Further remarks on asymptotic normality of likelihood and conditional analyses

D.A.S. FRASER and P. McDUNNOUGH

York University and University of Toronto

Key words and phrases: Likelihood, asymptotic normality, conditional analysis, confidence distribution, structural distribution.

AMS 1980 subject classifications: Primary 62E20, secondary 62A05.

ABSTRACT

Under weak conditions the normalized likelihood with or without weight function almost surely converges to a normal density function: for a real parameter or a vector parameter; with or without the assumption of independent identical distributions. Applications arise for confidence intervals, confidence distributions, structural distributions, and conditional analyses with transformation and structural models.

RÉSUMÉ

On montre, sous des hypothèses assez générales, que la fonction de vraisemblance normalisée (pondérée ou non) converge presque sûrement vers une fonction de densité normale. Tel est le cas, par exemple, pour une fonction de vraisemblance dépendant d’un paramètre réel ou vectoriel, peu importe qu’elle soit induite par des données indépendantes et identiquement distribuées ou non. Les résultats obtenus s’appliquent dans le contexte des intervalles de confiance, des distributions structurelles, ainsi que dans le cadre d’analyses conditionnelles basées sur des modèles structurels ou de transformation.

1. INTRODUCTION

Many statistical models with data summarize their inferences about an unknown \( \theta \) in the form of what can be termed a confidence distribution. Two well-known examples are the Bayesian model, which generates a posterior distribution, and the transformation or structural model, which yields conditional confidence intervals summarized by a structural distribution. Often the confidence distribution may be presented in terms of a density that is proportional to a weighted likelihood \( w(\theta)L(\theta) \), where \( w(\theta) \geq 0 \) is a weighting function satisfying

\[
\int w(\theta)L(\theta)\,d\theta < \infty,
\]

with \( L(\theta) \) representing the observed likelihood function.

The following problem was examined in Brenner, Fraser, and McDunnough (1982): given an independently and identically distributed sample \( x_1, \ldots, x_n \) from a location family, does limiting normal shape of the likelihood at the maximum imply limiting normality of the confidence distribution? Conditions were given under which it did, together with a general inversion theorem for transferring such properties to the distribution for conditional analyses.
This paper generalizes these results to the case of vector parameters and distributions not identical or independent, for any model and weight function for which (*) holds. Moreover, the arguments involved are simpler and the conditions milder.

2. INDEPENDENT IDENTICAL DISTRIBUTIONS; A REAL PARAMETER

Let \( x_1, \ldots, x_n \) be independent and identically distributed with density function \( f(x| \theta) \), where the parameter \( \theta \) takes values in \( \Omega = \mathbb{R} \); the case where \( \Omega \) is an open interval can be included by reparametrizing by means of a continuously differentiable function. The observed likelihood function is given by

\[
L_n(\theta) = f(x_1| \theta) \ldots f(x_n| \theta),
\]

and the log-likelihood function by

\[
\ell_n(\theta) = \ln L_n(\theta).
\]

There is no requirement that the \( x \)'s be real-valued; indeed, the nature of the variable \( x \) is immaterial to the argument.

Consider the following three assumptions.

**Assumption I.** For any \( \delta > 0 \),

\[
\lim_{n \to \infty} \frac{1}{n} \sup_{|s - \theta| \leq \delta} \left[ L_n(s) - L_n(\theta) \right] < 0
\]

almost surely.

This is a common assumption that ensures the existence and consistency of the MLE \( \hat{\theta} \). For independent identical distribution, Wald (1949) showed that it held under the assumption \( \mathbb{E}(\sup_s |\ln f(x| s)|) < \infty \), with \( \Omega \) compact.

**Assumption II.** \( l_i(\theta) \) is twice continuously differentiable with

\[
0 < \mathbb{E}(-l_i''(\theta)) < \infty.
\]

Again this is a common assumption which deals with the unit Fisher information \( \mathbb{E}(-l_i''(\theta)) \). If we set

\[
\sigma_n^{-2} = \mathbb{E}(-l_i''(\theta)),
\]

for independent identical distributions,

\[
\sigma_n^{-2} = n\mathbb{E}(-l_i''(\theta)) \to \infty.
\]

**Assumption III.** For each \( \varepsilon > 0 \), there exists \( \delta > 0 \) such that

\[
\lim_{n \to \infty} \sup_{|s - \theta| < \delta} \left| l_i''(s) - l_i''(\theta) \right| < \varepsilon
\]

almost surely \([\sigma_n^2 \text{ given in (1)}]\). This is a common continuity assumption on \( l_i''(\theta) \). For example, for independent identical distributions, if Assumptions I and II hold, then Assumption III is implied by the weaker condition

\[
\lim_{n \to \infty} \mathbb{E}\left| l_i'(s) - l_i'(\theta) \right| = 0.
\]

