Edgeworth Expansions and Saddlepoint Approximations

Robert Zimmerman

Report for STA4508H - Topics in Likelihood Inference

April 16, 2022

1 Introduction

Let $U_1, U_2, \ldots, U_n \overset{iid}{\sim} F_U$ with $\mathbb{E}[U_i] = \mu$ and $\mathrm{Var}(U_i) = \sigma^2 < \infty$, and let $X_n = (\sum_i U_i - n\mu)/\sqrt{n\sigma^2}$ be the standardized sample mean of the U_i's. The classical (Lindeberg-Lévy) central limit theorem (CLT) — perhaps the most important theorem in statistics — says that $X_n \xrightarrow{d} Z \sim \mathcal{N}(0,1)$. The standard proof of this result uses characteristic functions, a technique dating to the work of Laplace [Fischer, 2010]: one simply writes $\varphi_{X_n}(t) = \left(\varphi_{(U_i-\mu)/\sigma}(t/\sqrt{n})\right)^n$ and expands the inner quantity as $1 - \sigma^2 t^2/2n + O(n^{-3/2})$, so that the whole expression converges pointwise to $e^{-t^2/2} = \varphi_Z(t)$. With pointwise convergence of the characteristic functions established, the proof concludes with an appeal to the inversion theorem [Lévy, 1926], which roughly asserts a one-to-one correspondence between distributions and characteristic functions.[1] Lévy proved that the convergence of F_{X_n} to Φ is uniform under relatively weak regularity conditions, and Gnedenko [1954] showed the same for the convergence of the densities f_{X_n} to ϕ.

The CLT was put on rigorous footing in the early 1900s, and since then, various extensions of it have been developed to the extent that modern probability theorists typically refer to "CLTs" instead of "*the* CLT". It is, of course, an asymptotic result; while it is easy to see that convergence is slower when F_U is highly skewed or kurtotic, the classical CLT gives no indication of how much $F_{X_n}(x)$ and $\Phi(x)$ will differ for finite n and for any given $x \in \mathbb{R}$. The study of convergence rates for the CLT began in the 1940s with Berry and Esseen, who independently established that $\sup_{x\in\mathbb{R}} |F_{X_n}(x) - \Phi(x)| \le C\rho/(\sigma^3\sqrt{n})$, where $\rho = \mathbb{E}\left[|U_i|^3\right]$ and C is a universal constant [Berry, 1941, Esseen, 1943, Bergström, 1944] which has been steadily reduced over the subsequent years. While such results are certainly informative, their applicability to practical statistical inference is limited because they concern only the *absolute error* of the approximation, rather than the *relative error*. Thus, as Wallace [1958] notes, "Berry's bound on the error is usually intolerable except for very large errors."

A different line of work, originating several decades earlier, carefully adds "correction terms" to $\Phi(x)$ in order to reduce the error of its approximation to $F_{X_n}(x)$. According to Wallace [1958], an *asymptotic expansion* of a sequence $\{f_n(\cdot) : \mathcal{X} \to \mathbb{R}\}_{n\ge 1}$ is a formal power series $\sum_{j=0}^{\infty} n^{-j/2} A_j(\cdot)$ whose partial sums provide an approximation of any $f_n(\cdot)$.[2] The asymptotic expansion is *valid* when $\left|f_n(x) - \sum_{j=0}^{r} n^{-j/2} A_j(x)\right| \le n^{-(r+1)} C_r(x)$ for any r, where $C_r(x)$ is constant in n. In other words — to paraphrase Wallace [1958] — validity means that the error committed by approximating $f_n(y)$ by the rth partial sum is of the same order of the $(r+1)$th neglected term. The label "formal" is applied to the power series because in many cases, $\sum_{j=0}^{\infty} n^{-j/2} A_j(x)$ will diverge for some or all x, and the utility of the power series arises from its partial sums, an insight that originates with Stirling, Euler, and Maclaurin [Dingle, 1973].

According to Fischer [2010], the application of asymptotic expansions to probability began with Bessel and was advanced by Chebyshev, who was apparently the first to exploit the Hermite polynomials.[3] It was Charlier and Edgeworth, working contemporaneously, who set the stage for the modern theory. Charlier, starting from Laplace's innovations and apparently unaware of the work of his contemporaries, developed a

[1] One can obtain a proof using moment generating functions in essentially the same fashion.

[2] While Barndorff-Nielsen and Cox [1989] provide a considerably more general definition, the definition here suffices for our exposition.

[3] Thiele and Hausdorff also made contributions, which were not properly recognized at the time [Fischer, 2010].

series expansion for general "error probabilities" which implied

$$f_{X_n}(x) = \phi(x)\left(1 + \sum_{j=3}^{\infty} \frac{c_j}{j!} H_j(x)\right),$$

where $H_j(\cdot)$ is the jth Hermite polynomial and the coefficients c_j depend on the cumulants κ_j and n. The coefficients satisfy $c_j = O(n^{-(j-2)/2})$ only for $j < 6$ and become quite irregular thereafter, rendering the so-called Charlier A-series (or the Charlier-Gram series) of limited use as an asymptotic expansion. Charlier's "proof" that his expansion was valid was deeply flawed [Cramér, 1972]; moreover, Szeg [1939] later showed that the series itself converges only when the underlying distribution F_U has quickly decaying tails.

At about the same time Edgeworth, who in turn was apparently unaware of Charlier's work [Fischer, 2010], established an expansion that was essentially a rearrangement of Charlier's series, taking the form

$$f_{X_n}(x) = \phi(x)\left(1 + \sum_{j=1}^{\infty} \frac{q_j(x)}{n^{j/2}}\right),$$

where each $q_j(\cdot)$ is a linear combination of $H_i(\cdot)$s and powers of κ_is, but importantly, is free of n [Edgeworth, 1905]. In contrast to Charlier's series, Edgeworth's series truly was valid under relatively weak conditions — a result first rigorously shown by Cramér [1928] — and the series proved ripe for extension and modification. Edgeworth also provided a glimpse into future statistical applications of his series. While Cramér added several theoretical results on Edgeworth expansions to the literature over the subsequent years, there were no major developments in the area of (statistical) asymptotic expansions for several more decades, with some notable exceptions: Daniels introduced the saddlepoint approximation (to be detailed below) in Daniels [1954], while Chambers [1967] developed Edgeworth expansions for the multivariate case in the relatively modern context of Monte Carlo simulations. The lone book on the topic was the reference book of Bhattacharya and Rao [1976], which compiled most results known at the time. It appears that the potential utility of Edgeworth expansions and saddlepoint approximations was not appreciated by the statistical community as a whole.

