Lecture Notes in Statistics **88**

Edited by P. Bickel, P. Diggle, S. Fienberg, U. Gather,
I. Olkin, S. Zeger

John E. Kolassa

Series Approximation Methods in Statistics
Third Edition

Springer

John E. Kolassa
Department of Statistics
Rutgers University
Piscataway, NJ 08854
kolassa@stat.rutgers.edu

Library of Congress Control Number: 2005939010

ISBN-10: 0-387-31409-1 e-ISBN 0-387-32227-2
ISBN-13: 978-0387-31409-9

Printed on acid-free paper.

Printed in the United States of America. (SBA)

9 8 7 6 5 4 3 2 1

springer.com

Preface

This book was originally compiled for a course I taught at the University of Rochester in the fall of 1991, and is intended to give advanced graduate students in statistics an introduction to Edgeworth and saddlepoint approximations, and related techniques. Many other authors have also written monographs on this subject, and so this work is narrowly focused on two areas not recently discussed in theoretical text books. These areas are, first, a rigorous consideration of Edgeworth and saddlepoint expansion limit theorems, and second, a survey of the more recent developments in the field.

In presenting expansion limit theorems I have drawn heavily on notation of McCullagh (1987) and on the theorems presented by Feller (1971) on Edgeworth expansions. For saddlepoint notation and results I relied most heavily on the many papers of Daniels, and a review paper by Reid (1988). Throughout this book I have tried to maintain consistent notation and to present theorems in such a way as to make a few theoretical results useful in as many contexts as possible. This was not only in order to present as many results with as few proofs as possible, but more importantly to show the interconnections between the various facets of asymptotic theory.

Special attention is paid to regularity conditions. The reasons they are needed and the parts they play in the proofs are both highlighted.

Computational tools are presented to help in the derivation and manipulation of these expansions. The final chapter contains input for calculating many of the results here in Mathematica (R), a symbolic algebra and calculus program. Mathematica is a registered trademark of Wolfram Research, Inc.

This book is organized as follows. First, the notions of asymptotics and distribution approximations in general are discussed, and the present work is placed in this context. Next, characteristic functions, the basis of all of the approximations in this volume, are discussed. Their use in the derivation of the Edgeworth series, both heuristically and rigorously, is presented. Saddlepoint approximations for densities are derived from the associated Edgeworth series, and investigated. Saddlepoint distribution function approximations are presented. Multivariate and conditional counterparts of many of these results are presented, accompanied by a discussion of the extent to which these results parallel univariate results and the points where multivariate results differ. Finally, these results are applied to questions of the distribution of the maximum likelihood estimator, approximate ancillarity, Wald and likelihood ratio testing, Bayesian methods, and resampling methods.

Much of this volume is devoted to the study of lattice distributions, because in representing a departure from regularity conditions they represent an interesting variation on and completion of the mathematical development of the rest of the material, because they arise very frequently in generalized linear models for discrete data and in nonparametric applications, and because many of my research interests lie in this direction. In the interest of not unnecessarily burdening those who wish to avoid the additional complication of lattice distributions, I have tried to place the lattice distribution material as late in each chapter as possible. Those who wish may skip these sections without sacrificing much of the foundation for the rest of the book, but I recommend this material as both useful and inherently interesting.

Prerequisites and Notation

A knowledge of undergraduate real and complex analysis, on the level of Rudin (1976), Chapters 1-9, and Bak and Newman (1996), Chapters 1-12, is presupposed in the text. In particular, an understanding of continuity, differentiation, and integration in the senses of Riemann and Riemann-Stieltjes, is needed, as is an understanding of basic limit theorems for these integrals. An understanding of complex contour integration is also required. Lebesgue integration and integration of differential forms are not required. A background in matrix algebra of comparable depth is also presupposed, but will not be required as frequently.

As far as is possible, statistical parameters will be denoted by Greek characters. In general, upper-case Latin characters will in general refer to random variables, lower-case Latin characters will refer to potential observed values for the corresponding random variables, and bold face will in general denote vector or matrix quantities. Lower case Gothic characters refer to integer constants, and capital Gothic characters refer to sets. For example $\mathfrak{R}$, $\mathfrak{C}$, and $\mathfrak{Z}$ represent the sets of real numbers, complex numbers, and integers, respectively.

I have been unable to follow these conventions entirely consistently; for instance, densities and cumulative distribution functions are generally denoted by f and F, and cumulant generating functions are generally denoted by $\mathcal{K}$. Additionally, estimates of parameter values under various assumptions are denoted by the parameter value with a hat or tilde accent, and random estimators are denoted by upper case counterparts. For instance, ω will represent the signed root of the likelihood ratio, $\hat{\omega}$ will represent its fitted value, and $\hat{\Omega}$ will represent the random variable of which w is an observed value.

Unless stated otherwise, all integrals are Riemann-Stieltjes integrals, and the limiting operation implied by non-absolutely integrable improper integrals are given explicitly. The symbols $\Re$ and $\Im$ are functions returning the real and imaginary parts of complex numbers respectively. The Gamma function is denoted by $\Gamma(x)$, and ψ is the di–gamma function $(d/dx)\log(\Gamma(x))$. All logarithms are with respect to the natural base.

Acknowledgments

This edition was completed with support from NIH grant CA63050. I would like to thank Martin Tanner, who suggested and encouraged this project, and whose NIH grant CA35464 supported me for part of the time while I wrote the first edition of this book. I also thank Dr. Robert Griggs, on whose NIH training grant T32NS07338 I began this project. Thanks are also due to David Oakes, Lu Cui, Guang-qin Ma, Shiti Yu, Debajyoti Sinha, and especially Moonseong Heo, who attended the lectures that this book is based on, and who suggested improvements. I am especially indebted to Peter McCullagh and O.E. Barndorff–Nielsen, who read and commented in detail on an early draft of these notes, and others who read and commented more generally on part or all of this work, including Bradley Efron, David Cox, Mitchell Gail, Scott Zeger, Nancy Reid, and Michael Wichura. The efforts of the editor for this series, Martin Gilchrist, and the production staff at Springer–Verlag, are very much appreciated, as are the comments of three anonymous referees. I am very grateful for the support of family and friends while writing this book.

Thanks are also due to those who helped with the second and third editions, including students who used and commented on preliminary versions of the second and third editions: Chris Roe, Bo Yang, Peng Huang, Raji Natarajan, Aiyi Liu, Kelly Zou, Leila Chukhina, Ramprasath Lakshminarasimhan, Lars Nordmann, Juan Zhang, the editor John Kimmel, Anne Meagher, and anonymous reviewers. Special thanks are due to Yodit Seifu and Greg Elmer, whose sharp eye caught many typographical errors. NSF grant DMS 0505499 and the Rutgers University sabbatical leave program supported preparation of the third edition.

