ASSESSING THE ACCURACY OF THE MAXIMUM LIKELIHOOD ESTIMATOR: OBSERVED VERSUS EXPECTED FISHER INFORMATION

BY

BRADLEY EFRON and DAVID HINKLEY

TECHNICAL REPORT NO. 108
MARCH 8, 1978

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT MPS74-21416

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
ASSESSING THE ACCURACY OF THE MAXIMUM LIKELIHOOD ESTIMATOR:
OBSERVED VERSUS EXPECTED FISHER INFORMATION

BY

BRADLEY EFRON and DAVID HINKLEY

TECHNICAL REPORT NO. 108
MARCH 8, 1978

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT MPS74-21416

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
ASSESSING THE ACCURACY OF THE MAXIMUM LIKELIHOOD ESTIMATOR: OBSERVED VERSUS EXPECTED FISHER INFORMATION

Bradley Efron and David Hinkley

Abstract

This paper, which replaces Technical Report No. 100, concerns approximations to the variance of the maximum likelihood estimator in one parameter families. The traditional approximation is \( 1/J_\hat{\theta} \), where \( \hat{\theta} \) is the maximum likelihood estimator and \( J_\theta \) is the Fisher information. Many writers, including Fisher himself, have argued in favor of the variance estimate \( 1/I(\hat{\theta}) \), where \( I(\hat{\theta}) \) is the "observed Fisher information", i.e. minus the second derivative of the log likelihood function, evaluated at \( \hat{\theta} \). We give a frequentist justification for preferring \( 1/I(\hat{\theta}) \) to \( 1/J_\hat{\theta} \). The former is shown to approximate the conditional variance of \( \hat{\theta} \) given an appropriate ancillary statistic (which to a first approximation is \( I(\hat{\theta}) \) itself). A large number of examples are used to supplement a small amount of theory. We also consider conditional versus unconditional confidence intervals for the maximum likelihood estimator.

Key words and phrases: Likelihood, Ancillary, Fisher Information, Conditional Inference, Location Parameter, Cauchy Distribution, Curved Exponential Family
ASSESSING THE ACCURACY OF THE MAXIMUM LIKELIHOOD ESTIMATOR: OBSERVED VERSUS EXPECTED FISHER INFORMATION

Bradley Efron and David Hinkley

1. Introduction. In 1934, Sir Ronald Fisher's work on likelihood reached maturity. He had earlier advocated the maximum likelihood estimator as a statistic with least large-sample information loss, and had computed the approximate loss. Now, in 1934, Fisher showed that in certain special cases, namely the location and scale models, all of the information in the sample was recoverable by using an appropriately conditioned sampling distribution for the maximum likelihood estimator. This marks the beginning of exact conditional inference based on exact ancillary statistics, although the notion of ancillary statistics had appeared in Fisher's 1925 paper on statistical estimation.

Beyond the explicit details of exact conditional distributions for special cases, the 1934 paper contains the following intriguing claim about the general case:

"When these [log likelihood] functions are differentiable successive portions of the [information] loss may be recovered by using as ancillary statistics, in addition to the maximum likelihood estimate, the second and higher differential coefficients at the maximum" (Fisher, 1934, p. 300).

To this may be coupled an earlier statement

"The function of the ancillary statistic is analogous to providing a true, in place of an approximate, weight for the value of the estimate" (Fisher, 1925, p. 724).

There are no direct calculations by Fisher to clarify the above remarks, other than calculations of information loss. But one may infer that approximate conditional inference based on the maximum likelihood
estimate is claimed to be possible using observed properties of the like-
lihood function. To be specific, if we take for granted that inference
is accomplished by attaching a standard error to the maximum likelihood
estimate, then Fisher's remarks suggest that we use a conditional variance
approximation, based on the observed second derivative of the log like-
lihood function, as opposed to the usual unconditional variance approxi-
mation "one over the Fisher information". Our main topics in this paper are the
appropriateness and easy calculation of such a conditional variance approxi-
mation in the single parameter case. We begin with a simple illustrative
example borrowed from Cox (1958).

An experiment is conducted to measure a constant $\theta$. Independent
unbiased measurements $y$ of $\theta$ can be made with either of two instru-
ments, both of which measure with normal error: instrument $k$ produces
independent errors with a $\mathcal{N}(0, \sigma_k^2)$ distribution $k=0, 1$, where $\sigma_0^2$ and $\sigma_1^2$ are known and unequal. When a measurement $y$ is obtained a
record is also kept of the instrument used, so that after a series of
$n$ measurements the experimental results are of the form $(a_1, y_1),
(a_2, y_2), \ldots, (a_n, y_n)$, where

$$a_j = k \text{ if } y_j \text{ is obtained using instrument } k \text{ (k=0, 1)}.\$$

The choice between instruments for the $j^{th}$ measurement is made at random
by the toss of a fair coin,

$$\text{Prob}(a_j=0) = \text{Prob}(a_j=1) = \frac{1}{2}.$$

Throughout this paper, $x$ will denote the entire set of experimental
results available to the statistician in this case $(a_1, y_1), \ldots, (a_n, y_n)$. 
The log likelihood function \( \ell_\theta(x) \), \( \ell_\theta \) for short, is the (natural) log of the density function, thought of as a function of \( \theta \). In this case

\[
\ell_\theta(x) = \text{constant} - \frac{n}{j=1} \log \sigma_{a_j} - \frac{1}{2} \sum_{j=1}^{n} \frac{(y_j - \theta)^2}{\sigma_{a_j}}^2 . \tag{1.1}
\]

From this we obtain the maximum likelihood estimator as the weighted mean

\[
\hat{\theta} = \left( \frac{\sum_{j} y_j / \sigma_{a_j}^2}{\sum_{j} \sigma_{a_j}^2} \right) .
\]

If we denote first and second derivatives of \( \ell_\theta(x) \) with respect to \( \theta \) by \( \ddot{\ell}_\theta(x) \) and \( \dddot{\ell}_\theta(x) \), \( \ddot{\ell}_\theta \) and \( \dddot{\ell}_\theta \) for short, then the total Fisher information for this experiment is

\[
J_\theta = \text{Var}\{\dddot{\ell}_\theta(x)\} = \text{E}\{-\dddot{\ell}_\theta(x)\} = \frac{n}{2} \left( \frac{1}{\sigma_0^2} + \frac{1}{\sigma_1^2} \right) .
\]

Standard theory shows that \( \hat{\theta} \) is asymptotically normally distributed with mean \( \theta \) and variance

\[
\text{Var}\{\hat{\theta}\} = \frac{1}{J_\theta} . \tag{1.2}
\]

In this particular example \( J_\theta \) does not depend on \( \theta \), so that the variance approximation (1.2) is known. If this were not so we would use one of the two approximations (Cox and Hinkley, 1974, p. 302)

\[
\frac{1}{J_\theta} \quad \text{or} \quad \frac{1}{I(x)} , \tag{1.3}
\]

where

\[
I(x) \equiv -\frac{\partial^2 \ell_\theta(x)}{\partial \theta^2} \bigg|_{\theta=\hat{\theta}(x)} .
\]
The quantity $I(n)$ is aptly called the "observed Fisher information" by some writers, as distinguished from $\mathcal{J}_0$, the "expected Fisher information". (This last name is useful even though $E\mathcal{J}_0 \neq \mathcal{J}_0$ in general.) In the example above

$$I(n) = \frac{a}{\sigma_2^2} + \frac{n-a}{\sigma_1^2}$$

where $a = \Sigma a_j$, the number of times instrument 1 was used.

Approximation (1.2), one over the expected Fisher information, would presumably never be applied in practice, because after the experiment is carried out it is known that instrument 1 was used a times and that instrument 0 was used n-a times. With ancillary statistic a fixed at its observed value, $\hat{\theta}$ is normally distributed with mean $\theta$ and variance

$$\text{Var}(\hat{\theta}|a) = \left(\frac{a}{\sigma_2^2} + \frac{n-a}{\sigma_1^2}\right)^{-1}, \quad (1.4)$$

not (1.2). But now notice that, whereas (1.2) involves an average property of the likelihood, the conditional variance (1.4) is a corresponding property of the observed likelihood: (1.4) is equal to one over the observed Fisher information $I(n)$.

It is quite clear in this case that the conditional variance $\text{Var}(\hat{\theta}|a)$ is more meaningful than $\text{Var}(\hat{\theta})$ in assessing the accuracy of $\hat{\theta}$ as an estimator of $\theta$. To take an extreme situation, suppose that $n = 10$, $\sigma_0^2 = 1$, $\sigma_1^2 = .01$, and $a = 0$. Then $\frac{1}{\mathcal{J}_0} = 0.002$, but $\frac{1}{I(n)} = 0.1$, fifty times greater. The statistician was unlucky to draw the bad measuring instrument all ten times, but since he knows he was unlucky he is obliged to use the relevant variance estimate for $\hat{\theta}$.
Cox's example is misleadingly neat in that \( \text{Var}(\hat{\theta}|a) \) exactly equals \( 1/I(\tilde{\theta}) \). Nevertheless, a version of this relationship applies, as an approximation, to general one parameter estimation problems. The central topic of this paper is the accuracy of the approximation

\[
\text{Var}(\hat{\theta}|a) \approx \frac{1}{I(\tilde{\theta})}, \tag{1.5}
\]

where \( a \) is an ancillary or approximately ancillary statistic which affects the precision of \( \hat{\theta} \) as an estimator of \( \theta \). (To a first approximation, \( a \) will be equivalent to \( I(\tilde{\theta}) \) itself. It is exactly so in Cox's example.) This approximation was suggested, never too explicitly, by Fisher in his fundamental papers on ancillarity and estimation. In complicated situations such as that considered in Cox (1958) it is a good deal easier to compute \( I(\tilde{\theta}) \) than \( I(\hat{\theta}) \). There are also philosophical advantages to (1.5). It is "closer to the data" than \( 1/I(\hat{\theta}) \), and tends to agree more closely with Bayesian and fiducial analyses. In Cox's example of the two measuring instruments, for instance, an improper uniform prior for \( \theta \) on \((-\omega, \omega)\) gives \( \text{Var}(\theta|x) = 1/I(\tilde{\theta}) \), in agreement with (1.5).

To demonstrate that (1.5) has validity in more realistic contexts, consider the estimation of the center \( \theta \) of a standard Cauchy translation family. For random samples of size \( n \) the Fisher information is \( J_{\theta} = \frac{1}{2} n \). When \( n = 20 \), then \( \hat{\theta} \) has approximate variance 0.1, in accordance with (1.2). (The exact variance is about 0.115 according to Efron (1975, p. 1210).) In a Monte Carlo experiment 14,048 Cauchy samples of size 20, with \( \theta = 0 \), were obtained, and Figure 1.1 plots the resulting estimated conditional variances of \( \hat{\theta} \) given \( I(\tilde{\theta}) \) versus
1/I(\bar{x}). Samples were grouped according to values of 1/I(\bar{x}). For example, 224 of the 14,048 samples had 1/I(\bar{x}) in the range .170 - .180, averaging .175, and the 224 values of \hat{\theta}^2 had mean .201 and standard error .023. This gives the estimate \text{Var}(\hat{\theta}|1/I(\bar{x}) = .175) = .201 + .023 plotted in Figure 1.1 (since we know E(\hat{\theta}|I(\bar{x})) = 0 by symmetry).

