Asymptotic Distribution of P Values in Composite Null Models

Author(s): James M. Robins, Aad van der Vaart and Valerie Ventura

Source: *Journal of the American Statistical Association*, Vol. 95, No. 452 (Dec., 2000), pp. 1143-1156

Published by: American Statistical Association

Stable URL: http://www.jstor.org/stable/2669750

Accessed: 28/09/2013 22:32

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at
http://www.jstor.org/page/info/about/policies/terms.jsp

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

American Statistical Association is collaborating with JSTOR to digitize, preserve and extend access to *Journal of the American Statistical Association.*

http://www.jstor.org

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

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 θ, one must somehow get rid of θ, (e.g., by estimating it) to calculate a p value. Various proposals have been suggested to "remove" θ, 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 θ, the posterior predictive p value of Guttman 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 particular 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 $\mathbf{x}_{\text{obs}}$ of a random variable $\mathbf{X}$. We posit a parametric "null" model, H_0: $f(\mathbf{x};\theta), \theta \in \Theta \subset R^p$, for the density of $\mathbf{X}$, and wish to investigate the compatibility of the null model with the observed data $\mathbf{x}_{\text{obs}}$. We do so by comparing the distribution of a given test statistic $T = t(\mathbf{X})$ with its observed value $t_{\text{obs}} = t(\mathbf{x}_{\text{obs}})$, using the p value

$$p(\mathbf{x}_{\text{obs}}) \equiv \Pr{}^{m(\cdot)}[t(\mathbf{X}) > t_{\text{obs}}] \tag{1}$$

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

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

$$f(\mathbf{x};\psi,\theta), (\psi,\theta) \in \Psi \times \Theta, \tag{2a}$$

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

$$f(\mathbf{x};0,\theta) = f(\mathbf{x};\theta), \qquad \theta \in \Theta, \tag{2b}$$

whereas $\psi \neq 0$ characterizes alternatives to H_0. When $f(\mathbf{x};\psi,\theta)$ 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 $f(\mathbf{x};\psi,\theta)$ is simply used to represent alternatives to H_0 that are substantively important to detect, or when no alternative model is specified,

James M. Robins is Professor of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, MA 02115 (E-mail: *robins@epinet.harvard.edu*). Aad van der Vaart is Professor of Statistics, Faculty of Mathematics and Computer Science, Vrije Universiteit, Amsterdam, 1081 HV The Netherlands (E-mail: *aad@cs.vu.nl*). Valérie Ventura is Visiting Assistant Professor, Carnegie Mellon University, Pittsburgh, PA 15213 (E-mail: *vventura@stat.cmu.edu*). This work was supported in part by National Institutes of Health grant R01A132475. The authors thank George Casella and Martin Tanner for organizing these contributions to JASA, and to the associate editor and to three referees for helpful criticisms.

© 2000 American Statistical Association
Journal of the American Statistical Association
December 2000, Vol. 95, No. 452, Theory and Methods

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

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 $f(\mathrm{x}; \theta)$, there is universal agreement that $m(\mathrm{x})$ should be $f(\mathrm{x}; \theta)$. Then $p(\mathbf{X})$ is uniform when H_0 is true, where $p(\mathbf{X})$ denotes the random variable whose observed value is $p(\mathbf{x}_{\text{obs}})$. When the parameters space Θ is not a singleton (i.e., H_0 is composite), one must eliminate the unknown "nuisance" parameter θ 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 $m_{\text{plug}}(\mathrm{x}|\mathbf{x}_{\text{obs}}) = f(\mathrm{x}; \hat{\theta}_{\text{obs}})$, where $\hat{\theta}_{\text{obs}}$ maximizes $f(\mathrm{x}_{\text{obs}}; \theta)$; note that we write $m(\cdot)$ in (1) as $m_{\text{plug}}(\cdot|\mathbf{x}_{\text{obs}})$ to stress its dependence on the observed data $\mathbf{x}_{\text{obs}}$. The reference densities for p_{plug} and for other candidate p values based on the statistic $t(\mathbf{X})$ 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 θ. These include the prior predictive p value p_{prior} of Box (1980) and the posterior predictive p value p_{post} of Guttman (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 predictive 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_{cplug}, which uses the maximizer $\hat{\theta}_{\text{cMLE,obs}}$ of the conditional likelihood $f(\mathbf{x}_{\text{obs}}|t_{\text{obs}}; \theta)$ 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 $t(\mathbf{X})$ by a discrepancy $t(\mathbf{X}, \theta)$, a function of the data $\mathbf{X}$ and of the parameter θ, so that

$$p_{\text{dis}} = p_{\text{dis}}(\mathbf{x}_{\text{obs}}) = \Pr^{m_{\text{dis}}(\cdot)}[t(\mathbf{X}, \theta) > t(\mathbf{x}_{\text{obs}}, \theta)],$$

with $m_{\text{dis}}(\mathbf{x}, \theta|\mathbf{x}_{\text{obs}}) = f(\mathbf{x}; \theta)\pi_{\text{post}}(\theta|\mathbf{x}_{\text{obs}})$.

1.3 Desirable Sampling Properties of Candidate p Values

First, we present some terminology. We call the random variable $p(\mathbf{X})$ a *candidate p value* if it has range [0, 1]; if it is also uniform under H_0, then we say that $p(\mathbf{X})$ is a *frequentist p value.* When a candidate p value is not uniform, we say that it is conservative (anticonservative) at θ if $\Pr[p(\mathbf{X}) < t]$ is less (greater) than t for all $t < 1/2$ when H_0: $\mathbf{X} \sim f(\mathrm{x}; \theta), \theta \in \Theta$ 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 $p(\mathbf{x}_{\text{obs}})$ to denote surprise or incompatibility because, in analogy with the noncomposite case, we act as if $p(\mathbf{X})$ 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°, 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 $\mathbf{X}$ arise from n mutually independent random variables. Then m in (1) can be chosen so that $p(\mathbf{X})$ is an *asymptotic frequentist p value;* that is, one whose distribution converges in law to a $U[0, 1]$ distribution under H_0: $\mathbf{X} \sim f(\mathrm{x}; \theta)$ 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

Table 1. Reference Densities for the Various Candidate p Values

Method	*Reference density*
Plug-in (p_{plug})	$m_{\text{plug}}\ (\mathbf{x}\|\mathbf{x}_{\text{obs}}) = f(\mathbf{x};\ \hat{\theta}_{\text{obs}})$
Prior predictive (p_{prior})	$m_{\text{prior}}\ (\mathbf{x}) = \int f(\mathbf{x};\ \theta)\pi(\theta)\ d\theta$
Posterior predictive (p_{post})	$m_{\text{post}}(\mathbf{x}\|\mathbf{x}_{\text{obs}}) = \int f(\mathbf{x};\ \theta)\pi_{\text{post}}(\theta\|\mathbf{x}_{\text{obs}})\ d\theta$
Partial posterior predictive (p_{ppost})	$m_{\text{ppost}}(\mathbf{x}\|\mathbf{x}_{\text{obs}}) = \int f(\mathbf{x};\ \theta)\pi_{\text{ppost}}(\theta\|\mathbf{x}_{\text{obs}})\ d\theta$
Conditional predictive (p_{cpred})	$m_{\text{cpred}}(\mathbf{x}\|\mathbf{x}_{\text{obs}}) = \int f(\mathbf{x}\|\hat{\theta}_{\text{cMLE[s}};\ \theta)\pi_{\text{cpred}}(\theta\|\mathbf{x}_{\text{obs}})\ d\theta$
Conditional plug-in (p_{cplug})	$m_{\text{cplug}}(\mathbf{x}\|\mathbf{x}_{\text{obs}}) = f(\mathbf{x};\ \hat{\theta}_{\text{cMLE,obs}})$
Discrepancy (p_{dis})	$m_{\text{dis}}(\mathbf{x},\ \theta\|\mathbf{x}_{\text{obs}}) = f(\mathbf{x};\ \theta)\pi_{\text{post}}(\theta\|\mathbf{x}_{\text{obs}})$

NOTE: The data model is $f(\mathbf{x};\ \theta)$, where θ has MLE $\hat{\theta}_{obs}$. The prior for θ is $\pi(\theta)$, and the posterior $\pi_{post}(\theta|\mathbf{x}_{obs}) \propto f(\mathbf{x}_{obs};\ \theta)\pi(\theta)$. The posterior in the conditional model $f(\mathbf{x}|t;\theta)$ is $\pi_{ppost}(\theta|\mathbf{x}_{obs}) \propto f\ (\mathbf{x}_{obs}|t_{obs};\ \theta)\pi(\theta)$, where $t\ =\ t(\mathbf{x})$ is the test statistic. The conditional MLE, $\hat{\theta}_{cMLE,obs}$, is the maximizer of $f(\mathbf{x}_{obs}|t_{obs};\ \theta)$. The posterior for θ in the "marginal" model, where only the statistic $\hat{\theta}_{cMLE}$ is available to the data analyst, is $\pi_{cpred}(\theta|\mathbf{x}_{obs}) \equiv \pi_{cpred}\ (\theta|\hat{\theta}_{cMLE,obs}) \propto f(\hat{\theta}_{cMLE,obs};\ \theta)\pi(\theta)$. Here $f(\hat{\theta}_{cMLE,obs},\ \theta)$ denotes the marginal density of the random variable $\hat{\theta}_{cMLE}$ evaluated at its observed value $\hat{\theta}_{cMLE,obs}$, and $f(\mathbf{x}|\ \hat{\theta}_{cMLE,obs};\ \theta)$ denotes the conditional density of $\mathbf{X}$ given $\hat{\theta}_{cMLE}$.

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

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 Θ, 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: $f(\mathbf{x};\theta), \theta \in \Theta$ restrict attention to statistics $t(\mathbf{X})$ 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 $\theta \in \Theta$. Then $m_T(t)$ in (1) is the density of $T = t(\mathbf{X})$ corresponding to F, which does not depend on the observed data $\mathbf{x}_{\text{obs}}$. In contrast, in the Bayesian p value and parametric bootstrap literature, the limiting distribution of $t(\mathbf{X})$ often depends on θ and the reference density $m_T(t|\mathbf{x}_{\text{obs}})$ depends on $\mathbf{x}_{\text{obs}}$, although in the bootstrap context attention is generally restricted to statistics $t(\mathbf{X})$ whose asymptotic mean is independent of θ. 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 $t(\mathbf{X})$ does not depend on θ.

