베이지안 방법을 포함한 일반적 통계 추론에 대한 상관모의를 이용한 평가방법

EVALUATION OF FREQUENTIST AND BAYESIAN INFERENCES BY RELEVANT SIMULATION

김 윤 태*

<초 록>

현실적으로 통계추론 방법의 적용시, 그 정당성이 보장되는 기본가정이외에도 추가적인 가정이 불가피하여, 본래의 정당성이 퇴색되는 경우가 흔히 발생한다. 따라서 이런 경우에는 통계추론의 평가가 필수적일 것이나, 많은 경우에 분석적 평가를 하기에는 너무 복잡하여, 특정상황을 상정한 모의분석 평가가 주류를 이루고 있다.

본 고에서는 보다 일반적 상황에서의 통계추론의 평가를 위해 브트스트랩방법과 같이 관찰값에 의존한 모의방법(observation-based simulation)을 이용한 평가방법을 제안한다. 우선 설득력 있는 평가요소로서 구간추정시 포함확률(coverage probability)와 같은 빈도성질(frequency property)를 선택하였다. 빈도성질은 고전적 통계추론은 물론 베이지안 통계추론을 대상으로도 의미있는 평가기준으로 판단되는 바, 이를 평가요소로서 선택하고, 이의 추정을 위한 방법과, 그 추정결과의 해석과 나아가 이를 기준으로 한 통계추론 결과의 조정 방법까지 일련의 절차에 대한 방법론을 제시하였다.

* 한국 국방 연구원, ytkim64@hanmir.com

I . Overview

A reasonable statistical inference procedure would be justified by some validity and performance proved under some basic assumptions. But in many real situations for statistical procedures, the validity and performance are affected by further assumptions and decisions inevitably involved in the procedures. Here are some examples:

• The classical strategy in the regression model having a suite of highly correlated covariates is to select a subset model, and then make the inference assuming the subset model. We call this kind of method a selection and estimation (S/E) procedure. All the justification of the inferences coming from the subset model can not be guaranteed since it ignores the uncertainty in the selected subset model;

• Inferences based on asymptotic theory could be justified only for infinitely many samples, but the real situation always involves finite samples;

• In a parametric model with nuisance parameters, an estimate of the nuisance parameter is often substituted for true value in the inference procedure. Ignoring the uncertainty in the estimated nuisance values could invalidate the inference;

• The performance of ridge estimators depends critically on the choice of the ridge constant; there is as yet no consensus as to an optimal choice; and

• More generally, the repeated sampling properties of Bayesian procedures depend on the choice of the prior distribution; data-dependent priors for pseudo-Bayes methods have been proposed for certain problems, but no general theory exists.

When we get an inference result from such statistical procedures, we may want to check whether or not the additional assumptions and decisions distort validity and performance of the inference. However, the situation can be too complicated to get an analytical evaluation for the inference in many cases. For example, the complicated selection step in the S/E procedure for regression modeling, makes it impossible to get an analytical evaluation tool. Simulation study could be an alternative way of evaluating such a complicated inference procedure, but the

evaluations coming from general simulation study would only apply to the specific simulated circumstances. For the simulation in the parametric model, the simulation result may depend on the chosen parameter value.

We suggest an evaluation methodology relying on an observation-based simulation for general frequentist and Bayesian inferences on parametric model. The suggested methodology can be applied to any inference procedure, no matter how complicated, as long as it can be codified for repetition on the computer. Unlike general simulation, the suggested methodology provides the evaluation result which does not depend on the parameter value.

Generally speaking, evaluation of an object has three main components:

- Which "factor" of the object is interesting for evaluation?
- How does one "measure" the factor?
- How does one "evaluate" the object based on the measurement for the factor?

Here are our suggestions for these components for the evaluation of the general inference. We also provide a method for fixing up the inference methods found to be invalid under the suggested evaluation methodology.

Our suggestion for the factors of interest are the ideal sampling properties of the inference. For examples, a valid confidence interval guarantees the coverage probability of including true parameter in the sampling space is larger than the nominal value. The justification for the P-value of the frequentist test of a simple null hypothesis is that repeated samplings of P-value under the null follow uniform distribution over (0,1). We will adopt this frequentist criterion for both frequentist and Bayesian inferences. It seems natural to expect the frequentist inferences to have the ideal sampling properties. Though Bayesian inference procedures are not designed for ideal sampling properties, it is also expected to satisfy them well in order to be a reasonable statistical framework.