In 1979, Barndorff-Nielsen and Cox (henceforth referred to as "the authors") published a paper [Barndorff-Nielsen and Cox, 1979] whose stated goal was to (re)-introduce Edgeworth expansions and saddlepoint approximations in both the univariate and multivariate cases, and to emphasize their possible uses for statistical applications via several real-world examples (and a multitude of theoretical ones). We investigate their findings in this report. In Section 2, we derive the Edgeworth expansion and saddlepoint approximations presented by the authors in the univariate and bivariate cases. In Section 3, we examine the statistical applications provided by the authors, starting with four simple case studies and then moving onto deeper results related to conditional inference and likelihood ratio tests (LRTs). Finally, in Section 4 we provide a brief assessment of the paper and its impact on future research.

2 Edgeworth and Saddlepoint approximations

In their 1979 paper, the authors simply state the univariate Edgeworth expansion and briefly describe how the univariate saddlepoint approximation follows, with a view to preparing the reader for the bivariate and multivariate generalizations presented thereafter, from which the statistical applications of the approximations follow. As pointed out by Reid [1988], the primary challenge in deriving the multivariate extensions arises from introducing concise and readable notation for the error terms, because the number of terms in each summand explodes combinatorially with the dimension being considered.

In the two following subsections, we briefly explain how each approximation is derived in the univariate case, and then state the bivariate extension; together, these are enough to describe expansions of conditional densities, which are the main tools used in the statistical applications of the paper. We omit the general multivariate versions presented by the authors, since they are direct extensions of the bivariate ones. Throughout, we mostly maintain the original notation used by the authors despite some of their unconventional choices,[4] since consistently transcribing the mathematical content into modern form is quite challenging, and unnecessary for the purposes of this report. Much of their notation used in the multivariate setup was superseded by tensor notation (and Einstein notation in particular) by McCullagh [1987], which the authors themselves adopted in Barndorff-Nielsen and Cox [1989].

2.1 The direct Edgeworth expansion

Let $U_1, U_2, \ldots, U_n \overset{iid}{\sim} f_U(\cdot)$ with $\mathbb{E}[U] = \kappa_1$, $\mathrm{Var}(U) = \kappa_2$, higher-order cumulants $\kappa_3, \kappa_4, \ldots$, and cumulant generating function

$$K_U(t) = \log\left(\mathbb{E}\left[e^{tU}\right]\right) = \sum_{j=1}^{\infty} \kappa_j \frac{t^j}{j!}$$

If $X_n = (\sum_i U_i - n\kappa_1)/\sqrt{n\kappa_2}$ is the standardized sample mean, then the cumulant generating function of X_n is given by

$$K_{X_n}(t) = \sum_{j=2}^{\infty} \frac{n\kappa_j}{j!} \left(\frac{t}{\sqrt{n\kappa_2}}\right)^j = \frac{t^2}{2} + \rho_3 \frac{t^3}{6\sqrt{n}} + \rho_4 \frac{t^4}{24n} + O\Big(n^{-3/2}\Big),$$

where the standardized cumulants are defined as $\rho_j := \kappa_j / \kappa_2^{j/2}$, $j \geq 3$. To obtain the mgf, we exponentiate, extract a factor of $e^{t^2/2}$, write the remaining factor as a second-order Taylor series, and expand, relegating all powers of $1/\sqrt{n}$ beyond 3 to the error term:

$$\begin{aligned} M_{X_n}(t) &= \exp\left(\frac{t^2}{2}\right) \sum_{k=0}^{\infty} \frac{1}{k!} \left(\rho_3 \frac{t^3}{6\sqrt{n}} + \rho_4 \frac{t^4}{24n} + O\Big(n^{-3/2}\Big)\right)^k \\ &= \exp\left(\frac{t^2}{2}\right) \left[1 + \rho_3 \frac{t^3}{6\sqrt{n}} + \rho_4 \frac{t^4}{24n} + \rho_3^2 \frac{t^6}{72n} + O\Big(n^{-3/2}\Big)\right] \\ &= \mathcal{L}\{\phi H_0\}(-t) + \frac{\rho_3}{6\sqrt{n}} \mathcal{L}\{\phi H_3\}(-t) + \frac{\rho_4}{24n} \mathcal{L}\{\phi H_4\}(-t) + \frac{\rho_3^2}{72n} \mathcal{L}\{\phi H_6\}(-t) + O\Big(n^{-3/2}\Big) \end{aligned}$$

where $\mathcal{L}\{\phi H_j\}(t) = \int_{-\infty}^{\infty} e^{-tx} \phi(x) H_j(x)\,\mathrm{d}x$ is the (two-sided) Laplace transform of $\phi(x)H_j(x)$, and $H_j(x)$ is the jth Hermite polynomial.[5] Taking the inverse Laplace transform and factoring out $\phi(x)$ yields the density

$$f_{X_n}(x) = \phi(x) \left(1 + \frac{\rho_3}{6\sqrt{n}} H_3(x) + \frac{\rho_4}{24n} H_4(x) + \frac{\rho_3^2}{72n} H_6(x) + O\Big(n^{-3/2}\Big)\right). \tag{1}$$

In (1) we have the basic Edgeworth approximation for the standardized sample mean in one dimension (or as the authors call it, the *direct* Edgeworth expansion), which clearly agrees with the classical CLT

[4]For example, they regard the mgf of a random variable X as $M_X(t) = \mathbb{E}\left[e^{-tX}\right]$ instead of $M_X(t) = \mathbb{E}\left[e^{tX}\right]$, a choice that was gently criticized by Daniels in his discussion of the paper and tacitly revised by the authors in Barndorff-Nielsen and Cox [1989] only ten years later. Henry Daniels seemed to be amusingly particular about the notational choices made by his former student Cox. He also quibbles with the use of y for the size of a random sample in his contribution to the discussion of Cox [1958].

[5]The Laplace transforms – or Fourier transforms, when one uses characteristic functions instead of mgfs – are rarely written out explicitly like this in the literature on Edgeworth expansions; in the 1979 paper, the authors omit it entirely and simply call it "inverting". However, we find that this representation makes it easy to see how the inverse Laplace transform immediately recovers the density.

upon taking $n \to \infty$. Of course, the foregoing procedure is not valid for any arbitrary density $f_U(\cdot)$; in the appendix of the paper, the authors cite Bhattacharya and Rao [1976] for regularity conditions on the characteristic function of U which validate the expansions here and in the rest of the paper.

