5. RELATED TECHNIQUES

5.1 INTRODUCTION

In this chapter, a number of related techniques will be presented emphasizing their relationship to small sample asymptotics. Section 2 looks at an approach developed by Frank Hampel which has a number of desirable properties.

Next the relationship of small sample asymptotics to saddlepoint and large deviations is presented. We then turn to the work of Durbin and Barndorff-Nielsen and attempt to relate their work in the case of sufficiency and/or exponential families to the techniques of small sample asymptotics. To conclude the chapter, computations are done in the case of logistic regression to contrast the various approaches.

5.2. HAMPEL'S TECHNIQUE

In the paper, Hampel (1973), many of the motivating ideas for small sample asymptotics are laid down. Both authors were introduced to the topic via the paper and it is important to acknowledge its influence. Although the results turn out to be closely related to saddlepoint results, they were developed independently of the saddlepoint work of Daniels (1954). The approach proposed by Hampel is very interesting and probably has yet to be fully exploited. Our purpose here is to present the ideas and suggest some possible future directions. The initial development follows Hampel (1973) very closely, especially p. 111, 112.

Hampel's approach differs in several ways from typical classical approaches. The first is that the density of the estimate, rather than a standardized version of it, is approximated. A second feature is that we use low-order expansion in each point separately and then integrate the results rather than use a high-order expansion around a single point. It is this feature which really distinguishes small sample asymptotics (and saddlepoint techniques) from classical asymptotic expansions. The local accuracy from the first one or two terms is effectively transferred to a selected grid of points yielding the same accuracy globally. It is the availability of cheap computing which makes feasible this use of local techniques. A fairly simple approximation requiring non-trivial computation is carried out at a number of grid points. This is of course exactly the type of problem which is ideally suited to computer computations.

The third difference concerns the question of what to expand. Hampel argues effectively that the most natural and simple quantity to study is the derivative of the logarithm of the density, namely f_n'/f_n. There are at least four reasons why this seems reasonable.

(i) The form of the expansion of f_n'/f_n is such that the first term is proportional to n and the first two terms are linear in n. This contrasts with more complicated relationships coming from f_n or the cumulative.

(ii) Since our expansions are local in nature, it makes sense to focus on a feature of a distribution which is not affected by shifts or addition or deletion of mass elsewhere. Neither f_n or the cumulative satisfy these properties. f_n'/f_n is the first and simplest quantity with these local properties.

(iii) We can view the normal distribution as playing a very special and basic role in probability, in many ways analogous to the role of the straight line in geometry. For the normal, it is f'/f which has a particularly simple form, namely a linear function of x. By expanding f_n'/f_n locally, we are, in a sense, linearizing a function locally.

(iv) An expansion of f_n'/f_n will not give the constant of integration for f_n but forces us to determine it numerically. As has been noted in Remark 3.2, in approximating the density of mean, the order of the error is improved from n^{-1} to $n^{-3/2}$ by renormalization. Using the f_n'/f_n scale emphasizes the renormalization in a natural way.

To contrast this approach to the techniques developed so far, namely, it is useful to consider a specific problem. Consider the situation of approximating f_n'/f_n for the mean of n independent observations. The presentation is similar to that found in Field (1985). The development for the more important case of M-estimates of location is given in Field and Hampel (1982).

Assume that f_n'/f_n is to be approximated at a point t. The conjugate density is $h_t(x) = c(t)\exp\{\alpha(t)(x-t)\}f(x)$ and $\alpha(t)$ is the solution of

$$\int (x-t)h_t(x)dx = 0 \quad \text{or} \quad \int (x-t)\exp\{\alpha(t)(x-t)\}f(x)dx = 0.$$

In order to guarantee the existence of $\alpha(t)$ and its derivatives up to order 4, assume that $\int x^r e^{\alpha x} f(x)dx$ exists for r up to 5.

Now we obtain a centering results (cf 4.4, 4.23) as follows:

$$\begin{aligned} f_n(t) &= n\int_{\dots}\int f(nt - \sum_1^{n-1} x_i)\prod_1^{n-1} f(x_i)d\mathbf{x} \\ &= nc^{-n}(t)\int_{\dots}\int h_t(nt - \sum_1^{n-1} x_i)\prod_1^{n-1} h_t(x_i)d\mathbf{x} \\ &= c^{-n}(t)h_{t,n}(t) \end{aligned}$$

where $h_{t,n}(t)$ is the density of $\bar{X}$ with underlying density h_t. Now we use a normal approximation to $\bar{X}$ under h_t. Recall $E_{h_t}\bar{X} = t$ and $\operatorname{var}_{h_t}\bar{X} = \int (x-t)^2 h_t(x)dx/n \equiv \sigma^2(t)/n$. Hence $h_{t,n}(t)$ can be approximated by $n^{1/2}/\sqrt{2\pi}\sigma(t)$ and $h_{t,n}'/h_{t,n}(t)$ by $\sigma'/\sigma(t)$ each with errors of order $1/n$. The term of order $n^{-1/2}$ disappears since we are evaluating the density at the mean.

From this

$$\begin{aligned} f_n'/f_n(t) &= -nc'/c(t) - \sigma'/\sigma(t) + 0(1/n) \\ &= -n\alpha(t) - \sigma'/\sigma(t) + 0(1/n). \qquad (5.1) \end{aligned}$$

To illustrate the behavior of $f_n'/f_n(t)$, we examine its behavior for the case of the uniform density on $[-1,1]$ and for the extreme value density, $f(x) = \exp\{x - \exp(-x)\}$. For the uniform case, it is possible to compute the exact value of f_n'/f_n. The following Exhibit 5.1 compares the exact and approximate values. See Exhibit 3.7 and 3.8 for results on the density and distribution function.

	$n = 5$		$n = 20$	
t	**Exact**	**Approximate**	**Exact**	**Approximate**
0.00	0.0000	0.0000	0.0000	0.0000
0.05	-0.6608	-0.6546	-3.9140	-3.9150
0.10	-1.3270	-1.3230	-5.8540	-5.8560
0.20	-2.6950	-2.7270	-11.9200	-11.5200
0.30	-4.1520	-4.1680	-18.4500	-18.4500
0.40	-5.7700	-5.7890	-25.8200	-25.8200
0.50	-7.6610	-7.7420	-34.6100	-34.6200
0.70	-13.5200	-13.3300	-63.0700	-63.0500
0.90	-40.0000	-40.0000	-190.0000	-190.0000

Exhibit 5.1
Exact and approximate results for $f_n'/f_n(t)$
uniform observations.

From Exhibit 5.1, it is clear that even for $n = 5$, the approximation is very accurate over the whole range. The following plots (Exhibit 5.2, 5.3) demonstrate the approach to normality as n increases. As n increases, f_n'/f_n becomes smoother, on the one hand, and only a smaller increasingly steep central part contains most of the mass of the distribution. To calculate the curves the graphs are plotted for values of t corresponding to middle 99.8% of the density (i.e. tails of .001). In order to compute these percentiles, it is convenient to use the tail area approximation of Lugannani and Rice (1980) given by (3.27). It is worth noting that although the exact tail area requires an integration over the range (t, ∞), we can obtain a very accurate approximation with the values only at the point t. This results in considerable saving of computational effort.

The graph for the uniform (Exhibit 5.2) only shows the upper part of the graph since f_n is symmetric. It is clear that at $n = 40$, the graph shows very little deviation from a straight line indicating close agreement with the normal. We can argue that such a diagram makes it very easy to see how quickly a density is approaching its normal approximation. The second graph (Exhibit 5.3) for the extreme value density shows a density for the mean which is decidedly asymmetric at least for values of $n = 5$ and 10. However for $n = 40$, we have good agreement with the normal.