Now if we set

\[
\hat{\sigma}_n^2 = \sigma_n^2|_{\theta = \hat{\theta}},
\]
then Assumptions I, II, III imply
\[ \frac{L_n(\hat{\theta} + \hat{\sigma}_n t)}{L_n(\hat{\theta})} \to e^{-t^2/2} \]
(2)
almost surely. The convergence is uniformly for \( t \) bounded; see Theorem 4.1 in Brenner, Fraser, and McDunnough (1982). These assumptions also imply the asymptotic normality of \( \hat{\theta} \). Note too that Assumption III implies that \( \hat{\sigma}_n^2(-l''(\theta)) \to 1 \) almost surely, thus providing an alternative means of scaling in (2).

We can now present the main result of this section.

**Theorem 2.1.** For a sample \( x_1, \ldots, x_n \) from \( f(x | \theta) \), Assumptions I, II, III imply
\[ \frac{w(\hat{\theta} + \hat{\sigma}_n t) L_n(\hat{\theta} + \hat{\sigma}_n t)}{\hat{\sigma}_n \int w(s) L_n(s) \, ds} \to (2\pi)^{-1/2} e^{-t^2/2} \]
almost surely, where \( w(\theta) \geq 0 \) satisfies (\*) and is continuous and nonzero at the true \( \theta \).

**Proof:** As a consequence of (2) and the succeeding remarks it suffices to show that
\[ \frac{\hat{\sigma}_n \int w(s) L_n(s) \, ds}{w(\hat{\theta}) L_n(\hat{\theta})} \to \sqrt{2\pi} \]
almost surely and use \( w(\hat{\theta} + \hat{\sigma}_n t)/w(\hat{\theta}) \to 1 \) almost surely. For this we can rewrite
\[ \frac{\hat{\sigma}_n \int w(s) L_n(s) \, ds}{w(\hat{\theta}) L_n(\hat{\theta})} = \int \frac{w(\hat{\theta} + \hat{\sigma}_n t) L_n(\hat{\theta} + \hat{\sigma}_n t)}{w(\hat{\theta}) L_n(\hat{\theta})} \, dt. \]
The dominated-convergence theorem gives
\[ \int_A \frac{w(\hat{\theta} + \hat{\sigma}_n t) L_n(\hat{\theta} + \hat{\sigma}_n t)}{w(\hat{\theta}) L_n(\hat{\theta})} \, dt \to \int_A e^{-t^2/2} \, dt \]
almost surely for any bounded \( A \). Accordingly there remains to show only the following: for suitable \( \delta > 0 \)
\[ \lim_{\delta \to 0} \lim_{A \to R} \int_{B_{\delta}} \frac{w(\hat{\theta} + \hat{\sigma}_n t) L_n(\hat{\theta} + \hat{\sigma}_n t)}{w(\hat{\theta}) L_n(\hat{\theta})} \, dt \to 0 \]
(3)
almost surely, and
\[ \lim_{\delta \to 0} \int_{B_{\delta}} \frac{w(\hat{\theta} + \hat{\sigma}_n t) L_n(\hat{\theta} + \hat{\sigma}_n t)}{w(\hat{\theta}) L_n(\hat{\theta})} \, dt \to 0 \]
(4)
almost surely, where
\[ B_{\delta} = \{ t : |\hat{\theta} + \hat{\sigma}_n t - \theta| < \delta \} \cap A^c, \]
\[ B_\delta = (A \cup B_{\delta})^c. \]
Now (3) is implied by
\[ \lim_{\delta \to 0} \lim_{A \to R} \int_{B_{\delta}} \frac{L_n(\hat{\theta} + \hat{\sigma}_n t)}{L_n(\hat{\theta})} \, dt \to 0 \]
almost surely. For this we have
\[ \frac{L_n(\hat{\theta} + \hat{\sigma}_n t)}{L_n(\hat{\theta})} = \exp \left\{ -\frac{t^2}{2} \hat{\sigma}_n^2(-l''(\theta)) + t^2 \hat{\sigma}_n^2 r_n \right\}. \]
where by Assumption III, \( \hat{\sigma}^2 \sup_{n \in B_n} |r_n| \) can almost surely be made arbitrarily small by choice of a suitable \( \delta \); thus we have
\[
L_n(\hat{\theta} + \hat{\sigma}_n t) / L_n(\hat{\theta}) \leq e^{-ct^2} \quad (c > 0)
\]
almost surely. We then note that
\[
\lim_{A \to \mathbb{R}} \lim_{n \to \infty} \int_{B_{\delta_n}} e^{-ct^2} dt = 0
\]
amost surely, which thus establishes (3).