All the interesting ideal sampling properties of inferences could be expressed as the conditions for parameter dependent frequentist risks over a portion of parameter space. For example, achievement of the nominal coverage probability of a ν-level confidence interval $ci_{\nu}(X)$ for some interesting function phi(θ) of the parameter with parameter space Θ can be expressed as

$$E(I[ci_{\nu(X)} \ni \phi(\theta)] \mid \theta = \theta^{*}) \geq \nu, \forall \theta^{*} \in \Theta.$$

Relevant simulation methodology(RSM) is suggested for parameter free measurement for the risks of interesting ideal sampling properties. The main idea of RSM is to integrate out the parameter in the risk with a relevant measure of parameter to common sense and observations. Our choice for the relevant measure is based on posterior probability of parameter with non-informative prior. The basic idea of RSM, emphasizing the relevant region of parameter to the common sense and observation is not a new one. There are many simulation studies emphasizing the simulation results on the parameter region the researcher feels relevant. Bootstrap is also a well-known observation-based simulation methodology. The Bayesian procedures integrating over the parameter space with posterior distribution of parameter seem to be closer to the our RSM. The Bayesian posterior predictive distribution is an example of such Bayesian procedures. Another example can be found in the construction of empirical Bayesian (EB) confidence intervals.

The simulation-based RSM could provide not only an estimate for the risk, but also a measure of precision, such as standard error. Using these quantities, formal inference procedures like confidence intervals and hypothesis tests could be provided to determine whether the risk condition is satisfied.

Additionally, we provide the adjustment methods for the inferences founded to be invalid with respect to an ideal sampling property by the suggested evaluation scheme. Details of the methods for fixing up invalid P-values and interval estimators to have their ideal sampling properties are provided. We will call the presented evaluation methodology for general inference "evaluation scheme using RSM".

Ⅱ. Ideal sampling properties

As explained in Chapter Ⅰ, the first component in general evaluation scheme is the ""factor"" to be measured for the evaluation. The "ideal sampling property" of inference is chosen to be the "factor" for the evaluation of general inference. We will define the "ideal sampling properties" of inferences with a unified format, i.e. some conditions in terms of the parameter-dependent frequentist risk in the joint probability space of parameter and data. And also some rationale for choosing the "ideal sampling property" as the "factor" will be provided.

There are some ideal sampling properties for each inference which we are expecting the inference to have. For convenience, we categorize the ideal sampling properties into two kinds: "validity properties" and "performance properties". A reasonable inference should be justified to be valid in a sense and also have enough performance for obtaining a meaningful information.

Before going further, we may need some set-up and definitions of terms. Suppose x_{obs} is an observation of the n-dimensional random variable X following the very general parametric distribution having density $f(x \mid \theta)$, where θ is r-dimensional unknown parameter in parameter space Θ. Let's assume we are interested in the quantity expressed as a function of θ,

$$\phi(\theta): \Theta \rightarrow \Phi,$$

where Φ is a subspace of s-dimensional real space R^s. For example, in an epidemiological regression model for identifying the risk components to the mortality, $\phi(\theta)$ could be the coefficients of the potential risk components, or the relative risk which is a real valued function of the coefficients, or the future value of the mortality.

The statistical inference $\psi(X)$ for $\phi(\theta)$ can be defined as a random function from the sample space X to the action space A,

$$\psi(X): X \rightarrow A.$$

The inference types; point estimation, set (or interval) estimation, and hypothesis test can be defined by the specific action spaces as follows:

·Point estimation $\psi_P(X)$:

$$\psi_P(X): X \rightarrow \Phi;$$

·Set (or interval) estimation $\psi_I(X)$:

$$\psi_I(X): X \rightarrow \{\text{subset of } \Phi\}.$$

For convenience, $\phi_I(X)$ is assumed to be the interval estimator having $\phi_I^L(X)$ and $\phi_I^U(X)$ for the lower and upper bound respectively;

·Hypothesis test resulting in P-value or posterior mean $\phi_T(X)$:

$$\phi_T(X): X \rightarrow [0,1];$$

and

·Hypothesis test resulting in the decision about the rejection for null hypothesis $\phi_V(X)$:

$$\phi_V(X): X \rightarrow \{0,1\}.$$

Here 0 or 1 represents rejection or the acceptance of the null hypothesis respectively.

