Geometric Understanding of Likelihood Ratio Statistics

Jianqing FAN, Hui-Nien HUNG, and Wing-Hung WONG

It is well known that twice a log-likelihood ratio statistic follows asymptotically a chi-square distribution. The result is usually understood and proved via Taylor's expansions of likelihood functions and by assuming asymptotic normality of maximum likelihood estimators (MLEs). We obtain more general results by using a different approach: the Wilks type of results hold as long as likelihood contour sets are fan-shaped. The classical Wilks theorem corresponds to the situations in which the likelihood contour sets are ellipsoidal. This provides a geometric understanding and a useful extension of the likelihood ratio theory. As a result, even if the MLEs are not asymptotically normal, the likelihood ratio statistics can still be asymptotically chi-square distributed. Our technical arguments are simple and easily understood.

1. INTRODUCTION

One of the most celebrated folk theorems in statistics is that twice the logarithm of a maximum likelihood ratio statistic is asymptotically chi-square distributed. This result, due to Wilks (1938), is proved via a Taylor expansion of a likelihood function and by assuming that the maximum likelihood estimator (MLE) is asymptotically normal (see also Wald 1941; Wilks 1962; and heuristics given in popular textbooks such as Cox and Hinkley 1974 and Kendall and Stuart 1979, among others). Although this understanding is insightful, it has three drawbacks. First, the likelihood function must be sufficiently smooth to admit a second-order Taylor expansion. Second, the MLE must be asymptotically normal and this itself relies on Taylor expansions and the central limit theorem. Third, assumptions on the independence of observations are typically made. Rigorous technical proofs of the first two steps are by no mean simple. This is probably why rigorous statements and heuristic proofs are suppressed in many popular graduate textbooks (see, e.g., Bickel and Doksum 1977, p. 229; Casella and Berger 1990, p. 381; Lehmann 1986, p. 486).

We contend that much simpler insight to the Wilks theorem is available. If the contour sets of a likelihood function around an MLE are of fan shape, then the Wilks type of results hold. The classical Wilks theorem corresponds to the situations in which the contour sets are ellipsoidal. In general, the asymptotic normality of the MLE is not required, and the asymptotic distribution of the MLE need not exist. One can easily construct an example in which the MLE is not asymptotically normal, but a Wilks type of results hold; see Examples 1 and 2 in Section 3. An additional benefit is that our technical arguments are simple and can be understood without much probability background.

We begin with the simplest case, in which the null hypothesis consists of only one point,

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{versus} \quad H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0, \tag{1}$$

with $\boldsymbol{\theta}$ a vector of unknown parameters in an Euclidean space. Let $\mathbf{X} \sim f(\mathbf{x}; \boldsymbol{\theta})$ be a random vector from which a sample of data is drawn, and let $l(\boldsymbol{\theta}; \mathbf{x}) = \log f(\mathbf{x}; \boldsymbol{\theta})$ be the log-likelihood function. Let $\hat{\boldsymbol{\theta}}$ denote the MLE. Set

$$W(\boldsymbol{\theta}_0, \mathbf{X}) = l(\hat{\boldsymbol{\theta}}, \mathbf{X}) - l(\boldsymbol{\theta}_0, \mathbf{X}), \tag{2}$$

which is the log-likelihood ratio statistic for the test of hypotheses (1). Our idea is simple; it uses some simple tools of Bayesian statistics. We assign a continuous prior density $\pi(\cdot)$ for the parameter $\boldsymbol{\theta}$. We can then easily show that the posterior distribution of $W(\boldsymbol{\theta}, \mathbf{X})$ given $\mathbf{X}$ is asymptotically a gamma distribution, independent of the prior distribution. This implies that the marginal distribution of W is also asymptotically a gamma distribution. Because the result holds for every continuous prior distribution, it must follow that the distribution of W given $\boldsymbol{\theta}$ is asymptotically a gamma distribution. A similar kind of argument has been made before by, for example, Bickel and Ghosh (1990) and Dawid (1991). In particular, when the shape parameter of the gamma distribution is one-half of an integer, the random variable $2W$ follows asymptotically a chi-square distribution. In other words, the Wilks theorem is a specific case of our generalized results.

The foregoing arguments can be readily extended to the cases in which the null hypothesis contains nuisance parameters. The key to our success relies on the regenerating property of gamma distributions.

The article is organized as follows. Section 2 derives the posterior distributions of likelihood ratio statistics. Section 3 presents the sampling distributions of the likelihood ratio statistics from the posterior distributions. Section 4 extends the arguments to the cases in which the null hypothesis contains nuisance parameters.