Contents

1

Asymptotics in General

Many authors have examined the use of asymptotic methods in statistics. Serfling (1980) investigates applications of probability limit theorems for distributions of random variables, including theorems concerning convergence almost surely, to many questions in applied statistics. Le Cam (1969) treats asymptotics from a decision-theoretic viewpoint. Barndorff–Nielsen and Cox (1989) present many applications of the density and distribution function approximations to be described below in a heuristic manner. Hall (1992) investigates Edgeworth series with a particular view towards applications to the bootstrap. Field and Ronchetti (1990) treat series expansion techniques in a manner that most closely parallels this work; I have included more detailed proofs and discussion of regularity conditions, and a survey of the use of Barndorff–Nielsen's formula. Their work covers many aspects of robustness and estimating equations not included here. Skovgaard (1990) explores characteristics of models making them amenable to asymptotic techniques, and derives the concept of an analytic statistical model. He also investigates convergence along series indexed by more general measures of information than sample size. Jensen (1995) presents a range of topics similar to that presented here, but with a different flavor.

The question of convergence of various approximations to distribution functions and densities will be considered with as much attention to regularity conditions and rigor as possible. Bhattacharya and Rao (1976) and Bhattacharya and Denker (1990) rigorously treat multivariate Edgeworth series; the present work begins with the univariate case as a pedagogical tool. Valuable recent review articles include those of Reid (1996) and Skovgaard (1989).

1.1. Probabilistic Notions

Serfling (1980) describes various kinds of probabilistic limits often used in statistics. Probabilists generally model random variables as functions of an unobserved sample point ω lying in some sample space Ω, with a measure $\mathrm{P}[\cdot]$ assigning to each set in a certain class $\mathcal{F}$ of subsets of Ω a probability between 0 and 1. Generally, sequences of related random variables are modeled as a sequence of functions of ω, and a large part of asymptotic theory is concerned with describing the behavior of such sequences.

The strongest types of convergence are convergence almost surely, and convergence in mean. Random variables X_n are said to converge almost surely, or converge with probability 1, to Y if $\mathrm{P}[X_n(\omega) \to Y(\omega)] = 1$. This notion of convergence crucially involves the sample point ω, and concerns behavior of the functions X_n at ω for all n, or at least for all n sufficiently large, simultaneously. For instance, the strong law of large numbers implies that if X_n is a Binomial random variable representing the number of successes in n independent trials, with each trial resulting in a success with probability π and a failure with probability $1-\pi$, then X_n/n converges almost surely to π. Using this result, however, requires conceptually constructing a sample space Ω on which all of these random variables exist simultaneously. The natural sample space for X_n is $\Omega_n = \{0,1\}^n$ consisting of sequences of zeros and ones of length n. In this case $X_n(\omega) = \sum_i \omega_i$, and probabilities are defined by assigning each $\omega \in \Omega_n$ the probability 2^{-n}. The common sample space Ω must be constructed as a sort of infinite product of the Ω_n; this expanded sample space then bears little relation to the simple Ω_n describing the specific experiment considered.

Random variables X_n are said to converge to Y in r-th mean, for some $r \in (0,\infty)$ if $\mathrm{E}[|X_n(\omega) - Y(\omega)|^r] \to 0$. This type of convergence is less concerned with the simultaneous behavior of the random variables for each ω and more concerned about the overall relationship between $X_n(\omega)$ and $Y(\omega)$ globally on Ω for fixed n. The relative values of X_n and Y for a fixed ω play the central role.

A weaker form of convergence is convergence in probability. The variables X_n converge to Y in probability, if for every $\epsilon > 0$, then $\lim_{n\to\infty} \mathrm{P}[|X_n - Y| < \epsilon] = 1$. As with convergence in r-th mean, convergence in probability concerns the overall relationship between $X_n(\omega)$ and $Y(\omega)$ globally on Ω for fixed n, but in a weaker sense.

Random variables X_n are said to converge in distribution to Y, if

$$\lim_{n\to\infty} \mathrm{P}[X_n(\omega) \le x] = \mathrm{P}[Y(\omega) \le x]$$

for all x continuity points of $\mathrm{P}[Y(\omega) \le x]$. Of these various convergence notions convergence in distribution is the weakest, in the sense that convergence almost surely and convergence in r-th mean both imply convergence in distribution. It is also weakest in the sense that it relies the least heavily on the classical measure theoretic probability notions. In the binomial example above, then, one can show that $(X_n - n\pi)/\sqrt{n\pi(1-\pi)}$ converges in distribution to a standard normal variable Y, without having to envision, even conceptually, a probability space upon which Y and even one of the X_n are simultaneously defined.

If F_n is the cumulative distribution function for X_n and F is the cumulative distribution function for Y then the criterion for convergence in distribution can be written as $F_n(x) \to F(x)$ as $n \to \infty$ for all x at which F is continuous. Often times the limiting distribution F is then used to approximate the distribution F_n in cases when n is considered sufficiently large.

This course will concentrate on variants of the idea of convergence in distribution, and will involve deriving easily–calculated approximations G_n to F_n.

At this point it may be useful to introduce order notation. Suppose f and g are two functions of a parameter, and one wishes to describe how much they differ as the parameter approaches some limiting value. One might begin by assessing whether the difference converges to zero or whether the differences are bounded in the limit; a refined analysis might describe the rate at which the difference converges to zero or diverges from zero. The notation $f(n) = g(n) + o(h(n))$ means $(f(n) - g(n))/h(n) \to 0$; the notation $f(n) = g(n) + O(h(n))$ means $(f(n) - g(n))/h(n)$ is bounded as n approaches some limiting value. Usually the implied limit is as $n \to \infty$ if n is a discrete quantity like sample size. If n is a continuous quantity the implied limit is often as $n \to 0$. For example, $1 + 2t + t^3 = 1 + 2t + o(t^2)$ as $t \to 0$, and $(n + \log(n) - 1)/n^2$ may be described alternatively as $1/n + o(1/n)$ or $1/n + \log(n)/n^2 + O(1/n^2)$. As another example, the numbers $a_0, a_1, \ldots, a_l$ are the value and first l derivatives of a function f at zero if and only if $f(t) = \sum_{j=0}^{l} a_l t^l/l! + o(t^l)$ as $t \to 0$.