The two major schools of statistics, frequentist and Bayesian have their own suggestions for the inferences. Here are the notations and definitions of the frequentist and Bayesian inferences for $\phi(\theta)$. The definitions of inferences shown here are minimum requirements rather than rigorous definitions. Remember that frequentists consider θ as a constant, while Bayesians treat it as a random variable:

· $pe(X)$: frequentist general point estimator having small risk defined by

$$E(L(pe(X), \phi(\theta))), \forall \theta \in \Theta,$$

where $L(\cdot)$ is an appropriate loss function;

· $ci_\nu(X)$: frequentist nu-level confidence interval such that

$$\Pr(ci_\nu(X) \ni \phi(\theta)) \geq \nu, \forall \theta \in \Theta.$$

For convenience, we will assume that $ci_\nu(X)$ is the interval (not the general region) having lower and upper bound denoted by $ci_\nu^L(X)$ and $ci_\nu^U(X)$ respectively;

· $\underline{pv}(X)$: frequentist P-value for the hypothesis test of $H_0: \phi(\theta) \in \Phi_0$. Generally P-value for a sample point x is the smallest test size for which the sample point will lead to the rejection of H_0. Specifically for a family of tests with level-α rejection region S_α satisfying

(a) $\exists\, \Theta^* \subset \Theta_0 = \{\theta \mid \phi(\theta) \in \Phi_0\} s.t. \Pr_\theta(X \in S_\alpha) = \alpha, \forall\, \theta \in \Theta^*$,

(b) $S_\alpha \subset S_{\alpha'}, \forall\, \alpha, \alpha' \in (0,1)$ and $\alpha \leq \alpha'$,

the P-value is defined by

$$\underline{pv}(X) = \inf\{\alpha : X \in S_\alpha\};$$

For example, suppose $X = (x_1, \cdots, x_n)$ is a random sample from $N(\mu, \sigma^2)$. The usual T test for the hypothesis $H_0: \mu \leq \mu_0, H_1: \mu > \mu_0$ is defined by the rejection region $S_\alpha = \{X \mid t(X) > T_{\alpha, n-1}\}$, where $t(X) = \dfrac{\overline{X} - \mu_0}{S(X)/\sqrt{n}}$, $T_{\alpha, n-1}$ is the $(1-\alpha)$ quantile of T-distribution with n-1 degree of freedom. The S_α in this test satisfies the two conditions with $\Theta^* = \{(\mu, \sigma^2) \mid \mu = 0\}$, so the $\underline{pv}(X)$ can be defined as $\inf\{\alpha : X \in S_\alpha\}$.

· $ht_\alpha(X)$: frequentist hypothesis test decision function with size α defined by

$$ht_\alpha(X) = I[\underline{pv}(X) > \alpha],$$

where $I[\cdot]$ is the identity function. $ht(x) = 0$ leads to rejection of the null hypothesis;

· $\underline{pb}(X)$: Bayesian general point estimator having small expected risk defined by

$$\underline{pb}(X) = E(L(\underline{pe}(X), \phi(\theta)) \mid X).$$

The optimal Bayesian point estimator using a squared loss is the posterior mean,

$$\underline{pm}(X) = E(\phi(\theta) \mid X);$$

· $cr_\nu(X)$: Bayesian ν-level credible interval s.t.

$$\Pr(cr_\nu(X) \ni \phi(\theta) \mid X) \geq \nu.$$

For convenience, we will assume that $cr_\nu(X)$ is the interval (not the general region) having lower and upper bound denoted by $cr_\nu^L(X)$ and $cr_\nu^U(X)$ respectively;

· $\underline{pp}(X)$: Bayesian posterior probability for the hypothesis test defined by

$$\underline{pp}(X) = \Pr(\phi(\theta) \in \Phi_0 \mid X);$$

and

· $ht_\eta(X)$: Bayesian hypothesis test decision function with the critical value η defined by

$$ht_\eta(X) = I[\underline{pp}(X) > \eta].$$

The null hypothesis is rejected when $ht_\eta(X) = 0$.

Considering the parameter θ as random quantity, the ideal sampling properties of inferences can be explained by the conditional probabilistic properties on θ over a subspace of Θ in the joint space of (X, θ).