Figure 1.1 strongly suggests the relationship

\text{Var}(\hat{\theta}|I(\bar{x})) \approx 1/I(\bar{x}). \quad (1.6)

This is a weakened version of (1.5). In translation families I(\bar{x}) is ancillary, but it is only a function of the maximal ancillary \bar{a}, the configuration statistic, i.e. the n-1 spacings between the ordered values x(1) < x(2) < ... < x(n). The stronger statement (1.5) is verified for translation families in Sections 2 and 8. We prefer (1.5) to (1.6) because \text{Var}(\hat{\theta}|\bar{a}) is more relevant than \text{Var}(\hat{\theta}|I(\bar{x})) as a measure of precision for \hat{\theta}.

The implications of (1.5) are considerable. If I(\bar{x}) = 15 in the Cauchy example, a not very remarkable event since \text{Prob}(I(\bar{x}) > 15) = 0.05, then the approximate 95% confidence interval for \theta is

\hat{\theta} \pm 1.96/\sqrt{15} \quad (1.7)

rather than

\hat{\theta} \pm 1.96/\sqrt{10} \quad (1.8)

as suggested by (1.2). The latter interval is too wide, having conditional coverage probability of 98% rather than the normal 95%. Given the equally unremarkable event I(\bar{x}) = 6, interval (1.8) is too narrow, having conditional coverage probability of only 87%. These numerical comparisons presuppose accuracy of normal approximations, which is justified in Section 4.
Figure 1.1. 14,048 random samples of size 20 were generated from a Cauchy translation family centered at $\theta = 0$. The conditional variance of the m.l.e. $\hat{\theta}$, given the observed Fisher information $I(x)$, is plotted versus $1/I(x)$. The fact that the points lie along the 45° line is an example of the general phenomenon discussed in this paper.
The purpose of this paper is to justify (1.5) for a wide variety of one parameter problems. The justification consists of detailed numerical results for several special examples involving moderate sample sizes, in addition to the general asymptotic theory. The results are presented in the following order. Section 2 gives an outline of the theory for translation families; Section 3 contains two detailed examples of this theory. Section 4 deals with confidence interval interpretations for the results of Section 2. Section 5 outlines the more complicated theory appropriate for nontranslation problems; Section 6 follows with an example. Sections 7 and 8 present details of the asymptotic theory. Section 9 contains brief concluding remarks.

2. Translation Families.

2.1. Conditional Variance Approximations. The theory of ancillarity and conditional inference for translation families was developed by Fisher (1934). Here we will use Fisher's theory to justify (1.5), and its higher order corrections, in translation families. Section 4 contains the analogous results for approximate normal confidence limits based on (1.5). In Section 5 the general one parameter problem is reduced to approximate translation form by a transformation argument. The treatment in this section is presented in outline form, more careful calculations being reserved for Section 8.

Suppose then that \( x_1, x_2, \ldots, x_n \) are independent and identically distributed (i.i.d.) with density

\[
f_\theta(x) = f_0(x - \theta) .
\]

The data vector \( x \) can be reduced to the sufficient statistic \((\hat{\theta}, \check{a})\), where \( \hat{\theta}(x) \) is the m.l.e. and \( \check{a}(x) \) is the ancillary configuration statistic, representable as the spacings

\[
x(2) - x(1), x(3) - x(2), \ldots, x(n) - x(n-1) .
\]

8
Because $\sim$ is ancillary, its density $g(a)$ does not depend on $\theta$. The conditional density (with respect to Lebesgue measure) of $\hat{\theta}$ given $\sim$ is of the translation form

$$f_{\theta}(\hat{\theta}|\sim) = h_{a}(\hat{\theta} - \theta).$$

(2.1)

The Jacobian of the transformation from the ordered $x_i$ values to $(\hat{\theta}, a)$ is a constant not depending on $\sim$, which implies that the density of $\sim$ can be written

$$f_{\theta}(\sim) = cg(a) h_{a}(\hat{\theta} - \theta)$$

(2.2)

for some constant $c$.

The likelihood function $\text{LIKE}_{\sim}(\theta)$ is $f_{\theta}(\sim)$ thought of as a function of $\theta$, with $\sim$ fixed at its observed value. Fisher's (1934) main result relates $f_{\theta}(\hat{\theta}|\sim) = h_{a}(\hat{\theta} - \theta)$ to $\text{LIKE}_{\sim}(\theta)$: for any value of $t = \hat{\theta}(\sim) - \theta$,

$$h_{a}(t) \text{ LIKE}_{\sim}(\hat{\theta}(\sim) - t)$$

$$h_{a}(0) = \frac{\text{LIKE}_{\sim}(\hat{\theta}(\sim))}{\text{LIKE}_{\sim}(\hat{\theta}(\sim))}.$$ 

(2.3)

This result, which is derived immediately from (2.2), looks simple but is in fact a powerful computational tool. Given the data vector $x$, it is computationally easy to plot the shape of the likelihood function $\text{LIKE}_{\sim}(\theta)$. Reflecting this curve about its maximum point $\hat{\theta}(x)$ then gives the conditional sampling density $f_{\theta}(\hat{\theta}|\sim)$, which might otherwise be thought difficult to compute. (The word "shape" is necessary here since (2.3) determines $h_{a}(t)$ only relative to its maximum $h_{a}(0)$. Integration is necessary to determine the correct multiple, that which satisfies $\int_{-\infty}^{\infty} h_{a}(t)dt = 1$.) Fisher's tour de force was completed by
noting that fully informative frequentist inferences about \( \theta \) should certainly be made conditional on the ancillary \( a \), so that the likelihood theory leads easily and naturally to the appropriate frequentist theory.

To see how (2.3) applies to the phenomenon pictured in Figure 1.1 suppose, for the moment, that \( \text{LIKE}_{x}(\theta) \) happens to be perfectly normal-shaped; that is,

\[
\frac{\text{LIKE}_{x}(\theta)}{\text{LIKE}_{x}(\hat{\theta}(x))} = e^{-\frac{c_{2}}{2}(\theta - \hat{\theta}(x))^2}
\]

(2.4)

for some positive constant \( c_{2} \). Letting \( \theta - \hat{\theta}(x) \equiv t \) in (2.4), and using (2.3), gives

\[
\frac{h_{a}(t)}{h_{a}(0)} = e^{-\frac{c_{2}t^2}{2}}
\]

which implies

\[
f_{\hat{\theta}|a}(\hat{\theta}) = h_{a}(\hat{\theta} - \theta) = \frac{1}{\sqrt{2\pi/c_{2}}} e^{-\frac{c_{2}}{2}(\hat{\theta} - \theta)^2}
\]

(2.5)

(since in this case we know that choosing \( h_{a}(0) = (2\pi/c_{2})^{-1/2} \) results in \( \int_{-\infty}^{\infty} h_{a}(t)dt = 1. \)) In other words, a normal-shaped likelihood function implies that, conditional on \( a, \hat{\theta} \) is normally distributed with mean \( \theta \) and variance

\[
\text{Var}(\hat{\theta}|a) = \frac{1}{c_{2}}.
\]

(2.6)

In the notation of Section 1, where \( l_{\theta}(x) \) is the log likelihood function \( \log \text{LIKE}_{x}(\theta) \), and dots indicate differentiation with respect to \( \theta \), (2.4) gives
\[ c_2 = -\frac{\partial^2}{\partial \theta^2} \equiv I(x), \quad (2.7) \]

so (2.6) is an exact form of (1.5),

\[ \text{Var}(\hat{\theta} | \tilde{z}) = \frac{1}{I(x)}. \quad (2.8) \]

What if the likelihood function is not perfectly normal-shaped? As \( n \) gets large the likelihood will approach normality, assuming some mild regularity conditions on the form of \( f_\theta(x) \), and we can use this fact to obtain asymptotic expansions for the conditional mean and variance of \( \hat{\theta} \) given \( \tilde{z} \). These expansions involve the higher derivatives of the log likelihood function, say

\[ \bar{\lambda}_{\theta}^{(j)} \equiv \frac{\partial^j \bar{\lambda}(x)}{\partial \theta^j} \bigg|_{\theta = \hat{\theta}(\tilde{z})}, \quad j = 3, 4, \ldots, \quad (2.9) \]

all of which are zero in the normal case (2.4). (We will also use the notation \( \bar{\lambda}_{\theta}^{(2)} = \bar{\lambda}_{\theta}^{(2)} \), where convenient.) Notice that (2.2) implies

\[ \bar{\lambda}_{\theta}^{(j)} = \frac{\partial^j \log h_a(\hat{\theta} - \theta)}{\partial \theta^j} \bigg|_{\theta = \hat{\theta}} = (-1)^j \frac{\partial^j \log h_a(t)}{\partial t^j} \bigg|_{t = 0}, \quad (2.10) \]

which says that the \( \bar{\lambda}_{\theta}^{(j)} \) are functions of \( x \) only through \( \tilde{z} \), and are themselves ancillary statistics. This same statement applies to the observed Fisher information \( I(x) = -\bar{\lambda}_{\theta} \). Sometimes we will emphasize this fact by writing "I(\tilde{z})" instead of "I(\tilde{z})".

**Lemma 1.** In translation families satisfying the regularity conditions stated in Section 8,

\[ \text{Var}(\hat{\theta} | \tilde{z}) = \frac{1}{I(\tilde{z})} \left[ 1 + \frac{1}{2} \left( \bar{\lambda}_{\theta}^{(4)} \right) + \frac{(\bar{\lambda}_{\theta}^{(3)})^2}{(I(\tilde{z}))^2} \right] + o_p \left( \frac{1}{n} \right), \quad (2.11) \]
\[
E(\tilde{\theta} | a) = \theta - \left\{ \frac{1}{2} \frac{\ell(\cdot)^3}{(I(x))^2} \right\} + o_p \left( \frac{1}{n} \right), \quad (2.12)
\]

and
\[
E\{ (\tilde{\theta} - \theta)^2 | a \} = \frac{1}{I(x)} \left[ 1 + \frac{1}{2} \left( \frac{\ell(\cdot)^4}{(I(x))^2} + \frac{5}{4} \left( \frac{\ell(\cdot)^3}{(I(x))^3} \right)^2 \right) + o_p \left( \frac{1}{n} \right) \right]. \quad (2.13)
\]

The proof of Lemma 1, which is an elaboration of the argument leading from a normal-shaped likelihood to (2.8), is given in Section 8, along with the appropriate regularity conditions.

The bracketed terms in (2.11) - (2.13) are of order \( o_p(n^{-1}) \) or smaller. In particular (2.11) can be written
\[
\text{Var}(\tilde{\theta} | a) = \frac{1}{I(x)} \left[ 1 + o_p(n^{-1}) \right], \quad (2.14)
\]
which verifies (1.5). Lemma 2, in Section 2.2, provides the final justification for (1.5) being an improvement over \( 1/J_0 \). The approximate normality of the likelihood function, which is used to prove Lemma 1, also ensures that the conditional distribution of \( \tilde{\theta} \) is approximately normal, given Fisher's result (2.3). Results directly related to conditional confidence intervals for \( \theta \) are described in Section 4.

In special cases the higher order terms in (2.11) can be evaluated, giving expressions for \( \text{Var}(\tilde{\theta} | a) \) more accurate than (1.5). This is demonstrated by the two examples of Section 3 and the brief discussion of the Cauchy translation problem in Section 8.

Even though the maximal ancillary \( a \) consists of \( I(x) \) plus the higher order derivatives \( \ell(\cdot)^j \), \( j=3, 4, \ldots \), the conditional variance \( \text{Var}(\tilde{\theta} | a) \) is asymptotically equivalent to a function of just \( I(x) \),
namely \( 1/I(x) \). To put it another way, \( I(x) = -\lambda \) recaptures most of the information lost by considering only \( \theta(x) \) instead of the full sample \( x \). Roughly speaking, the pair \( (\hat{\theta}, \tilde{\lambda}) \) is the sufficient statistic for the two parameter exponential family which best approximates the family \( f_\theta(x) \) near the true value of \( \theta \), see Section 8 of Efron (1975).