In two dimensions, let $(U_1, V_1), \ldots, (U_n, V_n) \overset{iid}{\sim} f_{U,V}(\cdot,\cdot)$, with cumulants κ_{lm} and standardized cumulants ρ_{lm}. Putting $X_n = (\sum_i U_i - n\kappa_{10})/\sqrt{n\kappa_{20}}$ and $Y_n = (\sum_i V_i - n\kappa_{01})/\sqrt{n\kappa_{02}}$, the mgf is exponentiated and manipulated in the same fashion as before. Applying this technique directly, however, presents a problem: in the univariate case, the term of order $O(\sqrt{n})$ in the expansion of $K_{X_n}(t)$ vanished because $\kappa_1(X_n) = \mathbb{E}[X_n] = 0$, but in $K_{X_n,Y_n}(t)$ it remains, because $\rho_{11} = \kappa_{11}/\sqrt{\kappa_{01}\kappa_{10}} \neq 0$ in general. To remedy this, the authors take a clever sidestep by defining $Y_n' = (Y_n - \rho_{11}X_n)/\sqrt{1-\rho_{11}^2}$, which is uncorrelated with X_n by construction. Then the basic Edgeworth expansion yields

$$f_{X_n,Y_n'}(x,y) = \phi_2(x,y)\left(1 + \frac{(\mathbf{H}^T\boldsymbol{\rho})^{[3]}(x,y)}{6\sqrt{n}} + \frac{(\mathbf{H}^T\boldsymbol{\rho})^{[4]}(x,y)}{24n} + \frac{\left[(\mathbf{H}^T\boldsymbol{\rho})^{[3]}\right]^2(x,y)}{72n} + O\left(n^{-3/2}\right)\right), \tag{2}$$

where and $\phi_2(\cdot,\cdot)$ is the standard bivariate normal density and

$$\left[(\mathbf{H}^T\boldsymbol{\rho})^{[j]}\right]^l(x,y) := \sum_{\substack{k_0+\cdots+k_j=l \\ k_0,\ldots,k_j\geq 0}} \binom{l}{k_0,\cdots,k_j} \prod_{i=0}^{j}\left[\binom{j}{i}\rho_{i,j-i}\right]^{k_i} H_{k_i i}(x)H_{k_i(j-i)}(y) \tag{3}$$

with $(\mathbf{H}^T\boldsymbol{\rho})^{[j]}(x,y) := \left[(\mathbf{H}^T\boldsymbol{\rho})^{[j]}\right]^1(x,y)$.[6] Letting $y' = (y-\rho_{11}x)/\sqrt{1-\rho_{11}^2}$, a simple change of variables applied to (2) gives the bivariate Edgeworth expansion

$$f_{X_n,Y_n}(x,y) = \frac{\phi_2(x,y')}{\sqrt{1-\rho_{11}^2}}\left(1 + \frac{(\mathbf{H}^T\boldsymbol{\rho}')^{[3]}(x,y')}{6\sqrt{n}} + \frac{(\mathbf{H}^T\boldsymbol{\rho}')^{[4]}(x,y')}{24n} + \frac{\left[(\mathbf{H}^T\boldsymbol{\rho}')^{[3]}\right]^2(x,y')}{72n} + O\left(n^{-3/2}\right)\right), \tag{4}$$

where ρ_{ij}' in the analogue of (3) is the (i,j)th standardized cumulant of (X_n, Y_n'). The authors note that the $\left[(\mathbf{H}^T\boldsymbol{\rho}')^{[j]}\right]^k(x,y')$ terms can be written in terms of (x,y), but the required transformation uses generalized Hermite polynomials and is not elaborated upon.

The authors end by noting briefly how the Edgeworth expansion of a conditional density arises from inserting the univariate and bivariate expansions into the right side of $f_{Y_n|X_n}(y \mid x) = f_{X_n,Y_n}(x,y)/f_{X_n}(x)$. This expansion is then given by the authors without any derivation whatsoever, but the details can be inferred[7] from the standard formula $\left(\sum_{k=0}^\infty b_k x^k\right)/\left(\sum_{k=0}^\infty a_k x^k\right) = a_0^{-1}\sum_{k=0}^\infty c_k x^k$, where $c_k = b_k - a_0^{-1}\sum_{j=1}^k c_{k-j}a_j$ (as stated in Gradshteyn and Ryzhik [2014], using $\sqrt{n}$ in place of x). Some tedious but straightforward accounting yields

$$f_{Y_n|X_n}(y \mid x) = \frac{\phi(y')}{\sqrt{1-\rho_{11}^2}}\left(1 + \frac{1}{6\sqrt{n}}\left((\mathbf{H}^T\boldsymbol{\rho}')^{[3]}(x,y') - \rho_{3,0}H_3(r)\right) + \frac{1}{n}A(x,y') + O\left(n^{-3/2}\right)\right), \tag{5}$$

where

$$A(x,y') = \frac{(\mathbf{H}^T\boldsymbol{\rho}')^{[4]}(x,y') - \rho_{40}H_3(x)}{24} + \frac{\left[(\mathbf{H}^T\boldsymbol{\rho}')^{[3]}\right]^2(x,y') - \rho_{30}^2H_6(x)}{72} - \frac{\rho_{30}H_3(x)(\mathbf{H}^T\boldsymbol{\rho}')^{[3]}(x,y') - \rho_{30}^2H_3^2(x)}{36}.$$

[6]The authors describe the implementation of (3) in words, rather than providing an explicit formula as we do here. It follows from two applications of the multinomial theorem, followed by the relevant adjustments to the involved Hermite polynomials.

[7]Typically when power series representations of functions are divided, care must be taken to ensure that the ratio actually converges in some neighborhood; however, in the case of asymptotic expansions, convergence is of no concern and the manipulations can be done in a rather carefree fashion.

2.2 The saddlepoint approximation

The Edgeworth expansions (1) and (4) certainly approximate f_{X_n} and f_{X_n,Y_n} more accurately than ϕ and ϕ_2, respectively, but their behaviour can be unpalatable in the tails of the distributions. For example, they can yield negative values in the tails — an infamous drawback noted by Daniels [1954]. Apart from the problem of negative values, the approximations can be quite poor away from the mean, which corresponds to $x = 0$ in (1); writing that equation as $f_{X_n}(x) = \phi(x)\left(1 + \rho_3 H_3(x)/6\sqrt{n} + O(n^{-1})\right)$ and noting that every non-constant polynomial is unbounded, we see that when $|x| \gg 0$, the convergence to 1 of the second term as $n \to \infty$ will be slowed considerably by the relatively large value of $H_3(x)$, unless one is fortunate enough to start with a distribution for which $\rho_3 = 0$ (i.e., with zero skewness) — but even then, the problem would remain for higher-order cumulants.

