Asymptotic Distribution of $P$ Values in Composite Null Models

James M. Robins, Aad van der Vaart, and Valérie Ventura

We investigate the compatibility of a null model $H_0$ with the data by calculating a $p$ value; that is, the probability, under $H_0$, that a given test statistic $T$ exceeds its observed value. When the null model consists of a single distribution, the $p$ value is readily obtained, and it has a uniform distribution under $H_0$. On the other hand, when the null model depends on an unknown nuisance parameter $\theta$, one must somehow get rid of $\theta$ (e.g., by estimating it) to calculate a $p$ value. Various proposals have been suggested to “remove” $\theta$, each yielding a different candidate $p$ value. But unlike the simple case, these $p$ values typically are not uniformly distributed under the null model. In this article we investigate their asymptotic distribution under $H_0$. We show that when the asymptotic mean of the test statistic $T$ depends on $\theta$, the posterior predictive $p$ value of Gutman and Rubin, and the plug-in $p$ value are conservative (i.e., their asymptotic distributions are more concentrated around 1/2 than a uniform), with the posterior predictive $p$ value being the more conservative. In contrast, the partial posterior predictive and conditional predictive $p$ values of Bayarri and Berger are asymptotically uniform. Furthermore, we show that the discrepancy $p$ value of Meng and Gelman and colleagues can be conservative, even when the discrepancy measure has mean 0 under the null model. We also describe ways to modify the conservative $p$ values to make their distributions asymptotically uniform.

KEY WORDS: Asymptotic relative efficiency; Bayesian $p$ values; Bootstrap tests; Goodness of fit; Model checking.

1. INTRODUCTION

Bayarri and Berger (1999, 2000) proposed two new “Bayesian” $p$ values, the conditional predictive $p$ value and the partial posterior predictive $p$ value. They claimed that for checking the adequacy of a parametric model, these new $p$ values are often superior to the “plug-in” (i.e., parametric bootstrap) $p$ value and to previously proposed “Bayesian” $p$ values: the prior predictive $p$ value of Box (1980), the posterior predictive $p$ value of Guttman (1967) and Rubin (1984), and the discrepancy $p$ value of Gelman, Carlin, Stern, and Rubin (1995), Gelman, Meng, and Stern (1996), and Meng (1994). Their claim of superiority is based on extensive investigations of the small-sample properties of the various candidate $p$ values in specific examples. In this article, we investigate their large-sample properties and find that our asymptotic results indeed confirm the superiority of the conditional predictive and partial posterior predictive $p$ values.

In Section 2 we state two theorems that characterize the asymptotic distributions of the candidate $p$ values; their proofs are given in Section 5. In Section 3 we study three examples that vividly illustrate the advantages of Bayarri and Berger’s proposals. For certain models, however, the new $p$ values may be difficult to compute, and alternative approaches would be useful. One such approach, discussed in Section 4, is to appropriately modify the test statistic or discrepancy measure to make the plug-in, posterior predictive, and discrepancy $p$ values asymptotically uniform. This approach is particularly successful for the discrepancy $p$ value, in that we derive a test based on a partial discrepancy measure that is both asymptotically uniform and locally most powerful against prespecified alternatives. But this discrepancy can itself be difficult to compute in complex models. The remainder of this section is devoted to a broad overview of the main concerns and results of the article.

1.1 Overview

Suppose that we have observed a realization $x_{\text{obs}}$ of a random variable $X$. We posit a parametric “null” model, $H_0$: $f(x; \theta), \theta \in \Theta \subset \mathbb{R}^p$, for the density of $X$, and wish to investigate the compatibility of the null model with the observed data $x_{\text{obs}}$. We do so by comparing the distribution of a given test statistic $T = t(X)$ with its observed value $t_{\text{obs}} = t(x_{\text{obs}})$, using the $p$ value

$$p(x_{\text{obs}}) \equiv \Pr(m(\cdot) | t(X) > t_{\text{obs}})$$

as a measure of compatibility, where $m(x) \equiv m_X(x)$ is a reference density for $X$ and $m(t) \equiv m_X(t)$ the corresponding marginal density of $T$, and the superscript $m(\cdot)$ signifies that $X$ has density $m(x)$.

One approach to testing compatibility of the null model $f(x; \theta)$ is to embed it into a larger parametric model,

$$f(x; \psi, \theta), (\psi, \theta) \in \Psi \times \Theta,$$

in which $\psi = 0$ corresponds to the null model $H_0$; that is,

$$f(x; 0, \theta) = f(x; \theta), \quad \theta \in \Theta,$$

whereas $\psi \neq 0$ characterizes alternatives to $H_0$. When $f(x; \psi, \theta)$ truly represents all alternatives thought likely to be true when $H_0$ is not true, Bayesian statisticians tend to forego the use of $p$ values in lieu of Bayes factors or a full Bayesian analysis. But when $f(x; \psi, \theta)$ is simply used to represent alternatives to $H_0$ that are substantively important to detect, or when no alternative model is specified,
many Bayesian statisticians join with their frequentist counterparts and use \( p \) values as measures of compatibility.

1.2 Candidate \( p \) Values

To calculate the \( p \) value (1), a reference density \( m \) must be chosen. If the null model consists of a single density \( f(x; \theta) \), there is universal agreement that \( m(x) \) should be \( f(x; \theta) \). Then \( p(X) \) is uniform when \( H_0 \) is true, where \( p(X) \) denotes the random variable whose observed value is \( p(x_{obs}) \). When the parameters space \( \Theta \) is not a singleton (i.e., \( H_0 \) is composite), one must eliminate the unknown “nuisance” parameter \( \theta \) to obtain a reference density \( m \) (in (1)). Bayarri and Berger (2000) considered various candidates for \( m \), each resulting in a different candidate \( p \) value. For example, the plug-in \( p \) value \( p_{plug} \) uses \( m_{plug}(x_{obs}) = f(x; \theta_{obs}) \), where \( \theta_{obs} \) maximizes \( f(x_{obs}; \theta) \); note that we write \( m(\cdot) \) in (1) as \( m_{plug}(\cdot|X_{obs}) \) to stress its dependence on the observed data \( X_{obs} \). The reference densities for \( p_{plug} \) and for other candidate \( p \) values based on the statistic \( t(X) \) are reported in Table 1. Most of the \( p \) values considered by Bayarri and Berger are called “Bayesian” \( p \) values, because they assume a (possibly improper) prior density \( \pi(\theta) \) for \( \theta \). These include the prior predictive \( p \) value \( p_{prior} \) of Box (1980) and the posterior predictive \( p \) value \( p_{post} \) of Guttmann (1967) and Rubin (1984), which use the prior and posterior predictive densities as references. Bayarri and Berger (1999, 2000) added two new proposals, the partial posterior predictive \( p \) value \( p_{ppost} \) and the conditional posterior \( p \) value \( p_{cpred} \). We also study two additional candidate \( p \) values that were not considered by Bayarri and Berger (2000): the conditional plug-in \( p \) value \( p_{plug} \), which uses the maximizer \( \theta_{MLE, obs} \) of the conditional likelihood \( f(x_{obs}|x_{obs}; \theta) \) as a plug-in, and the discrepancy \( p \) value \( p_{dis} \) of Gelman et al. (1995), Gelman et al. (1996), and Meng (1994) which replaces the test statistic \( t(X) \) by a discrepancy \( t(X, \theta) \), a function of the data \( X \) and of the parameter \( \theta \), so that

\[
p_{dis} = p_{dis}(x_{obs}) = P_{m(\cdot)}(\{t(X, \theta) > t(x_{obs}, \theta)\}),
\]

with \( m_{dis}(x, \theta|x_{obs}) = f(x; \theta)p_{post}(\theta|x_{obs}) \).

1.3 Desirable Sampling Properties of Candidate \( p \) Values

First, we present some terminology. We call the random variable \( p(X) \) a candidate \( p \) value if it has range \([0, 1]\); if it is also uniform under \( H_0 \), then we say that \( p(X) \) is a frequentist \( p \) value. When a candidate \( p \) value is not uniform, we say that it is conservative (anticonservative) at \( \theta \) if \( Pr[p(X) < t] \) is less (greater) than \( t \) for all \( t < 1/2 \) when \( H_0: X \sim f(x; \theta), \theta \in \Theta \) is true. Finally, a candidate \( p \) value is globally conservative (anticonservative) if it is conservative (anticonservative) for all \( \theta \in \Theta \).

This terminology was motivated by the following considerations. All candidate \( p \) values in Table 1 have range \([0, 1]\), but because \( H_0 \) is composite, they may not be uniformly distributed, even when \( H_0 \) is true. Yet in practice, we use small values of \( p(x_{obs}) \) to denote surprise or incompatibility because, in analogy with the noncomposite case, we act as if \( p(X) \) was \( U[0,1] \) under \( H_0 \). Seriously anticonservative candidate \( p \) values may cause us to discard the null model even when it is quite compatible with the data, whereas seriously conservative candidates may cause us to fail to discard models that are grossly incompatible with the data. Examples are given in Section 3.

The essential point is that a \( p \) value is useful for assessing compatibility of the null model with the data only if its distribution under the null model is known to the analyst; otherwise, the analyst has no way of assessing whether or not observing \( p = .25 \), say, is surprising, were the null model true. That we specify that distribution to be uniform is largely a matter of convention. A useful analogy is as follows. It is a matter of convention whether temperatures are reported on the centigrade versus the Fahrenheit scale; however, if we are told that the temperature is 30\(^\circ\)C, then it is essential that we are also told the scale if we are to know whether to plan to go swimming or skiing.

Hence, for frequentist testing purposes, we should require that candidate \( p \) values be frequentist \( p \) values. This requirement is generally unfulfillable, with the exception of special models, many of which were discussed by Bayarri and Berger (2000), but often can be approximately satisfied in large samples, particularly when, as we assume, the data \( X \) arise from \( n \) mutually independent random variables. Then \( m \) in (1) can be chosen so that \( p(X) \) is an asymptotic frequentist \( p \) value; that is, one whose distribution converges in law to a \( U[0,1] \) distribution under \( H_0: X \sim f(x; \theta) \) for all \( \theta \in \Theta \), as \( n \to \infty \).

We next argue that Bayesian statisticians who use \( p \) values to assess the compatibility of a model with the data have

Table 1. Reference Densities for the Various Candidate \( p \) Values

<table>
<thead>
<tr>
<th>Method</th>
<th>Reference density</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plug-in ( (p_{plug}) )</td>
<td>( m_{plug}(x_{obs}) = f(x; \theta_{obs}) )</td>
</tr>
<tr>
<td>Prior predictive ( (p_{prior}) )</td>
<td>( m_{prior}(x) = \int f(x; \theta)p(\theta)d\theta )</td>
</tr>
<tr>
<td>Posterior predictive ( (p_{post}) )</td>
<td>( m_{post}(x_{obs}) = \int f(x; \theta)p_{post}(\theta</td>
</tr>
<tr>
<td>Partial posterior predictive ( (p_{ppost}) )</td>
<td>( m_{ppost}(x_{obs}) = \int f(x; \theta)p_{ppost}(\theta</td>
</tr>
<tr>
<td>Conditional predictive ( (p_{cpred}) )</td>
<td>( m_{cpred}(x_{obs}) = \int f(x; \theta)p_{cpred}(\theta</td>
</tr>
<tr>
<td>Conditional plug-in ( (p_{plug}) )</td>
<td>( m_{plug}(x_{obs}) = f(x; \theta_{MLE, obs}) )</td>
</tr>
<tr>
<td>Discrepancy ( (p_{dis}) )</td>
<td>( m_{dis}(x, \theta</td>
</tr>
</tbody>
</table>

NOTE: The data model is \( f(x; \theta) \), where \( \theta \) has MLE \( \hat{\theta}_{MLE} \). The prior for \( \theta \) is \( \pi(\theta) \), and the posterior \( \pi_{post}(\theta|x_{obs}) \) is \( \int f(x_{obs}|x_{obs}; \theta)p(\theta|x_{obs})d\theta \). The posterior in the conditional model \( f(x|\theta) \) is \( \pi_{post}(\theta|x_{obs})f(x_{obs}|x_{obs}; \theta)p(\theta|x_{obs})d\theta \). The conditional MLE \( \hat{\theta}_{MLE, obs} \) is the maximizer of \( f(x_{obs}|x_{obs}; \theta)p(\theta|x_{obs})d\theta \). The posterior \( \pi_{post}(\theta|x_{obs}) \) is \( \pi_{post}(\theta|x_{obs})f(x_{obs}|x_{obs}; \theta)p(\theta|x_{obs})d\theta \). Here \( \hat{\theta}_{MLE, obs} \) denotes the marginal density of the random variable \( \hat{\theta}_{MLE, obs} \) evaluated at its observed value \( \hat{\theta}_{MLE, obs} \), and \( f(x; \theta_{MLE, obs}) \) denotes the conditional density of \( X \) given \( \hat{\theta}_{MLE, obs} \).
should require them to be asymptotic frequentist \( p \) values. For if the goal is to check the model rather than the prior, then any procedure should perform adequately whatever the prior, including point-mass priors. This would imply that \( p \) values should be required to be frequentist \( p \) values, a requirement which, as mentioned earlier, usually cannot be fulfilled. But because as the sample size increases, the data dominate any prior with support on all of the \( \Theta \), Bayesians should both expect and require that any model checking procedure should perform adequately in the limit as \( n \to \infty \). Of course, not all Bayesians would agree with this argument; Bayarri and Berger (1999), Box (1980), Evans (1997), and Meng (1994), have provided some alternate viewpoints.