Towards (4) we note that
\[
\frac{L_n(\hat{\theta} + \hat{\sigma}_n t)}{L_n(\hat{\theta})} = \frac{f(x_1 \mid \hat{\theta} + \hat{\sigma}_n t)}{f(x_1 \mid \hat{\theta})} \frac{L_n(\hat{\theta} + \hat{\sigma}_n t)}{L_n(\hat{\theta})} = \frac{L(\hat{\theta} + \hat{\sigma}_n t)}{L(\hat{\theta})} \frac{L(\hat{\theta} + \hat{\sigma}_n t)}{L(\hat{\theta})},
\]
where
\[
L_{\alpha-1}(\theta) \propto f(x_2 \mid \theta) \cdots f(x_n \mid \theta).
\]
From Assumption I we have
\[
L_{\alpha-1}(\hat{\theta} + \hat{\sigma}_n t) / L_{\alpha-1}(\hat{\theta}) \leq e^{-dt}, \quad t \in B_n, \text{ some } d > 0,
\]
amost surely, so that using (*) we obtain
\[
\int_{B_n} w(\hat{\theta} + \hat{\sigma}_n t) \frac{L_n(\hat{\theta} + \hat{\sigma}_n t)}{L_n(\hat{\theta})} dt \leq \frac{e^{-dt}}{w(\hat{\theta}) L(\hat{\theta})} \int_{B_n} w(\hat{\theta} + \hat{\sigma}_n t) L(\hat{\theta} + \hat{\sigma}_n t) dt \leq \frac{e^{-dt}}{w(\hat{\theta}) L(\hat{\theta}) \hat{\sigma}_n} \int w(s) L(\hat{\theta}) ds \to 0
\]
amost surely. Q.E.D.

Notice that Theorem 2.1 implies that almost surely the (confidence-structural) distribution for \( T = (\theta - \hat{\theta}) / \hat{\sigma}_n \) converges weakly to the standard normal; this follows from a theorem due to Scheffé (1947). Furthermore, Walker’s result (Walker 1969) follows easily from the present analysis. Walker showed that the posterior distribution for \( T = (\theta - \hat{\theta}) / \hat{\sigma}_n \), derived from a proper prior, converges weakly in probability to a normal distribution. The weaker convergence-in-probability result is obtained by replacing “almost surely” in Assumptions I, II, III with “in probability” and noting the equivalence of almost sure convergence and convergence in probability for suitably extracted subsequences.

3. THE MULTIPARAMETER CASE WITH INDEPENDENT IDENTICAL DISTRIBUTIONS

The case with parameter space \( \mathbb{R}^k \) is closely analogous to that for a real parameter. Accordingly our treatment is brief.

For Assumption I we interpret \( |\theta| \) as Euclidian \( k \)-dimensional distance and refer to it as Assumption I’. We also modify Assumptions II and III:

ASSUMPTION II’. \( L_n(\theta) \) is twice continuously differentiable with
\[
0 < \det\{ -E L_n(\theta) \} < \infty.
\]

For this we let
ASYMPTOTIC NORMALITY

\[ l_n^*(\theta) = \left[ \frac{\partial^2}{\partial \theta_i \partial \theta_j} L_n(\theta) \right] \]

be the matrix of second-order derivatives. We also set

\[ \mathcal{L}_n^{-1} = \mathcal{E}(-l_n^*(\theta)) = n \mathcal{E}(-l_n^*(\theta)) \]

and \( \hat{\mathcal{L}}_n = \mathcal{L}_n \) as evaluated at \( \hat{\theta} \).

ASSUMPTION III'. For each \( \varepsilon > 0 \), there exists \( \delta > 0 \) such that

\[ \lim \det(\mathcal{L}_n) \sup_{|s|,|r| < \delta} \| l_n^*(s) - l_n^*(\theta) \| < \varepsilon \]

almost surely (note that for a matrix \( A = [a_{ij}] \), \( \| A \| = \max_{i,j} |a_{ij}| \)).