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

$$p_{\text{sup}}(\mathbf{x}_{\text{obs}}) = \sup_{\theta \in \Theta} pr^{f(\mathbf{X};\theta)}[t(\mathbf{X}) > t(\mathbf{x}_{\text{obs}})].$$

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

Many of the test statistics considered in the Bayesian p value literature have asymptotic means that depend on the parameter θ; three examples illustrate this in Section 3. In the remainder of the article, we study the consequences of allowing the asymptotic mean of $t(\mathbf{X})$ to depend on θ. We restrict attention to statistics $t(\mathbf{X})$ 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 $t(\mathbf{X})$ varies with θ, 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_{cplug} 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 α-level test based on p_{post} against local Pitman alternatives is always less than the power of the test based on p_{plug}, itself less than the power of the tests based on p_{ppost} and p_{cpred}. 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 α-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 $t(\mathbf{X})$ are asymptotically correlated, regardless of the sign of the correlation. Furthermore, we show that $\hat{\theta}$ and $t(\mathbf{X})$ will be asymptotically correlated whenever the asymptotic mean of $t(\mathbf{X})$ depends on θ. Thus, as pointed out by Berger and Bayarri (1999, 2000) and Evans (1997), the problem with these p values is that $t(\mathbf{X})$ is effectively used twice: first to estimate θ, and again to assess lack of fit. In contrast, the conditional MLE, $\hat{\theta}_{\text{cMLE}}$, and $t(\mathbf{X})$ are always asymptotically uncorrelated, which is why $p_{\text{ppost}}, p_{\text{cpred}}$, and p_{cplug} are not conservative. But although p_{ppost} and p_{cpred} are asymptotically uniform, p_{cplug} is anticonservative, because it fails to properly account for the variability of $\hat{\theta}_{\text{cMLE}}$. In proposing p_{ppost} and p_{cpred}, both motivated through Bayesian arguments, Bayarri and Berger have solved the "frequentist" math problem of finding a reference density $m(\cdot|\mathbf{x}_{\text{obs}})$ such that the p value (1), based on an arbitrary statistic $t(\mathbf{X})$ with a limiting normal distribution, is an asymptotic frequentist p value. What is curious is that the obvious frequentist guesses, $m_{\text{plug}}(\cdot|\mathbf{x}_{\text{obs}})$ and $m_{\text{cplug}}(\cdot|\mathbf{x}_{\text{obs}})$, fail.

The preceding considerations do not apply to the discrepancy p value. Specifically, we show that p_{dis} can be seriously conservative even when the discrepancy $t(\mathbf{X}, \theta)$ has asymptotic mean 0 under $f(\mathbf{x};\theta)$ 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 $\mathbf{X} \equiv \mathbf{X}_n = (X_1, \ldots, X_n)$, 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

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

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 ψ_n indicates that the unidimensional 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

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

local Pitman alternatives, where $k_n \to k \in R^1$ as $n \to \infty$. Note that when $\psi_n = 0$, we frequently write the data model $f(\mathbf{x}; \psi_n, \theta) = f(\mathbf{x}; 0, \theta)$ more simply as $f(\mathbf{x}; \theta)$ and, for notational convenience, often suppress the subscript n denoting sample size in quantities such as $\mathbf{X} \equiv \mathbf{X}_n$.

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

$$n^{1/2}\left[\frac{t(\mathbf{X}) - \nu_n(k_n/\sqrt{n}, \theta)}{\sigma(\theta)}\right] \rightsquigarrow \mathrm{N}(0,1), \tag{3a}$$

where $\rightsquigarrow$ 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 θ_0 and k^*.

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

$$\dot{\nu}_\theta(\theta) = \lim_{n\to\infty} \partial \nu_n(0, \theta)/\partial\theta \tag{3b}$$

and

$$\dot{\nu}_\psi(\theta) = \lim_{n\to\infty} \partial \nu_n(\psi, \theta)/\partial\psi|_{\psi=0} \tag{3c}$$

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

We also need to define the scalars

$$\Omega(\theta) = \dot{\nu}_\theta(\theta)' i_{\theta\theta}^{-1}(\theta) \dot{\nu}(\theta) \tag{4}$$

and

$$\omega(\theta) = \dot{\nu}_\psi(\theta) - \dot{\nu}_\theta(\theta)' i_{\theta\theta}^{-1}(\theta) i_{\theta\psi}(\theta) \tag{5}$$

and the noncentrality parameter

$$\mathrm{NC}(\theta) = \omega(\theta)/[\sigma^2(\theta) - \Omega(\theta)]^{1/2}, \tag{6}$$

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

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 $f(\mathbf{x}; k_n/\sqrt{n}, \theta)$, each candidate p value can be written as

$$p(\mathbf{X}) = 1 - \Phi(Q) + o_P(1),$$

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

Plug-in: $\tau^2_{\text{plug}}(\theta) = [\sigma^2(\theta) - \Omega(\theta)]/\sigma^2(\theta)$
Posterior predictive: $\tau^2_{\text{post}}(\theta) = [\sigma^2(\theta) - \Omega(\theta)]/[\sigma^2(\theta) + \Omega(\theta)]$
Partial posterior predictive: $\tau^2_{\text{ppost}}(\theta) = 1$
Conditional predictive: $\tau^2_{\text{cpred}}(\theta) = 1$
Conditional plug-in: $\tau^2_{\text{cplug}}(\theta) = \sigma^2(\theta)/[\sigma^2(\theta) - \Omega(\theta)]$

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 $\dot{\nu}_\theta(\theta) = 0$, because then $\Omega(\theta)$ in (4) is 0. When $\dot{\nu}_\theta(\theta) \neq 0$, with $\tau^2(\theta) < 1$ ($\tau^2(\theta) > 1$), the p value is conservative (anticonservative). Hence p_{cplug} is anticonservative, whereas p_{ppost} and p_{plug} are conservative, with p_{ppost} being the more conservative because $\tau^2_{\text{ppost}}(\theta) < \tau^2_{\text{plug}}(\theta)$.

Remark 3: Efficiency. Let $\chi(\alpha) = I[p(X) < \alpha]$ denote the nominal α-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

$$\beta(\alpha, k, \theta) = \lim_{n\to\infty} E_{k/\sqrt{n},\theta}[\chi(\alpha)], \tag{7}$$

where $E_{k/\sqrt{n},\theta}$ refers to expectations under $f(\mathbf{x}; k/\sqrt{n}, \theta)$. The asymptotic representation of $p(\mathbf{X})$ given in Theorem 1 implies that $\chi(\alpha)$ has $\beta(\alpha, k, \theta) = 1 - \Phi[z_{1-\alpha}\tau^{-1}(\theta) - k\mathrm{NC}(\theta)]$. In model $f(\mathbf{x}; \psi, \theta)$, a locally most powerful asymptotic α-level test $\chi_{\text{eff}}(\alpha)$ of the hypothesis $\psi = 0$ has asymptotic power $1 - \Phi[z_{1-\alpha} - k\mathrm{NC}_{\text{eff}}(\theta)]$, where $\mathrm{NC}_{\text{eff}}(\theta)$ is the efficient noncentrality parameter $\mathrm{NC}_{\text{eff}}(\theta) = \{i_{\psi\psi}(\theta) - i_{\psi\theta}(\theta)' i_{\theta\theta}^{-1}(\theta) i_{\theta\psi}(\theta)\}^{1/2}$, with $i_{\psi\psi}(\theta) = \lim_{n\to\infty} n^{-1} E_\theta[S_\psi(\theta)^2]$ (see van der Vaart 1998, chap. 15). The following lemma indicates that a sufficient condition for $\chi_{\text{ppost}}(\alpha)$ and $\chi_{\text{cpred}}(\alpha)$ to be locally most powerful at a particular value θ^* of θ is that $t(\mathbf{X})$ is asymptotically equivalent to an affine transformation of $S_\psi(\theta^*)$, because then $\mathrm{NC}(\theta^*) = \mathrm{NC}_{\text{eff}}(\theta^*)$; the proof is in Section 5.

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

Remark 4: Actual Asymptotic Level and Relative Power and Efficiency. The asymptotic actual α-level of test $\chi(\alpha)$ is $\beta(\alpha, k, \theta)$ evaluated at $k = 0$, which we denote by actual(α, θ). As a function of α, actual(α, θ) is the cdf of the asymptotic distribution of $p(\mathbf{X})$ under H_0. We define the asymptotic relative power (ARP) of test $\chi(\alpha)$, denoted by ARP $=$ ARP(α, β, θ), as its asymptotic power under alternative $f(\mathbf{x}; k_{\text{ppost}}/\sqrt{n}, \theta)$, with k_{ppost} chosen such that the asymptotic power of $\chi_{\text{ppost}}(\alpha)$ is β; that is, with k_{ppost}

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

such that $\beta_{\text{ppost}}(\alpha, k_{\text{ppost}}, \theta) = \beta$. Finally, the asymptotic relative efficiency ARE = ARE (α, β, θ) of candidate test $\chi(\alpha)$ is the limit, as $n_{\text{cand}} \to \infty$, of the ratio $n_{\text{ppost}}/n_{\text{cand}}$, where n_{ppost} and n_{cand} are the sample sizes needed for tests $\chi_{\text{ppost}}(\alpha)$ and candidate test $\chi(\alpha)$ to both have power β under the alternative $f(\mathbf{x}; k/\sqrt{n}, \theta)$.

Corollary 1. For each candidate $p(\mathbf{X})$ of Theorem 1, the actual asymptotic α-level of the test $\chi(\alpha)$ is

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

If $\dot{\nu}_\psi(\theta) \neq 0$, then the ARP and ARE are

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

and

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

Note that neither ARP(α, β, θ) nor ARE(α, β, θ) depends on NC(θ), and so these two quantities are the same for all local alternatives nesting the model $f(\mathbf{x}; \theta)$. When $\tau^2(\theta) < 1/3$, ARP$(\alpha, 1-\alpha, \theta) < \alpha$, so the asymptotic local power of the test $\chi(\alpha)$ is less than α, 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 $t(\mathbf{X}, \theta)$ be a discrepancy measure and, for a given fixed θ_0, let $t(\mathbf{X}) = t(\mathbf{X}, \theta_0)$. Then, under $f(\mathbf{x}; k_n/\sqrt{n}, \theta_0), p_{\text{dis}}(\mathbf{X})$ based on $t(\mathbf{X}, \theta^*)$ and $p_{\text{post}}(\mathbf{X})$ based on $t(\mathbf{X})$ have the same limiting distribution.

Moreover, if we redefine $\sigma^2(\theta)$ to be the asymptotic variance of $n^{1/2}t(\mathbf{X}, \theta)$ under $f(\mathbf{x}; 0, \theta)$ and $\dot{\nu}_\theta(\theta) = \lim_{n\to\infty} \partial\nu_n(0, \theta^*\colon \theta)/\partial\theta^*_{|\theta^*=\theta}$ and $\dot{\nu}_\psi(\theta) = \lim_{n\to\infty} \partial\nu_n(\psi, \theta\colon \theta)/\partial\psi_{|\psi=0}$ where $\nu_n(\psi, \theta^*\colon \theta)$ is the asymptotic mean of $n^{1/2}t(\mathbf{X}, \theta)$ under $f(\mathbf{x}; \psi, \theta^*)$, then Theorem 1 holds for $p_{\text{dis}}(\mathbf{X})$ 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 $\mathbf{X} = (X_1, \ldots, X_n)$, with the X_i's being mutually independent. Consider the null model $X_i \sim \text{N}(\gamma v_i, c^2), \theta = (\gamma, c^2)'$, and $\mathbf{v} = (v_1, \ldots, v_n)'$ a vector of known constants. Let the test statistic be $t(\mathbf{X}) = n^{-1}\sum_i X_i w_i = n^{-1}\mathbf{X}'\mathbf{w}$, where $\mathbf{w} = (w_1, \ldots, w_n)'$ is another vector of known constants. We assume that $\mathbf{v}'\mathbf{v} = \mathbf{w}'\mathbf{w} = 1$ and $\mathbf{v}'\mathbf{1} = \mathbf{w}'\mathbf{1} = 0$, where $\mathbf{1}$ is the n-vector of 1's. Then the mean of $t(\mathbf{X}), E_\theta(t(\mathbf{X})) = \rho\gamma$, depends on $\theta = (\gamma, c^2)'$ whenever $\rho \neq 0$, where $\rho = n^{-1}\sum_i w_i v_i = n^{-1}\mathbf{w}'\mathbf{v}$. Figure 1 shows the asymptotic cdf's actual(α) = actual(α, θ) of various candidate p values under the null model, for several choices of ρ. These depend on $\mathbf{w}$ and $\mathbf{v}$ only through ρ, and they do not depend on the value of θ 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 ρ of the regressors approaches 1. In contrast, p_{cplug} is anticonservative. In particular, as $\rho \to 1$, actual$_{\text{cplug}}(\alpha, \theta)$ converges to 1/2 for all $\alpha < 1/2$.

