1

Fundamental approximations

This chapter introduces the basic saddlepoint approximations without the burden of undue formality or rigor. The chapter is designed to present the most fundamental and straightforward of the techniques with some associated factual information so the reader may begin using the methods in practical settings with only minimal theoretical understanding. For this reason, the presentation emphasizes implementation and computation and keeps theoretical discussion to a minimum.

1.1 Univariate densities and mass functions

1.1.1 Preliminaries

The most fundamental saddlepoint approximation was introduced by Daniels (1954) and is essentially a formula for approximating a density or mass function from its associated moment generating function or cumulant generating function. We present this expression after first defining the relevant quantities involved.

Suppose continuous random variable X has density $f(x)$ defined for all real values of x. The *moment generating function* (*MGF*) of density f (also for X) is defined as the expectation of $\exp(sX)$ or

$$M(s) = E\left(e^{sX}\right) = \int_{-\infty}^{\infty} e^{sx} f(x)\, dx$$

over values of s for which the integral converges. With real values of s, the convergence is always assured at $s = 0$. In addition, we shall presume that M converges over an open neighborhood of zero designated as (a, b), and that, furthermore, (a, b) is the largest such neighborhood of convergence. This presumption is often taken as a requirement for the existence of the MGF in many textbooks. For simplicity we shall conform to this convention but then relax the assumption when later applications demand so. The *cumulant generating function* (*CGF*) of f (also X) is defined as

$$K(s) = \ln M(s) \qquad s \in (a, b). \tag{1.1}$$

The terminology arises from the fact that the Taylor series coefficients of M and K (the collection of higher order derivatives evaluated at $s = 0$) give rise to the moments and cumulants respectively for random variable X. A simple exercise shows that the kth derivative of M evaluated at $s = 0$ is $M^{(k)}(0) = E(X^k)$, the kth moment of X. Also, using these

results, one can easily show that the first two cumulants of X are given as $K'(0) = E(X)$ and $K''(0) = \text{var}(X)$ with higher order derivatives leading to the more complicated higher order cumulants.

When random variable X is discrete and integer valued rather than continuous, the same functions can be defined and we mention only those aspects that differ. Assume X has mass function $p(k)$ for integer k and define its MGF as

$$M(s) = E\left(e^{sX}\right) = \sum_{k=-\infty}^{\infty} e^{sk} p(k) \qquad s \in (a, b),$$

with (a, b) as the maximal neighborhood of convergence about zero. The CGF is defined as in (1.1) and the moments and cumulants are again the Taylor coefficients of M and K.

One of the more important facts discussed in probability courses and proven in mathematical analysis is the one-to-one (1-1) correspondence that exists between the collection of probability density and mass functions and their associated MGFs (assuming the latter exist). Indeed, the distribution of a random variable is often determined using this correspondence by recognizing the MGF as that associated with a particular density or mass function. Unfortunately, settings where such recognition may be used are more often the exception than the rule. Quite often MGFs or CGFs of random variables can be determined out of context but their density/mass functions cannot. In such instances, highly accurate approximations to these density/mass functions can be computed by using the saddlepoint methods introduced in Daniels (1954) which are based entirely on the known CGFs.

An example of this is the determination of the distribution for a sum of independent random variables. Suppose $X_1, \ldots, X_n$ is a sequence of independent variables for which X_i has MGF M_i and CGF K_i defined over (a_i, b_i). The CGFs of $X = \sum_{i=1}^n X_i$ and $\bar{X} = X/n$ are

$$K_X(s) = \sum_{i=1}^{n} K_i(s) \qquad K_{\bar{X}}(s) = \sum_{i=1}^{n} K_i(s/n) \qquad s \in (\max_i a_i, \min_i b_i).$$

Recognition of the form for the CGFs specifies their associated distributions. For example, if X_i has the Binomial (m_i, θ) mass function

$$p_i(k) = \binom{m_i}{k} \theta^k (1-\theta)^{m_i - k} \qquad k = 0, \ldots, m_i \tag{1.2}$$

with CGF

$$K_i(s) = m_i \ln\{\theta(e^s - 1) + 1\} \qquad s \in (-\infty, \infty), \tag{1.3}$$

then $K_X(s) = \sum_{i=1}^n K_i(s)$ is recognized from (1.3) as the CGF of a Binomial $(m_\cdot, \theta)$ mass function with $m_\cdot = \sum_i m_i$. Uniqueness of the CGF specifies that X must have this distribution and this is indicated by writing $X \sim \text{Binomial}(m_\cdot, \theta)$.

