergence* $G_{(k)}(y) \Rightarrow \Phi(y)$ *where* $G_{(k)}(y)$ *is the distribution function of the normalized variable* $Y''_{(k_n)} \equiv \left[Y_{(k_n)} - b_n\right]/a_n$, *with* a_n *and* b_n *defined above in terms of* k_n. *By proposition 9.16 of book 2, if* c_n *and* d_n *are sequences that satisfy:*

$$c_n \to c, \text{ and } d_n \to d,$$

then

$$G_{(k)}(c_n y + d_n) \Rightarrow \Phi(cy + d),$$

or equivalently:

$$\Pr\left[Y''_{(k_n)} \leq c_n y + d_n\right] \to \Pr[Z \leq cy + d],$$

Define:

$$c_n = \frac{\sqrt{q(1-q)/n}}{a_n}, \qquad d_n = \frac{q - b_n}{a_n},$$

and note that

$$\Pr\left[Y''_{(k_n)} \leq c_n y + d_n\right] = \Pr\left[Y'_{(k_n)} \leq y\right] = F_{(k_n)}(y),$$

and thus the proof is complete by showing that $c = 1$ *and* $d = 0$.

But $k_n/n - q \to 0$ *by assumption, and thus* $(k_n - 1)/(n-1) \to q$ *and* $(n - k_n)/(n-1) \to 1 - q$ *obtain:*

$$c_n = \frac{\sqrt{q(1-q)/n}}{a_n} = \sqrt{\frac{n-1}{k_n - 1}\frac{n-1}{n-k_n}\frac{n-1}{n}q(1-q)} \to 1,$$

and

$$d_n = \frac{q - b_n}{a_n} = \left(q - \frac{k_n - 1}{n-1}\right)\left(\frac{(k_n - 1)(n - k_n)}{(n-1)^2}\right)^{-1/2} \to 0,$$

Thus $c = 1$, $d = 0$ *and* $F_{(k_n)} \Rightarrow \Phi$. ■

5.2.6 Limit Theorem on Quantiles of General Distribution Functions

In this section we will see that the above corollary to Smirnov's limit theorem can be generalized from the uniform distribution to a statement on the quantiles of order statistics of general distributions which have inverse functions F^{-1} which are differentiable at q. With the aid of the Δ-Method of proposition 8.40 of book 2, we now apply the Smirnov quantile result to such general distribution functions.

Proposition 5.19 *Let* $\{X_{(k)}\}_{k=1}^n$ *be the order statistics from a distribution function* F *which is continuous and strictly increasing in a neighborhood of* $F^{-1}(q)$ *for* $0 < q < 1$, *and differentiable at* $F^{-1}(q)$ *with* $F'\left(F^{-1}(q)\right) \neq 0$. *Given a sequence* $\{k_n\}$ *so that* $k_n/n \to q$ *as* $n \to \infty$, *define:*

$$X'_{(k_n)} \equiv \frac{X_{(k_n)} - F^{-1}(q)}{\left[\sqrt{q(1-q)/n}\right] \Big/ F'(F^{-1}(q))},$$

and let $F_{(k_n)}$ denote the distribution function of $X'_{(k_n)}$.

Then as $n \to \infty$,

$$F_{(k_n)} \Rightarrow \Phi, \tag{5.18}$$

where $\Phi(x)$ is the distribution function of the standard normal.

Proof. *Let $\{Y_{(k)}\}_{k=1}^n$ be the order statistics from a uniform distribution on $[0,1]$ and define $X_{(k)} = F^*(Y_{(k)})$ where F^* is the left continuous inverse of F. Then by proposition 4.8 of book 2 $\{X_{(k)}\}_{k=1}^n$ have distribution function F, and are by definition order statistics. If k_n is a sequence as defined above, then corollary 5.18 applies to obtain in the notation of weak convergence of random variables:*

$$c_n \left(Y_{(k_n)} - q\right) \Rightarrow Z,$$

where $c_n = 1/\sqrt{q(1-q)/n}$ and Z is a standard normal random variable. Because $c_n \to \infty$, the Δ-Method of proposition 8.40 of book 2 states that if $F^*(y)$ is differentiable at $y = q$, then:

$$c_n \left[F^*(Y_{(k_n)}) - F^*(q)\right] \Rightarrow (F^*)'(q) Z.$$

By remark 3.24 of book 2 $F^*(q) = F^{-1}(q)$, and since $F'\left(F^{-1}(q)\right) \neq 0$:

$$\left(F^{-1}\right)'(q) = 1/F'(F^{-1}(q)).$$

Thus with $F^*(Y_{(k_n)}) = X_{(k_n)}$, the Δ-Method obtains:

$$c_n \left[X_{(k_n)} - F^{-1}(q)\right] \Rightarrow Z/F'(F^{-1}(q)),$$

which is equivalent to 5.18. ■

Example 5.20 *For n large, $X'_{(k_n)}$ is approximately normally distributed. For example, $\left|X'_{(k_n)}\right| \leq 1.96$ provides an approximation to a 95% confidence interval recalling that the 95th quantile of the standard normal is $z_{0.95} = 1.96$. Defining $k_n = \lfloor qn \rfloor$, the greatest integer less than or equal to n, this then yields a confidence interval for the quantile of $F^{-1}(q)$ of X with unknown distribution function F using the estimate $X_{(k_n)}$:*

$$\left|F^{-1}(q) - X_{(k_n)}\right| \leq 1.96\sqrt{q(1-q)/n} \Big/ F'(F^{-1}(q)).$$

Of course, F' is also unknown generally, and thus $F'(F^{-1}(q)) \approx f(X_{(k_n)})$ would need to be estimated. Using the normal density, it is reasonable to assume that $\varphi(X_{(k_n)}) < f(X_{(k_n)})$ and thus by replacing this unknown term by $\varphi(X_{(k_n)})$ provides an upper bound for this inequality and a conservative confidence interval.

5.2.7 A Limit Theorem on Exponential Order Statistics

In this section we prove a limiting result on the average "gap" between high order statistics of the standard exponential distribution that in essence is a corollary of the central limit theorem and the Rényi representation theorem. This result will be important in the section below on extreme value theory to prove that the **Hill estimator** γ_H of the index for a distribution function $F \in D(G_\gamma)$ with $\gamma > 0$ converges in probability to γ as the sample size $n \to \infty$. Given a sample $\{X_i\}_{i=1}^n$ with distribution function F and $\{X_{(j)}\}_{j=1}^n$ the associated order statistics, the Hill estimator is based on the $k+1$ largest variates, $\{X_{(n-j)}\}_{j=0}^k$. The proof of the convergence $\gamma_H \to_P \gamma$ will require that $k \to \infty$ as $n \to \infty$, as assumed below, but also that $k/n \to 0$.

For the current section let $\{X_i\}_{i=1}^n$ be a sample of standard exponential variates, so $\lambda = 1$, with order statistics $\{X_{(j)}\}_{j=1}^n$. Define the random variable:

$$Y_{k,n} = \frac{1}{k} \sum\nolimits_{j=0}^{k-1} \left[X_{(n-j)} - X_{(n-k)} \right],$$

which equals the average of the k "gaps" between $X_{(n-k)}$ and higher order variates $X_{(n-j)}$ for $j = 0$ to $k-1$. The next results states that if properly normalized, $Y_{k,n}$ is asymptotically normal as $n \to \infty$ if also $k \to \infty$. An an example, if $k \approx qn$ for $0 < q < 1$, then $k \to \infty$ as $n \to \infty$, though the proof does not require that k and n increase proportionately.

Proposition 5.21 *Let $\{X_{(j)}\}_{j=1}^n$ be an ordered sample from a standard exponential distribution. With $Y_{k,n}$ defined above, define the normalized average variate $Y'_{k,n}$ by:*

$$Y'_{k,n} = \frac{Y_{k,n} - 1}{1/\sqrt{k}},$$

and let $F_{k,n}$ denote the distribution function of $Y'_{k,n}$.

Then as $n \to \infty$ and $k \to \infty$,

$$F_{k,n} \Rightarrow \Phi, \tag{5.19}$$

where $\Phi(x)$ is the distribution function of the standard normal.

Proof. *By 2.20 in the statement of corollary 2.25:*

$$X_{(j)} = \sum\nolimits_{i=1}^{j} \frac{E_i}{n - i + 1},$$

where $\{E_i\}_{i=1}^n$ *denotes a sample of standard exponential random variables. Hence for* $j < k$, *and using an index substitution* $l = k + i - n$:

$$\begin{aligned} X_{(n-j)} - X_{(n-k)} &= \sum\nolimits_{i=n-k+1}^{n-j} \frac{E_i}{n-i+1} \\ &= \sum\nolimits_{l=1}^{k-j} \frac{E_{l+n-k}}{k-l+1}. \end{aligned}$$

Now define the sample of standard exponential random variables $\{E'_l\}_{l=1}^k$ *by:*

$$\{E'_l\}_{l=1}^k \equiv \{E_{l+n-k}\}_{l=1}^k = \{E_i\}_{i=n-k+1}^n,$$

and thus:

$$X_{(n-j)} - X_{(n-k)} = \sum\nolimits_{l=1}^{k-j} \frac{E'_l}{k-l+1}.$$

Another application of the Rényi representation theorem yields that the collection:

$$\left\{X'_{(k-j)}\right\}_{j=0}^{k-1} \equiv \left\{\sum\nolimits_{l=1}^{k-j} \frac{E'_l}{k-l+1}\right\}_{j=0}^{k-1},$$

is the complete set of order statistics for the standard exponential based on a sample of size k.

Hence,

$$Y_{k,n} = \frac{1}{k}\sum\nolimits_{j=0}^{k-1} X'_{(k-j)} = \frac{1}{k}\sum\nolimits_{i=1}^{k} X'_i,$$

is the average of k *standard exponentials. Since* $E[Z_{k,n}] = 1$ *and* $Var[Z_{k,n}] = 1/k$, *the result follows from the central limit theorem above because* $k \to \infty$ *as* $n \to \infty$. ■

Remark 5.22 *The central limit theorem of proposition 5.14 above is adequate for this application because as assumed for that proof, the exponential distribution has a moment generating function as given in 3.59 with* $\alpha = 1$.

5.3 Laws of Large Numbers

In this section we study convergence results for a sum of independent random variables, $\{X_n\}_{n=1}^\infty$, defined on a probability space $(\mathcal{S}, \mathcal{E}, \lambda)$. We begin by briefly reviewing the notion of **tail event** from definition 5.34 of book 2 and recall that the convergence set for such a series is a tail event. Further, we restate the **Kolmogorov Zero-One law** of proposition 5.36, and hence conclude that the convergence sets of series have probability 0

or probability 1. We then study **laws of large numbers,** extending the results of section 5.1 of book 2. The **weak laws of large numbers** will generalize **Bernoulli's theorem** of proposition 5.3 and give specific conditions which ensure that the summation, $\sum_{n=1}^{\infty} X_n$, converges in probability, while the **strong laws of large numbers** generalize **Borel's theorem** of proposition 5.9 and provide conditions which ensure that this series converges with probability 1.

5.3.1 Tail Events and the Kolmogorov $0-1$ Law

Given a random variable X defined on a probability space $(\mathcal{S}, \mathcal{E}, \lambda)$, recall definition 3.43 of book 2 concerning $\sigma(X)$, the **sigma algebra generated by** X. This is defined as the smallest sigma algebra with respect to which X is measurable. Of course $\sigma(X) \subset \mathcal{E}$ for any such X by the measurability requirement in the definition of random variable. By exercise 3.44 of book 2:

$$\sigma(X) \equiv \{X^{-1}(A) | A \in \mathcal{B}(\mathbb{R})\}.$$

Similarly, if $\{X_n\}$ is a finite or infinite collection of random variables, $\sigma(X_1, X_2, ...)$, **the sigma algebra generated by** $\{X_n\}$ is the smallest sigma algebra with respect to which each X_n is measurable.

As in definition 3.47 of book 2, random variables $\{X_n\}_{n=1}^{\infty}$ defined on a probability space $(\mathcal{S}, \mathcal{E}, \lambda)$ are said to be **independent random variables** if $\{\sigma(X_n)\}_{n=1}^{\infty}$ are independent sigma algebras in the sense of definition 1.15 of book 2. That is, given any **finite** index subcollection $J = (j(1), j(2), ..., j(n))$, and $\{B_{j(i)}\}_{i=1}^{n}$ with $B_{j(i)} \in \sigma(X_{j(i)})$:

$$\mu\left(\bigcap_{i=1}^{n} B_{j(i)}\right) = \prod_{i=1}^{n} \mu\left(B_{j(i)}\right).$$

We now recall definition the **tail sigma algebra,** $\mathcal{T}$, associated with an arbitrary collection of random variables, $\{X_n\}_{n=1}^{\infty}$.

Definition 5.23 *Given a probability space* $(\mathcal{S},\mathcal{E},\lambda)$ *and countable collection of random variables* $\{X_n\}_{n=1}^{\infty}$*, the* ***tail sigma algebra associated with*** $\{X_n\}_{n=1}^{\infty}$***,*** *denoted* $\mathcal{T} \equiv \mathcal{T}\left(\{X_n\}_{n=1}^{\infty}\right)$*, is defined by:*

$$\mathcal{T} = \bigcap_{n=1}^{\infty} \sigma(X_n, X_{n+1}, X_{n+2}, ...), \tag{5.20}$$

where $\sigma(X_n, X_{n+1}, X_{n+2}, ...)$ *is the sigma algebra generated by* $\{X_j\}_{j=n}^{\infty}$*. A* ***tail event*** *is any set* $A \in \mathcal{T}$*.*

Remark 5.24 *Of course, $\mathcal{T} \subset \mathcal{E}$ since $\sigma(X_n) \subset \mathcal{E}$ for all n.*

Example 5.25 (Convergence Sets) *Given a countable collection of random variables $\{X_n\}_{n=1}^{\infty}$, define the* ***convergence set*** *by:*

$$A = \left\{ \sum_{n=1}^{\infty} X_n(s) \text{ converges} \right\},$$

where we suppress the usual $s \in \mathcal{S}$ condition in this set notation. Then $A \in \mathcal{T} \equiv \mathcal{T}\left(\{X_n\}_{n=1}^{\infty}\right)$, and this is intuitively plausible since the convergence of a series does not depend on any finite number of terms. More formally, recall that a series of real numbers, $\sum_{n=1}^{\infty} a_n$, converges if and only if this series satisfies the ***Cauchy convergence criterion*** *named for* ***Augustin-Louis Cauchy*** *(1789 – 1857).*

Definition 5.26 *A series $\sum_{n=1}^{\infty} a_n$ satisfies the* ***Cauchy criterion*** *if given any $\epsilon > 0$ there is an N so that $\left|\sum_{j=m}^{n} a_j\right| < \epsilon$ for all $n, m \geq N$.*

We can then define the convergence set as follows, using rational ϵ :

$$A = \bigcap\nolimits_{\epsilon \in \mathbb{Q}} \bigcup\nolimits_{N} \bigcap\nolimits_{n \geq m \geq N} \left\{ \left| \sum\nolimits_{j=m}^{n} X_j(s) \right| < \epsilon \right\}. \tag{5.21}$$

Because $\left\{ \left|\sum_{j=m}^{n} X_j(s)\right| < \epsilon \right\} \subset \sigma(X_m, X_{m+1}, X_{m+2}, ...)$, it follows that $A \subset \sigma(X_m, X_{m+1}, X_{m+2}, ...)$ for all m and hence $A \in \mathcal{T}$.

Kolmogorov's zero-one law is named for **Andrey Kolmogorov** (1903 – 1987). It states that if $\{X_n\}_{n=1}^{\infty}$ are **independent random variables**, then for any $A \in \mathcal{T}$ the measure of A is predictable, but not precisely. The proof is found in book 2, proposition 5.36.

Proposition 5.27 (Kolmogorov's zero-one law) *Given a probability space $(\mathcal{S},\mathcal{E},\lambda)$ and independent random variables, $\{X_n\}_{n=1}^{\infty}$, then for any $A \in \mathcal{T} \equiv \mathcal{T}\left(\{X_n\}_{n=1}^{\infty}\right)$,*

$$\lambda(A) = 0 \text{ or } \lambda(A) = 1. \tag{5.22}$$

Corollary 5.28 (Kolmogorov's zero-one law) *Given a probability space $(\mathcal{S},\mathcal{E},\lambda)$ and independent random variables $\{X_n\}_{n=1}^{\infty}$, then:*

$$\lambda\left[\left\{ \sum\nolimits_{n=1}^{\infty} X_n(s) \text{ converges} \right\}\right] = 0 \text{ or } 1. \tag{5.23}$$

Proof. *This follows from example 5.25 and Kolmogorov's zero-one law.* ∎

Remark 5.29 *In the next two sections we continue the study of convergence of series. The weak laws of large numbers will provide additional information on the weaker notion of convergence in probability, while the strong laws of large numbers will provide additional information on when such series converge with probability one.*

5.3.2 Weak Laws of Large Numbers (WLLNs)

Weak laws of large numbers, often abbreviated WLLNs, identify conditions on a random variable sequence which ensure **converge in probability** to a specified random variable. Recall definition 5.11 of book 2, using the simpler set notation, for example:

$$\Pr(|Y_n - Y| \geq \epsilon) \equiv \lambda\{s \in \mathcal{S} | \, |Y_n - Y| \geq \epsilon\}.$$

Definition 5.30 *Given a probability space $(\mathcal{S}, \mathcal{E}, \lambda)$ and random variables Y and $\{Y_n\}_{n=1}^{\infty}$, we say that Y_n* ***converges to*** *Y* ***in probability,*** *denoted $Y_n \to_P Y$, if for every $\epsilon > 0$,*

$$\lim_{n \to \infty} \Pr(|Y_n - Y| \geq \epsilon) = 0. \tag{5.24}$$

Remark 5.31 *Since for all ϵ :*

$$\Pr(|Y_n - Y| < \epsilon) + \Pr(|Y_n - Y| \geq \epsilon) = 1,$$

$Y_n \to_P Y$ if and only if for every $\epsilon > 0$:

$$\lim_{n \to \infty} \Pr(|Y_n - Y| < \epsilon) = 1.$$

Weak laws are most often stated in the context of:

$$Y_n = S_n/n \equiv \sum\nolimits_{j=1}^{n} X_j/n,$$

with $\{X_j\}_{j=1}^{\infty}$ a sequence of independent random variables. When random variables are assumed identically distributed we can then address convergence in probability to $\mu \equiv E[X]$, though this assumption is not necessary and there are other versions of this result. The average of the first n random variables also denoted:

$$\bar{X}_n \equiv \sum\nolimits_{j=1}^{n} X_j/n.$$

We begin with the simplest version of a weak law and one with the simplest proof. It states that when the sequence $\{X_j\}_{j=1}^{\infty}$ is independent and identically distributed, with finite mean and variance, then the associated average sequence $\{\bar{X}_n\}$, converges in probability to the mean.

Proposition 5.32 (WLLN 1) *Let $\{X_j\}_{j=1}^{\infty}$ be a sequence of independent, identically distributed random variables with finite mean, μ, and variance, σ^2. Then*

$$\sum_{j=1}^{n} X_j/n \to_P \mu. \tag{5.25}$$

Proof. *Recall Chebyshev's inequality in 3.70, for which $Var\left(\sum_{j=1}^{n} X_j/n\right)$ is required. Because:*

$$Var\left(\sum_{j=1}^{n} X_j/n\right) = \sigma^2/n, \qquad E\left[\sum_{j=1}^{n} X_j/n\right] = \mu,$$

it follows that:

$$\Pr\left(\left|\sum_{j=1}^{n} X_j/n - \mu\right| \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2},$$

and 5.25 follows from 5.24. ■

Example 5.33 *A special case of this version of the weak law is **Bernoulli's theorem** presented in proposition 5.3 of book 2, which stated that the average of independent, identically distributed binomial random variables $\{X_j^B\}_{j=1}^{\infty}$, converges in probability to the binomial probability $p = E[X_j^B]$.*

Perhaps surprisingly, while the existence of σ^2 provides a very simple proof, the above proposition remains true with only the assumption of the existence of the first moment μ. But the proof is significantly harder, principally because of the need for the general Lebesgue dominated convergence theorem of book 5 to be applied to integrals used in 3.7. This theorem is needed for a technical result on "truncations" of random variables which will appear elementary, but resists direct validation. In cases where $E[X]$ as defined in 3.7 can be transformed to a Lebesgue integral as in 3.10, the Lebesgue dominated convergence theorem of proposition 2.61 of book 3 suffices, but this transformation itself requires the integration theory of book 5.

Lemma 5.34 (On Truncation) *Let X be a random variable on $(\mathcal{S}, \mathcal{E}, \lambda)$ and assume that $E[X] < \infty$. Given $n > 0$ define a random variable Y, **the truncation of X at** n, by:*

$$Y = \begin{cases} X, & |X| \leq n \\ 0, & |X| > n. \end{cases}$$

Then for any $\epsilon > 0$ there exists N so that for all $n \geq N$:

$$E\left[|X - Y|\right] < \epsilon,$$

where E is defined as in 3.7.

Proof. *Let Y_n denote the truncation of X at n and define $Z_n \equiv |X - Y_n|$. Then $Z_n \to 0$ pointwise on $\mathcal{S}$, and $|Z_n| \leq |X|$ with $|X|$ integrable by definition of $E[X] < \infty$. Thus by the general version of Lebesgue's dominated convergence theorem in book 5, $E\left[Z_n\right] \to 0$ and the result follows.* ■

Proposition 5.35 (WLLN 2) *Let $\{X_j\}_{j=1}^{\infty}$ be a sequence of independent, identically distributed random variables with finite mean, $\mu \equiv E[X]$. Then 5.25 is satisfied..*

Proof. *We prove that given $\epsilon > 0$, for any δ there exists N so that for $n \geq N$:*

$$\Pr\left[\left|\sum_{j=1}^{n} X_j/n - E[X]\right| \geq \epsilon\right] < \delta,$$

and we do this by truncation. To simplify notation let $\bar{X}_n \equiv \sum_{j=1}^{n} X_j/n$. Given $\lambda > 0$ to be specified below, by lemma 5.34 there exists $M = M(\lambda)$ so that $E\left[|X - Y|\right] < \lambda$ for any truncation Y of X at m if $m \geq M(\lambda)$. Let Y_j denote the corresponding truncation of X_j and $\bar{Y}_n \equiv \sum_{j=1}^{n} Y_j/n$. Then by the triangle inequality:

$$\left|\bar{X}_n - E[X]\right| \leq \left|\bar{X}_n - \bar{Y}_n\right| + \left|\bar{Y}_n - E[Y]\right| + \left|E[Y] - E[X]\right|,$$

and so:

$$\begin{aligned}\Pr\left[\left|\bar{X}_n - E[X]\right| \geq \epsilon\right] \;\leq\; & \Pr\left[\left|\bar{X}_n - \bar{Y}_n\right| \geq \epsilon/3\right] \\ & + \Pr\left[\left|\bar{Y}_n - E[Y]\right| \geq \epsilon/3\right] + \Pr\left[\left|E[Y] - E[X]\right| \geq \epsilon/3\right].\end{aligned}$$

This last statement follows by considering the defining sets in $\mathcal{S}$, noting that $A \subset \cup B_j$ follows from $\cap \widetilde{B}_j \subset \widetilde{A}$, and using subadditivity of measures.

We now prove that each of these probabilities on the right can be made small by making λ small.

1. *By the triangle inequality:*

$$E\left[\left|\bar{X}_n - \bar{Y}_n\right|\right] \leq \sum_{j=1}^{n} E\left[|X_j - Y_j|\right]/n = E\left[|X - Y|\right] < \lambda,$$

and thus by the Markov inequality in 3.76, for $n \geq M(\lambda)$:

$$\Pr\left[\left|\bar{X}_n - \bar{Y}_n\right| \geq \epsilon/3\right] \leq 3\lambda/\epsilon.$$

2. *For the second probability,* $\{Y_j\}$ *are independent because* $\{X_j\}$ *are independent (an application of proposition 3.56 of book 2), with finite mean and variance. The mean result follows because* $|Y_j| \leq |X_j|$, *while the variance result follows from* $\sigma^2 = \mu_2' + \mu^2$*:*

$$Var[Y_j] \leq E[Y_j^2] \leq ME[|Y_j|] < ME[|X|].$$

Thus $Var[\bar{Y}_n] \leq ME[|X|]/n$ *by independence. Then by 3.70 since* $E[Y] = E[\bar{Y}]$:

$$\Pr\left[\left|\bar{Y}_n - E[Y]\right| \geq \epsilon/3\right] \leq 9ME[|X|]/n\epsilon^2.$$

3. *Finally, by the triangle inequality:*

$$|E[Y] - E[X]| \leq E\left[|X - Y|\right] < \lambda,$$

and $\Pr\left[|E[Y] - E[X]| \geq \epsilon/3\right] = 0$ *if* $\lambda < \epsilon/3$.

Combining, if $\lambda < \epsilon/3$:

$$\Pr\left[\left|\bar{X}_n - E[X]\right| \geq \epsilon\right] \leq 3\lambda/\epsilon + 9ME[|X|]/n\epsilon^2.$$

Choosing $\lambda < \epsilon\delta/6$ *satisfies this initial constraint and makes* $3\lambda/\epsilon < \delta/2$ *for* $n \geq M(\lambda)$. *With this* λ *and* $M(\lambda)$ *the second term,* $9ME[|X|]/n\epsilon^2$, *can then be made less than* $\delta/2$ *for* $n > 18M(\lambda)E[|X|]/\delta\epsilon^2$, *so combining constraints on* n *completes the proof.* ■

For the final generalization of proposition 5.32, we again assume that second moments exist, and generalize to an independent sequence of random variables, $\{X_j\}_{j=1}^{\infty}$,with arbitrary distributions. But we will need to make an assumption on $E\left[\sum_{j=1}^n X_j\right]$ and $Var\left(\sum_{j=1}^n X_j\right)$ as functions of n. Denote:

$$m_n = \sum\nolimits_{j=1}^n \mu_j, \qquad s_n^2 = \sum\nolimits_{j=1}^n \sigma_j^2.$$

The assumption of identically distributed in the above **WLLN 1** proposition implies that $m_n/n = \mu$ and $s_n^2/n^2 = \sigma^2/n$. This is generalized next.

