Sankhyā : The Indian Journal of Statistics
Special Issue on Bayesian Analysis
1998, Volume spl, Series , Pt. 1 , pp. 161-175

SAMPLE SIZE DETERMINATION USING POSTERIOR PREDICTIVE DISTRIBUTIONS

By DONALD B. RUBIN
Harvard University, Cambridge
and
HAL S. STERN
Iowa State University, Ames

SUMMARY. A statistical model developed from scientific theory may "fail to fit" the available data if the scientific theory is incorrect or if the sample size is too small. The former point is obvious but the latter is more subtle. In the latter case, the hypothesized model may fail to fit in the sense that it is viewed as unnecessarily complicated, and so the investigators settle upon a simpler model that ignores structure hypothesized by scientific theory. We describe a simulation-based approach for determining the sample size that would be required for distinguishing between the simpler model and the hypothesized model assuming the latter is correct. Data are simulated assuming the hypothesized model is correct and compared to posterior predictive replications of the data, which are drawn assuming the simpler model is correct. This is repeated for a number of sample sizes. The Bayesian approach offers two especially nice features for addressing a problem of this type: first, we can average over a variety of plausible values for the parameters of the hypothesized model rather than fixing a single alternative; second, the approach does not require that we restrict attention to a limited class of regular models (e.g., t-tests or linear models). The posterior predictive approach to sample size determination is illustrated using an application of finite mixture models to psychological data.

1. **Introduction**

Consider a situation in which scientific theory suggests a probability model specifying the interrelationships among a number of random variables, but the complexity of that model is not supported by the available data. In that case, a simpler model that describes the existing data would likely be considered

---

AMS (1991) *subject classification.* 62F15, 62K99, 62P15.
Key words and phrases. Bayesian inference, finite mixture model, power calculation, Markov chain Monte Carlo, study design.
* Research supported by the National Science Foundation (NSF) under grand award DMS-9404479.

adequate. We denote the simpler model by M_o and the hypothesized model by M_h; there is no assumption that the models are nested (i.e., we do not assume that M_o can be obtained from M_h by fixing certain parameters). Two plausible explanations for the data's support of M_o are that either the proposed model M_h is invalid, or the size (structure) of the existing data set is insufficient for discerning the aspects of M_h that M_o is lacking. The main emphasis in this paper is on determining the sample size needed in a future study to provide a reasonable probability of detecting the weaknesses of M_o with respect to the science reflected in M_h.

Our basic approach is to propose a sample size for this future study and generate many data sets of this size under the hypothesized model (using a proper prior distribution to sample plausible parameter values and the hypothesized data model to sample data values). The simpler model is fit to each simulated data set, and the need to add more scientifically-motivated structure is then assessed. By repeating this process with a variety of sample sizes, we can determine the minimum sample size needed to be confident that the simpler model will be found lacking, assuming the hypothesized model provides a better description of the phenomenon being studied. The fit of a statistical model to data can be assessed in a variety of ways. The main tools used here are posterior predictive distributions of discrepancies selected to diagnose particular failures of the model (Rubin, 1984; Gelman *et al.*, 1996).

The simulation-based approach requires that statistical models be fit to a large number of simulated data sets; since each fitting of a model to a simulated data set may require Markov chain Monte Carlo methods, the computational work required to determine a suitable sample size can be quite demanding. Of course, it is expensive to carry out a study, and careful planning can provide substantial benefits and eventual savings.

In Section 2 we review the posterior predictive approach to model assessment and describe our approach for determining sample size using posterior predictive distributions. In addition, we briefly describe several extensions and modifications of the procedure, including the changes that are required to accommodate alternative model assessment approaches such as Bayes factors. Section 3 applies our approach to an example involving the use of finite mixture models in psychology. Research concerning infant temperaments hypothesized a latent class model with four classes but a two-class model provided an adequate fit based on a study of 93 infants (Stern *et al.*, 1994, 1995; Rubin and Stern, 1994). Section 4 includes additional discussion concerning the relationship of this approach to traditional power analyses, and Section 5 provides final remarks.