Jianqing Fan and Wing-Hung Wong are Professors, Department of Statistics, University of California, Los Angeles, CA 90095 (E-mail: *jfan@stat.ucla.edu* and *whwong@stat.ucla.edu*). Fan is also Professor, Department of Statistics, Chinese University of Hong Kong. Hui-Nien Hung is Associate Professor, Institute of Statistics, National Chiao-Tong University, Taiwan (E-mail: *hhung@stat.nctu.edu.tw*). Fan's research was partially supported by National Science Foundation (NSF) grant DMS-9803200 and National Security Agency grant 96-1-0015. Hung's research was partially supported by the National Science Council of Taiwan grant NSC892118M009007. Wong's work was partially supported by the Mathematical Sciences Research Institute of the Chinese University of Hong Kong and by NSF grant DMS-9703918. The authors would like to thank the editor, an associate editor, and two referees for valuable comments on an earlier draft of this article, which led to an improved presentation.

Journal of the American Statistical Association
September 2000, Vol. 95, No. 451, Theory and Methods

2. POSTERIOR DISTRIBUTION

Assume that $\boldsymbol{\theta} \in R^p$ has a prior density $\pi(\boldsymbol{\theta})$. Then the posterior density of $\boldsymbol{\theta}$ given $\mathbf{X} = \mathbf{x}$ is given by

$$\frac{\exp\{l(\boldsymbol{\theta}, \mathbf{x})\}\pi(\boldsymbol{\theta})}{\int_\Theta \exp\{l(\boldsymbol{\theta}, \mathbf{x})\}\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}} = \exp\{-W(\boldsymbol{\theta}, \mathbf{x})\}\pi(\boldsymbol{\theta})/g_n(\mathbf{x}),$$

where

$$g_n(\mathbf{x}) = \int_\Theta \exp\{-W(\boldsymbol{\theta}, \mathbf{x})\}\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

Let $S_w = \{\boldsymbol{\theta} \in \Theta : W(\boldsymbol{\theta}, \mathbf{x}) = w\}$ be a likelihood contour set. Our aim is to show that the posterior distribution of $W(\boldsymbol{\theta}, \mathbf{X})$ given $\mathbf{X} = \mathbf{x}$ is asymptotically gamma distributed if the likelihood contour set can be approximated as

$$S_w \approx \hat{\boldsymbol{\theta}} + a_n w^r S \tag{3}$$

for a sequence of $a_n \to 0, r > 0$ and a surface S in R^p. This is an extension of classical conditions on the Wilks theorem in which the likelihood contour sets are approximated by ellipses.

Condition (3) is not rigorous. To formally state the result, we assume that there exists a function h on R^p such that

$$h(t\boldsymbol{\theta}) = t^{1/r}h(\boldsymbol{\theta}), \quad \forall\, t > 0,$$

and

$$S = \{\boldsymbol{\theta} : h(\boldsymbol{\theta}) = 1\}. \tag{4}$$

This and (3) imply heuristically that

$$W(\boldsymbol{\theta}, \mathbf{x}) \approx h(a_n^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})) = a_n^{-1/r}h(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}).$$

Let $W_n^*(\boldsymbol{\theta}, \mathbf{x}) = a_n^{1/r}W(\boldsymbol{\theta}, \mathbf{x})$. The formal conditions can be expressed as follows:

A1. There exist a function $m(\cdot)$ and a constant N such that when $n > N$,

$$\inf\{W_n^*(\boldsymbol{\theta}, \mathbf{x}) : \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\| > \delta\} \geq m(\delta) > 0 \quad \forall\, \delta > 0.$$

Moreover the MLE $\hat{\theta}$ is a stochastically bounded sequence.

A2. There exists a function $h(\cdot)$ such that the likelihood contour sets are fan-shaped in the following sense:

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \sup_{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\| \leq \delta} \left| \frac{W_n^*(\boldsymbol{\theta}, \mathbf{x}) - h(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})}{h(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})} \right| = 0.$$

A3. The function h satisfies $h(t\boldsymbol{\theta}) = t^{1/r}h(\boldsymbol{\theta})$ and $\inf\{h(\boldsymbol{\theta}) : \|\boldsymbol{\theta}\| = 1\} > 0$.

Condition A2 is a rigorous condition of (3). The latter can be intuitively understood as saying that the unit volume around surface S_w is proportional to $w^{(rp-1)}$; namely,

$$\lim_{\Delta \to 0} \Delta^{-1} V\{\boldsymbol{\theta} : W(\boldsymbol{\theta}, \mathbf{X}) \in w \pm \Delta/2\}$$

$$\text{is proportional to } w^{(rp-1)}, \tag{5}$$

where $V(A)$ denotes the volume of a set A. In other words, as long as the contour sets $\{S_w\}$ are rigid for all $w > 0$, the Wilks type of results hold.

