6. APPROXIMATE CONFIDENCE INTERVALS IN MULTIPARAMETER PROBLEMS

6.1. INTRODUCTION

As has been noted in the previous chapters, a limiting aspect of the saddlepoint approximation has been the computational complexity in higher dimensions. In a problem with p parameters, we are often interested in approximating the marginal density of the estimate of one of the parameters or the density of some test statistics. We could in principle use the approximation to obtain the density on a grid in p-dimensions and then integrate over appropriate regions of p-dimensional space. However, at each grid point, we have to solve a non-linear system of p equations to obtain α. If p is three or more this is not really a feasible method numerically.

In this chapter, we introduce a technique developed by Tingley and Field (1990) designed to overcome this problem. In fact the approach to be presented will be based on a nonparametric bootstrap and will allow us to obtain correct one or two-sided second order confidence intervals without specifying an underlying density. One of the tools that is essential is the tail area approximation due to Lugannani and Rice (1980). This approximation eliminates the integration of the approximate density in calculating tail areas.

The next section discusses the tail area approximation while the third section demonstrates how to compute confidence intervals which are both robust and nonparametric.

6.2. TAIL AREA APPROXIMATION

Up to this point, we have developed approximations for densities of estimates. However in many situations, there is more interest in approximating the cumulative distribution. In particular, for confidence intervals and testing procedures, it is the tail area of the distribution which is of interest. In this section we will develop a tail area approximation for the case of the univariate mean. This approximation will then be used in the next section to construct confidence intervals in the multiparameter situation. The tail area approximation is based on uniform asymptotic expansions and was developed for tail areas by Lugannani and Rice (1980). Both Daniels (1987) and Tingley (1987) have placed the result in the context of small sample or saddlepoint approximations. Our development is similar to that of Daniels and Tingley but with some notational changes.

Consider the situation where we have n independent, identically distributed random variables, $X_1, X_2, \cdots, X_n$ and we want to approximate $P(\bar{X} \geq x_0)$ for some point x_0. Based on our previous approximation for the density of $\bar{X}$ we could approximate the upper tail area by $\int_{x_0}^{\infty} kc^{-n}(x)/\sigma(x)dx$ where k is the normalizing constant. In order to evaluate this integral, we have to evaluate $c(x)$ and $\sigma(x)$ (and hence $\alpha(x)$) over a grid of points from x_0 to ∞. As noted by Daniels (1987), the integration can be made simpler by a change of variables. The upper tail area can be written as $\int_{x_0}^{\infty} kc^{-n}(\alpha(x))/\sigma(\alpha(x))dx$. The monotone transformation $y = \alpha(x)$, gives the integral $\int_{\alpha(x_0)}^{\infty} kc^{-n}(y)\sigma(y)dy$ which avoids the necessity of computing the saddlepoint $\alpha(x)$ for every ordinate. Related work by Robinson (1982) gives a tail area approximation based on the Laplace approximation to the integral above.

The Lugannani-Rice tail area approximation will be based on the values of $c(x_0)$ and $\sigma(x_0)$ so we only require one evaluation of α, namely $\alpha(x_0)$. This is a considerable simplication in computation and, as we shall see, gives remarkable accuracy.

We assume X_i has a density f, a mean of 0 and a cumulant generating function $K(t)$ defined by

$$e^{K(t)} = E(e^{tX_1}).$$

The Fourier inversion formula gives the density of $\bar{X}$ as

$$f_n(x) = \frac{n}{2\pi} \int_{-\infty}^{\infty} e^{n(K(iu)-iux)} du.$$

Reversing the order of integration gives

$$\begin{aligned} P(\bar{X} \geq x_0) &= \int_{x_0}^{\infty} \frac{n}{2\pi} \int_{-\infty}^{\infty} e^{n(K(iu)-iux)} du dx \\ &= \frac{n}{2\pi} \int_{-\infty}^{\infty} e^{n(K(iu)-iux_0)} du/iu \\ &= \frac{n}{2\pi i} \int_{-i\infty}^{i\infty} e^{n(K(z)-zx_0)} dz/z. \end{aligned} \tag{6.1}$$

We denote the solution of the equation, $K'(z) = x_0$ by α. Note that α is the saddlepoint and in the notation of the previous chapters would be denoted by $\alpha(x_0)$. The next step is to make a change of variables from z to t where

$$K(z) - zx_0 = t^2/2 - \gamma t + \rho.$$

γ and ρ are chosen so that $t = \gamma$ is the image of the saddlepoint $z = \alpha$ and the origin is preserved. This implies that $\rho = K(0) = \log \int f(x)dx = 0$ and $-\gamma^2 = 2(K(\alpha) - \alpha x_o)$. γ takes the same sign as α i.e. $\gamma = sgn(x_0)\sqrt{-2(K(\alpha) - \alpha x_0)}$.

As we will show later on in the section, the choice of this transformation implies that a local normal approximation is being used. Under this transformation, (6.1) becomes

$$P(\bar{X} \geq x_0) = \frac{n}{2\pi i} \int_{b-i\infty}^{b+i\infty} e^{n(t^2/2-\gamma t)} G_0(t) dt/t$$

where $G_0(t) = (t/z)(dz/dt)$.

