Model Checking and Model Improvment (chapter for Gilks, Richardson, and Spiegelhalter book)

Andrew Gelman
Department of Statistics
University of California
Berkeley, CA 94720

Xiao-Li Meng
Department of Statistics
University of Chicago
Chicago, IL 60637

0.1 Introduction

Markov chain simulation, and Bayesian ideas in general, allow a wonderfully flexible treatment of probability models. In this chapter, we discuss two related ideas: (1) checking the fit of a model to data, and (2) improving a model by adding substantively meaningful parameters. Model improvement by expansion is also an important technique in assessing the sensitivity of inferences to untestable assumptions. We illustrate both these methods with an example of a mixture model fit to experimental data from psychology using the Gibbs sampler.

Any Markov chain simulation is conditional on an assumed probability model. As the applied chapters of this book illustrate, these models can be complicated and generally rely on inherently unverifiable assumptions. From a practical standpoint, then, it is important to explore how inferences of substantive interest depend on the assumptions, and to test the assumptions where possible.

0.2 Model checking using posterior predictive simulation

Bayesian prior-to-posterior analysis conditions on the whole structure (likelihood and prior distribution) of a probability model and can yield very misleading inferences when the model is far from reality. A good Bayesian analysis, therefore, should at least check to see if the fit of the model to the data is so poor that the model should be rejected without other evidence. In the classical setting, this checking is often facilitated by a goodness-of-fit test, which quantifies the extremeness of the observed value of a selected measure of discrepancy (e.g., differences between observations and predictions) by calculating a tail-area probability given that the model under consideration is true. In this section, we discuss how to employ a Bayesian test of model fit using the posterior predictive distribution.

In the classical approach, test statistics cannot depend on any unknown quantities and so comparisons are actually made between the data and the best-fitting model from within a family of models (typically obtained via maximum likelihood). The p-value for the test is determined based on the sampling distribution of the data under the model. The main technical problem with the classical method is that, in general, the p-value depends on the unknown parameters unless we restrict attention to test statistics that are pivotal. Unfortunately, as is well known, most useful statistics are not pivotal, especially with complex models. The Bayesian formulation naturally allows the test statistic to be a function of both data and unknown parameters, and thus allow more direct comparisons between the sample and population characteristics. In the frequentist setting, using parameter-dependent test statistics was largely promoted recently by Tsui and Weerahandi (1989), who called such test statistics “test variables.”

Here we discuss model checking using posterior predictive tests, which was first proposed and applied by Guttman (1967) and Rubin (1981, 1984). Recently, Gelman, Meng, and Stern (1993) provide a general discussion of this method, with special emphasis of using test variables rather than traditional test statistics. They also present several applications and some theoretical properties of the method and compare to the prior predictive test of Box (1980). Meng (1994) presents a similar discussion of the use of posterior predictive p-values for testing hypotheses of parameters within a given model.

The posterior predictive model checking goes as follows. Let y be the observed data, θ be the vector of unknown parameters in the model (including any hierarchical parameters), $p(y|\theta)$ be the likelihood, and $p(\theta|y)$ be the posterior distribution. We assume that we have already obtained draws $\theta_1, \dots, \theta_N$ from the posterior distribution, possibly using Markov chain simulation. We now draw simulations from hypothetical replications of the data, which we label $y_1^{\text{rep}}, \dots, y_N^{\text{rep}}$. For each $i = 1, \dots, N$, we draw y_i^{rep} from the sampling distribution given the simulated parameters θ_i. Creating simulations y^{rep} is the old idea of comparing data to simulations from a model, with the Bayesian twist that the parameters of the model are themselves drawn from their posterior distribution.

If the model is reasonably accurate, the hypothetical replications should look similar to the observed data y. Formally, one can compare the data to the predictive distribution by choosing a *test variable*, $T(y, \theta)$, and computing the p-value, the proportion of cases in which the simulated test variables exceeds the realized value:

$$\text{estimated } p\text{-value} = \frac{1}{N} \sum_{i=1}^{N} I_{T(y_i^{\text{rep}}, \theta_i) > T(y, \theta_i)},$$

where I is the indicator function. We call $T(y, \theta)$ a "realized value" because it is *realized* by the observed data y, although it cannot be observed when T depends on unknown parameters. In the special case that the test variable depends only on data and not on parameters, and thus can be written $T(y)$, we call it a *test statistic*, as in the classical usage.