Theorem 1. Suppose that $\pi(\boldsymbol{\theta})$ is a bounded positive continuous function. Then, under the regularity conditions A1–A3, the conditional distribution of the log-likelihood ratio statistic W given that $\mathbf{X} = \mathbf{x}$ has an asymptotic gamma distribution with shape parameter rp and scale parameter 1, namely,

$$\mathcal{L}(W|\mathbf{X} = \mathbf{x}) \to \text{gamma}(rp).$$

The proof of this theorem uses the following lemma, whose proof is similar to that of the Laplace approximation (Tierney and Kadane 1986). The proof is elementary but somewhat tedious.

Lemma 1. Under conditions A1–A3, we have

$$g_n(\mathbf{x}) = \pi(\hat{\boldsymbol{\theta}})a_n^p V(\mathcal{O})\Gamma(rp + 1)(1 + o(1)),$$

where $V(\mathcal{O})$ is the volume of the set $\mathcal{O} = [0, 1] \times S$ with S given by (4).

Proof. By condition A1, when $n > N$,

$$\int_{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\| > \delta} \exp\{-W(\boldsymbol{\theta}, \mathbf{x})\}\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} \leq \exp\{-a_n^{-1/r}m(\delta)\}.$$

For any given $\varepsilon > 0$, when δ is sufficiently small, we deduce from conditions A2 and A3 that

$$\begin{aligned}
&\int_{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\| \leq \delta} \exp\{-W(\boldsymbol{\theta}, \mathbf{x})\}\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\
&\quad \geq \int_{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\| \leq \delta} \exp\{-(1 + \varepsilon)a_n^{-1/r}h(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\}\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\
&\quad \geq \int_{\boldsymbol{\theta} \in R^p} \exp\{-(1 + \varepsilon)a_n^{-1/r}h(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\}\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\
&\qquad - \exp\{-(1 + \varepsilon)a_n^{-1/r}\delta^{1/r}c\},
\end{aligned}$$

where $c = \inf\{h(\eta) : \|\eta\| = 1\}$. By a change of variables, the first term is given by

$$a_n^p(1 + \varepsilon)^{-rp} \int_{\boldsymbol{\theta} \in R^p} \exp\{-h(\boldsymbol{\theta})\}\pi(\hat{\boldsymbol{\theta}} + a_n(1 + \varepsilon)^{-r}\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

By invoking the dominated convergence theorem, the foregoing expression is bounded from below by

$$\begin{aligned}
&\pi(\hat{\boldsymbol{\theta}})(1 - \varepsilon)a_n^p(1 + \varepsilon)^{-rp} \int_{\boldsymbol{\theta} \in R^p} \exp\{-h(\boldsymbol{\theta})\}\, d\boldsymbol{\theta} \\
&\quad = \pi(\hat{\boldsymbol{\theta}})(1 - \varepsilon)a_n^p(1 + \varepsilon)^{-rp} \int_0^\infty \exp\{-t\}\, d\{V(t^r\mathcal{O})\} \\
&\quad = \pi(\hat{\boldsymbol{\theta}})(1 - \varepsilon)a_n^p(1 + \varepsilon)^{-rp}V(\mathcal{O})rp \int_0^\infty e^{-t}t^{rp-1}\, dt.
\end{aligned}$$

By letting $n \to \infty$, and then $\delta \to 0$ and $\varepsilon \to 0$, we have

$$\liminf_{n \to \infty} a_n^{-p}\pi(\hat{\boldsymbol{\theta}})^{-1}g_n(\mathbf{x}) \geq V(\mathcal{O})\Gamma(rp + 1).$$

In a similar vein, we can easily show that

$$\limsup_{n \to \infty} a_n^{-p}\pi(\hat{\boldsymbol{\theta}})^{-1}g_n(\mathbf{x}) \leq V(\mathcal{O})\Gamma(rp + 1).$$

This concludes the proof of Lemma 1.