### 1.4 Centering of Test Statistics

The following discussion applies to all of our candidate \( p \) values except the discrepancy. In most statistics texts, discussions of the asymptotic distribution of tests of fit for a null model \( H_0 \): \( f(x; \theta), \theta \in \Theta \) restrict attention to statistics \( t(x) \) such as the score, likelihood ratio, or Wald test of the hypothesis \( \psi = 0 \) in a larger model (2a)–(2b), or general chi-squared goodness-of-fit statistics, which are asymptotically pivotal with distribution \( F \), often a chi-squared or a standard normal distribution independent of \( \theta \). Then \( m_{\psi}(t) = 0 \) in (1) is the density of \( T = t(X) \) corresponding to \( F \), which does not depend on the observed data \( x_{\text{obs}} \). In contrast, in the Bayesian \( p \) value and parametric bootstrap literature, the limiting distribution of \( t(x) \) often depends on \( \theta \) and the reference density \( m_{\psi}(t|X_{\text{obs}}) \) depends on \( X_{\text{obs}} \), although in the bootstrap context attention is generally restricted to statistics \( t(X) \) whose asymptotic mean is independent of \( \theta \). We show in Theorem 1 that under regularity conditions, all of the aforementioned candidate \( p \) values, with the exception of the prior predictive \( p \) value, are asymptotic frequentist \( p \) values when the asymptotic mean of \( t(X) \) does not depend on \( \theta \).

**Remark 1.** We have not yet discussed the most common definition of a frequentist \( p \) value,

\[
p_{\text{sup}}(X_{\text{obs}}) = \sup_{\theta \in \Theta} p(t(x; \theta) > t(X_{\text{obs}})).
\]

The \( p \) value \( p_{\text{sup}}(X) \), like \( p_{\text{prior}}(X) \), need not be an asymptotic frequentist \( p \) value if the limiting distribution of \( t(X) \) depends on \( \theta \), even if the asymptotic mean of \( t(X) \) does not vary with \( \theta \). For this reason, we do not consider either \( p_{\text{sup}}(X) \) or \( p_{\text{prior}}(X) \) further.

Many of the test statistics considered in the Bayesian \( p \) value literature have asymptotic means that depend on the parameter \( \theta \); three examples illustrate this in Section 3. In the remainder of the article, we study the consequences of allowing the asymptotic mean of \( t(X) \) to depend on \( \theta \). We restrict attention to statistics \( t(X) \) with a normal limiting distribution. Extensions to statistics with limiting chi-squared or folded normal distributions are immediate, and asymptotic results for statistics with other limiting distributions will be pursued elsewhere.

If the asymptotic mean of \( t(X) \) varies with \( \theta \), then the plug-in and posterior predictive \( p \) values will be conservative even as \( n \to \infty \), with the former always the less conservative, whereas the conditional plug-in \( p \) value \( p_{\text{plug}} \) will be anticonservative. In contrast, under regularity conditions, the partial posterior predictive and the conditional predictive \( p \) values are asymptotic frequentist \( p \) values. Further, the asymptotic power of the nominal \( \alpha \)-level test based on \( p_{\text{post}} \) against local Pitman alternatives is always less than the power of the test based on \( p_{\text{plug}} \), itself less than the power of the tests based on \( p_{\text{post}} \) and \( p_{\text{pred}} \). In fact, we show that in certain examples, the asymptotic relative efficiency (ARE) of the partial posterior predictive or conditional predictive test compared to a locally efficient likelihood ratio or score test is 1, whereas the ARE of the posterior predictive test can be \( 10^{-2} \) or less, and, consequently, its power much less than the nominal \( \alpha \)-level; see Section 3 for examples.

In the proof of Corollary 3 in Section 5, we show that the posterior predictive and plug-in \( p \) values are conservative when the maximum likelihood estimator (MLE) \( \hat{\theta} \) and \( t(X) \) are asymptotically correlated, regardless of the sign of the correlation. Furthermore, we show that \( \theta \) and \( t(X) \) will be asymptotically correlated whenever the asymptotic mean of \( t(X) \) depends on \( \theta \). Thus, as pointed out by Berger and Bayarri (1999, 2000) and Evans (1997), the problem with these \( p \) values is that \( t(X) \) is effectively used twice: first to estimate \( \theta \), and again to assess lack of fit. In contrast, the conditional MLE, \( \hat{\theta}_{\text{MLE}}, \) and \( t(X) \) are always asymptotically uncorrelated, which is why \( p_{\text{post}}, p_{\text{pred}}, \) and \( p_{\text{plug}} \) are not conservative. But although \( p_{\text{post}} \) and \( p_{\text{pred}} \) are asymptotically uniform, \( p_{\text{plug}} \) is anticonservative, because it fails to properly account for the variability of \( \hat{\theta}_{\text{MLE}} \). In proposing \( p_{\text{post}} \) and \( p_{\text{pred}} \), both motivated through Bayesian arguments, Bayarri and Berger have solved the “frequentist” math problem of finding a reference density \( m_{\psi}(X) \) such that the \( p \) value (1), based on an arbitrary statistic \( t(X) \) with a limiting normal distribution, is an asymptotic frequentist \( p \) value. What is curious is that the obvious frequentist guesses, \( m_{\text{plug}}(X_{\text{obs}}) \) and \( m_{\text{plug}}(X_{\text{obs}}) \), fail.

The preceding considerations do not apply to the discrepancy \( p \) value. Specifically, we show that \( p_{\text{dis}} \) can be seriously conservative even when the discrepancy \( |t(X, \theta)| \) has asymptotic mean 0 under \( f(x; \theta) \) for all \( \theta \in \Theta \).

### 2. LARGE-SAMPLE RESULTS

To formally study the large-sample properties of our candidate \( p \) values, we consider the following canonical setup. At sample size \( n \), the data are \( X \equiv X_n = (X_1, \ldots, X_n) \), where the \( X_i \) are mutually independent random variables, each following a parametric model \( f_i(x; \psi_n, \theta) \) with \( \psi_n \in \Psi \subset R^1 \) and \( \theta \in \Theta \subset R^p \). Thus the likelihood is

\[
f(x; \psi_n, \theta) = \prod_{i=1}^{n} f_i(x_i; \psi_n, \theta).
\]

Unlike Robins (1999), we do not assume the \( X_i \) to be identically distributed to allow for regression models in which the regressors are regarded as fixed constants, as in Example A of Section 3. The subscript \( n \) in \( \psi_n \) indicates that the uni-dimensional nuisance parameter is allowed to vary with \( n \); that is, \( \psi_n = 0 \) for all \( n \) under \( H_0 \), and \( \psi_n = k_n/\sqrt{n} \) under...
local Pitman alternatives, where $k_n \to k \in \mathbb{R}^d$ as $n \to \infty$. Note that when $\psi_n = 0$, we frequently write the data model $f(x; \psi_n, \theta) = f(x; 0, \theta)$ more simply as $f(x; \theta)$ and, for notational convenience, often suppress the subscript $n$ denoting sample size in quantities such as $X \equiv X_n$.

Attention is restricted to univariate test statistics $t(X)$ that are asymptotically normal with asymptotic mean $\nu_n(k_n/\sqrt{n})$ and asymptotic variance $\sigma^2(\theta)$ under the null and local alternatives; that is, we assume that when $X \sim f(x; k_n/\sqrt{n})$,

$$n^{1/2} \left[ t(X) - \nu_n(k_n/\sqrt{n}) \right] / \sigma(\theta) \sim \text{N}(0,1),$$

(3a)

where $\sim$ denotes convergence in distribution. Note that because local alternatives are contiguous (van der Vaart 1998, chap. 6), the asymptotic variance $\sigma^2(\theta)$ does not depend on $k$. We also assume that (3a) holds for sequences $\theta = \theta_0 + k^* \sqrt{n}$ for any fixed $\theta_0$ and $k^*$.

We further assume that the functions $\nu_n$ are continuously differentiable in a neighborhood of $(0, \theta)$, with partial derivatives converging to limits as $n \to \infty$. Thus

$$\dot{\nu}_\theta(\theta) = \lim_{n \to \infty} \partial \nu_n(0, \theta) / \partial \theta$$

(3b)

and

$$\dot{\psi}_\theta(\theta) = \lim_{n \to \infty} \partial \nu_n(\psi, \theta) / \partial \psi|_{\psi=0}$$

(3c)

both exist. Note that $\dot{\nu}_\theta(\theta)$ is a vector that is nonzero only when the asymptotic mean $\nu_n(0, \theta)$ of $t(X)$ under $H_0$ depends on $\theta$. In Theorem 3, we prove that under mild additional conditions, $\dot{\nu}_\theta(\theta)$ and $\dot{\psi}_\theta(\theta)$ are equal to the asymptotic covariances of $n^{1/2}(t(X) - \nu_n(0, \theta))$ with $n^{-1/2} S_\theta(\theta) = n^{-1/2} \partial \log f(x; \psi, \theta) / \partial \psi|_{\psi=0}$ and $n^{-1/2} \mathbf{S}_\theta(\theta) = n^{-1/2} \partial \log f(x; 0, \theta) / \partial \theta$, where $S_\theta(\theta)$ and $\mathbf{S}_\theta(\theta)$ are the scores for $\psi$ and $\theta$ at $\psi = 0$.

We also need to define the scalars

$$\Omega(\theta) = \dot{\nu}_\theta(\theta)' i_{\dot{\theta}}(\theta) \dot{\nu}_\theta(\theta)$$

(4)

and

$$\omega(\theta) = \dot{\psi}_\theta(\theta) - \dot{\nu}_\theta(\theta)' i_{\dot{\theta}}(\theta) i_{\dot{\psi}}(\theta)$$

(5)

and the noncentrality parameter

$$\text{NC}(\theta) = \omega(\theta)/[\sigma^2(\theta) - \Omega(\theta)]^{1/2},$$

(6)

where $i_{\dot{\theta}}(\theta) = \lim_{n \to \infty} n^{-1} E_\theta[-\partial^2 \log f(x; 0, \theta) / \partial \theta \partial \theta] = \lim_{n \to \infty} n^{-1} E_\theta[\mathbf{S}_\theta(\theta)' \mathbf{S}_\theta(\theta)]$ and $i_{\dot{\psi}}(\theta) = \lim_{n \to \infty} n^{-1} E_\theta[-\partial^2 \log f(x; \psi, \theta) / \partial \psi \partial \psi]|_{\psi=0} = \lim_{n \to \infty} n^{-1} E_\theta[S_\theta(\theta)' \mathbf{S}_\theta(\theta)]$. Here and elsewhere, $E_\theta$ denotes expectation with respect to $f(x, \theta)$. Note that $\Omega(\theta)$ and $\sigma(\theta)$, in contrast with $\omega(\theta)$ and $\text{NC}(\theta)$, depend only on the null model $f(x; \theta)$.