Here are the validity sampling properties of inferences:

• Coverage probability of the general interval estimator $\psi_I(X)$:

$$\Pr(\psi_I(X) \ni \phi(\theta) \mid \theta = \theta^*) \geq \nu, \forall \theta^* \in \Theta;$$

• Size of the general hypothesis test decision function $\psi_T(X)$:

$$\Pr(\psi_T(X) = 0 \mid \theta = \theta^*) < \alpha, \quad \forall \theta^* \in \Theta_0,$$

where $\Theta_0 = \{\theta \mid \phi(\theta) \in \Phi_0\}$; and

• Uniformity of pv(X) under the null hypothesis:

$$\Pr(pv(X) < k \mid \theta = \theta^*) = k, \quad \forall k \in (0,1), \quad \forall \theta^* \in \Theta^*,$$

where pv(X) is specifically for a family of tests with level-α rejection region S_α satisfying

(a) $\exists \Theta^* \subset \Theta_0 = \{\theta \mid \phi(\theta) \in \Phi_0\}$ $s.t.$ $\Pr_\theta(X \in S_\alpha) = \alpha, \forall \theta \in \Theta^*$,

(b) $S_\alpha \subset S_{\alpha'}, \forall \alpha, \alpha' \in (0,1), \alpha < \alpha'$.

For example, the T test for $H_0: \mu \leq \mu_0$, $H_1: \mu > \mu_0$ in $N(\mu, \sigma^2)$ is a family of test satisfying the above conditions with $\Theta^* = \{(\mu, \sigma^2) \mid \mu = 0\}$. So we can expect the repetitions of pv(X) from the T test under the subsample space Θ^* follow the uniform distribution over (0,1).

The first two validity properties are the definitions themselves, and the uniformity of P-value can be easily shown from the definition of P-value . Note that uniformity of P-value under the null hypothesis automatically guarantees the satisfaction of the nominal size of the test. The uniformity of the P-value seems to be a critical aspect of the frequentist hypothesis test which provides a rationale of the test procedure, allowing for its common interpretation across problems. As indicated in Chapter I, there have been many suggestions adjusting the P-value to achieve uniformity when the classical P-value is not uniform. Robinson et al(1999), Bayarri and Berger(1999), Meng(1994), Rubin(1996) have suggested and

evaluated various adjusted P-values in the presence of nuisance parameters.

Here are the performance sampling properties of inferences:

• Small error of $\psi_P(X)$:

$$D(\psi_P(X), \phi(\theta) \mid \theta = \theta^*) \text{ is small}, \quad \forall \theta^* \in \Theta,$$

where D is a distance measure between two random variables. For example, a risk function like MSE would be appropriate for D,

$$E((\psi_P(X) - \phi(\theta))^2 \mid \theta = \theta^*) \text{ is small}, \quad \forall \theta^* \in \Theta;$$

• Short length of $\psi_I(X) = (\psi_I^L(X), \psi_I^U(X))$:

$$D(\psi_I^L(X), \psi_I^U(X) \mid \theta = \theta^*) \text{ is small}, \quad \forall \theta^* \in \Theta.$$

For example,

$$E(\|\psi_I^L(X) - \psi_I^U(X)\| \mid \theta = \theta^*) \text{ is small}, \quad \forall \theta^* \in \Theta,$$

where $\|\cdot\|$ is the norm;

• Small pv(X) under the alternative:

$$D(pv(X), 0 \mid \theta = \theta^*) \text{ is small}, \quad \forall \theta^* \in \Theta_1,$$

where $\Theta_1 = \{\theta \mid \phi(\theta) \in \Phi_1\}$. For example,

$$E(pv(X) \mid \theta = \theta^*) \text{ is small}, \quad \forall \theta^* \in \Theta_1;$$

and

• Small pp(X) under the alternative, large pp(X) under the null

hypothesis:

$$D(pp(X), 0 \mid \theta = \theta^*) \text{ is small}, \quad \forall \theta^* \in \Theta_1,$$
$$D(pp(X), 1 \mid \theta = \theta^*) \text{ is small}, \quad \forall \theta^* \in \Theta_0.$$

For example,

$$E(\mid pp(X) - I[\phi(\theta) \in \Phi_0] \mid \mid \theta = \theta^*) \text{ is small}, \quad \forall \theta^* \in \Theta.$$

The expected length of the interval estimator and some variations were used as a performance measure for interval estimators in the literature. For example, Gross(1976) used it in measuring the "robustness of efficiency" for confidence interval robustness when dealing with long-tailed symmetric distributions. He pointed out that expected length could be criticized on various grounds, including the valid remark that since it is an average, it may perform poorly as an estimate of location for long-tail distributions.

The expected P-value under the alternative could be a measure of performance of the test procedure on the alternative. Harold and Easter(1999) suggested the expected P-value (EPV) as an alternative performance measure when it is difficult to calculate the power function. Similarly the expectation of pp(X) can be used for the performance measure of the Bayesian hypothesis test on both null and alternative hypothesis.

