Model Checking

Anita Wang

2017 - 2018

Model Checking

"All models are wrong but some models are useful"

--- George E. P. Box

- Checking the model is crucial to statistical analysis.
- Bayesian prior-to-posterior inferences assume the whole structure of a probability model and can yield misleading inferences when the model is poor.
- A good Bayesian analysis, therefore, should include at least some check of the adequacy of the fit of the model

Model Checking

- So far we have looked at a number of models and examined them with example data sets. Do the models used accurately describe the data used?
- In standard analyses, we will often check model assumptions. For example, in standard regression we will check for
  - Correct form of the regression function (e.g. linear vs quadratic)
  - Constant variance of the residuals
  - Independence and normality of the residuals

Model Checking

- Basic question: How sensitive are our posterior inferences to our modelling assumptions?
- Student sleeping time: Will the following models give significantly different answers about the probability of heavy sleepers?
- Original Model:
  - Data model: y is the number of heavy sleepers

$$y|\theta \sim Bin(n, \theta)$$

  - Prior: θ the probability of heavy sleepers

$$\theta \sim Beta(\alpha, \beta)$$

Model Checking

- Alternative model 1:
  - Data model: y is the number of heavy sleepers

$$y|\theta \sim Bin(n, \theta)$$

  - Prior: θ the probability of heavy sleepers

$$logit(\theta) \sim N(\mu, \sigma^2)$$

where $logit(\theta) = \log\frac{\theta}{1-\theta}$

- Alternative model 2:
  - Data model: y is the number of heavy sleepers

$$y|\theta \sim Beta - bin(n, \alpha, \beta)$$

  - Prior: (α, β) probability parameters

$$\alpha, \beta \sim Gamma(\gamma_\alpha, \delta_\alpha)Gamma(\gamma_\beta, \delta_\beta)$$

Model Checking

- Note that we will not be trying to answer the question of whether our model is correct or not. We are interested in whether the inaccuracies matter.
- One approach to build a super-model that contains all of our models of interest as special cases. This approach usually isn't taken as it is usually difficult to build this super-model and computation is usually infeasible, assuming you can build the model.
- Instead we will base these checks on the posterior predictive distribution. Does our data look like our fitted model says it should.

Model Checking

- Check on the posterior predictive distribution:
  - *External validation*: future data is compared with the posterior predictive distribution.
  - *Internal validation*: observed data is compared with the posterior predictive distribution.

- Compare the outliers and graphs of the future (existing) data and simulated data from posterior predictive distribution $p(\tilde{y}|y)$