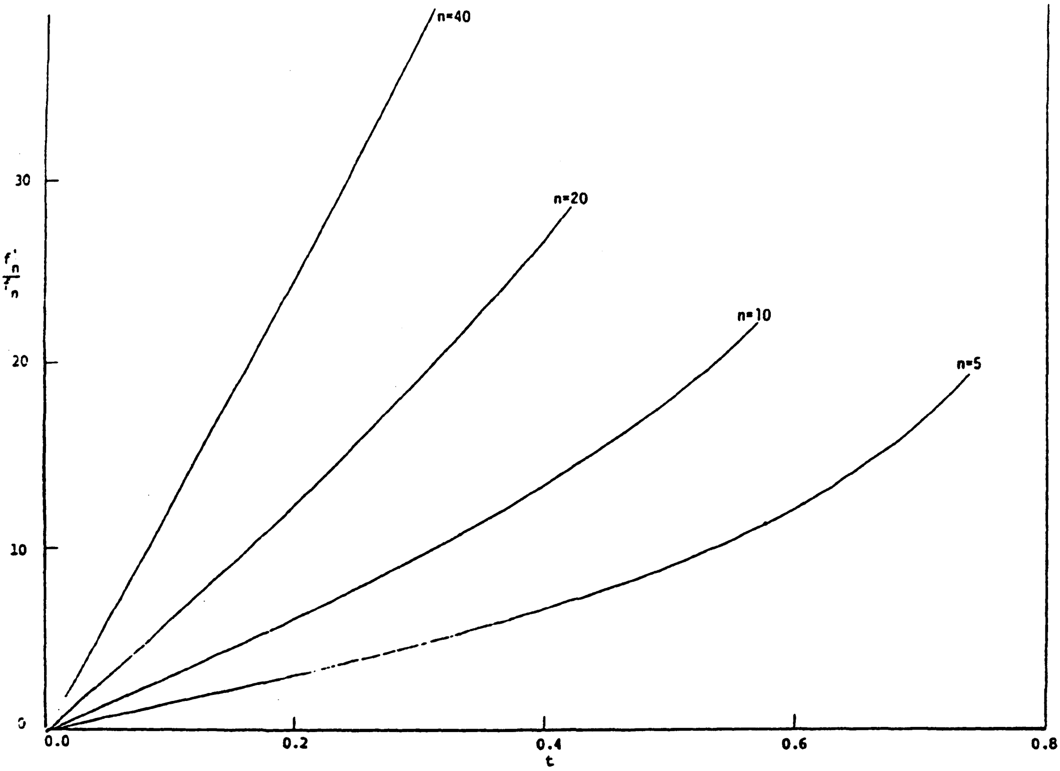

Exhibit 5.2
f'_n/f_n for mean of uniform observations

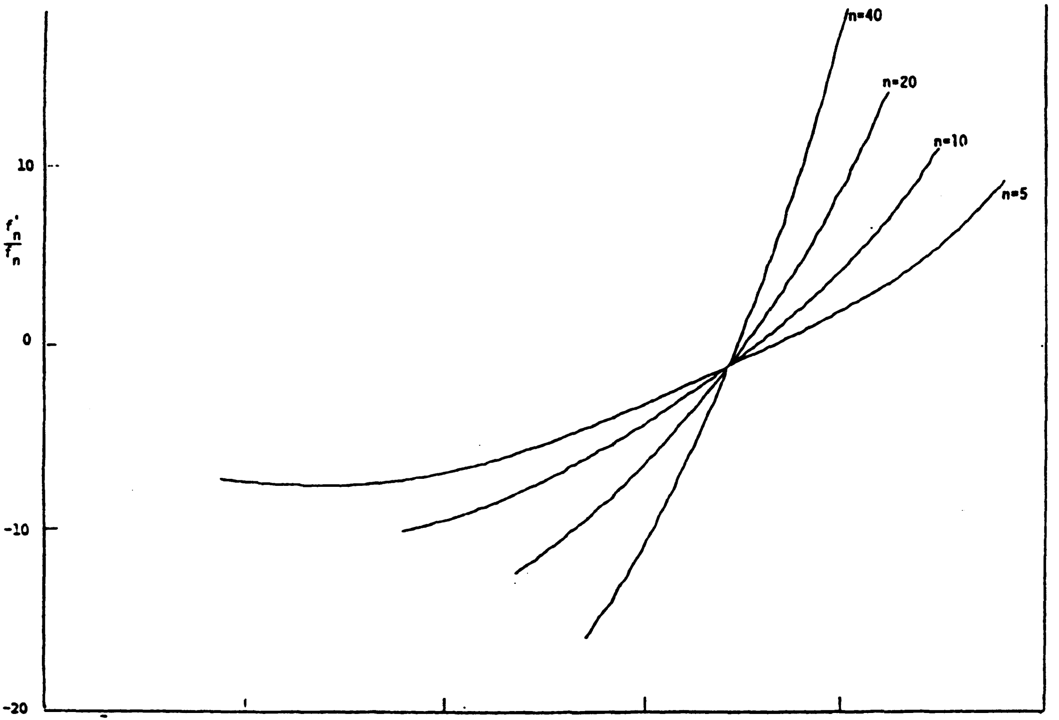

Exhibit 5.3
f'_n/f_n for mean of extreme observations

In order to assess the quality of the approximation, we begin by considering the density of $n^{1/2}(\bar{X}-t)/\sigma(t)$ where we now assume that the X_i's are distributed according to the conjugate density, $h_t(x)$. Denote this density of $n^{1/2}(\bar{X}-t)/\sigma(t)$ by $s_n(x;t)$. Now $s_n(0;t) = n^{-1/2}c_n(t)\sigma(t)f_n(t)$. But this implies that

$$\frac{f_n'}{f_n}(t) = -n\alpha(t) - \frac{\sigma'}{\sigma}(t) \quad \text{if} \quad \frac{\partial s_n(0;t)/\partial t}{s_n(0;t)} = 0.$$

Recalling that $s_n(0;t)$ is the density of a normalized sum at its expected value, we have an Edgeworth expansion as follows:

$$s_n(0;t) = (\pi/2)^{1/2}(1 + 0(1/n)).$$

The quality of the approximation will be determined by how closely the density of $n^{1/2}(\bar{X} - t)/\sigma(t)$ matches that of a standard normal in a neighborhood of t. Recent work by Field and Massam (1987) develops a diagnostic function based on this observation. The diagnostic function has some similarity to the diagnostic function for normality proposed by Efron (1981). As will be demonstrated in chapter 6, we can think of the small sample approximation being based on a local transformation to normality. Efron uses a similiar , but a less general transformation, as a means of constructing confidence intervals (cf. Efron (1987)). Section 6.3 gives a construction based on small sample approximations.

We now turn to a brief discussion of $\alpha(t)$. Recall that for the mean $\alpha(t)$ solves $\int(x-t)\exp\{\alpha(t)(x-t)\}f(x)dx = 0$. We note that $\alpha(t)$ uniquely determines f. For if f_1 and f_2 both give rise to the same $\alpha(t)$, then $\alpha(t) = c'(t)/c(t)$ which implies $\log c(t) = \int_\mu^t \alpha(s)ds$ so that $c(t)$ is the same for both f_1 and f_2. The relationship between $\alpha(t)$ and $c(t)$ can be obtained by differentiating the equation for $\alpha(t)$. From this

$$\int \exp(\alpha(t)(x-t))f_1(x)dx = \int \exp(\alpha(t)(x-t))f_2(x)dx.$$

Since these expressions are Laplace transforms, this implies the equality of f_1 and f_2.