2.2. Statistical Curvature and Comparison of Variance Approximations.

How different, numerically, is the conditional variance approximation \( 1/I(x) \) from the unconditional approximation \( 1/J_\theta \)? An asymptotic answer can be given in terms of the statistical curvature \( \gamma_\theta \), as defined by Efron (1975). This is a quantity which is zero for one parameter exponential families (in which \( I(x) \) always equals \( J_\theta \), so that the two variance approximations are identical), and positive for nonexponential families, such as the Cauchy translation family. Suppose each \( x_i \) has density function \( f_\theta(x) \), not necessarily of translation form. The definition of \( \gamma_\theta \) is given in terms of the moments

\[
\gamma_{jk}^{(\theta)} \equiv E \left\{ \left[ \frac{\partial \log f_\theta(x)}{\partial \theta} \right]^j \left[ \frac{\partial^2 \log f_\theta(x)}{\partial \theta^2} + E \left( \frac{\partial \log f_\theta(x)}{\partial \theta} \right)^2 \right]^k \right\}, \quad (2.15)
\]

assuming these moments exist, namely

\[
\gamma_\theta \equiv [(\nu_{02}^{(\theta)} \nu_{20}^{(\theta)} - \nu_{11}^{(\theta)})^2/\nu_{20}^{(\theta)}]^{1/2}. \quad (2.16)
\]

(This quantity is invariant under any monotone reparameterization, say \( \theta \rightarrow \phi \), a fact we use in Section 5.)
Lemma 2. If \( x_1, x_2, \ldots, x_n \) are i.i.d. with density function \( f_\theta(x) \) satisfying the regularity conditions stated in Section 8, then as \( n \to \infty \),

\[
\sqrt{n} \left[ I(x) / J_\theta - 1 \right] \to N(0, \gamma^2_\theta) \quad (2.17)
\]

(proof given in Section 8).

In a translation family, \( J_\theta = J \) and \( \gamma_\theta = \gamma \) are constants, so that (2.17) can be written as

\[
\frac{I(x)}{J} = 1 + O \left( \frac{1}{\sqrt{n}} \right) \quad (2.18)
\]

Combined with (2.14), this easily leads to

\[
\frac{\text{Var}(\hat{\theta} | a)}{\text{Var}(\hat{\theta} | a) - \frac{1}{J}} = O \left( \frac{1}{\sqrt{n}} \right),
\]

which shows that (1.5) is a valid and useful asymptotic approximation.

Granting that \( \text{Var}(\hat{\theta} | a) \) is a more meaningful measure of variance than \( \text{Var}(\hat{\theta}) \), \( 1/I(x) \) is a better variance approximation than \( 1/J \) by a half order of magnitude, in the usual exaggerated sense of asymptotic comparisons. The numerical results for the Cauchy translation problem and the two examples of Section 3 show that the improvement can be substantial even for moderate sample sizes.

Suppose we are interested in estimating a monotone function of \( \theta \), say \( \sigma = \sigma(\theta) \), rather than \( \theta \) itself. It is easy to verify that the observed Fisher information for \( \sigma \), say \( I^{(\sigma)}(x) \), is related to that for \( \theta \) by the formula

\[
I^{(\sigma)}(x) = I(x) \left( \frac{d \hat{\theta}}{d \sigma} \right)^2 \quad (2.20)
\]
The expected Fisher information transforms in the same way, \( J_\sigma^{(c)} = J_\hat{\theta} \cdot (d\hat{\theta}/d\hat{\sigma})^2 = J \cdot (d\hat{\theta}/d\hat{\sigma})^2 \). (Since the m.l.e. maps the same way as does the parameter, \( \hat{\sigma} = \sigma(\hat{\theta}) \), the notation \( d\hat{\theta}/d\hat{\sigma} \) is unambiguous, and equals \( d\theta/d\sigma |_{\sigma=\hat{\sigma}} \). A standard expansion argument proceeding from Lemmas 1 and 2 shows that (1.5) is valid for \( \hat{\sigma} \), in the sense of (2.19),

\[
\frac{\text{Var}(\hat{\sigma} | a)}{\text{Var}(\hat{\sigma} | a)} = \frac{1}{I^{(c)}(x)} = \frac{1}{J^{(c)}(\sigma)} = 0 \left( \frac{1}{\sqrt{n}} \right).
\]

(2.21)

If we wish to compare confidence intervals for \( \hat{\theta} \), conditional versus unconditional as at (1.7), (1.8), the ratio of lengths is

\[
\frac{\text{length of unconditional interval}}{\text{length of conditional interval}} = \left( \frac{I(x)}{J_\theta} \right)^{1/2}.
\]

Lemma 2 implies that \( [I(x)/J_\theta]^{1/2} \) has, asymptotically, a normal distribution with mean 1 and standard deviation \( \gamma_\theta / (2 \sqrt{n}) \). For the Cauchy translation family \( \gamma_\theta^2 = 2.5 \), so that with \( n = 20 \) the standard deviation of \( [I(x)/J_\theta]^{1/2} \) is approximately 0.28. We expect large variability in \( [I(x)/J_\theta]^{1/2} \) in this situation, which is indeed the case. Increasing \( n \) to 80 in the Cauchy problem reduces the standard deviation of \( [I(x)/J_\theta]^{1/2} \) to 0.14, so that conditioning effects become considerably less important.

3. Examples of Translation Families. We illustrate the theory of the preceding section using two particularly simple examples of symmetric translation families due to Fisher (1974).

**Example 3.1. Fisher's Normal Circle.** The first example is the "circle model" shown in Figure 3.1. The data consists of a two-dimensional normally
distributed vector $\tilde{z}$, covariance matrix the identity, whose mean vector is known to lie on a circle of given radius $\rho$ centered at the origin; that is,

$$x \sim \mathcal{N}_2(\rho \begin{pmatrix} \cos \theta \\ \sin \theta \end{pmatrix}, I).$$  \hspace{1cm} (3.1)$$

Having observed $\tilde{z}$, we wish to estimate the unknown $\theta$. (Note that given $n$ independent observations $\tilde{x}_1, \ldots, \tilde{x}_n$ on this model, with say $\rho = \rho_0$, the sufficient statistic $x = \sum_{i=1}^{\tilde{x}} / \sqrt{n}$ satisfies (3.1) with $\rho = \rho_0 \sqrt{n}$.)

![Diagram of Fisher's normal circle](image)

**Figure 3.1.** Fisher's normal circle: $\tilde{z}$ is bivariate normal with covariance $I$ and mean vector on the circle of radius $\rho$. The data vector $\tilde{x}$ is observed to have polar coordinates $(\hat{\theta}, r\rho)$, $r$ being the ancillary statistic, and $\hat{\theta}$ the m.l.e.
If the data vector $z$ has polar coordinates $(\theta, r \rho)$ then $\hat{\theta}$ is the m.l.e. of $\theta$, and $r = \|z\|/\rho$ is ancillary. The density is of the form (2.2), with $z$ replaced by $r$ (even though (3.1) doesn't look like a standard translation problem), so that we can apply the theory of Section 2. The density $g(r)$ is noncentral chi, while the conditional density $f_\theta(\hat{\theta}|r) = h_r(\hat{\theta}-\theta)$ is the "circular normal",

$$c e^r \rho^2 \cos(\hat{\theta}-\theta).$$

(3.2)

(Here we are assuming that $\hat{\theta}$ given $\theta$ ranges from $\theta-\pi$ to $\theta+\pi$, for the sake of symmetric definition. The constant $c$ equals $2\pi I_0(\rho^2 r)$ using standard Bessel function notation.)

Now we can apply lemma 1 of Section 2. From (3.2) we calculate

$$\mathcal{L}(\hat{\theta}|j) = (-1)^{j/2} \rho^2 r, \quad j=2, 4, 6, \ldots$$

(3.3)

and $\mathcal{L}(\hat{\theta}|j) = 0$ for $j$ odd. The Fisher information $\mathcal{J}_\theta$ is constant,

$$\mathcal{J}_\theta \equiv \mathcal{J} = \rho^2.$$  

(3.4)

Using $I(x) = -\frac{\mathcal{L}}{\mathcal{J}} = r \mathcal{J}$, $\mathcal{L}(\hat{\theta}|3) = 0$, $\mathcal{L}(\hat{\theta}|4) = r \mathcal{J}$, from (3.3) and (3.4), (2.11) can be written

$$\text{Var}[\hat{\theta}|r] \approx \frac{1}{r \mathcal{J}} (1 + \frac{1}{2r \mathcal{J}}).$$

(3.5)

The exact conditional variance of $\hat{\theta}$ given the ancillary statistic $r$ is calculated from (3.2) to be

$$\text{Var}[\hat{\theta}|r] = \frac{\int_{-\pi}^{\pi} t^2 \exp(r \mathcal{J} \cos(t)) dt}{\int_{-\pi}^{\pi} \exp(r \mathcal{J} \cos(t)) dt}.$$  

(3.6)
Figure 3.2 compares (3.6) with (3.5). For values of $\rho^2 = J \geq 8$ it can be shown that at least 95% of the realizations of $rJ (= I(x))$ will be greater than 4. We see that approximation (1.5), $\text{Var}(\hat{\theta}|r) \approx 1/rJ$, is quite acceptable in the range $1/rJ \leq 0.25$, and that the improved approximation (3.5) derived from lemma 1 is very accurate.

![Graph](image)

**Figure 3.2.** The exact conditional variance (curve) compared with approximation (2.11) (•) for the circle and hyperbola models. If the Fisher information $J$ exceeds 8, then at least 95% of cases have $1/rJ < .025$. In this range approximation (2.11) is excellent.
Example 3.2. Fisher's Gamma Hyperbola. Fisher's hyperbola model, introduced in connection with his famous "problem of the Nile", involves two independent scaled gamma variables whose means are restricted to lie on an hyperbola. Thus we observe \( z = (x_1, x_2) \) such that

\[
x_1 = e^{\theta G_{m,1}}, \quad x_2 = e^{-\theta G_{m,2}},
\]

(3.7)

where \( G_{m,i} \) indicates a variable with density \( x^{m-1} e^{-x}/\Gamma(m) \) on \((0, \infty)\).

The Fisher information for \( \theta \) is calculated to be

\[
\mathcal{J} = 2m.
\]

(3.8)

The maximum likelihood estimate of \( \theta \) is

\[
\hat{\theta} = \frac{1}{2} \log(x_1/x_2).
\]

(3.9)

This is illustrated in Figure 3.3.

![Figure 3.3. Fisher's hyperbola model: \( z \) is a pair of independent gamma variables of index \( m \) and scale parameters constrained to lie on the solid hyperbola. The data vector \( z \) determines \( \hat{\theta} \) and the orbit hyperbola (---) for the ancillary statistic \( r \).](image-url)
The ancillary statistic in the hyperbola model is

\[ r = \sqrt{x_1 x_2} = m, \quad (3.10) \]

the level curves of which are hyperbolas "parallel" to the curve of possible mean vectors, as shown in Figure 3.4. It has density

\[ g(r) = \frac{2^m}{\Gamma^2(m)} r^{2m-1} \int_{-\infty}^{\infty} e^{-2mr \cosh(t)} dt, \]

the conditional density of \( \hat{\theta} \) given \( r \) being

\[ f_{\theta | r}(\hat{\theta}) = \frac{e^{-2mr \cosh(\hat{\theta} - \theta)}}{\int_{-\infty}^{\infty} e^{-2mr \cosh(t)} dt}. \quad (3.11) \]

In other words, this is another nonobvious example of form (2.2).