2. **Sample size determination**

2.1 *Posterior predictive model assessment.* Because posterior predictive model assessment (Rubin 1984, Gelman *et al.* 1996) plays a crucial role in this ap-

proach to determining sample size, it is reviewed briefly here. Suppose that we have fit a model M_o with parameters θ_o to data y_{obs}. Covariates, if any, are considered fixed and do not appear explicitly in the notation. Inferences concerning the parameters of the model are based on the posterior distribution, $p(\theta_o|y_{obs}, M_o) \propto p(y_{obs}|\theta_o, M_o)\, p(\theta_o|M_o)$. We make no assumptions about the prior distribution $p(\theta_o|M_o)$ other than it leads to a proper posterior distribution; thus we allow for the possible use of improper prior distributions on θ_o. Let y^{rep} denote a hypothetical replication of the current data set, y_{obs}, i.e., a new sample generated by the probability model M_o using parameter values θ_o drawn from the posterior distribution of the parameters. The posterior distribution is used so that diagnostic effort is focussed on values of θ_o that are most plausible given the data y_{obs}. Formally, y_{obs} and y^{rep} are assumed to be independent and identically distributed under M_o given θ_o. In contrast, Box (1980) defines hypothetical replications using the prior distribution of θ_o which might be called prior predictive replications in our terminology. The relevant reference distribution for our definition of y^{rep} is its posterior predictive distribution,

$$\begin{aligned} p(y^{rep}|M_o, y_{obs}) &= \int p(y^{rep}|\theta_o, M_o, y_{obs})\, p(\theta_o|M_o, y_{obs})\, d\theta_o \\ &= \int p(y^{rep}|\theta_o, M_o)\, p(\theta_o|M_o, y_{obs})\, d\theta_o, \end{aligned}$$

which averages the sampling distribution of y^{rep} over the posterior distribution of the model parameters, θ_o, assuming M_o is true. Under M_o, the replications differ from the observed data due only to the variation inherent in the sampling distribution of the data given θ_o and the posterior uncertainty about the values of the parameters θ_o.

Let $T(y;\theta)$ be a discrepancy assessing the degree to which some model with parameters θ fails to fit data y. The model does not appear explicitly in this notation. Ordinarily T is chosen to measure discordance of the data and model with respect to specific structure that it is thought may be present in reality but not in the model under consideration. Model assessment typically uses several different discrepancy measures to consider different aspects of the model. For example, we might choose a vector T to be the sufficient statistics of a more complex model than the one we are fitting. In the mixture model example of Section 3, we focus on a scalar T that is an overall measure of fit instead of using discrepancy measures sensitive to individual model features; the overall measure appears to be adequate for detecting misspecification of the number of classes in the model. Note that T is allowed to depend on the unknown parameters of the model and is thus not restricted to be a test statistic. We assess the fit of our model by comparing the posterior distribution of $T(y_{obs};\theta_o)$ to the posterior predictive distribution of $T(y^{rep};\theta_o)$. In some cases these distributions can be obtained analytically, but in general practice we typically rely on simulation to compare the distributions, especially with complex models.

The use of simulation is especially natural when Markov chain Monte Carlo (MCMC) is used to obtain draws from the posterior distribution, $p(\theta_o|y_{obs}, M_o)$. Let $\theta_o^{(k)}, k = 1, \ldots, K$ represent a sample from the posterior distribution. Then we can draw a single replication, $y^{rep(k)}$, from the data model, $p(y|M_o, \theta_o^{(k)})$, for each posterior draw. The K draws from the posterior distribution of $T(y_{obs}; \theta_o^{(k)})$ and the K draws from the posterior predictive distribution of $T(y^{rep(k)}; \theta_o^{(k)})$ can be displayed in a scatterplot of $(T(y_{obs}; \theta_o^{(k)}),\ T(y^{rep(k)}; \theta_o^{(k)})), k = 1, \ldots, K$. The tail-area probability,

$$P_T(y_{obs}, M_o) \equiv \Pr(T(y^{rep}; \theta_o) \geq T(y_{obs}; \theta_o) \mid y_{obs}, M_o), \qquad \ldots(1)$$

is one summary of the model assessment with extremely large or small values indicating that the observed data is not similar to the data sets that would be expected under M_o with respect to the discrepancy measure T. This probability can be estimated using the proportion of the K simulated $(y^{rep(k)}, \theta_o^{(k)})$ pairs for which the event, $T(y^{rep(k)}; \theta_o^{(k)}) \geq T(y_{obs}; \theta_o^{(k)})$, occurs. If T is a traditional test statistic (i.e., it does not depend on θ_o), then the observed value $T(y_{obs})$ is compared to the K simulated draws from the posterior predictive distribution, $T(y^{rep(k)}), k = 1, \ldots, K$.