Let the alternative model be $\text{N}(\psi w_i + \gamma v_i, c^2)$. Then the score, $\mathbf{S}_\psi(\theta) = n(t(\mathbf{X}) - \rho\gamma)/c^2$, is an affine transformation of $t(\mathbf{X})$, and so $\chi_{\text{ppost}}(\alpha)$ is locally most powerful for testing $\psi = 0$. Figure 2 displays actual(α, θ), ARP(α, β, θ), and ARE(α, β, θ) as functions of ρ, with $\alpha = .05$ and $\beta = .80$, for several candidate p values. These functions do not depend on the value of $\theta = (\gamma, c^2)'$ 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 $\rho \to 1$, both power functions fall below the nominal α-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

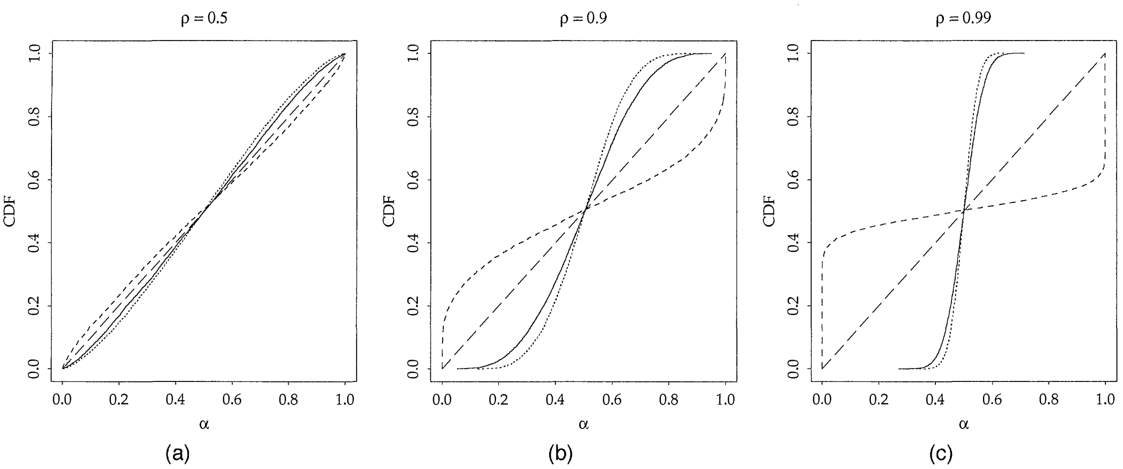

Figure 1. Example A. Asymptotic cdfs of Candidate p Values, for (a) ρ = .5, (b) ρ = .9, and (c) ρ = .99. (— plug, · · · post, - - - cplug, — — ppost, cpred).

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

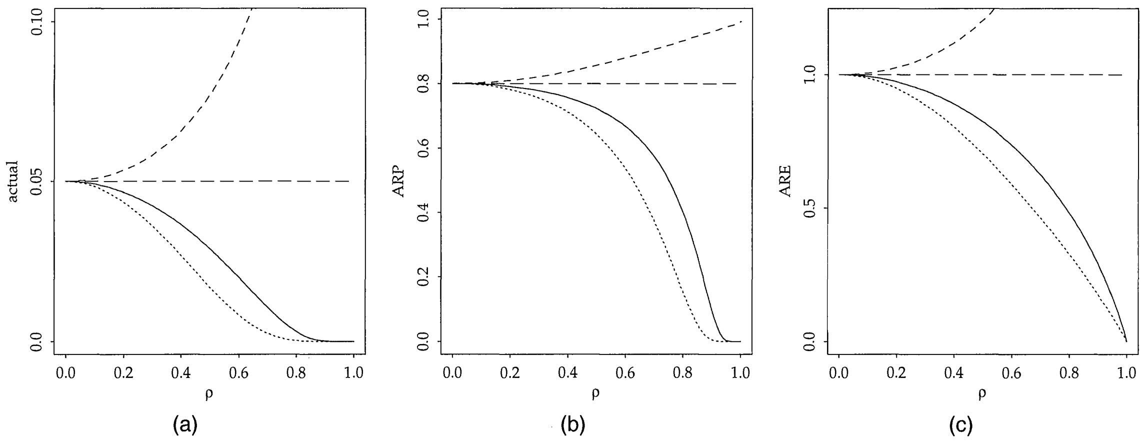

*Figure 2. Example A. (a) Actual(α, θ), (b) ARP(α, β, θ), and (c) ARE(α, β, θ) as Functions of ρ for α = 5% and β = 80%. The vertical axes of (a) and (c) were truncated for quality of display; note that the actual level of p_{cplug}(**X**) → 1/2 as ρ → 1, and its ARE → ∞. (— plug, ··· post, - - - cplug, —— ppost, cpred).*

the "correlation" ρ of the regressors, which is ancillary. For example, suppose that $p_{\text{post}} = .25$ 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, ρ would be the temperature scale, p_{post} the temperature, and the decision whether to use or discard the null model that to go swimming or skiing. However, it is rare for ρ to even be reported by the analyst, in which case reaching an appropriate decision would be impossible. Yet even if ρ 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 $t(\mathbf{X}, \theta) = n^{-1}\mathbf{S}_{\psi}(\theta) = (c^2 n)^{-1}\sum_i (X_i - \gamma v_i) w_i$. Note that the mean of $t(\mathbf{X}, \theta)$ under the null model, $X_i \sim \text{N}(\gamma v_i, c^2)$, is 0 for all θ. Nonetheless, because $t(\mathbf{X}) > t(\mathbf{x}_{\text{obs}})$ iff $t(\mathbf{X}, \theta) > t(\mathbf{x}_{\text{obs}}, \theta)$, it follows that $p_{\text{dis}}(\mathbf{X}) = p_{\text{post}}(\mathbf{X})$ with probability 1 under both the null and alternative models. Thus the curves for $p_{\text{dis}}(X)$ and $p_{\text{post}}(X)$ 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 iid draws from a $\text{N}(\mu, c^2)$ distribution, with a noninformative uniform prior on $(\mu, \log c)$. To look for incompatibility of the data with the $\text{N}(\mu, c^2)$ model in the left tail of the distribution, they computed a posterior predictive p value with $t(\mathbf{X}) = \min\{X_i; i = 1, \ldots, n\}$, the first-order statistic (Gelman et al. 1995, p. 166). A reasonable alternative choice for $T = t(\mathbf{X})$ would be the empirical qth quantile of $\mathbf{X}, T = \hat{Z}_q \equiv \sup\{t; n^{-1}\sum_i I(X_i \le t) < q\}$, 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 $T = \hat{Z}_q$ under the normal null model is the population qth quantile of a $\text{N}(\mu, c^2)$ distribution; that is, $z_q(\theta) = z_q(\mu, c^2) = z_q c + \mu$, with $z_q = z_q(0, 1)$, which depends on $\theta = (\mu, c^2)$. Hence we expect both the plug-in and posterior predictive p values to be asymptotically conservative. Figure 3 shows the curves actual$(.05, \theta)$, ARP$(.05, .8, \theta)$, and ARE$(.05, .8, \theta)$ for our candidate p values as a function of q, none of which depends on the value of θ generating the data. Note that as discussed in Corollary 1 we did not have to specify the alternative model $f(\mathbf{x}; \psi, \theta)$ under which the ARP and ARE are calculated.

A natural discrepancy measure generalizing the test statistic $t(\mathbf{X}) = \hat{Z}_q$ is $t(\mathbf{X}, \theta) = \hat{Z}_q - z_q(\theta)$, the difference between the empirical and the true qth quantiles of the null model. Because $t(\mathbf{X}) > t(\mathbf{x}_{\text{obs}})$ iff $t(\mathbf{X}, \theta) > t(\mathbf{x}_{\text{obs}}, \theta), p_{\text{dis}}(\mathbf{X})$ and $p_{\text{post}}(\mathbf{X})$ 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

$$t(\mathbf{X}, \theta) = |\hat{Z}_{1-q} - \mu| - |\mu - \hat{Z}_q|$$

with $q = .1$, to check whether or not the magnitude of skewness, as measured by $t(\mathbf{x}_{\text{obs}}, \theta)$, was compatible with a $\text{N}(\mu, c^2)$ distribution. Note that under the null model $\text{N}(\mu, c^2), t(\mathbf{X}, \theta)$ has asymptotic (and exact) mean 0 for all

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

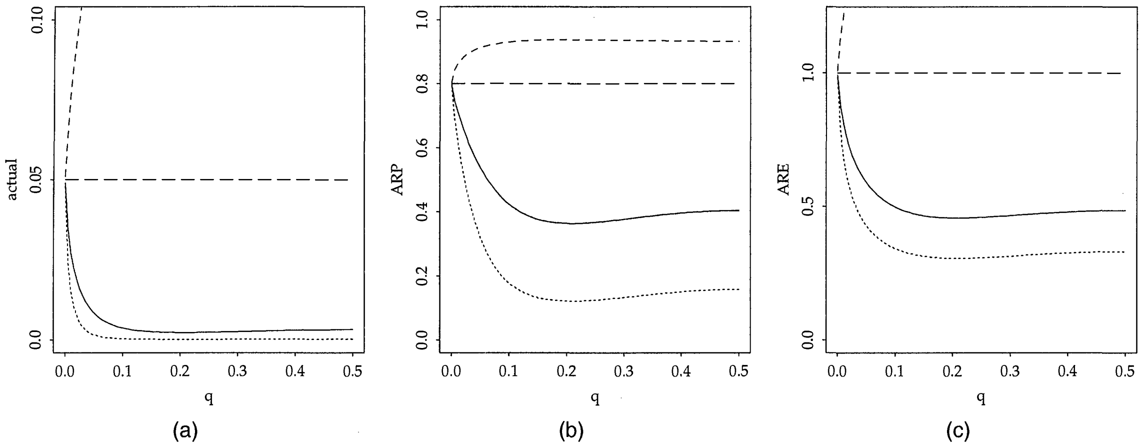

Figure 3. Example B. (a) Actual(α, θ) for α = 5% as a Function of q, (b) ARP(α, β, θ) and (c) ARE(α, β, θ), for α = 5% and β = 80%. The vertical axes of (a) and (c) were truncated for quality of display. (— plug, · · · post, - - - cplug, —— ppost, cpred).