In practice, we often can visually examine the posterior predictive distribution of the test variable as it compares to the realized value. If the test variable depends only on data and not on the parameters, θ, then one can plot a histogram of the posterior predictive distribution of $T(y^{\text{rep}})$ and compare it to the observed value, $T(y)$. A good fit is indicated by an observed value near the center of the histogram. If the test variable is a function of data and parameters, one can plot a scatterplot of the realized values, $T(y, \theta_i)$, versus the predictive values, $T(y_i^{\text{rep}}, \theta_i)$. A good fit is indicated by about half the points in the scatterplot falling above the 45° line and half falling below.

The test variable can be any function of data and parameters. It is most useful to choose a test variable that measures some aspect of the data that might not accurately be fit by the model. For example, if one is concerned with outliers in a normal regression model, a sometimes useful test variable is the proportion of residuals that are more than three standard deviations away from zero. If one is concerned about overall lack of fit in a contingency table model, a χ^2 discrepancy measure can be used (see Gelman, Meng, and Stern, 1993). One of course can use more than one test variable to check different aspects of the fitness. An advantage of Monte Carlo is that the same set of simulations of $(\theta, y^{\mathrm{rep}})$ can be used for checking the posterior predictive distribution of many test variables.

A model does not fit the data if the realized values for some meaningful test variable are far from the predictive distribution; the discrepancy cannot reasonably be explained by chance if the tail-area probability is close to 0 or 1. The p-values are actual posterior probabilities and can therefore be interpreted directly—*not* as the posterior probability of the model being true. The role of predictive model checking is to assess the practical fit of a model, *not* to estimate the "probability that the model is true," whatever that means. We may choose to work with an invalidated model but we should be aware of its deficiencies. On the other hand, a lack of rejection should not be interpreted as "acceptance" of the model, but rather as a sign that the model adequately fits the aspect of the data being tested.

Major failures of the model can be addressed by expanding the model, as we discuss in the next section. Lesser failures may also suggest model improvements or might be ignored in the short term if the failure appears not to affect the main inferences. In some cases, even extreme p-values may be ignored if the misfit of the model is substantively small compared to variation within the model. It is important not to interpret p-values as numerical "evidence." For example, a p-value of 0.00001 is virtually no stronger, in practice, than 0.001; in either case, the aspect of the data measured by the test variables is inconsistent with the model. A slight improvement in the model (or correction of a data coding error!) could bring either p-value to a reasonable range (between 0.05 and 0.95, say).

0.3 Model improvement via expansion and averaging

We will address two issues that are mathematically and statistically essentially the same: *expanding* a model that is already set up on the computer and has been estimated from the data, and *averaging* several competing models that have been estimated from the same data. The latter problem is sometimes posed as "model selection," but from the perspectives of both Bayesian methodology and Markov chain computation, it is more natural to think of averaging over a mixture of several models, rather than choosing a single distribution with certainty.

There are three natural reasons to expand a model. First, if the model clearly does not fit the data or prior knowledge, it should be improved in some way, possibly by adding new parameters to allow a better fit. Second, if a modeling assumption is particularly questionable, it may be relaxed. For example, a set of parameters that are fixed to equality may be replaced by a random effects model. Third, a model may be embedded into a larger model to address more general applied questions; for example, an study previously analyzed on its own may be inserted into a hierarchical population model (e.g., Rubin, 1981). The goal of our model expansion is not merely to fit the data, but rather to improve the model to better capture the substantive structures. Thus, when one adds new parameters, a key requirement is that these parameters should have clear substantive meaning. This point is illustrated in our example in the next section.

All these applications of model expansion have the same mathematical structure: the old model, $p(\theta)$, is replaced by a new model, $p(\theta, \phi)$, in which both θ and ϕ may be vectors. In Bayesian notation, the posterior distribution $p(\theta|y)$ is replaced by $p(\theta, \phi|y)$, with a prior distribution required for the additional parameters ϕ. Assuming that the original model in θ has already been programmed for Markov chain simulation, one can immediately use the Gibbs sampler to draw samples from the joint distribution, $p(\theta, \phi|y)$. The step of drawing from $p(\theta|\phi, y)$ is just the problem of drawing from $p(\theta|y)$ in the individual model identified by ϕ. The only new step required is sampling from among the models, given data and parameters: $p(\phi|\theta, y)$. If the model class has no simple form, this additional simulation can be performed using a Metropolis step.