Probabilistic versions of this order notation also exist. For two sequences of random variables U_n and V_n defined on the same probability space, we say that $V_n = O_p(U_n)$ if for any $\epsilon > 0$ there exist M_ϵ and N_ϵ such that $\mathrm{P}\left[|V_n/U_n| > M_\epsilon\right] < \epsilon$ for $n > N_\epsilon$. We say that $V_n = o_p(U_n)$ if V_n/U_n converges in probability to zero.

1.2. The Nature of Asymptotics

Wallace (1958) discusses the question of approximating a cumulative distribution function F_n depending on an index n by a function G_n also depending on n, in general terms. Often one estimates $F_n(x)$ as the truncation of a nominal infinite series, whose coefficients depend on n. That is,

$$G_{\mathrm{j},n}(x) = \sum_{j=0}^{\mathrm{j}} A_j(x) a_{j,n}, \tag{1}$$

where $a_{j,n}$ decreases in n. In the case of sample means often $a_{j,n} = n^{-j/2}$. In many cases, the difference between the target cumulative distribution function and the approximation is of a size comparable to the first term neglected; that is, $|F_n(x) - G_{\mathrm{j},n}(x)| < c(x) a_{\mathrm{j}+1,n}$, or in other words,

$$F_n(x) = G_{\mathrm{j},n}(x) + O(a_{\mathrm{j}+1,n}). \tag{2}$$

In general an expression line (1) does not imply that

$$F_n(x) = \sum_{j=0}^{\infty} A_j(x) a_{j,n}. \tag{3}$$

Approximations of the form (1) are useful even when the infinite sum (3) does not converge, as noted by Cramér (1925).

As an example, consider the standard normal survival function. If j is a non-negative integer, and $x > 0$, then integrating by parts shows that

$$\begin{aligned}\int_x^\infty \phi(y) y^{-2j}\, dy &= -\int_x^\infty [-y\phi(y)]\, \frac{dy}{y^{2j+1}} \\ &= -\left(\phi(y)/y^{2j+1}\right)\Big|_x^\infty - (2j+1)\int_x^\infty \phi(y)\, \frac{dy}{y^{2j+2}} \\ &= \phi(x)/x^{2j+1} - (2j+1)\int_x^\infty \phi(y)\, \frac{dy}{y^{2j+2}}.\end{aligned}$$

Applying this identity recursively, for any k,

$$\bar{\Phi}(x) = \phi(x)\sum_{j=0}^{k} x^{-1-2j}(-1)^j a_j + (-1)^{k+1} a_{j+1}\int_x^\infty \phi(y)\, \frac{dy}{y^{2k+2}}, \tag{4}$$

for $a_j = \begin{cases} 1 & \text{if } j = 0 \\ \prod_{i=0}^{j-1}(2i+1) & \text{otherwise} \end{cases}$, for any $x > 0$. Then

$$\bar{\Phi}(\sqrt{n}x) = \frac{\phi(\sqrt{n}x)}{\sqrt{n}}\left[\sum_{j=0}^{k} x^{-1-2j} n^{-j}(-1)^j a_j + O(n^{-j-1})\right]. \tag{5}$$

Thus (5) constitutes an asymptotic expansion for $\bar{\Phi}(\sqrt{n}x)$. Furthermore, the associated infinite series is alternating; since the integral in (4) is positive,

$$\bar{\Phi}(x)\begin{cases} \leq \phi(x)\sum_{j=0}^{k} x^{-1-2j}(-1)^j a_j & \text{if } k \text{ even} \\ \geq \phi(x)\sum_{j=0}^{k} x^{-1-2j}(-1)^j a_j & \text{if } k \text{ odd} \end{cases}. \tag{6}$$

Hence

$$\phi(x)\{1/x - 1/x^3\} \leq \bar{\Phi}(x) \leq \phi(x)\{1/x - 1/x^3 + 3/x^5\}. \tag{7}$$

However, the a_j increase quickly enough that when the finite sum is transformed into an infinite series, the series does not converge for any x.

As a second example, consider Stirling's asymptotic expansion for the Gamma function

$$e^x x^{\frac{1}{2}-x}\frac{\Gamma(x)}{\sqrt{2\pi}} = 1+\frac{x^{-1}}{12}+\frac{x^{-2}}{288}-\frac{139x^{-3}}{51840}-\frac{571x^{-4}}{2488320}+\frac{163879x^{-5}}{209018880}+ \frac{5246819x^{-6}}{75246796800}-\frac{534703531x^{-7}}{902961561600}-\frac{4483131259x^{-8}}{86684309913600}+ \frac{432261921612371x^{-9}}{514904800886784000}+\mathrm{O}\left(x^{-10}\right);$$

this is a valid asymptotic expansion as $x \to \infty$, but fixing x and letting the number of terms included increase to infinity eventually degrades performance (Fig. 1).

Error in Stirling's Series as the Order of Approximation Increases

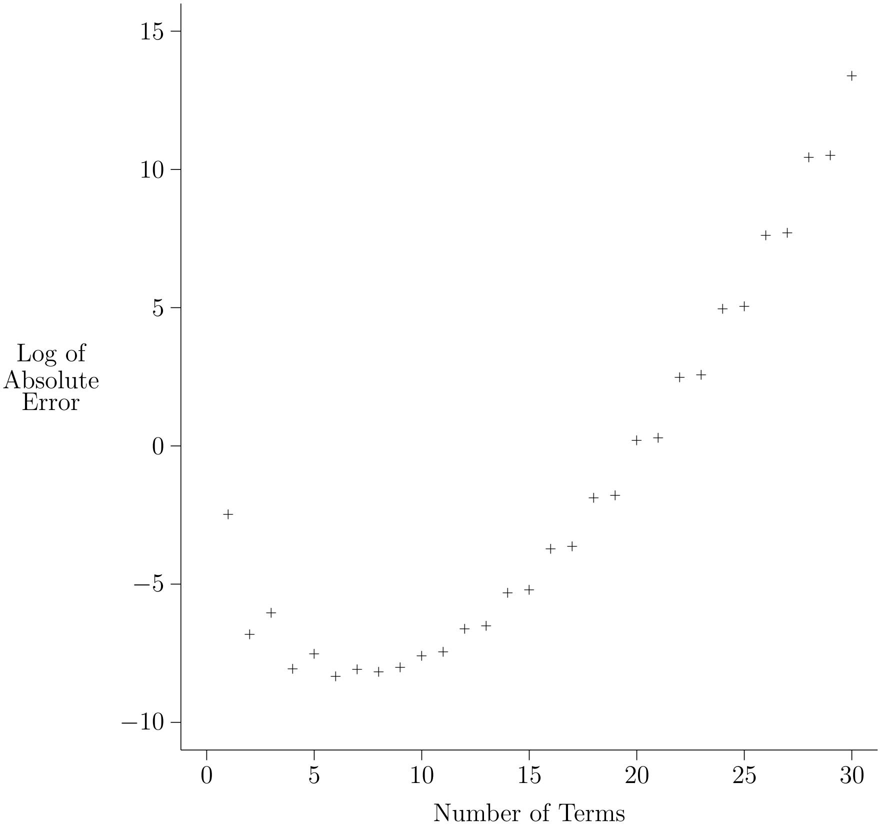

Fig. 1.