2.2 *Technique for determining sample size.* Let M_h denote a model hypothesized to be true based on scientific theory, and let θ_h denote the parameters of the model. We use $p(y|\theta_h, M_h)$ to represent the probability distribution for y implied by the model and $p(\theta_h|M_h)$ to represent the prior distribution for its parameters. The prior distribution must be a proper distribution since we will simulate from it as part of our approach. Suppose that we currently believe a simpler model M_o with parameters θ_o is adequate. Recall that the prior distribution on θ_o is not required to be a proper distribution because our draws of θ_o will always be from its posterior distribution. Although we assume that M_o is a "smaller" or less complex model than M_h, there is no requirement that θ_o be a function of θ_h. The preference for the simpler model M_o can arise, as it did in the mixture model example of the next section, when the posterior distribution of M_h given observed data y_{obs} is characterized by a great deal of uncertainty (perhaps including a large number of comparable modes), whereas the posterior distribution of M_o is well defined and appears to provide an adequate fit to the data. In this situation it is natural to ask whether a larger sample might find M_o invalid and favor the more sophisticated model M_h. The existence of observed data y_{obs} is not required; the study design question may be motivated without such data.

We now describe a general simulation-based approach for determining the sample size required to ascertain whether our preference for M_o indicates a failure of the scientific theory or not. We use N to denote the sample size under consideration. Data sets y of size N are simulated from the marginal distribution

of y under the scientifically motivated model M_h,

$$p(y|M_h) = \int p(y|\theta_h, M_h)\, p(\theta_h|M_h)\, d\theta_h.$$

For each simulated data set, we fit the simpler model M_o and carry out an assessment of the fit of M_o using the posterior predictive approach of Section 2.1. Specifically, we generate posterior draws from $p(\theta_o|y, M_o)$ and posterior predictive draws from $p(y^{rep}|y, M_o)$. In this context the replicate data refers to a replication of the simulated data y. The fit of the model M_o is then assessed using discrepancy $T(y;\theta)$. The entire model fitting and model assessment process is repeated for a number of simulated data sets of size N. To assess how well the simple model M_o fits according to T when data sets of size N are simulated under M_h, we define the tail-area probability

$$Q_T(N, M_o, M_h) \equiv \int P_T(y, M_o)\, p(y|M_h)\, dy, \qquad \ldots (2)$$

where $P_T(y, M_o)$ is the tail-area probability (1) that results when discrepancy measure T is used with data y to evaluate model M_o. The tail-area $Q_T(N, M_o, M_h)$ measures the likelihood that $T(y^{rep};\theta_o) \geq T(y;\theta_o)$ for data sets of size N generated under the model M_h. The expression (2) makes it clear that Q is a weighted average of model assessment tail-areas over data sets a priori thought likely to occur under M_h. Note that (2) is actually an integral that averages over $(\theta_h, y, \theta_o, y^{rep})$, with the integration over θ_h implicit in the definition of $p(y|M_h)$ and the integrations over (θ_o, y^{rep}) implicit in the definition of $P_T(y, M_o)$.

In practice $Q_T(N, M_o, M_h)$ is estimated via simulation. For each set of simulated data, y, we may evaluate $P_T(y, M_o)$ to any desired precision by drawing as needed from the posterior predictive distribution of y^{rep} given y under model M_o. Thus, we may choose a large number of posterior predictive draws for each simulated data set, or we may choose a small (perhaps only one) draw from the posterior predictive distribution of each simulated data set. We have used a single draw from each simulated data set because we are not interested in detailed information about $P_T(y, M_o)$ for any one data set. In addition, this allows us to ignore the correlation among the posterior predictive draws generated via a single Markov chain Monte Carlo analysis (although this could be addressed).

The technique we use to evaluate the efficacy of a particular sample size N is described fully by the following steps.

1. Fix a sample size N and a proper prior distribution on the parameters of the hypothesized model, $p(\theta_h|M_h)$.

2. For $k = 1, \ldots, K$:

(a) Simulate θ_h from its prior distribution.

(b) Simulate data $y^{(k)}$ (sample size N) from the hypothesized model M_h with parameters θ_h drawn in (a).

(c) Obtain a single draw from the posterior distribution of the null model's (M_o) parameters, $p(\theta_o|y^{(k)}, M_o)$.

(d) Obtain a posterior predictive draw of $y^{rep(k)}$ using the drawn value of θ_o.

3. Compare the distributions of $T(y^{rep(k)};\theta_o)$ and $T(y^{(k)};\theta_o)$, perhaps by estimating the tail-area probability $Q_T(N, M_o, M_h)$.