Proposition 5.36 (WLLN 3) *Let* $\{X_j\}_{j=1}^{\infty}$ *be a sequence of independent random variables with finite means,* μ_j, *and variances,* σ_j^2. *Then with the above notation, if* $m_n/n \to \mu$ *for some* μ, *and* $s_n^2/n^2 \to 0$, *then 5.25 is satisfied.*

Proof. *Again simplifying notation let* $\bar{X}_n \equiv \sum_{j=1}^n X_j/n$. *Then by the triangle inequality:*

$$\left|\bar{X}_n - \mu\right| \leq \left|\bar{X}_n - m_n/n\right| + |m_n/n - \mu|,$$

and so again by considering defining sets in $\mathcal{S}$*:*

$$\Pr\left(\left|\bar{X}_n - \mu\right| < \epsilon\right) \geq \Pr\left(\left|\bar{X}_n - m_n/n\right| < \epsilon - |m_n/n - \mu|\right).$$

By Chebyshev's inequality with $E\left[\bar{X}_n\right] = m_n/n$ *and* $Var\left[\bar{X}_n\right] = s_n^2/n^2$:

$$\Pr\left(\left|\bar{X}_n - m_n/n\right| < \epsilon - |m_n/n - \mu|\right) \geq 1 - \frac{s_n^2/n^2}{\left[\epsilon - |m_n/n - \mu|\right]^2}.$$

But $|m_n/n - \mu| \to 0$ *and* $s_n^2/n^2 \to 0$, *and thus as* $n \to \infty$,

$$\Pr\left(\left|\bar{X}_n - m_n/n\right| < \epsilon - |m_n/n - \mu|\right) \to 1,$$

and hence:

$$\Pr\left(\left|\bar{X}_n - \mu\right| < \epsilon\right) \to 1.$$

■

Remark 5.37 (Further Generalizations) *We can eliminate the assumption that* $m_n/n \to \mu$, *and then prove that:*

$$\sum\nolimits_{j=1}^n \left(X_j - \mu_j\right)/n \to_P 0.$$

Also, if it is assumed that $\mu_j = \mu$ *for all* j *then it is enough to assume that* $\sigma_n^2 \leq B$ *for all* n.

Similarly, we can replace the independence assumption in any version of the weak law by the assumption that correlations satisfy $\rho_{ij} \leq 0$ *for all* i, j, *since then* $Var\left[\sum_{j=1}^n X_j/n\right] \leq s_n^2/n^2$ *and so the proofs go through without change. Even in the nonnegative correlation case, a positive result is possible if we assume for example that for some* $0 < r < 1$, $\rho_{ij} \leq r^{|i-j|}$ *and* $\sigma_n^2 \leq B$ *for all* n.

Exercise 5.38 *Let* $\{X_j\}_{j=1}^\infty$ *be a sequence of random variables with the same finite mean,* μ, *and variances* σ_j^2 *with* $\sigma_j^2 \leq B$ *for all* n. *Show that:*

1. *If independent, then 5.25 remains true.*
2. *If for some* $0 < r < 1$, $\rho_{ij} \leq r^{|i-j|}$, *then 5.25 remains true. Note: There is no independence assumption here.*

5.3.3 Strong Laws of Large Numbers (SLLNs)

Strong laws of large numbers, often abbreviated SLLNs, identify conditions on a random variable sequence which ensure **converge almost everywhere** to a specified random variable. This notion is also called **convergence almost surely** and **convergence with probability** 1. Recall definition 5.15 of book 2:

Definition 5.39 *Given a probability space* $(\mathcal{S}, \mathcal{E}, \lambda)$ *and random variables* Y *and* $\{Y_n\}_{n=1}^{\infty}$, *we say that* Y_n ***converges to*** Y ***almost everywhere,*** *denoted* $Y_n \rightarrow_{a.e.} Y$, *if:*

$$\lambda\left[\left\{\lim_{n\rightarrow\infty} \sum\nolimits_{j=1}^{n} Y_n = Y\right\}\right] = 1. \tag{5.26}$$

Notation 5.40 *Correspondingly,* ***convergence almost surely*** *is denoted* $Y_n \rightarrow_{a.s.} Y$, *and* ***convergence with probability*** 1 *denoted* $Y_n \rightarrow_1 Y$.

Strong laws, as was the case for weak laws, are most often stated in the context of:

$$Y_n = S_n/n \equiv \sum\nolimits_{j=1}^{n} X_j/n,$$

with $\{X_j\}_{j=1}^{\infty}$ a sequence of independent random variables. When identically distributed we can then address convergence almost everywhere to $\mu \equiv E[X]$, though this assumption is not necessary and there are other versions of this result. The average of the first n random variables also denoted $\bar{X}_n$ as noted above.

Weak and strong laws can be related as follows. Let $\{X_j\}_{j=1}^{\infty}$ be a sequence of independent, identically distributed random variables defined on a probability space $(\mathcal{S}, \mathcal{E}, \lambda)$. Define the set $A_n(\epsilon)$ as in 5.2 of book 2, suppressing the notation $s \in \mathcal{S}$:

$$A_n(\epsilon) \equiv \left\{\left|\sum\nolimits_{j=1}^{n} X_j/n - \mu\right| \geq \epsilon\right\}.$$

In the first two versions of the weak law above it was proved that for such independent and identically distributed random variables and any $\epsilon > 0$:

$$\lambda\left[A_n(\epsilon)\right] \rightarrow 0 \text{ as } n \rightarrow \infty.$$

But as noted in book 2, such sets are not necessarily nested and it is possible to have both $A_{n+1}(\epsilon) - A_n(\epsilon) \neq \emptyset$ and $A_n(\epsilon) - A_{n+1}(\epsilon) \neq \emptyset$. Hence there is no apparent limit event in $\mathcal{S}$ associated with weak laws.

For the strong laws, the goal is to determine the measure of the set C_S on which this series converges as in 5.7 of book 2:

$$C_S \equiv \left\{\sum_{j=1}^{n} X_j/n \to \mu\right\} = \left\{\sum_{j=1}^{n} (X_j - \mu)/n \to 0\right\}.$$

This convergence set is a tail event as noted in example 5.25 above and hence for independent random variable has measure 0 or 1 by Kolmogorov's zero-one law. The strong laws identify conditions under which the measure of this convergence set is 1.

Note that this convergence set is closely related to the $A_n(\epsilon)$-sets, and generalizing 5.10 of proposition 5.8 of book 2:

Proposition 5.41 *With the notation above:*

$$\lambda\{C_S\} = 1 \iff \lambda[\limsup_n A_n(\epsilon)] = 0 \text{ for all } \epsilon > 0. \tag{5.27}$$

Proof. *The essence of the idea is similar to that used in 5.21 to define a convergence set:*

$$\begin{aligned} C_S &= \bigcap_{j=1}^{\infty} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \left\{\left|\sum_{j=1}^{n} X_j/n - \mu\right| < 1/j\right\} \\ &= \bigcap_{j=1}^{\infty} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \widetilde{A_n}(1/j). \end{aligned}$$

Recalling the definition of limit superior in definition 2.1 of book 2, this can be rewritten:

$$\widetilde{C}_S = \bigcup_{j=1}^{\infty} \limsup A_n(1/j).$$

Thus if $\lambda[C_S] = 1$, then $\lambda[\limsup_n A_n(\epsilon)] = 0$ for all $\epsilon = 1/j$ and therefore all $\epsilon > 0$ since $\{\limsup A_n(\epsilon)\}$ increases as ϵ decreases. On the other hand, by this nested property and continuity from below of the measure λ,

$$\lambda\left[\widetilde{C}_S\right] = \lim_{j\to\infty} \lambda[\limsup_n A_n(1/j)],$$

and so $\lambda[\limsup_n A_n(1/j)] = 0$ for all j implies that $\lambda[C_S] = 1$. ■

Remark 5.42 (SLLN Proof Strategy) *To prove a strong law we must by the above proposition show that for any $\epsilon > 0$,*

$$\lambda[\limsup_n A_n(\epsilon)] = 0.$$

A powerful method for proving this statement is to use the first ***Borel-Cantelli theorem*** *in proposition 2.6 of book 2, which stated:*

$$\sum_{n=1}^{\infty} \lambda(A_n) < \infty \Longrightarrow \lambda(\limsup A_n) = 0.$$

Consequently, if we show that for any $\epsilon > 0$ *the series* $\sum_{n=1}^{\infty} \lambda(A_n(\epsilon)) = \sum_{n=1}^{\infty} \Pr(A_n(\epsilon))$ *is convergent, then* $\lambda(\limsup A_n)(\epsilon) = 0$ *for all* $\epsilon > 0$ *and the convergence set* C_S *then has probability* 1.

We develop two strong laws, the first will have excess assumptions to more easily demonstrate the application of Borel-Cantelli. Much like the assumption of the existence of σ^2 in the first version of the weak law, we will require the existence of fourth moments to facilitate the application of Chebyshev's inequality.

Proposition 5.43 (SLLN 1) *Let* $\{X_j\}_{j=1}^{\infty}$ *be a sequence of independent, identically distributed random variables with finite fourth central moment,* μ_4. *Then with* μ *denoting the common mean,* $\sum_{j=1}^{n} X_j/n$ *converges to* μ *with probability* 1 :

$$\Pr\left\{\sum\nolimits_{j=1}^{n} X_j/n \to \mu\right\} = 1. \tag{5.28}$$

In other words, as $n \to \infty$:

$$\sum\nolimits_{j=1}^{n} X_j/n \to_{a.s.} \mu.$$

Proof. *The assumed existence of a fourth moment assures the existence of all lower order moments as noted in the introduction to section 3.2.4, and so* μ *is well defined. Also, with*

$$S_n \equiv \sum\nolimits_{j=1}^{n} (X_j - \mu),$$

a calculation produces:

$$\begin{aligned} E\left[S_n^4\right] &= E\left[\left[\sum\nolimits_{j=1}^{n} (X_j - \mu)\right]^4\right] \\ &= \sum\nolimits_{j=1}^{n} E(X_j - \mu)^4 + \binom{4}{2} \sum\nolimits_{i=1}^{n} \sum\nolimits_{j=i+1}^{n} E(X_j - \mu)^2 E(X_i - \mu)^2. \end{aligned}$$

This follows because all terms with odd powers have zero expectation by independence. Now by Lyapunov's inequality in 3.90,

$$E(X_j - \mu)^2 E(X_i - \mu)^2 \leq E(X_j - \mu)^4,$$

and hence

$$E\left[S_n^4\right] \leq \left(3n^2 - 2n\right)\mu_4.$$

Finally by Chebyshev's inequality,

$$\begin{aligned}\Pr\left[A_n(\epsilon)\right] &\equiv \Pr\left\{S_n \geq n\epsilon\right\} \\ &\leq \left(3n^2 - 2n\right)\mu_4/\left(n\epsilon\right)^4.\end{aligned}$$

This implies that $\sum_{n=1}^{\infty}\Pr\left[A_n(\epsilon)\right]$ *is a convergent series for any* $\epsilon > 0$ *and the result now follows by the first Borel-Cantelli theorem.* ■

Example 5.44 *A special case of this version of the strong law was* ***Borel's theorem*** *presented in proposition 5.9 of book 2, which stated that the average of independent, identically distributed binomial random variables* $\{X_j^B\}_{j=1}^{\infty}$, *converged with probability* 1 *to the binomial probability* $p = E[X_j^B]$. *All moments are finite for the binomial, so the above assumption that* $\mu_4 < \infty$ *is readily satisfied.*

Like the weak law, it is also true that the strong law is valid for independent and identically distributed random variables assuming only the existence of a first moment. We provide a somewhat weaker version of this result that requires a second moment. But this version is also somewhat more general as it is also applicable to independent random variables which need not be identically distributed. For its proof we will use Kolmogorov's inequality in 3.81, and modify the definition of $A_n(\epsilon)$ to a related set $A'_n(\epsilon)$.

Proposition 5.45 (SLLN 2) *Let* $\{X_j\}_{j=1}^{\infty}$ *be a sequence of mutually independent random variables with means* $\left\{\mu_j\right\}_{j=1}^{\infty}$ *and variances* $\left\{\sigma_j^2\right\}_{j=1}^{\infty}$ *with* $\sum_{j=1}^{\infty}\sigma_j^2/j^2 < \infty$. *Then:*

$$\Pr\left\{\sum\nolimits_{j=1}^{n}\left(X_j - \mu_j\right)/n \to 0\right\} = 1, \tag{5.29}$$

or in other words, as $n \to \infty$:

$$\sum\nolimits_{j=1}^{n}\left(X_j - \mu_j\right)/n \to_{a.s.} 0.$$

Proof. *Given* $\epsilon > 0$ *and* $Y_j \equiv X_j - \mu_j$, *define the event* $A'_n(\epsilon)$:

$$A'_n(\epsilon) = \left\{\left|\sum\nolimits_{j=1}^{k} Y_j/k\right| \geq \epsilon \ \text{ for at least one } k \text{ with } 2^{n-1} < k \leq 2^n\right\}. \tag{5.30}$$

The first step is to prove that

$$\sum\nolimits_{n=1}^{\infty}\Pr[A'_n(\epsilon)] < \infty,$$

and to this end note that $\Pr[A'_n(\epsilon)] \leq \Pr[A''_n(\epsilon)]$ *with:*

$$A''_n(\epsilon) = \left\{ \max_{2^{n-1}<k\leq 2^n} \left| \sum_{j=1}^{k} Y_j \right| \geq 2^{n-1}\epsilon \right\}.$$

This is because if $\left|\sum_{j=1}^{k} Y_j/k\right| \geq \epsilon$ *for any* k *with* $2^{n-1} < k \leq 2^n$, *then* $\max_{2^{n-1}<k\leq 2^n} \left|\sum_{j=1}^{k} Y_j\right| \geq 2^{n-1}\epsilon$ *by definition. By Kolmogorov's inequality in 3.81, the probability of this event is bounded:*

$$\Pr[A''_n(\epsilon)] \leq \frac{1}{\epsilon^2 2^{2n-2}} \sum_{j=1}^{2^n} \sigma_j^2.$$

Hence since $\Pr[A'_n(\epsilon)] \leq \Pr[A''_n(\epsilon)]$:

$$\begin{aligned} \sum_{n=1}^{\infty} \Pr[A'_n(\epsilon)] &\leq \frac{4}{\epsilon^2} \sum_{n=1}^{\infty} \frac{1}{2^{2n}} \sum_{j=1}^{2^n} \sigma_j^2 \\ &= \frac{4}{\epsilon^2} \sum_{j=1}^{\infty} \sigma_j^2 \sum_{2n\geq j}^{\infty} \frac{1}{2^{2n}} \\ &\leq \frac{4}{\epsilon^2} \sum_{j=1}^{\infty} \sigma_j^2/j^2. \end{aligned}$$

In this calculation was used that for $j \geq 4$:

$$\sum_{2n\geq j}^{\infty} 1/2^{2n} \leq \sum_{k=j}^{\infty} 1/2^k = 1/2^{j-1} \leq 2/j^2.$$

Hence, $\sum_{n=1}^{\infty} \Pr[A'_n(\epsilon)] < \infty$.

Thus by the first Borel-Cantelli theorem the convergence of this series implies that for all $\epsilon > 0$:

$$\Pr[\limsup_n A'_n(\epsilon)] = 0,$$

or

$$\Pr\left[\bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} A'_n(\epsilon)\right] = 0.$$

But $\bigcup_{j=2^{n-1}+1}^{2^n} A_j(\epsilon) = A'_n(\epsilon)$, *so this implies that*:

$$\Pr\left[\bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} A_n(\epsilon)\right] = 0,$$

and so $\Pr[\limsup_n A_n(\epsilon)] = 0$. *Proposition 5.41 then obtains 5.29.* ■

Remark 5.46 **1**. *If* $\sum_{j=1}^{n} \mu_j/n \to \mu$ *or if* $\mu_j = \mu$ *for all* j, *then 5.29 can be restated as 5.28.*

2. *The assumption in the Strong Law, that* $\sum_{j=1}^{\infty} \sigma_j^2/j^2 < \infty$, *is apparently an assumption about the growth rate of* σ_j^2 *as* $j \to \infty$. *For example, if* $\sigma_j^2 = \sigma^2$, *the assumption of no growth, then since* $\sum_{j=1}^{\infty} 1/j^2 < \infty$ *the Strong Law applies. This is the case with independent, identically distributed random variables. On the other hand, if* $\sigma_j^2 = j\sigma^2$, *whereby the standard deviation grows like* $\sqrt{j}$, *the Strong Law does not apply since* $\sum_{j=1}^{N} 1/j \to \infty$ *with* N. *Consequently, linear variance growth, or equivalently, square root growth in standard deviation, is just a bit too fast for the Strong Law to apply, but if* $\sigma_j^2 = j^a\sigma^2$ *for any* $a < 1$, *the Strong Law applies.*

5.3.4 Limit Theorem on Quantiles

In this section we present a result that is a corollary to the strong law of large numbers and addresses a question related to the probability 1 convergence of a sequence of high quantile order statistics. Let F be a distribution function, $\{X_i\}_{i=1}^n$ a sample with associated order statistics, $\{X_{(j)}\}_{j=1}^n$, where as usual, $X_{(j)} \leq X_{(j+1)}$. The quantile of the order statistic $X_{(j)}$ is by definition $q_j \equiv j/n$. In section 5.2.6 the corollary to Smirnov's limit theorem addressed the case where F^{-1} was assumed to be differentiable at q for $0 < q < 1$, and k_n a sequence with $k_n/\mathrm{n} \to q$ as $n \to \infty$. Then with $X_{(k_n)}$ properly normalized:

$$X'_{(k_n)} \equiv \frac{X_{(k_n)} - q}{\sqrt{q(1-q)/n}},$$

the distribution function $F_{(k_n)}$ was seen to be asymptotically standard normal. In this section we address the case where $k_n/\mathrm{n} \to 1$, and address this case in the context of the strong law. That is, we seek to prove a limit result with probability 1.

Remark 5.47 *In the context of the notational conventions of extreme value theory below and the development of the Hill estimator, for which* $k_n \equiv n - k'_n$, *the assumption of this section is equivalent to the assumption there that* $k'_n/n \to 0$.

Recall as defined in 9.6 in book 2, given a distribution function F,

$$x^* \equiv \inf\{x | F(x) = 1\}, \qquad x^* \equiv \infty \text{ if } F(x) < 1 \text{ for all } x. \tag{5.31}$$

The goal of this section is to prove the following proposition. It will be applied in the development of the Hill estimator where $F(x) = 1 - 1/x$ for $x \geq 1$, and hence $x^* = \infty$.

Proposition 5.48 *Let $\{X_j\}_{j=1}^{\infty}$ be a sequence of independent, identically distributed random variables on a probability space $(\mathcal{S},\mathcal{E},\lambda)$ with a continuous distribution function F. For each n let $\{X_{(j)}\}_{j=1}^{n}$ be the order statistics associated with $\{X_j\}_{j=1}^{n}$, and $X_{(k_n)}$ a given variate from this collection. If $k_n/n \rightarrow 1$, then $X_{(k_n)} \rightarrow x^*$ with probability 1 :*

$$X_{(k_n)} \rightarrow_{a.s.} x^*. \tag{5.32}$$

Proof. *For $r < x^*$, define the random variable $Z_i = \chi_{(r,\infty)}(X_i)$. Then Z_i is binomially distributed with $p \equiv \Pr[1] = 1 - F(r)$, and thus $E[Z_i] = 1 - F(r)$ and $Var[Z_i] = F(r)(1 - F(r))$. By version SLLN2 of the strong law, as $n \rightarrow \infty$:*

$$\frac{1}{n}\sum_{i=1}^{n} Z_i \rightarrow_{a.s.} 1 - F(r).$$

Because F is continuous and $F(r) \rightarrow 1$ as $r \rightarrow x^*$, for any $r < x^*$ it follows that $\frac{1}{n}\sum_{i=1}^{n} Z_i$ converges with probability 1 to a value that exceeds 0.

We now argue by contradiction. By 5.27 of proposition 5.41 above, if $x^* < \infty$, 5.32 is true if and only if for any $\epsilon > 0$,

$$\lambda\left[\limsup_n\{\left|X_{(k_n)} - x^*\right| \geq \epsilon\}\right] = 0.$$

This is equivalent to the statement that 5.32 is true if and only if for any $r < x^*$,

$$\lambda\left[\limsup_n\{X_{(k_n)} \leq r\}\right] = 0. \tag{(*)}$$

We prove $(*)$ by contradiction, and show that this obtains the desired result even if $x^* = \infty$.

To develop the contradiction, assume that for some $r < x^*$ that

$$\lambda\left[\limsup_n\{X_{(k_n)} \leq r\}\right] > 0.$$

Then for any $s \in \limsup_n\{X_{(k_n)} \leq r\}$ there are infinitely many values of n for which:

$$\frac{n - k_n}{n} = \frac{1}{n}\sum_{i=1}^{n} \chi_{(X_{(k_n)},\infty)}(X_i) > \frac{1}{n}\sum_{i=1}^{n} \chi_{(r,\infty)}(X_i).$$

However, the assumption that $\frac{k_n}{n} \rightarrow 1$ obtains that the right hand summation converges to 0, and this contradicts the above derivation that this summation converges to $1 - F(r) > 0$. Thus $(*)$ is valid for all $r < x^*$ and hence if $\{r_j\}$ is an enumeration of rationals with $r_j \rightarrow x^*$, where $x^* = \infty$ is allowed:

$$\lambda\left[\bigcup_j \limsup_n\{X_{(k_n)} \leq r_j\}\right] = 0.$$

Equivalently:

$$\lambda\left[\bigcap_j \liminf_n\{X_{(k_n)} > r_j\}\right] = 1,$$

and the proof of 5.32 is complete. ■

Corollary 5.49 *Let $\{X_j\}_{j=1}^{\infty}$ be a sequence of independent, identically distributed random variables on a probability space $(\mathcal{S},\mathcal{E},\lambda)$ with a continuous distribution function F. For each n let $\{X_{(j)}\}_{j=1}^{n}$ be the order statistics associated with $\{X_j\}_{j=1}^{n}$, and $X_{(k_n)}$ a given variate from this collection. If $k_n/n \to 0$, then $X_{(k_n)} \to x_*$ with probability 1 :*

$$X_{(k_n)} \to_{a.s.} x_*, \tag{5.33}$$

where $x_* = \sup\{x|F(x) = 0\}$, with $x_* \equiv -\infty$ if $F(x) > 0$ for all x.
Proof. *Left as an exercise. Hint: Follow the proof of proposition 5.48 but let $Z_i^r = \chi_{(-\infty,r]}(X_i)$ for $r > x_*$. Now prove by contradiction that for all such r, $\lambda\left[\limsup_n\{X_{(k_n)}(s) \geq r\}\right] = 0$.* ■

5.4 Convergence of Empirical Distribution Functions

In chapter 4 was discussed the problem of generating samples from a given distribution function F. In contrast to this exercise one may be confronted with observations which are deemed to be an independent collection of realizations of some event, and the question arises as to the identification of the underlying distribution function, or at least some of its properties. For example, such observations could be the IQs of voters, or non-voters, or the weekly total rainfall in a given rain forest, as well as examples found everywhere in finance. In the investment markets one finds the daily and other period returns of a variety of investment classes or individual securities, as well as data on financial variables such as inflation rates, interest rates on various fixed income securities, default rates, foreign currency exchange rates, commodity prices, and so forth. In the insurance markets one observes mortality rates by age, as well as various morbidity rates related to disability, hospitalization, and disability, or claims rates on various categories of automobile, homeowners and commercial insurance policies.

Of course any such quest to identify an underlying distribution function or some of its properties is based on the necessary assumption that the given

data in fact represent observations of a random variable with distribution function that is more or less stable, or **stationary**, through time. This is essentially a philosophical issue and one usually avoided by a reliance on one's intuitive understanding of the process underlying the given data, rather than a formal analysis which seeks to prove that such a distribution function must exist.

For example, if one looked at a series of observations of the month-end value of a given equity index over the last 25 years, it would hardly seem logical to define this as a random variable and then attempt to identify the associated distribution function. With the exception of market corrections, almost everyone's expectation is that over time the value of this index will continue to increase if for no other reason, inflation, but also as economists would argue, productivity. So while we can formally identify a "distribution function" for the given historical data, it is not clear that such an effort will produce something of value, if by "of value" we mean, providing some predictive insights for the future.

On the other hand, if we instead convert this price series into a monthly return series, then there would seem to be more of an argument that the distribution of returns over time had some stability, and hence some predictability, despite the fact that such observations can potentially provide very different insights when grouped into subperiods. This observation forces the analyst to choose between several competing interpretations:

1. There is no single underlying distribution function for the given data because, for example, different periods have different distributions and the timing and magnitude of the change between distributions is unpredictable.

2. There is an underlying distribution function but no single period reveals all of its qualities, and more data over longer periods is needed to reveal the ultimate truths.

3. There was an underlying distribution function, but due to a significant event, the distribution in the future can be expected to be different.

Unfortunately, there is no universally accepted approach to resolving which of these interpretations is correct in a given situation, or if indeed there are other interpretations. Ultimately, such a data analysis and the assumptions made to justify this analysis are part of the quantitative analyst's model building process, within which many other assumptions may also be

made. Are the assumptions valid? Is the model correct? What questions can the model answer?