To see the central limit theorem from this perspective, note that $\alpha(t) = t$ for the standard normal. Hence the central limit theorem requires

$$\lim_{n\to\infty} \alpha_n^*(t) = t$$

where $\alpha_n^*(t)$ corresponds to $\sqrt{n}(\bar{X}-\mu)/\sigma_0$, $E_f X_i = \mu$, $\sigma_0 = var_f X_i$. The first step is to express $\alpha_n^*(t)$ in terms of $\alpha(t)$ corresponding to f. $\alpha_n^*(t)$ must satisfy

$$E_f\left[\left(\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma_0} - t\right)\exp\left(\sigma_n^*(t)\left(\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma_0} - t\right)\right)\right] = 0$$

or

$$\frac{1}{\sigma_0\sqrt{n}}\sum_{i=1}^n E_f\left[\left(X_i - \mu - \frac{\sigma_0 t}{\sqrt{n}}\right)\exp\left(\frac{\alpha_n^*(t)}{\sigma_0\sqrt{n}}\left(X_i - \mu - \frac{\sigma_0 t}{\sqrt{n}}\right)\right)\right] \times$$

$$\prod_{j\neq i} E_f \exp\left(\frac{\alpha_n^*(t)}{\sigma_0\sqrt{n}}\left(X_j - \mu - \frac{\sigma_0 t}{\sqrt{n}}\right)\right) = 0.$$

But we can make the first term in the product equal to 0 by setting $\alpha_n^*(t)/\sigma_0\sqrt{n} = \alpha(\mu + \sigma_0 t/\sqrt{n})$. Therefore $\alpha_n^*(t) = \sigma_0\sqrt{n}\alpha(\mu + \sigma_0 t/\sqrt{n})$. Expanding α about μ, we obtain

$$\alpha_n^*(t) = \sigma_0\sqrt{n}\alpha(\mu) + t\alpha'(\mu)\sigma_0^2 + \sigma_0^3 t^2 \alpha''(\tilde{\mu}/2\sqrt{n}) \quad \text{for} \quad \mu < \tilde{\mu} \leq \mu + \sigma_0 t/\sqrt{n}.$$

We assume that $\alpha''(\tilde{\mu})$ is well behaved near μ. Using the fact that $\alpha(\mu) = 0$, $\alpha'(\mu) = 1/\sigma_0^2$, we have $\lim_{n\to\infty} \alpha_n^*(t) = t$ as required for the central limit theorem.

Although we have chosen to focus on $\alpha(t)$ as the transform, it is also possible to view $\log c(t)(= \int_\mu^t \alpha(s)ds)$ as a transform. If we let $K(\alpha) \equiv \log E_f \exp(\alpha X)$ denote the cumulant generating function, then the Legendre transform $K^*(t)$ of $K(\alpha)$ is defined as

$$K^*(t) = \sup_\alpha\{\alpha t - K(\alpha)\}.$$

The maximizing value of α is obtained by solving

$$K'(\alpha) = t \qquad \text{or}$$

$$\int (x - t) \exp(\alpha) f(x) dx = 0.$$

Hence

$$\begin{aligned} \alpha = \alpha(t) \quad \text{and} \quad K^*(t) &= \alpha(t)t - \log \int \exp(\alpha(t)x) f(x) \\ &= \log \int \exp(\alpha(t)(x - t)) f(x) dx \\ &= -\log c(t), \\ K^{*''}(t) = \alpha'(t) &= 1/K''(\alpha(t)), \end{aligned}$$

and the saddlepoint approximation can be written as

$$f_n(t) = (2\pi/n)^{1/2}(K^{*''}(t))^{1/2} \exp(-nK^*(t))[1 + O(1/n)].$$

A nice development of saddlepoint approximations using the Legendre transformation can be found in McCullagh (1987, Chapter 6).

Although f_n'/f_n seems in many ways to be the most natural quantity to approximate, the arguments become awkward if we move from the mean, either to M-estimates or multidimensional problems (cf. Field and Hampel 1982). In these cases we end up approximating both f_n and f_n' and then taking the quotient of the results. Given the approximation to f_n, there seems to be no practical argument to do all the work to approximate f_n'/f_n. The fact that f_n'/f_n gives us the correct integrating factor can be carried over to approximating f_n by a renormalization of the approximation.

It is quite possible that f_n'/f_n can be approximated directly in which case this would be an attractive alternative. This may involve a more geometric approach than the one we have been using. Insight in this direction could be very helpful in obtaining a deeper understanding of the mechanics of the approximation and could lead to new proofs and deeper understandings of the central limit theorems.

5.3. RELATIONSHIP TO EDGEWORTH EXPANSION AND LARGE DEVIATIONS

In this section, our formula for f_n'/f_n and the saddlepoint method are compared to the classical methods of Edgeworth expansion (Cramer, 1946, pp. 229, 223, 133 etc.; Daniels,

1954) and large deviations (Richter, 1957; Feller, 1971; Cramer, 1938). In addition, from the formula for f_n'/f_n a new variant of the classical methods is derived. Using this variant the connections between the methods is made very clear. The development follows closely that of Field and Hampel (1982, section 10).

To make comparisons easy, we consider the case of the arithmetic mean i.e. $\psi(x) = x$. Let $X_1, \cdots, X_n$ be independent observations from a density f satisfying regularity conditions required for the classical expansions; for example, conditions 1 and 2 of Richter (1957) p. 208, which require that the moment generating function of f exists in an interval and that $\sum_{i=1}^{n} X_i$ has a bounded density. Assume that $EX_i = 0$ and put $\text{var}X_i = \sigma^2$, $EX_i^3/\sigma^3 = \lambda_3$, $EX_i^4/\sigma^4 - 3 = \lambda_4$. Write $T_n = \bar{X}$ with density $f_n(t)$. Now $ET_n = 0$, $\text{var}T_n = \sigma^2/n$, $\lambda_3(T_n) = \lambda_3/\sqrt{n}$, $\lambda_4(T_n) = \lambda_4/n$.

Before proceeding, it is helpful to look at the situations to which these methods are directed. Both the Edgeworth and large deviation expansions approximate the density of $\sqrt{n}\bar{X}/\sigma$, which at a point x equals $f_n(x\sigma/\sqrt{n})\sigma/\sqrt{n}$. In the Edgeworth expansion $x = 0(1)$ while in large deviations $x = 0(\sqrt{n})$. Writing $t = \sigma x/\sqrt{n}$, the methods can be compared as follows:

f_n'/f_n and saddlepoint :	$t = 0(1)$
Large deviations up to order $k-2$:	$t = 0(n^{-1/k}), \quad k > 2$
Edgeworth :	$t = 0(n^{-1/2})$.

In deriving f_n'/f_n and the saddlepoint approximation at t, the underlying density is recentered around t using a conjugate (or associated) distribution, $h_t(x) = c(t)\exp\{\alpha(t)(x - t)\}f(x)$ with $\int(x-t)h_t(x)dx = 0$. This centering is equivalent to a shift in the f'/f space:

$$h_t'(x)/h_t(x) = f'(x)/f(x) + \alpha(t).$$

The Edgeworth expansion is used locally at 0 for each centered density, h_t, in both f_n'/f_n and the saddlepoint approximation. In fact, the saddlepoint approximation, which is, except for a constant, the integrated version of f_n'/f_n, only uses the first term of the Edgeworth expansion, the normal approximation, at each t. It is remarkable that this simple device yields the very good accuracy even in the extreme tails that has been shown in the previous sections. On the other hand, the Edgeworth or large deviation expansion are not recentered. To take our comparisons a step further, only the local behavior of f_n'/f_n and the saddlepoint approximation at $t = 0$ is considered.

Starting from formula (5.1), we have, for the arithmetic mean, $f_n'/f_n(t) = -n\alpha(t) - \beta(t) - \gamma(t)/n \cdots$ where $\alpha(t)$, $\beta(t) = \sigma'/\sigma(t)$, $\gamma(t)$ corresponds to terms of order $1/n$ which we will not need explicitly. Put $\alpha(t) = \alpha(0) + \sum_{v=1}^{\infty} \alpha^{(v)}(0)t^v/v!$, $\beta(t) = \beta(0) + \sum_{v=1}^{\infty} \beta^{(v)}(0)t^v/v!$ and $\gamma(t) = \gamma(0) + \sum_{v=1}^{\infty} \gamma^{(v)}(0)t^v/v!$. Recall that in both Edgeworth and large deviations $t \to 0$ as $n \to \infty$, so that these expansions make sense. By integrating, we obtain

$$\begin{aligned}\log f_n = \log f_n(0) &- n\alpha(0)t - \frac{n}{2}\alpha'(0)t^2 - \frac{n}{6}\alpha''(0)t^3 \cdots \\ &- \beta(0)t - \frac{1}{2}\beta'(0)t^2 - \frac{1}{6}\beta''(0)t^3 \\ &- \frac{1}{n}\gamma(0)t - \frac{1}{2n}\gamma'(0)t^2 \cdots.\end{aligned}$$