For instance, suppose we are interested in the sensitivity of a data analysis with possible outliers to a model with an assumed normal distribution. A natural expansion would be to replace the normal by a Student-t with unknown degrees of freedom, ν. For a Bayesian analysis, a prior distribution must be assumed for ν; a reasonable noninformative prior density is uniform in $1/\nu$; or, after transformation, $p(\nu) \propto 1/\nu^2$, with the t distributions restricted to the range $\nu \geq 1$, so that the limits are the Cauchy and normal distributions. (The uniform prior density on ν may seems reasonable at first, but it actually has essentially all its mass "near" $\nu = \infty$, which corresponds to the normal distribution.) The Markov chain simulation for the Student-t model is now obtained by altering the simulations based on the normal model to simulations conditional on ν, and an added step to draw ν. Since the t distributions are not conjugate, it is probably easiest to use the Metropolis algorithm to correct for draws based on the normal distribution at each Gibbs step.

Additional technique is required to average over, or choose among, a set of models that do not follow a single parametric family. With no parameters in common, the above Gibbs approach makes no sense. Carlin and Chib (1993) present a clever solution, defining a joint distribution on the space of models and the union of all the model parameters and applying the Gibbs sampling to a larger space. When averaging over models that are not part of a single parametric family, or models that have different sets of parameters, the above methods only work when the individual models have proper prior distributions. Kass and Raftery (1994) discuss this point and review some methods that have been proposed for circumventing this difficulty.

0.4 Example: hierarchical mixture modeling for reaction-time data

Neither model checking or model expansion is enough by itself. Once lack of fit has been found, the next step is to find a model that improves the fit, with the requirement that the new model should at least as interpretable substantively as the old one. On the other hand, model expansion alone can never reveal lack of fit of the larger, expanded model. In this example, we illustrate how model checking and expansion can be used in tandem.

The data and the basic model, fit by the Gibbs sampler

Belin and Rubin (1990) describe an experiment from psychology measuring thirty reaction times for each of seventeen subjects: eleven non-schizophrenics and six schizophrenics. Belin and Rubin (1994) fit several probability models using maximum likelihood; we describe their approach and fit related Bayesian models. Computation with the Gibbs sampler allows us to fit more realistic hierarchical models to the dataset.

The data are presented in Figure 0.1. It is clear that the response times are higher on average for schizophrenics. In addition, the response times for at least some of the schizophrenic individuals are considerably more variable than the response times for the non-schizophrenic individuals. Current psychological theory suggests a model in which schizophrenics suffer from an attentional deficit on some trials, as well as a general motor reflex retardation; both aspects lead to a delay in the schizophrenics' responses, with motor retardation affecting all trials and attentional deficiency only some.

To address the questions of scientific interest, the following basic model was fit. Response times for non-schizophrenics are thought of as arising from a normal random effects model, in which the responses of person $i = 1, \ldots, 11$ are normally distributed with distinct person mean α_i and common variance σ^2_{obs}. To reflect the attentional deficiency, the responses for each schizophrenic individual $i = 12, \ldots, 17$ were fit to a two-component mixture: with probability $(1 - \lambda)$, there is no delay and the response is normally distributed with mean α_i and variance σ^2_{obs}, and with probability λ, responses are delayed, with observations having a mean of $\alpha_i + \tau$ and the same variance. Because the reaction times are all positive and their

