John E. Kolassa

Series Approximation Methods in Statistics

Third Edition

Springer

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

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

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 $\mathfrak{R}$ and $\mathfrak{I}$ 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

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 $\mathrm{P}\left[\cdot\right]$ 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 $\mathrm{P}\left[X_n(\omega) \to Y(\omega)\right] = 1$. 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 $\mathrm{E}\left[|X_n(\omega) - Y(\omega)|^r\right] \to 0$. 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 $\lim_{n \to \infty} \mathrm{P}\left[|X_n - Y| < \epsilon\right] = 1$. 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

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

for all x continuity points of $\mathrm{P}\left[Y(\omega) \le x\right]$. 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

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

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
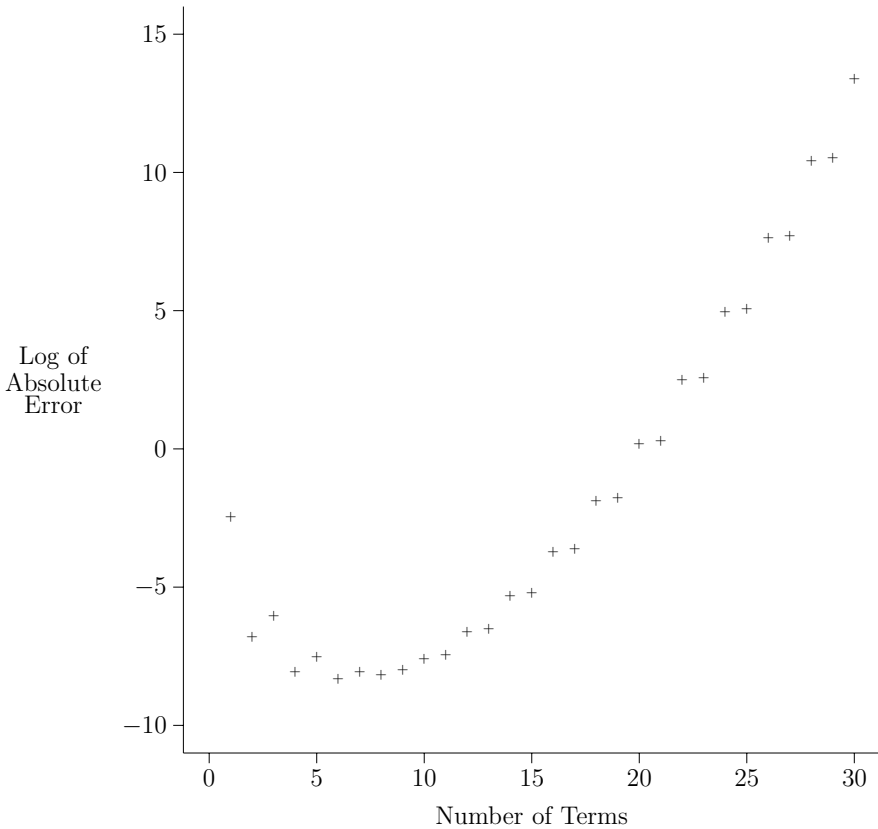
Fig. 1.

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 $G_{j,n}$ 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

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

$\zeta_X(\beta)$ 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 $X = a_1 X_1 + a_2 X_2 + b$, then

$$\begin{aligned}\zeta_X(\beta) &= \mathrm{E}\left[\exp(i[a_1 X_1 + a_2 X_2 + b]\beta)\right] \\ &= \exp(ib\beta)\mathrm{E}\left[\exp(ia_1 X_1 \beta)\right]\mathrm{E}\left[\exp(ia_2 X_2 \beta)\right] = \exp(ib\beta)\zeta_{X_1}(a_1\beta)\zeta_{X_2}(a_2\beta). \end{aligned} \quad (9)$$

Hence $\zeta_X(\beta) = \prod_{j=1}^{n} \zeta_{X_j}(a_j\beta)$ if $X = \sum_{j=1}^{n} a_j X_j$ 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 ‾ 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

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

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 $\beta \in \Re$; $\exp(\beta^2/2)$ also is defined and differentiable for $\beta \in \Re$. Since these functions coincide on a set of points converging to 0, then $\mathcal{M}_X(\beta) = \exp(\beta^2/2)$ for $\beta \in \Re$ (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 $(-1/2, 1/2)$:

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

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,* $\left| \exp(iy) - \sum_{k=0}^{l} (iy)^k/k! \right| \leq \min \left(\dfrac{|y|^{l+1}}{(l+1)!}, \dfrac{2\,|y|^l}{l!} \right).$

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,

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

and hence

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

$$\text{Q.E.D}$$

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.

<div align="right">Q.E.D</div>

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,

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

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.

<div align="right">Q.E.D</div>

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 $\beta \in \mathfrak{R}$.*

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 $\beta \in (-\epsilon, \epsilon) \times i\mathfrak{R}$. Choose $\gamma \in (\mathfrak{R}(\beta) - (\epsilon - |\mathfrak{R}(\beta)|)/2, \mathfrak{R}(\beta) + (\epsilon - |\mathfrak{R}(\beta)|)/2) \times i\mathfrak{R}$. 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 $(-\epsilon, \epsilon) \times i\Re$. Hence $\mathcal{M}_Y$ has derivatives of all orders, and Lemma 2.3.4 applies.

$$\text{Q.E.D}$$

Alternatively, we might avoid using Lemma 2.3.4 by noting that the proof to Lemma 2.3.6 shows that $\mathcal{M}'_Y(\beta) = \mathrm{E}\,[Y \exp(\beta Y)]$; 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*

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

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

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

Let $S(\theta) = \int_0^\theta \sin(x)/x \ dx$. Since the quantities $q_k = \int_{(k-1)\pi}^{k\pi} \sin(x)/x \ dx$ form an alternating sequence with terms whose absolute values are decreasing, then $S(\infty) = \lim_{\theta \to \infty} S(\theta)$ is finite. By the dominated convergence theorem,

$$\lim_{\Theta \to \infty} I_\Theta = \int_{-\infty}^\infty \frac{1}{2\pi} (2S(\infty) \operatorname{sgn}(x - b_1) - 2S(\infty) \operatorname{sgn}(x - b_2)) \ dF(x)$$

Since $\operatorname{sgn}(x - b_1) - \operatorname{sgn}(x - b_2)$ is 1 for $x \in (b_1, b_2)$ and 0 for $x \in [b_1, b_2]^c$, then for $\Theta > \max(|b_1|, |b_2|)$, $I_\Theta = c \int_{b_1}^{b_2} \ dF(x) = c(F(b_2) - F(b_1))$, for $c = S(\infty)/\pi$, which will be shown to equal 1. The result follows.

Q.E.D

The first factor in the integrand in (17) has a removable singularity at zero; that is, although it is undefined for $\beta = 0$, if the value $b_2 - b_1$ is substituted for $(\exp(-i\beta b_1) - \exp(-i\beta b_2))/i\beta$ when $\beta = 0$ the resulting factor is a differentiable complex function.

A version of this theorem for recovering density functions also exists.

Theorem 2.4.2: *The density of a random variable with the characteristic function ζ satisfying*

$$\int_{-\infty}^\infty |\zeta(\beta)| \ d\beta < \infty \tag{18}$$

exists and is given by

$$\frac{1}{2\pi} \int_{-\infty}^\infty \exp(-i\beta y)\zeta(\beta) \ d\beta; \tag{19}$$

furthermore, the constant c in the previous theorem has the value 1.

Proof: By (18) one can replace the limit of proper integrals as $\Theta \to \infty$ by the corresponding improper integral over the real line in (17). Furthermore, the bounded convergence theorem assures that the right hand side of (17) is continuous in b_1 and b_2. Then

$$
\begin{aligned}
F(b_2) - F(b_1) &= \frac{c}{2\pi} \int_{-\infty}^\infty \frac{\exp(-i\beta b_1) - \exp(-i\beta b_2)}{i\beta} \zeta(\beta) \ d\beta \\
&= \frac{c}{2\pi} \int_{-\infty}^\infty \int_{b_1}^{b_2} \exp(-i\beta y) \ dy \ \zeta(\beta) \ d\beta \\
&= \int_{b_1}^{b_2} \frac{c}{2\pi} \int_{-\infty}^\infty \exp(-i\beta y)\zeta(\beta) \ d\beta \ dy.
\end{aligned}
$$

Interchange of the order of integration is justified by Fubini's Theorem using the absolute convergence of the integral (18). By comparing the last line with the normal characteristic function, we find that $c = 1$. Hence the density is (19). The bounded convergence theorem implies that the resulting density is continuous.

Q.E.D

A useful corollary to Theorem 2.4.2 may formulated for functions that are not necessarily densities.

Corollary 2.4.3: *Suppose that two functions (not necessarily a density and characteristic function) are related by the equation $\zeta(\beta) = \int_{-\infty}^{\infty} \exp(i\beta x) f(x) \, dx$, and that both are uniformly integrable. Then $f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-i\beta x)\zeta(\beta) \, d\beta$.*

Proof: Apply Theorem 2.4.2 to the positive and negative parts of $f(x)$ separately, after rescaling to make each part integrate to 1.

<div align="right">Q.E.D</div>

A condition like integrability of the first derivative of the density implies the condition on integrability of the characteristic function (18). The presence of probability atoms, or singleton sets of positive probability, implies that condition (18) is violated.

The converse of Theorem 2.4.2 is not true. The existence of a density need not imply absolute convergence of the integral of the characteristic function. The question of minimal conditions on a function ζ to make it a characteristic function, and particularly the characteristic function of a distribution with a density, arises in functional analysis. The interested reader should consult Rudin (1973). We will instead ask the weaker question of what characteristic functions correspond to distributions whose convolutions with itself eventually have densities. For the present purposes this later question is more important, in that most of the density expansion theorems will hold as the number of summands in the quantity whose distribution is approximated is sufficiently large; hence it is necessary to consider when the quantity being approximated actually exists.

Lemma 2.4.4: *The characteristic function ζ of a random variable satisfies*

$$\int_{-\infty}^{\infty} |\zeta(\beta)|^{\mathfrak{r}} \, d\beta < \infty$$

for some $\mathfrak{r} > 1$ if and only if there exists an integer j such that the density of a j-fold convolution of the random variable with itself exists and is bounded.

Proof: The "only if" clause holds by the preceding theorem applied to $j = \lceil \mathfrak{r} \rceil$, where $\lceil \mathfrak{r} \rceil$ is the smallest integer at least as large as $\mathfrak{r}$. The "if" clause holds since for any $\Theta > 0$,

$$\int_{-\Theta}^{\Theta} \left|\zeta^{2j}(\beta)\right| \, d\beta = \int_{-\Theta}^{\Theta} \left[\int_{-\infty}^{\infty} f_j(x) \exp(i\beta x) \, dx\right] \left[\int_{-\infty}^{\infty} f_j(y) \exp(-i\beta y) \, dy\right] \, d\beta$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\Theta}^{\Theta} f_j(x) \exp(i\beta x) f_j(y) \exp(-i\beta y) \, d\beta \, dy \, dx;$$

the interchange in order of integration holds by Fubini's Theorem. Then

$$\int_{-\Theta}^{\Theta} \left|\zeta^{2j}(\beta)\right| \, d\beta = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\Theta}^{\Theta} f_j(x) f_j(x+z) \exp(-i\beta z) \, d\beta \, dx \, dz$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_j(x) f_j(x+z) \frac{\exp(-i\Theta z) - \exp(i\Theta z)}{-iz} \, dx \, dz.$$

Imaginary parts in the previous integral cancel, leaving

$$\int_{-\Theta}^{\Theta} \left| \zeta^{2j}(\beta) \right| \, d\beta = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_j(x) f_j(x+z) 2 \sin(\Theta z) z^{-1} \, dx \, dz.$$

Changing variables from z to $v = \Theta z$, and splitting the range of integration for v,

$$\int_{-\Theta}^{\Theta} \left| \zeta^{2j}(\beta) \right| \, d\beta = \int_{-\infty}^{\infty} \int_{-\pi}^{\pi} f_j(x) f_j(x+v/\Theta) \frac{2 \sin(v)}{v} \, dv \, dx +$$
$$\int_{-\infty}^{\infty} \int_{(-\pi,\pi)^c} f_j(x) f_j(x+v/\Theta) \frac{2 \sin(v)}{v} \, dv \, dx.$$

The first integral above is bounded by the maximum value of f_j times the integral of $2 \sin(v)/v$ over $(-\pi, \pi)$; the second integrand is bounded by $f_j(x) f_j(x+v/\Theta) 2\pi^{-1}$, and hence the integrating with respect to v and then with respect to x shows that the second integral can be bounded by $2\pi^{-1}$, which is independent of Θ.

Q.E.D

The inversion integrand in (19) is the derivative of the inversion integrand in (17), evaluated with $b_1 = b_2$. This will become important in our discussion of saddlepoint series.

2.5. Bounding Distribution Function Differences using Fourier Transforms

The following material is adapted from Feller (1971). Suppose X_n is a sequence of random variables, with cumulative distribution functions F_n and characteristic functions ζ_n, and suppose X has the cumulative distribution function F and characteristic function ζ, such that $\zeta_n \to \zeta$ pointwise. Then $F_n(x) \to F(x)$ for all x where F is continuous. This in known as weak convergence. A classical elegant probabilistic proof can be found in Billingsley (1986), §26. A classical analytic proof can be found in Feller (1971), §VIII.1. In the present context higher order approximations to $F_n(x)$ as n varies are considered. To generate these approximations additional regularity conditions will be required.

Series expansions will be generated by approximating the appropriate Fourier inversion integral. The first step is to investigate what changes in the cumulative distribution function arise from small changes in a characteristic function. Fundamental is a theorem relating differences in cumulative distribution functions to differences in characteristic functions. Suppose a cumulative distribution function H with expectation 0 is to be approximated. Approximations G having properties similar to those of a cumulative distribution function of a continuous random variable will be considered. Specifically, G will be required to have a Fourier transform $\xi_n(\beta) = \int_{-\infty}^{\infty} \exp(i\beta x) \, dG$ whose derivative will have the correct value of 0 at 0, and the limiting values of $G(x)$ as the argument becomes large will be the same as for a cumulative distribution function. Heuristically the difference between H and G

will be bounded by bounding the difference in inversion integrals (17) to obtain

$$H(b_2) - H(b_1) - G(b_2) + G(b_1)$$
$$= \lim_{\Theta \to \infty} \int_{-\Theta}^{\Theta} \frac{\exp(-i\beta b_1) - \exp(-i\beta b_2)}{2\pi i \beta}(\zeta(\beta) - \xi_n(\beta)) \ d\beta. \qquad (20)$$

The coarsest bound for (20), obtained by bounding the numerator of the integrand by 2, is

$$|H(b_2) - H(b_1) - G(b_2) + G(b_1)| \leq \lim_{\Theta \to \infty} \frac{1}{2\pi} \int_{-\Theta}^{\Theta} \frac{2}{|\beta|} |\xi(\beta) - \zeta(\beta)| \ d\beta. \qquad (21)$$

The integrand in (21) has a removable singularity at 0 and hence the integral in (21) is finite for Θ finite. The integral over $(-\infty, \infty)$, however, need not converge, and so some device is necessary to allow the limit in (21) to be truncated after some finite Θ. In fact a result which explicitly bounds the error incurred when the cumulative distribution functions on the left hand side of (20) are replaced by smoothed versions, is proved. The random variable added to smooth these cumulative distribution functions has a distribution which will vary with Θ. Its characteristic function disappears outside of $(-\Theta, \Theta)$. Since the Fourier transform of a sum of random variables is the product of the individual Fourier transforms, the Fourier transforms of the smoothed variables will then be zero outside of a bounded region, and the limit in (21) can be replaced by the improper integral. The amount of error incurred in the substitution will depend on Θ. The Edgeworth convergence theorems for cumulative distribution functions later in this book will then be proved by balancing the contribution of this error term with other terms, and so Θ is here left unspecified.

Lemma 2.5.1:   *Suppose H is a cumulative distribution function, and G is a function with the proper limits at infinity, such that G' exists and $|G'| \leq m < \infty$. Let $D = H - G$ and $D^{\Theta}(y) = \int_{-\infty}^{\infty} D(y-x)v(x\Theta)\Theta \ dx$ be D smoothed by adding on a little bit of a continuous variate with density $v(x) = \pi^{-1}(1 - \cos(x))/x^2$. Let $\eta = \sup |D|$ and $\eta_{\Theta} = \sup \left|D^{\Theta}\right|$. Then $\eta_{\Theta} \geq \eta/2 - 12m/(\pi\Theta)$.*

Proof: Unlike most of the results in this volume, the present lemma provides a lower bound for a maximum difference, rather than an upper bound. The proof will proceed by finding a point x_0 at which the difference should, heuristically, be largest, and use that difference as the lower bound for the maximal difference.

Construct a smoothing variable V with density

$$v(x) = (1 - \cos(x))/(\pi x^2) \qquad (22)$$

on $(-\infty, \infty)$, and characteristic function

$$\omega(\beta) = \begin{cases} 1 - |\beta| & \text{if } |\beta| \leq 1 \\ 0 & \text{if } |\beta| > 1 \end{cases}. \qquad (23)$$

Verification that this is indeed the characteristic function is left as an exercise.

The proof is completed by constructing a neighborhood to the left of x_0 where D is bounded below. Since D is right continuous, there exists x_0 such that $|D(x_0)| = \eta$. Without loss of generality assume $D(x_0) = \eta$. Refer to Fig. 2. Note that $D(x_0 + s) \geq \eta - ms$ for $s > 0$, since H is non-decreasing and the derivative of G is bounded by m. The supporting line $\eta - ms$ hits the x axis at $x_2 = x_0 + \eta/m$. Let x_1 be midway between these. Call the supporting function just defined

$$D^L = \begin{cases} \eta/2 - (x - x_1)m & \text{if } x \in (x_0, x_2) \\ -\eta & \text{otherwise.} \end{cases}$$

Bound below the convolution D^Θ at x_1 by the convolution of D^L with V/Θ, where V has density v. Note that $\mathrm{P}\left[|V/\Theta| \leq v\right] \leq 4 \int_{v\Theta}^\infty (\pi x^2)^{-1}\ dx = 4/(\pi v\Theta)$. Since on (x_0, x_2) the support function D^L is a constant plus a linear term, and the linear part convolves to 0 by symmetry, the contribution to the convolution here is at least $\eta/2$ times the minimum probability of the interval, $1 - 4/(\pi\Theta(x_2 - x_1))$, and since the probability mass for V/Θ outside this region is no more than $4/(\pi\Theta(x_2 - x_1))$ the entire convolution is at least $\eta/2 \times [1 - 4/(\pi\Theta(\eta/2m))] - \eta \times (4/\pi\Theta(\eta/2m)) = \eta/2 - 12m/(\pi\Theta)$.

Q.E.D

Feller, 1971, §XVI.4 presents the following Smoothing Theorem.

Theorem 2.5.2: *Suppose H is a cumulative distribution function with expectation 0 and characteristic function ζ, and G is differentiable, its Fourier transform ξ has a derivative ξ' which takes the value 0 at 0, and G has the limits 0 and 1 as its argument approaches $\pm\infty$ respectively. Then*

$$|H(x) - G(x)| \leq \frac{1}{\pi} \int_{-\Theta}^{\Theta} |\xi(\beta) - \zeta(\beta)|\ /\ |\beta|\ d\beta + 24 \max(|G'|)/(\pi\Theta). \tag{24}$$

Proof: Both H and G will be convolved with a rescaled v to create differentiable functions. The preceding lemma shows that a bound on the smoothed differences yields a bound on the original differences.

Denote by H^Θ and G^Θ the convolutions of H and G with the smoothing variable V/Θ, where V is as in Lemma 2.5.1. Since V/Θ is a continuous variable, these have derivatives h^Θ and g^Θ. Letting $b_1 \to -\infty$ and replacing b_2 by x in (21), and noting that the characteristic functions of H^Θ and G^Θ disappear outside of $(-\Theta, \Theta)$,

$$\left|D^\Theta\right| \leq \frac{1}{2\pi} \int_{-\infty}^{\infty} |\zeta(\beta) - \xi(\beta)|\, |\omega(\beta/\Theta)|\ d\beta/\, |\beta|\,.$$

Bound $|\omega(\beta/\Theta)|$ by 1. Then $\eta^\Theta \leq \pi^{-1} \int_{-\Theta}^{\Theta} |(\zeta(\beta) - \xi(\beta))/\beta|\ d\beta$ and the assertion follows.

Q.E.D

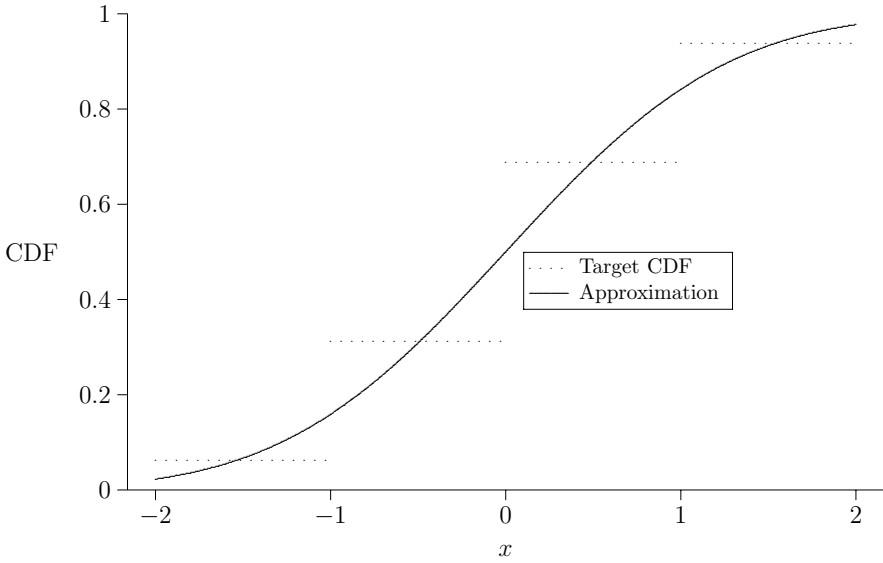*Geometric Construction for Smoothing Lemma; Centered Binomial (4 , 0.5) Example*
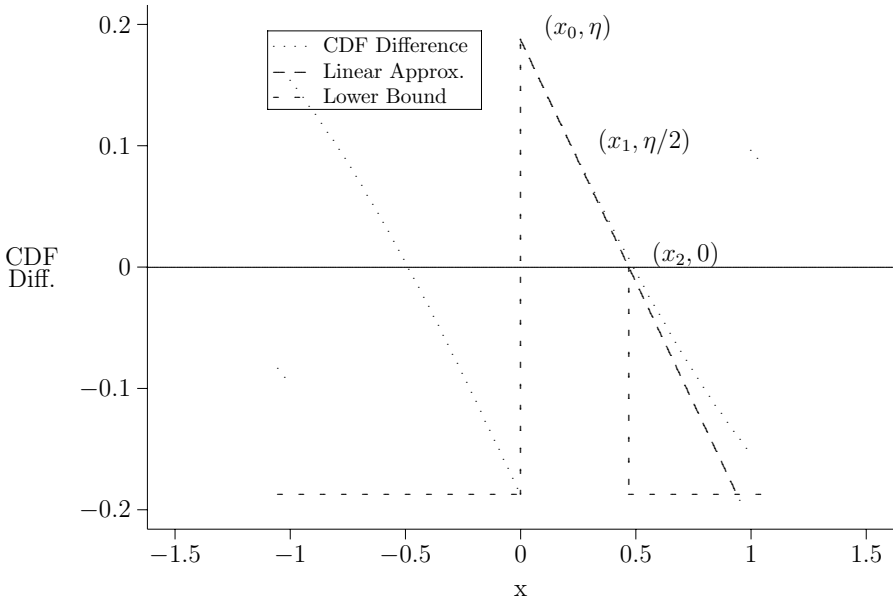

Fig. 2a

Fig. 2b

A bounding inequality similar to (21) for densities can be derived from (19). If H and G have Fourier transforms $\xi(\beta)$ and $\zeta(\beta)$ and derivatives h and g, then from (19),

$$|h(x) - g(x)| \leq \frac{1}{2\pi} \int_{-\infty}^{\infty} |\xi(\beta) - \zeta(\beta)| \ d\beta. \tag{25}$$

The next step is to measure variation between the true cumulant generating function and the approximation based on its Taylor expansion that will be then used to generate cumulative distribution function approximations. Henceforth this result will be known as the Series Theorem.

Theorem 2.5.3: *For any complex α and v, and any l a non-negative integer,*

$$\left| \exp(\alpha) - \sum_{k=0}^{l} \frac{v^k}{k!} \right| \leq \max(\exp(|\alpha|), \exp(|v|)) \left(|\alpha - v| + \left| \frac{v^{l+1}}{(l+1)!} \right| \right).$$

Proof: By the triangle inequality,

$$\left| \exp(\alpha) - \sum_{k=0}^{l} \frac{v^k}{k!} \right| \leq |\exp(\alpha) - \exp(v)| + \left| \sum_{k=l+1}^{\infty} \frac{v^k}{k!} \right|.$$

The second term is bounded by

$$\left| \sum_{k=l+1}^{\infty} \frac{v^k}{k!} \right| \leq \left| \frac{v^{l+1}}{(l+1)!} \right| \sum_{k=0}^{\infty} |v^k| \frac{(l+1)!}{(k+l+1)!}$$

$$\leq \left| \frac{v^{l+1}}{(l+1)!} \right| \sum_{k=0}^{\infty} \frac{|v^k|}{k!} \leq \left| \frac{v^{l+1}}{(l+1)!} \right| \max(\exp(|v|), \exp(|\alpha|)).$$

Also, by Taylor's Theorem, for some $\theta \in (0,1)$,

$$|\exp(\alpha) - \exp(v)| = |v - \alpha| \exp(\theta\alpha + (1-\theta)v)$$
$$\leq |v - \alpha| \max(\exp(|\alpha|), \exp(|v|)).$$

$$\text{Q.E.D}$$

The following result using Theorem 2.5.3 will be used in the construction of Edgeworth series.

Corollary 2.5.4: *Suppose that $\zeta(\boldsymbol{\beta})$ is a function $\mathfrak{R}^{\mathfrak{k}} \to \mathfrak{C}$ with $\mathfrak{g}$ derivatives in a neighborhood of $\mathbf{o}$, such that $\zeta(\mathbf{o}) = 1$, and let*

$$\kappa^{s_1 \cdots s_{\mathfrak{k}}} = (-i)^{s_1 + \cdots + s_{\mathfrak{k}}} \frac{d}{d\beta_{s_1}} \cdots \frac{d}{d\beta_{s_{\mathfrak{g}}}} \log(\zeta(\boldsymbol{\beta}))|_{\beta = \mathbf{o}}$$

for every vector $\boldsymbol{s} \in \{1, \ldots, \mathfrak{k}\}^{\mathfrak{g}}$. Suppose that $\kappa^{jk} = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{otherwise} \end{cases}$. Let

$$\xi(\boldsymbol{\beta}) = \exp(-\|\boldsymbol{\beta}\|^2/2) \sum_{k=0}^{\mathfrak{g}-2} \left[1 + \sum_{j=3}^{\mathfrak{g}} \sum_{s \in \{1,\ldots,\mathfrak{k}\}^j} \frac{i^j \kappa^{s_1 \cdots s_j} \beta_{s_1} \cdots \beta_{s_{\mathfrak{g}}}}{n^{j/2-1} j!} \right]^k / k!.$$

Then for every $\epsilon > 0$, there exists $\delta > 0$ and a constant C such that

$$\left| \exp(n[\log(\zeta(\boldsymbol{\beta}/\sqrt{n}))]) - \xi(\boldsymbol{\beta}) \right| \leq \exp(-\|\boldsymbol{\beta}\|^2/4) \left[\frac{\epsilon \|\boldsymbol{\beta}\|^{\mathfrak{g}}}{n^{\mathfrak{g}/2-1}} + \frac{C^{\mathfrak{g}-1} \|\boldsymbol{\beta}\|^{3(\mathfrak{g}-1)}}{(\mathfrak{g}-1)! n^{\mathfrak{g}/2-1/2}} \right],$$

for $\boldsymbol{\beta} \in \mathfrak{R}^{\mathfrak{k}}$ such that $\|\boldsymbol{\beta}\| < \delta \sqrt{n}$.

Proof: Set $\alpha^*(\boldsymbol{\beta}) = \log(\zeta(\boldsymbol{\beta})) + \sum_j (\beta_j)^2/2$, and

$$\upsilon^*(\boldsymbol{\beta}) = \sum_{j=3}^{\mathfrak{g}} \sum_{s \in S(j)} i^j \kappa^{s_1 \cdots s_{\mathfrak{k}}} \beta_{s_1} \cdots \beta_{s_{\mathfrak{k}}} / j!.$$

Set $\alpha(\boldsymbol{\beta}) = n\alpha^*(\boldsymbol{\beta}/\sqrt{n})$, and $\upsilon(\boldsymbol{\beta}) = n\upsilon^*(\boldsymbol{\beta}/\sqrt{n})$. Then $\alpha(\boldsymbol{\beta}) - \upsilon(\boldsymbol{\beta})$ has $\mathfrak{g}$ continuous derivatives at 0 and all are 0. Hence there exists $\delta > 0$ such that if $\|\boldsymbol{\beta}\| < \delta$ then

$$|\alpha^*(\boldsymbol{\beta}) - \upsilon^*(\boldsymbol{\beta})| < \epsilon \|\boldsymbol{\beta}\|^{\mathfrak{g}}, \tag{26}$$

and if $\|\boldsymbol{\beta}\| < \sqrt{n}\delta$ then $\alpha(\boldsymbol{\beta}) - \upsilon(\boldsymbol{\beta}) < \epsilon \|\boldsymbol{\beta}\|^{\mathfrak{g}} n^{1-\mathfrak{g}/2}$. Furthermore require that $\|\boldsymbol{\beta}\| < \delta$ imply that $\left| \zeta(\boldsymbol{\beta}) + \sum_j \beta_j^2/2 \right| < \|\boldsymbol{\beta}\|^2/4$; hence $\|\boldsymbol{\beta}\| < \delta\sqrt{n}$ implies that

$$|\alpha(\boldsymbol{\beta})| = \left| n\zeta(\boldsymbol{\beta}/\sqrt{n}) + \sum_j \beta_j^2/2 \right| < \|\boldsymbol{\beta}\|^2/4.$$

Also, since the first two derivatives of $\upsilon^*(\boldsymbol{\gamma})$ at zero are 0, and third derivative of $\upsilon^*(\boldsymbol{\gamma})$ is bounded on $\{\boldsymbol{\gamma} | \|\boldsymbol{\gamma}\| < \delta\}$, then $\|\boldsymbol{\beta}\| < \delta\sqrt{n}$ implies that

$$\upsilon(\boldsymbol{\beta}) = n\upsilon^*(\boldsymbol{\beta}/\sqrt{n}) < C \|\boldsymbol{\beta}\|^3 / \sqrt{n}$$

for

$$C = \sup_{\|\boldsymbol{\gamma}\| \leq \delta, i,j,k \in \{1,\mathfrak{k}\}} \mathfrak{k}^3 \left| \upsilon^{*ijk}(\boldsymbol{\gamma}) \right|. \tag{27}$$

The result follows from Theorem 2.5.3.

<div align="right">Q.E.D</div>

Corollary 2.5.4 is generally applied with ζ the characteristic function of a possibly multivariate distribution. The coefficients κ with superscripts correspond to the cumulants of the associated distribution.

Lemma 2.3.1 represented a result on a topic similar to that of Lemma 2.5.3. Lemma 2.3.1 bounded the difference between the exponential of a real number and its Taylor expansion, while Lemma 2.5.3 bounds the difference between the exponential of a general complex number and the value of its Taylor expansion evaluated

at a neighboring complex number. Hence Lemma 2.3.1 requires more restrictive conditions, and in return avoids the absolute values applied to the argument of the exponential in the bound of Lemma 2.5.3.

2.6. The Berry–Esseen Theorem

Before proceeding to cumulative distribution function approximations representing refinements of the normal approximation, we assess the error incurred when using the normal approximation. Here of interest is not just the asymptotic order of the error, but actual numeric bounds on the error. Theorems addressing this question are generally known as Berry–Esseen theorems.

Theorem 2.6.1: *Let $\{Y_n\}$ be a sequence of independent and identically distributed random variables with zero mean and unit variance. Let $Z = \sum_{j=1}^{n} Y_j / \sqrt{n}$, and $F_Z(z) = \mathrm{P}\left[Z \leq z\right]$. Suppose further that $\rho = \mathrm{E}\left[|Y_1|^3\right] < \infty$. Then $|F_Z - \Phi| \leq C\rho/\sqrt{n}$, where C is a constant independent of the distributions of the variables Y_n whose value can be taken to be 3.*

Proof: This theorem will be proved by using the Smoothing Theorem. Use the series expansion techniques introduced earlier to bound the characteristic function as $|\zeta(\beta)| \leq 1 - \beta^2/2 + |\beta|^3 \rho/6$ if $\beta^2/2 \leq 1$. The integrand is bounded over finite regions. The boundary of the finite region, Θ, is chosen to balance the contributions from the two terms in (24). Choose $\Theta = (4/3)\sqrt{n}/\rho$. Since $\mathrm{E}\left[|Z^2|\right]^{3/2} \leq \mathrm{E}\left[|Z|^3\right]$ by Jensen's inequality, $\rho > 1$ and $\Theta \leq (4/3)\sqrt{n} < \sqrt{2n}$. Hence the theorem is trivially true for $n < 10$. Using the Smoothing Theorem,

$$\pi \left|F_Z - \Phi\right| \leq \int_{-\Theta}^{\Theta} \left|\zeta^n(\beta/\sqrt{n}) - \exp(-\beta^2/2)\right| / |\beta| \ d\beta + 24 \max(|\phi|)/\Theta, \qquad (28)$$

where $\max(|\phi|) = 1/\sqrt{2\pi} \leq 2/5$. (This corresponds to a lower bound on π of $25/8 = 3.125$.) Using Lemma 2.3.1, for l an integer and z real,

$$\left|e^{iz} - \sum_{j=0}^{l}(iz)^j/j!\right| \leq \min(|z|^{l+1}/(l+1)!, 2\,|z|^l/l!);$$

hence $|\zeta(\beta) - 1 + \beta^2/2| \leq |\beta|^3 \rho/6 \ \forall \beta \in \mathfrak{C}$, yielding, after rescaling, $|\zeta(\beta/\sqrt{n}) - 1 - \beta^2/(2n)| \leq |\beta|^3 \rho n^{-3/2}/6$ and $|\zeta(\beta/\sqrt{n})| \leq 1 - \beta^2/(2n) + |\beta|^3 \rho/(6n^{3/2})$ if $|\beta| \leq \Theta$. Hence

$$\left|\zeta(\beta/\sqrt{n})\right| \leq 1 - \beta^2/(2n) - |\beta|^2\,4/(3 \times 6n)$$
$$= 1 - 5/(18n)\beta^2 \leq \exp((-5/18n)\beta^2) \text{ if } |\beta| \leq \Theta$$

and

$$\left|\zeta(\beta/\sqrt{n})\right|^{n-1} = \exp((-5(n-1)/18n)\beta^2) \text{ if } n \geq 10$$
$$\leq \exp(-\beta^2/4).$$

For α and υ complex, with $|\alpha| \geq |\upsilon|$,

$$|\alpha^n - \upsilon^n| = |\alpha - \upsilon| \left| \sum_{j=0}^{n-1} \alpha^j \upsilon^{n-1-j} \right| \leq |\alpha - \upsilon| \sum_{j=0}^{n-1} |\alpha^j| |\upsilon^{n-1-j}| = |\alpha - \upsilon| \, n \, |\alpha|^{n-1} \, .$$

Hence the integrand in (28) is bounded by $|\zeta(\beta/\sqrt{n}) - \exp(-\beta^2/2n)| \exp(-\beta^2/4) \times n |\beta|^{-1}$, and

$$\left| \zeta(\beta/\sqrt{n}) - \exp(-\beta^2/2n) \right| \leq \left| \zeta(\beta/\sqrt{n}) - 1 + \beta^2/(2n) \right| + \left| 1 - \beta^2/(2n) - \exp(-\beta^2/(2n)) \right|$$
$$\leq |\beta|^3 \rho/(6\sqrt{n}) + |\beta|^4/(8n)$$

Hence the integrand is bounded by

$$n \left(\frac{|\beta|^2 \, \rho}{6\sqrt{n}} + \frac{|\beta|^3}{8n} \right) \exp(-\beta^2/4)(4/3\sqrt{n})/\Theta \leq \left(\frac{2}{9} \beta^2 + \frac{1}{18} |\beta|^3 \right) \exp(-\beta^2/4)/\Theta.$$

This can be integrated using integration by parts over $(-\Theta, \Theta)$ to show

$$\pi \Theta \, |F_Z(z) - \Phi(z)| \leq (8/9)\sqrt{\pi} + 8/9 + 10 < 4\pi$$

Hence the theorem holds with $C = 3$.

<div align="right">Q.E.D</div>

The bound C is constructed rather crudely. It may be replaced by Esseen's original bound, 7.59, Esseen's unpublished bound, 2.9, or Wallace's unpublished bound 2.05.

Esseen (1956) showed by constructing a binomial counterexample that C must be greater than $(3 + \sqrt{10})/(6\sqrt{2\pi}) \approx .40$.

The constant C is required to hold for all distributions F and all sample sizes n. If uniformity is required only over all F as $n \to \infty$ then C may be replaced by a lower value. Beek (1972) shows that $C = 0.7975$ fulfills these less stringent requirements.

2.7. Inversion of Characteristic Functions for Lattice Variables

Now consider inversion integrals for lattice distributions. Suppose that X is a random variable taking values on the lattice $\{a + \Delta \mathfrak{Z}\}$. Assume further, without loss of generality, that $a = 0$ and $\Delta = 1$. The limiting operation (17) can be taken along any sequence of Θ going to infinity. Taking the limit along Θ that are integer multiples of 2π plus π will simplify the calculations of this section. Choose $x \in \{a + \Delta \mathfrak{Z}\}$.

$$
\begin{aligned}
\mathrm{P}\left[X = x\right] &= \lim_{\Theta \to \infty} \frac{1}{2\pi} \int_{-\Theta}^{\Theta} \frac{\exp(-i\beta(x - \frac{1}{2})) - \exp(-i\beta(x + \frac{1}{2}))}{i\beta} \zeta(\beta) \ d\beta. \\
&= \lim_{\Theta \to \infty} \frac{1}{2\pi} \int_{-\Theta}^{\Theta} \exp(-i\beta x) \frac{\exp(\frac{1}{2} i\beta) - \exp(-\frac{1}{2} i\beta)}{i\beta} \zeta(\beta) \ d\beta. \\
&= \lim_{\Theta \to \infty} \frac{1}{2\pi} \int_{-\Theta}^{\Theta} \frac{\exp(-i\beta x) \sinh(i\beta/2)}{i\beta/2} \zeta(\beta) \ d\beta.
\end{aligned}
$$

The same integral is involved in the inversion here as was involved in (19), with the introduction of $\sinh(iu/2)/(u/2)$. Generation of an asymptotic approximation to this integral will differ, however, because in the lattice case the tails of ζ do not die out. We need to express this as an integral over one period of ζ. The correct expression will be shown to be

$$
\mathrm{P}\left[X = x\right] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \exp(-i\beta x) \zeta(\beta) \ d\beta. \tag{29}
$$

This can be derived by working backwards. First, substitute in the definition of the characteristic function:

$$
\frac{1}{2\pi} \int_{-\pi}^{\pi} \exp(-i\beta x) \zeta(\beta) \ d\beta = \frac{1}{2\pi} \int_{-\pi}^{\pi} \sum_{j=-\infty}^{\infty} \mathrm{P}\left[X = x + j\right] \exp(i\beta(x + j) - i\beta x) \ d\beta.
$$

Now, separate out the term with zero multiplying $i\beta$ in the exponential from the other terms, and note that each of these other terms integrates to zero:

$$
\frac{1}{2\pi} \int_{-\pi}^{\pi} \exp(-i\beta x) \zeta(\beta) \ d\beta = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[\mathrm{P}\left[X = x\right] + \sum_{j \neq 0} \mathrm{P}\left[X = x + j\right] \exp(i\beta j) \right] \ d\beta
$$

$$
= \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{P}\left[X = x\right] \ d\beta = \mathrm{P}\left[X = x\right]
$$

or by proceeding from (17), and noting that characteristic function has period 2π. This periodicity allows the limit of the integral from $-\Theta$ to Θ to be transformed into an integral over $(-\pi, \pi]$ of an infinite sum. This infinite sum is then shown to equal the integrand in (29).

Of more interest are inversion integrals for tail probabilities. For x_0, $x \in \mathfrak{Z}$,

$$
\begin{aligned}
\mathrm{P}\left[x_0 > X \geq x\right] &= \sum_{y=x}^{x_0-1} \frac{1}{2\pi} \int_{-\pi}^{\pi} \zeta(\beta) \exp(-i\beta y) \ d\beta \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} \zeta(\beta) \sum_{y=x}^{x_0-1} \exp(-i\beta y) \ d\beta \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} \zeta(\beta) \frac{\exp(-i\beta x) - \exp(-i\beta x_0)}{1 - \exp(-i\beta)} \ d\beta.
\end{aligned}
\tag{30}
$$

Letting $x_0 \to \infty$,

$$
\begin{aligned}
\mathrm{P}\left[X \geq x\right] &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \zeta(\beta) \frac{\exp(-i\beta(x - \frac{1}{2}))}{2 \sinh(i\beta/2)} \ d\beta \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} \zeta(\beta) \frac{\exp(-i\beta(x - \frac{1}{2}))}{\beta} \frac{\beta/2}{\sinh(i\beta/2)} \ d\beta.
\end{aligned}
\tag{31}
$$

Unfortunately the result (31) is not strictly correct, since it assumes that $\exp(-i\beta x_0)$ converges to zero uniformly as $x_0 \to \infty$. In a later chapter on saddlepoint cumulative distribution function approximations attention will be restricted to cases in which $\zeta(\beta)$ exists for complex β in a neighborhood of 0, and the above manipulations will be justified by deforming the path of integration to make the real part of the argument of the exponential function in $\exp(-i\beta x_0)$ negative. In general (30) is true, and will be used to develop Edgeworth series expansions for lattice random variable cumulative distribution functions.

These equations are easily extended to cases with general positive Δ and general a. If X has characteristic function ζ and is confined to the lattice $\{a + \Delta\mathfrak{Z}\}$ then $X^* = (X - a)/\Delta$ has characteristic function $\zeta(\beta/\Delta) \exp(-i\beta a/\Delta)$ and satisfies the requirements for applying (29), (30), and (31). Hence

$$
\begin{aligned}
\mathrm{P}\left[X = x\right] &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \exp(-i\beta(x/\Delta - a/\Delta))\zeta(\beta/\Delta) \exp(-i\beta a/\Delta) \ d\beta \\
&= \frac{1}{2\pi} \int_{-\pi/\Delta}^{\pi/\Delta} \exp(-i\beta x)\zeta(\beta)\Delta \ d\beta.
\end{aligned}
\tag{32}
$$

Summing (32), interchanging the order of summation and integration, and using the standard formula for the sum of a geometric series,

$$
\mathrm{P}\left[x_0 > X \geq x\right] = \frac{1}{2\pi} \int_{-\pi/\Delta}^{\pi/\Delta} \zeta(\beta) \frac{\exp(-i\beta x) - \exp(-i\beta x_0)}{1 - \exp(-i\Delta\beta)} \Delta \ d\beta
\tag{33}
$$

These inversion integrals will be used to generate Edgeworth and saddlepoint series for lattice distributions in later chapters.

2.8. Exercises

1. Justify (15) by showing that if $f_0(\beta)$ is a function with n derivatives on an interval containing 0, and $f_k(\beta)$ is defined recursively by $f_k(\beta) = (f_{k-1}(\beta) - f_{k-1}(-\beta))/(2\beta)$ then the derivative of order n of f_0 at 0 is $\lim_{\beta \to 0} f_n(\beta)$. Consider using L'Hospital's Rule (See Rudin (1976), p. 109 and exercises.) Apply this to the function $f_0(\beta) = \mathrm{E}[\exp(iy\beta)]$, noting that the resulting $f_k(\beta) = \mathrm{E}[g_k(\beta, y)]$, to demonstrate Lemma 2.3.4.

2. Consider a random variable X taking values on the set of non-zero integers, assigning to the integer j, whether positive or negative, the probability c/j^2. (Hint: This problem can most easily be done using Fourier series methods; Rudin (1976) on page 191 outlines an approach.)

   a. What is the value of c?

   b. Show that $\mathrm{E}[|X|] = \infty$, and hence that $\mathrm{E}[X]$ does not exist, even though the distribution is symmetric about zero.

   c. Calculate $\zeta_X(\beta)$, and discuss its derivative at zero.

3. Show that the function defined in (23) is actually the characteristic function associated with the density in (22).

4. Show that the density determined in (19) is continuous.

5. Verify that the characteristic function for the Cauchy distribution is as given in (13), possibly making use of the inversion theorem. The right hand side of (13) is the density of the double exponential distribution rescaled and evaluated at β.

6. Prove that the quantity R defined in (11) is actually the radius of convergence of the power series representation of the moment generating function $\mathcal{M}_X(\beta)$.

7. Calculate the characteristic function for the following distributions. Calculate the radius of convergence for the series expansion about the origin, if it exists, and if it doesn't exist explain why.

   a. The exponential distribution.

   b. The logistic distribution, with the cumulative distribution function given by $(1 + \exp(-x))^{-1}$. (Hint: You may have to use tabulated definite integral values to do this. The answer is $(i\pi\beta)/\sin(i\pi\beta)$ for $\beta \in (-1, 1)$.)

8. Let U be a random variable uniform on $(-\frac{1}{2}, \frac{1}{2})$.

   a. Show that the cumulant generating function for U is $\log(\sinh(\frac{1}{2}\beta)/(\frac{1}{2}\beta))$, where

   $$\sinh(x) = \tfrac{1}{2}(\exp(x) - \exp(-x)).$$

   b. Show that $\mathcal{K}'(\beta) = -1/\beta + \frac{1}{2} + (\exp(\beta) - 1)^{-1}$.

   c. Let B_k be the coefficients in the power series expansion

   $$\beta/(\exp(\beta) - 1) = \sum_{k=0}^{\infty} B_k \beta^k / k! . \tag{34}$$

These constants are known as the Bernoulli numbers (Haynsworth and Goldberg (1965), p. 804). Show that $(\exp(\beta) - 1)^{-1} - 1/\beta = \sum_{k=1}^{\infty} B_k \beta^{k-1}/k!$, and hence that $\mathcal{K}(\beta) = \frac{1}{2}\beta + \sum_{k=1}^{\infty} B_k \beta^k/(kk!)$. Show that the cumulant of order k is B_k/k.

9. Give the moments for the standard lognormal distribution (ie., $Y = \exp(X)$, X standard normal). Comment on convergence properties of the formal power series associated with the characteristic function. Comment on the possibility that there exists another distribution with these same moments.

10. Define the inverse Gaussian distribution to have a density proportional to

$$x^{-3/2} \exp(-\tfrac{1}{2}(\xi/x + \psi x))$$

for positive x. The parameters ξ and ψ are also both positive.

   a. Show that the normalizing constant is $\exp(\sqrt{\psi\xi})\sqrt{\xi/(2\pi)}$.

   b. Show that the cumulant generating function for this distribution is

$$\sqrt{\psi\,\xi} - \sqrt{(\psi - 2\beta)\,\xi}.$$

What is the functional inverse of this function? Justify the name inverse Gaussian.

3
Edgeworth Series

The accuracy of the normal approximation was assessed in §2. The normal approximation was motivated by treating the log of the characteristic function as approximately quadratic and inverting this quadratic approximation. Deviations from this normal approximation were measured by the third derivative of the characteristic function at 0. This chapter considers higher-order approximations to cumulative distribution functions and densities. Not surprisingly, these can also be expressed as approximate inversions of the characteristic function. Two heuristic motivations for the Edgeworth series are presented. Their correctness for densities and distribution functions is demonstrated using the methods of the previous chapter. Regularity conditions are investigated and discussed. Examples are given. The standards for assessment of accuracy of these methods are discussed and criticized. The Edgeworth series is inverted to yield the Cornish–Fisher expansion. Extensions of results from the standard application of means of independent and identically distributed continuous random variables to non-identically distributed and lattice cases are presented. Parallels in the lattice case with classical Sheppard's corrections are developed.

Heuristically, if $Z = \sum_{j=1}^{n} Y_j / \sqrt{n}$ is the standardized sum of n independent and identically distributed unit variance zero mean random variables $Y_1, \ldots, Y_n$, then its characteristic function is

$$\exp(-\beta^2/2 + \sum_{j=3}^{\infty} \kappa_j^n (i\beta)^j / j!), \tag{35}$$

where κ_j^n is the cumulant of Z of order j, with $\kappa_j^n = n^{(2-j)/2} \kappa_j$. The Central Limit Theorem in its most basic form, and the Edgeworth series defined below, rely on the heuristic notion of approximating the distribution of Z by truncating the infinite sum in the exponent above, approximating the quantity $\exp(\sum_{j=3}^{j} \kappa_j^n (i\beta)^j / j!)$ as a polynomial $p(\beta)$, and calculating that approximate cumulative distribution function whose Fourier transform is approximately $\exp(-\beta^2/2)p(\beta)$. The result is approximately a cumulative distribution function; it need not be monotonic, nor need it take values in $[0, 1]$. Whether this matters will be discussed below.

After presenting heuristic arguments indicating what the expansion should be, careful attention is paid to quantify both the error incurred in truncating (35) to $\exp(-\beta^2/2)p(\beta)$, and then to the way in which this translates to error in the Fourier inversion. This yields a rigorous proof that the order of error in the resulting Edgeworth series is what the heuristic arguments indicate.

Applications considered in this chapter derive primarily from the Edgeworth series to the cumulative distribution function; these are presented in §§3.12f and in the exercises.

3.1. Heuristic Development

Formally, the Edgeworth series approximation to a density f_X is constructed as a modification of a baseline density f_Y. The following formal construction is due to Davis (1976). Let X and Y be two random variables, such that Y has density f_Y, and X has density f_X. Suppose further that X and Y can be constructed on a common probability space, such that $Z = X - Y$ is independent of Y. Conditionally on $Z = z$, X has density $f_Y(x - z)$, and expanding f_Y as a power series about x, $f_Y(x - z) = \sum_{j=0}^{\infty} f_Y^{(j)}(x)(-z)^j/j!$. Hence the unconditional density of X is $f_X(x) = \sum_{j=0}^{\infty} f_Y^{(j)}(x)(-1)^j \mu_j^*/j!$, where μ_j^* are the moments of Z, again assuming that such a construction is possible. Writing

$$h_j(x) = (-1)^j f_Y^{(j)}(x)/f_Y(x), \tag{36}$$

we observe that

$$f(x) = f_Y(x) \sum_{j=0}^{\infty} h_j(x)\mu_j^*/j!. \tag{37}$$

The "moments" μ_j^* are the moments of whatever distribution is necessary to add to Y to get X. The cumulant of order j associated with these moments is the cumulant of order j associated with X minus the corresponding cumulant for Y, since $Y + Z = X$ and Z is independent of Y. The functions h_i are ratios of the derivatives of the baseline density to the density itself. In the case a normal baseline these are polynomials.

The problem with this construction is that in general the random variable Z described above cannot be constructed. For instance, if f_Y is the normal density with the same mean and variance as the density f_X to be approximated, the variance of the variable Z is necessarily zero; hence all cumulants of Z must be zero.

3.2. A Less Untrue but Still Heuristic Development

McCullagh (1987), Chapter 5, presents an alternative heuristic construction of the Edgeworth series. Suppose that f_Y and f_X have moment generating functions $\mathcal{M}_Y(\beta)$ and $\mathcal{M}_X(\beta)$. Express their ratio as a power series in β:

$$\mathcal{M}_X(\beta) = \mathcal{M}_Y(\beta) \sum_{j=0}^{\infty} \beta^j \mu_j^*/j!;$$

this equation defines the coefficients μ_j^*. The inverse Laplace transform will now be applied term-wise to this sum to obtain an approximation to f_X.

The equation defining the coefficients μ_j^* can be expressed as

$$\log\left(\sum_{j=0}^{\infty} \beta^j \mu_j^*/j!\right) = \mathcal{K}_X(\beta) - \mathcal{K}_Y(\beta), \tag{38}$$

where $\mathcal{K}_X$ and $\mathcal{K}_Y$ are cumulant generating functions associated with f_X and f_Y respectively. Heuristically, μ_j^* are the moments that would be associated with the distribution whose cumulants κ_j^* are the difference between the cumulants of f_X and f_Y; rigorously they are the coefficients of the formal power series obtained when the power series defined by $\exp(z)$ is composed with the power series whose coefficients are the cumulant differences. Since these cumulant differences need not be the cumulants of any distribution, following McCullagh (1987), Chapter 5 we call these pseudo-cumulants and we call the coefficients μ_j^* pseudo-moments. In general these pseudo-moments will not be the moments of any real random variable; μ_2^* will generally be zero, and in some applications it may be negative.

By the definition of $\mathcal{M}_Y(\beta)$ the first term in the inversion is $f_Y(x)$, and

$$\int_{-\infty}^{\infty} \exp(\beta x)\frac{df_Y}{dx}\,dx = \exp(x\beta)f_Y(x)|_{-\infty}^{\infty} - \beta\int_{-\infty}^{\infty}\exp(x\beta)f_Y(x)\,dx = -\beta\mathcal{M}_Y(\beta),$$

where $f_Y^{(j)}(x)$ represents the derivative of order j of $f_Y(x)$ with respect to x. This calculation is predicated on the assumption that $\lim_{x\to\pm\infty}\exp(x\beta)f_Y(x) = 0$ for all β in the domain of $\mathcal{M}_Y$. Weak assumptions, such as uniform continuity of f_Y, will insure this. Similarly

$$\int_{-\infty}^{\infty}\exp(\beta x)f_Y^{(j)}(x)\,dx = \beta^j(-1)^j\mathcal{M}_Y(\beta). \tag{39}$$

Applying the inverse Laplace transform, in Corollary 2.4.3,

$$f_Y^{(j)}(x) = \frac{1}{2\pi i}\int_{-i\infty}^{+i\infty}\exp(-x\beta)\beta^j(-1)^j\mathcal{M}_Y(\beta)\,d\beta. \tag{40}$$

Hence at least formally, $f_X(x)$ can be expressed as

$$\sum_{j=0}^{\infty} f_Y^{(j)}(x)(-1)^j\mu_j^*/j! = \sum_{j=0}^{\infty} h_j(x)\mu_j^*/j!. \tag{41}$$

Then (37), where $h_j(x)$ is given in (36), will be used as a formal expansion for $f_X(x)$.

The density approximation (37) gives rise to a cumulative distribution function approximation that is calculated very easily. Suppose F_Y is the cumulative distribution function associated with f_Y. As will be argued later, the coefficient of the leading term in (37), μ_0^*, is 0. Integrating (37) term-wise we obtain the formal series

approximation to $F(x)$,

$$F_Y(x) + \sum_{j=1}^{\infty} f_Y^{(j-1)}(x)(-1)^j \mu_j^*/j! = F_Y(x) - f_Y(x) \sum_{j=1}^{\infty} h_{j-1}(x)\mu_j^*/j!. \qquad (42)$$

Here the first term arises from integrating the the density $f_Y(x)$ by itself. Other terms arise from lowering the order of the derivatives involved by one. Since the order of derivatives changes but the power of (-1) does not, when expressing the integrated approximation in terms of the functions h_j to obtain (42) the sign changes, explaining the subtraction rather than addition of the terms of order one and higher.

Note that (41) and (42) were formed from the relationship (39), which holds only when $\mathcal{M}_0(\beta)$ exists for some $\beta \neq 0$. For the sake of completeness, the counterpart of (39) in terms of characteristic functions, which holds generally, is

$$\int_{-\infty}^{\infty} \exp(i\beta z) f_0^k(z) \; dz = (-i\beta)^k \zeta_0(\beta), \qquad (43)$$

where ζ_0 is the characteristic function of f_0, and f_0^k represents the derivative of order k of f_0.

3.3. Choice of Baseline

Typically the target distribution depends on some parameter like sample size; here we denote this by n. The resulting moments and cumulants for the target density will also depend on n; when necessary denote the dependence by a superscript n. Although (37) is expressed in terms of pseudo-moments, it is often easier to think of it with the pseudo-moments expressed in terms of the pseudo-cumulants, since the pseudo-cumulants are more directly calculated from the target and baseline distributions, and since the dependence of the pseudo-cumulants on n is simpler. As above, in the case of standardized sums of independent and identically distributed random variables, $\kappa_l^n = \kappa_l^1 n^{1-l/2}$. These series expansion methods are also often used in the case of sums of non-independent and identically distributed random variables if $\kappa_l^n = O(n^{1-l/2})$. Note especially that κ_1^n increases rather than decreases as $n \to \infty$. Often the terms in the resulting series can be divided into those that do not decrease as n gets large, and those which get small as n gets large. In the cases when $\kappa_l^n = O(n^{1-l/2})$, an approximating distribution which matches the first and second cumulants of the target distribution makes zero all terms involved in the pseudo-moments, and hence involved in (37), to vanish as $n \to \infty$. When viewing these methods as extensions of the Central Limit Theorem, it is natural to use the approximating normal density as the baseline. This works well as long as $\kappa_j^n = o(1)$ for $j \geq 3$. When the baseline distribution is Gaussian, the series (37), truncated after a chosen number of terms, is called the Gram–Charlier series. Expansions based on other than the normal distribution are rare. On occasions expansions based on the χ^2 distribution are used.

3.4. Calculation of Hermite Polynomials and Pseudo-moments

When the approximating normal baseline function is used, the functions h_j defined in (36) and used in (37) are

$$h_j(x) = (-1)^j \left[\frac{d^j}{dx^j} \exp(-x^2/2) \right] / \exp(-x^2/2).$$

These are called Hermite polynomials. The first few are tabulated in Table 2.

Table 2: *The First Six Standard Hermite Polynomials*

$h_1(x) = x$	$h_3(x) = x^3 - 3x$	$h_5(x) = x^5 - 10x^3 + 15x$
$h_2(x) = x^2 - 1$	$h_4(x) = x^4 - 6x^2 + 3$	$h_6(x) = x^6 - 15x^4 + 45x^2 - 15$

Note the similarity between these equations and the equations for moments in terms of cumulants from Table 1. If in Table 1, x is substituted for κ_1 and -1 is substituted for κ_2, and zero is substituted for all other cumulants, the quantities μ_j are the Hermite polynomials. Recall that the moments are coefficients in the formal power series $\exp(\sum_{j=1}^{\infty} \kappa_j \beta^j / j!)$; McCullagh (1987) exhibits the functions $h_j(x)$ as coefficients in the formal power series $\exp(\beta x - \beta^2/2)$. This relationship extends to the case of a normal baseline with non-zero mean or non-unit variance, and to the multivariate case as well.

The pseudo-moments are easily calculated from the underlying baseline cumulants λ_j and target cumulants κ_j. One begins by taking the difference $\kappa_j - \lambda_j$ between respective cumulants. The cumulant to moment conversion formulas in reverse yield expressions for the pseudo-moments μ_j^* as a sum of products of the third and higher order cumulants of the target distribution. That is, the pseudo-moments μ_j^* satisfy the series equation

$$\exp(\sum_{j=1}^{\infty}(\kappa_j - \lambda_j)\beta^j / j!) = \sum_{j=1}^{\infty} \mu_j^* \beta^j / j!. \tag{44}$$

These are the same as those μ_j^* given previously in (38), when the approximating distribution is Gaussian, with mean and variance matching the target distribution. The leading term in both cumulant generating function series expansions are zero; the pseudo-cumulant of order zero is then always zero and the leading term in the exponentiated series, μ_0^*, is always 1. When a normal baseline distribution is used the first two pseudo-cumulants are generally zero and the rest are the unmodified corresponding cumulants of the target distribution. If κ_j is of size $O(n^{(2-j)/2})$, the various terms in the pseudo-moments can be sorted according to their order in n. These are given in Table 3.

Table 3: *Pseudomoments as a Function of Cumulants*

Order in n

	$O(n^{-1/2})$	$O(n^{-1})$	$O(n^{-3/2})$	$O(n^{-2})$	$O(n^{-5/2})$	$O(n^{-3})$	$O(n^{-7/2})$
$\mu_3^* =$	κ_3						
$\mu_4^* =$		κ_4					
$\mu_5^* =$			κ_5				
$\mu_6^* =$		$10\kappa_3^2$		κ_6			
$\mu_7^* =$			$35\kappa_4\kappa_3$		κ_7		
$\mu_8^* =$				$35\kappa_4^2+$ $56\kappa_5\kappa_3$		κ_8	
$\mu_9^* =$			$280\kappa_3^3$		$84\kappa_6\kappa_3+$ $126\kappa_5\kappa_4$		κ_9

3.5. The Expansion Theorem

This section contains a statement of the Edgeworth series theorem for distribution functions and densities of standardized sums. For the present, we may restrict attention to the case in which $\mu = 0$ and $\sigma = 1$. Proofs will follow in the next three sections.

The formal Edgeworth expansion (37) is generally truncated to a finite number of terms, by using only the first j cumulants. When the cumulant of order j is of size $O(n^{(2-j)/2})$ for $j \leq \mathsf{j}$, then the resulting error is of size at least $O(n^{(1-\mathsf{j})/2})$. This error is at least as large of some of the contributions to the pseudomoments, as noted in §3.3. For example, if $\mathsf{j} = 4$, then pseudomoments through order 6 are required to capture all of the terms of size $O(1/n)$. However, μ_6^* includes the term κ_6^n, which is of size $O(1/n^2)$, when it exists. The Edgeworth series does not include the term with $h_6(x)\kappa_6^n$.

More generally, let $\mu_{j,\mathsf{j}}^*$ be those terms in μ_j^* that are not $o(n^{(2-\mathsf{j})/2})$. These are the sums across the rows in Table 3, of items in columns 1 through $\mathsf{j}-1$. Applying this truncation to (37) and (42), with the baseline distribution taken to be the standard Gaussian distribution, we obtain the Edgeworth density and cumulative distribution function approximations are respectively

$$e_\mathsf{j}(x;\boldsymbol{\kappa}^n) = \phi(x)\sum_{j=0}^{\mathsf{j}} h_j(x)\mu_{j,\mathsf{j}}^*/j!, \quad E_\mathsf{j}(x;\boldsymbol{\kappa}^n) = \Phi(x) - \phi(x)\sum_{j=1}^{\mathsf{j}} h_{j-1}(x)\mu_{j,\mathsf{j}}^*/j!. \quad (45)$$

Recall again that $\mu_{0,\mathsf{j}}^* = 1$, $\mu_{1,\mathsf{j}}^* = 0$, and $\mu_{2,\mathsf{j}}^* = 0$.

This notation $E_\mathsf{j}(z,\boldsymbol{\kappa}^n)$ and $e_\mathsf{j}(z,\boldsymbol{\kappa}^n)$ is chosen to reflect the fact that the Edgeworth approximation depends on n only through the cumulants of the target dis-

tribution.

After substituting in the expressions for pseudo-moments in terms of pseudo-cumulants, and collecting terms according to their power in n, we find

$$E_j(z, \boldsymbol{\kappa}^n) = \Phi(z) - \phi(z)[\kappa_3^n h_2(z)/6 + (\kappa_4^n h_3(z)/24 + 10\kappa_3^{n2} h_5(z)/720)$$
$$+ (\kappa_5^n h_4(z)/120 + 35\kappa_3^n \kappa_4^n h_6(z)/5040 + 280\kappa_3^{n3} h_8(z)/362880) + \cdots]$$

and

$$e_j(z, \boldsymbol{\kappa}^n) = \phi(z)[1 + \kappa_3^n h_3(z)/6 + (\kappa_4^n h_4(z)/24 + 10\kappa_3^{n2} h_6(z)/720)$$
$$+ (\kappa_5^n h_5(z)/120 + 35\kappa_3^n \kappa_4^n h_7(z)/5040 + 280\kappa_3^{n3} h_9(z)/362880) + \cdots].$$

The above definition of $E_j(z, \boldsymbol{\kappa}^n)$ is appropriate only when the first two cumulants of the summands are 0 and 1 respectively, forcing $\kappa_1^n = 0$ and $\kappa_2^n = 1$ for all n. For general cumulants $\boldsymbol{\kappa}$, calculate the invariants $\boldsymbol{\rho}$ as in §2.1, and define

$$E_j(z, \boldsymbol{\kappa}^n) = E_j((z - \kappa_1^n)/\sqrt{\kappa_2^n}, \boldsymbol{\rho}^n). \tag{46}$$

When Z is the sum of n independent and identically distributed summands, each with cumulants $\kappa_{Y,j}$, divided by $\sqrt{n}$, then $\kappa_1^n = \sqrt{n}\kappa_{Y,1}$ and $\kappa_2^n = \kappa_{Y,2}$. As we shall see later, however, (46) is applicable in cases when the cumulants of Z depend on n in a more complicated way.

Theorem 3.5.1: *Suppose Y_j are independent and identically distributed random variables, with mean 0 and variance 1, and that j is an integer greater than or equal to two, such that Y_j has a cumulant of order j. When $j > 2$ suppose further that*

$$\begin{cases} |\zeta(\beta)| < 1 \; \forall \beta \neq 0 & \text{if } j = 3, \text{ or} \\ \limsup_{|\beta| \to \infty} |\zeta(\beta)| < 1 & \text{if } j > 3. \end{cases} \tag{47}$$

Let $Z = \sum_{j=1}^n Y_j/\sqrt{n}$, and let F_Z be the cumulative distribution function for Z. Let $\boldsymbol{\kappa}^n$ be the set of cumulants of Z. When $E_j(z, \boldsymbol{\kappa}^n)$, as in (45), is used to approximate F_Z, the absolute error is uniformly of order $o(n^{1-j/2})$. If, furthermore, $\int_{-\infty}^{\infty} |\zeta(\beta)|^{\mathfrak{r}} \, d\beta < \infty$ for some $\mathfrak{r} \geq 1$, then the density f_n exists for Z for $n \geq \mathfrak{r}$. When $e_j(z, \boldsymbol{\kappa}^n)$, as in (45), is used to approximate $f_n(z)$, the absolute error is uniformly of order $o(n^{1-j/2})$.

3.6. Rigorous Construction for Cumulative Distribution Functions

Density approximations created by multiplying the normal density times sums of multiples of Hermite polynomials have Fourier transforms of the form the normal characteristic function times a polynomial. In §3.3 a choice for the definition of the Hermite polynomials and their multipliers was motivated by truncating the power series that resulted from dividing the characteristic function to be inverted by the approximating normal characteristic function after a fixed number of terms, and inverting the Fourier transform term-wise. In §3.4 these Hermite polynomials

and coefficients were calculated. The error incurred by this series truncation is measured by (25) and (24). Use of these inequalities requires bounding the difference between the true and approximating characteristic functions. In §2.5, Corollary 2.5.4 bounded the portion of the integrals in (25) and (24) near the origin.

The following exposition is based on Chapter XVI of Feller (1971). It makes rigorous the arguments of §3.2, with the characteristic function substituted in place of the moment generating function. Define the approximation $E_j^*(x, \boldsymbol{\kappa}^n)$ to be that function having the Fourier transform ξ_n of Corollary 2.5.4. Then $E_j^*(x, \boldsymbol{\kappa}^n) = E_j(x, \boldsymbol{\kappa}^n) + o(n^{1-j/2})$. Since the cumulative distribution functions can be recovered from their characteristic functions using (17), using the smoothing lemma the difference between F_Z and $E_j^*(x, \boldsymbol{\kappa}^n)$ can be bounded by the principal value of the integral

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{2}{|\beta|} |\zeta_n(\beta) - \xi_n(\beta)| \, d\beta, \tag{48}$$

defined to be

$$\lim_{\Theta \to \infty} \frac{1}{2\pi} \int_{-\Theta}^{\Theta} \frac{2}{|\beta|} |\zeta_n(\beta) - \xi_n(\beta)| \, d\beta.$$

Because the integral (48) is derived from (17), which need not converge absolutely, this argument will proceed indirectly.

Three ideas go into this proof, the Smoothing Theorem, the Series Theorem, and a fact about what kind of functions have Fourier transforms looking like polynomials times the normal characteristic function. The Smoothing Theorem is used to convert the possibly non-absolutely convergent integral (48) to an integral over a finite range, to which can be applied a wider range of integral approximation techniques. The Series Theorem is a really key idea here, since it allows truncation from the infinite series to a finite sum.

The approximation $E_j(x, \boldsymbol{\kappa}^n)$ defined earlier is then formed by discarding terms of size $o(n^{1-j/2})$. For example, if $j = 4$, then $E_j^*(x, \boldsymbol{\kappa}^n)$ contains a term involving κ_4^{n2}, and this term is omitted from $E_j(x, \boldsymbol{\kappa}^n)$. I need to prove that for any $\epsilon > 0$, there exists $N(\epsilon) > 0$ such that $|E_j(x, \boldsymbol{\kappa}^n) - F_Z(x)| \le \epsilon n^{1-j/2}$ whenever $n > N(\epsilon)$.

Apply the Smoothing Theorem with the target cumulative distribution function F_Z in place of H, and the cumulative distribution function approximation represented by $E_j^*(x, \boldsymbol{\kappa}^n)$ in place of G. The integral (48) heuristically representing error has three ranges. The areas far out in the tails, the area very close to the origin, and the area in between will all be handled separately. Apply Corollary 2.5.4, with $\mathfrak{g} = j$, to determine δ such that the integrand is bounded by

$$\exp(-\beta^2/4) \left[\frac{\epsilon |\beta|^j}{n^{j/2-1}} + \frac{C^{j-1} |\beta|^{3(j-1)}}{(j-1)! n^{j/2-1/2}} \right] \frac{1}{|\beta|} \tag{49}$$

for β such that $|\beta| < \delta\sqrt{n}$.

a. Far in the tails ($|\beta| > \Theta$) the contribution of the error in approximating the inversion integral is accounted for by the remainder term in the Smoothing

Theorem, Theorem 2.5.2. A bound on the error in inversion begins by choosing Θ large enough to make this remainder term small enough. Take Θ in the Smoothing Theorem to be $n^{j/2-1}/(\pi\epsilon)$. Then the final term in (24), representing the contribution in the tails of (48), is

$$24 \sup_{x \in \mathfrak{R}} |e_j(x, \boldsymbol{\kappa}^n)| \, \epsilon n^{1-j/2}. \tag{50}$$

b. Near the origin ($|\beta| < \delta\sqrt{n}$), the contribution to the integral of this region is bounded by the integral of (49) over $\mathfrak{R}$. This integral is

$$n^{1-j/2}[\epsilon 2^{j/2}\mu_{j-1} + 2^{3j/2-1}\mu_{3j-2}C^{j-1}/(\sqrt{n}(j-1)!)], \tag{51}$$

for μ_j the absolute moment of order j for the standard normal distribution. ϵ was arbitrary, shows that this contribution is of order $o(n^{1-j/2})$.

c. For intermediate values of β ($\beta \in \pm[\delta\sqrt{n}, \Theta]$), regularity conditions (47) on the characteristic function are used to bound the integrand. In the case where $j = 3$ the weaker condition suffices, since Θ is proportional to $\sqrt{n}$ and hence the set of $\beta/\sqrt{n}$ such that $|\beta| \in [\delta\sqrt{n}, \Theta]$ is fixed as n varies. The contribution of this part of the range of integration is bounded by the bound on the integrand, times the length of the range of integration, which is bounded. When $j > 3$ then Θ increases faster than $\sqrt{n}$ and the stronger condition in (47) is necessary. In any event, the contribution to the integral (48) in this region is bounded by $2q^nC^\dagger\Theta$ and approaches 0 geometrically, at a rate that does not depend on ϵ.

Q.E.D

3.7. Rigorous Construction for Density Functions

The following proof is almost identical to the last. Since the cumulative distribution functions can be recovered from their characteristic functions using (19), the difference between $f_Z(z)$ and $e_j(z, \boldsymbol{\kappa}^n)$ can be bounded by the integral

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} 2 \left|\zeta_n(\beta) - \xi_n(\beta)\right| \, d\beta, \tag{52}$$

where ξ_n is the characteristic function associated with the approximate distribution function $e_j(z, \boldsymbol{\kappa}^n)$. Because the integral (52) converges absolutely, this argument will not need the Smoothing Theorem. The Series Theorem is still a key idea here.

The integral (52) representing error is the same as (48) of the previous proof, except that the factor of $|\beta|^{-1}$ is gone. The range of integration will now be split in two parts. The area very close to the origin, and the area farther out will be handled separately.

a. Very near the origin ($|\beta| < \delta\sqrt{n}$ for δ determined as before), exactly the same arguments, with the same value of δ, imply that the integrand is bounded by

$$\exp(-\beta^2/4) \left[n^{1-j/2}\epsilon |\beta|^j + n^{1/2-j/2}C^{j-1} |\beta|^{3(j-1)}/(j-1)!\right],$$

for β such that $|\beta| < \delta\sqrt{n}$. This is exactly the same bound as (49), except that the factor of $|\beta|^{-1}$ is gone, since it was missing in (52). When integrated over $(-\delta\sqrt{n}, \delta\sqrt{n})$ the result is still of order $o(n^{1-j/2})$.

b. For more extreme values of β ($|\beta| > \delta\sqrt{n}$), the integrability condition on the cumulant generating function is used to bound the integrand. As will be shown in the next section, this condition implies the stronger condition in (47). There exists $q < 1$ such that if $|\beta| > \delta\sqrt{n}$ then $|\zeta(\beta/\sqrt{n})| < q$. Hence the contribution of (52) from this part of the range of integration can be bounded by

$$q^{n-\mathfrak{r}} \int_{-\infty}^{\infty} |\zeta(\beta)|^{\mathfrak{r}} \ d\beta + \int_{|\beta|>\delta\sqrt{n}} |\xi_n(\beta)| \ d\beta,$$

and approaches 0 geometrically.

3.8. Regularity Conditions

In these proofs an integral involving the characteristic function was bounded by bounding the integrand near the origin. The regularity conditions were used to ensure that contributions to the integral in the range of integration far from the origin are negligible. The regularity conditions needed for various variants of the approximation theorem are:

1. Distribution functions when $j = 2$: This results in the Central Limit Theorem. No regularity conditions, apart from the existence of the mean and variance, are needed. In the cumulative distribution function proof, T does not depend on n when $j = 2$, and hence one can pick n sufficiently large so that parts (a) and (b) of the range of integration for the error integral cover $\mathfrak{R}$; hence part (c) need not be considered.

2. Distribution functions when $j = 3$: Here we require $|\zeta(\beta)| < 1 \ \forall \beta \neq 0$. This suffices for the cumulative distribution function Edgeworth proof, since an upper bound on $|\zeta(\beta)|$ over the whole line is unnecessary. One need bound it only over (δ, T). This requirement is equivalent to requiring the summands have a non-lattice distribution.

3. Distribution functions when $j > 3$: Here we require the stronger condition

$$\limsup_{|\beta|\to\infty} |\zeta(\beta)| < 1,$$

since T increases faster than $\sqrt{n}$. This condition is known as Cramér's condition, and rules out perverse examples like cumulative distribution functions that do all of their jumping on a Cantor set, as well as examples where the summands have probability masses on $0, 1, a$ where a is irrational. Examples similar to the latter arise, for instance, in likelihood ratio testing and will be addressed in a later chapter. Both of these examples are of non-lattice singular distributions.

4. Density functions for all orders: We require that $\int_{-\infty}^{\infty} |\zeta(\beta)^{\mathfrak{r}}| \ d\beta < \infty$ for some $\mathfrak{r} \geq 1$. This condition is needed for density proofs, since the Smoothing Theorem is not used to truncate the range of integration for the error integral.

This condition implies that the distribution has a density at every point on $\mathfrak{R}$. Lemma 2.4.4 shows that the finiteness of the integral of the absolute value of the characteristic function over $\mathfrak{R}$ implies the existence of a density, since the characteristic function of a sum of independent and identically distributed random variables is the characteristic function of the summands raised to the power equal to the number of summands. This condition is both necessary and sufficient for a bounded density to exist for some convolution of the random variable with itself.

Condition 4 implies condition 3, since condition 4 implies the existence of a density for Z when $n = \mathfrak{r}$, and the existence of this density implies condition 3 by the Riemann–Lebesgue Theorem:

Theorem 3.8.1: *If g is a real function such that $\int_{-\infty}^{\infty} |g(x)| \ dx < \infty$, and if its Fourier transform is given by $\zeta(\beta) = \int_{-\infty}^{\infty} \exp(i\beta x) g(x) \ dx$, then $\lim_{|\beta| \to \infty} \zeta(\beta) = 0$.*

Proof: Choose $\epsilon > 0$. Since $\int_{-\infty}^{\infty} |g(x)| \ dx < \infty$, there exists a step function $g^*(x) = \sum_{j=1}^{J} g_j I_{(a_j, a_{j+1}]}(x)$, where $I_{(a_j, a_{j+1}]}(x)$ is the function taking the value 1 on $(a_j, a_{j+1}]$ and zero otherwise, such that $\int_{-\infty}^{\infty} |g(x) - g^*(x)| \ dx < \epsilon$. Let $\zeta^*(\beta) = \int_{-\infty}^{\infty} \exp(i\beta x) g^*(x) \ dx$. Then

$$|\zeta^*(\beta) - \zeta(\beta)| \leq \int_{-\infty}^{\infty} |g(x) - g^*(x)| \ dx < \epsilon.$$

Since

$$\int_{a_j}^{a_{j+1}} \exp(i\beta x) \ dx = (\exp(i\beta a_{j+1}) - \exp(i\beta a_j))/(i\beta),$$

and $|\exp(i\beta a_{j+1}) - \exp(i\beta a_j)| \leq a_{j+1} - a_j$, then

$$|\zeta^*(\beta)| \leq \sum_j g_j (a_{j+1} - a_j)/|\beta| = \int_{-\infty}^{\infty} g^*(x) \ dx/|\beta|,$$

and hence

$$\zeta(\beta) < \zeta^*(\beta) + \epsilon \leq \int_{-\infty}^{\infty} g^*(x) \ dx/|\beta| + \epsilon < \int_{-\infty}^{\infty} g(x) \ dx/|\beta| + 2\epsilon$$

for all $\epsilon > 0$, and $\zeta(\beta) \leq \int_{-\infty}^{\infty} g(x) \ dx/|\beta|$.

Q.E.D

3.9. Some Examples

The gamma and beta distributions illustrate the behavior of Edgeworth series in approximating distributions that differ from normal in certain fundamental ways. These will be displayed in Fig. 3.

A $\Gamma(p, \lambda)$ random variable is one with the density

$$f(x) = \lambda^p x^{p-1} \exp(-\lambda x)/\Gamma(p).$$

Its characteristic function is $\zeta(\beta) = 1/(1 - i\beta/\lambda)^p$, and its cumulant generating function is $\mathcal{K}(\beta) = p\log(\lambda) - p\log(\lambda - \beta)$. Hence the cumulant of order j is $\kappa_j = p\Gamma(j)/\lambda^j$, and the convolution of n $\Gamma(p, \lambda)$ densities is a $\Gamma(np, \lambda)$ density. The standardized mean has a $\Gamma(np, \lambda/\sqrt{n})$ distribution. Note that $\int_{-\infty}^{\infty} |\zeta(\beta)|^\nu \, d\beta < \infty$ if and only if $\nu > 1/p$. This example illustrates the effect of the condition on the integrability of the characteristic function. Part of the claim of the density approximation theorem then is that the density of the standardized sum of n independent replicates of such a random variable exists and is continuous for all $n > 1/p$. In this example, one can explicitly calculate the density of such sums. Those sums with $n < 1/p$ have a density that approaches infinity as $x \to 0$, and sums with $n = 1/p$ have a density discontinuous at zero. Thus, the integrability condition successfully indicates all sums with a continuous density.

Fig. 3a shows the poor performance of the Edgeworth series using two, three, and four cumulants when $p = 1$, yielding the exponential distribution. In this case the target density is discontinuous at zero. The highest order approximation, using four cumulants, behaves poorly in the upper tail. Since it involves a higher order polynomial than the other approximations, oscillations in the tail are more severe. Near the ordinate 3 this approximation is negative. Fig. 3b shows that the performance of the Edgeworth series involving the third and fourth cumulants is much improved for standardized sums of 10 such variables. Distinguishing the $E_4(x, \kappa^{*10})$ approximation from the true density visually is difficult for most of the range plotted. Figs. 3c and 3d show improved behavior when $p = 2$.

A $B(p, q)$ random variable is one with the density

$$f(x) = x^{p-1}(1 - x)^{q-1}/B(p, q)$$

for $x \in (0, 1)$, where $B(p, q)$ is the beta function $B(p, q) = \Gamma(p)\Gamma(q)/\Gamma(p + q)$. Its moment of order l is $\Gamma(p + l)\Gamma(p + q)/(\Gamma(p)\Gamma(p + q + l))$. From these moments the cumulants can be calculated. Fig. 4 illustrates the behavior of the Edgeworth series when $p = q = 4$.

If $p = q$ the distribution is symmetric about $\frac{1}{2}$. Hence the odd cumulants are zero, and the normal approximation $E_2(x, \kappa^{*n})$ and the next order Edgeworth approximation $E_3(x, \kappa^{*n})$ coincide exactly. The Edgeworth approximation $E_4(x, \kappa^{*n})$ involving kurtosis represents a distinct improvement.

Additional examples occur in later sections on Edgeworth approximation for non-identically distributed and lattice random variables, and in the exercises of this chapter.

3.10. Relative versus Absolute Error

While absolute error is bounded by our convergence theorem, the relative error, defined as the ratio of the approximation error to what is being approximated, can behave very badly. The first term omitted can be expressed as the normal density approximation times a polynomial, and hence the relative error can usually be
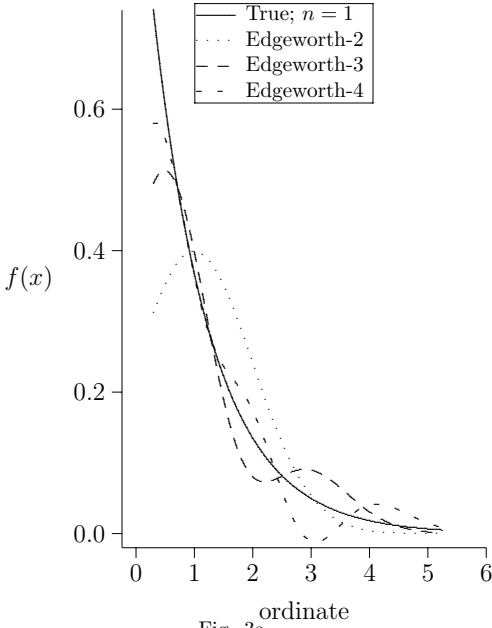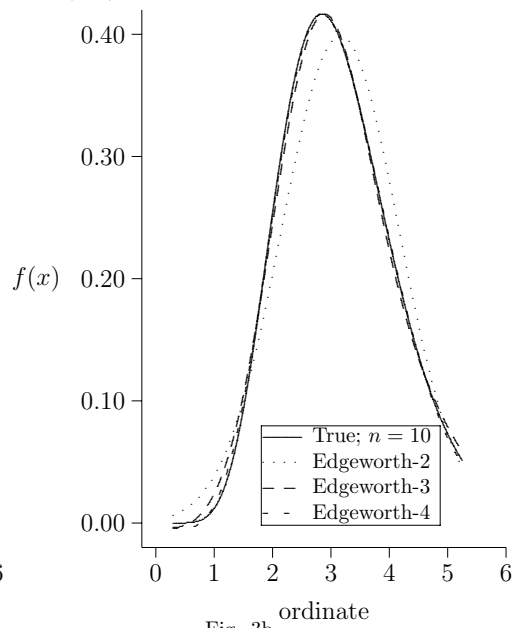
Calculations for Gamma(1,1)

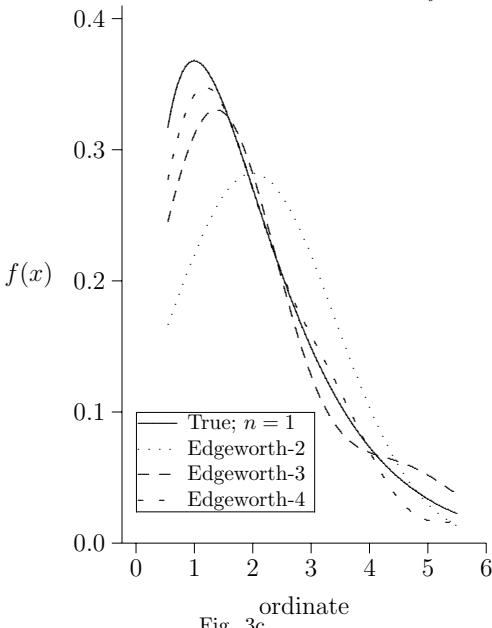Fig. 3a.

Fig. 3b.

Calculations for Gamma(1,2)

Fig. 3c.

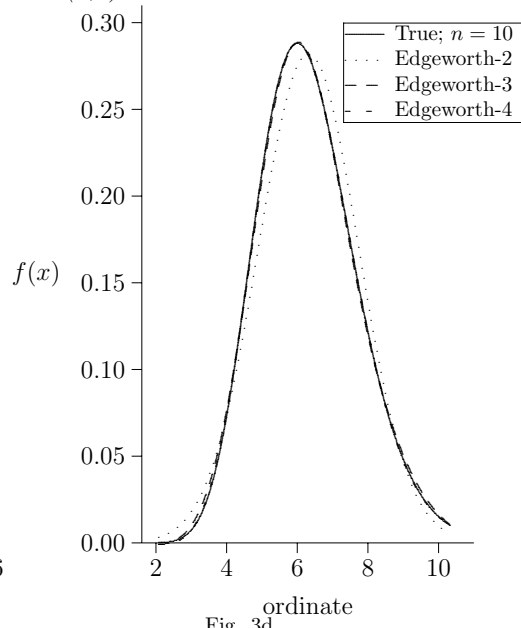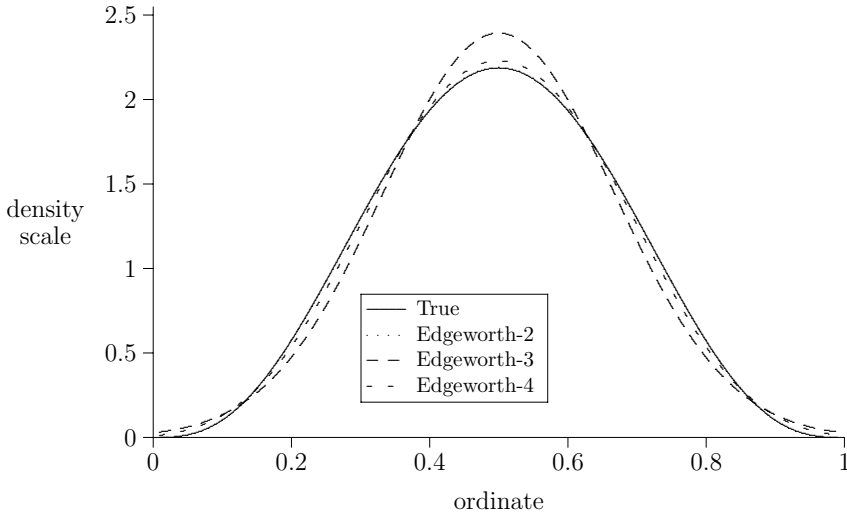Fig. 3d.

Calculations for Beta(4,4);n=1

Fig. 4

expected to be approximately polynomial. In those cases where the approximation becomes negative the relative error exceeds 100%.

3.11. The Non-Identically Distributed Case

Chambers (1967) provides an algorithm for constructing Edgeworth series for random vectors. He gives conditions for Edgeworth approximation to the density to hold other than in the independent and identically distributed standardized mean case using the two conditions:

1. The cumulants are of the same order in n as they are for standardized means;
2. A condition on the size of the integral of the characteristic function in the tails of the distribution.

When leaving the case of independent and identically distributed summands the construction of the Edgeworth series from the cumulants of the resulting variable, rather than from the cumulants of the original summands, becomes important, since these are either no longer the same across summands, or they cumulate in a non-standard manner. Regularity conditions are simpler in the case of independent summands, and are considered below.

As an example, consider the distribution of a least-squares estimate T for the parameter θ in a simple linear regression model $Y_j = \gamma + \theta z_j + E_j$, where the errors E_j are independent and identically distributed with mean zero and j cumulants λ_l. Then $T = \sum_{j=1}^{n}(z_j - \bar{z})Y_j / \sum_{j=1}^{n}(z_j - \bar{z})^2$, and the cumulants κ_l^n are given by

$\sum_{j=1}^{n}(z_j - \bar{z})^l \kappa_l / (\sum_{j=1}^{n}(z_j - \bar{z})^2)^l$. In applications where the ordinates z_j do not vary much the estimator should have a distribution as well approximated by an Edgeworth series as is the distribution of means of the errors. A theorem of this section shows that this is true.

When the summands comprising the sum whose distribution is to be approximated by an Edgeworth series do not have equal variances, the reduction to the case of summands with unit variance cannot be made without loss of generality. We must then rescale the sum by dividing by its standard deviation, as was done in (46). This requires that the cumulants be replaced by their invariants, before subtracting off the expectation and variance of the approximating normal distribution (viz., 0 and 1).

First, a lemma adapted from Bhattacharya and Rao (1976):

Lemma 3.11.1: *Suppose Y_j are independent random variables, with characteristic functions $\zeta_j(\beta)$, means zero and variances σ_j^2, and j is an integer greater than or equal to two, such that for all j, Y_j has a cumulant of order j. Let $\varsigma_n^2 = n^{-1}\sum_{j=1}^{n}\sigma_j^2$, and $\bar{\omega}_{n,\mathsf{j}} = n^{-1}\sum_{j=1}^{n}\mathrm{E}\left[\left|Y_j^{\mathsf{j}}\right|\right]\varsigma_n^{-\mathsf{j}}$.   Let ζ be the characteristic function of $\sum_{j=1}^{n}Y_j$. Then there exist K and $\delta > 0$ not depending on the distribution of the summands such that if $|\beta| \leq \delta n^{1/2 - 1/\mathsf{j}}\bar{\omega}_{n,\mathsf{j}}$ then*

$$\left| \log(\zeta(\beta/(\sqrt{n}\varsigma_n))) + \beta^2/2 - \sum_{j=1}^{n} v_j(\beta/(\sqrt{n}\varsigma_n)) \right| < Kn^{1-\mathsf{j}/2}\bar{\omega}_{n,\mathsf{j}} |\beta|^{\mathsf{j}}. \qquad (53)$$

Proof: Let $v_j(\beta) = \sum_{l=3}^{\mathsf{j}}\kappa_{j,l}(i\beta)^l/l!$, and let $C_j = \mathrm{E}\left[\left|Y_j^{\mathsf{j}}\right|\right]^{1/\mathsf{j}}$. By Jensen's Inequality $\sigma_j \leq C_j$. The argument proceeds by applying Lemma 2.5.3. Note that

$$\left| \sum_{j=1}^{n} \log(\zeta_j(\beta/(\sqrt{n}\varsigma_n))) + \sigma_j^2\beta^2/[2(n\varsigma_n^2)] - v_j(\sigma_j\beta/(\sqrt{n}\varsigma_n)) \right| = O(|\beta|^{\mathsf{j}})$$

for $\beta < \varsigma_n\delta$, for some $\delta > 0$. Let $\beta^* = \beta/(\sqrt{n}\varsigma_n)$. For every non-zero complex a and b, $|\log(a) - \log(b)| \leq |a - b| \max(|1/a|, |1/b|)$. Letting $a = \zeta_j(\beta^*)$ and $b = \exp(-\sigma_j^2\beta^{*2}/2 + v(\beta^*))$, one needs to establish a neighborhood of zero in which a and b are bounded away from zero, and one needs to provide a bound on $|a - b|$. By (14), with $\mathsf{j} = 1$, $|\zeta_j(\beta^*) - 1| < 2|\beta^*|^2 \sigma_j^2 \leq 2|\beta^*|^2 C_j^2$. Hence for $\beta^* < \frac{1}{2}$, $|a| > \frac{1}{2}$. Also, $|b - 1| \leq \exp(p_3(|\beta^*|)) - 1$. Hence there exists $\delta < \frac{1}{2}$ such that $|\beta^*| \leq \delta/C_j$ implies $|b - 1| \leq \frac{1}{2}$ and $a > \frac{1}{2}$. The factor $|a - b|$ may be bounded by

$$\left| \zeta_j(\beta^*) - \sum_{l=2}^{\mathsf{j}}\mu_{j,l}(i\beta^*)^l/l! \right| + \left| \sum_{l=3}^{\mathsf{j}}\mu_{j,l}(i\beta^*)^l/l! - \exp\left(\sum_{l=3}^{\mathsf{j}}\kappa_{j,l}(i\beta^*)^l/l! \right) \right|.$$

The first term above can be bounded by (14), to yield a bound of $2|\beta^*|^{\mathsf{j}} C_j^{\mathsf{j}}/\mathsf{j}!$. The second term has a bound of the form $|\beta^*|^{\mathsf{j}} C_j^{\mathsf{j}} \exp(p_1(C_j|\beta^*|))p_2(|C_j\beta^*|)$, with p_1 and

p_2 polynomials not depending on the distributions of summands. Hence there exists K depending only on p_1, p_2, and p_3, such that if $|\beta| \leq \delta\sqrt{n}n^{-1/\mathsf{j}}/[n^{-1}\sum_j(C_j^{\mathsf{j}}/(\varsigma_n^{\mathsf{j}}))]^{1/\mathsf{j}}$ then

$$\left| \log(\zeta(\beta/(\sqrt{n}\varsigma_n))) + \beta^2/2 - \sum_{j=1}^{n} \upsilon_j(\beta/(\sqrt{n}\varsigma_n)) \right| < Kn^{-\mathsf{j}/2}\sum_{j=1}^{n}[C_j^{\mathsf{j}}/\varsigma_n^{\mathsf{j}}]\,|\beta|^{\mathsf{j}}.$$

Q.E.D

The quantities The following theorem demonstrating the validity of the Edgeworth series depends on the bounds on the average standardized moments $\bar{\omega}_{n,\mathsf{j}}$. This condition guarantees that invariant j of the desired distribution is of size no larger than $O(n^{(2-j)/2})$, for $j \leq \mathsf{j}$. Under this condition, the Edgeworth approximation is valid to the order expected.