Write $\log f_n(0) = \log(\sqrt{n/2\pi\sigma^2})(1 + w_1/n + \cdots)$. Observing that $\alpha(0) = 0$, since $EX = 0$, we have

$$f_n(t) = \sqrt{n/2\pi\sigma^2}\exp\{w_1/n + \cdots\}\exp\{-n\alpha'(0)t^2/2\}\exp\{-\frac{n}{6}\alpha''(0)t^3$$
$$-\frac{n}{24}\alpha'''(0)t^4 \cdots - \beta(0)t - \frac{1}{2}\beta'(0)t^2 - \frac{1}{6}\beta''(0)t^3$$
$$-\frac{1}{n}\gamma(0)t - \frac{1}{2n}\gamma'(0)t^2 \cdots\}. \tag{5.2}$$

Note that the various terms can all be expressed in terms of $\sigma, \lambda_3, \lambda_4 \cdots$. This can be done directly by differentiating or from (5) and (6) in Richter (1957). However this re-expression of terms is easiest to see by means of comparison with the Edgeworth series which we do later in the section. Note that the infinite series (2.6) in Daniels (1954) has effectively $t = 0$ (after recentering) so that only the expansion of $f_n(0)$ remains in (5.2), i.e.$\sqrt{n/2\pi\sigma^2}e^{w_1/n + \cdots}$. In 2.6 the exponential is expanded to give $1 + w_1/n + \cdots$, so that for a finite piece of the series, negative approximated densities can result. However, the saddlepoint approximation, which ignores w_1/n etc., is always positive and has been seen to be very accurate even in the extreme tails . With regards to the constant of integration, $\log f_n(0)$, numerical results indicate that accuracy can be improved over the expansion used above by evaluating this numerically as $[2\int\limits_0^\infty f_n(t)dt]^{-1}$.

Key pages for the connection between large deviation and saddlepoint approximations are Richter (1957), p. 212 and 214. Note that in Richter's formulas, there are several misprints. On the bottom of p. 213, the formula for I, should have been $\varphi_3^2 t^6/2(3!)^2$ (the 2 is missing) and top of p.214, $\varphi_4(z_0)/8 - \frac{5}{24}\varphi_3^2(z_0)$ instead of $\varphi_4(z_0)/9 - 5\varphi_3^2(z_0)/12$.

A key formula for the connection between saddlepoint and Edgeworth approximations is (4.3) in Daniels (1954) where the Hermite polynomials differ from those in Cramer (1946) by a factor of $(-1)^n$.

Consider now the Edgeworth expansion which is an expansion for $n^{1/2}t$ = constant > 0 (i.e. at each fixed multiple of the standard deviation of T_n). We proceed by expanding the exponents in (5.2). Remembering nt^2 = constant, groups of terms of equal order are:

Constant, nt^2	(together with $\sqrt{n}$, the normal approximation)
nt^3, t	(skewness only in addition)
$nt^4, t^2, 1/n$	(skewness and excess).

The expansion of the exponentials in (5.2) up to this order yields

$$f_n(t) \cong \left(\frac{n}{2\pi\sigma^2}\right)^{1/2} \exp(-n\alpha'(0)t^2/2)\left\{1 - \frac{\alpha''(0)}{6}nt^3 - \beta(0)t\right.$$
$$- \frac{\beta'(0)}{2}t^2 + \frac{w_1}{n} - \frac{\alpha'''(0)}{24}nt^4 + \frac{\alpha''(0)\beta(0)}{6}nt^4 + \frac{\beta^2(0)}{2}t^2$$
$$\left. + \frac{\alpha''(0)^2}{72}n^2t^6\right\}.$$

To match with the Edgeworth expansion, make the substitution $x = \sqrt{n}t/\sigma$. The Edgeworth expansion is an expression for fixed x in powers of $n^{-1/2}$. From (2.9) we obtain the expansion for the standardized density

$$h_n(x) = \phi(x)\left\{1 + \frac{\lambda_3(T_n)}{3!}(x^3 - 3x) + \frac{\lambda_4(T_n)}{4!}(x^4 - 6x^2 + 3) \right.$$
$$\left. + \frac{\lambda_3^2(T_n)}{72}(x^6 - 15x^4 + 45x^2 - 15)\right\}$$

where $\phi(x)$ is the standard normal density.

Since $\lambda_3(T_n) = \lambda_3/\sigma n$ etc., we obtain by matching terms $\alpha(0) = 0$, $\alpha'(0) = 1/\sigma^2$, $\alpha''(0) = -\lambda_3/\sigma^3$, $\alpha'''(0) = -\lambda_4/\sigma^4 + 3\lambda_3^2/\sigma^4$, $\beta(0) = \lambda_3/2\sigma$, $\beta'(0) = \lambda_4/2\sigma^2 - \lambda_3/\sigma^2$, $w_1 = \lambda_4/8 - 5\lambda_3^2/24$.

From Theorem 2 and (7) of Richter (1957), the first two terms of the large deviation expansion yield

$$f_n(t) \cong \sqrt{\frac{n}{2\pi\sigma^2}}e^{-nt^2/2\sigma^2}\exp\left\{\frac{\lambda_3}{6\sigma^3}nt^3 + \frac{\lambda_4}{4!\sigma^4}nt^4 - \frac{\lambda_3}{8\sigma^4}nt^4\right\}. \tag{5.3}$$

This expansion corresponds to an extreme case of asymptotic direction in which $n \to \infty$ first and then $t \to 0$ or the limiting case of $n^c t = \text{const}$ for $c \to 0$. This corresponds to keeping only the leading constant term and the expansion of $\alpha(t)$ and leads precisely to formula (5.3) above.

Hence the large deviation approximation for the density is nothing but the expansion of $\alpha(t)$ totalling ignoring the other terms in (5.2). Its value at $t = 0$ coincides with that of the saddlepoint approximation, but since for $t \neq 0$, it does not readjust σ as the latter does (which amounts to keeping $\beta(t)$ in the local expression), even the full infinite large deviation series would (apart from a constant) correspond to using only the first order term in the integrated f_n'/f_n approximation or the equivalent saddlepoint approximation. Results in Hampel (1973) show that this will give a poor numerical fit; the finite pieces of the series such as (5.3) above are only an approximation to this poor fit. The versions of large deviations for the cumulative, instead of the density, such as (6.23) in Feller (1971), probably contain the additional approximating error of the normal tail area by a function of the normal density and are likely to be still worse.

If we keep all terms of order $n^{1/2}t = \text{constant}$, then the full version of (5.2) up to this order is

$$f_n(t) \cong \sqrt{\frac{n}{2\pi\sigma^2}}\; e^{-nt^2/2\sigma^2}\exp\left\{\frac{\lambda_3}{6\sigma^3}nt^3 + \frac{\lambda_4}{24\sigma^4}nt^4 - \frac{\lambda_3}{8\sigma^4}nt^4 - \frac{\lambda_3}{2\sigma}t \right.$$
$$\left. - \frac{\lambda_4}{4\sigma^2}t^2 + \frac{\lambda_3^2}{2\sigma^2}t^2 + \frac{\lambda_4}{8n} - \frac{5\lambda_3^2}{24n}\right\}. \tag{5.4}$$

The new formula (5.4) is what large deviations ought to be to give any hope of decent numerical results. It compares closely to Edgeworth, as the only difference is in the finite expansion of the exponent in Edgeworth. However the new formula can never be negative while the Edgeworth can. On the other hand, if $\lambda_4 - 3\lambda_3^2 \geq 0$, and not $\lambda_3 = \lambda_4 = 0$, formula (5.4) will eventually explode for large $|t|$, as will large deviations. Also the strict norming of the total probability, which Edgeworth expansions often pay for with negative densities, will be lost. But all this hardly matters if one cuts off for large $|t|$ when all these approximations, based on expansions at 0, will be bad anyway. Of greater interest is the behavior for small

$|t|$ and then it can be hoped that (5.4) is slightly better than Edgeworth. This has been supported by a numerical example with X_i exponential, $n = 4$ when for $|x| = |\sqrt{n}t/\sigma| \leq 1.5$, the error was rougly halved and outside $|x| = 2$, both approximations were bad as is to be expected.