On the other hand, when $x = 0$ the latter issue completely disappears, simply because the odd Hermite polynomials vanish at 0. This results in a pleasing expansion in powers of n rather than $\sqrt{n}$; in particular, the one-term approximation is accurate to order $O(n^{-2})$. In turn, this suggests a modification: supposing that we are interested in approximating $f_{X_n}(x)$ at some particular $x = x^*$, construct a *new* density $g(\cdot)$ such that $g(0) = f_{X_n}(x^*)$ — but which otherwise preserves the basic characteristics of the original density $f_{X_n}(\cdot)$ — and then apply an Edgeworth expansion to $g(\cdot)$. At first glance, one might think to choose the shifted density $g(\cdot) = f_{X_n}(\cdot + x^*)$; such a choice would be doomed to fail, however, because then $X_n^* \sim g$ would have $\kappa_1(X_n^*) \neq 0$ in general, resulting in an extra term in its cumulant generating function.

We fare much better if instead we *exponentially tilt* $f_{X_n}(\cdot)$ — that is, we embed it inside the natural exponential family $\{f_{X_n}(\cdot;\lambda) : \lambda \in \mathbb{R}\}$ with $f_{X_n}(x;\lambda) = \exp(\lambda x - nK_U(\lambda)) f_{X_n}(x)$, in which the particular choice $\lambda = 0$ returns the original density. According to Butler [2007], this technique was first presented in Esscher [1932], and is sometimes called the *Esscher transform*. In slightly more generality (which will be relevant later) suppose that $U_1, \ldots, U_n \sim f_U(x;\lambda_0)$ with cumulants $\kappa_{1(0)}, \kappa_{2(0)}, \ldots$, so that now $X_n = (\sum_i U_i - n\kappa_{1(0)})/\sqrt{n\kappa_{2(0)}}$, and our aim is to approximate $f_{X_n}(x^*;\lambda_0)$. Putting $r = r(x) = \sqrt{n\kappa_{2(0)}}x + n\kappa_{1(0)}$, simple manipulations yield

$$f_{X_n}(x;\lambda_0) = \exp\left(nK_U(\lambda) - nK_U(\lambda_0) + r(\lambda - \lambda_0)\right) f_{X_n}(x;\lambda) \tag{6}$$

We are thus free to optimize the choice of λ, and since we intend to perform an Edgeworth expansion of $f_{X_n}(x;\lambda)$, the best choice λ^* is that which makes x^* equal to the mean of $f_{X_n}(\cdot;\lambda^*)$. Using (6), this means that we insist on

$$x^* = \int x \cdot \exp\left(-nK_U(\lambda^*) + nK_U(\lambda_0) - r(\lambda^* - \lambda_0)\right) f_{X_n}(x;\lambda_0)\,\mathrm{d}x = \sqrt{\frac{n}{\kappa_{2(0)}}}\left(K_U'(\lambda^*) + \kappa_{1(0)}\right), \tag{7}$$

or equivalently,

$$r^* = -nK_U'(\lambda^*), \tag{8}$$

where the involvement of the derivative $K_U'(\cdot)$ follows from a standard result about exponential families. Such a λ^* is found by solving (either exactly or approximately) the equation $K_U'(\lambda^*) = \sqrt{\kappa_{2(0)}/n}x^* + \kappa_{1(0)}$. The authors point out that from a statistical perspective this corresponds to finding the MLE of λ given the "data" x^* under the exponential family model $f_{X_n}(\cdot;\lambda)$; this correspondence will have consequences in hypothesis testing applications.

With λ^* chosen as such, an Edgeworth expansion of $f_{X_n}(x^*;\lambda^*)$ applied to (6) yields

$$\begin{aligned} f_{X_n}(x^*;\lambda_0) &= \exp\left(nK_U(\lambda^*) - nK_U(\lambda_0) + r^*(\lambda^* - \lambda_0)\right) \frac{\phi(0)}{\sqrt{\kappa_2^*/\kappa_{2(0)}}} \left(1 + \frac{\rho_4^*}{24n}H_4(0) + \frac{\rho_3^{*2}}{72n}H_6(0) + O(n^{-2})\right) \\ &= \frac{\exp\left(nK_U(\lambda^*) - nK_U(\lambda_0) - nK'(\lambda^*)(\lambda^* - \lambda_0)\right)}{\sqrt{2\pi\kappa_2^*/\kappa_{2(0)}}} \left(1 + \frac{3\rho_4^* - 5\rho_3^{2*}}{24n} + O(n^{-2})\right), \end{aligned} \tag{9}$$

and (9) is the basic saddlepoint approximation. As the authors note, there is an alternative method of deriving (9) which was shown by Daniels [1954]. One writes the density as the inverse Fourier transform of the characteristic function $\psi_{S_n^*}(t)$ and deforms the contour of integration to pass through the solution λ^* of (7), thereby defining a saddlepoint in the complex plane [Platt, 1996]. It is to this fact that the saddlepoint approximation owes its name (although the authors themselves also used the term "tilted Edgeworth expansion", which did not seem to catch on in the literature).

As with the univariate case, the bivariate saddlepoint approximation exploits the fact that the bivariate Edgeworth expansion (4) evaluated at the point $(0, 0)$ gives an approximation with error of order $O(n^{-1})$. Writing $\boldsymbol{\theta} = (\lambda, \psi)^\top$, we again begin by embedding $f_{X_n,Y_n}(\cdot, \cdot; \boldsymbol{\theta}_0)$ inside the natural exponential family $\{f_{X_n,Y_n}(\cdot, \cdot; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathbb{R}^2\}$, where $f_{X_n,Y_n}(x, y; \boldsymbol{\theta}) = \exp\left(\boldsymbol{\theta}^\top(x, y) - nK_{U,V}(\boldsymbol{\theta})\right) f_{X_n,Y_n}(x, y; \boldsymbol{\theta}_0)$. Now putting $r = r(x) = \sqrt{n\kappa_{20(0)}}x + n\kappa_{10(0)}$ and $s = s(y) = \sqrt{n\kappa_{02(0)}}y + n\kappa_{01(0)}$, we have in analogy with (6) that

$$f_{X_n,Y_n}(x, y; \boldsymbol{\theta}_0) = \exp\left(nK_{U,V}(\boldsymbol{\theta}) - nK_{U,V}(\boldsymbol{\theta}_0) + (r, s)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\right) f_{X_n,Y_n}(x, y; \boldsymbol{\theta}). \tag{10}$$