$\theta = (\mu, c^2)$, because the X_i have a symmetric distribution centered at μ. A natural test statistic related to the discrepancy $t(\mathbf{X}, \theta)$ is $t(\mathbf{X}) = \hat{Z}_{1-q} + \hat{Z}_q$ because, on a set with probability going to 1, $t(\mathbf{X}, \theta) = \hat{Z}_{1-q} + \hat{Z}_q - 2\mu$ under the null model and any local alternative. On this set, $t(\mathbf{X}) > t(\mathbf{x}_{\text{obs}})$ iff $t(\mathbf{X}, \theta) > t(\mathbf{x}_{\text{obs}}, \theta)$, and so $p_{\text{dis}}(X)$ based on $t(\mathbf{X}, \theta)$, and $p_{\text{post}}(X)$ based on $t(\mathbf{X})$, have the same asymptotic distribution. Figure 4 shows actual$(.05, \theta)$, ARP$(.05, .8, \theta)$, and ARE$(.05, .8, \theta)$ for our candidate p values based on $t(\mathbf{X})$, 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 $\sigma(\theta), \Omega(\theta)$, and NC$(\theta)$ were obtained for the test statistics and discrepancies in Examples A–C.

Example A. Here $\theta' = (\theta_1, \theta_2) = (\gamma, c^2), X_i \sim \text{N}(\psi w_i + \gamma v_i, c^2)$, and $t(\mathbf{X}) = n^{-1}\mathbf{X}'\mathbf{w}$. We assume that at each sample size n, the vectors of constants $(\mathbf{v}, \mathbf{w}) = (\mathbf{v}_n, \mathbf{w}_n)$ are chosen such that $n^{-1}\mathbf{v}'\mathbf{v}, n^{-1}\mathbf{w}'\mathbf{w}$, and $\rho = n^{-1}\mathbf{w}'\mathbf{v}$ do not depend on n. Then the asymptotic mean of $t(\mathbf{X})$ is $\nu_n(\psi, \theta) = \psi + \theta_1\rho$. Also, $\dot{\nu}_\theta(\theta) = (\rho, 0)', \dot{\nu}_\psi(\theta) = 1, \sigma^2(\theta) = \theta_2, i_{\theta\theta}(\theta) = \text{diag}(\theta_2^{-1}, 1/2\theta_2^2), i_{\psi\theta}(\theta)' = (\theta_2^{-1}\rho, 0), \Omega(\theta) = \rho^2\theta_2$, and $\omega(\theta) = 1 - \rho^2$. For the discrepancy, $t(\mathbf{X}, \theta) = n^{-1}\mathbf{S}_\psi(\theta) = n^{-1}\theta_2^{-1}\sum_i(X_i - \theta_1 v_i)w_i$, $\nu_n(\psi, \theta^*\colon \theta) = E_{\psi,\theta^*}\{t(\mathbf{X}, \theta)\} = (\theta_2^*)^{-1}\{\psi + (\theta_1^* - \theta_1)\rho\}$. Hence $\dot{\nu}_\theta(\theta) = (\rho, 0)'\theta_2^{-1}, \dot{\nu}_\psi(\theta) = \theta_2^{-1}, \sigma^2(\theta) = \theta_2^{-1}, \Omega(\theta) = \theta_2^{-1}\rho^2$, and $\omega(\theta) = (1 - \rho^2)\theta_2^{-1}$.

Example B. Here $t(\mathbf{X}) = \hat{Z}_q$, and under $f(\mathbf{x}; \theta), X_i \sim \text{N}(\mu, c^2)$ with $\theta' = (\theta_1, \theta_2) = (\mu, c^2)$. Then $\nu_n(0, \theta) = \theta_2^{1/2}z_q + \theta_1$, and hence $\dot{\nu}_\theta(\theta) = (1, \theta_2^{-1/2}z_q/2), i_{\theta\theta}(\theta) = \text{diag}(1/\theta_2, 1/2\theta_2^2)$, and $\Omega(\theta) = \theta_2(1 + z_q^2/2)$. Further, it is well known that if the X_i are iid under any law $f(x; \theta)$, then

$$n^{1/2}(\hat{Z}_q - z_q(\theta)) \\ = n^{-1/2} - f[z_q(\theta); \theta]^{-1}\sum_i I(X_i < z_q(\theta)) + o_P(1).$$

Hence $\sigma^2(\theta) = f(z_q(\theta); \theta)^{-2}\text{var}_\theta\{I[X_i < z_q(\theta)]\}$, which, for our $\text{N}(\theta_1, \theta_2)$ model, evaluates to

$$\sigma^2(\theta) = \theta_2\phi^{-2}(z_q)q(1 - q),$$

where ϕ is the standard normal density. Further, under $f(\mathbf{x}; k_n/\sqrt{n}, \theta)$, it follows from Theorem 3 that $\dot{\nu}_\psi(\theta) =$

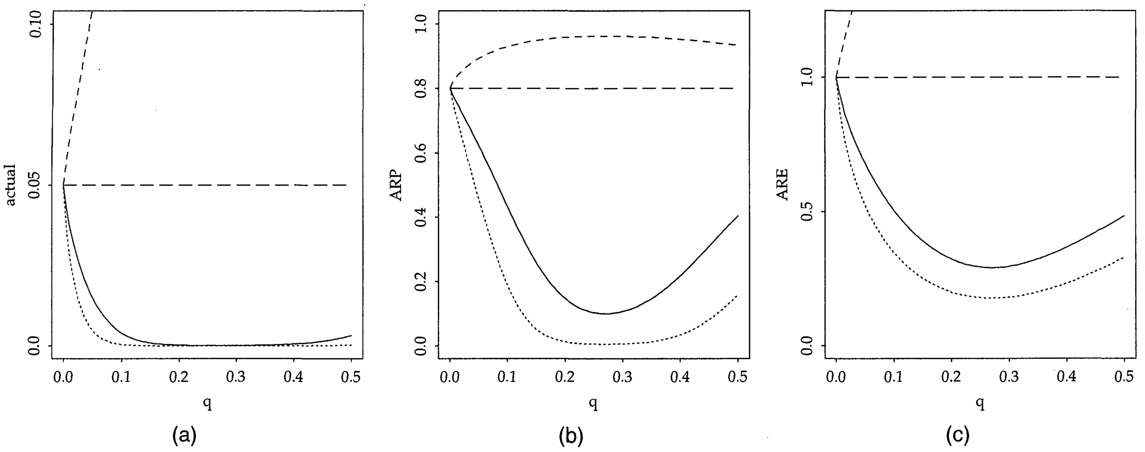

Figure 4. Example C. (a) Actual(α, θ), (b) ARP(α, β, θ), and (c) ARE(α, β, θ) as Functions of q for α = 5% and β = 80%. The vertical axes of (a) and (c) were truncated for quality of display. (— plug, · · · post, - - - cplug, —— ppost, cpred).

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

$E_\theta[-f(z_q(\theta);\theta)^{-1}I(X_i < z_q(\theta))S_{\psi,i}(\theta)]$, where $S_{\psi,i}(\theta)$ is the contribution of subject i to $\mathbf{S}_\psi(\theta)$. For the discrepancy $t(\mathbf{X},\theta) = \hat{Z}_q - z_q(\theta)$, all of the relevant foregoing quantities are the same as for $t(\mathbf{X}) = \hat{Z}_q$, but with $\nu_n(0,\theta^*{:}\ \theta) = z_q(\theta^*) - z_q(\theta) = \theta_2^{*1/2} z_q + \theta_1^* - (\theta_2^{1/2} z_q + \theta_1)$.

Example C. In this example, $t(\mathbf{X}) = \hat{Z}_q + \hat{Z}_{1-q}$, where $X_i \sim \mathrm{N}(\mu, c^2), \theta' = (\theta_1,\theta_2) = (\mu, c^2)$; so as in Example B, $i_{\theta\theta}(\theta) = \mathrm{diag}(1/\theta_2, 1/2\theta_2^2)$. It follows from the results of Example B that $\nu_n(0,\theta) = \theta_2^{1/2}(z_q + z_{1-q}) + 2\theta_1$, and so that $\dot{\nu}_\theta(\theta) = (2,0)$, because $z_q + z_{1-q} = 0$. Thus $\Omega(\theta) = 4\theta_2$. Further, it follows from (4) that $\sigma^2(\theta) = \mathrm{var}_\theta\{-f(z_q(\theta);\theta)^{-1}I(X_i < z_q(\theta)) - f(z_{1-q}(\theta);\theta)^{-1}I(X_i < z_{1-q}(\theta))\}$, which for our normal model evaluates to $\sigma^2(\theta) = 2q\phi^{-2}(z_q)$. All the relevant foregoing quantities are the same for the discrepancy measure $t(\mathbf{X},\theta) = |\hat{Z}_{1-q} - \mu| - |\mu - \hat{Z}_q|$ as for $t(\mathbf{X}) = \hat{Z}_q + \hat{Z}_{1-q}$, but with $\nu_n(0,\theta^*{:}\ \theta) = 2(\theta_1^* - \theta_1)$.

4. ALTERNATIVES TO p_{cpred} AND p_{ppost}

It follows from Theorems 1 and 2 and Corollary 1 that from an asymptotic frequentist viewpoint, p_{cpred} and p_{ppost} are preferred to our other candidate p values when the asymptotic mean of $t(\mathbf{X})$ depends on θ. 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 $t(\mathbf{X})$ of departure from the null model; the preferred, or intuitive, choice may happen to have a nonconstant asymptotic mean. So suppose that our choice $t(\mathbf{X})$ satisfies (3a)–(3c), with $\dot{\nu}_\theta(\theta) \neq 0$. Now p_{cpred} and p_{ppost} 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 $t(\mathbf{X})$ 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 $t(\mathbf{X})$ and $t(\mathbf{X},\theta)$

One alternative to computing p_{cpred} or p_{ppost} based on $t(\mathbf{X})$ is to compute p_{plug} or p_{post} based on the statistic $\hat{t}(\mathbf{X}) = t(\mathbf{X}) - \nu_n(\hat{\theta})$, where $\nu_n(\hat{\theta}) \equiv \nu_n(0,\hat{\theta})$, because, by a Taylor expansion, $\hat{t}(\mathbf{X})$ is asymptotically normal with constant asymptotic mean 0 and asymptotic variance

$$c^2(\theta) = \sigma^2(\theta) - \dot{\nu}_\theta(\theta)' i_{\theta\theta}^{-1}(\theta)\dot{\nu}_\theta(\theta)$$

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

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

where $\mathbf{X}^{(k)} = (X_1^{(k)}, \ldots, X_n^{(k)})$ are K independent draws from $f(\mathbf{X}; \hat{\theta}_{\mathrm{obs}})$, and $\hat{t}(\mathbf{x}_{\mathrm{obs}}) = t(\mathbf{x}_{\mathrm{obs}}) - \nu_n(\hat{\theta}_{\mathrm{obs}})$. A potential drawback of this approach is that to evaluate $\hat{t}(\mathbf{X}^{(k)}) = t(\mathbf{X}^{(k)}) - \nu_n(\hat{\theta}^{(k)})$, the maximizer $\hat{\theta}^{(k)}$ of $f(\mathbf{X}^{(k)};\theta)$ must be recomputed for each simulated dataset $\mathbf{X}^{(k)}$. 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_{post}, the posterior distribution of θ 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 θ for each simulated dataset, two additional approaches may be considered, both of which give p values that are asymptotically equivalent to $p_{\mathrm{ppost}}(\mathbf{X})$ under both the null model and local alternatives. The first approach is to replace $\hat{t}(\mathbf{X})$ by the asymptotically pivotal N(0, 1) random variable $\hat{\hat{t}}(\mathbf{X}) = \hat{t}(\mathbf{X})/c(\hat{\theta})$. We then obtain an asymptotic frequentist p value p_{pivot} based on $\hat{\hat{t}}(\mathbf{X})$ by using $m(t) = \mathrm{N}(0,\ 1)$ in (1); specifically, $p_{\mathrm{pivot}} = 1 - \Phi[\hat{\hat{t}}(\mathbf{x}_{\mathrm{obs}})]$. The second alternative is to calculate a discrepancy p value based on the discrepancy $\hat{t}(\mathbf{X},\theta) = t(\mathbf{X}) - \nu_n(\theta) - \dot{\nu}_\theta(\theta)' i_{\theta\theta}^{-1}(\theta) n^{-1}\mathbf{S}_\theta(\theta)$.