We are now ready to state our main theorem, which we subsequently interpret in a series of remarks.

**Theorem 1.** Subject to the assumptions of Theorems 3 and 4, under law $f(x; k_n/\sqrt{n}, \theta)$, each candidate $p$ value can be written as

$$p(X) = 1 - \Phi(Q) + o_p(1),$$

where $\Phi(1)$ denotes a random variable converging to 0 in probability, $\Phi$ is the standard normal cdf, and $Q = q(X) = N(k_{\mu}(\theta), \tau^2(\theta))$, with $\mu(\theta) = \tau(\theta) \text{NC}(\theta)$. The values of $\tau^2(\theta)$ for our candidates are as follows:

**Plug-in:** $\tau^2_{\text{plug-in}}(\theta) = [\sigma^2(\theta) - \Omega(\theta)]/\sigma^2(\theta)$

**Posterior predictive:** $\tau^2_{\text{post}}(\theta) = [\sigma^2(\theta) - \Omega(\theta)]/\sigma^2(\theta)$

**Partial posterior predictive:** $\tau^2_{\text{post}}(\theta) = 1$

**Conditional predictive:** $\tau^2_{\text{cpred}}(\theta) = 1$

**Conditional plug-in:** $\tau^2_{\text{plug-in}}(\theta) = \sigma^2(\theta)/[\sigma^2(\theta) - \Omega(\theta)]$

**Remark 2: Asymptotic Frequentist p Values.** Theorem 1 implies that a candidate $p$ value is an asymptotic frequentist $p$ value under $H_0$ (i.e., $k = 0$) if and only if $\tau^2(\theta) = 1$. Hence all candidate $p$ values referred to in Theorem 1 are asymptotic frequentist $p$ values when $\psi(\theta) = 0$, because then $\Omega(\theta)$ in (4) is 0. When $\Omega(\theta) \neq 0$, with $\tau^2(\theta) < 1$ ($\tau^2(\theta) > 1$), the $p$ value is conservative (anticonservative). Hence $p_{\text{plug-in}}$ is anticonservative, whereas $p_{\text{post}}$ and $p_{\text{plug}}$ are conservative, with $p_{\text{post}}$ being the more conservative because $\tau^2_{\text{post}}(\theta) < \tau^2_{\text{plug-in}}(\theta)$.

**Remark 3: Efficiency.** Let $\chi(\alpha) = I[p(X) < \alpha]$ denote the nominal $\alpha$-level test that rejects $H_0$ whenever $\chi(\alpha) = 1$; here $I[A]$ is the indicator function for event $A$ that takes value 1 if $A$ is true and 0 otherwise.

The asymptotic power of test $\chi(\alpha)$ is

$$\beta(\alpha, k, \theta) = \lim_{n \to \infty} E_{k/\sqrt{n}, \theta}[\chi(\alpha)],$$

(7)

where $E_{k/\sqrt{n}, \theta}$ refers to expectations under $f(x; k/\sqrt{n}, \theta)$. The asymptotic representation of $p(X)$ given in Theorem 1 implies that $\chi(\alpha)$ has $\beta(\alpha, k, \theta) = 1 - \Phi[z_{1-\alpha} \tau^{-1}(\theta) - k\text{NC}(\theta)]$. In model $f(x; \psi, \theta)$, a locally most powerful asymptotic $\alpha$-level test $\chi_{\text{asympt}}(\alpha)$ of the hypothesis $\psi = 0$ has asymptotic power $1 - \Phi[z_{1-\alpha} \tau^{-1}(\theta) - k\text{NC}(\theta)]$, whereas $\text{NC}(\theta)$ is the efficient noncentrality parameter $\text{NC}(\theta) = \{i_{\dot{\theta}}(\theta) - i_{\dot{\psi}}(\theta) \dot{\psi}(\theta) \dot{\psi}(\theta)\}^{1/2}$, with $i_{\dot{\theta}}(\theta) = \lim_{n \to \infty} n^{-1} E_\theta[S_\theta(\theta)]$ (see van der Vaart 1998, chap. 15). The following lemma indicates that a sufficient condition for $\chi_{\text{post}}(\alpha)$ and $\chi_{\text{cpred}}(\alpha)$ to be locally most powerful at a particular value $\theta^*$ of $\theta$ is that $t(X)$ is asymptotically equivalent to an affine transformation of $S_\theta(\theta^*)$, because then $\text{NC}(\theta^*) = \text{NC}(\theta^*)$; the proof is in Section 5.

**Lemma 1.** Under regularity conditions, if $n^{1/2} t(X) = a n^{-1/2} S_\theta(\theta^*) + b + o_p(1)$ under $f(x; \theta^*)$, for some constants $a$ and $b$ with $a \neq 0$, then $\text{NC}(\theta^*) = \text{NC}(\theta^*)$.

**Remark 4: Actual Asymptotic Level and Relative Power and Efficiency.** The asymptotic actual $\alpha$-level of test $\chi(\alpha)$ is $\beta(\alpha, k, \theta)$ evaluated at $k = 0$, which we denote by $\alpha(\alpha, \theta)$. As a function of $\alpha$, actual $\alpha(\alpha, \theta)$ is the cdf of the asymptotic distribution of $p(X)$ under $H_0$. We define the asymptotic relative power (ARP) of test $\chi(\alpha)$, denoted by $\text{ARP}(\alpha, \beta, \theta)$, as its asymptotic power under alternative $f(x; k_{p_{\text{post}}}/\sqrt{n}, \theta)$, with $k_{p_{\text{post}}}$ chosen such that the asymptotic power of $\chi_{\text{post}}(\alpha)$ is $\beta$; that is, with $k_{p_{\text{post}}}$...
such that \( \beta_{\text{ppost}}(\alpha, k_{\text{ppost}}, \theta) = \beta \). Finally, the asymptotic relative efficiency \( \text{ARE} = \frac{\text{ARE}(\alpha, \beta, \theta)}{\text{ARE}(\alpha, \beta, \theta)} \) of candidate test \( \chi(\alpha) \) is the limit, as \( n_{\text{cand}} \rightarrow \infty \), of the ratio \( n_{\text{ppost}} / n_{\text{cand}} \), where \( n_{\text{ppost}} \) and \( n_{\text{cand}} \) are the sample sizes needed for tests \( \chi_{\text{ppost}}(\alpha) \) and candidate test \( \chi(\alpha) \) to both have power \( \beta \) under the alternative \( f(x; k, \sqrt{n}, \theta) \).

**Corollary 1.** For each candidate \( p(X) \) of Theorem 1, the actual asymptotic \( \alpha \)-level of the test \( \chi(\alpha) \) is

\[
\text{actual}(\alpha, \theta) = 1 - \Phi[z_{1-\alpha} \tau^{-1}(\theta)].
\]

If \( \nu_\phi(\theta) \neq 0 \), then the ARP and ARE are

\[
\text{ARP}(\alpha, \beta, \theta) = 1 - \Phi[-z_0 + z_{1-\alpha} (\tau^{-1}(\theta) - 1)]
\]

and

\[
\text{ARE}(\alpha, \beta, \theta) = \frac{(1 + z_{1-\alpha} / z_0)^2}{(1 + \tau^{-1}(\theta) z_{1-\alpha} / z_0)^2}.
\]

Note that neither \( \text{ARP}(\alpha, \beta, \theta) \) nor \( \text{ARE}(\alpha, \beta, \theta) \) depends on \( \text{NC}(\theta) \), and so these two quantities are the same for all local alternatives nesting the model \( f(x; \theta) \). When \( \tau^2(\theta) < 1/3 \), \( \text{ARP}(\alpha, 1-\alpha, \theta) < \alpha \), so the asymptotic local power of the test \( \chi(\alpha) \) is less than \( \alpha \), even though the test \( \chi_{\text{ppost}}(\alpha) \) has asymptotic power \( \beta = 1 - \alpha \).

**Remark 5: Discrepancy p Values.** In Section 5 we prove that the discrepancy \( p \) value and the posterior predictive \( p \) value are related as follows.

**Theorem 2.** Let \( t(X, \theta) \) be a discrepancy measure and, for a given fixed \( \theta_0 \), let \( t(X) = t(X, \theta_0) \). Then, under \( f(x; k, \sqrt{n}, \theta_0) \), \( p_{\text{dis}}(X) \) based on \( t(X, \theta^*) \) and \( p_{\text{post}}(X) \) based on \( t(X) \) have the same limiting distribution.

Moreover, if we redefine \( \sigma^2(\theta) \) to be the asymptotic variance of \( n^{1/2} t(X, \theta) \) under \( f(x; \theta) \), and \( \nu_\phi(\theta) = \lim_{n \rightarrow \infty} \nu_n(\theta, \theta^*; \theta) / \theta^* \nu_n(\theta, \theta^*; \theta) \), then \( \nu_n(\theta, \theta^*; \theta) \) is the asymptotic mean of \( n^{1/2} t(X, \theta) \) under \( f(x; \theta) \), then Theorem 1 holds for \( p_{\text{dis}}(X) \) with \( \tau^2_{\text{dis}} = \tau^2_{\text{post}} \).

3. **EXAMPLES**

In this section we use Theorems 1 and 2 and Corollary 1 to compare the asymptotic properties of our candidate \( p \) values in three examples. We first report results, and then give their derivation.

3.1 **Example A**

Suppose that \( X = (X_1, \ldots, X_n) \), with the \( X_i \)'s being mutually independent. Consider the null model \( X_i \sim N(\gamma_1, \gamma_2^2), \theta = (\gamma_1, \gamma_2^2), \) and \( v = (v_1, \ldots, v_n) \) a vector of known constants. Let the test statistic be \( t(X) = n^{-1} \sum X_i w_i = n^{-1} X w \), where \( w = (w_1, \ldots, w_n) \) is another vector of known constants. We assume \( v'v = w'w = 1 \) and \( v'1 = w'1 = 0 \), where \( n \) is the \( n \)-vector of \( 1 \)'s. Then the mean of \( t(X), E_\theta(t(X)) = \rho \gamma \), depends on \( \theta = (\gamma_1, \gamma_2^2) \) whenever \( \rho \neq 0 \), where \( \rho = n^{-1} \sum w_i v_i = n^{-1} w'v \). Figure 1 shows the asymptotic cdf’s actual(\( \alpha \)) = actual(\( \alpha, \theta \)) of various candidate \( p \) values under the null model, for several choices of \( \rho \). These depend on \( w \) and \( v \) only through \( \rho \), and they do not depend on the value of \( \theta \) generating the data. As expected, the plug-in and posterior predictive \( p \) values are conservative when \( \rho \neq 0 \), with the plug-in the less conservative. Indeed, both are converging to a point mass at \( 1/2 \) as the empirical correlation \( \rho \) of the regressors approaches 1. In contrast, \( p_{\text{plug}} \) is anticonservative. In particular, as \( \rho \rightarrow 1 \), actual(\( \alpha, \theta \)) converges to \( 1/2 \) for all \( \alpha < 1/2 \).