Perhaps the best summary comment on this matter is one attributed to the statistician **George E. P. Box** (1919 – 2013), and while there are many versions of his dictum, a commonly cited version is the quote:

"All models are wrong, but some are useful."

5.4.1 Definition and Basic Properties

Let $\{X_j\}_{j=1}^n$ denote a collection of observations, or a **realization,** from a **sample** of a random variable X defined on a probability space $(\mathcal{S}, \mathcal{E}, \lambda)$, and with an unknown but assumed-to-exist distribution function F. As detailed in this chapter's introduction, $\{X_j\}_{j=1}^n$ can be thought of as the first n-components of a point in the infinite dimensional space $(\mathbb{R}^{\mathbb{N}}, \sigma(\mathbb{R}^{\mathbb{N}}), \mu_{\mathbb{N}})$, but importantly in this application it is assumed that all X_j are **identically distributed.**

Note that the "sample" $\{X_j\}_{j=1}^n$ will have two interpretations in this section. First, consistent with chapter 4 these are independent, identically distributed random variables defined on a probability space such as $(\mathbb{R}^{\mathbb{N}}, \sigma(\mathbb{R}^{\mathbb{N}}), \mu_{\mathbb{N}})$. Second, when we are working with a realization or a given sample of these variates we are actually working with the numerical values $\{X_j(s)\}_{j=1}^n$ for some fixed $s \in \mathbb{R}^{\mathbb{N}}$. Thus the premise of this section is that we are given such a numerical sample, but in order to evaluate properties of the empirical distribution function developed it is necessary to take a broader view, and that is, that the empirical distribution function is defined on $\{X_j(s)\}_{j=1}^n$ for all $s \in \mathbb{R}^{\mathbb{N}}$. So we will denote the given "sample" by $\{X_j\}_{j=1}^n$, largely dropping the "(s)," where this notation will represent both the random variables, as well as the numerical sample, with the interpretation obtained by the given context.

Given a numerical sample, one approach to visualizing F learned in grade school is through the construction of a **histogram** which is intended to shed some light on the underlying **density function**, f, which is thereby also assumed to exist. In this construction one assigns a probability of $1/n$ to each observation, and this is formally justified in proposition 4.9 of book 2, and summarized in the introduction to chapter 4. Specifically, if F is a given distribution function such a sample could be produced by $\{F^*(Y_j)\}_{j=1}^n$, where $\{Y_j\}_{j=1}^n$ is a numerical sample from the uniform continuous distribution on $[0, 1]$ and F^* denotes the left continuous inverse of F. Since each such Y_j is

uniformly distributed in the sense that $\Pr\{Y_j \in [c,d]\} = d-c$ for any interval $[c,d] \subset [0,1]$, it follows that the value of each variate $F^*(Y_j)$ is equally likely, and thus one logically assigns a probability of $\frac{1}{n}$ to each such value to achieve a true density function for the sample.

A graphical depiction of this data with each observation given probability $1/n$ then provides some visual clues on the density function f. Specifically, if $\{X_{(j)}\}_{j=1}^n$ are the associated order statistics, then for each j,

$$\int_{X_{(j)}+a_j}^{X_{(j)}+b_j} f(x)dx = 1/n,$$

where $\{[a_j, b_j]\}$ are disjoint intervals with union equal to the assumed range of X. In this representation, f is assumed constant on each interval, and specifically, $f(x) = 1/[n(b_j - a_j)]$ on $[a_j, b_j]$. Thus a histogram is not a discrete probability density function but a piecewise continuous approximation to the unknown density $f(x)$. The goal of this construction is to provide some clues about the "shape" of f. By requiring $a_j = b_{j-1}$, which reflects the assumption that the domain of f is connected, the various intervals will have different lengths and this uniform probability assumption will get translated into the various values for $f(x)$.

As an alternative approach to the density function f, it is also common to group sample variates into **bins,** or intervals of equal length. Specifically, we define disjoint intervals or bins $\{B_k\}_{k=1}^N$, with each interval having the same length, and $\cup_{k=1}^N B_k$ the assumed connected range of X. We then graph $f \equiv n_k/n$ as constant over each bin B_k, where n_k equals the number of variates in bin B_k. Since these rectangles have a common base, the magnitudes of the various n_k/n are observed as the relative heights of the rectangles that define this approximating step function. While there is no unique approach to the determination of the best binning structure, there have been many methods developed based on various assumptions on the underlying density f.

A third alternative approach which we pursue here, is to work with the so-called **empirical distribution function,** defined as follows.

Definition 5.50 *Given a numerical sample $\{X_j\}_{j=1}^n$ of a random variable X defined on a probability space $(\mathcal{S}, \mathcal{E}, \lambda)$, the associated* ***empirical distribution function*** $F_n(x) \equiv F_n(x|\ \{X_j\})$ *is defined by:*

$$F_n(x) = \frac{1}{n} \sum\nolimits_{j=1}^{n} \chi_{(-\infty,x]}(X_j). \tag{5.34}$$

Given $\{X_j\}_{j=1}^n$, we can look at $F_n(x)$ from at least two perspectives based on comments above:

1. For fixed numerical sample $\{X_j\}_{j=1}^n$, $F_n(x) \equiv F_n(x|\ \{X_j\})$ is indeed a distribution function in that it is increasing and right-continuous, and when defined as limits: $F_n(-\infty) = 0$ and $F_n(\infty) = 1$. In fact, $F_n(x)$ is continuous as a function of x except at $\{X_j\}_{j=1}^n$ with:

$$F_n(X_j) = F_n(X_j^-) + \frac{1}{n},$$

where $F_n(X_j^-)$ denotes the left limit of $F_n(x)$ at X_j.

2. For fixed x and $\{X_j\}_{j=1}^n$ interpreted as random variables, then $\{Y_j\}_{j=1}^n$ defined by $Y_j \equiv \chi_{(-\infty,x]}(X_j)$ are also **independent, standard binomial** variables defined on the sample space of the introduction:

$$Y_j : (\mathbb{R}^{\mathbb{N}}, \mathcal{B}(\mathbb{R}^{\mathbb{N}}), \mu_{\mathbb{N}}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}), m),$$

with m Lebesgue measure. Here Y_j is defined in terms of the jth component of $s \equiv (X_1, X_2, ...) \in \mathbb{R}^{\mathbb{N}}$. Thus $Y_j = 1$ if $X_j \leq x$ and $Y_j = 0$ otherwise.

That each $Y_j(s)$ is a random variable is seen as follows. Given $A \in \mathcal{B}(\mathbb{R})$, the Borel sigma algebra, then $Y_j^{-1}(A) \subset \mathbb{R}^{\mathbb{N}}$ is given by:

$$Y_j^{-1}(A) = \begin{cases} \emptyset, & 0, 1 \notin A, \\ \{s|X_j \in (x, \infty)\}, & 0 \in A,\ 1 \notin A, \\ \{s|X_j \in (-\infty, x]\}, & 0 \notin A,\ 1 \in A, \\ \mathbb{R}^{\mathbb{N}}, & 0, 1 \notin A. \end{cases}$$

Further, since $\{X_j\}_{j=1}^n$ are identically distributed with distribution function F, recalling 5.2 and 5.1:

$$\lambda_{\mathbb{N}}\left[Y_j^{-1}(A)\right] = \begin{cases} 0, & 0, 1 \notin A, \\ 1 - F(x), & 0 \in A,\ 1 \notin A, \\ F(x), & 0 \notin A,\ 1 \in A, \\ 1, & 0, 1 \notin A. \end{cases}$$

The independence of $\{Y_j\}_{j=1}^n$ is assigned as an exercise.

Thus, for each x,

$$F_n(x) \equiv \frac{1}{n}\sum\nolimits_{j=1}^{n} Y_j$$

is a random variable on this probability space, and in fact is a general binomial with parameters n and $p = F(x)$.

Exercise 5.51 *Prove that for fixed x, that $\{Y_j\}_{j=1}^n$ defined above are independent random variables. Note that this is true even if $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}(\mathbb{R}^{\mathbb{N}}), \mu_{\mathbb{N}})$ is defined relative to the general model for component spaces in the introduction, which does not require that X_j be identically distributed.*

When $\{X_j\}_{j=1}^n$ are identically distributed random variables with distribution function F it follows that $\{\chi_{(-\infty,x]}(X_j)\}_{j=1}^n$ are identically distributed standard binomials with common binomial probability $p = \lambda\left[X^{-1}((-\infty, x])\right] = F(x)$. Hence, in this case of identically distributed $\{X_j\}_{j=1}^n$, we have that **for each x:**

$$E\left[F_n(x)\right] = F(x), \qquad Var\left[F_n(x)\right] = F(x)(1 - F(x))/n. \tag{5.35}$$

Based on interpretation 2, the following proposition presents two results which are corollaries to the earlier results of this section. For these results we fix x and consider $F_n(x)$ as a random variable on the probability space $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}(\mathbb{R}^{\mathbb{N}}), \mu_{\mathbb{N}})$, and so in this context $F(x)$ as a constant.

Proposition 5.52 *Let $\{X_j\}_{j=1}^n$ be independent, identically distributed random variables with distribution function F, and define $F_n(x)$ as in 5.34. Then for each x :*

1. *As $n \to \infty$, $F_n(x)$ converges with probability 1 to $F(x)$:*

$$F_n(x) \to_{a.s.} F(x).$$

In other words,

$$\mu_{\mathbb{N}}\{F_n(x) \to F(x)\} = 1.$$

2. *Define the normalized random variable as in 5.7:*

$$Y_n = \frac{F_n(x) - F(x)}{\sqrt{F(x)(1 - F(x))/n}},$$

and let G_n denote the distribution function of Y_n. Then as $n \to \infty$:

$$G_n \Rightarrow \Phi,$$

where Φ denotes the distribution function of the standard normal.

Proof. *The proof of* 1 *is an immediate consequence of the strong law of large numbers version SLLN 1, while the proof of* 2 *is an application of the central limit theorem using 5.35.* ■

The above proposition provides the important insight that numerical samples are informative of their underlying distributions at each x. Not only does $F_n(x)$ converge to $F(x)$ with probability 1, but also we can in theory make probability statements about $F(x)$ based on the numerical sample value $\{X_j\}_{j=1}^n$ and value of $F_n(x)$. As is typical in estimating confidence intervals about unknown binomial parameters where we also do not know the standard deviation, we can approximate the standard deviation with the sample value:

$$\hat{\sigma} \equiv \sqrt{F_n(x)(1-F_n(x))/n},$$

and now use a **Student's T distribution** with $n-1$ **degrees of freedom** in place of the normal as noted in part 2 of example 1.22. This distribution is named for **William Sealy Gosset** (1876 – 1937) who used the pen name of Student.

This then produces the familiar $100(1-\alpha)\%$ confidence interval for $F(x)$:

$$F_n(x)+t_{\alpha/2}^{(n-1)}\sqrt{F_n(x)(1-F_n(x))/n} \leq F(x) \leq F_n(x)+t_{1-\alpha/2}^{(n-1)}\sqrt{F_n(x)(1-F_n(x))/n},$$

where $t_{\alpha/2}^{(n-1)}$ and $t_{1-\alpha/2}^{(n-1)}$ are the associated quantiles of this distribution, and by symmetry, $t_{\alpha/2}^{(n-1)} = -t_{1-\alpha/2}^{(n-1)}$. Once $n \approx 100$ or so, these quantiles are quite close to those of the standard normal. For smaller values of n, Student's T has fatter tails than the normal in that as $|x| \to \infty$, the density function with ν degrees of freedom satisfies:

$$f_\nu^T(x) = O\left(|x|^{-(\nu+1)}\right).$$

While for each x one can estimate individual values of $F(x)$ from a sample as above and be confident of probability 1 convergence: $F_n(x) \to_{a.s.} F(x)$, this theory does not immediately support any conclusions about the extent to which F_n converges to F more generally. Indeed, for each x we have that there is a measurable set $A_x \subset \mathbb{R}^{\mathbb{N}}$ with $\mu_{\mathbb{N}}(A_x) = 0$ and $F_n(x) \to F(x)$ for $s \in \mathbb{R}^{\mathbb{N}} - A_x$. But as there are uncountably many such x it is possible that the union of exceptional sets, $\cup_x A_x$, is not measurable, and even if measurable is it possible in theory that $\mu_{\mathbb{N}}(\cup_x A_x) > 0$. Consequently, it is in theory possible that $\mathbb{R}^{\mathbb{N}} - \cup_x A_x$, the set on which $F_n(x) \to F(x)$ for all x, is not measurable or measurable with probability less than 1.

Remark 5.53 *It is tempting to think that right continuity of distribution functions would simplify matters. Specifically, define* $A = \cup_{x\in\mathbb{Q}} A_x$ *so* A *is the union of* A_x *for rational* x*. Then since* $\mathbb{Q}$ *is countable it follows that* $A \subset \mathbb{R}^{\mathbb{N}}$ *is measurable with* $\mu_{\mathbb{N}}(A) = 0$*. And by definition, for all* $s \in \mathbb{R}^{\mathbb{N}} - A$*,* $F_n(x) \to F(x)$ *for all* $x \in \mathbb{Q}$*. It then seems compelling by right continuity that for all* $s \in \mathbb{R}^{\mathbb{N}} - A$ *that* $F_n(x) \to F(x)$ *for all* $x \in \mathbb{R} - \mathbb{Q}$ *as well.*

What is clear is that if $x \in \mathbb{R} - \mathbb{Q}$ *and if* $\{x_m\} \subset \mathbb{Q}$ *with* $x_m > x$ *and* $x_m \to x$*, then:*

- *For every* $s \in \mathbb{R}^{\mathbb{N}}$ *and every* n*,* $F_n(x_m) \to F_n(x)$ *by right continuity.*
- $F(x_m) \to F(x)$ *by right continuity,*
- $F_n(x_m) \to F(x_m)$ *for all* $x_m \in \mathbb{Q}$ *and* $s \in \mathbb{R}^{\mathbb{N}} - A$ *by probability* 1 *convergence.*

Can we then conclude that $F_n(x) \to F(x)$ *for such* $x \in \mathbb{R} - \mathbb{Q}$ *and* $s \in \mathbb{R}^{\mathbb{N}} - A$*?*

Unfortunately, no. In order to verify this deduction formally, it is necessary to make an assumption on the uniformity of convergence in x *of* $F_n \to F$ *for each* $s \in \mathbb{R}^{\mathbb{N}} - A$*. For example, if* $F(x) = 1$ *on* $[0,1]$ *and* $F_n(x) = x^{1/n}$ *on* $[0,1]$*, then all functions are right continuous at* $x = 0$*,* $F_n(x) \to F(x)$ *for all* $x > 0$*, and yet it is apparent that* $F_n(0) \nrightarrow F(0)$*.*

We now turn to a positive result.

5.4.2 The Glivenko-Cantelli Theorem

While not at all apparent as noted in remark 5.53, the next result proves that outside a set $A \in \sigma(\mathbb{R}^{\mathbb{N}})$ with $\mu_{\mathbb{N}}(A) = 0$, that $F_n(x) \to F(x)$ **uniformly in** x. It is known as the **Glivenko-Cantelli theorem,** named for **Valery Glivenko** (1896 – 1940) and **Francesco Paolo Cantelli** (1875 – 1966), who independently derived this result and published in the same journal in 1933. In order to state this result it is necessary to introduce a measure of the maximum distance between $F_n(x)$ and $F(x)$ for each $s \in \mathbb{R}^{\mathbb{N}}$. To this end, define:

$$D_n(s) = \sup_x |F_n(x) - F(x)|. \tag{5.36}$$

Although $|F_n(x) - F(x)|$ is a random variable defined on $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}(\mathbb{R}^{\mathbb{N}}), \mu_{\mathbb{N}})$ for every x, the definition of $D_n(s)$ requires the supremum over

uncountably many x and hence is not necessarily a measurable function. However, by right continuity we verify as an exercise that:

$$\sup_x |F_n(x) - F(x)| = \sup_{x\in\mathbb{Q}} |F_n(x) - F(x)|, \qquad ((*))$$

and hence as a supremum of countably many measurable functions, D_n is a random variable on $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}(\mathbb{R}^{\mathbb{N}}), \mu_{\mathbb{N}})$.

Exercise 5.54 *Verify* $(*)$. *Hint: Distribution functions are monotonic.*

Fixing a sample, which is to say fixing $s \in \mathbb{R}^{\mathbb{N}}$, $D_n(s)$ defined above is called a **Kolmogorov-Smirnov statistic,** named for **Andrey Kolmogorov** (1903 – 1987) and **Nikolai Smirnov** (1900 – 1966). This statistic was used in the **Kolmogorov-Smirnov test** or **K-S test,** to test the "goodness of fit" between an empirical distribution and an assumed distribution, as well as to test the equivalence of two empirical distributions.

The Glivenko-Cantelli theorem states that $D_n(s) \to 0$ outside a set in $\mathbb{R}^{\mathbb{N}}$ of measure 0. Immediate corollaries of this result are:

1. Outside an $\mathbb{R}^{\mathbb{N}}$-set of measure 0, F_n converges weakly to F: $F_n \Rightarrow F$.

2. Outside an $\mathbb{R}^{\mathbb{N}}$-set of measure 0, $F_n(x) \to F(x)$ uniformly in x.

We now state this result, suppressing the more complicated notation of $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}(\mathbb{R}^{\mathbb{N}}), \mu_{\mathbb{N}})$ and using instead, $(\mathcal{S}, \mathcal{E}, \lambda)$.

Proposition 5.55 (Glivenko-Cantelli theorem) *Let* $\{X_j\}_{j=1}^n$ *be independent, identically distributed random variables on* $(\mathcal{S}, \mathcal{E}, \lambda)$ *with distribution function* F, *and define* $D_n(s)$ *as in 5.36. Then as* $n \to \infty$, $D_n(s) \to 0$ *with probability* 1 :

$$D_n(s) \to_{a.s.} 0. \qquad (5.37)$$

Proof. *Define*

$$F_n(x^-) = \frac{1}{n}\sum\nolimits_{j=1}^{n} \chi_{(-\infty,x)}(X_j(s)),$$

and note that $F_n(x^-)$ *is again a binomial random variable but with probability* $p^- = \lambda\left[X^{-1}((-\infty, x))\right] = F(x^-)$. *Another application of the strong law of large numbers proves that for each* x, $F_n(x^-) \to F(x^-)$ *outside a set of measure* 0, *which we denote by* B_x.

With F^* *denoting the left continuous inverse of* F *defined in 1.19, let* $x_{k/m} = F^*(k/m)$ *for integer* m, *and* $1 \le k \le m$. *Then by proposition 3.19 of book 2,* $F\left(F^*(y)^-\right) \le y \le F\left(F^*(y)\right)$ *for all* $y \in (0, 1)$. *Letting* $y = k/m$ *it*

follows that $F\left(x_{k/m}^{-}\right) \leq k/m \leq F\left(x_{k/m}\right)$ *for* $1 \leq k \leq m-1$. *This obtains that:*

$$F\left(x_{(k-1)/m}^{-}\right) \leq (k-1)/m \leq F\left(x_{(k-1)/m}\right) \leq F\left(x_{k/m}^{-}\right) \leq k/m \leq F\left(x_{k/m}\right),$$

and thus for $2 \leq k \leq m-1$ *that* $F(x_{k/m}^{-}) - F(x_{(k-1)/m}) \leq 1/m$. *With* $k = 2$ *and* $k = m-1$ *these inequalities also yield directly that* $F(x_{1/m}^{-}) \leq 1/m$ *and* $(m-1)/m \leq F(x_{(m-1)/m})$, *and it then follows that* $F(x_{m/m}) \geq F(x_{m/m}^{-}) \geq 1 - 1/m$.

Now define

$$D_{n,m}(s) = \max\left[\max_{1\leq k\leq m}\left|F_n(x_{k/m}) - F(x_{k/m})\right|, \max_{1\leq k\leq m}\left|F_n(x_{k/m}^{-}) - F(x_{k/m}^{-})\right|\right].$$

We claim that for any m :

$$D_n(s) \leq D_{n,m}(s) + 1/m. \qquad ((*))$$

To see this, let $(k-1)/m \leq x < k/m$ *where* $2 \leq k \leq m-1$. *Then by the above inequalities:*

$$\begin{aligned} F_n(x) - F(x) &\leq F_n(x_{k/m}^{-}) - F(x_{(k-1)/m}) \leq F_n(x_{k/m}^{-}) - F(x_{k/m}^{-}) + 1/m, \\ F_n(x) - F(x) &\geq F_n(x_{(k-1)/m}) - F(x_{k/m}^{-}) \geq F_n(x_{(k-1)/m}) - F(x_{(k-1)/m}) - 1/m, \end{aligned}$$

and so:

$$F_n(x_{(k-1)/m}) - F(x_{(k-1)/m}) - 1/m \leq F_n(x) - F(x) \leq F_n(x_{k/m}^{-}) - F(x_{k/m}^{-}) + 1/m.$$

The same analysis applies when $k = 1, m$, *and the claim is proved.*

Now let $A = \bigcup_{k,m}\left(A_{x_{k/m}} \cup B_{x_{k/m}}\right)$, *where* $A_{x_{k/m}} \subset \mathcal{S}$ *is the exceptional set of measure zero outside of which* $F_n(x_{k/m}) \to F(x_{k/m})$, *and* $B_{x_{k/m}} \subset \mathcal{S}$ *is similarly defined relative to* $F_n(x_{k/m}^{-}) \to F(x_{k/m}^{-})$. *Then* A *is a countable union of sets of measure zero and hence* $\lambda(A) = 0$. *Also, if* $s \in \mathcal{S} - \mathcal{A}$, *then* $D_{n,m}(s) \to 0$ *for any* m *as* $n \to \infty$ *by two applications of the strong law of large numbers, and hence by* $(*)$, $\lim_{n\to\infty} D_n(s) \leq 1/m$ *for every* m, *proving 5.37.* ■

5.4.3 Distributional Estimates for $D_n(s)$

In this section we provide some additional results on the empirical distribution function. These results are presented without proof as the

tools required for a formal development are deep and intricate and would require a great deal of additional theory. As noted above, the Glivenko-Cantelli theorem was independently derived by Valery Glivenko and Francesco Paolo Cantelli and published in the same journal in 1933. Also in that same journal, **Andrey Kolmogorov** (1903 – 1987) published what has come to be known as **Kolmogorov's theorem** which gives an asymptotic distributional result for $D_n(s)$ as $n \to \infty$ when F is assumed to be continuous.

Remark 5.56 *It should be noted that there are a number of important results by Kolmogorov that have come to be known by his name.*

Proposition 5.57 (Kolmogorov's theorem) *Let $\{X_j\}_{j=1}^n$ be independent, identically distributed random variables on $(\mathcal{S}, \mathcal{E}, \lambda)$ with **continuous** distribution function F. Let $F_{D_n}(t)$ denote the distribution function of D_n defined as in 5.36. Then for all $t > 0$:*

$$F_{D_n}(t/\sqrt{n}) \to 1 - 2\sum_{k=1}^{\infty}(-1)^{k+1}e^{-2k^2t^2} \tag{5.38}$$

as $n \to \infty$, and this convergence is uniform in t.

Remark 5.58 *The distribution function in 5.38 is called the **Kolmogorov distribution function** and is defined to equal 0 when $t < 0$. When $t = 0$ the limit of this function is also 0 as $t \to 0^+$, and thus the Kolmogorov distribution function is continuous.*

Kolmogorov derived this result by first proving that the limiting distribution in 5.38 is independent of F for continuous distributions, and then explicitly derived the limiting distribution for the uniform distribution function on $[0, 1]$, F_U, associated with the density function defined in . This reduction utilizes the tools already developed so we provide this detail.

Proposition 5.59 (Distribution-free property of D_n) *The distribution for D_n defined in 5.36 is the same for all continuous distribution functions F, and thus in particular can be evaluated with $F = F_U$, the continuous, uniform distribution function of 1.18 defined on $[0, 1]$.*
Proof. *By proposition 4.8 of book 2, if Y has a continuous uniform distribution then $F^*(Y)$ has distribution function $F(x)$. Thus with $x = F^*(y)$:*

$$\begin{aligned} D_n(s) &= \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \\ &= \sup_{y\in[0,1]} |F_n(F^*(y)) - F(F^*(y))| . \end{aligned}$$

From proposition 3.22 of book 2, $F(F^*(y)) = y$ *if and only if* F *is continuous, and thus:*

$$D_n(s) = \sup_{y\in[0,1]} \left|F_n(F^*(y)) - y\right|.$$

By 5.34:

$$F_n(F^*(y)) = \frac{1}{n}\sum\nolimits_{j=1}^{n} \chi_{(-\infty,F^*(y)]}(X_j),$$

and since F *is increasing:*

$$X_j \leq F^*(y) \text{ if and only if } F(X_j) \leq F\left(F^*(y)\right).$$

Again from proposition 3.22, $F\left(F^*(y)\right) = y$ *and so:*

$$F_n(F^*(y)) = \frac{1}{n}\sum\nolimits_{j=1}^{n} \chi_{(-\infty,y]}\left(F(X_j)\right).$$

By proposition 4.9 of book 2, if F *is continuous then the distribution function of* $F(X)$ *is the continuous uniform distribution:*

$$F_{F(X)}(y) = y = F_U(y),$$

and with $Y_j \equiv F(X_j)$ *if* $\{X_j\}_{j=1}^n$ *are independent then* $\{Y_j\}_{j=1}^n$ *are independent. Thus:*

$$\begin{aligned} F_n(F^*(y)) &= \frac{1}{n}\sum\nolimits_{j=1}^{n} \chi_{(-\infty,y]}\left(Y_j\right) \\ &\equiv F_n^U(y), \end{aligned}$$

and:

$$D_n(s) = \sup_{y\in[0,1]} \left|F_n^U(y) - y\right|.$$

■

Several years after Kolmogorov's result, **Nikolai Smirnov** (1900 – 1966) derived in 1939 a comparable limit theorem for the one-sided measure:

$$D_n^+(s) = \sup_x \left[F_n(x) - F(x)\right], \tag{5.39}$$

and in the same year developed results for the difference between two empirical distributions. For example, on $D_n^+(s)$ he proved:

Proposition 5.60 (Smirnov's theorem) *Let* $\{X_j\}_{j=1}^n$ *be independent, identically distributed random variables on* $(\mathcal{S}, \mathcal{E}, \lambda)$ *with continuous distribution function* F*. Let* $F_{D_n^+}(t)$ *denote the distribution function of* D_n^+ *as defined as in 5.39. Then for all* $t > 0$ *:*

$$F_{D_n^+}(t/\sqrt{n}) \to 1 - e^{-2t^2} \tag{5.40}$$

as $n \to \infty$*, and this convergence is uniform in* t*.*

In 1956 this earlier work was used to develop a type of large deviation estimate as discussed in the next section, but for D_n with finite n. The first result is called the **Dvoretzky–Kiefer–Wolfowitz inequality,** or the **Dvoretzky–Kiefer–Wolfowitz theorem** and named for **Aryeh (Arie) Dvoretzky** (1916 – 2008), **Jack Kiefer** (1924 – 1981) and **Jacob Wolfowitz** (1910 – 1981). This original result stated the inequality below in 5.41 with an undefined coefficient C. This result was then improved in 1990 by **Pascal Massart** by the derivation of a sharp estimate that $C = 2$.