Aiming to approximate (10) at the point $(x^*, y^*)^\top$, we now choose $\boldsymbol{\theta}^*$ such that

$$\begin{pmatrix} x^* \\ y^* \end{pmatrix} = \sqrt{n}\left(\nabla K_{U,V}(\boldsymbol{\theta}^*) - \begin{pmatrix} \kappa_{10(0)} \\ \kappa_{01(0)} \end{pmatrix}\right) \odot \begin{pmatrix} \sqrt{\kappa_{20(0)}} \\ \sqrt{\kappa_{02(0)}} \end{pmatrix}, \tag{11}$$

or equivalently $\mathbf{t}^* = -n\nabla K_{U,V}(\boldsymbol{\theta}^*)$, where $\mathbf{t}^* := (r^*, s^*)^\top$. Applying the bivariate Edgeworth expansion (4) of $f_{X_n,Y_n}(x, y; \boldsymbol{\theta}^*)$ to the rightmost term in (10) then yields

$$f_{X_n,Y_n}(x^*, y^*; \boldsymbol{\theta}_0) = \frac{\exp\left(nK_{U,V}(\boldsymbol{\theta}^*) - nK_{U,V}(\boldsymbol{\theta}_0) + \mathbf{t}^T(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)\right)}{2\pi\sqrt{|\nabla\nabla^T K_{U,V}(\boldsymbol{\theta}^*)| \,/\, |\nabla\nabla^T K_{U,V}(\boldsymbol{\theta}_0)|}} \left(1 + \frac{(\mathbf{H}^T \boldsymbol{\rho}^{*\prime})^{[4]}(x^*, y^{*\prime})}{24n} + O(n^{-2})\right). \tag{12}$$

Obtaining a saddlepoint approximation for the conditional density $f_{Y_n|X_n}(\cdot \mid \cdot; \boldsymbol{\theta}_0)$ is less straightforward. Assuming an exponentially-tilted joint density of the form

$$f_{X_n,Y_n}(x, y; \boldsymbol{\theta}_0) = \exp\left(nK_{U,V}(\lambda'', \psi'') - nK_{U,V}(\lambda_0, \psi_0) + \mathbf{t}^\top(\boldsymbol{\theta}'' - \boldsymbol{\theta}_0)\right) f_{X_n,Y_n}(x, y; \lambda'', \psi'') \tag{13}$$

for some λ'', ψ'' as in (10) automatically imposes a marginal density of the form

$$f_{X_n}(x; \boldsymbol{\theta}_0) = \exp\left(nK_{U,V}(\lambda', \psi_0) - nK_{U,V}(\lambda_0, \psi_0) + r(\lambda' - \lambda_0)\right) f_{X_n}(x; \lambda', \psi_0) \tag{14}$$

for some λ'. There are thus several ways to choose the free parameters $\lambda', \lambda'', \psi''$ that are involved; moreover, there is nothing to preclude combining an Edgeworth expansion of $f_{X_n,Y_n}(\cdot, \cdot; \boldsymbol{\theta}_0)$ with a saddlepoint approximation of $f_{X_n}(\cdot; \boldsymbol{\theta}_0)$, or vice-versa.

The authors discuss two approaches to evaluating the conditional density $f_{X_n,Y_n}(\cdot \mid \cdot; \boldsymbol{\theta}_0)$ at $(y^* \mid x^*)$. The first approach applies separate saddlepoint approximations to $f_{X_n}(x; \lambda', \psi_0)$ and $f_{X_n,Y_n}(x, y; \lambda'', \psi'')$, which means choosing $(\lambda'', \psi'')^\top = \boldsymbol{\theta}^*$ as implicitly defined in (11). For the remaining parameter, we choose the $\lambda' = \lambda^*_{(0)}$ obtained by solving (7) in λ^*, but with the expectation now taken with respect to $f_{X_n}(\cdot; \lambda, \psi_0)$, holding ψ_0 fixed; equivalently, $\lambda^*_{(0)}$ is the MLE of λ under the hypothesis $\psi = \psi_0$ [Pedersen, 1979]. With these choices respectively inserted into (13) and (14), we divide the former by the latter and obtain

$$\begin{aligned} &f_{Y_n|X_n}(y^* \mid x^*; \boldsymbol{\theta}_0) \\ &= \frac{\exp\left(nK_{U,V}(\lambda^*, \psi^*) - nK_{U,V}(\lambda^*_{(0)}, \psi_0) + r^*(\lambda^* - \lambda^*_{(0)}) + s^*(\psi^* - \psi_0)\right)}{\sqrt{2\pi\, |\nabla\nabla^T K_{U,V}(\psi^*, \lambda^*)| \,/\, \kappa^*_{20(0)}\kappa_{02(0)}}} \left(1 + \frac{B(r^*, s^{*\prime})}{24n} + O(n^{-2})\right), \end{aligned} \tag{15}$$

where $B(r, s') = (\mathbf{H}^T \hat{\boldsymbol{\rho}}')^{[4]}(r, s') - 3\hat{\rho}_{4,0} + 5\hat{\rho}^2_{3,0}$, the latter two terms being calculated with respect to $\lambda^*_{(0)}$, as in (9). The authors refer to this equation as the *double saddlepoint approximation*.

The second approach instead chooses $\psi'' = \psi_0$ and $\lambda' = \lambda'' = \lambda^*_{(0)}$, where $\lambda^*_{(0)}$ is again obtained as the MLE of λ under $\psi = \psi_0$. These choices render the exponential terms in (13) and (14) *identical*, so that they entirely cancel one another out upon division, leaving us with

$$f_{Y_n|X_n}\left(y^* \mid x^*; \boldsymbol{\theta}_0\right) = \frac{f_{X_n,Y_n}\left(x^*, y^*; (\lambda^*_{(0)}, \psi_0)\right)}{f_{X_n}\left(x^*; (\lambda^*_{(0)}, \psi_0)\right)}.$$

We now apply Edgeworth expansions to both numerator and denominator, yielding an expansion in powers of $1/\sqrt{n}$. The authors do not state the resulting expression in the paper, but it is given by Pedersen [1979], and is essentially obtained via (4) and (1) and the same division rule for infinite series that was noted toward the end of Section 2.1. The authors call this variation the *single saddlepoint approximation* (or the *mixed Edgeworth saddlepoint approximation*).