Let the alternative model be \( N(\psi w_i, \gamma v_i, \psi^2) \). Then the score, \( S_\psi(\theta) = n(t(X) - \rho \gamma) / \psi^2 \), is an affine transformation of \( t(X) \), and \( \chi_{\text{ppost}}(\alpha) \) is locally most powerful for testing \( \psi = 0 \). Figure 2 displays actual(\( \alpha, \theta \)), \( \text{ARP}(\alpha, \beta, \theta) \), and \( \text{ARE}(\alpha, \beta, \theta) \) as functions of \( \rho \), with \( \alpha = 0.05 \) and \( \beta = 0.80 \), for several candidate \( p \) values. These functions do not depend on the value of \( \theta = (\gamma_1, \gamma_2^2) \) generating the data. Figure 2(b) shows that when \( \rho \neq 0 \), the power of both the plug-in and posterior predictive tests are less than .80, with the latter the smaller; and as \( \rho \rightarrow 1 \), both power functions fall below the nominal \( \alpha \)-level of .05 as they converge to 0.

What is disturbing is that the performances of both the plug-in and posterior predictive \( p \) values and tests depend on

![Figure 1. Example A. Asymptotic cdfs of Candidate p Values, for (a) \( \rho = .5 \), (b) \( \rho = .9 \), and (c) \( \rho = .99 \).](https://example.com/figure1.png)

![Figure 2. Example B. Asymptotic cdfs of Candidate p Values, for (a) \( \rho = .5 \), (b) \( \rho = .9 \), and (c) \( \rho = .99 \).](https://example.com/figure2.png)
the “correlation” $\rho$ of the regressors, which is ancillary. For example, suppose that $p_{post} = .25$ was reported by the data analyst. If we know that $\rho = 0$, from inspection of Figures 1 and 2, then we could conclude that the data and model appear compatible, but would reach the opposite conclusion if $\rho = .99$. In our previous weather analogy, $\rho$ would be the temperature scale, $p_{post}$ the temperature, and the decision whether to use or discard the null model that to swimming or skiing. However, it is rare for $\rho$ to even be reported by the analyst, in which case reaching an appropriate decision would be impossible. Yet even if $\rho$ were reported, $p_{post}$ would not be interpretable by the consumer as either compatible or incompatible with the data without the benefit of the additional detailed mathematical analysis that we used to create the plots in Figures 1 and 2. Such analysis is beyond the capabilities of most consumers of statistical reports. It would be as if an American schoolchild was told that the temperature was 30°C but had never been taught the centigrade scale.

Consider now the discrepancy $p$ value based on the average score $t(X, \theta) = n^{-1}S_n(\theta) = (c^2 n)^{-1} \sum_{i=1}^n (X_i - \gamma_{n,i}) w_i$. Note that the mean of $t(X, \theta)$ under the null model, $X_i \sim N(\gamma_{n,i},c^2)$, is 0 for all $\theta$. Nonetheless, because $t(X) > t(x_{obs})$ iff $t(X, \theta) > t(x_{obs}, \theta)$, it follows that $p_{dis}(X) = p_{post}(X)$ with probability 1 under both the null and alternative models. Thus the curves for $p_{dis}(X)$ and $p_{post}(X)$ in Figures 1 and 2 are indistinguishable. And hence when $\rho \neq 0$, the discrepancy $p$ value is conservative, and the corresponding test is inefficient. This result can also be derived directly from Theorem 2 or Corollary 2 below.

The curves in Figures 1 and 2 remain unchanged under the submodel in which the variance $c^2$ is known and thus $\theta = \gamma$. Results of Bayarri and Berger (2000) imply that under this submodel, the cdf’s, ARPs and AREs in Figures 1 and 2 are exact at each sample size $n$, under the noninformative prior $\pi(\gamma) \propto 1$.

### 3.2 Example B

Stigler (1977) provided data on Simon Newcomb’s $n = 66$ measurements for estimating the speed of light, with each measurement $X_i$ recorded as a deviation from 24,800 nanoseconds. Gelman et al. (1995, sec. 2.2) modeled these data as $n \text{iid}$ draws from a $N(\mu, \sigma^2)$ distribution, with a noninformative uniform prior on $(\mu, \log \sigma)$. To look for incompatibility of the data with the $N(\mu, \sigma^2)$ model in the left tail of the distribution, they computed a posterior predictive $p$ value with $t(X) = \min \{X_i; i = 1, \ldots, n\}$, the first-order statistic (Gelman et al. 1995, p. 166). A reasonable alternative choice for $T = t(X)$ would be the empirical $q$th quantile of $X$, $T = Z_q \equiv \sup \{t; n^{-1} \sum_{i=1}^n I(X_i \leq t) < q\}$, for a small value of $q$ (say, $q = .05$), which we use in place of the first-order statistic because, in contrast to the latter, it is asymptotically normal and covered by our large-sample theory. The asymptotic mean of $T = Z_q$ under the normal null model is the population $q$th quantile of a $N(\mu, \sigma^2)$ distribution; that is, $z_q(\theta) = z_q(\mu, \sigma^2) = z_q \sigma + \mu$, with $z_q = z_q(0, 1)$, which depends on $\theta = (\mu, \sigma^2)$. Hence we expect both the plug-in and posterior predictive $p$ values to be asymptotically conservative. Figure 3 shows the curves actual(.05, $\theta$), $\text{ARF(.05, }\theta)$, and $\text{ARE(.05, }\theta)$ for our candidate $p$ values as a function of $q$, none of which depends on the value of $\theta$ generating the data. Note that as discussed in Corollary 1 we did not have to specify the alternative model $f(\chi; \psi, \theta)$ under which the ARP and ARE are calculated.

A natural discrepancy measure generalizing the test statistic $t(X) = Z_q$ is $t(X, \theta) = Z_q - z_q(\theta)$, the difference between the empirical and the true $q$th quantiles of the null model. Because $t(X) > t(x_{obs})$ iff $t(X, \theta) > t(x_{obs}, \theta), p_{dis}(X)$ and $p_{post}(X)$ are equal with probability 1, and so they have identical distributions.

### 3.3 Example C

Gelman et al. (1995, pp. 171–172) also analyzed Newcomb’s speed of light data using a discrepancy $p$ value based on

$$t(X, \theta) = |Z_{1-q} - \mu| - |\mu - Z_q|$$

with $q = .1$, to check whether or not the magnitude of skewness, as measured by $t(x_{obs}, \theta)$, was compatible with a $N(\mu, \sigma^2)$ distribution. Note that under the null model $N(\mu, \sigma^2), t(X, \theta)$ has asymptotic (and exact) mean 0 for all
θ = (μ, c^2), because the X_i have a symmetric distribution centered at μ. A natural test statistic related to the discrepancy t(X, θ) is t(X) = Z_{-1,q} + Z_q because, on a set with probability going to 1, t(X, θ) = Z_{1-q} + Z_q - 2μ under the null model and any local alternative. On this set, t(X) > t(x_{obs}) if t(X, θ) > t(x_{obs}, θ), and so p_{dis}(X) based on t(X, θ), and p_{post}(X) based on t(X), have the same asymptotic distribution. Figure 4 shows actual(0.05, θ), ARP(0.05, 8, θ), and ARE(0.05, 8, θ) for our candidate p values based on t(X), as functions of q, none of which depends on the value of θ generating the data.

3.4 Derivation of the Results

We now show how the quantities σ(θ), Ω(θ), and NC(θ) were obtained for the test statistics and discrepancies in Examples A–C.

Example A. Here θ' = (θ_1, θ_2) = (γ, c^2), X_i ~ N(ψ w_i + γ v_i, c^2), and t(X) = n^{-1} X'w. We assume that at each sample size n, the vectors of constants (v, w) = (v_n, w_n) are chosen such that n^{-1} v' v, n^{-1} w' w, and ρ = n^{-1} w' v do not depend on n. Then the asymptotic mean of t(X) is ν_n(ψ, θ) = ψ + θ_1 ρ. Also, ν_θ(θ) = (θ, 0)', ν_θ(θ) = n^{-1} σ^2(θ) = θ_2, i_θ(θ) = diag(θ_2^{-1}, b_0 W), i_θ(θ)' = (θ_2^{-1} ρ, 0), Ω(θ) = ρ^2 θ_2, and ω(θ) = 1 - ρ^2. For the discrepancy, t(X, θ) = n^{-1} S_{θ}(θ) = n^{-1} θ_2^{-1} (W - θ_1 v w)' ψ_θ(θ) = θ_2^{-1} θ_2^{-1} \psi (θ_2^{-1} ψ + (θ_2^{-1} - θ_1) ρ). Hence ν_θ(θ) = (θ_2^{-1} θ_2^{-1}, θ_2^{-1} θ_2^{-1}, θ_2^{-1} θ_2^{-1}, Ω(θ) = θ_2^{-1} θ_2^{-1}, and ω(θ) = (1 - ρ^2) θ_2^{-1}.

Example B. Here t(X) = Z_q, and under f(x; θ), X_i ~ N(μ, c^2) with θ' = (θ_1, θ_2) = (μ, c^2). Then ν_n(0, θ) = θ_2^{-1} z_q + θ_1, and hence ν_θ(θ) = (1, θ_2^{-1} z_q/2, θ_2^{-1} θ_2^{-1} z_q/2), Ω(θ) = θ_2^{-1} (1 + z_q^2/2). Further, it is well known that if the X_i are iid under any law f(x; θ), then

\[ n^{-1/2} (Z_q - z_q(θ)) = n^{-1/2} - f[z_q(θ); θ]^{-1} \sum_i I(X_i < z_q(θ)) + o_P(1). \]

Hence \( σ^2(θ) = f(z_q(θ); θ)^{-2} \var{I(X_i < z_q(θ))} \), which, for our N(θ_1, θ_2) model, evaluates to

\[ σ^2(θ) = θ_2 \phi^2(z_q(q1 - q), \phi \text{ is the standard normal density. Further, under} f(x; k_n/\sqrt{n}, θ), it follows from Theorem 3 that \]

\[ i_θ(θ) = \]
Example C. In this example, $t(X) = Z_q + \hat{Z}_{1-q}$, where $X_i \sim N(\mu, \sigma^2)$, $	heta' = (\theta_1, \theta_2)$, and so as in Example B, $\nu_0(0, \theta')$ in (3a), $\nu_0(0, \theta')$ in (4), and $\nu_0(0, \theta')$ in (5). Further, it follows from (4) that $\sigma^2(\theta) = \text{var}_{\theta}(-f(z_q(\theta); \theta))^{-1}I(X_i < z_q(\theta)) - f(z_{1-q}(\theta); \theta))^{-1}I(X_i < z_{1-q}(\theta))$, which for our normal model evaluates to $2q^2 - 2$. All the relevant foregoing quantities are the same for the discrepancy measure $t(X, \theta) = |\hat{Z}_{1-q} - \mu| - |\mu - \hat{Z}_q|$ as for $t(X) = Z_q + \hat{Z}_{1-q}$, but with $\nu_0_n(0, \theta'; \theta) = 2(\theta_1 + \theta_2)$.

4. ALTERNATIVES TO $p_{\text{pred}}$ AND $p_{\text{post}}$

It follows from Theorems 1 and 2 and Corollary 1 that from an asymptotic frequentist viewpoint, $p_{\text{pred}}$ and $p_{\text{post}}$ are preferred to our other candidate $p$ values when the asymptotic mean of $t(X)$ depends on $\theta$. This result still leaves open the question of why we should want to use a test statistic with nonconstant asymptotic mean. Bayarri and Berger (2000) argued that this is desirable because it does not restrict the choice of possible measures $t(X)$ of departure from the null model; the preferred, or intuitive, choice may happen to have a nonconstant asymptotic mean. So suppose that our choice $t(X)$ satisfies (3a)–(3c), with $\nu_0(0, \theta') \neq 0$. Now $p_{\text{pred}}$ and $p_{\text{post}}$ are sometimes difficult to compute (Bayarri and Berger 1999, 2000; Pauler 1999), and alternative approaches might be useful. We consider two; the first is to replace $t(X)$ with a closely related test statistic that has a constant asymptotic mean; the second approach is to adjust (i.e., calibrate) those candidate $p$ values with nonuniform asymptotic distributions. We also describe how to modify a discrepancy measure so that the discrepancy $p$-value is asymptotically uniform.