Indeed given any discrepancy $t(\mathbf{X},\theta)$ with $E_\theta[t(\mathbf{X},\theta)] = 0$ the discrepancy p-value $\hat{p}_{\mathrm{dis}}$ based on the modified discrepancy

$$\hat{t}(\mathbf{X},\theta) = t(\mathbf{X},\theta) - \dot{\nu}_\theta(\theta)' i_{\theta\theta}^{-1}(\theta) n^{-1}\mathbf{S}_\theta(\theta),$$

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

A drawback of these latter approaches is that they can require computation or estimation of $\sigma^2(\theta), \nu_n(\theta), \dot{\nu}_\theta(\theta)$, and/or $i_{\theta\theta}(\theta)$, which may be computationally difficult. But an important advantage of the last approach is that if we take $t(\mathbf{X},\theta)$ equal to $n^{-1}\mathbf{S}_\psi(\theta)$, then $\hat{t}(\mathbf{X},\theta)$ becomes the "efficient score" discrepancy

$$\hat{t}(\mathbf{X},\theta) = n^{-1}\mathbf{S}_\psi(\theta) - i_{\psi\theta}(\theta)' i_{\theta\theta}^{-1}(\theta) n^{-1}\mathbf{S}_\theta(\theta) \qquad (8)$$

and the test $\hat{\chi}_{\mathrm{dis}}$ based on $\hat{p}_{\mathrm{dis}}$ is a locally most powerful asymptotic α-level test of $\psi = 0$.

4.2 Adjusted p Values

Another class of alternatives to p_{cpred} or p_{ppost} are the adjusted (i.e., calibrated) p values $p_{\mathrm{post,adj}}, p_{\mathrm{plug,adj}}$, and $p_{\mathrm{cplug,adj}}$, where for any candidate p value $p(\mathbf{X})$ with observed value $p = p(\mathbf{x}_{\mathrm{obs}})$,

$$p_{\mathrm{adj}} \equiv p_{\mathrm{adj}}(\mathbf{x}_{\mathrm{obs}}) \equiv F_{p(\mathbf{X})}[p; \hat{\theta}_{\mathrm{obs}}],$$

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

where $F_{p(\mathbf{X})}(u;\theta)$ is the cdf of $p(\mathbf{X})$ when $\mathbf{X} \sim f(\mathbf{x};\theta)$ (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, $p_{\text{plug,adj}}$ 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 $p(\mathbf{X}) = 1 - \Phi(Q) + o_p(1)$, under H_0, of Theorem 1 implies that $p_{\text{adj,anal}} = 1 - \Phi[\tau^{-1}(\theta)z_{1-p}]$ is a simple analytic approximation to p_{adj}. It is easy to show that for any of our candidate p values, both $p_{\text{adj}}(\mathbf{X})$ and $p_{\text{adj,anal}}(\mathbf{X})$ have the same limiting distribution as $p_{\text{ppost}}(\mathbf{X})$ and $p_{\text{cpred}}(\mathbf{X})$ under $f(\mathbf{x}; k_n/\sqrt{n}, \theta)$.

Given a candidate p value defined via (1) with reference density $m(t|\mathbf{x}_{\text{obs}}) \equiv m_T(t|\mathbf{x}_{\text{obs}})$, define the adjusted reference density evaluated at t_{obs} to be

$$m_{\text{adj}}(t_{\text{obs}}|\mathbf{x}_{\text{obs}}) = f_{p(\mathbf{X})}[p; \hat{\theta}_{\text{obs}}]m(t_{\text{obs}}|\mathbf{x}_{\text{obs}}),$$

where $f_{p(\mathbf{X})}(u;\theta) = \partial F_{p(\mathbf{X})}(u;\theta)/\partial u$ and $p \equiv p(t_{\text{obs}})$. Then p_{adj} is obtained via (1) with $m(t) = m_{\text{adj}}(t|\mathbf{x}_{\text{obs}})$. These densities $m_{\text{adj}}(t|\mathbf{x}_{\text{obs}})$ 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 $p_{\text{ppost}}(\mathbf{X})$ and $p_{\text{cpred}}(\mathbf{X})$ under the null model and local alternatives. Thus the corresponding nominal α-level tests $\chi(\alpha)$ have the same asymptotic power as the tests $\chi_{\text{ppost}}(\alpha)$ and $\chi_{\text{cpred}}(\alpha)$ based on $t(\mathbf{X})$; that is, $\text{ARP}(\alpha,\beta,\theta) = \beta$ and $\text{ARE}(\alpha,\beta,\theta) = 1$. 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_{ppost} and p_{cpred} will be preferred to p_{post} and p_{plug} in small samples, even when the mean of $t(\mathbf{X})$ does not depend on θ. 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 θ 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

$$p(\mathbf{X}) = 1 - \Phi(Q) + o_p(1) \tag{9}$$

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

$$\begin{aligned} p &= p(\mathbf{x}_{\text{obs}}) \\ &= \int_\Theta \Pr[t(\mathbf{X},\theta) > t(\mathbf{x}_{\text{obs}},\theta);\theta]\pi(d\theta|\mathbf{x}_{\text{obs}}). \end{aligned} \tag{10}$$

By taking $t(\mathbf{X},\theta) = t(\mathbf{X})$, we obtain the nondiscrepancy p values. Here $\pi(d\theta|\mathbf{x}_{\text{obs}}) = \pi(\theta|\mathbf{x}_{\text{obs}})d\theta$ is given in Table 1 for p_{ppost} and p_{post}. We take $\pi_{\text{dis}}(\cdot|\mathbf{x}_{\text{obs}}) = \pi_{\text{post}}(\cdot|\mathbf{x}_{\text{obs}})$. For p_{plug} and p_{cplug}, we take $\pi(d\theta|\mathbf{x}_{\text{obs}})$ to be the degenerate distribution that places all of its mass on $\hat{\theta}_{\text{obs}}$ and $\hat{\theta}_{\text{cMLE,obs}}$. When we take $\pi(\cdot|\mathbf{x}_{\text{obs}})$ in (10) to be $\pi_{\text{cpred}}(\cdot|\mathbf{x}_{\text{obs}})$, 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_{cpred}. Hence it will suffice to prove Theorem 1 for p_{acpred} in lieu of p_{cpred}. It will be useful to have an expression for the random p value $p(\mathbf{X})$; specifically,

$$p(\mathbf{X}) = \int_\Theta \Pr[t(\mathbf{X}^{\text{new}},\theta) > t(\mathbf{X},\theta)|\mathbf{X};\theta]\pi(d\theta|\mathbf{X}), \tag{11}$$

where $\mathbf{X}^{\text{new}}$ is drawn from $f(\mathbf{x};\theta)$ independently of $\mathbf{X}$.

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

Actually, because the measures $\pi(d\theta|\mathbf{X})$ concentrate on small (shrinking) neighborhoods of θ_0, the dependence of $t(\mathbf{X},\theta)$ on θ 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 $\tilde{\theta}_n = \theta_0 + O_p(n^{-1/2})$,

$$\frac{n^{1/2}\{t(\mathbf{X},\tilde{\theta}_n) - \nu_n(0,\theta_0\colon \tilde{\theta}_n)\}}{\sigma(\theta_0\colon \theta_0)} - \frac{n^{1/2}\{t(\mathbf{X},\theta_0) - \nu_n(0,\theta_0\colon \theta_0)\}}{\sigma(\theta_0\colon \theta_0)} \overset{P_{\theta_0}}{\to} 0, \tag{12}$$

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

$$\|\pi(\cdot|\mathbf{X}) - \text{N}(\theta^\dagger(\mathbf{X}), \mathbf{\Sigma}(\theta_0)/n)\| \overset{P_{\theta_0}}{\to} 0 \tag{13}$$

and

$$n^{1/2}(\theta^\dagger(\mathbf{X}) - \theta_0) = O_{P_{\theta_0}}(1). \tag{14}$$

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

Theorem 3. Suppose that (12)–(14) and (3) hold with $t(\mathbf{X},\theta)$ and $\nu_n(0,\theta\colon \theta)$ replacing $t(\mathbf{X})$ and $\nu_n(0,\theta)$. Then,

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

Table 2. Values of $\theta^\dagger$ and Σ

Method	$\theta^\dagger(\mathbf{X})$	$\Sigma(\theta_0)$
Plug-in	$\hat\theta$	0
Posterior predictive and discrepancy	$\hat\theta$	$i_{\theta\theta}^{-1}(\theta_0)$
Conditional predictive and approximate conditional predictive	$\hat\theta_{\text{cMLE}}$	$i_{c,\theta\theta}^{-1}(\theta_0) \equiv \{i_{\theta\theta}(\theta_0) - \sigma^{-2}(\theta_0)\nu_\theta(\theta_0)^{\otimes 2}\}^{-1}$
Partial posterior predictive	$\hat\theta_{\text{cMLE}}$	$i_{c,\theta\theta}^{-1}(\theta_0)$
Conditional plug-in	$\hat\theta_{\text{cMLE}}$	0

under $\mathbf{X} \sim f(\mathbf{x}\colon \theta_0)$,

$$p(\mathbf{X}) = 1 - \Phi(Q) + o_p(1), \tag{15}$$

where

$$Q = \frac{\begin{aligned} n^{1/2}\{t(\mathbf{X}, \theta_0) - \nu_n(0, \theta_0\colon \theta_0)\} \\ -\dot\nu_\theta(\theta_0)' n^{1/2}\{\theta^\dagger(\mathbf{X}) - \theta_0\}\end{aligned}}{[\sigma^2(\theta_0) + \dot\nu_\theta(\theta_0)'\boldsymbol{\Sigma}(\theta_0)\dot\nu_\theta(\theta_0)]^{1/2}}. \tag{16}$$

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 $t(\mathbf{X}^{\text{new}}, \theta) > t(\mathbf{X}, \theta)$ is algebraically equivalent to

$$\begin{aligned} &n^{1/2}\{t(\mathbf{X}^{\text{new}}, \theta) - \nu_n(0, \theta_0\colon \theta)\} \\ &\quad - n^{1/2}\{\nu_n(0, \theta\colon \theta_0) - \nu_n(0, \theta_0\colon \theta_0)\} \\ &\quad > n^{1/2}\{t(\mathbf{X}, \theta) - \nu_n(0, \theta_0\colon \theta)\} \\ &\quad - n^{1/2}\{\nu_n(0, \theta\colon \theta_0) - \nu_n(0, \theta_0\colon \theta_0)\}. \end{aligned} \tag{17}$$