3 Applications to statistical inference

Throughout their development of the Edgeworth expansions and saddlepoint approximations, the authors offer several brief examples of these methods applied to simple densities, in order to provide intuition to the reader rather than to showcase the power of the approximations in real-life statistical situations. For example, after introducing the univariate saddlepoint approximation, the authors repeat a popular example given by Daniels [1954]: if $Y_1, \dots, Y_n \overset{iid}{\sim} \text{Exp}(1)$, then the saddlepoint approximation to the density of $n\bar{Y}_n \sim \text{Gamma}(1, n)$ is

$$f_{n\bar{Y}_n}(y) = \frac{y^{n-1}e^{-y}}{\sqrt{2\pi/n}(n/e)^n}\left(1 + O(n^{-1})\right).$$

Remarkably, the denominator in the leading term is Stirling's approximation of $(n-1)!$, and moreover the density would be exact upon renormalization.[8]

Following their exposition of the double and single saddlepoint approximations in the bivariate case, the authors give two more simple examples of how their approximations of conditional densities lead to expressions which are exact, up to a normalizing constant. Interesting as they are, basic examples like these — in which the density being approximated is already available in closed form for comparison — are by construction unsuitable in any real-world situation, simply because no one would want to use an approximation when the exact result is already at hand. Hence, we omit further discussion of these, and instead, we report on the statistical applications provided by the authors in order to highlight the power of the approximations to statistical inference.

3.1 Some special cases of statistical interest

Following their main exposition of the Edgeworth and saddlepoint approximations, the authors demonstrate their statistical applicability with four relatively simple examples, all of which deal with exponential families.

The first involves modelling a time-inhomogeneous Poisson process in which, given event occurrences at times $0 < t_1 < \cdots < t_n < T$, one seeks to test whether a log-quadratic rate function more suitably describes the data than a log-linear one. To this end, the authors test the hypothesis that $\psi = 0$ in the rate function $\exp\left(\alpha + \beta t + \psi t^2\right)$, conditional on n events having occurred in the interval $(0, T)$.[9] Conditioning on the

[8] In his discussion of the paper, Daniels asserted that the only densities for which the renormalized saddlepoint approximation is exact are the normal, inverse Gaussian, and gamma. A rigorous proof of this claim was published several years later by Blæsild and Jensen [1985].

[9] Notably, Pedersen [1979] applies the single saddlepoint approximation to the same model in a more detailed study.

number of events $r = \sum_{i=1}^n t_i$ effectively fixes the parameter α, reducing the setup to a two-dimensional problem with sufficient statistic $\left(\sum_{i=1}^n t_i, \sum_{i=1}^n t_i^2\right)$. The authors apply the direct Edgeworth expansion in (5) to $s = \sum_{i=1}^n t_i^2$, yielding an approximate normal density from which confidence intervals and p-values can be extracted. The authors apply this method to the record of major freezes of Lake Constance in Western Europe, described in Steinijans [1976], who had performed the same hypothesis test on this data under an unrefined bivariate normal approximation to $\left(\frac{1}{n}\sum_{i=1}^n t_i, \frac{1}{n}\sum_{i=1}^n t_i^2\right)$. While both studies reject the hypothesis that $\psi = 0$, the p-value obtained using the double saddlepoint approximation is about 0.008, while that from the bivariate normal approximation is effectively zero at $\Phi^{-1}(-25.2)$. The authors attribute this difference to Steinijans' application of the bivariate normal approximation to an inappropriate range. In fact, the discussion of the authors' paper includes a note from a representative of the Forestry Commission confirming an error in a formula used by Steinijans; once the mistake is accounted for, the discrepancy between the two p-values is quite minimal.

The next simple example involves the von Mises distribution (or as the the authors call it, the *circular normal distribution*). The aim is to compare the "basic" distribution, where $f(\alpha) \propto \exp(\lambda_1 \cos\alpha + \lambda_2 \sin\alpha)$, with an extension[10] where $f(\alpha) \propto \exp(\lambda_1 \cos\alpha + \lambda_2 \sin\alpha + \psi_1 \cos 2\alpha + \psi_2 \sin 2\alpha)$. As in the previous example, given iid data $A_1, A_2, \ldots, A_n$, the authors obtain an approximate conditional density of the random vector $(\sum_{i=1}^n \cos 2A_i, \sum_{i=1}^n \sin 2A_i)$ given $r = (\sum_{i=1}^n \cos A_i, \sum_{i=1}^n \sin A_i)$ by applying a 4-dimensional extension of the single saddlepoint approximation. Following some rather tedious calculations, the authors show that to a first-order approximation, what results is a bivariate normal density; from this two statistics can be extracted, which, under the assumption $\psi_1 = \psi_2 = 0$, are independently normally distributed with known variances. From these, an approximate hypothesis test of $\psi_1 = \psi_2 = 0$ can thus be constructed on the basis of a $\chi^2_{(2)}$ test statistic.

The third example involves testing for conformity with the gamma distribution (with unknown shape and scale) – an exponential family distribution with sufficient statistics $(\sum_{i=1}^n X_i, \sum_{i=1}^n \log(X_i))$ – against a more complicated exponential family distribution with sufficient statistics $\left(\sum_{i=1}^n X_i, \sum_{i=1}^n \log(X_i), \sum_{i=1}^n X_i^2\right)$. Once again, the authors use the single saddlepoint method to approximate a conditional density, this time of $\sum_{i=1}^n X_i^2 \mid (\sum_{i=1}^n X_i, \sum_{i=1}^n \log(X_i))$. To a first-order approximation, what results is a univariate normal density; thus, a statistic can be constructed which follows a standard normal distribution, yielding a straightforward z-test. Interestingly, the numerator of the test statistic contains a difference of the sample squared coefficient of variation and the MLE of that under the hypothesized gamma distribution; the authors therefore refer to this construction as a "dispersion test". However, they also note that the test performs rather slowly, based on the results of a simulation study by their colleague D. Pregibon, and hence they do not advise its use in practical situations.

The last example pertains to a bioassay logistic regression model, in which three dose levels $d \in \{-1, 0, 1\}$ are administered to n patients each, yielding responses of $Y_1^{(d)}, \ldots, Y_n^{(d)} \overset{iid}{\sim} \text{Bernoulli}(\text{expit}(-\lambda - \psi d))$, for each d. To perform inference on the slope parameter ψ, the authors define the statistics $A_d = \sum_{i=1}^n Y_i^{(d)}$ and find that the joint probability mass function of (A_{-1}, A_0, A_1) – which is simply a product of binomial masses – can be written in terms of the total number of respondents $R = \sum_{d=-1}^1 A_d$ and the difference $S = A_1 - A_{-1}$, specifically as a function of $\lambda R - \psi S$. From here, the authors are finally in a position to apply their double saddlepoint method, and by doing so they obtain an approximation to the conditional mass function of $S \mid R$. The exact conditional distribution is available in closed form, which is compared with the double saddlepoint approximation in a table with $n = 16$ and $\lambda = 1$; however, no derivations of the approximation (nor of the exact mass function) are included. In most cases, the relative errors of the saddlepoint approximation lie between 2% and 7%.