If, however, these distributions are changed so they do not have common parameter θ and $X_i \sim \text{Binomial}(m_i, \theta_i)$, then the mass function of X becomes rather intractable. This particular computation arises in reliability analysis when determining the reliability of a k-out-of-m. heterogeneous system (Høyland and Rausand, 1994, p. 130). Suppose the system

consists of $m.$ independent components among which m_i have common reliability θ_i for $i = 1, \ldots, n$. Variable X represents the number of components working in the system. Suppose the system functions if and only if at least k of the components work. Then, $\Pr(X \geq k)$ is the reliability of the structure. This computation is illustrated as Example 4 in sections 1.1.6 and 1.2.4.

Four suggestions for computing the mass function of X in this reliability context are:

(1) enumerate the exact probabilities,
(2) use a normal density approximation,
(3) use brute force simulation, and
(4) use a saddlepoint approximation.

Option (1) may lead to intractable computations apart from small values of n, and (2) may not result in the desired accuracy particularly when $\{m_i\}$ are small and $\{\theta_i\}$ are not near $1/2$. Option (3) can be time consuming, even with the speed of modern computers. This same option in a continuous setting, when used to approximate a density, also requires kernel density smoothing techniques which can be inaccurate even when applied with relatively large simulations. The saddlepoint option (4) is examined below and is shown to result in highly accurate approximation without the need for placing constraints or guidelines on the values of $\{m_i\}$ and $\{\theta_i\}$. Another advantage of saddlepoint methods is that the required computational times are essentially negligible as compares with simulation.

1.1.2 Saddlepoint density functions

For continuous random variable X with CGF K and unknown density f, the saddlepoint density approximation to $f(x)$ is given as

$$\hat{f}(x) = \frac{1}{\sqrt{2\pi K''(\hat{s})}} \exp\{K(\hat{s}) - \hat{s}x\}. \tag{1.4}$$

Symbol $\hat{s} = \hat{s}(x)$ denotes the unique solution to the equation

$$K'(\hat{s}) = x \tag{1.5}$$

over the range $\hat{s} \in (a, b)$, and is an implicitly defined function of x. Expression (1.5) is referred to as the *saddlepoint equation* and $\hat{s}$ the *saddlepoint* associated with value x. The approximation is meaningful for values of x that are interior points of $\{x : f(x) > 0\} = \mathcal{X}$, the *support* of density f (or of random variable X). We adopt the convention of referring to $\hat{f}$ as the *saddlepoint density* even though it isn't really a density since generally

$$c = \int_{\mathcal{X}} \hat{f}(x)\, dx \neq 1.$$

The *normalized saddlepoint density*

$$\bar{f}(x) = c^{-1} \hat{f}(x) \quad x \in \mathcal{X}$$

is however a proper density on $\mathcal{X}$.

1.1.3 Examples

(1) Normal (0, 1) *density*

A standard normal distribution has CGF $K(s) = s^2/2$ defined for $s \in (-\infty, \infty)$. The saddlepoint equation is explicit in this case as $\hat{s} = x$. Simple computation shows that

$$\hat{f}(x) = \phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2}x^2\right) \qquad x \in \mathcal{X} = (-\infty, \infty) \tag{1.6}$$

and the saddlepoint approximation exactly reproduces the standard normal density ϕ.