Note that the following result does not require that F be continuous.

Proposition 5.61 (Dvoretzky–Kiefer–Massart–Wolfowitz theorem) *Let $\{X_j\}_{j=1}^n$ be independent, identically distributed random variables on $(\mathcal{S}, \mathcal{E}, \lambda)$ with distribution function F. With D_n defined as in 5.36:*

$$\Pr[D_n > t] \leq 2e^{-2nt^2} \quad t > 0. \tag{5.41}$$

Equivalently,

$$\Pr[D_n \leq t] \geq 1 - 2e^{-2nt^2}, \quad t > 0.$$

Example 5.62 *The DKMW theorem allows the estimation of a "confidence band" about the entire empirical distribution function which will contain the theoretical distribution function with the degree of confidence implied by $1 - 2e^{-2nt^2}$. For example, if $n = 100$ and a 95% confidence band for $F(x)$ is desired, we choose t so that $1 - 2e^{-200t^2} = 0.95$ producing $t = 0.135\,81$. Then with $F_n(x)$ denoting the empirical distribution function based on the given sample of 100, 5.41 implies the 95% confidence band:*

$$\max(F_n(x) - 0.135\,81, 0) \leq F(x) \leq \min(F_n(x) + 0.135\,81, 1).$$

This confidence band implies that if $A \subset \mathcal{S}$ is defined by:

$$A = \{\max(F_n(x, s) - 0.135\,81, 0) \leq F(x) \leq \min(F_n(x, s) + 0.135\,81, 1)\},$$

then

$$\lambda[A] \geq 0.95,$$

and hence since the original sample point s_0 is assumed random,

$$\Pr[s_0 \in A] \geq 0.95.$$

In other words, there is a 95% probability that the given sample s_0 will produce a distributional band which contains $F(x)$ for all x.

Remark 5.63 *Because of the theoretical bounds on* $F(x)$ *of* 0 *and* 1, *the bands for* F *produced with the DKMW theorem are primarily useful for intermediate values of* x, *and less useful in the tails of the distribution when these a priori bounds will be achieved. In general, for a* $100\alpha\%$ *confidence band the solution to* $1 - 2e^{-2nt^2} = \alpha$ *is given by:*

$$t = \left[\frac{1}{2n} \ln\left(\frac{2}{1-\alpha}\right)\right]^{0.5}.$$

If k *is the largest integer with* $k/n \leq t$, *then the lower bound of* 0 *is achieved when* $x \leq X_{(k)}$, *and the upper bound of* 1 *is achieved when* $x \geq X_{(n-k)}$. *Thus the confidence interval is informative only for* $X_{(k)} < x < X_{(n-k)}$, *an interval that decreases as* α *increases.*

Chapter 6

Estimating Tail Events 2

In this chapter we continue the investigations begun in chapter 9 of book 2 using properties of expectations, and also for the first section results on the moment generating function.

6.1 Large Deviation Theory 2

In book 2, large deviation theory was introduced and the main result derived there can be summarized as follows:
If $\{X_j\}_{j=1}^n$ *are independent and identically distributed,* $S_n \equiv \sum_{j=1}^n X_j$, *and* $\pi_n = \pi_n(t) \equiv \ln[\Pr\{S_n \geq nt\}]$, *then for any* $t > 0$:

1. $\lim_{n\to\infty}(\pi_n/n)$ *exists and equals* $\sup_m(\pi_m/m)$.
2. *Denoting by* $\pi(t) = \sup_m(\pi_m/m)$, *then* $-\infty \leq \pi(t) \leq 0$ *and for all* n :
$$\Pr\{S_n \geq nt\} \leq e^{n\pi(t)}.$$
3. *When* $-\infty < \pi(t) < 0$, *it follows that* $\Pr\{S_n \geq nt\} \to 0$ *exponentially fast as* $n \to \infty$.

In this section we develop additional insights on the bounding function, $\pi(t)$.

To do so, assume that the moment generating function $M_X(t)$ exists in an interval $(-t_0, t_0)$ with $t_0 > 0$, and recall Markov's inequality in 3.76:

$$\Pr[|X| \geq t] \leq E[|X|]/t.$$

From this it follows that for $\theta > 0$,

$$\begin{aligned} \Pr\{S_n \geq nt\} &= \Pr\{\exp(\theta S_n) \geq \exp(\theta nt)\} \\ &\leq \frac{E[\exp(\theta S_n)]}{\exp(\theta nt)}. \end{aligned}$$

By 3.35:

$$E[\exp(\theta S_n)] = M_X^n(\theta),$$

and so

$$\Pr\{S_n \geq nt\} \leq \exp[-n[\theta t - \ln M_X(\theta)]].$$

Since this inequality is true for all $\theta > 0$,

$$\Pr\{S_n \geq nt\} \leq \inf_{\theta \geq 0}[\exp[-n(\theta t - \ln M_X(\theta))]],$$

or to better reveal the exponent function,

$$\Pr\{S_n \geq nt\} \leq \exp\left[-n \sup_{\theta \geq 0}[\theta t - \ln M_X(\theta)]\right].$$

Thus in the case where the distribution function of X has a moment generating function on $(-t_0, t_0)$ with $t_0 > 0$, a candidate can be identified for the exponential decay function $\pi(t)$ associated with $\Pr\{S_n \geq nt\}$:

$$\pi(t) = -\sup_{\theta \geq 0}[\theta t - \ln M_X(\theta)]. \tag{6.1}$$

Summary 6.1 *If $S_n = \sum_{j=1}^{n} X_j$ with $\{X_j\}_{j=1}^n$ independent and identically distributed, and $M_X(\theta)$ exists on $(-t_0, t_0)$ with $t_0 > 0$, then for any $t > 0$,*

$$\Pr\{S_n \geq nt\} \leq \exp\left[-n \sup_{\theta \geq 0}[\theta t - \ln M_X(\theta)]\right]. \tag{6.2}$$

Remark 6.2 *The upper bound in 6.2 is called the **Chernoff bound,** named for **Herman Chernoff** (1923 –). This bound is of the same type as **Bernstein's inequality** in 5.4 of book 2, there derived for a sum of independent binomials and named for **Sergei Natanovich Bernstein** (1880 – 1968), who developed several specialized inequalities of this type.*

There are two immediate questions that arise from the **Chernoff bound:**

1. Is $\sup_{\theta \geq 0}(\theta t - \ln M_X(\theta))$ positive, so that 6.2 then provides an estimate of the exponential decay rate for $\Pr\{S_n \geq nt\}$?

2. If positive, is this estimate of exponential decay at least close to the best possible?

Before investigating, we consider two examples.

Example 6.3 1. *Binomial* X_j :

From 3.49, $M_B(\theta) = \left(1 + p(e^\theta - 1)\right)$. *With*

$$\begin{aligned}\Gamma_B(\theta) &\equiv \theta t - \ln[M_B(\theta)] \\ &= \theta t - \ln\left(1 + p(e^\theta - 1)\right),\end{aligned}$$

it is an exercise to check that $\Gamma''_B(\theta) < 0$ *for all* θ *and hence* $\Gamma_B(\theta)$ *is concave and has a maximum when* $\Gamma'_B(\theta) = 0$. *A calculation shows that this equation is solvable for any* t *with* $0 < t < 1$, *and has a solution at* $\theta^* \equiv \theta^*(t)$ *defined by:*

$$\theta^* = \ln\left[\frac{t(1-p)}{p(1-t)}\right].$$

But then $\theta^* > 0$ *if and only if* $t > p = E[X]$, *and in this case it follows that* $\sup_{\theta \geq 0} \Gamma_B(\theta) = \Gamma_B(\theta^*)$*:*

$$\Gamma_B(\theta^*(t)) = t \ln\left(\frac{t}{p}\right) + (1-t)\ln\left(\frac{1-t}{1-p}\right). \tag{6.3}$$

As a function of t, $\Gamma_B(\theta^*(t))$ *is strictly increasing for* $p < t < 1$ *since:*

$$\frac{d\Gamma_B(\theta^*(t))}{dt} = \ln\frac{t}{1-t} - \ln\frac{p}{1-p} > 0,$$

and since $\Gamma_B(\theta^*(p)) = 0$ *it follows that* $\Gamma_B(\theta^*(t)) > 0$. *Hence for* t *with* $p < t < 1$:

$$\Pr\{S_n \geq nt\} \leq \exp\left[-n\Gamma_B(\theta^*(t))\right], \tag{6.4}$$

and $\Pr\{S_n \geq nt\}$ *decreases exponentially as* $n \to \infty$.

When $p < t < 1$ *the decay rate in 6.3 increases with* t *because* $\frac{d}{dt}\Gamma_B(\theta^*(t)) > 0$, *and,* $\Gamma_B(\theta^*(t))$ *has positive second derivative and is hence a convex function of* t.

Finally, when $0 \leq t \leq p$ *then* $\theta^*(t) < 0$ *and so* $\sup_{\theta \geq 0} \Gamma_B(\theta) = \Gamma_B(0) = 0$ *and this upper bound in 6.4 is not useful.*

2. ***Normal*** X_j :

From 3.66 $M_N(\theta) = \exp\left(\mu\theta + \frac{1}{2}\sigma^2\theta^2\right)$. *Defining* $\Gamma_N(\theta)$ *as above it again follows that* $\Gamma_N''(\theta) < 0$ *for all* θ *and hence there is a maximum when* $\Gamma_N'(\theta) = 0$. *A calculation produces that this occurs for any* $t > 0$ *at* $\theta^* \equiv \theta^*(t) > 0$:

$$\theta^* = \frac{t-\mu}{\sigma^2}.$$

However, as above $\theta^* > 0$ *if and only if* $t > \mu = E[X]$. *We then have that* $\sup_{\theta\geq 0}\Gamma_N(\theta) = \Gamma_N(\theta^*)$ *where:*

$$\Gamma_N(\theta^*(t)) = \frac{(t-\mu)^2}{2\sigma^2}, \tag{6.5}$$

and hence for $t > \mu$*:*

$$\Pr\{S_n \geq nt\} \leq \exp\left[-n\Gamma_N(\theta^*(t))\right]. \tag{6.6}$$

Since $\Gamma_N(\theta^*(t)) > 0$, $\Pr\{S_n \geq nt\}$ *decreases exponentially as* $n \to \infty$. *As a function of* t *for* $t > \mu$, *it is apparent that* $\Gamma_N(\theta^*(t))$ *is increasing and convex.*

Definition 6.4 *Let*

$$\Gamma(\theta) = \theta t - \ln M_X(\theta).$$

If there exists $\theta^*(t) > 0$ *that satisfies*

$$\Gamma(\theta^*(t)) = \sup_{\theta\geq 0}\Gamma(\theta),$$

then $\Gamma(\theta^*(t))$ *is called the* ***rate function for*** X.

Remark 6.5 *When* $\theta^*(t) > 0$ *exists, it follows from 6.2 that:*

$$\Pr\{S_n \geq nt\} \leq \exp\left[-n\Gamma(\theta^*(t))\right], \tag{6.7}$$

and thus decreases exponentially as $n \to \infty$.

We are now ready to address the two questions posed before the above example. The first result is that when $M_X(\theta)$ exists and $t > E[X]$, the bound in 6.2 or 6.7 is meaningful in the sense that it produces exponential decay as $n \to \infty$. To then prove concavity of $\Gamma(\theta)$ will require a general version of Lebesgue's dominated convergence theorem from book 5, and a special case of the change of variables result found there which we derive below. The use of Lebesgue's result should not jeopardize the intelligibility of this derivation since the book 5 dominated convergence theorem will seem quite familiar after reading the book 3 version.

Proposition 6.6 *Assume that $M_X(\theta)$ exists on $(-\theta_0, \theta_0)$ with $\theta_0 > 0$. If $t > E[X]$ then $\Gamma(\theta) \equiv \theta t - \ln M_X(\theta) > 0$ for some $\theta > 0$ and hence $\sup_{\theta \geq 0} (\theta t - \ln M_X(\theta)) > 0$. In addition, $\Gamma''(\theta) < 0$ and so $\Gamma(\theta)$ is a concave function.*

Proof. *Because $M_X(\theta) > 0$ it follows that $\Gamma(\theta) = \theta t - \ln M_X(\theta)$ is infinitely differentiable on the interval $(-\theta_0, \theta_0)$ on which $M_X(\theta)$ is assumed to exist, and $\Gamma(0) = 0$. Also, $\Gamma'(\theta) = t - M_X'(\theta)/M_X(\theta)$ and so $\Gamma'(0) = t - E[X]$. Hence if $t > E[X]$ then $\Gamma'(0) > 0$ and by the definition of derivative as a limit, there is an interval $(0, \theta') \subset (0, \theta_0)$ so that $\Gamma(\theta) > 0$ for $\theta \in (0, \theta')$.*

To prove concavity assume that $\theta \in (-\theta_0, \theta_0)$ is given for which $M_X(\theta)$ is assumed to exist and define a function F_θ by the Riemann-Stieltjes integral:

$$F_\theta(y) = \frac{1}{M_X(\theta)} \int_{-\infty}^{y} e^{\theta x} dF, \tag{6.8}$$

where F is the distribution function of X. Since $e^{\theta x}$ is continuous and F bounded and increasing, this integral exists for all y by proposition 4.24 of book 3. It is an exercise to check that $F_\theta(y)$ is continuous, increasing, and has the appropriate limits at $\pm\infty$, and thus is a distribution function.

Let Y be a random variable with distribution function F_θ, assured to exist by proposition 3.6 of book 2. We now prove that $E[Y^n]$ exists for all n and:

$$E[Y^n] \equiv \int_{-\infty}^{\infty} y^n dF_\theta = \frac{M_X^{(n)}(\theta)}{M_X(\theta)},$$

where $M_X^{(n)}(\theta)$ is the nth derivative of $M_X(\theta)$. Note that $M_X(\theta)$ is infinitely differentiable by proposition 3.24. Once proved this assures concavity of $\Gamma(\theta)$:

$$\begin{aligned} \Gamma''(\theta) &= \left(\frac{M_X'(\theta)}{M_X(\theta)}\right)^2 - \frac{M_X''(\theta)}{M_X(\theta)} \\ &= -Var[Y] \\ &\leq 0, \end{aligned}$$

and so $\Gamma'(\theta)$ is a decreasing function and $\Gamma(\theta)$ is a concave function by proposition 3.37.

Now the first step toward the above formula for $E[Y^n]$ is to prove that for all $\theta \in (-\theta_0, \theta_0)$:

$$M_X^{(n)}(\theta) = \int_{-\infty}^{\infty} y^n e^{\theta y} dF. \tag{6.9}$$

This would be obvious if we could assume the validity of differentiation under the integral sign, but there is no apparent way to justify this. So we proceed directly by proving that for $n \geq 0$*:*

$$M_X^{(n)}(\theta) = \int_{-\infty}^{\infty} y^n e^{\theta y} dF \Rightarrow M_X^{(n+1)}(\theta) = \int_{-\infty}^{\infty} y^{n+1} e^{\theta y} dF,$$

and then noting that when $n = 0$ *that* $M_X^{(0)}(\theta) \equiv M_X(\theta)$*.*

The existence of $M_X^{(n+1)}(\theta)$ *assures by definition that as* $m \to \infty$*:*

$$\begin{aligned} m\left[M_X^{(n)}(\theta + 1/m) - M_X(\theta)\right] &\to M_X^{(n+1)}(\theta), \\ m\left[M_X^{(n)}(\theta) - M_X(\theta - 1/m)\right] &\to M_X^{(n+1)}(\theta), \end{aligned}$$

where we can assume m *is large enough so that* $\theta \pm 1/m \in (-\theta_0, \theta_0)$*. Now define:*

$$f_m^+(y) = y^n e^{\theta y} m\left[e^{y/m} - 1\right], \qquad f_m^-(y) = y^n e^{\theta y} m\left[1 - e^{-y/m}\right],$$

and note that:

$$\begin{aligned} m\left[M_X^{(n)}(\theta + 1/m) - M_X(\theta)\right] &= \int_{-\infty}^{\infty} f_m^+(y) dF, \\ m\left[M_X^{(n)}(\theta) - M_X(\theta - 1/m)\right] &= \int_{-\infty}^{\infty} f_m^-(y) dF. \end{aligned}$$

Both integrals on the right are well defined and bounded by definition since $\theta \pm 1/m \in (-\theta_0, \theta_0)$*. Then from*

$$m\left[1 - e^{-y/m}\right] \leq y \leq m\left[e^{y/m} - 1\right]$$

it follows that:

$$\min\left[\left|f_m^+(y)\right|, \left|f_m^-(y)\right|\right] \leq \left|y^{n+1} e^{\theta y}\right| \leq \max\left[\left|f_m^+(y)\right|, \left|f_m^-(y)\right|\right],$$

and since all bounding functions are integrable,

$$\int_{-\infty}^{\infty} \left|y^{n+1} e^{\theta y}\right| dF < \infty$$

for all $\theta \in (-\theta_0, \theta_0)$*.*

Let $g_m(y) \equiv \min[f_m^+(y), f_m^-(y)]$ *and note that since* $f_m^{\pm}(y) \to y^{n+1}e^{\theta y}$ *as* $m \to \infty$, *so too* $g_m(y) \to y^{n+1}e^{\theta y}$ *as* $m \to \infty$. *Thus by the general version of Lebesgue's dominated convergence theorem in book 5:*

$$\int_{-\infty}^{\infty} g_m(y)dF \to \int_{-\infty}^{\infty} y^{n+1}e^{\theta y}dF.$$

If n *is even then* $f_m^-(y) \le f_m^+(y)$ *and thus* $g_m(y) = f_m^-(y)$ *and this obtains by the above results that* $M_X^{(n+1)}(\theta) = \int_{-\infty}^{\infty} y^{n+1}e^{\theta y}dF$. *When* n *is odd,* $f_m^-(y) \le f_m^+(y)$ *for* $y \ge 0$ *and* $f_m^-(y) \ge f_m^+(y)$ *for* $y \le 0$, *but the same conclusion will follow if it can be proved that as* $m \to \infty$:

$$\int_0^{\infty} \left|f_m^+(y) - f_m^-(y)\right| dF \to 0.$$

Since $f_m^+(y) - f_m^-(y) \to 0$ *pointwise, each term converging to* $y^{n+1}e^{\theta y}$, *this result will again follow by Lebesgue's dominated convergence theorem if the integrand is so dominated. A Taylor series analysis show that for* $m \ge m_0$:

$$f_m^+(y) - f_m^-(y) \le 2y^{n+1}e^{(\theta+1/m)y} \le 2y^{n+1}e^{(\theta+1/m_0)y},$$

and this integrable as noted above if $\theta + 1/m_0 \in (-\theta_0, \theta_0)$.

Thus 6.9 is proved and for the next step we express this result as follows, recalling that $M_X(\theta) > 0$:

$$\frac{M_X^{(n)}(\theta)}{M_X(\theta)} = \frac{1}{M_X(\theta)} \int_{-\infty}^{\infty} y^n e^{\theta y}dF. \qquad ((*))$$

To show that the expression in $(*)$ *equals* $E[Y^n]$, *we begin by applying corollary 4.22 of book 3 to the above* dF_θ*-Riemann-Stieltjes integral representation of* $E[Y^n]$. *Specifically, for any* $\epsilon > 0$ *there exists a partition* $\{(y_{j-1}, y_j)\}_{j=-\infty}^{\infty}$ *of* $\mathbb{R}$ *and so that with arbitrary tags* $\tilde{y}_j \in (y_{j-1}, y_j)$ *this integral can be approximated within* ϵ *by a Riemann-Stieltjes summation:*

$$\left| \int_{-\infty}^{\infty} y^n dF_\theta - \sum_{j=-\infty}^{\infty} (\tilde{y}_j)^n \left[F_\theta(y_j) - F_\theta(y_{j-1})\right] \right| < \epsilon.$$

While corollary 4.22 applied to integrals over bounded $[a, b]$, $\mathbb{R}$ *is a countable union of such intervals and corollary 4.22 can be applied to approximate the associated integrals with an error of* $\epsilon/2^j$ *and this produces the desired total error of* ϵ. *If necessary we further refine this partition so that over each such interval* (y_{j-1}, y_j), *if* M *and* m *denote the maximum and minimum values of* y^n :

$$|M - m| \le \epsilon e^{-|\theta|y_j} M_X(\theta)/2^{j+1}. \qquad ((**))$$

This is possible by corollary 4.22 since the above approximation also applies to all refinements of the original partition.

Now:

$$F_\theta(y_j) - F_\theta(y_{j-1}) \equiv \frac{1}{M_X(\theta)} \int_{y_{j-1}}^{y_j} e^{\theta y} dF,$$

and corollary 4.22 of book 3 can again be applied to produce a partition $\{(y_{j_k-1}, y_{j_k})\}_{k=1}^{n_j}$ *of each* (y_{j-1}, y_j) *so that for arbitrary tags* $y'_{jk} \in (y_{j_k-1}, y_{j_k})$*:*

$$\left| [F_\theta(y_j) - F_\theta(y_{j-1})] - \frac{1}{M_X(\theta)} \sum_{k=1}^{n_j} e^{\theta y'_{jk}} [F(y_{j_k}) - F(y_{j_k-1})] \right| < \epsilon/2^{j+1}.$$

Combining obtains:

$$\left| \int_{-\infty}^{\infty} y^n dF_\theta - \frac{1}{M_X(\theta)} \sum_{j=-\infty}^{\infty} \sum_{k=1}^{n_j} (\tilde{y}_j)^n e^{\theta y'_{jk}} [F(y_{j_k}) - F(y_{j_k-1})] \right| < 2\epsilon.$$

This latter summation is seen to be a Riemann-Stieltjes summation for the integral in $(*)$*, except the* $\tilde{y}_j$ *tags are generally not in the* (y_{j_k-1}, y_{j_k}) *intervals. But by the above construction of* $\{(y_{j-1}, y_j)\}_j$ *we can replace these tags by the above subinterval tags at a cost of at most* ϵ *:*

$$\begin{aligned}
&\frac{1}{M_X(\theta)} \left| \sum_{j=-\infty}^{\infty} \left[\sum_{k=1}^{n_j} \left[(\tilde{y}_j)^n - \left(y'_{jk}\right)^n \right] e^{\theta y'_{jk}} [F(y_{j_k}) - F(y_{j_k-1})] \right] \right| \\
\leq\ & \epsilon \left| \sum_{j=-\infty}^{\infty} \left[\sum_{k=1}^{n_j} e^{-|\theta| y_j} e^{\theta y'_{jk}} [F(y_{j_k}) - F(y_{j_k-1})] \right] / 2^{j+1} \right| \\
\leq\ & \epsilon \sum_{j=-\infty}^{\infty} [F(y_j) - F(y_{j-1})] / 2^{j+1} \\
\leq\ & \epsilon.
\end{aligned}$$

The first inequality is from $(**)$*, the second since* $\theta y'_{jk} < |\theta| y_j$ *for all* k*, and finally* $[F(y_j) - F(y_{j-1})] \leq 1$*.* ■

Remark 6.7 *It will be seen in book 5 that the above transition from the* dF_θ*-integral representation for* $E[Y^n]$ *to the* dF*-integral representation is a special case of a general change of variables result. Recall from 6.8:*

$$F_\theta(y) = \frac{1}{M_X(\theta)} \int_{-\infty}^{y} e^{\theta x} dF,$$

and thus as for change of variables in Riemann integrals, at least notationally:

$$dF_\theta = \frac{1}{M_X(\theta)} e^{\theta y} dF.$$

Thus:

$$E[Y^n] \equiv \int_{-\infty}^{\infty} y^n dF_\theta = \frac{1}{M_X(\theta)} \int_{-\infty}^{\infty} y^n e^{\theta y} dF \tag{6.10}$$

seems obvious, again at least notationally. In book 5 such change of variable results will be justified more generally.