The effect of the transformation is to take the term which must be approximated out of the exponent where errors can grow very rapidly into the main part of the integrand. In order to obtain the terms in the expansion that we need, $G_0(t)$ is replaced by a linear approximation as follows. To begin, write

$$G_0(t) = a_0 + a_1 t + t(t - \gamma)H_0(t).$$

From this representation, $a_0 = G_0(0)$ and $a_1 = (G_0(0) - G_0(\gamma))/\gamma$. We now have to evaluate $G_0(0)$ and $G_0(\gamma)$. G_0 has removable singularities at both these points.

Starting with $K(z) - zx_0 = t^2/2 - \gamma t$, we obtain

$$\frac{dz}{dt} = \frac{t-\gamma}{K'(z)-x_0} \quad \text{so that} \quad \lim_{t\to 0}\frac{dz}{dt} \text{ exists.}$$

Hence

$$\lim_{t\to 0} G_0(t) = \lim_{t\to 0}\frac{t}{z}\lim_{t\to 0}\frac{dz}{dt} = \lim_{t\to 0}\frac{1}{dz/dt}\lim_{t\to 0}\frac{dz}{dt} = 1$$

so that $a_0 = 1$. Also

$$\lim_{t\to\gamma} G_0(t) = \lim_{t\to\gamma}\frac{t}{z}\lim_{t\to\gamma}\frac{dz}{dt} = \frac{\gamma}{\alpha(K''(\alpha))^{1/2}}$$

since

$$\lim_{t\to\gamma}\left(\frac{dz}{dt}\right) = \lim_{t\to\gamma}\frac{(t-\gamma)}{K'(z)-x_0} = \lim_{t\to\gamma}\frac{-1}{K''(z)\frac{dz}{dt}}$$

and

$$\lim_{t\to\gamma}\frac{dz}{dt} = \frac{1}{(K''(\alpha))^{1/2}}$$

so that

$$a_1 = -\frac{1}{\gamma} + \frac{1}{\alpha(K''(\alpha))^{1/2}}.$$

Approximating $G_0(t)$ by $a_0 + a_1 t$, we obtain

$$P(\bar{X} \geq x_0) \approx \frac{n}{2\pi i}\int_{b-i\infty}^{b+i\infty}\frac{1}{t}\, e^{-t^2/2-\gamma t}dt$$

$$+ \frac{n}{2\pi i}\left(-\frac{1}{\gamma} + \frac{1}{\alpha(K''(\alpha))^{1/2}}\right)\int_{b-i\infty}^{b+i\infty} e^{t^2/2-\gamma t}dt$$

$$\approx 1 - \Phi(\sqrt{n}\gamma) + \frac{e^{\sqrt{n}\gamma}}{\sqrt{2n\pi}}\left\{\frac{1}{\alpha(K''(\alpha))^{1/2}} - \frac{1}{\gamma}\right\}.$$

The next step is to rewrite the formula in more familiar terms.

$$\gamma^2 = -2(K(\alpha) - \alpha x_0) = -2\left(\log\int e^{\alpha(x-x_0)}f(x)dx\right) = 2\log c(x_0)$$

and $\gamma = sgn(x_0)\sqrt{2\log c(x_0)}$. Similarly $K''(\alpha) = \sigma^2(\alpha)$. The approximation for the tail area becomes

$$P(\bar{X} \geq x_0) = 1 - \Phi(sgn(x_0)\sqrt{2n\log c(x_0)})$$

$$+ \frac{c^{-n}(t)}{\sqrt{2\pi}}\left\{\frac{1}{\alpha(x_0)\sigma(x_0)\sqrt{n}} - \frac{1}{sgn(x_0)\sqrt{-2n\log c(x_0)}}\right\}$$

for $x_0 > E(X)$. (6.3)

Daniels (1987) shows that the formula is exact if we add an error term $0(n^{-3/2})$ in the curly brackets of the second term in (6.3) (cf. Daniels formula (4.9)).

To demonstrate the accuracy of the approximation, we can consider situations in which the small sample approximation for the mean is exact. Then the error in the tail area when computed with (6.3) can be attributed to the error of the tail area approximation. The following numerical results in Exhibit 6.1 are taken from Daniels (1987) and Field and Wu (1988).

	$n=1$			$n=5$	
x_0	**Exact**	(6.3)	x_0	**Exact**	(6.3)
.5	.6065	.6043	.2	.99634	.99633
1	.3679	.3670	.6	.8153	.8152
3	.0498	.0500	1	.4405	.4405
5	$.0^2674$	$.0^2681$	2	.0293	.0293
7	$.0^3912$	$.0^2926$	3	$.0^3857$	$.0^3858$
9	$.0^3123$	$.0^3126$	4	$.0^4169$	$.0^4170$
			5	$.0^6267$	$.0^6268$

Exhibit 6.1a
Tail area for mean under exponential
$f(x) = e^{-x}$, $x \geq 0$, $K(z) = -\log(1-z)$

	$n=3$			$n=5$	
x_0	**Exact**	(6.3)	x_0	**Exact**	(6.3)
.33	.9645	.9638	.2	$.9^4466$	$.9^4460$
.67	.6782	.6724	.6	.8334	.8315
1	.3927	.3848	1	.4147	.4108
1.67	.1156	.1108	2	.0378	.0328
3.33	$.0^2548$	$.0^2505$	4	$.0^3148$	$.0^3141$
6.67	$.0^4174$	$.0^4155$	5	$.0^5994$	$.0^5937$