The log derivatives, from (3.11), are

\[ \lambda^{(j)}_{\theta} = -2mr = -rj, \quad j = 2, 4, 6, \ldots, \quad (3.12) \]

\[ \lambda^{(j)}_{\theta} = 0 \] for \( j \) odd, so \( I(x) = rj \) as in the circle model, and formula (2.11) gives

\[ \text{Var}(\hat{\theta} | r) \approx \frac{1}{rj} \left[ 1 - \frac{1}{2rj} \right]. \quad (3.13) \]

This differs only in the sign of the second term from the corresponding formula (3.5) for the circle model. The actual conditional variance \( \text{Var}(\hat{\theta} | r) \), obtained by integrating (3.11), can also be expressed as a function of \( rj \). The comparison of (3.13) with the actual conditional variance (Figure 3.2) is almost exactly the same as for the circle model, except that here the deviations from the line \( \text{Var}(\hat{\theta} | r) = 1/I(x) = 1/rj \) go in the opposite direction.
4. Conditional Confidence Intervals for the Location Parameter. Our results so far have been presented mainly in terms of variances, it being understood that these are of most interest in conjunction with a normal approximation for \( \hat{\theta} - \theta \). The expansion theory of Section 2 can be directly expressed in terms of conditional confidence intervals, an idea we now pursue explicitly.

As before, consider first the situation where \( \text{LIKE}_X(\theta) \) happens to be perfectly normal-shaped, so that (2.14) holds with \( c_2 = I(\hat{\theta}) = I(\theta) \). There are two consequences of this relating to standard confidence interval methods. First, \( \hat{\theta} \) has an exact normal distribution conditional on \( \hat{\tau} \) so that

\[
u(\hat{\tau}) = I(\hat{\tau}) (\hat{\theta} - \theta)^2
\]

is exactly a \( \chi^2_1 \) variable conditional on \( \hat{\tau} \). If the upper \( \alpha \) point of \( \chi^2_1 \) is denoted \( \chi^2_1(\alpha) \), then level \( \alpha \) conditional limits on \( \theta \) are

\[
\hat{\theta} \pm \left\{ \chi^2_1(\alpha)/I(\hat{\tau}) \right\}^{1/2}.
\]

The other standard method of setting confidence limits is based on

\[
v(\hat{\tau}) = 2(\hat{\theta}(\hat{x}) - \theta(\hat{x}))
\]

which also has an exact \( \chi^2_1 \) distribution conditional on \( \hat{\tau} \).

Although in general the likelihood function is not exactly normal shaped, it is approximately so for large \( n \), and the same expansion methods used to confirm (1.5) also show that \( u(\hat{\tau}) \) and \( v(\hat{\tau}) \) defined above are asymptotically \( \chi^2_1 \) conditional on \( \hat{\tau} \). More formally, we have the following result, proved in Section 8.
Lemma 3. For translation families satisfying the regularity conditions in Section 8, the statistics \( u(\tilde{x}) \) and \( v(\tilde{x}) \) defined by (4.1) and (4.2) satisfy

\[
\text{prob}\{u(\tilde{x}) \geq u_0 | A\} = (1 + d_1 - d_2) \, \text{pr}(\chi_1^2 \geq u_0) - d_1 \, \text{pr}(\chi_5^2 \geq u_0) + d_2 \, \text{pr}(\chi_7^2 \geq u_0) + o_p(n^{-1}) \quad (4.3)
\]

\[
\text{prob}\{v(\tilde{x}) \geq u_0 | A\} = (1 + d_1 - d_2) \, \text{pr}(\chi_1^2 \geq u_0) + (d_2 - d_1) \, \text{pr}(\chi_3^2 \geq u_0) + o_p(n^{-1}) \quad (4.4)
\]

where

\[
d_1 = \frac{1}{8} \left( \frac{\ell(\hat{\theta}, \tilde{\theta})^{(4)}}{I(a)} \right)^2, \quad d_2 = \frac{5}{24} \left( \frac{\ell(\hat{\theta}, \tilde{\theta})^{(3)}}{I(a)} \right)^2.
\]

Because \( d_1 \) and \( d_2 \) are both \( o_p(n^{-1}) \), (4.3) and (4.4) imply

\[
I(a)\big(\hat{\theta} - \theta\big)^2 | A = \chi_1^2 + o_p(n^{-1}), \quad 2(\tilde{\theta} - \theta) = \chi_1^2 + o_p(n^{-1}). \quad (4.5)
\]

Note that the latter result is a conditional version of Wilks' famous theorem, and establishes that a standard method has the correct conditional properties.

The results (4.5) are superior to the unconditional result

\[
\text{J}_\theta(\hat{\theta} - \theta)^2 = \chi_1^2 + o_p(n^{-1}) \quad (4.6)
\]

in a sense similar to (2.19). As we pointed out in Section 2.2, the degree of superiority is determined by the curvature.

To investigate the practical validity of (4.5), we return to the Cauchy translation problem discussed in Section 1. We generated 10,000 samples of size \( n = 20 \) and computed the empirical frequencies with which
Figure 4.1. Empirical error rates of 95% confidence intervals for Cauchy location based on 24,000 samples of size n=20

- $2(l_0 - l_0)$
- $I(a)(\hat{\theta} - \theta)^2$
- $J(\hat{\theta} - \theta)^2$

2 standard errors of estimation
\[ \phi(\hat{\theta} - \theta)^2, \quad I(\theta)(\hat{\theta} - \theta)^2, \quad 2(\lambda - \lambda_0) \]
exceeded \( \chi^2(p) \) for \( p = 0.1, 0.05, \) and 0.01, broken down by interval values of \( I(\theta) \). Figure 4.1 graphs the results for \( p = .05 \), which show convincing evidence in support of (4.5) and dramatic conditional effects on the unconditional statistic (4.6).

5. Nontranslation Families. This section discusses an example of a nontranslation problem in which a version of (1.5) can be seen to hold. We will use this example to introduce definitions appropriate for general nontranslation problems. The example is totally artificial, being in fact a simple variant of Fisher's circle model, but furnishes a useful starting point because of its simplicity. Two nontranslation problems of a more realistic nature are discussed in Section 6, again showing (1.5) at work.

We have not been able to provide a general theoretical justification for these results, and pathological counterexamples are easy to construct, but nevertheless the examples suggest that (1.5), suitably interpreted, has wide validity.

Figure 5.1 illustrates a model in which the data vector is bivariate normal, covariance matrix the identity, and with mean vector constrained to lie on a spiral (instead of Fisher's circle),

\[ x \sim N_2(\beta_0, I), \quad \beta_0 = \begin{pmatrix} \cos \theta + \rho_0 \sin \theta \\ -\sin \theta + \rho_0 \cos \theta \end{pmatrix}, \quad (5.1) \]

\[ \rho_0 \equiv \rho_0 + \theta, \quad \theta \geq -\rho_0. \quad (5.2) \]

The spiral is generated by the end of a thread unwinding from a circular spool of radius 1. By definition the thread has length \( \rho_0 \) at \( \theta = 0 \), which implies that it has length \( \rho_\theta = \rho_0 + \theta \) at \( \theta \). At \( \theta = -\rho_0 \), we have
\( \rho_\theta = 0 \), which accounts for the restriction in (5.2). We wish to assign some measure of accuracy to the m.l.e. \( \hat{\theta} \), on the basis of the observed \( \bar{x} \). (The sample size is \( n=1 \), but under repeated sampling essentially the same model holds with \( \bar{x} \) replaced by \( \bar{x} = \Sigma \bar{x}_i / n \).) It is easy to calculate that the Fisher information and curvature for the spiral model are

\[
J_\theta = \rho_\theta^2, \quad \gamma_\theta = 1 / \rho_\theta,
\]

(5.3)

and that having observed \( x \), the m.l.e. \( \hat{\theta}(x) \) is the angular coordinate of the thread upon which \( \bar{x} \) lies. The vector \( \beta_\theta \) is the closest point to \( \bar{x} \) on the spiral of possible mean vectors. When \( \rho_\theta \) is large the curvature is small, and we expect small conditioning effects, the reverse being true when \( \rho_\theta \) is small.

The geometry of Figure 5.1, and familiarity with the bivariate normal distribution, suggests that

\[
Q(x) = \text{signed distance of } \bar{x} \text{ from } \beta(\hat{\theta})
\]

(5.4)

should be approximately ancillary, with a limiting \( \mathcal{N}(0,1) \) distribution as \( \rho_\theta \) gets large. We will take the sign of \( Q(x) \) positive if \( \bar{x} \) is closer to the spool than \( \beta_\theta \), and negative if it is farther from the spool. Figure 5.1 shows that the level curve \( Q(x) = q \) is a parallel spiral having thread everywhere \( q \) units shorter than that for the mean vector.

We intend to use \( Q(x) \) as an approximate ancillary, conditioning upon the observed value of \( Q \) as we did upon \( a \) in the translation case.*

*Both Pierce and Cox suggest this use of \( Q \) in the discussion following Efron (1975).
Table 5.1 displays the marginal density of $Q(x)$ for four values of $\theta$ and nine values of $Q=q$. The four $\theta$ values are chosen such that $\rho_\theta = \rho_0 + \theta = \sqrt{8}, \sqrt{16}, \sqrt{32}, \sqrt{64}$; it is irrelevant which combinations of $\rho_0$ and $\theta$ are used to get these values of $\rho_\theta$. The density would be constant across rows if $Q(x)$ were a genuine ancillary. We see that it is nearly constant, tending toward the $N(0,1)$ density as $\rho_\theta \to \infty$.

![Figure 5.1](image)

Figure 5.1. The spiral model. The normal data vector $x$ has mean vector $\beta_0$ lying on the logarithmic spiral unwinding from a circular spool of radius 1. The asymptotic ancillary statistic $Q(x)$ is constant along a parallel spiral $Q=q$ whose "thread" is always $q$ units shorter than that of the corresponding mean vector. The m.l.e. $\hat{\theta}$ is the angular coordinate of the thread upon which $x$ lies.
The marginal densities in Table 5.1 were obtained by numerical integration of the bivariate normal density along the spiral \( Q(\tau) = q \). To avoid certain definitional problems, for each \( \rho_\theta \) the integration was restricted to points on this spiral with angular coordinate in the interval \( \theta + \pi \). Notice that if \( \rho_\theta - q < \pi \), the spiral runs into the central spool before the lower limit \( \theta - \pi \) is reached. This "end effect" seriously distorts a few of the more extreme calculations, as indicated in the tables.