Proof of Theorem 1. Note that

$$P\,(W<w|\mathbf{X}=\mathbf{x})=g_n(x)^{-1}\int_{\{\theta:W(\boldsymbol{\theta},\mathbf{x})<w\}} \times e^{-W(\boldsymbol{\theta},\mathbf{x})}\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}.$$

We need only evaluate the integral term in the foregoing expression. The arguments follow similar lines to those in the proof of Lemma 1. More precisely, by condition A2, we have

$$\int_{W(\boldsymbol{\theta},\mathbf{x})<w}\exp\{-W(\boldsymbol{\theta},\mathbf{x})\}\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}$$
$$\leq \int_{\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\|>\delta}\exp\{-W(\boldsymbol{\theta},\mathbf{x})\}\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}$$
$$+\int_{\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\|\leq\delta,W(\boldsymbol{\theta},\mathbf{x})<w}\exp\{-W(\boldsymbol{\theta},\mathbf{x})\}\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}.$$

By condition A1, the first integral is bounded by $\exp\{-a_n^{-1/r}m(\delta)\}$. By condition A2 and a change of variable, the second integral is bounded by

$$\int_{h(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})<w(1+\varepsilon)a_n^{1/r}}\exp\{-a_n^{-1/r}(1-\varepsilon)h(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\}\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}$$
$$=a_n^p(1-\varepsilon)^{-rp}\int_{h(\boldsymbol{\theta})<w(1+\varepsilon)(1-\varepsilon)}\times\exp\{-h(\boldsymbol{\theta})\}\pi(\hat{\boldsymbol{\theta}}+a_n^{1/r}(1-\varepsilon)\boldsymbol{\theta})\,d\boldsymbol{\theta}$$
$$\leq\pi(\hat{\boldsymbol{\theta}})(1+\varepsilon)a_n^p(1-\varepsilon)^{-rp}\int_0^{w(1-\varepsilon^2)}e^{-t}\,d\{V(t^r\mathcal{O})\}$$
$$=\pi(\hat{\boldsymbol{\theta}})(1+\varepsilon)a_n^p(1-\varepsilon)^{-rp}rpV(\mathcal{O})\int_0^{w(1-\varepsilon^2)}\times e^{-t}t^{rp-1}\,dt.$$

Using the same method, we have

$$\int_{W(\boldsymbol{\theta},\mathbf{x})<w}\exp\{-W(\boldsymbol{\theta},\mathbf{x})\}\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}\geq\pi(\hat{\boldsymbol{\theta}})(1-\varepsilon)a_n^p(1+\varepsilon)^{-rp}rp\int_0^{w(1-\varepsilon^2)}e^{-t}t^{rp-1}\,dt.$$

Thus we have

$$\lim_{n\to\infty}a_n^{-p}\pi(\hat{\boldsymbol{\theta}})\int_{W(\boldsymbol{\theta},\mathbf{x})<w}\exp\{-W(\boldsymbol{\theta},\mathbf{x})\}\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}=V(\mathcal{O})\int_0^w e^{-t}t^{rp-1}\,dt.$$

This, together with Lemma 1, prove Theorem 1.

By noting that if $Y\sim\text{gamma}(s)$, then $2Y\sim\chi^2_{2s}$, we have the following result.

Corollary 1. Under the regularity conditions A1–A3, we have

$$\mathcal{L}\,(2W|\mathbf{X}=\mathbf{x})\to\chi^2_{2rp},$$

provided that $2rp$ is an integer.

3. SAMPLING DISTRIBUTION

In this section we derive the asymptotic distributions of likelihood ratio statistics from the frequentist viewpoint. Let Θ^0 be a bounded open set in an Euclidean space. To stress its dependence on n, we let $P_n(W\leq w|\boldsymbol{\theta})$ denote the sampling distribution. To apply the asymptotic posterior distribution established in Theorem 1, conditions A1 and A2 are required in the following sense [weaker than requiring conditions A1 and A2 to hold almost surely]:

A1*. There exists a function $m(\cdot)$ such that

$$P\{d_n(\mathbf{X},\delta)>m(\delta)\}\to1$$

where $d_n(\mathbf{x},\delta)=\inf\{W_n^*(\boldsymbol{\theta},\mathbf{x}):\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\|>\delta\}$. Moreover, the MLE $\hat{\boldsymbol{\theta}}$ is stochastically bounded.

A2*. There exists a function $h(\cdot)$ such that

$$\sup_{\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\|\leq\delta}\left|\frac{W_n^*(\boldsymbol{\theta},\mathbf{X})-h(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})}{h(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})}\right|\xrightarrow{P}0.$$

as $n\to\infty$ and then $\delta\to0$.