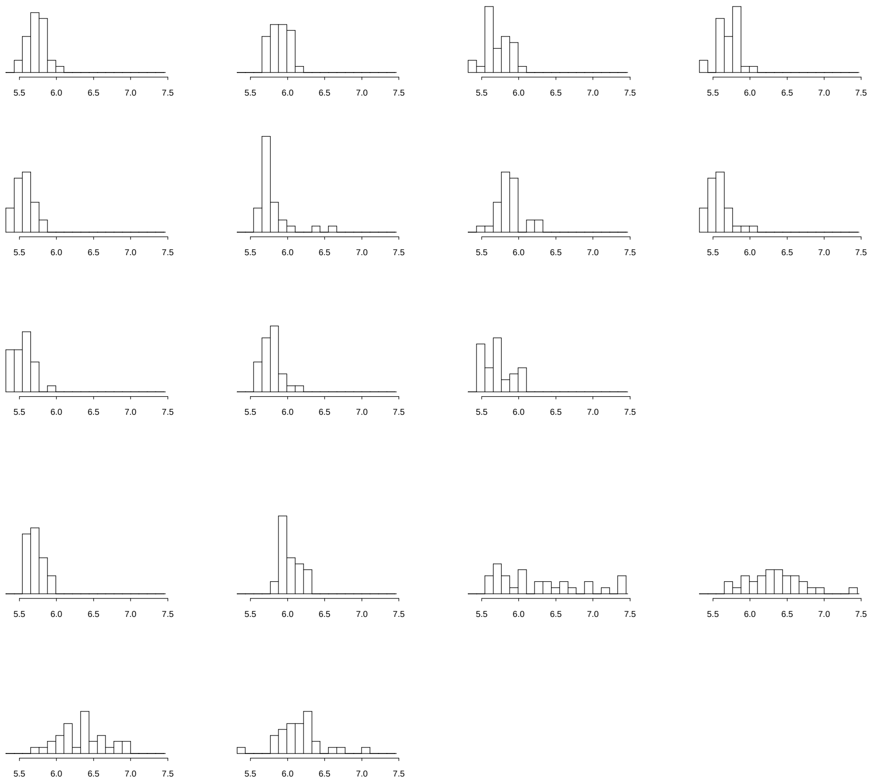

Figure 0.1. *(a) Log response times for 11 non-schizophrenic individuals, (b) Log response times for 6 schizophrenic individuals. All histograms are on a common scale.*

distributions are positively skewed, even for non-schizophrenics, the model was fit to the logarithms of the reaction time measurements. We modify the basic model of Belin and Rubin (1994) to incorporate a hierarchical parameter β measuring motor retardation. Specifically, variation among individuals is modeled by having the person means α_i follow a normal distribution with mean $\nu + \beta S_i$ and variance σ^2_α, where ν is the overall mean response time of non-schizophrenics, and S_i is an observed indicator that is 1 if person i is schizophrenic and 0 otherwise.

Letting y_{ij} be the jth response of individual i, the model can be written in the following hierarchical form:

$$y_{ij}|\alpha_i, z_{ij}, \phi \quad \sim \quad N(\alpha_i + \tau z_{ij}, \sigma^2_{obs}),$$

$$\begin{aligned} \alpha_i | z, \phi &\sim N(\nu + \beta S_i, \sigma_\alpha^2), \\ z_{ij} | \phi &\sim \text{Bernoulli}(\lambda S_i), \end{aligned}$$

where $\phi = (\log(\sigma_\alpha^2), \beta, \text{logit}(\lambda), \tau, \nu, \log(\sigma_{obs}^2))$, and z_{ij} is an unobserved indicator variable that is 1 if measurement j on person i arose from the delayed component and 0 if it arose from the undelayed component. The indicator random variables z_{ij} are not necessary to formulate the model, but allow convenient computation of the modes of (α, ϕ) using the iterative ECM algorithm (Meng and Rubin, 1993) and simulation using the Gibbs sampler. For the Bayesian analysis, the parameters σ_α^2, β, λ, τ, ν, and σ_{obs}^2 are assigned a joint uniform prior distribution, except that λ is restricted to the range $[0.001, 0.999]$, τ is restricted to be positive to identify the model, and σ^2 and σ_{obs}^2 are of course restricted to be positive.

We found the modes of the posterior distribution using the ECM algorithm and then used a mixture of multivariate Student-t densities centered at the modes as an approximation to the posterior density. We drew ten samples from the approximate density using importance resampling and then used those as starting points for ten parallel runs of the Gibbs sampler, which adequately converged (in the sense of potential scale reductions $\widehat{R}$ less than 1.1 for all model parameters; see Gelman, 1994) after 200 steps, after discarding the first half of each simulated sequence. We were left with a set of 1000 simulation draws of the vector of model parameters. Details of the Gibbs sampler implementation and the convergence monitoring appear in Gelman et al. (1994) and Gelman and Rubin (1992), respectively.