Now, by (13) and (14), we can asymptotically ignore all θ that are not within $O(n^{-1/2})$ of θ_0. Hence by (12), the left side of (17) is approximately

$$\begin{aligned} &n^{1/2}\{t(\mathbf{X}^{\text{new}}, \theta_0) - \nu_n(0, \theta_0\colon \theta_0)\} \\ &\quad - n^{1/2}\{\nu_n(0, \theta\colon \theta_0) - \nu_n(0, \theta_0\colon \theta_0)\} \\ &= n^{1/2}\{t(\mathbf{X}^{\text{new}}, \theta_0) - \nu_n(0, \theta\colon \theta_0)\}. \end{aligned}$$

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

$$\sqrt{n}\{t(\mathbf{X}, \theta_0) - \nu_n(0, \theta_0\colon \theta_0)\} - \dot\nu_\theta(\theta_0)'\sqrt{n}(\theta - \theta_0). \tag{18}$$

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

$$\begin{aligned} &\sqrt{n}\{t(\mathbf{X}^{\text{new}}, \theta_0) - \nu_n(0, \theta\colon \theta_0)\} + \dot\nu_\theta(\theta_0)'\sqrt{n}(\theta - \theta_0) \\ &\qquad > \sqrt{n}\{t(\mathbf{X}, \theta_0) - \nu_n(0, \theta_0\colon \theta_0)\}. \end{aligned} \tag{19}$$

Now conditional on $(\mathbf{X}, \theta)$, by (3a), the first term on the left side of (19) is approximately $\mathrm{N}(0, \sigma^2(\theta))$, and, given X, by (13), the second term on the left side of (19) is converging to a $\mathrm{N}(\dot\nu_\theta(\theta_0)' n^{1/2}(\theta^\dagger(\mathbf{X}) - \theta_0), \dot\nu_\theta(\theta_0)'\boldsymbol{\Sigma}(\theta_0)\dot\nu_\theta(\theta_0))$ distribution. Further, because the first term has mean 0 conditional on $(\mathbf{X}, \theta)$, the two terms are conditionally uncorrelated given $\mathbf{X}$. Hence, given $\mathbf{X}$, the left side of (19) is asymptotically

$$\begin{aligned} &\mathrm{N}(\dot\nu_\theta(\theta_0)'\sqrt{n}(\theta^\dagger(\mathbf{X}) - \theta_0), \sigma^2(\theta_0) \\ &\qquad + \dot\nu_\theta(\theta_0)'\boldsymbol{\Sigma}(\theta_0)\dot\nu_\theta(\theta_0)). \end{aligned} \tag{20}$$

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

Under the assumption that the distribution of $\mathbf{X}$ under (ψ_n, θ_0) and $(0, \theta_0)$ are contiguous (which will quite generally be the case), the expansion of Theorem 3 is also valid under $\mathbf{X} \sim f(\mathbf{x}; \psi_n, \theta_0)$. It remains to show that the distribution of Q in (16) converges to the $\mathrm{N}(u(\theta_0), \tau^2(\theta_0))$ distribution given in Theorems 1 and 2. For this, we need the joint distribution of $t(\mathbf{X}, \theta_0)$ and $\theta^\dagger(\mathbf{X})$.

In all of the examples that we consider, we will have for given "influence functions" $B_i(\theta\colon \theta)$ that for all θ in a neighborhood of θ_0, with $\mathbf{X} \sim f(\mathbf{x}; \theta_0)$,

$$\begin{aligned} &n^{1/2}[t(\mathbf{X}, \theta) - \nu_n(0, \theta_0\colon \theta)] \\ &\qquad = n^{-1/2}\sum_i B_i(\theta_0\colon \theta) + o_P(1) \end{aligned} \tag{21}$$

for some mean 0 $B_i(\theta_0\colon \theta) = b_i(X_i, \theta_0\colon \theta)$,

$$n^{1/2}(\hat\theta - \theta_0) = i_{\theta\theta}^{-1}(\theta_0) n^{-1/2}\mathbf{S}_\theta(\theta_0) + o_P(1), \tag{22}$$

$$\begin{aligned} &n^{1/2}(\hat\theta_{\text{cMLE}} - \theta_0) \\ &\qquad = i_{c,\theta\theta}(\theta_0)^{-1}\{n^{-1/2}\mathbf{S}_{c\theta}(\theta_0)\} + o_P(1) \end{aligned} \tag{23}$$

where

$$\mathbf{S}_{c\theta}(\theta_0) \equiv \mathbf{S}_\theta(\theta_0) - \dot\nu_\theta(\theta_0)\sigma^{-2}(\theta_0) n^{1/2} T_{\text{std}}(\theta_0), \tag{24}$$

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

and $i_{c,\theta\theta}(\theta_0)$ 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 $t(\mathbf{X}, \theta)$ and $\theta^\dagger$ 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 $i_{\theta\theta}^{-1}(\theta_0)$ and $i_{c,\theta\theta}(\theta_0)^{-1}$. To obtain the limit distribution under alternatives (ψ_n, θ_0), we make the further assumption

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

that as $n \to \infty$, for $(k_n', k_n^*)' \to (k', k^*)' \equiv h$,

$$\log \frac{f(\mathbf{X}; k_n/\sqrt{n}, \theta_0 + k_n^*/\sqrt{n})}{f(\mathbf{X}; \theta_0)} \\ = h' n^{-1/2} (\mathbf{S}_\psi(\theta_0), \mathbf{S}_\theta(\theta_0)')' - \frac{1}{2} h' i(\theta_0) h + o_P(1) \quad (25)$$

and

$$n^{-1/2} (\mathbf{S}_\psi(\theta_0), \mathbf{S}_\theta(\theta_0)')' \overset{\theta_0}{\rightsquigarrow} \mathrm{N}(0, i(\theta_0)), \qquad (26)$$

where

$$i(\theta_0) = \begin{pmatrix} i_{\psi\psi}(\theta_0) & i_{\psi\theta}(\theta_0)' \\ i_{\psi\theta}(\theta_0) & i_{\theta\theta}(\theta_0) \end{pmatrix}.$$

and we assume the sum on the right side of (21) satisfies the Lindeberg condition. Equations (25)–(26) imply that the model $f(\mathbf{x}; \psi, \theta)$ 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 $t(\mathbf{X}, \theta)$ and model (2a)–(2b), suppose that both (21)–(26) and the assumptions of Theorem 3 hold. Then the following obtain:

a. $$\dot{\nu}_\psi(\theta_0) = \mathrm{cov}_{\theta_0}^A (T_{\mathrm{std}}(\theta_0), n^{-1/2} \mathbf{S}_\psi(\theta_0)), \\ \dot{\nu}_\theta(\theta_0) = \mathrm{cov}_{\theta_0}^A (T_{\mathrm{std}}(\theta_0), n^{-1/2} \mathbf{S}_\theta(\theta_0)), \qquad (27)$$

where $\mathrm{cov}_{\theta_0}^A(S_1, S_2)$ denotes the asymptotic covariance of S_1 and S_2 under $(\psi = 0, \theta_0)$.

b. When $\mathbf{X} \sim f(\mathbf{x}; k_n/\sqrt{n}, \theta_0)$, $(T_{\mathrm{std}}(\theta_0), \dot{\nu}_\theta(\theta_0)' n^{1/2} (\hat{\theta} - \theta_0))$ converges to a normal distribution with mean $k(\dot{\nu}_\psi(\theta_0), \dot{\nu}_\theta(\theta_0)' i_{\theta\theta}^{-1}(\theta_0)$ $i_{\psi\theta}(\theta_0))$ and covariance matrix

$$\begin{pmatrix} \sigma^2(\theta_0) & \Omega(\theta_0) \\ \Omega(\theta_0) & \Omega(\theta_0) \end{pmatrix}, \qquad (28)$$

and thus $T_{\mathrm{std}}(\theta_0) - \dot{\nu}(\theta_0)' n^{1/2} \{\hat{\theta} - \theta_0\}$ converges to a $\mathrm{N}(k\omega(\theta_0), \ \sigma^2(\theta_0) - \Omega(\theta_0))$ distribution. Further, $(T_{\mathrm{std}}(\theta_0), \dot{\nu}_\theta(\theta_0)' n^{1/2} (\hat{\theta}_{\mathrm{cMLE}} - \theta_0))$ converges to a normal distribution with mean $k(\dot{\nu}_\psi(\theta_0), \dot{\nu}_\theta(\theta_0)' i_{c,\theta\theta}^{-1}(\theta_0)$ $i_{c,\psi\theta}(\theta_0))$ and covariance matrix

$$\begin{pmatrix} \sigma^2(\theta_0) & 0 \\ 0 & \Omega_c(\theta_0) \end{pmatrix}, \qquad (29)$$

where $i_{c,\psi\theta}(\theta_0) = \mathrm{cov}_{\theta_0}^A [n^{-1/2} \mathbf{S}_\psi(\theta_0), n^{-1/2} \mathbf{S}_{c\theta}(\theta_0)]$ and $\Omega_c(\theta_0) = \dot{\nu}_\theta(\theta_0)' i_{c,\theta\theta}^{-1}(\theta_0) \dot{\nu}_\theta(\theta_0)$. Thus $T_{\mathrm{std}}(\theta_0) - \dot{\nu}_\theta(\theta_0)' n^{1/2} \{\hat{\theta}_{\mathrm{cMLE}} - \theta_0\}$ converges to a $\mathrm{N}(k\omega_c(\theta_0), \sigma^2(\theta_0) + \Omega_c(\theta_0))$ distribution, where $\omega_c(\theta_0) = \dot{\nu}_\psi(\theta_0) - \dot{\nu}_\theta(\theta_0)' i_{c,\theta\theta}^{-1}(\theta_0) i_{c,\psi\theta}(\theta_0)$.

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

$$n^{-1/2} \mathbf{S}_{c\theta}(\theta_0) = n^{-1/2} \mathbf{S}_\theta(\theta_0) - \mathrm{cov}_{\theta_0}^A (\mathbf{S}_\theta(\theta_0), T_{\mathrm{std}}(\theta_0)) \\ \times \{\mathrm{var}_{\theta_0}^A (T_{\mathrm{std}}(\theta_0))\}^{-1} T_{\mathrm{std}}(\theta_0).$$

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_{acpred} substituted for p_{cpred} 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 $\dot{\nu}_\theta(\theta_0)' n^{1/2} (\hat{\theta} - \theta_0)$ 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 $\dot{\nu}_\theta(\theta_0)$ is nonzero.] Furthermore, it is immediate from Theorems 3 and 4 that expansion (16) holds with $Q \sim \mathrm{N}(k\omega_c(\theta_0)/\{\sigma^2(\theta_0) + \Omega_c(\theta_0)\}^{1/2}, 1)$ for $p_{\mathrm{ppost}}(\mathbf{X})$ and $p_{\mathrm{acpred}}(\mathbf{X})$ and $Q \sim \mathrm{N}(k\omega_c(\theta_0)/\sigma(\theta_0), \{\sigma^2(\theta_0) + \Omega_c(\theta_0)\}/\sigma^2(\theta_0))$ for $p_{\mathrm{cplug}}(\mathbf{X})$. But some algebra shows that $\{\sigma^2(\theta_0) + \Omega_c(\theta_0)\}/\sigma^2(\theta_0) = \sigma^2(\theta_0)/\{\sigma^2(\theta_0) - \Omega(\theta_0)\}$ and $\omega_c(\theta_0)/\{\sigma^2(\theta_0) + \Omega_c(\theta_0)\}^{1/2} = \mathrm{NC}(\theta_0)$, which proves the corollary.