Theorem 2. If $P_n(W\leq w|\boldsymbol{\theta})$ is equicontinuous in $\boldsymbol{\theta}\in\Theta^0$ for every w, then under conditions A1*, A2*, and A3,

$$\mathcal{L}(W|\boldsymbol{\theta}_0)\to\text{gamma}(rp),\quad\forall\,\boldsymbol{\theta}_0\in\Theta^0.$$

Proof. Let $\{\pi_m(\boldsymbol{\theta})\}$ be a sequence of bounded continuous prior distributions of $\boldsymbol{\theta}$ that shrink to the point $\boldsymbol{\theta}_0$. Denote the marginal distribution of W by $a_{n,m}(w)=\int P_n(W<w|\boldsymbol{\theta})\pi_m(\boldsymbol{\theta})\,d\boldsymbol{\theta}$. Recall that a sequence converges in probability if and only if for any subsequence there exists a further subsequence that converges almost surely. By Theorem 1, we have $\mathcal{L}_m(W|\mathbf{X})\xrightarrow{P}\text{gamma}(r)$ for all m, where $\mathcal{L}_m(W|\mathbf{X})$ is the conditional distribution of W under the prior π_m. By using the dominated convergence theorem,

$$\lim_{n\to\infty}a_{n,m}(w)=\Gamma(w,rp),\tag{6}$$

where $\Gamma(w,rp)$ is the cumulative distribution of gamma(rp). By the equicontinuity assumption, we have

$$\lim_{m\to\infty}a_{n,m}(w)=P_n(W<w|\boldsymbol{\theta}_0),\quad\text{and the convergence is uniform in } n.\tag{7}$$

It follows from (6) and (7) that

$$\lim_{n\to\infty}P_n(W<w|\boldsymbol{\theta}_0)=\lim_{n\to\infty}\lim_{m\to\infty}a_{n,m}(w)=\lim_{m\to\infty}\lim_{n\to\infty}a_{n,m}(w)=\Gamma(w,rp).$$

This completes the proof.

Remark 1. In many cases (see Examples 1–3), the sampling distribution of W is independent of $\boldsymbol{\theta}$ and hence is equicontinuous. This condition is used to show the existence of the limit distribution of W and can be replaced by the assumption that the limiting distribution of $a_n^{-1}(\hat{\boldsymbol\theta} - \boldsymbol\theta)$ exists. To see this, under the latter assumption, by conditions A2 and A3, the likelihood ratio statistic $W = h(a_n^{-1}(\hat{\boldsymbol\theta} - \boldsymbol\theta))(1 + o_P(1))$ converges weakly. Let $G(w, \theta)$ be the limit of $P_n(W < w|\boldsymbol\theta)$. Then by (6), $EG(w, \boldsymbol\theta) = \Gamma(w, rp)$ for every continuous prior π. Hence $G(w, \boldsymbol\theta) = \Gamma(w, rp)$ for almost all $\boldsymbol\theta \in \Theta^0$.

Theorems 1 and 2 reveal that the Wilks type of results hold as long as the likelihood contour is of fan shape. It provides a good geometric interpretation for conditions for the validity of the approximation being studied. The classical Wilks theorem is usually derived under the conditions similar to Cramér's (see, e.g., conditions C1–C5 of Le Cam and Yang 1990, p. 102). These conditions imply that the log-likelihood function can be locally approximated by a quadratic function (see Le Cam and Yang 1990) and the likelihood contour is ellipsoidal,

$$S_w \approx \hat{\boldsymbol\theta} + (2w/n)^{1/2} S,$$

where $S = \{\boldsymbol\theta : \boldsymbol\theta^T \boldsymbol\Sigma \theta = 1\}$ and $\boldsymbol\Sigma$ is the Fisher information matrix at the true underlying parameter. Hence this is a specific case of our results.

We contend that the Wilks type of result holds for a much larger class of likelihood contours. The shapes of the likelihood contours do not have to be ellipsoidal, and the radii are not necessarily proportional to $w^{1/2}$. Hence the shape parameter of the gamma distribution does not need to be $p/2$, where p is the number of parameters, and the MLE is not required to be asymptotically normal.

Example 1. Suppose that we have a random sample of size n from the exponential distribution model

$$\mathbf{X} = \boldsymbol\theta + \varepsilon,$$

where $\mathbf{X}, \boldsymbol\theta$, and ε are p-dimensional vectors and the components of ε are independent having the standard exponential distribution. Then the MLE is $\hat{\boldsymbol\theta} = \min(\mathbf{X}_1, \dots, \mathbf{X}_n)$, where the operator "min" is applied componentwise. Thus $n(\hat{\boldsymbol\theta} - \boldsymbol\theta) \sim \varepsilon$, which is not asymptotically normal. The likelihood ratio statistic

$$W(\boldsymbol\theta, \mathbf{X}) = n \sum (\hat{\boldsymbol\theta} - \boldsymbol\theta) \sim \text{gamma}(p),$$

where the operator $\sum$ is applied to the components of the vector. Hence

$$2W(\boldsymbol\theta, \mathbf{X}) \sim \chi^2_{2p}.$$

Note that the degree of freedom is $2p$ instead of p. The likelihood contour in this case is

$$S_w = \left\{\boldsymbol\theta : n \sum (\hat{\boldsymbol\theta} - \boldsymbol\theta) = w, \hat{\boldsymbol\theta} \geq \boldsymbol\theta\right\} = \hat{\boldsymbol\theta} + (w/n)S,$$

where $S = \{\boldsymbol\theta : \sum \boldsymbol\theta_i = -1, \boldsymbol\theta_i \leq 0\}$ is a hypertriangle. Conditions A1*, A2*, and A3 are satisfied with $h(\boldsymbol\theta) = -\sum_{i=1}^p \boldsymbol\theta_i \mathbf{I}(\boldsymbol\theta_i \leq 0)$ and $r = 1, a_n = 1/n$.