Exhibit 6.1b
Tail area for mean under inverse normal
$f(x;\mu) = \mu \exp\{-(x-\mu)^2/2x\}/(2\pi)^{1/2}x^{3/2}$
$K(z) = \mu(1-(1-z)^{1/2})$, $\mu = 1$

As can be seen the relative error introduced by (6.3) is small with the maximum being just above 10% in the tables above. This demonstrates that the tail area approximation works remarkably well for small sample sizes and in the extreme tails. The following example for chi square random variables uses (6.3) iteratively to compute the upper percentiles for x_5^2/df. We used the tail area routine from S to check our results in this case. S uses an algorithm developed by Goldstein (1973) to approximate the tail area. In that paper there are estimates of the maximum relative error up to an upper tail area of .0005. For the situation with $n = 5$, the following results in Exhibit 6.2 were obtained.

Upper tail area	.05	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}
Computed from S	2.2141	3.0173	4.1030	5.1490	6.1697	7.1616	8.1056
Computed from (6.3)	2.2150	3.0186	4.1049	5.1514	6.1740	7.1807	8.1760

Exhibit 6.2
Percentiles for χ_5^2/df

The maximum relative errors for the algorithm used by S are 10^{-2} for 5 degrees of freedom for tail areas up to 5×10^{-4}. For the tail areas in this range, we observe relative errors less than 4×10^{-4}, implying that approximation (6.3) again works very well.

As a final example we consider an example where the small sample approximation is not exact so that errors arise both from this and the tail area approximation; see Exhibit 6.3.

	$n = 1$			$n = 5$	
x_0	Exact	(6.3)	x_0	Exact	(6.3)
.2	.4	.3838	.2	.2250	.2249
.4	.3	.2750	.4	.0620	.0618
.6	.2	.1791	.6	$.0^2833$	$.0^2824$
.6	.1	.0948	.8	$.0^3260$	$.0^3255$

Exhibit 6.3
Tail area for mean under uniform
$f(x) = 1/2, -1 \leq x \leq 1, K(z) = \log(\sinh z/z)$

For this example, the tail area approximation works very well and the results from (6.3) are very close to those obtained by numerical integration.

If we consider the situation for robust location with a monotone score function, Daniels (1983) demonstrates that formula (6.3) remains valid. The following Exhibit 6.4 again demonstrates that (6.3) gives the same accuracy as numerical integration.

		5% Contaminated normal				Cauchy		
	t	Exact	Integrated saddlepoint	Lugannani-Rice	t	Exact	Integrated saddlepoint	Lugannani-Rice
	0.1	.46331	.46229	.46282	1	.25000	.28082	.28197
	1.0	.17601	.18428	.18557	3	.10242	.12397	.13033
n=1	2.0	.04674	.07345	.07082	5	.06283	.08392	.09086
	2.5	.03095	.06000	.05682	7	.04517	.06484	.07210
	3.0	.02630	.05520	.05190	9	.03522	.05327	.06077
	0.1	.42026	.42009	.42024	1	.11285	.11458	.11400
	1.0	.02799	.02779	.02799	3	.00825	.00883	.00881
n=5	2.0	.00414	.00413	.00416	5	.00210	.00244	.00244
	2.5	.00030	.00043	.00043	7	.00082	.00105	.00104
	3.0	.00018	.00031	.00031	9	.00040	.00055	.00055
	0.1	.39403	.39393	.39399	1	.05422	.05447	.05427
	1.0	.00538	.00535	.00537	3	.00076	.00078	.00078
n=9	2.0	.000018	.000018	.000018	5	.000082	.000088	.000088
	2.5	.000004	.000005	.000005	7	.000018	.000021	.000021
	3.5	.000002	.000003	.000003	9	.000006	.000006	.000007

Exhibit 6.4 (cf. Daniels, 1983)
Tail probability of Huber's M-estimate with $k = 1.5$

It is helpful in understanding the mechanism of the approximation to consider the case where the X_i's are normal with mean μ and variance σ^2. In this case,

$$K(z) = \mu z + \frac{\sigma^2}{2} z^2$$

and $K'(z) = x_0$ gives $\alpha = (x_0 - \mu)/\sigma^2$.
The transformation from z to t is given by $z = t/\sigma$ and the function $G_0(t) = 1$. Hence $a_0 = 1$, $a_1 = 0$ and $H_0(t) \equiv 0$ and the only non-zero term in the approximation is $1 - \Phi(\sqrt{n}\gamma)$ where $\gamma = (x_0 - \mu)/\sigma$ i.e. the approximation (6.3) gives

$$P(\bar{X} \geq x_0) = 1 - \Phi(\sqrt{n}(x_0 - \mu)/\sigma)$$

for normal random variables and is, of course, exact.

In the general case, the function $K(z) - zx_0$ is not parabolic and the Lugannani-Rice approximation proceeds by distorting $K(z) - zx_0$ so that it is parabolic. The first term in the approximation, $1 - \Phi(\sqrt{n}\gamma)$ comes from the normal approximation. The next terms arise from the non-linearity of the transformation of $K(z) - zx_0$ to a parabola. It should be noted that a different normal (or parabolic) approximation is used for each point x_0. This approach of a local normal approximation is of course the same as that used in deriving the small sample approximation for the density. The classic Edgeworth and the Fisher-Cornish inversion come from approximating $K(z)$ globally by a polynomial which behaves like $K(z)$ at the origin and matches derivatives at the origin. The local nature of (6.3), ensures the good accuracy of (6.3) even for very small n.