(2) Gamma $(\alpha, 1)$ *density*

The density in this instance is

$$f(x) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x} \qquad x > 0$$

with CGF

$$K(s) = -\alpha \ln(1-s) \quad s \in (-\infty, 1).$$

This leads to the explicit saddlepoint expression $\hat{s} = 1 - \alpha/x$ for $x > 0$. The term

$$K''(\hat{s}) = \alpha(1-\hat{s})^{-2} = x^2/\alpha$$

so that for $x > 0$,

$$\begin{aligned} \hat{f}(x) &= \frac{1}{\sqrt{2\pi x^2/\alpha}} \exp\{-\alpha \ln(1-\hat{s}) - \hat{s}x\} \\ &= \frac{1}{\sqrt{2\pi x^2/\alpha}} (x/\alpha)^{\alpha} \exp(-x+\alpha) \\ &= \left(\sqrt{2\pi}\alpha^{\alpha-1/2}e^{-\alpha}\right)^{-1} x^{\alpha-1}e^{-x}. \end{aligned} \tag{1.7}$$

The shape of $\hat{f}$ in (1.7) is the same as that of f but differs from f in the normalization constant. Using Stirling's approximation for $\Gamma(\alpha)$,

$$\hat{\Gamma}(\alpha) = \sqrt{2\pi}\alpha^{\alpha-1/2}e^{-\alpha} \simeq \Gamma(\alpha), \tag{1.8}$$

then

$$\hat{f}(x) = \frac{\Gamma(\alpha)}{\hat{\Gamma}(\alpha)} f(x) \qquad x > 0 \tag{1.9}$$

and differs by a constant relative error determined as the relative error of $\hat{\Gamma}(\alpha)$ in approximating $\Gamma(\alpha)$. The normalized saddlepoint density is exact in this setting.

(3) Normal–Laplace convolution

Consider plotting the density of $X = X_1 + X_2$ where $X_1 \sim$ Normal (0, 1) independently of $X_2 \sim$ Laplace (0, 1) with density

$$f(x) = \tfrac{1}{2}e^{-|x|} \qquad x \in (-\infty, \infty). \tag{1.10}$$

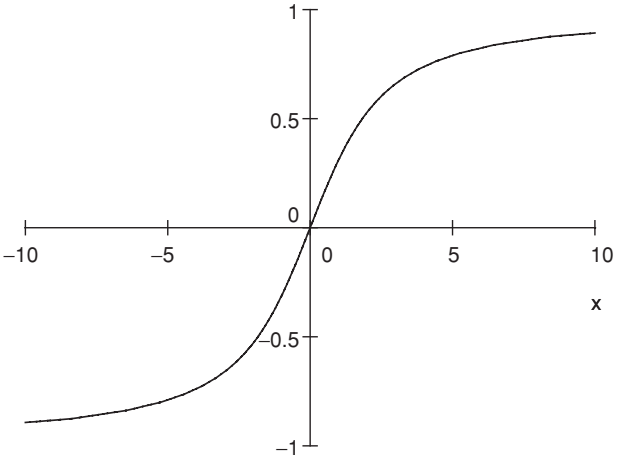

Figure 1.1. $\hat{s}(x)$ vs. x.

The CGF of X takes the particularly simple form

$$K(s) = \tfrac{1}{2}s^2 - \ln(1 - s^2) \qquad s \in (-1, 1)$$

and the saddlepoint equation is

$$K'(\hat{s}) = \hat{s}\left(1 + \frac{2}{1 - \hat{s}^2}\right) = x \qquad x \in (-\infty, \infty). \tag{1.11}$$

The saddlepoint solution is a root of a cubic polynomial that admits one real root and a complex conjugate pair of roots. The unique real root $\hat{s} = \hat{s}(x)$ has been determined numerically and a plot of $\hat{s}$ vs. x is shown in figure 1.1.

The plot shows that the saddlepoint $\hat{s}(x)$ is an odd function in x, a fact confirmed by noting that K' in (1.11) is also odd. Inspection of the plot, as well as consideration of the saddlepoint equation in (1.11) reveal that $\hat{s}(x)$ approaches asymptote $\hat{s} = -1$ as $x \to -\infty$ and asymptote $\hat{s} = 1$ when $x \to \infty$; thus for $x \in (-\infty, \infty)$, the saddlepoint equation can always be solved to find a saddlepoint within $(-1, 1)$. Figure 1.2 shows a comparative plot of the "true" density f (solid line) with the unnormalized saddlepoint density $\hat{f}$ (dotted line), and the normalized saddlepoint density $\bar{f}$ (dashed line).