Steps (2.c) and (2.d) can be replaced by multiple draws as discussed at the end of the previous paragraph. Some practical issues related to the implementation of this technique (e.g., selection of prior distributions) are found in the discussion of the example in Section 3. For now we restrict attention to a single important computational issue.

As has been mentioned earlier, it will often be the case the draws from the posterior distribution in step (2.c) will be obtained using MCMC. This raises the issue of how convergence of the MCMC algorithm can be judged since each iteration of step 2 requires analysis of a new data set. In our examples, we used the initial analysis (whereby sample data were used to choose the model M_o) to determine an approximate burn-in period beyond which draws are likely to represent the desired posterior distribution and used that burn-in period throughout. The multiple sequence approach of Gelman and Rubin (1992) was used to assess convergence in the initial analysis. Of course, when the prior distribution on θ_h is vague, then the simulated data sets may vary enough to require additional work on diagnosing convergence of the MCMC algorithms.

2.3 *Sample size determination and Bayes factors.* The use of posterior predictive model checks has been emphasized in our approach to determining sample size. The posterior predictive approach focuses on the current best fitting model and asks what sample size would be needed to cast doubts on its validity. Given the focus on two models (M_o and M_h) it may seem natural to ask what sample size would be needed to produce a reasonable expectation that the Bayes factor would favor M_h. The Bayes factor treats the two models symmetrically when choosing between them (Kass and Raftery, 1995). In fact, this requires only a minor modification of the technique described in the previous section. Given a simulated data set from M_h, we would evaluate the Bayes factor comparing M_h and M_o for the given data set (this would replace steps (2.c) and (2.d)). Each simulated data set would lead to a single Bayes factor and step 3 would examine the distribution of Bayes factors obtained from the simulated data sets of a given size. The tail area formula (2) would not be relevant. The prior distribution for θ_o would be required to be a proper distribution in order for the Bayes factor to be defined, which is one potential disadvantage of using Bayes factors for model selection. A more important issue is that Bayes factors address the relative fit of two models without addressing how well either actually fits the data. Posterior predictive model checks focus on the latter question by allowing the discrepancy $T(y;\theta)$ to be general and not necessarily determined by the models M_o and M_h.

Table 1. DATA FROM INFANT TEMPERAMENT STUDY INCLUDES 93 INFANTS IN A $4 \times 3 \times 3$ CONTINGENCY TABLE.

Motor activity category, M	Cry category, C	Number of observations in fear category $F=1$	$F=2$	$F=3$	Row total
1	1	5	4	1	10
1	2	0	1	2	3
1	3	2	0	2	4
2	1	15	4	2	21
2	2	2	3	1	6
2	3	4	4	2	10
3	1	3	3	4	10
3	2	0	2	3	5
3	3	1	1	7	9
4	1	2	1	2	5
4	2	0	1	3	4
4	3	0	3	3	6

3. Application to psychology example

3.1 *A finite mixture model for infant temperament data.* There has been some debate among developmental scholars in psychology about whether temperamental qualities of infants and children (irritability, activity level) should be conceptualized as continua or as categories. Most statistical analyses assume the former; the authors were involved in a study designed to explore appropriate statistical analyses assuming the categorical view is correct (Stern *et al.*; 1994, 1995). The data are measurements of 93 infants during test batteries of sensory stimuli at age 4 months and age 14 months. A number of measurements were made but these were reduced for purposes of analysis to 3 categorical variables, motor activity (M) at 4 months of age (four levels), crying (C) at 4 months of age (three levels), and fearfulness (F) at 14 months of age (three levels), defining the $4 \times 3 \times 3$ contingency table shown in Table 1. It is evident in the table that the three variables are not independent. Infants who exhibited low levels of both motor activity and irritability at four months had significantly fewer fears at 14 months than the infants who exhibited high levels of motor activity and irritability.