We now consider the so-called index of large deviations, namely

$$-\lim_{n\to\infty} \log P(T_n > t)/n$$

where T_n is an estimate. If T_n is a M-estimate of location with a monotone score function, then using a simple Laplacian approximation to the integrated density approximation, we have

$$P(T_n > t) = c^{-n}(t)/[\sigma(t)\alpha(t)(2\pi n)^{1/2}][1 + 0(1/n)].$$

This approximation works as well as (6.3) in the extreme tails. The extra terms in (6.3) are an adjustment for the situation when t is near the singularity in the integral at the mean. From this it follows that

$$-\lim_{n\to\infty} \log P(T_n > t)/n = \log c(t)$$

In addition, if T_n converges almost everywhere to $T(F)$, then the Bahadur half slope is given by

$$\lim_{n\to 0} \frac{1}{n} \log \int_{T_n}^{\infty} f_n(u)du = \lim_{n\to\infty} (-\log c(T_n))$$

$$= -\log c(T(F))$$

(see Bahadur 1971, section 7).

Remark 6.1
We can now complete the proof of (4.9). Let $g_n(t) = (n/2\pi)^{1/2}c^{-n}(t)A(t)/\sigma(t)$ denote the approximation to the density. The Laplacian approximation above for $\int_t^\infty g_n(t)dt$ gives that

$$\lim_{n\to\infty} n\int_t^\infty g_n(t)dt = \lim_{n\to\infty} \frac{n^{1/2}\exp(-n\log c(t))}{\sigma(t)\alpha(t)(2\pi)^{1/2}}\left(1+0\left(\frac{1}{n}\right)\right).$$

Since $\log c(t) > 0$ for $t > \theta_0$, it follows directly that the above limit is 0 as required to give $\int_{\delta_2+\theta_0}^\infty g_n(t)dt = 0(1/n)$.

Remark 6.2
It is worth noting that in the tail area approximation (6.3), we get very accurate results for the tail area beyond t using only the characteristics of the conjugate density at t. This suggests that the conjugate density is the natural mechanism for centering and accurately captures the tail area behavior.

To conclude this section we look at some results of Dinges (1986a,b,c) in which he tries to develop a coherent theory for distributions close to the normal. He obtains an approximation for the tails of various distributions including Student's t, Beta, Inverse normal, and Gamma. Moreover, he applies his technique to approximate the tail area of the distribution of the average of n iid random variables and that of M-estimators of location. Here we will discuss the connection between this approximation and those discussed so far.

The key point is the concept of Wiener germ which is defined as follows in Dinges (1986b).

Let U be a neighborhood of t_0. A *Wiener germ* of order m on U with center t_0 is a family of densities $\{f_\epsilon(t)|\epsilon \to 0\}$ of the following form

$$f_\epsilon(t) = (2\pi\epsilon)^{-p/2}\exp\{-K^*(t)/\epsilon\}\cdot D(t)\cdot\exp\{\epsilon\cdot S(\epsilon,t)\}, \tag{6.4}$$

where

a) $K^*(t)$, the entropy function , is $(m+1)$-times continuously differentiable with $K^*(t_0) = 0$, $K^*(t) \geq 0$, $K^{*\prime\prime}(t)$ positive definite;
b) $D(t)$, the modulating density , is positive and m−times continuously differentiable;
c) $\epsilon\cdot S(\epsilon,t) = \epsilon\cdot S_1(t)+\epsilon^2 S_2(t)+\cdots+\epsilon^{m-1}S_{m-1}(t)+\epsilon^m R(\epsilon,t)$, with the correcting functions $S_j(t)$ $(m-j)$-times continuously differentiable and the remainder term $R(\epsilon,t)$ uniformly bounded on U;
d)

$$\int_{t'}^{t''} f_\epsilon(t)dt = 1 - 0(\epsilon^m), \tag{6.5}$$

for some $t' < t_0 < t''$, $[t',t''] \subseteq U$.

The following interpretation will help to clarify this notion. Consider a diffeomorphism $V_0 : G \to U \subset \mathbf{R}^p$, where G is a neighborhood of the origin. Let $V_1(\cdot),\cdots,V_m(\cdot)$ be differentiable mappings on G. For small ϵ consider the mappings

$$V(\epsilon,\cdot) = V_0(\cdot)+\epsilon V_1(\cdot)+\cdots+\epsilon^m V_m(\cdot) \tag{6.6}$$

and denote by $W(\epsilon,\cdot) : U \to G_\epsilon$ its inverse. If the V_j are sufficiently smooth, there exist $W_0(t),\cdots,W_m(t)$ such that

$$W(\epsilon,t) = W_0(t) + \epsilon W_1(t) + \cdots + \epsilon^m W_m(t) + 0(\epsilon^{m+1}). \tag{6.7}$$

Then the distribution restricted to G_ϵ of a normal random variable with expectation 0 and covariance matrix ϵI is mapped by $V(\epsilon,\cdot)$ into a distribution on U with a density of the form $f_\epsilon(t)$ given by (6.4). Near the center t_0 the densities $f_\epsilon(t)$ are similar to the normal densities with mean t_0 and covariance matrix $\epsilon(K^{*\prime\prime}(t_0))^{-1}$.