Note the distinction between asymptotic expansions and convergence of series like (3). The main concern in this volume is the behavior of the difference between F_n and $G_{\mathrm{j},n}$ as n, rather than j, increases. An asymptotic expansion, then, is a formal series like (3) which when truncated after any number of terms j as in (1) exhibits the behavior described by (2).

2

Characteristic Functions and the Berry–Esseen Theorem

This chapter discusses the role of the characteristic function in describing probability distributions. Theorems allowing the underlying probability function to be reconstructed from the characteristic function are presented. Results are also derived outlining the sense in which inversion of an approximate characteristic function leads to an approximate density or distribution function. These results are applied to derive Berry–Esseen theorems quantifying the error incurred in such an approximation. Finally, the relationship between the characteristic function and moments and cumulants is investigated.

2.1. Moments and Cumulants and their Generating Functions

The characteristic function for a random variable X taking values in $\mathfrak{R}$ is defined to be

$$\zeta_X(\beta) = \mathrm{E}\left[\exp(i\beta X)\right]; \tag{8}$$

$\zeta_X(\beta)$ is also known as the Fourier transform of the distribution. The characteristic function always exists for $\beta \in \mathfrak{R}$, and if the density for X is symmetric then $\zeta(\beta) \in \mathfrak{R}$ for all β. It is called "characteristic" because in a sense described below it characterizes the distribution uniquely.

Additionally, the characteristic function has properties convenient when considering transformations of random variables. First, if X_1 and X_2 are independent random variables, a_1, a_2, and b are constants, and $X = a_1X_1 + a_2X_2 + b$, then

$$\begin{aligned}\zeta_X(\beta) &= \mathrm{E}\left[\exp(i[a_1X_1 + a_2X_2 + b]\beta)\right] \\ &= \exp(ib\beta)\mathrm{E}\left[\exp(ia_1X_1\beta)\right]\mathrm{E}\left[\exp(ia_2X_2\beta)\right] = \exp(ib\beta)\zeta_{X_1}(a_1\beta)\zeta_{X_2}(a_2\beta). \end{aligned} \tag{9}$$

Hence $\zeta_X(\beta) = \prod_{j=1}^n \zeta_{X_j}(a_j\beta)$ if $X = \sum_{j=1}^n a_jX_j$ and the variables X_j are independent, and $\zeta_{\bar{X}}(\beta) = \zeta_{X_1}(\beta/n)^n$ if $\bar{X}$ is the mean of n independent and identically distributed random variables X_j.

The distribution of the random variable $\sum_{j=1}^{n} X_j$ where $a_j = 1$ for all j is called the convolution of the distributions of the random variables X_j.

One can recover the moments of a distribution from its characteristic function. By differentiating (8) l times and evaluating the result at zero, one sees that

$$\zeta_X^{(l)}(0) = \mathrm{E}\left[i^l X^l \exp(i \times 0 \times X)\right] = i^l \mathrm{E}\left[X^l\right], \tag{10}$$

assuming that the orders of integration and differentiation can be interchanged.

The relationship between the characteristic function and moments of the underlying distribution unfortunately involves i. In chapters following this one, we will make use of a more direct generating function, the moment generating function, defined to be $\mathcal{M}_X(\beta) = \mathrm{E}\left[\exp(\beta X)\right]$, which is (8) with the i removed. The function $\mathcal{M}_X(\beta)$, with β replaced by $-\beta$, is called the Laplace transform for the probability distribution of X. Unlike characteristic functions, these need not be defined for any real $\beta \neq 0$. The range of definition is convex, however, since for any x the function $\exp(x\beta)$ is convex. That is, if $p \in (0,1)$, then

$$\begin{aligned}\mathcal{M}_X(p\gamma + (1-p)\beta) &= \mathrm{E}\left[\exp((p\gamma + (1-p)\beta)X)\right] \\ &\leq \mathrm{E}\left[p\exp(\gamma X) + (1-p)\exp(\beta X)\right] \\ &= p\mathcal{M}_X(\gamma) + (1-p)\mathcal{M}_X(\beta).\end{aligned}$$

Hence if $\mathcal{M}_X(\beta) < \infty$ for any real $\beta \neq 0$, then $\mathcal{M}_X(\beta) < \infty$ for all β in an interval $\mathfrak{Q}$ containing 0, although 0 may lie on the boundary of $\mathfrak{Q}$. The function $\mathcal{M}_X$ may only be defined for values of β lying on one side or the other of the origin. If $\mathcal{M}_X(\beta) < \infty$ on some open interval containing 0, then all moments are finite, and $\mathcal{M}_X$ has a power series expansion about 0 of the form $\mathcal{M}_X(\beta) = \sum_{j=0}^{\infty} \mu_j \beta^j / j!$. The counterpart of (10) is $\mathcal{M}_X^{(l)}(0) = \mathrm{E}\left[X^l\right]$. This will be demonstrated in §2.3. The radius of convergence of this series is given by

$$R = \min(\sup(\{\beta : \beta > 0, \mathcal{M}_X(\beta) < \infty\}, \sup(\{-\beta : \beta < 0, \mathcal{M}_X(\beta) < \infty\})). \tag{11}$$

Proof of this claim is left as an exercise.

Furthermore, if $\beta \in \mathfrak{Q} \times i\mathfrak{R}$, then $\mathrm{E}\left[|\exp(\beta X)|\right] < \infty$, and hence $\mathcal{M}_X$ exists for these β as well. Conversely, since $|\exp(\beta X)| = \exp(\Re(\beta)X)$, then $\mathrm{E}\left[|\exp(\beta X)|\right] < \infty$ implies that $\Re(\beta) \in \mathfrak{Q}$, and $\beta \in \mathfrak{Q} \times i\mathfrak{R}$.