Studies of humans and other mammalian species suggest possible biological mechanisms under which qualitatively different groups of infants would be expected. Stern *et al.* (1995) provide some discussion and references to the related science, which essentially suggests that four distinct subpopulations exist. Stern *et al.* applied a finite mixture model (e.g., see Everitt and Hand, 1981; Titterington, Smith, and Makov, 1985) to analyze these data. Let n_{ijk} denote the number of infants with $M = i,\ C = j,\ F = k$, and let π_{ijk} represent the probability that an infant is in that cell of the contingency table. In addition we imagine that

there is a latent or unobserved variable with S categories with the proportion of infants in the s-th class denoted by π_s and we let $\pi_{ijk|s}$ denote the distribution of (M, C, F) in the s-th class. Under this model, the joint distribution of the three observed random variables is a mixture of the joint distributions within the classes, $\pi_{ijk} = \sum_{s=1}^{S} \pi_s \pi_{ijk|s}$. In this particular case, it is assumed that the three random variables (M, C, F) are conditionally independent given the class to which an infant belongs so that $\pi_{ijk} = \sum_{s=1}^{S} \pi_s \pi_{i|s} \pi_{j|s} \pi_{k|s}$.

Table 2. MAXIMUM LIKELIHOOD PARAMETER ESTIMATES FOR TWO-CLASS MODEL FIT TO THE INFANT TEMPERAMENT DATA.

	Class	
Parameter	$s = 1$	$s = 2$
$\pi_s = \Pr(s)$	.50	.50
$\pi_{i=1\|s} = \Pr(M = 1\|s)$	.22	.14
$\pi_{i=2\|s} = \Pr(M = 2\|s)$	.60	.19
$\pi_{i=3\|s} = \Pr(M = 3\|s)$	.12	.40
$\pi_{i=4\|s} = \Pr(M = 4\|s)$	.06	.27
$\pi_{j=1\|s} = \Pr(C = 1\|s)$	.71	.28
$\pi_{j=2\|s} = \Pr(C = 2\|s)$	.08	.31
$\pi_{j=3\|s} = \Pr(C = 3\|s)$	.21	.41
$\pi_{k=1\|s} = \Pr(F = 1\|s)$	.74	.00
$\pi_{k=2\|s} = \Pr(F = 2\|s)$	.26	.32
$\pi_{k=3\|s} = \Pr(F = 3\|s)$	.00	.68

Table 2 shows the maximum likelihood estimates of the model parameters assuming two classes. The estimates are obtained using the EM algorithm (Dempster *et al.* 1977), which is described for mixture models of this type by Goodman (1974ab); Bayesian posterior inferences with a vague prior distribution are similar. Infants in the first class are characterized by low levels of motor activity, low levels of cry, and as they grow up, low levels of fear, whereas infants in the second class are categorized by high levels of each variable. Each of the two classes appears to comprise about half the population.

The results of Table 2 support the idea that the population can be viewed as a mixture of categorical types. Rubin and Stern (1995) and Gelman *et al.* (1996) apply the posterior predictive approach to these data to determine if there is any evidence that the two-class model is inadequate. Details of the MCMC algorithm used to to fit the models and carry out the model assessment are provided there. The discrepancy that is used is based on the likelihood ratio,

$$T(y_{obs}; \theta) = 2 \sum_{ijk} n_{ijk} \log(n_{ijk}/\pi_{ijk}),$$

which compares the estimated counts under the mixture model to the observed counts (which can be thought of as corresponding to a saturated model containing one parameter for each nonempty cell). We use y_{obs} as generic notation for data (here cell counts n_{ijk}), and θ as generic notation for parameters (here cell probabilities π_{ijk}). The traditional likelihood ratio test statistic replaces the parameters, π_{ijk}, by their maximum likelihood estimates, but our discrepancy uses posterior draws instead. Table 3 gives the traditional test statistic for models with one through four classes along with the posterior predictive tail-area probability for assessing the fit of the one- and two-class models. The traditional likelihood ratio test statistic and its associated reference distribution are of limited use since the regularity conditions needed to derive the asymptotic chi-squared reference distribution are not satisfied. In this case the posterior predictive approach and the "invalid" traditional approach both identify the two-class model as the smallest model that provides an adequate fit. An additional factor in favor of the two-class model is that the posterior distribution under the three- or four-class model is multimodal and appears to be extremely widely dispersed. These results suggest that evidence for the four-class model, favored by a priori theory, can not be found in these data, at least as summarized by the likelihood ratio discrepancy. This conclusion, however, simply may be an indication that the sample size was inadequate to detect the weaknesses of the two-class model.

Table 3. ASSESSING THE FIT OF FINITE MIXTURE MODELS TO THE INFANT TEMPERAMENT DATA RELATIVE TO THE SATURATED MODEL. POSTERIOR PREDICTIVE TAIL-AREA PROBABILITIES WERE NOT OBTAINED FOR THE THREE-CLASS AND FOUR-CLASS MODELS.