An example of Wiener germ is provided by Daniels' result (3.19) written in terms of Legendre transform (see section 5.2). In fact, given a sequence of random variables $X_1,\cdots,X_n$, the densities of the mean $\bar{X}_n = n^{-1}\sum_{i=1}^{n} X_i$ follow a Wiener germ along $\epsilon = \epsilon_n = \frac{1}{n}$. In this case the entropy function K^* is the Legendre transform of the cumulant generating function, i.e. $K^*(t) = \sup\{\tilde{\alpha}t - K(\tilde{\alpha}) : \tilde{\alpha}\} = \alpha K'(\alpha) - K(\alpha)$ with $\alpha = \alpha(t)$ determined by $K'(\alpha) = t$, and the modulating density is $D(t) = (K^{*\prime\prime}(t))^{1/2}$.

Dinges gives then the following expansion for the tail area of a Wiener germ of order m

$$\int_t^\infty f_\epsilon(y)dy = \Phi\Big\{-\epsilon^{1/2}[W_0(t) + \epsilon W_1(t) + \cdots + 0(\epsilon^m)]\Big\}, \tag{6.8}$$

uniformly in the interval $[t_0 - c\sqrt{\epsilon}, t'']$, where c is an arbitrary but fixed positive number and Φ is the cumulative of the standard normal distribution. The coefficients of the expansion (6.8) $W_0, W_1, \cdots$ have a direct interpretation as the coefficients in the expansion (6.7) of the inverse W of the mapping $V(\epsilon,\cdots)$. They can be computed iteratively from the entropy function K^*, the modulating density D, and the correcting functions. For instance,

$$W_0(t) = (2K^*(t))^{1/2}$$
$$W_1(t) = -\frac{1}{W_0(t)}\log\big[(K^{'}(t))^{1/2}/W_0'(t)\big].$$

When (6.8) is applied to the distribution of the mean $\bar{X}_n$ of n iid random variables one obtains the following expansion ($\epsilon = \epsilon_n = 1/n$)

$$P[\bar{X}_n > t] = \Phi\Big\{-\sqrt{n}\big[W_0(t) + \frac{1}{n}W_1(t) + \cdots o(n^{-m})\big]\Big\}$$

or

$$-\frac{1}{\sqrt{n}}\Phi^{-1}\{P[\bar{X}_n > t]\} = W_0(t) + \frac{1}{n}W_1(t) + \cdots o(n^{-m}), \tag{6.9}$$

where

$$W_0(t) = (2K^*(t))^{1/2} = \alpha\cdot K'(\alpha) - K(\alpha),$$
$$W_1(t) = \frac{1}{W_0(t)}\log[w(t)/W_0(t)],$$
$$w(t) = K^{*\prime}(t)\cdot(K^{*\prime\prime}(t))^{1/2},$$

and $\alpha(t)$ is determined by the saddlepoint equation $K'(\alpha) = t$ or equivalently $K^{*\prime}(t) = \alpha$. In other words, one obtains up to the term of order $1/n$ in (6.9).

$$P[\bar{X}_n > t] \sim \Phi\Big\{-\sqrt{n}W_0(t) - \frac{1}{\sqrt{n}W_0(t)}\log[w(t)/W_0(t)]\Big\}. \tag{6.10}$$

A similar result can be obtained for M-estimators of location.

The expansion (6.9) should be compared with large deviations types of results where the quantity $\frac{1}{n}\log\{P[\bar{X}_n > t]\}$ is expanded. Since

$$\frac{1}{2}[\Phi^{-1}(p)]^2 \sim \log\frac{1}{p} \quad \text{when} \quad p \to 0,$$

Dinges argues from (6.9) that $-\Phi^{-1}\{P[\bar{X}_n > t]\}/\sqrt{n}$ is the quantity that ought to be expanded. Finally, note that a term expansion of formula (6.10) is similar to Lugannani and Rice (1980) expansion (see 6.3) given by (in this notation)

$$P[\bar{X}_n > t] \sim \Phi(-\sqrt{n}W_0(t)) + \frac{1}{\sqrt{n}}\phi(\sqrt{n}W_0(t))\left[\frac{1}{w(t)} - \frac{1}{W_0(t)}\right]$$

where ϕ is the standard normal density.

6.3. CONFIDENCE INTERVALS

In many estimation situations, it is of substantial interest to compute confidence intervals for parameters of interest. We have developed techniques for approximating densities in the previous chapters and our aim here is to use those approximations to construct confidence intervals. Consider the setting of section 4.5, namely that of an independent identically distributed sample $X_1, \cdots, X_n$ drawn from a population F_η involving an unknown p-dimensional parameter η. η is estimated by an M-estimate $\hat{\eta}$ as the solution of

$$\sum_{i=1}^{n} \psi_j(x_i, \eta) = 0, \quad j = 1, \cdots, p.$$

We want to construct an interval for a real-valued parameter $\theta = \theta(\eta)$. θ will be estimated by $\hat{\theta} = \theta(\hat{\eta})$. In a linear regression problem $\eta = (\beta_0, \beta_1, \sigma)$ where β_0 is the intercept and β_1 the slope and we are often interested in $\theta(\eta) = \beta_1$ as a parameter of interest. In a testing context θ may be the value of a test statistic. In this case, we are more interested in computing tail areas for $\hat{\theta}$. The confidence intervals will be constructed via test statistics and tail areas so that if the interest is in P-values, they can easily be obtained.