To complete the proof of Theorem 1, it only remains to show that $p_{\mathrm{cpred}}(\mathbf{X})$ and $p_{\mathrm{acpred}}(\mathbf{X})$ have the same limiting distribution. The key observation is that, as discussed earlier, $\hat{\theta}_{\mathrm{cMLE}}$ and $t(\mathbf{X})$ are asymptotically uncorrelated, so that in large samples the conditional distribution given $\hat{\theta}_{\mathrm{cMLE}}$ and unconditional distribution of $t(\mathbf{X})$ 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,

$$\sup_{|\theta - \theta_0| \le (c/\sqrt{n})} \sup_t |\Pr\{t(\mathbf{X}) \le t | \hat{\theta}_{\mathrm{cMLE}}; \theta\} \\ - \Pr\{t(\mathbf{X}) \le t; \theta\}| \overset{P_{\theta_0}}{\to} 0,$$

then $p_{\mathrm{cpred}}(\mathbf{X})$ and $p_{\mathrm{acpred}}(\mathbf{X})$ have the same limiting distribution under $f(\mathbf{x}; k_n/\sqrt{n}, \theta_0)$.

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 $T = t(\mathbf{X})$ satisfies (3), then we would expect the marginal model $f_T(t; \ \psi, \ \theta) \equiv f(t; \ \psi, \ \theta)$ to be LAN. That is, under $T \sim$

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

$$f(t;\theta_0) \equiv f(t;0,\theta_0),$$

$$\log\frac{f(T;k_n/\sqrt{n},\theta_0+k_n^*/\sqrt{n})}{f(T;0,\theta_0)} = h'\dot{\nu}(\theta_0)\sigma^{-2}(\theta_0)T_{\mathrm{std}}(\theta_0) - \frac{1}{2}\,[h'\dot{\nu}(\theta_0)]^2/\sigma^2(\theta_0) + o_P(1). \quad (30)$$

Together with the similar expansion (25) for the unconditional model for $\mathbf{X}$, we obtain, on noting $f(\mathbf{X}) = f(\mathbf{X}|T)f(T)$, that

$$\log\frac{f(\mathbf{X}|T;k_n/\sqrt{n},\theta_0+k_n^*/\sqrt{n})}{f(\mathbf{X}|T;0,\theta_0)} = h'n^{-1/2}\mathbf{S}_c(\theta_0) - \frac{1}{2}h'i_c(\theta_0)h + o_P(1), \quad (31)$$

where

$$S_c(\theta_0) \equiv (S_{c\psi}(\theta_0), S_{c\theta}(\theta_0)')',$$

$$S_{c\psi}(\theta_0) = S_\psi(\theta_0) - \dot{\nu}_\psi(\theta_0)\sigma^{-2}(\theta_0)n^{1/2}T_{\mathrm{std}}(\theta_0)$$

and

$$\mathbf{i}_c(\theta_0) = \begin{pmatrix} i_{c\psi\psi}(\theta_0) & i_{c\psi\theta}(\theta_0)' \\ i_{c\psi\theta}(\theta_0) & i_{c\theta\theta}(\theta_0) \end{pmatrix}$$

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

Note that if $n^{1/2}\{T - \nu_n(k_n/\sqrt{n},\theta_0 + k_n^*/\sqrt{n})\}$ were exactly distributed $\mathrm{N}(0,\sigma^2(\theta_0))$ under $f_T(t;k_n/\sqrt{n},\theta_0 + k_n^*/\sqrt{n})$ with $\nu_n(k/\sqrt{n},\theta_0 + k_n^*/\sqrt{n}) = \nu_n(0,\theta_0) + \dot{\nu}_\psi(\theta_0)k_n/\sqrt{n} + \dot{\nu}_\theta(\theta_0)'k_n^*/\sqrt{n}$, 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 $\hat{\theta}_{\mathrm{cMLE}}$ maximizing $f(\mathbf{X}|T;\theta)$ 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 these results is a proof of $\sqrt{n}$ consistency of $\hat{\theta}_{\mathrm{cMLE}}$ 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 $t(\mathbf{X}) = n^{-1}\mathbf{S}_\psi(\theta^*)$ because, from its definition, $\mathrm{NC}(\theta^*)$ will be the same for a given statistic $t_1(\mathbf{X})$ and all affine transformations of $t_1(\mathbf{X}), t(\mathbf{X}) = at_1(\mathbf{X}) + b + o_P(1)$, with $a \neq 0$. Now, by Theorem 4, for $t(\mathbf{X}) = n^{-1}S_\psi(\theta^*)$,

$$\dot{\nu}_\psi(\theta^*) = i_{\psi\psi}(\theta^*)$$

and

$$\dot{\nu}_\theta(\theta^*) = i_{\psi\theta}(\theta^*),$$

which proves the lemma.

APPENDIX A: PROOF OF THEOREM 3

By (14), we have that with $\mathbf{X} \sim f(\mathbf{x};\theta_0)$, for all $\varepsilon > 0$, there exists a constant c_ε such that

$$\Pr_{\theta_0}\{E_{\mathrm{N}(\theta^\dagger(\mathbf{X}),\boldsymbol{\Sigma}(\theta_0)/n)}[I\{\|\theta-\theta_0\| \le c_\varepsilon/\sqrt{n}\}|\mathbf{X}] \ge 1-\varepsilon\} \ge 1-\varepsilon, \quad (\mathrm{A.1})$$

where $E_{\mathrm{N}(\mu,\Sigma)}$ refers to expectation with respect to a normal distribution with mean μ and variance matrix $\boldsymbol{\Sigma}$. Equation (A.1) says that with large probability, X is such that when $\theta \sim \mathrm{N}(\theta^\dagger(\mathbf{X}),\boldsymbol{\Sigma}(\theta_0)/n), \theta$ lies in the ball of radius $c_\varepsilon/\sqrt{n}$ around θ_0 with high probability.

Now, because the total variation norm $\|P-Q\|$ is also equal to $2\sup_f\{|\int f\,dP - \int f\,dQ|: 0 \le f \le 1\}$ and $\theta \mapsto \Pr(t(\mathbf{X}^{\mathrm{new}},\theta) \le t(\mathbf{X},\theta)|\mathbf{X};\theta)$ is uniformly bounded by 1, we have, by (13), that

$$p(\mathbf{X}) = \int_\Theta \Pr[t(\mathbf{X}^{\mathrm{new}},\theta) > t(\mathbf{X},\theta)|\mathbf{X};\theta]\phi(\theta;\theta^\dagger(\mathbf{X}),\boldsymbol{\Sigma}(\theta_0)/n)\,d\theta + o_{P_{\theta_0}}(1), \quad (\mathrm{A.2})$$

where $\phi(\theta;\mu,\boldsymbol{\Sigma})$ is the density of a $\mathrm{N}(\mu,\boldsymbol{\Sigma})$ random variable and, in (A.2), 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 $E_{\mathrm{N}(\theta^\dagger(\mathbf{X}),\boldsymbol{\Sigma}(\theta_0)/n)}[I\{\|\theta-\theta_0\| > c_\varepsilon/\sqrt{n}\}|\mathbf{X}]$, which, with probability at least $1-\varepsilon$, is less than ε. Let A_ε be the event on which this is true, so that $\Pr_{\theta_0}(A_\varepsilon) \ge 1-\varepsilon$. We can write the integrand as Pr [Eq. (17)$|\mathbf{X};\theta$]. By (12), we have that for $\theta_n = \theta_0 + O(1/\sqrt{n}), n^{1/2}(t(\mathbf{X}^{\mathrm{new}},\theta_n) - \nu_n(0,\theta_0\colon \theta_n)) - n^{1/2}(\nu_n(0,\theta_n\colon \theta_0) - \nu_n(0,\theta_0\colon \theta_0)) = n^{1/2}(t(\mathbf{X}^{\mathrm{new}},\theta_0) - \nu_n(0,\theta_n\colon \theta_0)) + o_{P_{\theta_0}}(1)$. By the contiguity assumption in (3), this is true also for the remainder term $o_{P_{\theta_n}}(1)$. By (3a), $1/\sigma(\theta_0)$ times the right side of the last equality is asymptotically standard normal under θ_n. Thus for every c,

$$\sup_{|\theta-\theta_0|\le(c/\sqrt{n})}\ \sup_t \Big|\Pr(n^{1/2}\{t(\mathbf{X}^{\mathrm{new}},\theta) - \nu_n(0,\theta_0\colon\ \theta)\} - n^{1/2}\{\nu_n(0,\theta\colon\ \theta_0) - \nu_n(0,\theta_0\colon\ \theta_0)\} \le t;\theta) - \Phi(t/\sigma(\theta_0))\Big| \to 0.$$

Next, by (3b) holding at θ_0 and (12), for every $\tilde{\theta}_n = \theta_0 + O_P(1/\sqrt{n})$,

$$\Bigg|\Phi\left(\frac{n^{1/2}(t(\mathbf{X},\tilde{\theta}_n) - \nu_n(0,\theta_0\colon\ \tilde{\theta}_n)) - n^{1/2}(\nu_n(0,\tilde{\theta}_n\colon\ \theta_0) - \nu_n(0,\theta_0\colon\ \theta_0))}{\sigma(\theta_0)}\right)$$

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

$$- \Phi\left(\frac{n^{1/2}(t(\mathbf{X},\theta_0) - \nu_n(0,\theta_0\colon \theta_0)) - \dot{\nu}_\theta(\theta_0)' n^{1/2}(\tilde{\theta}_n - \theta_0)}{\sigma(\theta_0)}\right)\Bigg|$$

$$\le \big|n^{1/2}(t(\mathbf{X},\tilde{\theta}_n) - \nu_n(0,\theta_0\colon \tilde{\theta}_n)) - n^{1/2}(t(\mathbf{X},\theta_0) - \nu_n(0,\theta_0\colon \theta_0))\big|\frac{\|\phi\|_\infty}{\sigma(\theta_0)} + \big|n^{1/2}(\nu_n(0,\tilde{\theta}_n\colon \theta_0) - \nu_n(0,\theta_0\colon \theta_0)) - \dot{\nu}_\theta(\theta_0)' n^{1/2}(\tilde{\theta}_n - \theta_0)\big|\frac{\|\phi\|_\infty}{\sigma(\theta_0)} \overset{P_{\theta_0}}{\to} 0,$$

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

$$\sup_{\|\theta-\theta_0\|\le(c_\varepsilon/\sqrt{n})} \Bigg|\Pr(t(\mathbf{X}^{\text{new}},\theta) \le t(\mathbf{X},\theta)|\mathbf{X};\theta) - \Phi\left(\frac{n^{1/2}(t(\mathbf{X},\theta_0) - \nu_n(0,\theta_0\colon \theta_0)) - \dot{\nu}_\theta(\theta_0)' n^{1/2}(\theta - \theta_0)}{\sigma(\theta_0)}\right)\Bigg| \overset{P_{\theta_0}}{\to} 0.$$