		"Edgeworth		Edgeworth				Large	
t	Exact	exponent	%		%	Normal	%	deviations	%
		(5.4)	error		error		error	(5.3)	error
-0.25	0	.00168		-.03920		.0350		.0002	
0	0	.04596		.02175		.1080		.0105	
0.25	0.24525	.29544	20.5	.02175	21.4	.2590	5.6	.0105	-56.1
0.5	.72179	.70413	-2.4	.69231	-4.1	.4840	-11.9	.3848	-46.7
0.75	.89617	.89126	-.5	.88857	-.8	.7042	-21.4	.6869	-23.4
1	.78147	.78143	-.005	.78126	-.02	.7979	2.1	.7979	2.1
1.25	.56150	.56358	.4	.56584	.8	.7042	25.4	.7162	27.6
1.5	.35694	.36151	1.3	.36968	3.6	.4840	35.6	.5371	50.5
1.75	.20852	.20305	-2.6	.20052	-3.8	.2590	24.2	.3313	58.9
2	.11451	.08952	-21.8	.09373	-18.1	.1080	-5.7	.1507	31.6
2.5	.03027	.00340	-88.7	.04026	33.0	.0088	-70.9	.0051	-83.2
3	.00708	1.10^{-6}	-100	.00889	25.6	.0003	-95.8	1.10^{-6}	-100
3.5	.00152	7.10^{-14}	-100	.00045	-70.4	3.10^{-6}	-100	4.10^{-14}	-100
4	.00031	2.10^{-17}	-100	6.10^{-6}	-98.1	1.10-8	-100	3.10^{-28}	-100
5	.00001	4.10^{087}	-100	3.10^{-11}	-100	2.10^{-11}	-100	8.10^{-89}	-100

Exhibit 5.4

Density of the mean of 4 exponential for various approximations
$E\bar{X}_4 = 1$, $var\bar{X}_4 = 1/4$, $\lambda_3(\bar{X}_4) = 1$, $\lambda_4(\bar{X}_4) = 3/2$.

The approximation of this using f_n'/f_n is exact everywhere while the saddlepoint approximation has a constant relative error of +2.1% everywhere.

5.4. APPROXIMATING THE DENSITY OF SUFFICIENT ESTIMATORS AND MAXIMUM LIKELIHOOD ESTIMATORS IN EXPONENTIAL FAMILIES

We consider the problem of approximating the density of sufficient estimators with the aim of relating the results to those of small sample asymptotics. The development is based closely on the important work by Durbin (1980a) and we will use his notation. Durbin assumes that we have a matrix of observations $\mathbf{y} = (y_1, \cdots, y_n)^T$ (not necessarily iid) where each y_i is of dimension ℓ and has density

$$f(\mathbf{y}, \theta) = G(\mathbf{t}, \theta)H(\mathbf{y})$$

where $\mathbf{t}$ is m-dimensional and is the value of an estimate T_n of θ.

Durbin assumes that a transformation $y_1, \cdots, y_n \rightarrow t_1, \cdots, t_m,\ u_{m+1}, \cdots, u_{m\ell}$ exists so that the density of $T_n, g(\mathbf{t}, \theta) = G(\mathbf{t}, \theta)H_1(\mathbf{t})$.

The basic equation for the first step comes from rewriting

$$f(\mathbf{y}, \theta) = g(\mathbf{t}, \theta)h(\mathbf{y}) \quad \text{where} \quad h(\mathbf{y}) = H(\mathbf{y})/H_1(\mathbf{t}).$$

Since this result holds for any θ, we obtain

$$g(\mathbf{t},\theta_0) = \frac{f(\mathbf{y},\theta_0)}{f(\mathbf{y},\theta)} g(\mathbf{t},\theta). \tag{5.5}$$

The approximation results by making an appropriate choice of θ and then approximating the last term on the right hand side using an appropriate Edgeworth expansion.

Although Durbin considers 4 cases, we will focus on case 4, the most general, in order to facilitate comparisons. In this development, we assume $\mathbf{T}_n$ satisfies the conditions required for the expansion (28) of Durbin to hold. This general Edgeworth expansion requires that certain regularity conditions hold and that the cumulants are of the correct order. Conditions given in sections 4.2 or 4.5 are typical of what is required. The first step is to choose the value of θ on the right hand side of (5.5). This is done by using the value $\tilde{\theta}$ such that $E(\mathbf{T}_n|\theta) = \mathbf{t}$, i.e.

$$\int (\mathbf{u}-\mathbf{t}) G(\mathbf{u},\tilde{\theta}) H_1(\mathbf{u}) d\mathbf{u} = 0.$$

When we use an Edgeworth expansion for $g(\mathbf{u},\tilde{\theta})$, the odd order terms disappear since we are expanding at the mean of $\mathbf{T}_n$. Substituting the expansion gives (21) of Durbin (1980a) i.e.

$$g(\mathbf{t},\theta_0) = (n/2\pi)^{m/2} |D_n(\mathbf{t})|^{-1/2} f(\mathbf{y},\theta_0)/f(\mathbf{y},\tilde{\theta})$$
$$\times \left\{ 1 + \sum_{k=2}^{[r/2]} n^{-k+1} P_{n,2k}(0,t) + o(n^{-r/2+1}) \right\}$$

where $D_n(\theta) = nE\{\mathbf{T}_n - E(\mathbf{T}_n)\}\{\mathbf{T}_n - E(\mathbf{T}_n)\}^T$ and $P_{n,j}(\mathbf{x},\theta)$ is a generalized Edgeworth polynomial defined by

$$\frac{|D_n(\theta)|^{-1/2}}{(2\pi)^{m/2}} \exp\{-\mathbf{x}^T D_n^{-1}(\theta)\mathbf{x}\} P_{n,j}(\mathbf{x},\theta)$$
$$= \frac{1}{(2\pi)^m} \int_{R_m} \exp\left\{ -i\mathbf{z}^T\mathbf{x} - \frac{1}{2}\mathbf{z}^T D_n(\theta)\mathbf{z} \right\} \pi_{nj}(\mathbf{z},\theta) d\mathbf{z}$$

(see Durbin 1980a for details).
Using only the first term we have

$$g(\mathbf{t},\theta_0) = (n/2\pi)^{m/2} |D_n(\mathbf{t})|^{-1/2} f(\mathbf{y},\theta_0)/f(\mathbf{y},\tilde{\theta}) \{1 + 0(n^{-1})\}. \tag{5.6}$$

If we compare this approximation to (4.25) we can see similarities. The expression for $c^{-n}(\mathbf{t})$ has been replaced by $f(\mathbf{y},\theta_0)/f(\mathbf{y},\tilde{\theta})$ and $|detA||det\Sigma|^{-1/2}$ has been replaced by $|D_n(\mathbf{t})|$. To explore this further, recall that in small sample asymptotics, there are two steps; the recentering, followed by the use of the first term in an Edgeworth expansion, namely a normal approximation, Equation (5.5) corresponds to (4.23) relating the density under f with the density under the conjugate distribution.

Using the notation of this section, the centering formula from chapter 4 would require (equations are written in univariate form for ease of notation)

$$\int (u-t) \exp\{\alpha(t)(u-t)\} G(u,\theta_0) H_1(u) du = 0.$$

Because of the sufficiency, Durbin uses

$$\int (u-t)G(u,\tilde{\theta})H_1(u)du = 0$$

to recenter where $\tilde{\theta}$ is analagous to $\alpha(t)$. As can be seen, the expression $\exp\{\alpha(t)(u-t)\}G(u,\theta)$ has been replaced by $G(u,\tilde{\theta})$ so that we are in effect using a different conjugate density. In each case, the recentering requires a rescaling of the new density. In chapter 4, this is done with $c^{-n}(t)$, while in (5.5) this is achieved with $f(y,\theta_0)/f(y,\tilde{\theta})$. In summary, the existence of a sufficient statistic enables us to recenter the density without the necessity of knowing the cumulant generating function. The approximation step is essentially the same in both the small sample asymptotics and Durbin's approach.