Under Assumptions I', II', III' we clearly have

\[ \mathcal{L}_n \to 0, \quad \hat{\mathcal{L}}_n \to 0 \]

and also

\[ \frac{L_n(\theta + \hat{\mathcal{L}}_n t)}{L_n(\hat{\theta})} \to \exp \left\{ -\frac{|t|^2}{2} \right\} \]

almost surely. The generalization of Theorem 2.1 is then given as

**Theorem 3.1** If \( w(\theta) \equiv 0 \) satisfies (*) and is continuous and nonzero at the true \( \theta \), then for a sample \( x_1, \ldots, x_n \) from \( f(x | \theta) \), Assumptions I', II', III' imply that

\[ \frac{w(\hat{\theta} + \hat{\mathcal{L}}_n t) L_n(\hat{\theta} + \hat{\mathcal{L}}_n t)}{\det(\hat{\mathcal{L}}_n) \int w(s) L_n(s) \ ds} \to (2\pi)^{-t/2} e^{-|t|^2/2} \]

almost surely.

The proof of Theorem 3.1 follows along the lines of that of Theorem 2.1. For example, as in Theorem 2.1, we need only show

\[ \frac{\det(\hat{\mathcal{L}}_n) \int w(s) L_n(s) \ ds}{w(\hat{\theta}) L_n(\hat{\theta})} \to (2\pi)^{t/2} \]

almost surely, and use \( w(\hat{\theta} + \hat{\mathcal{L}}_n t) / w(\hat{\theta}) \to 1 \) almost surely. As before, apply the dominated-convergence theorem for any bounded set \( A \subset \mathbb{R}^t \), and then define sets \( B_{b,n}, B_n \) by

\[ B_{b,n} = \{ t : |\hat{\theta} + \hat{\mathcal{L}}_n t - \theta | < \delta \} \cap A^\circ, \]

\[ B_n = (A \cup B_{b,n})^\prime. \]

Now use a multivariate version of Taylor’s theorem and Assumption III' to get, for each \( t \) in \( B_{b,n} \),

\[ l_n(\hat{\theta} + \hat{\mathcal{L}}_n t) - l_n(\hat{\theta}) \leq -c' |t|^2 \quad (c' > 0) \]

almost surely, from which the appropriate analogue to (3) follows. The corresponding version of (4) is also straightforward to demonstrate by using Assumption I' and the argument employed in the real-parameter case.

As in the one-dimensional case, we can use Scheffé’s theorem to show that the confidence distribution \( \hat{\mathcal{L}}_n(\theta - \hat{\theta}) \) almost surely converges weakly to the standard normal. It
is also of interest to note that any continuously differentiable function of \( \theta \), say a component parameter, will almost surely have a limiting normal "marginal confidence distribution".

4. DISTRIBUTIONS NOT INDEPENDENT OR IDENTICAL

As Heyde and Johnstone (1978) have noted, Walker’s proof of certain large-sample properties of posterior distributions (based on proper priors) is easily carried over to arbitrary discrete-time stochastic processes. The same carryover arises for the results we have been discussing. For convenience we examine the case of a real parameter with \( \Omega = \mathbb{R} \); the multiparameter case follows immediately.

For each \( n \) we suppose \( (X_1, \ldots, X_n) \) has density \( f_n(x \mid \theta) \) with respect to a \( \sigma \)-finite measure not dependent on \( \theta \); this leads to an observed likelihood \( L_n(\theta) \) and log-likelihood \( l_n(\theta) \) functions. For this we remark that a careful examination of the proofs in Walker (1969) and Heyde and Johnstone (1979) reveal that discrete time is not essential. What is needed is a concept of an ordering on the assembly of data and that large amounts of data yield accurate estimates. Similar remarks apply to the results of this section.

We now record assumptions that correspond closely to those used for the previous case.

ASSUMPTION I*. There exists \( a_n \downarrow 0 \) such that for any \( \delta > 0 \)

\[
\lim_{n \to \infty} a_n \sup_{s, |s - \theta| < \delta} [l_n(s) - l_n(\theta)] < 0
\]

almost surely.

As with Assumption I, this gives consistency of the MLE \( \hat{\theta} \) and also provides an indication of its rate of convergence.

ASSUMPTION II*. \( l_n(\theta) \) is twice continuously differentiable with

\[
\hat{\sigma}_n^{-2} = -l''_n(\hat{\theta}) \to \infty
\]

almost surely.

ASSUMPTION III*. For each \( \epsilon > 0 \), there exists \( \delta > 0 \) such that

\[
\lim_{n \to \infty} \hat{\sigma}_n^2 \sup_{s, |s - \theta| < \delta} |l''_n(s) - l''_n(\theta)| < \epsilon
\]

almost surely, where \( \hat{\sigma}^2 \) is defined in Assumption II*.