<table>
<thead>
<tr>
<th>( Q=q )</th>
<th>( \rho_\theta = \sqrt{8} )</th>
<th>( \rho_\theta = \sqrt{16} )</th>
<th>( \rho_\theta = \sqrt{32} )</th>
<th>( \rho_\theta = \sqrt{64} )</th>
<th>( N(0,1) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2</td>
<td>.07</td>
<td>.07</td>
<td>.06</td>
<td>.06</td>
<td>.05</td>
</tr>
<tr>
<td>-1.5</td>
<td>.16</td>
<td>.15</td>
<td>.15</td>
<td>.14</td>
<td>.13</td>
</tr>
<tr>
<td>-1</td>
<td>.29</td>
<td>.27</td>
<td>.26</td>
<td>.26</td>
<td>.24</td>
</tr>
<tr>
<td>-0.5</td>
<td>.39</td>
<td>.38</td>
<td>.37</td>
<td>.36</td>
<td>.35</td>
</tr>
<tr>
<td>0</td>
<td>.41</td>
<td>.40</td>
<td>.40</td>
<td>.40</td>
<td>.40</td>
</tr>
<tr>
<td>1</td>
<td>.19</td>
<td>.21</td>
<td>.22</td>
<td>.23</td>
<td>.24</td>
</tr>
<tr>
<td>1.5</td>
<td>.08*</td>
<td>.10</td>
<td>.11</td>
<td>.12</td>
<td>.13</td>
</tr>
<tr>
<td>2</td>
<td>.02*</td>
<td>.04</td>
<td>.04</td>
<td>.05</td>
<td>.05</td>
</tr>
</tbody>
</table>

Table 5.1. The marginal density of the asymptotic ancillary statistic \( Q(\tau) \), for four values of \( \theta \) chosen so that \( \rho_\theta = \sqrt{8}, \sqrt{16}, \sqrt{32}, \sqrt{64} \). If \( Q(\tau) \) were exactly ancillary the density would not depend on \( \theta \), and so would be constant across the rows of the table. The last column gives the \( N(0,1) \) density, corresponding to the limiting case \( \rho_\theta \to \infty \). The asterisk indicates cases substantially distorted by the end effects described in the text.
The observed Fisher information $I(\tilde{x})$ is calculated to be

$$I(\tilde{x}) = (1 - \gamma_{\tilde{\theta}} Q(\tilde{x})) J_{\tilde{\theta}} = \rho_{\tilde{\theta}} (\rho_{\tilde{\theta}} - q), \tag{5.5}$$

where $\hat{\theta} = \hat{\theta}(\tilde{x})$ and $q = Q(\tilde{x})$. We will also use the notation $I(q, \hat{\theta}) \equiv I(\tilde{x})$ to emphasize the partition of $x$ into the approximate ancillary $Q(\tilde{x}) = q$ and the m.l.e. $\hat{\theta}$. Rather than directly verifying that $\text{Var}\{\hat{\theta}|q\} \approx 1/I(q, \hat{\theta})$, which is in fact true, we will first make a "variance stabilizing" transformation of parameters, to put the problem in an appropriate translation form, where we can expect our approximation theory to work better. Fraser (1964) makes a similar effort using a different technique.

For a fixed value of $q$ consider the transformation $\theta \to \phi_q$ defined by the differential relationship

$$\frac{d\phi_q}{dq} = \sqrt{|I(q, \theta)|} = \sqrt{\rho_{\theta} (\rho_{\theta} - q)}. \tag{5.6}$$

Equation (2.20) shows that

$$I(\phi_q)(\tilde{x}) = \frac{I(\tilde{x})}{\rho_{\theta} (\rho_{\theta} - q)} = 1 \text{ for } Q(\tilde{x}) = q. \tag{5.7}$$

In terms of the new parameter $\phi_q$, the observed Fisher information is one for every $\tilde{x}$ on the level curve $Q(\tilde{x}) = q$. (Since we intend to make conditional statements given $Q(\tilde{x}) = q$, it causes no trouble to use a different transformation for each value of $q$.) Relation (1.5) becomes simply

$$\text{Var}\{\phi_q|q\} \approx 1. \tag{5.8}$$

Notice that if (5.8) holds true, then transforming back to $\hat{\theta}$, by the usual "delta" theory, gives
\[
\text{Var}\{\hat{\theta} | q\} \approx \text{Var}\{\hat{\phi} | q\} \left( \frac{d\hat{\phi}}{d\hat{\theta}_q} \right)^2 \approx \left( \frac{d\hat{\phi}}{d\hat{\phi}_q} \right)^2 = \frac{1}{I(x)},
\]

(5.9)

which is (1.5). There is one more level of approximation in (5.9) than in (5.8) which, to reiterate, is one reason for making the transformation (5.6).

Table 5.2 shows that the quantity \( \hat{\phi}_q - \phi_q \) does indeed have nearly the right mean and variance, 0 and 1 respectively, for the cases considered. The worst case is \( q=0, \rho_\theta = \sqrt{8} \), for which the variance is 1.10. The cases \( q = 1.5, 2, \rho_\theta = \sqrt{8} \) look terrible, but that is due to the end effect previously mentioned.

<table>
<thead>
<tr>
<th>Q=( q )</th>
<th>( \rho_\theta = \sqrt{8} )</th>
<th>( \rho_\theta = \sqrt{16} )</th>
<th>( \rho_\theta = \sqrt{32} )</th>
<th>( \rho_\theta = \sqrt{64} )</th>
<th>N(0, 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2</td>
<td>1.04</td>
<td>.02</td>
<td>1.02</td>
<td>.01</td>
<td>1.01</td>
</tr>
<tr>
<td>-1.5</td>
<td>1.05</td>
<td>.02</td>
<td>1.02</td>
<td>.01</td>
<td>1.01</td>
</tr>
<tr>
<td>-1</td>
<td>1.06</td>
<td>.02</td>
<td>1.03</td>
<td>.01</td>
<td>1.01</td>
</tr>
<tr>
<td>-0.5</td>
<td>1.08</td>
<td>.02</td>
<td>1.04</td>
<td>.01</td>
<td>1.02</td>
</tr>
<tr>
<td>0</td>
<td>1.10</td>
<td>.02</td>
<td>1.04</td>
<td>.01</td>
<td>1.02</td>
</tr>
<tr>
<td>0.5</td>
<td>1.08</td>
<td>.01</td>
<td>1.04</td>
<td>.01</td>
<td>1.02</td>
</tr>
<tr>
<td>1</td>
<td>* .99*</td>
<td>.05*</td>
<td>1.05</td>
<td>.00</td>
<td>1.02</td>
</tr>
<tr>
<td>1.5</td>
<td>* .79*</td>
<td>.22*</td>
<td>1.07</td>
<td>.00</td>
<td>1.02</td>
</tr>
<tr>
<td>2</td>
<td>* .58*</td>
<td>.49*</td>
<td>1.06*</td>
<td>.01*</td>
<td>1.03</td>
</tr>
</tbody>
</table>

Table 5.2. The variance (boldface) and mean of \( \hat{\phi}_q - \phi_q \), for \( \rho_\theta = \sqrt{8}, \sqrt{16}, \sqrt{32}, \sqrt{64} \). The asterisk indicates cases substantially distorted by end effects discussed in the text.
Other moments of \( \hat{\phi}_q - \phi_q \) were calculated, all of which indicated good agreement with a standard normal distribution. For example, \( \mathbb{E} |\hat{\phi}_q - \phi_q| \) equaled \( \mathbb{E} |N(0,1)| \) to within 4%, the worst case again being \( q=0 \), \( \rho_0 = \sqrt{6} \). Another advantage of variance stabilizing transformations is that they tend to improve normality. In our three examples, \( \hat{\phi}_q - \phi_q \) was more nearly normal than \( \hat{\theta} - \theta \). This suggests forming conditional (on \( Q(\tilde{x}) = q \)) confidence intervals for \( \theta \) by computing \( \hat{\phi}_q \pm z(\alpha/2) \), where \( z(\alpha/2) \) is the upper \( \alpha/2 \) point for \( N(0,1) \), and transforming back to the \( \theta \) scale. This method agrees with \( \hat{\theta} \pm z(\alpha/2)/I(\tilde{x})^{1/2} \) to first order, but can give quite different results for small sample sizes.

To summarize the results for the spiral model, \( Q(\tilde{x}) \) contains very little direct information about \( \theta \), but its observed value considerably influences the variance of \( \hat{\theta} \). For example (5.5) shows that if \( \rho_0 = \sqrt{16} \), \( I(q,\hat{\theta}) \) varies from 24 to 8 as \( q \) varies from -2 to 2, causing a three-fold change in the variance approximation \( 1/I(\tilde{x}) \) for \( \hat{\theta} \). In other words, \( Q(\tilde{x}) \) acts like an effective ancillary statistic.

We now extend the definitions of \( Q(\tilde{x}) \) and \( \phi_q \) to an arbitrary one parameter family, say \( \mathcal{F} = \{ f_\theta(x), \theta \in \Theta \}, \Theta \) an interval of \( \mathbb{R} \), satisfying the regularity conditions of Section 8.

Define

\[
Q(\tilde{x}) = \frac{1 - I(\tilde{x})/\frac{\partial^2}{\partial \theta^2}}{\gamma_{\hat{\theta}}} , \tag{5.10}
\]

(which agrees with (5.5)). Lemma 2 shows that

\[
\sqrt{n} Q(\tilde{x}) \rightarrow N(0,1) \tag{5.11}
\]
as \( n \to \infty \), assuming that \( \gamma_\theta \) is a continuous function of \( \theta \), i.e. \( Q(x) \) is asymptotically ancillary. We have already mentioned that \((\hat{\theta}, \bar{\lambda}_\theta)\) acts like the sufficient statistic for the two parameter exponential family which best approximates \( \mathcal{F} \) near any given point \( \theta \) in \( \Theta \). The statistic \( Q(x) \) is the function of \((\hat{\theta}, \bar{\lambda}_\hat{\theta})\) linear in \( \bar{\lambda}_\hat{\theta} \) for \( \hat{\theta} \) fixed, which is asymptotically ancillary. The definition of \( Q(x) \) is also motivated by the obvious geometrical considerations of Figure 5.1, generalized in Section 8.

From (5.10) we can write \( I(x) = I(q, \hat{\theta}) = (1-\gamma_\theta q)J_{\hat{\theta}} \) as before, since \( I(x) \) is a function of \( \hat{\theta} \) and the observed value \( q = Q(x) \).

The general definition\* we will use for \( \phi_q \) is

\[
\frac{d\phi}{d\theta} = \left[ \frac{|I(q, \theta)|}{n} \right]^{1/2},
\]

(5.12)

where \( n \) is the sample size of an i.i.d. sample \( x_1, x_2, \ldots, x_n \) from some member of \( \mathcal{F} \). In making this definition, \( q \) is considered fixed and \( \theta \) variable. The mapping \( \theta \mapsto \phi_q \) is monotonic over intervals of \( \Theta \) where \( I(q, \theta) \) doesn't change sign. The possible definitional difficulties at points where \( I(q, \theta) = 0 \) do not cause trouble in our examples. A discussion of the special nature of such points is given in Section 5 of Efron (1976).

By (2.20) and (5.12), the observed Fisher information for \( \phi_q \) is

\[
I_{\phi_q}(x) = \frac{I(x) n}{I(q, \hat{\theta})} = n, \quad \text{for} \quad Q(x) = q.
\]

(5.13)

\*In a translation family definition (5.12) automatically produces a linear function of the translation parameter, \( \phi_q = c_q + d_q \theta \), if the original "\( \theta \)" is any smooth monotonic function of the translation parameter.
That is, \( I_q^\sim(x) \) is constant on the level surface \( Q(x) = q \). (The choice of the constant equal \( n \) keeps \( \phi_q \) and \( \theta \) the same order of magnitude.) In the two examples of Section 6, as in the spiral example, we verify that \( Q(x) \) is close to ancillary, and that

\[
\text{Var}(\hat{\phi}_q | q) \approx 1/n
\]  

(5.14)
in accordance with (1.5).

The transformation \( \theta + \phi_q \) defined by (5.12) is mainly of theoretical and conceptual convenience. Practical evidence certainly suggests that the likelihood is often more normal on \( \phi_q \) scale. But the derivation of \( \phi_q \) is often difficult and usually requires approximation (see Section 6). Moreover, if as we believe the results of Section 4 generalize, then

\[
2(\hat{\lambda}_q - \lambda_q) \big| q = 2(\hat{\lambda}_q - \lambda_q) \big| q = \chi_1^2 + o_p(n^{-1})
\]  

(5.15)
so that confidence limits for \( \theta \) can be derived directly from \( \lambda_q(x) \).