A slightly more general definition of $\mathcal{M}_X(\boldsymbol{\beta})$ is $\lim_{R\to\infty} \int_{-\infty}^{R} \exp(x\beta) dF_X(x)$, when $\Re(\beta) > 0$, and analogously when $\Re(\beta) < 0$. This limit might exist for some β for which $\mathrm{E}\left[|\exp(\beta X)|\right] = \infty$. Widder (1946) proves the following:

Lemma 2.1.1: *If $\lim_{R\to\infty} \int_{-\infty}^{R} \exp(ix\boldsymbol{\beta}) dF_X(x)$ exists and is finite, then $\beta \in \bar{\mathfrak{Q}} \times i\mathfrak{R}$. Here $\bar{\cdot}$ applied to a set denotes closure; since $\mathfrak{Q}$ is an interval, the closure represents the set together with its supremum and infimum.*

Proof: It suffices to show that if $\lim_{R\to\infty} \int_{-\infty}^{R} \exp(x\beta) dF_X(x)$ exists and is finite for $\Re(\beta) > 0$, and if $\gamma \in \mathfrak{R}$ and $\gamma \in [0, \Re(\beta))$, then $\lim_{R\to\infty} \int_{-\infty}^{R} \exp(x\gamma) dF_X(x)$

exists and is finite. To see this, let $G(x) = \int_{-\infty}^{x} \exp(\beta y)\, dF_X(y)$. Then, integrating by parts, $\int_{-\infty}^{R} \exp(x\gamma) dF_X(x) = \int_{-\infty}^{R} \exp(x[\gamma - \beta]) dG(x) = G(R) \exp(R[\gamma - \beta]) + (\beta - \gamma) \int_{-\infty}^{R} \exp(x[\gamma - \beta]) G(x)\, dx$. Since $\lim_{R\to\infty} G(R) = \mathcal{M}_X(\beta)$ and $\Re(\gamma - \beta) < 0$, the first term converges to zero. Furthermore, the second integral is bounded.

Q.E.D

The moment generating function will be used for two purposes below. In the remainder of this section it will be used to define a sequence of numbers providing a characterization of a distribution that is more useful than the moments, and in later chapters on saddlepoint approximations, the real part of the moment generating function argument will be used to index an exponential family in which the distribution of interest is embedded, and the imaginary part will be used to parameterize the characteristic function of that distribution.

Normal approximations to densities and distribution functions make use of the expectation, or first moment, and the variance, or second central moment, of the distribution to be approximated. The Berry–Esseen theorem, to be discussed below, which assesses the quality of this approximation involved also the third central moment. Thus far, then, for the purposes of approximating distributions it seems sufficient to describe the distribution in terms of its first few moments.

For reasons that will become clear later, it is desirable to use, rather than moments, an alternate collection of quantities to describe asymptotic properties of distributions. These quantities, which can be calculated from moments, are called cumulants, and can be defined using the power series representation for the logarithm of the characteristic function. Since manipulation of logarithms of complex numbers presents some notational complications, however, we will instead derive cumulants from a real-valued function analogous to the characteristic function, called the moment generating function. Its logarithm will be called the cumulant generating function. Much of the material on this topic can be found in McCullagh (1987), §2.5-2.7, and Kendall, Stuart, and Ord (1987), §3.

Since $\mathcal{M}_X(\beta)$ is real and positive for all β, we can define the cumulant generating function $\mathcal{K}_X(\beta) = \log(\mathcal{M}_X(\beta))$. Let κ_j be its derivative of order j at zero. If derivatives of all orders exist, the formal expansion of $\mathcal{K}_X$ about $\beta = 0$ is $\sum_{j=0}^{\infty} \kappa_j \beta^j / j!$. Since $\mathcal{M}_X(0) = 1$, $\kappa_0 = 0$. The coefficients κ_j for $j > 0$ are called cumulants or semi-invariants.

These terms will be justified below. An important feature of the cumulant generating function is the simple way the cumulant generating function for an affine transformation of independent random variables is related to the underlying cumulant generating functions. Substituting β for $i\beta$ in (9), and taking logs, shows that the cumulant generating function of an affine transformation of one variable is given by

$$\mathcal{K}_{aX+b}(\beta) = \mathcal{K}_X(a\beta) + b\beta,$$

and hence if κ_j and λ_j are the cumulants of X and $aX + b$, $\lambda_j = a^j \kappa_j$ for $j > 1$ and $\lambda_1 = a\kappa_1 + b$. Thus cumulants of order two and higher are invariant under translation, and vary in a regular way with rescaling. This justifies the name semi-invariant.

If X has cumulants $\boldsymbol{\kappa} = (\kappa_1, \kappa_2, \ldots)$, then the cumulant of order j of $(X - \kappa_1)/\sqrt{\kappa_2}$ is $\rho_j = \kappa_j \kappa_2^{-j/2}$ for $j > 1$, and 0 for $j = 1$. Call these quantities the invariants. These do not change under affine transformations of X. Now consider linear combinations of more than one variable. Choose any X and Y independent. Then substituting β for $i\beta$ in (9), and taking logs,

$$\mathcal{K}_{X+Y}(\beta) = \mathcal{K}_X(\beta) + \mathcal{K}_Y(\beta),$$

and hence if κ_j, λ_j, and ν_j are cumulants of X, Y, and $X + Y$ respectively, $\nu_j = \kappa_j + \lambda_j$. Thus the cumulants "cumulate", giving rise to the name.

If Y_j are independent and identically distributed, and $Z = (1/\sqrt{n}) \sum_{j=1}^n Y_j$, then

$$\mathcal{K}_Z(\beta) = n\mathcal{K}_Y(\beta/\sqrt{n}) = \sum_{j=0}^{\infty} \kappa_j n^{(2-j)/2} \beta^j / j!.$$

Using the fact that $\log(1+z) = \sum_{j=1}^{\infty} (-1)^{j-1} z^j / j$, convergent for $|z| < 1$, one can express cumulants in terms of moments, and using $\exp(z) = \sum_{j=0}^{\infty} z^j / j!$, convergent for all z, one can express moments in terms of cumulants. These results are tabulated in Table 1.