The major advance brought about by Durbin's approach is that approximation (5.6) does not require that the observations be independent. This makes the results very useful in time series settings. It is the existence of the sufficient statistic which provides the essential simplication to make the technique feasible for the case of dependent observations. In several of the cases Durbin considers, he assumes $E[\mathbf{T}_n|\theta] = \theta$ up to order $1/n$, so that $\tilde{\theta} = \mathbf{t}$ and the approximation becomes

$$\hat{g}(\mathbf{t},\theta_0) = (n/2\pi)^{m/2}|D(\mathbf{t})|^{-1/2} f(\mathbf{y},\theta_0)/f(\mathbf{y},\mathbf{t}).$$

We now turn to the special case of the exponential family with n independent observations $x_1,\cdots,x_n$. The density is given by

$$f(x,\theta) = \exp\{\theta u(x) - K(\theta) + d(x)\}.$$

The argument here is for a one-dimensional θ but it carries through routinely for the p-dimensional case. Although there are several approaches which all lead to the same result, we will use Theorem 4.1 to illustrate the connection of the small sample asymptotic approach to that of Durbin and of Barndorff-Nielsen.

We want to approximate the density of the maximum likelihood estimate T_n (or $\hat{\Theta}$) at a point t (or $\hat{\theta}$) with an underlying density $f(x;\theta_0)$. We will develop the formula in terms of T_n and t in order to remain consistent with the notation to date. The score function is $\psi(x,t) = u(x) - K'(t)$ and $E_\theta u(x) = K'(\theta)$. The conjugate density is $h_t(x) = c(t)\exp\{\alpha(t)(u(x)-K'(t)) + \theta_0 u(x) - K(\theta_0) + d(x)\}$. The centering condition requires that $E_{h_t}(u(x) - K'(t)) = 0$. If we choose $\alpha(t) = t - \theta_0$, then h_t is exponential with parameter t and $E_{h_t}u(x) = K'(t)$ as required . Now

$$c^{-1}(t) = \exp\{K(t) - tK'(t)\}/\exp\{K(\theta_0) - \theta_0 K'(t)\}$$

$$a(t) = E_{h_t}[\partial(u(x) - K'(t))/\partial t] = -K''(t)$$

$$\sigma^2(t) = E_{h_t}(u(x) - K'(t))^2 = K''(t).$$

From this, approximation (4.8) becomes

$$\hat{g}_n(t) = (n/2\pi)^{1/2}\frac{\exp\{nK(t) - ntK'(t)\}}{\exp\{nK(\theta_0) - n\theta_0 K'(t)\}}(K''(t))^{1/2}.$$

To compare with formula (13) of Reid (1988), we replace t by $\hat{\theta}$ and note that in $L(\hat{\theta}) = \exp\{\hat{\theta}\Sigma u(x_i) - nK(\hat{\theta}) + \Sigma d(x_i)\}$, $\hat{\theta}$ satisfies the condition that $\Sigma u(x_i) = nK'(\hat{\theta})$ i.e. the maximum likelihood equation. Hence

$$\hat{g}_n(\hat{\theta}) = (n/2\pi)^{1/2}\{L(\theta_0)/L(\hat{\theta})\}(K''(\hat{\theta}))^{1/2} \tag{5.7}$$

with the convention that $\Sigma u(x_i) = nK'(\hat{\theta})$. If we replace $(K''(\hat{\theta}))^{1/2}$, the expected information by $j(\hat{\theta}) = -\partial^2 \log L(\theta)/\partial\theta^2|_{\theta=\hat{\theta}}$, then we obtain formula 13 of Reid (1988). (5.7) is often referred to as Barndorff- Nielsen's formula and appears in Barndorff-Nielsen (1980, 1983). It is the same as both the small sample approximation and the approximation given by Durbin for sufficient statistics. In order to put the results in historical perspective, Henry Daniels had noted this result expressed in (5.7) in a discussion of a paper by Cox (1958). It is interesting to see how many of the results used today come directly from the pioneering work of Henry Daniels (1954).

To conclude this section, we show the form of approximation (4.25) for a curved exponential family. The development is based on work by Hougaard (1985) and we use his setting and notation. Assume

$$\begin{aligned} f(x,\theta) &= \exp\{\theta' \mathrm{t}(x) - K(\theta) + h(x)\} \\ &= \exp\{\theta' \mathrm{t}(x) + h(x)\}/\phi(\theta). \end{aligned}$$

The parameter θ is a function $\theta(\beta)$ of a p-dimensional parameter β. We are interested in approximating the density of $\hat{\beta}$, the maximum likelihood estimate of β. This setting includes non-linear regression with normal errors, logistic regression and log-linear models. $\hat{\beta}$ is obtained as the solution of

$$n(\bar{t} - \tau(\theta(\beta))' d\theta/d\beta = 0 \quad \text{where} \quad \tau(\theta) = E_\theta \mathrm{t}(X) = K'(\theta).$$

The equation which α must satisfy is

$$\int \psi_r(x,\beta)\exp\{\Sigma\alpha_j\psi_j(x,\beta)\}f(x)dx = 0, \quad r = 1,\cdots,p$$

or equivalently

$$\int \frac{d}{d\alpha_r}\exp\{\Sigma\alpha_j\psi_j(x,\beta)\}f(x)dx = 0.$$

If we substitute, the left hand side becomes

$$\int \frac{d}{d\alpha_r}\exp\left[(\mathrm{t}(x) - \tau(\theta))'\frac{d\theta}{d\beta}\alpha + \mathrm{t}(x)'\theta_0 - K(\theta_0) + h(x)\right]dx.$$

Assuming we can interchange integration and differentiation, it is possible to carry out the integration to obtain

$$\frac{d}{d\alpha_r}\left[\exp\left(-\tau(\theta)'\frac{d\theta}{d\beta}\alpha\right)\exp\left(K\left(\theta_0 + \frac{d\theta}{d\beta}\alpha\right) - K(\theta_0)\right)\right].$$

To have this derivative zero, it suffices to have the derivative of the log equal to 0. i.e.

$$\left(\tau\left(\theta_0 + \frac{d\theta}{d\beta}\alpha\right)\right) - \tau(\theta(\beta))'d\theta/\partial\beta_r = 0.$$

Since the conjugate is exponential with parameter $\theta^* = \theta_0 + d\theta/d\beta\alpha$, it is straightforward to compute c^{-1}, Σ, A. The resulting approximation is

$$\begin{aligned} f_n(\beta) =&(n/2\pi)^{p/2} \exp\left\{ n\big(K(\theta_0) - K(\theta^*) - \tau(\theta)' \frac{d\theta}{d\beta}\alpha \right\} \\ &\left[\left| \frac{d\theta'}{d\beta} \frac{d^2 K}{d\theta^2} \frac{d\theta}{d\beta} - \{\tau(\theta^*) - \tau(\theta)\}' \frac{d^2\theta}{d\beta^2} \right| \left| \frac{d\theta'}{d\beta} \frac{d^2 K}{d\theta^{*2}} \frac{d\theta}{d\beta} \right|^{-1/2} \right] \\ &\times \left[1 + 0(1/n)\right] \end{aligned} \tag{5.8}$$

where $\theta^* = \theta_0 + d\theta/\partial\beta\alpha$ with α given by $(\tau(\theta^*) - \tau(\theta))' d\theta/d\beta = 0$.
In the last section of this chapter, this formula will be applied to an example. It is important to recall that (5.8) is the small sample approximation (4.25) applied to a curved exponential family. The form of the exponential leads to simplification in that the integration in (4.25) can be carried out explicitly. Result (5.8) is given in Hougaard (1985) as Theorem 1 which he in turn attributes to Skovgaard (1985).

5.5. CONDITIONAL SADDLEPOINT

Consider the situation where we require an approximation to the density of a statistic T_1 given that $T_2 = t_2$. In what follows we assume that we can approximate the density of $T = (T_1, T_2)$ as well as the density of T_1. The most direct way to proceed is to use a small sample approximation $\hat{g}_n(t_1, t_2)$ for the joint density, a small sample approximation $\hat{g}_n(t_2)$ for the density of T_2 and then divide the two approximations to give an approximation to the conditional density. In the literature, this approximation is referred to as the double saddlepoint approximation. To be specific we would choose $(\alpha_1(\mathbf{t}), \alpha_2(\mathbf{t}))$ in the joint conjugate to center (T_1, T_2) at the point $\mathbf{t}$ and $\alpha(t_2)$ in the conjugate for T_2 to center T_2 at t_2. Note that it may be a non-trivial process to compute $\alpha(t_2)$ for the marginal density of T_2. If T_1 and T_2 are both means, the score functions can be solved independently of each other making $\alpha(t_2)$ easy to compute. In other cases, this is not usually the case and there is no clear way to proceed.