We emphasize that (5.15) has not been proved. Confidence limits for \( \theta \) can also be determined by taking the quadratic approximation in \( (\hat{\theta} - \theta) \) to

\[
n^{1/2}(\hat{\phi}_q - \phi_q) = \int_\theta \sqrt{|I(q,t)|} dt.
\]

6. Example of a Nontranslation Family. We illustrate the theory of Section 5 for a simple nontranslation example, using Monte Carlo methods to estimate the conditional properties of maximum likelihood estimates.

**Example 6.1. Bivariate normal correlation.** Let \( X_i = (X_{1i}, X_{2i}) \), \( i=1, \ldots, n \), independent bivariate normal pairs with zero mean and covariance matrix
\[ \Sigma_{\theta} = \begin{pmatrix} 1 & \theta \\ \theta & 1 \end{pmatrix}. \]

The two-dimensional sufficient statistic is

\[ s_1 = \sum_{i=1}^{n} X_{1i} X_{2i}, \quad s_2 = \sum_{i=1}^{n} (X_{1i}^2 + X_{2i}^2), \]

and the derivative of the log likelihood function is

\[ \frac{\partial \ell_{\theta}}{\partial \theta} = \frac{n (1-\theta^2) - \theta s_2 + (1+\theta^2)s_1}{(1-\theta^2)^2}. \tag{6.1} \]

Calculations for the Fisher information and curvature (2.16) are straightforward, giving

\[ J_{\theta} = n \frac{1+\theta^2}{(1-\theta^2)^2}, \quad \gamma_{\theta}^2 = \frac{4(1-\theta^2)^2}{(1+\theta^2)^3}. \]

Some numerical evaluations of both \( J_{\theta} \) and \( \gamma_{\theta}^2 \) are given in Table 6.1. A qualitative interpretation of the curvature values is that our two-dimensional exponential family model is highly nonlinear for small \( |\theta| \), but nearly linear as \( \theta \to \pm 1 \). The effect of replacing \( J_{\theta} \) by \( I(x) \) is potentially large for small \( |\theta| \).

| \( |\theta| \) | 0     | 0.1   | 0.2   | 0.4   | 0.6   | 0.8   | 0.9   | 0.95  | 1  |
|--------------|------|------|------|------|------|------|------|------|----|
| \( J_{\theta}/n \) | 1    | 1.03 | 1.13 | 1.64 | 3.32 | 12.65| 50.14| 200  | \infty |
| \( \gamma_{\theta}^2 \) | 4.00 | 3.81 | 3.28 | 1.81 | 0.65 | 0.12 | 0.24 | 0.0055 | 0 |
| \( \phi_{\theta} \) | 0    | 0.100| 0.204| 0.435| 0.737 | 1.235| 1.727| 2.217 | \infty |

Table 6.1. Information, curvature, and parameter \( \phi_{\theta} \) for the \( N_2(0, \Sigma_{\theta}) \) model.
The variance-stabilizing transformation \( \theta \rightarrow \phi_q \) defined by (5.12) is equivalent to

\[
\phi_q = n^{-1/2} \int_0^\theta \sqrt{\frac{1}{t}} (1-q\gamma_t)^{1/2} \, dt.
\]

As in many examples, it is difficult to evaluate this transformation exactly, but a good approximation can be obtained by substituting

\[
(1-q\gamma_t)^{1/2} \approx 1 - 1/2 q\gamma_t;
\]

recall that \( q \) is \( O(n^{-1/2}) \). In the present case this substitution leads simply to

\[
\phi_q = \phi_0 - q \tan^{-1} \theta
\]

where

\[
\phi_0 = 2^{1/2} \tanh^{-1}(\alpha_\theta 2^{1/2}) - \tanh^{-1}(\alpha_\theta), \quad \alpha_\theta = \theta(1+\theta^2)^{-1/2}.
\]

The normalizing effect of the transformation \( \theta \rightarrow \phi_q \) is illustrated by plots of likelihoods and their normal approximations in Figure 7.1 for a small data set with \( n = 20, s_1 = 12, s_2 = 35 \). In each case likelihoods are graphed relative to their maxima. The observed informations \( (I(x) \) and \( n) \) are used as variance inverses in the normal approximations, which are centered on the maximum likelihood estimates.
Figure 6.1. Likelihood functions for $\theta$ and $\phi_q$ (relative to maxima) when $n=20$, $S_1 = 12$, and $S_2 = 35$. Here $\hat{\theta} = 0.7185$, $-l_{\hat{\theta}} = 122.77$, $q = 0.101$, $\hat{\phi}_q = 0.928$. ---- normal approximation with variance $1/l_{\hat{\theta}}$.

Our interest is in whether or not $Q(x)$ defined by (5.5) is approximately ancillary, and whether or not (5.14) is accurate. Notice that $\phi_o$ is the transformation which makes $J_\theta = n$, so that the superiority of $I(x)$ over $J_\theta$ as a measure of precision conditional on $Q(x) = q$ may be judged by comparing the conditional variances of $\hat{\phi}_q$ and $\hat{\phi}_o$. The preceding theory would indicate that conditional on $q$

$$\frac{\text{var}(\hat{\phi}_q | q) - \frac{1}{n}}{\text{var}(\hat{\phi}_o | q) - \frac{1}{n}} = 0(n^{-1/2}),$$

but we have not proved this.
The likelihood equation $\hat{\lambda}_0 = 0$ has three solutions, two of which may be complex; the frequency of multiple real zeroes increases with curvature and with $\theta$. We computed $\hat{\theta}$ as a solution to the likelihood equation by iterating from an efficient estimate of $\theta$ (i.e. not the sample correlation). We simulated samples for $n$ between 15 and 40, with $\theta$ ranging between 0 and 0.9. The numbers of samples were 10,000, 50,000, and 10,000 for $n = 15, 25, \text{ and } 40$ respectively. In each case results were recorded for twenty interval values of $q$ in the 99% range $-2 \leq q \leq +2$, there being approximately the same number of samples for each $q$ interval.

From the simulation results it was quickly apparent that the range of values of $\theta$ for which the approximate theory of Section 5 is accurate depends markedly on sample size. For that reason we have chosen to give a comprehensive set of illustrations here. First, Figure 6.2 contains normal plots of the empirical distributions of $Q(\bar{z})$, a separate graph for each sample size. Several $\theta$-cases are indistinguishable, but clearly as $|\theta| \to 1$ the approximate ancillarity of $Q(\bar{z})$ breaks down.

Figure 6.3 contains plots of empirical conditional variances of both $\hat{\phi}_q$ and $\hat{\phi}_o$ for six representative cases. Standard errors for the estimated variances are indicated. These graphs confirm the theory to a remarkable degree. Particularly striking are the deviations from $n^{-1}$ of the conditional variances of $\hat{\phi}_o$.

The approximation (6.2) is remarkably accurate for the conditional mean of $\hat{\phi}_q$, which implies that the conditional mean of $\hat{\phi}_o$ deviates from $\phi_o$. Figure 6.4 illustrates a typical case.
Figure 6.2. Empirical c.d.f. of approximate ancillary $Q$ in the $N_2(0, \Sigma_\theta)$ model, plotted against c.d.f. of $N(0,1)$ distribution. Note that several cases are indistinguishable. Each case is based on at least 10,000 samples.
Figure 6.3. Conditional variances of $\hat{\phi}_q$ (solid dots) for the $\mathcal{N}_2(0, \Sigma_\theta)$ model. Also plotted are the conditional variances of $\hat{\phi}_0$ (open dots), the unconditional variance stabilized parameter. Number of samples used were 10,000 ($n=15$), 50,000 ($n=25$), and 10,000 ($n=40$). The theoretical approximation $1/n$ for the variance $\hat{\phi}_q$ is indicated by the horizontal broken line.
In reviewing the above empirical results, the reader should note that even for \( n = 40 \) conditioning on \( q \) is likely to have an appreciable effect, because the coefficient of variation of \( I(x) \) is as high as \( 0.3 \) (its value at \( \theta=0 \)). Thus at \( n = 40 \) the unconditional variance approximation \( 1/\hat{\theta} \) can easily be off by a factor of two from the conditional variance approximation.

![Graph showing empirical values of conditional mean of \( \hat{\phi} \) for \( n=25, \theta=0.5 \) from 50,000 samples. Theoretical approximation (6.2).](image)

**Figure 6.4.** Empirical values of conditional mean of \( \hat{\phi} \) for \( n=25, \theta=0.5 \) from 50,000 samples. ---- theoretical approximation (6.2).
7. Curved Exponential Families. The definition of the asymptotic ancillary statistic $Q(x)$ at (5.10) is motivated by the geometry of curved exponential families. This section gives a brief description of the geometry involved. More details appear in Efron (1975, 1976).

We begin with a $k$ dimensional exponential family $\mathcal{G}$, with density functions of the form

$$g_{\alpha}(x) = e^{\alpha'x - \psi(\alpha)}, \quad \alpha \in A, \quad x \in \mathcal{X},$$

(7.1)

$\psi(\alpha)$ being a normalizing constant. The natural parameter space $A$ and the sample space $\mathcal{X}$ are both subsets of $\mathbb{R}^k$, $A$ being convex. Corresponding to each $\alpha$ is the mean vector and covariance matrix of $x$,

$$\beta \equiv \mathbb{E}_\alpha x, \quad \mathbb{V}_\alpha \equiv \text{Cov}_\alpha x.$$  

(7.2)

The mapping from $\alpha$ to $\beta$ is one to one, so $\mathcal{G}$ can just as well be indexed by $\beta$ as by $\alpha$. The space $B \equiv \{\beta(\alpha): \alpha \in A\}$ is not necessarily convex.

A curved exponential family $\mathcal{F}$ is a one parameter subset of $\mathcal{G}$, with typical density function say

$$f_{\theta}(x) = e^{\alpha^t_\theta x - \psi_\theta}, \quad \psi_\theta \equiv \psi(\alpha_\theta).$$

(7.3)

Here $\theta$ is a real parameter contained in $\Theta$, an interval of $\mathbb{R}^1$, and the mapping $\theta \mapsto \alpha_\theta$ is assumed to be continuously twice differentiable; $\mathcal{F}$ is fully described by the curve $\mathcal{F}_A \equiv \{\alpha_\theta: \theta \in \Theta\}$ through $A$, or equivalently by the curve $\mathcal{F}_B \equiv \{\beta_\theta = \beta(\alpha_\theta): \theta \in \Theta\}$ through $B$. All of our examples, except for those in Section 1, involve curved exponential families.
(The Cauchy translation problem may be thought of as a limiting case of a
curved exponential family, as remarked at the end of Section 5 of Efron (1975).)

We will use the notation \( \mathcal{L}_\theta = \mathcal{L}_{\alpha_\theta} \), and also \( \dot{\alpha}_\theta = d\alpha_\theta / d\theta \),
\[ \ddot{\alpha}_\theta = d^2\alpha_\theta / d\theta^2. \]

Suppose \( x_1, x_2, \ldots, x_n \) is an i.i.d. sample from some member \( f_\theta \)
of \( \mathcal{F} \). The average vector \( \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i / n \) is a sufficient statistic for
\( \theta \), and it is easy to verify that the derivatives of the log likelihood
function are
\[ \dot{\mathcal{L}}_\theta (\bar{x}) = n\dot{\alpha}_\theta (\bar{x} - \beta_\theta), \quad \ddot{\mathcal{L}}_\theta (\bar{x}) = n\ddot{\alpha}_\theta (\bar{x} - \beta_\theta) - \mathcal{J}_\theta, \quad (7.4) \]
where \( \mathcal{J}_\theta = n\dot{\alpha}_\theta \mathcal{L}_\theta \dot{\alpha}_\theta \) is the Fisher information.