[10] Gatto and Jammalamadaka [2007] refers to this as a "Generalized von Mises" distribution, itself a specification of the general family of densities proportional to exponentials of trigonometric polynomials $f(\alpha) \propto \exp\left(\sum_{j=1}^k a_j \cos j\alpha + b_j \cos j\alpha\right)$

3.2 Applications to conditional likelihood inference and likelihood ratio tests

The authors' first major statistical example pertains to conditional likelihood inference, in which inference for one subset of unknown model parameters is performed while regarding the remaining unknowns as nuisance parameters. In exponential family models, nuisance parameters can be removed via conditioning on sufficient statistics [Davison, 2003]. Thus, *approximate* conditional inference can be carried out by leveraging approximate conditional densities via the double saddlepoint approximation. The authors demonstrate this for the two-parameter case, which aligns with the discussion that leads to (15).

With the same setup — but dropping the subscripts on λ_0 and ψ_0 for notational simplicity — we are now interested in approximating the conditional likelihood function of ψ in terms of the observed sufficient statistics $s = \sum_i Y_i$ and $r = \sum_i X_i$, conditioning on the latter equality and regarding λ as the nuisance parameter. Writing $\hat{\lambda}_\psi$ for the MLE of λ holding ψ fixed, taking the logarithm of the leading term of (15) yields an approximate conditional log-likelihood

$$\ell(\psi; s \mid r) \approx \frac{1}{2}\log(\kappa_{20}) - nK_{U,V}(\hat{\lambda}_\psi, \psi) - r\hat{\lambda}_\psi - s\psi + C, \tag{16}$$

where κ_{20} is calculated based on $(\hat{\lambda}_\psi, \psi)$ and C is free of λ. Differentiating with respect to ψ yields an approximate conditional likelihood equation of the form

$$n\kappa_{01} = s - \frac{\kappa_{11}\kappa_{30} - \kappa_{20}\kappa_{21}}{2\kappa_{20}^2} \tag{17}$$

where the joint cumulants κ_{ij} are evaluated at $(\hat{\lambda}_\psi, \psi)$. The authors observe that (17) is essentially the unconditional likelihood equation from (7), adjusted by the correction term on the right.[11] To illustrate, the authors consider a 2×2 contingency table with entries x_{ij} and unknown cell probabilities π_{ij}. With the parameters $\lambda = (\log(\pi_{22}/\pi_{12}), \log(\pi_{22}/\pi_{21}))$ and $\psi = \log(\pi_{12}\pi_{21}/\pi_{11}\pi_{22})$, the authors define sufficient statistics $s = x_{11}$ and $\mathbf{r} = (x_{1\cdot}, x_{\cdot,1})$ and work out the details to determine the approximate conditional likelihood equation $\ell(\psi; s \mid \mathbf{r})$, showing that obtaining the required maximizer of ψ essentially amounts to solving a quadratic equation in π_{11}. They apply the idea to a study by Fisher [Fisher, 1935, 1962] on twins of criminals and demonstrate that for that particular application, the approximate conditional likelihood equation (as a function of ψ) — and its maximizer $\hat{\psi}$ — are nearly indistinguishable from their exact counterparts.

Another important application of the saddlepoint approximation involves approximating the distribution of the LRT statistic under the null hypothesis. The result is in essence a generalization of Wilks' theorem. Starting with the assumption that $Y_1, \ldots, Y_n$ are independent and identically distributed according to the exponentially tilted density $f(y;\lambda) = e^{-y\lambda}f(y)/M(\lambda)$, the goal is to test $H_0 : \lambda = \lambda_0$ using an LRT. The LRT statistic is

$$P_n = -2n\bar{Y}_n(\lambda - \lambda_0) - 2n(K(\lambda) - K(\lambda_0))$$

and $K'(\lambda) = -\bar{Y}_n$. We write p for the observed value of P_n. From (9) and two changes of variables (first from S_n^* to $n\bar{Y}_n$, and then from $n\bar{Y}_n$ to P_n), we obtain the saddlepoint approximation

$$f_{P_n}(p; \lambda_0) = \frac{e^{-p/2}}{2\sqrt{2\pi n}} \sum_{\lambda \in A_p} \frac{1}{|\lambda - \lambda_0|\sqrt{K''(\hat{\lambda})}} \left(1 + \frac{c}{n} + O(n^{-2})\right),$$

where c is a function of λ_0 and $A_p = \{\lambda : -2n\bar{y}_n(\lambda - \lambda_0) - 2n(K(\lambda) - K(\lambda_0)) = p\}$, and further manipulations[12] yield

$$f_{P_n}(p; \lambda_0) = \frac{e^{-p/2}}{\sqrt{2\pi p}} \left(1 + \frac{c(1-p)}{n} + O(n^{-3/2})\right).$$

[11] It is rather unclear how the authors arrive at this equation, despite their hint that it follows from differentiation of (16) with respect to ψ. We could not find a reproduction of this formula in our review of the subsequent literature.

[12] The authors simply refer to these manipulations as "Taylor expansions" without providing details. This is quite an understatement of the fully rigorous argument, given by, e.g., Theorem 8.2.1 of Kolassa [2006].

One defect of the above approximation is that its mean is $1 - 2c/n + O(n^{-3/2})$, which, although correct in the limit, is somewhat unsatisfactory as an approximation of the true mean. However, making the further change of variables from P_n to $P_n' := P_n/(1-c/n)$ and then Taylor expanding (to second-order) the resulting exponential and the square-root terms yields

$$f_{P_n'}(p';\lambda_0) = \frac{e^{-p'/2}}{\sqrt{2\pi p'}} e^{-cp'/2n} \sqrt{1-c/n} \left(1 + \frac{c(1-p(1-c/n))}{n} + O(n^{-3/2})\right) = \frac{e^{-p'/2}}{\sqrt{2\pi p'}} \left(1 + O(n^{-3/2})\right).$$

Thus, $P_n' \dot\sim \chi^2_{(1)}$ to order $n^{-3/2}$, and indeed, this approximation evidently has a much more palatable mean of $1 + O(n^{-3/2})$. The authors generalize the above slightly to the two-parameter case. The idea of rescaling the LRT statistic — in this case, by the factor $(1 - c/n)$ — to produce a better approximation to the asymptotic chi-square distribution dates back to Bartlett [1937]; the factor is usually called the *Bartlett correction* in his honour. This derivation via the saddlepoint approximation is considerably simpler than the notoriously complicated method published earlier by Lawley [1956].

4 Discussion

This is a well-written, deep, and groundbreaking paper. While relatively little of the underlying theory included was completely new, the paper was perhaps the first to compile the basic results on Edgeworth expansions and saddlepoint approximations into a concise and complete exposition, emphasizing their potential for statistical uses — particularly with regard to conditional inference and likelihood ratio tests for exponential family models. The greatest contributions of the paper are the new connections to statistical inference, particularly to conditional likelihood (which we addressed in STA4508H) and to LRTs.