In the case where we have sufficient estimators, Durbin provides a method of approximating the conditional density (section 4, Durbin 1980a). Assume $\mathbf{T}$ is a sufficient estimate of θ with bias of order n^{-1} at most and that $T = (T_1, T_2)$ and $\theta = (\theta_1, \theta_2)$. By sufficiency, the joint density of (T_1, T_2) can be written as $g(t_1, t_2; \theta_1, \theta_2)$ and $f(x, \theta) = g(t_1, t_2; \theta_1, \theta_2)h(x)$.

Durbin considered the case where the conditional density of T_1 given T_2 depends only on θ_1. Using arguments as in section 5.4, we have

$$g(t_1, t_2; \theta_{10}, \theta_2) = \frac{f(x; \theta_{10}, \theta_2)}{f(x; t_1, t_2)} g(t_1, t_2; t_1, t_2)$$

where θ_{10} is the particular value at which we require the approximation. Similarly, by sufficiency

$$g_2(t_2; \theta_{10}, \theta_2) = \frac{f(x; \theta_{10}, \theta_2)}{f(x; \theta_{10}, t_2)} g_2(t_2; \theta_{10}, t_2).$$

Dividing the equations yields

$$g(t_1 | t_2, \theta_{10}) = \frac{f(x; \theta_{10}, t_2)}{f(x; t_1, t_2)} \frac{g(t_1, t_2; t_1, t_2)}{g_2(t_2; \theta_{10}, t_2)}.$$

Effectively the sufficiency and unbiasedness has enabled us to center the densities at t_1 and t_2. It remains only to replace $g(t_1,t_2;t_1,t_2)$ and $g_2(t_2;\theta_{10},t_2)$ by their normal approximations to obtain an approximation with an error term of order n^{-1}; cf. also Skovgaard (1987).

Another case of interest is when we have an ancillary statistic. This case has been studied at length by Barndorff-Nielsen in several papers (cf. Barndorff-Nielsen 1983, 1984, 1986). To be specific, assume T_1 is an estimate of θ_1 and we have a minimal sufficient statistic (T_1,T_2) where T_2 is ancillary. Hence $f(x,\theta_1) = g(t,\theta_1)h(x) = g_1(t_1|t_2,\theta_1)g_2(t_2)h(x)$ since T_2 is ancillary. We can replace θ_1 by t_1 in the expression above and then divide expressions to obtain as Durbin does,

$$g_1(t_1|t_2,\theta_1) = \frac{f(x,\theta_1)}{f(x,t_1)}g(t_1|t_2,t_1).$$

The final step is to approximate $g(t_1|t_2,t_1)$ by its normal approximation. One way to do this is to use the limiting variance $D(\theta_1)$ of $\sqrt{n}(T_1-\theta_1)$ under the conditional distribution of T_1 given T_2. This gives us the approximation

$$g_1(t_1|t_2,\theta_1) = \left(\frac{n}{2\pi}\right)^{m_1/2}|D(t_1)|^{-1/2}\frac{f(x,\theta_1)}{f(x,t_1)}\{1+0(n^{-1})\} \tag{5.9}$$

where m_1 is the dimension of θ_1. This is exactly expression (27) of Durbin (1980a).

To relate this formula to the extensive work of Barndorff-Nielsen, it is useful to rewrite (5.9) as

$$f_{\hat\theta|A}(\hat\theta|a;\theta) = c(\theta,a)|j(\hat\theta)|^{1/2}\left\{L(\theta;\hat\theta,a)/L(\hat\theta;\hat\theta,a)\right\}\left\{1+0\left(\frac{1}{n}\right)\right\} \tag{5.10}$$

(cf. Reid 1988, formula 15).

Although the notation is different, the two formulas are the same except that in (5.10) the approximate density has been renormalized. The original work by Barndorff-Nielsen focussed on exponential families and transformation models and showed that formula (5.10) is exact in a number of cases.

McCullagh (1984) considers a fairly general situation and shows that formula (5.10) is generally valid (cf. (37) and the preceeding argument in McCullagh). However it is not easy to see how to construct the required second-order locally ancillary statistic A which is needed to carry out computations.

The simplest example where (5.10) is exact is that of the location/scale family. In this case the ancillary $\mathbf{a}$ is given by $\mathbf{a} = (a_1,\cdots,a_n)$, $a_i = (x_i-\hat\mu)/\hat\sigma$ where $(\hat\mu,\hat\sigma)$ is the maximum likelihood estimate of (μ,σ). Fisher (1934) showed that the density for $(\hat\mu,\hat\sigma)$ given the ancillary $\mathbf{a}$ can be written as

$$f(\hat\mu,\hat\sigma|\mathbf{a};\mu,\sigma) = c_0(\mathbf{a})\hat\sigma^{n-2}f(\mathbf{x};\mu,\sigma) \tag{5.11}$$

where the $\mathbf{x}$ on the right hand side is expressed as

$$\mathbf{x} = (\hat\mu+\hat\sigma a_1,\cdots,\hat\mu+\hat\sigma a_n).$$

To see how this is related to formula (5.10), note that $L(\mu,\sigma;\mathbf{x}) = f(\mathbf{x};\mu,\sigma)$ so that $L(\hat\theta;\hat\theta,\mathbf{a}) = \hat\sigma^{-n}f(\mathbf{a};0,1)$. Also $|j(\hat\theta)| = D(\mathbf{a})\hat\sigma^{-4}$ where

$$D(\mathbf{a}) = \{\Sigma g''(a_i)\}\{n+\Sigma a_i^2 g''(a_i)\} - \{\Sigma a_i g''(a_i)\}^2$$

with $g(\mathbf{x}) = -\log f(\mathbf{x})$. Hence

$$|j(\hat{\theta})|^{1/2}\{L(\theta;\hat{\theta},\mathbf{a})/L(\hat{\theta};\hat{\theta},\mathbf{a})\} = D^{1/2}(\mathbf{a})\hat{\sigma}^{n-2}f(\mathbf{x};\mu,\sigma)/f(\mathbf{a};0,1)$$

where $\mathbf{x}$ is expressed as above.
Formula (5.11) can be written as

$$c(\mathbf{a})\hat{\sigma}^{n-2}f(\mathbf{x};\mu,\sigma) \quad \text{where} \quad c(\mathbf{a}) = c_0(\mathbf{a})f(\mathbf{a};0,1)/D^{1/2}(\mathbf{a})$$

and we see that for location/scale (5.10) is exact. The reader is referred to Barndorff-Nielsen for other examples where exactness holds.

It is interesting to consider the relationship between the approximation conditional on an ancillary statistic and the unconditional approximation. Although this relationship is not clear the situation of observations from a normal with mean θ and variance $b^2\theta^2$ with b^2 known is one in which computations could be carried out relatively easily. The conditional formula is given in Reid (1988, cf formula 17) and is known to be exact. The unconditional formula can be worked out with some effort. $\alpha(t)$ can be computed explicitly so it appears the approximation can be given explicitly and the two approximations compared.

There is a need for more research to establish connections between the basic small sample approximations and the work of Barndorff-Nielsen.

5.6. NONTRIVIAL APPLICATION IN THE EXPONENTIAL FAMILY: LOGISTIC REGRESSION

In this section, we consider the example of logistic regression through the origin. Since the model falls within the exponential family, we might expect the approximation to be very straightforward. However, as we shall see, there are some complications in obtaining useful results. Consider the usual set-up for logistic regression through the origin with

$$P[Y=1] = e^{\beta x}/(1+e^{\beta x}).$$

We want to approximate the density of $\hat{\beta}$, the maximum likelihood estimate of β. In this example, we consider X to be random with a density $f(x)$. The situation in which x is considered fixed has no essential differences except for some increased complexity in notation.