Theorem 3.11.2: *Suppose Y_j are independent random variables, with characteristic functions $\zeta_j(\beta)$, means zero and variances σ_j^2, and j is an integer greater than or equal to two, such that for all j, Y_j has a cumulant of order j. Assume that $\bar{\omega}_{n,\mathsf{j}}$ remains bounded as n increases, and that ς_n^2 as defined in Lemma 3.11.1 is bounded away from 0. Further assume that*

$$\prod_{j=1}^{n} \zeta_j(\beta) = o(n^{-a}) \ \textit{uniformly for } |\beta| > \delta \ \forall\delta, a > 0. \tag{54}$$

Then the approximation $E_{\mathsf{j}}(z, \boldsymbol{\rho}^n)$ formed as in (46) by calculating the associated invariants $\boldsymbol{\rho}_j^n$, and omitting terms of size $O(n^{1-\mathsf{j}/2})$, satisfies $F_Z(z) = E_{\mathsf{j}}(z, \boldsymbol{\kappa}^n) + O(n^{1-\mathsf{j}/2})$, where F_Z is the cumulative distribution function of $Z = \sum_{j=1}^n Y_j/\sqrt{n\varsigma_n^2}$.

Proof: This proof will parallel the development of §3.6. Bound (53) replaces (26), and (54) replaces Cramér's condition. The minimum bound on ς_n^2 guarantees that the region near the origin in which the true characteristic function is well–approximated by the Fourier transform of the Edgeworth series expands at the correct rate. The rest of the theorem proceeds as before.

Q.E.D

Returning to the example of regression estimators, T has an Edgeworth approximation valid to $O(n^{1-\mathsf{j}/2})$ if

$$\sup_n n^{-1}\sum_{j=1}^{n}(z_j - \bar{z})^{\mathsf{j}}\kappa_{\mathsf{j}}/(n^{-1}\sum_{j=1}^{n}(z_j - \bar{z})^2\kappa_2)^{\mathsf{j}/2} < \infty, \ \inf_n n^{-1}\sum_{j=1}^{n}(z_j - \bar{z})^2\kappa_2 > 0.$$

These conditions are trivially true if z_j takes on a finite number of values in a fixed proportion. They are also true with probability one if the regression ordinates are drawn from a population with a finite moment of order j and the regression is considered conditional on the observed collection of ordinates, by the strong law of large numbers.

3.12. Cornish–Fisher Expansions

Related to the problem of approximating tail areas is the inverse problem of approximating quantiles of a distribution. Let Z be the standardized sum of n independent and identically distributed continuous random variables with mean zero and variance one, and a finite fourth moment. Let F_Z be the cumulative distribution function of Z. For each $\alpha \in (0, 1)$, there exists at least one solution to $F_Z(y) = \alpha$. Denote this solution by z_α^n. This section provides a series expansion for z_α^n. The result is called the Cornish–Fisher expansion (Cornish and Fisher, 1937, Barndorff–Nielsen and Cox, 1989).

Let $z_\alpha = \Phi^{-1}(\alpha)$. Note that the sequence z_α^n converges to z_α, since otherwise for any $\epsilon > 0$ the distribution F_Z either assigns probability less than $\Phi^{-1}(\alpha)$ to the interval $(-\infty, z_\alpha + \epsilon]$, or assigns probability less than $1 - \Phi^{-1}(\alpha)$ to the interval $[z_\alpha - \epsilon, \infty)$, for infinitely many n. This contradicts the central limit theorem.

Applying the Edgeworth cumulative distribution function approximation using the first four cumulants,

$$F_Z(z) = \Phi(z) - \phi(z)\{h_2(z)\kappa_3^n/6 + h_3(z)\kappa_4^n/24 + \kappa_3^{n2}h_5(z)/72\} + C_n(z)/n$$
$$= \Phi(z) - \phi(z)\left\{\frac{h_2(z)}{6}\frac{\kappa_3}{\sqrt{n}} + \frac{h_3(z)}{24}\frac{\kappa_4}{n} + \frac{h_5(z)}{36}\frac{\kappa_3^2}{2n}\right\} + C_n(z)/n.$$

Here $C_{z,n}$ converges to zero uniformly in z, for z in a bounded region. We solve $F_Z(z_\alpha^n) = \alpha$. Expanding in a power series about z_α,

$$F_Z(z_\alpha^n) - \alpha = \phi(z_\alpha)\left\{ -\frac{h_2(z_\alpha)}{6}\frac{\kappa_3}{\sqrt{n}} - \frac{h_3(z_\alpha)}{24}\frac{\kappa_4}{n} - \frac{h_5(z_\alpha)}{36}\frac{\kappa_3^2}{2n} + \right.$$
$$\left(1 + \frac{h_3(z_\alpha)}{6}\frac{\kappa_3}{\sqrt{n}} + \frac{h_4(z_\alpha)}{24}\frac{\kappa_4}{n} + \frac{h_6(z_\alpha)}{36}\frac{\kappa_3^2}{2n}\right)(z_\alpha^n - z_\alpha) -$$
$$\left.\left(h_1(z_\alpha) + \frac{h_4(z_\alpha)}{6}\frac{\kappa_3}{\sqrt{n}} + \frac{h_5(z_\alpha)}{24}\frac{\kappa_4}{n} + \frac{h_7(z_\alpha)}{36}\frac{\kappa_3^2}{2n}\right)\frac{(z_\alpha^n - z_\alpha)^2}{2}\right\} + O((z_\alpha^n - z_\alpha)^3) + \frac{C_n(z_\alpha^n)}{n}.$$

Setting $F_Z(z_\alpha^n) - \alpha$ equal to 0, and neglecting terms of size $o(n^{-1/2})$, $z_\alpha^n - z_\alpha = (\kappa_3/[6\sqrt{n}])h_2(z_\alpha) + o(1/\sqrt{n}) = O(1/\sqrt{n})$. Setting $F_Z(z_\alpha^n) - \alpha$ to zero and dropping terms of size $o(1/n)$:

$$o\left(\frac{1}{n}\right) = \frac{h_2(z_\alpha)\kappa_3}{6\sqrt{n}} + \frac{h_3(z_\alpha)}{24}\frac{\kappa_4}{n} + \frac{h_5(z_\alpha)}{36}\frac{\kappa_3^2}{2n} - \left(1 + \frac{h_3(z_\alpha)\kappa_3}{6\sqrt{n}}\right)(z_\alpha^n - z_\alpha)$$
$$+ \frac{h_1(z_\alpha)(z_\alpha^n - z_\alpha)^2}{2}.$$

Solving for z_α^n one finds that

$$z_\alpha^n = z_\alpha + \frac{\kappa_3}{\sqrt{n}} \frac{h_2(z_\alpha)}{6} + \frac{\kappa_4}{n} \frac{h_3(z_\alpha)}{24}$$

$$+ \frac{\kappa_3^2}{2n} \left(\frac{h_5(z_\alpha) - 2h_3(z_\alpha)h_2(z_\alpha) + h_1(z_\alpha)h_2^2(z_\alpha)}{36} \right) + o\left(\frac{1}{n}\right)$$

$$= z_\alpha + \frac{\kappa_3}{\sqrt{n}} \frac{h_2(z_\alpha)}{6} + \frac{\kappa_4}{n} \frac{h_3(z_\alpha)}{24} - \frac{\kappa_3^2}{2n} \frac{(2z_\alpha^3 - 5z_\alpha)}{18} + o\left(\frac{1}{n}\right).$$

For standardized sums of random variables with arbitrary first and second cumulants,

$$z_\alpha^n = \sqrt{n}\kappa_1 + \sqrt{\kappa_2} \left[z_\alpha + \frac{\rho_{Z,3}}{\sqrt{n}} \frac{h_2(z_\alpha)}{6} + \frac{\rho_{Z,4}}{n} \frac{h_3(z_\alpha)}{24} - \frac{\rho_{Z,3}^2}{2n} \frac{(2z_\alpha^3 - 5z_\alpha)}{18} \right] + o(1/n),$$

where the coefficients $\rho_{Z,j}$ are the invariant cumulants of §2.1. If the summands have a fifth cumulant, $o(1/n)$ may be replaced by $O(n^{-3/2})$, since an additional term in the series might be constructed. Let C_2, C_3, and C_4 represent z_α^n with terms of size $o(1)$, $o(1/\sqrt{n})$, and $o(1/n)$ dropped, respectively.

Fisher and Cornish (1960) provide tables aiding the construction of higher order approximations in the same spirit. Hall (1983) provides an extension to these methods to smooth transformations of random variables with known cumulants; these generalized expansions rely only on the cumulants of the underlying distribution and derivatives of the transformation; limiting properties are demonstrated with full rigor.

Koning and Does (1988) consider the distribution of Spearman's Rank Correlation, used for nonparametric bivariate inference. Spearman's Rank Correlation is an affine transformation of $T = \sum_{j=1}^n jR_j/(n+1)$ where R_j are ranks of n independently and identically distributed random variables. This distribution is difficult to calculate exactly, but moments, and hence cumulants and the Cornish–Fisher expansion, are readily available. Koning and Does (1988) give Fortran code implementing this expansion for T. This code is available from STATLIB. In this example the summands are identically distributed but not independent. Also, T has a lattice distribution, while the Cornish–Fisher expansion is continuous. None the less, the Cornish–Fisher expansion nominally to order $o(1/n)$ performed very well. An example of its application is given in Fig. 5.

As a second example, Barndorff–Nielsen and Cox (1989) approximate quantiles for the χ_n^2 distribution; the cumulants of the square of a standard normal random variable are $\kappa_l = 2^{l-1}(l-1)!$ and the standardized cumulants are $\rho_{Z,l} = 2^{l/2-1}(l-1)!$; $\rho_{Z,3} = 2\sqrt{2}$ and $\rho_{Z,4} = 12$. The χ_n^2 distribution is the convolution of n such distributions. True tail probabilities and the Cornish–Fisher expansions are given in Fig. 6.
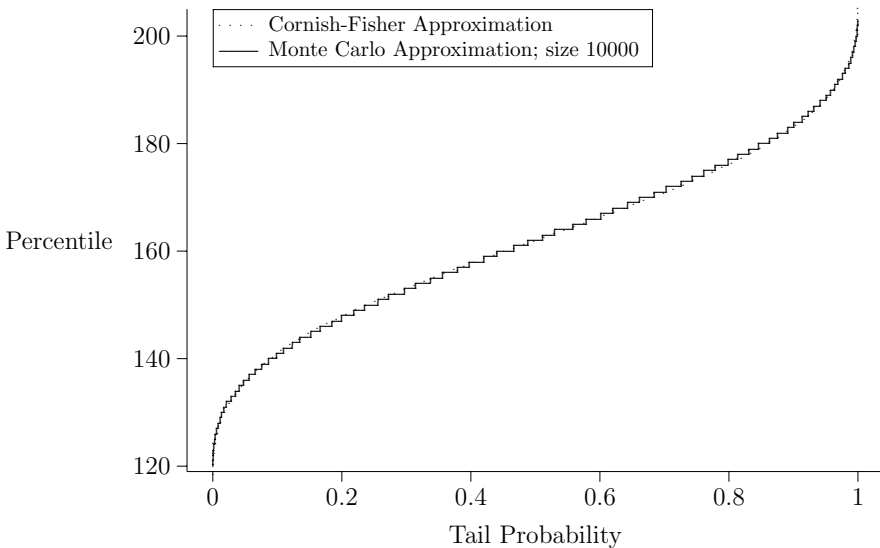
Inverse CDF of the Spearman Statistic T for Sample Size 8

Fig. 5.

Konishi, Niki, and Gupta (1988) apply these methods to independently but non-identically distributed sums of random variables. Withers (1988) and Hall (1988) apply these methods to the empirical cumulative distribution function instead of the true cumulative distribution function in order to avoid resampling in bootstrap applications. Such methods are discussed in §9.3 in the context of saddlepoint approximations.

3.13. The Lattice Case

As before, suppose that $\{Y_j\}$ are independent and identically distributed random variables, where for notational simplicity assume that the summands have zero mean and unit variance, and $Z = \sum_{l=1}^{n} Y_l / \sqrt{n}$. Consider again the problem of approximating the cumulative distribution function F_Z of Z. Now, however, suppose that the summands Y_j take values on the lattice $\{a + \Delta \mathfrak{Z}\}$ where $\mathfrak{Z}$ is the set of integers. This violates even the weaker regularity condition of §3.5.

One might be interested in this problem for three reasons. First, many distributions in applied statistics are lattice distributions. These include distributions of sufficient statistics in logistic regression, contingency table applications, and non-parametric inference. Hypothesis testing and confidence interval generation require the approximation of tail probabilities in these cases. Secondly, lattice distributions have applications to fields like queuing theory. Thirdly, the results of this section
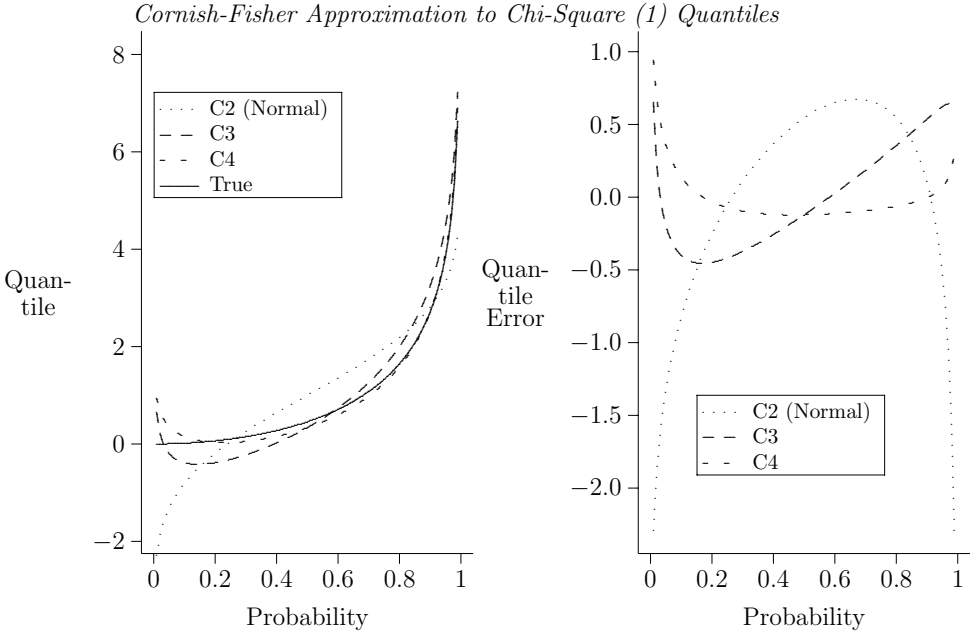
Cornish-Fisher Approximation to Chi-Square (1) Quantiles

Fig. 6a.

Cornish-Fisher Approximation to Chi-Square (10) Quantiles
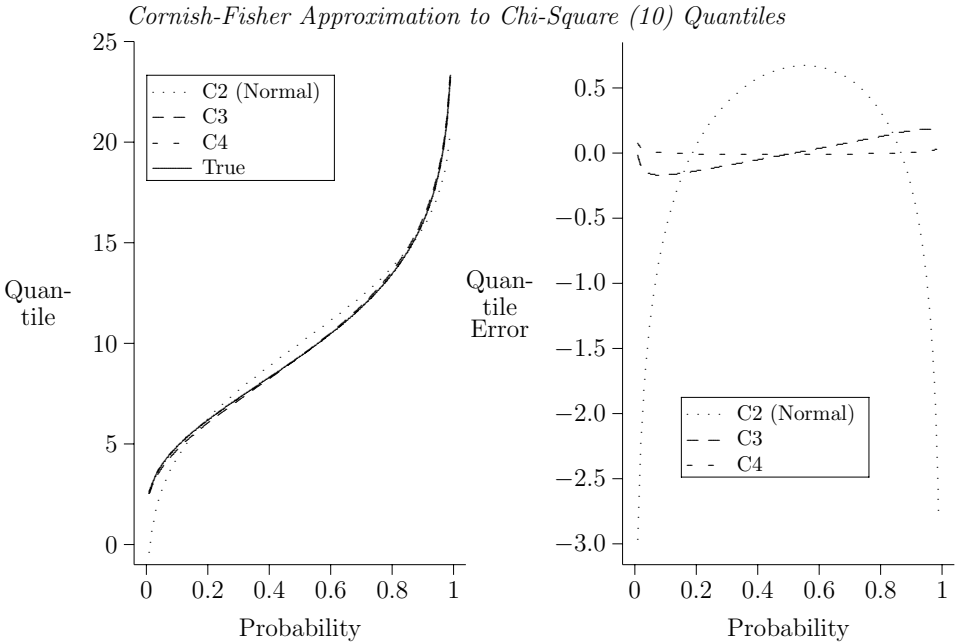

Fig. 6b.

tie together a number of seemingly unrelated ideas in probability; specifically, they extend the Edgeworth series to the next (and perhaps last) area of regular problems. Asymptotic expansions for lattice distributions exist, and putting them in the same framework as Edgeworth series has theoretical appeal.

A bit of reflection yields reasons why unmodified Edgeworth series will not suffice here. They clearly will not work uniformly, since the cumulative distribution function has jumps of order $O(1/\sqrt{n})$. Feller (1971) shows that the Edgeworth series evaluated only at continuity-corrected points $z^+ = z + \Delta/(2\sqrt{n})$ will yield results accurate to $o(1/\sqrt{n})$.

Esseen (1945) derives an approximation based on the Edgeworth series plus correction terms containing derivatives of the Edgeworth series, involving the functions Q_l given by

$$Q_l(y) = \begin{cases} (l-1)! g_l \sum_{j=1}^{\infty} \cos(2\pi j y)/(2^{l-1}(\pi j)^l) & \text{if } l \text{ is even} \\ (l-1)! g_l \sum_{j=1}^{\infty} \sin(2\pi j y)/(2^{l-1}(\pi j)^l) & \text{if } l \text{ is odd,} \end{cases} \tag{55}$$

with the constants g_l given by

$$g_l = \begin{cases} +1 & \text{if } l = 4j+1 \text{ or } l = 4j+2 \text{ for some integer } j \\ -1 & \text{if } l = 4j+3 \text{ or } l = 4j \text{ for some integer } j. \end{cases}$$

These functions have the following properties:
1. Q_l has period 1,
2. Q_l is piecewise polynomial,
3. Q_l is continuous for $l > 1$ on $\mathfrak{R}$.

Let B_l be the polynomial versions of Q_l on $[0, 1)$. The first few of these polynomials are:

$$B_1(y) = y - \frac{1}{2}, \ B_2(y) = y^2 - y + \frac{1}{6}, \ B_3(y) = y^3 - \frac{3y^2}{2} + \frac{y}{2}.$$

These polynomials are known as the Bernoulli polynomials and are defined by the relation

$$i\beta \frac{\exp(-i\beta y)}{1 - \exp(-i\beta)} = \sum_{l=0}^{\infty} B_l(y)(-i\beta)^l/l!$$

for $|\beta| < 2\pi$ (Haynsworth and Goldberg, 1965, p. 804).

In the presence of j cumulants, Esseen (1945) constructs an approximation valid uniformly to $o(n^{1-\mathsf{j}/2})$, the same order as the original Edgeworth series in the non-lattice case:

Theorem 3.13.1: *If $\{Y_n\}$ are independent and identically distributed random variables with zero mean and unit variance taking values on the lattice $\{a + \Delta\mathfrak{Z}\}$, with finite cumulant of order j, and if $Z = \sum_{l=1}^{n} Y_l/\sqrt{n}$, then Esseen's series*

$$E_{\mathsf{j}}(z; \boldsymbol{\kappa}^n) + D_{\mathsf{j}}(z; \boldsymbol{\kappa}^n) \tag{56}$$

is valid uniformly to $o(n^{1-j/2})$, where $D_j(z; \boldsymbol{\kappa}^n)$ is a discontinuous correction term

$$\sum_{l=1}^{j-2} \frac{1}{l!} (\Delta_n)^l Q_l((z-z_n)/(\Delta_n))(-1)^l \frac{d^l}{dz^l} E_j(z, \boldsymbol{\kappa}^n), \tag{57}$$

$\Delta_n = \Delta/\sqrt{n}$, $E_j(z, \boldsymbol{\kappa}^n)$ *is the Edgeworth series as in (46), and z_n is any point on the lattice $\{\sqrt{n}a + \Delta 3/\sqrt{n}\}$.*

Proof: Let $\zeta_n(\beta)$ be the characteristic function of Z. The inversion integral for lattice tail probabilities with separation $\Delta_n = \Delta/\sqrt{n}$ in (33) is

$$P[z > Z \geq z_0] = \frac{1}{2\pi} \int_{-\pi/\Delta_n}^{\pi/\Delta_n} \zeta_n(\beta) \frac{\exp(-i\beta z_0) - \exp(-i\beta z_n)}{1 - \exp(-i\Delta_n \beta)} \Delta_n \, d\beta$$

$$= \frac{1}{2\pi} \int_{-\pi/\Delta_n}^{\pi/\Delta_n} \zeta_n(\beta) \left(\frac{\exp(-i\beta z_0)}{1 - \exp(-i\Delta_n \beta)} - \exp(-i\beta z) \frac{\exp(-i\beta(z_0 - z))}{1 - \exp(-i\Delta_n \beta)} \right) \Delta_n \, d\beta, \tag{58}$$

where z_n is the minimal lattice point greater than or equal to z. This proof will be performed in four steps. First, $\zeta_n(\beta)$ in (58) will be replaced by $\xi_n(\beta)$ of §3.6. Second, the remaining factors in (58) will be expanded in a power series and truncated. Third, the inversion integral (58) will be extended over $\mathfrak{R}$. Fourth, the result will be shown to agree with (56).

Since the function $iy/(1 - \exp(iy))$ is bounded for $y \in [-\pi, \pi]$, there exists C such that $2\Delta_n/(1 - \exp(-i\beta\Delta_n))$ can be bounded by $C/|\beta|$, for $|\beta| \leq \pi/\Delta_n$. The quantity C may be chosen independently of n and Δ. As in §3.6, choose δ such that $|(\zeta_n(\beta) - \xi_n(\beta))/\beta|$ is bounded by (49) for $|\beta| < \delta\sqrt{n}$. Then for $|\beta| < \delta\sqrt{n}$ the difference between the integrand of (58) and this integrand with ζ_n replaced by ξ_n is also of the form (49), and uniformly,

$$\left| \frac{1}{2\pi} \int_{-\delta\sqrt{n}}^{\delta\sqrt{n}} \frac{(\zeta_n(\beta) - \xi_n(\beta))}{\beta} \frac{\beta}{1 - \exp(-i\Delta_n \beta)} \Delta_n \, d\beta \right| = o(n^{1-j/2}).$$

Over the region $\delta \leq |\beta| \leq \pi/\Delta$ the function $|\zeta_Y(\beta)|$ is bounded away from 1, and hence over the region $\delta\sqrt{n} \leq |\beta| \leq \pi/\Delta_n$, the function $\zeta_n(\beta\sqrt{n})$ converges uniformly and geometrically to 0. Also, $\xi_n(\beta)$ converges uniformly and geometrically to 0. Hence

$$\left| \frac{1}{2\pi} \int_{-\pi/\Delta_n}^{\pi/\Delta_n} (\zeta_n(\beta) - \xi_n(\beta)) \frac{1}{1 - \exp(-i\Delta_n \beta)} \Delta_n \, d\beta \right| = o(n^{1-j/2}). \tag{59}$$

A polynomial approximation will now be substituted for the terms of form $\exp(-i\beta y) i\beta(1 - \exp(-i\beta))^{-1}$ in (58). Choose $\delta > 0$ and $C_1 > 0$ such that

$$\left| i\beta^* \frac{\exp(-i\beta^* y)}{1 - \exp(-i\beta^*)} - \sum_{l=0}^{j-2} B_l(y)(-i\beta^*)^l/l! \right| < C_1 |\beta^*|^{j-1}$$

whenever $|\beta^*| < \delta$ and $y \in [0, 1)$. Set $\beta^* = \beta\Delta_n$. Choose any real u. Then

$$\left| \frac{1}{2\pi} \int_{-\pi/\Delta_n}^{\pi/\Delta_n} \frac{\xi_n(\beta)}{\beta} \left[\frac{\exp(-i\beta(\Delta_n y + u))\Delta_n\beta}{1 - \exp(-i\Delta_n\beta)} - \exp(-iu\beta) \sum_{l=0}^{\mathsf{j}-2} \frac{B_l(y)(-i\beta\Delta_n)^l}{l!} \right] d\beta \right|$$

$$\leq \frac{C_1}{2\pi} \int_{-\pi/\Delta_n}^{\pi/\Delta_n} \left| \xi_n(\beta)\beta^{\mathsf{j}-2} \right| d\beta \Delta_n^{\mathsf{j}-1},$$

where C_1 is independent of n, z, z_n, and δ. The integral on the right-hand side above can be bounded independently of n, z, z_n, and δ, yielding

$$\left| \int_{-\pi/\Delta_n}^{\pi/\Delta_n} \frac{\xi_n(\beta)}{2\pi\beta} \left[\frac{\exp(-i\beta(\Delta_n y + u))\Delta_n\beta}{1 - \exp(-i\Delta_n\beta)} - \exp(-iu\beta) \sum_{l=0}^{\mathsf{j}} \frac{B_l(y)(-i\beta\Delta_n)^l}{l!} \right] d\beta \right|$$

$$= O(n^{(1-\mathsf{j})/2}). \tag{60}$$

Applying (60) twice, once with $u = z$ and $y = (z_n - z)/\Delta_n$, and once with $u = z_0$ and $y = 0$, and applying the triangle inequality,

$$\mathrm{P}\left[z > Z \geq z_0\right] = \frac{1}{2\pi} \int_{-\pi/\Delta_n}^{\pi/\Delta_n} \frac{\xi_n(\beta)}{-i\beta}$$

$$\times \sum_{l=0}^{\mathsf{j}} \left(\exp(-i\beta z)B_l\left(\frac{z_n - z}{\Delta_n}\right) - \exp(-i\beta z_0)B_l\left(0\right) \right) \frac{(-i\beta\Delta_n)^l}{l!} \, d\beta + o(n^{1-\mathsf{j}/2}).$$

The third step in the proof is to prove that the above inversion integral is unchanged to $o(n^{1-\mathsf{j}/2})$ when the range of integration is extended to the entire real line. Since the above integrand on the compliment of the above range of integration may be bounded by $\exp(-n\pi^2/(4\Delta^2))$ times an integrable function, times a polynomial in n, the change occurring in extending the range of integration is of the proper order. By (43),

$$\mathrm{P}\left[z > Z \geq z_0\right] = \sum_{l=0}^{\mathsf{j}} \frac{\Delta_n^l}{l!} \left(B_l\left(\frac{z_n - z}{\Delta_n}\right) E_{\mathsf{j}}^{*(l)}(z; \boldsymbol{\kappa}^n) - B_l\left(0\right) E_{\mathsf{j}}^{*(l)}(z_0; \boldsymbol{\kappa}^n) \right) + o(n^{1-\mathsf{j}/2}),$$

where $E_{\mathsf{j}}^{*(l)}(z; \boldsymbol{\kappa}^n) = d^l/(dz^l)E_{\mathsf{j}}^*(z; \boldsymbol{\kappa}^n)$. Letting $z_0 \to \infty$, and discarding terms of size $o(n^{1-\mathsf{j}/2})$, gives (56).

$$\mathrm{Q.E.D}$$

Even at the continuity-corrected points the correct asymptotic approximation is not the usual Edgeworth series. Fig. 7 demonstrates the behavior of this approximation.

Bhattacharya and Rao (1976) present a lattice analogue for Theorem 3.11.2:

Theorem 3.13.2: *Suppose Y_j are independent random variables, with characteristic functions $\zeta_j(\beta)$, means zero and variances σ_j^2, and j is an integer greater than or equal to two, such that for all j, Y_j has a cumulant of order j. Assume that $\bar{\omega}_{n,\mathsf{j}}$ as*

defined in Theorem 3.11.2 remains bounded as n increases, and that ς_n^2 as defined in Lemma 3.11.1 is bounded away from zero. Further assume that each summand takes values in the lattice $\{a + \Delta 3\}$, and that this lattice is the minimal one supporting the entire sequence of random variables. Form the Edgeworth series $E_j(z, \boldsymbol{\kappa}^n)$ using (46) as in Theorem 3.11.2. Then (56) approximates $F_Z(z)$ with an accuracy of $O(n^{1-j/2})$, where F_Z is the cumulative distribution function of $Z = \sum_{j=1}^n Y_j / \sqrt{n\varsigma_n^2}$, (57) is calculated with $\Delta_n = \Delta/\sqrt{n}\varsigma_n^2$, and explicit dependence on sample size in (57) is replaced by a dependence through derivatives of $E_j(z, \boldsymbol{\kappa}^n)$.

Proof: This theorem parallels the development of §3.6.   Bound (53) replaces (26), causing (59) to hold with $O(n^{1-j/2})$ replacing $o(n^{1-j/2})$. The rest of the theorem proceeds as before.

<div align="right">Q.E.D</div>

As an example, Albers, Bickel, and van Zwet (1976) explore Edgeworth expansions for cumulative distribution functions of nonparametric one-sample location general score test statistics. For independent and identically distributed random variables $Y_1, \cdots, Y_n$ known to have an absolutely continuous distribution, tests of the hypothesis that their common median is zero may be constructed from the indicators V_i taking the value one if the Y_i is positive and zero otherwise, if the observations are ordered according to the magnitude of their absolute values. Let the general score test statistic be $T_n = \sum_j^n a_j^{(n)} V_j$ for constants $a_j^{(n)}$. In this case the summands are independent but not identically distributed, unless all of the $a_j^{(n)}$ are identical. The first four cumulants are

$$\kappa_1 = \sum_{j=1}^n a_j^{(n)}/2, \ \ \kappa_2 = \sum_{j=1}^n a_j^{(n)2}/4, \ \ \kappa_3 = 0, \ \ \kappa_4 = -\sum_{j=1}^n a_j^{(n)4}/8.$$

A simple example of this test statistic sets $a_j^{(n)} = j$, yielding the Wilcoxon signed rank statistic (Hettmansperger, 1984). In this case,

$$\kappa_1 = \frac{n(n+1)}{4}, \ \ \kappa_2 = \frac{n(n+1)(n+\frac{1}{2})}{12}, \ \ \kappa_3 = 0, \ \ \kappa_4 = -\frac{n(n+1)(n+\frac{1}{2})(n^2+n+\frac{1}{3})}{120},$$

and generally $\kappa_j = O(n^{j+1})$. Then if $\rho_{Y,j}$ is the invariant of order j,

$$\rho_{Y,j} = O(n^{j+1-3j/2}) = O(n^{1-j/2}),$$

and hence $T_n/\sqrt{\kappa_2}$ has invariants of the same order in n as is usual in Edgeworth series applications. Specifically, Theorem 3.13.2 applies. Furthermore, $\Delta_n = O(n^{-3/2})$, indicating that $E_4(z, \boldsymbol{\kappa}^n)$ is valid to $O(n^{-3/2})$. Albers, Bickel, and van Zwet (1976) derive the error bound of $O(n^{-5/4})$ for more general scores $a_j^{(n)}$.

As a further example, recall the distribution of T, related to Spearman's Rank Correlation, discussed in §3.12. Its Cornish–Fisher expansion to $o(1/n)$ is dependent on the validity of the standard Edgeworth series to the same order. As the minimal

lattice for T has spacings $1/(n+1)$, and $\mathrm{Var}\,[T] = n^2(n-1)/144$, $T/\sqrt{\mathrm{Var}\,[T]}$ takes values on a lattice with spacings $O(n^{-5/2})$. In this case errors induced by the lattice nature of the summands at continuity–corrected points are of order $O(n^{-5})$. An Edgeworth series theorem for dependent variables is needed in this case. Prášková–Vizková (1976) addresses this question as well.

Robinson (1982) addresses this problem of calculating tail probabilities for score statistics. Details will follow in a later chapter devoted to resampling and permutation methods. Kong and Levin (1996) address the problem of approximating multivariate distribution functions of sufficient statistics arising in logistic regression, and demonstrate that if the lattice nature of the problem is not too strong, conventional Edgeworth series methodology applies.

3.14. Sheppard's Corrections and Edgeworth Series

The preceding section presented an asymptotic approximation to tail areas for lattice variables that involved adding a correction to the standard Edgeworth series for continuous variables. This section discusses modifications to the cumulants that make the standard Edgeworth series valid for lattice variables. One might try constructing an asymptotically correct series expansion by smoothing the original lattice distribution to a continuous distribution whose cumulative distribution function agrees with the original cumulative distribution function at continuity-corrected points z^+, and using Edgeworth approximation techniques to approximate the smoothed cumulative distribution function. One method for doing this is as follows:

1. Construct a smoothed variable Z^* by choosing U uniform on $[-\frac{1}{2}, \frac{1}{2}]$. Set $Z^* = Z + (\Delta/\sqrt{n})U$. Then Z^* has a density, and furthermore its cumulative distribution function agrees with F_n at continuity-correction points.
2. The cumulant of order k of $(\Delta/\sqrt{n})U$ is $\epsilon_k^n = (\Delta/\sqrt{n})^k B_k/k$, for $k \geq 2$; the constants B_k are known as the Bernoulli numbers and are derived directly from the series expansion for the characteristic function for U as part of the exercises for Chapter 2.

Since Z^* is not a standardized sum the Edgeworth theorems do not necessarily apply. The density for Z^* has jumps of order $O(1/\sqrt{n})$ at continuity corrected points, but any approximation by a continuous function must have errors that do not get smaller in n as more cumulants are used.

The previous smoothing attempt failed because the density constructed was not differentiable. One might change the distribution of the added smoother U to produce a result with a density having more derivatives.

A simpler smoothing method is as follows: If D represents a continuous random variable C grouped to a lattice with spacing Δ, then the cumulants $\kappa_{D,k}$ of D are given by $\kappa_{C,k} - \epsilon_k^n$ where $\kappa_{C,k}$ are the cumulants of C. These modifications to cumulants of grouped data to recover cumulants of the underlying continuous variate are known as Sheppard's corrections, and are tabulated in Table 4. To obtain

Sheppard-corrected cumulants, subtract the last column from the appropriate raw cumulant, after multiplying by the appropriate power of the lattice spacing.

<center>Table 4: Bernoulli Numbers and Sheppard's Corrections</center>

k	B_k	B_k/k
any odd integer	0	0
2	1/6	1/12
4	−1/30	−1/120
6	1/42	1/252

Heuristically, then, $\zeta_C \zeta_{\Delta U} \approx \zeta_{C, \Delta U / \sqrt{n}}$. and that the continuous variable C and the smoothing ΔU are approximately independent in the sense measured by multiplicative nature of the characteristic functions, for Δ small. Calculations are performed using calculus of finite differences:

Theorem 3.14.1: *Suppose the random variable C has a continuous density h on $\Re$, and that the first $2l$ derivatives of $h(x)$ exist and converge to zero as $x \to \pm\infty$. This condition is known as high contact. Let $D = \Delta \langle \frac{C}{\Delta} \rangle$ be C rounded to the nearest lattice point, and define $U = (C - D)/\Delta$ to be the standardized difference between C and the nearest lattice point, which takes values on the interval $(-\frac{1}{2}, \frac{1}{2}]$. Here $\langle . \rangle$ denotes the nearest integer, with the convention that integers plus one half are rounded upwards. Let $\zeta_C(\beta_1)$, $\zeta_U(\beta_2)$, and $\zeta_{C,U}(\beta_1, \beta_2)$ be the characteristic functions of C, U, and the pair (C, U) respectively. Then as $\Delta \to 0$,*

$$\zeta_{C,U}(\beta_1, \beta_2) = \zeta_C(\beta_1)\zeta_U(\beta_2) + O(\Delta^{2l}) \text{ and } \zeta_U(\beta_2) = \sinh(\tfrac{1}{2}i\beta_2)/(\tfrac{1}{2}i\beta_2) + O(\Delta^{2l}).$$

Proof: By definition,

$$\zeta_{C,U}(\beta_1, \beta_2) = \int_{-\infty}^{\infty} \exp[i(s\beta_1 + (s/\Delta - \langle s/\Delta \rangle \beta_2))]h(s) \, ds$$

$$= \int_{-\infty}^{\infty} \exp[i(s(\beta_1 + \beta_2/\Delta) - \langle s/\Delta \rangle \beta_2)]h(s) \, ds$$

$$= \sum_{j \in \mathfrak{Z}} \exp(-ij\beta_2) \int_{(j-1/2)\Delta}^{(j+1/2)\Delta} \exp[is(\beta_1 + \beta_2/\Delta)]h(s) \, ds.$$

Also,

$$\zeta_C(\beta_1)\zeta_U(\beta_2) = \left[\int_{-\infty}^{\infty} e^{is\beta_1} h(s) \, ds \right] \left[\sum_{j \in \mathfrak{Z}} e^{-ij\beta_2} \int_{(j-1/2)\Delta}^{(j+1/2)\Delta} e^{is\beta_2/\Delta} h(s) \, ds \right]$$

$$= \left[\int_{-\infty}^{\infty} \exp(is\beta_1)h(s) \, ds \right] \left[\sum_{j \in \mathfrak{Z}} \int_{-\frac{1}{2}}^{\frac{1}{2}} \exp(it\beta_2)h(\Delta(t + j))\Delta \, dt \right].$$

Hence $\zeta_{C,U}(\beta_1, \beta_2) - \zeta_C(\beta_1)\zeta_U(\beta_2)$ may be expressed as

$$\sum_{j \in 3} \int_{-\frac{1}{2}}^{\frac{1}{2}} \left[e^{i(\Delta(j+t)\beta_1 + t\beta_2)} - \int_{-\infty}^{\infty} e^{is\beta_1 + it\beta_2} h(s) \; ds \right] h(\Delta(j+t))\Delta \; dt$$

$$= \sum_{j \in 3} \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{it\beta_2} \left[\int_{-\infty}^{\infty} \left\{ e^{i\Delta(j+t)\beta_1} - e^{is\beta_1} \right\} h(s) \; ds \right] h(\Delta(j+t))\Delta \; dt$$

$$= \sum_{j \in 3} e^{-ij\beta_2} \int_{(j-1/2)\Delta}^{(j+1/2)\Delta} e^{it\beta_2/\Delta} \left[\int_{-\infty}^{\infty} \left\{ e^{it\beta_1} - e^{is\beta_1} \right\} h(s) \; ds \right] h(t) \; dt = \sum_{j \in 3} f(j)$$

where

$$f(j) = e^{-ij\beta_2} \int_{(j-1/2)\Delta}^{(j+1/2)\Delta} e^{it\beta_2/\Delta} [\exp(it\beta_1) - \zeta_C(\beta_1)] h(t) \; dt.$$

Note that f depends also on β_1 and β_2.

If $f : \mathfrak{R} \to \mathfrak{R}$ is a function with $2l$ continuous derivatives, and $\lim_{t\to\pm\infty} f^{(k)}(t)$ exists for all $k \le 2l$, then, by the Euler–Maclaurin Summation Formula (Haynsworth and Goldberg, 1965, page 806),

$$\sum_{j \in 3} f(j) = \int_{-\infty}^{\infty} f(t) \; dt + \lim_{t\to\infty} \left[\tfrac{1}{2}(f(t) + f(-t)) + \sum_{k=1}^{l-1} \frac{B_{2k}}{(2k)!} (f^{(2k-1)}(t) - f^{(2k-1)}(-t)) \right]$$

$$+ \frac{B_{2l}}{(2l)!} \sum_{j \in 3} f^{(2l)}(j + \delta_0)$$

for some $\delta_0 \in (0,1)$. The first term on the right hand side above is

$$\int_{-\infty}^{\infty} f(x) \; dx = \int_{-\infty}^{\infty} e^{-ix\beta_2} \int_{(x-1/2)\Delta}^{(x+1/2)\Delta} e^{it\beta_2/\Delta} [\exp(it\beta_1) - \zeta_C(\beta_1)] h(t) \; dt \; dx$$

$$= \int_{-\infty}^{\infty} \int_{t/\Delta - \frac{1}{2}}^{t/\Delta + \frac{1}{2}} e^{-ix\beta_2} e^{it\beta_2/\Delta} [\exp(it\beta_1) - \zeta_C(\beta_1)] h(t) \; dx \; dt$$

$$= \int_{-\infty}^{\infty} (e^{i\beta_2/2} - e^{-i\beta_2/2})/(i\beta_2) [\exp(it\beta_1) - \zeta_C(\beta_1)] h(t) \; dt = 0.$$

To evaluate the second term, derivatives of f are needed. Since

$$f(x) = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{it\beta_2} g(t + x) \; dt,$$

where

$$g(t) = \Delta \int_{-\infty}^{\infty} (e^{it\Delta\beta_1} - e^{is\beta_1}) h(s) \; ds h(\Delta t) = \Delta h(\Delta t)[e^{it\Delta\beta_1} - \zeta_C(\beta_1)],$$

then

$$f^{(k)}(x) = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{it\beta_2} g^{(k)}(t + x) \; dt \to 0$$

as $x \to \pm\infty$ for $k < 2l$, by the high contact condition for h. Also,

$$\sum_{j \in 3} f^{(2l)}(j + \delta_0) \leq \Delta^{2l} \sum_{j \in 3} \int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_{k=0}^{2l} h^{(k)}(\Delta(t + j))\Delta \, dt$$

$$= \Delta^{2l} \sum_{k=0}^{2l} \int_{-\infty}^{\infty} h^{(k)}(\Delta t)\Delta \, dt = O(\Delta^{2l}).$$

This now proves the first claim of the theorem.

The characteristic function ζ_U of the smoothing variable U must now be calculated.

$$\zeta_U(\beta_2) = \sum_{j \in 3} \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{it\beta_2} h(\Delta(t + j))\Delta \, dt = \sum_{j \in 3} f(j),$$

where

$$f(j) = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{it\beta_2} h(\Delta(t + j))\Delta \, dt.$$

The first term in the Euler–Maclaurin Summation Formula is

$$\int_{-\infty}^{\infty} f(x) \, dx = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{it\beta_2} [\int_{-\infty}^{\infty} h(\Delta(t + x))\Delta \, dx] \, dt = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{it\beta_2} \, dt = \frac{\sinh(\frac{1}{2}i\beta_2)}{\frac{1}{2}i\beta_2}.$$

Also, $f^{(k)}(x) = \Delta^k \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{it\beta_2} h^{(k)}(\Delta(t+x))\Delta \, dt$. Again the integrability of derivatives of h insures that $\lim_{t \to \pm\infty} f^{(k)}(t) = 0$. Hence

$$\zeta_U(\beta_2) = \frac{\sinh(\frac{1}{2}i\beta_2)}{\frac{1}{2}i\beta_2} + \Delta^{2l} \sum_{j \in 3} \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{it\beta_2} h^{(2l)}(\Delta(t + j))\Delta \, dt = \frac{\sinh(\frac{1}{2}i\beta_2)}{\frac{1}{2}i\beta_2} + O(\Delta^{2l}).$$

Q.E.D

3.15. Continuity-Corrected Edgeworth Series

The smoothing argument presented earlier suggests using an Edgeworth series with cumulants adjusted by Sheppard's corrections to approximate the lattice distribution F_Z. Kolassa and McCullagh (1990) prove that by evaluating such an Edgeworth expansion at continuity-corrected points, the resulting errors are as small as are usually obtained in an Edgeworth approximation.

Suppose $\{Y_i\}$ is an independent and identically distributed collection of random variables in $\mathfrak{R}$, with mean 0, variance σ^2, and cumulants $\boldsymbol{\kappa} = (\kappa_j, j = 1, 2, 3, \ldots)$. Suppose further that these variables are confined to the lattice $a + \Delta 3$ almost surely. Let $Z = \sum_{i=1}^{n} Y_i / \sqrt{n}$. Then Z has cumulants $\boldsymbol{\kappa}^n = (\kappa_j^n, j = 1, 2, 3, 4, \ldots)$, with $\kappa_j^n = \kappa_j n^{(2-j)/2}$. Let $\lambda_j^n = \kappa_j - \epsilon_j n^{-1}$ be the adjusted cumulants. Let $\boldsymbol{\lambda}^n$ denote the

infinite vector $(0, \lambda_2^n, \lambda_3^n, \lambda_4^n, \ldots)$, and let $E_j(z; \boldsymbol{\lambda}^n)$ be the Edgeworth series defined in (42), with pseudo-moments as in (44), but with λ_j^n's in place of the κ_j^n's. Then

$$E_j(z, \boldsymbol{\lambda}^n) = \psi_{j,z}\left(\exp\left[\tfrac{1}{2}(\lambda_2 - 1)\beta^2 + \sum_{j=3}^{j} \frac{1}{j!}\lambda_k^n \beta^j\right]\right) + o(n^{-\frac{j-2}{2}}).$$

For instance, one constructs $E_4(z^+, \boldsymbol{\lambda}^n)$ from

$$\Phi(z^+) - \phi(z^+)\left(\frac{h_2(z^+)\kappa_3}{\sqrt{n}} - \frac{h_1(z^+)\epsilon_2}{n} + \frac{h_3(z^+)}{n}\left(\frac{\epsilon_2^2}{8n} + \frac{\epsilon_4}{24n} + \frac{\kappa_4}{24}\right)\right),$$

but by definition $\psi_{4,z}\left(\exp\left[\tfrac{1}{2}(\lambda_2 - \kappa_2)\beta^2 + \sum_{j=3}^{4}\frac{1}{j!}\lambda_k^n \beta^j\right]\right)$ omits the terms of size $o(n^{-1})$. The expression $E_4(z^+, \boldsymbol{\lambda}^n) = \Phi(z^+) - \phi(z^+)\left(\frac{h_2(z^+)\kappa_3}{\sqrt{n}} - \frac{h_1(z^+)\epsilon_2}{n} + \frac{\kappa_4 h_3(z^+)}{24n}\right)$ has more terms than $E_4(z^+, \boldsymbol{\kappa}^n)$, since second cumulants are no longer matched; instead, following Chambers (1967), pick the baseline variance to be the asymptotic variance. Since $\lambda_2^n - \sigma^2 = -1/12n$, the exponent on the left hand side of (44) now has a quadratic term.

Theorem 3.15.1: *Suppose that a random variable Z, with zero mean and unit variance, is supported on a lattice with spacing $\Delta_n = \Delta/\sqrt{n}$. When the cumulative distribution function of Z is approximated using an Edgeworth series, with the cumulants adjusted according to Sheppard's corrections, and the result is evaluated at continuity – corrected lattice points, the result is valid to the same order in n as in standard applications of the Edgeworth series; that is, $F_Z(z^+) = E_j(z^+; \boldsymbol{\lambda}^n) + o(n^{-\frac{j-2}{2}})$ for any lattice point z, with $z^+ = z + \Delta/(2\sqrt{n})$.*

Proof: The following is similar to the treatment of series expansions found in Chambers (1967), pages 368-371. The proof is completely algebraic and proceeds by the manipulations of power series generating functions for cumulants.

   This proof exploits a relation between power series and Edgeworth-like series. Let $P[\beta]$ be the set of formal power series in the variable β, of form $\sum_{j=0}^{\infty} \alpha_j \beta^j$, where the coefficients α_j are polynomials in $1/\sqrt{n}$. Consider this set to contain all conceivable exponentiated differences of the true cumulant generating function from the approximating normal cumulant generating function, convergent or not; the α_j are the possible cumulant differences. Choose an ordinate z at which to evaluate an Edgeworth series. Let $\psi_{j,z} : P[\beta] \to \mathfrak{R}$ be the function that maps $\sum_{j=0}^{\infty} \alpha_j \beta^j$ to the Edgeworth-like series evaluated at z:

$$\alpha_0 \Phi(z) - \phi(z)\sum_{j=1}^{\infty} \alpha_j h_{j-1}(z),$$

with all terms of size $o(n^{-\frac{j-2}{2}})$ discarded. This will not be an actually Edgeworth series unless $\alpha_0 = 1$ and some other restrictions among the coefficients hold. Effectively, we apply the Fourier inversion operation termwise, and discard terms that

are small enough, to obtain the Edgeworth series. The usual Edgeworth series for the standardized mean of independent and identically distributed random variables, with cumulants $\boldsymbol{\kappa}^n$, is

$$E_{\mathrm{j}}(z, \boldsymbol{\kappa}^n) = \psi_{\mathrm{j},z} \left(\exp \left[\sum_{j=3}^{k} \frac{1}{j!} \kappa_j^n \beta^j \right] \right).$$

The function $\psi_{\mathrm{j},z}$ has the following linearity property:

$$\psi_{\mathrm{j},z}(\gamma_1 p_1(\beta) + \gamma_2 p_2(\beta)) = \gamma_1 \psi_{\mathrm{j},z}(p_1(\beta)) + \gamma_2 \psi_{\mathrm{j},z}(p_2(\beta)).$$

Also,

$$\psi_{\mathrm{j},z}(\beta^j p(\beta)) = (-1)^j D^j \psi_{\mathrm{j},z}(p(\beta)). \tag{61}$$

Esseen's series (56) is a linear combination of derivatives of the Edgeworth series. The periodicity of the functions B_j causes all of the multipliers for these derivatives to be independent of the specific continuity-corrected point under consideration. Using (61) express the derivatives of the Edgeworth series as ψ_{j,z^+} applied to a multiple of a power of β times the power series

$$\exp \left[\sum_{j=3}^{k} \frac{1}{j!} \kappa_j^n \beta^j \right] \tag{62}$$

to generate the Edgeworth series. Using the linearity of ψ_{j,z^+}, express this as ψ_{j,z^+} applied to the product of a new series times the series (62). This new series will now be expressed as the exponential of the series with Sheppard's corrections.

Substituting into Esseen's series (56),

$$F_Z(z) = E_{\mathrm{j}}(z, \boldsymbol{\kappa}^n) + \sum_{j=1}^{\mathrm{j}-2} \Delta_n^j \frac{1}{j!} B_j \left(\frac{(z - z_n)}{\Delta_n} \right) (-1)^j \frac{d^j}{dx^j} E_{\mathrm{j}}(z, \boldsymbol{\kappa}^n) + o(n^{-\frac{\mathrm{j}-2}{2}}).$$

Choose any integer J. Let $z^+ = (na + J\Delta + \Delta/2)/\sqrt{n}$. Then

$$(z^+ - z_n)/\Delta_n = J - \lfloor na/\Delta \rfloor + \tfrac{1}{2},$$

which is an integer plus half. Hence for all positive integers j, $B_j((z^+ - z_n)/\Delta_n) = B_j(\tfrac{1}{2})$, by the periodicity of B_j. Straight-forward calculations show that

$$B_j(0) = B_j \text{ and } B_j(\tfrac{1}{2}) = (2^{1-j} - 1)B_j, \tag{63}$$

where again, B_j is the Bernoulli number of order j. Then

$$F_Z(z^+) = \sum_{j=0}^{j-2} n^{-j/2}\frac{1}{j!}B_j g_j(2^{1-j}-1)D^j E_j(z^+;\boldsymbol{\kappa}^n) + o(n^{-\frac{i-2}{2}})$$

$$= \sum_{j=0}^{j-2} n^{-j/2}\frac{1}{j!}B_j(2^{1-j}-1)(-1)^j\psi_{j,z^+}\left(\beta^j \exp\left[\sum_{j=3}^{k}\frac{1}{j!}\kappa_j^n\beta^j\right]\right) + o(n^{-\frac{i-2}{2}})$$

$$= \psi_{j,z^+}\left(\sum_{j=0}^{j-2} n^{-j/2}\frac{1}{j!}B_j(2^{1-j}-1)\beta^j \exp\left[\sum_{j=3}^{k}\frac{1}{j!}\kappa_j^n\beta^j\right]\right) + o(n^{-\frac{i-2}{2}})$$

$$= \psi_{j,z^+}\left(\exp\left[\sum_{j=2}^{\infty}\frac{-B_j}{j\,j!}n^{-\frac{j}{2}}\beta^j\right]\exp\left[\sum_{j=3}^{k}\frac{1}{j!}\kappa_j^n\beta^j\right]\right) + o(n^{-\frac{i-2}{2}})$$

since

$$\sum_{s=0}^{\infty}(2^{1-s}-1)\frac{B_s\beta^s}{s!n^{s/2}} = \frac{(\beta/\sqrt{n})\exp[(\beta/\sqrt{n})/2]}{\exp(\beta/\sqrt{n})-1} = \exp\left[\sum_{j=2}^{\infty}\frac{-B_j}{j\,j!}n^{-j/2}\beta^j\right]. \quad (64)$$

The last equality in (64) follows directly from Exercise 9 of Chapter 2. The first equality will be proved as part of the exercises in this chapter. Powers of (-1) drop out of the second of these equations, because the odd Bernoulli numbers are zero. Sheppard's correction for the cumulant of order j is $-B_j/(jn^{j/2})$. Let $\boldsymbol{\lambda}^n = (\lambda_k^n)$ be the cumulants $\boldsymbol{\kappa}^n$ adjusted by Sheppard's correction.

Hence

$$F_Z(z^+) = \psi_{j,z^+}\left(\exp\left[\sum_{j=3}^{j}\frac{1}{j!}\lambda_k^n\beta^j\right]\right) + o(n^{-\frac{i-2}{2}}) = E_j(z^+,\boldsymbol{\lambda}^n) + o(n^{-\frac{i-2}{2}}).$$

Q.E.D

3.16. Exercises

1. Justify the first equality in (64). Distributing two terms in the factor $(2^{1-s}-1)$ into two separate infinite series should help.
2. Justify the first equality (63) by evaluating the Fourier series expressions (55) for $B_k(0)$, substituting this expression in for B_k on the right hand side of (34), and showing that the result is the left hand side of (34).
3. Justify the second equality(63) by evaluating the Fourier series expressions expression (55) for $B_k(\frac{1}{2})$, and comparing it to the similar expansion for $B_k(0)$.
4. Refer to the example of calculating an Edgeworth expansion for a $\Gamma(p,\lambda)$ random variable, as in §3.9. Prove that $\int_{-\infty}^{\infty}|\zeta(\beta)|^k\,d\beta < \infty$ if and only if $k > 1/p$.
5. Justify relation (43).
6. Justify claim (60).

7. Analytically or numerically calculate the cumulative distribution function for the convolution of 3, 4, 5, and 6 random variables uniform on $[0, 1]$. Compare tail probabilities one, two, and three standard deviations above the mean with the Edgeworth estimates.

8. Let F_n be the cumulative distribution function of the sum of n independent random variables each with cumulative distribution function

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \Phi(\sqrt{x}) & \text{otherwise} \end{cases}.$$

   Kolassa (1992) constructs a test statistic $S(\boldsymbol{X}, \boldsymbol{\theta})$ whose cumulative distribution function is bounded by F_n if the parameter vector $\boldsymbol{\theta}$ governed a mechanism generating a set of data summarized by $\boldsymbol{X}$ and consisting of n data points. If s satisifes $F_n(s_n) = .95$ then $\{\boldsymbol{\theta}|S(\boldsymbol{X}, \boldsymbol{\theta}) \leq s_n\}$ represents a 95% confidence region for $\boldsymbol{\theta}$. Calculate s_5, s_{10}, and s_{20} numerically, and compare with the Cornish–Fisher expansions to orders $o(1)$, $o(1/\sqrt{n})$, and $o(1/n)$. Also, compare bounds on the coverage rates of the resulting regions when s_n is calculated using the normal approximation, by applying the Edgeworth approximation incorporating the kurtosis term.

9. Compare the accuracies of Cornish–Fisher and normal approximations to the Mann–Whitney–Wilcoxon statistic (Hettmansperger, 1984) percentage points. These approximations are available as FORTRAN subroutines in the STATLIB archive, at `http://statlib@lib.stat.cmu.edu`. Select software from the journal *Applied Statistics*. Algorithm 62 calculates the distribution of the Mann-Whitney-Wilcoxon statistic, and Algorithm 234 calculats the Cornish Fisher inversion of the Mann-Whitney-Wilcoxon statistic. Compare the accuracy of the various Cornish–Fisher approximations for a variety of sample sizes. Algorithm 234 takes as input standard normal quantiles. You may also need Algorithm 111, calculating these quantiles; this algorithm is coded as function PPND. Let m be the size of the smaller of the two groups. Let n be the other group size. Let N be the sum of group sizes. Here N is an input to Algorithm 234, and n is the input to Algorithm 62. Algorithm 62, coded as subroutine `UDIST`, calculates probabilities for the each of the potential statistic values, starting with the minimal value. You'll have to supply a real function that gives the ranks, which in this case is the identity function. We really want the tail probablities; modify the code as directed in the comments. The minimal statistic value is $m(m+1)/2$, and so after modifying the code entry i in the output correspond to a sum of ranks of $m(m+1)/2 + i - 1$. Algorithm 234, coded as the subroutine `PPRANK`, takes as the definition of the statistic the rank sum divided by $N + 1$. Hence the Cornish–Fisher expansion evaluated at the probabilities in component i of the results from Algorithm 62 to $(m(m+1)/2 + i - 1)/(N+1)$. Finaly, Algorithm 234 contains a mistake: The constant `CTA1` should have an exponent of -1 rather than 0.

4
Saddlepoint Series for Densities

In many statistical applications, approximations to the probability that a random variable exceeds a certain threshold value are important. Such approximations are useful, for example, in constructing tests and confidence intervals, and for calculating p-values. Edgeworth series converge uniformly quickly over the entire possible range of the random variable, when error is measured in an absolute sense. Often times, relative error behavior is more important than absolute error behavior; an error of .005 is of little importance when considering tests of approximate size .05 but is of great importance when considering tests of approximate size .001. Saddlepoint methodology is a method for achieving in many cases uniform bounds on relative error over the range of the distribution. This work was pioneered by Daniels (1954).

This chapter considers the problem of estimating the density of a random variable X, whose distribution depends on n, and whose cumulant generating function is represented by $\mathcal{K}_X$. Throughout this chapter implicitly assume that such a density exists for sufficiently large n. This condition will be strong enough for the introductory material heuristically motivating the saddlepoint density approximation as a rescaled approximation to a member of the natural exponential family containing the true density. When motivating the saddlepoint density as the asymptotic expansion of a complex integral over a properly deformed path, integrability of the characteristic function for sufficiently large n will be assumed, as was done for Edgeworth series density theorems. Recall that the equivalence of these assumptions was given in §2.4. Additional regularity conditions will be added when considering uniform relative error behavior of the saddlepoint approximation.

The saddlepoint approximation is first explored heuristically using an exponential family, and a parallel is developed between the saddlepoint and the maximum likelihood estimator. The question of existence of the saddlepoint is addressed. The saddlepoint approximation is derived using saddlepoint integration techniques applied to the density inversion integral. Advantages of saddlepoint techniques including uniform relative error are discussed. Possibilities of improving the accuracy of the saddlepoint approximation by determining a multiplicative correction

to normalize the approximate density to integrate to unity are considered. Finally, numeric hints for applying saddlepoint approximations are given.

The following chapter will concern saddlepoint distribution function approximations. Dividing material on saddlepoint approximations into two chapters, the first for density approximations and the second for distribution function approximations, while Edgeworth density and distribution function approximations are included in one chapter, may appear as a clumsy disruption of a natural parallelism between Edgeworth and saddlepoint approximations. I have chosen to organize the material in this manner precisely because while saddlepoint density approximations may be derived from Edgeworth density approximations, the same does not hold for distribution function approximations. In general Edgeworth distribution function approximations can be expressed as integrals of Edgeworth density approximations, and in general saddlepoint density approximations may not be analytically integrated.

The inversion results presented here will have their primary applications in the following chapter.

4.1. Expression using Exponential Families

Suppose that an approximation to a density $f_X(x)$ is desired, based on its cumulant generating function $\mathcal{K}_X(\gamma)$. Suppose further that the distribution of X depends on a parameter n, and that one desires to approximate this density far in the tails. Edgeworth series may work poorly in this case. An alternative is to embed the density of interest in an exponential family, and choose a density in this exponential family to estimate. Specifically, let $f_X(x, \gamma) = \exp(x\gamma - \mathcal{K}_X(\gamma))f_X(x)$. Here the density of interest is embedded in an exponential family having the cumulant generating function evaluated at τ: $\mathcal{K}_X(\tau + \gamma) - \mathcal{K}_X(\gamma)$. An approximation for any member of this family easily results in an approximation for the original density, since members of this family differ only by a factor of $\exp(x\gamma - \mathcal{K}_X(\gamma))$. Next one determines which member of this family is easiest to approximate. Recall that Edgeworth series are most accurate for ordinates near the mean, and least accurate for ordinates in the tails. One would expect that the easiest member of this family to approximate is the one whose mean is equal to the ordinate at which we wish to evaluate the approximation. The mean of the generic member of this family whose parameter is γ is $\mathcal{K}'_X(\gamma)$. Let $\hat{\gamma}$ be the parameter associated with the chosen member of the exponential family. Then $\hat{\gamma}$ is given by the solution to

$$\mathcal{K}'_X(\hat{\gamma}) = x. \tag{65}$$

One might think of this solution as the maximum likelihood estimator for the exponential family model based on the data x.

Once the element of the exponential family to be approximated is chosen, one must choose an approximation for it. Start with the Gaussian approximation to

$f_X(x; \hat{\gamma})$:

$$\exp(-(x - \mathcal{K}'_X(\hat{\gamma}))^2 \mathcal{K}''_X(\hat{\gamma})^{-1}/2)/\sqrt{2\pi\mathcal{K}''_X(\hat{\gamma})} = 1/\sqrt{2\pi\mathcal{K}''_X(\hat{\gamma})},$$

implying the approximation to $f_X(x)$:

$$g_X(x) = \exp(\mathcal{K}_X(\hat{\gamma}) - \hat{\gamma}x)/\sqrt{2\pi\mathcal{K}''_X(\hat{\gamma})}. \tag{66}$$

Although the standard normal approximation to $f_X(x; \hat{\gamma})$ is used, the resulting approximation to $f_X(x)$ uses information from all of the cumulants, because of the way in which $\hat{\gamma}$ is chosen. Unlike the higher-order Edgeworth expansions, this approximation is fortunately always positive. This approximation has a very simple form, but often not calculable explicitly, since an analytic expression for the saddlepoint $\hat{\gamma}$ is not always available.

To improve the accuracy of (66) one might use a higher-order Edgeworth expansion to $f_X(x; \hat{\gamma})$:

$$
\begin{aligned}
g_X(x) &= \frac{\exp(K_X(\hat{\gamma}) - \hat{\gamma}x)}{\sqrt{K''_X(\hat{\gamma})}} \phi(0) \sum_{j=0}^{\infty} h_j(0)\mu_j^*(\hat{\gamma})/j! \\
&= \frac{\exp(K_X(\hat{\gamma}) - \hat{\gamma}x)}{\sqrt{K''_X(\hat{\gamma})}} \phi(0) \left[1 + \frac{\hat{\rho}_{X,3}h_3(0)}{6} + \left(\frac{\hat{\rho}_{X,4}h_4(0)}{24} + \frac{10\hat{\rho}_{X,3}^2 h_6(0)}{720} \right) \right. \\
&\quad \left. + \left(\frac{\hat{\rho}_{X,5}h_5(0)}{120} + \frac{35\hat{\rho}_{X,3}\hat{\rho}_{X,4}h_7(0)}{5040} \right) + \cdots \right] \\
&= \frac{\exp(K_X(\hat{\gamma}) - \hat{\gamma}x)}{\sqrt{K''_X(\hat{\gamma})}} \phi(0)[1 + \hat{\rho}_{X,4}/8 - 5\hat{\rho}_{X,3}^2/24 + \cdots] \tag{67}
\end{aligned}
$$

using $h_4(0) = 3, h_6(0) = -15$. Here $\hat{\rho}_{X,j} = \mathcal{K}_X^{(j)}(\hat{\gamma})/(\mathcal{K}_X^{(2)}(\hat{\gamma}))^{j/2}$. When X is the average of n independent and identically distributed random variables, $\hat{\rho}_{X,j} = O(n^{(2-j)/2})$.

The above approximation (66) is derived from the cumulant generating function $\mathcal{K}_X$ for the associated random variable. The nature of the dependence of $\mathcal{K}_X$ on n was not important in this heuristic development. When results involving any mathematical rigor are desired, however, assumptions about the form of $\mathcal{K}_X$, or almost equivalently about the distribution of X, are needed. The rigorous development for the Edgeworth series of §§3.6 and 3.7 involved standardized sums of independent and identically distributed random variables. Limit theorems for other standardized sums were also considered in §3. When developing the saddlepoint approximations, it will be useful to take the mean of independent and identically distributed random variables as the prototypical distribution to be approximated. Reasons for this will be given shortly.

When the distribution of X, the mean of independent and identically distributed copies of a random variable Y, is desired, usually an expression in terms

of the cumulant generating function of the individual addends is preferred. Since $\mathcal{K}_X(\gamma) = n\mathcal{K}_Y(\gamma/n)$, the saddlepoint $\hat{\gamma}$ satisfies $\mathcal{K}'_Y(\hat{\gamma}/n) = x$, and the approximation (66) becomes $g_X(x) = \sqrt{n}\exp(n[\mathcal{K}_Y(\hat{\gamma}/n) - (\hat{\gamma}/n)x])/\sqrt{2\pi\mathcal{K}''_Y(\hat{\gamma}/n)}$. The leading factor $\sqrt{n}$ arises from $\sqrt{\mathcal{K}''_X(\hat{\gamma}/n)}$. If $\hat{\beta} = \hat{\gamma}/n$, then the usual saddlepoint density approximation is obtained:

$$g_X(x) = \sqrt{n}\exp(n[\mathcal{K}_Y(\hat{\beta}) - \hat{\beta}x])/\sqrt{2\pi\mathcal{K}''_Y(\hat{\beta})}. \tag{68}$$

The saddlepoint $\hat{\beta}$ is given by the solution to

$$\mathcal{K}'_Y(\hat{\beta}) = x. \tag{69}$$

For the Edgeworth series the natural scaling for sums of random variables was the standardized sum giving constant variance as n increases. For saddlepoint approximations the natural scaling is the mean, since it gives an expression for $\hat{\beta}$ independent of n, and because the term in the exponent in (68) is n times a factor independent of n. The saddlepoint approximation may be expressed in terms of the Legendre transform $\mathcal{K}^*_Y$ of $\mathcal{K}_Y$: For x in the range of $\mathcal{K}'_Y$ define $\mathcal{K}^*_Y(x) = \hat{\beta}x - \mathcal{K}_Y(\hat{\beta})$ for $\hat{\beta}$ satisfying (69). Then $g_X(x) = \sqrt{n}\exp(-n\mathcal{K}^*_Y(x))/\sqrt{2\pi\mathcal{K}''_Y(\hat{\beta})}$. Since $(d/dx)\mathcal{K}^*_Y(x) = \hat{\beta}(x)$ and $d\hat{\beta}/dx = 1/\mathcal{K}''_Y(\hat{\beta})$, then

$$(d^2/dx^2)\mathcal{K}^*_Y(x) = 1/\mathcal{K}''_Y(\hat{\beta}) > 0$$

for all x. Hence $\mathcal{K}^*_Y$ strictly concave as a function of x. These relations will be useful when assessing the error in the renormalized saddlepoint density approximation.

Since $\rho_{X,j} = n^{1-j}\mathcal{K}_Y^{(j)}(\hat{\beta})/(n^{-1}\mathcal{K}_Y^{(2)}(\hat{\beta}))^{j/2}$, the counterpart of (67) in terms of $\mathcal{K}_Y$ is:

$$f_X(x) = \sqrt{n}\frac{\exp(n[\mathcal{K}_Y(\hat{\beta}) - \hat{\beta}x])}{\sqrt{2\pi\mathcal{K}''_Y(\hat{\beta})}}[1 + b(\beta)/[2n] + O(n^{-2})], \tag{70}$$

for

$$b(\beta) = \hat{\rho}_{Y,4}/4 - 5\hat{\rho}^2_{Y,3}/12. \tag{71}$$

Here

$$\hat{\rho}_{Y,j} = \mathcal{K}_Y^{(j)}(\beta)/(\mathcal{K}_Y^{(2)}(\beta))^{j/2}. \tag{72}$$

This justifies the following result:

Theorem 4.1.1: *Suppose that f_Y is a density function for a random variable Y, with characteristic function $\mathcal{K}_Y$ defined in an open interval $\mathcal{I}$ containing zero, and one desires an approximation to the density f_X of X, the mean of n independent and identically distributed copies of Y. Suppose that equation (69) has a solution in $\mathcal{I}$. Then the expansion (70) represents a valid asymptotic expression for f_X. Furthermore, general higher-order terms in (70) are generated using an easy algorithm:*

1. *The saddlepoint $\hat{\beta}$ is determined.*
2. *The cumulant generating function $n\mathcal{K}_Y(\beta/(\sqrt{n\mathcal{K}''_Y(\hat{\beta})}))$ is expanded about $\hat{\beta}$.*
3. *The quadratic term is removed.*
4. *The resulting series is exponentiated, to form a bivariate series for*

$$\exp\left(n\mathcal{K}_Y\left(\beta/(n\sqrt{\mathcal{K}''_Y(\hat{\beta})})\right) - (\beta - \hat{\beta})^2/2\right)$$

in β about $\hat{\beta}$, and in $1/\sqrt{n}$ about 0.
5. *Powers of $\beta - \hat{\beta}$ are replaced by the corresponding Hermite polynomial evaluated at 0.*

Proof: As this is merely the result of the Edgeworth series theorem for density functions, Theorem 3.5.1, applied to the density proportional to $\exp(\beta x) f_Y(x)$, it suffices to verify the conditions of Theorem 3.5.1. Moments of all orders for this distribution exist, by the arguments of §2.1. Furthermore, since the density f_Y exists, by Lemma 2.4.4, the integrability condition is also fulfilled.

Q.E.D

Here the cumulants of the standardized distribution are varying, and the point at which the Hermite polynomials are evaluated is fixed. Hence uniform asymptotic properties do not follow from the Edgeworth theorem already proved, since the distributions that the Edgeworth series approximation theorems are applied to vary with both x and n. We can say that the series is valid asymptotically as an approximation for the density of X at x as $n \to \infty$ and x changes so as to keep the saddlepoint fixed. In the case of an approximation to the mean of independent and identically distributed random variables, this corresponds to fixing the value of the mean as n varies. Contrast this with the Edgeworth series approximation, in which the value of $n^{-1/2}\sum_j Y_j$ is held fixed.

Later this chapter conditions will be given under which far more can be said about asymptotic bounds on the error for approximations of the form (66), or extensions of the form (70). Heuristically this works because under lax assumptions the standardized cumulants are bounded as $\hat{\beta}$ varies. Hence good relative error behavior rather than merely absolute error behavior can be expected. Recall that for the Edgeworth series, as the ordinate varied over $\mathfrak{R}$, the point at which Hermite polynomials were evaluated also varied over $\mathfrak{R}$, inducing poor relative error properties. Kolassa (1991) shows that an approximate solution to the saddlepoint equation (69), with consequent modifications to (70), provides a compromise between Edgeworth and saddlepoint methodologies, in that good relative error properties are preserved while possibly intractable equations are avoided.

Blackwell and Hodges (1959), without referencing Daniels (1954), note that when Y_j are supported on the integers, and on no coarser lattice, then (70) also represents

$$n\mathrm{P}\left[X = x\right](1 + O(1/n^2)) \tag{73}$$

for $X = \sum_{j=1}^{n} Y_j/n$. Proof of this fact will deferred until §4.5.

Jing and Robinson (1994) note that if $Z = g(X)$, and g is monotonic and differentiable, then the density f_Z of Z evaluated at z may be approximated by

$$\sqrt{n}\frac{\exp(n[\mathcal{K}_Y(\hat{\beta}) - \hat{\beta}g^{-1}(z)])}{g'(g^{-1}(z))\sqrt{2\pi\mathcal{K}_Y''(\hat{\beta})}}[1 + b(\beta)/[2n] + O(n^{-2})], \tag{74}$$

and (69) is evaluated at $x = g^{-1}(z)$.

4.2. Examples

The following three examples are cited by Daniels (1954): Consider the saddlepoint density approximation to means of independent and identically distributed random variables, each with a normal distribution with mean μ and unit variance. The cumulant generating function for each summand is $\mathcal{K}_Y(\beta) = \mu\beta + \beta^2/2$, and hence the saddlepoint satisfies $\hat{\beta} + \mu = x$. Then $\hat{\beta} = x - \mu$. Then approximation is $g_X(x) = \exp(-n(x - \mu)^2/2)/n\sqrt{2\pi}$. In this case, the approximation is exactly the correct distribution.

As a further example, consider the gamma distribution. Recall that a $\Gamma(\lambda, \alpha)$ random variable has density

$$f_X(x) = \lambda^{\alpha}x^{\alpha-1}\exp(-\lambda x)/\Gamma(\alpha)$$

and the cumulant generating function $\mathcal{K}_X(\beta) = \alpha\log(\lambda) - \alpha\log(\lambda - \beta)$. The saddlepoint $\hat{\beta}$ satisfies $x = -\alpha/(\lambda - \beta)$, and hence $\hat{\beta} = \lambda - \alpha/x$. Then $\mathcal{K}_X''(\hat{\beta}) = \alpha/(\lambda-\hat{\beta})^2 = x^2/\alpha$, $\mathcal{K}_X'''(\hat{\beta}) = 2\alpha/(\lambda-\hat{\beta})^3 = 2x^3/\alpha^2$, $\mathcal{K}_X''''(\hat{\beta}) = 6\alpha/(\lambda-\hat{\beta})^4 = 6x^4/\alpha^3$, and $b(\hat{\beta}) = \frac{1}{4}(6x^4/\alpha^3)\alpha^2/x^4 - 5(4x^6/\alpha^4)(\alpha^3/x^6)/12 = -\frac{1}{6}\alpha^{-1}$.

$$f_X(x) = \sqrt{n}\frac{\exp(n(\alpha\log(\lambda) - \alpha\log(\lambda - \lambda + \alpha/x) - x\lambda + \alpha))}{x\sqrt{2\pi/\alpha}}$$

$$= \sqrt{n}\lambda^{n\alpha}\alpha^{-n\alpha}x^{n\alpha-1}\exp(n\alpha - x\lambda n)\sqrt{\alpha/2\pi}[1 - \frac{1}{12}\alpha^{-1}]$$

$$= (n\lambda)^{n\alpha}x^{n\alpha-1}\exp(n\alpha - nx\lambda)(n\alpha)^{-n\alpha+1/2}/\sqrt{2\pi}[1 - \frac{1}{12}\alpha^{-1}]$$

From the cumulant generating function one finds that the convolution of n densities of the form $\Gamma(\alpha, \lambda)$ is a $\Gamma(n\alpha, \lambda)$ density, and hence/ the mean has a $\Gamma(n\alpha, n\lambda)$ distribution. The saddlepoint approximation agrees with the exact density except that the gamma function is replaced by its Sterling approximation $\Gamma(n\alpha) \approx \exp(-n\alpha)(n\alpha)^{n\alpha-1/2}\sqrt{2\pi}$.

As a final example, the moment generating function of the uniform distribution on $(-1/2, 1/2)$ is

$$\mathcal{M}_U(\beta) = (\exp(\beta/2) - \exp(-\beta/2))/\beta = \sinh(\beta/2)/(\beta/2),$$

and the cumulant generating function $\mathcal{K}_U$ is the logarithm of $\mathcal{M}_U$. Then

$$\mathcal{K}'_U(\beta) = (\exp(\beta/2) + \exp(-\beta/2))/(\exp(\beta/2) - \exp(-\beta/2)) - 1/\beta$$

and $\mathcal{K}''_U(\beta) = 1/\beta^2 - 4/(\exp(\beta/2) - \exp(-\beta/2))^2$. The solution to the saddlepoint equation must either be approximated, or obtained numerically. Here the latter approach is taken. Results show that far in the tails saddlepoint outperforms both the normal and Edgeworth-4 approximations. See Fig. 8.

Errors in Approximations to Densities for the Mean of 6 Uniform $[-1, 1]$.
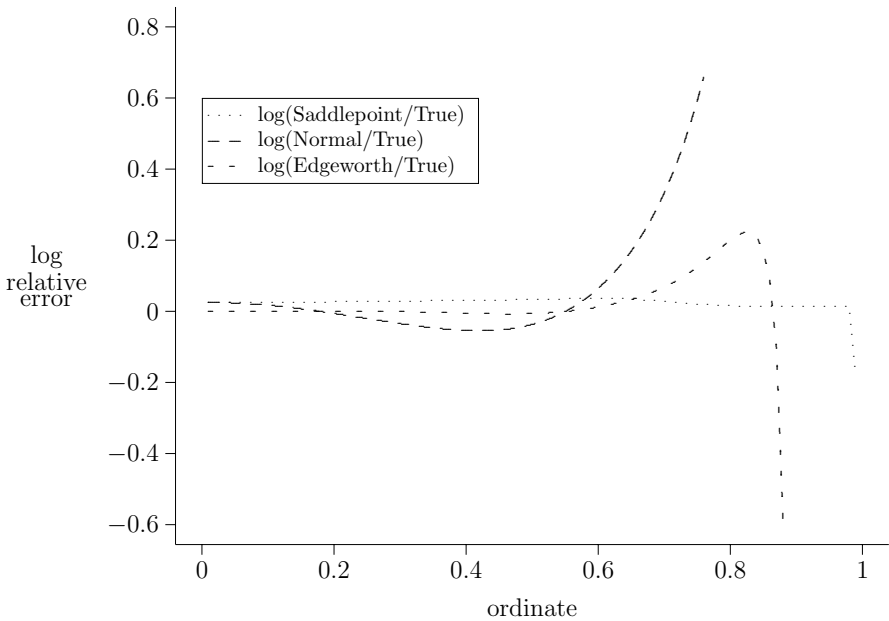
Fig. 8.

4.3. Conditions for the Existence of the Saddlepoint

The preceding discussion presupposes the existence of $\hat{\beta}$ solving (69). Daniels (1954) investigates under what circumstances such a $\hat{\beta}$ exists. These conditions will depend on the form of the interval $\mathfrak{Q}$ on which the cumulant generating function $\mathcal{K}_Y$ is defined, and also on the form of the interval of support for F_Y. Express $\mathfrak{Q}$ as an interval with end points $-c_1 \le c_2$ where c_1 and c_2 are non-negative. Saddlepoint approximation methods are only considered when these endpoints are not both zero. Express the interval of support for the distribution F_Y as $[-x_1, x_2]$, where values of

∞ for x_1 or x_2 or both are allowed. Then a solution of the saddlepoint equation exists for all values of $x \in (-x_1, x_2)$ if

$$\lim_{\beta \to -c_1} \mathcal{K}'_Y(\beta) = -x_1, \quad \lim_{\beta \to c_2} \mathcal{K}'_Y(\beta) = x_2, \tag{75}$$

since $\mathcal{K}'_Y$ is continuous, and $\mathcal{K}''_Y > 0$. Hence $\mathcal{K}'_Y$ is strictly increasing. Define the conditions $C_i : \lim_{\beta \to (-1)^i c_i} \mathcal{K}'_Y(\beta) = (-1)^i x_i$. When x_i are infinite, condition $C_1 \cap C_2$ is known as steepness (Barndorff–Nielsen, 1978). Define the conditions C'_i as $c_i = \infty$. Note that C'_i is implied by $x_i < \infty$. As we will see below, the condition $x_i < \infty$ forces $c_i = \infty$; however, the following lemma and theorem will be stated in terms of the slightly more immediate conditions. The following lemma holds.

Lemma 4.3.1: *Condition C'_i implies C_i for $i = 1, 2$.*

Proof: This lemma is proved for $i = 2$; the case when $i = 1$ is similar. Consider the two cases.

1. x_2 is finite. Note that

$$x_2 - \mathcal{K}'_Y(\beta) = \int_{-x_1}^{x_2} (x_2 - y) \exp(\beta y) \, dF(y) \Big/ \int_{-x_1}^{x_2} \exp(\beta y) \, dF(y) \geq 0.$$

   Showing that $\lim_{\beta \to \infty} [x_2 - \mathcal{K}'_Y(\beta)] \leq 0$ implies C_2 in this case. For any $z < x_2$,

$$\begin{aligned}
x_2 - \mathcal{K}'_Y(\beta) &\leq \frac{(x_2 - z) \int_z^{x_2} \exp(\beta(y-z)) \, dF(y) + \int_{-x_1}^{z} (x_2 - y) \exp(\beta(y-z)) \, dF(y)}{\int_z^{x_2} \exp(\beta(y-z)) dF(y) + \int_{-x_1}^{z} \exp(\beta(y-z)) dF(y)} \\
&= \frac{(x_2 - z) + \int_{-x_1}^{z} (x_2 - y) \exp(\beta(y-z)) \, dF(y) / \int_z^{x_2} \exp(\beta(y-z)) \, dF(y)}{1 + \int_{-x_1}^{z} \exp(\beta(y-z)) dF(y) / \int_z^{x_2} \exp(\beta(y-z)) \, dF(y)} \\
&\to x_2 - z
\end{aligned}$$

   Hence $\lim_{\beta \to \infty} \mathcal{K}'_Y(\beta) = x_2$.
2. $x_2 = \infty$. Then $\mathcal{K}'_Y(\beta) = \int_{-x_1}^{\infty} y \exp(\beta y) \, dF(y)$. Set $a^* = \int_0^{\infty} y^2 \, dF(y) / \int_0^{\infty} y \, dF(y)$. Then

$$\mathcal{K}'_Y(\beta) \geq \int_{-x_1}^{0} y \exp(\beta y) \, dF(y) + \int_0^{\infty} y \, dF(y) \exp(\beta a^*).$$

   The limit of the first term is zero as $\beta \to \infty$, and the second term diverges to ∞.

<div align="right">Q.E.D</div>

This lemma leads immediately to the following theorem:

Theorem 4.3.2:   *If X is a random variable supported on the possibly infinite interval $[-x_1, x_2]$, whose cumulant generating function is defined on an interval with end points $-c_1$ and c_2, and if for $i = 1$ and 2, $x_i < \infty$ or $c_i = \infty$, then a solution of the saddlepoint equation exists for all values of $x \in (-x_1, x_2)$.*

Proof: Conditions C'_1 and C'_2 implies (75), by Lemma 4.3.1.

Q.E.D

Daniels (1954) presents an example in which these conditions fail.

In this section we have considered only solutions to the saddlepoint approxima-
tion for $x \in (-x_1, x_2)$. If $x = -x_1$ or $x = x_2$ one may take the density approximation
to be 0, since x sits at the edge of the support of the distribution. When consider-
ing cumulative distribution function approximations this choice will be even easier,
since by definition the cumulative distribution function evaluated at x will be 0 if
$x = -x_1$ and 1 if $x = x_2$. When considering double saddlepoint approximations to
conditional densities and distribution functions, other cases with conditioning ran-
dom variables on the edges of their ranges of definition will be more troublesome.
This question will be addressed later.

4.4. Construction of the Steepest Descent Curve

The saddlepoint density approximations (66) and (70) for a random variable X
with characteristic function $\zeta_X(\gamma)$ and cumulant generating function $\mathcal{K}_X(\gamma)$ may
be derived by using the Fourier inversion formula and approximating the integral
involved. Recall from (19) that

$$f_X(x) = \frac{1}{2\pi} \int_{-i\infty}^{+i\infty} \zeta_X(\gamma) \exp(-i\gamma x) \, d\gamma = \frac{1}{2\pi i} \int_{-i\infty}^{+i\infty} \exp(\mathcal{K}_X(\gamma) - \gamma x) \, d\gamma. \quad (76)$$

The variable of integration was changed from β to γ, since later in this section
the scale of this variable of integration will be changed, and the variable β will be
reserved for this changed scale to conform more closely to standard notation. This
expression will be manipulated to yield (70). First, denote the real domain of $\mathcal{K}_X$ by
$\mathcal{I}_X$. Then within the set $\mathcal{I}_X \times i\Re$ the integrand above is differentiable, and the path
of integration can be chosen to be any path from $-i\infty$ to $+i\infty$, by the closed curve
theorem of complex variables (Bak and Newman (1996), §8.1). That is, the path
of integration need not be along the complex axis, but along any path diverting to
positive and negative infinity along the complex axis, but having perhaps non-zero
real components.

The next step is to show that the path of integration in (76) can be replaced by
a path along which the real value of the variable of integration is fixed at any value
in the range of definition of the cumulant generating function; that is, that

$$\lim_{T \to \infty} \int_{-iT}^{iT} \exp(\mathcal{K}_X(\gamma) - \gamma x) \, d\gamma = \lim_{T \to \infty} \int_{c-iT}^{c+iT} \exp(\mathcal{K}_X(\gamma) - \gamma x) \, d\gamma \quad (77)$$

for all $c \in \mathcal{I}_X$. By the closed curve theorem, all that is required is to show that

$$\lim_{T \to \infty} \int_{iT}^{c+iT} \exp(\mathcal{K}_X(\gamma) - \gamma x) \, d\gamma = 0 \text{ and } \lim_{T \to \infty} \int_{-iT}^{c-iT} \exp(\mathcal{K}_X(\gamma) - \gamma x) \, d\gamma = 0. \quad (78)$$

At each point γ along the two paths of integration parallel to the real axis, the
integrand is the characteristic function of a random variable with density $f_X(y-x) \times$
$\exp(-\mathcal{K}_X(\Re(\gamma)) + \Re(\gamma)(y-x))$, evaluated at $\Im(\gamma)$. Here y is a dummy argument for

the density, and $\Im(\gamma) = \pm T$. By the Riemann–Lebesgue Theorem, Theorem 3.8.1 of §3.8, this is bounded by $|T|^{-1}$. Hence the integrands of (78) converge uniformly to zero, and so the limits in (78) hold.

Use of steepest descent methods requires that (78) be manipulated so that the exponent of the resulting integrand is approximately quadratic near $\hat{\beta}$. Calculations begin with an expression of the density inversion integral (19) in terms of the cumulant generating function for one summand in the standardized mean of independent and identically distributed components:

$$
\begin{aligned}
f_X(x) &= \frac{1}{2\pi i} \int_{-i\infty}^{+i\infty} \exp(\mathcal{K}_X(\gamma) - \gamma x) \; d\gamma \\
&= \frac{n}{2\pi i} \exp(n[\mathcal{K}_Y(\hat{\beta}) - \hat{\beta}x]) \int_{-i\infty}^{+i\infty} \exp(n[\mathcal{K}_Y(\beta) - \beta x - \mathcal{K}_Y(\hat{\beta}) + \hat{\beta}x]) \; d\beta. \quad (79)
\end{aligned}
$$

The second equality above follows from the change of variables $\gamma = \beta n$. Relation (77) ensures that (79) can be replaced by

$$
f_X(x) = \frac{n \exp(n[\mathcal{K}_Y(\hat{\beta}) - \hat{\beta}x])}{2\pi i} \int_{\hat{\beta}-i\infty}^{\hat{\beta}+i\infty} \exp(n[\mathcal{K}_Y(\beta) - \beta x - \mathcal{K}_Y(\hat{\beta}) + \hat{\beta}x]) \; d\beta. \quad (80)
$$

Since the integrand is a characteristic function evaluated at $\Im(\beta)$, when this integration is performed along the path parallel to the complex axis, the result recovered is the Edgeworth series times the leading factor, yielding (70).

The preceding heuristic development of the saddlepoint series implies that a better path of integration runs through the previously defined saddlepoint $\hat{\beta}$. Relation (77) shows that this integration may be performed along the path $\Re(\beta) = \hat{\beta}$. As will be shown later in this section, the integrand in (78) along this path is the characteristic function of the tilted distribution used in the heuristic development of the saddlepoint density approximation in §4.1, and the result will be (70) as before. This derivation of the saddlepoint series using the shifted vertical path of integration will give the sharpest results, and will be crucial when considering uniformity in the behavior of approximation errors. Never the less, one might ask whether local changes in the path of integration might produce sharper results. To answer this question by showing that the heuristically optimal results are (70), to exhibit the saddlepoint development as introduced by Daniels (1954), and to set the stage for further developments in which these methods will be very useful, this section and the next present the development of saddlepoint series using the method of steepest descent.

This approach can be motivated from a purely analytic point of view, by recourse to saddlepoint integration techniques (Mathews and Walker, 1964). Pick $\hat{\beta}$ along the real axis minimizing $\mathcal{K}(\beta) - \beta x$, and pick the path through $\hat{\beta}$ along which the real part of $\mathcal{K}(\beta) - \beta x$ decreases the fastest as one moves away from the real axis, at least near $\hat{\beta}$. This path is called the path of steepest descent. As will be explained below, along this path $\mathcal{K}(\beta) - \beta x$ is real. See Fig. 9. The steepest descent curve in

this figure is denoted by the most finely dotted line. The surface of the real part of $\mathcal{K}$ above the complex plane then has paths through $\hat{\beta}$ along which it rises in either direction, and paths along which it falls in either direction. The surface has the shape of a saddle, giving the name saddlepoint to $\hat{\beta}$. Using the same ideas as found in the proof of the asymptotic validity of Edgeworth series, as n increases, the contribution to the integral of any bit outside of an increasingly small circle about the saddlepoint vanishes exponentially, and the remainder is determined by the derivatives of $\mathcal{K}$ at $\hat{\beta}$.

Path of Integration on β scale, x= 2
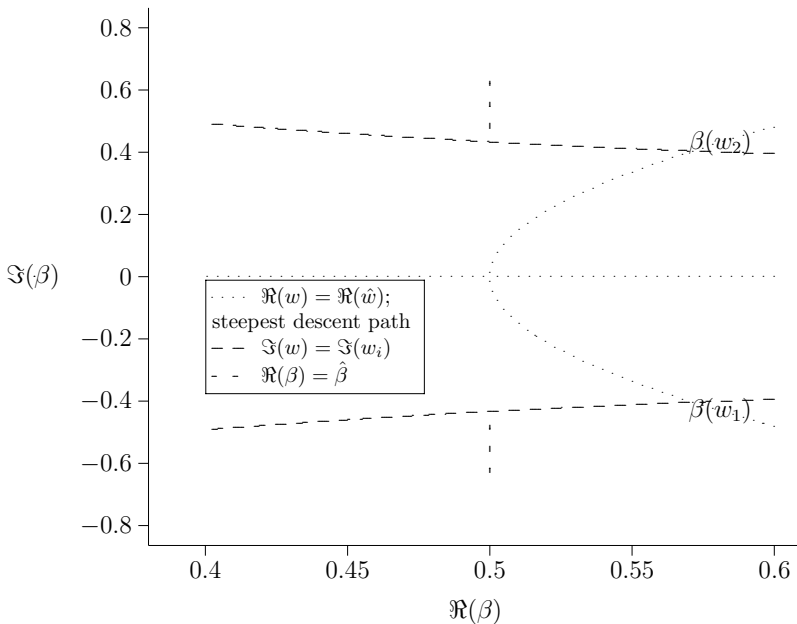
Fig. 9.