Model checking using posterior predictive simulations

The model was chosen to accurately fit the unequal means and variances in the two groups of subjects in the study, but there was still some question about the fit to individuals. In particular, the measurements for the first two schizophrenics seem much less variable than the last four. Is this difference "statistically significant," or could it be explained as a random fluctuation from the model? To compare the observations to the model, we compute, s_i, the standard deviation of the 30 log reaction times y_{ij}, for each schizophrenic individual $i = 12, \ldots, 17$. We then defined three test statistics—the smallest, largest, and average of the six values s_i—which we label $T_{\min}(y)$, $T_{\max}(y)$, and $T_{\text{avg}}(y)$, respectively. These are test statistics and not just test variables because they are defined in terms of data alone. Examination of Figure 0.1 suggests that $T_{\min}$ is too low and $T_{\max}$ too high than would be predicted from the model. The third test statistic, T_{avg}, is included as a comparison; we expect it to be very close to the model's predictions, since it is essentially estimated by the model parameters σ^2, τ, and λ. For the data in Figure 0.1, the observed values of the test statistics are $T_{\min}(y) = 0.11$, $T_{\max}(y) = 0.58$, and $T_{\text{avg}}(y) = 0.30$.

To perform the posterior predictive model check, we simulate a predictive dataset from the normal-mixture model, for each of the 1000 simulation draws of the parameters from the posterior distribution. For each of those 1000 simulated datasets $y^{\rm rep}$, we compute the test statistics $T_{\rm min}(y^{\rm rep})$, $T_{\rm max}(y^{\rm rep})$, and $T_{\rm avg}(y^{\rm rep})$. Figure 0.2 displays histograms of the 1000 simulated values of the each statistic, with the observed values, $T(y)$, indicated by vertical lines. The observed data y are clearly atypical of the posterior predictive distribution—$T_{\rm min}$ is too low and $T_{\rm max}$ is too high—with estimated p-values within 1/1000 of 0 and 1. In contrast, $T_{\rm avg}$ is well fit by the model, with a p-value of 0.72. More important than the p-values is the poor fit on the absolute scale: the observed minimum and maximum within-schizophrenic standard deviations are off by factors of two compared to the posterior model predictions.

Expanding the model

Following Belin and Rubin (1992), we try to better fit the data by including two additional parameters in the model: to allow for some schizophrenics to have no attentional delays and for delayed observations to be more variable than undelayed observations. The two new parameters are ω, the probability that each schizophrenic individual has attentional delays, and σ^2_{obs2}, the variance of attention-delayed measurements. Both these parameters are given uniform prior densities (we use the uniform density on ω because it is proper and the uniform density on σ^2_{obs2} to avoid the singularities at $\sigma^2 = 0$ that occur with uniform prior densities on the log scale for hierarchical and mixture models). For computational purposes, we also introduce another indicator variable for each individual i, W_i, that is 1 if the individual can have attention delays and 0 otherwise. The indicator W_i is automatically 0 for non-schizophrenics and, for each schizophrenic, is 1 with probability ω.

The model we have previously fit can be viewed as a special case of the new model, with $\omega = 1$ and $\sigma^2_{obs2} = \sigma^2_{obs}$ previously. It is quite simple to fit the new model by just adding three new steps in the Gibbs sampler to update ω, σ^2_{obs2}, and W (parameterized to allow frequent jumps between the states $W_i = 1$ and $W_i = 0$ for each schizophrenic individual i. In addition, the Gibbs sampler steps for the old parameters must be altered somewhat to be conditional on the new parameters. We do not give the details here but just present the results. We use ten randomly selected draws from the previous posterior simulation as starting points for ten parallel runs of the Gibbs sampler. Because of the added complexity of the model, we ran the simulations for 500 steps, and discarded the first half of each sequence, leaving a set of 2500 draws from the posterior distribution of the larger model. The potential scale reductions $\widehat{R}$ for all parameters were less than 1.1, indicating approximate convergence.