4.1 Modifications of $t(X)$ and $t(X, \theta)$

One alternative to computing $p_{\text{pred}}$ or $p_{\text{post}}$ based on $t(X)$ is to compute $p_{\text{plug}}$ or $p_{\text{post}}$ based on the statistic $\hat{t}(X) = t(X) - \nu_0(\theta)$, where $\nu_0(\theta) \equiv \nu_0(0, \theta')$, because, by a Taylor expansion, $\hat{t}(X)$ is asymptotically normal with constant asymptotic mean 0 and asymptotic variance

$$c^2(\theta) = \sigma^2(\theta) - \nu_0(\theta)'i\sigma^2_0(\theta)\nu_0(\theta)$$

under the null model. It then follows from Theorem 1 that $p_{\text{plug}}(X)$ and $p_{\text{post}}(X)$ calculated using $\hat{t}(X)$ are asymptotic frequentist $p$ values with limited distribution under $f(x; k_1/n, \theta)$, equal to that of $p_{\text{post}}(X)$ and $p_{\text{pred}}(X)$, because $NC(\theta)$ is the same for $\hat{t}(X)$ and for $t(X)$. But because in general $\hat{t}(X)$ is not asymptotically pivotal [as its asymptotic variance $c^2(\theta)$ depends on $\theta$], $p_{\text{plug}}$ and $p_{\text{post}}$ often will be calculated by simulation using the fact that, for example,

$$p_{\text{plug}} = \lim_{K \to \infty} K^{-1} \sum_{k=1}^{K} I(\hat{t}(X^{(k)}) > \hat{t}(X_{\text{obs}})),$$

where $X^{(k)} = (X_1^{(k)}, \ldots, X_n^{(k)})$ are $K$ independent draws from $f(X; \hat{\theta}_{\text{obs}})$, and $\hat{t}(X_{\text{obs}}) = t(X_{\text{obs}}) - \nu_0(\hat{\theta}_{\text{obs}})$. A potential drawback of this approach is that to evaluate $t(X^{(k)}) = t(X^{(k)}) - \nu_0(\hat{\theta}(k))$, the maximizer $\hat{\theta}(k)$ of $t(X^{(k)}; \theta)$ must be recomputed for each simulated dataset $X^{(k)}$. The computational difficulties could be overcome by substituting for $\hat{\theta}(k)$ a single Newton-step estimator starting from the original MLE $\hat{\theta}$. Similarly, in the case of $p_{\text{post}}$, the posterior distribution of $\theta$ must be recomputed for each dataset $X^{(k)}$, which also may be computationally impractical. Again, the computational difficulties could be overcome by substituting an easy-to-compute normal approximation to the posterior.

To avoid having to recompute (either exactly or approximately) the MLE or the posterior density of $\theta$ for each simulated dataset, two additional approaches may be considered, both of which give $p$ values that are asymptotically equivalent to $p_{\text{post}}(X)$ under both the null model and local alternatives. The first approach is to replace $\hat{t}(X)$ by the asymptotically pivotal $\text{N}(0,1)$ random variable $\bar{t}(X) = t(X)/c(\theta)$. We then obtain an asymptotic frequentist $p$ value $p_{\text{pivot}}$ based on $\bar{t}(X)$ by using $m(t) = \text{N}(0,1)$ in (1); specifically, $p_{\text{pivot}} = 1 - \Phi[\bar{t}(X_{\text{obs}})]$. The second alternative is to calculate a discrepancy $p$ value based on the discrepancy $\hat{t}(X, \theta) = t(X, \theta) - \nu_0(\theta) - \nu_0(\theta)'i\sigma^2_0(\theta)n^{-1}S_\theta(\theta)$. Indeed given any discrepancy $t(X, \theta)$ with $E_\theta[\hat{t}(X, \theta)] = 0$ the discrepancy $p$-value $\hat{X}_{\text{dis}}$ based on the modified discrepancy

$$\hat{t}(X, \theta) = t(X, \theta) - \nu_0(\theta)'i\sigma^2_0(\theta)n^{-1}S_\theta(\theta)$$

is, by Theorem 3, uncorrelated with $S_\theta(\theta)$ and thus an asymptotic frequentist $p$ value with $\nu_0(\theta)$ as defined in Theorem 2. As emphasized by Meng (1994), neither the MLE nor the posterior distribution of $\theta$ needs to be recomputed when calculating $p_{\text{dis}}$ by simulation.

A drawback of these latter approaches is that they can require computation or estimation of $\sigma^2(\theta), \nu_0(\theta), \nu_0(\theta)'i\sigma^2_0(\theta), \nu_0(\theta)'i\sigma^2_0(\theta)n^{-1}S_\theta(\theta)$, and/or $\nu_0(\theta)'i\sigma^2_0(\theta)n^{-1}S_\theta(\theta)$, which may be computationally difficult. But an important advantage of the last approach is that if we take $t(X, \theta)$ equal to $n^{-1}S_\theta(\theta)$, then $t(X, \theta)$ becomes the “efficient score” discrepancy

$$\hat{t}(X, \theta) = n^{-1}S_\theta(\theta) - \nu_0(\theta)'i\sigma^2_0(\theta)n^{-1}S_\theta(\theta)$$

and the test $\hat{X}_{\text{dis}}$ based on $p_{\text{dis}}$ is a locally most powerful asymptotic $\alpha$-level test of $\psi = 0$.

4.2 Adjusted $p$ Values

Another class of alternatives to $p_{\text{pred}}$ or $p_{\text{post}}$ are the adjusted (i.e., calibrated) $p$ values $p_{\text{adj}}$, $p_{\text{adj}}(X_{\text{obs}})$, and $p_{\text{adj}}(X_{\text{obs}})$, where for any candidate $p$ value $p(X)$ with observed value $p = p(X_{\text{obs}})$,

$$p_{\text{adj}} = p_{\text{adj}}(X_{\text{obs}}) = F_p(p(X_{\text{obs}})).$$
where \( F_p(X; u; \theta) \) is the cdf of \( p(X) \) when \( X \sim f(x; \theta) \) (Davison and Hinkley 1997, p. 132). Beran (1987) introduced adjusted \( p \) values in bootstrap context as a means of calibrating asymptotic uniform \( p \) values so that they become second-order correct, whereas we use them here to render asymptotically nonuniform candidate \( p \) values first-order correct. When estimated by simulation, \( p_{plug,adj} \) is precisely the "double parametric bootstrap" \( p \) value of Beran (1987) and Davison and Hinkley (1997, p. 177); the computation burden of such simulations can perhaps be alleviated by recycling (Newton and Geyer 1994). A double bootstrap simulation may be avoided because the representation \( p(X) = 1 - \Phi(Q) + o_p(1) \), under \( H_0 \), of Theorem 1 implies that \( p_{adj,anna}(X) = 1 - \Phi(\frac{1}{\sqrt{n}}(Z_{1,p} - p)) \) is a simple analytic approximation to \( p_{adj} \). It is easy to show that for any of our candidate \( p \) values, both \( p_{adj}(X) \) and \( p_{adj,anna}(X) \) have the same limiting distribution as \( p_{post}(X) \) and \( p_{cpred}(X) \) under \( f(x; \theta_n, \sqrt{n}) \).

Given a candidate \( p \) value defined via (1) with reference density \( m(t|x_{obs}) = m_T(t|x_{obs}) \), define the adjusted reference density evaluated at \( t_{obs} \) to be

\[
m_{adj}(t_{obs}|x_{obs}) = f_p(X)[m_T(t|x_{obs})],
\]

where \( f_p(X)(u; \theta) = \partial F_p(X)(u; \theta)/\partial \theta \) and \( p = p(t_{obs}) \). Then \( p_{adj} \) is obtained via (1) with \( m(t) = m_{adj}(t|x_{obs}) \). These densities \( m_{adj}(t|x_{obs}) \) are additional solutions to the frequentist math problem of Section 1.3.

To summarize this section, we have proposed alternative candidate \( p \) values that have the same limiting distribution as \( p_{post}(X) \) and \( p_{cpred}(X) \) under the null model and local alternatives. Thus the corresponding nominal \( \alpha \)-level tests \( \chi_{post}(\alpha) \) and \( \chi_{cpred}(\alpha) \) based on \( t(X) \); that is, ARP(\( \alpha, \beta, \theta, \theta^* \)) = \( \beta \) and ARE(\( \alpha, \beta, \theta, \theta^* \)) = 1. The choice as to which \( p \) value to use in practice will depend both on the relative ease with which each can be calculated and on their second-order asymptotic and small-sample nonasymptotic distributional properties. These topics are beyond the scope of this article, although example 2.1 of Berger and Bayarri (2000) suggests that \( p_{post} \) and \( p_{cpred} \) will be preferred to \( p_{post} \) and \( p_{plug} \) in small samples, even when the mean of \( t(X) \) does not depend on \( \theta \). When one has a specific alternative model in mind, a major advantage of the discrepancy \( p \) value based on the efficient score discrepancy (8) is that it is guaranteed to be locally most powerful whatever the value of \( \theta \) generating the data.

5. PROOFS OF THEOREMS 1 AND 2 AND LEMMA 1

The first theorem in this section, Theorem 3, derives the asymptotic expansion

\[
p(X) = 1 - \Phi(Q) + o_p(1)
\]

for a particular random variable \( Q \). Theorem 3 and corollary 3 allow us to deduce that \( Q \) has the \( N(k\mu(\theta), \tau^2(\theta)) \) distribution specified in Theorems 1 and 2. All of our candidate \( p \) values, including \( p_{dis} \) but excluding \( p_{cpred} \), can be written as

\[
p = p(x_{obs}) = \int_0^\infty \text{Pr}[t(X, \theta) > t(x_{obs}, \theta; \theta) | d\theta|x_{obs}] (10)
\]

By taking \( (X, \theta) = (X, \theta) \), we obtain the nondiscrepancy \( p \) values. Here \( \pi(d\theta|x_{obs}) = \pi(\theta|x_{obs})d\theta \) is given in Table 1 for \( p_{post} \) and \( p_{post} \). We take \( \pi_{dis}(x_{obs}) = \pi_{post}(x_{obs}) \). For \( p_{plug} \) and \( p_{plug} \), we take \( \pi(d\theta|x_{obs}) \) to be the degenerate distribution that places all of its mass on \( \theta_{obs} \) and \( \hat{\theta}_{MLE,obs} \). When we take \( \pi(\theta|x_{obs}) \) in (10) to be \( \pi_{pred}(\theta|x_{obs}) \), we obtain a new \( p \) value, the approximate conditional predictive \( p \) value, \( p_{acpred} \), that we show in Lemma 2 is asymptotically equivalent to \( p_{pred} \). Hence it will suffice to prove Theorem 1 for \( p_{acpred} \) in lieu of \( p_{pred} \). It will be useful to have an expression for the random \( p \) value \( p(X) \); specifically,

\[
p(X) = \int_0^\infty \text{Pr}[t(X^new, \theta) > t(X, \theta)| \theta; \pi(d\theta|X)] (11)
\]

where \( X^new \) is drawn from \( f(x; \theta) \) independently of \( X \).

We prove Theorems 1 and 2 together. The asymptotic normality of \( t(X, \theta) \) remains a basic assumption. Now if \( t(X, \theta) \) is allowed to depend on an additional parameter \( \theta \), then it is natural to allow the asymptotic mean and variance of \( t(X, \theta^*) \) under \( \theta \) also depend on \( \theta^* \). Thus we assume (3), but with \( \nu_n(\psi, \theta) \) replaced by \( \nu_n(\psi, \theta^*; \theta^*) \) and \( \sigma^2(\theta) \) replaced by \( \sigma^2(\theta; \theta^*) \). These quantities are the asymptotic mean and variance of \( t(X, \theta^*) \) under \( \theta \).