Steepest descent methods involve integration along paths through $\hat{\beta}$ along which the real part decreases fastest near $\hat{\beta}$. The first objective is to determine whether such paths exist. The quantity multiplying n in the exponent of (80),

$$h(\beta) = [\mathcal{K}_Y(\beta) - \beta x - \mathcal{K}_Y(\hat{\beta}) + \hat{\beta}x],$$

is an analytic function that is zero and has zero first derivative at $\hat{\beta}$, by the choice of $\hat{\beta}$. Specifically, a function $\omega : \mathfrak{R} \to \mathfrak{C}$ must be constructed that is continuous and piecewise differentiable, such that $\lim_{a \to -\infty} \omega(a) = \hat{\beta} - i\infty$, $\lim_{a \to +\infty} \omega(a) = \hat{\beta} + i\infty$,

and $h(\omega(a))$ falls off as quickly per distance traveled along ω as along any other path through $\hat{\beta}$, in a neighborhood of $\hat{\beta}$. The line integral in (80) is then defined by splitting up this path into its differentiable parts, using the change of variables theorem to express the integral along each part as the sum of two standard one-dimensional integrals, representing the real and complex parts of the transformed integrand

$$\exp(n[\mathcal{K}_Y(\beta(a)) - \beta(a)x - \mathcal{K}_Y(\hat{\beta}) + \hat{\beta}x])(d\beta/da) \ da,$$

performing the separate integrations, and summing the parts. The second objective is to characterize this path.

Since $h(\beta)$ is analytic, it has a convergent power series representation $h(\beta) = \sum_{k=0}^{\infty} c_k(\beta - \hat{\beta})^k$ for $\left|\beta - \hat{\beta}\right| < \delta$ where $\delta > 0$. By construction $c_0 = c_1 = 0$, and c_2 is real and positive. Then $h(\beta)/(\beta - \hat{\beta})^2$ is analytic for $\left|\beta - \hat{\beta}\right| < \delta$, and its value at $\hat{\beta}$ is c_2. By the composition of functions theorem from complex analysis, $\sqrt{2h(\beta)/(\beta - \hat{\beta})^2}$ is analytic for $\left|\beta - \hat{\beta}\right| < \delta_1$, where $\delta_1 > 0$, and has the value $\sqrt{2c_2} > 0$ at $\hat{\beta}$. An alternate square root function with the opposite sign could also be defined. Let

$$\omega(\beta) = \hat{\omega} + (\beta - \hat{\beta})\sqrt{2[\mathcal{K}_Y(\beta) - \beta x - \mathcal{K}_Y(\hat{\beta}) + \hat{\beta}x]/(\beta - \hat{\beta})^2}, \qquad (81)$$

where

$$\hat{\omega} = \sqrt{2(\hat{\beta}x - \mathcal{K}_Y(\hat{\beta}))}, \qquad (82)$$

with the sign the same as that of $\hat{\beta}$. Then $\omega(\beta)$ is an analytic function of β for $\left|\beta - \hat{\beta}\right| < \delta_1$, takes the value $\hat{\omega}$ at $\hat{\beta}$, and has first derivative $\sqrt{2c_2} > 0$ at $\hat{\beta}$. By the inverse function theorem for complex variables, there exist $\delta_2 > 0$ and a function $\beta(\omega)$ analytic for $|\omega - \hat{\omega}| < \delta_2$, such that $\beta(\omega)$ is the inverse of $\omega(\beta)$ in (81).

Through this ball $|\omega - \hat{\omega}| < \delta_2$ runs the path $\Re(\beta(\omega)) = \hat{\beta}$. Integrate instead along the line $\Re(\omega) = \hat{\omega}$, the path of steepest descent. Again, recall that this is the same path along which $\mathcal{K}_Y(\beta) - \beta x - \mathcal{K}_Y(\hat{\beta}) + \hat{\beta}x$ is real and negative. Within $|\omega - \hat{\omega}| < \delta_2$ will be constructed three curves: The line of steepest descent, along $\Re(\omega) = \hat{\omega}$, and two segments along which the imaginary part of ω is fixed, leading back to the path $\Re(\beta(\omega)) = \hat{\beta}$. Construct these as follows: Let ω_1 and ω_2 be the two places that the path $\Re(\beta(\omega)) = \hat{\beta}$ crosses the circle $|\omega - \hat{\omega}| = \delta_2$, where $\Im(\omega_1) < 0$, $\Im(\omega_2) > 0$. The three paths are then

1. $\Im(\omega) = \Im(\omega_1)$ from $\Re(\omega) = \Re(\omega_1)$ to $\Re(\omega) = \Re(\hat{\omega})$.
2. $\Re(\omega) = \Re(\hat{\omega})$ from $\Im(\omega) = \Im(\omega_1)$ to $\Im(\omega) = \Im(\omega_2)$.
3. $\Im(\omega) = \Im(\omega_2)$ from $\Re(\omega) = \Re(\hat{\omega})$ to $\Re(\omega) = \Re(\omega_2)$.

This reparameterization $\beta(\omega)$ will make finding the path of steepest descent easy. Since $\frac{1}{2}(\omega(\beta) - \hat{\omega})^2 = \mathcal{K}_Y(\beta) - \beta x - \mathcal{K}_Y(\hat{\beta}) + \hat{\beta}x$, the path along which the exponent of (80) decreases most quickly, at least very close to $\hat{\beta}$, is the path along

which $\omega - \hat{\omega}$ is pure imaginary. Let $\omega(a) = \beta(\hat{\omega} + ia)$ for $\Im(\omega_1) \le a \le \Im(\omega_2)$, representing path 2 above. Along this path ω, when $|a| < \delta_2$, the exponent of (80) is real and negative, since it is the square of a pure imaginary quantity. This gives a second characterization of the path of steepest descent. Complete this path with segments 1 and 3 from above, and segments from $\beta(\omega_2)$ to $\hat{\beta} + i\infty$ and from $\hat{\beta} + i\infty$ to $\beta(\omega_1)$, along which $\Re(\beta) = \hat{\beta}$. See Fig. 9.

For the purposes so far, the choice of $\hat{\omega}$ was irrelevant. Choosing $\hat{\omega} = 0$ would have been as satisfactory. The particular choice of $\hat{\omega}$ will only become critical in later sections when saddlepoint expansions for the cumulative distribution function are constructed. At that point the the fact that the current choice of $\hat{\omega}$ implies not only $\omega(\hat{\beta}) = \hat{\omega}$, but also $\omega(0) = 0$, will become important.

Having determined the path of steepest descent, parameterizing (80) in terms of this path yields

$$f_X(x) = \frac{n}{2\pi i} \exp(n[\mathcal{K}_Y(\hat{\beta}) - \hat{\beta}x]) \int_{\hat{\omega}-i\infty}^{\hat{\omega}+i\infty} \exp(\frac{n}{2}[\omega - \hat{\omega}]^2) \frac{d\beta}{d\omega} d\omega,$$

plus an error arising from points outside of the circle of radius δ_2; this error converges to zero geometrically quickly.

4.5. Rigorous Derivation using Steepest Descent

Daniels (1954) uses a special case of Watson's Lemma, a theorem from complex variables (Jeffreys, 1962), in conjunction with the method of steepest descent of the previous section, to show that the order of the Edgeworth series carries over but only pointwise for each value of the mean.

Lemma 4.5.1: *If $\varpi(\omega)$ satisfies $\varpi(\omega) = \sum_{j=0}^{2j} a_j \omega^j / j! + O(\omega^{2j+1})$ as $\omega \to 0$, and A and B are positive numbers, then*

$$\left(\frac{n}{2\pi}\right)^{\frac{1}{2}} \int_{-A}^{B} \exp(-\frac{n}{2}\omega^2) \varpi(\omega) \, d\omega = \sum_{j=0}^{j} \varpi^{(2j)}(0) / [(2n)^j j!] + O(n^{-j-1/2}), \qquad (83)$$

provided the integrand on the left hand side of (83) converges absolutely for some n. Specifically, there exist constants D_j and N_{j+1} independent of n such that for $n > N_{j+1}$,

$$\left| \left(\frac{n}{2\pi}\right)^{\frac{1}{2}} \int_{-A}^{B} \exp(-\frac{n}{2}\omega^2) \varpi(\omega) \, d\omega - \sum_{j=0}^{j} \varpi^{(2j)}(0) / [(2n)^j j!] \right| \le D_j / n^{j+1/2}.$$

Proof: Let $E_{j+1}(\omega) = \varpi(\omega) - \sum_{j=0}^{2j} a_j \omega^j / j!$. Choose $a < A$, $b < B$, and D_j such that on $(-a, b)$, $|E_{j+1}(\omega)| \le \frac{1}{2} D_j \omega^{2j+1}$. Choose N_{j+1} such that $\frac{1}{2} D_j n^{-j-1/2} \ge$

$\exp(-\max(a,b)n)\varpi^{(j)}(0)$, if $n > N_{j+1}$. Then

$$\left(\frac{n}{2\pi}\right)^{\frac{1}{2}} \int_{-A}^{B} \exp(-\frac{n}{2}\omega^2)\varpi(\omega) \ d\omega = \sum_{j=0}^{2j} \left(\frac{n}{2\pi}\right)^{\frac{1}{2}} \int_{-\infty}^{\infty} \exp(-\frac{n}{2}\omega^2)\frac{\varpi^{(j)}(0)}{j!}\omega^j \ d\omega -$$

$$\sum_{j=0}^{2j} \left(\frac{n}{2\pi}\right)^{\frac{1}{2}} \int_{-\infty}^{-A} \exp(-\frac{n}{2}\omega^2)\frac{\varpi^{(j)}(0)}{j!}\omega^j \ d\omega - \sum_{j=0}^{2j} \left(\frac{n}{2\pi}\right)^{\frac{1}{2}} \int_{B}^{\infty} \exp(-\frac{n}{2}\omega^2)\frac{\varpi^{(j)}(0)}{j!}\omega^j \ d\omega$$

$$+ \left(\frac{n}{2\pi}\right)^{\frac{1}{2}} \int_{-A}^{-a} \exp(-\frac{n}{2}\omega^2)\varpi(\omega) \ d\omega + \left(\frac{n}{2\pi}\right)^{\frac{1}{2}} \int_{b}^{B} \exp(-\frac{n}{2}\omega^2)\varpi(\omega) \ d\omega$$

$$+ \left(\frac{n}{2\pi}\right)^{\frac{1}{2}} \int_{-a}^{b} \exp(-\frac{n}{2}\omega^2)E_{j+1}(\omega) \ d\omega. \qquad (84)$$

The proof is completed by showing that

1. Even terms in the first summation of (84) are the corresponding terms given in the statement of the theorem.
2. Odd terms in the first summation of (84) are zero.
3. The final term in (84) is bounded by $D_j n^{-j-1/2} \int_{-\infty}^{\infty} \exp(-\omega^2/2) |\omega|^{2j+1} \ d\omega$.
4. All other terms of (84) converge to zero geometrically, at rate $\exp(-\max(a,b)n)$.

Q.E.D

A rescaled version of Lemma 4.5.1 of more immediate use is:

Lemma 4.5.2: *If $\varpi(\omega)$ is analytic in a neighborhood of $\omega = \hat{\omega}$ containing the path $(-Ai + \hat{\omega}, Bi + \hat{\omega})$ with $\infty \geq A > 0$ and $\infty \geq B > 0$, then*

$$i^{-1}\left(\frac{n}{2\pi}\right)^{\frac{1}{2}} \int_{-Ai+\hat{\omega}}^{Bi+\hat{\omega}} \exp(\frac{n}{2}(\omega-\hat{\omega})^2)\varpi(\omega) \ d\omega = \sum_{j=0}^{\infty} \frac{(-1)^j \varpi^{(2j)}(\hat{\omega})}{(2n)^j j!}$$

is an asymptotic expansion in powers of n^{-1}, provided the integral converges absolutely for some n; that is,

$$\left|\left(\frac{n}{2\pi}\right)^{\frac{1}{2}} \int_{-Ai+\hat{\omega}}^{Bi+\hat{\omega}} \exp(\frac{n}{2}(\omega-\hat{\omega})^2)\varpi(\omega) \ d\omega - \sum_{j=0}^{j}(-1)^j\frac{\varpi^{(2j)}(\hat{\omega})}{(2n)^j j!}\right| \leq D_j/n^{j+1}.$$

Proof: This lemma follows immediately from Lemma 4.5.1 and the usual change of variables theorem in integration. Since terms in the expansion of all orders in n exist, the remainder term is of the same order as the last term omitted.

Q.E.D

The rescaled Watson's Lemma will be applied to the center portion of the integral in (80), after changing the path of integration as in §4.4 and parameterizing using (81). This reparameterization introduces the factor $\varpi_x(\omega) = d\beta/d\omega$, which by the arguments of §4.4 is analytic at $\hat{\omega}$. One can prove a result that is stronger than that of Theorem 4.1.1 using Watson's lemma; the following result is stronger

because it addresses uniformity properties of the approximation as the ordinate x changes, it includes cases when the distribution of X is not continuous for small n more cleanly, and it yields the formula of Blackwell and Hodges (1959) directly.

Theorem 4.5.3:   *Suppose that:*
1. *The random variables $Y_1, \ldots, Y_n$ are independent and identically distributed, each with cumulant generating function $\mathcal{K}_Y$ defined on an open interval containing 0, and each with cumulative distribution function F_Y.*
2. *An ordinate x is chosen such that the equation $\mathcal{K}_Y'(\hat{\beta}) = x$ has a solution.*
3. *There exists $N \in \mathbf{3}$ such that the characteristic function for the tilted density*

$$\zeta_N(s) = \exp(N[\mathcal{K}_Y(\beta + is)x - \mathcal{K}_Y(\beta)])$$

is absolutely integrable for each real $\beta \in [0, \hat{\beta}]$: $\int_{-\infty}^{\infty} |\zeta_N(s)| \ ds < \infty$.
Then the mean $X = (Y_1 + \cdots + Y_n)/n$ has a density f_X if $n \geq N$, and the result (70) is a valid asymptotic expansion for $f_X(x)$, in that the error incurred by keeping only the first j terms can be bounded by a quantity of the form $D_{\mathsf{j}}/n^{\mathsf{j}+1}$. Furthermore, if (3) is replaced
3' *Y_j is supported on an integer lattice,*
then (70) is a valid asymptotic expansion for $n\mathrm{P}\,[X = x]$.

Proof: As indicated above, f_X is given by the inversion integral (80). This is integrated along the five-part path described in §4.4.

Recall that the third part is the integral along the path of steepest descent:

$$\sqrt{\frac{n}{2\pi}} \frac{\exp(n[\mathcal{K}_Y(\hat{\beta}) - \hat{\beta}x])}{\sqrt{\mathcal{K}_Y''(\hat{\beta})}} i^{-1} \sqrt{\frac{n}{2\pi}} \int_{\hat{\omega}-i\delta_2}^{\hat{\omega}+i\delta_2} \exp(\frac{n}{2}(\omega - \hat{\omega})) \frac{\sqrt{\mathcal{K}_Y''(\hat{\beta})}d\beta}{d\omega} d\omega.$$

Lemma 4.5.2 is applied to this integral. Verification that

$$\frac{d\beta}{d\omega} = \frac{\left[1 - \frac{1}{3}\hat{\rho}_{Y,3}(\omega - \hat{\omega}) + (\frac{5}{24}\hat{\rho}_{Y,3}^2 - \frac{1}{8}\rho_{Y,4})(\omega - \hat{\omega})^2 + O((\omega - \hat{\omega})^3)\right]}{\sqrt{\mathcal{K}''(\hat{\beta})}} \tag{85}$$

and hence that the approximation agrees with (70), will be outlined in the exercises.

The second and fourth parts of the paths of integration contribute a pure imaginary portion. Since the complete resulting integral is real, imaginary contributions must cancel out and can be ignored.

The first and fifth integrals, along rays to $\hat{\beta} \pm i\infty$, are treated exactly as in the Edgeworth series proof for densities, part b, in §3.7. Along the line $\Re(\beta) = \hat{\beta}$ the integrand of (80) is the characteristic function of the sum of n random variables whose probability distribution is given by

$$dF_Y^*(y) = [\exp(\hat{\beta}y) \ dF_Y(y)]/\int_{-\infty}^{\infty} \exp(\hat{\beta}z) \ dF_Y(z),$$

evaluated at $\Im(\beta)$. By condition 3, the limit of the characteristic function for this distribution as the argument diverges to $\pm\infty$ is zero. Hence outside of a fixed interval about the origin the absolute value of the characteristic function is bounded away from one, and the integral over the first and fifth parts of the path is bounded by the least upper bound of the individual characteristic functions over this region, which is less than 1, to the power $n - N$, times a convergent integral. Hence the contribution converges to zero geometrically.

The result for the lattice case is derived by noting that the inversion integral for lattice probabilities, (32), is identical to that for densities, (19), except for the factor of Δ, and for the different range of integration. Because the saddlepoint result is obtained by truncating the range of integration in (19) to a small open set containing the real line, then the same arguments justify the expansion of Blackwell and Hodges (1959).

<div align="right">Q.E.D</div>

4.6. Conditions for Uniformly Bounded Relative Error

The preceding steepest descent methods imply that the resulting series are, for a given ordinate, the best approximate cumulant generating function inversion approximations available that require only derivatives of the cumulant generating function at one point. Since the resulting error bounds involve only the invariants of the summands. one might conjecture that suitable regularity conditions on their behavior imply uniform relative error bounds of the asymptotic expansion. This conjecture is true. Unfortunately steepest descent methods do not provide an easy way to establish this uniformity. Two difficulties arise:

1. The error involved in truncating the power series expansion for $d\beta/d\omega$ near the origin was governed by the behavior of higher-order derivatives of $d\beta/d\omega$. Bounding these coefficients away from the real axis using only behavior along this axis is difficult.

2. The radius of convergence for $d\beta/d\omega$ determined the geometric rate at which parts of the inversion integral away from $\hat{\beta}$ converged to zero. This radius is not constant, and in fact for extreme $\hat{\beta}$ may converge to zero. Consider the logistic distribution. The tilted cumulant generating function $\mathcal{K}(r + is) - (r + is)x$ is, after simplification,

$$i \arcsin\left(\frac{\sqrt{2}\,(s\,\cosh(\pi\,s)\,\sin(\pi\,r) - r\,\cos(\pi\,r)\,\sinh(\pi\,s))}{\sqrt{(r^2 + s^2)\,(\cosh(2\,\pi\,s) - \cos(2\,\pi\,r))}}\right) - \frac{3}{2}\log(2)$$

$$-\log(\cosh(2\pi s) - \cos(2\pi r)) + \tfrac{1}{2}\log\left(\left(r^2 + s^2\right)(\cosh(2\pi s) - \cos(2\pi r))\right) - rx - isx,$$

and at $r = 1$ the imaginary part is

$$-s\,x + \arcsin\left((\sqrt{2}\,\sinh(\pi\,s))/(\sqrt{(1 + s^2)\,(\cosh(2\,\pi\,s) - 1)})\right).$$

The value of s making this imaginary part zero can be plotted as a function of x (Fig. 10). The length of the portion of the path of integration to which arguments involving the derivatives of the integrand apply is shrinking to zero. Fig. 11 illustrates this in the case of the logistic distribution.

Height where Steepest Descent Curve for Logistic Density Hits Boundary vs. x
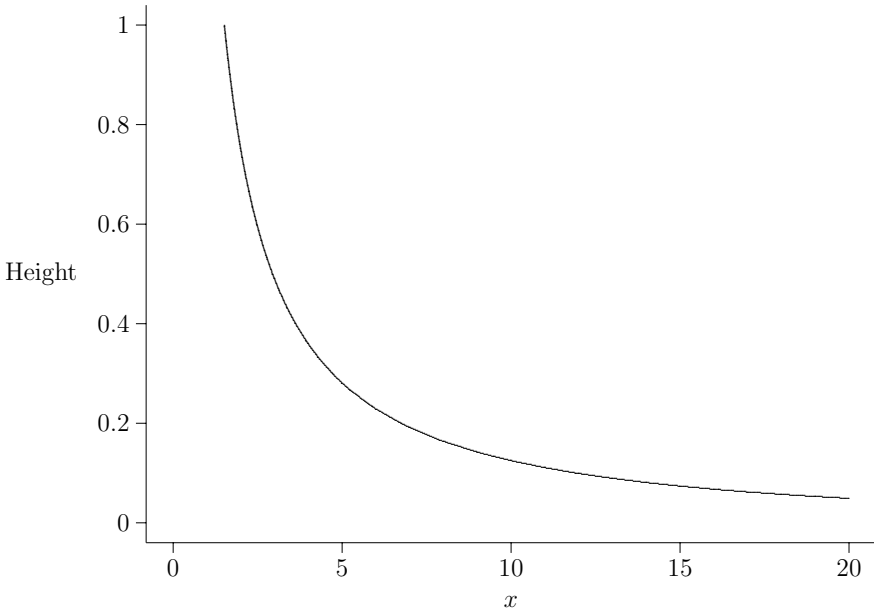

Fig. 10.

A different tactic must be employed demonstrating that the series (70) has desired uniformity properties.

Theorem 4.6.1: *Assume the conditions of Theorem 4.5.3. Suppose further that:*

4. *The standardized absolute moments* $\mathrm{E}\left[|Y_1|^j \exp(Y_1\hat\beta)\right]/\mathrm{E}\left[|Y_1|^2 \exp(Y_1\hat\beta)\right]^{j/2}$ *are bounded for each j as $\hat\beta$ varies.*

5. *For some integer $\mathfrak{r}$, $\int_{-i\infty}^{i\infty}\left|\zeta_{\hat\beta}(\gamma)\right|^{\mathfrak{r}} d\gamma \le c_1 < \infty$, for all $\hat\beta$ in the domain of the cumulant generating function, where*

$$\zeta_{\hat\beta}(\gamma) = \exp(\mathcal{K}_Y(\hat\beta + i\gamma/\sigma(\hat\beta)) - \mathcal{K}_Y(\hat\beta) - ix\gamma/\sigma(\hat\beta))$$

is the characteristic function for each tilted distribution.

6. *$\zeta_{\hat\beta}(\gamma)$ converge uniformly to some function as $\hat\beta$ goes to its extreme value,*

Steepest Descent Curve for Logistic Distribution
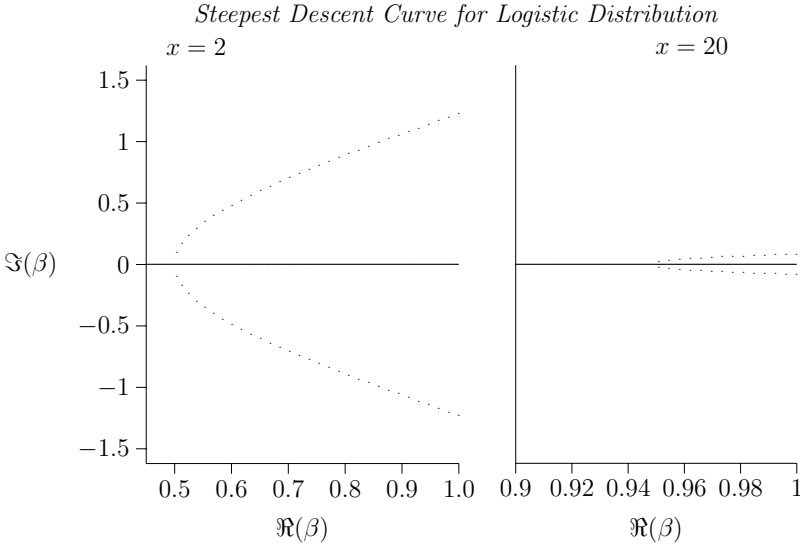

Fig. 11a.   Fig. 11b.

Then the bound on the error in the asymptotic series (70) can be chosen independently of x and $\hat{\beta}$, for all x for which $\hat{\beta}$ is defined.

Proof: Condition 4 in a sense refines the prior condition 2, in that attention is restricted to ordinates for which $\hat{\beta}$ exists, and condition 5 extends the prior condition 3. The present condition 4 is slightly stronger than that considered by Daniels (1954), and implies boundedness of the invariants. The result presented is due to Jensen (1988).

Refer to the Edgeworth series density result, Theorem 3.5.1, as proved in §3.7. Condition 1 of Theorem 4.5.3 shows that Corollary 2.5.4, with the cumulants replaced by the invariance, can be applied for $\mathfrak{g} = \mathfrak{j} + 1$ and $\epsilon = 1$, yielding C from (27) uniformly bounded. The result follows if we can demonstrate that δ may of §3.7 may also be chosen independently of x and $\hat{\beta}$. Consequently, all quantities involved in bounding the portion of the error integral near the origin also can be taken independent of x and $\hat{\beta}$.

Condition 2 shows that away from the real axis, all of the characteristic functions being approximately inverted converge to 0, and condition 3 shows that this must happen uniformly. Outside of $(-\delta\sqrt{n}, \delta\sqrt{n})$ the characteristic function to be inverted can be bounded by $q^{n-\mathfrak{r}}$ times something integrable, and whose integral is bounded, by condition 2. The constant $q < 1$ is independent of x and $\hat{\beta}$, and the result follows.

Q.E.D

Uniformity in (70) refers to the accuracy of the series multiplying the first-

order saddlepoint density approximation. Hence this uniformity refers not to the absolute error of each partial sum in (70), but also to the relative error of the first-order approximation. If $\sqrt{\mathcal{K}''_Y(\beta)}$ remains bounded below, absolute uniformity is recovered, although this theorem demonstrates no improvement over the Edgeworth series result, when considering analytically proven uniform behavior. If $\sqrt{\mathcal{K}''_Y(\beta)}$ goes to infinity then in fact the resulting uniform behavior of the absolute error is better than that for Edgeworth series. Bear in mind that this is all from an analytic point of view; in practice the saddlepoint approximations behave much better than the Edgeworth series.

4.7. Cases with Uniformly Bounded Relative Error

Daniels (1954) lists forms of summand densities such that the relative error of the saddlepoint approximation to the distribution of means of independent and identically distributed copies is uniformly bounded. These are characterized by how the tails behave as the ordinate goes to limiting values. Consider a distribution defined on $(-x_1, x_2)$ and examine behavior as the ordinate approaches x_2. In the first three examples $x_2 = \infty$.

1. Gamma: $f(x) \sim A x^{\alpha-1} \exp(-cx)$ where $\alpha > 0$, $c > 0$.
2. $f(x) \sim A \exp(\psi x^\alpha - cx)$; $\psi > 0$, $c > 0$, $0 < \alpha < 1$. Note Daniels' paper reads $0 < x < 1$.
3. Stable laws: $f(x) \sim A \exp(-\psi x^\alpha)$; $\psi > 0$, $\alpha > 1$.
4. Beta: $f(x) \sim A (x_2 - x)^{\alpha-1}$ on $(-x_1, x_2)$; $\alpha > 0$.
5. $f(x) \sim A \exp(-\psi(x_2 - x)^{-\gamma})$ on $(-x_1, x_2)$, $\psi > 0$, $\gamma > 0$.

Jensen (1988) presents conditions on density functions that imply the conditions of the theorem of the previous section. These are applied to the cases presented by Daniels.

4.8. The Effect of Renormalization

The quality of the saddlepoint approximation can often be improved by multiplying the density approximation by a constant to make it integrate to 1. Recall from §4.2 that the saddlepoint approximation to a normal density is exact, and the saddlepoint approximation to a gamma density is exact up to a constant of proportionality. There are very few densities with this property, as the following theorem of Blæsild and Jensen (1985) shows.

Theorem 4.8.1: *Suppose that Y_j are independent and identically distributed random variables, and suppose that the saddlepoint approximation to the density of the mean of n such variables is exactly proportional to the true density, for every $n \in \mathfrak{Z}$. Then the distribution of the Y_j is either gamma, normal, or inverse Gaussian.*

Proof: Since §4.2 contains verification that the normal and gamma densities are approximated exactly up to a multiplicative constant, it remains only to prove that

the inverse Gaussian also shares this property, and that no other densities share this property. Verifying the exactness of the approximation to the inverse Gaussian is left as an exercise.

Barndorff–Nielsen and Klüppelberg (1999) consider the multivariate case of this exactness phenomenon.

From (70), the relative error incurred when using the saddlepoint approximation is

$$
\frac{3\hat{\rho}_{Y,4} - 5\hat{\rho}_{Y,3}^2}{24n} + \frac{385\hat{\rho}_{Y,3}^4 - 630\hat{\rho}_{Y,3}^2\hat{\rho}_{Y,4} + 105\hat{\rho}_{Y,4}^2 + 168\hat{\rho}_{Y,3}\hat{\rho}_{Y,5} - 24\hat{\rho}_{Y,6}}{1152n^2} + \cdots, \quad (86)
$$

where $\hat{\rho}_{Y,l}$ are given by (72) and depend on $\hat{\beta}$. For this error to be constant, each term in its series (86) must be constant as a function of $\hat{\beta}$. This theorem will be proved by examining the first two terms in (86), setting them equal to a constant, treating them like a system of differential equations for $\mathcal{K}_Y$, and showing that the only solution is the cumulant generating function associated with a normal, gamma, or inverse Gaussian distribution.

If the first term is zero, $\hat{\rho}_{Y,4}/8 - 5\hat{\rho}_{Y,3}^2/24 = 0$ and

$$
3\mathcal{K}_Y^{(4)}(\hat{\beta}) = 5\mathcal{K}_Y^{(3)}(\hat{\beta})^2/\mathcal{K}_Y^{(2)}(\hat{\beta}). \quad (87)
$$

This represents a second-order non-linear differential equation for $\mathcal{K}_Y''(\hat{\beta})$. We seek solutions for which $\mathcal{K}_Y''$ has a power series representation near zero, and for which $\mathcal{K}_Y''(0) > 0$. Expressing $\mathcal{K}_Y''(\hat{\beta})$ as a power series, and substituting into (87) yields a recursion for the power series coefficients of $\mathcal{K}_Y''$ in which higher-order coefficients are completely determined by the first two. The function $\mathcal{K}_Y^{*}{}''(\beta) = (a\beta + b)^{-3/2}$ solves (87), for all $b > 0$. Since $\mathcal{K}_Y^{*}{}''(0) = b^{-3/2}$ and $\mathcal{K}_Y^{*}{}'''(0) = -(3/2)ab^{-5/2}$, the set of possible cumulant generating functions satisfying (87) is contained in the set generated from $\mathcal{K}_Y^{*}$ with a and b varied. Hence solutions of the form $(a\beta + b)^{-3/2}$ for all $b > 0$ exhaust the class of solutions of interest. Showing that these solutions represent cumulant generating functions of normal and inverse Gaussian distributions is left as an exercise.

If the first term is constant but not zero, (87) no longer holds. An alternate equation is derived by examining the conditions of constancy for the coefficients of n^{-1} and n^{-2} in (86), and their derivatives. The following rule for differentiating the $\hat{\rho}_{Y,l}$ is useful:

$$
\begin{aligned}
d\hat{\rho}_{Y,l}/d\hat{\beta} &= \mathcal{K}_Y^{(l)}(\hat{\beta})(-l/2)\mathcal{K}_Y^{(2)}(\hat{\beta})^{-(l+2)/2}\mathcal{K}_Y^{(3)}(\hat{\beta}) + \mathcal{K}_Y^{(l+1)}(\hat{\beta})\mathcal{K}_Y^{(2)}(\hat{\beta})^{-l/2} \\
&= (-l/2)\sqrt{\mathcal{K}_Y''(\hat{\beta})}\hat{\rho}_{Y,l}\hat{\rho}_{Y,3} + \sqrt{\mathcal{K}_Y''(\hat{\beta})}\hat{\rho}_{Y,l+1}. \quad (88)
\end{aligned}
$$

Setting the first two derivatives of $\hat{\rho}_{Y,4}/8 - 5\hat{\rho}_{Y,3}^2/24$ to 0, setting $(\hat{\rho}_{Y,4}/8 - 5\hat{\rho}_{Y,3}^2/24)^2$ to a constant, and $385\hat{\rho}_{Y,3}^4 - 630\hat{\rho}_{Y,3}^2\hat{\rho}_{Y,4} + 105\hat{\rho}_{Y,4}^2 + 168\hat{\rho}_{Y,3}\hat{\rho}_{Y,5} - 24\hat{\rho}_{Y,6}/1152$ to a

different constant, results in the equations

$$15\,\hat{\rho}_{Y,3}^3 - 16\hat{\rho}_{Y,3}\hat{\rho}_{Y,4} + 3\hat{\rho}_{Y,5} = 0$$

$$\frac{-135\,\hat{\rho}_{Y,3}^4}{2} + 101\,\hat{\rho}_{Y,3}^2\hat{\rho}_{Y,4} - 16\,\hat{\rho}_{Y,4}^2 - \frac{47\,\hat{\rho}_{Y,3}\,\hat{\rho}_{Y,5}}{2} + 3\,\hat{\rho}_{Y,6} = 0$$

$$25\hat{\rho}_{Y,3}^4 - 30\hat{\rho}_{Y,3}^2\hat{\rho}_{Y,4} + 9\hat{\rho}_{Y,4}^2 = c_2^2$$

$$385\,\hat{\rho}_{Y,3}^4 - 630\,\hat{\rho}_{Y,3}^2\,\hat{\rho}_{Y,4} + 105\,\hat{\rho}_{Y,4}^2 + 168\,\hat{\rho}_{Y,3}\,\hat{\rho}_{Y,5} - 24\,\hat{\rho}_{Y,6} = c_1.$$

Multiplying these by $20\,\hat{\rho}_{Y,3}/3$, 8, 23/9, and 1, and summing them yields

$$\frac{16\,\hat{\rho}_{Y,3}^2\left(5\,\hat{\rho}_{Y,3}^2 - 3\hat{\rho}_{Y,4}\right)}{9} = \frac{16\,\hat{\rho}_{Y,3}^2\,c_2}{9}$$

on the left hand side, and a constant on the right. Hence in this case, $\hat{\rho}_{Y,3}$ is constant, and $\mathcal{K}_Y''$ is determined by the first order differential equation $\mathcal{K}_Y^{(3)}(\hat{\beta}) = c\mathcal{K}_Y''(\hat{\beta})^{3/2}$ for some constant c. A solution to this equation is $\mathcal{K}_Y^{*}{}''(\hat{\beta}) = (-\frac{1}{2}c\hat{\beta}+b)^{-2}$ for $b > 0$. For the same reasons as in the first case solutions of this form include all solutions of interest. Showing that these solutions represent cumulant generating functions of normal and gamma distributions is left as an exercise.

Q.E.D

Near exactness up to a multiplicative constant is fairly common for saddlepoint density approximations. Such near exactness holds if the coefficients in (70) vary slowly enough. Recovering the constant multiplier and normalizing has potential for improving the behavior of the saddlepoint approximation. This constant may be derived by numerically integrating the approximation, and Durbin (1980) shows that in some cases the order of the error is divided by $\sqrt{n}$.

Theorem 4.8.2: *Suppose that the conditions of Theorem 4.6.1 hold. Then the saddlepoint density approximation (68) may be rescaled so that its relative error is $O(n^{-3/2})$ as the value of the sum standardized to constant variance is held fixed. This rescaling may be calculated by integrating (68) and dividing by the result.*

Proof: Without loss of generality assume that $\mathrm{E}\,[Y] = 0$. From (70),

$$f_n(x) = \sqrt{n}\frac{\exp(n[\mathcal{K}_Y(\hat{\beta}) - \hat{\beta}x])}{\sqrt{2\pi\mathcal{K}_Y''(\hat{\beta})}}[1 + b(\hat{\beta})/(2n) + O(n^{-2})].$$

Here c is a sum of products of invariants; hence its derivative is the sum of products of invariants and $\sqrt{\mathcal{K}_Y''(\hat{\beta})}$, by (88). The derivative of $b(\hat{\beta})$ with respect to x is $db/dx = b'(\hat{\beta})/\sqrt{\mathcal{K}_Y''(\hat{\beta})}$, which is bounded on a set of the form $(-\delta, \delta)$ for some $\delta > 0$. By the strict concavity of the Legendre transform, $\mathcal{K}_Y(\hat{\beta}) - \hat{\beta}x$ is bounded

away from zero for x outside of $(-\delta, \delta)$. Using Taylor's theorem, for $x^\dagger \in [0, x]$,

$$f_n(x) = \sqrt{n}\frac{\exp(n[\mathcal{K}_Y(\hat{\beta}) - \hat{\beta}x])}{\sqrt{2\pi\mathcal{K}_Y''(\hat{\beta})}}[1 + b(0)/(2n) + y(db/dx)(x^\dagger)/(2n\sqrt{n}) + O(n^{-2})]$$

(89)

where $y = x\sqrt{n}$. For any fixed y, then $x \in (-\delta, \delta)$ for sufficiently large n. Hence multiplying (68) by $1 + b(0)/(2n)$ gives the standardization resulting in a relative error of size $O(n^{-3/2})$. Furthermore, since $\sqrt{n}\bar{X}$ has a limit in distribution, the limit of

$$\int_{-\infty}^{\infty} \sqrt{n}\frac{\exp(n[\mathcal{K}_Y(\hat{\beta}) - \hat{\beta}x])}{\sqrt{2\pi\mathcal{K}_Y''(\hat{\beta})}}(\sqrt{n}x)(db/dx)(x^\dagger)\ dx$$

is finite. Integrating both sides of (89),

$$1 = (1 + b(0)/(2n))\int_{-\infty}^{\infty} \sqrt{n}\frac{\exp(n[\mathcal{K}_Y(\hat{\beta}) - \hat{\beta}x])}{\sqrt{2\pi\mathcal{K}_Y''(\hat{\beta})}}\ dx(1 + O(n^{-3/2})).$$

Hence $\int_{-\infty}^{\infty} \sqrt{n}\exp(n[\mathcal{K}_Y(\hat{\beta}) - \hat{\beta}x])/\sqrt{2\pi\mathcal{K}_Y''(\hat{\beta})}\ dx = (1 + b(0)/(2n))(1 + O(n^{-3/2}))$.

Q.E.D

4.9. Numeric Hints for Solving Saddlepoint Equations

In many applications the saddlepoint equation (69) cannot be solved analytically, even when the solution $\hat{\beta}$ exists. In these cases, saddlepoint methods can usually still be applied by solving (69) numerically.

Often one uses Newton–Raphson derivative-based methods to calculate the saddlepoint. These methods in general will work well, since the function $\mathcal{K}_Y(\beta) - \beta x$ to be minimized is convex. The second derivative exists and is always positive. An iterative procedure is then used to calculate the saddlepoint. An initial approximation β_0 is chosen. In general this will not exactly solve $\mathcal{K}_Y'(\hat{\beta}) = x$, and so the initial approximation must be modified. Approximating the saddlepoint equation linearly in the neighborhood of β_0, the saddlepoint $\hat{\beta}$ satisfies $\mathcal{K}_Y'(\beta_0) + \mathcal{K}_Y''(\beta_0)(\hat{\beta} - \beta_0) \approx x$. The solution to this equation is

$$\hat{\beta} = (x - \mathcal{K}_Y'(\beta_0))/\mathcal{K}_Y''(\beta_0) + \beta_0. \tag{90}$$

The solution to this equation is then used as an updated approximate solution, and (90) is iterated until the approximation is considered close enough to the true solution to (69). This convergence may be assessed by referring to the difference $|x - \mathcal{K}_Y'(\beta_0)|$, or by referring to changes in subsequent values of β_0. The former method is preferred, because in cases where $\mathcal{K}_Y$ is defined only on a finite range, and diverges to infinity as its argument approaches the end of this range, $\mathcal{K}_Y(\hat{\beta})$ and $\mathcal{K}_Y''(\hat{\beta})$ are far more sensitive to small errors in $\hat{\beta}$ near the ends of the range than in the middle. These quantities $\mathcal{K}_Y(\hat{\beta})$ and $\mathcal{K}_Y''(\hat{\beta})$ are needed to calculate the

saddlepoint density approximation. Hence a policy of terminating iterations of (90) when changes in $\hat{\beta}$ become small on an absolute scale is problematic, since it is likely to be unnecessarily fine for moderate values of $\hat{\beta}$ and too coarse for extreme values of $\hat{\beta}$.

The Newton–Raphson method is not guaranteed to yield a solution $\hat{\beta}$. The method depends on the cumulant generating function being approximately quadratic, at least in a region containing the true and approximate saddlepoint. If the cumulant generating function is exactly quadratic one iteration of (90) will result in an exact solution to the saddlepoint equation. If the second derivative of the cumulant generating function decreases as one moves from β_0 to $\hat{\beta}$, iterations of (90) will tend to result in steps toward $\hat{\beta}$ that are too small. In this case the iterative saddlepoint approximations typically move monotonically toward $\hat{\beta}$. If the second derivative of the cumulant generating function increases as one moves from β_0 to $\hat{\beta}$, iterations of (90) will tend to result in steps toward $\hat{\beta}$ that overshoot $\hat{\beta}$. In this case the iterative saddlepoint approximations typically first cross to the other side of $\hat{\beta}$, and then move monotonically toward $\hat{\beta}$. If the second derivative of $\mathcal{K}_Y$ is lower on either side of the saddlepoint than it is at the saddlepoint, successive saddlepoint approximations may move from one side of $\hat{\beta}$ to the other while never converging.

Since the Newton–Raphson method performs best when the cumulant generating function is approximately quadratic, rescaling the cumulant generating function to make it closer to quadratic speeds convergence. Especially to be avoided are cases in which the cumulant generating function exists only on a finite interval, and in which Newton's method may give iterations falling outside this domain of definition. In this case rescaling the cumulant generating function argument often avoids such problems.

For example, consider the logistic distribution. Exercise 7 of §2.8 gave this cumulant generating function as $\mathcal{K}_Y(\beta) = \log(\pi\beta/\sin(\pi\beta))$. Its derivatives are

$$\mathcal{K}_Y'(\beta) = 1/\beta - \pi\cot(\pi\beta)$$
$$\mathcal{K}_Y''(\beta) = -1/\beta^2 + \pi^2\csc^2(\pi\beta)$$

for $\beta \in (-1, 1)$. Solving the saddlepoint equation in this case by iterating (90) gives rise to two problems. First, note that the values at zero are $\mathcal{K}_Y'(0) = 0$ and $\mathcal{K}_Y''(0) = \pi^2/3$. Hence for $x > \pi^2/3$, when starting from $\beta_0 = 0$, one iteration of (90) gives a second saddlepoint approximation outside the domain of $\mathcal{K}_Y$. Second, supposing that for a large x a more reasonable starting value β_0 is found. Since $\mathcal{K}_Y$ is far from quadratic, the Newton–Raphson method can be expected to converge very slowly.

Solving instead the equation $\mathcal{K}_Y'(2\tan^{-1}(\hat{s})/\pi) = x$ using the Newton–Raphson method, and setting $\hat{\beta} = 2\tan^{-1}(\hat{s})/\pi$, avoids both of these problems. The new quantity to be solved for can take any value on the real line, so no problem of

provisional approximations lying outside the domain of the cumulant generating function arises. Furthermore, no poles near the end points slow the convergence. See Fig. 12.

Alternatively, one might use the secant method to solve for the saddlepoint. Begin with two initial values, β_0 and β_1, are determined such that $\mathcal{K}'_Y(\beta_0) < x < \mathcal{K}'_Y(\beta_1)$. If $\mathcal{K}'(\beta)$ were linear, the saddlepoint would then be given by

$$\beta_2 = (x[\beta_1 - \beta_0] + \beta_0\mathcal{K}'_Y(\beta_1) - \beta_1\mathcal{K}'_Y(\beta_0))/(\mathcal{K}'_Y(\beta_1) - \mathcal{K}'_Y(\beta_0)).$$

Graphically, one determines $\hat{\beta}$ by observing where the secant joining $(\beta_0, \mathcal{K}'_Y(\beta_0))$ to $(\beta_1, \mathcal{K}'_Y(\beta_1))$ crosses the horizontal line through x. One then evaluates $\mathcal{K}'_Y(\beta_2)$ and determines whether $\mathcal{K}'_Y(\beta_2) > x$, implying β_2 replaces β_1, or whether $\mathcal{K}'_Y(\beta_2) < x$, implying β_2 replaces β_0. This process continues until β_0 and β_1 are close enough to consider the common value the saddlepoint.

This secant method generally only produces the correct answer in one iteration if $\mathcal{K}'_Y$ is linear, as is the case with the Newton–Raphson method. Like the Newton–Raphson method, if $\mathcal{K}''_Y$ varies quickly enough, convergence can be very very slow, although divergence is not possible for the secant method.

4.10. Exercises

1. Complete the proof of Watson's Lemma by following the steps:
   a. Prove the first and second assertions in the proof of Watson's Lemma in §4.5, by using integration by parts, or other methods, to show that $\int_{-\infty}^{\infty} \exp(-\omega^2/2)\omega^k \, d\omega = \int_{-\infty}^{\infty} \exp(-\omega^2/2)(k-1)\omega^{k-2} \, d\omega$ for any integer k, and hence for any even integer k, $\int_{-\infty}^{\infty} \exp(-\omega^2/2)\omega^k \, d\omega = k!/((k/2)!)2^{-k/2}$.
   b. Prove the third assertion in the proof of Watson's Lemma in §4.5.
   c. Prove the fourth assertion in the proof of Watson's Lemma in §4.5.
2. Show that the expansion resulting from applying the modified version of Watson's Lemma to $d\beta/d\omega$ gives the results (70) by following the steps:
   1. Show that if a function $f(\beta)$ has a series expansion $1 + \sum_{j=1}^{\infty} a_j\beta^j$, then $\sqrt{f(\beta)}$ has the expansion

   $$1 + \tfrac{1}{2}a_1\beta^1 + (\tfrac{1}{2}a_2 - \tfrac{1}{8}a_1^2)\beta^2 + (\tfrac{1}{2}a_3 - \tfrac{1}{4}a_1a_2 + \frac{1}{16}a_1^3)\beta^3 + \cdots.$$

   2. Use the preceding result to show that if a function $f(\beta)$ has a series expansion $\sum_{j=2}^{\infty} a_j\beta^j$, then $\sqrt{f(\beta)}$ has the expansion

   $$\sqrt{a_2}(\beta + \tfrac{1}{2}a_3\beta^2 + (\tfrac{1}{2}a_4 - \tfrac{1}{8}a_3^2)\beta^3 + (\tfrac{1}{2}a_5 - \tfrac{1}{4}a_3a_4 + \frac{1}{16}a_3^3)\beta^4 + \cdots. \quad (91)$$

   3. Show that if a function $f(\beta)$ has a series expansion $\sum_{j=1}^{\infty} a_j\beta^j$, and $f(\beta) = s$, then
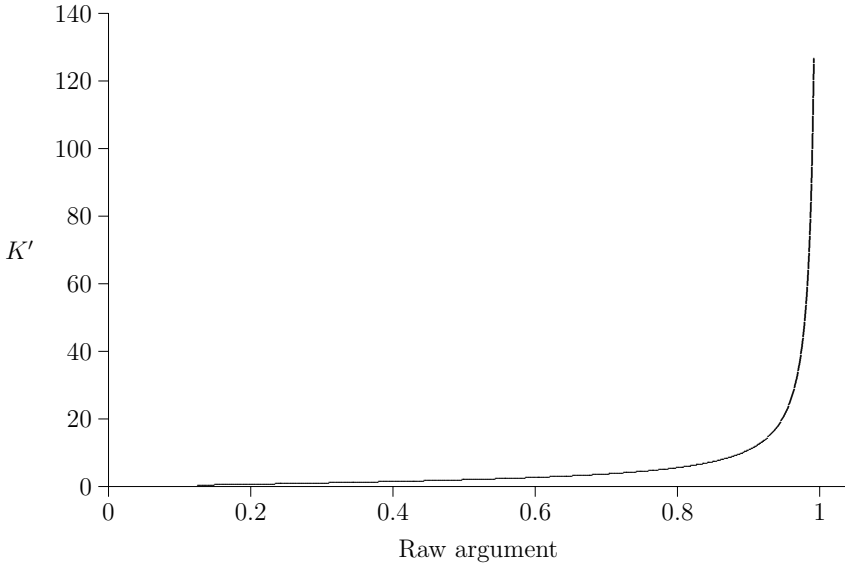 $$\beta = a_1^{-1}s - a_1^{-3}a_2s^2 + a_1^{-5}(2a_2^2 - a_1a_3)s^3 + \cdots. \quad (92)$$

Derivative of Logistic Cumulant Generating Function

Fig. 12a.
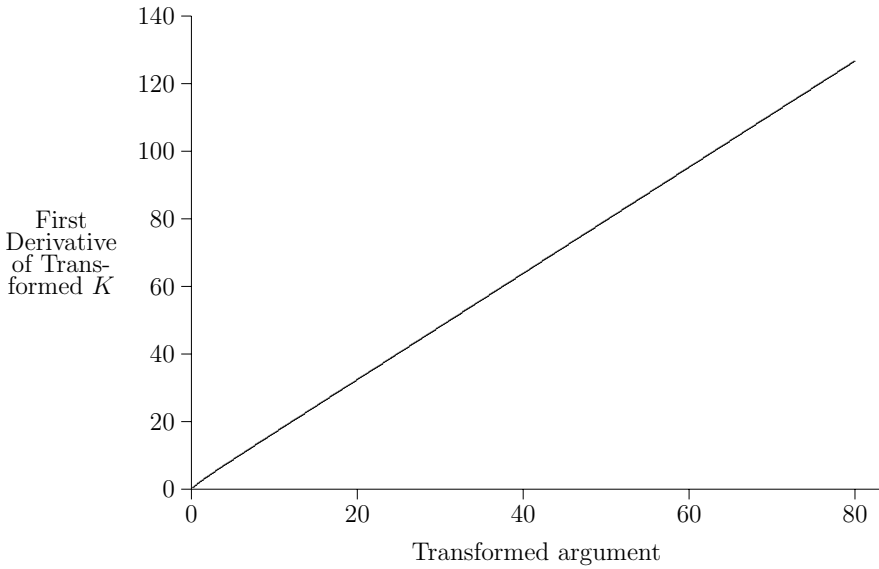
Fig. 12b.

4. Use the result (91) to get a series expansion for ω in terms of β and derivatives of $\mathcal{K}_Y(\hat\beta)$, and use (92) to give an expansion for β in terms of ω. Take a derivative of this to get the expansion for $d\beta/d\omega$.

3. Complete the proof of the theorem of §4.8 using the following steps.

   1. Show that normal and inverse Gaussian distributions have cumulant generating functions satisfying $\mathcal{K}''(\beta) = (a\beta + b)^{-3/2}$ for $b > 0$.
   2. Show that normal and gamma distributions have cumulant generating functions satisfying $\mathcal{K}''(\beta) = (-\frac{1}{2}c\hat\beta + b)^{-2}$ for $b > 0$.
   3. Show that the saddlepoint approximation to the inverse Gaussian distribution is exact.

5

Saddlepoint Series for Distribution Functions

Recall from §3 that calculating distribution function approximations from Edgeworth density approximations was a simple matter. The Edgeworth series for the density is a linear combination of derivatives of the normal distribution function, and hence is easily integrated to give a corresponding cumulative distribution function approximation. This cumulative distribution function approximation inherits many good properties from the density approximation.

Throughout this chapter, assume that the random variable X results as the mean of n independent and identically distributed random variables, each of which has a cumulant generating function $\mathcal{K}_Y(\beta)$ defined in an open interval containing zero. Hence X also has a cumulant generating function $\mathcal{K}_X(\beta)$. Until noted to the contrary, assume that X has a density f_X.

Integrals involving saddlepoint density approximations are in general not tractable analytically. One might approximate the cumulative distribution function by integrating the saddlepoint density (68) from $-\infty$ to x numerically. Many techniques for numeric integration exist (Thisted, 1988). Most of these techniques involve calculating the integrand at a large number of points $x_0, \ldots, x_j$ between $-\infty$ and x, evaluating the integrand at each of these points to obtain $y_0, \ldots, y_j$, and estimating the integral as a linear combination of $y_0, \ldots, y_j$. Each evaluation of the saddlepoint density may require an iterative solution to (69). Daniels (1987) notes that this numeric integration of the saddlepoint density can be simplified by changing variables from x to $\hat{\beta}$:

$$
\begin{aligned}
Q_X(x) &= \mathrm{P}\left[X \geq x\right] \\
&= \int_x^{\max(X)} f_X(z)\, dz &\qquad (93) \\
&= \int_x^{\max(X)} \exp(\mathcal{K}_X(\hat{\beta}(z)) - z\hat{\beta}(z))/\sqrt{2\pi\mathcal{K}_X''(\hat{\beta}(z))}\, dz + O(n^{-1}) &\qquad (94) \\
&= \int_{\hat{\beta}(x)}^{\max(\hat{\beta})} \exp(\mathcal{K}_X(\beta) - \beta\mathcal{K}_X'(\beta))\sqrt{\mathcal{K}_X''(\beta)/2\pi}\, d\beta + O(n^{-1})
\end{aligned}
$$

to avoid having to calculate $\hat{\beta}$ for each value x_i, as would be required to numerically evaluate (94). Although quite good results can be obtained in this way, this method has the inelegant drawback of embodying both numeric and analytic approximation. The rest of this chapter will be concerned with alternative analytic methods for deriving saddlepoint cumulative distribution function approximations. No strong heuristic parallels exist to guide us here. Principle results are derived directly from the appropriate inversion integrals. This task is complicated by a pole at 0 in the integrand. Approximations are derived ignoring and accounting for this pole, and are compared. Finally, extensions of these techniques to lattice distributions are presented. Reid (1996) provides a recent thorough review of these questions.

Rather than approximating the cumulative distribution function of a random variable directly, we begin by calculating tail areas for the random variable. This is done because some of the saddlepoint cumulative distribution function approximations degenerate for ordinates near the mean, as will be shown in the next sections.

Many analytic approximations to tail areas begin with integration of the exact expression for the density (19). This results in a double integral, with respect to the dummy ordinate representing potential values of the random variable, and with respect to the cumulant generating function argument. Fubini's Theorem justifies interchanging the order of integration by selecting a path of integration C for which all points have a positive real part bounded away from zero; hence

$$Q_X(x) = \int_x^\infty \left[\frac{1}{2\pi i} \int_C \exp(\mathcal{K}_X(\beta) - \beta y) \, d\beta \right] \, dy$$
$$= \frac{1}{2\pi i} \int_C \left[\int_x^\infty \exp(\mathcal{K}_X(\beta) - \beta y) \, dy \right] \, d\beta.$$

The integral over $[x, \infty)$ can be replaced by a limiting operation:

$$Q_X(x) = \lim_{b \to \infty} \frac{1}{2\pi i} \int_C \exp(\mathcal{K}_X(\beta))(\exp(-\beta x) - \exp(-\beta b)) \, d\beta / \beta \qquad (95)$$
$$= \frac{1}{2\pi i} \int_C \exp(\mathcal{K}_X(\beta) - \beta x) \, d\beta / \beta$$
$$= \frac{1}{2\pi i} \int_C \exp(n \left[\mathcal{K}_Y(\beta) - \beta x \right]) \, d\beta / \beta \qquad (96)$$

The result (96) of the limiting operation in (95) also requires a path for β with real part bounded above zero. This inversion integral (96) lacks the factor n appearing in the density inversion integral (79) from the linear transformation creating $d\beta$, since it is eliminated by an identical factor in β.

5.1. A Large-Deviations Result

Bahadur and Ranga Rao (1960) approximate tail areas in a manner similar to that used to derive (94). Instead of forming the Edgeworth series after tilting the distribution separately for each y in the range of integration, one tilted Edgeworth series is calculated; this is chosen so that the mean is at x. Recall that when it can be constructed, the Edgeworth series e_{j} has error uniformly of order $o(n^{1-\mathrm{j}/2})$ in general, but in this case, since Edgeworth series of all orders can be constructed, e_{j} has error uniformly of order $o(n^{1-\mathrm{j}/2})$. Let $\hat{\omega} = \sqrt{2(\hat{\beta}x - \mathcal{K}_Y(\hat{\beta}))}$, with the sign the same as that of $\hat{\beta}$, exactly as in (82). Let $\hat{\sigma} = \sqrt{\mathcal{K}_X''(\hat{\beta})} = \sqrt{n\mathcal{K}_Y''(\hat{\beta})}$. From (93),

$$
\begin{aligned}
Q_X(x) &= \exp(\mathcal{K}_X(\hat{\beta}) - \hat{\beta}x) \int_x^\infty \exp(-(\hat{\beta}y - \hat{\beta}x)) \exp((\hat{\beta}y - \mathcal{K}_X(\hat{\beta}))) f_X(y) \; dy \\
&= \exp(-\frac{n}{2}\hat{\omega}^2) \int_x^\infty \exp(-(\hat{\beta}y - \hat{\beta}x)) f_X(y; \hat{\beta}) \; dy \\
&= \exp(-\frac{n}{2}\hat{\omega}^2) \int_x^\infty \frac{\exp\left(-(\hat{\beta}y - \hat{\beta}x)\right) \left[e_{\mathrm{j}}((y - x)/\hat{\sigma}; \hat{\boldsymbol{\rho}}_X) + o(n^{1-\mathrm{j}/2})\right]}{\hat{\sigma}} \; dy \\
&= \exp(-\frac{n}{2}\hat{\omega}^2) \int_0^\infty \exp\left(-\hat{\beta}y\hat{\sigma}\right) \left[e_{\mathrm{j}}(y; \hat{\boldsymbol{\rho}}_X) + o(n^{1-\mathrm{j}/2})\right] \; dy \qquad (97)
\end{aligned}
$$

Here $\hat{\boldsymbol{\rho}}_X$ are the invariants of §2.1, after tilting, and implicitly involve n. Bahadur and Ranga Rao (1960) expand the integral in (97) as an asymptotic series in n, effectively using Watson's Lemma, which is justified for any $\hat{\beta} > 0$, to obtain

$$
\frac{\phi(\sqrt{n}\hat{\omega})}{\sqrt{n}\hat{z}} \left\{ 1 + (1/n) \left[\frac{1}{8}\hat{\rho}_{Y,4} - \frac{5}{24}\hat{\rho}_{Y,3}^2 - \frac{\hat{\rho}_{Y,3}}{2\hat{z}} - \frac{1}{\hat{z}^2} \right] + O(n^{-2}) \right\}. \qquad (98)
$$

Approximation (98) may also be derived from (96). The same change of variables as in (81) is used:

$$
\omega(\beta) = \hat{\omega} + (\beta - \hat{\beta})\sqrt{2[\mathcal{K}_Y(\beta) - \beta x - \mathcal{K}_Y(\hat{\beta}) + \hat{\beta}x]/(\beta - \hat{\beta})^2},
$$

for ω near $\hat{\omega}$. Then

$$
Q_X(x) = \frac{1}{2\pi i} \exp(n[\mathcal{K}_Y(\hat{\beta}) - \hat{\beta}x]) \int_{\hat{\omega}-i\infty}^{\hat{\omega}+i\infty} \exp(\frac{n}{2}[\omega - \hat{\omega}]^2) \varpi(\omega)\beta^{-1} \frac{d\beta}{d\omega} \; d\omega. \quad (99)
$$

Since the factor of β^{-1} does not have an exponent of n this term is included with $d\beta/d\omega$ in the function $\varpi(\omega)$ of Lemma 4.5.1. The function ϖ has a pole at 0. Furthermore, if $x = \mathrm{E}[Y]$, and hence $\hat{\beta} = 0$, this pole coincides with $\hat{\omega}$. Otherwise $\varpi(\omega)$ is analytic at $\hat{\omega}$, and Watson's Lemma can be used. As before the value of the integral is determined by the derivatives of ϖ at $\hat{\omega}$. These derivatives will be calculated by expanding $\varpi(\omega) = (1/\beta)(d\beta/d\omega)$ about $\hat{\omega}$.

To expand $\varpi(\omega)$, first integrate the expansion in (85) to obtain

$$\beta = \hat{\beta} + \frac{\left[(\omega - \hat{\omega}) - \frac{1}{6}\hat{\rho}_{Y,3}(\omega - \hat{\omega})^2 + (\frac{5}{72}\hat{\rho}_{Y,3}^2 - \frac{1}{24}\rho_{Y,4})(\omega - \hat{\omega})^3 + O((\omega - \hat{\omega})^4)\right]}{\sqrt{\mathcal{K}''(\hat{\beta})}}.$$

(100)

Divide (85) by (100) to obtain

$$\varpi(\omega) = \left(1 + \frac{(\omega - \hat{\omega})^2 \left(24\,\hat{z}^{-2} + 12\,\hat{z}^{-1}\,\hat{\rho}_{Y,3} + 5\,\hat{\rho}_{Y,3}^2 - 3\,\hat{\rho}_{4,Y}\right)}{24} + (\omega - \hat{\omega})^4\right)/\hat{z},$$

(101)

where

$$\hat{z} = \hat{\beta}\sqrt{\mathcal{K}_Y''(\hat{\beta})},$$

(102)

and $\hat{\rho}_{Y,3}$ and $\hat{\rho}_{Y,4}$ are invariants of the distribution with density proportional to $f_Y(y)\exp(y\hat{\beta} - \mathcal{K}_Y(\hat{\beta}))$, defined in (72). The above series has odd order terms in $\omega - \hat{\omega}$ removed, since these do not affect the value of the integral as approximated by Watson's Lemma.

Alternatively, rather than expanding the composition of the logarithm function with $\beta(\omega)$, one might expand the series $d\beta/d\omega$ and $\beta(\omega)$, and calculate the resulting quotient.

Substituting (101) into (99), and then using Lemma 4.5.1 gives the tail probability approximation (98). This is valid only for $x > 0$, since the calculations resulting in (96) only hold for integration paths C consisting of points with positive real parts. One expects the expansion to become inaccurate as x moves close to 0, since as $x \to 0$ the radius of convergence for ϖ shrinks, and the error incurred in the outer parts of the inversion integral becomes more important.

Jensen (1988) includes uniform relative error bounds for (98) for ordinates with $\hat{\beta}$ bounded away from 0 as a consequence of the similar result for densities presented in §4.6, under the same conditions. Since (98) diverges to infinity as $\hat{\beta} \to 0$, it is inappropriate for tail areas corresponding to ordinates near the mean.

5.2. Direct Saddlepoint Series for Distribution Functions

Robinson (1982) presents a much different approach. He expresses the tail probability $Q_X(x)$ as (97), but unlike Bahadur and Ranga Rao (1960), explicitly accounts for the end of the range of integration at zero, by expressing (97) as the sum of integrals of the form $I_j(a) = \int_0^\infty \exp(-ay)h_j(y)\phi(y)\,dy$, in essence deriving the Laplace transform of the Edgeworth series.

Results of this section will be derived both by approximate integration of a density approximation, and later by an approximation to the complex integral (96) giving the tail probability exactly.

Robinson (1982) develops his approximation in cases more general than the mean of independent and identically distributed random variables. Suppose that

X is the mean of observations Y_j, satisfying the conditions of Lemma 3.11.1 and Theorem 3.11.2. By analogy with §4.1, let $\hat{\gamma}$ solve $\mathcal{K}'_X(\hat{\gamma}) = x$, and let

$$\hat{z} = \hat{\gamma}\sqrt{\mathcal{K}''_X(\hat{\gamma})/n}, \quad \hat{\omega} = \sqrt{2[\mathcal{K}_X(\hat{\gamma}) - x\hat{\gamma}]/n}. \tag{103}$$

When X is the mean of independent and identically distributed summands, then $\hat{z}$ and $\hat{\omega}$ are the same as before. Relation (99) continues to hold under these relaxed conditions with the broader definitions of $\hat{z}$ and $\hat{\omega}$.

Theorem 5.4.1: *Suppose that a random variable X with a distribution depending on a parameter n has a cumulant generating function on an interval $\mathcal{I}_n$ such that $[\inf_{\beta \in \mathcal{I}_n^c} |\beta|^2] \sup_{\beta \in \mathcal{I}_n} \mathcal{K}''_X(0)/n$ is bounded away from zero. Suppose further that the density of X is approximated by a sum of form $e_j(x; \boldsymbol{\rho}^n)$ with error that is uniformly of order $o(n^{(2-j)/2})$. Then*

$$Q_X(x) = \exp(-\frac{n}{2}\hat{\omega}^2) \sum_{j=0}^{j} \mu^*_{j,j} I_j(\sqrt{n}\hat{z}) + o(n^{1-j/2})$$

for $\hat{\omega}$ and $\hat{z}$ as in (103), and for $\mu^*_{j,j}$ the pseudomoments of Table 3, calculated from the invariants $\boldsymbol{\rho}^n$ instead of the cumulants $\boldsymbol{\kappa}^n$, and with terms of size $o(n^{1/2-j/2})$ deleted.

Proof: The result follows by integrating (97). The error remains of order stated since the convergence is uniform.

Q.E.D

Using integration by parts, and noting that $(d/dy)h_j(y)\phi(y) = -h_{j+1}(y)\phi(y)$, one finds that $I_j(a) = (2\pi)^{-1/2}h_{j-1}(0) - aI_{j-1}(a)$ and $I_0(a) = \exp(a^2/2)(1 - \Phi(a))$. Then

$$I_1(a) = (2\pi)^{-1/2} - a \exp(a^2/2)\left(1 - \Phi(a)\right),$$
$$I_2(a) = -a\left((2\pi)^{-1/2} + a\exp(a^2/2)\left(1 - \Phi(a)\right)\right),$$
$$I_3(a) = a^2\left((2\pi)^{-1/2} - a\exp(a^2/2)\left(1 - \Phi(a)\right)\right) - (2\pi)^{-1/2}.$$

When j in (97) is 3, then $\mu^*_{0,3} = 1$, $\mu^*_{1,3} = \mu^*_{2,3} = 0$, and $\mu^*_{3,3} = \hat{\rho}_{X,3}$, and the resulting approximation is

$$Q^\circ_X(x) = \exp\left(n(\hat{z}^2 - \hat{\omega}^2)/2\right) \times$$
$$\left[[1 - \Phi(\sqrt{n}\hat{z})]\left(1 - n^{3/2}\hat{\rho}_{X,3}\hat{z}^3/6\right) + \phi(\sqrt{n}\hat{z})\sqrt{n}\hat{\rho}_{X,3}(n\hat{z}^2 - 1)/(6\sqrt{n})\right], \tag{104}$$

and

$$Q_X(x) = Q^\circ_X(x) + \phi(\sqrt{n}\hat{\omega})O(1/n). \tag{105}$$

When X is the mean of independent and identically distributed random variables, each with third invariant $\rho_{Y,3}$, then

$$Q^\circ_X(x) = \exp\left(n(\hat{z}^2 - \hat{\omega}^2)/2\right) \times$$

$$\left[[1 - \Phi(\sqrt{n}\hat{z})] \left(1 - n\hat{\rho}_{Y,3}\hat{z}^3/6 \right) + \phi(\sqrt{n}\hat{z})\hat{\rho}_{Y,3}(n\hat{z}^2 - 1)/(6\sqrt{n}) \right], \quad (106)$$

with $\hat{\omega}$ as in (82) and $\hat{z}$ as in (102).

The remaining discussion of approximations of this type will center on (106), since it is simpler, but most of the comments apply to (104) as well. Approximation $Q_X^\circ(x)$ is valid to the same order as the Edgeworth series it uses, since the error rate of $e_{\mathrm{j}}(y; \boldsymbol{\kappa}(\hat{\beta}))$ is uniform. The invariant $\hat{\rho}_{Y,3}$ is for the tilted distribution; multipliers involving n arise because the invariants of the tilted mean occurring in (97) have been replaced by the tilted invariants of the original summands. The method is adaptable to situations in which the Edgeworth series is valid, and particularly has applications outside the realm of sums of independent and identically distributed random variables.

Both approximations (98) and (104) are derived from the integral (97). The uniform error behavior of the Edgeworth series and the bounded convergence theorem might lead one to believe that the integral represented by (97) will converge uniformly at least on intervals bounded above zero. This is not necessarily true, as the coefficients of the Edgeworth series depend on the ordinate. One expects the accuracy of (104) to deteriorate as $\hat{\beta} \to 0$, since the quantity multiplying the error term increases as $\hat{\beta}$ decreases. Unlike the previous large-deviations result (98), however, (104) is still defined when $\hat{\beta} = 0$.

Daniels (1987) notes that (106) may be formally derived from (96) by expanding the exponent in β about $\hat{\beta}$, retaining the quadratic term in the exponent, and exponentiating the rest of the series, while retaining β in the denominator unexpanded. That is,

$$Q_X(x) = \frac{\exp(-n\hat{\omega}^2/2)}{2\pi i}$$
$$\times \int_C \exp\left(n \left[\mathcal{K}_Y''(\hat{\beta})(\beta - \hat{\beta})^2/2 \right] \right) (1 + \frac{n}{6} \mathcal{K}_Y'''(\hat{\beta})(\beta - \hat{\beta})^3 + \cdots) \frac{d\beta}{\beta}, (107)$$

where terms represented by $\cdots$ are powers of $\beta - \hat{\beta}$ times an inverse factorial, times pseudo-moments calculated from third and higher order derivatives of $\mathcal{K}_Y$ at $\hat{\beta}$, as in §3.2. Kolassa (2003a) shows that the infinite series implied in (107) may be truncated and integrated termwise to yield an error that is of the same order as the first omitted term. The resulting integrated truncated series be expressed as a linear combination of integrals of the form

$$J_k(v) = \int_{\hat{\omega}-i\infty}^{\hat{\omega}+i\infty} \frac{\exp(n[\frac{1}{2}\gamma^2 - \gamma v])(\gamma - v)^k}{2\pi i \gamma} \, d\gamma;$$

specifically, when the sum in (107) is truncated to only those terms given explicitly, then

$$Q_X^\circ(x) = \exp(n[\hat{z}^2 - \hat{\omega}^2]/2)(J_0(\hat{z}) + \frac{n}{6}\hat{\rho}_{Y,3}J_3(\hat{z})).$$

These integrals $J_k(v)$ can be calculated analytically, by noting that

$$J_k(v) = \int_{\hat{\omega}-i\infty}^{\hat{\omega}+i\infty} \frac{\exp(n[\frac{1}{2}\gamma^2 - \gamma v])(\gamma - v)^{k-1}}{2\pi i} \, d\gamma - v J_{k-1}(v),$$

and by applying (96) to the standard normal distribution, to see that

$$J_0(v) = 1 - \Phi(v). \tag{108}$$

Relation (106) results from substitution of J_0 and J_3. The exercises invite exploration of the relationship between I_j and $J_j(v)$.

The results of this section involve two difficulties. As $v \to 0$, these integrals become larger, and one expects the accuracy of approximating the infinite series by a truncated series to deteriorate. Furthermore, if $Q^\dagger$ represents (106) calculated for the distribution of $-X$, then $Q^\circ(x) \neq 1 - Q^\dagger(x)$, illustrating that (106) is not the appropriate expansion when $\hat{\beta} < 0$.

Robinson (1982) notes that his approach and that of Bahadur and Ranga Rao (1960) differ in that his series contains terms involving the normal distribution function and the equivalent terms in the Bahadur and Ranga Rao (1960) series have this replaced by a first order approximation. The results of Robinson (1982) correspond to inversion integrals with β in the denominator of the integrand, while Bahadur and Ranga Rao (1960) expand β^{-1} about $\hat{\beta}$, removing β from the denominator. Heuristically, in the work of Bahadur and Ranga Rao (1960), factors like $\Phi(x)$ are replaced by approximations.

5.3. Less Direct Saddlepoint Series for Distribution Functions

Lugannani and Rice (1980) provide an alternate expansion again using the methods of steepest descent. The form of this expansion is very simple, and it is valid over a wide range of possible ordinates. Again the same steepest descent path (81) is used, and the transformed distribution function inversion integral is still (99). An expansion is now developed for this integral that will differ from (98) in that it will explicitly account for the singularity in $\varpi(\omega) = (1/\beta)(d\beta/d\omega)$ at zero. This will illustrate why care has been taken in the selection of $\hat{\omega}$. Until now the only value of ω of interest has been that value corresponding to $\beta = \hat{\beta}$. In the development of this section the value of ω corresponding to $\beta = 0$ is also of interest, since explicit allowance for the singularity here is desired.

Theorem 5.3.1: *Suppose that a random variable X arises as the mean of n independent and identically distributed random variables, each with cumulant generating function $\mathcal{K}_Y$. Consider any x for which the saddlepoint equation (69) has a solution, and let $\hat{\beta}$ be that solution. Let $Q_X(x) = \mathrm{P}\,[X \geq x]$, and let the Lugannani and Rice approximation to $Q_X(x)$ be*

$$Q_X^*(x) = 1 - \Phi(\sqrt{n}\hat{\omega}) + \phi(\sqrt{n}\hat{\omega})(1/\hat{z} - 1/\hat{\omega})/\sqrt{n}, \tag{109}$$

where again $\hat{\omega}$ is given by (82) and $\hat{z}$ is given in (102). When $\hat{\beta} = 0$, replace $1/\hat{z} - 1/\hat{\omega}$ by $\frac{1}{6}\hat{\rho}_{Y,3}$. Then

$$Q_X(x) = Q_X^*(x) + \phi(\sqrt{n}\hat{\omega})(1 + O(n^{-3/2})). \tag{110}$$

Proof: Express the factor $\varpi(\omega)$ in the integrand of (99) as the sum of two terms, $\varpi^*(\omega) = (1/\beta)(d\beta/d\omega) - (1/\omega)$ and $\varpi^\dagger(\omega) = 1/\omega$, and integrate the two terms separately. In the first instance the function is differentiable for ω in a region about $\hat{\omega}$ which does not shrink as $\hat{\omega}$ approaches zero. To see this, note that $\omega\varpi^*(\omega)$ is analytic and hence $\varpi^*(\omega) = \sum_{j=-1}^{\infty} c_j\omega^j$ for some coefficients $\{c_j\}$. Also, $\lim_{\omega\to 0} \omega\varpi^*(\omega) = \lim_{\omega\to 0}(\omega/\beta)(d\beta/d\omega) - 1 = 0$ by the definition of first derivative. This implies that $c_{-1} = 0$ and $\varpi^*(\omega)$ is analytic at 0. Hence the use of Watson's Lemma does not have the problems associated with it as does the process yielding (98).

As noted in §4.5, $(d\beta/d\omega)(\hat{\omega}) = 1/\sqrt{\mathcal{K}_Y''(\hat{\beta})}$, and hence $\varpi^*(\hat{\omega}) = 1/\hat{z} - 1/\hat{\omega}$. Hence Watson's Lemma shows that

$$\frac{\exp(-\frac{n}{2}\hat{\omega}^2)}{\sqrt{2\pi}} \int_{\hat{\omega}-iA}^{\hat{\omega}+iB} \exp\left(\frac{n}{2}[\omega - \hat{\omega}]^2\right)\left(\frac{1}{\beta}\frac{d\beta}{d\omega} - \frac{1}{\omega}\right)d\omega = \frac{\exp(-\frac{n}{2}\hat{\omega}^2)}{\sqrt{2n\pi}}\left(\frac{1}{\hat{z}} - \frac{1}{\hat{\omega}} + O(n^{-1})\right)$$

$$= \phi(\sqrt{n}\hat{\omega})\left(\frac{1}{\hat{z}} - \frac{1}{\hat{\omega}} + O(n^{-1})\right)/\sqrt{n}.$$

Integration of $\varpi^\dagger(\hat{\omega})$ can be done exactly; (108) implies that

$$\frac{\exp(-\frac{n}{2}\hat{\omega}^2)}{2\pi} \int_{\hat{\omega}-i\infty}^{\hat{\omega}+i\infty} \exp(\frac{n}{2}[\omega - \hat{\omega}]^2)d\omega/\omega = 1 - \Phi(\sqrt{n}\hat{\omega}).$$

When $\hat{\beta} = 0$, both $\hat{\omega}$ and $\hat{z}$ are zero, and $1/\hat{z} - 1/\hat{\omega}$ has a singularity. This singularity is removable; that is, as $\hat{\beta} \to 0$, the limit of $1/\hat{z} - 1/\hat{\omega}$ exists. Yang and Kolassa (2002) present a smooth method for removing this singularity. Note that $\hat{\omega}^2 = 2[\mathcal{K}_Y'(\hat{\beta})\hat{\beta} - \mathcal{K}_Y(\hat{\beta})] = \mathcal{K}_Y''(0)\hat{\beta}^2 + 2\mathcal{K}_Y^{(3)}(0)\hat{\beta}^3/3 + O(\hat{\beta}^4)$, $\hat{z}^2 = \mathcal{K}_Y''(0)\hat{\beta}^2 + \mathcal{K}_Y^{(3)}(0)\hat{\beta}^3 + O(\hat{\beta}^4)$, and hence $\hat{z}/\hat{\omega} - 1 = \frac{1}{6}\hat{\rho}_{Y,3}\hat{z} + o(\hat{\beta})$.

Q.E.D

When $\hat{\beta} = 0$, (109), after the modification in the statement of the theorem, equals Robinson's approximation (106).

When (109) is applied to calculate tail areas for the mean of n variables following a logistic distribution, even when $n = 1$, it performs very well. See Fig. 13a. Fig. 13b exhibits this approximation when applied to the mean of ten independent and identically distributed variables. In this case, the cumulant generating function was given in exercise 7 of §2.8, and its derivatives were given in in 4.9. Fig. 13 might then be drawn by selecting a set of ordinates $\mathfrak{X}$, and for each $x \in \mathfrak{X}$, solve (69) for $\hat{\beta}$, calculate $\hat{\omega}$ and $\hat{z}$ from x and $\hat{\beta}$, calculate the invariants $\hat{\rho}_{Y,j}$, and then apply (98), (104), and (109). Solving (69) might require some computational effort; in this case,

Fig. 13 might more easily be constructed by selecting a set of saddlepoints $\mathfrak{T}$, and for each $\hat{\beta} \in \mathfrak{T}$, calculating x from (69). One might then proceed as before.

If Q^*_{-X} represents Q^* calculated for the distribution of $-X$, then $Q^*_{-X}(x) = 1 - Q^*_X(-x)$. Hence the Lugannani and Rice approximation to upper tails applies equally well to lower tails, unlike the previous tail probability approximations.

One might try to add terms to (109) in order to construct an approximation whose error converges to zero more quickly as n increases. Kolassa (1996b) provides such an approximation for conditional distributions; these results are discussed in §7.4. An unconditional higher–order counterpart to (109) is a special case of the expansion of §7.4.

The following section presents an alternate derivation of (109) that avoids construction of the exact inversion integral for the tail probability.

5.4. The Equivalence of the Series of Lugannani and Rice and Robinson

The approximations (109) and (104), when applied to sums of independent and identically distributed continuous random variables, agree to $O(n^{-1})$, since both agree with the true tail probabilities to this order. Robinson justifies the use of (104) in more general cases when the tail probabilities are given by an Edgeworth series. In these cases no generalization of (109) is available. In this section I give conditions under which (109) and (104) agree to $O(n^{-1})$, allowing the use of (109) as an alternative to (104). This proof makes use of the derivatives of the two approximations; before proving the theorem, I will construct these derivatives. Differentiating $\hat{\omega}^2/2 = \hat{\beta}\mathcal{K}_n{}'(\hat{\beta}) - \mathcal{K}_n(\hat{\beta})$, $\hat{z}^2 = \hat{\beta}^2\mathcal{K}_n{}''(\hat{\beta})$, and $\mathcal{K}_n{}'(\hat{\beta}) = x$ with respect to x, and solving, yields

$$\frac{d\hat{\omega}}{dx} = \hat{z}\sqrt{\mathcal{K}_n{}''(\hat{\beta})}/\hat{\omega}\frac{d\hat{\beta}}{dx},\ \frac{d\hat{z}}{dx} = \sqrt{\mathcal{K}_n{}''(\hat{\beta})}[1 + \tfrac{1}{2}\hat{z}\hat{\rho}_{n,3}]\frac{d\hat{\beta}}{dx},\ \frac{d\hat{\beta}}{dx} = 1/\mathcal{K}_n{}''(\hat{\beta}). \quad (111)$$

Hence $\frac{d}{dx}Q^*_X(x) = \sqrt{n}\phi(\sqrt{n}\hat{\omega})\{-1 + (\hat{\omega}^{-3}\hat{z} - \hat{z}^{-2}[1+\tfrac{1}{2}\hat{z}\hat{\rho}_{n,3}])/n\}/\sqrt{\mathcal{K}_n{}''(\hat{\beta})}$. Similarly,

$$\frac{d}{dx}Q^\circ_X(x) = \frac{\sqrt{n}\phi(\sqrt{n}\hat{\omega})}{\sqrt{\mathcal{K}_n{}''(\hat{\beta})}}\left\{ -1 + \frac{3\,\hat{\rho}_{n,3}^2 + n^2\,\hat{z}^4\,\hat{\rho}_{n,3}^2 - n\,\hat{z}^2\left(\hat{\rho}_{n,3}^2 - 2\,\hat{\rho}_{n,4}\right) - 2\,\hat{\rho}_{n,4}}{12\,n} + \right.$$
$$\left. \frac{n^{3/2}\,\hat{z}^3\,\bar{\Phi}(\sqrt{n}\,\hat{z})\left(n\,\hat{z}^2\,\hat{\rho}_{n,3}^2 + 2\,\hat{\rho}_{n,4}\right)}{12\,n\,\phi(\sqrt{n}\,\hat{z})}\right\}$$

Kolassa (2005) proves the following result:

Theorem 5.4.1: *Suppose that $\mathcal{K}_n$ is a sequence of complex functions analytic on $\mathcal{I} \times i\mathfrak{R}$, for $\mathcal{I} \subset \mathfrak{R}$ a compact set such that the interior of $\mathcal{I}$ contains 0. Suppose that $\mathcal{K}_n{}''(0) = 1$ for all n, and that $\mathcal{K}_n$ restricted to $\mathcal{I}$ is convex. Let $\hat{\rho}_{n,j} = \mathcal{K}_n{}^{(j)}(\hat{\beta})/\mathcal{K}_n{}''(\hat{\beta})^{j/2}$, and suppose further that $|\hat{\rho}_{n,j}| \leq B_j$ for $\hat{\beta} \in \mathcal{I}$, and $j = 3, 4$. Let $\hat{\beta}$ be a function of x satisfying $\mathcal{K}_n{}'(\hat{\beta}) = x$, and let $\hat{z}(x)$ and $\hat{\omega}(x)$*

Saddlepoint Approximations to the CDF of a Logistic Variable
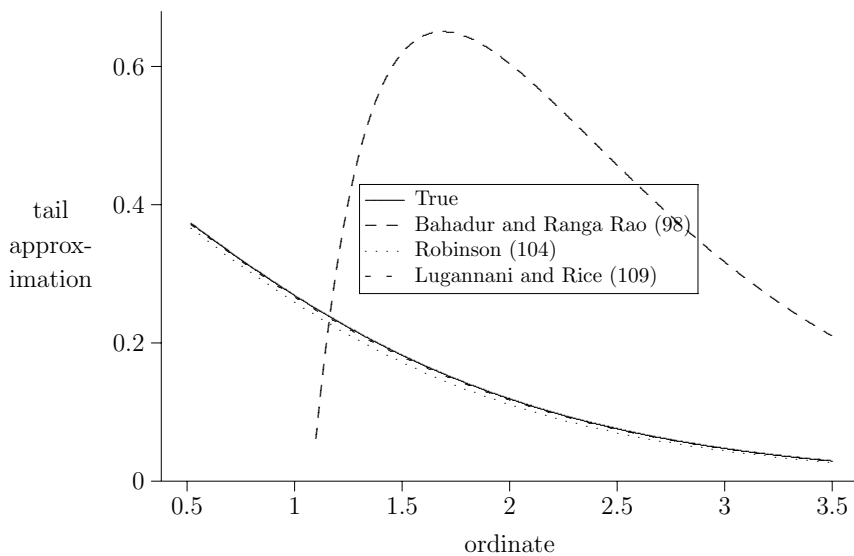

Fig. 13a.

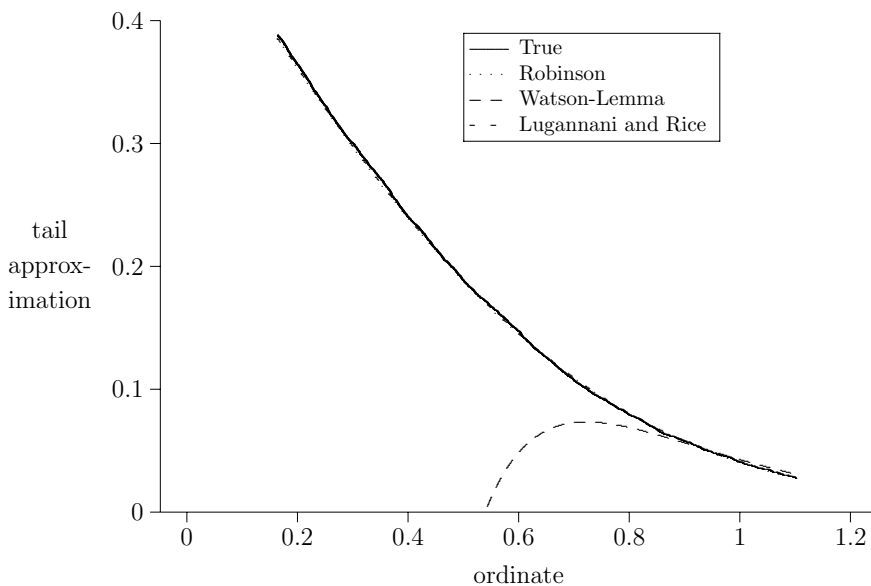*Saddlepoint Approximations to the CDF of Mean of 10 Logistic Variables*

Fig. 13b.

be as in (103). Then (104) and (109) agree to order $O(1/n)\phi(\sqrt{n}\hat\omega)$; this order is uniform on $\mathcal{I} \cap [0, \infty)$.

Proof: The conditions of this theorem put bounds in standardized derivatives of $\mathcal{K}_n(\hat\beta)$, of the form $\mathcal{K}_n^{(j)}(\hat\beta)/\mathcal{K}_n''(\hat\beta)^{j/2}$ During the course of this proof it will be necessary to bound similar derivatives, but a different argument to $\mathcal{K}_n''$ in some of the factors making up the denominator. The first step in the proof bounds the variation in $\mathcal{K}_n''(\beta)$. For some $\delta > 0$ and $B_3 > 0$, three cases are treated separately. **Case 1:** $\hat\beta \le n^{-1/2}\delta$. **Case 2:** $1/(2B_3) \ge \hat\beta \ge \delta n^{-1/2}$. **Case 3:** $\hat\beta \ge 1/(2B_3)$.

Case 3 is handled by examining the asymptotic expansion for $\bar\Phi(x)$ in powers of $1/x$. Cases 1 and 2 are handled by examining the bound on the error of the Taylor series for the difference in the two approximations. Treatment of cases 1 and 2 differ according to what is allowed to occur in denominators of the terms in the difference bound.

<div align="right">Q.E.D</div>

5.5. Approximate Sample-Space Integration

As an alternative to the derivation of saddlepoint cumulative distribution function approximations via inversion of the cumulant generating function, this section discusses derivation by integrating the saddlepoint density approximation. First, consider the following special case of a result due to Temme (1982), giving an asymptotic expansion for the integral of a normal density times another function, over part of the reals:

Lemma 5.5.1:   *Suppose that $H(w)$ is a function with two derivatives, such that $\sup_{-K \le w \le K} |H''(w)| < \infty$. Choose $\epsilon > 0$. Choose $u \in [0, K - \epsilon]$. Then*

$$\sqrt{n} \int_u^{K-\epsilon} \phi(\sqrt{n}w)H(w) \, dw = \left(H(0) + \frac{H''(0)}{2n}\right) \bar\Phi(\sqrt{n}u) +$$

$$\frac{\phi(\sqrt{n}u)}{\sqrt{n}} \left\{ \frac{H(u) - H(0)}{u} + O\left(\frac{1}{n}\right) \right\}.$$

Proof: With a bit of algebra,

$$\sqrt{n} \int_u^K \phi(\sqrt{n}w)H(w) \, dw = H(0)[\Phi(\sqrt{n}K) - \Phi(\sqrt{n}u)]$$

$$+ \int_u^K \sqrt{n}w\phi(\sqrt{n}w)\frac{H(w) - H(0)}{w} \, dw.$$

After integration by parts,

$$\sqrt{n} \int_u^K \phi(\sqrt{n}w)H(w) \, dw = H(0)[\bar\Phi(\sqrt{n}u) - \bar\Phi(\sqrt{n}K)]$$

$$- \frac{1}{\sqrt{n}} \phi(\sqrt{n}w)[H(w) - H(0)]/w \Big|_u^K + \frac{1}{\sqrt{n}} \int_u^K \phi(\sqrt{n}w)\frac{d}{dw}\frac{H(w) - H(0)}{w} \, dw.$$

The final term may be bounded by $[\Phi(\sqrt{n}K) - \Phi(\sqrt{n}u)]\sup_{w\in[0,K]}|H''(w)|/n$; this bound is of size $\phi(\sqrt{n}u)O(1/n)$. Similarly,

$$\sqrt{n}\int_u^K \phi(\sqrt{n}w)\frac{d}{dw}\frac{H(w)-H(0)}{w}\,dw = \tfrac{1}{2}H''(0)\bar{\Phi}(\sqrt{n}u) + O(1/\sqrt{n})\phi(\sqrt{n}u).$$

<div align="right">Q.E.D</div>

The preceding lemma will be used to integrate a saddlepoint density to obtain a saddlepoint cumulative distribution function approximation. Since the exponent of the saddlepoint density approximation is not exactly the sample size times a quadratic, exponentiated, the following corollary will be useful.

Corollary 5.5.2: *Suppose that g is a twice–differentiable function defined on $[0,K]$, with $g(0) = 0$, $g'(0) = 0$, and $g''(v) > 0$ for all v, and suppose that G has two bounded derivatives on $[0,K]$. Then for any $x \in [0,K]$,*

$$\int_x^K \frac{\sqrt{n}\exp(-ng(v))}{\sqrt{2\pi}}G(v)\,dv = \frac{G(0)}{\sqrt{g''(0)}}\left[1 + \frac{5\,a_3{}^2 - 3a_4 - 12a_3\,b_1 + 12b_2}{12n}\right] \times$$

$$\bar{\Phi}(\sqrt{2ng(x)}) + n^{-1/2}\phi(\sqrt{2ng(x)})\left[\frac{G(x)}{g'(x)} - \frac{G(0)}{\sqrt{2g(x)g''(0)}} + O(1/n)\right], \quad (112)$$

where $a_j = g^{(j)}(0)(g^{(2)}(0))^{-j/2}$, and $b_j = G^{(j)}(0)(g^{(2)}(0))^{-j/2}G(0)^{-1}$.

Proof: Let $w = \sqrt{2g(v)}$. Then $\frac{dw}{dv}w = g'(v)$; hence $\frac{dv}{dw} = w/g'(v) = \sqrt{2g(v)}/g'(v)$, and $\frac{dv}{dw}\big|_{w=0} = 1/\sqrt{g''(0)}$. After changing variables,

$$\int_x^\infty \frac{\sqrt{n}\exp(-ng(v))}{\sqrt{2\pi}}G(v)\,dv = \int_{\sqrt{2g(x)}}^\infty \sqrt{n}\phi(\sqrt{n}w)G(v(w))\frac{dv}{dw}\,dw.$$

Applying Lemma 5.5.1, with $u = \sqrt{2g(x)}$ as above, and $H(w) = G(v(w))\frac{dv}{dw}$. Then $[H(u)-H(0)]/u = G(x)/g'(x)-G(0)/[\sqrt{2g''(0)g(x)}]$, and $H(0) = G(0)/g''(0)$, gives the result.

<div align="right">Q.E.D</div>

These results apply to the integration of the saddlepoint density approximation (68) over a tail region to obtain a saddlepoint tail probability approximation. Recall from the discussion surrounding Theorem 4.1.1 that standard regularity conditions generally guarantee the existence of the saddlepoint density approximation only on a convex set containing zero, and so the range of integration must be truncated. A lemma that is a simpler version of Theorem 15 of Petrov (1975), Chapter 3, is needed.

Lemma 5.5.3: *Suppose that X is a random variable with cumulant generating function $\mathcal{K}_X$. Then $P[X \geq x] \leq \exp(\mathcal{K}_X(\beta) - \beta x)$.*

Proof: For $\beta > 0$, let $Y = \exp(x\beta)$ if $X > x$, and 0 otherwise. Since $\exp(X\beta) \geq Y$, then $\exp(\mathcal{K}(\beta)) = \mathrm{E}\left[\exp(X\beta)\right] \geq \mathrm{E}\left[Y\right] = \exp(x\beta)\mathrm{P}\left[X \geq x\right]$.