Before performing posterior predictive checks, it makes sense to compare

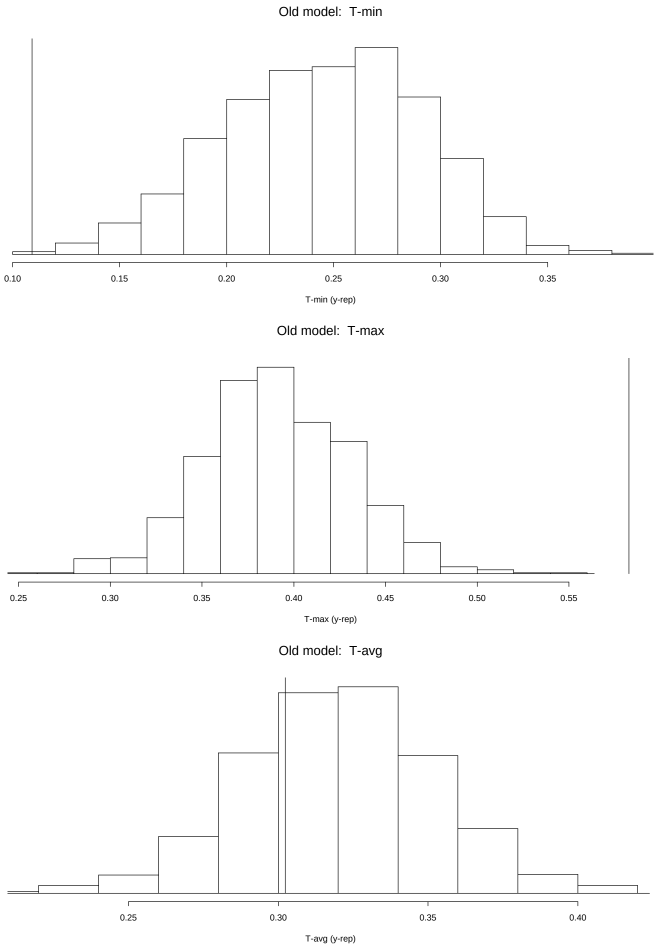

Figure 0.2. *Posterior predictive distributions and observed values for three test statistics: the smallest, average, and largest observed within-schizophrenic variances. The vertical line on each histogram represents the observed value of the test statistic.*

the old and new models in their posterior distributions for the parameters. Table 0.1 displays posterior medians and 95% intervals (from the Gibbs sampler simulations) for the parameters of most interest to the psychologists:

- λ, the probability that an observation will be delayed, for an individual subject to attentional delays;
- ω, the probability that a schizophrenic will be subject attentional delays;

	Old model			New model		
Parameter	2.5%	median	97.5%	2.5%	median	97.5%
λ	0.07	0.12	0.18	0.46	0.64	0.88
ω		fixed at 1		0.24	0.56	0.84
τ	0.74	0.85	0.96	0.21	0.42	0.60
β	0.17	0.32	0.48	0.07	0.24	0.43

Table 0.1. *Posterior quantiles for parameters of interest under the old and new mixture models for the reaction time experiment.*

- τ, the attentional delay (on the log scale);
- β, the average log response time for the non-delayed observations of schizophrenics minus the average log response time for non-schizophrenics.

Table 0.1 shows a significant difference between the parameter estimates in the two models. Since the old model is nested within the new model, the changes suggest that the improvement in fit is significant. It is a strength of the Bayesian approach, as implemented by iterative simulation, that we can model so many parameters and compute summary inferences for all of them.

Checking the new model

The expanded model is an improvement, but how well does it fit the data? We expect that the new model should show an improved fit to the test statistics considered in Figure 0.2, since the new parameters were added explicitly for this purpose. We emphasize that all the new parameters here have substantive interpretations in psychology. To check the fit of this new model, we use posterior predictive simulation of the same test statistics under the new posterior distribution. The results are displayed in Figure 0.3.

Once again, the vertical lines indicates the observed test statistic. Comparing to Figure 0.2, the observed values are in the same place, but the posterior predictive distributions have moved closer to them. (The histograms of Figures 0.2 and 0.3 are not plotted on a common scale.) However, the fit is by no means perfect in the new model: the observed values of $T_{\min}$ and $T_{\max}$ are still in the periphery, and the estimated p-values of the two test statistics are 0.98 and 0.03. (The average variance statistic, T_{avg}, is still fit well by the expanded model, with a p-value of 0.81.) The lack of fit is more visible on Figure 0.3, and the p-values provide probability summaries. We are left with an improved model that still shows some lack of fit, suggesting directions for improved modeling and data collection.