The ideal sampling properties are usually used as the strategy for choosing a specific frequentist inference, so it is very natural to expect the frequentist inferences to have ideal sampling properties. The main justification of the Bayesian inferences is not based on sampling behaviors, but posterior probability structure. A Bayesian may not be as interested in the sampling properties, but it is also expected to satisfy the ideal sampling properties well in order for the Bayesian inference to be a reasonable statistical framework. Actually there have been many suggestions for the inference procedure constructed by Bayesian framework and having a good frequentist sampling properties.

A well-known difficulty in frequentist inference is the presence of nuisance parameters. The plug-in approach, substituting estimates of nuisance parameters

for the true values would be a typical frequentist method for this situation. But this plug-in approach is ignoring the uncertainty of the estimates and so distorts the designated validity and performance of the inferences. The Modified profile likelihood function suggested by Cox(1993) which accounts for the effect of the estimating nuisance parameter would be one frequentist solution to the nuisance parameter problem. Some Bayesian alternatives to solve this problem have been suggested.

It is also well known that the classical plug-in P-value for hypothesis test in the presence of nuisance parameters is not satisfactory with the ideal sampling property, "uniformity under the null point". Some Bayesian P-values have been suggested that account for the uncertainty of the estimates in the Bayesian framework including: posterior predictive P-value (Guttman(1967), Rubbin(1984)), discrepancy P-value (Meng(1994), Gelman at al(1996)) and partial posterior predictive and conditional predictive P-values (Bayarri and Berger(1999)). Bayarri and Berger(1999) compared the uniformity of various P-values under the null hypothesis in the various set-up. Robin(1999) evaluated asymptotic uniformity of P-values under the null hypothesis.

Interval estimation seems best place for communication of frequentist and Bayesian inference procedures. It is well known that, in regular cases, ν-level Bayesian credible regions have approximately ν coverage probability in repeated sampling.

The first-order approximation to confidence interval from Bayesian credible interval can be achieved to be independent of the prior distribution. The second-order approximation can be constructed by the Jeffrey's prior (see Welch and Peers(1963). However, it is problematic that improper priors yield undesirable improper posteriors for certain mixture models. Recently Wasserman(2000) suggested a data-dependent proper prior producing the intervals with second-order correct frequentist coverage on the mixture model,

$$f(x \mid \rho) = \sum_{j=1}^{k} p_j g(x \mid \theta_j),$$

where $\rho = (\theta_1, \ldots, \theta_k, p_1, \ldots, p_k)$, with $p_j \geq 0, j = 1, \ldots, k, \Sigma_j p_j = 1$ and normal density $g(x \mid \theta_j)$.

The suggested data-dependent prior provided a way of doing valid frequentist inference in such a mixture model since it does not depend on subjective input. For two-sided intervals, Severni(1993) and Sweeting(1999) obtained the formulae for third-order correct confidence interval assuming the model $f(x;\theta)$ depending on a real scalar parameter θ.

Ⅲ. Methodology

1. Measurement

All of the interesting ideal sampling properties defined in Chapter Ⅱ can be expressed by a common property of the following frequentist risk function, the conditional expectations relying on parameter values, over a subparameter space,

$$E(L(\psi(X), \phi(\theta)) \mid \theta = \theta^*), \quad \forall \theta^* \in \Theta^*,$$

where L is a real-valued loss (or gain) function, Θ^* is a subset of the parameter space Θ and both L and Θ^* are defined appropriately for each ideal sampling property. For example, achievement of the nominal coverage probability of ν-level confidence interval $ci_\nu(X)$ can be expressed as

$$E(L(ci_\nu(X), \phi(\theta)) \mid \theta = \theta^*) \geq \nu, \quad \forall \theta^* \in \Theta^*,$$

where $L(ci_\nu(X), \phi(\theta)) = I[ci_\nu(X) \ni \phi(\theta)]$ and $\Theta^* = \Theta$.

The uniformity of P-value for test with the rejection region S_α can be expressed by,

$$E(L_k(pv(X), \phi(\theta)) \mid \theta = \theta^*) = k, \text{ and } \forall k \in (0,1), \forall \theta^* \in \Theta^*,$$

where $L_k = I[pv(X) < k]$, and $\Theta^* \subset \Theta_0$ s.t. $\Pr_\theta(X \in S_\alpha) = \alpha, \forall \theta \in \Theta^*$.