<div align="right">Q.E.D</div>

These results may now be used to provide an alternate proof to Theorem 5.3.1. Assume without loss of generality that $\mathrm{E}\left[X\right] = 0$. If the saddlepoint tail probability approximation (70) to a random variable X is defined for $x \in [0, K]$ (and perhaps for other x), and if x is an element of the smaller interval $[0, K - \epsilon]$, then Lemma 5.5.3 shows that $\mathrm{P}\left[X \geq x\right] = \mathrm{P}\left[K \geq X \geq x\right] + O(\exp(\mathcal{K}_X(\beta) - \beta x)/n^{3/2})$. We will now apply Corollary 5.5.2 with $g(x) = \hat{\beta}x - \mathcal{K}_Y(\hat{\beta})$, and $G(x) = \mathcal{K}_Y''(\hat{\beta})^{-1/2}[1 + (\hat{\rho}_4/8 - 5\hat{\rho}_3^2/24)/n]$. In order to prove a slightly weaker result than Theorem 5.3.1, in which the relative error is of size $O(1/n)$, one need only observe that $G(0)/\sqrt{g''(0)} = 1$, and that u of Corollary 5.5.2 is the same as $\hat{\omega}$ of (82), and $g'(x)/G(x)$ is the same as $\hat{z}$ of (102). To capture the full result of Theorem 5.3.1, one must verify that the term of size $O(1/n)$ multiplying $\bar{\Phi}(\sqrt{n}u)$ in (112) exactly balances the $O(1/n)$ term in $G(0)$. Verifying this is an exercise in rather tedious algebra, and is suggested in the exercises.

Future applications of Corollary 5.5.2 will also require an argument like Lemma 5.5.3 to complete the argument; this level will not always be given in the sequel.

5.6. An Adjusted Approximation

Again suppose that X is the mean of n independent and identically distributed random variables, each with cumulant generating function $\mathcal{K}_Y$. Note that $\hat{\omega}$ (first defined in (82)) is the value of $\hat{\Omega} = \sqrt{2(\hat{\beta}(X)X - \mathcal{K}_Y(\hat{\beta}(X)))}$, the signed root of the log likelihood ratio statistic, corresponding to the potential data value x. As before, $\hat{\beta}(X)$ satisfies $\mathcal{K}_Y'(\hat{\beta}(X)) = X$. Since $\hat{\omega}$ is an increasing function of x, tail probabilities for $\hat{\Omega}$ satisfy $\mathrm{P}\left[\hat{\Omega} \geq \hat{\omega}\right] = Q_X^*(x(\hat{\omega}))(1 + O(n^{-1}))$, for $x(\hat{\omega})$ the inverse of relationship (82), and where $\hat{z}$ is considered as a function of $\hat{\omega}$. Recall that $Q_X^*(x(\hat{\omega})) = \bar{\Phi}(\sqrt{n}\hat{\omega}) + \phi(\sqrt{n}\hat{\omega})(1/\hat{z} - 1/\hat{\omega})/\sqrt{n}$; that is, the approximation consists of the normal tail probability associated with the observed statistic value, plus a correction of smaller order. Practitioners might find an alternative approximation useful; if an additive correction to the normal tail probability were found that is applied to the statistic value rather than to the probability, the corrected statistic value might be reported as part of routine data analyses, and significance levels could be approximated using standard normal tables.

This section produces such a modified saddlepoint tail probability approximation. It also presents a proof that the approximations presented in §§5.2 and 5.3 are asymptotically equivalent. Of course, since both of these approximations are accurate to $O(1/n)$ when the regularity conditions for both approximations are satisfied.

Barndorff–Nielsen (1986, 1990a) provides corrections additive on the ordinate scale instead, in order to facilitate conditional inference. Jensen (1992) demonstrates

that these corrections hold in great generality, by demonstrating the following result:

Lemma 5.6.1: *Let g_1 and g_2 be smooth increasing functions from $\Re$ to $\Re$, such that $g_1(0) = g_2(0) = 0$ and $g_1'(0) = g_2'(0) > 0$. Let $g_3(s) = g_1(s)^{-1}\log(g_2(s)/g_1(s))$, and let $g_4(s) = g_1(s) - g_3(s)/n$. Then for $s_0 > 0$,*

$$\Phi\left(\sqrt{n}g_4(s)\right) = \Phi\left(\sqrt{n}g_1(s)\right) - \frac{\phi\left(\sqrt{n}g_1(s)\right)}{\sqrt{n}}\left\{\frac{1}{g_2(s)} - \frac{1}{g_1(s)} + \frac{e_{1,n(s)}}{g_1(s)n}\right\},$$

with

$$\sup\{e_{1,n}(s)|0 \le s \le s_0, n > 0\} < \infty, \quad \sup\{\sqrt{n}e_{1,n}(s)|0 \le s \le s_0/\sqrt{n}, n > 0\} < \infty. \tag{113}$$

Proof: For $s < s_0/\sqrt{n}$, the quantities $g_1(s)\sqrt{n}$, $g_2(s)\sqrt{n}$, $|g_1(s)/g_2(s) - 1|\, s^{-1}$, and $g_1(s)^{-1}|g_1(s)/g_2(s) - 1|$ are bounded. The power series expansion for $\log(z)$ is $(z-1) - \frac{1}{2}(z-1)^2 + \frac{1}{3}z^{*-3}(z-1)^3$ for z^* between 1 and z, so there exist bounded functions $e_{2,n}(s)$ and $e_{3,n}(s)$ such that

$$g_4(s) - g_1(s) = \frac{n^{-1}}{g_1(s)}\left\{\left(\frac{g_1(s)}{g_2(s)} - 1\right) - \frac{1}{2}\left(\frac{g_1(s)}{g_2(s)} - 1\right)^2 + e_{2,n}(s)\left(\frac{g_1(s)}{g_2(s)} - 1\right)^3\right\},$$

and

$$\Phi(\sqrt{n}g_4(s)) - \Phi(\sqrt{n}g_1(s)) = \sqrt{n}(g_4(s) - g_1(s))\times$$
$$\phi(\sqrt{n}g_1(s))\left[1 - \frac{n}{2}(g_4(s) - g_1(s))g_1(s)\right] + e_{3,n}(s)n^{3/2}(g_4(s) - g_1(s))^3.$$

Substituting,

$$\Phi(\sqrt{n}g_4(s)) - \Phi(\sqrt{n}g_1(s)) = \frac{\phi(\sqrt{n}g_1(s))}{\sqrt{n}g_1(s)}\left\{\left(\frac{g_1(s)}{g_2(s)} - 1\right) - \frac{1}{2}\left(\frac{g_1(s)}{g_2(s)} - 1\right)^2\right.$$
$$\left. + e_{2,n}(s)\left(\frac{g_1(s)}{g_2(s)} - 1\right)^3\right\}\left[1 - \frac{1}{2}\left(\frac{g_1(s)}{g_2(s)} - 1\right) + \frac{1}{4}\left(\frac{g_1(s)}{g_2(s)} - 1\right)^2 - \frac{e_{2,n}(s)}{2}\left(\frac{g_1(s)}{g_2(s)} - 1\right)^3\right]$$
$$+ e_{3,n}(s)n^{3/2}(g_4(s) - g_1(s))^3.$$

Thus the second relation in (113) holds.

For $s > s_0/\sqrt{n}$, replace the difference in the normal cumulative distribution function values by an integral of the density over the interval between the ordinates, replace the density using the first relationship of Table 2, and use integration by

parts, to find

$$\Phi\left(\sqrt{n}g_4(s)\right) - \Phi\left(\sqrt{n}g_1(s)\right) = \int_{\sqrt{n}g_1(s)}^{\sqrt{n}g_4(s)} \phi(x)\, dx = -\int_{\sqrt{n}g_1(s)}^{\sqrt{n}g_4(s)} \phi'(x)x^{-1}\, dx$$

$$= \frac{\phi(\sqrt{n}g_1(s))}{\sqrt{n}g_1(s)} - \frac{\phi(\sqrt{n}g_4(s))}{\sqrt{n}g_4(s)} - \int_{\sqrt{n}g_1(s)}^{\sqrt{n}g_4(s)} \phi(x)x^{-2}\, dx$$

$$= \frac{\phi(\sqrt{n}g_1(s))}{\sqrt{n}g_1(s)} - \frac{\phi(\sqrt{n}g_4(s))}{\sqrt{n}g_4(s)} + g_3(s)\frac{\phi(\sqrt{n}g_1(s))}{\sqrt{n}g_1(s)^2} - \int_{\sqrt{n}g_1(s)}^{\sqrt{n}g_4(s)} \left[\frac{\phi(x)}{x^2} - \frac{\phi(\sqrt{n}g_1(s))}{ng_1(s)^2}\right]\, dx$$

$$= \frac{\phi(\sqrt{n}g_1(s))}{\sqrt{n}g_1(s)} - \frac{\phi(\sqrt{n}g_4(s))}{\sqrt{n}g_4(s)} + g_3(s)\frac{\phi(\sqrt{n}g_1(s))}{\sqrt{n}g_1(s)^2}$$

$$\quad - \phi(\sqrt{n}g_1(s))\int_{\sqrt{n}g_1(s)}^{\sqrt{n}g_4(s)} \int_{\sqrt{n}g_1(s)}^{x} \frac{\exp(ng_1^2(s)/2 - u^2/2)(2+u^2)}{u^3}\, du\, dx.$$

The final integral may be expressed as

$$\frac{\phi(\sqrt{n}g_1(s))}{\sqrt{n}g_1(s)}\int_{\sqrt{n}g_1(s)}^{\sqrt{n}g_4(s)} \int_{\sqrt{n}g_1(s)}^{x} \frac{\exp(\frac{1}{2}(ng_1(s) - \sqrt{n}u)(g_1(s) + u/\sqrt{n}))\sqrt{n}g_1(s)(2+u^2)}{u^3}\, du\, dx.$$

The integrand above is bounded, and hence the integral is bounded by a constant times $n\,|g_4(s) - g_1(s)|^2$, which in turn is bounded by a constant times $1/n$. Note that

$$\frac{\phi(\sqrt{n}g_4(s))}{\sqrt{n}g_4(s)} = \frac{\phi(\sqrt{n}g_1(s))}{\sqrt{n}g_1(s)}\exp(-\tfrac{1}{2}g_3(s)/\sqrt{n})/(1 + n^{-1}g_3(s)/g_1(s));$$

this allows expansion of the remaining terms, demonstrating the result.

Q.E.D

Lemma 5.6.1 justifies the approximation

$$F_X(x) = F_{\hat{\Omega}}(\hat{\omega}) = \Phi(\sqrt{n}\hat{\omega}^*) + \phi(\sqrt{n}\hat{\omega})O(n^{-1}), \tag{114}$$

for $\hat{\omega}^* = \hat{\omega} + (n\hat{\omega})^{-1}\log(\hat{z}/\hat{\omega})$; (114) is known as the r^* approximation, since the symbol r is often used in place of $\hat{\omega}$.

5.7. Secant Approximations

Consider a random variable X_n with density $f_n(x) = c_n\sqrt{n}\exp(-np(x))q(x)$, with $p(x) \to \infty$ as $x \to a$ or $x \to b$, and with its unique minimum at $\hat{x}$, and with q positive. Tail probabilities might be calculated for X under conditions weaker than those assumed so far. Change variables to obtain $y(x) = (x-\hat{x})\sqrt{2(p(x) - p(\hat{x}))/(x - \hat{x})^2}$. This is similar to the change of variables in (81). Let $Y_n = y(X_n)$. Then $dy/dx = p'(x)/y$, the density of Y_n is given by $\exp(-np(\hat{x}) - \frac{n}{2}y^2)g(y)$, with $g(y) = q(x(y)) \times y/p'(x(y))$, and the integral to be approximated is $\int_{y_0}^{\infty} \exp(-\frac{n}{2}y^2)g(y)\, dy$. The function $g(y)$ could be expressed as a power series in y, and the resulting integrand could

be treated as in (97). Such a treatment would have the disadvantages outlined at the end of §5.3. Skates (1993) presents an alternative to (109) for such cases, and refers to it as the secant approximation. No cumulant generating function is assumed to exist in this example, and hence usual saddlepoint inversion techniques do not apply. The hard analytical work for these results is contained in the following lemma.

Lemma 5.7.1:   *If g is a function of bounded variation defined on $\mathfrak{R}$, having the expansion $g(y) = \sum_{j=1}^{2\mathsf{j}} a_j y^j / j! + O(y^{2\mathsf{j}+1})$ as $y \to 0$, and j continuous derivatives in a neighborhood of y_0, such that*

$$I_{n,0}(y_0) = \sqrt{n} \int_{y_0}^{\infty} \phi(\sqrt{n}y) g(y) \ dy \qquad (115)$$

is finite for some integer n, and if $I_{0,\mathsf{j}}(-\infty)$ is as in (115) with the range of integration replaced by the whole real line, then

$$\frac{I_{n,0}(y_0)}{I_{n,0}(-\infty)} = 1 - \Phi(\sqrt{n}y_0) + \left(\sum_{j=0}^{\mathsf{j}} a_l n^{-l} + O(n^{-\mathsf{j}-\frac{1}{2}}) \right) \phi(\sqrt{n}y_0)/\sqrt{n}, \qquad (116)$$

for quantities a_l defined in terms of y_0 below. Furthermore, if the functions m_l as defined below, and their derivatives, exist and are bounded on $\mathfrak{R}$, then the error term given above is uniform.

Proof: By symmetry about 0, it suffices to prove this for $y_0 > 0$. Let

$$I_{n,j}(y_0) = \sqrt{n} \int_{y_0}^{\infty} \phi(\sqrt{n}y) g_j(y) \ dy.$$

Following Bleistein (1966), Skates (1993) notes that when the function $g(y)$ is approximated by the affine function of y taking the correct value both at $y(\hat{x})$ and at zero, the integral of the remainder is of a lower order in n. Specifically, let $g_0(y) = g(y)$, and recursively, let $m_j(y) = (g_j(y) - g_j(0))/y$ and $g_{j+1}(y) = m'_j(y)$. Since $g_j(y) = g_j(0) + y m_j(y_0) + y [m_j(y) - m_j(y_0)]$, then for $j \leq \mathsf{j}$,

$$I_{n,j}(y_0) = \sqrt{n} \int_{y_0}^{\infty} \phi(\sqrt{n}y) \{ g_j(0) + y m_j(y_0) + y [m_j(y) - m_j(y_0)] \} \ dy$$

$$= g_j(0)(1 - \Phi(\sqrt{n}y_0)) + \frac{m_j(y_0)}{\sqrt{n}} \phi(\sqrt{n}y_0) +$$

$$\sqrt{n} \int_{y_0}^{\infty} \phi(\sqrt{n}y) y [m_j(y) - m_j(y_0)] \ dy.$$

Integrating by parts, using the identity $\int_a^b u \ dv = uv|_a^b - \int_a^b v \ du$ with $v = \phi(\sqrt{n}y)$,

and $dv = ny\phi(\sqrt{n}y)$ yields for $j < \mathfrak{j}$,

$$I_{n,j}(y_0) = g_j(0)(1 - \Phi(\sqrt{n}y_0)) + n^{-1/2}m_j(y_0)\phi(\sqrt{n}y_0)$$

$$- \left[\exp(-\frac{n}{2}y^2)[m_j(y) - m_j(y_0)]/n\right]_{y_0}^{\infty} + n^{-1}\sqrt{n}\int_{y_0}^{\infty}\phi(\sqrt{n}y)m_j'(y)\,dy$$

$$= g_j(0)(1 - \Phi(\sqrt{n}y_0)) + n^{-1/2}m_j(y_0)\phi(\sqrt{n}y_0) + n^{-1}I_{n,j+1}(y_0),$$

since $m_j'(y) = g_{j+1}(y)$. This integration by parts is valid only in a formal sense; strictly speaking, m_j need not be differentiable outside a neighborhood of y_0. The contribution to the integral outside of this neighborhood, however, converges to 0 geometrically. Inductively, m_l has $\mathfrak{j} - l$ continuous derivatives in a neighborhood of y_0, and g_l has $\mathfrak{j} - l + 1$ continuous derivatives in a neighborhood of y_0. This integration by parts can also be performed in the Riemann-Stieltjes sense for $j = \mathfrak{j}$; here $I_{n,\mathfrak{j}+1}(y_0) = \sqrt{n}\int_{y_0}^{\infty}\phi(\sqrt{n}y)m_{\mathfrak{j}}(dy)$. Since $I_{n,\mathfrak{j}+1}(y_0) = \phi(\sqrt{n}y_0)O(1)$, then $I_{n,0}(y_0) = \sum_{j=0}^{\mathfrak{j}}g_j(0)n^{-j}(1-\Phi(\sqrt{n}y_0))+(\sum_{j=0}^{\mathfrak{j}}m_j(y_0)n^{-j}+O(n^{-\mathfrak{j}-1/2}))\phi(\sqrt{n}y_0)/\sqrt{n}$. This uses the fact that for $s > 0$,

$$1 - \Phi(s) \le \phi(s)/s. \tag{117}$$

Still to be determined are the coefficients $g_j(0)$. Since $g(y) = \sum_{j=0}^{\mathfrak{j}}g^{(j)}(0)y^j/j! + o(y^{\mathfrak{j}})$, then $g_0(0) = g(0)$, by definition, and by induction

$$m_l(y) = \sum_{j=1+2l}^{\mathfrak{j}}g^{(j)}(0)(j+1)\cdots(j+1-2l)y^{j-1-2l}/(j+1)! + o(y^{\mathfrak{j}-1-2l})$$

$$g_l(y) = \sum_{j=2l}^{\mathfrak{j}}g^{(j)}(0)(j+1)\cdots(j+1-2l)y^j/(j+1)! + o(y^{\mathfrak{j}-j}),$$

and hence $g_l(0) = g^{(2l)}(0)2^{-l}/l!$. By Lemma 4.5.1,

$$I_{n,\mathfrak{j}}(-\infty) = \sum_{j=0}^{\mathfrak{j}}g^{(2j)}(0)/((2n)^j j!) + O(n^{-\mathfrak{j}-1/2}).$$

Let the series $\sum_{j=0}^{\mathfrak{j}}a_l n^{-l}$ be the quotient of $\sum_{j=0}^{\mathfrak{j}}m_j(y_0)n^{-j}$ and $\sum_{j=0}^{\mathfrak{j}}g^{(2j)}(0) \times (j!)^{-1}(2n)^{-j}$, satisfying $m_k(y_0) = \sum_{j=0}^{k}g^{(2j)}(0)/(2^j j!)a_{k-j}$. Hence (116) holds. Furthermore, under the last conditions above, uniformity holds since none of the error terms depended on y_0 except through values of m_j.

Q.E.D

Now to apply this lemma to the problem at hand:

Theorem 5.7.2:   *If X_n has the density $f_n(x) = c_n\sqrt{n}\exp(-np(x))q(x)$ on (a,b), where*

1. *p has a single local minimum at $\hat{x}$, and $p(x) \to \infty$ as $x \to a$ or $x \to b$.*
2. *p and q have expansions $p(x) = \sum_{j=0}^{2\mathfrak{j}+1}p_j(x - \hat{x})^j + o((x - \hat{x})^{2\mathfrak{j}+1})$ and $q(x) = \sum_{j=0}^{2\mathfrak{j}}q_j(x - \hat{x})^j + o((x - \hat{x})^{2\mathfrak{j}})$.*

3. $p^{(j+1)}$ and $q^{(j)}$ are continuous in a neighborhood $(x - \delta, x + \delta)$ in which p' has no zeros.

4. $I_n(y) = \int_{\hat{x}}^{b} \exp(-np(x))q(x) \, dx$ converges absolutely

then $Q_X(x_0) = \sum_{j=0}^{j-1} g_j(0)n^{-j}(1 - \Phi(\sqrt{n}y_0)) + \sum_{j=0}^{j} m_j(y_0)n^{-j-1/2}\phi(\sqrt{n}y_0) + O(n^{-j})$. Furthermore, if the functions m_l as defined below, and their derivatives, exist and are bounded on $\mathfrak{R}$, then the error term given above is uniform.

Proof: Apply Lemma 5.7.1 to $g(y) = yq(x(y))/p'(x(y))$. The coefficients $g_j(0)$ and $m_j(y_0)$ are calculated based on the functions $p(x)$ and $q(x)$.

$$\text{Q.E.D}$$

Without loss of generality $p(x)$ and $q(x)$ may be chosen such that $p(\hat{x}) = 0$, $p''(\hat{x}) = 1$, and $q(\hat{x}) = 1$. Then $g_0(0) = 1$, $m_0(y) = (g(y) - 1)/y$, and $a_0 = m_0(y_0) = q(x_0)/p'(x_0) - 1/y_0$. Hence

$$Q_X(x_0) = 1 - \Phi(\sqrt{n}y_0) + \left((q(x_0)/p'(x_0) - 1/y_0) + O(n^{-1/2}) \right) \phi(\sqrt{n}y_0).$$

When the cumulant generating function of a distribution exists, Lemma 5.7.1 can also be used to calculate integrals of the form (97). This alternate method also produces (109).

5.8. Saddlepoint Series for Lattice Variables

Now consider saddlepoint cumulative distribution function approximations for lattice distributions. As before, suppose that $\{Y_n\}$ are independent and identically distributed random variables, but explicitly drop the assumption that they have a density. Rather, suppose that they take values on the lattice $\{a + \Delta\mathfrak{Z}\}$, and that $X = \sum_{k=1}^{n} Y_k/n$. Assume further, without loss of generality, that $a = 0$ and $\Delta = 1$. Again an approximation to the distribution function F_X of X is desired. Recall from §2.7 that characteristic functions for lattice distributions are periodic, thus changing the form of the inversion integrals. This periodicity is inherited by cumulant generating functions.

Consider the inversion integral (33) for $P[x_0 > X \geq x]$. If we deform the path of integration to run through $\mathfrak{R}(\beta) = c$ for some $c > 0$, then one can pass the limit as $x_0 \to \infty$ through the integral, to obtain

$$P[X \geq x] = \frac{1}{2\pi i} \int_{c-i\pi n}^{c+i\pi n} \frac{\exp(n\mathcal{K}_Y(\beta/n) - \beta(x - 1/[2n]))}{\beta} \frac{\beta/[2n]}{\sinh(\beta/[2n])} \, d\beta$$

$$= \frac{1}{2\pi i} \int_{c-i\pi}^{c+i\pi} \frac{\exp(n[\mathcal{K}_Y(\beta) - \beta(x - 1/[2n])])}{\beta} \frac{\beta/2}{\sinh(\beta/2)} \, d\beta$$

$$= \frac{1}{2\pi i} \int_{c-i\pi}^{c+i\pi} \frac{\exp(-\mathcal{K}_U(\beta) + n[\mathcal{K}_Y(\beta) - \beta(x - 1/[2n])])}{\beta} \, d\beta$$

Here $\mathcal{K}_U = \log(\sinh(\beta/2)) - \log(\beta/2)$ is the cumulant generating function for the uniform variable on $(-1/2, 1/2)$. One might be tempted to use usual saddlepoint

cumulative distribution function methods (98), (106), or (109), with continuity correction and the cumulant generating function corrected by subtracting $\mathcal{K}_U(\beta)$; however, since this does not have the factor n the behavior of this approximation is an open question. Daniels (1987) includes $\exp(-\mathcal{K}_U(\beta))$ in the n-independent function ϖ in Watson's Lemma. Approximation (109) still satisfies (110), with $\hat{z} = 2\sqrt{\mathcal{K}_Y''(\hat{\beta})}\sinh(\hat{\beta}/2)$, and $\hat{\beta}$ satisfying

$$\mathcal{K}_Y'(\hat{\beta}) = x - 1/[2n]. \tag{118}$$

One might again try to add terms to (109) in order to construct an approximation whose error, in the lattice case, converges to zero more quickly as n increases. Kolassa (1996b) provides such an approximation for conditional distributions; these results are discussed in §7.4. An unconditional higher–order counterpart to (109) is a special case of the expansion of §7.4.

In order to adapt approximation (106) of Robinson (1982), first calculate $\hat{\beta}$ from the continuity–corrected data, as in (118). Now note that $Q_X(x)$ is still given by (108) if the integrand is multiplied by

$$\frac{\beta}{2\sinh(\beta/2)} = \frac{\hat{\beta}}{2\sinh(\hat{\beta}/2)} - \tfrac{1}{4}\left(\hat{\beta}\coth(\hat{\beta}/2) - 2\right)\operatorname{csch}(\hat{\beta}/2)\left(\beta - \hat{\beta}\right) + O(\beta - \hat{\beta})^2.$$

The resulting approximation, a special case of the multivariate approximation presented by Kolassa (2003a), is

$$\tfrac{1}{2}\hat{\beta}\operatorname{csch}(\hat{\beta}/2)\left(\exp([\hat{z}^2 - \hat{\omega}^2]\,n/2)\,\bar{\Phi}(\hat{z}\,\sqrt{n})\left(\hat{\beta}\coth(\hat{\beta}/2)/2 - \hat{z}^3\,n\,\rho_{Y,3}/6\right) + \right.$$

$$\left. \phi(\hat{\omega}\,\sqrt{n})\left(1/\hat{z} - \tfrac{1}{2}\hat{\beta}\coth(\hat{\beta}/2)/\hat{z} + \rho_{Y,3}[\hat{z}^2\,n - 1]/6\right)/\sqrt{n}\right).$$

5.9. Exercises

1. Numerically calculate the cumulative distribution function for the convolution of 5, 10, and 20 copies of the distribution of Kolassa (1992) whose cumulative distribution function is given by

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \Phi(\sqrt{x}) & \text{otherwise} \end{cases}$$

at its mean and one, two, and three standard deviations above the mean, and compare to the Robinson and Lugannani and Rice approximations. You may find it useful to manipulate the gamma distribution to do the exact calculations. You can get Fortran routines by sending electronic mail twice to `statlib@lib.stat.cmu.edu`, with `send 147 from griffiths-hill` and `send acm291 from griffiths-hill` as subject lines.

2. Verify formula (117), by considering the integral representation of $1 - \Phi(s)$, and changing variables so that $\phi(s)/s$ may be factored out, leaving an integrand depending on s. This integrand may be bounded by something that integrates to unity.

3. Suppose that X is a random variable with a binomial distribution, with 15 trials and success probability π. In the below calculations, make sure that you properly account for discreteness of X.

   1. Use the Lugannani and Rice approximation to approximate the observed significance level for the hypothesis $\pi = \frac{1}{2}$ vs. the alternative $\pi > \frac{1}{2}$, if 9 successes are observed.

   2. Use the Lugannani and Rice approximation to approximate the standard confidence interval for π, if 9 successes are observed.

4. Describe the relationship between the integrals I_j and $J_j(v)$ of §5.2.

5. Verify that the term of size $O(n^{-3/2})$ in the expansion of Theorem 5.3.1, calculated according to the method of §5.5, agrees with the value given in (109).

6. Prove (111).

6

Multivariate Expansions

Edgeworth and saddlepoint expansions also have analogues for distributions of random vectors. As in the univariate case these expansions will be derived with reference to characteristic functions and cumulant generating functions, and hence these will be defined first. Subsequently Edgeworth density approximations will be defined. Just as in the univariate case, the Edgeworth approximation to probabilities that a random vector lies in a set is the integral of the Edgeworth density over that set; however, since sets of interest are usually not rectangular, theorems for the asymptotic accuracy of these approximations are difficult to prove. These proofs are not presented here. Approximation for variables on a multivariate lattice are discussed. Multivariate saddlepoint approximations are also defined, by a multivariate extension of steepest descent methods. These methods are also used to approximate conditional probabilities.

Results presented in this chapter will have their primary applications in justifying techniques presented in the following two chapters.

6.1. Multivariate Generating Functions

Define the multivariate characteristic function of $\boldsymbol{X} = (X^1, ..., X^{\mathfrak{k}})^\top$ as

$$\zeta_{\boldsymbol{X}}(\boldsymbol{\beta}) = \mathrm{E}\left[\exp(i\boldsymbol{\beta}^\top \boldsymbol{X})\right] = \mathrm{E}\left[\exp(i\sum_{j=1}^{\mathfrak{k}} \beta_j X^j)\right].$$

This always exists over $\mathfrak{R}^{\mathfrak{k}}$. Define the multivariate moment generating function of $\boldsymbol{X}$ as $\mathcal{M}_{\boldsymbol{X}}(\boldsymbol{\beta}) = \mathrm{E}\left[\exp(\boldsymbol{\beta}^\top \boldsymbol{X})\right]$. As in the univariate case, $\mathcal{M}_{\boldsymbol{X}}$ exists at $\mathbf{o} = (0, \dots, 0)$. Saddlepoint expansions will be available when $\mathcal{M}_{\boldsymbol{X}}$ exists in a neighborhood $\mathfrak{Q} \subset \mathfrak{R}^{\mathfrak{k}}$ of $\mathbf{o}$.

Many of the useful properties of univariate moment generating functions hold in the multivariate case as well. For instance, if $\boldsymbol{A}$ and $\boldsymbol{B}$ are matrices and $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent random vectors such that the combination $\boldsymbol{A}\boldsymbol{X} + \boldsymbol{B}\boldsymbol{Y}$ is defined,

(that is, $\boldsymbol{A}$ and $\boldsymbol{B}$ have the same number of rows, and the numbers of columns equal the lengths of $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively), then

$$
\begin{aligned}
\mathcal{M}_{\boldsymbol{AX}+\boldsymbol{BY}}(\boldsymbol{\beta}) &= \mathrm{E}\left[\exp(\boldsymbol{\beta}^\top(\boldsymbol{AX}+\boldsymbol{BY}))\right] = \mathrm{E}\left[\exp(\boldsymbol{\beta}^\top\boldsymbol{AX})\exp(\boldsymbol{\beta}^\top\boldsymbol{BY}))\right] \\
&= \mathrm{E}\left[\exp(\boldsymbol{\beta}^\top\boldsymbol{AX})\right]\mathrm{E}\left[\exp(\boldsymbol{\beta}^\top\boldsymbol{BY}))\right] = \mathcal{M}_{\boldsymbol{X}}(\boldsymbol{A}^\top\boldsymbol{\beta})\mathcal{M}_{\boldsymbol{Y}}(\boldsymbol{B}^\top\boldsymbol{\beta}).\text{(119)}
\end{aligned}
$$

Expand the moment generating function as

$$
\mathcal{M}_{\boldsymbol{X}}(\boldsymbol{\beta}) = \sum_{j=0}^{\infty} \sum_{s\in S(j)} \frac{1}{j!}\mu^{s_1\cdots s_j}\beta_{s_1}\cdots\beta_{s_j}
$$

where $S(j) = \{1,\ldots,\mathfrak{k}\}^j$, the set of vectors of integers with j components and all entries between 1 and $\mathfrak{k}$. This series by itself is not enough to uniquely define the μ's, since all coefficients μ with the same set of indices permuted correspond to the same product of elements of $\boldsymbol{\beta}$. Adding the requirement that the μ's have the same value if their indices are permuted, however, does allow the above expansion to uniquely determine the coefficients. As in the univariate case, $\mu^{s_1\cdots s_j} = \mathrm{E}\left[X^{s_1}\cdots X^{s_j}\right]$. For example, $\mu^1 = \mathrm{E}\left[X^1\right]$, $\mu^2 = \mathrm{E}\left[X^2\right]$, $\mu^{11} = \mathrm{E}\left[X^1 X^1\right]$, and $\mu^{12} = \mathrm{E}\left[X^1 X^2\right]$.

One may also define the multivariate cumulant generating function $\mathcal{K}_{\boldsymbol{X}}(\boldsymbol{\beta})$ using the moment generating function in analogy with the univariate case: $\mathcal{K}_{\boldsymbol{X}}(\boldsymbol{\beta}) = \log(\mathcal{M}_{\boldsymbol{X}}(\boldsymbol{\beta}))$. Again this function will generate the cumulants. Expand $\mathcal{K}_{\boldsymbol{X}}$ as

$$
\mathcal{K}_{\boldsymbol{X}}(\boldsymbol{\beta}) = \sum_{j=1}^{\infty} \sum_{s\in S(j)} \frac{1}{j!}\kappa_{\boldsymbol{X}}^{s_1\cdots s_j}\beta_{s_1}\cdots\beta_{s_j}, \tag{120}
$$

with the same definition for the inner sum. As above, require that coefficients $\kappa_{\boldsymbol{X}}$ with the same indices up to permutation have the same value. This is enough to define the multivariate cumulants $\kappa_{\boldsymbol{X}}$ uniquely via the series expansion. Alternatively, as a generalization of (12),

$$
\kappa_{\boldsymbol{X}}^{s_1\cdots s_j} = \frac{d^j}{d\beta_{s_1}\cdots d\beta_{s_j}}\mathcal{K}_{\boldsymbol{X}}(\boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{o}} = (-i)^{s_1+\cdots+s_j}\frac{d^j}{d\beta_{s_1}\cdots d\beta_{s_j}}\log(\zeta_{\boldsymbol{X}}(\boldsymbol{\beta}))|_{\boldsymbol{\beta}=\boldsymbol{o}}
$$

.

As in the univariate case, these cumulants may be expressed in terms of the moments. For example, $\kappa_{\boldsymbol{X}}^1 = \mathrm{E}\left[X^1\right]$, $\kappa_{\boldsymbol{X}}^2 = \mathrm{E}\left[X^2\right]$, $\kappa_{\boldsymbol{X}}^{11} = \mathrm{Var}\left[X^1\right]$, and $\kappa_{\boldsymbol{X}}^{12} = \mathrm{Cov}[X^1, X^2]$. If the components of $\boldsymbol{X}$ are independent, then

$$
\begin{aligned}
\mathcal{K}_{\boldsymbol{X}}(\boldsymbol{\beta}) &= \log\left(\mathrm{E}\left[\exp\left(\sum_{j=1}^{\mathfrak{k}}\beta_j X^j\right)\right]\right) = \log\left(\prod_{j=1}^{\mathfrak{k}}\mathrm{E}\left[\exp(\beta_j X^j)\right]\right) \\
&= \sum_{j=1}^{\mathfrak{k}}\log\left(\mathrm{E}\left[\exp(\beta_j X^j)\right]\right) = \sum_{j=1}^{\mathfrak{k}}\mathcal{K}_{X^j}(\beta_j)
\end{aligned}
$$

Hence in the case of independent components mixed cumulants are zero. That is, $\kappa_{\boldsymbol{X}}^{s_1\cdots s_j} = 0$ if the indices are not all the same.

Multivariate cumulants have some of the same invariance properties as have univariate cumulants, as described in §2.1. Applying (119) with $\boldsymbol{B} = \mathbf{o}$, and taking logarithms, $\mathcal{K}_{\boldsymbol{AX}}(\boldsymbol{\beta}) = \mathcal{K}_{\boldsymbol{X}}(\boldsymbol{A}^\top \boldsymbol{\beta})$. If λ denotes the cumulants of the linear combination $\boldsymbol{AX}$, and if the element in the i, j element of $\boldsymbol{A}^\top$ is a_i^j, then

$$\mathcal{K}_{\boldsymbol{AX}}(\boldsymbol{\beta}) = \sum_{j=1}^{\infty} \sum_{\boldsymbol{s} \in S(j)} \frac{1}{j!} \kappa_{\boldsymbol{X}}^{s_1 \cdots s_j} (\sum_{r_1} a_{s_1}^{r_1} \beta_{r_1}) \cdots (\sum_{r_j} a_{s_j}^{r_j} \beta_{r_j})$$

$$= \sum_{j=1}^{\infty} \sum_{\boldsymbol{r} \in S(j)} \frac{1}{j!} \sum_{\boldsymbol{s} \in S(j)} \left[\kappa_{\boldsymbol{X}}^{s_1 \cdots s_j} a_{s_1}^{r_1} \cdots a_{s_j}^{r_j} \right] \beta_{r_1} \cdots \beta_{r_j}.$$

Hence $\lambda^{r_1 \cdots r_j} = \sum_{\boldsymbol{s} \in S(j)} \left[\kappa_{\boldsymbol{X}}^{s_1 \cdots s_j} a_{s_1}^{r_1} \cdots a_{s_j}^{r_j} \right]$.

6.2. Multivariate Inversion Integrals

We require the following multivariate extension of the univariate theorem on inverting characteristic functions to recover the underlying cumulative distribution function, presented in §2.4. This proof is also found in Billingsley (1986).

Theorem 6.2.1:   *If a probability distribution* $\mathrm{P}\,[\cdot]$ *on* $\mathfrak{R}^{\mathfrak{k}}$ *corresponds to a characteristic function* ζ, *and the rectangle* $\mathfrak{B} = \{\boldsymbol{x} \in \mathfrak{R}^{\mathfrak{k}} | x^j \in [b^{1j}, b^{2j}] \ \forall j = 1, \ldots, \mathfrak{k}\}$ *has a boundary*

$$\partial\mathfrak{B} = \{\boldsymbol{x} \in \mathfrak{R}^{\mathfrak{k}} | x^j \in [b^{1j}, b^{2j}] \ \forall j = 1, \ldots, \mathfrak{k}, x^j = b^{1j} \text{ or } b^{2j} \text{ for some } j\}$$

that is assigned zero probability, then

$$\mathrm{P}\,[\mathfrak{B}] = \lim_{\Theta \to \infty} \int_{-\Theta}^{\Theta} \cdots \int_{-\Theta}^{\Theta} \prod_{j=1}^{\mathfrak{k}} \frac{\exp(-i\beta_j b^{1j}) - \exp(-i\beta_j b^{2j})}{i\beta_j} \zeta(\boldsymbol{\beta}) \frac{d\boldsymbol{\beta}}{(2\pi)^{\mathfrak{k}}}. \qquad (121)$$

The factors in the product over j *in the integrand have a removable singularity at* $\beta_j = 0$; *as before substitute the value* $b^{2j} - b^{1j}$.

Proof: Let

$$I_\Theta = \int_{-\Theta}^{\Theta} \cdots \int_{-\Theta}^{\Theta} \prod_{j=1}^{\mathfrak{k}} \frac{\exp(-i\beta_j b^{1j}) - \exp(-i\beta_j b^{2j})}{i\beta_j} \zeta(\boldsymbol{\beta}) \frac{d\boldsymbol{\beta}}{(2\pi)^{\mathfrak{k}}}$$

be the quantity inside the limit operation in (121). After expressing this characteristic function as an integral over possible values of $\boldsymbol{x}$,

$$I_\Theta = \int_{-\Theta}^{\Theta} \cdots \int_{-\Theta}^{\Theta} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{j=1}^{\mathfrak{k}} \frac{\exp(i\beta_j(x^j - b^{1j})) - \exp(i\beta_j(x^j - b^{2j}))}{i\beta_j} \, dF(\boldsymbol{x}) \, \frac{d\boldsymbol{\beta}}{(2\pi)^{\mathfrak{k}}}.$$

The integrand above is bounded by $\prod_{j=1}^{\mathfrak{k}} |b^{2j} - b^{1j}|$ by Theorem 2.5.3. The order of integration can therefore be interchanged, and the complex exponential can be

expressed in terms of sines and cosines:

$$I_\Theta = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{-\Theta}^{\Theta} \cdots \int_{-\Theta}^{\Theta} \prod_{j=1}^{\ell} \left(\frac{\cos(\beta_j(x^j - b^{1j})) + i \sin(\beta_j(x^j - b^{2j}))}{i\beta_j} \right.$$
$$\left. - \frac{\cos(\beta_j(x^j - b^{2j})) + i \sin(\beta_j(x^j - b^{2j}))}{i\beta_j} \right) \frac{d\boldsymbol{\beta}}{(2\pi)^{\ell}} \, dF(\boldsymbol{x}).$$

Since the cosine terms involve odd order terms in β_j these terms integrate to zero:

$$I_\Theta = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{-\Theta}^{\Theta} \cdots \int_{-\Theta}^{\Theta} \prod_{j=1}^{\ell} \frac{\sin(\beta_j(x^j - b^{1j})) - \sin(\beta_j(x^j - b^{1j}))}{\beta_j} \frac{d\boldsymbol{\beta}}{(2\pi)^{\ell}} \, dF(\boldsymbol{x})$$

$$= \int_{-\Theta}^{\Theta} \cdots \int_{-\Theta}^{\Theta} \prod_{j=1}^{\ell} \frac{2S(\beta \, |x^j - b^{1j}|) \operatorname{sgn}(x^j - b^{1j}) - 2S(\beta \, |x^j - b^{2j}|) \operatorname{sgn}(x^j - b^{2j})}{2\pi} \, dF(\boldsymbol{x})$$

where again $S(\theta) = \int_0^\theta \sin(x)/x \, dx$. Since the integrand is bounded, one can pass to the limit:

$$\lim_{\Theta \to \infty} I_\Theta = (S(\infty)/\pi)^{\ell} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{j=1}^{\ell} (\operatorname{sgn}(x^j - b^{1j}) - \operatorname{sgn}(x^j - b^{2j})) \, dF(\boldsymbol{x})$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{j=1}^{\ell} (\operatorname{sgn}(x^j - b^{1j}) - \operatorname{sgn}(x^j - b^{2j})) \, dF(\boldsymbol{x}),$$

since $S(\infty)/\pi = 1$ as was found in §2.4.

<div align="right">Q.E.D</div>

A theorem for recovering multivariate density functions also exists:

Theorem 6.2.2:   *If a random vector has the characteristic function ζ satisfying*

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} |\zeta(\boldsymbol{\beta})| \, d\boldsymbol{\beta} < \infty \tag{122}$$

then its density is given by

$$f_Y(\boldsymbol{y}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp(-i\boldsymbol{\beta}\boldsymbol{y})\zeta(\boldsymbol{\beta}) \frac{d\boldsymbol{\beta}}{(2\pi)^{\ell}}. \tag{123}$$

Proof: By (122) one can replace the limit of proper integrals as $\Theta \to \infty$ by the corresponding improper integral over the real line in (121). As in §2.4 replace the difference in exponentials by an integral. Then

$$F(\boldsymbol{b}^2) - F(\boldsymbol{b}^1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{b^{11}}^{b^{21}} \cdots \int_{b^{1\ell}}^{b^{2\ell}} \exp(-i\boldsymbol{\beta}\boldsymbol{y}) \, d\boldsymbol{y}\zeta(\boldsymbol{\beta}) \frac{d\boldsymbol{\beta}}{(2\pi)^{\ell}}$$

$$= \int_{b^{11}}^{b^{21}} \cdots \int_{b^{1\ell}}^{b^{2\ell}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp(-i\boldsymbol{\beta}\boldsymbol{y})\zeta(\boldsymbol{\beta}) \frac{d\boldsymbol{\beta}}{(2\pi)^{\ell}} \, d\boldsymbol{y}$$

Interchange of the order of integration is justified by the absolute convergence of the integral (122). Hence the density is (123).

<div align="right">Q.E.D</div>

The bounded convergence theorem implies that the resulting density is continuous. A condition like integrability of the first derivative of the density implies the integrability condition of the cumulant generating function (122). The presence of probability atoms implies that this condition (122) is violated.

The inversion integrand in (123) is the $\mathfrak{k}$-order derivative of the inversion integrand in (121) with respect to each component of $\boldsymbol{b}$, and evaluated with $b^{1j} = b^{2j}$. This will become important in our discussion of saddlepoint series.

For completeness, and because it will be useful in the later development of multivariate saddlepoint density approximations, the following analogue of Lemma 2.4.4 is stated:

Lemma 6.2.3:   *The characteristic function ζ of a random vector satisfies*

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} |\zeta(\boldsymbol{\beta})|^{\mathfrak{r}} \ d\boldsymbol{\beta} < \infty$$

for some $\mathfrak{r} > 1$ if and only if the density of a j-fold convolution of the random variable with itself exists and is bounded, for some integer j.

Proof: Left as an exercise.

There is also a multivariate inversion counterpart to (32) for the probability mass function for a random vector $\boldsymbol{X}$ confined to a multivariate lattice:

Theorem 6.2.4:   *Suppose that a random vector $\boldsymbol{X}$ of length $\mathfrak{k}$ is confined to a generalized lattice in the following sense: The distribution of $X^{\mathfrak{k}}$ is confined to a lattice of form $a^{\mathfrak{k}} + \Delta_{\mathfrak{k}} 3$, and for each $j < \mathfrak{k}$ the distribution of X^j is confined to a lattice of form $a^j(x^{j+1}, \ldots, x^{\mathfrak{k}}) + \Delta_j 3$. Then*

$$\prod_{j=1}^{\mathfrak{k}} \frac{\Delta_j}{2\pi} \int_{-\pi/\Delta_1}^{\pi/\Delta_1} \cdots \int_{-\pi/\Delta_{\mathfrak{k}}}^{\pi/\Delta_{\mathfrak{k}}} \exp(-i\boldsymbol{\beta}\boldsymbol{x})\zeta(\boldsymbol{\beta}) \ d\boldsymbol{\beta} = \mathrm{P}\left[\boldsymbol{X} = \boldsymbol{x}\right]. \qquad (124)$$

Proof: By Fubini's theorem,

$$\prod_{j=1}^{\mathfrak{k}} \frac{\Delta_j}{2\pi} \int_{-\pi/\Delta_1}^{\pi/\Delta_1} \cdots \int_{-\pi/\Delta_{\mathfrak{k}}}^{\pi/\Delta_{\mathfrak{k}}} \sum_{l_1=-\infty}^{\infty} \cdots \sum_{l_{\mathfrak{k}}=-\infty}^{\infty} \mathrm{P}\left[\boldsymbol{X} = \boldsymbol{x} + \Delta l\right] \exp(i\boldsymbol{\beta}(\boldsymbol{x} + \Delta l) - i\boldsymbol{\beta}\boldsymbol{x}) \ d\boldsymbol{\beta}$$

$$= \prod_{j=1}^{\mathfrak{k}} \frac{\Delta_j}{2\pi} \int_{-\pi/\Delta_1}^{\pi/\Delta_1} \cdots \int_{-\pi/\Delta_{\mathfrak{k}}}^{\pi/\Delta_{\mathfrak{k}}} \mathrm{P}\left[\boldsymbol{X} = \boldsymbol{x}\right] \ d\boldsymbol{\beta}$$

The left hand side of the above equation is equal to the left hand side of (124) after expanding the definition of the characteristic function; the right hand side of the above equation is trivially equal to the right hand side of (124).

<div align="right">Q.E.D</div>

The above generalized lattice definition generalizes the conventional definition as the set

$$\boldsymbol{a} + \Delta_1 \mathfrak{Z} \times \cdots \times \Delta_{\mathfrak{k}} \mathfrak{Z}$$

in that it allows the starting point for the lattice to depend on other components of the vector $\boldsymbol{x}$. This generalization will prove useful in a later chapter on conditional cumulative distribution function approximations for lattice random variables.

Finally, Kolassa (2003a) demonstrates the following analogue to (96):

Lemma 6.2.5: *Suppose that $T^1, \ldots, T^{\mathfrak{k}}$ has a cumulant generating function $\mathcal{K}$, and is either continuous (that is, $\boldsymbol{T}$ has a bounded density, in which case $\psi(\beta)$ is set to β, $K = \infty$, and $\boldsymbol{t} = \boldsymbol{t}^*$) or confined to an integer lattice (in which case $\psi(\beta)$ is set to $2 \sinh(\beta/2)$, $K = \pi$, and $\boldsymbol{t} = \boldsymbol{t}^* - \frac{1}{2}\boldsymbol{1}$). Choose $\boldsymbol{c} > \boldsymbol{o}$ in the domain of $\mathcal{K}$. Then the relation*

$$\mathrm{P}\left[\boldsymbol{T} \geq \boldsymbol{t}\right] = \int_{\boldsymbol{c}-iK}^{\boldsymbol{c}+iK} \frac{\exp(\mathcal{K}(\boldsymbol{\beta}) - \boldsymbol{\beta}^\top \boldsymbol{t}^*)}{(2\pi i)^{\mathfrak{k}} \prod_{j=1}^{\mathfrak{k}} \psi(\beta_j)} \, d\boldsymbol{\beta}, \tag{125}$$

holds.

Proof: When $\boldsymbol{T}$ has a density, standard Fourier inversion techniques imply that for $K = \infty$,

$$f_{\boldsymbol{T}}(\boldsymbol{t}) = \int_{\boldsymbol{c}-iK}^{\boldsymbol{c}+iK} (2\pi i)^{-\mathfrak{k}} \exp(\mathcal{K}(\boldsymbol{\beta}) - \beta_j t^j) \, d\boldsymbol{\beta},$$

and

$$\int_{\boldsymbol{c}-i\infty}^{\boldsymbol{c}+i\infty} \int_{\boldsymbol{u} \geq \boldsymbol{t}} \left| \exp(\mathcal{K}(\boldsymbol{\beta}) - \beta_j u^j) \right| \, d\boldsymbol{u} d\boldsymbol{\beta} = \int_{\boldsymbol{c}-i\infty}^{\boldsymbol{c}+i\infty} \int_{\boldsymbol{u} \geq \boldsymbol{t}} \left| \exp(\mathcal{K}(\boldsymbol{\beta})) \right| \exp(-c_j u^j) \, d\boldsymbol{u} d\boldsymbol{\beta}$$

$$= \left[\int_{\boldsymbol{c}-i\infty}^{\boldsymbol{c}+i\infty} \left| \exp(\mathcal{K}(\boldsymbol{\beta})) \right| \, d\boldsymbol{\beta} \right] \left[\int_{\boldsymbol{u} \geq \boldsymbol{t}} \exp(-c_j u^j) \, d\boldsymbol{u} \right].$$

The first factor above is finite since $\boldsymbol{T}$ has a density, and the second is $\exp(c_j t^j) \times [\prod_{j=1}^{\mathfrak{k}} c_j]^{-1}$. By Fubini's theorem, the result follows by interchanging the order of integration with respect to $\boldsymbol{t}$ and with respect to $\boldsymbol{\beta}$, as long as all components of $\boldsymbol{c}$ are positive. When $\boldsymbol{T}$ is confined to an integer lattice, $K = \pi$ and the integration with respect to $\boldsymbol{u}$ is replaced by summation.

<div align="right">Q.E.D</div>

The requirement that $\boldsymbol{c} > \boldsymbol{o}$, needed to justify Fubini's theorem, is important; because of this requirement, tail probabilities associated with $\boldsymbol{t}$ for which one or more of the components of $\hat{\boldsymbol{\beta}}$ are negative must be calculated by applying the approximation to random vectors with some of their components negated, and differencing.

6.3. Edgeworth Series

In this section the multivariate counterpart of the Edgeworth approximations to densities and distribution functions are presented. These approximations can be expressed as refinements to the normal approximations, and so these normal approximations are presented first. Proofs are deferred until a later section.

Assume that $\boldsymbol{X}$ is a random vector such that $\mathrm{E}[\boldsymbol{X}] = \mathbf{o}$. Let Σ represent the covariance matrix of $\boldsymbol{X}$, and let $\Sigma^{-1} = [\kappa_{ij}]$ represent its inverse. If the distribution of $\boldsymbol{X}$ has a density, define the normal approximation to this density to be

$$\phi(\boldsymbol{x}, \Sigma) = \frac{\exp(-\frac{1}{2}\sum_{j=1}^{\mathfrak{k}}\sum_{i=1}^{\mathfrak{k}} x^j x^i \kappa_{ji})}{(2\pi)^{\mathfrak{k}/2}\sqrt{\det[\Sigma]}}.$$

Define its cumulative distribution function $F_{\boldsymbol{X}}(\boldsymbol{x})$ to be the normal approximation for the cumulative distribution function $\mathrm{P}\left[\boldsymbol{X} \leq \boldsymbol{x}\right] = \mathrm{P}\left[X^j \leq x^j\ \forall j\right]$ and define the normal approximation for the cumulative distribution function to be

$$\Phi(\boldsymbol{x}, \Sigma) = \int_{-\infty}^{x^1} \cdots \int_{-\infty}^{x^{\mathfrak{k}}} \phi(\boldsymbol{y}, \Sigma)\ d\boldsymbol{y}.$$

Just as in the univariate case, the normal approximation uses information in the first and second order cumulants to approximate the distribution of interest. Also as in the univariate case, higher order cumulants can be used to correct this series, producing a higher-order analogue to the Edgeworth series.

Multivariate Fourier inversion techniques will be used to approximate a density inversion integral. As before, the exponent will be divided into a quadratic term in $\boldsymbol{\beta}$, and a higher-order term in $\boldsymbol{\beta}$. The higher order term will be exponentiated, and the result will be integrated term by term.

Begin by defining higher-order analogues to pseudo-moments. Let $\kappa^{s_1\cdots s_j}$ represent the cumulants of a random vector $\boldsymbol{X}$, as defined in (120). Let the coefficients μ^* be defined from the power series expansion

$$1 + \sum_{j=3}^{\infty}\sum_{s\in S(j)} \mu^{*s}\beta_{s_1}\cdots\beta_{s_j}/j! = \exp\left(\sum_{j=3}^{\infty}\sum_{s\in S(j)} \kappa^s\beta_{s_1}\cdots\beta_{s_j}/j!\right)$$

$$= 1 + \left(\frac{1}{3!}\kappa^{ijk}\beta_i\beta_j\beta_k + \frac{1}{4!}\kappa^{ijkl}\beta_i\beta_j\beta_k\beta_l + \cdots\right)$$

$$+ \frac{1}{2}\left(\frac{1}{3!}\kappa^{ijk}\beta_i\beta_j\beta_k + \frac{1}{4!}\kappa^{ijkl}\beta_i\beta_j\beta_k\beta_l + \cdots\right)^2 + \cdots \qquad (126)$$

again with the convention that the μ^*'s have the same value if their indices are permuted, and again where the inner sum is over all j-long vectors of integers $\boldsymbol{s}$ between 1 and $\mathfrak{k}$. For a vector $\boldsymbol{s} = (s_1, \ldots, s_{\mathfrak{k}})$ let $\kappa^{\boldsymbol{s}}$ and $\mu^{*\boldsymbol{s}}$ denote the cumulants and pseudo-moments with superscripts $s_1, \ldots, s_{\mathfrak{k}}$. Although $\boldsymbol{s}$ is a vector these quantities do not depend on the order of the components of $\boldsymbol{s}$. Then, still assuming

$E[\boldsymbol{X}] = \mathbf{o}$, the characteristic function of $f_{\boldsymbol{X}}(\boldsymbol{x})$ is

$$\exp(\tfrac{1}{2}\boldsymbol{\beta}^{\top}\boldsymbol{\Sigma}\boldsymbol{\beta})\left\{1 + \sum_{j=1}^{\infty}\sum_{s\in S(j)}\frac{i^{j}}{j!}\mu^{*s}\beta_{s_1}\cdots\beta_{s_j}\right\}.$$

Applying (123) to the multivariate normal case, and differentiating,

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp(\beta\boldsymbol{z})\phi^{s_1\cdots s_l}(\boldsymbol{z};\Sigma)\,d\boldsymbol{z} = \beta_{s_1}\cdots\beta_{s_l}(-1)^{l}\exp(\tfrac{1}{2}\beta_j\Sigma^{jk}\beta_k), \qquad (127)$$

where again superscripts on $\phi(\boldsymbol{z};\Sigma)$ refer to derivatives with respect to the indicated components of its argument. Relation (127) generalizes (43). The density of $\boldsymbol{X}$ can be approximated by an Edgeworth series derived from the formal series expansion

$$\sum_{j=0}^{\infty}\sum_{s\in S(j)}\frac{1}{j!}\mu^{*s_1\cdots s_j}(-1)^{j}\frac{d^{j}}{dx^{s_1}\cdots dx^{s_j}}\phi(\boldsymbol{x},\Sigma) \qquad (128)$$

where the $\mu^{*s_1\cdots s_j}$ are sums of products of the joint cumulants. This can also be defined in terms of the generalized Hermite polynomials

$$h_{s_1,\ldots,s_j}(\boldsymbol{x};\mathcal{K}''(\hat{\boldsymbol{\beta}})) = (-1)^{j}\frac{d^{j}}{dx^{s_1}\cdots dx^{s_j}}\phi(\boldsymbol{x},\Sigma)/\phi(\boldsymbol{x},\Sigma). \qquad (129)$$

Integrating with respect to $\boldsymbol{x}$, the cumulative distribution function of $\boldsymbol{X}$ can be approximated by an Edgeworth series derived from the formal series expansion

$$\sum_{j=1}^{\infty}\sum_{s\in S(j)}\frac{1}{j!}\mu^{*s_1\cdots s_j}(-1)^{j}\frac{d^{j}}{dx^{s_1}\cdots dx^{s_j}}\Phi(\boldsymbol{x},\Sigma). \qquad (130)$$

Kolassa (2003a) provides recursive formulae for calculating derivatives of $\Phi(\boldsymbol{x},\Sigma)$.

When the distribution of $\boldsymbol{X}$ depends on a parameter n, the cumulants $\kappa_{\boldsymbol{X}}^{s}$ then also depend on n. When $\boldsymbol{X}$ is the standardized sum of independent and identically distributed components, and each of these components has as its $s_1\cdots s_j$ cumulant $\kappa_{\boldsymbol{Y}}^{s_1\cdots s_j}$, then $\kappa_{\boldsymbol{X}}^{s} = n^{1-|s|/2}\kappa_{\boldsymbol{Y}}^{s}$, and in this case discarding terms in (128) of size $o(n^{-j/2})$ leaves an Edgeworth series $e_{\mathrm{j}}(\boldsymbol{x},\boldsymbol{\kappa}^{n})$ such that

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = e_{\mathrm{j}}(\boldsymbol{x},\boldsymbol{\kappa}^{n}) + o(n^{-j/2}). \qquad (131)$$

Under other assumptions on the distribution for $\boldsymbol{X}$, $\kappa_{\boldsymbol{X}}^{s} = O(n^{1-|s|/2})$ (Chambers, 1967). Here $|\boldsymbol{s}|$ denotes the number of elements in $\boldsymbol{s}$. McCullagh (1987) treats this multivariate case in detail. Proofs demonstrating the order of the error are provided in §6.6.

6.4. Saddlepoint Approximations

As in the univariate case, the first order saddlepoint density approximation can be calculated directly from the normal approximation. Suppose that a random vector $\boldsymbol{X}$ has a density $f_{\boldsymbol{X}}(\boldsymbol{x})$ and a multivariate cumulant generating function $\mathcal{K}_{\boldsymbol{X}}(\boldsymbol{\gamma})$. Embed $\boldsymbol{X}$ in an exponential family, to define a density $f_{\boldsymbol{X}}(\boldsymbol{x}, \boldsymbol{\gamma}) = \exp(\boldsymbol{\gamma}^{\top}\boldsymbol{x} - \mathcal{K}_{\boldsymbol{X}}(\boldsymbol{\gamma}))f_{\boldsymbol{X}}(\boldsymbol{x})$. As before we desire an approximation $g_n(\boldsymbol{x})$ to $f_{\boldsymbol{X}}(\boldsymbol{x})$, and proceed by finding $\hat{\boldsymbol{\gamma}}$ dependent on $\boldsymbol{x}$ such that $f_{\boldsymbol{X}}(\boldsymbol{x}, \hat{\boldsymbol{\gamma}})$ has mean $\boldsymbol{x}$ and approximating $f_{\boldsymbol{X}}(\boldsymbol{x}, \hat{\boldsymbol{\gamma}})$ by a normal density with mean 0. Then $\hat{\boldsymbol{\gamma}}$ is defined by the multivariate saddlepoint equation $\mathcal{K}'_{\boldsymbol{X}}(\hat{\boldsymbol{\gamma}}) = \boldsymbol{x}$. The Gaussian approximation to $f_{\boldsymbol{X}}(\boldsymbol{x}, \hat{\boldsymbol{\gamma}})$ is $(2\pi)^{(-\ell/2)} \det\left[\mathcal{K}''_{\boldsymbol{X}}(\hat{\boldsymbol{\gamma}})\right]^{-1/2}$. By (131),

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \exp(\mathcal{K}_{\boldsymbol{X}}(\hat{\boldsymbol{\gamma}}) - \hat{\boldsymbol{\gamma}}^{\top}\boldsymbol{x})(2\pi)^{(-\ell/2)} \det\left[\mathcal{K}''_{\boldsymbol{X}}(\hat{\boldsymbol{\gamma}})\right]^{-1/2} + O(n^{-1/2}).$$

This derivation generalizes the univariate argument leading to (66). As in the univariate case, when the saddlepoint equation has a solution the only conditions to be checked are the those necessary to demonstrate the order of error of the underlying Edgeworth series. We will see that this requires only the existence of the proper cumulants, which is verified by noting that moments of all orders exist, and the integrability of some power of the characteristic function, which is verified by noting that the density $f_{\boldsymbol{X}}$ is assumed to exist. Also as in the univariate case no uniformity claim is made here, although the same heuristics lead one to expect good uniform error properties.

The solution to the saddlepoint equation is once again the maximum likelihood estimator, this time for the multivariate parameter $\boldsymbol{\gamma}$ in the exponential family defined by $f_{\boldsymbol{X}}(\boldsymbol{x}, \boldsymbol{\gamma})$. Existence of such a solution will be considered in §7.7.

Predictably, the higher-order saddlepoint approximations can be derived from the Edgeworth approximation. Suppose that $\boldsymbol{X}$ arises as the mean of n independent and identically distributed copies of $\boldsymbol{Y}$. Let $\hat{\boldsymbol{\beta}}$ be the solution to the saddlepoint equation defined in terms of the distribution of $\boldsymbol{Y}$,

$$\mathcal{K}'_{\boldsymbol{Y}}(\hat{\boldsymbol{\beta}}) = \boldsymbol{x}. \tag{132}$$

In the multivariate setting, as in the univariate setting, the factor multiplying $1/\sqrt{n}$ in the multiplicative correction factor to the baseline approximation is zero, since the associated Hermite polynomial is evaluated at $\boldsymbol{o}$. One might expect that $f_{\boldsymbol{X}}(\boldsymbol{x})$ might have an expansion of the form

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{\exp(n[\mathcal{K}_{\boldsymbol{Y}}(\hat{\boldsymbol{\beta}}) - \hat{\boldsymbol{\beta}}^{\top}\boldsymbol{x}])}{\sqrt{\det\left[\mathcal{K}''_{\boldsymbol{Y}}(\hat{\boldsymbol{\beta}})\right]}}\left(\frac{n}{2\pi}\right)^{\ell/2}\left[1 + \frac{b(\hat{\boldsymbol{\beta}})}{2n} + O\left(n^{-2}\right)\right]. \tag{133}$$

To obtain $b(\hat{\boldsymbol{\beta}})$, approximate the tilted distribution $f_{\boldsymbol{X}}(\boldsymbol{x}, \hat{\boldsymbol{\beta}})$ by

$$\sum_{j=0}^{\infty}\sum_{\boldsymbol{s}\in S(j)}\frac{1}{j!}\mu^{*s_1\cdots s_j}(-1)^j\frac{d^j}{dx^{s_1}\cdots dx^{s_j}}\phi(0, \mathcal{K}''(\hat{\boldsymbol{\beta}})), \tag{134}$$

where the cumulants involved in (134) are derivatives of $\mathcal{K}_X$ at $\hat{\boldsymbol{\beta}}$. One may easily check that the terms of order $O(1/\sqrt{n})$ are zero. Then identify the terms of order $O(n^{-1})$ and evaluate the generalized Hermite polynomials $h_{s_1,\ldots,s_j}(\boldsymbol{x}; \mathcal{K}''(\hat{\boldsymbol{\beta}}))$ of (129) at zero. The pseudo-moments $\mu^{*s_1\cdots s_j}$ were defined in (126). Here and below invoke the convention that an index appearing both as a subscript and as a superscript is summed over. After replacing products of components of $\boldsymbol{\beta}$ by the associated Hermite polynomials, terms of order $O(1/\sqrt{n})$ in (134) are

$$\frac{1}{3!}\mathcal{K}_X^{ijk}(\hat{\boldsymbol{\beta}})h_{ijk}(\boldsymbol{x}; \mathcal{K}''(\hat{\boldsymbol{\beta}}));$$

the terms of order $O(n^{-1})$ in (134) are

$$\frac{1}{4!}\mathcal{K}_X^{ijkl}(\hat{\boldsymbol{\beta}})h_{ijkl}(\boldsymbol{x}; \mathcal{K}''(\hat{\boldsymbol{\beta}})) + \frac{1}{2 \times 3!3!}\mathcal{K}_X^{ijk}(\hat{\boldsymbol{\beta}})\mathcal{K}_X^{lmn}(\hat{\boldsymbol{\beta}})h_{ijklmn}(\boldsymbol{x}; \mathcal{K}''(\hat{\boldsymbol{\beta}})).$$

Here dependence of $\mathcal{K}_X$ on n is suppressed, and the superscripts refer to derivatives with respect to components of $\boldsymbol{\beta}$. For some purposes it is preferable to have the pseudo-moments be symmetric in their indices; the coefficients of products of components of $\boldsymbol{\beta}$ in (126) do not have this property. In particular, the second term of size $O(n^{-1})$ is asymmetric. These terms may be made symmetric by summing over all permutations giving recognizably different index patterns and dividing by the number of such patterns. Were this necessary in the present context we would be required to add up over all $(6!/(3!3!))/2 = 10$ ways to partition $ijklmn$ into two groups of 3 each, and then divide by 10. This is not necessary for what follows. McCullagh (1987) gives formulas for the Hermite polynomials associated with an arbitrary covariance matrix Σ as equation (5.7); for the present purposes it suffices to note that

$$h_{ijk}(\mathbf{o}; \Sigma) = 0$$
$$h_{ijkl}(\mathbf{o}; \Sigma) = \kappa_{ij}\kappa_{kl}[3] = \kappa_{ij}\kappa_{kl} + \kappa_{ik}\kappa_{jl} + \kappa_{il}\kappa_{jk}$$
$$h_{ijklmn}(\mathbf{o}; \Sigma) = -\kappa_{ij}\kappa_{kl}\kappa_{mn}[15].$$

These polynomials will be evaluated with Σ equal to the covariance matrix $[\hat{\kappa}_{ij}]$ associated with the saddlepoint. Here we use the notation of McCullagh (1987) in denoting by quantity with subscripts followed by an integer in brackets the sum of all similar but distinct terms generated by permuting the indices. After replacing quantities involving $\mathcal{K}_X$ by counterparts involving $\mathcal{K}_Y$,

$$\frac{1}{2}b(\hat{\boldsymbol{\beta}}) = \frac{1}{4!}\mathcal{K}_Y^{ijkl}(\hat{\boldsymbol{\beta}})h_{ijkl}(\mathbf{o}; \mathcal{K}''(\hat{\boldsymbol{\beta}})) + \frac{1}{2 \times 3!3!}\mathcal{K}_Y^{ijk}(\hat{\boldsymbol{\beta}})\mathcal{K}_Y^{lmo}(\hat{\boldsymbol{\beta}})h_{ijklmo}(\mathbf{o}; \mathcal{K}''(\hat{\boldsymbol{\beta}}))$$
$$= \frac{1}{4!}\hat{\kappa}^{ijkl}(\hat{\kappa}_{ij}\hat{\kappa}_{kl}[3]) - \frac{10}{6!}\hat{\kappa}^{ijk}\hat{\kappa}^{lmo}(\hat{\kappa}_{ij}\hat{\kappa}_{kl}\hat{\kappa}_{mo}[15]).$$

Sorting the fifteen summands represented by $\hat{\kappa}_{ij}\hat{\kappa}_{kl}\hat{\kappa}_{mo}[15]$ into nine for which the indices for two factors are subsets of the indices of $\hat{\kappa}^{ijk}$ or $\hat{\kappa}^{lmo}$, and six for which

indices of all three factors are split between the indices of $\hat{\kappa}^{ijk}$ and $\hat{\kappa}^{lmo}$,

$$
\begin{aligned}
b(\hat{\boldsymbol{\beta}}) &= \frac{2}{4!}\hat{\kappa}^{ijkl}(\hat{\kappa}_{ij}\hat{\kappa}_{kl}[3]) - \frac{20}{6!}\left[(\hat{\kappa}^{ijk}\hat{\kappa}^{lmo}\hat{\kappa}_{ij}\hat{\kappa}_{kl}\hat{\kappa}_{mo}[9]) + (\hat{\kappa}^{ijk}\hat{\kappa}^{lmo}\hat{\kappa}_{il}\hat{\kappa}_{jm}\hat{\kappa}_{ko}[6])\right] \\
&= \frac{6}{4!}\hat{\kappa}^{ijkl}\hat{\kappa}_{ij}\hat{\kappa}_{kl} - \left[\frac{90}{6!}\hat{\kappa}^{ijk}\hat{\kappa}^{lmo}\hat{\kappa}_{ij}\hat{\kappa}_{kl}\hat{\kappa}_{mo} + \frac{60}{6!}\hat{\kappa}^{ijk}\hat{\kappa}^{lmo}\hat{\kappa}_{il}\hat{\kappa}_{jm}\hat{\kappa}_{ko}\right] \\
&= \tfrac{1}{4}\hat{\rho}_{\boldsymbol{Y},4} - \tfrac{1}{4}\hat{\rho}_{\boldsymbol{Y},13} - \tfrac{1}{6}\hat{\rho}_{\boldsymbol{Y},23},
\end{aligned} \tag{135}
$$

for

$$
\hat{\rho}_{\boldsymbol{Y},4} = \hat{\kappa}^{ijkl}\hat{\kappa}_{ij}\hat{\kappa}_{kl}, \ \ \hat{\rho}_{\boldsymbol{Y},13} = \hat{\kappa}^{ijk}\hat{\kappa}^{lmo}\hat{\kappa}_{ij}\hat{\kappa}_{kl}\hat{\kappa}_{mo}, \ \ \hat{\rho}_{\boldsymbol{Y},23} = \hat{\kappa}^{ijk}\hat{\kappa}^{lmo}\hat{\kappa}_{il}\hat{\kappa}_{jm}\hat{\kappa}_{ko}. \tag{136}
$$

For the sake of brevity all cumulants $\hat{\kappa}$ denote those of the distribution tilted to $\hat{\boldsymbol{\beta}}$ and are given by derivatives of $\mathcal{K}_{\boldsymbol{Y}}(\boldsymbol{\beta}) - \boldsymbol{\beta}^\top \boldsymbol{x}$ at $\hat{\boldsymbol{\beta}}$. For example, $\hat{\kappa}^1 = \partial \mathcal{K}_{\boldsymbol{Y}}/\partial\beta_1 - x^1$, and $\hat{\kappa}^{12} = \partial^2 \mathcal{K}_{\boldsymbol{Y}}/\partial\beta_1\partial\beta_2$. Let $\hat{\kappa}_{ij}$ refer to the i,j entry of the inverse of the matrix formed by $\hat{\kappa}^{ij}$. The quantities $\hat{\rho}_{\boldsymbol{Y},4}$, $\hat{\rho}_{\boldsymbol{Y},13}$, and $\hat{\rho}_{\boldsymbol{Y},23}$ are multivariate skewness and kurtosis measures proposed by Mardia (1970) and McCullagh (1986). Hence, as in the univariate case the terms of order $O(n^{-1/2})$ and $O(n^{-3/2})$ are zero, and (133) holds.

Jing and Robinson (1994) present a saddlepoint approximation to the density of a smooth transformation of a random vector $\boldsymbol{X}$ approximated by (133); this will be recounted in §6.7.

The saddlepoint approximation may be expressed in terms of the Legendre transform $\mathcal{K}_{\boldsymbol{Y}}^*$ of $\mathcal{K}_{\boldsymbol{Y}}$: For $\boldsymbol{x}$ in the range of $\mathcal{K}_{\boldsymbol{Y}}'$ define

$$
\mathcal{K}_{\boldsymbol{Y}}^*(\boldsymbol{x}) = \hat{\boldsymbol{\beta}}^\top \boldsymbol{x} - \mathcal{K}_{\boldsymbol{Y}}(\hat{\boldsymbol{\beta}}) \tag{137}
$$

for $\hat{\boldsymbol{\beta}}$ satisfying (132). Then $\mathcal{K}_{\boldsymbol{Y}}^{*'}(\boldsymbol{x}) = [\boldsymbol{x}^\top - \mathcal{K}_{\boldsymbol{Y}}'(\hat{\boldsymbol{\beta}})]\hat{\boldsymbol{\beta}}'(\boldsymbol{x}) + \hat{\boldsymbol{\beta}}(\boldsymbol{x}) = \hat{\boldsymbol{\beta}}(\boldsymbol{x})$, and $\mathcal{K}_{\boldsymbol{Y}}^{*''}(\boldsymbol{x}) = \hat{\boldsymbol{\beta}}'(\boldsymbol{x})$. Differentiating (132), $\mathcal{K}_{\boldsymbol{Y}}''(\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}}'(\boldsymbol{x}) = \boldsymbol{I}$, and hence

$$
\mathcal{K}_{\boldsymbol{Y}}^{*''}(\boldsymbol{x}) = \mathcal{K}_{\boldsymbol{Y}}''(\hat{\boldsymbol{\beta}}(\boldsymbol{x}))^{-1}. \tag{138}
$$

Hence

$$
f_{\boldsymbol{X}}(\boldsymbol{x}) = \exp(-n\mathcal{K}_{\boldsymbol{Y}}^*(\boldsymbol{x}))\sqrt{\det\left[\mathcal{K}_{\boldsymbol{Y}}^{*''}(\hat{\boldsymbol{\beta}})\right]}\left(\frac{n}{2\pi}\right)^{\ell/2}\left[1 + \frac{b(\hat{\boldsymbol{\beta}})}{2n} + O\left(n^{-2}\right)\right]. \tag{139}
$$

6.5. Multivariate Integral Expansion Theorems

Multivariate saddlepoint approximations may also be derived via a multivariate version of Watson's Lemma. The aim is to calculate integrals whose integrand is the exponential of a quadratic term times sample size, times a function defined on a set $\mathcal{Q}$ containing the minimum of the quadratic form. Two additional results are presented, one recasting the first result in terms of integrations in complex space along hyperplanes of constant real part, and the second presenting results with integrands with functions that are not exactly quadratic in the integrand.

Lemma 6.5.1:   *Suppose that θ has derivatives of order* $2j$, *bounded on* $[-\epsilon, \epsilon]^{\mathfrak{k}}$. *Let*
$\mathfrak{F}(s) = \{ \boldsymbol{v} \in \mathfrak{Z}^{\mathfrak{k}} | v_1, \cdots, v_{\mathfrak{k}} \geq 0, \sum v_j = s \}$, *and*

$$A_s = \sum_{\mathfrak{F}(s)} \frac{(-2)^{-\sum_{j=1}^{\mathfrak{k}} v_j}}{v_1! \cdots v_{\mathfrak{k}}!} \left[\frac{\partial^{2v_1 + \cdots + 2v_{\mathfrak{k}}}}{\partial^{2v_1} \omega_1 \cdots \partial^{2v_{\mathfrak{k}}} \omega_{\mathfrak{k}}} \theta \right] (\mathbf{o});$$

then

$$\left(\frac{n}{2\pi} \right)^{\mathfrak{k}/2} \int_{-\epsilon}^{\epsilon} \cdots \int_{-\epsilon}^{\epsilon} \exp(-\frac{n}{2} \sum_j \omega_j^2) \theta(\boldsymbol{\omega}) \ d\boldsymbol{\omega} = \sum_{s=0}^{j-1} A_s n^{-s} + O(n^{-j}).$$

Proof: Problem 1a of §4.10 shows that $\sum_{s=0}^{j-1} A_s n^{-s}$ is the value of the integral with $\theta(\boldsymbol{\omega})$ replaced by its Taylor approximation using derivatives of order $2(j-1)$ and lower. Hence it suffices to prove the lemma assuming all derivatives of order $2(j-1)$ and lower are zero.

Applying the univariate Taylor series approximation to $\theta(\lambda \boldsymbol{\omega})$ find that $\theta(\boldsymbol{\omega}) = \omega_j \omega_k \cdots \omega_m \theta^{jk...m}(\boldsymbol{\omega}^*)/(2j)!$, where the list $j, k, \ldots, m$ has $2j$ entries and $\boldsymbol{\omega}^*$ is a linear combination of $\mathbf{o}$ and $\boldsymbol{\omega}$. So

$$|\theta(\boldsymbol{\omega})| \leq (\sum_j \omega_j^2)^j \sqrt{\sup_{(j,k,...,m)\mathfrak{F}(2j), \boldsymbol{\omega}^* \in [-\epsilon, \epsilon]^{\mathfrak{k}}} |\theta^{jk...m}(\boldsymbol{\omega}^*)| \mathfrak{k}^j / (2j)!}.$$

The result follows directly.

Q.E.D

Theorem 6.5.2:   *Suppose that $\theta_j(\boldsymbol{\omega})$ are analytic functions from $\mathcal{Q} \subset \mathfrak{C}^{\mathfrak{k}}$ to $\mathfrak{C}$, and let*

$$f_n(\boldsymbol{\omega}) = \sum_{j=0}^{j} \theta_j(\boldsymbol{\omega})/n^j, \tag{140}$$

Take $\hat{\boldsymbol{\omega}} \in \mathcal{Q}$, such that $\hat{\boldsymbol{\omega}} + i[-\epsilon, \epsilon]^{\mathfrak{k}} \subset \mathcal{Q}$. Then

$$\left(\frac{n}{2\pi} \right)^{\mathfrak{k}/2} i^{-\mathfrak{k}} \int_{\hat{\omega}_1 - i\epsilon}^{\hat{\omega}_1 + i\epsilon} \cdots \int_{\hat{\omega}_{\mathfrak{k}} - i\epsilon}^{\hat{\omega}_{\mathfrak{k}} + i\epsilon} \exp(\frac{n}{2} \sum_j (\omega_j - \hat{\omega}_j)^2) f_n(\boldsymbol{\omega}) \ d\boldsymbol{\omega} = \sum_{s=0}^{j-1} A_s n^{-s} + O(n^{-j}),$$

for $\mathfrak{F}(j) = \{ \boldsymbol{v} \in \mathfrak{Z}^{\mathfrak{k}} | v_1, \cdots, v_{\mathfrak{k}} \geq 0, \sum v_j \leq s \}$, and

$$A_s = \sum_{\mathfrak{F}(j)} \frac{(-2)^{-\sum_{j=1}^{\mathfrak{k}} v_j}}{v_1! \cdots v_{\mathfrak{k}}!} \left[\frac{\partial^{2v_1 + \cdots + 2v_{\mathfrak{k}}}}{\partial^{2v_1} \omega_1 \cdots \partial^{2v_{\mathfrak{k}}} \omega_{\mathfrak{k}}} \theta_{s - \sum v_j} \right] (\hat{\boldsymbol{\omega}}). \tag{141}$$

Proof: This is just Lemma 6.5.1 applied term-wise, after changing the variable of integration.

Q.E.D

The next result shows that similar techniques may be used to approximate more general real integrals.

Lemma 6.5.3: *Suppose $\mathcal{B}(\boldsymbol{x})$ is a convex function minimized at $\hat{\boldsymbol{x}} \in ([-\epsilon, \epsilon]^{\mathfrak{k}})^{\circ}$, and with all order 6 derivatives existing. Suppose that $\theta(\boldsymbol{x})$ is a function such that $\left|\theta^{ijkl}(\boldsymbol{x})\right| \leq C_1 \exp(C_2 |\boldsymbol{x}|^2)$ for constants C_1 and C_2, for all fourth degree derivatives $\theta^{ijkl}(\boldsymbol{x})$. Then*

$$\left(\frac{n}{2\pi}\right)^{\mathfrak{k}/2} \int_{-\epsilon}^{\epsilon} \cdots \int_{-\epsilon}^{\epsilon} \exp(-\frac{n}{2}\mathcal{B}(\boldsymbol{x}))\theta(\boldsymbol{x}) \, d\boldsymbol{x} = \frac{\exp(-\frac{n}{2}\mathcal{B}(\hat{\boldsymbol{x}}))}{\det\left[\mathcal{B}''(\hat{\boldsymbol{x}})\right]^{1/2}}\{\theta(\hat{\boldsymbol{x}})(1+$$

$$[\tfrac{1}{8}\rho_4 - \tfrac{1}{8}\rho_{31} - \tfrac{1}{12}\rho_{32}]/n) - \tfrac{1}{2}(\hat{\theta}^{ij}\hat{\mathcal{B}}_{ij} - \hat{\mathcal{B}}^{ijk}\hat{\theta}^l\hat{\mathcal{B}}_{ij}\hat{\mathcal{B}}_{kl})/n + O(1/n^2)\}, (142)$$

for

$$\rho_4 = -\hat{\mathcal{B}}^{ijkl}\hat{\mathcal{B}}_{ij}\hat{\mathcal{B}}_{kl}, \ \rho_{31} = -\hat{\mathcal{B}}^{ijk}\hat{\mathcal{B}}^{lmp}\hat{\mathcal{B}}_{ij}\hat{\mathcal{B}}_{kl}\hat{\mathcal{B}}_{mp} \ and \ \rho_{32} = -\hat{\mathcal{B}}^{ijk}\hat{\mathcal{B}}^{lmp}\hat{\mathcal{B}}_{il}\hat{\mathcal{B}}_{jm}\hat{\mathcal{B}}_{kp},$$

and $\hat{\mathcal{B}}$ with superscripts the derivative of $\mathcal{B}$ with respect to the indicated components of $\boldsymbol{x}$, evaluated at $\hat{\boldsymbol{x}}$, $\hat{\mathcal{B}} = \mathcal{B}(\hat{\boldsymbol{x}})$, and $\hat{\mathcal{B}}_{kl}$ the element in row k and column l of the inverse of $\mathcal{B}''(\hat{\boldsymbol{x}})$; $\hat{\theta}$ with and without superscripts is defined similarly. The error is uniform for $\hat{\boldsymbol{x}}$ bounded away from the boundary of $[-\epsilon, \epsilon]^{\mathfrak{k}}$, if fourth derivatives of $\mathcal{B}$ are bounded.

Proof: First assume that $\mathcal{B}''(\hat{\boldsymbol{x}})$ is the identity matrix, and that $\hat{\boldsymbol{x}} = \boldsymbol{o}$. Let

$$P(\boldsymbol{x}) = 1 - \frac{\hat{\mathcal{B}}^{jkl}x_jx_kx_l}{6\sqrt{n}} - \frac{\hat{\mathcal{B}}^{jklm}x_jx_kx_lx_m}{24n} - \frac{\hat{\mathcal{B}}^{jklmp}x_jx_kx_lx_mx_p}{120n\sqrt{n}} - \frac{\hat{\mathcal{B}}^{jklmpq}x_jx_kx_lx_mx_px_q}{720n^2}.$$

Then by Corollary 2.5.4,

$$\left|\exp(-\frac{n}{2}\mathcal{B}(\boldsymbol{x})) - \exp(-\frac{n}{2}x_j\delta^{jk}x_i)\sum_{s=0}^{4}\frac{P(\boldsymbol{x})^s}{s!}\right| \leq \exp(-\frac{n}{4}x_j\delta^{jk}x_i)C_3(\delta^j|x_j|)^4, \ (143)$$

and

$$\left|\theta(\boldsymbol{x}) - \hat{\theta} - \hat{\theta}^ix_i - \tfrac{1}{2}\hat{\theta}^{ij}x_ix_j - \tfrac{1}{6}\hat{\theta}^{ijk}x_ix_jx_k\right| \leq |x_i|\,|x_j|\,|x_k|\,|x_l|\,C_1\exp(C_2\|\boldsymbol{x}\|^2).$$
$$(144)$$

Without loss of generality, assume that C_1 and C_2 are large enough that $|\theta(\boldsymbol{x})| \leq C_1\exp(C_2|\boldsymbol{x}|^2)$ as well. Multiplying all but the first quantities inside the absolute value signs on the left hand sides of (143) and (144), eliminating terms with odd numbers of factors of components of $\boldsymbol{x}$, noting every term involving $2k$ factors of components of $\boldsymbol{x}$ is of size n^{-k}, and deleting terms of size $O(n^{-2})$ and smaller, implies that the left side of (142) is

$$\left(\frac{n}{2\pi}\right)^{\mathfrak{k}/2}\int_{-\epsilon}^{\epsilon}\cdots\int_{-\epsilon}^{\epsilon}\exp(-\frac{n}{2}x_i\delta^{ij}x_j)\left[\hat{\theta}\left\{1 - \frac{n}{24}\hat{\mathcal{B}}^{ijkl}x_ix_jx_kx_l + \frac{n^2}{72}\hat{\mathcal{B}}^{ijk}\hat{\mathcal{B}}^{lmp}x_ix_jx_kx_lx_mx_p\right\}\right.$$

$$\left. - \frac{n}{6}\hat{\theta}^j\hat{\mathcal{B}}^{klm}x_jx_kx_lx_m - \tfrac{1}{2}\hat{\theta}^{jk}x_jx_k\right]\, d\boldsymbol{x} + O(n^{-2}) \tag{145}$$

Integrating term-wise yields the result. The calculations for the non–error part of the integral may be performed either by analogy with the integrals generating the multivariate Edgeworth series, using the inverse of (127) to express multidimensional integrals of an exponentiated quadratic form in $\boldsymbol{\beta}$ times powers of components of $\boldsymbol{\beta}$ in terms of derivatives of the multivariate normal density, or more directly. The next proof of the next lemma demonstrates this direct calculation for one of the terms above.

Proof when $\hat{\mathcal{B}}^{ij}$ is not the identity matrix follows by applying a Cholesky decomposition to $\hat{\mathcal{B}}^{ij}$, and changing variables to reduce the problem to the simpler case.

<div align="right">Q.E.D</div>

The proof of the previous lemma involved calculating multivariate integrals using the information about the Laplace transform of derivatives of the normal cumulative distribution function. The next lemma demonstrates that these quantities may also be calculated more directly.

Lemma 6.5.4:   *When $\hat{\mathcal{B}}^{ijk}$ is a symmetric array, in that the elements are unchanged if the indices are permuted, then*

$$\left(\frac{n}{2\pi}\right)^{1/2}\int_{-\epsilon}^{\epsilon}\cdots\int_{-\epsilon}^{\epsilon}\exp(-\tfrac{1}{2}x_i\delta^{ij}x_j)\frac{1}{72}\hat{\mathcal{B}}^{ijk}\hat{\mathcal{B}}^{lmp}x_ix_jx_kx_lx_mx_p\ d\boldsymbol{x} = \tfrac{1}{8}\rho_{31} + \tfrac{1}{12}\rho_{32}.$$

Proof: Any term involving a component of $\boldsymbol{x}$ to an odd power is zero, and so three cases must be considered. When all indices agree, the integral is $15\hat{\mathcal{B}}^{iii}\hat{\mathcal{B}}^{iii}/72$. This exactly corresponds to the sum of addends in $-\tfrac{1}{8}\rho_{31} - \tfrac{1}{12}\rho_{32}$ in which all superscripts agree. When there are two distinct indices, one must be appear twice and the other four times. The resulting integral is either $3\hat{\mathcal{B}}^{iii}\hat{\mathcal{B}}^{ijj}/72$, or $3\hat{\mathcal{B}}^{iij}\hat{\mathcal{B}}^{iij}/72$. The first pattern occurs six times in $\hat{\mathcal{B}}^{ijk}\hat{\mathcal{B}}^{lmp}x_ix_jx_kx_lx_mx_p$ (as $\hat{\mathcal{B}}^{iii}\hat{\mathcal{B}}^{ijj}$, $\hat{\mathcal{B}}^{iii}\hat{\mathcal{B}}^{jij}$, $\hat{\mathcal{B}}^{iii}\hat{\mathcal{B}}^{jji}$, $\hat{\mathcal{B}}^{jji}\hat{\mathcal{B}}^{iii}$, $\hat{\mathcal{B}}^{ijj}\hat{\mathcal{B}}^{iii}$, $\hat{\mathcal{B}}^{jij}\hat{\mathcal{B}}^{iii}$), and so the integral contributes $\tfrac{1}{4}\hat{\mathcal{B}}^{iii}\hat{\mathcal{B}}^{ijj}$. This pattern occurs twice in ρ_{31}, and not at all in ρ_{32}, and hence contributions on both sides of the equality match. The second pattern occurs nine times in $\hat{\mathcal{B}}^{ijk}\hat{\mathcal{B}}^{lmp}x_ix_jx_kx_lx_mx_p$ (as can be seen by rearranging the two sets of indices separately), and so has a multiplier of $\tfrac{3}{8}$, and occurs three times in ρ_{32}, and once in ρ_{31}, making the total contributions match on both sides of the equal sign.

When there are three distinct indices, each must appear three times, and resulting integral is either $\hat{\mathcal{B}}^{iik}\hat{\mathcal{B}}^{kjj}/72$, or $\hat{\mathcal{B}}^{ikj}\hat{\mathcal{B}}^{ikj}/72$. Each term of the first form appears 18 times (since in each factor the location for the non-duplicated index may be chosen 3 different ways, and the factors may be swapped). Each term of second form appears 36 times (since each set of indices may be rearranged 6 ways separately), Each of the first type of term appears twice in ρ_{31}, and each of the second type of term appears six times in ρ_{32}. Hence the coefficients agree.

<div align="right">Q.E.D</div>

Alternatively, one might prove Lemma 6.5.3 by defining new variables

$$y_j = \sqrt{2[\min(\mathcal{B}(\boldsymbol{z})|z_k = x_k \forall k \leq j) - \min(\mathcal{B}(\boldsymbol{z})|z_k = x_k \forall k < j)]},$$

for $j = 1, \mathfrak{k}$, and change variables from $\boldsymbol{z}$ to $\boldsymbol{y}$.

The relative error of the $O(1/n)$ approximation in the results of Lemma 6.5.3 is determined by ρ_4, ρ_{31}, and ρ_{32}; when these terms are small, the remaining terms tend to be small as well. In this way, Lemma 6.5.3 motivates a more heuristic approximation, which might be used whether or not $\mathcal{B}(\boldsymbol{x})$ is a multiple of sample size:

$$\left(\frac{1}{2\pi}\right)^{\mathfrak{k}/2} \int_{-\epsilon}^{\epsilon} \cdots \int_{-\epsilon}^{\epsilon} \exp(-\tfrac{1}{2}\mathcal{B}(\boldsymbol{x}))\theta(\boldsymbol{x}) \, d\boldsymbol{x} \approx \frac{\exp(-\tfrac{1}{2}\mathcal{B}(\hat{\boldsymbol{x}}))}{\det\left[\mathcal{B}''(\hat{\boldsymbol{x}})\right]^{1/2}}\theta(\hat{\boldsymbol{x}}). \tag{146}$$

This approximation (146) is known as Laplace's method, and is justified whenever the derivative ratios can be expected to be small. This quantity is of size $O(1/n)$ when $\mathcal{B}(\boldsymbol{x})$ is a multiple of sample size; (146) has also proved useful in other contexts as well.

The identification of $\mathcal{B}(\boldsymbol{x})$ and $\theta(\boldsymbol{x})$ on the left side of (146) is not entirely rigid; for example, if $\theta(\boldsymbol{x})$ is positive for all $\boldsymbol{x}$, then the integrand may also be written as $\exp(-\tfrac{1}{2}\mathcal{B}(\boldsymbol{x}) + \log(\theta(\boldsymbol{x})))$, and $\hat{\boldsymbol{x}}$ would then be chosen to minimize $\tilde{\mathcal{B}}(\boldsymbol{x}) - 2\log(\theta(\boldsymbol{x}))$. The determinant in the denominator of the right hand side of (146) will be changed. Similar results are generally obtained with a wide variety of identifications of $\mathcal{B}$ (and hence θ), as long as quantities like ρ_4, ρ_{31}, and ρ_{32} remain small. Kass (1988) discusses the relationship between Laplace's method and saddlepoint methods.

6.6. Error Analysis for the Edgeworth Series

This section contains a proof of the multivariate Edgeworth approximation for densities, and discusses the proof for distribution functions.

Theorem 6.6.1: *Suppose F is a cumulative distribution function, with mean $\boldsymbol{\mu}$ and positive definite variance matrix $\boldsymbol{\Sigma}$, and $\mathfrak{j}$ is an integer greater than or equal to two, such that F has all cumulants of order $\mathfrak{j}$. Suppose further that*

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} |\zeta(\boldsymbol{\beta})|^{\mathfrak{r}} \, d\boldsymbol{\beta} < \infty \text{ for some } \mathfrak{r} \geq 1. \tag{147}$$

Let $\boldsymbol{Y}_j$ be independent and identically distributed random vectors with distribution function F. Let $\boldsymbol{X} = (\sum_{j=1}^{n} \boldsymbol{Y}_j - n\boldsymbol{\mu})/\sqrt{n}$. Then $\boldsymbol{X}$ has a bounded density $f_{\boldsymbol{X}}$ for sufficiently large n, and when (128) is calculated using only the first $\mathfrak{j}$ cumulants, terms of order $o(n^{1-\mathfrak{j}/2})$ dropped, and the result $e_{\mathfrak{j}}(\boldsymbol{x}, \boldsymbol{\kappa}^n)$ is used to approximate the density $f_{\boldsymbol{X}}$, the absolute error is uniformly of order $o(n^{1-\mathfrak{j}/2})$.