The "true" density was computed using numerical convolution of the densities involved. A complicated exact expression in terms of special functions has been given in Johnson and Kotz (1970, chap. 1, Eq. (28)), however, numerical computation suggests that it is incorrect since it differs substantially from the numerical convolution. The normalization constant for $\bar{f}$ is most easily computed numerically by making the substitution $dx = K''(\hat{s})d\hat{s}$ so that

$$\int_{-\infty}^{\infty} \hat{f}(x)dx = \int_{-1}^{1} \sqrt{K''(\hat{s})/(2\pi)}\exp\{K(\hat{s}) - \hat{s}K'(\hat{s})\}d\hat{s}$$
$$\simeq 0.8903.$$

The graphical difference between the normalized saddlepoint approximation $\bar{f}$ and f is slight since $\hat{f}$ mostly captures the proper shape of f but not the correct scaling.

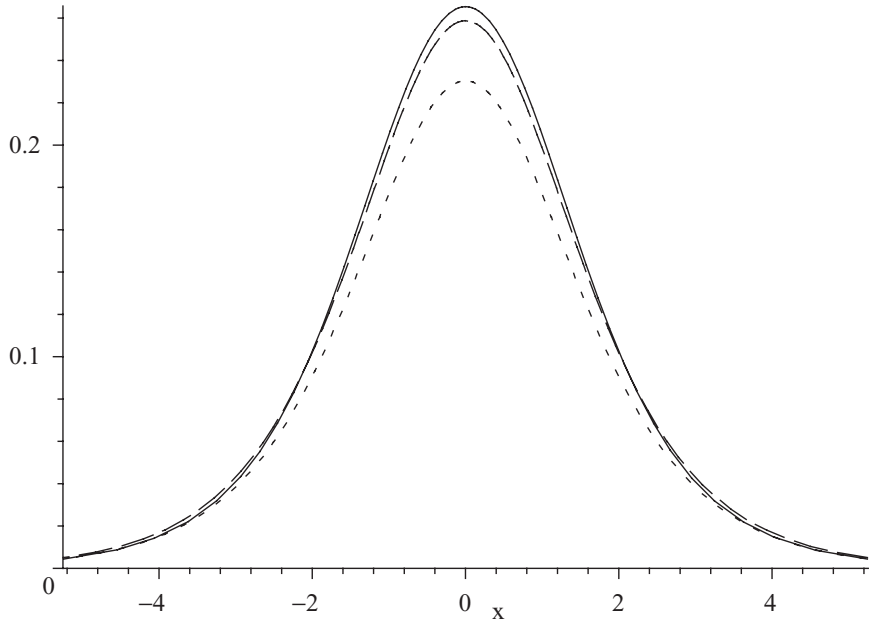

Figure 1.2. $f(x)$ (*solid*), $\bar{f}(x)$ (*dashed*), and $\hat{f}(x)$ (*dotted*) vs. x for the normal-Laplace convolution.

(4) Gumbel (0, 1) *density*

This distribution is also called the extreme value distribution and has CDF

$$F(x) = \exp(-e^{-x}) \quad x \in (-\infty, \infty),$$

with CGF

$$K(s) = \ln \Gamma(1 - s) \quad s \in (-\infty, 1).$$

Saddlepoint computation involves first and second derivatives of the $\ln \Gamma$ function which are the di- and tri-gamma functions respectively. Both functions are in the Maple V library and were computed using these routines. Figure 1.3 compares f, $\hat{f}$, and $\bar{f}$. The degree of accuracy of $\hat{f}$ is striking and $\hat{f}$ integrates to about 0.9793. The plot of $\bar{f}$ is virtually indistinguishable from f.

1.1.4 Remarks

A cursory understanding of the approximation in (1.4) and (1.5) requires clarification of a number of presupposed facts that can be supported with the examples above:

(1) Function K is always a strictly convex function when evaluated over (a, b) so $K''(\hat{s}) > 0$ and the square root is well-defined.