Example: speed of light

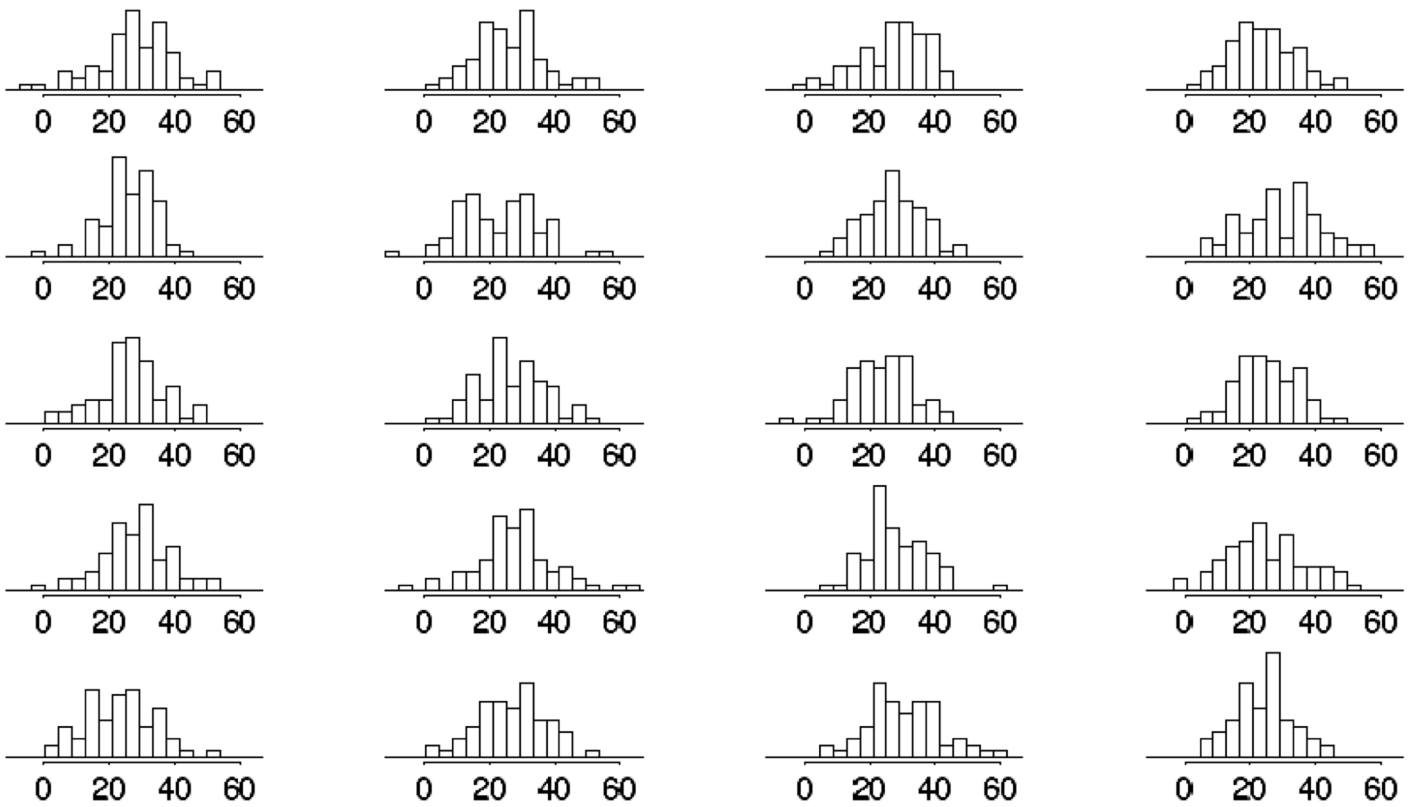


Figure 6.2 *Twenty replications, y^{rep}, of the speed of light data from the posterior predictive distribution, $p(y^{\text{rep}}|y)$; compare to observed data, y, in Figure 3.1. Each histogram displays the result of drawing 66 independent values $\tilde{y}_i$ from a common normal distribution with mean and variance (μ, σ^2) drawn from the posterior distribution, $p(\mu, \sigma^2|y)$, under the normal model.*

Example: speed of light

- Figure displays twenty histograms, each of which represents a single draw from the posterior predictive distribution of the values in Newcomb's experiment
- first drawing (μ, σ^2) from their joint posterior distribution, then drawing 66 values from a normal distribution with this mean and variance.
- All these histograms look different from the histogram of actual data in the next slide
- One way to measure the discrepancy is to compare the smallest value in each hypothetical replicated dataset to Newcomb's smallest observation, −44.

Example: speed of light

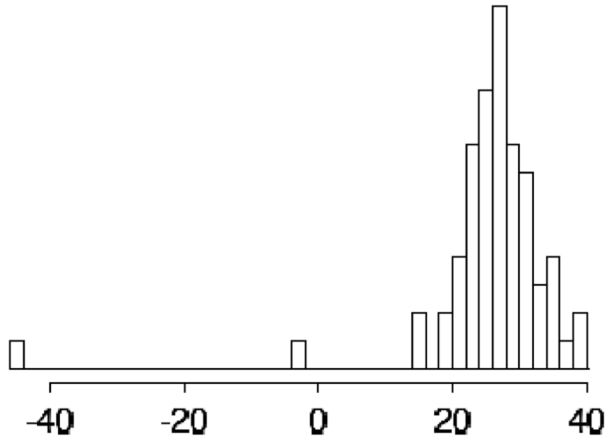


Figure 3.1 *Histogram of Simon Newcomb's measurements for estimating the speed of light, from Stigler (1977). The data are recorded as deviations from* 24,800 *nanoseconds.*

Example: speed of light

The histogram below shows the smallest observation in each of the 20 hypothetical replications; all are much larger than Newcomb's smallest observation

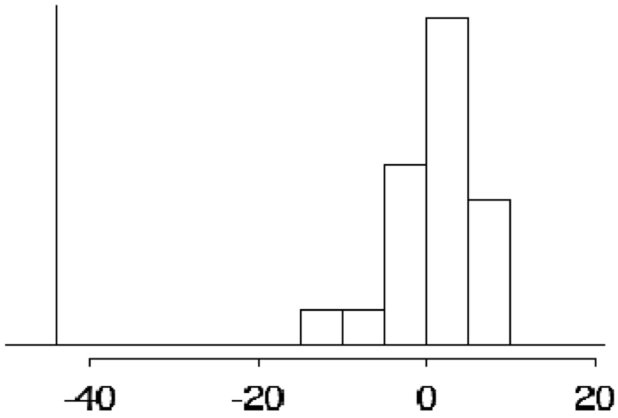
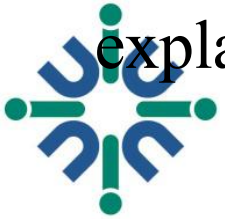