Proof: The following proof is almost identical to the univariate proof. Without loss of generality assume $\Sigma = \boldsymbol{I}$. Since the cumulative distribution functions can

be recovered from their characteristic functions using (19), the difference between $f_X(x)$ and $e_j(x, \kappa_X)$ can be bounded by the integral

$$\left(\frac{1}{2\pi}\right)^{\mathfrak{k}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} 2 \left| \zeta_X(\beta) - \xi_X(\beta) \right| \ d\beta, \qquad (148)$$

the analogue of (52). Here ξ_X is the Fourier transform of the approximate density $e_j(x, \kappa_X)$. Because the integral (148) converges absolutely, this argument will not need the Smoothing Theorem. The Series Theorem is still a key idea here.

The range of integration will now be split in two parts. The area very close to the origin, and the area farther out will be handled separately.

a. Very near the origin, Corollary 2.5.4 implies that the integrand is bounded by

$$\exp(-\left\|\beta\right\|^2/4) \left[\frac{\epsilon \left\|\beta\right\|^j}{n^{j/2-1}} + \frac{C^{j-1} \left\|\beta\right\|^{3(j-1)}}{(j-1)! n^{j/2-1/2}} \right],$$

for β such that $\left\|\beta\right\| < \delta\sqrt{n}$. When integrated over $(-\delta\sqrt{n}, \delta\sqrt{n})^{\mathfrak{k}}$ the result is still of order $o(n^{1-j/2})$. Integration is performed using a trivial application of Lemma 6.5.1.

b. For more extreme values of β ($\left\|\beta\right\| > \delta\sqrt{n}$), the integrability condition on the characteristic function is used to bound the integrand. There exists $q < 1$ such that if $\left\|\beta\right\| > \delta\sqrt{n}$ then $\left|\zeta(\beta/\sqrt{n})\right| < q$. Hence the contribution of (148) from this part of the range of integration can be bounded by

$$q^{n-p} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left|\zeta(\beta)\right|^p \ d\beta + \int \cdots \int_{\left\|\beta\right\| > \delta\sqrt{n}} \left|\xi_X(\beta)\right| \ d\beta,$$

and approaches 0 geometrically.

<div align="right">Q.E.D</div>

The following theorem is the analogue of the univariate Edgeworth cumulative distribution function theorem. As the flavor of its proof deviates from that of the rest of this volume, and as it will not be needed to prove later results, it is cited without complete proof.

Theorem 6.6.2: *Suppose F is a cumulative distribution function, with mean μ and positive definite variance matrix Σ, and j is an integer greater than or equal to two, such that F has all cumulants of order j. Suppose further that*

$$\begin{cases} \left|\zeta(\beta)\right| < 1 \ \forall \beta \neq 0 & \textit{if } j = 3, \textit{ or} \\ \limsup_{\left\|\beta\right\| \to \infty} \left|\zeta(\beta)\right| < 1 & \textit{if } j > 3. \end{cases} \qquad (149)$$

The condition when $j > 3$ is a multivariate version of Cramér's condition. Let Y_j be independent and identically distributed random vectors with distribution function F. Let $X = (\sum_{j=1}^{n} Y_j - n\mu)/\sqrt{n}$, and let F_X be the cumulative distribution function for X. When (130) is calculated using only the first j cumulants, terms of order $o(n^{1-j/2})$ dropped, and the result $E_j(x, \kappa^n)$ is used to approximate the distribution

given by $F_{\boldsymbol{X}}$, the absolute error is uniformly of order $o(n^{1-j/2})$, in the sense that not only

$$\sup_{\boldsymbol{x}} |E_j(\boldsymbol{x}, \boldsymbol{\kappa}^n) - F_{\boldsymbol{X}}(\boldsymbol{x})| = o(n^{1-j/2}), \tag{150}$$

but furthermore, that $\sup_C |\int_C dE_j(\boldsymbol{x}, \boldsymbol{\kappa}^n) - \int_C dF_{\boldsymbol{X}}(\boldsymbol{x})| = o(n^{1-j/2})$, where the supremum is taken over all convex Borel sets C.

The precise definition of a Borel set is unimportant for the present purposes; it suffices to note that this class of sets contains all open sets and all closed sets.

Note that (150) provides a result applicable when approximating probabilities of multivariate rectangles and their unions. Proof of this result is a simple extension of univariate results. A multivariate version of the Smoothing Theorem could be constructed to prove that the difference between $F_{\boldsymbol{X}}(\boldsymbol{x})$ and $E_j(\boldsymbol{x}, \boldsymbol{\kappa}^n)$ can be bounded by the principal value of the integral

$$\lim_{\Theta \to \infty} \frac{1}{2\pi} \int_{-\Theta}^{\Theta} \cdots \int_{-\Theta}^{\Theta} \frac{2}{\|\boldsymbol{\beta}\|} |\zeta_{\boldsymbol{X}}(\boldsymbol{\beta}) - \xi_{\boldsymbol{X}}(\boldsymbol{\beta})| \, d\boldsymbol{\beta}, \tag{151}$$

where ξ is the Fourier transform of $E_j(\boldsymbol{x}, \boldsymbol{\kappa}^n)$. Remember that ξ is not really a characteristic function since $E_j(\boldsymbol{x}, \boldsymbol{\kappa}^n)$ is not really a cumulative distribution function. Nevertheless, just as in the univariate case the cumulative distribution function inversion theorem arguments hold for ξ as well, justifying (151). As in the univariate case, the range of integration is divided into three parts: within a multivariate cube with edges of length δ, outside a multivariate cube with edges of length Θ, and in the region between these. The outer region again is eliminated through the use of the Smoothing Theorem. The inner region can be bounded just as in the multivariate density proof, and the intermediate region bounded using conditions (149). Proving the final result is more difficult. See Bhattacharya and Rao (1976).

Q.E.D

Result (150) is, however, too weak to be useful in the sequel, since of interest will be primarily approximately elliptically-shaped regions. Bhattacharya and Rao (1976) also present a direct analogy to Esseen's series, given in Theorem 3.13.2, for lattice variables. These theorems in general are more complicated, since of interest are probabilities of sets in $\boldsymbol{Y}$-space with complicated shapes. As in the univariate case, this series evaluated at continuity corrected points is equivalent to the multivariate Edgeworth series, to the proper order in n (Kolassa, 1989).

6.7. A Direct Derivation of the Multivariate Saddlepoint Approximation

The multivariate saddlepoint density might be derived directly from (123), using the results of the previous section. Consider specifically the case in which $f_{\boldsymbol{X}}$ represents the distribution of the mean of n independent and identically distributed copies of $\boldsymbol{Y}$. After changing variables,

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \left(\frac{n}{2\pi i}\right)^{\mathfrak{k}} \int_{-i\infty}^{+i\infty} \cdots \int_{-i\infty}^{+i\infty} \exp(n[\mathcal{K}_{\boldsymbol{Y}}(\boldsymbol{\beta}) - \boldsymbol{\beta}\boldsymbol{x}]) \, d\boldsymbol{\beta}. \tag{152}$$

Choose a differentiable function $\boldsymbol{\omega}(\boldsymbol{\beta})$ so that

$$\tfrac{1}{2}(\boldsymbol{\omega} - \hat{\boldsymbol{\omega}})^\top (\boldsymbol{\omega} - \hat{\boldsymbol{\omega}}) = \mathcal{K}_{\boldsymbol{Y}}(\boldsymbol{\beta}) - \boldsymbol{\beta}^\top \boldsymbol{x} - \mathcal{K}_{\boldsymbol{Y}}(\hat{\boldsymbol{\beta}}) + \hat{\boldsymbol{\beta}}^\top \boldsymbol{x} \qquad (153)$$

A specific choice for $\boldsymbol{\omega}(\boldsymbol{\beta})$ that will be both useful and easily shown to be analytic will be constructed in this section.

As these results will be used not only in the present chapter but also in a subsequent chapter on approximating conditional distributions, the function $\boldsymbol{\omega}$ will be constructed such that for each integer l, $\omega_1, \ldots, \omega_l$ depend only on $\beta_1, \ldots, \beta_l$. In essence, ω_l becomes the signed likelihood ratio statistic for the hypothesis that the canonical parameter associated with X^l is zero in the exponential family into which $(X^1, \cdots, X^l)$ is embedded, as in §6.4.

For real $\boldsymbol{\beta}$, and $1 \le l \le \mathfrak{k}$, let

$$-\tfrac{1}{2}(\omega_l - \hat{\omega}_l)^2 = \min(\mathcal{K}_{\boldsymbol{Y}}(\boldsymbol{\gamma}) - \gamma_j x^j | \gamma_j = \beta_j \,\forall j < l) -$$
$$\min(\mathcal{K}_{\boldsymbol{Y}}(\boldsymbol{\xi}) - \xi_j x^j | \xi_j = \beta_j \,\forall j \le l) \qquad (154)$$
$$-\tfrac{1}{2}\hat{\omega}_l^2 = \min(\mathcal{K}_{\boldsymbol{Y}}(\boldsymbol{\gamma}) - \gamma_j x^j | \gamma_j = 0 \,\forall j < l) -$$
$$\min(\mathcal{K}_{\boldsymbol{Y}}(\boldsymbol{\xi}) - \xi_j x^j | \xi_j = 0 \,\forall j \le l). \qquad (155)$$

We adopt the notation that indices appearing both as superscripts and subscripts are to be summed over. Determine the signs of the square roots in (154) by requiring that for real $\boldsymbol{\beta}$, ω_l is increasing in β_l, and determine the signs of the square roots in (155) by requiring that $\hat{\omega}_l$ have the same sign as the maximizing γ_l. This parameterization makes proving the following lemma easy. Recall that a function of one or more complex variables having a power series representation about a point is called analytic at that point.

Lemma 6.7.1:   *The function $\boldsymbol{\omega}(\boldsymbol{\beta})$ satisfying (154) for $\boldsymbol{\beta} \in \mathfrak{Q}$ is analytic for $\boldsymbol{\omega}$ in a neighborhood of $\hat{\boldsymbol{\omega}}$. Furthermore, for $j \in \{1, \ldots, \mathfrak{k}\}$, ω_j depends only on $\beta_1, \ldots, \beta_j$, and if the first j components of $\boldsymbol{\beta}$ are zero, so are the corresponding first j components of $\boldsymbol{\omega}$. Finally, $\boldsymbol{\omega}(\boldsymbol{\beta})$ satisfies (153).*

Proof: For each l the components $\gamma_{l+1}, \ldots, \gamma_{\mathfrak{k}}$ at which the first minimum in (154) occurs satisfy $\mathcal{K}_{\boldsymbol{Y}}^j(\beta_1, \ldots, \beta_l, \gamma_{l+1}, \ldots, \gamma_{\mathfrak{k}}) = x^j$ for $j = l + 1, \ldots, \mathfrak{k}$, and hence by the inverse function theorem for complex variables (Bochner and Martin, 1948) is a differentiable function of $\beta_1, \ldots, \beta_l$. A similar characterization holds for $\boldsymbol{\xi}$ involving the the second minimum in (154). Holding $\beta_1, \ldots, \beta_{l-1}$ fixed, the right hand side of (154) achieves a maximal value of 0 at $\beta_l = \gamma_l(\beta_1, \ldots, \beta_{l-1})$. The same argument as gave (81) demonstrates that the right side of (154) may be expressed as $(\beta_l - \gamma_l(\beta_1, \ldots, \beta_{l-1}))^2$ times a function $h_l(\boldsymbol{\beta})$ of $\boldsymbol{\beta}$ analytic at $\hat{\boldsymbol{\beta}}$ and such that $h_l(\hat{\boldsymbol{\beta}})$ is real and strictly positive. Then $\omega_l - \hat{\omega}_l = (\beta_l - \gamma_l(\beta_1, \ldots, \beta_{l-1}))\sqrt{2h_l(\boldsymbol{\beta})}$ describes the function ω_l of (154), where branch of the square root function assigning a positive result is used. Hence ω_l is analytic. Note that ω_l depends only on $\beta_1, \ldots, \beta_l$. Furthermore, when $\beta_1 = \ldots = \beta_l = 0$ then $-\tfrac{1}{2}(\omega_l - \hat{\omega}_l)^2 = -\tfrac{1}{2}\hat{\omega}_l^2$. Possible solutions

for ω_l are 0 or $2\hat{\omega}_2$; the second contradicts the requirement that ω_l be increasing in β_l.

The final claim holds by summing (154), and noting that all of the minimizations except for the first with $l = 1$ and the last with $l = \mathfrak{k}$ cancel. The first minimization with $l = 1$ is degenerate, since all values of the argument to $\mathcal{K}_{\boldsymbol{Y}}$ are set to $\boldsymbol{\beta}$, and the second minimum with $l = \mathfrak{k}$ is achieved at $\boldsymbol{\xi} = \hat{\boldsymbol{\beta}}$.

Q.E.D

Kolassa (1996a) demonstrates that this parameterization may be constructed in a ball whose size is bounded away from zero in an appropriate metric, as $\hat{\boldsymbol{\beta}}$ varies.

Changing variables in (152),

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \left(\frac{n}{2\pi i}\right)^{\mathfrak{k}} \exp(n[\mathcal{K}_{\boldsymbol{Y}}(\hat{\boldsymbol{\beta}}) - \hat{\boldsymbol{\beta}}^{\top}\boldsymbol{x}]) \int_{-i\infty}^{+i\infty} \cdots \int_{-i\infty}^{+i\infty} \exp(\frac{n}{2}(\boldsymbol{\omega} - \hat{\boldsymbol{\omega}})^{\top}(\boldsymbol{\omega} - \hat{\boldsymbol{\omega}})) \left|\frac{\partial\boldsymbol{\beta}}{\partial\boldsymbol{w}}\right| d\boldsymbol{\omega}. \tag{156}$$

By differentiating both sides of (153) with respect to $\boldsymbol{\beta}$, $\sum_l(\omega_l - \hat{\omega}_l)(\partial\omega_l/\partial\beta_j) = \mathcal{K}_{\boldsymbol{Y}}^j(\boldsymbol{\beta}) - x^j$ and $\sum_l(\partial\omega_l/\partial\beta_j)(\partial\omega_l/\partial\beta_{\mathfrak{k}}) + (\omega_l - \hat{\omega}_l)(\partial^2\omega_l/\partial\beta_j\partial\beta_k) = \mathcal{K}_{\boldsymbol{Y}}^{jk}(\boldsymbol{\beta})$. Hence

$$\det\left[\frac{\partial\boldsymbol{\omega}}{\partial\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}})\right] = \sqrt{\det\left[\mathcal{K}_{\boldsymbol{Y}}''(\hat{\boldsymbol{\beta}})\right]}. \tag{157}$$

Application of Theorem 6.5.2 yields the first order terms in (133). Higher-order terms will be calculated by evaluating derivatives of $d\boldsymbol{\beta}/d\boldsymbol{\omega}$; as (133) has already been completely derived in §6.4, and as these calculations are lengthy, the calculations will be deferred until §7.5.

Approximations to probabilities of random vectors supported on a lattice may be obtained as a byproduct of the calculations of this section. When the lattice has unit spacings, then (123) and (124) differ only in the range of integration, and by a fixed multiple involving n. As before, the portion of the inversion integral more than an arbitrarily small distance from the real axis contributes a term of relative size $\exp(-cn)$ for some $c > 0$. Approximation of these integrals to the error present in (136) is identical. This represents a multivariate extension of the approximation of Blackwell and Hodges (1959), who present the univariate density approximation (70) as an approximation to n times univariate lattice probability atoms (73). When the lattice for $\boldsymbol{Y}_j$ has unit spacings in each direction,

$$\mathrm{P}\left[\boldsymbol{X} = \boldsymbol{x}\right] = \frac{\exp(n[\mathcal{K}_{\boldsymbol{Y}}(\hat{\boldsymbol{\beta}}) - \hat{\boldsymbol{\beta}}^{\top}\boldsymbol{x}])}{(2\pi)^{\mathfrak{k}/2} n^{\mathfrak{k}/2} \det\left[\mathcal{K}_{\boldsymbol{Y}}''(\hat{\boldsymbol{\beta}})\right]^{1/2}} \left[1 + \frac{b(\hat{\boldsymbol{\beta}})}{2n} + O\left(n^{-2}\right)\right].$$

A multivariate analogue of (74), the saddlepoint density of a smooth transformation of a random variable, holds in higher dimensions. Various authors, including Phillips (1983), Tierney, Kass, and Kadane (1989), and Jing and Robinson (1994) consider this problem. When $\boldsymbol{Z} = \mathcal{G}(\boldsymbol{X})$, for a transformation $\mathcal{G}$ that is one-to-one

on an open set containing $\mathrm{E}\left[\boldsymbol{X}\right]$. In this case,

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = \frac{\exp(n[\mathcal{K}_{\boldsymbol{Y}}(\hat{\boldsymbol{\beta}}) - \hat{\boldsymbol{\beta}}^{\top}\mathcal{G}^{-1}(\boldsymbol{z})])}{\det\left[\mathcal{G}'(\mathcal{G}^{-1}(\boldsymbol{z}))\right]\sqrt{\det\left[\mathcal{K}_{\boldsymbol{Y}}''(\hat{\boldsymbol{\beta}})\right]}}\left(\frac{n}{2\pi}\right)^{\mathfrak{k}/2}\left[1 + \frac{b(\hat{\boldsymbol{\beta}})}{2n} + O\left(n^{-2}\right)\right]. \quad (158)$$

6.8. Saddlepoint Approximations to Marginal Distributions

When one can exhibit the cumulant generating function of random vector $\boldsymbol{X}$, approximations to marginal distributions of some components of $\boldsymbol{X}$ without regard to the other components may be derived in the same way as (133) was derived. The cumulant generating function for the marginal distribution is the cumulant generating function of the joint distribution with the components of the argument corresponding to components of $\boldsymbol{X}$ not of interest set to zero. This derived cumulant generating function is then used in (133) and (132).

When considering approximations to marginal distributions for transformed random vectors, and the joint cumulant generating function of the transformed vector is not available, this method is no longer adequate. Marginal approximations are obtained by integrating the saddlepoint density approximation using Lemma 6.5.3. This is done most easily by rephrasing (158) in terms of the Legendre transform, defined in (137), evaluated at the transformed ordinate. Recall (138), and let $\mathcal{K}_{\boldsymbol{Z}}^{\dagger}(\boldsymbol{z}) = \mathcal{K}_{\boldsymbol{Y}}^{*}(\mathcal{G}^{-1}(\tilde{\boldsymbol{z}}))$. Then

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = \exp(n\mathcal{K}_{\boldsymbol{Z}}^{\dagger}(\boldsymbol{z}))\frac{\sqrt{\det\left[\mathcal{K}_{\boldsymbol{Y}}^{*''}(\mathcal{G}^{-1}(\boldsymbol{z}))\right]}}{\det\left[\mathcal{G}'(\mathcal{G}^{-1}(\boldsymbol{z}))\right]}\left(\frac{n}{2\pi}\right)^{\mathfrak{k}/2}\left[1 + O\left(n^{-1}\right)\right]. \quad (159)$$

Suppose a density for components $\mathfrak{m}+1, \ldots, \mathfrak{k}$ of $\boldsymbol{Z}$ is desired. Let $\tilde{\boldsymbol{z}}$ minimize $\mathcal{K}_{\boldsymbol{Y}}^{\dagger}(\boldsymbol{z})$, subject to $\tilde{z}^{j} = z^{j} \,\forall j > \mathfrak{m}$. Apply Lemma 6.5.3 with $Q(z^{\mathfrak{m}+1}, \ldots, z^{\mathfrak{k}}) = \mathcal{K}_{\boldsymbol{Y}}^{\dagger}(\boldsymbol{z})$. In this case,

$$f_{Z^{\mathfrak{m}+1},\ldots,Z^{\mathfrak{k}}}(z^{\mathfrak{m}+1}, \ldots, z^{\mathfrak{k}}) = \frac{\exp(n\mathcal{K}_{\boldsymbol{Z}}^{\dagger}(\tilde{\boldsymbol{z}}))}{\det\left[\mathcal{G}'(\mathcal{G}^{-1}(\tilde{\boldsymbol{z}}))\right]}\sqrt{\det\left[\mathcal{K}_{\boldsymbol{Y}}^{*''}(\mathcal{G}^{-1}(\tilde{\boldsymbol{z}}))\right]/\det\left[\boldsymbol{C}\right]}\left(\frac{n}{2\pi}\right)^{(\mathfrak{k}-\mathfrak{m})/2}$$
$$\left[1 + O\left(n^{-1}\right)\right], \quad (160)$$

where $\boldsymbol{C}$ is the submatrix of $\mathcal{K}_{\boldsymbol{Z}}^{\dagger''}(\tilde{\boldsymbol{z}})$ with row and column indices no larger than $\mathfrak{m}$.

Approximation (160) is most closely related to that provided by Jing and Robinson (1994). Phillips (1983) and Tierney, Kass, and Kadane (1989) construct marginal approximations by applying Lemma 6.5.3 to the exact representation for a multivariate density, rather than to the saddlepoint approximation.

As an example, consider the distribution of $Z^{1} = \log(X^{1} + X^{2})$, when $\boldsymbol{X}$ is the mean of n independent vectors, each with independent components, with component j having an exponential distribution, with rate λ_{j}. In this case, $\mathcal{K}_{\boldsymbol{Y}}(\boldsymbol{\beta}) = -\log(1 - \beta_{1}/\lambda_{1}) - \log(1 - \beta_{2}/\lambda_{2})$, Then $\hat{\boldsymbol{\beta}} = (\lambda_{1} - 1/x_{1}, \lambda_{2} - 1/x_{2})$, and $\mathcal{K}_{\boldsymbol{Y}}^{*}(\boldsymbol{x}) = \lambda_{1}\,x^{1} +$

$\lambda_2\, x^2 - 2 - \log(\lambda_1\, x^1) - \log(\lambda_2\, x^2)$. Assume that $\lambda_2 < \lambda_1$, and let $\delta = \lambda_1 - \lambda_2$. Let $\mathcal{G}(\boldsymbol{x}) = (\log(x^1 + x^2), x^2/(x^1 + x^2))$. Then $\mathcal{G}^{-1}(\boldsymbol{z}) = (\exp(z^1)(1 - z^2), \exp(z^1)z^2)$, and

$$\mathcal{K}_{\boldsymbol{Z}}^{\dagger}(\boldsymbol{z}) = -2(z^1 + 1) - \log(\lambda^1\left(1 - z^2\right)) - \log(\lambda^2\, z^2) + e^{z^1}\lambda^1\left(1 - z^2\right) + e^{z^1}\lambda^2\, z^2.$$

Differentiating $\mathcal{K}_{\boldsymbol{Z}}^{\dagger}(\boldsymbol{z})$ with respect to z^2 and setting the result to zero,

$$\tilde{\boldsymbol{z}} = (z^1, \tfrac{1}{2} - \exp(-z^1)/\delta + \sqrt{\tfrac{1}{4} + \exp(-2z^2)\delta^{-1}}).$$

Furthermore,

$$\det\left[\mathcal{K}_{\boldsymbol{Y}}^{*\prime\prime}(\mathcal{G}^{-1}(\tilde{\boldsymbol{z}}))\right] = \frac{\delta^2\, e^{2\, z^1} + 4\left(2 + \sqrt{4 + \delta^2\, e^{2\, z^1}}\right)}{e^{4\, z^1}},$$

$\det[A] = 4 + \delta^2\, e^{2\, z^1} + 2\sqrt{4 + \delta^2\, e^{2\, z^1}}$, and $\det[\mathcal{G}'(\mathcal{G}^{-1}(\boldsymbol{z}))] = \exp(-2z^1)$. Approximation (160) may now be assembled. The behavior of this approximation when $n = 10$ is exhibited in Fig. 14. The exact value of the marginal density may also be calculated. Then

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = \frac{\exp(\exp(z^1)k\left(-\delta z^2 - \lambda_2\right))\lambda_1{}^k\,\lambda_2{}^k\, e^{2kz^1}\left(1 - z^2\right)^{k-1}z^{2k-1}}{\Gamma(k)^2}k^{2k},$$

and

$$
\begin{aligned}
f_{Z^1}(z^1) &= \int_0^1 f_{\boldsymbol{Z}}(z^1, z^2)\, dz^2 \\
&= \int_0^1 \frac{\exp(\exp(z^1)k\left(-\delta z^2 - \lambda_2\right))\lambda_1{}^k\,\lambda_2{}^k\, e^{2kz^1}\left(1 - z^2\right)^{k-1}z^{2k-1}}{\Gamma(k)^2}k^{2k}\, dz^2 \\
&= \frac{(\lambda_1\lambda_2\exp(2z^1))^k\exp(-\exp(z^1)k\lambda_2)}{\Gamma(k)\delta^k} \times \\
&\quad \sum_{j=0}^{k-1}\frac{\Gamma(k+j)}{j!\Gamma(k-j)}[-k\delta\exp(z^1)]^{-j}F_{k+j}(k\delta\exp(z^1)),
\end{aligned}
$$

where F_j is the cumulative distribution function associated with the gamma distribution with shape parameter j and scale parameter 1. The complete parameterization $\boldsymbol{z}$ was used to show that $z^2 = (2 + \delta\exp(z^1) - s)/(2\,\delta\exp(z^1))$ minimizes $\mathcal{K}_{\boldsymbol{Z}}^{\dagger}(\boldsymbol{z})$ with z^1 held fixed. Furthermore, $\mathcal{G}^{-1}(\boldsymbol{z}) = (\exp(z^1)(z^2 - 1), \exp(z^1)z^2)$, and hence the expression for $\tilde{\boldsymbol{x}}$ above holds.

DiCiccio and Martin (1991) apply Corollary 5.5.2 to an approximate univariate marginal density to obtain approximate marginal tail probabilities. Jing and Robinson (1994) apply Corollary 5.5.2 to a marginalized saddlepoint density approximation (160). This application requires the evaluation of $\det\left[\mathcal{K}_{\boldsymbol{Z}}^{\dagger\prime\prime}(\tilde{\boldsymbol{z}})\right]$ for $\tilde{\boldsymbol{z}}$ satisfying $\mathcal{K}_{\boldsymbol{Y}}^{\dagger\,k}(\boldsymbol{z}) = 0\ \forall k \leq \mathfrak{m}$ and $\tilde{z}^j = z^j\ \forall j > \mathfrak{m}$. The following result is useful in this situation:

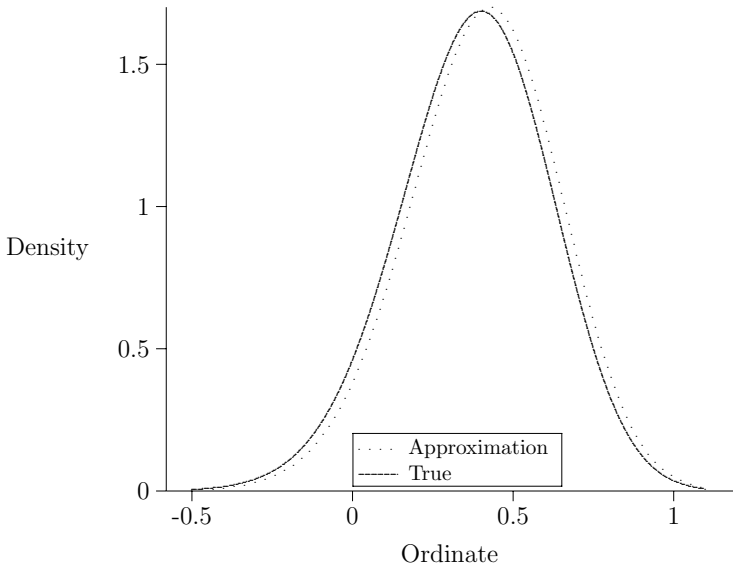*Approximating the Density of the Log of the Sum of Two Gamma Variables*

Fig. 14.

Lemma 6.8.1: *Suppose that a function $h(\boldsymbol{x}) : \mathfrak{R}^{\mathfrak{k}} \to \mathfrak{R}$ is twice differentiable, and that $x_j = y_j$ if $j \leq J$ and $h^k(\boldsymbol{x}) = 0$ for $k > J$. Let $g(\boldsymbol{y}) = h(\boldsymbol{x})$. Express*

$$h''(\boldsymbol{x}) = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^\top & \boldsymbol{C} \end{pmatrix},$$

for $\boldsymbol{A}$ a square matrix with J rows and columns, and the other submatrices with their implied dimensions. Then for $\boldsymbol{y}$ such that $h'(\boldsymbol{x}) = \boldsymbol{o}$, $g''(\boldsymbol{y}) = \boldsymbol{A} - \boldsymbol{B}^\top \boldsymbol{C}^{-1} \boldsymbol{B}$.

Proof: Let $x_m^j = \frac{d}{dy_m} x_j$. Note that $h^{kj}(\boldsymbol{x}) x_j^m = 0$ for $k > J$. Then

$$\begin{pmatrix} \boldsymbol{I} & \boldsymbol{o} \\ \boldsymbol{B}^\top & \boldsymbol{C} \end{pmatrix} \boldsymbol{x}'(\boldsymbol{y}) = \begin{pmatrix} \boldsymbol{I} \\ \boldsymbol{o} \end{pmatrix},$$

where $\boldsymbol{o}$ and $\boldsymbol{I}$ have their implied dimensions. Since

$$\begin{pmatrix} \boldsymbol{I} & \boldsymbol{o} \\ \boldsymbol{B}^\top & \boldsymbol{C} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{o} \\ -\boldsymbol{C}^{-1}\boldsymbol{B}^\top & \boldsymbol{C}^{-1} \end{pmatrix},$$

then

$$\boldsymbol{x}'(\boldsymbol{y}) = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{o} \\ -\boldsymbol{C}^{-1}\boldsymbol{B}^\top & \boldsymbol{C}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{I} \\ \boldsymbol{o} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I} \\ -\boldsymbol{C}^{-1}\boldsymbol{B} \end{pmatrix}.$$

Then $g^{ij}(\boldsymbol{y}) = h^k(\boldsymbol{x})x_k^{ij}(\boldsymbol{y}) + h^{kl}(\boldsymbol{x})x_k^i(\boldsymbol{y})x_l^m(\boldsymbol{y})$, and since $h^k(\boldsymbol{x}) = 0$,

$$g''(\boldsymbol{y}) = (\,\boldsymbol{I} \quad -\boldsymbol{B}^\top \boldsymbol{C}^{-1}\,) \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^\top & \boldsymbol{C} \end{pmatrix} \begin{pmatrix} \boldsymbol{I} \\ -\boldsymbol{C}^{-1}\boldsymbol{B} \end{pmatrix}$$

$$= (\,\boldsymbol{I} \quad -\boldsymbol{B}^\top \boldsymbol{C}^{-1}\,) \begin{pmatrix} \boldsymbol{A} - \boldsymbol{B}^\top \boldsymbol{C}^{-1}\boldsymbol{B} \\ \boldsymbol{o} \end{pmatrix},$$

which gives the result.

Q.E.D

Now apply (112) to the integration of (160), with $g(z^{\mathfrak{k}}) = -\mathcal{K}_Y^\dagger(\tilde{z})$ and $\mathfrak{m} = \mathfrak{k}-1$, and $G(z^{\mathfrak{k}}) = \det\left[\mathcal{G}'(\mathcal{G}^{-1}(\tilde{z}))\right]^{-1} \sqrt{\det\left[\mathcal{K}_Y^{*''}(\mathcal{G}^{-1}(\tilde{z}))\right]/\det\left[\boldsymbol{C}\right]}$. Translate $z^{\mathfrak{k}}$ so that $g'(0) = 0$. When $z^{\mathfrak{k}} = 0$, $\mathcal{K}_Y^{*'}(\mathcal{G}^{-1}(\tilde{z})) = 0$, and

$$\det\left[\mathcal{K}_Y^{*''}(\mathcal{G}^{-1}(\tilde{z}))\right] \det\left[\mathcal{G}'(\mathcal{G}^{-1}(\tilde{z}))\right]^{-2} = \det\left[\mathcal{K}_Z^{\dagger''}(\tilde{z})\right],$$

and so $G(0) = \sqrt{g''(0)}$, using Lemma 6.8.1. Let $\hat{\omega} = \sqrt{2g(x)}$, and $\hat{z} = g'(x)/G(x)$. Hence

$$\int_x^\infty f_{Z^{\mathfrak{k}}}(v)\ dv = \bar{\Phi}(\sqrt{n}\hat{\omega}) + n^{-1/2}\phi(\sqrt{n}\hat{\omega})\left[1/\hat{z} - 1/\hat{\omega} + O(1/\sqrt{n})\right].$$

6.9. Multivariate Saddlepoint Distribution Function Approximations

In this section I present a saddlepoint approximation for multivariate tail probabilities, and in the following section I present formulae for computing some of the components. Unless otherwise noted, results of this section and the next are due to Kolassa (2003a).

Theorem 6.9.1: *Suppose that $\boldsymbol{T}$ is the mean of n independent random vectors, each with cumulant generating function $\mathcal{K}$. Choose a compact subset $\mathcal{W}$ of the range of $\mathcal{K}'$. For $\boldsymbol{r}, \boldsymbol{m} \in \mathfrak{Z}^{\mathfrak{k}}$, and for $\boldsymbol{\Sigma}$ a positive definite matrix, let*

$$I(\boldsymbol{\Sigma}, \boldsymbol{r}, \boldsymbol{m}, \boldsymbol{t}, \hat{\boldsymbol{\beta}}) = \int_{\hat{\boldsymbol{\beta}}-i\infty}^{\hat{\boldsymbol{\beta}}+i\infty} \frac{\exp(\beta_j \boldsymbol{\Sigma}^{jk}\beta_k/2 - \beta_j\delta^{jk}t_k))}{(2\pi i)^{\mathfrak{k}}} \prod_{l=1}^{\mathfrak{k}}[\beta_l^{r_l}(\beta_l - \hat{\beta}_l)^{m_l}]\ d\boldsymbol{\beta}.$$

Here δ^{jk} is 1 if $j = k$ and zero otherwise. When $T_1, \ldots, T_{\mathfrak{k}}$ has a continuous distribution, let $\boldsymbol{s}$ be the vector such that $s_j = -1$ for all j. For $\boldsymbol{t} \in \mathcal{W}$, set

$$Q(\boldsymbol{t}) = \exp(n[\mathcal{K}(\hat{\boldsymbol{\beta}}) - \hat{\beta}_j t^j + \hat{\beta}_j \mathcal{K}^{jk}(\hat{\boldsymbol{\beta}})\hat{\beta}_k/2])\Big[I(n\mathcal{K}''(\hat{\boldsymbol{\beta}}), \boldsymbol{s}, \boldsymbol{o}, n\mathcal{K}''(\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}})$$

$$+ \tfrac{1}{6}\mathcal{K}^{jkl}(\hat{\boldsymbol{\beta}})I(n\mathcal{K}''(\hat{\boldsymbol{\beta}}), \boldsymbol{s}, \boldsymbol{e}_j + \boldsymbol{e}_k + \boldsymbol{e}_l, n\mathcal{K}''(\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}})/\sqrt{n}\Big], \tag{161}$$

where $\boldsymbol{e}_j$ *is the vector with every component* 0 *except for component* j*, which is* 1*. When* $\boldsymbol{T}$ *is supported on a unit lattice, let* $\alpha^j = 1/\hat{\beta}_j - \cosh(\hat{\beta}_j/2)/(2\sinh(\hat{\beta}_j/2))$*, and let*

$$Q(\boldsymbol{t}) = \exp(n[\mathcal{K}(\hat{\boldsymbol{\beta}}) - \hat{\beta}_j t^j + \hat{\beta}_j \mathcal{K}^{jk}(\hat{\boldsymbol{\beta}})\hat{\beta}_k/2]) \prod_{j=1}^{\mathfrak{k}} \frac{\hat{\beta}_j}{2\sinh(\hat{\beta}_j/2)} \Big[$$

$$I(n\mathcal{K}''(\hat{\boldsymbol{\beta}}), \boldsymbol{s}, \boldsymbol{o}, n\mathcal{K}''(\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}) + \alpha^j I(n\mathcal{K}''(\hat{\boldsymbol{\beta}}), \boldsymbol{s}, \boldsymbol{e}_j, n\mathcal{K}''(\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}})$$

$$+ \tfrac{1}{6}\mathcal{K}^{jkl}(\hat{\boldsymbol{\beta}}) I(n\mathcal{K}''(\hat{\boldsymbol{\beta}}), \boldsymbol{s}, \boldsymbol{e}_j + \boldsymbol{e}_k + \boldsymbol{e}_l, n\mathcal{K}''(\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}})/\sqrt{n} \Big]. \qquad (162)$$

Then

$$\mathrm{P}\,[\boldsymbol{T} \geq \boldsymbol{t}] = Q(\boldsymbol{t}) + \exp(n[\mathcal{K}(\hat{\boldsymbol{\beta}}) - \hat{\beta}_j t^j]) E_2(\boldsymbol{t})/n, \ \text{for} \ \sup_{\boldsymbol{t}\in\mathcal{W}} |E_2(\boldsymbol{t})| < \infty. \qquad (163)$$

Proof: Begin with (125). The first step is to reduce the range of integration in (125) from $\hat{\boldsymbol{\beta}} \pm i\infty$ to a bounded range on which the exponent in the integrand is well–behaved. Existence of a bounded density corresponding to the cumulant generating function $\mathcal{K}(\boldsymbol{\beta}) - \beta_j t^j - \mathcal{K}(\hat{\boldsymbol{\beta}}) + \hat{\beta}_j t^j$, and Theorem 6.2.3, show that the contribution to the relative error from outside a small region about $\hat{\boldsymbol{\beta}}$ is exponentially small, and one might choose $\epsilon > 0$ and $q < 1$ so that

$$\sup_{\boldsymbol{t}\in\mathcal{W}, (\beta_j-\hat{\beta}_j)\mathcal{K}^{jk}(\bar{\beta}_k-\hat{\beta}_k)>\epsilon^2} \left|\exp(\mathcal{K}(\boldsymbol{\beta}) - \beta_j t^j)\right| < q,$$

as was done in §3.7; here $\bar{\cdot}$ denotes complex conjugate, and $\hat{\boldsymbol{\beta}}(\boldsymbol{t})$ is the solution to $\mathcal{K}'(\hat{\boldsymbol{\beta}}) = \boldsymbol{t}$. Below the dependence of $\hat{\boldsymbol{\beta}}$ on $\boldsymbol{t}$ is not generally made explicit. Hence, using (125),

$$\mathrm{P}\,[\boldsymbol{T} \geq \boldsymbol{t}] = \oint \frac{\exp(n[\mathcal{K}(\boldsymbol{\beta}) - \beta_j t^j])}{(2\pi i)^{\mathfrak{k}} \prod_{j=1}^{\mathfrak{k}} \psi(\beta_j)} \ d\boldsymbol{\beta} + \exp(n[\mathcal{K}(\hat{\boldsymbol{\beta}}) - \hat{\beta}_j t^j]) E_1,$$

with $|E_1| \leq \exp(-nc)$ for some $c > 0$, and $\oint$ representing integration along paths with real parts given by $\hat{\boldsymbol{\beta}}$, and imaginary parts $\boldsymbol{\gamma}$ satisfying $\gamma_j \hat{\kappa}^{jk} \gamma_k \leq \epsilon^2$. Expand $\mathcal{K}$ about $\hat{\boldsymbol{\beta}}$, and let $\hat{\kappa}^{ij\cdots s} = \mathcal{K}^{ij\cdots s}(\hat{\boldsymbol{\beta}})$. Consider first the case when $\boldsymbol{T}$ has a density; in this case, ψ is the identity function. Then there exists $\boldsymbol{\beta}^\dagger$ between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$, such that

$$\mathrm{P}\,[\boldsymbol{T} \geq \boldsymbol{t}] = \exp(n[\mathcal{K}(\hat{\boldsymbol{\beta}}) - \hat{\beta}_j t^j]) \Big\{ \oint \frac{\exp(n[\gamma_j \hat{\kappa}^{jk} \gamma_k/2])}{(2\pi i)^{\mathfrak{k}} \prod_{j=1}^{\mathfrak{k}}(\hat{\beta}_j + \gamma_j)} \Big[1 + \frac{n}{6}\hat{\kappa}^{klm}\gamma_k\gamma_l\gamma_m +$$

$$\frac{n}{24}\kappa^{\dagger jklm}\gamma_j\gamma_k\gamma_l\gamma_m + \frac{n^2}{72}\kappa^{\dagger jkl}\kappa^{\dagger mpq}\gamma_j\gamma_k\gamma_l\gamma_m\gamma_p\gamma_q\Big] \ d\boldsymbol{\gamma} + E_1 \Big\}$$

$$= \exp(n[\mathcal{K}(\hat{\boldsymbol{\beta}}) - \hat{\beta}_j t^j]) \Big[\oint \frac{\exp(n[\gamma_j \hat{\kappa}^{jk} \gamma_k/2])}{(2\pi i)^{\mathfrak{k}} \prod_{j=1}^{\mathfrak{k}}(\hat{\beta}_j + \gamma_j)} \Big[1 +$$

$$\frac{n}{6}\hat{\kappa}^{klm}\gamma_k\gamma_l\gamma_m\Big] \, d\boldsymbol{\gamma} + E_2/n + E_1\bigg], \tag{164}$$

where $\kappa^{\dagger ij\cdots s} = \mathcal{K}^{ij\cdots s}(\boldsymbol{\beta}^\dagger)$, and

$$E_2 = \oint \frac{n\exp(n[\gamma_j\hat{\kappa}^{jk}\gamma_k/2])}{(2\pi i)^{\mathfrak{k}}\prod_{j=1}^{\mathfrak{k}}(\hat{\beta}_j+\gamma_j)}\left[\frac{\kappa^{\dagger jklm}}{24}\gamma_j\gamma_k\gamma_l\gamma_m + \frac{n\kappa^{\dagger jkl}\kappa^{\dagger mpq}}{72}\gamma_j\gamma_k\gamma_l\gamma_m\gamma_p\gamma_q\right] \, d\boldsymbol{\beta}.$$

The terms in E_2 will be shown to be of the correct size using Theorem 6.11.3, of §6.11. The case when $\boldsymbol{T}$ is similar.

<div align="right">Q.E.D</div>

Wang (1990b) provides a saddlepoint approximation to bivariate cumulative distribution functions.

6.10. Terms in Multivariate Distribution Function Approximations

Non-error terms in expression (164) will be integrated term-wise. Evaluation of $Q(\boldsymbol{t})$ requires evaluation of $I(\boldsymbol{\Sigma}, \boldsymbol{r}, \boldsymbol{m}, \boldsymbol{t}, \hat{\boldsymbol{\beta}})$ for $\boldsymbol{r}$ a vector of integers no smaller than -1. For any $\boldsymbol{r}, \boldsymbol{m} \in 3^{\mathfrak{k}}$, and $j \in \{1, \ldots, \mathfrak{k}\}$,

$$I(\boldsymbol{\Sigma}, \boldsymbol{r}, \boldsymbol{m}, \boldsymbol{t}, \hat{\boldsymbol{\beta}}) = I(\boldsymbol{\Sigma}, \boldsymbol{r} + \boldsymbol{e}_j, \boldsymbol{m} - \boldsymbol{e}_j, \boldsymbol{t}, \hat{\boldsymbol{\beta}}) - \hat{\beta}_j I(\boldsymbol{\Sigma}, \boldsymbol{r}, \boldsymbol{m} - \boldsymbol{e}_j, \boldsymbol{t}, \hat{\boldsymbol{\beta}}). \tag{165}$$

This recursion may be continued until for each j, $m_j = 0$. Alternatively, one might expand each of the factors $(\beta_j - \hat{\beta}_j)^{m_j}$ using the binomial theorem and integrate term-wise. Manipulating (23),

$$I(\boldsymbol{\Sigma}, \boldsymbol{r}, \boldsymbol{o}, \boldsymbol{t}, \hat{\boldsymbol{\beta}}) = \prod_{j=1}^{\mathfrak{k}}(-1)^{r_j+1}\frac{d^{r_j+1}}{(dt_j)^{r_j+1}}\bar{\Phi}(\boldsymbol{t}, \boldsymbol{\Sigma}). \tag{166}$$

The integrals $I(\boldsymbol{\Sigma}, \boldsymbol{r}, \boldsymbol{o}, \boldsymbol{t}, \hat{\boldsymbol{\beta}})$ will be evaluated by expressing them in closed form when $\boldsymbol{r}$ takes on only values in $\{-1, 0\}$, and by providing a recursive representation for other values of $\boldsymbol{r}$. For $\mathcal{A}, \mathcal{B} \subset \mathfrak{F} = \{1, \ldots, \mathfrak{k}\}$, and for vector $\boldsymbol{t}$ of length $\mathfrak{k}$ and $\mathfrak{k} \times \mathfrak{k}$ matrix $\boldsymbol{\Sigma}$, let $\boldsymbol{t}^{\mathcal{A}}$ be the components of $\boldsymbol{t}$ with indices in $\mathcal{A}$, let $\boldsymbol{\Sigma}^{\mathcal{A}}$ the elements of $\boldsymbol{\Sigma}$ with row and column indices in $\mathcal{A}$, if any, and $\boldsymbol{\Sigma}^{\mathcal{A},\mathcal{B}}$ the elements of $\boldsymbol{\Sigma}$ with row indices in $\mathcal{A}$ and column indices in $\mathcal{B}$, if any. Let $\boldsymbol{\Sigma}_{\mathcal{A},\mathcal{B}}$ be the corresponding entries in $\boldsymbol{\Sigma}^{-1}$. Let $\mathcal{A} = \{j | r_j = 0\}$.

Differentiation of the $\mathfrak{k}$ dimensional normal tail probability with respect to components whose indices are in $\mathcal{A}$ yields the marginal density for components in $\mathcal{A}$ times the conditional tail probability of components in $\mathcal{A}^c$ conditional on those in $\mathcal{A}$, all evaluated at $\boldsymbol{t}$, and hence

$$I(\boldsymbol{\Sigma}, \boldsymbol{r}, \boldsymbol{o}, \boldsymbol{t}, \hat{\boldsymbol{\beta}}) = \phi(\boldsymbol{t}^{\mathcal{A}}, \boldsymbol{\Sigma}^{\mathcal{A}})\bar{\Phi}(\boldsymbol{t}^{\mathcal{A}^c} + (\boldsymbol{\Sigma}_{\mathcal{A}}^c)^{-1}\boldsymbol{\Sigma}_{\mathcal{A}^c,\mathcal{A}}\boldsymbol{t}^{\mathcal{A}}, \boldsymbol{\Sigma}_{\mathcal{A}}^c)$$
$$\text{for } r_j \in \{0, -1\}\forall j, \ \mathcal{A} = \{j | r_j = 0\}. \tag{167}$$

If $\boldsymbol{r} \in 3^{\mathfrak{k}}$ such that $r_l \geq -1 \;\forall l$, then

$$I(\boldsymbol{\Sigma}, \boldsymbol{r} + \boldsymbol{e}_j, \mathbf{o}, \boldsymbol{t}, \hat{\boldsymbol{\beta}}) = -\frac{d}{dt_j} I(\boldsymbol{\Sigma}, \boldsymbol{r}, \mathbf{o}, \boldsymbol{t}, \hat{\boldsymbol{\beta}}) \tag{168}$$

If $\boldsymbol{r} \in 3^{\mathfrak{k}}$ such that $r_l \in \{-1, 0\} \;\forall l$ and $r_j = r_k = 0$, (168) implies

$$\begin{aligned}
I(\boldsymbol{\Sigma}, \boldsymbol{r} + \boldsymbol{e}_j, \mathbf{o}, \boldsymbol{t}, \hat{\boldsymbol{\beta}}) = {}& [(\boldsymbol{\Sigma}^{\mathcal{A}})^{-1} \boldsymbol{t}^{\mathcal{A}}]_j I(\boldsymbol{\Sigma}, \boldsymbol{r}, \mathbf{o}, \boldsymbol{t}, \hat{\boldsymbol{\beta}}) \\
& + \sum_{l \in \mathcal{A}^c} I(\boldsymbol{\Sigma}, \boldsymbol{r} + \boldsymbol{e}_l, \mathbf{o}, \boldsymbol{t}, \hat{\boldsymbol{\beta}}) [(\boldsymbol{\Sigma}_{\mathcal{A}}^c)^{-1} \boldsymbol{\Sigma}_{\mathcal{A}^c, \{j\}}]_l,
\end{aligned} \tag{169}$$

and

$$\begin{aligned}
I(\boldsymbol{\Sigma}, \boldsymbol{r} + \boldsymbol{e}_j + \boldsymbol{e}_k, \mathbf{o}, \boldsymbol{t}, \hat{\boldsymbol{\beta}}) = {}& [(\boldsymbol{\Sigma}^{\mathcal{A}})^{-1} \boldsymbol{t}^{\mathcal{A}}]_j I(\boldsymbol{\Sigma}, \boldsymbol{r} + \boldsymbol{e}_k, \mathbf{o}, \boldsymbol{t}, \hat{\boldsymbol{\beta}}) \\
& - (\boldsymbol{\Sigma}^{\mathcal{A}})_{jk}^{-1} I(\boldsymbol{\Sigma}, \boldsymbol{r}, \mathbf{o}, \boldsymbol{t}, \hat{\boldsymbol{\beta}}) \\
& + \sum_{m \in \mathcal{A}^c} I(\boldsymbol{\Sigma}, \boldsymbol{r} + \boldsymbol{e}_m + \boldsymbol{e}_k, \mathbf{o}, \boldsymbol{t}, \hat{\boldsymbol{\beta}}) [(\boldsymbol{\Sigma}_{\mathcal{A}}^c)^{-1} \boldsymbol{\Sigma}_{\mathcal{A}^c, \{m\}}]_j
\end{aligned} \tag{170}$$

and

$$\begin{aligned}
I(\boldsymbol{\Sigma}, \boldsymbol{r} + \boldsymbol{e}_j + \boldsymbol{e}_k + \boldsymbol{e}_l, \mathbf{o}, \boldsymbol{t}, \hat{\boldsymbol{\beta}}) = {}& (\boldsymbol{\Sigma}^{\mathcal{A}})_{jl}^{-1} I(\boldsymbol{\Sigma}, \boldsymbol{r} + \boldsymbol{e}_k, \mathbf{o}, \boldsymbol{t}, \hat{\boldsymbol{\beta}}) \\
- [(\boldsymbol{\Sigma}^{\mathcal{A}})^{-1} \boldsymbol{t}^{\mathcal{A}}]_j I(\boldsymbol{\Sigma}, \boldsymbol{r} + \boldsymbol{e}_k + \boldsymbol{e}_l, \mathbf{o}, \boldsymbol{t}, \hat{\boldsymbol{\beta}}) & + (\boldsymbol{\Sigma}^{\mathcal{A}})_{jk}^{-1} I(\boldsymbol{\Sigma}, \boldsymbol{r} + \boldsymbol{e}_l, \mathbf{o}, \boldsymbol{t}, \hat{\boldsymbol{\beta}}) \\
- \sum_{m \in \mathcal{A}^c} I(\boldsymbol{\Sigma}, \boldsymbol{r} + \boldsymbol{e}_m + \boldsymbol{e}_k + \boldsymbol{e}_l, \mathbf{o}, \boldsymbol{t}, \hat{\boldsymbol{\beta}}) & [(\boldsymbol{\Sigma}_{\mathcal{A}}^c)^{-1} \boldsymbol{\Sigma}_{\mathcal{A}^c, \{m\}}]_j.
\end{aligned} \tag{171}$$

Hence (165), (167), (169), (170), and (171) allow for the recursive calculation of the quantities in $Q(\boldsymbol{t})$.

Expanding $\mathcal{K}$ about $\mathbf{o}$ rather than $\hat{\boldsymbol{\beta}}$ causes (164) to be replaced by

$$\begin{aligned}
Q^*(\boldsymbol{t}) &= \int_{-i\infty}^{+i\infty} \frac{\exp(n[\beta_j \mathcal{K}^{jk}(\mathbf{o}) t_k/2 - \beta_j \delta^{jk} \hat{\beta}_k])}{(2\pi i)^{\mathfrak{k}} \prod_{j=1}^{\mathfrak{k}} \psi(\beta_j)} \left[1 + n \mathcal{K}^{klm}(\mathbf{o}) \beta_k \beta_l \beta_m \right] \, d\boldsymbol{\beta} + O(1/n), \\
&= I(n\mathcal{K}''(\mathbf{o}), \boldsymbol{s}, \mathbf{o}, n\boldsymbol{t}, \mathbf{o}) + \frac{\mathcal{K}^{jkl}(\mathbf{o})}{6\sqrt{n}} I(n\mathcal{K}''(\mathbf{o}), \boldsymbol{s} + \boldsymbol{e}_j + \boldsymbol{e}_k + \boldsymbol{e}_l, \mathbf{o}, n\boldsymbol{t}, \mathbf{o}) \\
&\quad + O(1/n).
\end{aligned}$$

Expansion Q^* is the well–known Edgeworth expansion for $\mathrm{P}\,[\boldsymbol{T} \geq \boldsymbol{t}]$, and is valid even when $\mathcal{K}(\boldsymbol{\beta})$ exists only for pure imaginary arguments. It is also valid when $\boldsymbol{T}$ is confined to a unit lattice and $t_1, \ldots, t_{\mathfrak{k}}$ is evaluated at continuity corrected points (Kolassa, 1989). In this case $\psi(\tau) = 2\sinh(\tau/2)$, and linear terms generated by expanding $\tau/\psi(\tau)$ are zero. Note that $1/\hat{\beta} - \cosh(\hat{\beta}/2)/2\sinh(\hat{\beta}/2)$ evaluated at $\hat{\beta} = 0$ is zero.

Approximation $Q(\boldsymbol{t})$ holds only when all components of $\hat{\boldsymbol{\beta}}$ are positive. Approximations for other $\boldsymbol{t}$ may be calculated recursively. Specifically, suppose that a vector with a negative subscript denotes that vector with the indicated component omitted, and suppose that $\boldsymbol{t}$ corresponds to a multivariate saddlepoint $\hat{\boldsymbol{\beta}}$ with

$\hat{\beta}_j < 0$. Then let $\boldsymbol{u} = (t_1, \ldots, t_{j-1}, -t_j, t_{j+1}, \ldots, t_{\mathfrak{k}})$, and define $\boldsymbol{U}$ analogously. Then $\mathrm{P}\left[\boldsymbol{T} \geq \boldsymbol{t}\right] = \mathrm{P}\left[\boldsymbol{T}_{-j} \geq \boldsymbol{t}_{-1}\right] - \mathrm{P}\left[\boldsymbol{U} \geq \boldsymbol{u}\right]$, and the saddlepoints associated with $\boldsymbol{u}$ has one fewer negative entries than does $\hat{\boldsymbol{\beta}}$, and $Q(\boldsymbol{t})$ may be applied to $\boldsymbol{U}$ and $\boldsymbol{T}_{-j}$.

The vector $\hat{\boldsymbol{\beta}}$ might be interpreted as the maximum likelihood estimator for $\boldsymbol{\beta}$ when $\boldsymbol{t}$ is embedded in the exponential family with density $f_{\boldsymbol{T}}(\boldsymbol{t})\exp(\boldsymbol{\beta}^\top\boldsymbol{t} - \mathcal{K}(\boldsymbol{\beta}))$. Lugannani and Rice (1980), Skovgaard (1987), and Wang (1990b) have developed approximations built around modifications of signed roots of likelihood ratio statistics when $d \leq 2$; the next section will review the two dimensional version for comparison with $Q(\boldsymbol{t})$, and argue why this approach is infeasible for higher dimensions.

Saddlepoint approximations to densities typically yield an error term that is relative; that is, the ratio of the true density to the approximation may be expressed as one plus a negative power of the sample size times a term that is uniformly bounded as n increases and $\boldsymbol{t}$ varies, at least within a compact set such as $\mathcal{W}$. Many authors, including Routledge and Tsao (1995), describe such results. Achieving a uniform relative bound on tail probability approximations is more difficult, even in one dimension. These approximations are typically of form $a(\boldsymbol{t}, n)\bar{\Phi}(\sqrt{n}v(\boldsymbol{t})) + b(\boldsymbol{t}, n)\phi(\sqrt{n}v(\boldsymbol{t}))$, for functions a, b, and v; see, for example, Robinson (1982) and Daniels (1987). The error term typically has a bound of form $n^{-\alpha}C\phi(\sqrt{n}v(\boldsymbol{t}))$, for some constant C. Unfortunately, $\bar{\Phi}(\sqrt{n}v(\boldsymbol{t}))/\phi(\sqrt{n}v(\boldsymbol{t})) \to 0$ as $n \to \infty$, and so uniformity of the relative error fails. The error bound in $Q(\boldsymbol{t})$ is of this form; the error is uniformly exponentially small, but not strictly speaking both relative and uniform.

6.11. Bounding Multiple Complex Integrals with Singularities

New results are needed to generate asymptotic expansions of multivariate complex integrals with singularities in their denominators. These arguments are more delicate than those above, since the integrals considered in this section do not converge absolutely. First, a class of integrals whose values can be bounded by comparison with results from §6.10 are investigated. These integrals have a quadratic exponential factor, and remaining factors equal to the inverse of a constant plus components of the variable of integration. Next, a result involving a more general function multiplying the integrand is presented. Finally, a factor representing sample size is inserted into the quadratic exponent, and similar techniques are used to demonstrate its asymptotic order.

Lemma 6.11.1: *Let $\mathcal{U} \subset \{1, \ldots, \mathfrak{k}\}$ be a collection of unique indices. possibly non-exhaustive and non-unique. Let $\mathcal{B}(\epsilon, \hat{\kappa}^{jk}) = \{\boldsymbol{\gamma} \in \mathfrak{R}^{\mathfrak{k}} | \gamma_j \hat{\kappa}^{jk} \gamma_k \leq \epsilon^2\}$. Suppose for every $s \in \mathcal{U}$, $|u_s| \leq 2\epsilon/\sqrt{\hat{\kappa}^{ss}}$. Let*

$$C(\mathcal{U}, \epsilon) = \sup_{u_s = 0 \forall s \in \mathcal{U}^c, |u_s| \leq 2\epsilon/\sqrt{\hat{\kappa}^{ss}} \forall s \in \mathcal{U}} \left| \int_{\mathcal{B}(\epsilon, \hat{\kappa}^{jk})} \frac{\exp(-\gamma_j \hat{\kappa}^{jk} \gamma_k / 2)}{(2\pi)^{\mathfrak{k}} \prod_{j \in \mathcal{U}} (u_j + i\gamma_j)} \, d\boldsymbol{\gamma} \right|. \qquad (172)$$

Then $C(\mathcal{U}, \epsilon) < \infty$.

Proof: The arguments of §6.10 show that $C(\mathcal{U}, \infty)$ may be expressed as sums, products, and ratios involving the multivariate normal cumulative distribution function and the multivariate normal density, all evaluated in a compact region around $\mathbf{o}$; hence the supremum above is finite. Furthermore, $C(\mathcal{U}, \epsilon) - C(\mathcal{U}, \infty)$ may be expressed as as an integral whose integrand is the power of components of $\boldsymbol{\gamma}$, times a bounded function.

Q.E.D

The previous result is too restrictive for our purposes, since the integral that will eventually be bounded includes integrals like (172), but with powers of γ_j as factors of the integrand, and additionally, some of the components of $\boldsymbol{u}$ will be larger than allowed above. Both of these complications will be accommodated by adding a general function of $\boldsymbol{\gamma}$ as a factor of the integrand. Our eventual use of the next lemma will be confined to g equal to the product of components of its arguments, but a more general statement is necessitated by the first two remarks in the proof simplifying the result without loss of generality.

Lemma 6.11.2: *Let $\mathcal{U} \subset \{1, \ldots, \mathfrak{k}\}$ be as in Lemma 6.11.1. Let $g(\boldsymbol{\xi}, \boldsymbol{u})$ be a function from $\{(\boldsymbol{\xi}, \boldsymbol{u}) | \Re(\boldsymbol{\xi}), \boldsymbol{u} \in \mathcal{B}(2\epsilon, \hat{\kappa}^{jk}), \Im(\boldsymbol{\xi}) \in \mathcal{B}(\epsilon, \hat{\kappa}^{jk})\}$ to $\mathfrak{C}$, such that all derivatives of g with respect to components of $\boldsymbol{\xi}$ have a uniform bound over its domain. Then*

$$\int_{\mathcal{B}(\epsilon, \hat{\kappa}^{jk})} \frac{\exp(-\gamma_j \hat{\kappa}^{jk} \gamma_k / 2)}{(2\pi)^{\mathfrak{k}} \prod_{j \in \mathcal{U}} (u_j + i\gamma_j)} g(\boldsymbol{u} + i\boldsymbol{\gamma}, \boldsymbol{u}) \ d\boldsymbol{\gamma} < K, \qquad (173)$$

where K is independent of u_s.

Proof: Without loss of generality, we may assume that for every $j \in \mathcal{U}$, $|u_s| \leq 2\epsilon/\sqrt{\hat{\kappa}^{ss}}$; otherwise, the corresponding factors in the numerator of the integrand could be moved into g. Furthermore, we may assume that $\hat{\kappa}^{jk} = 0$ if $j \in \mathcal{U}$ and $k \notin \mathcal{U}$; otherwise, γ_k may be redefined as a linear combination of γ_k and the γ_j for $j \in \mathcal{U}$. Finally, without loss of generality, assume that $\mathcal{U}$ consists of consecutive integers starting with 1.

Suppose that p is the minimal index such $g(\boldsymbol{\xi}, \boldsymbol{u}) = g(\boldsymbol{\gamma}, \boldsymbol{u})$ whenever $\gamma_s = \xi_s$ if $s \leq p$ for all $\boldsymbol{v}$. If $p = 0$, then g does not depend on any of the variables occurring in the denominator of the integrand, and, because of our assumptions about $\hat{\kappa}^{jk}$, the integral may be expressed as the product of two factors. One factor consists of integration with respect to γ_j for $j \in \mathcal{U}$, and the other consists of integration with respect to the remaining variables. The first integral is bounded by Lemma 6.11.1. The second is bounded by the arguments used in Theorem 6.5.2. Hence the theorem holds for $p = 0$.

During the above factoring of the integral, an exponentially small error is incurred, when transforming the original spherical range of integration into the product of two separate spherical ranges of integration of lower dimension.

Suppose that the theorem holds for p, and suppose that $p + 1 \in \mathcal{U}$. Let $\boldsymbol{\xi}$ be such that $\xi_j = \gamma_j$ if $j \neq p + 1$, and $\xi_{p+1} = -u_{p+1}$. Express (173) as the sum of two integrals, with $g(\boldsymbol{\xi}, \boldsymbol{u})$ and $g(\boldsymbol{\gamma}, \boldsymbol{u}) - g(\boldsymbol{\xi}, \boldsymbol{u})$ replacing $g(\boldsymbol{\gamma}, \boldsymbol{u})$. Apply the result for p to the first of these integrals, and note that the second term may be expressed as (173), with $g(\boldsymbol{\gamma}, \boldsymbol{u})$ replaced by $(g(\boldsymbol{\gamma}, \boldsymbol{u}) - g(\boldsymbol{\xi}, \boldsymbol{u}))/(u_{p+1} + i\gamma_{p+1})$, and $p + 1$ removed from $\mathcal{U}$. The result follows by induction.

<div align="right">Q.E.D</div>

The next result considers integrals with n multiplying the quadratic term in the exponent.

Theorem 6.11.3: *Let $\mathcal{U} \subset \{1, \ldots, \mathfrak{k}\}$ be a collection of unique indices, and let $\mathcal{T}$ be a collection of entries of $\mathcal{U}$, possibly non-exhaustive and non-unique. Let g be a function from $[-\epsilon, \epsilon]^{\mathfrak{k}} \times \mathfrak{R}^{\mathfrak{k}}$ to $\mathfrak{C}$, such that every derivative of g, with respect to components of its first argument, has a bound independent of $\hat{\boldsymbol{\beta}}$ and its arguments. Then*

$$\int_{\mathcal{B}(\epsilon, \hat{\kappa}^{jk})} \frac{\exp(-n\gamma_j \hat{\kappa}^{jk} \gamma_k / 2)}{(2\pi)^{\mathfrak{k}}} g(\boldsymbol{\gamma}, \hat{\boldsymbol{\beta}}) \frac{\prod_{j \in \mathcal{T}} \gamma_t}{\prod_{j \in \mathcal{U}}(\hat{\beta}_s + i\gamma_s)} \, d\boldsymbol{\gamma} = O(n^{(\#(\mathcal{U}) - \mathfrak{k} - \#(\mathcal{T}))/2})$$

Proof: This result will be proved by induction, in a manner similar to the previous proof. Again without loss of generality, one may assume that $\hat{\kappa}^{jk} = 0$ if $j \in \mathcal{U}$ and $k \in \mathcal{U}^c$. Suppose that g depends only on components of its argument that lie outside $\mathcal{U}$. Then the integral may be factored into two integrals, one over components of $\boldsymbol{\gamma}$ with components in $\mathcal{U}$, and the other over the remaining components of $\boldsymbol{\gamma}$. The first factor is of size $O(n^{(\#(\mathcal{U}) - \mathfrak{k})/2})$, using standard methods, and the second factor is

$$n^{-\#(\mathcal{T})/2} \int_{\mathcal{B}(\sqrt{n}\epsilon, \hat{\kappa}^{jk})} \frac{\exp(-Q/2)}{(2\pi)^{\mathfrak{k} - \#(\mathcal{U})}} \frac{\prod_{j \in \mathcal{T}} \gamma_j}{\prod_{j \in \mathcal{U}}(\sqrt{n}\hat{\beta}_j + i\gamma_j)} \prod_{j \in \mathcal{U}} d\gamma_j. \tag{174}$$

Here Q represents the contribution to $\gamma_j \hat{\kappa}^{jk} \gamma_k$ when the implicit sum over j and k is restricted to indices in $\mathcal{U}$. The contribution to the above integral from $\mathcal{B}(\epsilon, \hat{\kappa}^{jk})^c$ is bounded, and the remaining integral is bounded by Lemma 6.11.2. Now suppose the result holds when fewer than p of those components of its first argument which g actually depends on are in $\mathcal{U}$. Choose g dependent only on p components of its first arguments, with indices are in $\mathcal{U}$. Let m be one such index. Let $g_1(\boldsymbol{\gamma}, \hat{\boldsymbol{\beta}})$ be $g(\boldsymbol{\gamma}, \hat{\boldsymbol{\beta}})$ with $\gamma_m = 0$, and let $g_2(\boldsymbol{\gamma}) = (g(\boldsymbol{\gamma}, \hat{\boldsymbol{\beta}}) - g_1(\boldsymbol{\gamma}, \hat{\boldsymbol{\beta}}))/\gamma_m$. Then the result holds if both

$$\int_{-\epsilon}^{+\epsilon} \frac{\exp(-n\gamma_j \hat{\kappa}^{jk} \gamma_k / 2)}{(2\pi)^{\mathfrak{k}}} g_1(\boldsymbol{\gamma}) \frac{\prod_{j \in \mathcal{T}} \gamma_t}{\prod_{j \in \mathcal{U}}(\hat{\beta}_s + i\gamma_s)} \, d\boldsymbol{\gamma} = O(n^{(\#(\mathcal{U}) - \mathfrak{k} - \#(\mathcal{T}))/2})$$

$$\int_{-\epsilon}^{+\epsilon} \frac{\exp(-n\gamma_j \hat{\kappa}^{jk} \gamma_k / 2)}{(2\pi)^{\mathfrak{k}}} g_2(\boldsymbol{\gamma}) \frac{\prod_{j \in \mathcal{T}} \gamma_t}{\prod_{j \in \mathcal{U}, j \neq m}(\hat{\beta}_s + i\gamma_s)} \, d\boldsymbol{\gamma} = O(n^{(\#(\mathcal{U}) - \mathfrak{k} - \#(\mathcal{T}))/2}).$$

In both cases, g_l depends on fewer than p components with indices in represented in the denominator, and the result holds.

6.12. Exercises

1. Prove Lemma 6.2.3.

7
Conditional Distribution Approximations

Often inference for a subset of model parameters is desired, and the others are treated as nuisance parameters. The sampling distribution of quantities like p-values and confidence intervals generally depends on the nuisance parameter value. Among the many methods removing this dependence is conditional inference, in which sufficient statistics for nuisance parameters are conditioned on. Calculations involving these conditional distributions are often quite difficult. This chapter will develop methods for approximating densities and distribution functions for conditional distributions.

Consider full canonical exponential family distributions consisting of distributions with densities of the form

$$f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{v}) = \exp(\boldsymbol{y}^\top \boldsymbol{v} - \mathcal{H}_{\boldsymbol{Y}}(\boldsymbol{v}) - \mathcal{G}_{\boldsymbol{Y}}(\boldsymbol{y})). \tag{175}$$

Here the cumulant generating function of $\boldsymbol{Y}$ is $\mathcal{K}_{\boldsymbol{Y}}(\boldsymbol{\beta}; \boldsymbol{v}) = \mathcal{H}_{\boldsymbol{Y}}(\boldsymbol{\beta} + \boldsymbol{v}) - \mathcal{H}_{\boldsymbol{Y}}(\boldsymbol{v})$. In such families the conditional distribution of a subset of sufficient statistics conditional on the rest is also a canonical exponential family, depending only on the corresponding canonical parameters:

$$f_{\boldsymbol{Y}}(Y^1 = y^1, \cdots, Y^{\mathfrak{m}} = y^{\mathfrak{m}} | Y^{\mathfrak{m}+1} = y^{\mathfrak{m}+1}, \cdots, Y^{\mathfrak{k}} = y^{\mathfrak{k}}; \boldsymbol{v})$$

$$= \frac{\exp(y^1 v_1 + \cdots + y^{\mathfrak{k}} v_{\mathfrak{k}} - \mathcal{G}_{\boldsymbol{Y}}(\boldsymbol{Y}))}{\int_{-\infty}^{\infty} \cdot \int_{-\infty}^{\infty} \exp(y^1 v_1 + \cdots + y^{\mathfrak{k}} v_{\mathfrak{k}} - \mathcal{G}_{\boldsymbol{Y}}(\boldsymbol{Y})) \, dy^1 \cdots dy^{\mathfrak{m}}}.$$

This conditional distribution may be used for inference on a subset of model parameters, while treating the remaining model parameters as nuisance parameters.

Also note that if $\boldsymbol{X}$ is the mean of n independent and identically distributed random vectors with density (175), then the density for $\boldsymbol{X}$ may be expressed as

$$f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\theta}) = \exp(\boldsymbol{x}^\top \boldsymbol{\theta} - \mathcal{H}_{\boldsymbol{X}}(\boldsymbol{\theta}) - \mathcal{G}_{\boldsymbol{X}}(\boldsymbol{x})), \tag{176}$$

with $\boldsymbol{\theta} = n\boldsymbol{v}$, and so a similar conditioning argument can be used for inference. Furthermore, if $\boldsymbol{y}$ and $\boldsymbol{v}$ in (175) are of length n, with components of $\boldsymbol{Y}$ independent, so that $\mathcal{H}_{\boldsymbol{Y}}(\boldsymbol{v}) = \sum_{j=1}^{n} \mathcal{H}_{Y_j}(v_j)$, and if $\boldsymbol{Z}$ is a $n \times \mathfrak{k}$ matrix such that

$$\boldsymbol{v} = \boldsymbol{Z}\boldsymbol{\theta}, \tag{177}$$

then $\boldsymbol{T} = \boldsymbol{Z}^{\top}\boldsymbol{Y}$ is sufficient for $\boldsymbol{\theta}$, and the density or probabilities for $\boldsymbol{T}$ are given by

$$f_{\boldsymbol{T}}(\boldsymbol{t};\boldsymbol{\theta}) = \exp(\boldsymbol{t}^{\top}\boldsymbol{\theta} - \mathcal{H}_{\boldsymbol{T}}(\boldsymbol{\theta}) - \mathcal{G}_{\boldsymbol{T}}(\boldsymbol{t})), \qquad (178)$$

with $\mathcal{H}_{\boldsymbol{T}}(\boldsymbol{\theta}) = \sum_{j=1}^{n} \mathcal{H}_{Y_j}(\boldsymbol{z}_j^{\top}\boldsymbol{\theta})$, for $\boldsymbol{z}_j$ representing row j of $\boldsymbol{Z}$. Relations (175) and (177) represent the simplest examples of generalized linear models, including logistic and Poisson regression. The methods derived in this chapter will allow inference on one component of $\boldsymbol{\theta}$ without specification of the remaining components.

This chapter will consider approximations that aid in conditional inference for some components of a parameter vector that is a linear transformation of the canonical parameter in an exponential family as in (176), especially those arising under repeated sampling from a family like (175). The question of inference on a subset of model parameters that cannot be expressed as canonical parameters in an exponential family will be addressed in the next chapter.

7.1. Double Saddlepoint Density Approximations

In this section an approximation to the conditional density of the first $\mathfrak{m}$ components of a random vector $\boldsymbol{X} = (X^1, \ldots, X^{\mathfrak{k}})$ arising as the mean of n independent and identically distributed random vectors $\boldsymbol{Y}_j$, each with cumulant generating function $\mathcal{K}_{\boldsymbol{Y}}$. is derived. Let $\hat{\boldsymbol{\beta}}$ satisfy $\mathcal{K}'_{\boldsymbol{Y}}(\hat{\boldsymbol{\beta}}) = \boldsymbol{x}$, as in (132). Let $\tilde{\boldsymbol{\beta}}$ satisfy

$$\tilde{\beta}_1, \ldots, \tilde{\beta}_{\mathfrak{m}} = 0, \ \mathcal{K}_{\boldsymbol{Y}}^j(\tilde{\boldsymbol{\beta}}) = x^j \text{ for } j > \mathfrak{m}. \qquad (179)$$

Barndorff–Nielsen and Cox (1979) present the double saddlepoint approximation to the conditional density. Its name derives from the fact that the saddlepoint equations are solved twice, once in (132) to obtain $\hat{\boldsymbol{\beta}}$ and once in (179) to obtain $\tilde{\boldsymbol{\beta}}$. This approximation might also be derived from two applications of (133); the ratio of (133) calculated for $(X^{\mathfrak{m}+1}, \ldots, X^{\mathfrak{k}})$ and for $\boldsymbol{X}$, with terms of size $o(1/n)$ deleted, is valid to $O(n^{-2})$, and is

$$f_{X^1,\ldots,X^{\mathfrak{m}}|X^{\mathfrak{m}+1},\ldots,X^{\mathfrak{k}}}(x^1,\ldots,x^{\mathfrak{m}}|x^{\mathfrak{m}+1},\ldots,x^{\mathfrak{k}})$$

$$= n^{\mathfrak{m}/2} \frac{\det\left[\mathcal{K}''_{Y^{\mathfrak{m}+1},\ldots,Y^{\mathfrak{k}}}(\tilde{\boldsymbol{\beta}})\right]}{\det\left[\mathcal{K}''_{\boldsymbol{Y}}(\hat{\boldsymbol{\beta}})\right]} \frac{\exp(n[\mathcal{K}_{\boldsymbol{Y}}(\hat{\boldsymbol{\beta}}) - \mathcal{K}_{\boldsymbol{Y}}(\tilde{\boldsymbol{\beta}}) - (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^{\top}\boldsymbol{x}])}{(2\pi)^{\mathfrak{m}/2}}$$

$$\times \left(1 + n^{-1}\left[\tfrac{1}{8}(\hat{\rho}_{13} - \tilde{\rho}_{13}) + \tfrac{1}{12}(\hat{\rho}_{23} - \tilde{\rho}_{23}) - \tfrac{1}{8}(\hat{\rho}_{13} - \tilde{\rho}_{13})\right] + O(n^{-2})\right). \qquad (180)$$

Here the invariants $\hat{\rho}$ are exactly as in (136), and the invariants $\tilde{\rho}$ are as in (136), calculated from the derivatives of the cumulant generating function of $(X^{\mathfrak{m}+1}, \ldots, X^{\mathfrak{k}})$, evaluated at $\tilde{\boldsymbol{\beta}}$.

The two approximations whose ratio gives (180) also represent the multivariate Blackwell–Hodges approximations for random vectors $\boldsymbol{X}$ that are means of random vectors supported on a unit lattice, with an appropriately adjusted leading factor, from (142). Hence in this case, the right-hand side of (180) is also an asymptotic expansion for $\mathrm{P}\left[X^1 = x^1, \ldots, X^{\mathfrak{m}} = x^{\mathfrak{m}}|X^{\mathfrak{m}+1} = x^{\mathfrak{m}+1}, \ldots, X^{\mathfrak{k}} = x^{\mathfrak{k}}\right]$.

Jing and Robinson (1994) present approximations for conditional densities arising from smooth but nonlinear transformations of random vectors. Dividing (159) by (160),

$$f_{Z^1,\dots,Z^{\mathfrak{m}}|Z^{\mathfrak{m}+1},\dots,Z^{\mathfrak{k}}}(z^1,\dots,z^{\mathfrak{m}}|z^{\mathfrak{m}+1},\dots,z^{\mathfrak{k}})$$

$$= n^{\mathfrak{m}/2} \frac{\exp(n[\mathcal{K}_Y^{\dagger}(\hat{z}) - \mathcal{K}_Y^{\dagger}(\tilde{z})])}{(2\pi)^{\mathfrak{m}/2}} \sqrt{\frac{\det\left[\mathcal{K}_Z^{\dagger''}(\tilde{z})\right]}{\det\left[\mathcal{K}_Z^{\dagger''}(z)\right]\det[A]}} \left(1 + O(n^{-1})\right). \quad (181)$$

where again $\tilde{z}$ minimizes $\mathcal{K}_Y^*(\mathcal{G}^{-1}(z))$, $\mathcal{K}_Z^{\dagger}(z) = \mathcal{K}_Y^*(\mathcal{G}^{-1}(\tilde{z}))$, and A is the submatrix of $\mathcal{K}_Z^{\dagger''}(\tilde{z})$, with row and column indices no larger than $\mathfrak{m}$. DiCiccio, Martin, and Young (1993) apply similar techniques to an exact joint distribution, rather than to its saddlepoint approximation.

7.2. The Sequential Saddlepoint Approximation

Most statistical applications rely on the evaluation of tail probabilities rather than densities. Perhaps the most direct approach to calculating conditional tail probabilities involves calculating the cumulant generating function for the conditional distribution. Specifically, when desiring the distribution of $(X^1, \dots, X^{\mathfrak{m}})$ conditional on $(X^{\mathfrak{m}+1}, \dots, X^{\mathfrak{k}})$, one could calculate

$$\mathcal{K}_{X^1,\dots,X^{\mathfrak{m}}|X^{\mathfrak{m}+1},\dots,X^{\mathfrak{k}}}(\beta_1,\dots,\beta_{\mathfrak{m}})$$

$$= \log\left(\frac{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp(\sum_{j=1}^{\mathfrak{m}} x^j \beta_j) f_X(x) \, dx^1 \cdots dx^{\mathfrak{m}}}{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(x) \, dx^1 \cdots dx^{\mathfrak{m}}}\right). \quad (182)$$

This is often not obtainable in closed form even when the unconditional cumulant generating function is available.

In this case cumulant generating function derivatives may be obtained from log likelihood derivatives. This relationship will be further explored in the next chapter. Fraser, Reid, and Wong (1991) exploit this fact, and suggest approximating the marginal density in the denominator of (182) by its saddlepoint approximation, and approximating numerically the derivatives of this approximate conditional likelihood. The result of this method is called the sequential saddlepoint approximation. Consider the density (176), arising from (175) under repeated sampling. The cumulant generating function for the marginal distribution of $X_{-\mathfrak{m}} = (X^{\mathfrak{m}+1}, \dots, X^{\mathfrak{k}})$ is $\mathcal{H}_X(\theta + \beta) - \mathcal{H}_X(\theta)$, with $\beta_j = 0$ for $j \leq \mathfrak{m}$, and the multivariate saddlepoint satisfies $\mathcal{H}_X^i(\theta + \tilde{\beta}) = x^i$ for $i \geq \mathfrak{m}+1$ and $\tilde{\beta}_j = 0$ for $j \leq \mathfrak{m}$, where $\mathcal{H}_X^i$ is the partial derivative of $\mathcal{H}_X$ with respect to argument i. Let $\tilde{\theta} = \theta + \tilde{\beta}$. Then $\tilde{\theta}$ satisfies

$$\mathcal{H}_X^i(\tilde{\theta}) = x^i \text{ for } i \geq \mathfrak{m} + 1, \ \tilde{\theta}_j = \theta_j \text{ for } j \leq \mathfrak{m}. \quad (183)$$

The saddlepoint approximation to this marginal density $f_{X_{-\mathfrak{m}}}(x_{-\mathfrak{m}})$, treating X as

the mean of 1 independent and identically distributed copy of $\boldsymbol{X}$, is proportional to

$$\exp\left(\sum_{j=\mathfrak{m}+1}^{\ell}(\theta_j - \tilde{\theta}_j)x^j - \mathcal{H}_{\boldsymbol{X}}(\boldsymbol{\theta}) + \mathcal{H}_{\boldsymbol{X}}(\tilde{\boldsymbol{\theta}})\right)\bigg/\sqrt{\det\left[\mathcal{H}''_{\boldsymbol{X}_{-\mathfrak{m}}}(\tilde{\boldsymbol{\theta}})\right]},$$

where $\mathcal{H}''_{\boldsymbol{X}_{-\mathfrak{m}}}(\tilde{\boldsymbol{\theta}})$ is the matrix of second derivatives of $\mathcal{H}_{\boldsymbol{X}}$ with respect to arguments with indices greater than $\mathfrak{m}$. Dependence in sample size is implicit in the relationship between $\mathcal{H}_{\boldsymbol{X}}$ and $\mathcal{H}_{\boldsymbol{Y}}$. An approximation to the conditional density $f_{X^1,\ldots,X^{\mathfrak{m}}|\boldsymbol{X}_{-\mathfrak{m}}}(x^1,\ldots,x^{\mathfrak{m}}|\boldsymbol{x}_{-\mathfrak{m}}) = f_{\boldsymbol{X}}(\boldsymbol{x})/f_{\boldsymbol{X}_{-\mathfrak{m}}}(\boldsymbol{x}_{-\mathfrak{m}})$ is

$$\exp\left(\sum_{j=1}^{\mathfrak{m}}x^j\theta_j + \sum_{j=\mathfrak{m}+1}^{\ell}\tilde{\theta}_jx^j - \mathcal{H}_{\boldsymbol{X}}(\tilde{\boldsymbol{\theta}}) - \mathcal{G}_{X^1,\ldots,X^{\mathfrak{m}}|\boldsymbol{X}_{-\mathfrak{m}}}(\boldsymbol{X})\right)\bigg/\sqrt{\det\left[\mathcal{H}''_{\boldsymbol{X}_{-\mathfrak{m}}}(\tilde{\boldsymbol{\theta}})\right]},$$

where $\mathcal{G}_{X^1,\ldots,X^{\mathfrak{m}}|\boldsymbol{X}_{-\mathfrak{m}}}(\boldsymbol{X})$ is a function that might be quite difficult to evaluate, but fortunately will not be needed in the sequel. This corresponds to the conditional profile log likelihood of Cox and Reid (1987) after adjustment for non-orthogonality of parameters; also see Levin (1990). This thread, with more references, will be picked up later in §8.6. Approximately, then, the cumulant generating function for this conditional distribution is

$$-\sum_{j>\mathfrak{m}}\tilde{\theta}_jx^j + \mathcal{H}_{\boldsymbol{X}}(\tilde{\boldsymbol{\theta}}) - \tfrac{1}{2}\log\left(\det\left[\mathcal{H}''_{\boldsymbol{X}_{-\mathfrak{m}}}(\tilde{\boldsymbol{\theta}})\right]\right). \tag{184}$$

Derivatives of this cumulant generating function can be calculated numerically, and be used in conjunction with univariate saddlepoint formulae such as (66) and (109). Fraser, Reid, and Wong (1991) consider this argument with $\mathfrak{m} = 1$.

As an example, let X^1 and X^2 be independent exponential random variables, from the family $f_X(x;\theta) = \exp(x\theta - x + \log(1-\theta))$ for $x > 0$ and $\theta < 1$. Consider the distribution of $T^1 = X^1$ conditional on $T^2 = t^2$, where $T^2 = X^1 + X^2$. The joint cumulant generating function of $\boldsymbol{T} = (T^1, T^2)$ is $-\log(1 - \theta_1 - 2\theta_2 + \theta_2(\theta_1 + \theta_2))$. The saddlepoint equation for θ_1 fixed is $(2-\theta_1-2\theta_2)/(1-\theta_1-2\theta_2+\theta_2(\theta_1+\theta_2)) = t^2$; this equation has two solutions for θ_2:

$$\theta_2 = 1 - \tfrac{1}{2}\theta_1 - \left(1 \pm \sqrt{1 + \tfrac{1}{4}\theta_1^2(t^2)^2}\right)/t^2.$$

The solution taking $\pm$ as $-$ reduces to $\theta_2 = 1$ when $\theta_1 = 0$, and as this lies outside the range of definition for the full cumulant generating function, the solution with $\pm$ taken as $+$ is chosen. The approximate conditional profile log likelihood is

$$\tfrac{1}{2}\log(2 + \sqrt{4 + \theta_1^2(t^2)^2}) + \tfrac{1}{2}\log(\sqrt{4 + \theta_1^2(t^2)^2}) + \theta_1 t^1 - \tfrac{1}{2}\theta_1 t^2 - \tfrac{1}{2}\sqrt{4 + \theta_1^2(t^2)^2},$$

and hence $\mathcal{K}^{\dagger}_{T^1|T^2}(\theta_1) = 2 - \tfrac{1}{2}\log(4 + 2\sqrt{4 + {\theta_1}^2(t^2)^2}) - \tfrac{1}{4}\log(4 + \theta_1^2(t^2)^2) + \tfrac{1}{2}\theta_1 t^2 + \tfrac{1}{2}\sqrt{4 + {\theta_1}^2(t^2)^2}$ represents the cumulant generating function approximation (184). The true conditional distribution of T^1 is uniform over $(0, t^2)$, and if $t^2 = 1$, the

true cumulant generating function of this conditional distribution is $\mathcal{K}_{T^1|T^2}(\theta_1) = \log(2\sinh(\theta_1/2)/\theta_1) + \frac{1}{2}\theta_1$, the cumulant generating function of a uniform random variable on $(0,1)$. Fig. 15 illustrates how well $\mathcal{K}^{\dagger}_{T^1|T^2}(\theta_1)$ approximates $\mathcal{K}_{T^1|T^2}(\theta_1)$.

This example illustrates an interesting aspect of joint cumulant generating functions. The domain of $\mathcal{K}_{\boldsymbol{T}}$ is not the product of the domains of $\mathcal{K}_{T^1}$ and $\mathcal{K}_{T^2}$. Thus, while the domain of $\mathcal{K}_{T^1}$ is $(-\infty,1)$, the domain of $\mathcal{K}^{\dagger}_{T^1|T^2}(\theta_1)$ is $\mathfrak{R}$, as is the domain of $\mathcal{K}_{T^1|T^2}$.

An alternate expression for (184) may be derived by considering the problem of conditional inference based on a $\boldsymbol{T}$, a linear transformation of $\boldsymbol{X}$. Then the likelihood equations (183) can be expressed in terms of $\mathcal{H}_{\boldsymbol{X}}$ rather than $\mathcal{H}_{\boldsymbol{T}}$, which is an advantage when $\mathcal{H}_{\boldsymbol{X}}$ has a simple form.

Lemma 7.2.1: *Suppose that $\boldsymbol{X}$ has the density (176), and that $\boldsymbol{T} = \boldsymbol{BX}$, with $\boldsymbol{B}$ invertible. Express $\boldsymbol{B}$ as a partitioned matrix, with the first row, giving T^1, denoted by $\boldsymbol{c}$, and the remaining rows denoted by the $(\mathfrak{k}-1) \times \mathfrak{k}$ matrix $\boldsymbol{A}$. Suppose that $\boldsymbol{Ac}^{\top} = \boldsymbol{o}$, a vector of all zeros. Then $\boldsymbol{T}$ has a density of the form*

$$f_{\boldsymbol{T}}(\boldsymbol{t}) = \exp(\boldsymbol{t}^{\top}\boldsymbol{\eta} - \mathcal{H}_{\boldsymbol{T}}(\boldsymbol{\eta}) - \mathcal{G}_{\boldsymbol{T}}(\boldsymbol{t})),$$

with $\mathcal{H}_{\boldsymbol{T}}(\boldsymbol{\eta}) = \mathcal{H}_{\boldsymbol{X}}(\boldsymbol{B}^{\top}\boldsymbol{\eta})$. Suppose $\boldsymbol{t}$ is the observed value of $\boldsymbol{T}$, and $\boldsymbol{x}$ is the associated value for $\boldsymbol{X}$. Let $\hat{\boldsymbol{\eta}}$ and $\tilde{\boldsymbol{\eta}}$ satisfy $\mathcal{H}'_{\boldsymbol{T}}(\hat{\boldsymbol{\eta}}) = \boldsymbol{t}^{\top}$ and $\mathcal{H}'_{\boldsymbol{T}_{-1}}(\tilde{\boldsymbol{\eta}}) = \boldsymbol{t}^{\top}_{-1}$ with $\tilde{\eta}_1 = \eta_1$. Let $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ satisfy $\mathcal{H}'_{\boldsymbol{X}}(\hat{\boldsymbol{\theta}}) = \boldsymbol{x}^{\top}$ and $\mathcal{H}'_{\boldsymbol{X}_{-1}}(\tilde{\boldsymbol{\theta}})\boldsymbol{A} = \boldsymbol{x}^{\top}_{-1}\boldsymbol{A}$ with $\boldsymbol{c}^{\top}\tilde{\boldsymbol{\theta}}/(\boldsymbol{c}^{\top}\boldsymbol{c}) = \eta_1$. Then $\boldsymbol{B}^{\top}\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\theta}}$ and $\boldsymbol{B}^{\top}\tilde{\boldsymbol{\eta}} = \tilde{\boldsymbol{\theta}}$.

Proof: This result follows from straight-forward calculation.

<div align="right">Q.E.D</div>

The following lemma will also be useful:

Lemma 7.2.2: *Suppose that $\boldsymbol{V}$ is a symmetric invertible matrix of size $\mathfrak{k} \times \mathfrak{k}$, that $\boldsymbol{A}$ is a $\mathfrak{k} \times (\mathfrak{k}-1)$ matrix of full rank, and $\boldsymbol{c}$ is a vector of length $\mathfrak{k}$ such that $\boldsymbol{A}^{\top}\boldsymbol{c} = \boldsymbol{o}$. Let $\boldsymbol{B}$ be the $\mathfrak{k} \times \mathfrak{k}$ matrix whose first row is $\boldsymbol{c}$ and whose remaining rows are $\boldsymbol{A}$. Then $\det\left[\boldsymbol{BVB}^{\top}\right] = (\boldsymbol{c}^{\top}\boldsymbol{c})^2 \det\left[\boldsymbol{AVA}^{\top}\right]/(\boldsymbol{c}^{\top}\boldsymbol{V}^{-1}\boldsymbol{c})$.*

Proof: Expanding the product of partitioned matrices, and using standard determinant formulae for partitioned matrices (Hocking, 1985, Appendix A.II.1.2),

$$\det\left[(\boldsymbol{c}\,\boldsymbol{A}^{\top})^{\top}\boldsymbol{V}(\boldsymbol{c}\,\boldsymbol{A}^{\top})\right] = \det\left[\begin{pmatrix} \boldsymbol{c}^{\top}\boldsymbol{Vc} & \boldsymbol{c}^{\top}\boldsymbol{VA}^{\top} \\ \boldsymbol{AVc} & \boldsymbol{AVA}^{\top} \end{pmatrix}\right]$$

$$= \det\left[\boldsymbol{AVA}^{\top}\right]\left|\boldsymbol{c}^{\top}\boldsymbol{Vc} - \boldsymbol{c}^{\top}\boldsymbol{VA}^{\top}(\boldsymbol{AVA}^{\top})^{-1}\boldsymbol{AVc}\right|.$$

Letting $\boldsymbol{U}$ be a matrix such that $\boldsymbol{UU}^{\top} = \boldsymbol{V}$, this last factor can be expressed as

$$\left|\boldsymbol{c}^{\top}\boldsymbol{U}(\boldsymbol{I} - (\boldsymbol{AU})^{\top}(\boldsymbol{AU}(\boldsymbol{AU})^{\top})^{-1}\boldsymbol{AU})\boldsymbol{U}^{\top}\boldsymbol{c}\right|.$$

Sequential Saddlepoint Approximation to
the Uniform Cumulant Generating Function
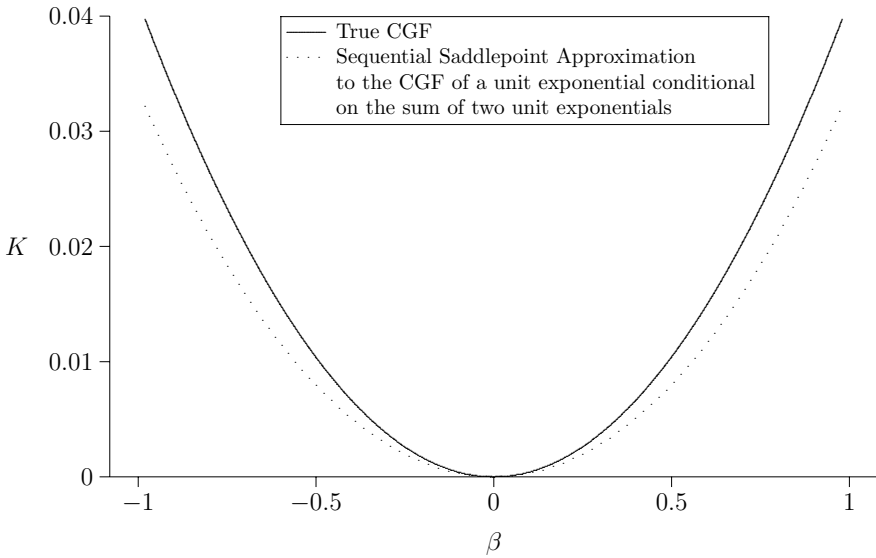
Fig. 15a.