(2) Consider the solvability of (1.5) for the saddlepoint. An appropriate choice of x guarantees that there is a unique solution to (1.5) as now explained. If $\mathcal{X}$ is the support of random variable X, defined as $\mathcal{X} = \{x : f(x) > 0\}$ for continuous density f, let $\mathcal{I}_{\mathcal{X}}$ be the

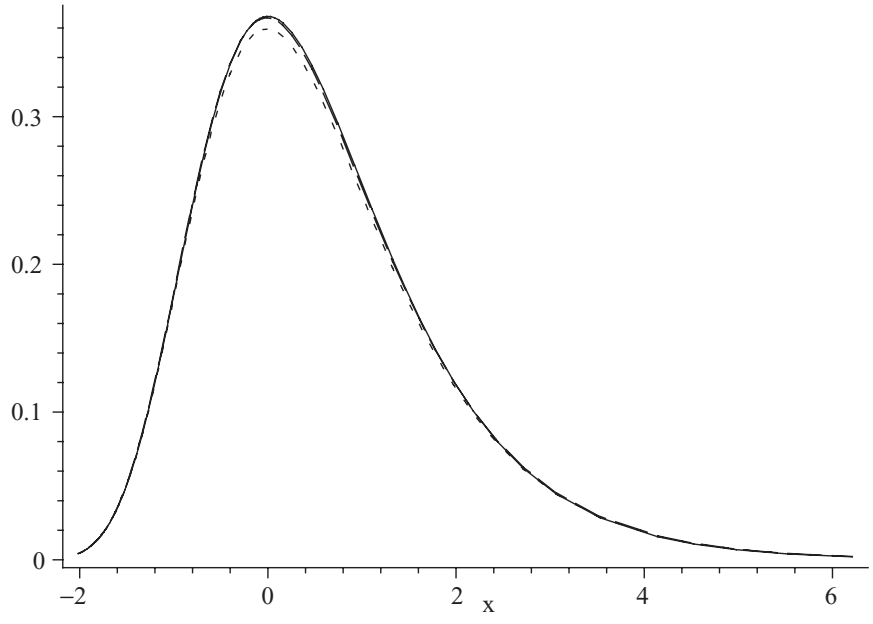

Figure 1.3. $f(x)$ (*solid*), $\bar{f}(x)$ (*dashed*), and $\hat{f}(x)$ (*dotted*) vs. x for the Gumbel density.

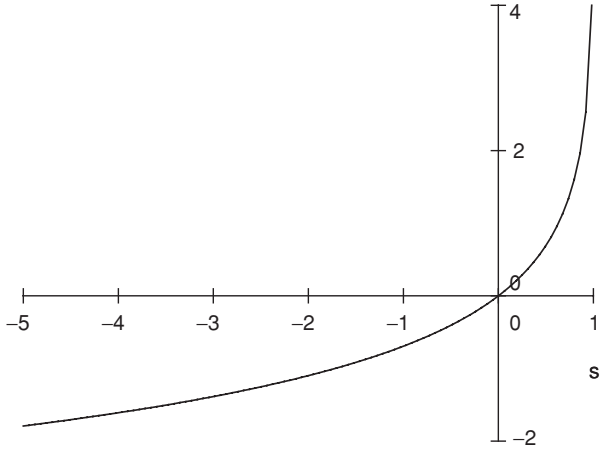

Figure 1.4. The CGF for an Exponential (1).

interior of the span of this support. For example, if X has a uniform density over support $\mathcal{X} = [0, 1] \bigcup (2, 3)$, then the span of the support is $[0, 3)$ and $\mathcal{I}_{\mathcal{X}} = (0, 3)$. The mapping $K' : (a, b) \to \mathcal{I}_{\mathcal{X}}$ is 1-1 and onto (a bijection), and K' is strictly increasing, as mentioned in remark (1), so that a unique solution exists. Other solutions to (1.5) may exist outside of (a, b), but such roots are not allowed since solutions to the saddlepoint equation are restricted to the neighborhood (a, b) about zero. If a value of $x \notin \mathcal{I}_{\mathcal{X}}$ is chosen, then a solution to (1.5) cannot exist in (a, b) but may exist outside of this range.