We suggest relevant simulation methodology (RSM) as an observation-based simulation methodology to measure the parameter dependent risks. The only condition for RSM is that the statistical procedure generating the inference can be codified, and repeated on the computer. The RSM has two main features; data-dependent integrated risk measurement and Monte Carlo approximation.

Instead of looking at each expectation on each parameter value to check the property above, we suggest using the integrated expectation over parameter space,

$$\int_{\Theta} E(L(\psi(X), \phi(\theta)) \mid \theta)\, \mu(d\theta),$$

where $\mu(d\theta)$ is a "relevant measure" defined on the parameter space to common sense and observation. The strategy using the relevant measure $\mu(d\theta)$ reflects the idea that the level of satisfaction of ideal sampling properties may depend on the parameter values. Plausible parameter values implied by the data and common sense should be emphasized in the evaluation. Therefore, the evaluation obtained through this measurement could be interpreted as the weighted average of the individual evaluations over the parameter space with relevant weight.

We will denote the "relevant distribution density for the parameter" corresponding to the relevant measure $\mu(d\theta)$ by $\pi^r(\theta)$ and the "relevant joint probability distribution" constructed by the product of $\pi^r(\theta)$ and $f(x \mid \theta)$ by $G^r(X, \theta)$. Naturally the expectation and probability in the "relevant space" will be referred as "relevant expectation" and "relevant probability" and denoted by $E^r(\cdot)$ and $\Pr^r(\cdot)$. Following these notational scheme, the data-dependent integrated risk measurement above can be expressed as the "relevant expectation over the subparameter space" and denoted by

$$E^r_{\Theta}(L(\psi(X), \phi(\theta)).$$

Note that $E^r(L(\psi(X), \phi(\theta))$, an expectation w.r.t. the relevant measure $\pi^r(\theta)$ would be relevant to the unknown $E(L(\psi(X), \phi(\theta))$, the expectation w.r.t. the true marginal distribution of parameter $\pi(\theta)$, but they would still be

far from each other.

Our choices for the relevant measure are based on the posterior distribution on the observation with the non-informative prior. Note that one could simply condition on $\theta = \hat{\theta}$, i.e. $\pi^r(\theta) = I[\theta = \hat{\theta}]$.

In many cases, the calculation of the integral above would be algebraically impossible, so we suggest using Monte Carlo simulation as an approximation.

1) Relevant distribution for parameter

Our basic approach to get the relevant distribution denoted by $\pi^r(\theta)$, as the relevant measure of the parameter is to rely on both information from data and knowledge about the parameter.

There are different ways one might try to objectively reflect the information in the data. Here are two alternatives; parametric bootstrap and posterior distribution with non-informative prior. The first one is using a degenerate relevant distribution as follows,

$$\pi^r(\theta) = I[\theta = \hat{\theta}],$$

where $\hat{\theta}$ is a reasonable estimate for θ. This reflects implicitly the consideration of θ as a fixed quantity rather than a random one and results in the parametric bootstrap methodology. Another alternative for the relevant distribution is the posterior distribution with non-informative prior,

$$\pi^r(\theta) = \pi_{NI}(\theta \mid X).$$

Our general suggestion is the posterior distribution with non-informative prior. It would be more reasonable to reflect the relevance of the parameters by averaging them in some way rather than depending only on the best looking parameter point. There would be no big difference between the evaluation results from two relevant distributions when the likelihood function is a nice continuous one, but generally depending on one point seems to be too risky. Furthermore, we suggest using various posterior distributions having a range of variances from concentrated distribution to diffuse one.

There are many ways to reflect the knowledge about the parameter in constructing the relevant distribution along with the information from data. We could first put the restrictions of the parameters based on past knowledge and then reflect the observed information within restrictions. In cases where knowledge about parameters could be accurately expressed using a prior distribution, the posterior distribution would be a naturally relevant distribution to both the knowledge and the data.

2) Monte Carlo approximation

We suggest Monte Carlo simulation for approximating the integral, when an analytic solution is impossible. The only condition for applying this approximation is that the statistical procedure generating the inference $\psi(x)$ can be codified, so that it can be repeated. Assuming the relevant measure is given by $\pi^r(\theta)$, the steps are as follows:

- Generate $\theta_i \in \Theta^*$ randomly from $\pi^r(\theta), i=1, \ldots, R$;
- Generate x_i randomly from $f(x \mid \theta_i), i=1, \ldots, R$;
- Calculate $\psi(x_i)$ by the codified procedure, and $\Phi(\theta_i)$, then $L(\psi(x_i), \Phi(\theta_i)), i=1, \ldots, R$;
- Approximate $\int_{\Theta^*} E(L(\psi(X), \phi(\theta)) \mid \theta)\mu(d\theta)$ by the average,

$$\frac{1}{R}\sum_{i=1}^{R} L(\psi(x_i), \Phi(\theta_i)).$$

This Monte Carlo approximation of the relevant expectation will be referred to as "relevant estimate" and denoted by

$$\widehat{E^r_\theta}\cdot(L(\psi(X), \phi(\theta))).$$

This approximation can be justified by the fact that $\widehat{G_{EDF}}(X, \theta)$, the empirical distribution function defined by the simulated samples, is a reasonable

approximation for the relevant joint space $G^r(X, \theta)$ and the integral with the simple function $\widehat{G_{EDF}}(X, \theta)$ having equal mass on the finite simulated points is just the average of the points, i.e.

$$\begin{aligned} & \int_{x,\Theta^*} E(L(\psi(X), \phi(\theta)) \mid \theta)\mu(d\theta) \\ = & \int_{x,\Theta^*} L(\psi(X), \phi(\theta)))dG^{r(X,\theta)} \\ \fallingdotseq & \int_{x,\Theta^*} L(\psi(X), \phi(\theta)))d\widehat{G_{EDF}}(X, \theta) \\ = & \frac{1}{R}\sum_{i=1}^{R} L(\psi(x_i), \Phi(\theta_i)). \end{aligned}$$

Furthermore, the standard error of this relevant estimate for the relevant expectation can be acquired by

$$S.D.(L_1, \ldots, L_R)/\sqrt{R},$$

where $S.D.$ stands for the standard deviation and $L_i = L(\psi(X_i), \Phi(\theta_i))$.

2. Evaluation

The final evaluation for the satisfaction of the ideal sampling properties would be completed by checking the conditions for the ideal sampling properties in terms of relevant expectations. For example, it should be checked if

$$E^r_\Theta(I(ci_\nu(X) \ni \phi(\theta))) \geq \nu,$$

for the ν-level confidence interval $ci_\nu(X)$, and

$$E^r_\Theta(I(pv(X) < k)) = k, \quad \forall k \in (0,1),$$

where $\Theta^* \subset \Theta_0$ s.t. $\Pr_\theta(X \in S_\alpha) = \alpha$, for the P-value $pv(X)$ of frequentist test

with the rejection region S_α satisfying the needed conditions.

Even though it may be possible in some cases to get the analytical estimate for the expectations, we are assuming Monte Carlo approximation, and so the relevant estimate and its standard error are always available. Using these quantities from the simulation with enough replications to apply the central limit theorem, we could construct normal theory based inferences for checking them. For example, we could construct 95% confidence interval for the relevant expectation for checking the nominal coverage probability or conduct a hypothesis test for checking it.

For the uniformity of P-value, instead of checking the infinitely many conditions over $k\in(0,1)$, a better way is suggested. Considering the meaning of the condition; uniformity of the repeated P-values, a lack-of-fit test for uniformity of the repeated P-values in the relevant joint space would be appropriate. The Kolmogorov-Smirnov (K.S.) test is a possible choice. K.S. test for the test of uniformity of the P-values is based on the statistic,

$$D_n = \sup_{(-\infty<x<\infty)} | F_n(x) - F_0(x) | ,$$

for H_0: $[F(x)=F_0(x)$ for all x] and H_1: $[F(x)=F_0(x)$ for at least one $x]$, where $F(x)$ is the empirical distribution function of the repeated P-values and $F_0(x)$ is the uniform distribution function over (0,1). We are specially interested in the uniformity on the region of sample space corresponding to small P-values since the P-values appeared in real applications would be small. The K.S. test is not designed with some emphasis on a specific sample space of the distribution. Some tests based on rank statistics might have better power for this situation. In this study, we used the K.S. test for convenience and the development of better test for this problem was left for further study.

3. Adjustment

In this section, we explain the method to fix up the inference. The aim is to adjust the original inference to have the ideal sampling property, not the construction of a new inference method. This inference adjustment provides not

only a valid inference procedure, but also a measurement for the invalidity of the original inference.

Let's assume the inference $\psi(X)$ is invalid with respect to an ideal validity property through the suggested "evaluation scheme using RSM". Generally speaking, if we can find a function $\eta(A): A \rightarrow A$ s.t. $\eta(\psi(X)): X \rightarrow A$ satisfies the needed ideal sampling property in our simulation scheme, then the composite function $\eta(\psi(X))$ would be the adjusted inference via the suggested simulation scheme.