Assume that X has a density $f(x)$. To follow the notation developed to date, $\psi(y,x,\beta)$ is the derivative of the log likelihood function and is given by

$$\psi(y,x,\beta) = yx - x\exp(x\beta)/(1+\exp(x\beta)).$$

A direct approach is to evaluate $f_n(t)$ using the approximation (4.8) for M-estimates. It is straightforward to verify that $\alpha(t) = t - \beta$ and the approximation can easily be computed.

If these computations are done and the results compared to the asymptotic results, there are some clear discrepancies. For instance if $f(x)$ is normal and $\beta = 0$, the variance for a fixed n can be computed from $f_n(t)$ and can be approximated on the basis of the asymptotic variance. The results are as follows:

	n				
Variance of $\hat{\beta}$	5	10	20	40	100
based on $f_n(t)$	19.6	2.2	.32	.12	.04
based on asymptotic variance	.8	.4	.2	.1	.04

It's clear that in the range of interest, namely 5–20, the approximation is giving results in which the density is much longer-tailed than is suggested by the asymptotic theory. To understand this problem, it is necessary to look at the equation for solving β i.e.

$$\Sigma y_i x_i - \Sigma x_i \exp(\beta x_i)/(1+\exp(\beta x_i) = 0.$$

If $y_i = 1$ for all positive x_i's and 0 for all negative x_i's , then $\beta = \infty$ is the maximum likelihood estimate. The observed long tail of the approximation for small to moderate n is due to a positive mass at $\pm\infty$. $f_n(t)$ is a smooth approximation to a mixture of a continuous density and point masses at $\pm\infty$.

In practice, the density of interest is the density of $\hat{\beta}$ conditional on $\hat{\beta}$ finite since the experimenter will only be interested in making inferences about β in the case that $\hat{\beta}$ is finite. We now proceed with adapting our basic approximation to handle this case.

To obtain an approximation, we need to find an appropriate conjugate density and verify a centering lemma. As a first step, consider the moment generating function $\hat{\beta}$ and $\Sigma\psi(y, x, \beta)$ conditional on $\hat{\beta}$ finite. The true value of β will be denoted as β_0. For ease of discussion, we assume that all the x_i's are greater than or equal to 0. It then follows that

$$\hat{\beta} \text{ finite } \Leftrightarrow 1 \leq \sum_{i=1}^{n} y_i \leq n-1.$$

Also

$$\begin{aligned} P(\hat{\beta} \text{ finite}) = 1 &- \left(\int \exp(\beta_0 x)/(1+\exp(\beta_0 x))f(x)dx\right)^n \\ &- \left(\int 1/(1+\exp(\beta_0 x))^n f(x)dx\right)^n \\ &= h(\beta_0) \quad \text{(say)}. \end{aligned}$$

The conditional density of (X_i, Y_i), $i = 1, \cdots, n$ given $\hat{\beta}$ finite is

$$\exp\left(\beta \sum_{i=1}^{n} x_i y_i\right) \Big/ \left[\prod_{i=1}^{n}(1+\exp(\beta x_i))\right] P(\hat{\beta} \text{ finite})$$

$$\text{if } 1 \leq \sum_{i=1}^{n} y_i \leq n-1$$

and 0 otherwise.

The conjugate density at a point β in this case will be a joint density in n dimensions given by:

$$h_\beta(\mathbf{x},\mathbf{y}) = c^n(\beta)\exp\left\{\beta_0 \sum_{i=1}^{n} y_i x_i + \alpha \sum_{i=1}^{n} \psi(y_i, x_i, \beta)\right\} \prod_{i=1}^{n} f(x_i)$$

$$I_{[1,n-1]}\left(\sum_{i=1}^{n} y_i\right) \Big/ \prod_{i=1}^{n}(1+\exp(\beta_0 x_i))$$

where I_A is the indicator function for the set A. α must satisfy the condition

$$E_{h_\beta}\left[\sum_{i=1}^{n}\psi(Y_i,X_i,\beta)\right]=0. \tag{5.12}$$

Note that to solve (5.12) in its current form requires an n-dimensional summation (over y_i's) and an n-dimensional integration (over x_i's) and so is not computationally feasible. We will show how (5.12) can be simplified to obtain a computationally manageable form.

By using a proof very similar to that given in section 4.5 for the centering result (4.23), it can be shown that a similar result holds in this case, namely

$$f_n(\beta)=c^{-n}(\beta)h_{\beta,n}(\beta)$$

where f_n represents the density of $\hat{\beta}$ under joint density of X and Y given by $f(x)$ and β_0 and $h_{\beta,n}(\cdot)$ represents the density of $\hat{\beta}$ under h_β. The densities are all conditional on $\hat{\beta}$ finite.

We now simplify (5.12). To begin let

$$g_\beta(x,y)=c(\beta)\exp\{\beta_0 yx+\alpha\psi(y,x,\beta)\}f(x)/(1+\exp(\beta_0 x_i)).$$

Now

$$h_\beta(\mathbf{x},\mathbf{y})=\prod_{i=1}^{n}g_\beta(x_i,y_i)I_{[1,n-1]}\left(\sum_{i=1}^{n}y_i\right).$$

To simplify notation, let Σ' denote the sum over all vectors $\mathbf{y}$ of 0's and 1's with $1\le\sum_{i=1}^{n}y_i\le n-1$. Now (5.12) becomes

$$\Sigma'\int_{\dots}\int\sum_{i=1}^{n}\psi(x_i,y_i,\beta)g_\beta(x_i,y_i)\prod_{j\neq i}g_\beta(x_j,y_j)d\mathbf{x}=0.$$

Let $\ell(y,\beta)=\int g_\beta(x,y)dx=\int\exp\{\beta_0 y+\alpha\psi(y,x,\beta)\}f(x)/(1+\exp\beta_0 x)dx$. We can now write our equation as

$$\Sigma'\sum_{i=1}^{n}\int\psi(x_i,y_i,\beta)g_\beta(x_i,y_i)\prod_{j\neq i}\ell(y_i,\beta)dx_i=0$$

or rearranging the summation, obtain

$$\sum_{r=1}^{n-1}\binom{n}{r}\Bigg[\int\sum_{i=1}^{r}\psi(x_i,1,\beta)g_\beta(x_i,1)(\ell(1,\beta))^{r-1}(\ell(0,\beta))^{n-r}dx_i$$
$$+\int\sum_{i=r+1}^{n}\psi(x_i,0,\beta)g_\beta(x_i,0)(\ell(1,\beta))^{r}(\ell(0,\beta))^{n-r-1}dx_i\Bigg]=0.$$

Simplifying, (5.12) can be written as:

$$\sum_{r=1}^{n-1}(\ell(1,\beta))^{r-1}(\ell(0,\beta))^{n-r-1}\binom{n}{r}\Bigg[r\ell(0,\beta)\int\psi(x,1,\beta)g_\beta(x,1)dx$$
$$+(n-r)\ell(1,\beta)\int\psi(x,0,\beta)g_\beta(x,0)dx\Bigg]=0.$$

Now instead of having to evaluate an n-dimensional integral, we have to evaluate four one-dimensional integrals making the computation of $\alpha(\beta)$ straightforward. The approximation to the density of $\hat{\beta}$ conditional on $\hat{\beta}$ finite, $f_n(\beta)$, is then given by

$$g_n(\beta) = \left(\frac{n}{2\pi}\right)^{1/2} c^{-n}(\beta)a(\beta)/\sigma(\beta)$$

where $\sigma^2(\beta)$ has the same form as the left-hand side of (5.12) with ψ^2 instead of ψ and a division by $P(\hat{\beta}$ finite) and $a(\beta)$ replaces ψ by $\partial\psi/\partial\beta$ and similarly includes a division by $P(\hat{\beta}$ finite). Finally

$$c^{-1}(\beta) = \sum_{r=1}^{n-1} \binom{n}{r} (\ell(\beta,1))^r (\ell(\beta,0))^{n-r} / P(\hat{\beta} \text{ finite}).$$

The extension of this approximation to higher dimension can be made if we retain the condition that the x_i's are positive. If the x_i's can be both positive and negative, the condition that $\hat{\beta}$ be finite involves both x_i and y_i and it becomes more difficult to simplify (5.12) to involve only one-dimensional integrals.