Figure 6.3 *Smallest observation of Newcomb's speed of light data (the vertical line at the left of the graph), compared to the smallest observations from each of the 20 posterior predictive simulated datasets displayed in Figure 6.2.*

Posterior Predictive Checking

- Idea: If the model fits, replicated data generated under the model should look similar to the observed data.
- If we see some discrepancy, is it due to model misspecification or due to chance.
- Approach: Generate L datasets, $y_1^{rep}, \ldots, y_L^{rep}$ from the posterior predictive distribution $p(y^{rep}|y)$. y^{rep} corresponds to replicated data.
- $\tilde{y}$ represents any future outcome whereas y^{rep} indicates a replication exactly like the observed y.
- For example, if the model has explanatory variables, x, they will be identical for y and y^{rep}, but $\tilde{y}$ may have its own explanatory variables, $\tilde{x}$.



Test quantities

- We measure the discrepancy between model and data by defining test quantities
- A test quantity (discrepancy measure), $T(y, \theta)$, is a scalar summary of parameters and data when comparing data to predictive simulations.
- We use the notation $T(y)$ for a test statistic, which is a test quantity that depends only on data
- In the Bayesian context, we can generalize test statistics to allow dependence on the model parameters under their posterior distribution

Tail-area probabilities

- The lack of fit of the data as compared to the posterior predictive distribution can be compared by a tail-area probability (e.g. p-value) of the test statistic $T(y, \theta)$. To calculate this probability we will use the replicates sampled from $p(y^{rep}|y)$.
- Classical p-value

$$p_C = P[T(y^{rep}) \geq T(y)|\theta]$$

where the probability is calculated over the distribution of y^{rep} given a fixed θ. In the classical testing setting θ would correspond to the null hypothesis value. It could also be a point estimate (say the MLE).

Posterior predictive *p*-values

- To evaluate the fit of the posterior distribution of a Bayesian model, we can compare the observed data to the posterior predictive distribution
- In the Bayesian approach, test quantities can be functions of the unknown parameters as well as data
- The Bayesian p-value is defined as the probability that the replicated data could be more extreme than the observed data,

$$p_B = P[T(y^{rep}, \theta) \geq T(y, \theta)|y]$$

$$= \iint I\big(T(y^{rep}, \theta) \geq T(y, \theta)\big) p(y^{rep}|\theta) p(\theta|y)\, dy^{rep} d\theta$$

Posterior predictive *p*-values

- Usually we can't calculate the Bayesian p-value exactly, but can do it by simulation. Suppose that we have L simulations of $\theta(\theta^1, \ldots, \theta^L)$ from the posterior distribution $p(\theta|y)$. Then for each of these θ samples, generate one sample y^{rep_l} from $p(y^{rep}|\theta^l)$.
- We want to compare each of the $T(y^{repl}, \theta^l)$ with $\mathrm{T}(y, \theta^l)$ Then

$$\hat{p}_B = \frac{1}{L}\sum_{l=1}^{L} I(T(y^{repl}, \theta^l) \geq \mathrm{T}(y, \theta^l))$$

(i.e. the proportion of samples where $T(y^{repl}, \theta^l) \geq \mathrm{T}(y, \theta^l)$) is an estimate of p_B.

Posterior predictive *p*-values

- Note that the test statistic $T(y, \theta)$ needs to be chosen to investigate deviations of interest. This is similar to choosing a powerful test statistic when conducting a hypothesis test
- For example, in the analysis of Newcomb's speed of light experiment, a worry was the effect of outliers. Thus $T(y, \theta)$ needs to be chosen to focus on this issue.
- Previously, we use $T(y,\theta) = \min y_i$ as test statistics to demonstrate the poor fit of the normal model to the speed of light data
- We try to use other test quantities

Posterior predictive *p*-values

- Variance:

The sample variance does not make a good test statistic because it is a sufficient statistic of the model

With non-informative prior, the posterior distribution will automatically be centered near the observed value.