The foregoing example provides evidence that a Wilks type of result continues to hold even though the MLE is not asymptotically normal. Such an example is not uncommon. We provide an additional one.

Example 2. Suppose that we have a random sample of size n from the uniform distribution on the p-dimensional hyperrectangle $[0, \boldsymbol\theta]$. Then the MLE is $\hat{\boldsymbol\theta} = \max(\mathbf{X}_1, \dots, \mathbf{X}_n)$, where the operator "max" is applied componentwise. The log-likelihood ratio statistic is

$$W(\boldsymbol\theta, \mathbf{X}) = n \sum_{i=1}^p \log(\boldsymbol\theta_i/\hat{\boldsymbol\theta}_i) \mathbf{I}(\boldsymbol\theta_i \geq \hat{\boldsymbol\theta}_i) \sim \text{gamma}(p).$$

Again, $\hat{\boldsymbol\theta}$ is not asymptotically normal, and the degree of freedom for $2W$ is $2p$ instead of p. The likelihood contour in this example is approximately a hypertriangle,

$$S_w \approx \hat{\boldsymbol\theta} + (w/n)S,$$

where $S = \{\boldsymbol\theta : \sum_{i=1}^p (\boldsymbol\theta_i/\boldsymbol\theta_{i0}) = 1, \boldsymbol\theta_i \geq 0\}$ with $(\boldsymbol\theta_{10}, \dots, \boldsymbol\theta_{p0})^T$ the true underlying parameters. Conditions A1*, A2*, and A3 are satisfied with $h(\boldsymbol\theta) = \sum_{i=1}^p (\boldsymbol\theta_i/\boldsymbol\theta_{i0}) \mathbf{I}(\boldsymbol\theta_i \geq 0)$ and $r = 1, a_n = 1/n$.

The Cramér condition for the Wilks theorem depends critically on parameterization. This drawback is attenuated under our formulation.

Example 3. Suppose that we parameterize a normal population as $\mathrm{N}(\boldsymbol\theta^3, \mathbf{I}_p)$, where $\boldsymbol\theta = (\boldsymbol\theta_1, \dots, \boldsymbol\theta_p)'$ is a p-dimensional unknown vector, $\boldsymbol\theta^3 = (\boldsymbol\theta_1^3, \dots, \boldsymbol\theta_p^3)'$, and $\mathbf{I}_p$ is a $p \times p$ identity matrix. Based on a random sample of size n, the MLE is given by $\hat{\boldsymbol\theta} = \bar{X}^{1/3}$, where $\bar{X}$ is the sample mean. Let $\mathbf{Z}$ be the p-dimensional standard normal random vector. It is clear that when the true parameter $\boldsymbol\theta = \mathbf{0}, n^{1/6}\hat{\boldsymbol\theta} \sim \mathbf{Z}^{1/3}$, which is not asymptotically normal. Hence the Cramér conditions do not hold under this parameterization. In this case the likelihood function can be approximated as

$$\begin{aligned} W(\boldsymbol\theta, \mathbf{X}) &= \frac{1}{2} n \|\bar{\mathbf{X}} - \boldsymbol\theta^3\|^2 \\ &= \frac{1}{2} n \sum_{i=1}^p (\bar{X}_i^{1/3} - \boldsymbol\theta_i)^2 \{9\boldsymbol\theta_{i0}^{4/3} + o_P(1)\} \end{aligned}$$

for $\boldsymbol\theta$ in a neighborhood of $\hat{\boldsymbol\theta}$, where $\boldsymbol\theta_0 = (\boldsymbol\theta_{10}, \dots, \boldsymbol\theta_{p0})^T$ are the true underlying parameters. The likelihood contour is fan-shaped,

$$S_w = \{\boldsymbol\theta : n\|\bar{X} - \boldsymbol\theta^3\|^2 = 2w\} \approx \bar{X}^{1/3} + (2w/n)^{1/2} S$$

where S is an ellipse given by $S = \{\boldsymbol\theta : \sum_{i=1}^p \boldsymbol\theta_i^2 \boldsymbol\theta_{i0}^{4/3} = 1/9\}$. Thus conditions A1*, A2*, and A3 hold with $h(\boldsymbol\theta) = 9/2 \sum_{i=1}^p \boldsymbol\theta_i^2 \boldsymbol\theta_{i0}^{4/3}, r = 1/2$, and $a_n = n^{-1/2}$.