Model description	Degrees of freedom	Likelihood ratio test statistic	Posterior predictive tail-area probability, P_T
Independence (= 1 class)	28	48.8	0.06
2 Latent Classes	20	14.1	0.74
3 Latent Classes	12	9.1	
4 Latent Classes	4	4.7	

3.2 *Exploring the effect of sample size.* Scientific theory suggests a four-class model and, moreover, the theory tells us what the infants in each class should be like. We should expect intermediate groups in addition to the (1) high motor, high cry, high fear group and the (2) low motor, low cry, low fear group identified by the two-class model. Specifically, one should expect to find high motor, low cry infants (and their opposites) having somewhat intermediate values of fear. It appears that in the current data set, the more fearful of these intermediate children are combined with the high fear children and the less fearful are combined with the low fear group. If the four-class model is accurate, how large a sample would be required to expect to find fault with the two-class

model? We apply the technique described in Section 2 to address this issue. It should be emphasized that the technique for estimating the required sample size does not depend in any way on the observed data set of the previous section, the data there merely motivate the question.

The first step required is to specify a prior distribution on the four-class model parameters. We specify a Dirichlet prior distribution for the parameters of each multinomial distribution in each class (recall that there are independent multinomial distributions for motor activity, cry and fear in each class). We also specify a Dirichlet prior distribution for the parameters of the multinomial distribution describing the size of the classes, $(\pi_s, s = 1, \ldots, 4)$. Moreover, we assume that all of these Dirichlet prior distributions are independent. The Dirichlet is quite convenient in this case because it is conjugate for the multinomial distributions in a number of the MCMC sampling steps. The Dirichlet distribution is usually parameterized in terms of a vector of prior sample sizes (e.g., see Gelman *et al.*, 1995). The prior sample sizes give information about the a priori expected values of the multinomial component probabilities. It is convenient here to characterize prior information in terms of a vector of prior sample proportions, α with $\sum_i \alpha_i = 1$ (where the number of elements of α is equal to the number of multinomial cells on which the Dirichlet is defined), and a prior overall sample size m. The usual parameterization is obtained upon multiplying α by m. Large m corresponds to a great deal of prior information and small m corresponds to vague prior information.

Table 4. Parameters of Dirichlet prior distributions for four-class model parameters assuming prior sample size equal to one.

Parameter		Dirchlet parameters α for parameters in class 1	2	3	4
Distribution of population across classes	$\pi_s = \Pr(s)$	.35	.25	.25	.15
Distribution of motor activity within classes	$\pi_{i=1\|s} = \Pr(M = 1\|s)$	.45	.05	.40	.10
	$\pi_{i=2\|s} = \Pr(M = 2\|s)$	.35	.15	.30	.20
	$\pi_{i=3\|s} = \Pr(M = 3\|s)$	.15	.35	.20	.30
	$\pi_{i=4\|s} = \Pr(M = 4\|s)$	.05	.45	.10	.40
Distribution of cry within classes	$\pi_{j=1\|s} = \Pr(C = 1\|s)$	.80	.05	.15	.60
	$\pi_{j=2\|s} = \Pr(C = 2\|s)$	.15	.15	.25	.25
	$\pi_{j=3\|s} = \Pr(C = 3\|s)$	.05	.80	.60	.15
Distribution of fear within classes	$\pi_{k=1\|s} = \Pr(F = 1\|s)$	.80	.05	.15	.60
	$\pi_{k=2\|s} = \Pr(F = 2\|s)$	.15	.15	.25	.25
	$\pi_{k=3\|s} = \Pr(F = 3\|s)$	.05	.80	.60	.15

Table 4 provides the vectors of expected proportions for our prior distributions (corresponding to $m = 1$); the prior sample size is varied in subsequent

discussion. Classes 1 and 2 are assumed to be similar to what was found in the original data; class 3 comprises children tending to have low motor scores, high cry scores, and intermediate-high fear scores, whereas class 4 comprises children with high motor scores, low cry scores and intermediate-low fear scores. The relative sizes of the four classes and the expected distribution within each are essentially educated guesses based on conversations with psychologists. We do not discuss these values further.