From the outset, the paper was very well-received — all of the eminent statisticians who contributed to the discussion acknowledged its value. It has also been highly influential; it has been cited almost ubiquitously by subsequent publications on Edgeworth expansions and saddlepoint approximations, and it helped to provoke an explosion of work in the 1980s, especially on approximating distributions themselves in the tails. Edgeworth expansions and saddlepoint approximations have been used in countless statistical applications since then, and at least three textbooks on the topic of such asymptotic expansions have been published since 1990. The research continues — for example, Tang and Reid [2021] have recently studied the saddlepoint approximation in high dimensions.

The paper does have one major shortcoming: its lack of detailed calculations. While explaining every step would be cumbersome, one wishes that the authors had included more detail in the appendix. In general, the derivation of the saddlepoint approximation is explained in too little detail, and is not much improved in Barndorff-Nielsen and Cox [1989]. The steps, which are not that lengthy, are much more clearly explained in Wallace [1958] and Reid [1988]. Other results provided without details, such as (17), are simply mystifying. In fairness, the publication constraints of the day may have precluded a great deal of detail; it was impossible to submit appendices of arbitrary length for online access, as we would do today. A second, more minor shortcoming of the paper involves the authors' unconventional notational choices, many of which they revised in their 1989 monograph on the subject.

Only by actually tracing through the mathematical details in works like this is it possible to truly appreciate how much tedious algebra had to be carried out behind the scenes in the days before computational tools. For example, Wallace [1958] notes that determining just one fourth-order term for an expansion related to the Behrens-Fisher problem required one hundred pages of algebra by Welch [Aspin, 1948]. Barndorff-Nielsen and Cox seem to understate the amount of computational work involved, presenting a real challenge to novices grappling with this landmark work. Then again, for mathematical geniuses like the authors, perhaps it really was that easy.

References

Hans Fischer. *A history of the central limit theorem: From classical to modern probability theory*. Springer Science & Business Media, 2010. 1, 2

Paul Lévy. Calcul des probabilités. *Revue de Métaphysique et de Morale*, 33(3), 1926. 1

BV Gnedenko. A local limit theorem for densities. In *Doklady Akad. Nauk SSSR (NS)*, volume 95, pages 5–7, 1954. 1

Andrew C Berry. The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the american mathematical society*, 49(1):122–136, 1941. 1

Carl-Gustav Esseen. *Determination of the maximum deviation from the Gaussian law*. Almqvist & Wiksell, 1943. 1

Harald Bergström. On the central limit theorem. *Scandinavian Actuarial Journal*, 1944(3-4):139–153, 1944. 1

David L Wallace. Asymptotic approximations to distributions. *The Annals of Mathematical Statistics*, 29(3):635–654, 1958. 1, 10

O.E. Barndorff-Nielsen and D.R. Cox. *Asymptotic techniques for use in statistics*. Monographs on statistics and applied probability (Series). Chapman and Hall, London, 1989. ISBN 0412314002. 1, 3, 10

Robert B Dingle. *Asymptotic expansions: their derivation and interpretation*. Academic Press, 1973. 1

Harald Cramér. Studies in the history of probability and statistics. xxviii on the history of certain expansions used in mathematical statistics. *Biometrika*, 59(1):205–207, 1972. 2

Gabor Szeg. *Orthogonal polynomials*, volume 23. American Mathematical Soc., 1939. 2

Francis Ysidro Edgeworth. The law of error. *Transactions of the Cambridge Philosophical Society*, 20:36–65,113–41, 1905. 2

Harald Cramér. On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal*, 1928(1):13–74, 1928. 2

Henry E Daniels. Saddlepoint approximations in statistics. *The Annals of Mathematical Statistics*, pages 631–650, 1954. 2, 5, 6, 7

John M Chambers. On methods of asymptotic approximation for multivariate distributions. *Biometrika*, 54(3-4): 367–383, 1967. 2

Rabi N Bhattacharya and R Ranga Rao. *Normal approximation and asymptotic expansions*. Wiley, 1976. 2, 4

Ole Barndorff-Nielsen and David R Cox. Edgeworth and saddle-point approximations with statistical applications. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(3):279–299, 1979. 2

Nancy Reid. Saddlepoint methods and statistical inference. *Statistical science*, pages 213–227, 1988. 2, 10

David R Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232, 1958. 3

Peter McCullagh. *Tensor methods in statistics*. Chapman and Hall/CRC, 1987. 3

Izrail Solomonovich Gradshteyn and Iosif Moiseevich Ryzhik. *Table of integrals, series, and products*. Academic press, 2014. 4

Ronald W Butler. *Saddlepoint approximations with applications*, volume 22. Cambridge University Press, 2007. 5

Fredrik Esscher. On the probability function in the collective theory of risk. *Skand. Aktuarie Tidskr.*, 15:175–195, 1932. 5

Robert William Platt. *An evaluation of saddlepoint approximations in the generalized linear model*. University of Washington, 1996. 6

Bo V Pedersen. Approximating conditional distributions by the mixed edgeworth-saddlepoint expansion. *Biometrika*, 66(3):597–604, 1979. 6, 7

Preben Blæsild and Jens Ledet Jensen. Saddlepoint formulas for reproductive exponential models. *Scandinavian journal of statistics*, pages 193–202, 1985. 7

Volker W Steinijans. A stochastic point-process model for the occurrence of major freezes in lake constance. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 25(1):58–61, 1976. 8

Riccardo Gatto and Sreenivasa Rao Jammalamadaka. The generalized von mises distribution. *Statistical Methodology*, 4(3):341–353, 2007. 8

Anthony Christopher Davison. *Statistical models*, volume 11. Cambridge university press, 2003. 9

Ronald A Fisher. The logic of inductive inference. *Journal of the Royal Statistical Society*, 98(1):39–82, 1935. 9

Ronald A Fisher. Confidence limits for a cross-product ratio. *Australian Journal of Statistics*, 1962. 9

John E Kolassa. *Series approximation methods in statistics*, volume 88. Springer Science & Business Media, 2006. 9

Maurice Stevenson Bartlett. Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, 160(901):268–282, 1937. 10

Derrick N Lawley. A general method for approximating to the distribution of likelihood ratio criteria. *Biometrika*, 43 (3/4):295–303, 1956. 10

Yanbo Tang and Nancy Reid. Laplace and saddlepoint approximations in high dimensions. *arXiv preprint arXiv:2107.10885*, 2021. 10

Alice A Aspin. An examination and further development of a formula arising in the problem of comparing two mean values. *Biometrika*, 35(1/2):88–96, 1948. 10