Another advantage of our new results is that they can accommodate situations in which different parameters can have different rates of convergence. We elaborate this further in the following remark.

Remark 2. In some cases different components of $\boldsymbol\theta$ can have different asymptotic behavior. For example, suppose that $\boldsymbol\theta = (\boldsymbol\theta_1, \boldsymbol\theta_2)$ and that X and Y are two independent random variables with distributions uniform$[0, \boldsymbol\theta_1]$ and $\mathrm{N}(\boldsymbol\theta_2, 1)$. To apply Theorems 1 and 2 to this kind of problem, we need to modify conditions A1–A3 as follows:

A1′. The function $W_n^*(\boldsymbol{\theta}, \mathbf{x}) = a_n W(\boldsymbol{\theta}, \mathbf{x})$ satisfies

$$\inf\{W_n^*(\boldsymbol{\theta}, \mathbf{x}) : \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\| > \delta\} \geq m(\delta) > 0 \quad \forall\, \delta > 0.$$

A2′. The likelihood function can be approximated by

$$\lim_{\delta \to 0} \lim_{n \to \infty} \sup_{\|\hat{\theta} - \theta\| < \delta} \left| \frac{W_n^*(\boldsymbol{\theta}, \mathbf{x}) - h(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})}{h(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})} \right| = 0.$$

A3′. The function h satisfies $h(t^{r_1}\boldsymbol{\theta}_1, t^{r_2}\boldsymbol{\theta}_2) = th(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ with $\boldsymbol{\theta}_1 \in R^{p_1}$ and $\boldsymbol{\theta}_2 \in R^{p_2}$. Further, $\inf\{h(\boldsymbol{\theta}) : \|\boldsymbol{\theta}\| = 1\} > 0$.

Then we need only replace the shape parameter rp in Theorem 1 by $r_1 p_1 + r_2 p_2$. The unit volume in (5) is now proportional to $r_1 p_1 + r_2 p_2 - 1$ in this case.

4. EXTENSION TO CASES WITH NUISANCE PARAMETERS

We now consider situations with composite null hypothesis. Partition the parameter vector $\boldsymbol{\xi}$ into two parts $\boldsymbol{\xi} = (\boldsymbol{\theta}^T, \lambda^T)^T$, where $\boldsymbol{\theta} \in R^p$ and $\boldsymbol{\lambda} \in R^q$. Under the null hypothesis that $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, the likelihood ratio test statistic is

$$W_1(\boldsymbol{\theta}_0, \mathbf{X}) = l(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\lambda}}, \mathbf{X}) - l(\boldsymbol{\theta}_0, \hat{\boldsymbol{\lambda}}_{\boldsymbol{\theta}_0}, \mathbf{X}),$$

where $\hat{\lambda}_{\boldsymbol{\theta}_0}$ is the MLE under the null hypothesis that $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

Decompose the likelihood ratio $W(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{X}) = l(\hat{\boldsymbol{\theta}}, \hat{\lambda}, \mathbf{X}) - l(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{X})$ as

$$W(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{X}) = W_1(\boldsymbol{\theta}, \mathbf{X}) + W_2(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{X}), \tag{8}$$

where

$$W_2(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{X}) = l(\boldsymbol{\theta}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{\theta}}, \mathbf{X}) - l(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{X}).$$

As in Section 2, we first consider the posterior distributions of $W(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{X})$ and $W_2(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{X})$. The regularity conditions are similar to A1–A3 and are stated as follows:

B1. The likelihood function $W(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{X})$ satisfies conditions A1–A3.

B2. For each given $\boldsymbol{\theta}$ in a bounded open set Θ^0, the likelihood function $W_2(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{X})$ satisfies conditions A1–A3.