The posterior predictive sample size technique (the P^2S^2 technique) of Section 2 is used to assess the sample size required to expect to find that a particular model is inadequate. Our main interest is in assessing the fit of the two-class model but, we also present results for the one-class model. Recall that with our data the one-class model was clearly invalid and even a sample of size 93 was sufficient to detect its inadequacies. The two-class model is of more interest since we have already seen that our particular sample of size 93 did not detect any problems with the two-class model, at least using the likelihood ratio discrepancy. Since the P^2S^2 technique requires a Bayesian analysis of each simulated data set under the simpler model M_o, we must specify a prior distribution for the parameters of the one-class and two-class models. When simulated data are analyzed using the one-class model, we use Dirichlet prior distributions with equal prior sample proportions. The prior distributions for the two-class model are taken to be identical to the first two columns of Table 4 with the expected proportion of the population in the first class equal to 0.55. In both cases, the prior sample size is fixed at one because this is a low-precision prior that guarantees proper posterior distributions.

We report $Q_T(N, M_o, M_h)$, the tail probability comparing the discrepancy between the model M_o (the one- or two-class model), and the simulated data drawn from the hypothesized true model M_h (a sample of size N from the four-class model), with the discrepancy between M_o and posterior predictive replications under M_o. Small values of Q indicate that the specified sample size provides a reasonable chance of identifying weaknesses in the simpler model, M_o. Table 5 gives values of Q for evaluating one- and two-class models for a range of sample sizes (100, 250, 500 and 1000), and a range of prior distribution precisions for M_h (the effective sample size of the Dirichlet prior distributions in Table 4 is varied from 1 to 100). The results are based on 1000 samples drawn from the four-class distribution. All of the sample sizes are sufficient to reject the one-class model that implies that M, C, F are independent. It appears, however, that a sample size of 1000 is required to expect to find the two-class model inadequate. As expected, increasing the sample size, N, increases (within the bounds of the simulations' binomial sampling variability) the probability that the incorrect (one-class or two-class) model will be found inadequate.

The behavior of Q as the prior sample size increases is a bit more subtle. In this instance, a high-precision prior distribution on the parameters of the four-class model ($m = 100$) suggests that a sample size bigger than 1000 is needed to

find the two-class model inadequate whereas the low-precision prior distribution ($m = 1$ or $m = 10$) suggests that 1000 is about the right sample size. This effect is a consequence of the values chosen for the prior distributions in Table 4. The four-class model with those expected parameter values generates data that is similar to data possible under a two-class model so that prior distributions with large prior sample sizes (little prior uncertainty) tend to simulate data sets in which the two extra classes are difficult to detect. By contrast, small prior sample sizes (greater prior uncertainty) increase the chance of obtaining simulated data sets for which four classes can be more easily detected. If the expected parameter values in the four classes are chosen so that the four subpopulations are more disparate, then we find the reverse pattern: smaller sample sizes are needed if high-precision prior distributions are used instead of low-precision prior distributions. Incidentally, it has not been possible to put this sample size calculation to the empirical test because sample sizes of 1000 are nearly impossible in infant temperament research.

Table 5. TAIL-AREA PROBABILITIES FOR THE LIKELIHOOD RATIO DISCREPANCY. ALL ESTIMATES ARE BASED ON 1000 SIMULATED DATA SETS OF THE SPECIFIED SAMPLE SIZE FROM THE FOUR-CLASS MODEL WITH PRIOR DISTRIBUTIONS SPECIFIED BY THE DIRICHLET PARAMETERS IN TABLE 4 AND THE PRIOR SAMPLE SIZE m. SMALL VALUES INDICATE THAT THE SIMPLER MODEL (ONE CLASS OR TWO CLASSES) IS LIKELY TO BE FOUND INADEQUATE BASED ON SAMPLES OF SIZE N.

	prior sample size					
sample	$m = 1$		$m = 10$		$m = 100$	
size, N	1-class	2-class	1-class	2-class	1-class	2-class
100	0.041	0.293	0.021	0.312	0.007	0.434
250	0.025	0.222	0.000	0.267	0.000	0.452
500	0.020	0.201	0.001	0.182	0.000	0.425
1000	0.015	0.146	0.000	0.085	0.000	0.328

4. **Additional issues**

4.1 *Relation to traditional power calculations.* The idea of determining in advance the sample size needed to provide a study with a reasonable probability of detecting an expected effect is certainly not a new one. Tables relating the sample size, expected effect, significance level and power (probability of rejecting the null hypothesis) at fixed alternatives are available for many standard statistical methods (e.g., two-sample tests of means and proportions). The existing tables, however, do not allow one to easily evaluate power against a fixed prior distribution of alternatives. In addition, no such tables exist for fitting mixture models or the other types of sophisticated models that are increasingly common in applied statistical work. The technique described here allows us to extend the traditional approach to determining power and sample size to these models.