Table 1: *Conversions between Moments and Cumulants*

$\kappa_1 = \mu_1,$	$\mu_1 = \kappa_1,$
$\kappa_2 = \mu_2 - \mu_1^2,$	$\mu_2 = \kappa_2 + \kappa_1^2,$
$\kappa_3 = \mu_3 - 3\mu_1\mu_2 + 2\mu_1^3,$	$\mu_3 = \kappa_3 + 3\kappa_1\kappa_2 + \kappa_1^3,$
$\kappa_4 = \mu_4 - 4\mu_1\mu_3 - 3\mu_2^2 + 12\mu_2\mu_1^2 - 6\mu_1^4,$	$\mu_4 = \kappa_4 + 4\kappa_1\kappa_3 + 3\kappa_2^2 + 6\kappa_2\kappa_1^2 + \kappa_1^4.$

Conversion in either direction, then, involves forming linear combinations of products of moments and cumulants, with the coefficients derived from the coefficients of the moments or cumulants in the appropriate generating function, and from the number of ways in which a particular term arises in the series exponentiation or logarithm. The constants arising in these conversion relations are more transparent in the multivariate case since the number of times symmetric terms arise in the transformed power series is explicitly recorded. See McCullagh (1987) for a further discussion of these transformations. Kendall, Stuart, and Ord (1987) §3.14 give transformations between these in both directions.

One can define cumulants even when $\mathcal{K}_X$ does not have a positive radius of convergence, using the same conversion rules as above, or using the definition

$$\kappa_j = (-i)^j \left. \frac{d^j \log(\zeta(\beta))}{d\beta^j} \right|_{\beta=0}, \tag{12}$$

and some suitable definition for complex logarithms, when this derivative exists. As we will see below, however, existence of derivatives of the characteristic function implies the existence of the moment of the corresponding order only if that order is even; the exercises provide a counterexample in which the first derivative of a characteristic function exists but the first moment does not. Loosely then we will say that moments to a certain order exist if and only if cumulants of the same order exist.

2.2. Examples of Characteristic and Generating Functions

The following are simple examples of calculating characteristic functions.

a. The normal distribution: Recall that the standard normal density has the form

$$f_X(x) = \exp(-x^2/2)/\sqrt{2\pi}.$$

Two options for presenting the corresponding characteristic function present themselves. One might evaluate the integral

$$\begin{aligned}\zeta_X(\beta) &= \int_{-\infty}^{\infty} \exp(ix\beta)\exp(-x^2/2)(2\pi)^{-1/2}\,dx\\ &= \int_{-\infty}^{\infty} (\cos(\beta x) + i\sin(\beta x))\exp(-x^2/2)\,dx/\sqrt{2\pi}\\ &= \int_{-\infty}^{\infty} \cos(\beta x)\exp(-x^2/2)\,dx/\sqrt{2\pi}.\end{aligned}$$

Alternatively one might calculate moments of the random variable and then use the power series expression for the characteristic function to construct $\zeta_X(\beta)$. The moments may be expressed as an integral, which may be evaluated using integration by parts to show:

$$\mu_l = \begin{cases} 0 & l \text{ odd} \\ 2^{-l/2} l!/(l/2)! & l \text{ even.} \end{cases}$$

Since the radius of convergence is infinite,

$$\zeta_X(\beta) = \sum_{l \text{ even}} (i\beta)^l/(2^{l/2}(l/2)!) = \sum_l (-\beta^2/2)^l/l! = \exp(-\beta^2/2).$$

The moment generating function is calculated more easily.

$$\begin{aligned}\mathcal{M}_X(\beta) &= \int_{-\infty}^{\infty} \exp(\beta x - x^2/2)\,dx/\sqrt{2\pi}\\ &= \exp(\beta^2/2)\int_{-\infty}^{\infty} \exp(-(x-\beta)^2/2)\,dx/\sqrt{2\pi} = \exp(\beta^2/2),\end{aligned}$$

for β real. Hence $\mathcal{M}(\beta)$ exists (and is hence differentiable) for $\beta \in \mathfrak{R}$; $\exp(\beta^2/2)$ also is defined and differentiable for $\beta \in \mathfrak{R}$. Since these functions coincide on a set of points converging to 0, then $\mathcal{M}_X(\beta) = \exp(\beta^2/2)$ for $\beta \in \mathfrak{R}$ (Bak and Newman, 1996, §6.3). The set on which $\mathcal{K}_X(\beta)$ exists is smaller, since the log function is not defined for arguments with a zero imaginary part and a non-positive real part. For instance, since $\exp((\pi+i)^2/2) = \exp((\pi^2-1)/2+i\pi) = -\exp((\pi^2-1)/2)$, then $\mathcal{K}_X(\pi+i)$ is not defined. However, on some neighborhood of $\mathfrak{R}$, $\mathcal{K}_X(\beta) = \beta^2/2$, and cumulants of order 3 and above are zero.

b. The uniform distribution on $(-1/2, 1/2)$:

$$\begin{aligned}\zeta_X(\beta) &= \int_{-1/2}^{1/2} (\cos(\beta x) + i\sin(\beta x))\ dx \\ &= [\sin(\beta/2) - \sin(-\beta/2) + i\cos(-\beta/2) - i\cos(\beta/2)]/\beta = 2\sin(\beta/2)/\beta.\end{aligned}$$

Calculation of its cumulant generating function is left as an exercise.

c. The Cauchy distribution: The density $f_X(x) = 1/(\pi(1+x^2))$ has the corresponding characteristic function

$$\zeta_X(\beta) = \exp(-|\beta|), \tag{13}$$

differentiable everywhere *except at* 0. Its derivation is left as an exercise. No moments of order greater than or equal to one exist for this distribution, but expectations of the absolute value of the random variable raised to positive powers less than one do exist. Kendall, Stuart, and Ord (1987) give these as $\mathrm{E}\left[|X|^c\right] = 1/\sin((1+c)\pi/2)$ for $|c| < 1$. For $\beta \in \mathfrak{R}$ such that $\beta \neq 0$, the integral

$$\zeta_X(\beta) = \int_{-\infty}^{\infty} \exp(x\beta)/(\pi(1+x^2))\ dx$$

is infinite, and so the cumulant generating function does not exist.

d. The Bernoulli distribution and the binomial distribution: If X is a Bernoulli variable taking the value 1 with probability π and the value 0 otherwise, then its characteristic function is $\zeta_X(\beta) = (1-\pi) + \pi\exp(i\beta)$, and if Y_n has the distribution of the sum of n independent such variables, its characteristic function is $\zeta_{Y_n}(\beta) = ((1-\pi) + \pi\exp(i\beta))^n$. The cumulant generating function is $\log((1-\pi) + \pi\exp(\beta))$.