Derivative of Sequential Saddlepoint Approximation
to the Uniform Cumulant Generating Function
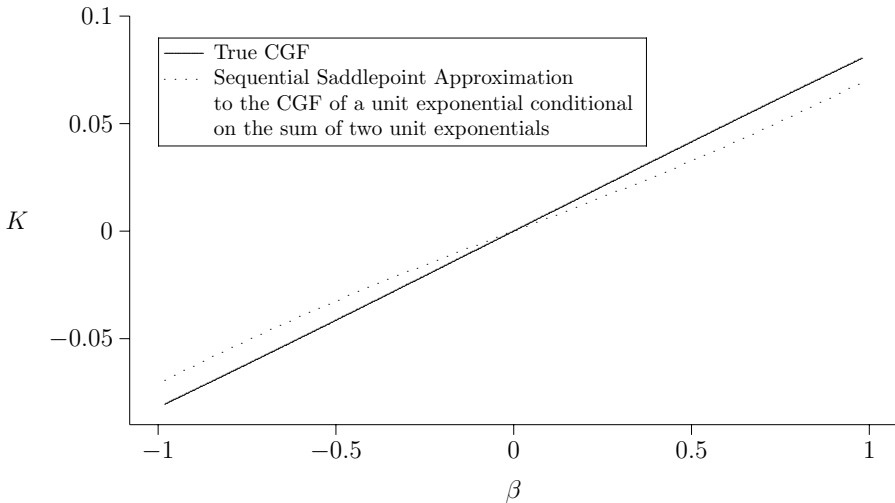
Fig. 15b.

The inner matrix is the projection matrix onto the space perpendicular to the space spanned by $\boldsymbol{AU}$; this space is spanned by $\boldsymbol{c}^\top \boldsymbol{U}^{-1\top}$, and the projection matrix onto this space can be expressed as

$$\boldsymbol{U}^{-1}\boldsymbol{c}(\boldsymbol{c}^\top \boldsymbol{V}^{-1}\boldsymbol{c})^{-1}\boldsymbol{c}^\top \boldsymbol{U}^{-1\top}.$$

Hence $\left|\boldsymbol{c}^\top \boldsymbol{V}\boldsymbol{c} - \boldsymbol{c}^\top \boldsymbol{V}\boldsymbol{A}^\top (\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^\top)^{-1}\boldsymbol{A}\boldsymbol{V}\boldsymbol{c}\right| = (\boldsymbol{c}^\top \boldsymbol{c})^2(\boldsymbol{c}^\top \boldsymbol{V}^{-1}\boldsymbol{c})^{-1}.$

Q.E.D

Then assuming the setup of Lemma 7.2.1, (184) can be expressed as

$$-\tilde{\boldsymbol{\theta}}^\top \boldsymbol{x} + \mathcal{H}_{\boldsymbol{X}}(\tilde{\boldsymbol{\theta}}) - \tfrac{1}{2}\log\left(\det\left[\mathcal{H}_{\boldsymbol{X}}''(\tilde{\boldsymbol{\theta}})\right]\right) - \tfrac{1}{2}\log(\boldsymbol{c}^\top \mathcal{H}_{\boldsymbol{X}}''(\tilde{\boldsymbol{\theta}})^{-1}\boldsymbol{c}).$$

Alternatively, McCullagh and Tibshirani (1990) attack the problem of inference in the presence of nuisance parameters by approximating marginal rather than conditional distributions, and derive an adjusted profile log likelihood.

Kolassa (2004) provides formulas that avoid numerical differentiation of (184), and also demonstrates that (184) might be used in conjunction with Theorem 6.9.1 to approximate conditional multivariate tail probabilities.

7.3. Double Saddlepoint Distribution Function Approximations

Skovgaard (1987) applies double saddlepoint techniques to the problem of approximating tail probabilities for conditional distributions, by approximating the saddlepoint inversion integral for the quantity $D(x^1|\boldsymbol{x}_{-1}) = \int_{x^1}^\infty f_n(y, x^2, \ldots, x^{\mathfrak{k}})\, dy$; the ratio

$$D(x^1|\boldsymbol{x}_{-1})/f_{n,X^2,\ldots,X^{\mathfrak{k}}}(x^2, \ldots, x^{\mathfrak{k}})$$

is by definition the conditional cumulative distribution function. Here $\boldsymbol{x}_{-1} = (x^2, \ldots, x^{\mathfrak{k}})$. An inversion integral representation for D is derived by replacing x^1 by a dummy integration variable y in (133), and then integrating with respect to y between x^1 and ∞, to yield, for $C_{n,\mathfrak{k}} = (n/[2\pi i])^{\mathfrak{k}} \exp(n[\mathcal{K}_{\boldsymbol{Y}}(\hat{\boldsymbol{\beta}}) - \hat{\boldsymbol{\beta}}^\top \boldsymbol{x}])$,

$$
\begin{aligned}
D(x^1|\boldsymbol{x}_{-1}) &= \left(\frac{n}{2\pi i}\right)^{\mathfrak{k}} \int_{x^1}^\infty \int_{-i\infty}^{+i\infty} \cdots \int_{-i\infty}^{+i\infty} \exp\left(n[\mathcal{K}_{\boldsymbol{Y}}(\boldsymbol{\beta}) - \beta_1(y - x^1) - \boldsymbol{\beta}^\top \boldsymbol{x}]\right)\, d\boldsymbol{\beta}\, dy \\
&= \left(\frac{n}{2\pi i}\right)^{\mathfrak{k}} \int_{-i\infty}^{+i\infty} \cdots \int_{-i\infty}^{+i\infty} \int_{x^1}^\infty \exp\left(n[\mathcal{K}_{\boldsymbol{Y}}(\boldsymbol{\beta}) - \beta_1(y - x^1) - \boldsymbol{\beta}^\top \boldsymbol{x}]\right)\, dy\, d\boldsymbol{\beta} \\
&= C_{n,\mathfrak{k}} \int_{-i\infty}^{+i\infty} \cdots \int_{-i\infty}^{+i\infty} \exp\left(n[\mathcal{K}_{\boldsymbol{Y}}(\boldsymbol{\beta}) - \mathcal{K}_{\boldsymbol{Y}}(\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta}^\top - \hat{\boldsymbol{\beta}})\boldsymbol{x}]\right)\, \frac{d\boldsymbol{\beta}}{n\beta_1}. \quad (185)
\end{aligned}
$$

The last equality in (185) follows only if the path of integration for the integral with respect to β_1 is deformed so that the real part of the path is bounded below by some positive number. After changing the variable of integration from $\boldsymbol{\beta}$ to $\boldsymbol{\omega}$ in the same way that generated (156), we obtain,

$$D(x^1|\boldsymbol{x}_{-1}) = C_{n,\mathfrak{k}} \int_{-i\infty}^{+i\infty} \cdots \int_{-i\infty}^{+i\infty} \exp\left(\frac{n}{2}(\boldsymbol{\omega} - \hat{\boldsymbol{\omega}})^\top (\boldsymbol{\omega} - \hat{\boldsymbol{\omega}})\right) \left|\frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{w}}\right| \frac{d\boldsymbol{\omega}}{n\beta_1(\boldsymbol{\omega})}. \quad (186)$$

Skovgaard expands D in terms of the multivariate saddlepoints $\hat{\boldsymbol{\beta}}$ for the full distribution of $\boldsymbol{X}$ and $\tilde{\boldsymbol{\beta}}$ for the distribution of the shorter random vector $(X^2,\ldots,X^{\mathfrak{k}})$, and in terms of derivatives of $\mathcal{K}_Y$ at these saddlepoints $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$. Analytic difficulties in expanding (186) are similar to those that arose in expanding (99), in that the integrand has a simple pole at $\beta_1 = 0$. As with the univariate cumulative distribution function approximation of Lugannani and Rice (1980), presented in §5.3, the singularity will be isolated into a term that can be integrated exactly, and the remaining integral will be approximated. As with the saddlepoint density, the expansion for D can be factored into two components, with one factor identical to the saddlepoint approximation to the density of $X^2,\ldots,X^{\mathfrak{k}}$. The second factor is then the conditional tail probability approximation of interest.

Let $g(\boldsymbol{\omega}) = \det\left[d\boldsymbol{\beta}/d\boldsymbol{\omega}\right]\omega_1/\beta_1(\boldsymbol{\omega})$. Expressing $g(\boldsymbol{\omega})$ as $g(0,\boldsymbol{w}_{-1}) + (g(\boldsymbol{\omega}) - g(0,\boldsymbol{w}_{-1}))$ separates out the singularity into a term that can be integrated exactly. Here $\boldsymbol{w}_{-1}$ denotes the vector $(\omega_2,\ldots,\omega_{\mathfrak{k}})$. These two terms will be integrated separately. That is, D is expressed as $D_1 + D_2$, where

$$
\begin{aligned}
D_1 &= C_{n,\mathfrak{k}} \int_{-i\infty}^{+i\infty} \cdots \int_{-i\infty}^{+i\infty} \exp(\frac{n}{2}(\boldsymbol{\omega}-\hat{\boldsymbol{\omega}})^\top(\boldsymbol{\omega}-\hat{\boldsymbol{\omega}}))\frac{g(0,\boldsymbol{w}_{-1})}{\omega_1}\,d\boldsymbol{\omega} \\
D_2 &= C_{n,\mathfrak{k}} \int_{-i\infty}^{+i\infty} \cdots \int_{-i\infty}^{+i\infty} \exp(\frac{n}{2}(\boldsymbol{\omega}-\hat{\boldsymbol{\omega}})^\top(\boldsymbol{\omega}-\hat{\boldsymbol{\omega}}))\frac{g(\boldsymbol{\omega})-g(0,\boldsymbol{w}_{-1})}{\omega_1}\,d\boldsymbol{\omega}.
\end{aligned}
\tag{187}
$$

The particular choice of $\boldsymbol{\beta}(\boldsymbol{\omega})$ of §6.7 implies that $\partial\boldsymbol{\beta}/\partial\boldsymbol{w}$ is triangular, and hence the Jacobian is the product of the derivatives; specifically, for any $j \in \{2,\ldots,\mathfrak{k}\}$,

$$
\left|\frac{\partial\boldsymbol{\beta}}{\partial\boldsymbol{w}}\right| = \det\left[\frac{\partial\beta_1,\ldots,\beta_{j-1}}{\partial\omega_1,\ldots,\omega_{j-1}}\right](\boldsymbol{\omega})\det\left[\frac{\partial\beta_j,\ldots,\beta_{\mathfrak{k}}}{\partial\omega_j,\ldots,\omega_{\mathfrak{k}}}\right](\omega_j,\ldots,\omega_{\mathfrak{k}}).
\tag{188}
$$

Since $\lim_{\omega_1\to 0}\omega_1/\beta_1 = \partial\beta_1/\partial\omega_1$, equation (188) implies that

$$
\lim_{\omega_1\to 0}\frac{\beta_1}{\omega_1}\left|\frac{\partial\boldsymbol{\beta}}{\partial\boldsymbol{w}}\right| = \frac{\partial\boldsymbol{\beta}_{-1}}{\partial\boldsymbol{w}_{-1}}(0,\boldsymbol{w}_{-1}).
$$

Since $g(0,\boldsymbol{w}_{-1})/\omega_1$ factors as the product of $1/\omega_1$ times a function of the other variables, the resulting integral D_1 may be factored the same way, and may be expressed as

$$
\begin{aligned}
D_1 = {}&\frac{\phi(n\hat{\omega}_1)}{i\sqrt{2\pi}} \int_{c-i\infty}^{c+i\infty} \exp\left(\frac{n}{2}(\omega_1-\hat{\omega}_1)^2\right)\frac{d\omega_1}{\omega_1}\left\{\left(\frac{n}{2\pi i}\right)^{\mathfrak{k}-1}\exp(n[\mathcal{K}_Y(\tilde{\boldsymbol{\beta}})-\tilde{\boldsymbol{\beta}}^\top\boldsymbol{x}])\times\right. \\
&\left.\int_{c-i\infty}^{c+i\infty}\cdots\int_{c-i\infty}^{c+i\infty}\exp\left(\frac{n}{2}\sum_{j=2}^{\mathfrak{k}}(\omega_j-\hat{\omega}_j)^2\right)\det\left[\frac{\partial\boldsymbol{\beta}_{-1}}{\partial\boldsymbol{w}_{-1}}(0,\boldsymbol{w}_{-1})\right]d\boldsymbol{w}_{-1}\right\}.
\end{aligned}
$$

By (156) the quantity in brackets is exactly the unconditional density of $(x^2,\ldots,x^{\mathfrak{k}})$; the first factor is $1 - \Phi(\sqrt{n}\hat{\omega}_1)$, by (108).

Then $(g(\boldsymbol{\omega}) - g(0, \boldsymbol{w}_{-1}))/\omega_1$ can be expressed as

$$\left(\frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{w}}(\boldsymbol{\omega})/\beta_1 - \frac{\partial \boldsymbol{\beta}_{-1}}{\partial \boldsymbol{w}_{-1}}(0, \boldsymbol{w}_{-1})/\omega_1 \right).$$

The integrals will not factor as those in the preceding paragraph did. However, the first-order Watson's Lemma approximation to this integral does factor into

$$D_2 = \left[\frac{n^{(\mathfrak{k}-1)/2} \exp(n[\mathcal{K}_{\boldsymbol{Y}}(\tilde{\boldsymbol{\beta}}) - \sum_{j=2}^{\mathfrak{k}} \tilde{\beta}_j x^j])}{(2\pi)^{(\mathfrak{k}-1)/2}} \det \left[\frac{\partial \beta_2, \dots, \beta_{\mathfrak{k}}}{\partial \omega_2, \dots, \omega_{\mathfrak{k}}} \right] (\hat{\omega}_2, \dots, \hat{\omega}_{\mathfrak{k}}) \right] \times$$

$$\left[\frac{\exp(n[\mathcal{K}_{\boldsymbol{Y}}(\hat{\boldsymbol{\beta}}) - \mathcal{K}_{\boldsymbol{Y}}(\tilde{\boldsymbol{\beta}}) - (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top \boldsymbol{x}])}{(2\pi)^{1/2}\sqrt{n}} \left(\det \left[\frac{\partial \beta_1}{\partial \omega_1} \right] (\hat{\boldsymbol{\omega}})/\hat{\beta}_1 - \frac{1}{\hat{\omega}_1} \right) \right] (1 + O(n^{-1})).$$

Again the first factor is the density approximation for $(X^2, \dots, X^{\mathfrak{k}})$ to $O(n^{-1})$. This implies that

$$1 - F_{x^1|\boldsymbol{x}_{-1}}(X^1|\boldsymbol{X}_{-1}) = 1 - \Phi(\sqrt{n}\hat{\omega}_1) + \frac{\phi(\sqrt{n}\hat{\omega}_1)}{\sqrt{n}} \left(\frac{\vartheta}{\hat{u}_1} - \frac{1}{\hat{\omega}_1} + O(n^{-1}) \right), \quad (189)$$

where

$$\hat{u}_1 = \hat{\beta}_1, \quad \hat{\omega}_1 = \operatorname{sgn}(\hat{\beta}_1)\sqrt{2[\hat{\boldsymbol{\beta}}^\top \boldsymbol{x} - \mathcal{K}_{\boldsymbol{Y}}(\hat{\boldsymbol{\beta}})] - 2[\tilde{\boldsymbol{\beta}}^\top \boldsymbol{x} - \mathcal{K}_{\boldsymbol{Y}}(\tilde{\boldsymbol{\beta}})]},$$

$$\vartheta = \det \left[\mathcal{K}''_{\boldsymbol{Y}_{-1}}(\tilde{\boldsymbol{\beta}}) \right]^{1/2} \det \left[\mathcal{K}''_{\boldsymbol{Y}}(\hat{\boldsymbol{\beta}}) \right]^{-1/2}, \quad (190)$$

$\mathcal{K}_{\boldsymbol{Y}_{-1}}$ is the $(\mathfrak{k} - 1) \times (\mathfrak{k} - 1)$ submatrix of the matrix of second derivatives of $\mathcal{K}$, corresponding to all components of $\boldsymbol{\beta}$ except the first, and Φ and ϕ are the normal distribution function and density respectively, since D is (189) times the saddlepoint density approximation for $(X^2, \dots, X^{\mathfrak{k}})$ to $O(n^{-1})$. This is the double saddlepoint approximation to the conditional distribution function.

Also of interest are inversion techniques random vectors $\boldsymbol{X}$ confined to a generalized lattice in the sense of Lemma 6.2.4. Let $D(x^1|\boldsymbol{x}_{-1}) = \sum_{x^1}^{\infty} p_n(y, x^2, \dots, x^{\mathfrak{k}})$ where p_n is the probability mass function for $\boldsymbol{X}$. Using the multivariate lattice probability inversion theorem (124), the counterpart of equation (185) is

$$D(x^1|\boldsymbol{x}_{-1}) = \sum_{y^1 = x^1}^{\infty} \left(\frac{n}{2\pi i} \right)^{\mathfrak{k}} \int_{-\pi/\Delta_1}^{\pi/\Delta_1} \cdots \int_{-\pi/\Delta_{\mathfrak{k}}}^{\pi/\Delta_{\mathfrak{k}}} \exp(n[\mathcal{K}_{\boldsymbol{Y}}(\boldsymbol{\beta}) - y^1 \beta_1 - \boldsymbol{\beta}_{-1}^\top \boldsymbol{x}_{-1}]) \, d\boldsymbol{\beta}$$

$$= \left(\frac{n}{2\pi i} \right)^{\mathfrak{k}} \int_{-\pi/\Delta_1}^{\pi/\Delta_1} \cdots \int_{-\pi/\Delta_{\mathfrak{k}}}^{\pi/\Delta_{\mathfrak{k}}} \frac{\exp(n[\mathcal{K}_{\boldsymbol{Y}}(\boldsymbol{\beta}) - (x^1 - \frac{1}{2}\Delta_1)\beta_1 - \boldsymbol{\beta}_{-1}^\top \boldsymbol{x}_{-1}])}{2 \sinh(\Delta_1 \beta_1/2)} \, d\boldsymbol{\beta}.$$

The last equality above follows only if the path of integration for the integral with respect to β_1 is deformed so that the real part of the path is bounded below by some positive number. The same analysis as above is performed with $g(\boldsymbol{\omega}) = \det [d\boldsymbol{\beta}/d\boldsymbol{\omega}] \, \omega_1/2 \sinh(\beta_1(\boldsymbol{\omega})/2)$. In the lattice case (189) holds, with

$$\hat{u}_1 = 2\sinh(\tfrac{1}{2}\Delta_1 \hat{\beta}_1)/\Delta_1, \quad \hat{\omega}_1 = \operatorname{sgn}(\hat{\beta}_1)\sqrt{2[\hat{\boldsymbol{\beta}}^\top \boldsymbol{x} - \mathcal{K}_{\boldsymbol{Y}}(\hat{\boldsymbol{\beta}})] - 2[\tilde{\boldsymbol{\beta}}^\top \boldsymbol{x} - \mathcal{K}_{\boldsymbol{Y}}(\tilde{\boldsymbol{\beta}})]},$$

$$\vartheta = \det \left[\mathcal{K}''_{Y_{-1}}(\tilde{\boldsymbol{\beta}}) \right]^{1/2} \det \left[\mathcal{K}''_{Y}(\hat{\boldsymbol{\beta}}) \right]^{-1/2}, \tag{191}$$

and x^1 is corrected for continuity when calculating $\hat{\boldsymbol{\beta}}$. That is, if possible values for X^1 are Δ_1 units apart, $\hat{\boldsymbol{\beta}}$ solves $\mathcal{K}'_{Y}(\hat{\boldsymbol{\beta}}) = \tilde{\boldsymbol{x}}$ where $\tilde{x}^j = x^j$ if $j \neq 1$ and $\tilde{x}^1 = x^1 - \frac{1}{2}\Delta_1$.

When $\hat{u}_1 = 0$, (189) is undefined; these can be replaced by their limiting values, which are given by Skovgaard (1987). When $\hat{u}_1$ is close to zero, (189) is numerically unstable. Yang and Kolassa (2002) provide a Taylor series replacement to (189) in this case.

These methods have been applied to distributions that are either wholly continuous or wholly lattice. Results in mixed cases are also not only possible but easily derived.

Reconsider the case in which $\boldsymbol{X}$ arises from the full exponential family (175). In this case the double saddlepoint distribution function approximation can be expressed in terms of the quantities in (175). Skovgaard's approximation (189) holds, with

$$\hat{\omega}_1 = \text{sgn}(\hat{\theta}_1 - \theta_1)\sqrt{2[\hat{\boldsymbol{\theta}}^\top \boldsymbol{x} - \mathcal{H}_{\boldsymbol{Y}}(\hat{\boldsymbol{\theta}}) - \tilde{\boldsymbol{\theta}}^\top \boldsymbol{x} + \mathcal{H}_{\boldsymbol{Y}}(\tilde{\boldsymbol{\theta}})]}, \;\; \hat{u}_1 = \hat{\theta}_1 - \theta_1,$$
$$\vartheta = \det \left[\mathcal{H}''_{\boldsymbol{Y}_{-1}}(\tilde{\boldsymbol{\theta}}) \right]^{1/2} \det \left[\mathcal{H}''_{\boldsymbol{Y}}(\hat{\boldsymbol{\theta}}) \right]^{-1/2} \tag{192}$$

and depends on the parameter vector $\boldsymbol{\theta}$ only through $\hat{\omega}_1$, the restricted maximum likelihood estimator $\tilde{\boldsymbol{\theta}}$, and the second derivative matrix of $\mathcal{H}_{\boldsymbol{Y}}$ at $\tilde{\boldsymbol{\theta}}$. The saddlepoint $\tilde{\boldsymbol{\theta}}$ is chosen to satisfy (183), and the saddlepoint $\hat{\boldsymbol{\theta}}$ is chosen to satisfy

$$\mathcal{H}^i_{\boldsymbol{Y}}(\hat{\boldsymbol{\theta}}) = X^i \text{ for all } i, \tag{193}$$

where the superscript denotes partial differentiation. Hence Skovgaard's approximation also depends only on the first component of the parameter vector and not on any of the other components. When X^1 has a lattice distribution approximation (189) holds with

$$\left. \begin{aligned} \hat{\omega}_1 &= \text{sgn}(\hat{\theta}_1 - \theta_1)\sqrt{2[\hat{\boldsymbol{\theta}}^\top \boldsymbol{x} - \mathcal{H}_{\boldsymbol{Y}}(\hat{\boldsymbol{\theta}}) - \tilde{\boldsymbol{\theta}}^\top \boldsymbol{x} + \mathcal{H}_{\boldsymbol{Y}}(\tilde{\boldsymbol{\theta}})]}, \\ \hat{u}_1 &= 2\sinh(\Delta_1(\hat{\theta}_1 - \theta_1)/2)/\Delta_1, \\ \vartheta &= \det \left[\mathcal{H}''_{\boldsymbol{Y}_{-1}}(\tilde{\boldsymbol{\theta}}) \right]^{1/2} \det \left[\mathcal{H}''_{\boldsymbol{Y}}(\hat{\boldsymbol{\theta}}) \right]^{-1/2} \end{aligned} \right\} \tag{194}$$

where $\hat{\boldsymbol{\theta}}$ is calculated from continuity-corrected data.

Alternate versions of (192) and (194) for conditional calculations involving random vectors that are linear transformations of more simple vectors are also desirable. Assume the definitions of Lemma 7.2.1. Suppose further that the distribution of T^1 conditional on all other components of $\boldsymbol{T}$ is desired. Then $\mathcal{H}_{\boldsymbol{T}}(\hat{\boldsymbol{\eta}}) - \hat{\boldsymbol{\eta}}^\top \boldsymbol{t} - \mathcal{H}_{\boldsymbol{T}}(\tilde{\boldsymbol{\eta}}) + \tilde{\boldsymbol{\eta}}^\top \boldsymbol{t} = \mathcal{H}_{\boldsymbol{X}}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}^\top \boldsymbol{x} - \mathcal{H}_{\boldsymbol{X}}(\tilde{\boldsymbol{\theta}}) + \tilde{\boldsymbol{\theta}}^\top \boldsymbol{x}$, and so the magnitude

of $\hat{\omega}_1$ is unchanged by substituting $\mathcal{H}_X$, $\hat{\boldsymbol{\theta}}$, and $\tilde{\boldsymbol{\theta}}$ for $\mathcal{H}_T$, $\hat{\boldsymbol{\eta}}$, and $\tilde{\boldsymbol{\eta}}$. Furthermore, $\hat{\eta}_1 = (\boldsymbol{c}^\top \boldsymbol{c})^{-1} \boldsymbol{c}^\top \hat{\boldsymbol{\theta}}$, and $\det [\mathcal{H}_T''(\hat{\boldsymbol{\eta}})] = \det \left[\mathcal{H}_X''(\hat{\boldsymbol{\theta}})\right] \det [\boldsymbol{B}]^2$. By Lemma 7.2.2, $\det \left[\mathcal{H}_{T_{-1}}''(\tilde{\boldsymbol{\eta}})\right] = \det \left[\boldsymbol{A}\mathcal{H}_X''(\tilde{\boldsymbol{\theta}})\boldsymbol{A}^\top\right] = \det [\mathcal{H}_X''(\tilde{\boldsymbol{\eta}})] (\boldsymbol{c}^\top \left[\mathcal{H}_X''(\tilde{\boldsymbol{\theta}})\right]^{-1} \boldsymbol{c}) \det [\boldsymbol{B}]^2 (\boldsymbol{c}^\top \boldsymbol{c})^{-2}$.
After accounting for n by replacing $\mathcal{H}_X$ by its expression in $\mathcal{H}_Y$, (189) holds, with $\boldsymbol{T}$ substituted for $\boldsymbol{X}$, and

$$\vartheta = \sqrt{\det \left[\mathcal{H}_Y''(\tilde{\boldsymbol{\theta}})\right] (\boldsymbol{c}^\top \left[\mathcal{H}_Y''(\tilde{\boldsymbol{\theta}})\right]^{-1} \boldsymbol{c})} \Big/ \left(\boldsymbol{c}^\top \boldsymbol{c} \sqrt{\det \left[\mathcal{H}_Y''(\hat{\boldsymbol{\theta}})\right]}\right),$$

$$u_1 = \boldsymbol{c}^\top \hat{\boldsymbol{\theta}} / \boldsymbol{c}^\top \boldsymbol{c} - \eta_1, \; \hat{\omega}_1 \text{ is as in (194)}, \tag{195}$$

when $\boldsymbol{Y}$ has a continuous distribution, and

$$\vartheta = \sqrt{\det \left[\mathcal{H}_Y''(\tilde{\boldsymbol{\theta}})\right] (\boldsymbol{c}^\top \left[\mathcal{H}_Y''(\tilde{\boldsymbol{\theta}})\right]^{-1} \boldsymbol{c})} \Big/ \left(\boldsymbol{c}^\top \boldsymbol{c} \sqrt{\det \left[\mathcal{H}_Y''(\hat{\boldsymbol{\theta}})\right]}\right),$$

$$u_1 = 2 \sinh \left(\Delta_1 [(\boldsymbol{c}^\top \hat{\boldsymbol{\theta}}) / (\boldsymbol{c}^\top \boldsymbol{c}) - \eta_1] / 2\right) / \Delta_1, \; \hat{\omega}_1 \text{ is as in (194)}, \tag{196}$$

when $\boldsymbol{Y}$ has a lattice distribution.

The following theorem summarizes the results of this this section:

Theorem 7.3.1: *Suppose that the random vector $\boldsymbol{X}$ is the mean of n independent and identically distributed random vectors, each of which has a cumulant generating function defined on an open neighborhood about $\mathbf{o}$. Take a potential value $\boldsymbol{x}$ for the mean of n independent and identically distributed copies of such vectors. If one of the following approximations is desired:*

a. *An expansion for the cumulative distribution function of X^1 conditional on $X^2, \ldots, X^{\natural}$ in terms of the cumulant generating function*

b. *An expansion for the cumulative distribution function of X^1 conditional on $X^2, \ldots, X^{\natural}$ in terms of the exponential family expression*

c. *An expansion for the cumulative distribution function of T^1 conditional on $T^2, \ldots, T^{\natural}$, where $\boldsymbol{T} = \boldsymbol{B}\boldsymbol{X}$ with $\boldsymbol{B}$ as above, in terms of the cumulant generating function of $\boldsymbol{X}$*

and one of the following conditions holds:

1. *A joint density exists for $\boldsymbol{X}$,*

2. *$\boldsymbol{X}$ is confined to a generalized lattice as in Lemma 6.2.4.*

Then (189), with components defined as indicated in Table 5, gives valid a asymptotic expansion.

Table 5: *Appropriate Double Saddlepoint Distribution Function Results*

Condition	Objective		
	a.	b.	c.
1.	(190)	(192)	(195)
2.	(191)	(194)	(196)

Proof: See the above discussion in this section.

<div align="right">Q.E.D</div>

When conditioning on variables taking values at the ends of their ranges, (179) may not have a finite solution. Additionally, when limits are taken in (180) as components of this vector converge to plus or minus infinity, the matrices whose determinants are taken become of less than full rank. Kolassa (1997) discusses modifications to (180) in this case. This will be revisited in §7.7.

Expressions (189) and (109) are formally equivalent. The adjusted r^* approximation of §5.6 may also be used here, following Lemma 5.6.1.

The calculations of this chapter are applied to conditional inference in one of the exercises; the application is due to Davison (1988).

Jing and Robinson (1994) calculate conditional tail probabilities for smooth transformations of random vectors satisfying certain non-lattice properties; they integrate (181) using Corollary 5.5.2.

7.4. Higher-Order Double Saddlepoint Approximations

Kolassa (1996b) derives a higher-order double saddlepoint distribution function approximation from (187). Recall that the first-order approximation to D_1 was exact; higher order terms will all come from D_2. Skovgaard (1987) suggests performing the integrals with respect to $\omega_2, \ldots, \omega_\ell$ to obtain the standard higher order saddlepoint approximation to the density of the conditioning variables, and then applying Temme's Theorem to the remaining univariate integral. The present approach of using the extended version of Temme's Theorem is analytically more straight-forward. Theorem 6.5.2 is applied once to D_2. The series expansion for D_2 will be expressed as the product of two factors. The first will be the series expansion for the density of $\boldsymbol{X}_{-1}$; the second will then be what is added to $1 - \Phi(\hat{\omega}_1)$ to produce the desired expansion for the conditional cumulative distribution function.

Use of Theorem 6.5.2 requires evaluation of the quantities A_s defined by (141). Since the integrand is free of n except in the exponent, the functions θ_s vanish unless $s = 0$. The inequality in the description of the set over which summation is taken in (141) becomes equality, yielding, after combining factors of i in Theorem 6.5.2 and in the leading factor of D_2, $D_2 = (n/2\pi)^{\ell/2} \exp(-\frac{n}{2} \sum_{j=1}^{\ell} \hat{\omega}_j^2) \sum_{s=0}^{\infty} A_s n^{-s}$, where

$$A_s = \sum_{v_1, \cdots, v_\ell \geq 0, \, \sum v_j = s} \frac{(-2)^{-s}}{\boldsymbol{v}!} \left[\frac{\partial^{2\boldsymbol{v}}}{\partial^{2\boldsymbol{v}}\boldsymbol{\omega}} \theta_0 \right] (\hat{\boldsymbol{\omega}}),$$

and $\theta_0(\boldsymbol{\omega}) = \det \left[\partial\boldsymbol{\beta}/\partial\boldsymbol{\omega} \right] / \beta^1 - \det \left[\partial\boldsymbol{\beta}_{-1}/\partial\boldsymbol{w}_{-1}(0, \boldsymbol{w}_{-1}) \right] / \omega_1$. Again employ the convention that indices appearing as superscripts and subscripts in a term are summed over. Let $h(\boldsymbol{\omega})$ be the function of $\boldsymbol{\omega}$ whose power series about $\hat{\boldsymbol{\omega}}$ is constructed from that of $\det \left[(\partial\boldsymbol{\beta}_{-1}/\partial\boldsymbol{w}_{-1})(0, \boldsymbol{w}_{-1}) \right]$, with all terms involving components of $\boldsymbol{\omega} - \hat{\boldsymbol{\omega}}$

to odd powers omitted. Let

$$g(\boldsymbol{\omega}) = \frac{\theta_0(\boldsymbol{\omega})}{h(\boldsymbol{\omega})} = \det\left[\frac{\partial\boldsymbol{\beta}}{\partial\boldsymbol{w}}(\boldsymbol{\omega})\right]/(h(\boldsymbol{\omega})\beta_1) - \det\left[\frac{\partial\boldsymbol{\beta}_{-1}}{\partial\boldsymbol{w}_{-1}}(0,\boldsymbol{w}_{-1})\right]/(h(\boldsymbol{\omega})\omega_1). \quad (197)$$

Then

$$\frac{(-2)^{-s}}{\boldsymbol{v}!}\left[\frac{\partial^{2\boldsymbol{v}}}{\partial^{2\boldsymbol{v}}\boldsymbol{\omega}}\theta_0\right](\hat{\boldsymbol{\omega}}) = \frac{1}{\boldsymbol{v}!}\sum_{0\leq u_j\leq 2v_j}\frac{2^{-s}(2\boldsymbol{v})!}{\boldsymbol{u}!(2\boldsymbol{v}-\boldsymbol{u})!}\left[\frac{\partial^{\boldsymbol{u}}}{\partial^{\boldsymbol{u}}\boldsymbol{\omega}}h\right](\hat{\boldsymbol{\omega}})\left[\frac{\partial^{2\boldsymbol{v}-\boldsymbol{u}}}{\partial^{2\boldsymbol{v}-\boldsymbol{u}}\boldsymbol{\omega}}g\right](\hat{\boldsymbol{\omega}}).$$

Recalling that h has only even-order terms,

$$A_s = \sum_{\substack{v_1,\cdots,v_{\ell}\geq 0 \\ \sum v_j = s}}\frac{(2\boldsymbol{v})!}{\boldsymbol{v}!}\sum_{0\leq u_j\leq v_j}\frac{(-2)^{-s}}{(2\boldsymbol{u})!(2\boldsymbol{v}-2\boldsymbol{u})!}\left[\frac{\partial^{2\boldsymbol{u}}}{\partial^{2\boldsymbol{u}}\boldsymbol{\omega}}h\right](\hat{\boldsymbol{\omega}})\left[\frac{\partial^{2\boldsymbol{v}-2\boldsymbol{u}}}{\partial^{2\boldsymbol{v}-2\boldsymbol{u}}\boldsymbol{\omega}}g\right](\hat{\boldsymbol{\omega}}). \quad (198)$$

The aim here, then, is to express the expansion $\sum_{s=0}^{j-1}A_s n^{-s}$ as

$$\sum_{s=0}^{j-1}B_s n^{-s} \times \sum_{s=0}^{j-1}C_s n^{-s} + O(n^{-j}),$$

where by Theorem 6.5.2, B_s are coefficients in the asymptotic expansion of the density for the conditioning variables:

$$B_s = \sum_{\substack{v_2,\cdots,v_{\ell}\geq 0 \\ \sum_{j=2}^{\ell} v_j = s}}\frac{(-2)^{-\sum_{j=2}^{\ell} v_j}}{v_2!\cdots v_{\ell}!}\left[\frac{\partial^{2v_2+\cdots+2v_{\ell}}}{\partial^{2v_2}\boldsymbol{\omega}_2\cdots\partial^{2v_{\ell}}\boldsymbol{\omega}_{\ell}}h\right](\hat{\boldsymbol{\omega}})$$

such that $f_{X^2,\cdots,X^{\ell}} = \sum_{s=0}^{j-1}B_s n^{-s} + O(n^{-j})$. Since

$$\left(1 - \Phi(\sqrt{n}\hat{\omega}_1) + \phi(\sqrt{n}\hat{\omega}_1)n^{-1/2}\sum_{s=0}^{j-1}C_s n^{-s}\right)\left(\phi(\sqrt{n}\hat{\omega}_{-1})n^{(1-\ell)/2}\sum_{s=0}^{j-1}B_s n^{-s}\right)$$

is an asymptotic expansion for D, $1 - \Phi(\sqrt{n}\hat{\omega}_1) + \phi(\sqrt{n}\hat{\omega}_1)n^{-1/2}\sum_{s=0}^{j-1}C_s n^{-s}$ is the required expansion for the conditional tail probability. Skovgaard (1987) provides such a decomposition for $j = 1$; that is, he factors the lead term $A_0 = \det\left[\frac{\partial\boldsymbol{\beta}_{-1}}{\partial\boldsymbol{w}_{-1}}(0,\boldsymbol{w}_{-1})\right] \times g(\hat{\boldsymbol{\omega}})$ into the lead term $B_0 = \det\left[\frac{\partial\boldsymbol{\beta}_{-1}}{\partial\boldsymbol{w}_{-1}}(0,\boldsymbol{w}_{-1})\right]$ in the asymptotic expansion for the density of the conditioning density times a factor $C_0 = g(\hat{\boldsymbol{\omega}}) = \sqrt{\det\left[\mathcal{K}''_{\boldsymbol{X}_{-1}}(0,\hat{\boldsymbol{\beta}}_{-1})\right]}/(\hat{\beta}_1\sqrt{\det\left[\mathcal{K}''_{\boldsymbol{X}}(\hat{\boldsymbol{\beta}})\right]}) - 1/\hat{\omega}_1$ which consequently contributes to the lead term in the asymptotic expansion for the conditional tail probability desired.

When $j = 2$, of concern are A_0 and A_1. When $s = 0$ or $s = 1$, $(2\boldsymbol{v})!$ in (198) is equal either to $(2\boldsymbol{u})!$ or $(2\boldsymbol{v}-2\boldsymbol{u})!$, or to both. The other quantity is 1. Also, $\boldsymbol{u}!$,

$\boldsymbol{v}$!, and $(\boldsymbol{v} - \boldsymbol{u})$! are all 1. Hence when $s = 0$ or $s = 1$,

$$
A_s = \sum_{\substack{v_1,\cdots,v_{\mathfrak{k}} \geq 0 \\ \sum v_j = s}} \sum_{0 \leq u_j \leq v_j} \frac{(-2)^{-s}}{(\boldsymbol{u})!(\boldsymbol{v} - \boldsymbol{u})!} \left[\frac{\partial^{2\boldsymbol{u}}}{\partial^{2\boldsymbol{u}}\boldsymbol{\omega}} h \right] (0, \boldsymbol{w}_{-1}) \left[\frac{\partial^{2\boldsymbol{v}-2\boldsymbol{u}}}{\partial^{2\boldsymbol{v}-2\boldsymbol{u}}\boldsymbol{\omega}} g \right] (\hat{\boldsymbol{\omega}}),
$$

indicating that $C_1 = \sum_{j=1}^{\mathfrak{k}} \frac{(-2)^{-1}}{1!} \frac{\partial^2}{\partial \omega_j^2} g(\hat{\boldsymbol{\omega}}) = -\frac{1}{2} \sum_{k=1}^{\mathfrak{k}} g^{kk}(\hat{\boldsymbol{\omega}})$, where g is given by (197), and superscripts on g indicate derivatives with respect to components of $\boldsymbol{\omega}$.

The next section shows that

$$
C_1 = -\frac{1}{\hat{z}} \left(\frac{1}{8}(\hat{\rho}_{13} - \tilde{\rho}_{13} - \hat{\rho}_4 + \tilde{\rho}_4) + \frac{1}{12}(\hat{\rho}_{23} - \tilde{\rho}_{23}) + \frac{1}{2}\frac{\hat{\kappa}_{1j}\hat{\kappa}^{ijk}\hat{\kappa}_{ik}}{\hat{\beta}_1} + \frac{\hat{\kappa}_{11}}{(\hat{\beta}_1)^2} \right) - \frac{1}{(\hat{\omega}_1)^3},
$$
$$(199)$$

where $\hat{z} = \hat{\beta}_1 \det\left[\frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{w}}(\hat{\boldsymbol{\omega}}) \right] / \det\left[\frac{\partial \boldsymbol{\beta}_{-1}}{\partial \boldsymbol{w}_{-1}}(0, \hat{\boldsymbol{\omega}}_{-1}) \right]$, and $\hat{\rho}_4$, $\hat{\rho}_{13}$, $\hat{\rho}_{23}$, $\tilde{\rho}_4$, $\tilde{\rho}_{13}$, and $\tilde{\rho}_{23}$ are as in §7.1. The second-order saddlepoint approximation to the conditional cumulative distribution function is now, to $O(n^{-5/2})$,

$$
1 - \tilde{F}(x^1|\boldsymbol{x}_{-1}) = 1 - \Phi(\sqrt{n}\hat{\omega}_1) + \frac{\phi(\sqrt{n}\hat{\omega}_1)}{\sqrt{n}} \left(\frac{1}{n\hat{\omega}_1^3} - \frac{1}{\hat{\omega}_1} + \frac{1}{\hat{z}} \left(1 + \frac{1}{n}[\frac{1}{8}(\hat{\rho}_4 - \tilde{\rho}_4) - \right. \right.
$$
$$
\frac{1}{8}(\hat{\rho}_{13} - \tilde{\rho}_{13}) - \frac{1}{12}(\hat{\rho}_{23} - \tilde{\rho}_{23}) - \frac{1}{2}\frac{\hat{\kappa}_{1k}\hat{\kappa}^{ijk}\hat{\kappa}_{ij}}{\hat{\beta}_1} - \frac{\hat{\kappa}^{11}}{(\hat{\beta}_1)^2}]) \right).
$$

In the lattice case,

$$
1 - \tilde{F}(x^1|\boldsymbol{x}_{-1}) = 1 - \Phi(\sqrt{n}\hat{\omega}_1) + \frac{\phi(\sqrt{n}\hat{\omega}_1)}{\sqrt{n}} \left(-\frac{1}{\hat{\omega}_1} + \frac{1}{\hat{z}} \left(1 + \frac{1}{n} [\right. \right.
$$
$$
\frac{1}{8}(\hat{\rho}_4 - \tilde{\rho}_4) - \frac{1}{8}(\hat{\rho}_{13} - \tilde{\rho}_{13}) - \frac{1}{12}(\hat{\rho}_{23} - \tilde{\rho}_{23})
$$
$$
- \frac{1}{4}\Delta\hat{\kappa}_{1k}\hat{\kappa}^{ijk}\hat{\kappa}_{ij}\coth(\frac{1}{2}\hat{\beta}_1\Delta_1) - \left(\frac{1}{4}\coth(\frac{1}{2}\hat{\beta}_1\Delta_1)^2 - \frac{1}{8}\right)\Delta^2\hat{\kappa}_{11}]\right) + \frac{1}{n\hat{\omega}_1^3} \right), \quad (200)
$$

where $\hat{z} = 2\sinh(\frac{1}{2}\hat{\beta}_1\Delta_1) \det\left[\frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{w}}(\hat{\boldsymbol{\omega}}) \right] / \det\left[\frac{\partial \boldsymbol{\beta}_{-1}}{\partial \boldsymbol{w}_{-1}}(0, \hat{\boldsymbol{\omega}}_{-1}) \right]$.

7.5. Derivatives of the Jacobian

In this section are presented calculations of derivatives of functions of $\partial\boldsymbol{\beta}/\partial\boldsymbol{\omega}$ necessary for for deriving via Temme's Theorem higher-order terms in the series expansions given in §6.4, §7.1, and §7.4. In this section again we employ the convention that an index appearing in a term both as a subscript and as a superscript is summed over.

Let $\hat{\beta}_r^{m\ldots o}$ be the derivative of component r of $\boldsymbol{\beta}$ with respect to $m,\ldots,o$, evaluated at $\hat{\boldsymbol{\omega}}$. Implicitly differentiating (153),

$$
\left.
\begin{aligned}
\delta_{m,n} &= \hat{\beta}_l^m \hat{\beta}_j^n \hat{\kappa}^{jl} + \hat{\kappa}^j \hat{\beta}_j^{mn} \\
0 &= \hat{\kappa}^{jli} \hat{\beta}_j^o \hat{\beta}_l^m \hat{\beta}_i^n + \hat{\kappa}^{ji} \hat{\beta}_j^{mo} \hat{\beta}_i^n [3] + \hat{\kappa}^j \hat{\beta}_j^{mno} \\
0 &= \hat{\kappa}^{ijkl} \hat{\beta}_j^o \hat{\beta}_l^m \hat{\beta}_i^n \hat{\beta}_k^p + \hat{\kappa}^{ijl} \hat{\beta}_j^o \hat{\beta}_l^m \hat{\beta}_i^{np} [6] + \hat{\kappa}_{ij} \hat{\beta}_j^{mo} \hat{\beta}_i^{np} [3] + \\
&\quad \hat{\kappa}_{ij} \hat{\beta}_j^{mop} \hat{\beta}_i^n [4] + \hat{\kappa}^j \hat{\beta}_j^{mnop}.
\end{aligned}
\right\}
\tag{201}
$$

Evaluating these at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, note that $-\tfrac{1}{3}\hat{\beta}_l^m \hat{\beta}_j^n \hat{\kappa}^{jlp} = \hat{\kappa}^{pj}\hat{\beta}_j^{mn}$ satisfies the second equality of (201); substituting into the third equality of (201),

$$
\left[-\hat{\kappa}^{ijkl} + \tfrac{5}{9}\hat{\kappa}^{gij}\hat{\kappa}_{gh}\hat{\kappa}^{hkl}[3]\right] \hat{\beta}_j^o \hat{\beta}_l^m \hat{\beta}_i^n \hat{\beta}_k^p = \hat{\kappa}^{ij}\hat{\beta}_j^{mop}\hat{\beta}_i^n[4] + \hat{\kappa}^j \hat{\beta}_j^{mnop}.
$$

Then $\tfrac{1}{4}\left[\tfrac{5}{9}\hat{\kappa}^{gij}\hat{\kappa}_{gh}\hat{\kappa}^{hkl}[3] - \hat{\kappa}^{ijkl}\right]\hat{\beta}_j^o \hat{\beta}_l^m \hat{\beta}_k^n = \hat{\kappa}^{ij}\hat{\beta}_j^{mon}$ satisfies this equation. Hence

$$
\hat{\beta}_r^{mn} = -\tfrac{1}{3}\hat{\kappa}_{ir}\hat{\kappa}^{ijk}\hat{\beta}_j^m \hat{\beta}_k^n, \quad \hat{\beta}_r^{mno} = \tfrac{1}{4}\hat{\kappa}_{ir}\left[\tfrac{5}{9}\hat{\kappa}^{gij}\hat{\kappa}_{gh}\hat{\kappa}^{hkl}[3] - \hat{\kappa}^{ijkl}\right]\hat{\beta}_j^o \hat{\beta}_l^m \hat{\beta}_k^n.
$$

Let α_q^r represent the matrix inverse of $\hat{\beta}_r^q$. By (201) $\alpha_q^r \hat{\kappa}_{ri} = \hat{\beta}_i^r \delta_{qr}$ and $\hat{\kappa}_{ij} = \hat{\beta}_i^g \delta_{gh}\hat{\beta}_i^h$. Also

$$
\alpha_q^r \hat{\beta}_r^{mn} = -\tfrac{1}{3}\alpha_q^r \hat{\kappa}_{ri}\hat{\kappa}^{ijk}\hat{\beta}_j^m \hat{\beta}_k^n = -\tfrac{1}{3}\delta_{rq}\hat{\kappa}^{ijk}\hat{\beta}_i^r \hat{\beta}_j^m \hat{\beta}_k^n
$$
$$
\alpha_q^r \hat{\beta}_r^{mt}\alpha_t^s \hat{\beta}_s^{no} = \tfrac{1}{9}\delta_{rq}\hat{\kappa}^{gij}\hat{\kappa}_{gh}\hat{\kappa}^{hkl}\hat{\beta}_j^o \hat{\beta}_j^m \hat{\beta}_k^n \hat{\beta}_i^r
$$
$$
\alpha_q^r \hat{\beta}_r^{mno} = \tfrac{1}{4}\left[\tfrac{5}{9}\hat{\kappa}^{gij}\hat{\kappa}_{gh}\hat{\kappa}^{hkl}[3] - \hat{\kappa}^{ijkl}\right]\hat{\kappa}_{ij}\hat{\beta}_l^m \hat{\beta}_k^n.
$$
$$
\alpha_q^r \hat{\beta}_r^{mq}\alpha_s^p \hat{\beta}_p^{ns} = \tfrac{1}{9}\hat{\kappa}^{ijg}\hat{\kappa}_{ig}\hat{\beta}_j^m \hat{\kappa}^{klh}\hat{\beta}_l^m \hat{\kappa}^{kh}.
$$

Standard matrix determinant differentiation formulae (Hocking, 1985, Appendix A.II.1.2) show that

$$
\frac{\partial\boldsymbol{\beta}}{\partial\boldsymbol{w}}(\boldsymbol{\omega}) = \frac{\partial\boldsymbol{\beta}}{\partial\boldsymbol{w}}(\hat{\boldsymbol{\omega}})\Big(1 + \alpha_q^r \hat{\beta}_r^{qm}(\omega_m - \hat{\omega}_m) +
$$
$$
\tfrac{1}{2}\left(\alpha_q^r \hat{\beta}_r^{qm}\alpha_p^s \hat{\beta}_s^{pn} + \alpha_q^r \hat{\beta}_r^{qmn} - \alpha_q^r \hat{\beta}_r^{sn}\alpha_s^t \hat{\beta}_t^{qm}\right)(\omega_m - \hat{\omega}_m)(\omega_n - \hat{\omega}_n)\Big) + O(\|\boldsymbol{\omega} - \hat{\boldsymbol{\omega}}\|^3)
$$
$$
= \frac{\partial\boldsymbol{\beta}}{\partial\boldsymbol{w}}(\hat{\boldsymbol{\omega}})\Big(1 - \tfrac{1}{3}\hat{\kappa}^{ijk}\hat{\kappa}_{ik}\hat{\beta}_j^m(\omega_m - \hat{\omega}_m) + \tfrac{1}{2}\Big[\tfrac{1}{4}\hat{\kappa}^{gij}\hat{\kappa}_{gh}\hat{\kappa}^{hkl}\hat{\kappa}_{ij} + \tfrac{1}{6}\hat{\kappa}^{gil}\hat{\kappa}_{gh}\hat{\kappa}^{hjk}\hat{\kappa}_{ij}
$$
$$
- \tfrac{1}{4}\hat{\kappa}^{ijkl}\hat{\kappa}_{ij}\Big]\hat{\beta}_k^m \hat{\beta}_l^n(\omega_m - \hat{\omega}_m)(\omega_n - \hat{\omega}_n)\Big) + O(\|\boldsymbol{\omega} - \hat{\boldsymbol{\omega}}\|^3).
\tag{202}
$$

Similarly,

$$\frac{1}{h(\boldsymbol{\omega})} = \det\left[\frac{\partial \boldsymbol{\beta}_{-1}}{\partial \boldsymbol{w}_{-1}}(\hat{\boldsymbol{\omega}})\right]^{-1}\left(1 + \tfrac{1}{2}\left[-\tfrac{1}{4}\tilde{\kappa}^{gij}\tilde{\kappa}_{gh}\tilde{\kappa}^{hkl}\tilde{\kappa}_{ij} - \tfrac{1}{6}\tilde{\kappa}^{gil}\tilde{\kappa}_{gh}\tilde{\kappa}^{hjk}\tilde{\kappa}_{ij}\right.\right.$$

$$\left.\left. + \tfrac{1}{4}\tilde{\kappa}^{ijkl}\tilde{\kappa}_{ij}\right]\tilde{\beta}_k^m\tilde{\beta}_l^n(\omega_m - \hat{\omega}_m)(\omega_n - \hat{\omega}_n)\right) + O(\|\boldsymbol{\omega} - \hat{\boldsymbol{\omega}}\|^3). \tag{203}$$

Furthermore,

$$\frac{\hat{\beta}_1}{\beta_1} = 1 - \frac{\hat{\beta}_1^m(\omega_m - \hat{\omega}_m)}{\hat{\beta}_1} + \left[\frac{\hat{\beta}_1^m\hat{\beta}_1^n}{(\hat{\beta}_1)^2} - \frac{\tfrac{1}{2}\hat{\beta}_1^{mn}}{\hat{\beta}_1}\right](\omega_m - \hat{\omega}_m)(\omega_n - \hat{\omega}_n) + O(\|\boldsymbol{\omega} - \hat{\boldsymbol{\omega}}\|^3)$$

$$= 1 - \frac{\hat{\beta}_1^m}{\hat{\beta}_1}(\omega_m - \hat{\omega}_m) + \left[\frac{\hat{\beta}_1^m\hat{\beta}_1^n}{(\hat{\beta}_1)^2} + \frac{\hat{\kappa}_{1i}\hat{\kappa}^{ijk}\hat{\beta}_j^m\hat{\beta}_k^n}{6\hat{\beta}_1}\right](\omega_m - \hat{\omega}_m)(\omega_n - \hat{\omega}_n)$$

$$+ O(\|\boldsymbol{\omega} - \hat{\boldsymbol{\omega}}\|^3). \tag{204}$$

Also needed for the expansion in the lattice case is:

$$\frac{1}{\sinh(\tfrac{1}{2}\beta^1\Delta)} = \frac{1}{\sinh(\tfrac{1}{2}\hat{\beta}^1\Delta)}\left(1 - \tfrac{1}{2}\hat{\beta}_1^m\Delta\coth(\tfrac{1}{2}\hat{\beta}^1\Delta)(\omega_m - \hat{\omega}_m) + \right.$$

$$\left(\hat{\beta}_1^m\hat{\beta}_1^n\Delta^2\left(\tfrac{1}{4}\coth(\tfrac{1}{2}\hat{\beta}^1\Delta)^2 - \tfrac{1}{8}\right)\right.$$

$$\left.\left. - \tfrac{1}{4}\hat{\beta}_1^{mn}\Delta\coth(\tfrac{1}{2}\hat{\beta}^1\Delta)\right)(\omega_m - \hat{\omega}_m)(\omega_n - \hat{\omega}_n)\right) + O(\|\boldsymbol{\omega} - \hat{\boldsymbol{\omega}}\|)^3. \tag{205}$$

Then (199) follows by multiplying (202), (203), and (204), and summing over all coefficients of squared components of $\boldsymbol{\omega} - \hat{\boldsymbol{\omega}}$. Also, (200) follows by multiplying (202), (203), and (205), and summing the appropriate coefficients. Furthermore, the unconditional multivariate saddlepoint density formula (133) follows from summing over the same coefficients in (202), and the conditional double saddlepoint density approximation (180) follows by multiplying (203) and (202) and summing over the appropriate coefficients.

7.6. An Example

Skovgaard illustrates (194) in the context of the hypergeometric distribution, which has applications to testing independence in 2×2 contingency tables. Suppose that X^i are independent random variables Poisson variables with intensities $\boldsymbol{\eta}$, for $i = 1, \ldots, 4$, and arranged into the following table:

X^1	X^2	T^2
X^3	X^4	
T^3		T^4

.

Of interest is whether $\eta_1\eta_4 = \eta_2\eta_3$; that is, whether the data come from a hypergeometric distribution. A test will be performed conditional on row and column marginals. This test will be performed by determining whether X^1, or equivalently, $\boldsymbol{c}^\top \boldsymbol{X}$, is too high or low to be consistent with the null hypotheses. Here $\boldsymbol{c} = (1, -1, 1, -1)^\top$. Approximation (196) will be used to approximate tail probabilities, with $n = 1$. Here $\boldsymbol{T} = \boldsymbol{B}\boldsymbol{X}$, where

$$
\boldsymbol{B} = \begin{pmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}.
$$

Then $\hat{\boldsymbol{\eta}}$, defined by (132), is $\log(\tilde{\boldsymbol{X}})$ where $\tilde{\boldsymbol{X}}$ is $\boldsymbol{X}$ corrected for continuity. Let $\boldsymbol{X}^*$ be the vector of fitted values, that is, the vector of cell values given by the product of the appropriate row and column mean, divided by the grand mean. Then $\boldsymbol{c}^\top \log(\boldsymbol{X}^*) = 0$ and $\boldsymbol{A}\boldsymbol{X}^* = \boldsymbol{A}\boldsymbol{X}$, and $\tilde{\boldsymbol{\eta}} = \log(\boldsymbol{X}^*)$. The cumulant generating function second derivative matrices are diagonal with elements $\tilde{\boldsymbol{X}}$ and $\boldsymbol{X}^*$, and hence their determinants and inverses are simple to calculate. The lattice spacing for $T^1 = \boldsymbol{c}^\top \boldsymbol{X}$ and $\boldsymbol{c}^\top \boldsymbol{c}$ are both 4, and hence cancel. Skovgaard (1987) then simplifies the double saddlepoint approximation to:

$$
1 - \Phi(\hat{\omega}) + \phi(\hat{\omega}) \left\{ \left(\frac{x_{1.}x_{.1}x_{2.}x_{.2}}{x_{..}\tilde{x}_{11}\tilde{x}_{12}\tilde{x}_{21}\tilde{x}_{22}} \right)^{1/2} (2\sinh(\mu))^{-1} - \hat{\omega}^{-1} \right\}
$$

where, in an abuse of notation, $\boldsymbol{x}$ is now the matrix of cell counts, and $\tilde{\boldsymbol{x}}$ is the continuity-corrected vector of cell counts. An index replaced with a . denotes summation. Also, $\mu = \log(\tilde{x}_{11}\tilde{x}_{22}/(\tilde{x}_{21}\tilde{x}_{12}))$ is the log of the observed continuity-corrected odds ratio, and

$$
\hat{\omega} = \mathrm{sgn}(\mu) \left\{ 2 \left[\sum_{\substack{l=1,2 \\ j=1,2}} \tilde{x}_{lj} \log(\tilde{x}_{lj}) - \sum_{l=1,2} x_{l.} \log(x_{l.}) - \sum_{j=1,2} x_{.j} \log(x_{.j}) + x_{..} \log(x_{..}) \right] \right\}^{\frac{1}{2}}.
$$

Alternatively one might proceed directly from (191), rather than taking advantage of the simplification offered by the methods of Lemma 7.2.1. If $\boldsymbol{\theta}$ are the canonical parameters associated with $\boldsymbol{T}$, then

$$
\eta_1 = \theta_1 + \theta_2 + \theta_2 + \theta_4, \eta_2 = \theta_2 + \theta_4, \eta_3 = \theta_3 + \theta_4, \eta_4 = \theta_4.
$$

The joint probability mass function of $\boldsymbol{T}$ is given by

$$
\frac{\exp(-\exp(\theta_1 + \theta_2 + \theta_3 + \theta_4) - \exp(\theta_2 + \theta_4) - \exp(\theta_3 + \theta_4) - \exp(\theta_4) + \boldsymbol{\theta}^\top \boldsymbol{t})}{t^1!(t^2 - t^1)!(t^3 - t^1)!(t^4 - t^2 - t^3 + t^1)!}.
$$

In this case $\mathcal{H}(\boldsymbol{\theta}) = \exp(\theta_1 + \theta_2 + \theta_3 + \theta_4) + \exp(\theta_2 + \theta_4) + \exp(\theta_3 + \theta_4) + \exp(\theta_4)$.

When testing whether $\theta_1 = 0$, $\tilde{\boldsymbol{\theta}}$ satisfies

$$\exp(\tilde{\theta}_2 + \tilde{\theta}_3 + \tilde{\theta}_4) + \exp(\tilde{\theta}_2 + \tilde{\theta}_4) + \exp(\tilde{\theta}_3 + \tilde{\theta}_4) + \exp(\tilde{\theta}_4) = t^4$$
$$\exp(\tilde{\theta}_2 + \tilde{\theta}_3 + \tilde{\theta}_4) + \exp(\tilde{\theta}_3 + \tilde{\theta}_4) = t^3$$
$$\exp(\tilde{\theta}_2 + \tilde{\theta}_3 + \tilde{\theta}_4) + \exp(\tilde{\theta}_2 + \tilde{\theta}_4) = t^2.$$

These equations have the solutions

$$\tilde{\theta}_2 = \log\left(\frac{t^2}{t^4 - t^2}\right), \quad \tilde{\theta}_3 = \log\left(\frac{t^3}{t^4 - t^3}\right) \quad \tilde{\theta}_4 = \log\left(\left(t^4 - t^2\right)\left(t^4 - t^3\right)/t^4\right).$$

From knowledge of $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ approximation (191) may easily be calculated.

Exact probabilities for atoms for the conditional distribution are given by

$$\mathrm{P}\left[T^1 = t^1 | T^2 = t^2, T^3 = t^3, T^4 = t^4\right] = \binom{t^2}{t^1}\binom{t^4 - t^2}{t^3 - t^1} / \binom{t^4}{t^3}.$$

Exact tail probabilities are quite difficult to compute when t^4, t^2, t^3, t^1, and $\min(t^2, t^3) - t^1$ are large. The double saddlepoint distribution function approximation is very accurate and is a simple function of the x^{ij} (Fig. 16).

Approximations to Hypergeometric Tail Probs. with Marginals 10 10 10 10
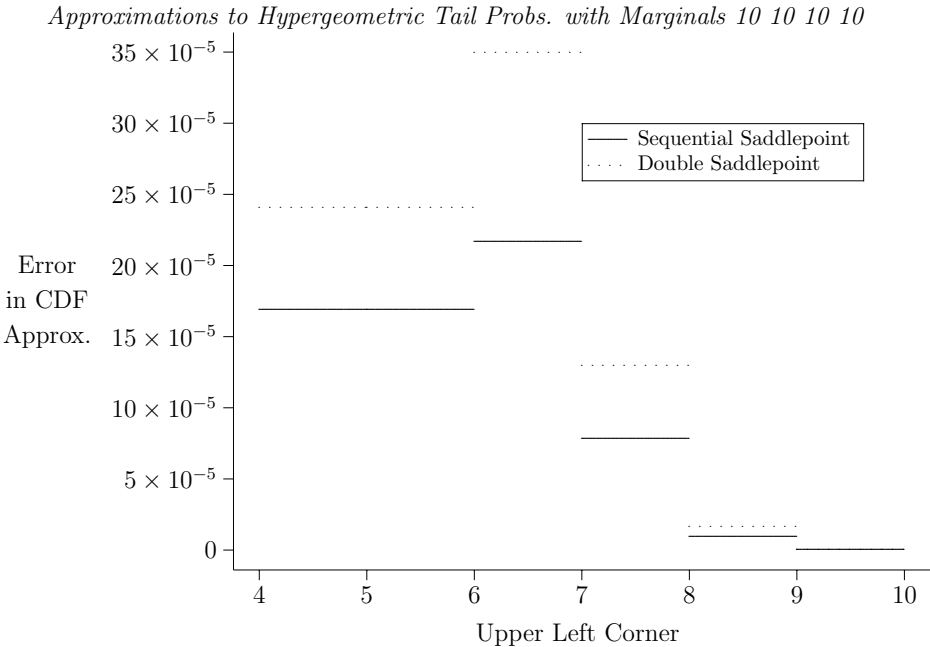
Fig. 16.

Skovgaard (1987) shows that the double saddlepoint approximation performs very well for large tables that are not extreme, but not so well for extreme tables.

The sequential saddlepoint approximation will also be applied to these data. As above, let $\boldsymbol{\theta}$ be the canonical parameter vector for $\boldsymbol{T}$, and let $\boldsymbol{\theta} = \boldsymbol{B}^\top \boldsymbol{\eta}$. Then if $\tilde{\boldsymbol{X}}$ is once again the vector of fitted table entries, this time under the null hypothesis that the log odds ratio is $4\theta_1$ rather than 0, then

$$\tilde{X}^1 = \tfrac{1}{2}\left(t^2 + t^3 - t^4(1 - \exp(4\theta_1))^{-1} + \right.$$

$$\left. \frac{\sqrt{(t^{2^2}+t^{3^2})(1 - \exp(4\theta_1))^2 + 2t^2t^3(1-\exp(4\theta_1)^2) + 2t^4(t^2+t^3)(\exp(4\theta_1)-1) + t^{4^2}}}{1-\exp(4\theta_1)} \right).$$

Other elements of $\tilde{\boldsymbol{X}}$ can be calculated from $\tilde{X}^1$, T^2, T^3, and T^4. The sequential saddlepoint approximation performs even better than the double saddlepoint approximation, but is not so easy to compute.

7.7. Corrections when the Saddlepoints Do Not Exist

Most of the material in this section was originally presented by Kolassa (1997). Consider the canonical generalized linear model (175) and (177). Suppose that the support of Y_j is contained in a finite interval, without loss of generality of form $[0, m_j]$. For example, when the Y_j are binomial variables, and (175) and (177) constitute a logistic regression model, then m_j is the number of Bernoulli trials comprising observation j. When Y_j is supported on a finite interval, the domain of $\mathcal{H}_{Y_j}$ is $\mathfrak{R}$. Suppose further that $\mathrm{P}[Y_j = 0]$ and $\mathrm{P}[Y_j = m_j]$ are positive for some, and hence all, values of v. Then for each j,

$$\lim_{v\to\infty} \mathcal{H}'_{Y_j}(v) = m_j \text{ and } \lim_{v\to-\infty} \mathcal{H}'_{Y_j}(v) = 0. \tag{206}$$

Finally, suppose that $\boldsymbol{Z}$ is of full rank.

For some $\boldsymbol{Y}$ one or more components of $\hat{\boldsymbol{\theta}}$ solving (193) may be infinite. This happens, for example, when all of the entries of $\boldsymbol{Z}$ are positive, and when $\boldsymbol{Y} = \mathbf{o}$. Barndorff–Nielsen (1978) gives necessary and sufficient conditions for finiteness of the solution to (193) (p. 151). Jacobsen (1989) gives more precise conditions specifically tailored to regression models.

This section considers circumstances under which finite solutions to (193) do not exist. Infinite solutions are expressed as rays in $\mathfrak{R}^{\mathfrak{k}}$, and contemporary methods for describing these rays are reviewed. Kolassa (1997) proposed modifications to the methods described in §7.3 for implementing the double saddlepoint conditional distribution function approximation of Skovgaard (1987), facilitating applications to conditional inference, both in the unidimensional case considered by Davison (1988) and in the multivariate case considered by Kolassa and Tanner (1994). In the latter application the dependence of the results only on the canonical sufficient statistics $\boldsymbol{T} = \boldsymbol{Z}^\top \boldsymbol{Y}$ is critical.

Albert and Anderson (1984) and Jacobsen (1989) both provide conditions for determining whether all parameter estimates are finite. They provide results that simplify to the following in the case of logistic regression:

Lemma 7.7.1:   *Choose $\boldsymbol{\lambda} \in \mathfrak{R}^{\mathfrak{k}}$. Let $\mathfrak{U} = \{\boldsymbol{u} \in \mathfrak{R}^{\mathfrak{k}} | \lim_{a \to \infty} \boldsymbol{t}^{\top}(\boldsymbol{\lambda} + \boldsymbol{u}a) - \mathcal{H}_T(\boldsymbol{\lambda} + \boldsymbol{u}a) > -\infty\}$. Then $\mathfrak{U}$ is independent of $\boldsymbol{\lambda}$, and is the set of vectors $\boldsymbol{u}$ satisfying*

$$\begin{cases} \boldsymbol{z}_j \boldsymbol{u} \geq 0 & \text{whenever } y_j = m_j, \\ \boldsymbol{z}_j \boldsymbol{u} \leq 0 & \text{whenever } y_j = 0, \\ \boldsymbol{z}_j \boldsymbol{u} = 0 & \text{whenever } y_j \in (0, m_j). \end{cases} \tag{207}$$

Proof: Note that $\boldsymbol{t}^{\top}(\boldsymbol{\lambda} + \boldsymbol{u}a) - \mathcal{H}_T(\boldsymbol{\lambda} + \boldsymbol{u}a) = \sum_j y_j \boldsymbol{z}_j^{\top}(\boldsymbol{\lambda} + \boldsymbol{u}a) - \mathcal{H}_{Y_j}(\boldsymbol{z}_j^{\top}(\boldsymbol{\lambda} + \boldsymbol{u}a))$. By (206), each of these summands is bounded as $a \to \infty$, and by the convexity of $\mathcal{H}_{Y_j}$, $\boldsymbol{u} \in \mathfrak{U}$ if and only if each of the summands is bounded below as $a \to \infty$. Furthermore, summand j is bounded below as $a \to \infty$ if and only if $\boldsymbol{z}_j^{\top} \boldsymbol{u}(y_j - \mathcal{H}_{Y_j}'(\boldsymbol{z}_j^{\top}(\boldsymbol{\lambda} + \boldsymbol{u}a))) \to 0$ as $a \to \infty$, by the convexity of $y_j \boldsymbol{z}_j^{\top}(\boldsymbol{\lambda} + \boldsymbol{u}a) - \mathcal{H}_{Y_j}(\boldsymbol{z}_j^{\top}(\boldsymbol{\lambda} + \boldsymbol{u}a))$. Hence $\boldsymbol{u} \in \mathfrak{U}$ if and only if $\lim_{a \to \infty} \boldsymbol{z}_j^{\top} \boldsymbol{u}(y_j - \mathcal{H}_{Y_j}'(\boldsymbol{z}_j^{\top}(\boldsymbol{\lambda} + \boldsymbol{u}a))) = 0$ for all j. This happens if and only if (207) holds.

Q.E.D

Albert and Anderson (1984) note that describing the $\boldsymbol{u}$ satisfying the above conditions may be achieved using linear programming methods. This method requires knowledge of $\boldsymbol{Y}$ rather than just $\boldsymbol{T}$. The following new result will serve as the basis for characterizing $\mathfrak{U}$ in terms of $\boldsymbol{T}$, by creating new variables $\boldsymbol{r}$ and $\boldsymbol{s}$ representing the contribution of $\boldsymbol{u}$ to the fitted probabilities for the individual observations; $\boldsymbol{r}$ and $\boldsymbol{s}$ are related to $\boldsymbol{Z}$ by the relation $\boldsymbol{s} - \boldsymbol{r} = \boldsymbol{Z}\boldsymbol{u}$. In the following sections we will determine which observations correspond to extreme fitted probabilities by maximizing the total number of positive entries in both $\boldsymbol{r}$ and $\boldsymbol{s}$ subject to constraints outlined in the next theorem.

Theorem 7.7.2:   *Suppose that $\boldsymbol{Z}$ is of full rank. Let $\boldsymbol{Q}$ be the matrix of maximal rank such that $\boldsymbol{Q}\boldsymbol{Z}\boldsymbol{o}$; note that $\boldsymbol{Q} = \boldsymbol{o}$ unless $n > \mathfrak{k}$. Then $\mathfrak{U} - \boldsymbol{o}$ is exactly the set of vectors*

$$\boldsymbol{u} = (\boldsymbol{Z}^{\top}\boldsymbol{Z})^{-1}\boldsymbol{Z}^{\top}(\boldsymbol{s} - \boldsymbol{r}), \tag{208}$$

such that $\boldsymbol{r}$ and $\boldsymbol{s}$ are column vectors of non-negative numbers such that $\boldsymbol{s} \neq \boldsymbol{r}$ and such that the following two conditions both hold:

$$(\boldsymbol{t}^{\top}(\boldsymbol{Z}^{\top}\boldsymbol{Z})^{-1}\boldsymbol{Z}^{\top} - \boldsymbol{m}^{\top})\boldsymbol{s} - \boldsymbol{t}^{\top}(\boldsymbol{Z}^{\top}\boldsymbol{Z})^{-1}\boldsymbol{Z}^{\top}\boldsymbol{r} = 0, \tag{209}$$

$$\boldsymbol{Q}(\boldsymbol{s} - \boldsymbol{r}) = \boldsymbol{o}, \tag{210}$$

where $\boldsymbol{T} = \boldsymbol{Z}^{\top}\boldsymbol{Y}$ are the sufficient statistics associated with $\boldsymbol{\beta}$, and $\boldsymbol{t}$ is the observed value of this vector. Furthermore, suppose that $\boldsymbol{r}$ and $\boldsymbol{s}$ satisfy (209) and (210), and that $\boldsymbol{t} = \boldsymbol{Z}\boldsymbol{y}$. If $r_j > 0$, then $y_j = 0$, and if $s_j > 0$, then $y_j = m_j$.

Proof: Let $I(x)$ be the function returning 1 if its argument is true and 0 otherwise. Let $f_j(\boldsymbol{u}) = \boldsymbol{z}_j \boldsymbol{u}(I(\boldsymbol{z}_j \boldsymbol{u} \geq 0) - y_j/m_j)$. Note that $f_j(\boldsymbol{u}) \geq 0$ for all j and $\boldsymbol{u}$, and by

Lemma 7.7.1, $\boldsymbol{u} \in \mathfrak{U}$ if and only if $f_j(\boldsymbol{u}) = 0$ for all $j \leq n$. Then

$$\sum_j m_j f_j(\boldsymbol{u}) = \sum_j n_j \boldsymbol{z}_j \boldsymbol{u} I(\boldsymbol{z}_j \boldsymbol{u} \geq 0) - \boldsymbol{t}^\top \boldsymbol{u}, \tag{211}$$

and $\boldsymbol{u} \in \mathfrak{U}$ if and only if (211) is zero. Take $\boldsymbol{u} \in \mathfrak{U}$ such that $\boldsymbol{u} \neq \boldsymbol{o}$. Choose $\boldsymbol{s}$ and $\boldsymbol{r}$ vectors of non-negative real numbers satisfying $\boldsymbol{s} - \boldsymbol{r} = \boldsymbol{Z}\boldsymbol{u}$, and such that for all j, either s_j or r_j is zero. Then (208) holds. Since $\boldsymbol{Z}$ is of full rank, $\boldsymbol{s} - \boldsymbol{r} \neq \boldsymbol{o}$. If (211) is zero then $\boldsymbol{m}^\top \boldsymbol{s} - \boldsymbol{t}^\top \boldsymbol{u} = 0$, and (209) follows. Equation (210) follows by multiplying $\boldsymbol{Z}\boldsymbol{u}$ by $\boldsymbol{Q}$.

Take $\boldsymbol{s}$ and $\boldsymbol{r}$ such that $\boldsymbol{s} \neq \boldsymbol{r}$ and such that (209) and (210) hold. Relation (210) implies that $(\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top)(\boldsymbol{s} - \boldsymbol{r}) = \boldsymbol{o}$; hence (209) and (210) together imply that $-(\boldsymbol{m}^\top - \boldsymbol{y}^\top)\boldsymbol{s} - \boldsymbol{y}^\top \boldsymbol{r} = 0$, and so for each j either $s_j = 0$ or $r_j = 0$. Hence if $\boldsymbol{u}$ is defined by (208), then s_j and r_j are the positive and negative parts of $\boldsymbol{z}_j \boldsymbol{u}$ respectively, and (209) implies that (211) is zero. The equation $-(\boldsymbol{m}^\top - \boldsymbol{y}^\top)\boldsymbol{s} - \boldsymbol{y}^\top \boldsymbol{r} = 0$ also proves the last claim of the theorem.

<div align="right">Q.E.D</div>

One choice for $\boldsymbol{Q}$ is $(\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top)$; a more efficient choice is constructed by using a trivial variation of the Gramm-Schmidt procedure to determine a set of vectors orthogonal to the columns of $\boldsymbol{Z}$. Clarkson and Jennrich (1991) begin with the raw success counts $\boldsymbol{y}$, and reparameterize the problem in such a way as to insure that $y_j < n_j$ for all $j \leq n$. Since for each $\boldsymbol{u} \in \mathcal{U}$ equivalent $\boldsymbol{s}$ and $\boldsymbol{r}$ satisfying (208), (209), and (210) also satisfy $-(\boldsymbol{m}^\top - \boldsymbol{y}^\top)\boldsymbol{s} - \boldsymbol{y}^\top \boldsymbol{r} = 0$, then $\boldsymbol{s} = \boldsymbol{o}$. They proceed using linear programming methods to provide a quick method of searching through all possible subsets of $\{1, \ldots, n\}$ to find the maximal number of positive entries in $\boldsymbol{r}$ satisfying (209) and (210). The algorithm of Kolassa and Tanner (1994) requires proceeding using only sufficient statistics, the above reparameterization is impossible. This problem is solved by allowing $\boldsymbol{s}$ to be non-negative, and determining the maximal number of positive entries in $\boldsymbol{r}$ and $\boldsymbol{s}$ satisfying (209) and (210).

Determine which covariate vectors $\boldsymbol{z}_j$ are associated with fitted observations of 0 or m_j, as follows: Use a modification of the simplex algorithm described by Clarkson and Jennrich (1991) to find vectors $\boldsymbol{r}$ and $\boldsymbol{s}$ with the maximal number of positive entries, for s_j and r_j nonnegative, subject to (209) and (210). Let $\boldsymbol{s}^*$ and $\boldsymbol{r}^*$ be the maximizing values of these vectors. Calculate $\boldsymbol{u}^*$ from $\boldsymbol{s}^*$ and $\boldsymbol{r}^*$ using (208). This method succeeds in identifying all problematic observations:

Lemma 7.7.3: *Let $\boldsymbol{s}^*$ and $\boldsymbol{r}^*$ be the values of the non-negative vectors $\boldsymbol{r}$ and $\boldsymbol{s}$, satisfying (209) and (210), with the maximal number of positive entries. If $\boldsymbol{v} \in \mathfrak{U}$, and if $j \in \{1, \cdots, n\}$ such that $r_j^* = s_j^* = 0$, then $\boldsymbol{z}_j \boldsymbol{v} = 0$.*

Proof: Suppose that there exists a vector $\boldsymbol{v} \in \mathfrak{U}$ satisfying (207), and an index j such that $r_j^* = s_j^* = 0$, and $\boldsymbol{z}_j \boldsymbol{v} \neq 0$. Then $\boldsymbol{u}^* + \boldsymbol{v} \in \mathfrak{U}$. Lemma 7.7.1 indicates that for each j, $\boldsymbol{z}_j \boldsymbol{u}^*$ and $\boldsymbol{z}_j \boldsymbol{v}$ cannot have opposite sign, since each is either zero or has the sign of $y_j/m_j - \frac{1}{2}$. Hence $\boldsymbol{Z}(\boldsymbol{u}^* + \boldsymbol{v})$ has more nonzero elements than $\boldsymbol{Z}\boldsymbol{u}^*$,

contradicting the optimality of $\boldsymbol{s}^*$ and $\boldsymbol{r}^*$.

<div align="right">Q.E.D</div>

Clarkson and Jennrich (1991) suggest using a modified simplex algorithm to solve the above optimization. That is, optimize an objective function of the form $\mathbf{1}\boldsymbol{r} + \mathbf{1}\boldsymbol{s}$ as usual, subject to (209) and (210), and to $\boldsymbol{r} \geq \boldsymbol{o}$ and $\boldsymbol{s} \geq \boldsymbol{o}$, except that when the algorithm suggests that increasing a component of $(\boldsymbol{r}, \boldsymbol{s})$ leads to an increase in the objective function subject to (209) and (210), instead record this fact, change the objective function multiplier for this component to zero, and proceed as before.

Kolassa and Tanner (1994) use the approximation of §7.3 to probabilities of form $\mathrm{P}\left[T_k \geq t_k | \boldsymbol{T}_{-k} = \boldsymbol{t}_{-k}\right]$ in conjunction with Gibbs sampling to calculate p-values for tests of *multivariate* null hypotheses involving components of the canonical parameter vector in a multivariate exponential family conditional on the sufficient statistics associated with the other parameters. Here $\boldsymbol{T}$ is a random vector and $\boldsymbol{T}_{-k}$ is the same vector with component k removed.

Approximation (189) is undefined when (183) fails to have a finite solution. In these cases, (193) also fails to have a finite solution. In this case the conditional distribution of T^j can be derived from the generalized linear model including only some of the observations. The design matrix in this reduced regression model, however, is of less than full rank. Some columns should be dropped to yield a full rank matrix.

Lemma 7.7.4: *Perform the optimization implied by Lemma 7.7.3 on the generalized linear model with covariate $1, \ldots, k$ omitted. Classify each of the covariate vectors into one of three cases. Assign an index j to $\mathcal{A}$ if $r_j^* > 0$, assign an index j to $\mathcal{B}$ if $s_j^* > 0$, and assign an index j to $\mathcal{C}$ if $s_j^* = r_j^* = 0$. Let $\boldsymbol{Z}^*$ be the regression matrix $\boldsymbol{Z}$ whose rows with indices in $\mathcal{C}$ are set to zero, and whose columns $k, \ldots, \ell$ are reduced to form a full–rank matrix. Let $\mathcal{D}$ represent the deleted columns. Let $\boldsymbol{W} = \boldsymbol{Z}^{*\top}\boldsymbol{Y}$, and let $\boldsymbol{w} = \boldsymbol{Z}^{*\top}\boldsymbol{y}$. Then $\mathrm{P}\left[T^1 = t^1, \ldots, T^k = t^k | T^{k+1} = t^{k+1}, \ldots\right] = \mathrm{P}\left[W^1 = w^1, \ldots, W^k = w^k | W^{k+1} = w^{k+1}, \ldots\right]$. Here $\boldsymbol{w}$ is calculated from $\boldsymbol{t}$ by subtracting $\boldsymbol{z}_j m_j$ for each $j \in \mathcal{B}$, and dropping components in $\mathcal{D}$.*

Proof: Let $\sum^{\dagger}$ represent the summation over all vectors $\boldsymbol{y}$ such that $\boldsymbol{Z}\boldsymbol{y} = \boldsymbol{t}$, and let $\sum^{\ddagger}$ be the larger summation over vectors $\boldsymbol{y}$ with the constraint arising from $t^1, \ldots, t^k$ removed. Then

$$\mathrm{P}\left[T^1 = t^1, \ldots, T^k = t^k | T^{k+1} = t^{k+1}, \ldots\right] = \sum\nolimits^{\dagger} \prod_{i=1}^{n} \mathrm{P}\left[Y_i = y_i\right] / \sum\nolimits^{\ddagger} \prod_{i=1}^{n} \mathrm{P}\left[Y_i = y_i\right]$$

$$= \sum\nolimits^{\dagger} \prod_{i \in \mathcal{C}}^{n} \mathrm{P}\left[Y_i = y_i\right] / \sum\nolimits^{\ddagger} \prod_{i \in \mathcal{C}}^{n} \mathrm{P}\left[Y_i = y_i\right], \tag{212}$$

since for all vectors $\boldsymbol{y}$ in both summations, the values of y_j are fixed at the same values for all $j \notin \mathcal{C}$. Since the non-zero rows of $\boldsymbol{Z}$ span the same space as those of

$\boldsymbol{W}$, $\sum^{\dagger}$ also represents the summation over all vectors $\boldsymbol{y}$ such that $\boldsymbol{Z}^*\boldsymbol{y} = \boldsymbol{w}$, and $\sum^{\ddagger}$ also represents the summation over vectors $\boldsymbol{y}$ with the constraint arising from w_1 removed.

<div align="right">Q.E.D</div>

Now conditional probabilities for $\boldsymbol{T}$ are exactly those of $\boldsymbol{W}$, and may be calculated from (189):

Corollary 7.7.5: *When $k = 1$, if the random vector $\boldsymbol{W}$ and the observed value $\boldsymbol{w}$ are defined as in Lemma 7.7.4, then $\mathrm{P}\left[W^1 \geq w^1 | \boldsymbol{W}_{-1} = \boldsymbol{w}_{-1}\right]$ can be calculated from (189) without further modification.*

Proof: Lemma 7.7.4 presents a reparameterization of the generalized linear model to avoid problems in the calculation of the saddlepoint for the reduced model. No further reparameterization is generally necessary in order to calculate the saddlepoint approximation for tail probabilities. Since t^1, and hence w_1, is corrected for continuity, then either $\boldsymbol{w}$ is in the interior of the range of $\mathcal{H}'_{\boldsymbol{W}}$, or outside of the closure of the range of $\mathcal{H}'_{\boldsymbol{W}}$. In the former case, both the full and reduced saddlepoints exist, and (189) may be applied to the distribution of $\boldsymbol{W}$ with no further modification. In the latter case, an additional maximization of $\boldsymbol{1}\boldsymbol{r} + \boldsymbol{1}\boldsymbol{s}$, using an unmodified version of the simplex algorithm (Press, *et. al.*, 1986), subject to (209) and (210), and to $\boldsymbol{r} \geq \boldsymbol{o}$ and $\boldsymbol{s} \geq \boldsymbol{o}$, indicates that $\boldsymbol{1}\boldsymbol{r} + \boldsymbol{1}\boldsymbol{s}$ is unbounded, and hence the conditional probability is either 0 or 1. One may choose between 0 and 1 by checking the value of t^k in question with a value known to lie in the conditional sample space. That is, if optimization indicates that $\mathrm{P}\left[T^k \geq t^k | \boldsymbol{T}_{-k} = \boldsymbol{t}_{-k}\right] \in \{0, 1\}$, if $\check{\boldsymbol{t}} = \boldsymbol{Z}^{\top}\boldsymbol{y}$ for some $\boldsymbol{y} \in \prod_j[0, m_j]$, if $\check{\boldsymbol{t}}_{-k} = \boldsymbol{t}_{-k}$, then

$$\mathrm{P}\left[T^k \geq t^k | \boldsymbol{T}_{-k} = \boldsymbol{t}_{-k}\right] = \begin{cases} 1 & \text{if } t^k < \check{t}^k \\ 0 & \text{if } t^k > \check{t}^k \end{cases}.$$

<div align="right">Q.E.D</div>

On the other hand, the conditional version of the Blackwell–Hodges formula, given by the right hand side of (180), requires more care, because there exist vectors $\boldsymbol{t}$ assigned positive probability, for which no finite full saddlepoint exists, even when the saddlepoint associated with the conditioning event exists and is finite. This occurs because no continuity correction is employed in (180). In this case, first apply the optimization in Lemma 7.7.4 on $(T^{\mathfrak{m}+1}, \ldots, T^{\mathfrak{k}})$, to eliminate observations for which the conditioning event forces the response variable Y_j to take on 0 or m_j, as was done to obtain (212). In many cases (180) may be applied to the reduced data set. In all cases the unconditional multivariate Blackwell–Hodges formula (142) may be applied to the denominator, but in some cases the full saddlepoint may fail to exist. To allow for these cases, repeat the optimization on the reduced observations, with covariates 1 through $\mathfrak{m}$ added, to obtain new sets $\mathcal{A}'$, $\mathcal{B}'$, and $\mathcal{C}'$, corresponding to observations in the reduced model in which the event $\boldsymbol{W} = \boldsymbol{w}$ forces response variable Y_j to take the values 0, m_j, and some other value,

respectively. Probabilities for Y_j for $j \in \mathcal{A}' \cup \mathcal{B}'$ may be factored out of $\sum^\dagger$. The remainder of the summation corresponds to an unconditional probability from a smaller regression model, as constructed in Lemma 7.7.4, in which (142) may be applied directly. These calculations are presented by Kolassa (2001).

7.8. Multivariate Conditional Approximations

The preceding sections described methods for approximating one-dimensional conditional probabilities. To approximate multidimensional conditional probabilities, one might integrate with respect to other components of the random vector in (185), to obtain approximations to probabilities of rectangular sets. Kolassa and Tanner (1994) introduce a Markov chain Monte Carlo method for approximating probabilities of less regular sets, including rejection regions for frequentist hypothesis tests. They call this method the Gibbs–Skovgaard algorithm. Further details are given by Kolassa (1994, 1995).

The Gibbs sampler is a popular Markov chain method useful for yielding a sample from a posterior or likelihood density. It was first introduced by Geman and Geman (1984) in the context of image reconstruction. See Tanner (1996) for background details and important references.

Let the symbol $p(\cdots|\cdots)$ denote the distribution of those random variables listed before the vertical line conditional on those listed after. To obtain a sample from the joint distribution $p(T^1, \cdots, T^j|T_{j+1}, \ldots, T_{\mathfrak{k}})$, the systematic scan Gibbs sampler iterates the following loop:

1) Sample $T^1_{(i+1)}$ from $p(T^1|T^2_{(i)}, \cdots, T^j_{(i)}, T_{j+1}, \ldots, T_{\mathfrak{k}})$.
2) Sample $T^2_{(i+1)}$ from $p(T^2|T^1_{(i+1)}, T^3_{(i)}, \cdots, T^j_{(i)}, T_{j+1}, \ldots, T_{\mathfrak{k}})$.
   $\vdots$
j) Sample $T^j_{(i+1)}$ from $p(T^j|T^1_{(i+1)}, \cdots, T^{j-1}_{(i+1)}, T_{j+1}, \ldots, T_{\mathfrak{k}})$.

For a sufficiently large value of I we can take $T^1_{(I)}, \cdots, T^j_{(I)}$ as a simulated observation from the equilibrium distribution $p(T^1, \cdots, T^j|T_{j+1}, \ldots, T_{\mathfrak{k}})$ of the Markov chain. Independently replicating this Markov chain produces an independent and identically distributed sample from the distribution of interest. Kolassa and Tanner (1994) sample instead from approximate conditional distributions of Theorem 7.3.1, to give a sample drawn approximately from the conditional distribution $p(T^1, \cdots, T^j|T_{j+1}, \ldots, T_{\mathfrak{k}})$, in order to perform conditional inference in certain generalized linear models (McCullagh and Nelder, 1989).

Kolassa (2004) suggests the use of (161) or (162), in conjunction with the approximate conditional cumulant generating function (184), for approximating the conditional probability of rectangular regions. Other authors, including Waterman and Lindsay (1996), present an analytic asymptotic approximation to continuous conditional distributions, resulting in accurate estimates and standard errors. The

results of Strawderman, Casella, and Wells (1996) might also be extended to examine these conditional distributions.

7.9. Exercises

1. Consider the hypergeometric examples given by Skovgaard (1987):

   a. $\begin{matrix} 85 & 2 \\ 75 & 35 \end{matrix}$ b. $\begin{matrix} 14 & 6 \\ 8 & 12 \end{matrix}$ c. $\begin{matrix} 5 & 3 \\ 1 & 9 \end{matrix}$ d. $\begin{matrix} 5 & 1 \\ 1 & 5 \end{matrix}$ e. $\begin{matrix} 6 & 0 \\ 0 & 6 \end{matrix}$

   a. Calculate the tail probability approximation. (Note that the sign of $\hat{\omega}$ in equation (5.5) of Skovgaard (1987) must be reversed.)
   b. Compare these with the tail probability approximations ignoring continuity corrections.
   c. Using this approximation or otherwise, determine values for x_{11} giving approximately the lower 2.5% point of the distribution, and compare the accuracy of the approximation embodying the continuity correction here. Do not put much effort into determining this point; a normal approximation will probably be good enough.
   d. Calculate the higher–order expansion of §7.4.