If we start with approximation (4.25) for $f_n(\mathbf{t})$, the density of $\hat{\eta}$, there are several problems arising in the construction of confidence intervals. The first is the choice of an appropriate test procedure to see whether a trial value θ_0 belongs to the interval or not. Given an appropriate test statistic, we need to compute a P-value on which to base our decision about θ_0. Although we have an approximation for the density of $\hat{\eta}$, to compute the marginal density of $\hat{\theta}$ requires evaluation of f_n over a p-dimensional space and then effectively integrating out $(p-1)$ dimensions. Each evaluation of f_n requires the solution of a system of p nonlinear equations. This procedure is feasible for $p = 2$ but becomes computationally infeasible for larger p's . We need to find a technique to reduce the computational complexity. Even without the computational difficulties, there still remains the problem of how to handle the nuisance parameters in constructing the interval for θ. As a final difficulty we may not want to specify F_η. The more natural way to overcome this is to replace F_η by an appropriate empirical density. This means that we are using a philosophy similar to that of the bootstrap but as we shall show, we avoid the resampling associated with the bootstrap. The procedure outlined here is based on Tingley and Field (1990).

To begin, we assume the assumptions in section 4.5 are met. Then from (4.24) we can write

$$(\hat{\eta} - \eta_0)_r = \sum_{j=1}^{p} b_{rj} \bar{Z}_j + o_p(1/\sqrt{n}) \tag{6.11}$$

where $\bar{Z}_j = \sum_{i=1}^{n} \psi_j(X_i, \eta_0)/n$ and $B = (b_{rj}) = -A(\eta_0)^{-1}$ and $A(\eta_0) = E_{\eta_0}\left[\frac{\partial \psi}{\partial \eta}(X, \eta_0)\right]$.

We now want to consider testing a point θ_0 to see whether it belongs to the interval or not. η_0 is unknown but satisfies $\theta(\eta_0) = \theta_0$. To construct a test statistic we expand $\hat{\theta} = \hat{\theta}(\eta)$ in a Taylor series expansion about η_0 using (6.11) and obtain

$$\hat{\theta} - \theta(\eta_0) = \sum_{i=1}^{n} g(X_i, \eta_0)/n + 0_p(1/n)$$

where $g(X_i, \eta_0) = \psi^T(X_i, \eta_0) B^T \partial\theta(\eta_0)/\partial\eta$. The random variables $g_i = g(X_i, \eta_0)$ are referred to as the configuration.

Since η_0 is assumed unknown, we actually work with the observed configuration $g_i = g(X_i, \hat{\eta})$. In order to keep the ideas clear we denote the observed $\hat{\eta}$ by $\hat{\eta}_{obs}$. Both $\hat{\eta}_{obs}$ and $\hat{\theta} = \theta(\hat{\eta}_{obs})$ are held fixed in what follows. Note that $\bar{g}_{obs} = 0$. Our test statistic is $\bar{g}$ which approximates $\theta - \theta_0$ with error $0_p(1/n)$. Although this may not appear to be accurate enough (cf. Hall, 1988), we are able to obtain intervals with good coverage by carefully approximating the density of $\bar{g}$. Since our test statistic is a mean, the small sample approximation for the density of a mean can be used. Before giving the coverage properties of our interval, we give the algorithm to construct the interval and illustrate it with an example.

Step 1:

Compute $\hat{\eta}_{obs}$, θ and $g_i = \psi^T(x_i, \hat{\eta}_{obs}) B^T \partial\theta(\eta)/\partial\eta$. $\hat{B}$ can be computed parametrically as the inverse of $E_{\eta_{obs}}\left[\frac{\partial \psi}{\partial \eta}(X, \hat{\eta}_{obs})\right]$ or nonparametrically as the inverse of $\hat{A} = \sum_{i=1}^{n} \frac{\partial \psi}{\partial \eta}(x_i, \eta)/n$.

Step 2:

Obtain an initial estimate of the distribution of the g_i. At this point either a parametric or nonparametric estimate can be used.

In this development we use a nonparametric estimate via the cumulant generating function $K(t)$.

Let $K(t) = \log\left(\sum_{i=1}^{n} e^{tg_i}/n\right)$.

Steps 1 and 2 must be computed once for every sample. Note that the mean of our estimated distribution, $K'(0) = \sum_{i=1}^{n} g_i/n = 0$. For each θ_0 under test we need to recenter the distribution at $\theta_0 - \hat{\theta}$, which is the approximate expected value of $g(Y, \hat{\eta}_{obs})$ under F_{η_0}. This is accomplished as in previous chapters by solving for α in the equation

$$\sum_{i=1}^{n} (g_i - (\theta_0 - \hat{\theta})) \exp\{\alpha(g_i - (\theta_0 - \hat{\theta}))\} = 0. \tag{6.12}$$

Now (6.12) is the centering result used before except that the density f has been replaced by the empirical distribution function.

Step 3:

For each θ_0 under test, compute $\alpha(\theta_0)$ as the solution of

$$\sum_{i=1}^{n}(g_i - (\theta_0 - \hat{\theta})(1-e))\exp\{\alpha(g_i - (\theta_0 - \hat{\theta})(1-e))\} = 0$$

where e is of order $1/n$.