Figure 7.1 illustrates two useful facts about solutions to the maxi-
mum likelihood equation \( \ddot{\mathcal{L}}_\theta (\bar{x}) = 0 \), both of which follow from (7.4).

i) Given \( \hat{\theta} \), the set of \( \bar{x} \) vectors for which \( \ddot{\mathcal{L}}_\theta (\bar{x}) = 0 \), i.e.
for which \( \hat{\theta} \) is a solution to the m.l.e. equations, is the \( k-1 \) dimen-
sional hyperplane through \( \beta_\theta \) orthogonal to \( \dot{\alpha}_\theta \), say
\[ \mathcal{L}_\hat{\theta} = \{ x : \dot{\alpha}_\theta (x - \beta_\hat{\theta}) = 0 \}. \quad (7.5) \]

ii) From (5.10) and (7.4),
\[ Q(\bar{x}) = \frac{n\ddot{\alpha}_\theta (x - \beta_\hat{\theta})}{\gamma_\hat{\theta} \mathcal{J}_\hat{\theta}}. \quad (7.6) \]
Therefore for a given \( \hat{\theta} \), the set of \( \bar{x} \) vectors for which \( Q(\bar{x}) = q \) is
the \( k-2 \) dimensional hyperplane contained in \( \mathcal{L}_\hat{\theta} \) and orthogonal to \( \dot{\alpha}_\theta \),
the projection of \( \dot{\alpha}_\theta \) into \( \mathcal{L}_\hat{\theta} \).
Figure 7.1. The geometry of maximum likelihood estimation in curved exponential families. The hyperplane $\mathcal{L}_{\hat{\theta}}$ consists of those vectors $\tilde{x}$ having $\hat{\theta}$ as a solution to the likelihood equations. The $k-1$ coordinates necessary to specify the location of $\tilde{x}$ in $\mathcal{L}_{\hat{\theta}}$ are asymptotically ancillary, under the proper definition, the first of them being $Q(\tilde{x})$.

If $\mathcal{Q}$ is two dimensional, as in Figures 3.1, 3.3, and 5.1, then for a given $\hat{\theta}$, the set $Q(\tilde{x}) = q$ is a single vector. It is shown in Section 5 of Efron (1976) that this vector $v$ is squared Mahalanobis distance $q^2$ from $\bar{x}_B$ in the inner product $Z_{\hat{\theta}}^{-1}: (v-\beta_{\hat{\theta}})' Z_{\hat{\theta}}^{-1}(v-\beta_{\hat{\theta}}) = q^2$. This generalizes the geometric description of $Q(\tilde{x})$ given in Figure (5.1) (where $Z_{\theta} = I$). A similar interpretation holds for higher dimensional families $\mathcal{Q}$. 

42
So far we have discussed two "coordinates" of \( x \) of particular interest, namely \( \hat{\theta} \) and \( q \). The remaining \( k-2 \) coordinates necessary to completely specify \( x \) are higher order ancillaries, corresponding to the \( \chi^2_j \) for the translation problem. We can replace \( x \) by the sufficient statistic \( (\hat{\theta}, \sim) \), where \( \sim \) represents the \( k-1 \) coordinates which locate \( x \) in \( \mathcal{L}_\theta \). The coordinate system for \( \sim \) rotates with \( \mathcal{L}_\theta \), so that the first coordinate of \( \sim \) always corresponds to \( Q(x) \). The second coordinate of \( \sim \) is essentially the component of \( \sim \) along that part of \( \alpha^{(3)}_\theta \) orthogonal to \( \hat{\alpha}_\theta \) and \( \sim_\theta \), orthogonal being defined with respect to the inner product \( \chi^2_\theta \). This process of definition can be continued so that each successive component of \( \sim \) is less important in describing local behavior of the likelihood function near \( \hat{\theta} \), and so that \( \sqrt{n} \sim + N_{k-1}(0, I) \) as \( n \to \infty \). In other words, \( \sim \) is asymptotically ancillary. In curved exponential families, Lemma 2 can be extended to give this stronger result.

8. Details of Theoretical Results. We describe here proofs of Lemmas 1 and 2 (stated in Section 2) and Lemma 3 (stated in Section 4), together with some incidental remarks about the Cauchy location example discussed in Sections 1 and 4.

Lemmas 1 and 3 relate to expansions for conditional expectations of the form \( E\{k(t)|\sim\} \), where the conditional density of \( t = \hat{\theta} - \theta \) is given by (2.1), (2.3). These expansions are deterministic numerical approximations of the form

\[
E\{k(t)|\sim\} = v(\sim) + r(\sim),
\]

where for a given \( s \) and any \( \epsilon > 0 \).
\[ \lim_{n \to \infty} \text{Prob}_\theta\{ |n^S r(a)| < \varepsilon \} = 0. \]

We write \( r(a) = o_p(n^{-g}) \) to express this. Most of the theory relies on standard asymptotic results for regular likelihoods, a particularly useful reference being Walker (1969). Sufficient conditions for each result are given at the end of the proof.

**Lemma 1.** Consider the conditional mean squared error of \( t = \hat{\theta} - \theta \).

By (2.1) and (2.3) we may write

\[
E(t^2 | \tilde{a}) = \frac{\int_{-\infty}^{\infty} t^2 h_a(t)dt}{\int_{-\infty}^{\infty} h_a(t)dt} = \frac{\int_{-\infty}^{\infty} t^2 \exp(\lambda_{\hat{\theta}} - \lambda_{\theta})dt}{\int_{-\infty}^{\infty} \exp(\lambda_{\hat{\theta}} - \lambda_{\theta})dt},
\]

(8.1)

where \( \lambda_{\hat{\theta}} \equiv \lambda_{\hat{\theta}(\tilde{a})} \). If both integrals here are finite, as we assume, then they may be approximated arbitrarily closely by the corresponding integrals truncated at \( t = \pm b(\tilde{a}) \) for suitably large finite \( b(\tilde{a}) \). We choose \( b(\tilde{a}) \) so that the error incurred in (8.1) is \( o_p(n^{-2}) \).

The next simplification follows from the fact that for arbitrary \( \delta > 0 \) there is a \( c_\delta > 0 \) such that

\[
\lim_{n \to \infty} \text{Prob}_\theta\{ \sup_{|t| > \delta} n^{-1}(\lambda_{\hat{\theta}} - \lambda_{\theta}) < -c_\delta \} = 1,
\]

a result essentially given by Walker (1969, Section ). This result implies that the contributions to the integrals in (8.1) from \( \delta < |t| \leq b(\tilde{a}) \) are \( o_p(e^{-\delta}) \), certainly \( o_p(n^{-k}) \) for all \( k \), for all \( \delta > 0 \). Our problem then reduces to computing the truncated integrals

\[
\text{NUM}(\delta, a) = \int_{-\delta}^{\delta} t^2 \exp(\lambda_{\hat{\theta}} - \lambda_{\theta})dt \quad (8.2)
\]

\[
\text{DEN}(\delta, a) = \int_{-\delta}^{\delta} \exp(\lambda_{\hat{\theta}} - \lambda_{\theta})dt
\]

44
for arbitrarily small $\delta$.

It will be convenient to write $c_j = (-1)^{j+1} \ell^{(j)}_{\hat{\theta}}$, $j = 1, 2, \ldots$,

where $c_2 = I(x)$ and $c_1 = 0$, assuming $\hat{\theta}$ to be a stationary point of $\ell_{\hat{\theta}}$. Then we have the Taylor expansion

$$
\ell_{\hat{\theta}-t} - \ell_{\hat{\theta}} = -1/2 c_2 t^2 - \frac{1}{6} c_3 t^3 - \frac{1}{24} c_4 t^4 (1 + \varepsilon_n), \quad (8.3)
$$

where

$$
\varepsilon_n = \frac{(\ell_{\hat{\theta}}^{(4)} - \ell_{\hat{\theta}}^{(4)})}{c_4} \quad \theta_1 \in (\hat{\theta} - t, \hat{\theta}). \quad (8.4)
$$

Under a continuity condition on $\ell_{\hat{\theta}}^{(4)}$, $\varepsilon_n$ will be $o_p(1)$. We now use (8.3) to expand the integrals in (8.2) about the leading normal density term. To do this, let

$$
z = t\sqrt{c_2}, \quad w = w(z) = \frac{1}{6} c_3 t^3 + \frac{1}{24} c_4 t^4 (1 + \varepsilon_n), \quad (8.5)
$$

and note that

$$
1 - w + \frac{1}{2} w^2 - \frac{1}{6} |w|^3 \leq e^{-w} \leq 1 - w + \frac{1}{2} + \frac{1}{6} |w|^3. \quad (8.6)
$$

Substitution of (8.5) and use of (8.6) for the first integral in (8.2) gives

$$
\int_{-\delta\sqrt{c_2}}^{\delta\sqrt{c_2}} z^2 (1 - w + \frac{1}{2} w^2 - \frac{1}{6} |w|^3) \phi(z) dz \leq \frac{c_2^{3/2} \text{NUM}(\delta, a)}{(2\pi)^{1/2}}
$$

$$
\int_{-\delta\sqrt{c_2}}^{\delta\sqrt{c_2}} z^2 (1 - w + \frac{1}{2} w^2 + \frac{1}{6} |w|^3) \phi(z) dz, \quad (8.7)
$$

where $\phi(z)$ is the $N(0,1)$ density. Next replace the integration limits $\pm \delta\sqrt{c_2}$ by $\pm \infty$, which incurs an error of $o_p(e^{-n})$ because $c_2 - o_p(n)$. 

45
Then the integrals in (8.7) simply involve the first twelve moments of the 
N(0,1) distribution. Using the fact that $c_3/c_2^2 = o_p(n^{-1})$, calculation 
of the bounds in (8.7) immediately leads to

$$\text{NUM}(\delta, a) = \left(\frac{2\pi}{3} \right)^{1/2} \frac{1}{c_2} \left\{ 1 - \frac{15}{24} \frac{c_4}{c_2^2} + \frac{105}{72} \frac{c_3^2}{c_2^2} + \frac{9}{4} o_p(n^{-2}) + o_p(n^{-1} \varepsilon_n) \right\}. \quad (8.8)$$

The corresponding evaluation of the second integral in (8.2) gives

$$\text{DEN}(\delta, a) = \left(\frac{2\pi}{3} \right)^{1/2} \frac{1}{c_2} \left\{ 1 - \frac{3}{24} \frac{c_4}{c_2^2} + \frac{5}{4} \frac{c_3^2}{c_2^2} + \frac{9}{4} o_p(n^{-2}) + o_p(n^{-1} \varepsilon_n) \right\}. \quad (8.9)$$

Finally, noting that the magnitudes of earlier truncation errors are
smaller than those in (8.8), (8.9), and that $\varepsilon_n = o_p(1)$, we substitute for
numerator and denominator in (8.1) and simplify to obtain (2.13).

The corresponding calculation for $E(t|a)$ is very similar and need
not be given here. The conditional variance (2.11) is simply

$$E(t^2|a) - (E(t|a))^2.$$

The essential conditions for these results to hold are, first, that
$E(\hat{\delta} - \delta)^2 < \infty$, so that the integrals in (8.1) exist. Secondly, the first
four derivatives of $\log f_\delta(x)$ with respect to $\delta$ exist in an open
neighborhood of the true $\theta$, and have finite expectations; the second
derivative has strictly negative expectation ($\delta > 0$). Also we require
a condition on $\lambda_\delta^{(4)}(x)$ to ensure that $\varepsilon_n$ in (8.4) is $o_p(1)$. Since
in (8.4) we have $|\hat{\theta} - \hat{\delta}| < \delta$, and $|\theta - \hat{\delta}| < \delta$ for suitably large $n$, it is
sufficient to assume that for $|\theta - \theta| < \delta$

$$|\lambda_\delta^{(4)}(x_1) - l_\delta^{(4)}(x_1)| < m_\delta(x_1, \theta), \quad \lim_{\delta \to 0} E_\theta m_\delta(x_1, \theta) = 0.$$
A similar condition was used by Walker (1969).