Conditions B1 and B2 admit similar geometric interpretation as that given in Section 2. In particular, the likelihood contour sets must be fan-shaped. These regularity conditions can be understood as

$$\lim_{\Delta \to 0} \Delta^{-1} V\{(\boldsymbol{\theta}, \boldsymbol{\lambda}) : W(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{X}) \in w \pm \Delta/2\} \quad \text{is proportional to } w^{r(p+q)-1},$$

and

$$\lim_{\Delta \to 0} \Delta^{-1} V\{\boldsymbol{\lambda} : W_2(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{X}) \in w \pm \Delta/2\} \quad \text{is proportional to } w^{rq-1}.$$

Theorem 3. Suppose that $\pi(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is a bounded positive continuous prior on a bounded open set $\Theta^0 \times \Lambda^0$. Then the posterior distribution of the likelihood ratio statistic $W_1(\boldsymbol{\theta}, \mathbf{X})$ has an asymptotic gamma distribution,

$$\mathcal{L}\{W_1(\boldsymbol{\theta}, \mathbf{X}) | \mathbf{X} = \mathbf{x}\} \to \text{gamma}(rp).$$

Proof. By Theorem 1, we have

$$\mathcal{L}\{W(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{X}) | \mathbf{X}\} \to \text{gamma}(r(p+q))$$

and

$$\mathcal{L}\{W_2(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{X}) | \mathbf{X}, \boldsymbol{\theta}\} \to \text{gamma}(rq).$$

Because r is independent of $\boldsymbol{\theta}$, conditioning on $\mathbf{X}$, $W_2(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{X})$ is asymptotically independent of $\boldsymbol{\theta}$ and hence independent of $W_1(\boldsymbol{\theta}, \mathbf{X})$. It follows that the characteristic functions satisfy

$$\begin{aligned} E\{\exp(itW)|\mathbf{X}\} &= E\{\exp(itW_1) E\{\exp(itW_2)|\mathbf{X}, \boldsymbol{\theta}\}|\mathbf{X}\} \\ &= E\{\exp(itW_1)|\mathbf{X}\}\phi(t, rq) + o(1), \end{aligned}$$

where $\phi(t, rq)$ is the characteristic function of gamma(rq). Thus

$$E\{\exp(itW_1)|\mathbf{X}\} = \phi(t, r(p+q) - rq) + o(1).$$

This completes the proof.

From the posterior distribution, we can similarly obtain the sampling distribution.

Theorem 4. If $\mathcal{L}\{W_1(\boldsymbol{\theta}, \mathbf{X}) | \boldsymbol{\theta}, \boldsymbol{\lambda}\}$ is equicontinuous in $(\boldsymbol{\theta}, \boldsymbol{\lambda}) \in \Theta^0 \times \Lambda^0$, an open set in the Euclidean space, or if the limiting distributions of the MLEs $(\hat{\boldsymbol{\theta}}^T, \hat{\boldsymbol{\lambda}}^T)^T$ and $\hat{\boldsymbol{\lambda}}_{\boldsymbol{\theta}}$ exist upon suitable normalization, then the likelihood ratio statistic has an asymptotic gamma distribution,

$$\mathcal{L}\{W_1(\boldsymbol{\theta}, \mathbf{X}) | \boldsymbol{\theta}, \boldsymbol{\lambda}\} \to \text{gamma}(rp),$$

for almost all $(\boldsymbol{\theta}, \boldsymbol{\lambda}) \in \Theta^0 \times \Lambda^0$.

Note that situations similar to those described in Remark 2 can also be accommodated in Theorem 4. For simplicity, we omit the details.

[Received March 1998. Revised August 1999.]

REFERENCES

Bickel, P. J., and Doksum, K. A. (1977), *Mathematical Statistics: Basic Ideas and Selected Topics*, Englewood Cliffs, NJ: Prentice-Hall.

Bickel, P. J., and Ghosh, J. K. (1990), "A Decomposition for the Likelihood Ratio Statistic and the Bartlett Correction—A Bayesian Argument," *The Annals of Statistics*, 18, 1070–1090.

Casella, G., and Berger, R. L. (1990), *Statistical Inference*, Belmont, CA: Duxbury.

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, London: Chapman and Hall.

Dawid, A. P. (1991), "Fisherian Inference in Likelihood and Frequential Frames of Reference" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 53, 79–109.

Kendall, M., and Stuart, A. (1979), *The Advanced Theory of Statistics, Vol.*

II: Inference and Relationship (4th ed.), New York: Macmillan.
Le Cam, L., and Yang, G. L. (1990), *Asymptotics in Statistics: Some Basic Concepts*, New York: Springer-Verlag.
Lehmann, E. L. (1986), *Testing Statistical Hypotheses* (2nd ed.), New York: Wiley.
Tierney, L., and Kadane, J. B. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, 81, 82–86.
Wald, A. (1941), "Asymptotically Most Powerful Tests of Statistical Hypotheses," *The Annals of Mathematical Statistics*, 12, 1–19.
Wilks, S. S. (1938), "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses," *The Annals of Mathematical Statistics*, 9, 60–62.
——— (1962), *Mathematical Statistics*, New York: Wiley.