As an application of this general result, or derived using the approach of the above proof, it follows that if $\theta \in (-\theta_0, \theta_0)$ *then* $M_Y(t)$ *exists for* $t \in (-t_0, t_0)$ *where* $(\theta - t_0, \theta + t_0) \in (-\theta_0, \theta_0)$*, and replacing* y^n *with* e^{ty} *in 6.10:*

$$M_Y(t) \equiv \int_{-\infty}^{\infty} e^{ty} dF_\theta = \frac{M_X(\theta + t)}{M_X(\theta)}. \tag{6.11}$$

Notation 6.8 *The distribution function* F_θ *in 6.8 is referred to as the* ***twisted distribution,*** *or the* ***exponentially tilted distribution,*** *associated with* F*.*

This proposition assures that if $t > E[X]$, $\sup_{\theta \geq 0} \ (\theta t - \ln M_X(\theta)) > 0$ and that $\Gamma(\theta)$ is a concave function. Hence since $\Gamma(\theta)$ is also differentiable, if we can solve $\Gamma'(\theta) = 0$ for given t with $\theta^*(t) > 0$ then $\sup_{\theta \geq 0} \Gamma(\theta) = \Gamma(\theta^*)$. This was the approach taken in example 6.3. Solving $\Gamma'(\theta) = 0$ is equivalent to solving the following for $\tilde{\theta}$:

$$t = \frac{M'_X(\tilde{\theta})}{M_X(\tilde{\theta})}, \tag{6.12}$$

noting that by concavity that it then follows that $\theta^*(t) \equiv \tilde{\theta}$.

Now the above proof shows that $M'_X(\theta)/M_X(\theta)$ is an increasing function since $[M'_X(\theta)/M_X(\theta)]' = Var[Y]$, and since $M'_X(0)/M_X(0) = \mu$ we conclude that "if" the equation in 6.12 is solvable for $t > \mu$ then it will be solvable with $\tilde{\theta} > 0$. That said, the above proposition does not assure that we can always determine the value of $\theta^*(t)$ which achieves this supremum using this approach because $M'_X(\theta)/M_X(\theta)$ may be bounded and hence 6.12 will be unsolvable for $t > \max[M'_X(\theta)/M_X(\theta)]$.

Example 6.9 *With density function* $f(x) = Cx^{-3}e^{-x}$ *on* $x \geq 1$*, where* C *is chosen so that* f *integrates to* 1*, note that* $M_X(\theta)$ *exists when* $\theta \leq 1$*. Further,* $M'_X(\theta)/M_X(\theta)$ *is increasing but bounded when* $\theta \leq 1$*, since*

$$\begin{aligned} \frac{M'_X(\theta)}{M_X(\theta)} &\leq \frac{M'_X(1)}{M_X(1)} \\ &= \int_1^{\infty} x^{-2} dx \Big/ \int_1^{\infty} x^{-3} dx \\ &= 2. \end{aligned}$$

Hence, the equation in 6.12 has no solution for $t > 2$.

We next turn to the accuracy of the bound in 6.2 and 6.7, and for this we have the **Cramér-Chernoff Theorem** named for **Harald Cramér** (1893 – 1985) and **Herman Chernoff** (1923 –). It states that if $\theta^*(t) > 0$ exists for given $t > \mu \equiv E[X]$, meaning $\sup_{\theta \geq 0} \Gamma(\theta) = \Gamma(\theta^*(t))$, then the bound in 6.7 is exact in the limit as $n \to \infty$. In other words, this bound is tight.

For this result, we assume that the distribution function F of X is absolutely continuous, recalling the discussion in section 1.1. The same proof works if F is a saltus function, replacing integrals below with summations. The key point of these assumptions is to assure that X has a density function, so in particular it is assumed that F has has no singular part.

Remark 6.10 *Note that a new function $\theta^{\#}(t)$ is introduced in the statement of this proposition which is defined in terms of $\sup_\theta \Gamma(\theta)$ while $\theta^*(t)$ is defined relative to $\sup_{\theta \geq 0} \Gamma(\theta)$. However, the first development in the proof is that $\theta^{\#}(t) = \theta^*(t)$ when $t > \mu$.*

Also, note that for the below proof with an absolutely continuous distribution function F, we must again appeal to the useful result of Fubini's theorem of book 5 which allows the evaluation of a multivariate integral in terms of iterated 1-dimensional integrals. When F is a saltus function with discrete density, the book 5 machinery is not needed.

Proposition 6.11 (Cramér-Chernoff Theorem) *Let $S_n = \sum_{j=1}^n X_j$ with $\{X_j\}_{j=1}^n$ independent and identically distributed with a distribution function containing no singular part, and where $M_X(\theta)$ exists on $(-\theta_0, \theta_0)$ with $\theta_0 > 0$. Assume that for given $t > \mu \equiv E[X]$, and that $\theta^{\#}(t) > 0$ exists with $\theta^{\#}(t) \in (0, \theta_0)$ and:*

$$\Gamma(\theta^{\#}(t)) = \sup_\theta \Gamma(\theta),$$

where $\Gamma(\theta) = \theta t - \ln M_X(\theta)$. Then for any such t and $\epsilon > 0$ there is an N so that for $n \geq N$:

$$\exp\left[-n\left(\Gamma(\theta^{\#}(t)) + \epsilon\right)\right] \leq \Pr\{S_n \geq nt\} \leq \exp\left[-n\Gamma(\theta^{\#}(t))\right]. \qquad (6.13)$$

Proof. *The upper bound in 6.13 is obtained by 6.7 if we can prove that for $t > \mu$ that $\theta^{\#}(t) = \theta^*(t)$, meaning:*

$$\sup_\theta \Gamma(\theta) = \sup_{\theta \geq 0} \Gamma(\theta).$$

Because $\Gamma(0) = 0$ *this will follow by showing that* $\Gamma(\theta) \leq 0$ *for* $\theta < 0$. *Now since* $f(x) \equiv e^{\theta x}$ *is convex,* $M_X(\theta) \geq e^{\theta\mu}$ *by Jensen's inequality in 3.80, while if* $\theta < 0$ *and* $t > \mu$ *obtain* $\exp[-\theta(t-\mu)] > 1$, *and so*:

$$\exp[-\theta(t-\mu)]M_X(\theta) > e^{\theta\mu}.$$

Then $\Gamma(\theta) \leq 0$ *is confirmed after taking logarithms.*

For the lower bound let $\epsilon > 0$ *be given. For this proof we use the density version of the twisted distribution function* F_θ *introduced in 6.8 in the proof of proposition 6.6. That* X *has an absolutely continuous distribution function* F *implies that there is an associated measurable density function* f. *Independence of* $\{X_j\}_{j=1}^n$ *assures that* $f(x_1, ..., x_n) = \Pi_{j=1}^n f(x_j)$ *by the proposition 3.53 of book 2 result on the associated distribution functions.*

For any $\delta > 0$:

$$\begin{aligned}\Pr\{S_n \geq nt\} &\geq \Pr\left\{nt \leq \sum\nolimits_{j=1}^n X_j \leq n(t+\delta)\right\} \\ &= \int \cdots \int_A f(x_1)...f(x_n)dx_1...dx_n\end{aligned}$$

where $A \equiv \left\{nt \leq \sum_{j=1}^n x_j \leq n(t+\delta)\right\}$. *Define the twisted density function,*

$$g(x) = \frac{\exp[\theta^* x]}{M_X(\theta^*)} f(x),$$

noting that $\theta^* \equiv \theta^*(t) = \theta^{\#}(t)$ *since* $t > \mu$. *Let* Y *be a random variable associated with the distribution function defined by* $g(x)$, *again assured to exist by proposition 3.6 of book 2.*

$$\begin{aligned}&\Pr\left\{\sum\nolimits_{j=1}^n X_j \geq nt\right\} \\ \geq\ &\frac{M_X^n(\theta^*)}{\exp[n(t+\delta)\theta^*]} \int \cdots \int_A \frac{\prod_{j=1}^n \exp[\theta^* x_j]}{M_X^n(\theta^*)} f(x_1)...f(x_n)dx_1...dx_n \\ \equiv\ &\frac{M_X^n(\theta^*)}{\exp[n(t+\delta)\theta^*]} \int \cdots \int_A g(x_1)...g(x_n)dx_1...dx_n \\ =\ &\frac{M_X^n(\theta^*)}{\exp[n(t+\delta)\theta^*]} \Pr\left\{nt \leq \sum\nolimits_{j=1}^n Y_j \leq n(t+\delta)\right\}.\end{aligned}$$

In the first step the inequality is $\Pr\left\{\sum_{j=1}^n X_j \geq nt\right\} \geq \Pr\left\{nt \leq \sum_{j=1}^n X_j \leq n(t+\delta)\right\}$, *with the latter represented as an integral, as well as* $\sum_{j=1}^n x_j \leq n(t+\delta)$ *by the definition of* A, *while the last also follows from the definition of* A.

Next, note that

$$\begin{aligned}\frac{M_X^n(\theta^*)}{\exp\left[n\left(t+\delta\right)\theta^*\right]} &= \exp\left[n\ln M_X(\theta^*) - n\left(t+\delta\right)\theta^*\right] \\ &= \exp\left[-n\Gamma(\theta^*)\right]\exp\left[-n\theta^*\delta\right].\end{aligned}$$

In addition, $E[Y] = M_X'(\theta^*)/M_X(\theta^*)$ *as derived in the prior proof, and since* θ^* *maximizes concave and differentiable* $\Gamma(\theta)$ *it follows that* $\Gamma'(\theta^*) = 0$. *A calculation with* $\Gamma(\theta) = \theta t - \ln M_X(\theta)$ *then produces* $E[Y] = t$.

Combining the pieces, let σ_Y *denote the standard deviation of* Y, *which exists since* $M_Y(\theta) = M_X(\theta^* + \theta)/M_X(\theta^*)$ *exists for* $|\theta| < \theta_0 - \theta^*$ *as noted in 6.11. Then:*

$$\begin{aligned}&\Pr\left\{\sum\nolimits_{j=1}^n X_j \geq nt\right\} \\ \geq\ &\exp\left[-n\Gamma(\theta^*)\right]\exp\left[-n\theta^{*\#}\delta\right] \times \\ &\Pr\left\{0 \leq \sum\nolimits_{j=1}^n \left(Y_j - E[Y]\right)/\left(\sigma_Y\sqrt{n}\right) \leq \delta\sqrt{n}/\sigma_Y\right\}.\end{aligned}$$

By the Central Limit theorem the probability expression converges to $1/2$ *as* $n \to \infty$ *for any* $\delta > 0$. *Choose* δ *so that* $\delta \leq \epsilon/\left(2\theta^*\right)$, *then since* $\exp\left[-n\theta^*\delta\right] \geq \exp\left[-n\epsilon/2\right]$ *determine* N_1 *so that:*

$$\begin{aligned}\exp\left[-n\theta^*\delta\right] &\geq \exp\left[-n\epsilon\right]\exp\left[n\epsilon/2\right] \\ &\geq 4\exp\left[-n\epsilon\right]\end{aligned}$$

for $n \geq N_1$. *For this same* δ, *let* N_2 *be defined so that for* $n \geq N_2$:

$$\Pr\left\{0 \leq \sum\nolimits_{j=1}^n \left(Y_j - E[Y]\right)/\left(\sigma_Y\sqrt{n}\right) \leq \delta\sqrt{n}/\sigma_Y\right\} \geq 1/4$$

Then for $n \geq \max[N_1, N_2]$:

$$\Pr\left\{\sum\nolimits_{j=1}^n X_j \geq nt\right\} \geq \exp\left[-n\left(\Gamma(\theta^{\#}) + \epsilon\right)\right].$$

■

Corollary 6.12 *Under the assumptions of the prior proposition, for* $t > \mu \equiv E[X]$:

$$\lim_{n\to\infty}\frac{1}{n}\ln\left[\Pr\left\{S_n \geq nt\right\}\right] = -\Gamma(\theta^*(t)). \tag{6.14}$$

Proof. *From the above result it follows that for any* $\epsilon > 0$ *there is an* N *so that for* $n \geq N$:

$$-\left(\Gamma(\theta^{\#}(t)) + \epsilon\right) \leq \frac{1}{n}\ln\left[\Pr\left\{S_n \geq nt\right\}\right] \leq -\Gamma(\theta^{\#}(t)).$$

The result follows by definition of limit, and because $\theta^{\#}(t) = \theta^{*}(t)$ *for* $t > \mu$.
■

6.2 Extreme Value Theory 2

Extreme value theory was introduced in book 2. One of the two central results of this theory is called the **Fisher-Tippett theorem** and named for the earliest developers – **Ronald Fisher** (1890 – 1962), and **Leonard Tippett** (1902 – 1985), known professionally as L. H. C. Tippett. This theorem is also called the **Fisher-Tippett-Gnedenko theorem** recognizing the later contributions of **Boris Gnedenko** (1912 – 1995). The second central result is the **Pickands–Balkema–de Haan theorem,** named for 1974-5 papers of **A. A. Balkema** and **Laurens de Haan,** and, **James Pickands III**, and sometimes called the **Gnedenko-Pickands–Balkema–de Haan theorem,** and even the **Gnedenko theorem** in recognition of the earlier 1943 paper of **Boris Gnedenko**.

The first result above was developed in proposition 9.30 of book 2 which we summarize here, while the multivarate version of this theorem can be seen in that book's proposition 9.52. In the one-dimensional version, this result address weak convergence of the distribution function of the maximum value of an independent collection of n-identically distributed random variables. In the notation of order statistics the distribution function of this variate is $F_{(n)}$, and by proposition 2.2, $F_{(n)}(x) = F^n(x)$ where F is the distribution function of the underlying random variable.

Proposition (Fisher-Tippett-Gnedenko theorem) *Let* F *be the distribution function of a random variable* X *and* F^n *the distribution function of* $M_n = \max_{m \leq n}\{X_m\}$ *for independent* $\{X_m\}_{m=1}^{n}$. *Assume that there exists sequences* $\{a_n\}_{n=1}^{\infty}$ *and* $\{b_n\}_{n=1}^{\infty}$ *where* $a_n > 0$ *for all* n, *so that* $F^n(a_n x + b_n) \Rightarrow G(x)$ *for a nondegenerate distribution function* G. *Then there are real constants* $A > 0$, B, *and* γ *so that* $G(x) = G_\gamma(Ax + B)$ *with* G_γ *defined as in 9.25 of book 2 for* $\gamma \neq 0$ by*:*

$$G_\gamma(x) = \exp\left(-(1+\gamma x)^{-1/\gamma}\right), \qquad 1 + \gamma x \geq 0. \tag{6.15}$$

When $\gamma = 0$, $G_\gamma(x)$ *is defined as in 9.26 of book 2, and equals the limit of* $G_\gamma(x)$ *as* $\gamma \to 0$:

$$G_0(x) \equiv \exp\left(-e^{-x}\right). \tag{6.16}$$

Remark 6.13 *The single family of distributions identified and parametrized by γ is called the* ***extreme value class of distributions****, and the distribution function $G_\gamma(ax+b)$ is called a* ***generalized extreme value distribution,*** *abbreviated* ***GEV.*** *The parameter $\gamma \in \mathbb{R}$ is called the* ***extreme value index.***

When $F^n(a_n x+b_n) \Rightarrow G_\gamma(Ax+B)$ we say that the distribution function F is in the ***domain of attraction of*** *G_γ, denoted $F \in \mathcal{D}(G_\gamma)$.*

In proposition 9.45 of book 2 was derived the **von Mises' Condition,** named for **Richard von Mises** (1883 – 1953), which identified how the extreme value index γ could be derived given twice continuously differentiable F. Recall that as defined in 5.31,

$$x^* = \inf\{x|F(x) = 1\},$$

with $x^* \equiv \infty$ if $F(x) < 1$ for all x.

Proposition (von Mises' Condition) *Let F be a twice continuously differentiable distribution function with $F'(x) > 0$ for some interval (x_0, x^*), where x^* is defined in 5.31. If:*

$$\lim_{x \to x^*} \left(\frac{1-F}{F'} \right)' (x) = \gamma, \tag{6.17}$$

then F is in the domain of attraction of G_γ, i.e., $F \in D(G_\gamma)$.

The von Mises' condition provides one approach to determining if a given distribution function is in the domain of attraction of G_γ for some γ. For example, with Φ denoting the standard normal distribution function it was shown in example 9.48 of book 2 that $\Phi \in \mathcal{D}(G_0)$.

When dealing with data sets from unknown distributions in finance and other disciplines, the question naturally arises as to how one can determine if the data is consistent with that from a distribution F with $F \in \mathcal{D}(G_\gamma)$ for some γ. While there are many approaches to this estimation problem, a popular and frequently used approach is the **Hill estimator,** introduced in 1975 by **Bruce M. Hill.**

The first of two sections below study this estimator and its properties. The final section returns to the Gnedenko-Pickands–Balkema–de Haan theorem introduced in the book 2 section 9.2.4 on Finance Applications, and is now addressed in more detail in the case of $\gamma > 0$.

6.2.1 The Hill Estimator, γ_H

The Hill estimator was introduced as an approach to estimating the exponent parameter of the **Pareto distribution** introduced in remark 9.40 of book 2 and named for **Vilfredo Pareto** (1848 – 1923). This distribution is also called the **power law** or a **distribution of Zipf type,** the latter named for **George Kingsley Zipf** (1902 – 1950). This distribution is defined on $x > x_0$ by:

$$F(x) = 1 - \left(\frac{x}{x_0}\right)^{-1/\gamma}, \tag{6.18}$$

where $\gamma > 0$. This model is reflective of many data observations made in finance and elsewhere, and is often parametrized with $\alpha = 1/\gamma$. But parametrized this way, it follows that

$$\left(\frac{1-F}{F'}\right)' = \gamma,$$

and using **von Mises' condition** in 6.17 such distributions satisfy $F \in \mathcal{D}(G_\gamma)$. Since $\gamma > 0$ here it follows that Pareto, power law, or Zipf-type distributions are in the domain of attraction of the **Fréchet class of distributions**, also called **Type II extreme value distributions** as noted in remark 9.14 of book 2.

The Hill estimator for $\gamma > 0$ is defined as follows. Let $\{X_i\}_{i=1}^n$ be a data sample of a given variate, and $\{X_{(j)}\}_{j=1}^n$ the associated order statistics as defined in chapter 2. The Hill estimator is based on the $k+1$ largest variates, $\{X_{(n-j)}\}_{j=0}^k$, and is defined as an average of k log ratios:

$$\gamma_H \equiv \frac{1}{k}\sum\nolimits_{j=0}^{k-1} \ln\left[\frac{X_{(n-j)}}{X_{(n-k)}}\right], \tag{6.19}$$

which is equivalently expressed in terms of differences:

$$\gamma_H \equiv \frac{1}{k}\sum\nolimits_{j=0}^{k-1} \left[\ln X_{(n-j)} - \ln X_{(n-k)}\right]. \tag{6.20}$$

Example 6.14 *Assume that the distribution function F is known to be a Pareto distribution as in 6.18 with $x_0 = 1$ for simplicity:*

$$F_P(x) = \begin{cases} 0, & x < 1, \\ 1 - x^{-1/\gamma}, & x \geq 1, \end{cases}$$

so $F_P \in \mathcal{D}(G_\gamma)$ by von Mises' condition as noted above. It is then not difficult to justify the estimator in 6.19. Given the ordered sample, $\{X_{(j)}\}_{j=1}^n$, define the conditional distribution function for $x \geq X_{(n-k)}$ as in definition 3.39 of book 2:

$$\begin{aligned} F_P\left(x|x \geq X_{(n-k)}\right) &= \frac{F_P(x) - F_P(X_{(n-k)})}{1 - F_P(X_{(n-k)})} \\ &= \left[X_{(n-k)}^{-\alpha} - x^{-\alpha}\right] / X_{(n-k)}^{-\alpha}. \end{aligned}$$

Here we temporarily denote the tail index parameter $\alpha = 1/\gamma$ for notational simplicity. The conditional density function is then given as the derivative of this differentiable function:

$$f_P\left(x|x \geq X_{(n-k)}\right) = \alpha x^{-(\alpha+1)} / X_{(n-k)}^{-\alpha}.$$

*Given a sample $\{X_j\}_{j=1}^n$ with an assumed density function $f(x;\alpha)$ which "reflects" and is therefore "conditional" on a parameter α to be estimated, the **conditional likelihood function,** and sometimes the **likelihood function** of the sample, is defined:*

$$L[\{X_j\}_{j=1}^n; \alpha] \equiv \prod_{j=1}^n f(X_j; \alpha).$$

*A logical objective is to maximize L as a function of α, producing the **conditional maximum likelihood estimate** for α, often called the **maximum likelihood estimate/estimator** or **MLE.** By maximizing L, the given parameter α provides a model which maximizes the probability of the observed sample among the family of distributions $\{f(x;\alpha)\}$ parametrized by α.*

Applying this approach to the sample $\{X_{(n-j)}\}_{j=0}^k$ and $f(x;\alpha) \equiv \alpha x^{-(\alpha+1)} / X_{(n-k)}^{-\alpha}$ obtains the likelihood function:

$$L\left[\{X_{(n-j)}\}_{j=0}^k; \alpha\right] = \alpha^k \prod_{j=0}^{k-1} X_{(n-j)}^{-(\alpha+1)} / X_{(n-k)}^{-\alpha}.$$

To maximize L as a differentiable function of α, or equivalently to maximize the logarithm of this function, we differentiate:

$$\frac{\partial \ln L}{\partial \alpha} = \frac{k}{\alpha} - \sum_{j=0}^{k-1} \ln X_{(n-j)} + k \ln X_{(n-k)}.$$

The maximum likelihood estimate equals the value of α that solves $\frac{\partial \ln L}{\partial \alpha} = 0$ if $\frac{\partial^2 \ln L}{\partial \alpha^2} < 0$, which is verifiable in this case. A calculation yields $\alpha = 1/\gamma_H$.

It is then checked as an exercise that if parametrized as a function of γ, one again has that $\frac{\partial \ln L}{\partial \gamma} = 0$ and $\frac{\partial^2 \ln L}{\partial \gamma^2} < 0$ when $\gamma = \gamma_H$.

Summary: *When the distribution function F is known to be a Pareto distribution, the Hill estimator γ_H is equal to the maximum likelihood estimate for the parameter γ.*

Of course for general $F \in \mathcal{D}(G_\gamma)$ with $\gamma > 0$ we cannot assert that F is Pareto. However, as was derived in proposition 9.38 of book 2, if $\gamma > 0$ then for all $x \geq 0$:

$$\lim_{t \to x^*} \frac{1 - F(t + x h_a(t))}{1 - F(t)} = (1 + \gamma x)^{-1/\gamma} .$$

Recall that $h_a(t) \equiv a\left(\frac{1}{1-F(t)}\right)$ where $a(t)$ is the normalizing function in that book's corollary 9.27. This is defined as $a(t) \equiv a_{\lfloor t \rfloor}$, with $\lfloor t \rfloor$ is the **greatest integer function** defined by $\lfloor t \rfloor = \max\{n | n \leq t\}$, and a_n the sequence given in the statement of the Fisher-Tippett-Gnedenko theorem. Also, as defined in 5.31, $x^* = \inf\{x | F(x) = 1\}$, with $x^* \equiv \infty$ if $F(x) < 1$ for all x.

To investigate the implications of this general result, note that

$$1 - \frac{1 - F(t + x h_a(t))}{1 - F(t)} = \frac{F(t + x h_a(t)) - F(t)}{1 - F(t)},$$

and the right hand expression equals the conditional distribution, $F(t + x h_a(t) | x > 0)$. In other words, for $F \in \mathcal{D}(G_\gamma)$ with $\gamma > 0$, the above limiting result asserts that the conditional distribution function: $F(t + x h_a(t) | x > 0)$ has limit as $t \to x^*$:

$$\lim_{t \to x^*} F(t + x h_a(t) | x > 0) = 1 - (1 + \gamma x)^{-1/\gamma} .$$

With a reparametrization of $y = 1 + \gamma x$, this implies that the following special conditional distribution is asymptotically Pareto, meaning:

$$\lim_{t \to x^*} F\left(t + h_a(t)\left(\frac{y-1}{\gamma}\right) | y > 1\right) = 1 - y^{-1/\gamma}.$$

In the next three subsections we develop the Hill estimator result in three steps:

1. **If $F \in D(G_\gamma)$ with $\gamma > 0$, then F is conditionally asymptotically Pareto.**

We will refine the analysis of the behavior of F and show in 6.28 below that:

$$\lim_{t\to\infty} \Pr[X \le tx | X > t] = 1 - x^{-1/\gamma}.$$

In other words, this conditional distribution of F will be shown to be asymptotically Pareto. From this result we will then derive an alternative integral formula for the parameter γ associated with such a function. This formula will generalize the von Mises' condition in that it will not require the differentiability needed for that approach, but requires the added restriction that $\gamma > 0$.

2. **If** $F \in D(G_\gamma)$ **with** $\gamma > 0$, **then** $\gamma_H \approx \gamma$.

   The integral formula for γ derived in step 1 will be approximated and will produce the estimate γ_H.

In the third section we will prove the main result for the Hill estimator:

3. **If** $F \in D(G_\gamma)$ **with** $\gamma > 0$, **then** $\gamma_H \to_P \gamma$ **as** $n \to \infty$.

   For this convergence in probability result it will be required that $k \to \infty$ and $k/n \to 0$ as $n \to \infty$. In other words, the Hill estimator γ_H must of necessity be based on increasingly high quantiles of X since the estimator uses ordered data in the $[n-k, n]$ range, which is equivalent to the $\left[1 - \frac{k}{n}, 1\right]$ quantile range.

The final discussion on the Hill estimator will address its asymptotic normality.