The correction term e will be discussed later and is effectively a calibration correction.

Step 4:

Approximate $P_{\eta_0}[\bar{g} > 0]$ by

$$\hat{P}_{\theta_0}[\bar{g} > 0] = \Phi(-\sqrt{2n_1 \log c(\theta_0)}) - \frac{c^{-n_1}(\theta_0)}{\sqrt{2\pi n_1}}\left[\frac{1}{a(\theta_0)\sigma(0)} - \frac{1}{\sqrt{2\log c(\theta_0)}}\right]$$

i.e. use the tail area approximation (6.3), where $n_1 = n - p$ or $n - (p-1)$. Include θ_0 in the $(1-2\epsilon)100\%$ interval if $\epsilon < \hat{P}_\theta[\bar{g} > 0] < 1 - \epsilon$. Note that (6.3) is used with cumulant generating function $K(t) = K(t + \alpha_0) - K(t)$ where $\alpha_0 = \alpha(\theta_0)$.

It is more efficient to work backwards and replace steps 3 and 4 by

Step 3′ :

Find α_1 solutions of $P[\bar{g} > 0] = \epsilon$ and $P[\bar{g} > 0] = 1 - \epsilon$ where

$$\hat{P}[\bar{g} > 0] = \Phi(-\sqrt{2n_1 \log c(\alpha)}) - \frac{c^{-n_1}(\alpha)}{\sqrt{2\pi n_1}}\left\{\frac{1}{\alpha\sigma(0)} - \frac{1}{\sqrt{2n_1 \log c(\alpha)}}\right\}.$$

We write c and s as depending on α in this case, rather than θ_0. It is still the integrating constant of the conjugate density.

Step 4′ :

Find θ_1 and θ_2 solutions of

$$\sum_{i=1}^{n}(g_i - (\theta_j - \hat{\theta})(1-e))\exp\{\alpha_j(g_i - (\theta_j - \hat{\theta})(1-e))\} = 0, \quad j = 1, 2.$$

The estimated $(1-2\epsilon)100\%$ confidence interval is (θ_1, θ_2).

We now consider using this algorithm for the location/scale problem. For the sample $X_1, \cdots, X_n$ from an unknown distribution, location μ and scale σ are estimated by the M-estimates $\hat{\mu}$, $\hat{\sigma}$, solutions of

$$\frac{1}{n}\sum_{i=1}^{n}\psi(X_i, \eta) = 0$$

where $\eta = (\mu, \sigma)$ and

$$\psi_1(x, \eta) = \psi_k\left(\frac{x-\mu}{\sigma}\right)$$

$$\psi_2(x, \eta) = \psi_k^2\left(\frac{x-\mu}{\sigma}\right) - \beta,$$

$$\psi_k(y) = \begin{cases} -k & \text{if } y \le -k \\ y & \text{if } |y| < k \\ k & \text{if } y \ge k, \end{cases}$$

$$\beta = (n-1)\int_{-\infty}^{\infty}\psi_k^2(y)\phi(y)dy/n,$$

and $\phi(y)$, $\Phi(y)$ are the density function and cumulative distribution, respectively, of a standard normal random variable. The estimate η is referred to in the literature as Huber's Proposal 2 (1981). The constant k is usually between 1 and 2. We have used, exclusively, $k = 1.0$.

If the matrix $A = E[\partial\psi/\partial\eta]$ is calculated under the normal model, then

$$A = -\frac{1}{\sigma}\begin{bmatrix}\delta & 0\\ 0 & 2\epsilon\end{bmatrix}$$

where $\delta = \Phi(k) - \Phi(-k)$ and

$$\epsilon = \int_{-k}^{k} y^2\phi(t)dy,$$

Then $B = -A^{-1}$ is estimated by

$$\hat{B} = \hat{\sigma}\begin{bmatrix}1/\delta & 0\\ 0 & 1/2\epsilon\end{bmatrix}$$

and

$$g_i = \psi^T(x_i, \hat{\eta})\hat{B}^T\partial\theta(\hat{\eta})/\partial\eta = \hat{\sigma}\psi_k\Big((x_i - \hat{\mu})/\hat{s}\Big)/\delta. \tag{6.13}$$

If the matrix A is estimated empirically, then

$$g_i = \frac{\hat{\sigma}}{a_{11}a_{22} - a_{12}^2}\left\{a_{22}\psi_k\left(\frac{x_i - \hat{\mu}}{\hat{\sigma}}\right) - \frac{a_{12}}{2}\left(\psi_k^2\left(\frac{x_i - \hat{\mu}}{\hat{\sigma}}\right) - \beta\right)\right\}. \tag{6.14}$$

where the a_{ij} are the elements of $\hat{A}$ (see step 1).

By looking at numerical results, it can be shown that (6.14) is non-robust in that a small shift in an observation can lead to a large change in the configuration. On that basis we recommend the use of (6.13).

The following table taken from Tingley and Field (1990) shows the results which are obtained using a Monte Carlo swindle and were blocked by generating 2000 samples of size 30 and then using subsets of these for samples of size 5, 10 and 20. The classical samples are the usual t-intervals while the small sample intervals are as given above with $k = 1$.