Assumptions I*, II*, III* imply that

\[
L_n(\hat{\theta} + \hat{\sigma}_n t)/L_n(\hat{\theta}) \to e^{-t^2/2}
\]

almost surely, with uniform convergence in a bounded region. By also assuming that Assumption II* holds with \( a_n = \hat{\sigma}_n^2 \), we obtain a stronger result.

**Theorem 4.1.** If \( w(\theta) \geq 0 \) satisfies (*) and is continuous and nonzero at the true \( \theta \), then Assumptions I* (with \( a_n = \hat{\sigma}_n^2 \)), II*, III* imply

\[
w(\hat{\theta} + \hat{\sigma}_n t)L_n(\hat{\theta} + \hat{\sigma}_n t) \hat{\sigma}_n \int w(s)L_n(s)ds \to (2\pi)^{-1/2}e^{-t^2/2}
\]

almost surely.

**Proof:** The proof of Theorem 2.1 carries over directly except for details for demonstrating (4). Accordingly we need only show that
\[
\lim_{n \to \infty} \int_{B_n} \frac{w(\hat{\theta} + \hat{\sigma}_n t)}{w(\hat{\theta})} \cdot \frac{L_n(\hat{\theta} + \hat{\sigma}_n t)}{L_n(\hat{\theta})} \, dt \to 0
\]

almost surely, where

\[B_n = \{ t : |\hat{\theta} + \hat{\sigma}_n t - \theta| \geq \delta \}.
\]

For this we write

\[
\frac{L_n(\hat{\theta} + \hat{\sigma}_n t)}{L_n(\hat{\theta})} = \frac{f(x_1 | \hat{\theta} + \hat{\sigma}_n t)}{f(x_1 | \hat{\theta})} \cdot \frac{L_{n-1}(\hat{\theta} + \hat{\sigma}_n t)}{L_{n-1}(\hat{\theta})} = \frac{L_i(\hat{\theta})}{L_i(\hat{\theta})} \cdot \frac{L_{n-1}(\hat{\theta} + \hat{\sigma}_n t)}{L_{n-1}(\hat{\theta})},
\]

where

\[
L_{n-1}(\theta) \propto \frac{f(x_1, \ldots, x_n | \theta)}{f(x_1 | \theta)}.
\]

For \( \delta_1 > 0 \) we have

\[
\lim_{n \to \infty} \sup_{s:|s - \theta| > \delta_1} \left[ L_{n-1}(s) - L_{n-1}(\theta) \right] = \lim_{n \to \infty} \sup_{s:|s - \theta| > \delta_1} \left[ \left( L_n(s) - L_n(\theta) \right) + L_n(\theta) \right] = \lim_{n \to \infty} \sup_{s:|s - \theta| > \delta_1} \left[ L_n(s) - L_n(\theta) \right] < 0.
\]

Thus for all \( t \in B_n \)

\[
L_{n-1}(\hat{\theta} + \hat{\sigma}_n t) / L_{n-1}(\hat{\theta}) \leq e^{-\delta_1^2 n}, \quad \text{some} \quad d > 0,
\]

almost surely, and

\[
\int_{B_n} \frac{w(\hat{\theta} + \hat{\sigma}_n t)}{w(\hat{\theta})} \cdot \frac{L_n(\hat{\theta} + \hat{\sigma}_n t)}{L_n(\hat{\theta})} \, dt \leq \frac{e^{-\delta_1^2 n}}{w(\hat{\theta})L_i(\hat{\theta}) \hat{\sigma}_n} \int w(s)L_i(\theta) \, ds \to 0
\]

almost surely. Q.E.D.

Note that in Theorem 4.1 the integration condition (*) need not hold for \( n = 1 \). Rather it suffices to have it hold for some \( n \), say \( k \), and modify the proof by writing

\[
\frac{L_n(\hat{\theta} + \hat{\sigma}_n t)}{L_n(\hat{\theta})} = \frac{L_k(\hat{\theta} + \hat{\sigma}_n t)}{L_k(\hat{\theta})} \cdot \frac{L_{n-k}(\hat{\theta} + \hat{\sigma}_n t)}{L_{n-k}(\hat{\theta})}
\]

and then proceeding as before.

REFERENCES


---

Received 23 January 1983  
Revised 12 September 1983  
Accepted 7 February 1984

Department of Mathematics  
York University  
Downsview, Ontario, M5J 1P3

Department of Statistics  
University of Toronto  
Toronto, Ontario M5S 1A1