Actually, because the measures \( \pi(d\theta|X) \) concentrate on small (shrinking) neighborhoods of \( \theta_{obs} \), the dependence of \( t(X, \theta) \) on \( \theta \) does not play a major role, as soon as we assume some natural continuity in this parameter. Specifically, we assume that for every random sequence \( \theta_n = \theta_0 + O_p(n^{-1/2}), \)

\[
n_{1/2} \{t(X, \theta_n) - \nu_n(0, \theta_0; \theta_0)\} / \sigma(\theta_0; \theta_0)
\]

\[
- n_{1/2} \text{Pr}(t(X, \theta_0) - \nu_n(0, \theta_0; \theta_0) | \sigma(\theta_0; \theta_0) = 0, \theta_0 = \theta_0; \theta_n = \theta_n) \}
\]

which we take as trivially true when \( t(X, \theta) = t(X) \). Further, we assume that for some \( p \)-vector–valued function \( \theta^*(X) \) on the sample space and some \( p \times p \) matrix \( \Sigma(\theta_0) \),

\[
||\pi(\cdot|X) - N(\theta^*(X), \Sigma(\theta_0)/n)|| = O_p(1)
\]

and

\[
n_{1/2} \{\theta^*(X) - \theta_0\} = O_p(1).
\]

Here \( || \cdot || \) is the total variation distance between two distributions \( P \) and \( Q \); \( ||P - Q|| = 2 \sup_{B \subseteq \mathbb{R}^p \cap \Theta} \{P(B) - Q(B)\} \), with the supremum taken over all Borel sets in \( \mathbb{R}^p \cap \Theta \). Let \( \sigma(\theta_0) = \sigma(\theta_0; \theta_0), \nu_0(\theta_0) = \lim_{n \to \infty} \partial \nu_n(0, \theta_0; \theta_0)/\partial \theta_0 = \theta_0 \) and \( \nu_0(\theta) = \lim_{n \to \infty} \partial \nu_n(\psi, \theta; \theta)/\partial \psi |_{\psi = 0} \). Then, under conditions discussed later, (13) and (14) hold for the choices of \( \theta^*(X) \) and \( \Sigma(\theta_0) \) specified in Table 2.

Theorem 3. Suppose that (12)–(14) and (3) hold with \( t(X, \theta) \) and \( \nu_n(0, \theta; \theta) \) replacing \( t(X) \) and \( \nu_n(0, \theta) \). Then,
under $X \sim f(x; \theta_0)$,

$$p(X) = 1 - \Phi(Q) + o_p(1),$$

(15)

where

$$n^{1/2}\{t(X, \theta_0) - \nu_n(0, \theta_0; \theta_0)\}$$

$$Q = \frac{n^{1/2}\{t(X, \theta_0) - \nu_n(0, \theta_0; \theta_0)\}}{\sigma^2(\theta_0) + \nu_\theta(\theta_0)' \Sigma(\theta_0) \nu_\theta(\theta_0)]^{1/2}.}

(16)

As this theorem is our main result, we give an informal proof that emphasizes the main idea. We give a formal proof in Appendix A.

**Informal Proof of Theorem 3.** Note that $t(X^{\text{new}}, \theta) > t(X, \theta)$ is algebraically equivalent to

$$n^{1/2}\{t(X, \theta_0) - \nu_n(0, \theta_0; \theta_0)\}$$

$$- n^{1/2}\{\nu_n(0, \theta_0) - \nu_n(0, \theta_0; \theta_0)\}$$

$$> n^{1/2}\{t(X, \theta) - \nu_n(0, \theta_0; \theta)\}$$

$$- n^{1/2}\{\nu_n(0, \theta_0; \theta) - \nu_n(0, \theta_0; \theta_0)\}.}

(17)

Now, by (13) and (14), we can asymptotically ignore all $\theta$ that are not within $O(n^{-1/2})$ of $\theta_0$. Hence by (12), the left side of (17) is approximately

$$n^{1/2}\{t(X^{\text{new}}, \theta_0) - \nu_n(0, \theta_0; \theta_0)\}$$

$$- n^{1/2}\{\nu_n(0, \theta_0; \theta) - \nu_n(0, \theta_0; \theta_0)\}$$

$$= n^{1/2}\{t(X^{\text{new}}, \theta_0) - \nu_n(0, \theta_0; \theta)\}.}

By (12) and the differentiability assumption in (3b), the right side of (17) is approximately

$$\sqrt{n}\{t(X, \theta_0) - \nu_n(0, \theta_0; \theta_0)\} - \nu_\theta(\theta_0)' \sqrt{n}(\theta - \theta_0).}

(18)

Hence the event $t(X^{\text{new}}, \theta) > t(X, \theta)$ is approximately equivalent to the event

$$\sqrt{n}\{t(X, \theta_0) - \nu_n(0, \theta_0; \theta_0)\} + \nu_\theta(\theta_0)' \sqrt{n}(\theta - \theta_0)$$

$$> \sqrt{n}\{t(X, \theta_0) - \nu_n(0, \theta_0; \theta_0)\}.}

(19)

Now conditional on $(X, \theta)$, by (3a), the first term on the left side of (19) is approximately $N(0, \sigma^2(\theta))$, and, given $X$, by (13), the second term on the left side of (19) is converging to a $N(\nu_\theta(\theta_0)'^\prime \Sigma(\theta_0) \nu_\theta(\theta_0))$ distribution. Further, because the first term has mean 0 conditional on $(X, \theta)$, the two terms are conditionally uncorrelated given $X$. Hence, given $X$, the left side of (19) is asymptotically

$$N(\nu_\theta(\theta_0)'^\prime \sqrt{n}(\theta - \theta_0), \sigma^2(\theta_0) + \nu_\theta(\theta_0)'^\prime \Sigma(\theta_0) \nu_\theta(\theta_0)).}

(20)

It follows that the conditional probability, given $X$, of the event $t(X^{\text{new}}, \theta) > t(X, \theta)$ is approximately $1 - \Phi(Q)$ with $Q$ as in (16), which concludes the proof.

Under the assumption that the distribution of $X$ under $(\psi_n, \theta_0)$ and $(\theta, \theta_0)$ are contiguous (which will quite generally be the case), the expansion of Theorem 3 is also valid under $X \sim f(x; \psi_n, \theta_0)$. It remains to show that the distribution of $Q$ in (16) converges to the $N(1, \tau^2(\theta_0))$ distribution given in Theorems 1 and 2. For this, we need the joint distribution of $(t(X, \theta_0)$ and $\theta^1(X)$.

In all of the examples that we consider, we will have for given “influence functions” $B_i(\theta; \theta)$ that for all $\theta$ in a neighborhood of $\theta_0$, with $X \sim f(x; \theta_0)$,

$$n^{1/2}[t(X, \theta) - \nu_n(0, \theta_0; \theta)]$$

$$= n^{1/2} \sum_i B_i(\theta_0; \theta) + o_P(1)

(21)

for some mean 0 $B_i(\theta_0; \theta) = b_i(X, \theta_0, \theta)$,

$$n^{1/2}(\theta - \theta_0) = \nu_\theta(\theta_0)n^{1/2}S_{\theta}(\theta_0) + o_P(1),

(22)

$$n^{1/2}(\theta_{\text{MLE}} - \theta_0)$$

$$= \nu_\theta(\theta_0)(\theta_0) - n^{1/2}S_{\theta}(\theta_0)) + o_P(1)

(23)

where

$$S_{\theta}(\theta_0) \equiv S_{\theta}(\theta_0) - \nu_\theta(\theta_0) \sigma^{-2}(\theta_0)n^{1/2}T_{\text{std}}(\theta_0),

(24)

$$T_{\text{std}}(\theta_0) = n^{1/2}\{t(X) - \nu_n(0, \theta_0; \theta_0)\},

and $\nu_\theta(\theta_0)$ is defined in Table 2. Equation (21) is the usual asymptotically linear expansion of an asymptotically normal statistic, showing that it behaves like a sample average, and (22) is the usual expansion of the MLE. Equation (23) is a conditional version of (22), which we discuss further later. Given the foregoing expansions, the joint limit distribution of $t(X, \theta)$ and $\theta^1$ under the null hypothesis $(\psi_n = 0, \theta_0)$ follows immediately from the multivariate central limit theorem (CLT) (where we need to assume the Lindeberg–Feller conditions to take care of the possible non-iid character of the terms in the sums). Note that the right sides of (22) and (23) are also sums. The expansions (22) and (23) imply that the asymptotic variance of the MLE and conditional MLE are $\nu_\theta(\theta_0)$ and $\nu_\theta(\theta_0)^{-1}$. To obtain the limit distribution under alternatives $(\psi_n, \theta_0)$, we make the further assumption
that as $n \to \infty$, for $(k'_n, k^*_n)' \to (k', k^*)' \equiv h$, 
$$
\log f(X; k_n, \sqrt{n}, \theta_0 + k^*_n/\sqrt{n}) 
= h'n^{-1/2}(S_{\varphi}(\theta_0), S_{\theta}(\theta_0))' - \frac{1}{2} h'^*i(\theta_0)h + o_P(1) 
$$
(25)
and
$$
n^{-1/2}(S_{\varphi}(\theta_0), S_{\theta}(\theta_0))' \quad \Rightarrow \quad \mathcal{N}(0, i(\theta_0)),
$$
(26)
where
$$
i(\theta_0) = \begin{pmatrix}
i_{\varphi}(\theta_0) 
\quad i_{\varphi}(\theta_0)
\end{pmatrix},
$$
and we assume the sum on the right side of (21) satisfies the Lindeberg condition. Equations (25)–(26) imply that the model $f(x; \psi, \theta)$ is locally asymptotically normal (LAN) at $(0, \theta_0)$. Therefore, we can apply LeCam’s third lemma to obtain the desired result (van der Vaart 1998). Specifically, we obtain the following theorem.

**Theorem 4.** Given $f(x, \theta)$ and model (2a)–(2b), suppose that both (21)–(26) and the assumptions of Theorem 3 hold. Then the following obtain:

a. $\hat{\psi}(\theta_0) = \text{cov}_{\theta_0}^{\varphi}(T_{\text{std}}(\theta_0), n^{-1/2}S_{\varphi}(\theta_0))$, 

$$
i(\theta_0) = \text{cov}_{\theta_0}^{\varphi}(T_{\text{std}}(\theta_0), n^{-1/2}S_{\theta}(\theta_0)),
$$
(27)
where $\text{cov}_{\theta_0}^{\varphi}(S_1, S_2)$ denotes the asymptotic covariance of $S_1$ and $S_2$ under $(\psi = 0, \theta_0)$.