Under stronger regularity conditions the remainder terms in (2.11) - (2.13) can be shown to be $o_p(n^{-2})$ rather than $o_p(n^{-1})$.

**Lemma 3.** These results are proved in much the same way as Lemma 1, using Laplace transforms. Because of the very similar natures of (4.3) and (4.4), we discuss only the second of these.

Consider, then, the log likelihood ratio statistic

$$v(x) = 2(\hat{\lambda}_{{\hat{\theta}}} - \lambda_0) = 2(\hat{\lambda}_{{\hat{\theta}}} - \lambda_{{\hat{\theta}} - t}) .$$

The conditional moment generating function of $v(x)$ is, by (2.1) and (2.3),

$$E(e^{sv(x)} |_a) = \frac{\int_{-\infty}^{\infty} \exp((1-2s)(\hat{\lambda}_{\hat{\theta}} - \lambda_0)) dt}{\int_{-\infty}^{\infty} \exp(\hat{\lambda}_{\hat{\theta}} - \lambda_0) dt} .$$

(8.10)

We assume $s < 1/2$. Approximation to both integrals is accomplished just as in Lemma 1, after which (8.3) is used for $|t| < \delta$ with $\varepsilon_n$ as in (8.4). A calculation parallel to that leading from (8.2) to (8.8) gives

$$\int_{-\delta}^{\delta} \exp((1-2s)(\hat{\lambda}_{\hat{\theta}} - \lambda_0)) dt$$

(8.11)

$$= \left(\frac{2\pi}{c_2}\right)^{1/2} \frac{1}{(1-2s)^{1/2}} \left[ 1 - \left( \frac{1}{8} \frac{c_4}{c_2^2} - \frac{5}{24} \frac{c_3}{c_2^3} \right) \left( \frac{1}{1-2s} \right) + o_p(n^{-1} \varepsilon_n) \right] ,$$

with the $c_j$ as in Lemma 1. Substitution of (8.11) in the numerator and denominator $(s=0)$ of (8.10) gives
\[ E(e^{sv(x)}|\hat{\gamma}) = \frac{1}{(1-2s)^{1/2}} \left\{ 1 + \frac{1}{8} - \frac{5}{24} \frac{c_3^2}{c_2} - \left( \frac{1}{8} \frac{c_4}{c_2} - \frac{5}{24} \frac{c_3^2}{c_2} \right) \left( \frac{1}{1-2s} \right) + o_p(n^{-1}) \right\} \] 

(8.12)

for \( s < 1/2 \); \( o_p(n^{-1}) \) term is a bound on a function of \( s \) for \( s < 1/2 - \eta \), \( \eta > 0 \).

Formal inversion of (8.12) gives the result (4.4). The necessary regularity conditions are clearly the same as those given for Lemma 1, and in addition we assume that \( E_0[\exp(sv(x))] < \infty \) for \( s < 1/2 \).

The \( o_p(n^{-1}) \) terms in (4.3) and (4.4) will be \( o_p(n^{-1}) \) under stronger regularity conditions.

**Corollary to Lemmas 1 and 3.** The second-order expansion results in Lemmas 1 and 3 help to explain the deviations from first-order approximations apparent in Figures 1.1 and 4.1 for the Cauchy case. To show this informally we argue as follows. By symmetry \( E_0[\ell_0^{(3)}|I(\gamma)] = 0 \), so that

\[ \{\ell_0^{(3)}\}^2 | I(\gamma) = o_p(n) . \]

Therefore, taking expectations with respect to \( \gamma \) conditional on \( I(\gamma) \) in Lemmas 1 and 3 gives

\[ \text{Var}(\hat{\gamma}|I(\gamma)) \approx \frac{1}{I(\gamma)} \left[ 1 + \frac{1}{2} \frac{K(\gamma)}{[I(\gamma)]^2} \right] \]

(8.13)

\[ \text{Prob}(u(\gamma) \leq c | I(\gamma)) \approx \left[ 1 + \frac{1}{8} \frac{K(\gamma)}{[I(\gamma)]^2} \right] \text{Prob}(\chi_{1/2}^2 \leq c) + \frac{1}{8} \frac{K(\gamma)}{[I(\gamma)]^2} \text{Prob}(\chi_{5/2}^2 \leq c) \]

(8.14)

\[ \text{Prob}(u(\gamma) \leq c | I(\gamma)) \approx \left[ 1 + \frac{1}{8} \frac{K(\gamma)}{[I(\gamma)]^2} \right] \text{Prob}(\chi_{1/2}^2 \leq c) + \frac{1}{8} \frac{K(\gamma)}{[I(\gamma)]^2} \text{Prob}(\chi_{3/2}^2 \leq c) \]

where \( K(\gamma) = E[\ell_0^{(4)}|I(\gamma)] \).
Now suppose that \( K(x) \approx b_1 n + b_2 I(x) \), so that

\[
E[I(\theta)] \approx E[I(\theta)] = b_1 + b_2 E[I(x)] = b_1 n + b_2 \cdot \theta.
\]

For the Cauchy distribution \( E[I(\theta)] = \theta \), so that \( b_1 \approx 0 \) and \( b_2 \approx 1 \).

The implied form of (8.13) is

\[
\text{Var}(\tilde{\theta} | I) \approx \frac{1}{I} \left( 1 + \frac{1}{2I} \right),
\]

which is a very good approximation to the empirical variances in Figure 1.1, and is for the majority of cases at \( n = 10 \). The implied forms of (8.14) are also very accurate. Note that (8.14) also explains the tendency for \( \nu(x) \) to be closer to \( \chi_1^2 \) than is \( u(x) \), because \( \chi_3^2 \) is stochastically smaller than \( \chi_5^2 \).

Lemma 2. Sufficient regularity conditions are stated in the last paragraph, following the formal derivation. First notice that since \( I(x)/J_\theta \) is invariant under monotone reparameterizations, by (2.20), we can change to the parameter \( \sigma \) defined by \( d\sigma/d\theta = (J_\theta/n)^{1/2} \), for which \( J_\sigma(\sigma) = n \). We might as well assume this parameterization to begin with, so \( J_\theta = n \) for all \( \theta \). This implies \( \nu_{20}(\theta) = 1 \) since by definition (2.15), \( \nu_{20}(\theta) \) is the Fisher information in a single observation \( x \). For notational convenience, let \( \theta = 0 \) be the true value of the parameter. Then we wish to show that

\[
\sqrt{n} \left( \frac{I(x)}{n} - 1 \right) = \frac{I(x) - n}{\sqrt{n}} \rightarrow N(0, \gamma_0^2)
\]

under i.i.d. sampling from \( f_0(x) \).
Let \( S(\hat{\gamma}) = \frac{-\ddot{\gamma}_0(\gamma) - \nu_{11}(0)\dot{\gamma}_0(\gamma)}{\sqrt{n}} \). Because \( \nu_{11}(0) \) is the regression coefficient of \( \ddot{\gamma}_0 \) on \( \dot{\gamma}_0 \), by definition (2.15), and the fact that \( \nu_{20}(0) = 1 \), it is easy to see from definition (2.16) that \( S(\gamma) \sim N(0, \gamma_0^2) \). The proof is completed by showing that \( S(\gamma) \) is asymptotically equivalent to \( \frac{I(\gamma)-\nu}{\sqrt{n}} \).

Notice that by differentiating \( \nu_{20}(\gamma) = 1 = E_\theta \{ -\ddot{\gamma}_\theta(x) \} \) with respect to \( \theta \) we get
\[
E_\theta \{ -\ddot{\gamma}_\theta^{(3)}(x) \} = \nu_{11}(\theta).
\] (8.16)

The strong law of large numbers then implies that \( -\ddot{\gamma}_\theta^{(3)}(x)/n = \nu_{11}(\theta)[1 + \varepsilon_n] \) where \( \varepsilon_n \xrightarrow{a.e.} 0 \). (In what follows, \( \varepsilon_n \) will stand for any sequence of random variables converging to zero almost everywhere.) The standard proof for the asymptotic normality of the m.l.e., as in Section 5f of Rao (1973) shows that \( \hat{\theta} = (\ddot{\gamma}_0(\gamma)/n)[1 + \varepsilon_n] \). These results imply that the expansion \( \ddot{\gamma}_\theta(x) = \ddot{\gamma}_0(x) + \hat{\theta} \dot{\gamma}_0^{(3)}(\gamma) + (\hat{\theta}^2/2) \dot{\gamma}_0^{(4)}(\gamma) \), \( \hat{\theta} \in (0, \hat{\theta}) \), can be rewritten as
\[
-\ddot{\gamma}_\theta^{(3)}(x)/\sqrt{n} = S(\gamma) - \varepsilon_n \nu_{11}(0) \dot{\gamma}_0^{(4)}(\gamma) \sqrt{n} \hat{\theta}^2 \frac{\dot{\gamma}_0^{(4)}(x)}{2n}.
\] (8.17)

Since \( \ddot{\gamma}_0(x)/\sqrt{n} \sim n(0,1) \) the term \( \varepsilon_n \nu_{11}(0) \dot{\gamma}_0^{(4)}(x)/\sqrt{n} \to 0 \). The last term in (8.17) is also negligible under a boundedness condition on \( \dot{\gamma}_0^{(4)}(\gamma) \), completing the proof of (8.15).

The regularity conditions needed in this proof are (i) the usual conditions for the asymptotic normality of the m.l.e., see Rao (1973), Section 5f; (ii) equality (8.16), or any regularity conditions justifying the differentiation under the integral sign leading to (8.16);
(iii) \[ |\xi_{(x)}(x)| < M(x) \text{ for } \theta \text{ in a neighborhood of } 0, \text{ where} \]
\[ E_0 M(x) < \infty. \] (Then \[ |\xi_{(x)}(x)/n| < \sum M(x)/n \text{ a.e. } E_0 M(x), \text{ justifying the} \]
last step in the proof.) We remark that these conditions are always satisfied in curved exponential families. The cruder result (2.18), which is all that is needed to get (2.19), requires less stringent conditions.

9. Concluding Remarks. The thrust of this paper has been to offer what we believe to be convincing evidence that there is a meaningful approximate conditional inference for single parameters based on the maximum likelihood estimate \( \hat{\theta} \) and the observed information \( I(x) \). For the most part the evidence has been of an empirical nature, although in the location case the theory has been found to flow directly from Fisher's calculations. In the nonlocation case the discussions of Sections 5 and 7 demonstrate the existence of a dominant approximate ancillary, together with a convenient framework (curved exponential families) for further work. We have not obtained a formal proof generalizing the results of Sections 2 and 4, although we do not doubt that this is possible.

Not all of the examples that we considered have been included here. We also looked at the \( N(\theta, c\theta^2) \) case (discussed by Hinkley (1977)); a two-parameter (linear model) version of Fisher's hyperbola; and the double-exponential location model, where the theory of Section 2 must fail, but does so in an intriguing manner.

We have not attempted to discuss the case of several parameters, which raises certain definitional problems. However the extension of our results to the case of the general (regular) location-scale model is straightforward, again because of duality between conditional distribution and likelihood as shown by Hinkley (1978).
REFERENCES