We can relate our posterior predictive sample-size (P^2S^2) technique to traditional approaches for determining power and sample size by carefully considering the elements that play a role in our technique. Recall that $Q_T(N, M_o, M_h)$ is actually an integral (or average) over four distributions:

- the prior distribution on the parameters of the hypothesized model M_h;
- the data distribution under the hypothesized model M_h;
- the posterior distribution of the parameters under the fitted model M_o;
- the posterior predictive distribution of replicate data under M_o.

A very precise prior distribution on the hypothesized model is nearly equivalent to specifying exact values for the parameters of M_h. In that case our technique is similar to the traditional idea of specifying a particular alternative and using repeated samples from that alternative to address power or sample size concerns. One disadvantage of this traditional approach in high-dimensional problems is that it can be difficult to specify a small but still reasonably complete set of alternatives. Our fully Bayesian P^2S^2 approach provides an opportunity to average over a prior distribution of alternatives. Or in analogy with power curves, we can model Q_T empirically as a function of values of the parameters θ_h. Another difference between the approach described here and the traditional one is that the full posterior distributions of discrepancy measures are available for assessing model fit; the traditional approach relies on formal test statistics with fixed binary decision rules for accepting or rejecting a model. In addition, the justification for the test statistics used in traditional procedures is often based on asymptotic theory valid only for regular and nested models, whereas our Bayesian approach using posterior predictive discrepancies obtains valid posterior inferences regardless of sample size and with no critical restrictions on the models that can be used.

4.2 *Other study design issues.* Sample size naturally has a substantial effect on our ability to distinguish between two competing models. There are often other possible refinements of a study that may help provide a better assessment of the propriety of a hypothesized model. Our P^2S^2 technique can be used easily to address such refinements. For example, in some instances, studies obtain data via two levels of sampling (e.g., classrooms are sampled and then students within the classrooms provide data). Our P^2S^2 approach can be used to assess the relative benefits of obtaining more classrooms and more students within each classroom. Also, we can speculate on the effect of measuring an additional variable. In the psychology mixture model example, we find that it would be possible to invalidate the two-class model using only 250 observations *if* there were four observable variables rather than three (the P^2S^2 results are not provided here). This result is quite speculative because it assumes that

we can identify another observable variable that provides independent, reliable information about infant temperaments.

5. Concluding comments

Model assessment and model choice are active areas of research among frequentist and Bayesian statisticians. The Bayesian P^2S^2 approach demonstrates that advances in computational technology and algorithms have made it possible to perform common statistical tasks, like sample size determination for model assessment and choice, for increasingly complex models.

Acknowledgements. The authors thank Jerome Kagan, Nancy Snidman and Doreen Arcus for the infant temperament study data, and the National Science Foundation (NSF) for support under grant award DMS-9404479. They also thank a referee and the guest editors for their help. The computing for this research was performed on equipment purchased with funds provided by an NSF SCREMS grant award DMS-9707740.

References

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society Series A*, 143, 383–430.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B*, 39, 1–38.

Everitt, B. S. and Hand, D. J. (1981). *Finite Mixture Distributions*. London: Chapman and Hall.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*, London: Chapman and Hall.

Gelman, A., Meng, X., and Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, 6, 733–807.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7, 457–511.

Goodman, L. A. (1974a). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61, 215–231.

Goodman, L. A. (1974b). The analysis of a system of qualitative variables when some of the variables are unobservable. Part I: A modified latent structure approach. *American Journal of Sociology*, 79, 1179–1259.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773-795.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, 12, 1151–1172.

Rubin, D. B. and Stern, H. S. (1994). Testing in latent class models using a posterior predictive check distribution. In *Latent Variables Analysis: Applications for Developmental Research*, A. von Eye and C. C. Clogg (eds.), pp. 420–438. Sage Publications: Thousand Oaks, CA.

Stern, H. S., Arcus, D., Kagan, J., Rubin, D. B., and Snidman, N. (1994). Statistical choices in infant temperament research." *Behaviormetrika*, 21, pp. 1–17.

– – ––– (1995). Using mixture models in temperament research. *International Journal of Behavioral Development*, Vol. 18, pp. 407-423.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley: New York.

Donald B. Rubin
Department of Statistics
Harvard University
Cambridge, MA 02138
e-mail : rubin@hustat.harvard.edu

Hal S. Stern
Department of Statistics
Iowa State University
Ames, IA 50011
e-mail : hstern@iastate.edu