b. When $X \sim f(x; k_n/\sqrt{n}, \theta_0)$, 

$$
(T_{\text{std}}(\theta_0), \hat{\psi}(\theta_0)'n^{1/2}(\hat{\theta} - \theta_0)) \text{ converges to a normal distribution with mean } \begin{pmatrix} k(\hat{\psi}(\theta_0), \hat{\psi}(\theta_0))'i_{\theta}(\theta_0) 
\quad i_{\theta}(\theta_0)' 
\end{pmatrix}
$$
and covariance matrix
$$
\begin{pmatrix}
\sigma^2(\theta_0) & 0 
0 & \Omega(\theta_0)
\end{pmatrix},
$$
(28)
and thus $T_{\text{std}}(\theta_0) - \hat{\psi}(\theta_0)'n^{1/2}(\hat{\theta} - \theta_0)$ converges to 

$$
\mathcal{N}(k_{\omega_{c}}(\theta_0), \sigma^2(\theta_0) - \Omega(\theta_0))
$$
(29)

where

$$
\text{cov}_{\theta_0}^{\varphi}(\theta(\theta_0)) = \text{cov}_{\theta_0}^{\varphi}[n^{-1/2}S_{\varphi}(\theta_0), n^{-1/2}S_{\theta}(\theta_0)]
$$
and $\Omega(\theta_0) = i_{\theta}(\theta_0)'i_{\theta}(\theta_0)\varphi(\theta_0)$. Thus $T_{\text{std}}(\theta_0) - \hat{\psi}(\theta_0)'n^{1/2}(\hat{\theta}_{\text{MLE}} - \theta_0)$ converges to a normal distribution with mean $\begin{pmatrix} k(\hat{\psi}(\theta_0), \hat{\psi}(\theta_0))'i_{\theta}(\theta_0) 
\quad i_{\theta}(\theta_0)' 
\end{pmatrix}$ and covariance matrix

$$
\begin{pmatrix}
\sigma^2(\theta_0) & 0 
0 & \Omega(\theta_0)
\end{pmatrix},
$$
Remark 6. A critical observation required in applying LeCam’s third lemma to obtain the results in Theorem 4(b) is that $0 = \text{cov}_{\theta_0}^{\varphi}(T_{\text{std}}(\theta_0), n^{-1/2}S_{\theta}(\theta_0))$, which is a consequence of the fact that, by (27), $n^{-1/2}S_{\theta}(\theta_0)$ is the residual from the asymptotic least squares projection of $n^{-1/2}S_{\varphi}(\theta_0)$ on the normalized test statistic $T_{\text{std}}(\theta_0)$, because

$$
n^{-1/2}S_{\varphi}(\theta_0) = n^{-1/2}S_{\theta}(\theta_0) - \text{cov}_{\theta_0}^{\varphi}(S_{\varphi}(\theta_0), T_{\text{std}}(\theta_0))
\times \{\text{var}_{\theta_0}(T_{\text{std}}(\theta_0))\}^{-1}T_{\text{std}}(\theta_0),
$$
As shown in Corollary 3, Theorems 1 and 2 follow from Theorems 3 and 4, provided that we can establish that (13) and (14) hold for the entries in Table 2. The first row of the table merely asserts asymptotic normality of the MLE and hence is valid under the usual conditions. The second row is the assertion of the Bernstein–Von Mises theorem and hence is valid under even weaker conditions. Primitive conditions to ensure the validity of the last three rows of Table 2 are less easily available. We do not provide such a set of conditions, but rather offer below an informal argument as to why these rows are expected to be correct.

**Corollary 3.** Under the assumptions of Theorem 4, if (13) and (14) hold for the entries in Table 2, then, with $p_{c_{\text{pred}}}$ substituted for $p_{c_{\text{pred}}}$ the $p$ values considered in Theorems 1 and 2 have the asymptotic expansions given therein.

**Proof of Corollary 3.** For the plug-in, posterior predictive, and discrepancy $p$ values, the proof is immediate from Theorems 3 and 4. [Note that if the off-diagonal entries in (28) were 0 rather than $\Omega(\theta_0)$, then, even if the variance of $\hat{\psi}(\theta_0)'n^{1/2}(\hat{\theta} - \theta_0)$ had remained nonzero, $Q$ for the partial posterior and discrepancy would have had variance 1, and the associated $p$ values would not be conservative. But the covariance $\Omega(\theta_0)$ is in fact nonzero whenever $\hat{\psi}(\theta_0)$ is nonzero.] Furthermore, it is immediate from Theorems 3 and 4 that expansion (16) holds with $Q \sim N(k_{\omega_{c}}(\theta_0))/\{\sigma^2(\theta_0) + \Omega(\theta_0)\}^{1/2}$, 1) for $p_{c_{\text{pred}}}(X)$ and $p_{c_{\text{pred}}}(X)$ and $Q \sim N(k_{\omega_{c}}(\theta_0)/\sigma(\theta_0), \{\sigma^2(\theta_0)/\Omega(\theta_0)\}$ for $p_{c_{\text{pred}}}(X)$. But some algebra shows that

$$
\{\sigma^2(\theta_0) + \Omega(\theta_0)\}/\sigma^2(\theta_0) = \sigma^2(\theta_0)/\{\sigma^2(\theta_0) - \Omega(\theta_0)\}
$$
and $\omega_{c}(\theta_0)/\{\sigma^2(\theta_0) + \Omega(\theta_0)\}^{1/2} = NC(\theta_0)$, which proves the corollary.

To complete the proof of Theorem 1, it only remains to show that $p_{c_{\text{pred}}}(X)$ and $p_{c_{\text{pred}}}(X)$ have the same limiting distribution. The key observation is that, as discussed earlier, $\hat{\theta}_{\text{MLE}}$ and $t(X)$ are asymptotically uncorrelated, so that in large samples the conditional distribution given $\theta_{\text{MLE}}$ and unconditional distribution of $t(X)$ are the same. Formally, we have the following lemma, whose proof is similar to aspects of the proof of Theorem 3 given in Appendix A and thus is omitted.

**Lemma 2.** If for every $c$,

$$
\sup_{|t - \theta_0| \leq c/\sqrt{n}} \sup_{t \leq t} \Pr\{t(X) \leq t|\hat{\theta}_{\text{MLE}}; \theta\} \quad \text{and} \quad -\Pr\{t(X) \leq t; \theta\} \quad \text{then} \quad p_{c_{\text{pred}}}(X) \text{ and } p_{c_{\text{pred}}}(X) \text{ have the same limiting distribution under } f(x; k_n/\sqrt{n}, \theta_0).
$$

The supposition of Lemma 2 would need to be checked on a case-by-case basis, as general regularity conditions for it are not known.

**Conditional Inference.** If our given statistic $T = t(X)$ satisfies (3), then we would expect the marginal model $f_T(t; \theta)$ to be LAN. That is, under $T \sim$
\[ f(t; \theta_0) \equiv f(t; 0, \theta_0), \]
\[ \frac{f(T; k_n^*/\sqrt{n}, \theta_0 + k_n^*/\sqrt{n})}{f(T; 0, \theta_0)} = h'\psi(\theta_0)^2/\sigma^2(\theta_0) + O_P(1). \]  
(30)

Together with the similar expansion (25) for the unconditional model for \( X \), we obtain, on noting \( f(X) = f(X|T)f(T) \), that
\[ \frac{f(X|T; k_n^*/\sqrt{n}, \theta_0 + k_n^*/\sqrt{n})}{f(X|T; 0, \theta_0)} = h'\psi(\theta_0)^2/\sigma^2(\theta_0) + O_P(1), \]  
(31)

where
\[ S_c(\theta_0) = (S_{\psi}(\theta_0), S_{\theta_0}(\theta_0))', \]
\[ S_{\psi}(\theta_0) = \psi(\theta_0) - \frac{1}{2} \frac{h'\psi(\theta_0)h}{\psi(\theta_0)} \]

is the asymptotic covariance matrix of \( n^{-1/2}S_c(\theta_0) \). The vector \( S_c(\theta_0) \) is referred to as the conditional score because it is the linear term in the expansion of the conditional density, and \( L(\theta_0) \) is the conditional information matrix. As noted earlier, \( n^{-1/2}S_c(\theta_0) \) is the residual from the asymptotic least squares projection of \( n^{-1/2}S(\theta_0) \) on the normalized test statistic \( T_{\text{null}}(\theta_0) \).

Note that if \( n^{-1/2}\{T - \nu_0(k_n^*/\sqrt{n}, \theta_0 + k_n^*/\sqrt{n}) \} \) were exactly distributed \( N(0, \psi^2(\theta_0)) \) under \( f_T(t; k_n^*/\sqrt{n}, \theta_0 + k_n^*/\sqrt{n}) \) with \( \nu_0(k_n^*/\sqrt{n}, \theta_0 + k_n^*/\sqrt{n}) = \nu_0(0, \theta_0) \) and \( \nu_0(0, \theta_0) \) then (30) would be exactly true without the \( O_P(1) \) term. But to establish (30) for general asymptotically normal statistics \( T \) requires additional regularity conditions, which we discuss in Appendix B. For example, we show that if \( T = n^{-1} \sum_i d(X_i) \) and the \( X_i \) are iid, then (30) holds if \( d(X_i) \) has either an absolutely continuous component or \( d(X_i) \) is discrete with finite support.

The expansion (31) is the basis for deriving the asymptotic distribution of the conditional MLE and the validity of the last three rows of Table 2. First, the expansion suggests that \( \theta_{\text{MLE}} \) maximizing \( \log f(X|T; \theta) \) satisfies (23). The expansion (23) is similar to the expansion (22) for the unconditional MLE, but with the conditional score and information substituted for the unconditional ones. Second, we may expect a conditional Bernstein–von Mises theorem to hold. Basically, what is lacking for a full proof of results is a proof of \( \sqrt{n} \) consistency of \( \theta_{\text{MLE}} \) and \( \sqrt{n} \) consistency of the conditional posterior. These are not trivial matters, but they are of a technical nature and do not add to our knowledge of the form of the limits. This form is determined by the expansion (31) only. We content ourselves with providing in Appendix B exact conditions for the validity of the structural expansion (31) and sketching in Appendix C a direct proof for Example B of Section 3.

**Proof of Lemma 1.** We only need to prove the lemma in the special case where \( t(X) = n^{-1}S_{\psi}(\theta^*) \) because, from its definition, \( NC(\theta^*) \) will be the same for a given statistic \( t_1(X) \) and all affine transformations of \( t_1(X), t(X) = a t_1(X) + b + O_P(1) \), with \( a \neq 0 \). Now, by Theorem 4, for \( t(X) = n^{-1}S_{\psi}(\theta^*) \),
\[ \hat{\nu}_\psi(\theta^*) = i_{\psi}(\theta^*) \]
and
\[ \hat{\nu}_\psi(\theta^*) = i_{\phi}(\theta^*), \]

which proves the lemma.

**APPENDIX A: PROOF OF THEOREM 3**

By (14), we have that with \( X \sim f(x; \theta_0) \), for all \( \varepsilon > 0 \), there exists a constant \( c_\varepsilon \) such that
\[ \Pr_{\theta_0} \left\{ \left| \frac{E_{\theta}(X; \Sigma(\theta_o)/n)}{\sqrt{n}} \right| t \right\} \leq c_\varepsilon \leq \frac{\sqrt{n}}{\sqrt{\sigma^2(\theta_0)}} \]

and
\[ \frac{f(X|T, \theta_0)}{f(X|T, \theta_0)} = \frac{f(X|T, \theta_0)}{f(X|T, \theta_0)} = \frac{f(X|T, \theta_0)}{f(X|T, \theta_0)} \]

where \( E_{\theta}(X; \Sigma(\theta_o)/n) \) refers to expectation with respect to a normal distribution with mean \( \mu \) and variance matrix \( \Sigma \). Equation (A1) says that with large probability, \( X \) is such that when \( \theta \sim \mathcal{N}(\theta(0), \Sigma(\theta(0))/n) \), \( \theta \) lies in the ball of radius \( c_\varepsilon \sqrt{n} \) around \( \theta_0 \) with high probability.

Now, because the total variation norm \( \|P - Q\| \) is also equal to
\[ \sup_{\|P - Q\|} \left\{ \int f dP - \int f dQ \right\} < \frac{1}{2} \]
and \( \Pr_{\theta_0} \left\{ \left| \frac{E_{\theta}(X; \Sigma(\theta_o)/n)}{\sqrt{n}} \right| t \right\} \leq c_\varepsilon \leq \frac{\sqrt{n}}{\sqrt{\sigma^2(\theta_0)}} \]

from (14), the integrand can be defined in an arbitrary way for \( \theta \notin \Theta \).

Now if we restrict the integral to the set \( \{ \theta: \|\theta - \theta_0\| < c_\varepsilon \sqrt{n} \} \), then it changes at most by \( \int_{\|\theta - \theta_0\| > c_\varepsilon \sqrt{n}} \left\{ \int f dP - \int f dQ \right\} \leq \frac{1}{2} \)

for \( \theta \notin \Theta \) and \( \|	heta - \theta_0\| < c_\varepsilon \sqrt{n} \). Thus, \( f(x; \theta_0) \) is the density of a \( \mathcal{N}(\mu, \Sigma) \) random variable and, in (A2), the integrand can be defined in an arbitrary way for \( \theta \notin \Theta \).