Figure 1.4 plots $K(s)$ vs. s for an Exponential (1) or Gamma (1,1) density. The slopes of the graph range from 0 as $s \downarrow -\infty$ to ∞ as $s \uparrow 1$ and span the range of support $(0, \infty)$. The corresponding s-values on the horizontal axis are the associated saddlepoints spanning

$(-\infty, 1)$, the convergence neighborhood of K. The choice of $x < 0$ outside of $\mathcal{X}$, the density's support, does not yield a solution to the saddlepoint equation since K' cannot be negative over $(-\infty, 1)$; nevertheless, a meaningless solution can be found for some $s \in (1, \infty)$.

A sum of two independent exponential random variables provides an example that illustrates the possibility of erroneous solutions to the saddlepoint equation. Suppose $X = X_1 + X_2$ where X_1 and X_2 are independent exponentials with means of 1 and 2. The MGF of X is

$$M_X(s) = (1-s)^{-1}(1-s/2)^{-1}$$

and the convergence strip is $(-\infty, 1)$. Solution to the saddlepoint equation

$$K'_X(\hat{s}) = \frac{1}{1-\hat{s}} + \frac{1}{2-\hat{s}} = x \in (0, \infty) \tag{1.12}$$

when restricted to $\hat{s} \in (-\infty, 1)$ is unique and is the smaller root of the quadratic equation determined in (1.12). The larger root is within the range $(1.5, 2)$ and is an erroneous solution. The mapping $K'_X : (-\infty, 1) \to (0, \infty)$ is a bijection, however, the mapping $K'_X : (1.5, 2) \to (0, \infty)$ is also a bijection, yielding erroneous saddlepoints when root finding is not restricted.

(3) Since $K'(0) = E(X)$, the saddlepoint associated with $x = E(X)$ must be $\hat{s} = 0$ by uniqueness. Using the monotonicity of K' then $\hat{s}(x)$ must have the same sign as $x - E(X)$ or

$$\operatorname{sgn}\{\hat{s}(x)\} = \operatorname{sgn}\{x - E(X)\}.$$

Furthermore, saddlepoints for x-values in the right (left) tail of f are necessarily close but smaller than b (greater than a).

Note that for the Exponential (1) example with $x > 0$, $\hat{s} = 1 - 1/x$ and $\operatorname{sgn}(\hat{s}) = \operatorname{sgn}(x-1)$.

1.1.5 Saddlepoint mass functions

For discrete integer-valued random variable X, the saddlepoint approximation for its mass function $p(k)$, based on CGF K, is the same expression as (1.4) and (1.5) when evaluated over the integer values of k. It is written as

$$\hat{p}(k) = \frac{1}{\sqrt{2\pi K''(\hat{s})}} \exp\{K(\hat{s}) - \hat{s}k\} \tag{1.13}$$

where

$$K'(\hat{s}) = k \tag{1.14}$$

and $k \in \mathcal{I}_{\mathcal{X}}$, the interior of the span of the support of X. Saddlepoint expression (1.13) is computable for any value in $\mathcal{I}_{\mathcal{X}}$ whether real or integer-valued, but the plot of $\hat{p}(k)$ is meaningful as an approximation to $p(k)$ only for integer-valued arguments.

1.1.6 Examples

(1) Poisson (λ)

For fixed and known λ, the CGF is

$$K(s) = \lambda(e^s - 1) \qquad s \in (-\infty, \infty).$$

The saddlepoint equation

$$K'(\hat{s}) = \lambda e^{\hat{s}} = k \qquad k = 1, 2, \ldots$$

has an explicit saddlepoint solution $\hat{s} = \ln(k/\lambda)$. The saddlepoint equation cannot be solved at $k = 0$ but note that $0 \notin \mathcal{I}_{\mathcal{X}} = (0, \infty)$ and lies on the boundary of support. The saddlepoint density is

$$\begin{aligned} \hat{p}(k) &= \frac{1}{\sqrt{2\pi k}} \exp\{k - \lambda - k \ln(k/\lambda)\} \\ &= \left(\sqrt{2\pi k} k^k e^{-k}\right)^{-1} \lambda^k e^{-\lambda} \\ &= \lambda^k e^{-\lambda} / \widehat{k!} \end{aligned} \tag{1.15}$$