1. If $F \in D(G_\gamma)$ with $\gamma > 0$, then F is Conditionally Asymptotically Pareto

The goal of this section is to derive an integral formula for the parameter γ for general $F \in D(G_\gamma)$ with $\gamma > 0$, and then in the next section show that when this integral is approximated the Hill estimator is obtained. We begin by recalling corollary 9.35 of book 2, that if $F \in D(G_\gamma)$ for any γ, then from 9.29 there we have for all $x > 0$:

$$\lim_{t\to\infty} \frac{U(tx) - U(t)}{a(t)} = \frac{x^\gamma - 1}{\gamma}, \tag{6.21}$$

where here the constant c of that result is integrated into $a(t)$. The function $U(t)$ is the left continuous inverse of $1/(1 - F(x))$:

$$U(t) \equiv \left(\frac{1}{1-F}\right)^*(t).$$

We begin by investigating and refining this limit.

Proposition 6.15 *If $\gamma > 0$, then $U(t) \to \infty$ as $t \to \infty$ and:*

$$\lim_{t\to\infty} \frac{U(t)}{a(t)} = \frac{1}{\gamma}. \tag{6.22}$$

Also, for $x > 0$:

$$\lim_{t\to\infty} \frac{U(tx)}{U(t)} = x^{\gamma}. \tag{6.23}$$

Proof. *Defining $V_t(x) \equiv \frac{U(tx)-U(t)}{a(t)}$, then:*

$$\frac{a(tx)}{a(t)} = \frac{V_t(xy) - V_t(x)}{V_{tx}(y)},$$

and by applying 6.21 to each V we conclude that

$$\lim_{t\to\infty} \frac{a(tx)}{a(t)} = x^{\gamma}. \tag{6.24}$$

Further, since:

$$\frac{U(tx)}{U(t)} = 1 + \frac{U(tx) - U(t)}{a(t)} \frac{a(t)}{U(t)},$$

it follows from 6.21 that the limit in 6.23 exists if and only if the limit in 6.22 exists and is nonzero.

But note that if the limit in 6.22 exists this limit would of necessity be nonzero. This follows since:

$$\frac{U(tx) - U(t)}{a(t)} = \frac{U(tx)}{a(tx)} \frac{a(tx)}{a(t)} - \frac{U(t)}{a(t)},$$

and since $U(tx)/a(tx)$ and $U(t)/a(t)$ must have this same limit, the limit in 6.22 is nonzero by 6.24 and 6.21. Finally, if the limit in 6.22 exists and is thus nonzero, the above identity and 6.21 obtain:

$$\lim_{t\to\infty} \frac{U(tx)}{U(t)} = 1 + \frac{x^{\gamma} - 1}{\gamma} \left[1 \Big/ \lim_{t\to\infty} \frac{U(t)}{a(t)} \right],$$

and so 6.23 is true if and only if 6.22 is true.

So the final challenge is to prove the existence of either limit in 6.22 or 6.23, and we verify the existence of the latter limit as well as the fact that $U(t) \to \infty$ as $t \to \infty$.

Let $Z > 1$, then an application of 6.24 yields $a\left(Z^k\right)/a\left(Z^{k-1}\right) \to Z^\gamma$ by defining $t = Z^{k-1}$. A similar application of 6.21 and defining $t = Z^{k-1}$ or $t = Z^k$ as appropriate obtains:

$$\begin{aligned} & \lim_{k\to\infty} \frac{U\left(Z^{k+1}\right) - U\left(Z^k\right)}{U\left(Z^k\right) - U\left(Z^{k-1}\right)} \\ = & \lim_{k\to\infty} \frac{U\left(Z^{k+1}\right) - U\left(Z^k\right)}{a\left(Z^k\right)} \Bigg/ \frac{U\left(Z^k\right) - U\left(Z^{k-1}\right)}{Z^\gamma a\left(Z^{k-1}\right)} \\ = & \ Z^\gamma. \end{aligned}$$

So for any $\epsilon > 0$ there is an $N \equiv N(\epsilon)$ so that for $k \geq N$:

$$Z^\gamma(1-\epsilon) \leq \frac{U\left(Z^{k+1}\right) - U\left(Z^k\right)}{U\left(Z^k\right) - U\left(Z^{k-1}\right)} \leq Z^\gamma(1+\epsilon). \qquad ((*))$$

Now for $n > N$,

$$U\left(Z^{n+1}\right) - U\left(Z^N\right) = \left[U\left(Z^N\right) - U\left(Z^{N-1}\right)\right] \sum_{j=N}^{n} \prod_{k=N}^{j} \frac{U\left(Z^{k+1}\right) - U\left(Z^k\right)}{U\left(Z^k\right) - U\left(Z^{k-1}\right)},$$

and so from $(*)$:

$$\begin{aligned} \lim_{n\to\infty} \left[U\left(Z^{n+1}\right) - U\left(Z^N\right)\right] & \geq \left[U\left(Z^N\right) - U\left(Z^{N-1}\right)\right] \lim_{n\to\infty} \sum_{j=N}^{n} \prod_{k=N}^{j} Z^\gamma(1-\epsilon) \\ & = \left[U\left(Z^N\right) - U\left(Z^{N-1}\right)\right] \lim_{n\to\infty} \sum_{j=N}^{n} \left[Z^\gamma(1-\epsilon)\right]^{j-N+1}. \end{aligned}$$

Hence $\lim_{n\to\infty} U\left(Z^{n+1}\right) = \infty$ if $Z^\gamma(1-\epsilon) > 1$, which is to say, $\epsilon < 1 - Z^{-\gamma}$, and this then proves that $U(t) \to \infty$ as $t \to \infty$.

Turning next to 6.23, since $U(t) \to \infty$ as $t \to \infty$:

$$\begin{aligned} \lim_{n\to\infty} \left[\sum_{k=N}^{n} \left[U\left(Z^{k+1}\right) - U\left(Z^k\right)\right] / U\left(Z^n\right)\right] & = \lim_{n\to\infty} \frac{U\left(Z^{n+1}\right) - U\left(Z^N\right)}{U\left(Z^n\right)} \quad ((**)) \\ & = \lim_{n\to\infty} \frac{U\left(Z^{n+1}\right)}{U\left(Z^n\right)}. \end{aligned}$$

Using the inequalities in $(*)$ applied to each term in the first summation yields:

$$\begin{aligned} & Z^\gamma(1-\epsilon) \lim_{n\to\infty} \sum_{k=N}^{n} \left[U\left(Z^k\right) - U\left(Z^{k-1}\right)\right] / U\left(Z^n\right) \\ \leq & \lim_{n\to\infty} \frac{U\left(Z^{n+1}\right)}{U\left(Z^n\right)} \\ \leq & \ Z^\gamma(1+\epsilon) \lim_{n\to\infty} \sum_{k=N}^{n} \left[U\left(Z^k\right) - U\left(Z^{k-1}\right)\right] / U\left(Z^n\right). \end{aligned}$$

Repeating the steps in $(**)$*, it follows that the limits of the summations in these bounds equal* 1*. Thus for any* $\epsilon < 1 - Z^{-\gamma}$ *:*

$$Z^{\gamma}(1-\epsilon) \leq \lim_{n\to\infty} \frac{U\left(Z^{n+1}\right)}{U\left(Z^{n}\right)} \leq Z^{\gamma}(1+\epsilon),$$

and this limit therefore equals Z^{γ} *for any* $Z > 1$*.*

Letting $t = Z^n$ *then suggests that* $\lim_{t\to\infty} U\left(tx\right)/U\left(t\right)$ *exists for all* $x = Z > 1$ *and equals the limit in 6.23 but this must be formalized. Given* $x > 1$ *and* $Z > 1$ *let* $n(x)$ *be such that* $Z^{n(x)} \leq x < Z^{n(x)+1}$*, then because* U *is increasing:*

$$\frac{U(Z^{n(x)}Z^{n(t)})}{U(Z^{n(t)+1})} \leq \frac{U\left(tx\right)}{U\left(t\right)} \leq \frac{U(Z^{n(x)+1}Z^{n(t)+1})}{U(Z^{n(t)})}.$$

For the left bound,

$$\begin{aligned}
\lim_{t\to\infty} \frac{U(Z^{n(x)}Z^{n(t)})}{U(Z^{n(t)+1})} &= \lim_{t\to\infty} \frac{U(Z^{n(x)}Z^{n(t)})}{U(Z^{n(t)})} \frac{U(Z^{n(t)})}{U(Z^{n(t)+1})} \\
&= Z^{(n(x)-1)\gamma} \\
&\geq (x/Z^2)^{\gamma}.
\end{aligned}$$

In this calculation, for example:

$$\frac{U(Z^{n(x)}Z^{n(t)})}{U(Z^{n(t)})} = \frac{U(W^{m(t)+1})}{U(Z^{m(t)})} \to W^{\gamma},$$

with $W \equiv Z^{n(x)}$ *and* $m(t) \equiv \frac{n(t)}{n(x)}$*. Similarly, for the right bound,*

$$\lim_{t\to\infty} \frac{U(Z^{n(x)+2}Z^{n(t)})}{U(Z^{n(t)})} \leq (xZ^2)^{\gamma},$$

and 6.23 follows for $x > 1$ *by letting* $Z \to 1$*.*

The final step to address $0 < x \leq 1$*. Note that the existence of the limit in 6.23 for* $x > 1$ *is sufficient to assure the existence of the limit in 6.22, and this with 6.21 assures 6.23 for all* $x > 0$*.* ■

Definition 6.16 *A function* f *is called* ***regularly varying at infinity with index*** α*,* $\alpha \in \mathbb{R}$*, if:*

$$\lim_{t\to\infty} \frac{f(tx)}{f(t)} = x^{\alpha} \tag{6.25}$$

for all $x \geq x_0 > 0$*. When* $\alpha = 0$*,* f *is said to be* ***slowly varying at infinity.*** *When* f *is* ***regularly varying at infinity with index*** α*, one often writes* $f \in RV_{\alpha}$*.*

Remark 6.17 *The result in 6.23 states that if $F \in D(G_\gamma)$ with $\gamma > 0$, then $U(y) \equiv \left(\frac{1}{1-F}\right)^*(y)$ is **regularly varying at infinity with index** γ. The result in 6.24 implies that the normalizing function $a(t)$ associated with such $F \in D(G_\gamma)$ is also regularly varying with index γ.*

Hence, if $F \in D(G_\gamma)$ with $\gamma > 0$, then both $a, U \in RV_\gamma$.

Before continuing in the current development we identify a corollary result to 6.23 which was promised in remark 9.36 in book 2, regarding the normalizing sequences a_n and b_n in the statement of the **Fisher-Tippett-Gnedenko theorem.**

Corollary 6.18 *If $F \in D(G_\gamma)$ with $\gamma > 0$, $a_n \equiv U(n)$ and $b_n = 0$, then $F^n(a_n x) \Rightarrow G_\gamma(x/\gamma - 1)$ as $n \to \infty$. In other words, for these normalizing sequences $A = 1/\gamma$ and $B = -1$ in the result of proposition 9.30 of book 2, and:*

$$\lim_{n\to\infty} F^n(U(n)x) = \exp\left(-x^{-1/\gamma}\right), \qquad x \geq 0. \tag{6.26}$$

Proof. *The limit in 6.23 restricted to integer $t = n$ can be expressed in terms of left continuous inverses as $H_n^*(x) \Rightarrow K^*(x)$ where $H_n(x) = 1/\{n[1 - F(xU(n))]\}$ and $K(x) = x^{1/\gamma}$. For example,*

$$K^*(x) = \inf\{y | y^{1/\gamma} \geq x\} = x^\gamma.$$

For $H_n(x)$, recalling that also $U(y) \equiv F^*(1 - 1/y)$:

$$\begin{aligned} H_n^*(x) &= \inf\{y | 1/\{n[1 - F(U(n)y)]\} \geq x\} \\ &= \inf\{y | F(U(n)y) \geq 1 - 1/nx\}. \end{aligned}$$

Letting $z = yU(n)$,

$$H_n^*(x) = \frac{1}{U(n)} \inf\{z | F(z) \geq 1 - 1/nx\} = \frac{U(nx)}{U(n)}.$$

By corollary 8.28 of the book 2, $H_n^*(x) \Rightarrow K^*(x)$ implies that $H_n(x) \Rightarrow K(x)$, and taking reciprocals this implies that for all $x \geq 0$:

$$\lim_{n\to\infty} n[1 - F(U(n)x)] = x^{-1/\gamma}.$$

Thus $1 - F(U(n)x) \to 0$ and so $1 - F(U(n)x) = -\ln F(U(n)x) + O(n^{-2})$ and this limit can be expressed:

$$\lim_{n\to\infty} \ln F^n(U(n)x) = -x^{-1/\gamma},$$

which is 6.26. ■

The next step in this development is to convert the limiting results for U to limiting results for the distribution function F. The following proposition states that if $F \in D(G_\gamma)$ with $\gamma > 0$, then the conditional distribution, $F(tx|t) \equiv \Pr[X \leq tx|X > t]$ and defined for $x > 0$, is asymptotically Pareto as $t \to \infty$.

Proposition 6.19 *A distribution function $F \in D(G_\gamma)$ with $\gamma > 0$ if and only if $x^* = \infty$ and for all $x > 0$:*

$$\lim_{t\to\infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-1/\gamma}. \tag{6.27}$$

Since $\Pr[X \leq tx|X > t] \equiv [F(tx) - F(t)] / [1 - F(t)]$, *6.27 can be stated:*

$$\lim_{t\to\infty} \Pr[X \leq tx|X > t] = 1 - x^{-1/\gamma}. \tag{6.28}$$

Proof. *Assume that $F \in D(G_\gamma)$ with $\gamma > 0$. By proposition 6.15 $U(t) \to \infty$ as $t \to \infty$ where by definition,*

$$\begin{aligned} U(t) &\equiv \left(\frac{1}{1-F}\right)^*(t) \\ &= F^*(1 - 1/t). \end{aligned}$$

Hence $F^*(1 - 1/t)$ is unbounded as $t \to \infty$ and $F(x) < 1$ for all x. Thus, $x^* = \infty$ by definition.

Now by proposition 3.19 of book 2:

$$F^*(F(t)) \leq t \leq F^*\left(F(t)^+\right),$$

and since $U\left(\frac{1}{1-F(t)}\right) = F^*(F(t))$ it follows that for any $\epsilon > 0$,

$$U\left(\frac{1-\epsilon}{1-F(t)}\right) \leq t \leq U\left(\frac{1+\epsilon}{1-F(t)}\right),$$

and hence

$$\frac{U\left(\frac{y}{1-F(t)}\right)}{U\left(\frac{1+\epsilon}{1-F(t)}\right)} \leq \frac{1}{t}U\left(\frac{y}{1-F(t)}\right) \leq \frac{U\left(\frac{y}{1-F(t)}\right)}{U\left(\frac{1-\epsilon}{1-F(t)}\right)}.$$

For $F \in D(G_\gamma)$, 6.23 yields that for all $\epsilon > 0$:

$$\left(\frac{y}{1+\epsilon}\right)^\gamma \leq \lim_{t\to\infty} \frac{1}{t}U\left(\frac{y}{1-F(t)}\right) \leq \left(\frac{y}{1-\epsilon}\right)^\gamma,$$

since $F(t) \to 1$, *and so:*

$$\lim_{t\to\infty} \frac{1}{t} U\left(\frac{y}{1-F(t)}\right) = y^{\gamma}. \qquad ((*))$$

Defining $g_n(y) = \frac{1}{n} U\left(\frac{y}{1-F(n)}\right)$ *and* $g(y) = y^{\gamma}$, *then* $g_n(y) \to g(y)$ *for all* $y > 0$ *implies that* $g_n^*(x) \to g^*(x)$ *for each continuity point of* g^* *by proposition 8.27 of book 2. A calculation shows that* $g^*(x) = x^{1/\gamma}$ *and* $g_n^*(x) = \frac{1-F(n)}{1-F(nx)}$, *and the result in 6.27 follows.*

Conversely, if $x^* = \infty$ *and 6.27 is satisfied then by reversing the above steps with corollary 8.28 of book 2 obtains* $(*)$. *Now let* $s = \frac{1}{1-F(t)}$. *Then the assumption that* $x^* = \infty$ *implies that* $s \to \infty$ *as* $t \to \infty$ *and also* $t = F^*(1 - 1/s) = U(s)$, *and so* $(*)$ *implies that*

$$\lim_{s\to\infty} \frac{U(sy)}{U(s)} = y^{\gamma}.$$

Now define the normalizing function $a(s) = \gamma U(s)$ *and* $b(s) = U(s)$, *then for all* $x > 0$:

$$\lim_{s\to\infty} \frac{U(sy) - b(s)}{a(s)} = \frac{y^{\gamma} - 1}{\gamma},$$

and so $F \in D(G_{\gamma})$. ■

Remark 6.20 *Recalling the terminology introduced in definition 6.16, the result in 6.27 states that if* $F \in D(G_{\gamma})$ *with* $\gamma > 0$, *then* $1 - F$ *is* ***regularly varying at infinity with index*** $-1/\gamma$. *Notationally, if* $F \in D(G_{\gamma})$ *with* $\gamma > 0$ *then* $1 - F \in RV_{-1/\gamma}$. *This can in fact be expressed in an even more descriptive way, that if* $F \in D(G_{\gamma})$ *for* $\gamma > 0$ *then as* $x \to \infty$:

$$F(x) = 1 - L(x)x^{-1/\gamma}, \qquad L \in RV_0, \tag{6.29}$$

which is to say that L *is* ***slowly varying at infinity****. This result was derived by Gnedenko, and follows from 6.27 by considering* $L(tx)/L(t)$.

Thus, if $F \in D(G_{\gamma})$ *for* $\gamma > 0$, *then 6.29 states that* F *has a* ***fat tail*** *in the sense that* $1 - F(x)$ *effectively decays like a power function, and this is a distributional observation most often identified in finance applications.*

We present one last result and corollary that improves the above proposition regarding the asymptotic Pareto-like behavior of $F \in D(G_{\gamma})$ with $\gamma > 0$. The proposition below sharpens the result above by providing a uniform estimate of convergence. This proposition is a special case of **Karamata's**

Representation theorem, named for **Jovan Karamata** (1902 – 1967), and while it is also true with modifications for $\gamma < 0$ and $\gamma = 0$, we do not develop this theory.

Proposition 6.21 (Karamata's Representation theorem) *Given a distribution function F, then $F \in D(G_\gamma)$ with $\gamma > 0$ if and only if there are positive measurable functions c and g, so that for all $t \in (t_0, \infty)$ with $t_0 > 0$:*

$$1 - F(t) = c(t) \exp\left[- \int_{t_0}^{t} \frac{ds}{g(s)} \right], \tag{6.30}$$

where

$$\lim_{t\to\infty} c(t) = c \in (0, \infty), \qquad \lim_{t\to\infty} \frac{g(t)}{t} = \gamma. \tag{6.31}$$

Remark 6.22 *Note that in the special case where $c(t) \equiv c$ and $g(t) \equiv t\gamma$, then 6.30 states that for $t > t_0$, $1 - F(t) \equiv c(t/t_0)^{-1/\gamma}$ and so F is exactly a Pareto distribution. Hence for all $t > t_0$,*

$$\frac{1 - F(tx)}{1 - F(t)} = x^{-1/\gamma},$$

giving a special case of the limiting result in 6.27 which there required $t \to \infty$ for general $F \in D(G_\gamma)$ with $\gamma > 0$.

But see also corollary 6.23 for the general case.

Proof. *If 6.30 is satisfied, then*

$$\frac{1 - F(tx)}{1 - F(t)} = \frac{c(tx)}{c(t)} \exp\left[- \int_{t}^{tx} \frac{ds}{g(s)} \right].$$

Given the above limits in 6.31, for any $\epsilon > 0$ there is a T so that for $t \geq T$:

$$\gamma - \epsilon \leq \frac{g(t)}{t} \leq \gamma + \epsilon.$$

Hence for $x \geq 1$:

$$\exp\left[- (\gamma - \epsilon)^{-1} \int_{t}^{tx} \frac{ds}{s} \right] \leq \exp\left[- \int_{t}^{tx} \frac{ds}{g(s)} \right] \leq \exp\left[- \int_{t}^{tx} (\gamma + \epsilon)^{-1} \frac{ds}{s} \right],$$

which results in:

$$x^{-1/(\gamma-\epsilon)} \leq \exp\left[- \int_{t}^{tx} \frac{ds}{g(s)} \right] \leq x^{-1/(\gamma+\epsilon)}$$

for $t \geq T$, $x \geq 1$. Since $\frac{c(tx)}{c(t)} \to 1$ for all $x \geq 1$, 6.30 obtains:

$$x^{-1/(\gamma-\epsilon)} \leq \lim_{t\to\infty} \frac{1-F(tx)}{1-F(t)} \leq x^{-1/(\gamma+\epsilon)}.$$

As ϵ is arbitrary, $F \in D(G_\gamma)$ with $\gamma > 0$ by proposition 6.19.

Conversely, assume that $F \in D(G_\gamma)$ with $\gamma > 0$ and define

$$d(t) = \frac{1-F(t)}{\int_t^\infty (1-F(x))\frac{dx}{x}}.$$

In proposition 6.25 of the next section we will prove that the integral in the definition of $d(t)$ is finite for all $t \geq 1$, and in 6.33 that $\lim_{t\to\infty} d(t) \to 1/\gamma$, a limit needed below. Assuming this for now:

$$\frac{d}{dt}\left(\ln \int_t^\infty (1-F(x))\frac{dx}{x}\right) = \frac{d(t)}{t},$$

and so if $t_0 \geq 1$,

$$\begin{aligned} -\int_{t_0}^t \frac{d(s)}{s}ds &= \ln \int_{t_0}^\infty (1-F(x))\frac{dx}{x} - \ln \int_t^\infty (1-F(x))\frac{dx}{x} \\ &= \ln \frac{1-F(t)}{d(t)} - \ln \frac{1-F(t_0)}{d(t_0)}. \end{aligned}$$

Rewriting:

$$\exp\left[-\int_{t_0}^t \frac{d(s)}{s}ds\right] = \frac{1-F(t)}{d(t)}\frac{d(t_0)}{1-F(t_0)},$$

and so

$$1-F(t) = d(t)\frac{1-F(t_0)}{d(t_0)}\exp\left[-\int_{t_0}^t \frac{d(s)}{s}ds\right].$$

Defining $c(t) = d(t)\frac{1-F(t_0)}{d(t_0)}$ and $\frac{1}{g(s)} = \frac{d(s)}{s}$ yields 6.30. Also, 6.31 follows from 6.33 proved below:

$$c \equiv \lim_{t\to\infty} c(t) = \frac{1}{\gamma}\int_{t_0}^\infty (1-F(x))\frac{dx}{x} \in (0,\infty),$$

and $\lim_{t\to\infty} \frac{g(t)}{t} = \gamma$ since $\lim_{t\to\infty} d(t) \to 1/\gamma$ by 6.33 as noted above. ■

The following corollary shows that for $F \in D(G_\gamma)$ with $\gamma > 0$, not only is F conditionally asymptotically Pareto, but the conditional distributions of F are bounded by Pareto distributions with arbitrarily close tail indexes of $\frac{1}{\gamma\pm\epsilon}$ if t is large enough.

Corollary 6.23 *Given a distribution function $F \in D(G_\gamma)$ with $\gamma > 0$ and c defined in 6.31, then for any $\epsilon > 0$ with $\epsilon < c/2$ there exists T so that for $t \geq T$ and all $x \geq 1$:*

$$(1-\epsilon)\, x^{-1/(\gamma-\epsilon)} \leq \frac{1-F(tx)}{1-F(t)} \leq (1+\epsilon)\, x^{-1/(\gamma+\epsilon)}. \tag{6.32}$$

Proof. *Given the above limits in 6.31, for any $\epsilon > 0$ with $\epsilon < 1$ there is a T so that for $t \geq T$:*

$$\gamma - \epsilon \leq \frac{g(t)}{t} \leq \gamma + \epsilon, \qquad c\,(1-\epsilon/3) \leq c(t) \leq c\,(1+\epsilon/3)\,.$$

As in the first part of the above proof we then have for $t \geq T$ and all $x \geq 1$:

$$\frac{1-\epsilon/3}{1+\epsilon/3} x^{-1/(\gamma-\epsilon)} \leq \frac{1-F(tx)}{1-F(t)} \leq \frac{1+\epsilon/3}{1-\epsilon/3} x^{-1/(\gamma+\epsilon)}.$$

The result now follows since $\frac{1+\epsilon/3}{1-\epsilon/3} \leq 1+\epsilon$ and $\frac{1-\epsilon/3}{1+\epsilon/3} \geq 1-\epsilon$. ■

2. If $F \in D(G_\gamma)$ with $\gamma > 0$, then $\gamma_H \approx \gamma$

In example 6.14 was noted that if $\gamma > 0$ and $F \in D(G_\gamma)$ is a Pareto distribution, then the Hill estimator γ_H is equal to the conditional maximum likelihood estimate for the parameter γ. More generally for general $F \in D(G_\gamma)$ with $\gamma > 0$, the conditional distribution $F(tx|t) \equiv \Pr[X \leq tx | X > t]$ defined for $x > 0$ is asymptotically Pareto as $t \to \infty$. So it seems logical that the Hill estimator should be applicable asymptotically as $t \to \infty$, and hence it should provide an approximation to the given γ for finite t. We derive this result in this section, but to do so requires an a new formula for γ.

The proposition below provides the needed formula for γ based on an integral of the distribution function F introduced above, and will support the justification that the Hill estimator γ_H approximates this exact value of γ for $F \in D(G_\gamma)$ with $\gamma > 0$. This formula for $\gamma > 0$ generalizes the von Mises' condition because it is valid for all such F without differentiability conditions. On the other hand, von Mises' condition is valid for all γ.

This proposition can also be formulated with modifications in the cases of $\gamma < 0$ and $\gamma = 0$, but we do not develop this theory.