Next, by (3b) holding at \( \theta_0 \) and (12), for every \( \theta_0 + O_P(1/\sqrt{n}) \),
\[ \Phi \left( \frac{n^{-1/2}(t(X, \theta_0) - \nu_0(0, \theta_0))}{\sigma(\theta_0)} \right) \]

This content downloaded from 192.12.88.148 on Sat, 28 Sep 2013 22:32:51 PM
All use subject to JSTOR Terms and Conditions
Now, we consider the expression for $\Phi(t(X, \theta_0) - \nu_n(0, \theta_0; \theta_0))$ and its convergence to zero. This can be rewritten as:

$$
\Phi\left(\frac{n^{1/2}(t(X, \theta_0) - \nu_n(0, \theta_0; \theta_0))}{\sigma(\theta_0)}\right)
$$

and

$$
\leq \left| n^{1/2}(t(X, \theta_0) - \nu_n(0, \theta_0; \theta_0)) \right| \left( \frac{\|\phi\|_{\infty}}{\sigma(\theta_0)} \right) \rightarrow 0.
$$

where $\|\phi\|_{\infty}$ is the maximum of a N(0, 1) density. By combining the two previous displays, we see that

$$
\sup_{|\theta - \theta_0| \leq \sigma(\theta_0) / \sqrt{n}} |Pr(t(X, \theta) - t(X, \theta_0; \theta_0))| \rightarrow 0.
$$

Now, by combining this display with (A.2), we obtain

$$
1 - p(X) = \int \left( \frac{n^{1/2}(t(X, \theta_0) - \nu_n(0, \theta_0; \theta_0))}{\sigma(\theta_0)} - \frac{\hat{\nu}_n(0, \theta_0; \theta_0) n^{1/2}(\theta - \theta_0)}{\sigma(\theta_0)} \right) d\theta
$$

and

$$
\times \phi(\theta; \theta^1(X, \Sigma(\theta_0)/n)) d\theta
$$

where $Q$ is given by (16).

**APPENDIX B: ASSUMPTIONS IMPLYING LOCAL ASYMPTOTIC NORMALITY**

**Lemma B.1.** Suppose that (21), (25), and (26) hold for a statistic $T = t(X)$. Furthermore, suppose that under $f(x; k/\sqrt{n}, \theta_0 + k^*/\sqrt{n})$,

$$
\sqrt{n}(T - \nu_n(k/\sqrt{n}, \theta_0 + k^*/\sqrt{n})
$$

converges in variation distance to a $N(0, \sigma^2(\theta_0))$ distribution for all $k \in R^1$, $k^* \in R^2$. Then (27) holds.

**Idea of Proof.** The lemma is essentially a consequence of theorem 4 of LeCam and Yang (1988), because in their terminology, our assumptions imply that $n^{1/2}(T - \nu_n(\psi, \theta))$ is distinguished in local experiments indexed by $\psi(\theta) = (k/\sqrt{n}, \theta_0 + k^*/\sqrt{n})$ with $\theta_0$ known. Details will be presented elsewhere.

If our statistic $T$ equals $n^{1/2} \sum d(X_i)$, then condition (B.1) is satisfied for $h = (k, k^*) = 0$ if $d(X_i)$ has a finite second moment and the distribution of $d(X_i)$ has an absolutely continuous component and the $X_i$ are iid. This follows by theorem XV.5.2 of Feller (1971). For any $h$, (B.1) will be true if we make these conditions uniform in $\theta$ running through a neighborhood of $\theta_0$. For more general asymptotically normal test statistics $T$, results such as (B.1) appear to be usually established as part of the derivation of an Edgeworth expansion for the distribution of $T$. (A discussion, with special attention to curved exponential families, and further references have been given in, for example, Ghosh 1994, chap. 2.) Results of this type are nontrivial. The use of the total variation norm makes (B.1) much more restrictive than convergence in law of $n^{1/2}(T - \nu_n(k/\sqrt{n}, \theta_0 + k^*/\sqrt{n}))$.

Condition (B.1) is certainly stronger than needed. It would be sufficient if the sequence $n^{1/2}(T - \nu_n(\psi, \theta))$ be distinguished in local experiments consisting of observing $T$ with parameter $(\psi, \theta) = (k/\sqrt{n}, \theta_0 + k^*/\sqrt{n})$, with $\theta_0$ being known, and $k \in R^3$, and $k^* \in R^5$. This concept was discussed by LeCam (1986), along with sufficient conditions, but the discussion is involved.

We now discuss an important special case in which (B.1) can be relaxed. If $T = T_n$ is lattice distributed with the span of lattice possibly depending on $n$, but not on $\theta$, then observing $T_n$ is statistically equivalent to observing a smooth version of $T_n$, if the smoothing is performed within the intervals generated by the lattice. In this case it can suffice to verify (B.1) for smooth versions of the law of $T_n$. We make this precise in the following theorem. We assume that $T_n = n^{-1} \sum d(X_i)$ with $d(X_i)$ taking its values in a grid of points $\ldots, a - s, a, a + s, a + 2s, \ldots$ for fixed numbers $a$ and $s$ (the span of the lattice). It appears that we can always arrange this without loss of generality. For example, if $d(X_i)$ is finitely discretely distributed, then it certainly suffices that $d(X_i)$ takes finitely many values in the rationals only. We assume that $a$ and $s$ do not depend on $\theta$.

**Theorem B.1.** Suppose that the $X_i$ are iid and $E_{\psi_n, \theta_0} [d(X_i)]^3 = O(1)$ for every $\psi_n \rightarrow 0$ and $\theta_0 \rightarrow \theta$. Assume that $\nu(\psi, \theta) = E_{\psi_n} [d(X_i)]$ is differentiable at $(0, \theta_0)$ and $\sigma^2(\psi, \theta) = var_{\psi, \theta} [d(X_i)]$ is continuous at $(0, \theta_0)$. Finally, assume that the distribution of $n^{1/2}t(X)$ under $f(x; \psi_n, \theta_0)$ converges in law to the distribution of $d(X_i)$ under $\nu(\psi, \theta) = (0, \theta_0)$. Then (30) holds.

The proof will be given elsewhere.

**APPENDIX C: CONDITIONAL INFERENCE IN EXAMPLE B OF SECTION 2**

Consistency and Asymptotic Normality of $\hat{\theta}_{MLE}$

For simplicity and without loss of generality, we consider the case where the variance $\sigma^2$ is known and equal to 1, so $\theta = \mu$. Thus $X_1, \ldots, X_n$ are iid $N(\mu, 1)$. Let $Z_{an} \equiv Z_{an},$ be the $n_{an}$-order statistic where $n_{an} \rightarrow n$ and $n_{an}$ is an integer. Then data from $Z_i; Z_{an} = x$ can be generated by generating iid data $Y_1, \ldots, Y_{an-1}$ from the density $I(y \leq x)\phi(y - \theta)/\Phi(x - \theta)$ and then generating $Y_{an+1}, \ldots, Y_n$ iid from the density $I(y > x)\phi(y - \theta)/\Phi(x - \theta)$ independently of the previous $Y_i$. Thus the likelihood function is proportional to

$$
\prod_{i=1}^{n_{an}-1} \phi(Y_i - \theta)/\Phi(x - \theta) \prod_{i=n_{an}+1}^{n} \phi(Y_i - \theta)/(1 - \Phi(x - \theta))
$$

(C.1)
Because the likelihood function is the product of the likelihood for two iid identified parametric models with common parameter \( \theta \), it follows that the maximizer \( \hat{\theta}_{\text{MLE}} \) of (C.1) is consistent for \( \theta \) for each fixed \( x \). Because the convergence of \( \hat{\theta}_{\text{MLE}} \) to \( \theta \) is uniform in \( x \) for \( x \) in a neighborhood of \( z_2(\theta) \), we conclude that \( \hat{\theta}_{\text{MLE}} \) is (unconditionally) consistent for \( \theta \). Now \( \hat{\theta}_{\text{MLE}} \) satisfies the conditional score equation

\[
\Psi(\hat{\theta}_{\text{MLE}}; X(nqn)) = 0,
\]

where \( X(j) \) is the \( j \)-th order statistic and

\[
\Psi(\theta, x) = n^{-1} \frac{\partial}{\partial \theta} \sum_{i=1}^{n-1} \log \phi(X(i) - \theta) - (nqn - 1) \log \Phi(x - \theta)
+ \sum_{i=nqn+1}^{n} \log \phi(X(i) - \theta) - (n - nqn) \log(1 - \Phi(x - \theta)).
\]

This implies that up to terms of \( O(1/n) \), \( \Psi(\theta, x) \) is approximated by

\[
\frac{n^{-1}}{q} \left( \sum_{i=1}^{n} (X(i) - \theta) - q \phi(x - \theta) \right)
\]

The approximation in (C.3) is obtained by adding \( X(nqn) - \theta \) and replacing \( nqn - 1 \) by \( nqn \).

By Taylor expansion of the score equation, we obtain

\[
n^{-1/2}(\hat{\theta}_{\text{MLE}} - \theta) = -n^{-1/2} \Psi(\theta, X(nqn))/\Psi(\hat{\theta}, X(nqn)) \tag{C.4}
\]

for \( \theta \) between \( \hat{\theta}_{\text{MLE}} \) and \( \theta \), where \( \Psi \) is the derivative of \( \Psi(\theta, x) \) with respect to \( \theta \). By \( \hat{\theta}_{\text{MLE}} \) consistent for \( \theta \), we have, from (C.3), by a further expansion around \( \theta \),

\[
n^{-1/2} \Psi(\theta, X(nqn)) + o_P(1)
\]

\[
= n^{-1/2} \sum_{i=1}^{n} \left( X(i) - \theta \right) - q \phi(x - \theta) \frac{1}{\Phi(x - \theta)} \frac{\Phi(x - \theta)}{\Phi(x - \theta) - \Phi(x - \theta)}
\]

\[
= n^{-1/2} S_{\theta}(\theta) - \frac{1}{\Phi(x - \theta)}\frac{\Phi(x - \theta)}{\Phi(x - \theta) - \Phi(x - \theta)} T_{\text{stat}}(\theta)
\]

and

\[
\Psi(\hat{\theta}, X(nqn)) = o_P(n^{-1/2} \Psi(\theta, x)) \tag{C.5}
\]

for each \( x \), we can use the Bernstein–von Mises theorem for independent random variables to conclude that

\[
E(\|\pi(\cdot|X(1), \ldots, X(nqn)) - N(\Delta(\theta, x), n^{-1} \Sigma(\theta, x))\|^2) \to 0. \tag{C.6}
\]

where \( P \) is the law of \( f(\mathcal{X}|T = x; \theta) = \Delta(\theta, x) + \Psi(\theta, x)/i(\theta, x) \), and \( i(\theta, x) = \text{var}(\Psi(\theta, x)|T = x) \). Because the convergence in (C.6) is uniform for \( x \) in a neighborhood of \( z_2(\theta) \), we can conclude that for every sequence \( x_n \to z_2(\theta) \),

\[
E(\|\pi(\cdot|X(1), \ldots, X_n, X(nqn)) - N(\Delta(\theta, x_n), n^{-1} i(\theta, x_n))\|^2) \to 0.
\]

This is sufficient to conclude that

\[
E(\|\pi(\cdot|X_1, \ldots, X_n, X(nqn)) - N(\Delta(\theta, X(nqn)), n^{-1} i(\theta, X(nqn)))\|^2) \to 0.
\]

By dominated convergence, this gives

\[
\lim_{n \to \infty} E(\|\pi(\cdot|X_1, \ldots, X_n, X(nqn)) - N(\Delta(\theta, X(nqn)), n^{-1} i(\theta, X(nqn)))\|^2) = 0.
\]

However, by (C.4), we can substitute \( \hat{\theta}_{\text{MLE}} \) for \( \Delta(\theta, X(nqn)) \) and \( i_{\text{MLE}}(\theta) \) for \( i(\theta, X(nqn)) \), concluding the proof.

[Received January 1999. Revised November 1999.]

REFERENCES