where

$$\widehat{k!} = \sqrt{2\pi} k^{k+1/2} e^{-k} \simeq k! \tag{1.16}$$

is Stirling's approximation to the factorial function. This factorial approximation is not exactly the equivalent of the gamma function approximation $\hat{\Gamma}(k+1)$ in (1.8) but they are related by

$$\widehat{k!} = k\hat{\Gamma}(k) \neq \hat{\Gamma}(k+1).$$

(This distinction has been a common source of confusion in the saddlepoint area.) Thus, $\hat{p}$ is related to the true mass function p by

$$\hat{p}(k) = \frac{k!}{\widehat{k!}} p(k) \quad k = 1, 2, \ldots \tag{1.17}$$

Normalization requires $p(0)$ to be known so that

$$\bar{p}(k) = \begin{cases} p(0) & k = 0 \\ \{1 - p(0)\} \hat{p}(k) / \sum_{j \geq 1} \hat{p}(j) & k \geq 1 \end{cases}$$

is a normalized saddlepoint approximation. In practical applications, the calculation of boundary probabilities like $p(0)$ is usually possible. The saddlepoint approximations are not exact in this setting and the relative error is that of Stirling's approximation to $k!$. Table 1.1 compares Poisson (1) probabilities for $k = 1, 3, \ldots, 19$ with the values given by $\hat{p}$ and $\bar{p}$. The values of the normalized density $\bar{p}$ are the same as the exact values to the four significant digit accuracy displayed and are listed along with the exact values. Unnormalized values of $\hat{p}$ are given in the third column with their associated relative percentage errors in the fourth column. The diminishing relative error with larger k might have been anticipated since it reflects the relative error of Stirling's approximation $\widehat{k!}$ in (1.16) which is known to decrease with k.

Table 1.1. *Poisson (1) probabilities p with saddlepoint approximations $\hat{p}$ (unnormalized) and $\bar{p}$ (normalized)*

k	$p(k) \simeq \bar{p}(k)$	$\hat{p}(k)$	% relative error
1	.3679	.3989	8.426
3	.06131	.06303	2.805
5	$.0^2 3066$	$.0^2 3117$	1.663
7	$.0^4 7299$	$.0^4 7387$	1.206
9	$.0^5 1014$	$.0^5 1023$	0.8875
11	$.0^8 9216$	$.0^8 9286$	0.7595
13	$.0^{10} 5908$	$.0^{10} 5946$	0.6432
15	$.0^{12} 2813$	$.0^{12} 2829$	0.5688
17	$.0^{14} 1034$	$.0^{14} 1039$	0.4836
19	$.0^{17} 3024$	$.0^{17} 3037$	0.4299

(2) Binomial (n, θ)

The binomial mass function in (1.2) with CGF (1.3) admits an explicit saddlepoint solution

$$\hat{s} = \ln\left\{\frac{k(1-\theta)}{(n-k)\theta}\right\} \quad k = 1, \ldots, n-1. \tag{1.18}$$

The term $K''(\hat{s}) = k(n-k)/n$ and the saddlepoint mass function works out to be

$$\hat{p}(k) = \widehat{\binom{n}{k}} \theta^k (1-\theta)^{n-k} \quad k = 1, \ldots, n-1 \tag{1.19}$$

where the notation

$$\widehat{\binom{n}{k}} = \frac{\widehat{n!}}{\widehat{k!}\,\widehat{(n-k)!}} \simeq \binom{n}{k}$$

is used based on Stirling's approximation in (1.16). The relationship of $\hat{p}(k)$ to $p(k)$ in (1.19) exhibits a structure like that of the Poisson example in (1.17), and which arises consistently in applications of saddlepoint approximations: In the range of examples where combinatorics, factorials or gamma functions are normalization constants, the saddlepoint approximations $\hat{f}$ and $\hat{p}$ are often equal to f or p times a factor consisting of ratios of various Stirling approximations. In this example, the relationship is

$$\hat{p}(k) = \widehat{\binom{n}{k}} \binom{n}{k}^{-1} p(k) \quad k = 1, \ldots, n-1.$$

(3) Negative Binomial (n, θ)

Suppose X is the number of failures occurring in a sequence of Bernoulli trials before the nth success where θ is the probability of a single success. Exercise 9 specifies the details