The Bernoulli example illustrates two points. First, this characteristic function has a non-zero imaginary part. In the three preceding examples the distributions were symmetric about zero, eliminating the imaginary part of the integral. This distribution is not symmetric, and so imaginary parts do not cancel out, since, for a fixed value of x, values of the summands or integrands in the expectation calculation are no longer in general the conjugates of the values at $-x$. Second, this characteristic function is periodic. This arises from the fact that possible values for the random variable are restricted to a lattice of equally spaced points. Most of the applications considered in this volume will involve either continuous or lattice distributions. Complications arising from other distributions will be discussed along with the regularity conditions for Edgeworth series. The Bernoulli distribution arises in applications involving testing a binomial proportion, including determining critical regions for the non–parametric sign test.

2.3. Characteristic Functions and Moments

We have seen that under certain regularity conditions, the set of derivatives of the characteristic function ζ_Y of a random variable Y determines the moments of Y, which in turn under certain regularity conditions determines the distribution of Y. Billingsley (1986) proves as Theorem 26.2 that two distinct real random variables cannot have the same characteristic function. A portion of this argument is summarized below, in providing an expansion for the characteristic function in terms of moments of the underlying distribution. Some lemmas link the existence of derivatives of the characteristic function at zero to the existence of moments of Y.

I first demonstrate an inequality for use in bounding the error when $\exp(iy)$ is approximated by partial sums of its Taylor expansion; this is given by Billingsley (1986), p. 297.

Lemma 2.3.1: *For any real y,* $\left|\exp(iy) - \sum_{k=0}^{l}(iy)^k/k!\right| \leq \min\left(\frac{|y|^{l+1}}{(l+1)!}, \frac{2\,|y|^l}{l!}\right).$

Proof: Integration by parts shows that

$$\int_0^x (y-s)^j \exp(is)\; ds = \frac{y^{j+1}}{j+1} + \frac{i}{j+1}\int_0^x (y-s)^{j+1}\exp(is)\; ds.$$

Furthermore, $\exp(iy) = 1 + i\int_0^y \exp(is)\; ds$, and hence by induction,

$$\exp(iy) = \sum_{k=0}^{j}(iy)^k/k! + (i^{j+1}/j!)\int_0^y (y-s)^j \exp(is)\; ds.$$

Note that

$$\left|\int_0^y (y-s)^j \exp(is)\; ds\right| \leq |y|^{j+1}.$$

By integration by parts,

$$\begin{aligned}\int_0^y (y-s)^j \exp(is)\; ds &= -iy^j + ij\int_0^y (y-s)^{j-1}\exp(is)\; ds\\ &= ij\int_0^y (y-s)^{j-1}(\exp(is)-1)\; ds\end{aligned}$$

and hence

$$\left|\int_0^y (y-s)^j \exp(is)\; ds\right| \leq 2j\,|y|^j.$$

Q.E.D

Lemma 2.3.1 will be used in this section and §2.6 to bound errors in approximating $\exp(y)$. The following result was presented by Billingsley (1986), §26. We will also need the following result limiting the size of tails of the distribution function.

Lemma 2.3.2: *If $G(y)$ is a cumulative distribution function, then*

$$\lim_{A\to\infty}\int_{-\infty}^{A} y\ dG(y)/A = 0.$$

Proof: If the result fails to hold, then there exists $\epsilon > 0$ and a sequence A_k such that $\lim_{k\to\infty} A_k = \infty$ and $\int_{-\infty}^{A_k} y\ dG(y) \geq \epsilon A_k$ for all k. In this case, $\sum_{j=1}^{k}(G(A_j) - G(A_{j-1}))A_j \geq A_k\epsilon$ for all k. Choose K so that $G(A_K) > 1-\epsilon/2$. Then

$$\sum_{j=1}^{K}(G(A_j) - G(A_{j-1}))A_j + \epsilon A_k/2 \geq A_k\epsilon$$

for all k. This is a contradiction, and the result holds.

Q.E.D

All of the preliminaries are completed for demonstrating that the existence of absolute moments of a random variable implies differentiability of the characteristic function to the same order.

Lemma 2.3.3: *If Y has a moment μ_l of order l (in the sense that $\mathrm{E}\left[|Y|^l\right] < \infty$), then the derivative of ζ of order l exists at zero, with $\zeta^{(l)}(0) = \mu_l i^l$, and $\zeta(\beta) = \sum_{k=0}^{l}\mu_k(i\beta)^k/k! + o(\beta^l)$.*

Proof: Let $F(y)$ be the distribution function for $|Y|$. Substituting the random variable βY for y in Lemma 2.3.1, and taking expectations on both sides,

$$\left|\zeta(\beta) - \sum_{k=0}^{l}\beta^k i^k \mu_k/k!\right| \leq |\beta|^l\, \mathrm{E}\left[\min(|\beta|\,|Y|^{l+1}/(l+1)!, 2\,|Y|^l/l!)\right]. \tag{14}$$

If $G(y) = \int_{-y}^{y}|y|^l\ dF(y)/\mathrm{E}\left[|Y|^l\right]$, then

$$\left|\zeta(\beta) - \sum_{k=0}^{l}\frac{\beta^k i^k\mu_k}{k!}\right| \leq \frac{\mathrm{E}\left[|Y|^l\right]|\beta|^l}{l!}\left[\frac{|\beta|}{l+1}\int_0^{2(l+1)/|\beta|} y\ dG(y) + 2G\left(\frac{2(l+1)}{|\beta|}\right)\right].$$

The second term in brackets above clearly goes to zero as $|\beta| \to 0$. The first term in brackets above goes to zero as $|\beta| \to 0$, by Lemma 2.3.2. Hence if Y has moments $\mu_1, ..., \mu_l$, then $\zeta(\beta) = \sum_{k=0}^{l}\mu_k(i\beta)^k/k! + o(\beta^l)$, and ζ is l-times differentiable at zero with the derivatives claimed.

Q.E.D

The previous result shows that if Y has a moment of a certain order, then the corresponding derivative of ζ_Y also exists. The next result is a partial converse.