Sample Size		Normal		Contaminated normal $.9N(0,1)+.1N(0,16)$		t_3	
		Classical	Small sample	Classical	Small sample	Classical	Small sample
5	2 sided coverage	.95	.903	.960	.916	.963	.921
	$P < R$	.975	.950	.980	.958	.98	.96
	s.d	-	.07	.010	.063	.02	.06
	log length	.98	.91	1.30	1.07	1.38	.15
10	2 sided coverage	.95	.929	.961	.929	.960	.931
	$P < R$	.975	.960	.981	.965	.98	.96
	s.d	-	.05	.015	.053	.03	.06
	log length	.39	.41	.73	.51	.79	.59
15	2 sided coverage	.95	.937	.960	.939	.955	.937
	$P < R$	.975	.97	.981	.971	.98	.96
	s.d	-	.04	.024	.041	.06	.06
	log length	-.05	-.02	.32	.08	.38	.14
20	2 sided coverage	.95	.943	.958	.941	.954	.939
	$P < R$	.975	.97	.979	.971	.98	.96
	s.d	-	.03	.033	.037	.06	.06
	log length	-.34	-.29	.05	-.20	.11	-.15

Exhibit 6.5
95% Confidence intervals for location

The results from the table show that the coverage is always slightly less than the required level of .95. This can be adjusted by fine-tuning the effective degrees of freedom or adjusting the shift condition. Work is continuing on the geometric interpretation of these adjustments. The results on the lengths show that for both the contaminated normal and t_3, the length of the classical interval is about 25% longer than that of the small sample interval. Results for the correlation coefficient and the percentiles of a t_3-density (first and tenth) are reported in Tingley and Field (1990). In both instances, the small sample intervals give actual coverage very close to the desired level. In particular for the correlation coefficient, the coverage is always better and the length shorter than that of the Fisher interval.

We now turn to a discussion of the procedure and its coverage. As can be noted from the algorithm, the interval is constructed by inverting the test $H_0 : \theta(\eta_0) = \theta_0$ where η_0 is unknown. Recall that in this argument, $\hat{\eta}_{obs}$ and $\hat{\theta} = \theta(\hat{\eta}_{obs})$ are being held fixed. With η_0, it can be verified that $E(g(X, \hat{\eta}_{obs})) = \theta_0 - \hat{\theta}_{obs}$ with error $O_p(1/n^{3/2})$. The approximation to the density of g under η_0 is given by the cumulant generating function $K(t)$ defined by

$$K(t) = \log\left(\frac{1}{n}\sum_{i=1}^{n} \exp(tg(x_i, \hat{\eta}_{obs}))\right).$$

However with $K(t)$, we have that the mean of $g(X, \hat{\eta}_{obs})$ is 0 while in fact the mean should be $\theta_0 - \hat{\theta}$. This suggests that $K(t)$ is not a good estimate of the density. The centering conditon (6.12) modifies our estimate $K(t)$ to $K_{\alpha_0}(t) = K(\alpha_0+t)-K(\alpha_0)$. The calculations involving g under η_0 with $\theta(\eta_0) = \theta_0$ are now done with the approximate density determined by $K_{\alpha_0}(t)$. Consider the one-dimensional exponential family $h^*(g;\alpha) = \exp(\alpha g - K(\alpha))h(g,\eta)$ where $h(g,\eta)$ is the density of g determined by K. Then the hypothesis $H_0 : \theta(\eta_0) = \theta_0$ is equivalent (to order $0_p(1/n^{3/2})$) to the hypothesis $H_0 : \alpha = \alpha_0$ in the exponential family h^*. For such an exponential family, we know that the uniformly most powerful test of $H_0 : \alpha = \alpha_0$ versus $H_1 : \alpha > \alpha_0$ is to reject H_0 if $\bar{g}_{obs}$ exceeds a critical value. Our confidence intervals are now constructed by computing $P_{\theta_0}[\bar{g} > 0]$ using the approximation in step 4 and inverting the test. Conditional on the density determined by $K(t)$ and the assumed one-parameter exponential family $h^*(g;\alpha)$, the confidence intervals constructed above are optimal.

The final step is to obtain results on the coverage. A straightforward application of an Edgeworth expansion gives bounds on the error of the cumulants from using $K_{\alpha_0}(t)$ and we obtain

$$P_{\eta_0}[\bar{g} < 0] = \hat{P}_{\theta_0}[\bar{g} < 0] + 0_p(1/n) \quad \text{if} \quad \theta(\eta_0) = \theta_0. \tag{6.15}$$

Expression (6.15) guarantees that the constructed intervals are second order correct for one-sided coverage. See Tingley and Field (1990) for a proof of (6.15) along with further discussion. Hall (1988) gives an excellent exposition on bootstrap confidence intervals.

The importance of the algorithm is that it resolves two problems with the usefulness of the small sample approximation in some practical settings. The first is that of the computational complexity in multiparameter problems. The use of the configuration reduces the multiparameter problem to a one-dimensional problem. Secondly by using $K(t)$ and $K_\alpha(t)$) we avoid the problem of specifying an underlying density and allow the data to determine the interval directly. In this sense, the procedure is closely analogous to the bootstrap.

Finally we note that there is no inherent difficulty in handling multiple regression, logistic regression or any procedure where the estimate is defined via a solution of a system of equations. Research is currently underway in applying the technique to a variety of situations.