Typically the qualified $\eta(\cdot)$ to be valid in the above sense would not be unique. So we may have to choose one among the possible alternatives which is supported by some rationale. A general rationale for the selection could be provided by the evaluation result with respect to another performance measure through the suggested evaluation scheme. Here are our choices for the adjusted P-value and adjusted confidence interval. The rational or the performance measure for these selections are not given.

For the adjustment of pv(X) to have the uniformity property, we could define $\eta(pv) = EDF(pv)$, where $EDF(pv)$ is the empirical distribution function of the repeated pv's under the null relevant space. It is obvious that $EDF(pv(X))$ is a valid inference satisfying the uniformity in the null relevant space. Notice that the uniformity of $EDF(pv(X))$ is acquired in the sense of average over the null relevant space, not for each parameter value. So the suggested adjustment $EDF(pv(X))$ can be invalid in the strict sense requiring the uniformity for each parameter value in the null space.

Confidence intervals $ci_\nu(X)$ can also be adjusted to have the nominal coverage probability by defining $\eta(ci_\nu) = ci_{adj^{-1}(\nu)}$, where $adj(\nu)$ is the function providing the real coverage level of the confidence interval with nominal coverage probability ν in the relevant space. This $adj(\nu)$ can be constructed in our simulation scheme and could be assumed as the non-decreasing function for the reasonable inference procedure.

Ⅳ. Conclusion

We believe the suggested "evaluation scheme using RSM" can serve as a valid tool for frequentist point-of-view evaluation of the implemented frequentist and Bayesian inference on the observation. This evaluation tool could serve many purposes:

- Justification of the implemented inference result on observation. For example, it can judge if a published P-value is valid with respect to the uniformity as long as the inference procedure can be codified and the data are available;
- Comparison of the competitive inferences on the real observation. For example, if we have conflicting inference result on the same observation, it can be helpful in deciding which one is more reliable;
- Interface from the Bayesian inference to the frequentist one. For example, it can suggest the confidence level of the Bayesian 95% credible region;
- Adjustment of frequentist inference. For example, the adjusted P-value can also be provided satisfying the uniformity criterion from a hypothesis test procedure found to be invalid; and
- A general tool for measuring general sampling properties as well as ideal frequentist sampling properties. For example, the RSM can be used in inspecting the sensitivity of the Bayesian inferences on the repeated samples. This sensitivity analysis of Bayesian inference for some reasonable priors would be a Bayesian point-of-view justification for a given Bayesian methodology.

Bibliography

Sellke T., Bayarri M. J. and Berger J. O. 1998. "Calibration of P-values for testing precise null hypothesis." Duke university.

Cox D. R. and Reid N. 1993. "A note on the calculation of adjusted profile likelihood." *Journal of the Royal Statistical Society B* 55(2): 467-471.

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.

Gross A. M. 1976. "Confidence interval robbustness with long-tailed symmetric distributions." *Journal of the American Statistical Association* 71: 409-416.

Guttman I. 1967. "The use of the concept of a future observation in goodness-of-fit problems" *Journal of the Royal Statistical Society B* 29: 83-100.

Harold S. and Ester S. C. 1999. "P values as random variables - expected p values" *The American statistician* 53: 326-331.

Meng X. L. 1994 "Posterior predictive p-values" *The Annals of Statistics* 22: 1142-1160.

Robins J. M., Vaart A. V. and Ventura V. 1999. "The asymptotic distribution of P-values in composite null models" *submitted to Journal of the American Statistical Association.*

Rubin D. B. 1984. "Bayesianly justifiable and relevant frequency calculation for the applied statistician" *The Annals of Statistics* 12: 1151-1172.

Rubin D. B. 1996 "Discussion of Posterior predictive assessment of model fitness via realized discrepancies" *Statistica Sinica* 6: 787-792.

Severini T. A. 1993. "Bayesian interval estimates which are also confidence intervals" *Journal of the Royal Statistical Society B* 55:533-540.

Sweeting T. J. 1999. "On the construction of Bayes-confidence region" *Journal of the Royal Statistical Society B* 61: 849-861.

Wasserman L. 2000 "Asymptotic inference for mixture models using data-dependent priors" *Journal of the Royal Statistical Society B* 62:159-180.

Welch B. L. and Peers H. W. 1963. "On formulae for confidence points based on integrals of weighted likelihoods" *Journal of the Royal Statistical Society B* 25: 318-329.