Remark 6.24 *To perhaps state the obvious, as the conclusions of the next proposition were used in the proof of the Karamata representation theorem, we note that we will not be using the earlier result in the following proof!*

Proposition 6.25 *A distribution function $F \in D(G_\gamma)$ with $\gamma > 0$ if and only if $F(x) < 1$ for all $x > 0$,*

$$\int_1^\infty (1 - F(x)) \frac{dx}{x} < \infty,$$

and

$$\lim_{t\to\infty} \frac{\int_t^\infty (1 - F(x)) \frac{dx}{x}}{1 - F(t)} = \gamma. \tag{6.33}$$

Proof. *If $F \in D(G_\gamma)$ with $\gamma > 0$, then $F(x) < 1$ for all $x > 0$ follows from proposition 6.19 which assured $x^* = \infty$. By 6.27, for any $\epsilon > 0$ there is a T so that for $t \geq T$:*

$$\frac{1 - F(te)}{1 - F(t)} \leq (1 + \epsilon)\, e^{-1/\gamma} \leq e^{\epsilon - 1/\gamma}.$$

Since $te^k \geq T$ for $k \geq 0$:

$$\frac{1 - F(te^n)}{1 - F(t)} = \prod\nolimits_{k=1}^n \frac{1 - F(te^k)}{1 - F(te^{k-1})} \leq e^{n(\epsilon - 1/\gamma)}.$$

Given $x > 1$, let $n = \lceil \ln x \rceil$, the least integer greater than or equal to $\ln x$. Then $\lceil \ln x \rceil < \ln x + 1$ and substituting $x = e^{\ln x}$:

$$\frac{1 - F(tx)}{1 - F(t)} \leq e^{n(\epsilon - 1/\gamma)} \leq e^{\epsilon - 1/\gamma} x^{\epsilon - 1/\gamma}.$$

Consequently, for any $t \geq T$, $\frac{1-F(tx)}{1-F(t)} \frac{1}{x}$ is dominated by the integrable function $e^{\epsilon - 1/\gamma} x^{\epsilon - 1/\gamma - 1}$, and so

$$\int_1^\infty \frac{1 - F(Tx)}{1 - F(T)} \frac{dx}{x} < \infty,$$

proving the integrability of $(1 - F(x))/x$. Also, by Lebesgue's dominated convergence theorem of proposition 2.61 of book 3 and 6.27:

$$\begin{aligned} \lim_{t\to\infty} \frac{\int_t^\infty (1 - F(x)) \frac{dx}{x}}{1 - F(t)} &= \lim_{t\to\infty} \int_1^\infty \frac{1 - F(tx)}{1 - F(t)} \frac{dx}{x} \\ &= \int_1^\infty x^{-1/\gamma - 1} dx \\ &= \gamma, \end{aligned}$$

which is 6.33.

Conversely, given 6.33 define as in the proof of proposition 6.21,

$$d(t) = \frac{1 - F(t)}{\int_t^\infty (1 - F(x)) \frac{dx}{x}},$$

and note that $\frac{d(t)}{t} = -\frac{b'(t)}{b(t)}$ *with* $b(t) = \int_t^\infty (1 - F(x)) \frac{dx}{x}$. *Consequently,*

$$\int_1^t \frac{d(x)}{x} dx = -\ln \left[\int_t^\infty (1 - F(x)) \frac{dx}{x} \right] + \ln \left[\int_1^\infty (1 - F(x)) \frac{dx}{x} \right],$$

and so

$$\begin{aligned} 1 - F(t) &= d(t) \int_t^\infty (1 - F(x)) \frac{dx}{x} \\ &= d(t) \int_1^\infty (1 - F(x)) \frac{dx}{x} \exp \left[- \int_1^t \frac{d(x)}{x} dx \right]. \end{aligned}$$

With a similar expression for $1 - F(ty)$*:*

$$\begin{aligned} \frac{1 - F(ty)}{1 - F(t)} &= \frac{d(ty)}{d(t)} \exp \left[- \int_t^{ty} \frac{d(x)}{x} dx \right] \\ &= \frac{d(ty)}{d(t)} \exp \left[- \int_1^y \frac{d(tx)}{x} dx \right]. \end{aligned}$$

Letting $t \to \infty$ *obtains from 6.33 that* $\frac{d(ty)}{d(t)} \to 1$ *and* $d(tx) \to 1/\gamma$, *and so*

$$\lim_{t \to \infty} \frac{1 - F(ty)}{1 - F(t)} = y^{-1/\gamma}.$$

Consequently it must be the case that $F(x) < 1$ *for all* $x > 0$ *since otherwise* $\lim_{t \to \infty} \frac{1 - F(ty)}{1 - F(t)} = 0$ *for all* $y > 1$. *Thus* $x^* = \infty$ *and* $F \in D(G_\gamma)$ *with* $\gamma > 0$ *by proposition 6.19.* ■

With most of the hard work done and the integral formula for γ derived in 6.33, we will now demonstrate that the formula in 6.19 for the Hill estimator, γ_H , approximates the value of this integral. To this end, we begin with a transformation of the formula in 6.33.

Proposition 6.26 *Given a distribution function* $F \in D(G_\gamma)$ *with* $\gamma > 0$,

$$\gamma = \lim_{t \to \infty} \frac{\int_t^\infty \ln (x/t) \, dF(x)}{1 - F(t)}, \tag{6.34}$$

where the integral in the numerator is interpreted as a Riemann-Stieltjes integral.

Proof. *Given $F \in D(G_\gamma)$ with $\gamma > 0$ define $h(x) \equiv 1 - F(x)$ and $k(x) \equiv \ln(x/t)$ for fixed $t \geq 1$. Since k is an increasing function, define the Riemann-Stieltjes integral of monotonic $h(x)$ relative to k over bounded intervals by proposition 4.19 and remark 4.20 of book 3. Then by proposition 4.30 of book 3, since $k(x)$ is continuously differentiable:*

$$\int_t^N h(x)dk = \int_t^N h(x)k'(x)dx.$$

The integral on the right is defined as a Riemann integral since $k'(x)$ is continuous, and $h(x)$ is monotonic and thus by proposition 3.15 of book 3 is differentiable almost everywhere and so continuous almost everywhere. Substituting for $h(x)$ and $k'(x) = 1/x$ obtains:

$$\int_t^N h(x)dk = \int_t^N (1 - F(x))\frac{dx}{x}.$$

Since the limit as $N \to \infty$ of the integral on the right exists by proposition 6.25:

$$\lim_{N\to\infty} \int_t^N h(x)dk = \int_t^\infty (1 - F(x))\frac{dx}{x}.$$

Using the integration by parts formula for Riemann-Stieltjes integrals of proposition 4.15 of book 3:

$$\int_t^N h(x)dk = h(N)k(N) - h(t)\,k(t) - \int_t^N k(x)dh.$$

Now $h(t)$ is finite and $k(t) = 0$ by definition. Also, by 6.32 if $F \in D(G_\gamma)$ with $\gamma > 0$, then for any $\epsilon > 0$ there is a $C_\epsilon = (1+\epsilon)\,t^{1/(\gamma+\epsilon)}$ so that

$$1 - F(N) \leq C_\epsilon N^{-1/(\gamma+\epsilon)},$$

and hence $h(N)k(N) = (1 - F(N))\ln\left(\frac{N}{t}\right) \to 0$ as $N \to \infty$. Combining:

$$\int_t^\infty (1 - F(x))\frac{dx}{x} = -\lim_{N\to\infty} \int_t^N k(x)dh = -\int_t^\infty \ln(x/t)d(1 - F),$$

where the integral on the right is again defined as a Riemann-Stieltjes integral.

By the definition of Riemann-Stieltjes integral, using $-(1 - F)$ or F as an integrator function produces the same result. This plus the result in 6.33 now produces 6.34. ■

Remark 6.27 *The formula in 6.34 can be interpreted in the context of expectations as defined in 3.1. To this end, let X be a random variable defined on a probability space $(\mathcal{S},\mathcal{E},\lambda)$ with distribution function $F \in D(G_\gamma)$ with $\gamma > 0$, which exists by proposition 3.6 of book 2, and let $t \geq 1$ be fixed. Define the conditional distribution function $F_t(x)$ for $x \geq t$ by:*

$$F_t(x) \equiv \frac{F(x) - F(t)}{1 - F(t)}.$$

Because $F(t)$ is a constant, by proposition 4.26 of book 3:

$$\frac{\int_t^\infty \ln(x/t)\, dF}{1 - F(t)} = \int_t^\infty \ln(x/t)\, dF_t(x),$$

and by 3.1:

$$\int_t^\infty \ln(x/t)\, dF_t(x) = E\left[\ln(X/t) \,|X \geq t\right].$$

Thus by 6.34:

$$\gamma = \lim_{t\to\infty} \int_t^\infty \ln(x/t)\, dF_t(x), \tag{6.35}$$

or equivalently:

$$\gamma = \lim_{t\to\infty} E\left[\ln(X/t) \,|X \geq t\right]$$

With γ defined in 6.35 as a limit as $t \to \infty$, an approximation can be achieved by evaluating this expression for large enough t. Given a sample of random variates $\{X_i\}_{i=1}^n$ with distribution $F \in D(G_\gamma)$ with $\gamma > 0$ and associated order statistics, $\{X_{(j)}\}_{j=1}^n$, it follows that for n large and k small:

$$\gamma \approx \frac{\int_{X_{(n-k)}}^\infty \ln\left(x/X_{(n-k)}\right) dF}{1 - F(X_{(n-k)})}.$$

This is a nice formula, but for estimation of γ it is not yet useful because F is unknown.

Proposition 6.28 *Given a distribution function $F \in D(G_\gamma)$ with $\gamma > 0$, and variates $\{X_i\}_{i=1}^n$ with distribution F and order statistics, $\{X_{(j)}\}_{j=1}^n$, then for n large and k small:*

$$\gamma \approx \frac{1}{k} \sum\nolimits_{j=0}^{k-1} \left[\ln\left(X_{(n-j)}\right) - \ln\left(X_{(n-k)}\right)\right] = \gamma_H. \tag{6.36}$$

Proof. *The goal is to approximate the integral in 6.35 which is based on the unknown distribution function F, with an integral based on the empirical*

distribution function F_n implied by the given sample $\{X_i\}_{i=1}^n$. This empirical distribution function was introduced in 5.34, and assigns a probability of $\frac{1}{n}$ to each variate:

$$F_n(x) = \frac{1}{n}\sum\nolimits_{j=1}^{n} \chi_{(-\infty,x]}(X_j).$$

By the ***Glivenko-Cantelli theorem*** *in proposition 5.55, $\sup_x |F_n(x) - F(x)| \to 0$ with probability 1 and thus it seems reasonable to approximate:*

$$\int_{X_{(n-k)}}^{\infty} \ln\left(x/X_{(n-k)}\right) dF \approx \int_{X_{(n-k)}}^{\infty} \ln\left(x/X_{(n-k)}\right) dF_n.$$

The justification for this approximation might be that since n can be chosen so that $\sup_x |F_n(x) - F(x)| < \epsilon$, for any term in the defining Riemann-Stieltjes summations, $\sup_x |\Delta F_n - \Delta F| < 2\epsilon$. However while this seems intuitively plausible it cannot be made rigorous with the tools at hand because $\ln\left(x/X_{(n-k)}\right)$ is unbounded. However this can be formalized by a result in the book 6 section, General Results on Weak Convergence of Measures, in the following way. Convergence in the Glivenko-Cantelli theorem assures that $F_n(x) \Rightarrow F(x)$ and hence $\mu_{F_n} \Rightarrow \mu_F$ for the associated Borel measures. The book 6 result will state that weak convergence of measures then assures the convergence of the Lebesgue-Stieltjes integrals, which for continuous integrands assures convergence of the Riemann-Stieltjes integrals.

We proceed assuming that this approximation of the dF-integral by the dF_n-integral improves as $n \to \infty$. Since F_n is a step function, $1-F(X_{(n-k)}) = k/n$ and the Riemann-Stieltjes integral with respect to F_n is obtained in proposition 4.30 of book 3. We thus conclude that for n and $X_{(n-k)}$ large enough:

$$\begin{aligned}
\gamma &\approx \frac{\int_{X_{(n-k)}}^{\infty} \ln\left(x/X_{(n-k)}\right) dF_n}{1 - F_n(X_{(n-k)})} \\
&= \frac{\sum_{j=0}^{k-1} \ln\left(X_{(n-j)}/X_{(n-k)}\right)/n}{k/n} \\
&= \frac{1}{k}\sum\nolimits_{j=0}^{k-1} \left[\ln\left(X_{(n-j)}\right) - \ln\left(X_{(n-k)}\right)\right].
\end{aligned}$$

This final summation is seen to be the formula for the Hill estimator in 6.19.

Thus if $F \in D(G_\gamma)$ with $\gamma > 0$, then

$$\gamma_H \approx \gamma.$$

■

As noted in the proof, this approximation would be expected to improve as $n \to \infty$ since then $F_n(x)$ improves as an approximation to $F(x)$ by the Glivenko-Cantelli theorem. But we also require that $X_{(n-k)} \to \infty$ because the integration formula for γ in 6.34 is only exact in the limit as $t \to \infty$.

In the next section it will be demonstrated that under more clearly defined conditions that $\gamma_H \to_P \gamma$, which is to say, γ_H converges in probability to γ as defined in 5.24.

3. If $F \in D(G_\gamma)$ with $\gamma > 0$, then $\gamma_H \to_P \gamma$ as $n \to \infty$

From the derivation of proposition 6.28 that $\gamma_H \approx \gamma$ for $F \in D(G_\gamma)$ with $\gamma > 0$, it was noted that this approximation should improve as the sample size n increases, but also as the initial order statistic $X_{(n-k)}$ used in γ_H increases without bound. The statement about n is clear from the Glivenko-Cantelli theorem, but when thinking about the order statistic $X_{(n-k)}$ we need to clarify what this second condition means. Recall how this requirement originated. The denominator of the integral formula for γ in 6.34 is $1 - F(t)$, and the limit $t \to \infty$ assures $1 - F(t) \to 0$. However $t \equiv X_{(n-k)}$ in the approximation above, and then $1 - F(X_{(n-k)}) = k/n$, so $X_{(n-k)} \to \infty$ requires $k/n \to 0$. In other words, the lowest order statistic used in the Hill estimator must be based on a quantile $q_{k_n} \equiv 1 - k/n$ with the property that $q_{k_n} \to 1$. This conclusion is also implied by corollary 5.18 which states that if $q_{k_n} \to q < 1$ as $n \to \infty$, then the normalized **uniform** variate $\left(Y_{(k_n)} - q\right) / \sqrt{\frac{q(1-q)}{n}}$ converges to the standard normal distribution, and hence the variate $Y_{(k_n)}$ cannot increase and approach 1. Indeed, one then has that $Y_{(k_n)} \to_P q$, which is to say that for any $\epsilon > 0$,

$$\Pr[\left|Y_{(k_n)} - q\right| > \epsilon] \to 0,$$

and it then follows that $X_{(n-k)} \equiv F^*(Y_{(k_n)}) \nrightarrow \infty$.

Now $n \to \infty$ and $k/n \to 0$ certainly allows for k to be bounded. But in the above proof k also reflected the number of terms in the approximating Riemann-Stieltjes summation, and thus logically we must have $k \to \infty$ in order for these sums to well-approximate the Riemann-Stieltjes integral. This requirement will be seen to be needed in the proof below in order to justify an application of the central limit theorem.

In summary, the proposition below will formalize the result that if $k \to \infty$ and $k/n \to 0$ as $n \to \infty$, then $\gamma_H \to_P \gamma$.

In order to prove this section's result a more general version of **Karamata's representation theorem** referenced above is needed. In the special case of proposition 6.21 this theorem stated that for $F \in D(G_\gamma)$ with $\gamma > 0$ that $1 - F(t)$ could be represented as an integral in 6.30, and with functions with limiting properties summarized in 6.31. As noted there, this result also applies to $F \in D(G_\gamma)$ with $\gamma < 0$ or $\gamma = 0$ but with appropriate modifications to the limiting properties of the functions in 6.31. But far beyond this, Karamata's representation theorem applies not only to distribution functions, but to all functions f that are **regularly varying at infinity with index** $\alpha \in \mathbb{R}$. As noted in 6.25, this terminology means that for all $x \geq x_0 > 0$:

$$\lim_{t\to\infty} \frac{f(tx)}{f(t)} = x^\alpha.$$

For example, the result in 6.23 states that if $F \in D(G_\gamma)$ with $\gamma > 0$, then U is **regularly varying at infinity with index** γ, while 6.27 states that $1 - F$ is **regularly varying at infinity with index** $-1/\gamma$.

Our application of this general result will again be for the case $\alpha > 0$. The importance of this more general representation theorem is that it can then be applied to the function $U(t) \equiv \left(\frac{1}{1-F}\right)^* (t)$, the left continuous inverse of $\frac{1}{1-F}$ for $F \in D(G_\gamma)$, which is not a distribution function. This representation theorem will then provide the needed uniform estimate of $U(tx)/U(t)$ for $t \in (t_0, \infty)$ and all $x \geq 1$, just as proposition 6.21 provided such a uniform estimate of $\frac{1-F(tx)}{1-F(t)}$ in 6.32 of corollary 6.23.

We state without proof this general result though again will only require the case $\alpha > 0$.

Proposition 6.29 (Karamata's Representation theorem) *A function f is regularly varying at infinity with index α if and only if there are positive measurable functions c and g so that for all $t \in (t_0, \infty)$ with $t_0 > 0$:*

$$f(t) = c(t) \exp\left[\int_{t_0}^{t} \frac{h(s)}{s} ds\right], \tag{6.37}$$

where

$$\lim_{t\to\infty} c(t) = c \in (0, \infty), \qquad \lim_{t\to\infty} h(t) = \alpha. \tag{6.38}$$

Proof. *See de Haan and Ferreira, Theorem B.1.6.* ■

The proof of the needed corollary now follows the proof of corollary 6.23.

Corollary 6.30 *Given a function f that is regularly varying at infinity with index α, then for any $\epsilon > 0$ with $\epsilon < 1$ there is a T so that for $t \geq T$ and all $x \geq 1$:*

$$(1-\epsilon)\, x^{(\alpha-\epsilon)} \leq \frac{f(tx)}{f(t)} \leq (1+\epsilon)\, x^{(\alpha+\epsilon)}. \tag{6.39}$$

Proof. *If 6.37 is satisfied for $t > t_0$, then for $x > 1$:*

$$\frac{f(tx)}{f(t)} = \frac{c(tx)}{c(t)} \exp\left[\int_t^{tx} \frac{h(s)}{s} ds\right].$$

Given the above limit in 6.38, for any $\epsilon > 0$ with $\epsilon < 1$ there is a T so that for $t \geq T$:

$$\alpha - \epsilon \leq h(t) \leq \alpha + \epsilon, \qquad c\,(1-\epsilon/3) \leq c(t) \leq c\,(1+\epsilon/3)\,.$$

Hence for $x > 1$:

$$\exp\left[(\alpha-\epsilon)\int_t^{tx} \frac{ds}{s}\right] \leq \exp\left[\int_t^{tx} \frac{h(s)}{s} ds\right] \leq \exp\left[\int_t^{tx} (\alpha+\epsilon)\, \frac{ds}{s}\right],$$

which is equivalent to

$$x^{(\alpha-\epsilon)} \leq \exp\left[\int_t^{tx} \frac{h(s)}{s} ds\right] \leq x^{(\alpha+\epsilon)}.$$

These bounds then produce

$$\frac{1-\epsilon/3}{1+\epsilon/3} x^{(\alpha-\epsilon)} \leq \frac{f(tx)}{f(t)} \leq \frac{1+\epsilon/3}{1-\epsilon/3} x^{(\alpha+\epsilon)}.$$

and the result in 6.39 now follows since $\frac{1+\epsilon/3}{1-\epsilon/3} \leq 1+\epsilon$ and $\frac{1-\epsilon/3}{1+\epsilon/3} \geq 1-\epsilon$. ■

We now turn to the result that if $F \in D(G_\gamma)$ with $\gamma > 0$, then as $n \to \infty$ the Hill estimator γ_H converges to γ in probability, recalling the definition in 5.24. To achieve this convergence result we will also require that $k \to \infty$ and the implied quantile $q_{k_n} \equiv 1 - k/n$ of the base order statistic $X_{(n-k)}$ converges to 1, or equivalently $k/n \to 0$.

Proposition 6.31 (Hill Estimator 1) *Let $\{X_i\}_{i=1}^n$ be independent and identically distributed random variables with distribution function $F \in D(G_\gamma)$ with $\gamma > 0$, and γ_H the Hill estimator defined in 6.19. Then if $k \to \infty$ and $k/n \to 0$ as $n \to \infty$, the estimator γ_H converges in probability to γ :*

$$\gamma_H \to_P \gamma. \tag{6.40}$$

Proof. *Applying corollary 6.30 to the left continuous inverse function* $U(t) \equiv \left(\frac{1}{1-F}\right)^*(t)$, *which is regularly varying at infinity with index* γ *by 6.23, obtains by 6.39 that for any* $\epsilon > 0$ *there is a* T *so that for* $t \geq T$ *and all* $x \geq 1$:

$$(1-\epsilon)\, x^{(\gamma-\epsilon)} \leq \frac{U(tx)}{U(t)} \leq (1+\epsilon)\, x^{(\gamma+\epsilon)},$$

and so

$$\ln(1-\epsilon) + (\gamma-\epsilon)\ln x \leq \ln U(tx) - \ln U(t) \leq \ln(1+\epsilon) + (\gamma+\epsilon)\ln x.$$

Recall that $U(t) \equiv F^*(1-1/t)$ *and so by proposition 4.9 of book 2,* $\{U(Y_j)\}_{i=1}^n$ *will be independent and have distribution function* F *if* $\{1-1/Y_j\}_{i=1}^n$ *are independent and uniformly distributed on* $[0,1]$. *By proposition 4.8 of book 2, if* $\{Y_j\}_{i=1}^n$ *are independent random variables with continuous distribution function* $G(y) = 1-1/y$ *on* $y \geq 1$, *then* $\{G(Y_j)\}_{i=1}^n = \{1-1/Y_j\}_{i=1}^n$ *will be uniformly distributed on* $[0,1]$ *and independent by proposition 4.9. And so for such variates,* $\{U(Y_j)\}_{i=1}^n$ *will have distribution function* F. *In summary,* γ_H *can be defined with 6.20 in terms of the order statistics of a sample* $\{U(Y_j)\}_{i=1}^n$:

$$\gamma_H = \frac{1}{k}\sum_{j=0}^{k-1}\left[\ln U(Y_{(n-j)}) - \ln U(Y_{(n-k)})\right],$$

where $\{Y_j\}_{i=1}^n$ *is a sample with distribution function* $G(y) = 1-1/y$.

To use the above bounds for $\ln U(tx) - \ln U(t)$, *let* $t = Y_{(n-k)}$. *Then by proposition 5.48, if* $k/n \to 0$ *then* $Y_{(n-k)} \to y^* = \infty$ *with probability* 1, *and so it can be assumed that for any* ϵ *and associated* T *that* $Y_{(n-k)} \geq T$ *eventually with probability* 1. *With* $x \equiv Y_{(n-j)}/Y_{(n-k)}$, *it then follows from the above bounds that:*

$$\ln(1-\epsilon) + (\gamma-\epsilon)\, Z_{k,n} \leq \gamma_H \leq \ln(1+\epsilon) + (\gamma+\epsilon)\, Z_{k,n}, \qquad ((*))$$

where

$$Z_{k,n} \equiv \frac{1}{k}\sum_{j=0}^{k-1}\ln\left[\frac{Y_{(n-j)}}{Y_{(n-k)}}\right].$$

We now show that

$$Z_{k,n} \to_P 1.$$

First note that if Y *has distribution function* $G(y) = 1-1/y$ *on* $y \geq 1$, *then* $X \equiv \ln Y$ *has a standard exponential distribution since for* $x \geq 0$:

$$F_X(x) = \Pr[\ln Y \leq x] = \Pr[Y \leq e^x] = 1 - e^{-x}.$$

Thus proposition 5.21 applies to $Z_{k,n}$. Specifically, if $F_{k,n}$ denotes the distribution function of $Z'_{k,n}$, defined by

$$Z'_{k,n} = \frac{Z_{k,n} - 1}{1/\sqrt{k}},$$

then as $n \to \infty$ and $k/n \to 0$, $F_{k,n}$ converges in distribution to the normal distribution:

$$F_{k,n} \Rightarrow \Phi.$$

So for any $\delta > 0$:

$$\Pr\left[|Z_{k,n} - 1| \geq \delta\right] = \Pr\left[\left|Z'_{k,n}\right| \geq \delta\sqrt{k}\right] \to 0,$$

since $k \to \infty$. Thus, $Z_{k,n} \to_P 1$.

This now obtains 6.40 as follows. Given $\delta > 0$:

$$\begin{aligned} \Pr\left[\gamma_H - \gamma \geq \delta\right] &\leq \Pr\left[\ln(1+\epsilon) + (\gamma+\epsilon) Z_{k,n} - \gamma \geq \delta\right] \\ &= \Pr\left[Z_{k,n} - 1 \geq (\delta - \epsilon - \ln(1+\epsilon)) / (\gamma+\epsilon)\right], \end{aligned}$$

and $(\delta - \epsilon - \ln(1+\epsilon)) / (\gamma+\epsilon) > 0$ for ϵ small and so $\Pr\left[\gamma_H - \gamma \geq \delta\right] \to 0$. Similarly,

$$\begin{aligned} \Pr\left[\gamma_H - \gamma \leq -\delta\right] &\leq \Pr\left[\ln(1-\epsilon) + (\gamma-\epsilon) Z_{k,n} - \gamma \leq -\delta\right] \\ &= \Pr\left[Z_{k,n} - 1 \leq (-\delta + \epsilon - \ln(1-\epsilon)) / (\gamma-\epsilon)\right], \end{aligned}$$

and $(-\delta + \epsilon - \ln(1-\epsilon)) / (\gamma-\epsilon) < 0$ for ϵ small and so $\Pr\left[\gamma_H - \gamma \leq -\delta\right] \to 0$. Thus

$$\Pr\left[|\gamma_H - \gamma| \geq \delta\right] \to 0,$$

which is 6.40. ■

Remark 6.32 *The above theorem has a converse, which we state without proof.*

Proposition 6.33 *Let $\{X_i\}_{i=1}^n$ be independent and identically distributed random variables with distribution function F, and γ_H the Hill estimator defined in 6.19. Assume that there exists a sequence $k_n \to \infty$ with $k_n/n \to 0$ and $k_{n+1}/k_n \to 1$ as $n \to \infty$ so that the estimator γ_H converges in probability to a constant γ :*

$$\gamma_H \to_P \gamma > 0.$$

Then $F \in D(G_\gamma)$.