2. Consider the logistic regression model:

$$y^j \sim B(\pi_j, n_j), \quad \pi^j = \text{alogit}(\lambda_j), \quad \lambda^j = \sum_{i=1}^{\partial} z_i^j \theta_j,$$

   and the data set

	Observation											
Variable	1	2	3	4	5	6	7	8	9	10	11	12
z_1^j	1	1	1	1	1	1	1	1	1	1	1	1
z_2^j	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1
n^j	8	10	9	9	10	10	9	10	10	9	10	10
y^j	3	4	2	3	2	2	1	1	2	1	1	1 ,

   where the first two rows contain covariates and the last two the number of trials and number of successes. This data is a subset of that collected by Berman, *et. al.* (1990) as part of an experiment to measure the effect of magnetic radiation on chicken embryos. Sufficient statistics for the parameter vector are $\boldsymbol{T} = \boldsymbol{Z}^\top \boldsymbol{Y}$, where $\boldsymbol{Z}$ is the transpose of the matrix formed by the first two rows of the table above, and $\boldsymbol{Y}$ is the column vector whose transpose is given in the final row.
   a. Graph the approximate conditional probability that T_2 meets or exceeds the observed value, as a function of θ_2, using (195).
   b. Graph the approximate conditional probability that T_2 fails to exceed the observed value, as a function of θ_2, using (195).

    c. From parts a and b construct an approximate 95% confidence interval for θ_2.

    d. Repeat a – c using (109) in conjunction with (184).

8

Applications to Wald, Likelihood Ratio, and Maximum Likelihood Statistics

The breath of this material is similar to that of Reid (1988). Of primary concern will be approximation to densities and distribution functions of maximum likelihood estimators, likelihood ratio statistics, and Wald statistics. Bartlett's correction for the distribution of likelihood ratio statistics is derived. Approximate ancillarity is also discussed.

8.1. The Distribution of the Wald Statistic

Consider the problem of testing a hypothesis specifying a parameter $\boldsymbol{\theta}$ in the model having the sufficient statistic vector $\boldsymbol{T} \in \mathfrak{R}^{\mathfrak{k}}$ using the test statistic $V(\boldsymbol{T}) = (\boldsymbol{T} - \boldsymbol{\mu})^{\top} \boldsymbol{\Sigma}^{-1} (\boldsymbol{T} - \boldsymbol{\mu})$, when

$$\mathrm{E}\left[\|\boldsymbol{T}\|^{s}\right] < \infty, \tag{213}$$

for some $s \geq 2$. This statistic a trivial example of a Wald statistic, and in a full exponential family is the score statistic. Suppose that $\boldsymbol{\mu} = \mathrm{E}_{\boldsymbol{\theta}}\left[\boldsymbol{T}\right]$ and $\boldsymbol{\Sigma} = \mathrm{Var}_{\boldsymbol{\theta}}\left[\boldsymbol{T}\right]$, the expectation and variance of $\boldsymbol{T}$ calculated at the postulated parameter value $\boldsymbol{\theta}$, exist and are finite. Suppose further that $\boldsymbol{\Sigma}$ is invertible. When $\boldsymbol{T}$ is approximately multivariate normal, V has a distribution that is approximately χ^2 on $\mathfrak{k}$ degrees of freedom. This may be seen by expressing $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Omega}^{\top} \boldsymbol{\Omega}$, and noting that $\boldsymbol{\Omega}(\boldsymbol{T} - \mathrm{E}_{\boldsymbol{\theta}}\left[\boldsymbol{T}\right])$ is approximately multivariate normal with mean $\mathbf{o}$ and all components independent with unit variance.

When $\boldsymbol{T}$ has a distribution dependent on n, such as when $\boldsymbol{T}$ is the standardized sum of independent and identically distributed random vectors $\boldsymbol{X}_i$, each with characteristic function ζ, then the vague notions of approximation above can be refined. Rigorous asymptotic approximation for tail probabilities of V can be derived with reference to the underlying distribution of $\boldsymbol{T}$. Let $\mathfrak{C}'_n = \{\boldsymbol{t} | (\boldsymbol{t} - \boldsymbol{\mu}_n)^{\top} \boldsymbol{\Sigma}_n^{-1} (\boldsymbol{t} - \boldsymbol{\mu}_n) \leq v\}$ be the elliptical set of $\boldsymbol{t}$ giving rise to $V \leq v$. Theorem 6.6.1 concerning multivariate Edgeworth series, or its analogues in non-independent and identically distributed cases, allows the approximation of probabilities of sets like $\mathfrak{C}'_n$, as long as $\boldsymbol{T}$ satisfies regularity conditions discussed below. These are satisfied, for instance, if $\boldsymbol{T}$ arises

from the likelihood for independent and identically distributed components, each of whose distribution is given by a density with the proper moments. The first order approximation is $\int \cdots \int \phi(\boldsymbol{t}; \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)\, d\boldsymbol{t}$ where the integral is taken over $\mathfrak{E}'$. After a change of variables this is expressed as $F_{\mathfrak{k}}(v)$, where $F_{\mathfrak{k}}$ is the χ^2 cumulative distribution function on $\mathfrak{k}$ degrees of freedom. The χ^2 approximation is accurate to $o(1)$, by the Central Limit Theorem. Under the hypothesis of an integrable characteristic function (147), and with the existence of a third moment as in (213) with $s = 3$ the χ^2 approximation is accurate to $O(1/\sqrt{n})$. If fourth order moments exist, that is, (213) holds with $s = 4$, the next term in the expansion can be added, resulting in an approximation valid to $O(n^{-1})$. However, by the symmetry of $\mathfrak{E}'_n$, the term of $O(1/\sqrt{n})$ is zero, implying that $\mathrm{P}\left[V \leq v\right] = F_{\mathfrak{k}}(v) + O(n^{-1})$.

    This expansion holds when $\boldsymbol{T}$ is continuous. Esseen (1945) remarks at the end of §7.4 that under a multivariate version of Cramér's condition of (149),

$$\limsup_{\|\boldsymbol{\beta}\| \to \infty} |\zeta(\boldsymbol{\beta})| < 1,$$

the next term in the asymptotic expansions for the probability of the ellipse is of order $O(n^{-1})$ in this case as well. These conditions allow the expansion with error of order $O(n^{-1})$ in the multivariate Edgeworth series.

    Bhattacharya and Rao (1976) prove, under Cramér's condition, crediting Rao (1960, 1961), convergence for the multivariate Edgeworth series for general convex sets, and get error terms of size $o(n^{-s/2})$ when terms of size $O(n^{-s/2})$ are included. This is their Corollary 20.4, and this is paraphrased as Theorem 6.6.2 in §6.6 of this volume. Earlier versions of these theorems included powers of $\log(n)$ in the error term; see their book, p. 222, for more bibliographic information. In the case of ellipses symmetric about the origin, the term of size $O(1/\sqrt{n})$ in the integral of the multivariate case over this region is zero, proving that the χ^2 approximation is valid to $O(n^{-1})$.

    The most important class of distributions for which Cramér's condition fails is the class of lattice distributions. A random vector has a lattice distribution if it takes on values in a translation and rotation of $\mathfrak{Z}^{\mathfrak{k}} = \{(z_1, ..., z_{\mathfrak{k}}) | z_j \in \mathfrak{Z} \forall j\}$. When instead $\boldsymbol{T}$ has a lattice distribution approximating probabilities of elliptical regions becomes trickier. Kolassa and McCullagh (1990) show that in the unidimensional lattice case, the Edgeworth series gives approximations valid to $O(n^{-1})$ if evaluated at continuity corrected points and using the third cumulant. The Edgeworth series gives higher order approximations if the cumulants are adjusted and fourth and higher order cumulants are used. In the multivariate case Esseen (1945) showed as Theorem 1 of §7 that $\mathrm{P}\left[\boldsymbol{T}_n \in \mathfrak{E}'_n\right] = F_{\mathfrak{k}}(v) + O(n^{-\mathfrak{k}/(\mathfrak{k}+1)})$ assuming only finite fourth moments for general $\boldsymbol{T}$. Esseen observed (Theorem 1 of §8) that one can approximate atoms of the lattice by the Edgeworth series density approximation using the third cumulants, times the volume of the lattice cell, to order $O(n^{-(\mathfrak{k}+2)/2})$. Estimating the probability of any set of constant volume involves adding $O(n^{\mathfrak{k}/2})$ of these errors, leaving an error of $O(n^{-1})$. Kolassa (2003b) suggests adding saddlepoint approximations to

probability atoms in order to approximate the probabilities of irregularly–shaped sets.

Sharper results are based on careful expansions for the distribution function of $\boldsymbol{T}$. Rao (1960, 1961) develops an analogue to the Esseen's series for the cumulative distribution function of a lattice distribution in the multivariate case. Error bounds here contain factors of $\log(n)$ later proved unnecessary by Bhattacharya and Rao (1976). Evaluation of this series is, however, difficult for non-rectangular and non-elliptical sets. Kolassa (1989) shows that this Rao series, when evaluated at midpoints of lattice cubes, is equivalent to the Edgeworth series at the same points, with cumulants adjusted by Sheppard's corrections, to the same order of error. Yarnold (1972) addresses the problem of evaluating the Rao series for convex sets, and in particular for standardized ellipses. The Yarnold approximation is the χ^2 approximation plus the difference between the actual number of points in the ellipse and the volume of the ellipse divided by the volume of a unit cube of the lattice, times the normal approximation to the density at each point on the ellipse boundary. Specifically, suppose $\boldsymbol{T}_n$ is the sum of independent and identically distributed distributed vectors $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$, divided by $\sqrt{n}$, and suppose that $\boldsymbol{X}_i$ are confined to a lattice with unit spacings, with zero mean and $\mathrm{Var}\,[\boldsymbol{X}_i] = \Sigma$. Choose a point $\boldsymbol{a}$ on the lattice. Then

$$\mathrm{P}\,[V(\boldsymbol{T}) \leq v] = F_{\mathfrak{k}}(v) + \left(N(nv) - \frac{(\pi n v)^{\mathfrak{k}/2} \det [\Sigma]^{1/2}}{\Gamma(\mathfrak{k}/2 + 1)} \right) \frac{\exp(-v/2)}{(2\pi n)^{\mathfrak{k}/2} \det [\Sigma]^{1/2}} + O(n^{-1}),$$

where $N(nv)$ is the number of vectors of integers $\boldsymbol{\mu}$ such that $(\boldsymbol{\mu} + n\boldsymbol{a})^\top \Sigma^{-1} (\boldsymbol{\mu} + n\boldsymbol{a}) < v$. In general calculating $N(nv)$ is computationally intensive. Furthermore, finding an order $O(n^{-1})$ correction term independent of v for use in the development of a Bartlett's correction, to be described below, seems very difficult.

Bartlett (1953, 1955) considers asymptotic approximations to the distributions of score statistics in non-exponential-family settings; other authors including Levin and Kong (1990) have addressed the same question.

8.2. Bartlett's Correction

For a sufficient statistic $\boldsymbol{T}$ with density $f_{\boldsymbol{T}}(\boldsymbol{t}, \boldsymbol{\theta})$ depending on a measure of sample size n, with the unknown parameter $\boldsymbol{\theta} \in \mathfrak{P} \subset \mathfrak{R}^{\mathfrak{d}}$, let $\ell_n(\boldsymbol{\theta}; \boldsymbol{t}) = \log(f_{\boldsymbol{T}}(\boldsymbol{t}, \boldsymbol{\theta}))$, and define the likelihood ratio statistic

$$W_n(\boldsymbol{\theta}) = 2(\ell_n(\hat{\boldsymbol{\theta}}; \boldsymbol{T}_n) - \ell_n(\boldsymbol{\theta}; \boldsymbol{T}_n)), \tag{214}$$

where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator of $\boldsymbol{\theta}$. Then under wide regularity conditions, $\mathrm{P}\,[W_n(\boldsymbol{\theta}) \leq w] \approx F_{\mathfrak{d}}(w)$ where $F_{\mathfrak{d}}$ is the χ^2 cumulative distribution function on $\mathfrak{d}$ degrees of freedom. Serfling (1980) demonstrates this by expanding $\ell_n(\hat{\boldsymbol{\theta}}; \boldsymbol{T}_n) - \ell_n(\boldsymbol{\theta}; \boldsymbol{T}_n)$ in $\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}$ about $\hat{\boldsymbol{\theta}}$. By the definition of $\hat{\boldsymbol{\theta}}$ the first order term

is zero; then

$$2(\ell_n(\hat{\boldsymbol{\theta}}; \boldsymbol{T}_n) - \ell_n(\boldsymbol{\theta}; \boldsymbol{T}_n)) = (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) j_n(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + o_p(1)$$
$$= (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) i_n(\boldsymbol{\theta})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + o_p(1) \sim \chi_{\mathfrak{d}}^2.$$

Here $j_n(\hat{\boldsymbol{\theta}})$ is the observed information matrix evaluated at the maximum likelihood estimator; that is, the negative of the matrix of second derivatives, and $i_n(\boldsymbol{\theta})$ is the expected information matrix evaluated at the true parameter vector, which is equal to the asymptotic inverse variance matrix. When data arise from the sum of independent and identically distributed random variables,

$$j_n = O_p(n), \ i_n = O_p(n), \ \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} = O_p(1/\sqrt{n}). \tag{215}$$

Relations (215) hold if:

a. The distribution of the maximum likelihood estimator is multivariate normal with the usual variance matrix to the correct order, which is true in turn if
    1. f has three derivatives for all data values,
    2. These are bounded by integrable functions near the true parameter values,
    3. For any $\boldsymbol{\theta}$, $0 < \mathrm{E}_{\boldsymbol{\theta}}\left[\partial f(Y, \boldsymbol{\theta})/\partial \boldsymbol{\theta}\right]^2 < \infty$.
b. $i_n(\boldsymbol{\theta})$ and $j_n(\boldsymbol{\theta})$ are sufficiently close; order $o_p(n)$ component-wise is sufficient. This is proved by noting that $j_n(\boldsymbol{\theta})$ obeys the law of large numbers.
c. $j_n(\boldsymbol{\theta})$ and $j_n(\hat{\boldsymbol{\theta}})$ are $o_p(n)$ apart;

this result follows from expanding about $\boldsymbol{\theta}$ and using dominated convergence methods.

Bartlett (1953) notes that generally the $\chi_{\mathfrak{d}}^2$ approximation to the distribution of the log likelihood ratio statistic does not hold to $O(1/n)$. Since $W_n(\boldsymbol{\theta})$ has approximately the $\chi_{\mathfrak{d}}^2$ distribution, $\mathrm{E}_{\boldsymbol{\theta}}\left[W_n(\boldsymbol{\theta})\right] \approx \mathfrak{d}$. As a first refinement to the elementary χ^2 approximation to the distribution of $W_n(\boldsymbol{\theta})$, consider multiplicative rescalings, called Bartlett's corrections, moving the entire distribution closer to the asymptotic approximation. This approach is presented below.

Suppose that $\boldsymbol{Y}_i$ are distributed independent and identically distributed, giving rise to the log likelihood

$$\ell_n(\boldsymbol{\theta}; \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n) = \sum_{i=1}^{n} \ell_n(\boldsymbol{\theta}; \boldsymbol{Y}_i),$$

where $\boldsymbol{\theta} \in \mathfrak{P} \subset \mathfrak{R}^{\mathfrak{d}}$. The next theorem investigates the existence and value of the multiplicative correction, and investigates the distribution of the transformed statistic. In general these questions are hard to answer, and in this section attention is restricted to the exponential family case, in which $\ell_n(\boldsymbol{\theta}; \boldsymbol{y}) = \boldsymbol{\theta}^\top \boldsymbol{T}(\boldsymbol{y}) - \mathcal{H}(\boldsymbol{\theta})$.

Theorem 8.2.1: *When the likelihood ratio test statistic $W_n(\boldsymbol{\theta})$ given by (214) arising from n independent and identically distributed random vectors $\boldsymbol{Y}_j$ from the*

exponential family (175) is divided by the factor $1-b(\boldsymbol{\theta})/n$, then the resulting statis-
tic $W'_n(\boldsymbol{\theta}) = W_n(\boldsymbol{\theta})/(1 - b(\boldsymbol{\theta})/n)$ has a χ^2 density to order $O(n^{-3/2})$. Here $b(\boldsymbol{\theta})$ is
as in (135).

Proof: *Recall that*

$$f_{\boldsymbol{T}}(\boldsymbol{t};\boldsymbol{\theta}) = (2\pi)^{-\partial/2} \det \left[j_n(\hat{\boldsymbol{\theta}}) \right]^{-1/2} \exp(\ell_n(\boldsymbol{\theta}) - \ell_n(\hat{\boldsymbol{\theta}}))(1 + b(\hat{\boldsymbol{\theta}})/(2n)) + O(n^{-2})$$

by (220), where $\boldsymbol{T}_n = \sum_{j=1}^n \boldsymbol{Y}_j$. As before,

$$f_{\hat{\boldsymbol{\theta}}}(\hat{\boldsymbol{\theta}};\boldsymbol{\theta}) = (2\pi)^{-\partial/2} \det \left[j_n(\hat{\boldsymbol{\theta}}) \right]^{1/2} \exp(-W(\boldsymbol{\theta})/2)(1 + b(\hat{\boldsymbol{\theta}})/(2n)) + O(n^{-2}) \quad (216)$$

here $j(\boldsymbol{\theta})$ is the observed information matrix, and $W(\boldsymbol{\theta})$ is calculated as in (214) with
$\boldsymbol{T}_n$ the sufficient statistic vector necessary to yield a maximum likelihood estimator
$\hat{\boldsymbol{\theta}}$. Since this is an exponential family, $j(\boldsymbol{\theta})$ is free of $\boldsymbol{T}$ and $j(\boldsymbol{\theta}) = \mathcal{H}''_{\boldsymbol{T}}(\boldsymbol{\theta})$.

   In the multivariate case this proof requires an integration over a $\partial-1$ dimensional
surface of $\hat{\boldsymbol{\theta}}$ on which W is constant. The proof given here is for the simpler
univariate case, and is found in Barndorff–Nielsen and Cox (1979). Since $W = 2((\hat{\theta} - \theta)T + n(\mathcal{H}(\theta) - \mathcal{H}(\hat{\theta}))$, and $T = n\mathcal{H}'(\hat{\theta})$, then $W = 2n((\hat{\theta} - \theta)\mathcal{H}'(\hat{\theta}) + \mathcal{H}(\theta) - \mathcal{H}(\hat{\theta}))$.
Then $dW/d\hat{\theta} = 2n(\hat{\theta} - \theta)\mathcal{H}''(\hat{\theta}) + 2n\mathcal{H}'(\hat{\theta}) - 2n\mathcal{H}'(\hat{\theta}) = 2n(\hat{\theta} - \theta)\mathcal{H}''(\hat{\theta})$, and

$$f_W(w;\theta) = \frac{1}{2\sqrt{2\pi w}} \exp(-\tfrac{1}{2}w) \sum \left| \frac{w/n}{(\hat{\theta} - \theta)^2 \mathcal{H}''(\hat{\theta})} \right|^{1/2} \left\{ 1 + \frac{b(\hat{\theta})}{2n} + O(n^{-2}) \right\}.$$

The sum above is taken over both values of $\hat{\theta}$ yielding the same value for w.

   Since $\mathcal{H}(\hat{\theta}) = \sum_{j=0}^{\infty} \mathcal{H}^{(j)}(\theta)(\hat{\theta} - \theta)^j/j!$, then

$$W/n = 2 \left[\mathcal{H}(\theta) + (\hat{\theta} - \theta) \sum_{j=1}^{\infty} \mathcal{H}^{(j)}(\theta)(\hat{\theta} - \theta)^{j-1}/(j-1)! - \sum_{j=0}^{\infty} \mathcal{H}^{(j)}(\theta)(\hat{\theta} - \theta)^j/j! \right]$$

$$= 2 \left[\sum_{j=2}^{\infty} \mathcal{H}^{(j)}(\theta)(\hat{\theta} - \theta)^j/(j(j-2)!) \right]. \quad (217)$$

Then

$$(W/n)/(\theta - \hat{\theta})^2 \mathcal{H}''(\hat{\theta}) = \frac{\sum_{j=2}^{\infty} \mathcal{H}^{(j)}(\theta)(\theta - \hat{\theta})^j/j!}{\sum_{j=2}^{\infty} \mathcal{H}^{(j)}(\theta)(\theta - \hat{\theta})^j/(j(j-2)!)},$$

and

$$\sqrt{(W/n)/(\theta - \hat{\theta})^2 \mathcal{H}''(\hat{\theta})} \quad (218)$$

has a power series expansion for $\hat{\theta}$ near θ. The lead term is one. If the square
root is taken to both sides of (217) and one solves for $\hat{\theta}$ then powers of $\sqrt{W/n}$ can
replace those of $\hat{\theta}$ in (218) yielding a power series for $\sqrt{(W/n)/(\theta - \hat{\theta})^2 \mathcal{H}''(\hat{\theta})}$ in
powers of $\sqrt{W/n}$. The solution will depend on whether θ is greater or less than $\hat{\theta}$.

The coefficient of the first order term of $\sqrt{W/n}$ in (218) will vary in sign according to which side of the true parameter $\hat{\theta}$ is on, since to first order (217) is quadratic. Higher order terms will also vary in magnitude. Hence

$$\sqrt{(W/n)/(\theta - \hat{\theta})^2 \mathcal{H}''(\hat{\theta})} = 1 \pm c_1\sqrt{W/n} + (c_2 \pm c_3)W/n + O(n^{-3/2}),$$

and $[(\theta - \hat{\theta})^2 \mathcal{H}''(\hat{\theta})]^{-1/2} = \sqrt{n/W}(1 \pm c_1\sqrt{W/n} + (c_2 \pm c_3)W/n + O(n^{-3/2}))$, and

$$
\begin{aligned}
f_W(w;\theta) &= \frac{q_1(w)}{2}\left[\sum\left(1 \pm c_1\sqrt{\frac{w}{n}} + (c_2 \pm c_3)\frac{w}{n}\right)\left(1 + \frac{b(\theta) + e_n\sqrt{\frac{w}{n}}}{2n}\right) + O(n^{-3/2})\right] \\
&= \frac{q_1(w)}{2}\left[\sum\left(1 + c_2\frac{w}{n}\right)\left(1 + \frac{b(\theta)}{2n}\right) + O(n^{-3/2})\right] \\
&= q_1(w)\left(1 + (2c_2 w + b(\theta))/2n + O(n^{-3/2})\right),
\end{aligned}
$$

where $q_1(w) = \phi(+\sqrt{w})/\sqrt{w}$ is the density of a χ^2 variable with one degree of freedom, and e_n is the derivative of $b(\hat{\theta})$ with respect to w/n, evaluated somewhere between θ and $\hat{\theta}$, and under the conditions of Theorem 4.6.1, this is bounded. As the summation is taken over the two values corresponding to the positive and negative values for $\pm$, these terms cancel. Since f_W must integrate to 1, and since the norming constants for the χ^2 distributions with 1 and 3 degrees of freedom are the same, $2c_2 + b(\theta) = 0$, and

$$f_W(w;\theta) = q_1(w)(1 + b(\theta)(1 - w)/2n + O(n^{-3/2})).$$

The density of $W' = W/(1 - b(\theta)/n)$ is $f_{W'}(w') = f_W(w'(1 - b(\theta)/n))\,(1 - b(\theta)/n)$. Hence

$$f_{W'}(w') = q_1(w')\exp(w'b(\theta)/2n)\,(1 - b(\theta)/n)^{1/2}\left(1 + \frac{b(\theta) - w'b(\theta)}{2n} + O(n^{-3/2})\right).$$

Then

$$
\begin{aligned}
f_{W'}(w') &= q_1(w')\left(1 + w'b(\theta)/[2n] + O(n^{-2})\right)\left(1 - b(\theta)/[2n] + O(n^{-2})\right) \times \\
&\quad \left(1 + (1 - w')b(\theta)/[2n] + O(n^{-3/2})\right) \\
&= q_1(w')\left(1 + O(n^{-3/2})\right).
\end{aligned}
$$

Q.E.D

In the general multivariate case, recall (216). If A is a positive definite matrix then for any $c > 0$, $\hat{\boldsymbol{\beta}}^\top A^\top A \hat{\boldsymbol{\beta}} = c$ is an ellipse in $\mathfrak{R}^\partial$, with surface area

$2\left(\pi^{\mathfrak{d}/2}\right)/\Gamma(\mathfrak{d}/2)\det[A]^{-1}c^{\mathfrak{d}/2-1/2}$. Hence, if $d\sigma$ is the surface area measure,

$$\int_{\hat{\boldsymbol{\theta}}\text{ yielding }w}\det\left[j(\hat{\boldsymbol{\theta}})\right]^{1/2}d\sigma=\int_{\hat{\boldsymbol{\theta}}|(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^{\top}j(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})=w}\det\left[j(\hat{\boldsymbol{\theta}})\right]^{1/2}d\sigma$$
$$\approx w^{\mathfrak{d}/2-1/2}/\Gamma(\mathfrak{d}/2).$$

As in the univariate case it is sufficient to show that an integrable term of order $O(n^{-1})$ exists. From the order $O(n^{-1})$ term in the multivariate saddlepoint expansion one can show that to order $O(n^{-3/2})$ the resulting density for W is $(1-\mathfrak{d}b(\boldsymbol{\theta})/n)q_{\mathfrak{d}}(w)+(\mathfrak{d}b(\boldsymbol{\theta})/n)q_{\mathfrak{d}+2}(w)$, and that the density of $W'=W/(1-b(\boldsymbol{\theta})/n)$ is $q_{\mathfrak{d}}(w')$ to order $O(n^{-3/2})$, where $q_{\mathfrak{d}}$ is the χ^2 density on $\mathfrak{d}$ degrees of freedom. Again $b(\boldsymbol{\theta})$ is given by (136). References for this material are Reid (1988), Barndorff–Nielsen and Cox (1979), and Barndorff–Nielsen and Cox (1984).

Another approach is presented by Lawley (1956), McCullagh and Cox (1986), and McCullagh (1987). This involves constructing approximately normal random variables $\boldsymbol{Y}$ such that $W=\boldsymbol{Y}^{\top}j(\boldsymbol{\theta})\boldsymbol{Y}$, and such that $\boldsymbol{Y}$ has a non-zero mean. The resulting W has a non-central χ^2 distribution with cumulants $\kappa_l=(1-c/n)^l2^{l-1}(l-1)!\mathfrak{d}+O(n^{-2})$, implying that W' has the same cumulants as the χ^2 distribution, to order $O(n^{-2})$. See also Cordeiro and Paula (1989).

The extension to cases of compound null hypotheses is more difficult. In many such cases, the parameter of interest $\boldsymbol{\theta}$ is written as $(\boldsymbol{\alpha},\boldsymbol{\beta})$, where the null hypothesis is $\boldsymbol{\alpha}=\boldsymbol{\alpha}_0$. Often the test statistic is $W(\boldsymbol{\alpha})=2(\ell_n(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})-\ell_n(\boldsymbol{\alpha},\tilde{\boldsymbol{\beta}}(\boldsymbol{\alpha})))$ where $\tilde{\boldsymbol{\beta}}(\boldsymbol{\alpha})$ maximizes ℓ_n for a fixed value of $\boldsymbol{\alpha}$. Barndorff–Nielsen and Cox (1984) do this on a case-by-case basis, by determining if necessary the existence of an appropriate ancillary statistic allowing use of the Barndorff–Nielsen's formula to estimate the density of the constrained maximum likelihood estimator. The Bartlett factor is determined from the difference in norming function for these two approximations; if $b_1(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})$ is the norming function for the density of the full maximum likelihood estimator, and $b_2(\boldsymbol{\alpha}_0,\tilde{\boldsymbol{\beta}}(\boldsymbol{\alpha}_0))$ is the norming constant for the density of the constrained maximum likelihood estimator, as in (216), then the Bartlett factor is $1-(b_1(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})-b_2(\boldsymbol{\alpha}_0,\tilde{\boldsymbol{\beta}}(\boldsymbol{\alpha}_0)))/(n\mathfrak{d})$ where $\mathfrak{d}$ is the length of $\boldsymbol{\alpha}$, the change in degrees of freedom as extra parameters are estimated. The functions b may often be calculated from saddlepoint expansions to the densities of the sufficient statistics.

Proofs that the Bartlett's correction is as above follow a general pattern:

1. Transform the saddlepoint density approximation,

$$f_W(w;\boldsymbol{\theta})=\frac{1}{2}\frac{\exp(-w/2)}{(\sqrt{2\pi w})^{\mathfrak{d}}}\int\det\left[j(\hat{\boldsymbol{\theta}})\right]^{-1/2}\mathfrak{d}s(\hat{\boldsymbol{\theta}})\left(1+b(\boldsymbol{\theta})/2n+O(n^{-3/2})\right).$$

2. Expand the integrand in $\sqrt{W/n}$. The constant term is 1, the first order term is 0, and the second term is of form b_2w/n. Find b_2 either using a careful expansion, or by using next correction term in the density of the sufficient statistic, and the fact that the corrected density integrates to $1+O(n^{-3/2})$.

3. Show that $(1 - c/2n) f_W(w(1 - c/2n))$ is a χ^2 density to order $O(n^{-3/2})$.

As an example consider testing $H_0 : \mu = \mu_0$ in the normal model with known variance; that is, suppose that $T = Y_1 \sim N(\mu, 1)$. Since $\ell_n(\mu; t) = (\mu t - \mu^2/2)$, then $W(\mu_0) = (t - \mu_0)^2$. Hence the distribution of $W(\mu_0)$ is exactly χ^2. The basic saddlepoint approximation to the density of T is exact, and hence is, $b(\mu) = 0$.

As another example consider testing the scale parameter in a gamma distribution with known shape α. The density here is $\gamma^\alpha y^{\alpha-1} \exp(-\gamma y)/\Gamma(\alpha)$; we assume that α is known. The log likelihood for an independent and identically distributed sample $Y_1, \dots, Y_n$ of size n is $- (\gamma \sum_i Y_i) + n\alpha \log(\gamma) + (\alpha - 1) \sum_i \log(Y_i) - n \log(\Gamma(\alpha))$, its derivative is $n(-T + \alpha/\gamma)$ for $T = \bar{Y}$, and the maximum likelihood estimator for γ is $\hat{\gamma} = \alpha/T$. Then $W(\gamma) = n[\alpha(1 - \log(T) + \log(\alpha)) + \gamma T - \log(\gamma))]$. From §4.2, $b(\beta)$ is $-\frac{1}{6}\alpha^{-1}$, which is independent of γ. Recall from §4.8 that the gamma distribution, with the normal and Inverse Gaussian, are the only distributions with constant third and fourth standardized cumulants. Bartlett's correction is

$$1 + 1/(6n\alpha). \tag{219}$$

The expectation of the likelihood ratio statistic may also be calculated exactly. Recall that the density of the mean of n independent $\Gamma(\gamma, \alpha)$ random variables has a $\Gamma(n\gamma, n\alpha)$ distribution. The likelihood for γ is now $- (n\gamma \bar{Y}) + \alpha n (\log(n) + \log(\gamma)) + (\alpha n - 1) \log(\bar{Y}) - \log(\Gamma(\alpha n))$, and the likelihood ratio statistic is

$$2n \left(-\gamma \bar{Y} + \alpha(1 - \log(\alpha)) + \alpha \log(\gamma) + \alpha \log(\bar{Y}) \right).$$

This expression is linear in $\bar{Y}$ and $\log(\bar{Y})$. The expectation of these quantities can be calculated analytically: $\mathrm{E}\left[\bar{Y}\right] = \alpha/\gamma$ and $\mathrm{E}\left[\log(\bar{Y})\right] = \psi(\alpha n) - \log(n\gamma)$, where $\psi(x) = \frac{d}{dx} \log(\Gamma(x))$ is the di-gamma function. Hence the expectation of $W(\gamma)$ is

$$2\alpha n \left(\log(\alpha) + \log(n) - \psi(\alpha n) \right).$$

The approximation (219) agrees with this exact value very closely (Fig. 17).

These derivations are difficult for lattice distributions. Examples of lattice distributions for T_n might arise when the underlying random variables $Y_1, \dots, Y_n$ for which T_n is a summary are lattice variables. Examples are:

1. When Y_i is a $\eth$-long vector of zeros except for a one in one place; multinomial probabilities are to be estimated, and the sufficient statistics are T_n are the sum of the Y_i giving cell totals.

2. When Y_i are response variables in a logistic regression corresponding to covariates z_i; the sufficient statistic is $T_n = Z^\top Y$ where Z is the matrix with the z_i as rows, and Y is the vector $(Y_1, \dots, Y_n)$. When the ratios of entries of Z are all rational, T_n is a lattice variable; otherwise it is not a lattice variable but has a singular distribution.

In the present circumstance the variable of real interest is $W(\boldsymbol{\theta})$ rather than T_n; in general $W(\boldsymbol{\theta})$ will not have a lattice distribution, but a singular non-lattice

Evaluation of Bartlett's Correction for Gamma(1,1) Mean Parameter
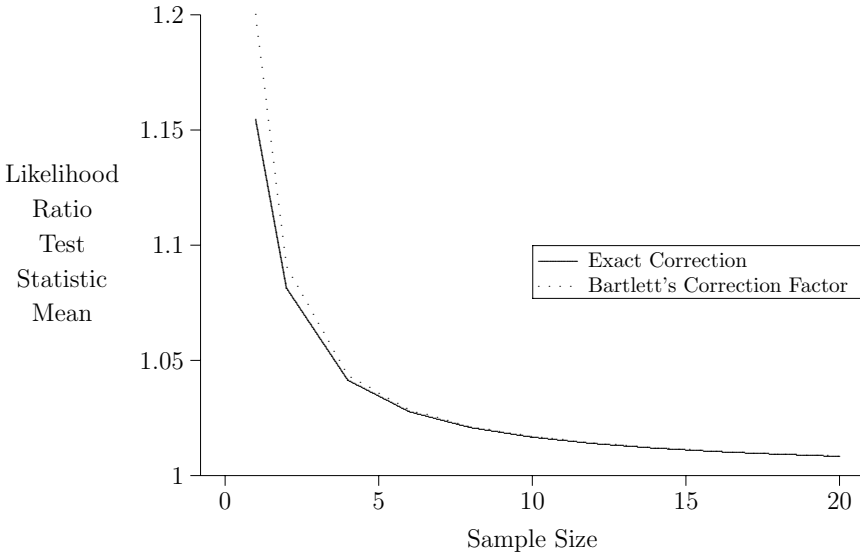
Fig. 17.

distribution in which $\limsup_{|t|\to\infty} \phi_{W(\theta)}(t) = 1$ and the Edgeworth series to order $O(n^{-1})$ and smaller is not valid, and for which no continuity correction exists to give the series better asymptotic properties. Since $W(\theta)$ does not have a density, transformation techniques like those used in the continuous case do not apply, and one is forced to proceed by using asymptotic techniques in T_n space rather than in $W(\theta)$-space. Of interest are calculations of quantities like $\mathrm{P}\left[T_n \in W(\theta)^{-1}[0, w]\right]$. Since these regions are approximately elliptical with the form $\mathfrak{E}' = \{t | t^\top \Sigma^{-1} t \leq c\}$, the ability to approximate probabilities for such sets should be an upper limit on the ability to approximate the cumulative distribution function of $W(\theta)$. Frydenberg and Jensen (1989) provide calculations indicating that while Bartlett's correction does not improve the asymptotic error rate, in many cases it does improve accuracy. Comments on the difficulties in approximating the distribution of the Wald statistic carry over to the likelihood ratio statistic as well.

As an example of Wald and likelihood ratio testing in multivariate discrete cases, consider the problem of simultaneously testing values of means in two independent Poisson distributions. Note the similarity in the Wald and likelihood ratio rejection regions (Fig. 18). Note also the irregular pattern of jumps in both cumulative distribution functions arising because points enter the expanding acceptance regions at irregular intervals (Fig. 19).

Wald and Likelihood Ratio Acceptance Regions for the the Poisson Distribution
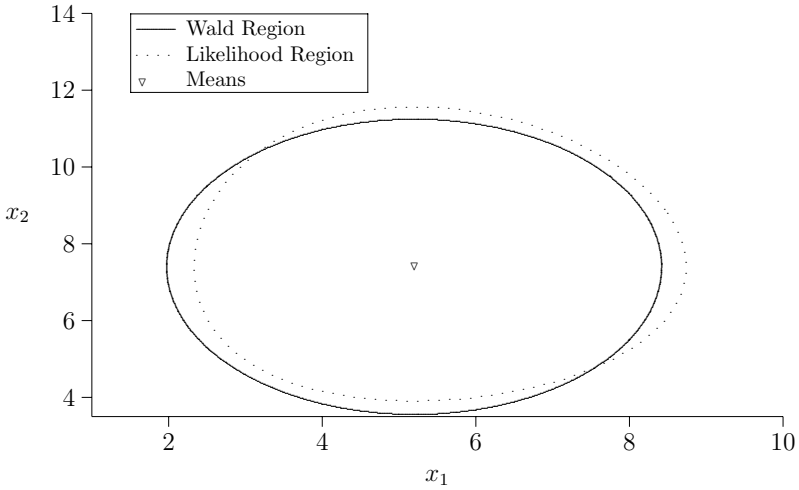
Fig. 18.

Wald and Likelihood Ratio CDFs for the Poisson Distribution
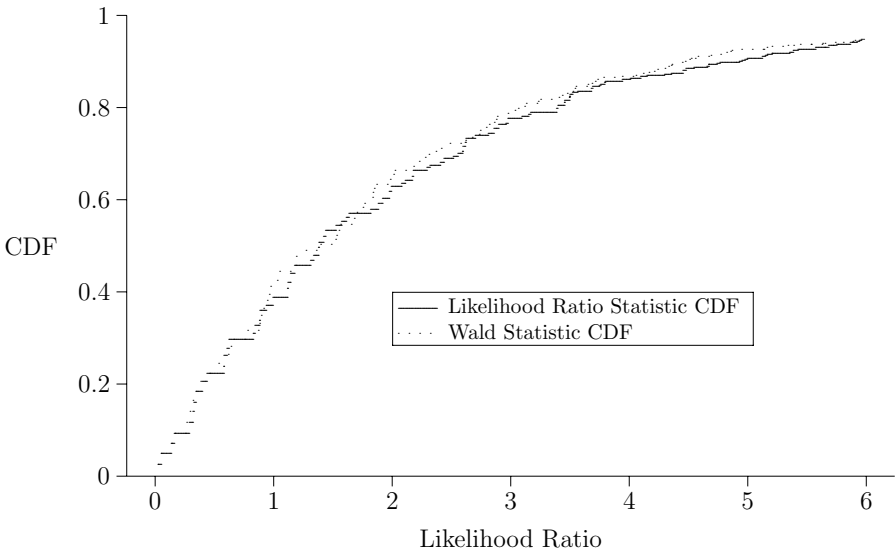
Fig. 19.

Bartlett's corrections may also be applied in less standard contexts; DiCiccio, Hall, and Romano (1991) apply it to empirical likelihoods.

8.3. Distributions of Maximum Likelihood Estimators

As in §7.2, consider the exponential family model for a random vector $\boldsymbol{T}$ whose density evaluated at $\boldsymbol{t}$ is: $f_{\boldsymbol{T}}(\boldsymbol{t};\boldsymbol{\theta}) = \exp(\boldsymbol{\theta}^{\top}\boldsymbol{t} - \mathcal{H}_{\boldsymbol{T}}(\boldsymbol{\theta}) - \mathcal{G}(\boldsymbol{t}))$, such that $\mathcal{H}_{\boldsymbol{T}}$ is twice differentiable and $\mathcal{H}'_{\boldsymbol{T}}$ is one to one. The associated cumulant generating function is $\mathcal{K}_{\boldsymbol{T}}(\boldsymbol{\beta},\boldsymbol{\theta}) = \mathcal{H}_{\boldsymbol{T}}(\boldsymbol{\beta}+\boldsymbol{\theta}) - \mathcal{H}_{\boldsymbol{T}}(\boldsymbol{\theta})$, and hence the saddlepoint $\hat{\boldsymbol{\beta}}$ is defined by $\mathcal{H}'_{\boldsymbol{T}}(\hat{\boldsymbol{\beta}}+\boldsymbol{\theta}) = \boldsymbol{t}$. The log likelihood is $\ell(\boldsymbol{\theta};\boldsymbol{t}) = \boldsymbol{\theta}^{\top}\boldsymbol{t} - \mathcal{H}_{\boldsymbol{T}}(\boldsymbol{\theta})$, and the maximum likelihood estimator is defined by $\mathcal{H}'_{\boldsymbol{T}}(\hat{\boldsymbol{\theta}}) = \boldsymbol{t}$. Hence $\hat{\boldsymbol{\beta}}+\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. Then by (133),

$$
\begin{aligned}
f_{\boldsymbol{T}}(\boldsymbol{t};\boldsymbol{\theta}) &= (2\pi)^{-\mathrm{t}/2} \det\left[\mathcal{K}''_{\boldsymbol{T}}(\hat{\boldsymbol{\beta}};\boldsymbol{\theta})\right]^{-1/2} \exp(\mathcal{K}_{\boldsymbol{T}}(\hat{\boldsymbol{\beta}};\boldsymbol{\theta}) - \hat{\boldsymbol{\beta}}^{\top}\boldsymbol{t})(1 + \tfrac{1}{2}b(\hat{\boldsymbol{\theta}})n^{-1} + O(n^{-2})) \\
&= (2\pi)^{-\mathrm{t}/2} \det\left[\mathcal{H}''_{\boldsymbol{T}}(\hat{\boldsymbol{\theta}})\right]^{-1/2} \exp(\mathcal{H}_{\boldsymbol{T}}(\hat{\boldsymbol{\theta}}) - \mathcal{H}_{\boldsymbol{T}}(\boldsymbol{\theta}) + (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\boldsymbol{t}) \times \\
&\quad (1 + \tfrac{1}{2}b(\hat{\boldsymbol{\theta}})n^{-1} + O(n^{-2})) \\
&= (2\pi)^{-\mathrm{t}/2} \det\left[j(\hat{\boldsymbol{\theta}})\right]^{-1/2} (\mathcal{L}(\boldsymbol{\theta};\boldsymbol{t})/\mathcal{L}(\hat{\boldsymbol{\theta}};\boldsymbol{t})) \left(1 + \tfrac{1}{2}b(\hat{\boldsymbol{\theta}})n^{-1} + O(n^{-2})\right). \quad (220)
\end{aligned}
$$

Here j is the observed information, and $\mathcal{L}$ is the likelihood. In these full exponential families, j is equal to i, the Fisher information $\mathrm{E}_{\boldsymbol{\theta}}\left[-\ell''\right]$. Since $\mathcal{H}'_{\boldsymbol{T}}(\hat{\boldsymbol{\theta}}) = \boldsymbol{t}$, $\mathcal{H}''_{\boldsymbol{T}}(\hat{\boldsymbol{\theta}})(d\hat{\boldsymbol{\theta}}/d\boldsymbol{t}) = \boldsymbol{I}$, and $(d\hat{\boldsymbol{\theta}}/d\boldsymbol{t}) = \mathcal{H}''_{\boldsymbol{T}}(\hat{\boldsymbol{\theta}})^{-1} = i(\hat{\boldsymbol{\theta}})^{-1}$. Let $\hat{\boldsymbol{\Theta}}$ be the random variable arising from solutions to $\mathcal{H}'_{\boldsymbol{T}}(\hat{\boldsymbol{\Theta}}) = \boldsymbol{T}$. Using standard change of variables techniques,

$$
f_{\hat{\boldsymbol{\theta}}}(\hat{\boldsymbol{\theta}};\boldsymbol{\theta}) = (2\pi)^{-\mathrm{t}/2} \det\left[i(\hat{\boldsymbol{\theta}})\right]^{1/2} \mathcal{L}(\boldsymbol{\theta};\boldsymbol{t})/\mathcal{L}(\hat{\boldsymbol{\theta}};\boldsymbol{t}) + O(n^{-1}). \qquad (221)
$$

Approximation (221), first noted by Daniels (1958), is a case of Barndorff–Nielsen's formula, and derived in various cases by Barndorff–Nielsen (1980, 1983) and Durbin (1980), who also also noted that arguments analogous to those in §4.8 indicate that after renormalizing, the error term in (221) can be replaced by $O(n^{-3/2})$. Formula (221) is also known as the p^* formula, from a series of papers primarily by Barndorff–Nielsen in which the density of the maximum likelihood estimator is denoted by p and the approximation is denoted by adding the star.

Reid (1988) considers the example of a gamma distribution with an unknown shape parameter. Suppose that X_i are independently distributed with density

$$
f(x;\mu,\nu) = (\nu/\mu)^{\nu} x^{\nu-1} \exp(-x\nu/\mu)\Gamma^{-1}(\nu).
$$

Then the sample $X_1,\ldots,X_n$ gives rise to the log likelihood

$$
\ell_n(\mu,\nu;\boldsymbol{X}) = n\nu\log(\nu) - n\nu\log(\mu) + (\nu-1)\sum\log(X_i) - (\nu/\mu)\sum X_i - n\log(\Gamma(\nu)),
$$

$\boldsymbol{T} = ((\sum\log X_i)/n, \ (\sum X_i)/n)$ is sufficient, and the maximum likelihood estimators

satisfy

$$\nu\,(-\mu + T_1)/\mu^2 = 0 \text{ and } 1 - T_1/\mu + T_2 - \log(\mu) + \log(\nu) - \psi(\nu) = 0,$$

where ψ is the di–gamma function defined to be $\psi(\nu) = (d/d\nu)\log(\Gamma(\nu))$, and

$$\hat{\mu} = T^1, \ \psi(\hat{\nu}) - \log(\hat{\nu}) = T^2 - \log(T^1). \tag{222}$$

The matrix of log likelihood second derivatives is

$$\frac{\partial^2 l}{\partial(\mu,\nu)\partial(\mu,\nu)^\top} = n \begin{pmatrix} \nu(\mu - 2T^2)/\mu^3 & (-\mu + T^1)/\mu^2 \\ (-\mu + T^1)/\mu^2 & 1/\nu - \psi'(\nu) \end{pmatrix}.$$

Its determinant is

$$\frac{n^2}{\mu^2}\left\{\left(1 - \frac{2T^2}{\mu}\right)(1 - \nu\psi'(\nu)) - \left(\frac{T^2}{\mu} - 1\right)^2\right\} = \frac{n^2}{\mu^2}\left\{\frac{2T^2\nu\psi'(\nu)}{\mu} - \nu\psi'(\nu) - \frac{(T^2)^2}{\mu^2}\right\}.$$

After using (222) to eliminate $\boldsymbol{T}$, $\det\left[j(\hat{\theta})\right]^{-1/2} = n\{\hat{\nu}\psi'(\hat{\nu}) - 1\}^{1/2}/\mu$.

Also, $\ell_n(\mu,\nu;\boldsymbol{X}) - \ell_n(\hat{\mu},\hat{\nu};\boldsymbol{X}) = n(\hat{\nu} - \hat{\mu}\nu/\mu + \nu\log(\hat{\mu}/\mu) + \log(\Gamma(\hat{\nu})/\Gamma(\nu)) + \nu\log(\nu/\hat{\nu}) + (\nu - \hat{\nu})\psi'(\hat{\nu}))$. The joint maximum likelihood estimator density approximation (221) is

$$n\exp\left(n\,(\hat{\nu} - \hat{\mu}\nu/\mu + \nu\log(\hat{\mu}/\mu) + \nu\log(\nu/\hat{\nu}) + \log(\Gamma(\hat{\nu})/\Gamma(\nu)) + (\nu - \hat{\nu})\psi'(\hat{\nu}))\right)\times$$

$$\sqrt{(\hat{\nu}\psi'(\hat{\nu}) - 1)/\hat{\mu}^2}$$

Let $g_1(\hat{\mu};\mu,\nu) = \exp(n(\nu\log(\hat{\mu}/\mu) - \hat{\mu}\nu/\mu - \log(\Gamma(\nu))))(n/\hat{\mu})$, and

$$g_2(\hat{\nu};\nu) = \exp(n(\hat{\nu} + \nu\log(\nu/\hat{\nu}) + (\nu - \hat{\nu})\psi'(\hat{\nu}) + \log(\Gamma(\hat{\nu}))))\sqrt{\hat{\nu}\psi'(\hat{\nu}) - 1}.$$

Hence $\hat{\mu}$ has approximately a gamma density with scale parameter μ and shape parameter $n\nu$, and is approximately independent of $\hat{\nu}$. One can easily verify that up to a multiplicative constant the density approximation for $\hat{\mu}$ is exact. Fig. 20 demonstrates that the approximation for the density of $\hat{\nu}$ is very close.

The numeric approximation is calculated by sampling 10,000 sets of 10 $\Gamma(1,1)$ random variables and calculating T^1 and T^2 for each set. The density for $\hat{\nu}$ might then be approximated by noting from (222) that since $\mathcal{H}_{\boldsymbol{T}}(\hat{\nu}) - \log(\hat{\nu}) = (T^2/n)^` - \log(T^1/n)$, the density of $\hat{\nu}$ is the density of $(T^2/n) - \log(T^1/n)$ times the first derivative of $\psi(\hat{\nu}) - \log(\hat{\nu})$. The density of $(T^2/n) - \log(T^1/n)$ can be approximated by fitting a spline to the midpoints of the bars of the histogram of $(T^2/n) - \log(T^1/n)$. However, since $\psi(\hat{\nu}) - \log(\hat{\nu})$ from (222) has a pole at zero a better approximation is derived by noting that $h(\hat{\nu}) = \exp((T^2/n) - \log(T^1/n))$, where $h(\hat{\nu}) = \exp(\psi(\hat{\nu}) - \log(\hat{\nu}))$. The density is then approximated by fitting a spline to the midpoints of the bars of the histogram of $\exp((T^2/n) - \log(T^1/n))$ and multiplying the result by h'. Also shown is the normal approximation to the

Density of the MLE of the Scale Parameter in a Gamma(1,1) Distribution
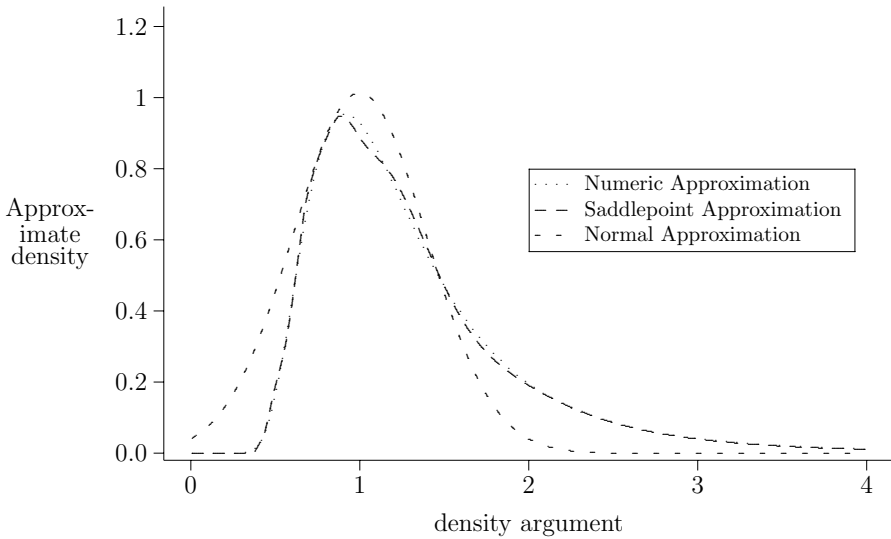
Fig. 20.

density of $\hat{\nu}$, based on the large sample approximations to the mean, ν, and to the variance, $n^{-1}(-1/\nu + \psi'(\nu))^{-1} = (n[\pi^2/6 - 1])^{-1}$.

Formula (221) holds in some non-exponential cases, as shown by Durbin (1980). Suppose $\boldsymbol{X}$ is a random vector of length n, from density $f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \mathfrak{P} \subset \mathfrak{R}^{\partial}$, and suppose $\boldsymbol{T} = \boldsymbol{t}(\boldsymbol{X})$ is a sufficient estimator of $\boldsymbol{\theta}$. By the definition of sufficiency, the density for $\boldsymbol{X}$ can be factored as $f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\theta}) = g(\boldsymbol{t}; \boldsymbol{\theta})h(\boldsymbol{x})$, and for any $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0 \in \mathfrak{P}$, $g(\boldsymbol{t}; \boldsymbol{\theta}_0) = g(\boldsymbol{t}; \boldsymbol{\theta})f(\boldsymbol{x}, \boldsymbol{\theta}_0)/f(\boldsymbol{x}, \boldsymbol{\theta})$. In particular,

$$g(\boldsymbol{t}; \boldsymbol{\theta}_0) = g(\boldsymbol{t}; \boldsymbol{t})[f(\boldsymbol{x}, \boldsymbol{\theta}_0)/f(\boldsymbol{x}, \boldsymbol{t})].$$

This is a non-cumulant generating function analogy to exponential tilting to the mean. Here since $\boldsymbol{t}$ estimates $\boldsymbol{\theta}$ one might expect $\boldsymbol{t}$ to lie near the mean of $g(.; \boldsymbol{t})$ and hence for $g(\boldsymbol{t}; \boldsymbol{t})$ to be easy to approximate. Durbin (1980) uses the normal approximation to the mean to show that

$$g(\boldsymbol{t}; \boldsymbol{\theta}_0) = (n/2\pi)^{\partial/2} \det[i(\boldsymbol{t})]^{1/2} \mathcal{L}(\boldsymbol{\theta}_0)/\mathcal{L}(\boldsymbol{t})\{1 + O(1/n)\}. \tag{223}$$

If $\boldsymbol{X}$ had a cumulant generating function $\mathcal{K}_{\boldsymbol{X}}$ then $i(\boldsymbol{t})$ could be replaced by $\mathcal{K}''_{\boldsymbol{X}}(\boldsymbol{t})$; if $\boldsymbol{X}$ came from an exponential family this would be $\mathcal{H}''_{\boldsymbol{X}}(\boldsymbol{t} + \boldsymbol{\theta})$. Durbin (1980) gives four cases in which (223) holds. The first and most important is:

Theorem 8.3.1: *Suppose that the sufficient statistic vector $\boldsymbol{T}$ with a distribution depending on n estimates $\boldsymbol{\theta}$ with a bias that is uniformly $O(n^{-1})$ in a neighborhood of*

$\boldsymbol{\theta}_0$, that $n\text{Var}\,[\boldsymbol{T}]$ converge to a positive definite matrix uniformly in a neighborhood of $\boldsymbol{\theta}_0$, and that $\boldsymbol{T}$ have a valid Edgeworth expansion to order $o(1/n)$. Then (221) holds with $\hat{\boldsymbol{\theta}} = \boldsymbol{t}$.

8.4. Ancillarity

More subtle expansions of use in conditional inference require a review of the background on sufficiency and ancillarity.   A statistic $\boldsymbol{T} = \boldsymbol{t}(\boldsymbol{X})$ calculated from raw data $\boldsymbol{X}$ and used to make inference on $\boldsymbol{\theta}$ is called sufficient if the density generating $\boldsymbol{X}$ can be factored as $g(\boldsymbol{t};\boldsymbol{\theta})h(\boldsymbol{x})$. The implication is that inference on $\boldsymbol{\theta}$ should be based on $\boldsymbol{T}$ and not on information contained in $\boldsymbol{X}$ but not reflected in $\boldsymbol{T}$. Examples are:

a. If X_i are independent and identically distributed observations from a $N(\mu, \sigma^2)$ distribution, then their density is

$$f_{\boldsymbol{X}}(\boldsymbol{x};\mu,\sigma) = \exp\left(-n\frac{\mu^2}{2\sigma^2} + \frac{\mu}{\sigma^2}\sum x_i - \frac{1}{\sigma^2}\sum_i x_i^2 - n\log(\sigma) - \frac{n}{2}\log(2\pi)\right).$$

   Hence $\boldsymbol{T} = (\sum_i X_i, \sum_i X_i^2)^{\top}$ is sufficient.

b. If $\boldsymbol{X}_i$ are independent and identically distributed random vectors from a natural multivariate exponential family, of the form $f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta}) = \exp(\boldsymbol{\theta}^{\top}\boldsymbol{x} - \mathcal{H}(\boldsymbol{\theta}) - \mathcal{G}(\boldsymbol{x}))$, then $\boldsymbol{T} = \sum_i \boldsymbol{X}_i$ is sufficient.

c. If X_i are independent draws form a binomial distribution with success probability π and number of trials N_i, with N_i known, then $\sum_i X_i$ is sufficient.

   For any statistical model the complete raw data set is always a sufficient statistic. The principal of simplifying analysis by examining a sufficient reduction of the data heuristically leads to choosing, among those sufficient statistics available, the smallest or minimal sufficient statistic, where the minimal sufficient statistic is defined to be any sufficient statistic that is a function of any other sufficient statistic.

   Related to the concept of sufficiency is the concept of ancillarity. If $\boldsymbol{T}$ is a minimal sufficient statistic and can be written as

$$\boldsymbol{T} = \begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{A} \end{pmatrix}$$

where the distribution of $\boldsymbol{A}$ does not depend on $\boldsymbol{\theta}$, then $\boldsymbol{A}$ is called ancillary. Many authors as early as Fisher (1934) have argued for inference conditional on ancillary statistics. For example, if X has a binomial distribution with success probability π and number of trials N, with N generated from some probability model not depending on π, then $\boldsymbol{T} = (X, Y)^{\top}$ is sufficient and Y is ancillary. Then $\mathrm{P}_{\pi}\left[X = x, Y = y\right] = \exp(x\log(\pi/(1-\pi)) + y\log(1-\pi) + \log(\binom{y}{x}))$.

8.5. Approximate Ancillarity

For a multivariate $(\mathfrak{k}, \mathfrak{d})$ curved exponential family assigning probabilities to a random vector $\boldsymbol{T}$ taking values in a convex set $\mathfrak{T} \subset \mathfrak{R}^{\mathfrak{k}}$ according to the density

$$f(\boldsymbol{t}; \theta) = \exp(\boldsymbol{t}^{\top} \boldsymbol{\eta}(\boldsymbol{\theta}) - \mathcal{H}(\boldsymbol{\theta}) - \mathcal{G}(\boldsymbol{t})) \text{ with } \boldsymbol{\eta}: \mathfrak{H} \subset \mathfrak{R}^{\mathfrak{d}} \to \mathfrak{R}^{\mathfrak{k}}, \ \mathfrak{H} \text{ open}, \quad (224)$$

a sufficient statistic is $\boldsymbol{T}$. In this case no shorter sufficient statistic is available, except when $\boldsymbol{\eta}$ is very simple, even though when $\mathfrak{k} > \mathfrak{d}$ the sufficient statistic vector is longer than the parameter vector. Efron (1975) discusses the geometry of such families, and produces an asymptotic expansion for the variance of the maximum likelihood estimator. This will be discussed later in the context of efficiency.

Often times greater asymptotic precision can be obtained by using conditioning to reduce the dimensionality of the conditioned sampling distribution to that of the unknown parameter. Assume that the curved exponential family can be embedded in a full exponential family; that is, suppose that there exists an open set $\mathfrak{Q} \supset \boldsymbol{\eta}(\mathfrak{H})$ such that

$$\mathcal{H}^{*}(\boldsymbol{\eta}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp(\boldsymbol{t}^{\top} \boldsymbol{\eta} - \mathcal{G}(\boldsymbol{t})) \ d\boldsymbol{t} < \infty \ \forall \boldsymbol{\eta} \in \mathfrak{Q}.$$

Then the cumulant generating function for the larger family exists. This assumption insures that the variance matrix for the sufficient statistic vector exists, and insures that the chain rule can be used for differentiating. It also insures that $\mathcal{H}^{*}$ is twice-differentiable, and that the second derivative matrix is positive-definite at all $\boldsymbol{\eta} \in \mathfrak{Q}$. Assume further that the full-model likelihood equation $\mathcal{H}^{*\prime}(\boldsymbol{\eta}) = \boldsymbol{t}$ has a solution in $\mathfrak{Q}$ for all $\boldsymbol{t} \in \mathfrak{T}$. Set $\boldsymbol{\tau}(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}[\boldsymbol{T}]$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathrm{Var}_{\boldsymbol{\theta}}[\boldsymbol{T}]$. Since $\boldsymbol{\tau}(\boldsymbol{\theta}) = \mathcal{H}^{*\prime}(\boldsymbol{\eta}(\boldsymbol{\theta}))$, then

$$\partial \boldsymbol{\tau}/\partial \boldsymbol{\theta} = \mathcal{H}^{*\prime\prime}(\boldsymbol{\eta}(\boldsymbol{\theta}))\partial \boldsymbol{\eta}/\partial \boldsymbol{\theta} = \boldsymbol{\Sigma}\partial \boldsymbol{\eta}/\partial \boldsymbol{\theta}. \quad (225)$$

The range of $\boldsymbol{\tau}(\boldsymbol{\theta})$ then represents the part of the data space that can be explained by the data. The observed information $j(\boldsymbol{\theta})$ and the Fisher information $i(\boldsymbol{\theta})$ in this family are given by

$$j(\boldsymbol{\theta}) = -\sum_{j}(T^{j} - \tau^{j})\partial^{2}\eta_{j}/(\partial \boldsymbol{\theta}^{\top}\partial \boldsymbol{\theta}) + (\partial \boldsymbol{\tau}/\partial \boldsymbol{\theta})^{\top}\partial \boldsymbol{\eta}/\partial \boldsymbol{\theta}$$

$$i(\boldsymbol{\theta}) = (\partial \boldsymbol{\tau}/\partial \boldsymbol{\theta})^{\top}\partial \boldsymbol{\eta}/\partial \boldsymbol{\theta} = (\partial \boldsymbol{\eta}/\partial \boldsymbol{\theta})^{\top}\boldsymbol{\Sigma}\partial \boldsymbol{\eta}/\partial \boldsymbol{\theta}. \quad (226)$$

Barndorff–Nielsen (1980) suggests taking as an ancillary $\boldsymbol{A}$ perpendicular to $\boldsymbol{\eta}(\hat{\theta})$ in the metric defined locally by the covariance matrix of $\boldsymbol{T}$. See Fig. 21. The desired ancillary should have a distribution such that as $\hat{\boldsymbol{\theta}}$ becomes more and more precise, then $\boldsymbol{A}$ converges to a non-degenerate distribution. In fact, it can be made to have approximately zero mean and unit variance, if it is given by

$$\boldsymbol{A} = \boldsymbol{B}(\hat{\boldsymbol{\theta}})(\boldsymbol{T} - \boldsymbol{\tau}(\hat{\boldsymbol{\theta}})), \quad (227)$$

where $\boldsymbol{B}(\boldsymbol{\theta}) = [(\partial \boldsymbol{\tau}/\partial \boldsymbol{\theta})^{\perp\top}\boldsymbol{\Sigma}(\partial \boldsymbol{\tau}/\partial \boldsymbol{\theta})^{\perp}]^{-\frac{1}{2}}(\partial \boldsymbol{\tau}/\partial \boldsymbol{\theta})^{\perp\top}$. This will be proven below.

The matrix $(\partial \boldsymbol{\tau} / \partial \boldsymbol{\theta})^{\perp}$ is chosen such that

$$(\partial \boldsymbol{\tau}/\partial \boldsymbol{\theta})^{\perp}(\partial \boldsymbol{\tau}/\partial \boldsymbol{\theta})^{\top} = 0 \ \text{and} \ \det \left[(\partial \boldsymbol{\tau}/\partial \boldsymbol{\theta})^{\perp} \boldsymbol{\Sigma}(\partial \boldsymbol{\tau}/\partial \boldsymbol{\theta})^{\perp \top}\right] = \det \left[i\right] / \det \left[\boldsymbol{\Sigma}\right];$$

(228)

in other words $\partial \boldsymbol{\tau}/\partial \boldsymbol{\theta}$ is chosen to have rows orthogonal according to an inner product determined by $\boldsymbol{\Sigma}$. The second condition in (228) is equivalent to

$$\det \left[\left(\begin{array}{c} \partial \boldsymbol{\eta}/\partial \boldsymbol{\theta} \\ (\partial \boldsymbol{\tau}/\partial \boldsymbol{\theta})^{\perp} \end{array} \right) \right] = \det \left[i\right] / \det \left[\boldsymbol{\Sigma}\right],$$

as is seen by expanding the determinant for

$$\left(\begin{array}{c} \partial \boldsymbol{\eta}/\partial \boldsymbol{\theta} \\ (\partial \boldsymbol{\tau}/\partial \boldsymbol{\theta})^{\perp} \end{array} \right)^{\top} \boldsymbol{\Sigma} \left(\begin{array}{c} \partial \boldsymbol{\eta}/\partial \boldsymbol{\theta} \\ (\partial \boldsymbol{\tau}/\partial \boldsymbol{\theta})^{\perp} \end{array} \right),$$

and using (225) and the second expression for $i(\boldsymbol{\theta})$. That such a matrix can be constructed for each value of $\boldsymbol{\theta}$ is an elementary result from linear algebra; that $(\partial \boldsymbol{\tau}/\partial \boldsymbol{\theta})^{\perp}$ can be expressed as a differentiable function of $\boldsymbol{\theta}$ is a deeper question. Construction of such a matrix will be outlined in the exercises. The key idea here is that while the matrix $\boldsymbol{B}(\hat{\boldsymbol{\theta}})$ is the function of the random quantity $\hat{\boldsymbol{\theta}}$ and hence is random, it varies far less than $\boldsymbol{T} - \boldsymbol{\tau}(\hat{\boldsymbol{\theta}})$.

To make these notions precise, suppose that $\boldsymbol{T}_n$ arises as the sum of n independent and identically distributed observations from the curved exponential family (224) nested within a full exponential family. Let $\boldsymbol{\tau}_1(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}_1(\boldsymbol{\theta})$ be the mean and variance functions for each summand. Then $\boldsymbol{A}_n = \boldsymbol{B}(\hat{\boldsymbol{\theta}}_n)(\boldsymbol{T}_n - n\boldsymbol{\tau}_1(\hat{\boldsymbol{\theta}}_n))/\sqrt{n}$.

One might consider how far $\boldsymbol{A}_n$ is from being exactly ancillary. Note that

$$\boldsymbol{A}_n = \boldsymbol{B}(\boldsymbol{\theta})(\boldsymbol{T}_n - n\boldsymbol{\tau}_1(\boldsymbol{\theta}))/\sqrt{n} + (\boldsymbol{B}(\hat{\boldsymbol{\theta}}_n) - \boldsymbol{B}(\boldsymbol{\theta}))(\boldsymbol{T}_n - n\boldsymbol{\tau}_1(\boldsymbol{\theta}))/\sqrt{n}$$
$$+ \sqrt{n}(\boldsymbol{B}(\hat{\boldsymbol{\theta}}_n) - \boldsymbol{B}(\boldsymbol{\theta}))(\boldsymbol{\tau}(\boldsymbol{\theta}) - \boldsymbol{\tau}(\hat{\boldsymbol{\theta}}_n)) + \sqrt{n}\boldsymbol{B}(\boldsymbol{\theta})(\boldsymbol{\tau}_1(\boldsymbol{\theta}) - \boldsymbol{\tau}_1(\hat{\boldsymbol{\theta}}_n))].$$

The first term on the right has a distribution independent of $\boldsymbol{\theta}$, with zero mean and unit variance, to order $O_p(1/\sqrt{n})$. The remaining terms are all of size $O_p(1/\sqrt{n})$. Furthermore, under weak regularity conditions the moments of $\boldsymbol{A}_n$ will be differentiable functions of $\boldsymbol{\theta}$, and hence for changes in $\boldsymbol{\theta}$ of order $1/\sqrt{n}$ the mean, variance, and third cumulant will be constant to order $O(n^{-3/2})$, $O(n^{-2})$, and $O(n^{-5/2})$. Cox (1980) refers to this property as second order local ancillarity.

As will be proved below, the general formula for expansions of statistics conditional on the ancillary vector $\boldsymbol{A}$ of (227) is

$$f(\hat{\boldsymbol{\theta}}|\boldsymbol{a}; \boldsymbol{\theta}) = (2\pi)^{-\mathfrak{k}/2} \det \left[\hat{j}\right]^{-\frac{1}{2}} (\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{t})/\mathcal{L}(\hat{\boldsymbol{\theta}}; \boldsymbol{t}))\{1 + O(1/n)\}.$$

In the case of full exponential $(\mathfrak{k}, \mathfrak{k})$ families, with the canonical parameterization (224) with $\mathfrak{d} = \mathfrak{k}$ and $\boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{\theta}$, this result is proved above as (221). Here the ancillary $\boldsymbol{A}$ is of length 0, and no effective conditioning happens.

Example of trivial curved exponential family.
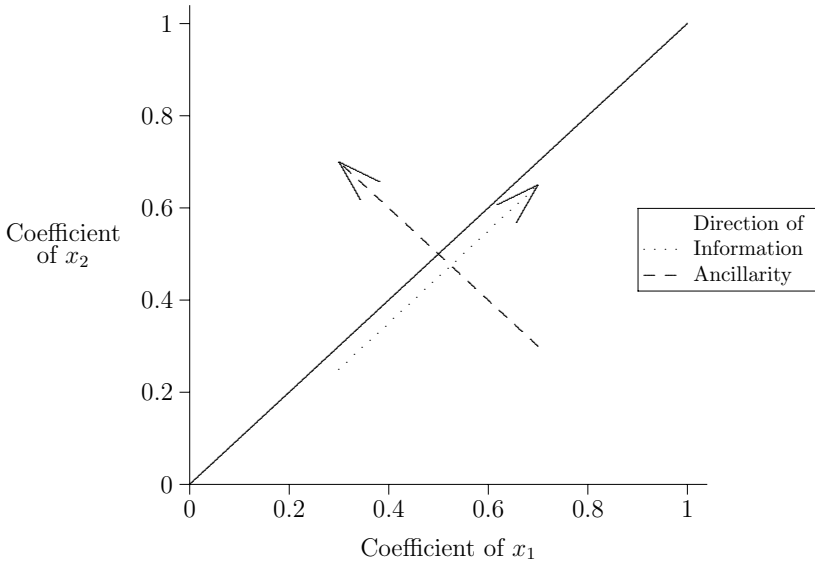
Fig. 21a.

Example of normal correlation curved exponential family.
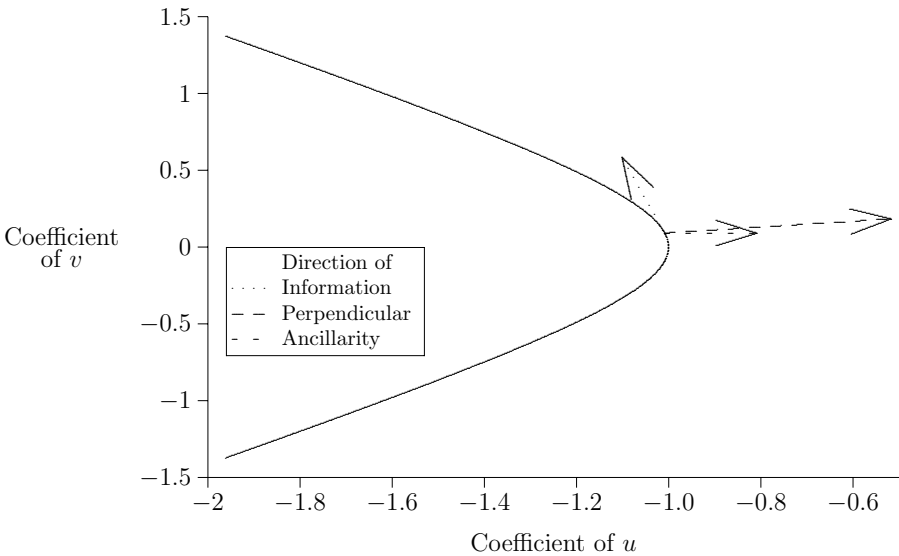
Fig. 21b.

When $\mathfrak{k} > \mathfrak{d}$ ancillarity really has teeth; conditioning on the ancillary is a requirement, not just an option, since a one-to-one map is needed to transform to the maximum likelihood estimator, and an approximate ancillary is the logical choice.

As an example of a curved exponential family, consider inference on a normal correlation presented by Barndorff–Nielsen (1980). Suppose that $(X_i, U_i)^\top$ are independent and identically distributed with distribution $N\left(0, \begin{pmatrix} 1 & \theta \\ \theta & 1 \end{pmatrix}\right)$. Then $\boldsymbol{T} = (\frac{1}{2}\sum_i(X_i^2 + U_i^2), \sum_i X_i U_i)$ is sufficient, and

$$f_{X,U}(\boldsymbol{x}, \boldsymbol{u}; \theta) = (2\pi)^{-n/2}(1 - \theta^2)^{-n/2}\exp(\theta t_2 - t_1/(1 - \theta)).$$

Straight forward calculations show that

$$\mathrm{E}\left[\boldsymbol{T}\right] = n\begin{pmatrix} 1 \\ \theta \end{pmatrix}, \ \mathrm{Var}\left[\boldsymbol{T}\right] = n\begin{pmatrix} 1 + \theta^2 & 3\theta \\ 3\theta & 1 + 2\theta^2 \end{pmatrix}.$$

Hence $\partial\boldsymbol{\tau}/\partial\theta = (0, n)$, and so we can take $(\partial\boldsymbol{\tau}/\partial\theta)^\perp = (1, 0)$, yielding an ancillary statistic $\boldsymbol{A} = (1 + \theta^2)^{-1/2}(T^1 - n)/\sqrt{n}$ that is an affine function of T^1.

The ancillary $\boldsymbol{A}$ is not uniquely determined by (227), because of the lack of uniqueness in the definition of perpendicular compliments used above. This non-uniqueness is a general drawback of inference conditional on ancillary statistics. Furthermore, calculation of any particular choice of $\boldsymbol{A}$ might prove troublesome. These problems might be avoided by constructing a statistic $\boldsymbol{S}$ such that $(\boldsymbol{S}, \boldsymbol{A})$ are sufficient statistics and such that $\boldsymbol{S}$ and $\boldsymbol{A}$ are approximately independent at some parameter value of interest. McCullagh (1984) shows that such a $\boldsymbol{S}$ is uniquely defined in one-dimensional problems, and is easily calculated from the likelihood ratio statistic. This work results in a more general proof of Barndorff–Nielsen's formula. The construction and proof crucially uses multivariate tensor notation beyond the scope of this work. McCullagh (1987) provides details.

8.6. Barndorff–Nielsen's Formula

This section presents the argument of Barndorff–Nielsen (1980) that Barndorff–Nielsen's formula holds for certain approximately ancillary statistics. First a lemma:

Lemma 8.6.1:   *If $\boldsymbol{A}$ is constructed as in (227), then the Jacobian* $\det\left[\partial\boldsymbol{T}/\partial(\hat{\boldsymbol{\theta}}, \boldsymbol{A})\right]$ *is given by*

$$\det\left[j(\hat{\boldsymbol{\theta}})\right]\det\left[i(\hat{\boldsymbol{\theta}})\right]^{-1/2}\det\left[\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})\right]^{1/2}.$$

Proof: The maximum likelihood estimator $\hat{\boldsymbol{\theta}}(\boldsymbol{t})$ satisfies $\boldsymbol{t}^\top(\partial\boldsymbol{\eta}/\partial\boldsymbol{\theta}) - \mathcal{H}'(\hat{\boldsymbol{\theta}}) = 0$. Since $\mathcal{H}'(\boldsymbol{\theta}) = \mathcal{H}^{*'}(\boldsymbol{\eta}(\boldsymbol{\theta}))(\partial\boldsymbol{\eta}/\partial\boldsymbol{\theta}) = \boldsymbol{\tau}(\boldsymbol{\theta})(\partial\boldsymbol{\eta}/\partial\boldsymbol{\theta})$, the expression relating $\boldsymbol{T}$ to $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{A}$ is

$$(\boldsymbol{T} - \boldsymbol{\tau})^\top\left(\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\theta}}\ \frac{\partial\boldsymbol{\tau}}{\partial\boldsymbol{\theta}}^\perp\right)\begin{pmatrix} \boldsymbol{I} & 0 \\ 0 & [(\partial\boldsymbol{\tau}/\partial\boldsymbol{\theta})^{\perp\top}\boldsymbol{\Sigma}(\partial\boldsymbol{\tau}/\partial\boldsymbol{\theta})^\perp]^{-\frac{1}{2}} \end{pmatrix} = (0, \boldsymbol{A}^\top),$$

with all functions of $\boldsymbol{\theta}$ evaluated at $\hat{\boldsymbol{\theta}}$. The quantity $(\partial\boldsymbol{\tau}/\partial\boldsymbol{\theta})^{\perp}$ is the result of taking the derivative of the mean function $\boldsymbol{\tau}$, and constructing a matrix perpendicular to it. Letting $\boldsymbol{A}^* = \boldsymbol{A}[(\partial\boldsymbol{\tau}/\partial\boldsymbol{\theta})^{\perp\top}\boldsymbol{\Sigma}(\partial\boldsymbol{\tau}/\partial\boldsymbol{\theta})^{\perp}]^{\frac{1}{2}}$,

$$(\boldsymbol{T}-\boldsymbol{\tau})^{\top}\left(\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\theta}}\,\frac{\partial\boldsymbol{\tau}^{\perp}}{\partial\boldsymbol{\theta}}\right) = (\mathbf{o},\boldsymbol{A}^*).$$

Differentiating these relations with respect to $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{A}^*$, we find that

$$\left(\frac{\partial\boldsymbol{T}}{\partial(\hat{\boldsymbol{\theta}},\boldsymbol{A}^*)} - (\partial\boldsymbol{\tau}/\partial\boldsymbol{\theta}\quad\mathbf{o})\right)^{\top}\left(\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\theta}}\,\frac{\partial\boldsymbol{\tau}^{\perp}}{\partial\boldsymbol{\theta}}\right) + \sum_j (T^j-\tau^j)\begin{pmatrix}\frac{\partial^2\eta_j}{\partial\boldsymbol{\theta}^{\top}\partial\boldsymbol{\theta}} & * \\ \mathbf{o} & \mathbf{o}\end{pmatrix} = \begin{pmatrix}\mathbf{o} & \mathbf{o} \\ \mathbf{o} & \boldsymbol{I}\end{pmatrix}.$$

The quantity $*$ above is the result of differentiating row j of $(\partial\boldsymbol{\tau}/\partial\boldsymbol{\theta})^{\perp}$ with respect to $\boldsymbol{\theta}$. Then

$$\frac{\partial\boldsymbol{T}}{\partial(\hat{\boldsymbol{\theta}},\boldsymbol{A}^*)}^{\top}\left(\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\theta}}\,\frac{\partial\boldsymbol{\tau}^{\perp}}{\partial\boldsymbol{\theta}}\right) = \begin{pmatrix}\mathbf{o} & \mathbf{o} \\ \mathbf{o} & \boldsymbol{I}\end{pmatrix} + \begin{pmatrix}\frac{\partial\boldsymbol{\tau}}{\partial\boldsymbol{\theta}} & \mathbf{o}\end{pmatrix}^{\top}\left(\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\theta}}\,\frac{\partial\boldsymbol{\tau}^{\perp}}{\partial\boldsymbol{\theta}}\right) - \sum_j (T^j-\tau^j)\begin{pmatrix}\frac{\partial^2\eta_j}{\partial\boldsymbol{\theta}^{\top}\partial\boldsymbol{\theta}} & * \\ \mathbf{o} & \mathbf{o}\end{pmatrix}$$

$$= \begin{pmatrix}\mathbf{o} & \mathbf{o} \\ \mathbf{o} & \boldsymbol{I}\end{pmatrix} + \begin{pmatrix}(\partial\boldsymbol{\tau}/\partial\boldsymbol{\theta})^{\top}(\partial\boldsymbol{\eta}/\partial\boldsymbol{\theta}) & * \\ \mathbf{o} & \mathbf{o}\end{pmatrix}^{\top} + \begin{pmatrix}j(\hat{\boldsymbol{\theta}})-i(\hat{\boldsymbol{\theta}}) & * \\ \mathbf{o} & \mathbf{o}\end{pmatrix}$$

$$= \begin{pmatrix}j(\hat{\boldsymbol{\theta}}) & * \\ \mathbf{o} & \boldsymbol{I}\end{pmatrix}, \tag{229}$$

by (226). Here $*$ marks parts of the matrix that can be ignored when taking its determinant since the corresponding lower block is $\mathbf{o}$. The determinant of (229) is $\det\left[j(\hat{\boldsymbol{\theta}})\right]$, and hence the determinant of $(\partial\boldsymbol{T}/\partial(\hat{\boldsymbol{\theta}},\boldsymbol{A}^*))$ is $\det\left[j(\hat{\boldsymbol{\theta}})\right]\det\left[\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})\right] \times \det\left[i(\hat{\boldsymbol{\theta}})\right]^{-1}$. The Jacobian of the transformation from $\boldsymbol{A}^*$ to $\boldsymbol{A}$ is $\det\left[i(\hat{\boldsymbol{\theta}})\right]^{1/2} \times \det\left[\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})\right]^{-1/2}$, and hence the Jacobian of the transformation from $\boldsymbol{t}$ to $(\hat{\boldsymbol{\theta}},\boldsymbol{A})$ is $\det\left[j(\hat{\boldsymbol{\theta}})\right]\det\left[i(\hat{\boldsymbol{\theta}})\right]^{-1/2}\det\left[\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})\right]^{1/2}$.

Q.E.D

The next lemma expresses the exact joint distribution of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ and an associated statistic $\boldsymbol{U}$ such that $(\hat{\boldsymbol{\theta}},\boldsymbol{U})$ are minimally sufficient, as the product of Barndorff–Nielsen's formula and an error term.

Lemma 8.6.2: *Suppose that $(\hat{\boldsymbol{\theta}},\boldsymbol{U})$ are minimally sufficient for the model (224), and $\boldsymbol{A}$ is a function of $\boldsymbol{U}$, possibly depending on $\hat{\boldsymbol{\theta}}$ as well. Let*

$$H(\hat{\boldsymbol{\theta}},\boldsymbol{U}) = \det\left[\partial\boldsymbol{T}/\partial(\hat{\boldsymbol{\theta}},\boldsymbol{U})\right]\sqrt{(2\pi)^{\eth}/\det\left[j(\hat{\boldsymbol{\theta}})\right]}\exp(\boldsymbol{t}^{\top}\boldsymbol{\eta}(\hat{\boldsymbol{\theta}})-\mathcal{H}(\hat{\boldsymbol{\theta}})-\mathcal{G}(\boldsymbol{t}))$$

Then $f_{\hat{\boldsymbol{\Theta}},A}(\hat{\boldsymbol{\theta}},\boldsymbol{a};\boldsymbol{\theta}) = (2\pi)^{-\eth/2}[H(\hat{\boldsymbol{\theta}},\boldsymbol{U})/(\partial\boldsymbol{U}/\partial\boldsymbol{A})](\hat{\boldsymbol{\theta}},\boldsymbol{a})\det\left[j(\hat{\boldsymbol{\theta}})\right]^{1/2}\mathcal{L}(\boldsymbol{\theta};\boldsymbol{t})/\mathcal{L}(\hat{\boldsymbol{\theta}};\boldsymbol{t})$.

Proof: This holds from a straight-forward application of the usual change of variables theorem.

<div align="right">Q.E.D</div>

Applying this to the affine ancillary calculated earlier:

Theorem 8.6.3:   *If $\boldsymbol{T}$ arises as the mean of n independent and identically distributed observations from the curved exponential family (224) nested within a full exponential family, and if $\boldsymbol{A}$ is constructed as in (227), then the expansion holds:*

$$f_{\hat{\Theta}|A}(\hat{\boldsymbol{\theta}}|\boldsymbol{a};\boldsymbol{\theta}) = (2\pi)^{-\eth/2} \det\left[j(\hat{\boldsymbol{\theta}})\right]^{1/2} \mathcal{L}(\boldsymbol{\theta};\boldsymbol{t})/\mathcal{L}(\hat{\boldsymbol{\theta}};\boldsymbol{t})(1+O_p(n^{-1})). \qquad (230)$$

Proof: By Lemma 8.6.2, with $\boldsymbol{A}$ constructed as before and $\boldsymbol{U} = \boldsymbol{A}$, it suffices to show that $H(\hat{\boldsymbol{\theta}}, \boldsymbol{A}) = 1 + O_p(n^{-1})$. The quantity $\mathcal{G}(\boldsymbol{t})$ is approximated by reference to the saddlepoint expansion for the full exponential family model $f_{\boldsymbol{T}}(\boldsymbol{t};\boldsymbol{\eta})$, which shows

$$\exp(\boldsymbol{t}^{\top}\boldsymbol{\eta} - \mathcal{H}^*(\boldsymbol{\eta}) - \mathcal{G}(\boldsymbol{t})) = c\det\left[\mathcal{H}^{*\prime\prime}(\tilde{\boldsymbol{\eta}})\right]^{-1/2}\exp(\boldsymbol{t}^{\top}(\boldsymbol{\eta}-\tilde{\boldsymbol{\eta}}) - \mathcal{H}^*(\boldsymbol{\eta}) + \mathcal{H}^*(\tilde{\boldsymbol{\eta}}))\times$$
$$\left\{1 + b(\tilde{\boldsymbol{\eta}})/[2n] + O(n^{-2})\right\}$$

where $\tilde{\boldsymbol{\eta}}$ satisfies $\mathcal{H}^{*\prime}(\tilde{\boldsymbol{\eta}}) = \boldsymbol{t}$. Then

$$\exp(-\mathcal{G}(\boldsymbol{t})) = c\det\left[\mathcal{H}^{*\prime\prime}(\tilde{\boldsymbol{\eta}})\right]^{-\frac{1}{2}}\exp(-\boldsymbol{t}^{\top}\tilde{\boldsymbol{\eta}} + \mathcal{H}^*(\tilde{\boldsymbol{\eta}}))\{1 + b(\tilde{\boldsymbol{\eta}})/[2n] + O(n^{-2})\}.$$

Hence by Lemma 8.6.1,

$$H(\hat{\boldsymbol{\theta}}, \boldsymbol{A}) = \left[\frac{\det\left[\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})\right]\det\left[j(\hat{\boldsymbol{\theta}})\right]}{\det\left[\mathcal{H}^{*\prime\prime}(\tilde{\boldsymbol{\eta}})\right]\det\left[i(\hat{\boldsymbol{\theta}})\right]}\right]^{\frac{1}{2}}\exp(\boldsymbol{t}^{\top}\boldsymbol{\eta}(\hat{\boldsymbol{\theta}}) - \mathcal{H}(\hat{\boldsymbol{\theta}}) - \boldsymbol{t}^{\top}\tilde{\boldsymbol{\eta}} + \mathcal{H}^*(\tilde{\boldsymbol{\eta}}))\times$$
$$\left\{1 + b(\tilde{\boldsymbol{\eta}})/[2n] + O(n^{-2})\right\} \qquad (231)$$

If $\boldsymbol{T}$ is the mean of independent and identically distributed random vectors from (224), then the matrix $\boldsymbol{\Sigma}$ in the definition of $\boldsymbol{B}$ has a factor of n^{-1}, and the functions $(\partial\boldsymbol{\tau}/\partial\boldsymbol{\theta})^{\perp}$ are unchanged. Then $\boldsymbol{T}$ can be expressed as an invertible function of $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{A}/\sqrt{n}$ near $(\boldsymbol{\theta}, \mathbf{o})$ with no other dependence on n. Since $\sqrt{n}(\boldsymbol{T} - \boldsymbol{\tau})$ converges in distribution, or in other words $\boldsymbol{T} - \boldsymbol{\tau} = O_p(1/\sqrt{n})$, then $\boldsymbol{A} = O_p(1)$. Expanding the quantity in large brackets in (231) about $(\hat{\boldsymbol{\theta}}, \mathbf{o})$ we determine an expansion of the form $1 + n^{-1/2}\sum_l A^l c_l(\hat{\boldsymbol{\theta}}, \mathbf{o}) + O(\|\boldsymbol{A}\|^2/n)$, where c_l are continuous in both arguments, and $\boldsymbol{A} = (A^1, \dots, A^{\mathfrak{k}-\eth})$. Since $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} + O_p(1/\sqrt{n})$, as will be shown in the section on the Bartlett's correction, $\hat{\boldsymbol{\theta}}$ can be replaced by $\boldsymbol{\theta}$, incurring an error no larger than $O_p(1/n)$. A similar treatment shows that the final factor in large brackets is of the form $c(\boldsymbol{A}) + O_p(1/n)$.

<div align="right">Q.E.D</div>

Formula (230) is known as Barndorff–Nielsen's formula. One might ask whether this formula might hold exactly.   Consider the case of $(\mathfrak{k}, \eth)$ curved exponential

families in which $\mathfrak{d} = \mathfrak{k} - 1$. Barndorff–Nielsen (1984) explores the applicability of this formula for other ancillary statistics, and notes that exactness holds for approximate ancillary statistics $\boldsymbol{V}$ satisfying $H(\hat{\boldsymbol{\theta}}, \boldsymbol{V}) = \partial \boldsymbol{U}/\partial \boldsymbol{V}$, and presents an example in which this differential equation can be solved to yield an ancillary for which Barndorff–Nielsen's formula holds exactly.

More generally when $\mathfrak{d} = \mathfrak{k} - 1$ Barndorff–Nielsen (1984) constructs an ancillary for which this formula holds to $O(n^{-3/2})$. This might be constructed in the above proof by multiplying the factors containing the errors, $\{1 + n^{-1/2} \sum_l A^l c_l(\hat{\boldsymbol{\theta}}, \mathbf{o}) + O(\|A\|^2/n)\}\{1 + b(\tilde{\boldsymbol{\eta}})/[2n] + O(n^{-2})\}$, extracting the next order term in n, and constructing an ancillary that sets this term to zero. This may be achieved as a linear combination of components of components of $\boldsymbol{A}$, times a scalar function of $\boldsymbol{A}$, because of the presence of the term $\sum_l A^l c_l(\hat{\boldsymbol{\theta}}, \mathbf{o})$ in the error. Barndorff–Nielsen (1984) suggests using a directed log likelihood ratio statistic formed by multiplying the log likelihood ratio statistic by the unit vector in the direction of the affine ancillary. Specifically, let $\hat{\Omega} = \sqrt{2n[\boldsymbol{T}^\top(\tilde{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}}) - \mathcal{H}(\hat{\boldsymbol{\theta}}) + \mathcal{H}^*(\tilde{\boldsymbol{\eta}})]}$, with the sign given by the sign of $\hat{\eta}_1$. Define the directed log likelihood ratio statistic to be

$$\left|\hat{\Omega}\right| \boldsymbol{A}/\|\boldsymbol{A}\|. \tag{232}$$

The appropriateness in the case of a one-dimensional affine ancillary is demonstrated in the following theorem. In this case a careful analysis of the above-mentioned linear combination is avoided by transforming to the density of the maximum likelihood estimator and the directed log likelihood ratio statistic immediately, and exploiting the occurrence of the likelihood ratio statistic in the exponent of the term in brackets in (231).