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

$$\Bigg|1 - p(\mathbf{X}) - \int \Phi\left(\frac{n^{1/2}(t(\mathbf{X},\theta_0) - \nu_n(0,\theta_0\colon \theta_0)) - \dot{\nu}_\theta(\theta_0)' n^{1/2}(\theta - \theta_0)}{\sigma(\theta_0)}\right) \times \phi(\theta;\theta^\dagger(\mathbf{X}),\boldsymbol{\Sigma}(\theta_0)/n)\,d\theta$$
$$\le 2E_{\mathrm{N}(\theta^\dagger(\mathbf{X}),\boldsymbol{\Sigma}(\theta_0/n))}[I\{\|\theta-\theta_0\| > c_\varepsilon/\sqrt{n}|X\}] + o_{P_{\theta_0}}(1) \le 2\varepsilon I(X \in A_\varepsilon) + I(X \notin A_\varepsilon) + o_{P_{\theta_0}}(1).$$

This being true for every $\varepsilon > 0$ implies that $1 - p(\mathbf{X})$ is asymptotically equivalent to

$$\int \Phi\left(\frac{n^{1/2}(t(\mathbf{X},\theta_0) - \nu_n(0,\theta_0\colon \theta_0))}{\sigma(\theta_0)} - \dot{\nu}_\theta(\theta_0)'\sigma^{-1}(\theta_0) n^{1/2}(\theta - \theta_0)\right) \times \phi(\theta;\theta^\dagger(\mathbf{X}),\boldsymbol{\Sigma}(\theta_0)/n)\,d\theta = \Phi(Q),$$

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 $T = t(\mathbf{X})$. Furthermore, suppose that under $f(\mathbf{x}; k/\sqrt{n}, \theta_0 + k^*/\sqrt{n})$,

$$\sqrt{n}\{T - \nu_n(k/\sqrt{n}, \theta_0 + k^*/\sqrt{n})\} \tag{B.1}$$

converges in variation distance to a $\mathrm{N}(0,\sigma^2(\theta_0))$ distribution for all $k \in R^1, k^* \in R^P$. 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 θ_0 known. Details will be presented elsewhere.

If our statistic T equals $n^{-1}\sum_i d(X_i)$, 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 general h, (B.1) will be true if we make these conditions uniform in θ running through a neighborhood of θ_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 that 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 θ_0 being known, and $k \in R^1$ and $k^* \in R^p$. 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 θ, then observing T_n is statistically equivalent to observing a smoothed 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 $T_n = n^{-1}\sum_i d(X_i)$ with $d(X_i)$ taking its values in a grid of points $\dots, a - s, a, a + s, a + 2s, \dots$ 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 θ.

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

The proof will be given elsewhere.

APPENDIX C: CONDITIONAL INFERENCE IN EXAMPLE B OF SECTION 2

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

For simplicity and without loss of generality, we consider the case where the variance c^2 is known and equal to 1, so $\theta = \mu$. Thus $X_1,\dots,X_n$ are iid $\mathrm{N}(\theta,1)$. Let $\hat{Z}_q \equiv \hat{Z}_{q_n}$ be the nq_n-order statistic where $q_n \to q$ and nq_n is an integer. Then data from $\mathbf{X}|\hat{Z}_{q_n} = x$ can be generated by generating iid data $Y_1,\dots,Y_{nq_n-1}$ from the density $I(y \le x)\phi(y-\theta)/\Phi(x-\theta)$ and then generating $Y_{nq_n+1},\dots,Y_n$ iid from the density $I(y > x)\phi(y-\theta)/\{1-\Phi(x-\theta)\}$ independently of the previous Y_i. Thus the likelihood function is proportional to

$$\prod_{i=1}^{nq_n-1}\phi(Y_i-\theta)/\Phi(x-\theta)\prod_{i=nq_n+1}^{n}\phi(Y_i-\theta)/\{1-\Phi(x-\theta)\}. \tag{C.1}$$

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

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

$$\Psi(\hat{\theta}_{\text{cMLE}}, X_{(nq_n)}) = 0,$$

where $X_{(j)}$ is the jth-order statistic and

$$\Psi(\theta, x) = n^{-1}\frac{\partial}{\partial\theta}\sum_{i=1}^{nq_n-1}\log\phi(X_{(i)}-\theta) - (nq_n-1)\log\Phi(x-\theta)$$

$$+ \sum_{i=nq_n+1}^{n}\log\phi(X_{(i)}-\theta) - (n-nq_n)\log(1-\Phi(x-\theta)). \quad \text{(C.2)}$$

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

$$n^{-1}\sum_{i=1}^{n}(X_{(i)}-\theta) - \frac{\Phi(x-\theta)-q}{\Phi(x-\theta)(1-\Phi(x-\theta))}\phi(x-\theta). \quad \text{(C.3)}$$

The approximation in (C.3) is obtained by adding $X_{(nq_n)} - \theta$ and replacing $nq_n - 1$ by nq_n.

By Taylor expansion of the score equation, we obtain

$$n^{1/2}(\hat{\theta}_{\text{cMLE}} - \theta) = -n^{1/2}\Psi(\theta, X_{(nq_n)})/\dot{\Psi}(\tilde{\theta}, X_{(nq_n)}) \quad \text{(C.4)}$$

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

$$\begin{aligned} n^{1/2}\Psi(\theta, X_{(nq_n)}) &+ o_P(1) \\ &= n^{-1/2}\sum_{i=1}^{n}(X_i-\theta) - \{q(1-q)\}^{-1}\phi^2(z_q)n^{1/2}(\hat{Z}_q - z_q(\theta)) \\ &= n^{-1/2}\mathbf{S}_\theta(\theta) - \dot{\nu}_\theta(\theta)\sigma^{-2}(\theta)T_{\text{std}}(\theta) \end{aligned}$$

and $\dot{\Psi}(\tilde{\theta}, X_{(nq_n)}) \xrightarrow{P} -i_{c,\theta\theta}(\theta)$, as required.

Proof of (13)–(14) for $\pi_{\text{ppost}}(\cdot|\mathbf{X})$ in Example B of Section 3

The posterior density for θ based on the conditional model f $(\mathbf{x}|T = x;\theta)$ with $T = \hat{Z}_q = X_{(nq_n)}$ is, by (C.1),

$$\frac{\prod_{i=1}^{nq_n-1}\phi(X_{(i)}-\theta)\prod_{i=nq+1}^{n}\phi(X_{(i)}-\theta)\pi(\theta) \times \Phi^{-(nq_n-1)}(x-\theta)\{1-\Phi^{n-(nq_n+1)}(x-\theta)\}^{-1}}{\int\prod_{i=1}^{nq_n-1}\phi(X_{(i)}-\theta)\prod_{i=nq+1}^{n}\phi(X_{(i)}-\theta)\pi(\theta) \times \Phi^{-(nq_n-1)}(x-\theta)\{1-\Phi^{n-(nq_n+1)}(x-\theta)\}^{-1}\,d\theta}. \quad \text{(C.5)}$$

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

$$\begin{aligned} &\|\pi(\cdot|X_{(1)},\ldots,X_{(nq_{n-1})},X_{(nq_n+1)},\ldots,X_{(n)},x) \\ &\qquad - \mathrm{N}(\Delta(\theta,x), n^{-1}\mathbf{\Sigma}(\theta,x))\| \xrightarrow{P_x} 0, \quad \text{(C.6)} \end{aligned}$$

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

$$\begin{aligned} &E(\|\pi(\cdot|X_1,\ldots,X_n,X_{(nq_n)}) \\ &\qquad - \mathrm{N}(\Delta(\theta,x_n), n^{-1}i(\theta,x_n))\||X_{(nq_n)} = x_n) \xrightarrow{P} 0. \end{aligned}$$

This is sufficient to conclude that

$$\begin{aligned} &E(\|\pi(\cdot|X_1,\ldots,X_n,X_{(nq_n)}) \\ &\qquad - \mathrm{N}(\Delta(\theta,X_{(nq_n)}), n^{-1}i(\theta,X_{(nq_n)}))\||X_{(nq_n)}) \xrightarrow{P} 0. \end{aligned}$$

By dominated convergence, this gives

$$\begin{aligned} &\|\pi(\cdot|X_1,\ldots,X_n,X_{(nq_n)}) \\ &\qquad - \mathrm{N}(\Delta(\theta,X_{(nq_n)}), n^{-1}i(\theta,X_{(nq_n)}))\| \xrightarrow{P} 0. \end{aligned}$$

However, by (C.4), we can substitute $\hat{\theta}_{\text{cMLE}}$ for $\Delta(\theta, X_{(nq_n)})$ and $i_{c,\theta\theta}(\theta)$ for $i(\theta, X_{(nq_n)})$, concluding the proof.

[Received January 1999. Revised November 1999.]

REFERENCES

Bayarri, M. J., and Berger, J. O. (1999), "Quantifying Surprise in the Data and Model Verification," in *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 53–67.

——— (2000), "*P* Values in Composite Null Models," *Journal of the American Statistical Association*, 95, 1127–1142.

Beran, R. J. (1988), "Pre-Pivoting Test Statistics: A Bootstrap View of Asymptotic Refinements," *Journal of the American Statistical Association*, 83, 687–697.

Box, G. E. P. (1980), "Sampling and Bayes Inference in Scientific Modeling and Robustness," *Journal of the Royal Statistical Society*, Ser. A, 143, 383–430.

Davison, A. C., and Hinkley, D. V. (1997), *Bootstrap Methods and Their Application*, Cambridge, U.K.: Cambridge University Press.

Evans, M. (1997), "Bayesian Inference Procedures Derived via the Concept of Relative Surprise," *Communications in Statistics*, 26, 125–143.

Feller, W. (1971), *An Introduction to Probability Theory and Its Applications*, Vol. 2, New York: Wiley.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995), *Bayesian Data Analysis*, New York: Wiley.

Gelman, A., Meng, X. L., and Stern, H. (1996), "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies" (with discussion), *Statistica Sinica*, 6, 733–807.

Ghosh, J. K. (1994), "Higher-Order Asymptotics," NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 4, Hayward, CA: Institute of Mathematical Statistics.

Guttman, I. (1967), "The Use of the Concept of a Future Observation in Goodness-of-Fit Problems," *Journal of the Royal Statistical Society*, Ser. B, 29, 83–100.

LeCam, L. (1986), *Asymptotic Methods in Statistical Decision Theory*, New York: Springer-Verlag.

LeCam, L., and Yang, G. (1988), "On the Preservation of Local Asymptotic Normality Under Information Loss," *The Annals of Statistics*, 16, 483–520.

Meng, X. L. (1994), "Posterior Predictive *p* Values," *The Annals of Statistics*, 22, 1142–1160.

Newton, M. A., and Geyer, C. J. (1994), "Bootstrap Recycling: A Monte Carlo Alternative to the Nested Bootstrap," *Journal of the American Statistical Association*, 89, 905–912.

Pauler, D. (1999), Discussion of "Quantifying Surprise in the Data and Model Verification" by M. J. Baylarri and J. O. Berger in *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 70–73.

Robins, J. M. (1999), Discussion of "Quantifying Surprise in the Data and Model Verification" by M. J. Baylarri and J. O. Berger in *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 67–70.

Rubin, D. B. (1984), "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician," *The Annals of Statistics*, 12, 1151–1172.

Stigler, S. M. (1977), "Do Robust Estimators Work With Real Data," *The Annals of Statistics*, 5, 1055–1077.

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge, U.K.: Cambridge University Press.

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