Proof. *See de Haan and Ferreira, Theorem 3.2.4.* ■

Asymptotic Normality of the Hill Estimator

Although we do not develop this theory here, it is known that under additional assumptions on the distribution function $F \in D(G_\gamma)$ with $\gamma > 0$, that the Hill estimator is asymptotically normally distributed. Recall that by corollary 9.35 of book 2 that if $F \in D(G_\gamma)$ then for $x > 0$:

$$\frac{U(tx) - U(t)}{a(t)} \to \frac{x^\gamma - 1}{\gamma}$$

as $t \to \infty$, with $U(t) \equiv \left(\frac{1}{1-F}\right)^* (t)$ the left continuous inverse of $\frac{1}{1-F}$. Here, c in the earlier formula is integrated into the definition of $a(t)$. The right hand limit is defined to be $\ln x$ when $\gamma = 0$, which equals $\lim_{\gamma \to 0} (x^\gamma - 1)/\gamma$.

In order to obtain the asymptotic normality result this function F must satisfy an additional assumption known as a **second-order condition,** which provides information on the rate of convergence in the above limit. Specifically, we say that $F \in D(G_\gamma)$ satisfies a second-order condition if there is a function $A(t)$ with $\lim_{t\to\infty} A(t) = 0$ and where $A(t)$ does not change sign for $t \geq T$ say, and a function $H(x)$, so that as $t \to \infty$:

$$\left[\frac{U(tx) - U(t)}{a(t)} - \frac{x^\gamma - 1}{\gamma}\right] \Big/ A(t) \to H(x). \tag{6.41}$$

By 6.22 this is equivalent to:

$$\left[\frac{U(tx)}{U(t)} - x^\gamma\right] \Big/ A(t) \to H(x).$$

For well-definedness, it is required that $H(x)$ is not a multiple of $\frac{x^\gamma - 1}{\gamma}$.

Much like the development surrounding the Fisher-Tippett-Gnedenko theorem, it turns out that when such a function $H(x)$ exists it must have a well defined structure, and when $\gamma > 0$, the assumption underlying the above Hill estimator theory, it is the case that:

$$H(x) = x^\gamma \frac{x^\rho - 1}{\rho},$$

where $\rho \leq 0$. When $\rho = 0$,

$$H(x) \equiv x^\gamma \ln x,$$

which is equal to $x^{\gamma} \lim_{\rho \to 0} (x^{\rho} - 1)/\rho$. The significance of the parameter ρ is that it is then the case that $A(t) \in RV_{\rho}$, meaning that $A(t)$ is regularly varying at infinity with index ρ.

We then have the following result.

Proposition 6.34 (Hill Estimator 2) *Let $F \in D(G_{\gamma})$ with $\gamma > 0$ and assume that F satisfies a second order condition where $\rho \leq 0$. Let $F_{\gamma'_H}$ denote the distribution function of the normalized Hill estimator*

$$\gamma'_H = \frac{\gamma_H - \gamma}{1/\sqrt{k}}.$$

Then if $k \to \infty$ and $k/n \to 0$ as $n \to \infty$, and

$$\lambda \equiv \lim_{n \to \infty} \sqrt{k} A(n/k)$$

is finite, then:

$$F_{\gamma_H} \Rightarrow N\left(\lambda/(1-\rho), \gamma^2\right), \tag{6.42}$$

where $N\left(\lambda/(1-\rho), \gamma^2\right)$ denotes the normal distribution with mean $\lambda/(1-\rho)$ and variance γ^2.

Proof. *See de Haan and Ferreira, Theorem 3.2.5.* ■

Remark 6.35 *Note that since $n/k \to \infty$ it follows that $A(n/k) \to 0$ by the definition of second order condition. However, the assumption that $k/n \to 0$ implies that k/n may approach 0 at any rate as a function of n, and it is thus also the case that n/k may approach ∞ at any rate as a function of n. So the value of the parameter $\lambda = \lim_{n \to \infty} \sqrt{k_n} A(n/k_n)$ can depend on the actual sequence $\{k_n\}$ of parameters used, and need not be finite.*

6.2.2 The Pickands–Balkema–de Haan Theorem: $\gamma > 0$

Like the Fisher-Tippett-Gnedenko theorem, the Pickands–Balkema–de Haan theorem addresses the question of the existence of a limiting distribution. Both begin with a distribution function F of a random variable X. The former theorem investigates $F^n(x)$, the distribution function of $M_n = \max_{m \leq n}\{X_m\}$ for independent $\{X_m\}_{m=1}^n$, and investigates one key question related to the limiting distribution of $F^n(x)$ as $n \to \infty$. Namely, if there exists sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$ where $a_n > 0$ for all n, and a nondegenerate distribution function $G(x)$ so that $F^n(a_n x + b_n) \Rightarrow G(x)$, what can be said about G? The answer provided by this theorem is proposition 9.30 of book 2 as noted above and states that if

such sequences and distribution exist, so $F \in \mathcal{D}(G)$ in the notation of **domains of attraction**, then there are real constants $A > 0$, B, and γ so that $G(x) = G_\gamma(Ax + B)$ with $G_\gamma(x)$ defined for $\gamma \neq 0$ by:

$$G_\gamma(x) = \exp\left(-(1+\gamma x)^{-1/\gamma}\right), \qquad 1 + \gamma x \geq 0.$$

When $\gamma = 0$, $G_\gamma(x)$ is defined on $\mathbb{R}$:

$$G_0(x) \equiv \exp\left(-e^{-x}\right).$$

Proposition 6.19 then provided a characterization of such $F \in \mathcal{D}(G_\gamma)$ for $\gamma > 0$ in 6.27, that for $x > 0$:

$$\lim_{t\to\infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-1/\gamma}.$$

This is restated in an even more descriptive way in 6.29, that if $F \in D(G_\gamma)$ for $\gamma > 0$, then as $x \to \infty$:

$$F(x) = 1 - L(x)x^{-1/\gamma}, \qquad L \in RV_0,$$

where $L \in RV_0$ means that L is **slowly varying at infinity** as in definition 6.16.

The Pickands–Balkema–de Haan theorem investigates another "tail" distribution, specifically the **conditional probability distribution of exceedances**, and the analysis underlying this result is often referred to as the **peaks over threshold** method. This investigation was initiated in book 2, where in that book's proposition 9.38 the following result was proved:

Proposition *If $F \in D(G_\gamma)$ for any γ, then:*

- *For all x with $x > -1/\gamma$ when $\gamma \geq 0$, where $-1/0 \equiv -\infty$, or,*
- *For $0 \leq x < -1/\gamma$ when $\gamma < 0$,*

and with x^* as defined in 5.31:

$$\lim_{t\to x^*} \frac{1 - F(t + x h_a(t))}{1 - F(t)} = (1+\gamma x)^{-1/\gamma}.$$

Here

$$h_a(t) \equiv a\left(\frac{1}{1 - F(t)}\right),$$

with $a(t)$ defined in terms of the normalizing function in 6.21. When $\gamma = 0$ the limit function is defined for all x as $\exp(-x)$.

This result can also be expressed as a conditional probability statement:

$$\lim_{t \to x^*} \Pr\left[X \leq t + x h_a(t) | X > t\right] = 1 - (1 + \gamma x)^{-1/\gamma}.$$

Alternatively, for t "large" relative to x^*:

$$\Pr\left[X \leq t + y | X > t\right] \approx 1 - \left(1 + \frac{\gamma}{h_a(t)} y\right)^{-1/\gamma}. \tag{6.43}$$

This limiting distribution is an example of the **generalized Pareto distribution,** $H_{\gamma,0,\beta(t)}(x)$ with $\beta(t) \equiv h_a(t)$.

Definition 6.36 *A **generalized Pareto distribution** $H_{\gamma,t,\beta}(x)$ is defined by:*

$$H_{\gamma,t,\beta}(x) \equiv 1 - \left(1 + \frac{\gamma}{\beta}(x - t)\right)^{-1/\gamma}, \tag{6.44}$$

and when $\gamma = 0$ defined as the limit of $H_{\gamma,t,\beta}(x)$ as $\gamma \to 0$:

$$H_{0,t,\beta}(x) \equiv 1 - \exp\left(-(x - t)/\beta\right). \tag{6.45}$$

The distribution $H_{\gamma,t,\beta}(x)$ is defined for $x \geq t$ when $\gamma \geq 0$, and for $t \leq x \leq t - \beta/\gamma$ when $\gamma < 0$.

Remark 6.37 *1. For the result below, we will be primarily interested in $H_{\gamma,0,\beta}(x)$ with $\gamma > 0$:*

$$H_{\gamma,0,\beta}(x) \equiv 1 - \left(1 + \frac{\gamma}{\beta} x\right)^{-1/\gamma}, \tag{6.46}$$

but introduced the more general notation because it is commonly cited.

2. Note that for $\gamma > 0$:

$$H_{\gamma,t,\gamma t}(x) \equiv 1 - \left(\frac{x}{t}\right)^{-1/\gamma}, \qquad x \geq t,$$

and:

$$H_{\gamma,0,\gamma t}(x) \equiv 1 - \left(1 + \frac{x}{t}\right)^{-1/\gamma}, \qquad x \geq 0,$$

*representing two common parametrizations for a standard **Pareto distribution.** In this context, it is common to represent the exponential index by α and so $\alpha = 1/\gamma > 0$.*

The asymptotic result above for the conditional distribution function can be improved as stated in proposition 9.44 of book 2, but there without proof. With the aid of corollary 6.23 to **Karamata's Representation theorem,** this earlier result can be proved in the case $\gamma > 0$ in which case $x^* = \infty$ by proposition 6.19.

Proposition 6.38 (Pickands–Balkema–de Haan theorem II) *Assume that F is in the* ***domain of attraction of*** G_γ, $F \in \mathcal{D}(G_\gamma)$ *with $\gamma > 0$, and given $t \geq 0$ define the conditional distribution function $F_t(y)$ for $y \geq 0$ by:*

$$F_t(y) \equiv \Pr\left[X \leq t + y | X > t\right] = \frac{F(t+y) - F(t)}{1 - F(t)}.$$

Then the approximation in 6.43 is uniform in y in the sense that there exists a positive function $\beta(t)$, so that:

$$\lim_{t\to\infty} \sup_{0\leq y<\infty} \left|F_t(y) - H_{\gamma,0,\beta(t)}(y)\right| = 0. \tag{6.47}$$

In fact 6.47 is true with $\beta(t) \equiv \gamma t$, so $H_{\gamma,0,\beta(t)}(y) \equiv H_{\gamma,0,\gamma t}(y)$, and thus $F_t(y)$ is asymptotically Pareto:

$$\Pr\left[X \leq t + y | X > t\right] \approx 1 - \left(1 + \frac{y}{t}\right)^{-1/\gamma} \text{ as } t \to \infty, \tag{6.48}$$

and the error in this approximation converges to 0 uniformly in $y \geq 0$.

In addition, if $\beta(t)$ is another function which satisfies 6.47, then $\beta(t)/\gamma t \to 1$ as $t \to \infty$. Thus $\beta(t) \equiv \gamma t$ is asymptotically unique.

Proof. *From corollary 6.23, given a distribution function $F \in D(G_\gamma)$ with $\gamma > 0$ and c defined in 6.31, then for any $\epsilon > 0$ with $\epsilon < c/2$ there is a T so that for $t \geq T$ and all $x \geq 1$, we have 6.32:*

$$(1-\epsilon)\, x^{-1/(\gamma-\epsilon)} \leq \frac{1 - F(tx)}{1 - F(t)} \leq (1+\epsilon)\, x^{-1/(\gamma+\epsilon)}.$$

Writing $t + y = t(1 + y/t)$, $y \geq 0$, then for $t \geq T$:

$$(1-\epsilon)\,(1+y/t)^{-1/(\gamma-\epsilon)} \leq \frac{1 - F(t+y)}{1 - F(t)} \leq (1+\epsilon)\,(1+y/t)^{-1/(\gamma+\epsilon)}.$$

Hence,

$$\begin{aligned}
&(1-\epsilon)\left(1 + \frac{y}{t}\right)^{-1/(\gamma-\epsilon)} - \left(1 + \frac{\gamma}{\beta(t)} y\right)^{-1/\gamma} \\
\leq\ & \frac{1 - F(t+y)}{1 - F(t)} - \left(1 + \frac{\gamma}{\beta(t)} y\right)^{-1/\gamma} \\
\leq\ & (1+\epsilon)\left(1 + \frac{y}{t}\right)^{-1/(\gamma+\epsilon)} - \left(1 + \frac{\gamma}{\beta(t)} y\right)^{-1/\gamma}.
\end{aligned}$$

This provides bounds for the difference between $F_t(y)$ *and* $H_{\gamma,0,\beta(t)}(y)$*, a difference we seek to evaluate in terms of 6.47.*

The proof of 6.47 will be completed by proving that for $\beta(t) \equiv \gamma t$*, that the supremum in* y *of both these bounds converges to* 0 *as* $t \to x^*$*. To this end we investigate the upper bound and leave the lower bound as an exercise. With* $\beta(t) \equiv \gamma t$ *the upper bound can be expressed:*

$$U(y) \equiv (1+\epsilon)\left(1+\frac{y}{t}\right)^{-1/(\gamma+\epsilon)} - \left(1+\frac{y}{t}\right)^{-1/\gamma}.$$

Letting $w = y/t$*,* $a \equiv 1/(\gamma+\epsilon)$ *and* $\delta \equiv \epsilon/\gamma$ *obtains for* $0 \le w \le \infty$*:*

$$U(w) = (1+\gamma\delta)(1+w)^{-a} - (1+w)^{-a(1+\delta)}.$$

Now $U(0) = \gamma\delta$*,* $U(\infty) = 0$*,* $U(w) \ge 0$ *and* $U(w)$ *can be differentiated to reveal that* $U'(w) \le 0$ *for all* w*, and thus:*

$$U(w) \le \gamma\delta = \epsilon.$$

Since ϵ *can be made arbitrarily small by choosing* t *large, the proof of 6.47 with* $\beta(t) \equiv \gamma t$ *is complete.*

Finally, assume that 6.47 is true for given $\beta(t)$*. Then*

$$\begin{aligned}\left|2^{-1/\gamma} - \left(1+\frac{\gamma}{\beta(t)}t\right)^{-1/\gamma}\right| &= \left|H_{\gamma,0,\gamma t}(t) - H_{\gamma,0,\beta(t)}(t)\right| \\ &\le \sup_{0\le y<\infty}\left|H_{\gamma,0,\gamma t}(y) - H_{\gamma,0,\beta(t)}(y)\right| \\ &\le \sup_{0\le y<\infty}\left|H_{\gamma,0,\gamma t}(y) - F_t(y)\right| + \sup_{0\le y<\infty}\left|F_t(y) - H_{\gamma,0,\beta(t)}(y)\right|.\end{aligned}$$

As $t \to \infty$ *the first supremum converges to* 0 *as proved above, while the second converges to* 0 *by assumption. Thus as* $t \to \infty$*:*

$$\left|2^{-1/\gamma} - \left(1+\frac{\gamma}{\beta(t)}t\right)^{-1/\gamma}\right| \to 0,$$

which obtains $\beta(t)/\gamma t \to 1$*.* ■

Remark 6.39 *While the Pareto distribution* $H_{\gamma,0,\gamma t}(y)$ *is the exact asymptotic limit for the conditional distribution function* $F_t(y)$ *as* $t \to \infty$*, it is common in applications to assume the more general model of the generalized Pareto distribution,* $H_{\gamma,0,\beta}(y)$*. Given the chosen threshold* t *and data set his approach provides two parameters to be determined by maximum likelihood or other estimation method rather than one. The desirability of two parameters is reinforced by the fact that the convergence to Pareto can be very slow indeed.*

Example 6.40 *As an illustration of the potential for slow convergence we follow Makarov (2007) with* $F(x) \equiv 1 - \frac{\ln x}{x}$. *Then since:*

$$\lim_{t\to\infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-1},$$

$F \in \mathcal{D}(G_1)$ *by 6.27. Also:*

$$F_t(y) = 1 - \frac{t}{t+y}\frac{\ln(t+y)}{\ln t}, \qquad H_{1,0,t}(y) = 1 - \frac{t}{t+y},$$

and $\sup_y [H_{1,0,t}(y) - F_t(y)]$ *is found by calculus to occur at* $\hat{y} \equiv (e-1)t$. *Thus*

$$\sup_y [H_{1,0,t}(y) - F_t(y)] = \frac{1}{e \ln t},$$

which converges to zero very slowly as $t \to \infty$.

References

I have listed below a number of textbook references for the mathematics and finance presented in this series of books. All provide both theoretical and applied materials in their respective areas that are beyond those developed here and are worth pursuing by those interested in gaining a greater depth or breadth of knowledge. This list is by no means complete and is intended only as a guide to further study. In addition, various published research papers have been identified in the some chapters in the locations where these results were discussed.

The reader will no doubt observe that the mathematics references are somewhat older than the finance references and upon web searching will find that several of the older texts in each category have been updated to newer editions, sometimes with additional authors. Since I own and use the editions below, I decided to present these editions rather than reference the newer editions which I have not reviewed. As many of these older texts are considered "classics", they are also likely to be found in university and other libraries.

That said, there are undoubtedly many very good new texts by both new and established authors with similar titles that are also worth investigating. One that I will at the risk of immodesty recommend for more introductory materials on mathematics, probability theory and finance is:

1. Reitano, Robert, R. *Introduction to Quantitative Finance: A Math Tool Kit.* Cambridge, MA: The MIT Press, 2010.

Topology, Measure, and Integration

2. Dugundji, James. *Topology.* Boston, MA: Allyn and Bacon, 1970.

3. Doob, J. L. *Measure Theory.* New York, NY: Springer-Verlag, 1994.

4. Edwards, Jr., C. H. *Advanced Calculus of Several Variables.* New York, NY: Academic Press, 1973.

5. Gemignani, M. C. *Elementary Topology.* Reading, MA: Addison-Wesley Publishing, 1967.

6. Halmos, Paul R. *Measure Theory.* New York, NY: D. Van Nostrand, 1950.

7. Royden, H. L. *Real Analysis,* 2nd Edition. New York, NY: The MacMillan Company, 1971.

8. Rudin, Walter. *Principals of Mathematical Analysis,* 3rd Edition. New York, NY: McGraw-Hill, 1976.

9. Rudin, Walter. *Real and Complex Analysis,* 2nd Edition. New York, NY: McGraw-Hill, 1974.

10. Shilov, G. E., and B. L. Gurevich. Integral, Measure & Derivative: A Unified Approach. New York, NY: Dover Publications, 1977.

Probability Theory & Stochastic Processes

11. Billingsley, Patrick. *Probability and Measure,* 3rd Edition. New York, NY: John Wiley & Sons, 1995.

12. Chung, K. L., and R. J. Williams. *Introduction to Stochastic Integration.* Boston, MA: Birkhäuser, 1983.

13. Davidson, James. *Stochastic Limit Theory.* New York, NY: Oxford University Press, 1997.

14. de Haan, Laurens, and Ana Ferreira. *Extreme Value Theory, An Introduction.* New York, NY: Springer Science, 2006.

15. Durrett, Richard. *Probability: Theory and Examples,* 2nd Edition. Belmont, CA: Wadsworth Publishing, 1996.

16. Durrett, Richard. *Stochastic Calculus, A Practical Intriduction.* Boca Raton, FL: CRC Press, 1996.

17. Feller, William. *An Introduction to Probability Theory and Its Applications,* Volume I. New York, NY: John Wiley & Sons, 1968.

18. Feller, William. *An Introduction to Probability Theory and Its Applications,* Volume II, 2nd Edition. New York, NY: John Wiley & Sons, 1971.

19. Friedman, Avner. *Stochastic Differential Equations and Application, Volume 1 and 2.* New York, NY: Academic Press, 1975.

20. Ikeda, Nobuyuki, and Shinzo Watanabe. Stochastic Differential Equations and Diffusion Processes. Tokyo, Japan: Kodansha Scientific, 1981.

21. Karatzas, Ioannis, and Steven E. Shreve. *Brownian Motion and Stochastic Calculus.* New York, NY: Springer-Verlag, 1988.

22. Kloeden, Peter E., and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations.* New York, NY: Springer-Verlag, 1992.

23. Nelson, Roger B. *An Introduction to Copulas,* 2nd Edition. New York, NY: Springer Science, 2006.

24. Øksendal, Bernt. *Stochastic Differential Equations, An Introduction with Applications,* 5th Edition. New York, NY: Springer-Verlag, 1998.

25. Protter, Phillip. *Stochastic Integration and Differential Equations, A New Approach.* New York, NY: Springer-Verlag, 1992.

26. Revuz, Daniel, and Marc Yor. *Continuous Martingales and Brownian Motion,* 3rd Edition. New York, NY: Springer-Verlag, 1999.

27. Rogers, L. C. G., and D. Williams. *Diffusions, Markov Processes and Martingales*, Volume 1, Foundations, 2nd Edition. Cambridge, UK: Cambridge University Press, 2000.

28. Rogers, L. C. G., and D. Williams. *Diffusions, Markov Processes and Martingales*, Volume 2, Itô Calculus 2nd Edition. Cambridge, UK: Cambridge University Press, 2000.

29. Schilling, René L. and Lothar Partzsch. *Brownian Motion: An Introduction to Stochastic Processes,* 2nd Edition. Berlin/Boston: Walter de Gruyter GmbH, 2014.

30. Schuss, Zeev, *Theory and Applications of Stochastic Differential Equations.* New York, NY: John Wiley and Sons, 1980.

Finance Applications

31. Etheridge, Alison. *A Course in Financial Calculus.* Cambridge, UK: Cambridge University Press, 2002.

32. Embrechts, Paul, Claudia Klüppelberg, and Thomas Mikosch. *Modelling Extremal Events for Insurance and Finance.* New York, NY: Springer-Verlag, 1997.

33. Hunt, P. J., and J. E. Kennedy. *Financial Derivatives in Theory and Practice,* Revised Edition. Chichester, UK: John Wiley & Sons, 2004.

34. McLeish, Don L. *Monte Carlo Simulation and Finance.* New York, NY: John Wiley, 2005.

35. McNeil, Alexander J., Rüdiger Frey, and Paul Embrechts. *Quantitative Risk Management: Concepts, Techniques, and Tools.* Princeton, NJ.:Princeton University Press, 2005.

Research Papers for Book 4

36. Balkema, A., de Haan, L. "Residual life time at great age." Annals of Probability, 2, 792–804, 1974.

37. Cantelli, F. P. "Sulla determinazione empirica delle leggi di probabilita." Giorn. Ist. Ital. Attuari 4, 221-424, 1933.

38. Dvoretzky, A., Kiefer, J.; Wolfowitz, J. "Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator." Annals of Mathematical Statistics, 27 (3): 642–669, 1956.

39. Fisher, R. A., Tippett, L. H. C. "Limiting forms of the frequency distribution of the largest or smallest member of a sample." Proc Camb Philos Soc 24:180–190, 1928.

40. Glivenko, V. "Sulla determinazione empirica della legge di probabilita." Giorn. Ist. Ital. Attuari 4, 92-99, 1933.

41. Gnedenko, B. "Sur la distribuion limite du terme maximum d'une série aléatoire." Ann Math 44:423–453, 1943.

42. Heyde, C. C. "On a Property of the Lognormal Distribution." In: Maller R., Basawa I., Hall P., Seneta E. (eds) Selected Works of C.C. Heyde. Selected Works in Probability and Statistics. Springer, New York, NY, 2010.

43. Hill, B. "A simple general approach to inference about the tail of a distribution." The annals of statistics, 3(5):1163–1174, 1975.

44. Kolmogorov, A. "Sulla determinazione empirica di una legge di distribuzione." G. Ist. Ital. Attuari. 4: 83–91, 1933.

45. Makarov, Mikhail. "Applications of exact extreme value theorem." Journal of Operational Risk, Volume 2/Number 1: 115–120, 2007.

46. Massart, P. "The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality." The Annals of Probability, 18 (3): 1269–1283, 1990.

47. Pickands, J. "Statistical inference using extreme order statistics." Annals of Statistics, 3, 119–131, 1975.

48. H. Scheffé, "A Useful Convergence Theorem for Probability Distributions," Ann. Math. Statistics, 18, 434–438, 1947.

49. Smirnov, N. V. "Sur les ´ecarts de la courbe de distribution empirique." (Russian, French summary). Rec. Math. Moscou (Mat. Sbornik), 6, 3–26, 1939.

50. Smirnov, N. V. "On the estimation of the discrepancy between empirical curves of distribution for two independent samples." Bull. Math. Univ. Moscou, 2(2), 1939.

51. Smirnov, N. V. "Table for estimating the goodness of fit of empirical distributions." Annals of Mathematical Statistics. 19: 279–281, 1948.

Index