Lemma 2.3.4: *If l is an even integer, and if the derivative of ζ of order l exists at zero, then Y has a moment of order l given by (10).*

Proof: This proof is essentially that given by Cramér (1946). Let $g_0(\beta, y) = \exp(\beta y)$ and $g_k(\beta, y) = ((\exp(\beta y) - \exp(-\beta y))/(2\beta))^k$ for $k > 0$. Furthermore, let

Z be the random variable taking the value of Y if $|Y| < K$ and zero otherwise. Exercise 1 of this chapter outlines a proof that if the derivative of order k of ζ at 0 exists, then

$$\zeta^{(k)}(0) = \lim_{\beta \to 0} \mathrm{E}\left[g_k(\beta, iY)\right]. \tag{15}$$

Hence

$$\zeta^{(l)}(0) = \lim_{\beta \to 0} \mathrm{E}\left[g_l(\beta, iY)\right] \geq \lim_{\beta \to 0} \mathrm{E}\left[g_l(\beta, iZ)\right] = \mathrm{E}\left[Z^l\right]$$

for all K. The last equality holds by the Bounded Convergence Theorem, since $|(\exp(i\beta y) - \exp(-i\beta y))/(2\beta)| \leq |y|$. By the Monotone Convergence Theorem, $\zeta^{(l)}(0) = \mathrm{E}\left[Y^l\right]$.

Q.E.D

If l is odd, the Monotone Convergence Theorem does not apply. The claim of the lemma does not hold for odd l. Left as an exercise is a counterexample in which the first derivative of the characteristic function of a random variable is defined at 0 even though the first absolute moment does not exist. The principal value of the associated sum is 0.

When derivatives of a characteristic function of all orders exist, one might consider constructing the power series representation for the function. The next theorem considers the radius of convergence for this series.

Theorem 2.3.5: *If Y has moments of all orders, and $R = 1/\limsup(\mu_l/l!)^{1/l}$ then ζ has the expansion $\sum_{k=0}^{\infty}(i\beta)^k\mu_k/k!$ valid on $|\beta| < R$. This radius R might be zero, in which case the series expansion holds for no $\beta \neq 0$, or R might be ∞, in which case the expansion holds for all $\beta \in \Re$.*

Proof: First, a note about notation. Given a sequence a_n of real numbers, $\limsup a_n$ is the upper bound on the set of limits of subsequences of the sequence a_n, if any convergent subsequences exist, and is infinite otherwise. If the sequence a_n converges, then $\lim_{n\to\infty} a_n = \limsup a_n$.

The theorem follows from (14).

Q.E.D

These results for characteristic functions also hold for moment generating functions; in fact, since the existence of a moment generating function is a rather restrictive condition, results for moment generating functions are stronger:

Lemma 2.3.6: *Suppose that the moment generating function $\mathcal{M}_Y(\beta)$ exists for β in a neighborhood $(-\epsilon, \epsilon)$ of 0. Then moment of Y of order k, $\mu_k = \mathrm{E}\left[Y^k\right]$ exist for all k and are given by $\mu_k = \mathcal{M}_Y^{(k)}(\beta)$.*

Proof: Choose $\beta \in (-\epsilon, \epsilon) \times i\Re$. Choose $\gamma \in (\Re(\beta) - (\epsilon - |\Re(\beta)|)/2, \Re(\beta) + (\epsilon - |\Re(\beta)|)/2)) \times i\Re$. Then

$$(\mathcal{M}_Y(\beta) - \mathcal{M}_Y(\gamma))/(\beta - \gamma) = \mathrm{E}\left[[\exp(Y\beta) - \exp(Y\gamma)]/(\beta - \gamma)\right], \tag{16}$$

and for some β^* between β and γ,

$$\left|\frac{\exp(Y\beta) - \exp(Y\gamma)}{\beta - \gamma}\right| = |Y \exp(Y\beta^*)|$$
$$\leq |Y| \exp(Y[|\Re(\beta)|/2 + \epsilon/2]) \leq \sup_{w\in\Re}(w \exp(w(|\Re(\beta)| - \epsilon)/2) \exp(Y\epsilon).$$

The supremum above is finite. Since this last quantity is integrable, the Dominated Convergence Theorem allows us to interchange expectation and $\lim_{\beta\to\gamma}$ to show that $\mathcal{M}_Y(\beta)$ is differentiable as a complex function on $(-\epsilon, \epsilon) \times i\mathfrak{R}$. Hence $\mathcal{M}_Y$ has derivatives of all orders, and Lemma 2.3.4 applies.

Q.E.D

Alternatively, we might avoid using Lemma 2.3.4 by noting that the proof to Lemma 2.3.6 shows that $\mathcal{M}'_Y(\beta) = \mathrm{E}\left[Y \exp(\beta Y)\right]$; the argument might be iterated to show that $\mathcal{M}_Y^{(k)}(\beta) = \mathrm{E}\left[Y^k \exp(\beta Y)\right]$.

2.4. Inversion of Characteristic Functions

The following theorem on inverting characteristic functions to recover the underlying cumulative distribution function is found in Billingsley (1986).

Theorem 2.4.1: *If a distribution function F corresponds to a characteristic function ζ and the points b_1 and b_2 have zero probability assigned to them then*

$$F(b_2) - F(b_1) = \lim_{\Theta\to\infty} \frac{1}{2\pi} \int_{-\Theta}^{\Theta} \frac{\exp(-i\beta b_1) - \exp(-i\beta b_2)}{i\beta} \zeta(\beta)\, d\beta. \tag{17}$$

Proof: Let $I_\Theta = \frac{1}{2\pi} \int_{-\Theta}^{\Theta} \frac{\exp(-i\beta b_1) - \exp(-i\beta b_2)}{i\beta} \zeta(\beta)\, d\beta$. Then

$$I_\Theta = \frac{1}{2\pi} \int_{-\Theta}^{\Theta} \int_{-\infty}^{\infty} \frac{\exp(-i\beta b_1) - \exp(-i\beta b_2)}{i\beta} \exp(i\beta x)\, dF(x)\, d\beta$$
$$= \frac{1}{2\pi} \int_{-\Theta}^{\Theta} \int_{-\infty}^{\infty} \frac{\exp(i\beta(x - b_1)) - \exp(i\beta(x - b_2))}{i\beta}\, d\beta\, dF(x).$$

Interchange of the order of integration is justified because the integrand is bounded and the set of integration is of finite measure. Expanding the complex exponential, one can express I_Θ as

$$\int_{-\Theta}^{\Theta} \int_{-\infty}^{\infty} \frac{\cos(\beta(x - b_1)) + i \sin(\beta(x - b_1)) - \cos(\beta(x - b_2)) - i \sin(\beta(x - b_2))}{2\pi i\beta}\, d\beta\, dF(x).$$

Since the cosine is an even function, the integral of terms involving cosine is zero, leaving

$$I_\Theta = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\Theta}^{\Theta} \frac{\sin(\beta(x - b_1)) - \sin(\beta(x - b_2))}{\beta}\, d\beta\, dF(x).$$