Theorem 8.6.4: *If $\boldsymbol{T}$ arises as the sum of n independent and identically distributed observations from the $(\mathfrak{k}, \mathfrak{k}-1)$ curved exponential family (224) nested within a full exponential family, and $\hat{\Omega}$ is given as above, then there exists an additive correction to $\hat{\Omega}$ depending on $\hat{\boldsymbol{\theta}}$ yielding $\hat{\Omega}^\dagger$ such that $\hat{\Omega}^\dagger$ has a marginal normal density to $O(n^{-1})$, and such that*

$$f_{\hat{\boldsymbol{\Theta}}|\hat{\Omega}^\dagger}(\hat{\boldsymbol{\theta}}|\hat{\omega}^\dagger; \boldsymbol{\theta}) = (2\pi)^{-\mathfrak{d}/2} \det\left[j(\hat{\boldsymbol{\theta}})\right]^{1/2} \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{t})/\mathcal{L}(\hat{\boldsymbol{\theta}}; \boldsymbol{t}) + O_p(n^{-3/2}).$$

Proof: In this theorem A is scalar, since the parameter vector has one fewer component than has the sufficient statistic. Again this theorem will be verified by examining $H(\hat{\boldsymbol{\theta}}, \hat{\Omega}/\sqrt{n})$. Then $\partial(\hat{\Omega}/\sqrt{n})/\partial \boldsymbol{T} = [\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}(\hat{\boldsymbol{\theta}})]/(\hat{\Omega}/\sqrt{n})$ when $\hat{\Omega} \neq 0$. Since $\hat{\Omega} \neq 0$ and $A \neq 0$ if $\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}(\hat{\boldsymbol{\theta}}) \neq 0$, and since $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}(\hat{\boldsymbol{\theta}})$ are differentiable functions of $\boldsymbol{T}$, the derivative is defined even when $\hat{\Omega} = 0$. Furthermore, as a consequence of (229),

$$\partial(\hat{\boldsymbol{\theta}}, A)/\partial \boldsymbol{T} = \begin{pmatrix} j^{-1}(\hat{\boldsymbol{\theta}}) & \mathbf{o} \\ * & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \dfrac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\theta}} & \dfrac{\partial \boldsymbol{\tau}^\perp}{\partial \boldsymbol{\theta}} \end{pmatrix}^\top,$$

and hence $\partial \hat{\boldsymbol{\theta}}/\partial \boldsymbol{T} = j^{-1}(\hat{\boldsymbol{\theta}})(\partial \boldsymbol{\eta}/\partial \boldsymbol{\theta})^\top$. Here i and j are expected and observed information for a sample of size 1, evaluated at $\boldsymbol{T}$. Therefore

$$\partial(\hat{\boldsymbol{\theta}}, \hat{\Omega}/\sqrt{n})/\partial \boldsymbol{T} = \begin{pmatrix} j^{-1}(\hat{\boldsymbol{\theta}})(\partial \boldsymbol{\eta}/\partial \boldsymbol{\theta})^\top \\ [\tilde{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}}]^\top/(\hat{\Omega}/\sqrt{n}) \end{pmatrix} = \begin{pmatrix} j(\hat{\boldsymbol{\theta}}) & 0 \\ 0 & (\hat{\Omega}/\sqrt{n}) \end{pmatrix}^{-1} \begin{pmatrix} (\partial \boldsymbol{\eta}/\partial \boldsymbol{\theta})^\top \\ [\tilde{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}}]^\top \end{pmatrix},$$

and

$$H(\hat{\boldsymbol{\theta}}, (\hat{\Omega}/\sqrt{n})) = \det\left[\begin{pmatrix} (\partial \boldsymbol{\eta}/\partial \boldsymbol{\theta})^\top \\ \sqrt{n}(\tilde{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}})^\top/\hat{\Omega} \end{pmatrix} \right]^{-1} \det\left[j(\hat{\boldsymbol{\theta}}) \right]^{1/2} \phi(\hat{\Omega}) \left\{ 1 + \frac{b(\tilde{\boldsymbol{\eta}})}{2n} + O(n^{-2}) \right\}.$$

The ratio $(\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}(\hat{\boldsymbol{\theta}}))/(\hat{\Omega}/\sqrt{n})$ must be expanded as a function of $\hat{\Omega}/\sqrt{n}$ and $\hat{\boldsymbol{\theta}}$. Only the lead term will be used. Differentiability of the function will first be demonstrated, and need only be demonstrated at $\hat{\Omega} = 0$. Since $\boldsymbol{T} = \mathcal{H}^{*\prime}(\tilde{\boldsymbol{\eta}})$, $\boldsymbol{T}$ is a differentiable function of $\tilde{\boldsymbol{\eta}}$, having a differentiable inverse. Hence $\hat{\boldsymbol{\eta}}$ is a differentiable function of $\tilde{\boldsymbol{\eta}}$. Since

$$\hat{\Omega}^2/n = -\mathcal{H}^{*\prime}(\tilde{\boldsymbol{\eta}})^\top(\hat{\boldsymbol{\eta}} - \tilde{\boldsymbol{\eta}}) - \mathcal{H}^*(\tilde{\boldsymbol{\eta}}) - \mathcal{H}^*(\hat{\boldsymbol{\eta}}) = \tfrac{1}{2}(\tilde{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}})^\top \mathcal{H}^{*\prime\prime}(\hat{\boldsymbol{\eta}})(\tilde{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}})$$
$$+ O(\|\tilde{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}}\|^3),$$

then $\hat{\Omega}/\sqrt{n}$ has a power series expansion in $\tilde{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}}$ near any $\hat{\boldsymbol{\eta}} \in \boldsymbol{\eta}(\mathfrak{H})$, and when $\hat{\boldsymbol{\eta}} = \tilde{\boldsymbol{\eta}}$ then $\hat{\Omega} = 0$. Hence $\det\left[\partial \boldsymbol{T}/\partial(\hat{\boldsymbol{\theta}}, \hat{\Omega}/\sqrt{n}) \right]$ has a power series representation as $\hat{\Omega} \to 0$. The lead term in this series will now be calculated. Choose any $\hat{\boldsymbol{\theta}} \in \mathfrak{H}$. By the definition of $\hat{\boldsymbol{\eta}}$, $[\mathcal{H}^{*\prime}(\tilde{\boldsymbol{\eta}}) - \mathcal{H}^{*\prime}(\hat{\boldsymbol{\eta}})](d\boldsymbol{\eta}/d\boldsymbol{\theta}) = 0$, and hence $(\tilde{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}})^\top \mathcal{H}^{*\prime\prime}(\hat{\boldsymbol{\eta}})(d\boldsymbol{\eta}/d\boldsymbol{\theta}) = O(\|\tilde{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}}\|^2)$. Then

$$\det\left[\begin{pmatrix} \left(\dfrac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\theta}}\right)^\top \\ [\tilde{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}}]^\top \end{pmatrix} \boldsymbol{\Sigma} \begin{pmatrix} \dfrac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\theta}} & \tilde{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}} \end{pmatrix} \right] = \det\left[\begin{pmatrix} i(\hat{\boldsymbol{\theta}}) & O(\|\tilde{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}}\|^2) \\ O(\|\tilde{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}}\|^2) & \dfrac{\hat{\Omega}^2}{n}(1 + O(\|\tilde{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}}\|)) \end{pmatrix} \right],$$

and at $\hat{\Omega} = 0$,

$$\det\left[\begin{pmatrix} (\partial \boldsymbol{\eta}/\partial \boldsymbol{\theta})^\top \\ (\tilde{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}})^\top/(\hat{\Omega}/\sqrt{n}) \end{pmatrix} \right]^{-1} = \det[\boldsymbol{\Sigma}]^{1/2}/\det\left[i(\hat{\boldsymbol{\theta}}) \right]^{1/2},$$

so from (231), $H(\hat{\boldsymbol{\theta}}, \hat{\Omega}/\sqrt{n}) = \phi(\hat{\Omega})\left(1 + c_1(\hat{\boldsymbol{\theta}})\hat{\Omega}/\sqrt{n} + c_2(\hat{\boldsymbol{\theta}})\hat{\Omega}^2/n + O(1/n\sqrt{n})\right)$, for some functions $c_1(\hat{\boldsymbol{\theta}})$ and $c_2(\hat{\boldsymbol{\theta}})$. Letting $\hat{\Omega}^\dagger = \hat{\Omega} - c_1(\hat{\boldsymbol{\theta}})/\sqrt{n}$, and applying Lemma 8.6.2, there exists a function $c_3(\hat{\boldsymbol{\theta}})$ such that

$$f_{\hat{\boldsymbol{\Theta}}, \hat{\Omega}^\dagger}(\hat{\boldsymbol{\theta}}, \hat{\omega}^\dagger; \boldsymbol{\theta}) = (2\pi)^{-\partial/2}[\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{t})/\mathcal{L}(\hat{\boldsymbol{\theta}}; \boldsymbol{t})]\phi(\hat{\omega}^\dagger)\{1 + c_3(\hat{\boldsymbol{\theta}}, \hat{\omega}^\dagger/\sqrt{n})/n + O(n^{-3/2})\}.$$

The additional cross terms added into the exponent of $\phi(\hat{\omega}^\dagger)$ is exactly that necessary to remove the term $c_1(\hat{\boldsymbol{\theta}})\hat{\omega}/\sqrt{n}$. The theorem is completed by noting that $c_3(\hat{\boldsymbol{\theta}}, \hat{\omega}^\dagger/\sqrt{n}) = c_3(\boldsymbol{\theta}, \hat{\omega}^\dagger/\sqrt{n}) + O(1/\sqrt{n})$.

Q.E.D

When $\mathfrak{d} < \mathfrak{k} - 1$ this construction may be applied sequentially to produce a multivariate ancillary with the same properties.

Barndorff–Nielsen (1986) applies these techniques to general statistical models represented by submanifolds of larger models. Barndorff–Nielsen (1990b) derives conditional cumulative distribution function approximations for the maximum likelihood estimator. Approximate integration techniques like those of §5.7 are used to integrate (230). The ancillary to be conditioned on is the directed likelihood ratio (232) for testing the the adequacy of (224) within the ambient full exponential family. The result is an approximation similar to (109):

Corollary 8.6.5: *Suppose that the conditions of Theorem 8.6.3 hold, and suppose that $\mathfrak{k} = 1$; that is, suppose that $\boldsymbol{\theta}$ is scalar. Let $\bar{\ell}(\theta; \hat{\theta}, \boldsymbol{a}) = \log(\mathcal{L}(\theta; \hat{\theta}, \boldsymbol{a}))/n$. Then $\bar{F}_{\hat{\boldsymbol{\Theta}}|A}(\hat{\boldsymbol{\theta}}|\boldsymbol{a}; \boldsymbol{\theta}) = [\bar{\Phi}(\sqrt{n}\hat{\omega}) + \phi(\sqrt{n}\hat{\omega})[1/\check{z} - 1/\hat{\omega}]/\sqrt{n}][1 + O_p(n^{-1})]$, with $\check{z} = [\bar{\ell}^{;1}(\hat{\theta}; \hat{\theta}, \boldsymbol{a}) - \bar{\ell}^{;1}(\theta; \hat{\theta}, \boldsymbol{a})]/\sqrt{j(\hat{\theta})}$, and the superscripts $;1$ on $\bar{\ell}^{;1}$ represent differentiation of the likelihood with respect to $\hat{\theta}$, after expressing $\boldsymbol{t}$ in terms of $\hat{\theta}$ and $\boldsymbol{a}$.*

Proof: Let $g(\hat{\theta}) = \bar{\ell}(\hat{\theta}; \hat{\theta}, \boldsymbol{a}) - \bar{\ell}(\theta; \hat{\theta}, \boldsymbol{a})$, and let $G(\hat{\theta}) = \left|j(\hat{\theta})\right|^{1/2}$. The data represented in $\bar{\ell}$ have been reexpressed in terms of $\hat{\theta}$ and $\boldsymbol{a}$. Clearly $g(\theta) = 0$. Also,

$$g'(\hat{\theta}) = \bar{\ell}^{;1}(\hat{\theta}; \hat{\theta}, \boldsymbol{a}) - \bar{\ell}^{;1}(\theta; \hat{\theta}, \boldsymbol{a}) + \bar{\ell}^{1;}(\hat{\theta}; \hat{\theta}, \boldsymbol{a}) = \bar{\ell}^{;1}(\hat{\theta}; \hat{\theta}, \boldsymbol{a}) - \bar{\ell}^{;1}(\theta; \hat{\theta}, \boldsymbol{a}),$$

where superscripts of the form $r;$ on $\bar{\ell}$ refer to derivatives with respect to θ_r and of form $;r$ refer to derivatives with respect to $\hat{\theta}_r$. In this scalar case, r may only take the value 1. Hence $g'(0) = 0$. Now differentiate $\bar{\ell}^{1;}(\hat{\theta}; \hat{\theta}, \boldsymbol{a}) = 0$ with respect to $\hat{\theta}$ to obtain $\bar{\ell}^{1;1}(\hat{\theta}; \hat{\theta}, \boldsymbol{a}) = -\bar{\ell}^{11;}(\hat{\theta}; \hat{\theta}, \boldsymbol{a})$. Furthermore,

$$g''(\hat{\theta}) = \bar{\ell}^{;1;1}(\hat{\theta}; \hat{\theta}, \boldsymbol{a}) + \bar{\ell}^{11}(\hat{\theta}; \hat{\theta}, \boldsymbol{a}) - \bar{\ell}^{11}(\theta; \hat{\theta}, \boldsymbol{a}) = j(\hat{\theta}) + \bar{\ell}^{11}(\hat{\theta}; \hat{\theta}, \boldsymbol{a}) - \bar{\ell}^{11}(\theta; \hat{\theta}, \boldsymbol{a}),$$

Hence $g''(\theta) = j(\theta)$. For $\hat{\theta}$ sufficiently close to θ, $g''(\hat{\theta}) > 0$. Apply Corollary 5.5.2.

Q.E.D

See §5.6 for an alternative rescaled signed root of the likelihood ratio statistic, derived from the r^* approximation.

8.7. Transformation Families

Consider the example presented by Reid (1988), citing Fisher (1934), in which $X_1, \ldots, X_n$ are generated independently with density $f((x - \mu)/\sigma)/\sigma$; here f is presumed known but the location and scale parameters μ and σ are to be estimated. The family of distributions generated as the unknown parameters are allowed to vary is not necessarily an exponential family. Such families are in general known as transformation families (Barndorff–Nielsen, 1978, Barndorff–Nielsen, 1983). In the case of transformation families Barndorff–Nielsen's formula also holds, as this example will illustrate. Once $\hat{\mu}$ and $\hat{\sigma}$ are determined, $\boldsymbol{A} = \left((X_{(1)} - \hat{\mu})/\hat{\sigma}, \ldots, (X_{(n)} - \hat{\mu})/\hat{\sigma}\right)$

are treated as ancillary. The transformation $(\hat{\mu}, \hat{\sigma}, a^1, ..., a^{t-2}) \to \boldsymbol{x}$ has a Jacobian of the form $\hat{\sigma}^{n-2}$ times a function of $\boldsymbol{a}$. Hence the joint density of $(\hat{\mu}, \hat{\sigma}, A_1, ..., A_{t-2})$ is $g(\boldsymbol{a})\hat{\sigma}^{n-2}\sigma^{-n} \prod_i f((a^i\hat{\sigma} + \hat{\mu} - \mu)/\sigma)$. Integrating with respect to $\hat{\mu}$ removes dependence on μ and also removes one factor of σ; integrating with respect to $\hat{\sigma}$ removes the remaining factors of σ, resulting in a distribution of $\boldsymbol{A}$ free of unknown parameters and justifying treating $\boldsymbol{A}$ as ancillary. Since $\mathcal{L}(\hat{\mu}, \hat{\sigma}; \boldsymbol{x}) = \hat{\sigma}^{-n} \prod_i f(a^i)$, and $\mathcal{L}(\mu, \sigma; \boldsymbol{x}) = \prod_i f((a^i\hat{\sigma} + \hat{\mu} - \mu)/\sigma)$, then the density of $(\hat{\mu}, \hat{\sigma})$ conditional on $\boldsymbol{A}$ can be written as $c(\boldsymbol{a}) \det [j(\hat{\mu}, \hat{\sigma})]^{-1} \mathcal{L}(\mu, \sigma; \boldsymbol{x})/\mathcal{L}(\hat{\mu}, \hat{\sigma}; \boldsymbol{x})$, thus verifying Barndorff–Nielsen's formula in this case.

The data X_i have the cumulative distribution function $F(g_{\mu,\sigma}(x))$, with F known, the function $g_{\mu,\sigma}(x) = (x - \mu)/\sigma$, and the parameters μ and σ unknown. Define an operation $\circ$ on $\mathfrak{G} = \{g_{\mu,\sigma} : \mu \in \mathfrak{R}, \sigma > 0\}$ to be composition. Then

$$(g_{\mu_2,\sigma_2} \circ g_{\mu_1,\sigma_1})(x) = g_{\mu_2,\sigma_2}((x - \mu_1)/\sigma_1) = (x - \mu_1 - \sigma_1\mu_2)/\sigma_1\sigma_2 = g_{\mu_1+\sigma_1\mu_2,\sigma_1\sigma_2}(x)$$

Hence $\mathfrak{G}$ has the following 4 properties.

a. $g_{\mu_2,\sigma_2} \circ g_{\mu_1,\sigma_1} \in \mathfrak{G}$.
b. $g_{0,1} \circ g_{\mu,\sigma} = g_{\mu,\sigma}$.
c. $g_{-\mu/\sigma,1/\sigma} \circ g_{\mu,\sigma} = g_{0,1}$.
d. $(g_{\mu_3,\sigma_3} \circ g_{\mu_2,\sigma_2}) \circ g_{\mu_1,\sigma_1} = g_{\mu_3,\sigma_3} \circ (g_{\mu_2,\sigma_2} \circ g_{\mu_1,\sigma_1})$.

A structure $(\mathfrak{G}, \circ)$ with these four properties is called a group; a group which is a set of functions on another space where the operation is composition is said to be a group acting on the other space. The general model, then, has X_i distributed independently with common cumulative distribution function $F(g^{-1}(x))$ for a known cumulative distribution function F on a sample space $\mathfrak{X}$ and an unknown $g \in \mathfrak{G}$ for the group $\mathfrak{G}$ acting on $\mathfrak{X}$. Equivalently, $X_i = g(Y_i)$ with Y_i drawn from F. The objective is to draw inference on which element $g \in \mathfrak{G}$ is involved. Generally, the group $\mathfrak{G}$ and cumulative distribution function F are said to comprise a transformation family. Fraser (1968, 1979) provides a fuller exposition. Barndorff–Nielsen's formula for all such models is exact. If $F(g(x))$ is different for each distinct $g \in \mathfrak{G}$, or equivalently, if $g(Y_i)$ has a different distribution for each $g \in \mathfrak{G}$, and one can find $\hat{g} \in \mathfrak{G}$ maximizing the likelihood, then the ancillary quantity is $\boldsymbol{A}$ with components $A^i = \hat{g}^{-1}(X_i)$. Barndorff–Nielsen (1980) treats the more general case when $g(y)$ is a distinct function of y for each $g \in \mathfrak{G}$, and Barndorff–Nielsen (1983) treats the still more general case when this assumption is dropped.

8.8. Exercises

1. Consider the logistic regression model of question 2 of §7.9. Sketch contours for the density of the maximum likelihood estimator assuming the true regression parameters are the observed value of the maximum likelihood estimator.

2. Give an expression for the distribution of the maximum likelihood estimator in a Poisson regression model, similar to the above but with y a Poisson variable whose mean is the exponential of linear covariates.

9

Other Topics

This chapter contains miscellaneous material applying saddlepoint and other distribution function approximations to statistical inference. An approximation for the root of an estimating equation is presented. Series approximation methods in Bayesian inference and resampling will also be discussed.

9.1. Applications to Estimating Equations

Many statistical estimates are constructed by finding the root of a function depending on the value of a parameter and the unknown data. For instance, maximum likelihood estimators are found by equating the sum of the derivatives of the log likelihood contributions for the various items sampled to zero. Daniels (1983) provides saddlepoint approximations to the distributions of roots of estimating equations.

Suppose $\boldsymbol{X}_j$ are independent and identically distributed according to a distribution indexed by a scalar parameter θ, for $j = 1, ..., n$. Let $M(\boldsymbol{x}, \theta)$ be a function such that $\mathrm{E}\left[M(\boldsymbol{X}, \theta)\right] = 0$ for all θ, and such that M is decreasing in θ. Then estimate θ as the root $\hat{\theta}$ of $\sum_{j=1}^{n} M(\boldsymbol{X}_j, \hat{\theta}) = 0$. The estimator $\hat{\theta}$ is known as an M-estimate. Serfling (1980) discusses the asymptotic properties of these estimators. Under general conditions, one can, for instance, demonstrate asymptotic normality.

Define $\bar{M}_n(a) = \sum_{j=1}^{n} M(\boldsymbol{X}_j, a)/n$; then by the monotonicity of M,

$$\mathrm{P}\left[\hat{\theta} > a\right] = \mathrm{P}\left[\bar{M}(a) > 0\right],$$

and the problem of tail areas for $\hat{\theta}$ reduces to finding probabilities that $\bar{M}_n(a) > 0$ as a varies. The question of uniformity of error bounds is more difficult in this case than in the case of cumulative distribution function approximation, since as a varies the distribution whose cumulative distribution function at 0 is approximated changes. Uniformity properties in this case are not as widely studied as they are in the case of tail probabilities from one distribution.

Let $\mathcal{K}(\beta; a)$ be the cumulant generating function of $\bar{M}_1$. This may be difficult to calculate. Then

$$\mathrm{P}\left[\hat{\theta} > a\right] = \mathrm{P}\left[\bar{M} > 0\right] = \frac{1}{2\pi i} \int_C \exp(n\mathcal{K}(\beta; a))\beta^{-1} \, d\beta, \qquad (233)$$

where the path C is chosen to have a positive real part at each point. Taking the derivative with respect to a, the density of $\hat{\theta}$ is $-(n/2\pi i)\int_C \exp(n\mathcal{K}(\beta; a))\mathcal{K}_2(\beta; a) \times \beta^{-1}\, d\beta$. Here $\mathcal{K}_2(\beta; a)$ represents the derivative of $\mathcal{K}$ with respect to a. Since $\mathcal{K}(0; a) = 0$ for all a, the singularity at zero is removable. Approximating this integral using steepest descent methods gives the density approximation

$$-\sqrt{\frac{n}{2\pi\mathcal{K}_{11}(\hat{\beta}; a)}}\mathcal{K}_2(\hat{\beta}; a)\hat{\beta}^{-1}\exp(n\mathcal{K}(\hat{\beta}; a)). \tag{234}$$

Here $\mathcal{K}_{11}(\beta; a)$ represents the second derivative of $\mathcal{K}$ with respect to a. Alternatively, the Robinson and Lugannani and Rice approximations may be used directly on (233) to calculate tail probabilities.

These methods may be applied to maximum likelihood estimators, giving approximations to the unconditional distribution of the estimator; recall that the Barndorff–Nielsen's formula in general gives an approximation to distribution of the maximum likelihood estimator conditional on an approximate ancillary statistic.

As a second example, consider the problem of drawing inference on the ratio ρ of two means from the ratio of the observations. If X and Y are exponential random variables with means μ and $\mu\rho$, then the joint likelihood is $-X/\mu - Y/(\mu\rho) - 2\log(\mu) - \log(\rho)$. The maximum likelihood estimator $\hat{\rho}$ of ρ is Y/X, and hence satisfies $M(X, Y, \hat{\rho}) = Y - \hat{\rho}X = 0$; note that $\mathrm{E}\left[M(X, Y, \rho)\right] = 0$. Then $\mathcal{K}(\beta, \hat{\rho}) = -\log(1 + \hat{\rho}\beta\mu) - \log(1 - \rho\beta\mu)$. Partial derivatives with respect to β are

$$\mathcal{K}_1(\beta, \hat{\rho}) = \frac{\mu(\rho - \hat{\rho} + 2\rho\hat{\rho}\beta\mu)}{(1 + \hat{\rho}\beta\mu)(1 - \beta\mu\rho)},$$

and

$$\mathcal{K}_{11}(\beta, \hat{\rho}) = \frac{\mu^2\left(\rho^2 + 2\hat{\rho}\beta\mu\rho^2 + \hat{\rho}^2\left(1 - 2\beta\mu\rho + 2\beta^2\mu^2\rho^2\right)\right)}{(1 + \hat{\rho}\beta\mu)^2(1 - \beta\mu\rho)^2}.$$

Then $\hat{\beta} = (\hat{\rho} - \rho)/(2\hat{\rho}\mu\rho)$, and so $\mathcal{K}_{11}(\hat{\beta}, \hat{\rho}) = 8\hat{\rho}^2\mu^2\rho^2(\hat{\rho} + \rho)^{-2}$. Also, $\mathcal{K}_2(\beta, \hat{\rho}) = -\beta\mu/(1 + \hat{\rho}\beta\mu)$, and $\mathcal{K}_2(\hat{\beta}, \hat{\rho}) = (\rho - \hat{\rho})\hat{\rho}^{-1}(\hat{\rho} + \rho)^{-1}$. Substituting into (234), the approximation to the density of $\hat{\rho}$ is

$$2^{-1+2n}\left(\frac{\hat{\rho}\rho}{(\hat{\rho} + \rho)^2}\right)^n\sqrt{n}/(\hat{\rho}\sqrt{\pi})$$

Field and Ronchetti (1990) discuss saddlepoint approximations for estimating equations in detail and provide many additional references.

9.2. Applications to Bayesian Methods

Consider the Bayesian paradigm in which π is a prior distribution on a parameter space $\mathfrak{P} \subset \mathfrak{R}^k$. Data represented by the random vector $\boldsymbol{T}$ is generated according to a density $f_{\boldsymbol{T}}(\boldsymbol{t}, \boldsymbol{\theta})$ depending on $\boldsymbol{\theta} \in \mathfrak{P}$. Let $\ell(\boldsymbol{\theta}; \boldsymbol{t}) = \log(f_{\boldsymbol{T}}(\boldsymbol{t}, \boldsymbol{\theta}))$ be the corresponding likelihood. The posterior density for $\boldsymbol{\theta}$ conditional on $\boldsymbol{T}$ is then

$$\frac{\exp(\ell(\boldsymbol{\theta}; \boldsymbol{t}))\pi(\boldsymbol{\theta})}{\int_{\mathfrak{P}} \exp(\ell(\boldsymbol{\theta}; \boldsymbol{t}))\pi(\boldsymbol{\theta}) \ d\boldsymbol{\theta}}.$$

An approximation to posterior moments

$$\int_{\mathfrak{P}} g(\boldsymbol{\theta})\exp(\ell(\boldsymbol{\theta}; \boldsymbol{t}))\pi(\boldsymbol{\theta}) \ d\boldsymbol{\theta} \Big/ \int_{\mathfrak{P}} \exp(\ell(\boldsymbol{\theta}; \boldsymbol{t}))\pi(\boldsymbol{\theta}) \ d\boldsymbol{\theta} \qquad (235)$$

for a function $g(\boldsymbol{\theta})$ is desired. For example, g might represent the value of one component of $\boldsymbol{\theta}$ when a marginal posterior distribution is desired. Tierney and Kadane (1986) apply Laplace's method to the problem of approximating moments of a function g of the parameter. This involves approximating two integrals of the form (146), with the functions $\mathcal{B}(\theta) = -\log(\pi(\theta)) - \ell(\theta, \boldsymbol{T})$ and $\mathcal{B}^*(\theta) = -\log(g(\theta)) - \log(\pi(\theta)) - \ell(\theta, \boldsymbol{T})$. The function in (146) multiplying the exponential is taken to be uniformly 1. This moment is approximated as the ratio of the Laplace approximation (146) to the two integrals in (235). If $\hat{\theta}$ minimizes $\mathcal{B}$ and $\tilde{\theta}$ minimizes $\mathcal{B}^*$, the approximation is $\exp(\mathcal{B}(\hat{\theta}) - \mathcal{B}^*(\tilde{\theta}))\sqrt{\det\left[\mathcal{B}''(\hat{\theta})\right] / \det\left[\mathcal{B}^{*}{}''(\tilde{\theta})\right]}$.

Asymptotics come into play when the distribution of $\boldsymbol{T}$ depends on a parameter n indexing sample size, and when n increases; fortunately the Laplace approximation works well for moderate values of n. As n increases, ℓ_n becomes the dominant term in both $\mathcal{B}$ and $\mathcal{B}^*$. Formal asymptotic treatments of these integrals rest on the fact that $\ell_n(\theta, \boldsymbol{T}) = O_p(n)$ and can be expressed as $n\ell(\theta) + h_n(\theta) - \log(\pi(\theta))$. By the weak law of large numbers, this holds if the likelihood arises from independent and identically distributed random variables.

Using an analogue to the method of steepest descent, these methods can be demonstrated accurate to $O(n^{-2})$; see Kass, Tierney, and Kadane (1990).

When these methods are applied to deriving the marginal posterior of θ_1, the first component of $\boldsymbol{\theta}$, one proceeds as follows: First, fix a value for θ_1 at which to evaluate the posterior. Let $\tilde{\boldsymbol{\theta}}(\theta_1)$ solve

$$\ell'_n(\tilde{\boldsymbol{\theta}}) + \pi'(\tilde{\boldsymbol{\theta}})/\pi(\tilde{\boldsymbol{\theta}}) = 0 \qquad (236)$$

where differentiation is with respect to all but the first components of $\boldsymbol{\theta}$, and the first component is fixed. Let $\hat{\boldsymbol{\theta}}$ solve (236) for all first order derivatives without θ_1 fixed. The posterior approximation is the ratio of the appropriate integral approximation,

and can be expressed as

$$
\frac{(2\pi)^{-1/2} \det\left[\ell_n^{(-)''}(\tilde{\boldsymbol{\theta}}(\theta_1))\right]^{-1/2} \exp(\ell_n(\tilde{\boldsymbol{\theta}}(\theta_1)) + \log(\pi(\tilde{\boldsymbol{\theta}}(\theta_1))))}{\det\left[\ell_n''(\hat{\boldsymbol{\theta}})\right]^{-1/2} \exp(\ell_n(\hat{\boldsymbol{\theta}}) + \log(\pi(\hat{\boldsymbol{\theta}})))},
$$

where $\ell_n^{(-)''}$ denotes the Hessian matrix with entries corresponding to the first component removed. When simplified this yields

$$
(2\pi)^{-1/2}\left(\det\left[\ell_n''(\hat{\boldsymbol{\theta}})\right] / \det\left[\ell_n^{(-)''}(\tilde{\boldsymbol{\theta}}(\theta_1))\right]\right)^{1/2} \exp(\ell_n(\tilde{\boldsymbol{\theta}}(\theta_1)) - \ell_n(\hat{\boldsymbol{\theta}}))\frac{\pi(\tilde{\boldsymbol{\theta}}(\theta_1))}{\pi(\hat{\boldsymbol{\theta}})}. \quad (237)
$$

Note the parallels between this and the double saddlepoint conditional density approximation (180).

Davison (1986) uses similar methods to approximate the unmarginalized posterior

$$
\pi(\boldsymbol{\theta}|\boldsymbol{T}) = \frac{f_{\boldsymbol{T}}(\boldsymbol{T};\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f_{\boldsymbol{T}}(\boldsymbol{T};\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\theta}.
$$

Laplace's methods are again used to approximate the denominator as:

$$
(2\pi)^{k/2} \det\left[\ell_n''(\hat{\boldsymbol{\theta}})\right]^{-1/2} \exp(\ell_n(\hat{\boldsymbol{\theta}}) + \log(\pi(\hat{\boldsymbol{\theta}}))),
$$

where $\hat{\theta}$ solves (236), and ratio is approximated as

$$
(2\pi)^{-k/2} \det\left[\ell_n''(\hat{\boldsymbol{\theta}})\right]^{1/2} \exp(\ell_n(\boldsymbol{\theta}) - \ell_n(\hat{\boldsymbol{\theta}}))\frac{\pi(\boldsymbol{\theta})}{\pi(\hat{\boldsymbol{\theta}})}(1 + O(n^{-1})).
$$

Note the formal equivalence to Barndorff–Nielsen's formula (221); in this case, however, it is $\boldsymbol{\theta}$ and not $\hat{\boldsymbol{\theta}}$ that is random. Wong and Li (1992) apply asymptotic methods to the problem of generating marginal distributions for marginal distributions of a non–linear function of the parameter vector.

As an example, Tierney and Kadane (1986) consider the analysis of a survival model for the Stanford heart transplant data; following Turnbull, Brown, and Hu (1974), they divide patients into groups $i = 1$ if they receive a transplant and $i = 2$ if they receive no transplant and model the survival time without transplant T_{ij} for individual j in group i as exponential with parameter ϕ_{ij}, and for those in group 1, the survival times after replacement Z_j as exponential $\theta\phi_{ij}$. Of particular interest is the posterior distribution for θ. Tierney and Kadane (1986) apply a uniform prior to this data, and compare the results of the approximation (237) to the values given by 60 point Gaussian quadrature and the normal approximation. They find close agreement between (237) and quadrature, and substantial disagreement between these and the normal approximation.

9.3. Applications to Resampling Methods

Suppose that $\{X_i\}$ are independent and identically distributed with unknown cumulative distribution function F. One desires to estimate a characteristic θ of F; for example, $\theta(F)$ might be the mean $\int x \, dF$. Often times one uses as an estimate $\theta(\hat{F})$, the corresponding characteristic of $\hat{F}$, the empirical cumulative distribution function; $\hat{F}(x) = \frac{1}{n}\#\{X_i \leq x\}$. The cumulative distribution function of the difference of the resulting estimate from the true value is given by $G(y) = \mathrm{P}_F\left[\theta(\hat{F}) - \theta(F) \leq y\right]$ and inference on $\theta(F)$ would ideally be done using $G(y)$. Usually $G(y)$ is unavailable, and the estimated cumulative distribution function $\hat{G}(y) = \mathrm{P}_{\hat{F}}\left[\theta(\hat{F}) - \theta(F) \leq y\right]$ is used in its place. Calculating quantities depending on $\hat{G}(y)$ exactly is often impractical; integrals with respect to $d\,\hat{G}(y)$ are often approximated using sampling techniques. Davison and Hinkley (1988) suggest instead using saddlepoint methods. The cumulant generating function associated with $\hat{F}$ is $\hat{\mathcal{K}}(\beta) = \log(\sum_{i=1}^{n} \exp(\beta X_i)/n)$.

Confidence intervals for $\theta(F)$ might be constructed as follows: Ideally one desires statistics $U(\boldsymbol{X})$ and $L(\boldsymbol{X})$ such that

$$\mathrm{P}_F\left[\theta(\hat{F}) - \theta(F) \leq L(\boldsymbol{X})\right] = \alpha_1, \quad \mathrm{P}_F\left[\theta(\hat{F}) - \theta(F) \geq U(\boldsymbol{X})\right] = \alpha_2$$

for all F under consideration; this implies that $(\theta(\hat{F}) + L(\boldsymbol{X}), \theta(\hat{F}) + U(\boldsymbol{X}))$ forms a $1 - \alpha_1 - \alpha_2$ confidence interval. As an approximation to this method, one might use L and U such that

$$\mathrm{P}_{\hat{F}}\left[\theta(\hat{F}) - \theta(F) \leq L\right] = \alpha_1, \quad \mathrm{P}_{\hat{F}}\left[\theta(\hat{F}) - \theta(F) \geq U\right] = \alpha_2. \tag{238}$$

The construction (238) may exhibit poor coverage properties, since L and U are chosen for one possible cumulative distribution function, the empirical one, rather than for all distribution functions under consideration. One solution, proposed by Efron (1982) for estimating the mean, is to embed $\hat{F}$ in the exponential family $\hat{F}(x, \theta)$ putting mass $\exp(X_i \eta)/\sum_j \exp(X_j \eta)$ on X_i. Here $\eta(\theta)$ is chosen so that

$$\sum_j X_j \exp(X_j \eta(\theta))/\sum_j \exp(X_j \eta(\theta)) = \theta.$$

The resulting likelihood is called the exponential empirical likelihood; Jing and Wood (1996) discuss its properties with respect to Bartlett's correction. Here then, $\theta(\hat{F}) = \bar{X}$. Within this exponential family, the sample mean is sufficient for θ, and $1 - \alpha_1 - \alpha_2$ confidence intervals for θ can be taken as those θ such that

$$\mathrm{P}_{F(.,\theta)}\left[\theta(\hat{F}) \geq t\right] = \alpha_1, \quad \mathrm{P}_{F(.,\theta)}\left[\theta(\hat{F}) \leq t\right] = \alpha_2. \tag{239}$$

Here t is the observed characteristic value. One might calculate these probabilities under the distribution $\hat{F}(.; \theta)$ rather than F; Davison and Hinkley (1988) suggest

approximating these quantities using saddlepoint methods. The cumulant generating function associated with $\hat{F}(.;\theta)$ is

$$\hat{\mathcal{K}}(\beta;\theta) = \log(\sum \exp[(\beta + \eta(\theta))X_i]) - \log(\sum \exp[\eta(\theta)X_i])$$

and

$$\hat{\mathcal{K}}'(\beta;\theta) = \sum X_i \exp[(\beta + \eta(\theta))X_i] / \sum \exp[(\beta + \eta(\theta))X_i].$$

The saddlepoint $\hat{\beta}$ is defined by $\hat{\mathcal{K}}'(\hat{\beta};\theta) = \bar{X}$. Hence $\hat{\beta} + \eta(\theta) = 0$. The Lugannani and Rice approximation (109) can then be used to approximate tail areas and hence approximate L and U in (239). Tail areas then have the form $1 - \Phi(\hat{\omega}) + \phi(\hat{\omega})[1/\hat{z} - 1/\hat{\omega}]$; here $\hat{z} = -\eta(\theta)\sqrt{\sum(X_i - \bar{X})^2}$ and $\hat{\omega} = \sqrt{2n[\log(\sum \exp(X_i\eta(\theta)/n)) - \eta(\theta)\bar{X}]}$, with $\hat{\omega}$ given the same sign as $\hat{z}$.

A quantity more nearly pivotal may be created by removing the effect of properties of the underlying distribution of higher order than the mean is approximate studentization. Suppose H is the cumulative distribution function of $\theta(\hat{F}) - \theta(F)$. Then $H^{-1}(\theta(\hat{F}) - \theta(F))$ is distributed uniformly on the unit interval. Confidence intervals might then be generated by comparing $\hat{H}^{-1}(\theta(\hat{F}) - \theta(F))$ to a uniform distribution, where $\hat{H}$ denotes the cumulative distribution function of $\theta(\hat{F}) - \theta(F)$. Although in general H would be analytically intractable even if $\hat{F}$ were not, in practice both are intractable, and $\hat{H}$ might be approximated through a two-step sampling scheme known as double-bootstrapping. Davison and Hinkley (1988) suggest using saddlepoint methodology here too to avoid sampling. Theoretical justifications for much of this is found in Wang (1990a).

Similarly, Davison and Hinkley (1988) suggest using the empirical cumulant generating function $\hat{\mathcal{K}}$ to do non-parametric inference on M-estimates.

Related are non-parametric permutation tests. Suppose that $X_1, ..., X_n \sim F(x)$, and one wishes to test whether F is symmetric about θ_0. The test is usually performed conditional on the size of deviations from the proposed symmetry point. Let

$$a_i = (X_i - \theta_0)/\sqrt{\sum_j (X_i - \theta_0)^2} \text{ and } T = \sum a_i. \tag{240}$$

The distribution of T conditional on the absolute values of the a_i is now that of $\sum V_i |a_i|$, where the V_i take the values of positive and negative 1 with probability half each. Asymptotic techniques can then be used to construct tests and generate p-values. Construction of the cumulant generating function is straight forward.

Suppose that

$$X_1, ..., X_{n_B} \sim F(x - \theta_1)$$
$$X_{n_B+1}, ..., X_{n_B+n_A} \sim F(x - \theta_2)$$

and we wish to test whether $\theta_1 - \theta_2 = \delta_0$. Here set $a_k = (Y_k - \bar{Y})/\sqrt{\sum_{i=1}^{n_B+n_A}(Y_i - \bar{Y})^2}$ where $Y_k = X_k - \delta_0$ if $k \leq n_B$ and X_k otherwise. Inference is conditional on

the a_k; here the test statistic is $\sum_{k=1}^{n_B} a_k$ whose conditional distribution is that of $\sum_{k=1}^{n_B} a_{S_k}$ where S_j are n_B random selections without replacement from $1, ..., n_B+n_A$. Robinson (1982) gives an approximation to the cumulant generating function and shows that the tail approximation (98) works well here.

Here the cumulative distribution function has jumps of regular size, since the underlying distribution has atoms all equally likely. Unfortunately the spacing between atoms is non-uniform, and in general the limit distribution will be a singular distribution but not a lattice distribution. The most straight-forward continuity correction is half the distance to the neighboring atom. Because finding the next larger atom might be quite difficult, Robinson suggests an approximate continuity correction using the average atom separation and the constant size of cumulative distribution function jumps.

9.4. Applications to Efficiency

In this section I present expansions of the form (1) for efficiencies of a method of inference. Suppose that two statistical procedures are considered, and that k_n observations are required to give the second procedure the same precision as is realized with n observations from the first procedure; then the relative efficiency of these two procedures defined to be k_n/n, and the asymptotic relative efficiency and the asymptotic deficiency are defined to be the first–order and second–order terms respectively in the expansion of k_n/n:

$$k_n/n = \text{ARE} + \text{ADEF}/n + o(n^{-1}). \tag{241}$$

Here equality of precision may refer to equality of mean square errors of two estimates, or powers of two tests for a similar alternative. The first procedure is preferable if $\text{ARE} > 1$, and the second procedure is preferable if $\text{ARE} < 1$. Fisher (1925) considered discrimination between estimation procedures when the asymptotic relative efficiency is unity. In this case the first procedure is preferable if $\text{ADEF} > 0$, and the second procedure is preferable if $\text{ADEF} < 0$. Relation (241) implies that $\text{ADEF} = \lim_{n\to\infty} k_n - n$. Hodges and Lehmann (1970) define the deficiency of the two procedures to be $k_n - n$; the asymptotic deficiency of the second procedure relative to the first is then ADEF, when this limit exists. For further details and references, see Kolassa (1996c). Consider assessing an asymptotically normal and asymptotically unbiased estimator of a parameter. Take as a definition of the efficiency of this estimator to be the ratio of the Cramér–Rao lower bound for its variance to the actual achieved variance. Kolassa (1996c) provides alternate definitions and examples. Estimates with efficiency closer to one are preferable to those with lower efficiency. Estimators whose asymptotic efficiency is unity are called first–order efficient.

One might also define the efficiency of one estimator relative to another to be the inverse of the ratio of their variances, and the asymptotic relative efficiency to be the limit of this inverse ratio. When the two estimators have variances approx-

imately proportional to n in large samples, this definition of asymptotic relative efficiency coincides with the definition in terms of relative sample sizes needed to give equivalent precision.

As a simple example, consider the problem of estimating a mean of a population with finite variance, using a sample of n independent observations. If one procedure estimates the mean as the sample mean, and the second procedure estimates the mean as the sample mean with the first observation ignored, then $k_n = n + 1$, the relative efficiency is $(n + 1)/n$, and the asymptotic relative efficiency is 1. The deficiency is then 1 for all values of n, and so the asymptotic deficiency is also 1.

Fisher (1925) argues heuristically that maximum likelihood estimators are first–order efficient, with variances, to first order, given by the Fisher information in the whole sample, and that loss in efficiency incurred by other estimators might be measured by the correlation of these other estimators to the maximum likelihood estimator, or alternately by the differences between the whole sample information and the information in the sampling distribution of the estimator. Other authors have made these claims rigorous, and some of these results will be reviewed below. Wong (1992) presents a more thorough rigorous review. Rao (1962, 1963) uses the correlation between estimators to build a definition for second order efficiency which is equivalent to the Fisher information difference, under certain regularity conditions.

Higher–order asymptotic expansions for the mean squared error for the maximum likelihood estimator can be generated. Since expansions for the mean squared error are related to expansions for the information content of the maximum likelihood estimator, and the information expansion is simpler, the information expansion will be considered first. Efron (1975) uses methods from differential geometry to define the statistical curvature γ_θ of an inference problem, at a potential parameter value θ_0, to be rate of change in $b(\theta) = \left(1 - \text{Cor}\left[\dot{l}(\theta, \boldsymbol{X}), \dot{l}(\theta_0, \boldsymbol{X})\right]^2\right)^{1/2}$ per unit change in the arc length, evaluated at $\theta = \theta_0$. The arc length between θ_0 and θ is defined by $(\text{E}_\theta\left[U(\theta_0, \boldsymbol{X})\right] - \text{E}_{\theta_0}\left[U(\theta_0, \boldsymbol{X})\right])/\text{Var}_{\theta_0}\left[U(\theta_0, \boldsymbol{X})\right]^{1/2}$. The quantity $b(\theta)$ has the interpretation as the unexplained fraction of standard deviation of $U(\theta, \boldsymbol{X})$ given $U(\theta_0, \boldsymbol{X})$. This curvature is related to the loss of efficiency when inference procedures designed for local alternatives are applied globally. Consider a family of distributions on a sample space $\mathfrak{T}$ parameterized by θ taking values in $\mathfrak{P} \subset \mathfrak{R}$, and suppose that $\boldsymbol{X} \in \mathfrak{T}^n$ is a vector of n independent and identically distributed variables X_j. Let $\ell_n(\theta; \boldsymbol{X})$ be the log likelihood for $\boldsymbol{X}$. Let $\ddot{\ell}_n(\theta, \boldsymbol{X})$ be the second derivative of the log likelihood with respect to θ. Let $i_n(\theta) = -\text{E}_\theta\left[\ddot{\ell}_n(\theta, \boldsymbol{X})\right]$ be the Fisher information in the sample $\boldsymbol{X}$, let $i_n^{\hat{\theta}}(\theta)$ be Fisher's information for the sampling distribution of $\hat{\theta}$, the maximum likelihood estimator for θ, and let $i_1(\theta)$ be the Fisher information for X_1. If γ_θ^1 is the curvature defined for the distribution of a random variable X_1, and γ_θ^n is the curvature calculated for the distribution of $\boldsymbol{X}$, then $\gamma_\theta^1 = \gamma_\theta^n/\sqrt{n}$. One may show that $\lim_{n\to\infty}(i_n(\theta) - i_n^{\hat{\theta}}(\theta)) = i_1(\theta)(\gamma_\theta^1)^2$,

and hence $i_n^{\hat{\theta}}(\theta)/n = i_1(\theta) - i_1(\theta)\gamma_\theta^1/n + o(n^{-1})$, giving an asymptotic expansion for the average information is contained in a maximum likelihood estimator. Efron (1975) also produced an asymptotic expansion for the variance of the maximum likelihood estimator at a parameter value θ_0, which contains the statistical curvature and additional terms involving the curvature of the bias of the result of one scoring iteration, and the bias of the maximum likelihood estimator at θ_0. These terms are all of size $O(n^{-2})$, and the error is of size $o(n^{-2})$. Asymptotic comparisons of powers of families of tests having exactly or approximately the same significance level have been examined by many authors. Generally these investigations have considered the problem of testing a null hypothesis that a statistical parameter θ takes a null value, of the from $H_0 : \theta = \theta_0$, using two competing tests T_n^1 and T_n^2, indexed by a parameter n generally indicating sample size. Their critical values t_n^1 and t_n^2 satisfy

$$\mathrm{P}\left[T_n^i \geq t_n^i; H_0\right] = \alpha \text{ for } i = 1, 2. \tag{242}$$

Single–number measures of efficiency often times compare powers of tests whose sizes, exactly or approximately, are fixed and identical. For consistent tests, and a fixed alternative hypothesis, distribution functions for the test statistics, or asymptotic approximations to these distribution functions, indicate an identical first–order asymptotic power of unity. Distinguishing between such tests, then, requires a local measure of relative efficiency such as Pitman efficiency, which is the ratio of sample sizes necessary to give the same power against a local alternative. That is, alternatives of the from $H_A : \theta = \theta_0 + \epsilon/c_n$, where $c_n \to \infty$, are considered, and the limit $\lim_n(k_n/n)$ is desired, where

$$\mathrm{P}\left[T_n^1 \geq t_n^1; H_A\right] = \mathrm{P}\left[T_{k_n}^2 \geq t_n^2; H_A\right]. \tag{243}$$

Hodges and Lehmann (1970) apply their concept of deficiency to the problem of comparing tests in cases in which sizes can be calculated exactly and in which the asymptotic relative efficiency is unity.

Often times exact expressions for the probabilities in (242) and (243) are unavailable. In such cases the critical value, as well as the power, usually must be approximated. Pfanzagl (1980) notes that asymptotic comparisons of power are only interesting when significance levels of the tests agree to the same asymptotic order, and achieves this equality of size through a process of studentization, in the presence of nuisance parameters. Such equality of size might be obtained using a Cornish–Fisher expansion to calculate the critical value for the test. Albers, Bickel, and van Zwet (1976) apply Edgeworth series to calculate the powers of nonparametric tests. The primary difficulty in such cases arises from the discrete nature of the distributions to be approximated. Pfaff and Pfanzagl (1985) present applications of Edgeworth and saddlepoint expansions to the problem of approximating power functions for test statistics with continuous distributions; they find that the Edgeworth series is more useful for analytic comparisons of power, while the saddle-

point methods give more accurate numerical results. Kolassa (1996c) gives further references.

Asymptotic expansions of cumulative distribution functions can also be used to calculate asymptotic relative efficiencies and deficiencies. Those applications discussed here are under local alternatives. First–order approximations are generally sufficient to calculate asymptotic relative efficiencies; deficiency calculations generally require that second order terms be included as well. Taniguchi (1991) calculates these powers to examine cases when the deficiency is zero and a third–order counterpart of (241) is required; comparisons may then be based on the final coefficient.

9.5. Exercises

1. Calculate the tail probability approximations for the root of an estimating equation by approximating the inversion integral (233) using
   a. The method of Robinson (§5.2).
   b. The method of Lugannani and Rice (§5.3).
2. Consider the following changes in systolic blood pressure after administration of a drug: -9, -4, -21, -3, -20, -31, -17, -26, -26, -10, -23, -33, -19, -19, -23 (Cox and Snell, 1981, p. 72). Assume that these observations come from a symmetric distribution with point of symmetry θ.
   a. Approximate the 5% and 95% points t_L and t_U of the null distribution of the statistic T in (240), assuming $\theta = \theta_0$, using perhaps (109).
   b. Invert the statement $P\left[t_L \leq T(\theta) \leq t_U\right] = 90\%$ to generate a 90% confidence interval for θ.

10
Computational Aids

This chapter contains code for doing some of the calculations presented here using the computer algebra package *Mathematica*. Lines surrounded by parenthesis star are comments. Material here is generally in the same order as it appears in the text, except that generally lattice material in the text was at the end of chapters; here it follows more naturally immediately after the continuous analogues. Code presented here is the minimal code necessary to perform many of the calculations in the text. Andrews and Stafford (1993) and Stafford and Andrews (1993) present more sophisticated *Mathematica* code.

```
(* Stirling's Expansion of §1.2.*)
gn[n_]=Normal[Series[Gamma[n],{n,Infinity,30}]]
uu=Series[Simplify[(gn[n] n^-n Exp[n] n^(1/2))],{n,Infinity,30}]
cl=CoefficientList[Normal[uu/Sqrt[2 Pi]]/.n->1/oon,oon]
n=1
approx=cl
pn=1
Do[pn=pn*n;approx[[i]]=cl[[i]]/pn+approx[[i-1]],{i,2,Length[approx]}]
error=approx-Gamma[n] n^-n Exp[n] n^(1/2)/Sqrt[2 Pi]

(* Define and exponentiate the MGF and CGF expansion of §2.1 for a
standardized sum.  Note rn is 1/sqrt(n).  Give conversions between
moments and cumulants. Give sample Fortran output.*)
K[x_]=Series[Sum[k[i]*x^i*rn^(i-2)/i!,{i,2,6}],{x,0,6}]
expK[x_]=Exp[K[x]]
M[x_]=Series[1+Sum[m[i]*x^i*rn^(i-2)/i!,{i,2,6}],{x,0,6}]
momentrule=Simplify[Solve[Simplify[LogicalExpand[M[x]==expK[x]]],
  {m[2],m[3],m[4],m[5],m[6]}]]
cumulantrule=Simplify[Solve[Simplify[LogicalExpand[M[x]==expK[x]]],
  {k[2],k[3],k[4],k[5],k[6]}]]
FortranForm[cumulantrule]
```

```
(* Define pseudocumulant g.f. of §3.4 and exponentiate to get
pseudomoments.  We need series to order 18 in x since we need powers of x
to order three times power in rn.*)
PK[x_]=Sum[k[i]*x^i*rn^(i-2)/i!,{i,3,18}]
PMlist=CoefficientList[Expand[Normal[Series[Exp[PK[x]],{x,0,18}]]],{x}]

(* Do the same thing to get the pseudomoments for the Edgeworth series
with Sheppard-corrected cumulants of §3.15.*)
PE[x_]=Sum[e[2*i]*x^(2*i)*rn^(2*i)/(2*i)!,{i,1,5}]
PMSlist=CoefficientList[,{x}]nd[Normal[Series[Exp[PK[x]+PE[x]],{x,0,12}]]]

(* Write Edgeworth Series without expanding phi, from §3.1.*)
Remove[phi]
hermitepolys=Table[h[i-1],{i,1,10}]
edgeworthseries=phi[x]*Series[Sum[PMlist[[i]]
   *hermitepolys[[i]],{i,1,Length[hermitepolys]}],{rn,0,6}]
Edgeworthseries=Phi[x]-phi[x]*Series[Sum[PMlist[[i]]
   *hermitepolys[[i-1]],{i,2,10}],{rn,0,3}]
TeXForm[Edgeworthseries]

(* Calculate Hermite Polynomials of §3.4 by constructing an empty list,
assigning phi to the first element, then recursively differentiating.*)

phipr=Table[NULL,{i,1,20}]
phi[x_]=Exp[-x^2/2]/Sqrt[2*Pi]
phipr[[1]]=phi[x]
Do[phipr[[i+1]]=D[phipr[[i]],x],{i,1,Length[phipr]-1}]
hermitepolys=Simplify[phipr/phi[x]]* ((-1)^(1+Range[Length[phipr]]))

(* Construct the Cornish--Fisher expansion of §3.12 by expanding the
Edgeworth Series about the normal approximation quantile and inverting.
Leave Hermite polynomials unexpanded (except for the first); express the
difference between the Edgeworth inverse and the Normal inverse as a
power series in rn.*)
Clear[phi]; Clear[z]; Clear[Phi]
ES[z_]=(Normal[Phi[z]-Series[Sum[PMlist[[i+1]]*(-1)^(i-1)
   *Derivative[i][Phi][z],{i,2,10}],{rn,0,2}]])
ESS=Simplify[(Expand[Normal[Series[ES[z+a]-Phi[a],{z,0,2}]]]//.
   {Derivative[k_][Phi][z_]->Derivative[k-1][phi][z],
    Derivative[k_][phi][z_]->phi[z]*h[k]*(-1)^k,h[0]->1})/phi[a]]
z=Sum[b[i]*rn^i,{i,1,3}]
ESSS=Expand[Series[ESS,{rn,0,2}]]
cfeqn=Simplify[ESSS==0]
cfrule=Solve[Simplify[LogicalExpand[cfeqn]],{b[1],b[2]}]
Do[cfrule=Simplify[cfrule/.h[i]->hermitepolys[[i+1]]],{i,0,7}]
CFapprox=x+z/.Join[cfrule[[1]],{b[3]->0}]
Clear[z]
```

```
(* Construct the saddlepoint expansion relative error in terms of invariants
by, as in §4.1, evaluating the Edgeworth series at 0 and adjusting the
coefficients for non-unit variance. The coefficients beta are invariants.*)

spre=Simplify[edgeworthseries/.{x->0,k[r_]->beta[that,r] }]

(* Calculations for the proof that no distributions other than only
normal, inverse Gaussian, and gamma distributions have saddlepoint
approximations exact up to a constant of proportionality, from §4.8.*)
f[x]=(a x+b)^(-3/2)
3*D[D[f[x],x],x]-5 D[f[x],x]^2/f[x]
alphas=Expand[CoefficientList[spre,rn]]
dbetarule=Derivative[1,0][beta][that,r_]->Sqrt[K''[that]]*(
  beta[that,r+1]- (r/2)*beta[that,3]*beta[that,r])
eq={Expand[24*(D[alphas[[3]],that]/.dbetarule)/Sqrt[K''[that]]],
    Expand[(D[eq1,that]/.dbetarule)/Sqrt[K''[that]]],
    Expand[(24*alphas[[3]])^2],
    Expand[3*384*alphas[[5]]]}
TeXForm[eq/.beta[that,r_]->beta[r]]
coefs={20*beta[that,3]/3,8,23/9,1}
TeXForm[coefs/.beta[that,r_]->beta[r]]
TeXForm[Simplify[Sum[coefs[[i]]*eq[[i]],{i,1,4}]]/.beta[that,r_]->beta[r]]

(* Calculations for the Bahadur and Ranga Rao and the Lugannani and Rice
expansions of §5.1 and §5.3.  nt is the number of terms; nt=2
indicates the n^(-3/2) term as well as the n^(-1/2) term.*)
nt=2
w[beta_]=Sqrt[2*(K[beta]-beta K'[betahat]-K[betahat]+betahat K'[betahat])]+what
bb=(Simplify[InverseSeries[Series[w[beta],{beta,betahat,2*nt-1}],w]/.
   Derivative[k_][K][betahat]:>r[k] Derivative[2][K][betahat]^(k/2)/;k>2])//.
   {K''[betahat]->sigma^2,Sqrt[K''[betahat]]->sigma,betahat->zhat/sigma,
   (sigma^a_)^b_->sigma^(a b)}/.sigma->Sqrt[K''[betahat]]
brr=Ndens[rtn what]*Sum[Simplify[(-1)^(jj-1)*
   SeriesCoefficient[Simplify[D[bb,w]/bb],2*(jj-1)]]*rtn^(1-2*jj),{jj,1,nt}]
landr=1-Ncdf[rtn what]+Ndens[rtn what]*Sum[(-1)^(jj-1)*Simplify[
   SeriesCoefficient[Simplify[D[bb,w]/bb-1/w],2*(jj-1)]]*rtn^(1-2*jj),{jj,1,nt}]

(* Commands to calculate terms in Robinson's expansion of §5.2.  I calculate
integrals that are the upper half of the Laplace transform of the normal density
times Hermite polynomials, both using recursion and from first principles.*)

II[j_]=If[j>0,(hermitepolys[[j]]/.x->0)/Sqrt[2 Pi]-rtn hz II[j-1],
  Ncdfbar[rtn hz]/(Ndens[rtn hz] Sqrt[2 Pi])]
III[j_,uh_]:=Simplify[Integrate[Exp[-u^2/2]u^j I^j/(2 Pi(I u+uh)),
   {u,-Infinity,Infinity}, GenerateConditions->False
   ]//.{(rtn^c_ hz^a_)^b_->hz^(a b) rtn^(c b),Erf[x_]->1-2*Ncdfbar[x*Sqrt[2]],
   Exp[hz^2 rtn^2/2]->1/(Sqrt[2 Pi] Ndens[hz rtn]) }]
robinson=Sqrt[2 Pi]Ndens[rtn hatw] (II[0]+rtn^(-1) rho[3] II[3]/6)
```

```
(* Commands to calculate Robinson's approximation via inversion.*)
AA[j_]=If[j>1,(j-1) AA[j-2],If[j==0,1,0]]
JJ[j_]=If[j>0,(-1)^((j-1)/2)rtn^(-j)Exp[-rtn^2 hz^2/2]AA[j-1]/Sqrt[2 Pi]-
   hz JJ[j-1], Ncdfbar[rtn hz]]

poly=(1+rtn^2 rho[3]*sigma^3*gamma^3/6)
cl=CoefficientList[poly,gamma]
Doapprox[cl_]:=Module[{},
   Sqrt[2 Pi]Ndens[rtn hw]* Sum[
   cl[[i]]*JJ[i-1]*Exp[rtn^2 hz^2/2]/(sigma)^(i-1),{i,1,Length[cl]}]
     //.{
     Exp[a_^2/2]->1/(Sqrt[2 Pi]Ndens[a]), Exp[-a_^2/2]->Sqrt[2 Pi]Ndens[a]}]
robinson2=Doapprox[cl]/.Exp[hz^2 rtn^2/2]->1/(Ndens[hz rtn] Sqrt[2 Pi])
(* Check that these coincide.*)
Simplify[robinson1-robinson2]
(* Commands to calculate Robinson's approximation via inversion, lattice case.*)
ncl=CoefficientList[Series[poly*Normal[Series[
   (beta+gamma)/(2 Sinh[(beta+gamma)/2]),{gamma,0,1}]] ,{gamma,0,3}],gamma]
lapprox=(ncl[[1]]*Apart[Doapprox[Simplify[ncl/ncl[[1]]]]])//.{
   rtn^2->n,beta sigma->hz}

(* Commands to calculate the Skates (1993) expansion of §5.7.*)
P[x_]=Series[Sum[p[i]*x^i/i!,{i,2,7}],{x,0,7}]
Pp[x_]=D[P[x],x]
p[2]=1
Q[x_]=Series[Sum[q[i]*x^i/i!,{i,0,5}],{x,0,5}]
q[0]=1
X[y_]=Simplify[InverseSeries[Simplify[Sqrt[2*P[y]]]]]/.x->y
g[y_]=Simplify[Q[X[y]] y/Pp[X[y]]]

(* Calculate the discrete analogues of §5.8. Some care must be taken in the
treatment of Sinh, since otherwise the expansion won't simplify usefully.
Introduce new symbols for hyperbolic sine and cosine that Mathematica won't try
to expand except according to the derivative rules below. These calculations
require definition of nt and tt[x] from above.*)

Derivative[i_][Ss][x_]:=Cc[x]/;OddQ[i]
Derivative[i_][Cc][x_]:=Ss[x]/;OddQ[i]
Derivative[i_][Ss][x_]:=Ss[x]/;EvenQ[i]
Derivative[i_][Cc][x_]:=Cc[x]/;EvenQ[i]

inside=Ss[Times[Rational[1,2],zhat,Power[Derivative[2][K][betahat],
   Rational[-1, 2]]]]-> 2 zhat/(Sqrt[Derivative[2][K][betahat]])
brr=Ndens[rtn what]*Sum[Simplify[(-1)^(jj-1)*rtn^(1-2*jj)*
   SeriesCoefficient[Simplify[D[bb,w]/(Ss[bb/2]/2)],2*(jj-1)]],{jj,1,nt}
   ]//.inside
landr=1-Ncdf[rtn what]+Ndens[rtn what]*Sum[(-1)^(jj-1)*Simplify[
   SeriesCoefficient[Simplify[D[bb,w]/(Ss[bb/2]/2)-1/w],2*(jj-1)]
   ]*rtn^(1-2*jj),{jj,1,nt}]//.inside
```

```
(* Commands for the example of §7.2.*)
exppsi[theta1_rtheta2_]=1-theta1-2 theta2+theta2(theta1+theta2)
Psi[theta1_,theta2_]=-Log[exppsi[theta1,theta2]]
tr=Simplify[Solve[D[Psi[theta1,theta2],theta2]==t2,theta2]][[1]]
lr={Log[a_*b_]->Log[a]+Log[b], Log[a_/c_]->Log[a]+-Log[c],
   Log[c_.*Power[a_,b_]]->b Log[a]+Log[c]}
var=Simplify[D[D[Psi[theta1,theta2], theta2],theta2]]
lvar=Log[var] //.lr
ep=Simplify[exppsi[theta1,theta2]/.tr]
K[theta1_]=Simplify[Simplify[(-theta2 t2-Log[ep]-(1/2) Log[var])/.tr]//.lr]
trans=( K[theta1]/.Times[Power[theta1, 2], Power[t2, 2]]->w^2-4)/.Sqrt[w^2]->w
Simplify[Simplify[trans]/.lr]

(* Commands for Reid's Gamma distribution ex. of §8.3.*)
L=(nu/mu)^(n*nu)*Exp[slx*(nu-1)]*Exp[-sx*nu/mu]/(Gamma[nu]^n)
l[mu_,nu_]=n(nu*Log[nu]-nu*Log[mu]+(nu-1)*T2-nu*T1/mu-Log[Gamma[nu]])
Simplify[{D[l[mu,nu],nu],D[l[mu,nu],mu]}]
Simplify[{ {D[D[l[mu,nu],nu],nu],D[D[l[mu,nu],mu],nu]},
  {D[D[l[mu,nu],nu],mu],D[D[l[mu,nu],mu],mu]}} ]

(* Calculate expected value of log(x).*)
eT2=Integrate[Simplify[Log[x]*L/.{n->1,sx->x,slx->Log[x]}],
   {x,0,Infinity}]
MLEdef={D[l[muhat,nuhat],nuhat]==0,D[l[muhat,nuhat],muhat]==0}
MLErules=Simplify[Solve[MLEdef, {T1,T2}]]
obsinfmat={{D[D[l[mu,nu],mu],mu],D[D[l[mu,nu],nu],mu]},
   {D[D[l[mu,nu],nu],mu],D[D[l[mu,nu],nu],nu]}}
det=Simplify[obsinfmat[[1,1]]*obsinfmat[[2,2]]-obsinfmat[[1,2]]^2/.
   {mu->muhat,nu->nuhat}
sqrtdet=Sqrt[Simplify[det/.MLErules]][[1]]
exponent=Simplify[((l[mu,nu]-l[muhat,nuhat])/.MLErules)/n][[1]]
TeXForm[Simplify[Exp[n*exponent]*sqrtdet]]

(* Commands to calculate the exact correction factor for the mean of
the Gamma likelihood ratio statistic with fixed shape, from §8.2.*)
Llk[x_,n_]=(tau*n)^(n*alpha)*x^(n*alpha-1)*Exp[-tau x*n]/Gamma[alpha*n]
lrules={
   Log[a_/b_]->Log[a]-Log[b], Log[a_*c_]->Log[a]+Log[c],
   Log[a_*c_/b_]->Log[a]+Log[c]-Log[b], Log[a_^b_]->b*Log[a]}
grules={Gamma[1+x_]->x*Gamma[x],PolyGamma[0,x_]->psi[x]}
llk[x_,n_]=Simplify[Log[Llk[x,n]]]//.lrules
tauhat=Solve[D[llk[x,n],tau]==0,tau]
lrs=Simplify[-2*(llk[x,n]-(llk[x,n]/.tauhat[[1]]))]/.lrules]
ex=Integrate[x*Llk[x,n],{x,0,Infinity}]/.grules
elx=Integrate[Log[x]*Llk[x,n],{x,0,Infinity}]
Elrs=Simplify[(lrs/.Log[x]->elx)/.x->ex/.lrules]/.grules
```

(* Perform calculations for the distribution of the ratio of means from Daniels
(1983) of §9.1. The first three lines calculate maximum likelihood estim-
ates. Moment and Cumulant generating functions and derivatives are defined,
the saddlepoint is calculated, and three factors for the approximation are
defined.*)

```
l=-x/mux-y/muy-Log[mux]-Log[muy]
newl:=l/.muy->mux*rho
Solve[{D[newl,mux]==0,D[newl,rho]==0},{mux,rho}]
Kr[t_,rho_]=-Log[1-mux*t]-Log[1+mux*rho*t]
M[t_,rho_]=1/((1-mux*t)*(1+mux*rho*t))
Kr1[t_,rho_]=D[Kr[t,rho],t]
Kr11[t_,rho_]=D[Kr1[t,rho],t]
Kr2[t_,rho_]=D[Kr[t,rho],rho]
rl=Simplify[Solve[D[Kr[t,rho],t]==0,t]][[1]]
FirstSq=Simplify[n/(2*pi*Kr11[t,rho])/.rl]
Secnd=Simplify[-((1/t)*Kr2[t,rho])/.rl]
Third=Simplify[(M[t,rho]^n)/.rl]
that=t/.rl
approx=Expand[Sqrt[FirstSq]*Secnd*Third]//.
   {Sqrt[a_^2*b_.]->a*Sqrt[b],Sqrt[b_./(c_.*a_^2)]->Sqrt[b/c]/a,
   Sqrt[a_^(-2)*b_]->Sqrt[b]/a}
TeXForm[approx]
Remove[that]
```

Bibliography

Albers, W., Bickel, P.J., and van Zwet, W.R. (1976), "Asymptotic Expansions for the Power of Distribution Free Tests in the One-Sample Problem," *Annals of Statistics*, **4**, 108–156.

Albert, A., and Anderson, J.A. (1984), "Maximum Likelihood Estimates in Logistic Regression," *Biometrika*, **71**, 1–10.

Andrews, D.F., and Stafford, J.E. (1993), "Tools for the symbolic computation of asymptotic expansions," *Journal of the Royal Statistical Society Series B*, **55**, 613–627.

Bahadur, R.R., and Ranga Rao, R. (1960), "On deviations of the sample mean," *Annals of Mathematical Statistics*, **31**, 1015–1027.

Bak, J., and Newman, D.J. (1996), *Complex Analysis*, New York: Springer–Verlag.

Barndorff–Nielsen, O.E. (1978), *Information and Exponential Families in Statistical Theory*, Chichester: John Wiley and Sons.

Barndorff–Nielsen, O.E. (1980), "Conditionality Resolutions," *Biometrika*, **67**, 293–310.

Barndorff–Nielsen, O.E. (1983), "On a Formula for the Distribution of the Maximum Likelihood Estimator," *Biometrika*, **70**, 343–365.

Barndorff–Nielsen, O.E. (1984), "On Conditionality Resolution and the Likelihood Ratio for Curved Exponential Models," *Scandinavian Journal of Statistics*, **11**, 157–170. Correction (1985).

Barndorff–Nielsen, O.E. (1986), "Inference on Full or Partial Parameters Based on the Standardized Signed Log Likelihood Ratio," *Biometrika*, **73**, 307–322.

Barndorff–Nielsen, O.E. (1990a), "A note on the standardized signed log likelihood ratio," *Scandinavian Journal of Statistics*, **17**, 157–160.

Barndorff–Nielsen, O.E. (1990b), "Approximate Interval Probabilities," *Journal of the Royal Statistical Society Series B*, **52**, 485–496.

Barndorff–Nielsen, O.E., and Cox, D.R. (1979), "Edgeworth and Saddlepoint Approximations with Statistical Applications," *Journal of the Royal Statistical Society Series B*, **41**, 279–312.

Barndorff–Nielsen, O.E., and Cox, D.R. (1984), "Bartlett Adjustments to the Likelihood Ratio Statistic and the Distribution of the Maximum Likelihood Estimator," *Journal of the Royal Statistical Society Series B*, **46**, 483–495.

Barndorff–Nielsen, O.E., and Cox, D.R. (1989), *Asymptotic Techniques for Use in Statistics*, London: Chapman and Hall.

Barndorff–Nielsen, O.E., and Klüppelberg, C. (1999), "Tail exactness of multivariate saddlepoint approximations," *Scandinavian Journal of Statistics*, **26**, 253–264.

Bartlett, M.S. (1953), "Approximate Confidence Intervals: II. More than One Unknown Parameter," *Biometrika*, **40**, 306–317.

Bartlett, M.S. (1955), "Approximate Confidence Intervals: III. A Bias Correction," *Biometrika*, **42**, 201–204.

Beek, P. van (1972), "An Application of Fourier Methods to the Problem of Sharpening the Berry–Esseen Inequality," *Zeitshrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **23**, 187–196.

Berman, E., Chacon, L., House, D., Koch, B.A., Koch, W.E., Leal, J., Løvtrup, S., Mantiply, E., Martin, A.H., Martucci, G.I., Mild, K.H., Monahan, J.C., Sandström, M., Shamsaifar, K., Tell, R., Trillo, M.A., Ubeda, A., and Wagner, P. (1990), "Development of Chicken Embryos in a Pulsed Magnetic Field," *Bioelectromagnetics*, **11**, 169–187.

Bhattacharya, R., and Denker, M. (1990), *Asymptotic Statistics*, Boston: Birkhäuser.

Bhattacharya, R.N., and Rao, R.R. (1976), *Normal Approximation and Asymptotic Expansions*, New York: Wiley.

Billingsley, P. (1986), *Probability and Measure*, New York: Wiley.

Blæsild, P., and Jensen, J.L. (1985), "Saddlepoint Formulas for Reproductive Exponential Models," *Scandinavian Journal of Statistics*, **12**, 193–202.

Blackwell, D., and Hodges, J.L. (1959), "The probability in the extreme tail of a convolution," *Annals of Mathematical Statistics*, **30**, 1113-1120.

Bleistein, N. (1966), "Uniform Asymptotic Expansions of Integrals with Stationary Point near Algebraic Singularity," *Communications on Pure and Applied Mathematics*, **19**, 353–370.

Bochner, S., and Martin, W.T. (1948), *Several complex variables*, Princeton: Princeton University Press.

Chambers, J.M. (1967), "On methods of asymptotic approximation for multivariate distributions," *Biometrika*, **54**, 367–383.

Clarkson, D.B., and Jennrich, R.I. (1991), "Computing Extended Maximum Likelihood Estimates for Linear Parameter Models," *Journal of the Royal Statistical Society Series B*, **53**, 417–426.

Cordeiro, G.M., and Paula, G.A. (1989), "Improved likelihood ratio statistics for exponential family nonlinear models (Corr: V78 p935)," *Biometrika*, **76**, 93–100.

Cornish, E.A., and Fisher, R.A. (1937), "Moments and Cumulants in the Specification of Distributions," *International Statistical Review*, **5**, 307–322.

Cox, D.R., and Reid, N. (1987), "Parameter Orthogonality and Approximate Conditional Inference," *Journal of the Royal Statistical Society Series B*, **49**, 1–39.

Cox, D.R., and Snell, E.J. (1981), *Applied Statistics*, London: Chapman and Hall.

Cox, D.R. (1980), "Local Ancillarity," *Biometrika*, **67**, 279–86.

Cramér, H. (1946), *Mathematical Methods of Statistics*, Princeton: Princeton University Press.

Cramér, H. (1925), "On Some Classes of Series Used in Mathematical Statistics," *Proceedings of the Sixth Scandinavian Congress of Mathematicians*, 399–425.

Daniels, H.E. (1954), "Saddlepoint Approximations in Statistics," *Annals of Mathematical Statistics*, **25**, 614–649.

Daniels, H.E. (1958), "Discussion of The Regression Analysis of Binary Sequences," *Journal of the Royal Statistical Society Series B*, **20**, 236–238.

Daniels, H.E. (1983), "Saddlepoint approximations for estimating equations," *Biometrika*, **70**, 89–96.

Daniels, H.E. (1987), "Tail Probability Approximations," *Review of the International Statistical Institute*, **55**, 37–46.

Davis, A.W. (1976), "Statistical distributions and multivariate Edgeworth populations," *Biometrika*, **63**, 661–670.

Davison, A. C. (1986), "Approximate predictive likelihood (Corr: V77 p667)," *Biometrika*, **73**, 323–332. Davison, A.C., and Hinkley, D.V. (1988), "Saddlepoint Approximations in Resampling Methods," *Biometrika*, **75**, 417–431.

Davison, A.C. (1988), "Approximate Conditional Inference in Generalized Linear Models," *Journal of the Royal Statistical Society Series B*, **50**, 445–461.

DiCiccio, T., Hall, P., and Romano, J. (1991), "Empirical likelihood is Bartlett-correctable," *Annals of Statistics*, **19**, 1053–1061.

DiCiccio, T.J., Martin, M.A., and Young, G.A. (1993), "Analytical approximations to conditional distribution functions," *Biometrika*, **80**, 781–790.

DiCiccio, T.J., and Martin, M.A. (1991), "Approximations of marginal tail probabilities for a class of smooth functions with applications to Bayesian and conditional inference," *Biometrika*, **78**, 891–902.

Durbin, J. (1980), "Approximations for Densities of Sufficient Estimators," *Biometrika*, **67**, 311–333.

Efron, B. (1982), *The Jackknife, the Bootstrap, and Other Resampling Plans*, Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B. (1975), "Defining the Curvature of a Statistical Problem (with Applications to Second Order Efficiency)," *Annals of Statistics*, **3**, 1189–1242.

Esseen, C.G. (1945), "Fourier Analysis of Distribution Functions," *Acta Mathematica*, **77**, 1–125.

Esseen, C.G. (1956), "A Moment Inequality with an Application to the Central Limit Theorem," *Skandinavisk Aktuarietidskrift*, **39**, 160–170.

Feller, W. (1971), *An Introduction to Probability Theory and its Applications*, New York: Wiley, **II**.

Field, C., and Ronchetti, E. (1990), *Small Sample Asymptotics*, Hayward, CA: Institute of Mathematical Statistics.

Fisher, R.A., and Cornish,E.A. (1960), "The Percentile Points of Distributions Having Known Cumulants," *Technometrics*, **2**, 209–225.

Fisher, R.A. (1925), "Theory of Statistical Estimation," *Proceedings of the Cambridge Philosophical Society*, **22**, 700–725.

Fisher, R.A. (1934), "Two New Properties of Mathematical Likelihood," *Proceedings of the Royal Society of London Series A*, **144**, 285–307.

Fraser, D.A.S., Reid, N., and Wong, A. (1991), "Exponential Linear Models: A Two–Pass Procedure for Saddlepoint Approximation," *Journal of the Royal Statistical Society Series B*, **53**, 483–492.

Fraser, D.A.S. (1968), *The Structure of Inference*, New York: Wiley.

Fraser, D.A.S. (1979), *Inference and Linear Models*, New York: McGraw-Hill.

Frydenberg, M., and Jensen, J.L. (1989), "Is the 'Improved Likelihood Ratio Statistic' Really Improved in the Discrete Case?," *Biometrika*, **76**, 655–661.

Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*, New York: Springer.

Hall, P. (1983), "Inverting an Edgeworth Expansion," *Annals of Statistics*, **11**, 569–576.

Haynsworth, E.V., and Goldberg, K. (1965), "Bernoulli and Euler Polynomials, Riemann Zeta Function," *Handbook of Mathematical Functions* (ed. Abramowitz, M., and Stegun, I.A.), New York: Dover.

Hettmansperger, T.P. (1984), *Statistical Inference Based on Ranks*, Melbourne, FL: Krieger.

Hocking, R.R. (1985), *The Analysis of Linear Models*, Monterey, CA: Brooks/Cole.

Hodges, J.L.Jr., and Lehmann, E.L. (1970), "Deficiency," *Annals of Mathematical Statistics*, **41**, 783–801.

Jacobsen, M. (1989), "Existence and Unicity of MLEs in Discrete Exponential Family Distributions," *Scandinavian Journal of Statistics*, **16**, 335–349.

Jeffreys, H. (1962), *Asymptotic Approximations*, London: Oxford University Press.

Jensen, J. L. (1992), "The modified signed likelihood statistic and saddlepoint approximations," *Biometrika*, **79**, 693–703.

Jensen, J.L. (1995), *Saddlepoint Approximations*, Oxford: Oxford Science Publications.

Jensen, J.L. (1988), "Uniform Saddlepoint Approximations," *Advances in Applied Probability*, **20**, 622–634.

Jing, B., and Robinson, J (1994), "Saddlepoint Approximations for Marginal and Conditional Probabilities of Transformed Variables," *Annals of Statistics*, **22**, 1115–1132.

Jing, B.Y., and Wood, A.T.A (1996), "Exponential Empirical Likelihood is not Bartlett Correctable," *Annals of Statistics*, **24**, 365–369.

Kass, R.E. (1988), "Discussion of Saddlepoint Methods and Statistical Inference," *Statistical Science*, **3**, 234–236.

Kass, R.E., Tierney, L., and Kadane, J.B (1990), "The validity of posterior expansions based on Laplace's method," *Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George A. Barnard*, 473–488.

Kendall, M.G., Stuart, A., and Ord, J.K. (1987), *Kendall's Advanced Theory of Statistics*, New York: Oxford, **I**.

Kolassa, J.E. (1989), *Topics in Series Approximations to Distribution Functions*, Chicago: Unpublished doctoral dissertation.

Kolassa, J.E. (1991), "Saddlepoint Approximations in the Case of Intractable Cumulant Generating Functions," *Selected Proceedings of the Sheffield Symposium on Applied Probability* (ed. Basawa, I.V. and Taylor, R.E.), Hayward, CA: Institute of Mathematical Statistics, 236–255.

Kolassa, J.E. (1992), "Confidence Intervals for Thermodynamic Constants," *Geochimica et Cosmochimica Acta*, **55**, 3543–3552.

Kolassa, J.E. (1994), "Small Sample Conditional Inference in Biostatistics," *Proceedings of the Symposium on the Interface* (ed. Sall, J., and Lehman, A.), Fairfax Station, VA: Interface Foundation of North America, **26**, 333–339.

Kolassa, J.E. (1995), "Monte Carlo Sampling in Multiway Contingency Tables," *Bulletin of the International Statistical Institute: Proceedings of the Fiftieth Session*, Beijing: International Statistical Institute, **II**, 471–480.

Kolassa, J.E. (1996a), "Uniformity of Double Saddlepoint Conditional Probability Approximations,".

Kolassa, J.E. (1996b), "Higher-Order Approximations to Conditional Distribution Functions," *Annals of Statistics*, **24**, 353–365.

Kolassa, J.E. (1996c), "Asymptotics, Higher–Order," *Encyclopedia of Statistical Sciences*, : .

Kolassa, J.E. (1997), "Infinite Parameter Estimates in Logistic Regression," *Scandinavian Journal of Statistics*, **24**, 523–530.

Kolassa, J.E. (2001), "Saddlepoint Approximation at the Edges of a Conditional Sample Space," *Statistics and Probability Letters*, **50**, 343–349.

Kolassa, J.E. (2003a), "Multivariate Saddlepoint Tail Probability Approximations," *Annals of Statistics*, **31**, 274–286.

Kolassa, J.E. (2003b), "Algorithms for Approximate Conditional Inference," *Statistics and Computing*, **13**, 121–126.

Kolassa, J.E. (2004), "Approximate Multivariate Conditional Inference Using the Adjusted Profile Likelihood," *Canadian Journal of Statistics*, **32**, 5–14.

Kolassa, J.E. (2005),, "A Proof of the Asymptotic Equivalence of Two Tail Probability Approximations," *Rutgers University Department of Statistics Technical Report*.

Kolassa, J.E., and McCullagh, P. (1990), "Edgeworth Series for Lattice Distributions," *Annals of Statistics*, **18**, 981–985.

Kolassa, J.E., and Tanner, M.A. (1994), "Approximate Conditional Inference in Exponential Families Via the Gibbs Sampler," *Journal of the American Statistical Association*, **89**, 697–702.

Kong, F., and Levin, B. (1996), "Edgeworth Expansions for the Conditional Distributions in Logistic Regression Models," *Journal of Statistical Planning and Inference*, **52**, 109–129.

Koning, A.J., and Does, R.J.M.M. (1988), "Approximating the Percentage Points of Simple Linear Rank Statistics with Cornish–Fisher Expansions," *Applied Statistics*, **37**, 278–284.

Konishi, S., Niki, N., and Gupta, A.K. (1988), "Asymptotic Expansions for the Distribution of Quadratic Forms in Normal Variables," *Annals of the Institute of Statistical Mathematics*, **40**, 279–296.

Lawley, D.N. (1956), "A General Method for Approximating to the Distribution of the Likelihood Ratio Criteria," *Biometrika*, **43**, 295–303.

Le Cam, L.M. (1969), *Théorie Asymptique de la Décision Statistisque*, Montreal: Université de Montréal.

Levin, B. (1990), "The Saddlepoint Correction in Conditional Logistic Likelihood Analysis," *Biometrika*, **77**, 275–285.

Levin, B., and Kong, F. (1990), "Bartlett's Bias Correction to the Profile Score Function is a Saddlepoint Corrrection," *Biometrika*, **77**, 219–221.

Lugannani, R., and Rice, S. (1980), "Saddle Point Approximation for the Distribution of the Sum of Independent Random Variables," *Advances in Applied Probability*, **12**, 475–490.

Mardia, K.V. (1970), "Measures of Multivariate Skewness and Kurtosis with Applications," *Biometrika*, **57**, 519–530.

Mathews, J., and Walker, R.L. (1964), *Mathematical Methods of Physics*, New York: W.A. Benjamin, Inc..

McCullagh, P. (1984), "Local Sufficiency," *Biometrika*, **71**, 233–244.

McCullagh, P. (1986), "The Conditional Distribution of Goodness-of-Fit Statistics for Discrete Data," *Journal of the American Statistical Association*, **81**, 104–107.

McCullagh, P. (1987), *Tensor Methods in Statistics*, London: Chapman and Hall.

McCullagh, P., and Cox, D.R. (1986), "Invariants and Likelihood Ratio Statistics," *Annals of Statistics*, **14**, 1419–1430.

McCullagh, P., and Nelder, J.A. (1989), *Generalized Linear Models*, London: Chapman and Hall.

McCullagh, P., and Tibshirani, R. (1990), "A Simple Method for the Adjustment of Profile Likelihoods," *Journal of the Royal Statistical Society Series B*, **52**, 325–344.

Petrov, V.V. (1975), *Sums of independent random variables*, New York: Springer–Verlag.

Pfaff, T., and Pfanzagl, J. (1985), "On the Accuracy of Asymptotic Expansions for Power Functions," *Journal of Statistical Computation and Simulation*, **22**, 1–25.

Pfanzagl, J. (1980), "Asymptotic Expansions in Parametric Statistical Theory," *Developments in Statistics* (ed. Krishnaiah, P.R.), New York: Academic Press, **3**, 1–98.

Phillips, P.C.B. (1983), "Marginal Densities of Instrument Variables Estimators in the General Single Equation Case," *Advances in Econometrics*, **2**, 1–24.

Prášková–Vizková, Z. (1976), "Asymptotic Expansion and a Local Limit Theorem for a Function of the Kendall Rank Correlation Coefficient," *Annals of Statistics*, **4**, 597–606.

Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T. (1986), *Numerical Recipes: The Art of Scientific Computing*, Cambridge: Cambridge University Press, 312–326.

Rao, C.R. (1962), "Efficient Estimates and Optimum Inference Procedures," *Journal of the Royal Statistical Society Series B*, **24**, 47–72.

Rao, C.R. (1963), "Criteria of Estimation in Large Samples," *Sankhyā*, **25**, 189–206.

Rao, R.R. (1960), *Some Problems in Probability Theory*, Calcutta: Unpublished doctoral thesis.

Rao, R.R. (1961), "On the Central Limit Theorem in R_k," *Bulletin of the American Mathematical Society*, **67**, 359–361.

Reid, N. (1988), "Saddlepoint Methods and Statistical Inference," *Statistical Science*, **3**, 213–238.

Reid, N. (1996), "Higher order results on the likelihood ratio statistic under model misspecification," *Canadian Journal of Statistics*, **?**, ?–?.

Robinson, J. (1982), "Saddlepoint approximations for permutation tests and confidence intervals," *Journal of the Royal Statistical Society Series B*, **44**, 91–101.

Routledge, R., and Tsao, M. (1995), "Uniform Validity of Saddlepoint Expansion on Complex Sets," *Canadian Journal of Statistics*, **23**, 425–431.

Rudin, W. (1973), *Functional Analysis*, New York: McGraw-Hill.

Rudin, W. (1976), *Principals of Mathematical Analysis*, New York: McGraw-Hill.

Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics*, New York: Wiley.

Skates, S.J. (1993), "On Secant Approximations to Cumulative Distribution Functions," *Biometrika*, **80**, 223–235.

Skovgaard, I.M. (1987), "Saddlepoint Expansions for Conditional Distributions," *Journal of Applied Probability*, **24**, 875–887.

Skovgaard, I.M. (1989), "A Review of Higher Order Likelihood Methods," *Bulletin of the International Statistical Institute: Proceedings of the Forty-seventh Session*, Paris: International Statistical Institute, **III**, 331–351.

Skovgaard, I.M. (1990), *Analytic Statistical Models*, Hayward, CA: Institute of Mathematical Statistics.

Stafford, J.E., and Andrews, D.F. (1993), "A symbolic algorithm for studying adjustments to the profile likelihood," *Biometrika*, **80**, 715–730.

Strawderman, R.L., Casella, G., and Wells, M.T. (1996), "Practical Small–Sample Asymptotics for Regression Problems," *Journal of the American Statistical Association*, **91**, 643–654.

Taniguchi, M. (1991), "Third-order asymptotic properties of a class of test statistics under a local alternative," *Journal of Multivariate Analysis*, **37**, 223–238.

Tanner, M.A. (1996), *Tools for Statistical Inference*, Heidelberg: Springer–Verlag.

Temme, N.M. (1982), "The Uniform Asymptotic Expansion of a Class of Integrals Related to Cumulative Distribution Functions," *SIAM Journal on Mathematical Analysis*, **13**, 239–252.

Thisted, R.A. (1988), *Elements of Statistical Computing*, New York: Chapman and Hall.

Tierney, L., and Kadane, J.B. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, **81**, 82–87.

Tierney, L., Kass, R.E., and Kadane, J.B. (1989), "Approximate Marginal Densities of Nonlinear Functions (Corr: V78 p233–234)," *Biometrika*, **76**, 425–433.

Turnbull, B.W., Brown, B.W., and Hu, M. (1974), "Survivorship Analysis of Heart Transplant Data," *Journal of the American Statistical Association*, **69**, 74–80.

Wallace, D.L. (1958), "Asymptotic Approximations to Distributions," *Annals of Mathematical Statistics*, **29**, 635–654.

Wang, S. (1990a), "Saddlepoint Approximations in Resampling Analysis," *Annals of the Institute of Statistical Mathematics*, **42**, 115–131.

Wang, S. (1990b), "Saddlepoint approximations for bivariate distributions," *Journal of Applied Probability*, **27**, 586–597.

Waterman, R.P., and Lindsay, B.G. (1996), "A Simple and Accurate Method for Approximate Conditional Inference in Linear Exponential Family Models," *Journal of the Royal Statistical Society Series B*, **?**.

Widder, D.V. (1946), *The Laplace Transform*, Princeton: Princeton University Press.

Withers, C.S. (1988), "Nonparametric Confidence Intervals for Functions of Several Distributions," *Annals of the Institute of Statistical Mathematics*, **40**, 727–746.

Wong, W.H. (1992), "On asymptotic efficiency in estimation theory," *Statistica Sinica*, **2**, 47–68.

Wong, W.H., and Li, B. (1992), "Laplace expansion for posterior densities of non-linear functions of parameters," *Biometrika*, **79**, 393–398.

Yang, B., and Kolassa, J.E. (2002), "Saddlepoint Approximation for the Distribution Function Near the Mean," *Annals of the Institute of Statistical Mathematics*, **54**, 743–747.

Yarnold, J. (1972), "Asymptotic Approximations for the Probability that a Sum of Lattice Random Vectors Lies in a Convex Set," *Annals of Mathematical Statistics*, **43**, 1566–1580.

Author Index

Subject Index

Lecture Notes in Statistics

For information about Volumes 1 to 132, please contact Springer-Verlag

160: Tomasz Rychlik, Projecting Statistical Functonals. viii, 184 pp., 2001.

161: Maarten Jansen, Noise Reduction by Wavelet Threholding. xxii, 224 pp., 2001.

162: Constantine Gatsonis, Bradley Carlin, Alicia Carriquiry, Andrew Gelman, Robert E. Kass, Isabella Verdinelli, and Mike West (Editors), Case Studies in Bayesian Statistics, Volume V. xiv, 448 pp., 2001.

163: Erkki P. Liski, Nripes K. Mandal, Kirti R. Shah, and Bikas K. Sinha, Topics in Optimal Design. xii, 164 pp., 2002.

164: Peter Goos, The Optimal Design of Blocked and Split-Plot Experiments. xiv, 244 pp., 2002.

165: Karl Mosler, Multivariate Dispersion, Central Regions and Depth: The Lift Zonoid Approach. xii, 280 pp., 2002.

166: Hira L. Koul, Weighted Empirical Processes in Dynamic Nonlinear Models, Second Edition. xiii, 425 pp., 2002.

167: Constantine Gatsonis, Alicia Carriquiry, Andrew Gelman, David Higdon, Robert E. Kass, Donna Pauler, and Isabella Verdinelli (Editors), Case Studies in Bayesian Statistics, Volume VI. xiv, 376 pp., 2002.

168: Susanne Rässler, Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches. xviii, 238 pp., 2002.

169: Yu. I. Ingster and Irina A. Suslina, Nonparametric Goodness-of-Fit Testing Under Gaussian Models. xiv, 453 pp., 2003.

170: Tadeusz Caliński and Sanpei Kageyama, Block Designs: A Randomization Approach, Volume II: Design. xii, 351 pp., 2003.

171: D.D. Denison, M.H. Hansen, C.C. Holmes, B. Mallick, B. Yu (Editors), Nonlinear Estimation and Classification. x, 474 pp., 2002.

172: Sneh Gulati, William J. Padgett, Parametric and Nonparametric Inference from Record-Breaking Data. ix, 112 pp., 2002.

173: Jesper Møller (Editor), Spatial Statistics and Computational Methods. xi, 214 pp., 2002.

174: Yasuko Chikuse, Statistics on Special Manifolds. xi, 418 pp., 2002.

175: Jürgen Gross, Linear Regression. xiv, 394 pp., 2003.

176: Zehua Chen, Zhidong Bai, Bimal K. Sinha, Ranked Set Sampling: Theory and Applications. xii, 224 pp., 2003

177: Caitlin Buck and Andrew Millard (Editors), Tools for Constructing Chronologies: Crossing Disciplinary Boundaries, xvi, 263 pp., 2004

178: Gauri Sankar Datta and Rahul Mukerjee, Probability Matching Priors: Higher Order Asymptotics, x, 144 pp., 2004

179: D.Y. Lin and P.J. Heagerty (Editors), Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data, vii, 336 pp., 2004

180: Yanhong Wu, Inference for Change-Point and Post-Change Means After a CUSUM Test, xiv, 176 pp., 2004

181: Daniel Straumann, Estimation in Conditionally Heteroscedastic Time Series Models, x, 250 pp., 2004

181: Daniel Straumann, Estimation in Conditionally Heteroscedastic Time Series Models, x, 250 pp., 2004

182: Lixing Zhu, Nonparametric Monte Carlo Tests and Their Applications, xi, 192 pp., 2005

183: Michel Bilodeau, Fernand Meyer, and Michel Schmitt (Editors), Space, Structure and Randomness, xiv, 416 pp., 2005

184: Viatcheslav B. Melas, Functional Approach to Optimal Experimental Design, vii., 352 pp., 2005

185: Adrian Baddeley, Pablo Gregori, Jorge Mateu, Radu Stoica, and Dietrich Stoyan, (Editors), Case Studies in Spatial Point Process Modeling, xiii., 324 pp., 2005

186: Estela Bee Dagum and Pierre A. Cholette, Benchmarking, Temporal Distribution, and Reconciliation Methods for Time Series, xiv., 410 pp., 2006