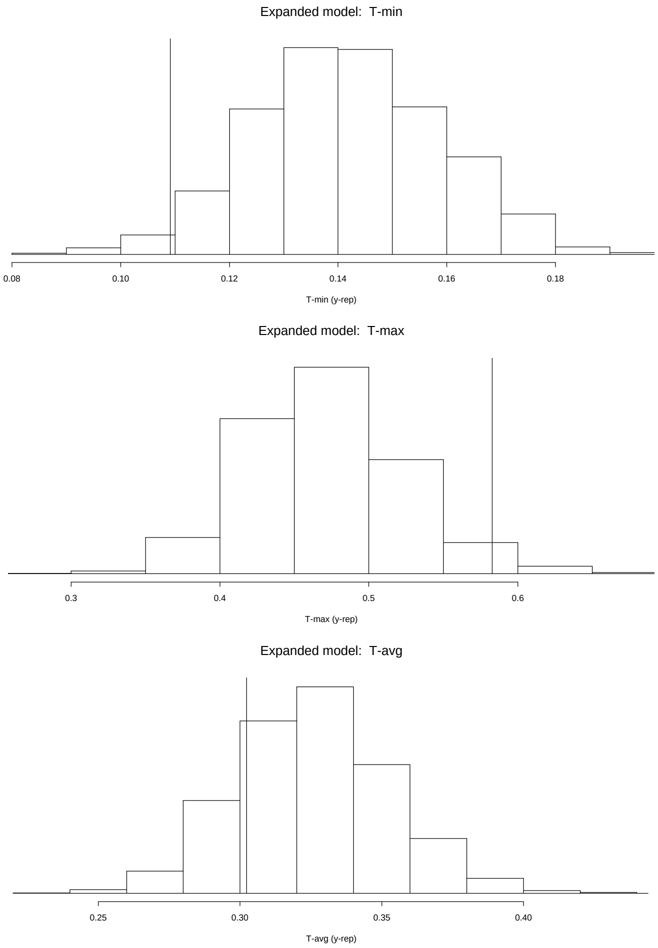

Figure 0.3. *Posterior predictive distributions, under the expanded model, for three test statistics: the smallest, average, and largest observed within-schizophrenic variances. The vertical line on each histogram represents the observed value of the test statistic.*

0.5 Some current research topics

In our experience (Gelman, Meng, and Stern, 1993), it appears that p-values for test variables that depend on both the data and the unknown parameters behave slightly differently than those based only on the data. This topic is worth exploring further, especially for examples such as hierarchical regression in which test variables are much more conveniently understood and computed when defined in terms of both data and parameters (for example, residuals from $X\beta$, rather than from $X\hat{\beta}$).

On another topic, we have treated model checking and model expansion as separate procedures. Various approaches have been suggested for combining the two—checking models by comparing them to formal alternatives (Gelfand, Dey, and Chang, 1992; Kass and Raftery, 1994). Another related area is theoretical studies of sensitivity analysis (e.g., Wasserman, 1992).

References

Belin, T. R., and Rubin, D. B. (1990). Analysis of a finite mixture model with variance components. *Proceedings of the American Statistical Association, Social Statistics Section*, 211–215.

Belin, T. R., and Rubin, D. B. (1994). The analysis of repeated-measures data on schizophrenic reaction times using mixture models. *Statistics in Medicine*, to appear.

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

Carlin, B. P., and Chib, S. (1993). Bayesian model choice via Markov chain Monte Carlo. Technical report, Division of Biostatistics, School of Public Health, University of Minnesota.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. In *Bayesian Statistics 4*, ed. J. M. Bernardo et al., 147–167. New York: Oxford University Press.

Gelman, A. (1994). Inference and monitoring convergence. In this volume.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1994). *Bayesian Data Analysis.* Chapman and Hall, to appear.

Gelman, A., Meng, X. L., and Stern, H. S. (1993). Bayesian model checking using tail area probabilities. Technical report, Department of Statistics, University of California, Berkeley.

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

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.

Kass, R. E., and Raftery, A. E. (1994). Bayes factors and model uncertainty. *Journal of the American Statistical Association*, to appear.

Meng, X. L. (1994). Posterior predictive p-values. *Annals of Statistics* **22**, to appear.

Meng, X. L., and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* **80**, 267–278.

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics* **6**, 377–401.

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

Tsui, K. W., and Weerahandi, S. (1989). Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. *Journal of American Statistical Association* **84**, 602–607.

Wasserman, L. (1992). Recent methodological advances in robust Bayesian inference. In *Bayesian Statistics 4*, ed. J. M. Bernardo et al., 438–502. New York: Oxford University Press.