- A test quantity sensitive to asymmetry

$$T(y, \theta) = |y_{(61)} - \theta| - |y_{(6)} - \theta|.$$

The 61st and 6th order statistics are chosen to represent approximately the 90% and 10% points of the distribution. The test quantity should be scattered about zero for a symmetric distribution.

Choosing test quantities

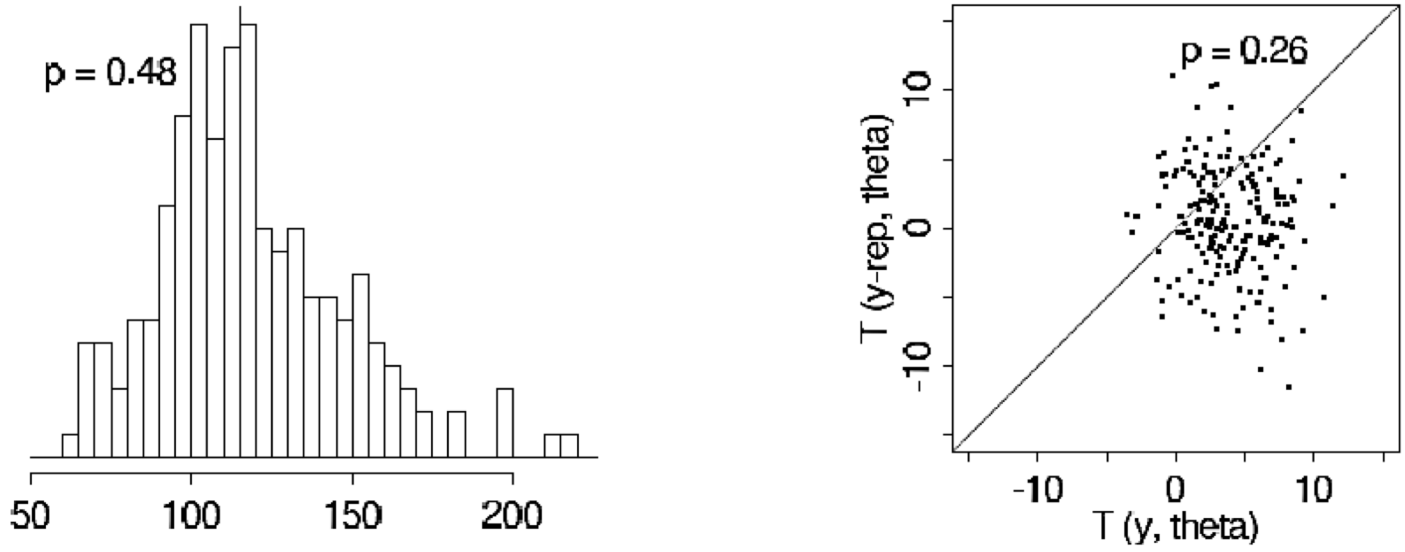
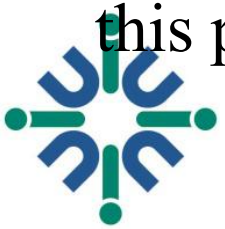


Figure 6.4 *Realized vs. posterior predictive distributions for two more test quantities in the speed of light example: (a) Sample variance (vertical line at 115.5), compared to 200 simulations from the posterior predictive distribution of the sample variance. (b) Scatterplot showing prior and posterior simulations of a test quantity:* $T(y,\theta) = |y_{(61)} - \theta| - |y_{(6)} - \theta|$ *(horizontal axis) vs.* $T(y^{\text{rep}},\theta) = |y^{\text{rep}}_{(61)} - \theta| - |y^{\text{rep}}_{(6)} - \theta|$ *(vertical axis) based on 200 simulations from the posterior distribution of* (θ, y^{rep})*. The p-value is computed as the proportion of points in the upper-left half of the scatterplot.*

Interpreting posterior predictive p-values

- Major failures of the model, typically corresponding to extreme tail-area probabilities (less than 0.01 or more than 0.99.
- Lesser failures might also suggest model improvements or might be ignored in the short term if the failure appears not to affect the main inferences.
- even extreme p-values may be ignored if the misfit of the model is substantively small compared to variation within the model.
- We typically evaluate a model with respect to several test quantities, and we should be sensitive to the implications of this practice.



Graphical posterior predictive checks

- Direct display of all the data
- Display of data summaries or parameter inferences.

This can be useful in settings where the dataset is large and we wish to focus on the fit of a particular aspect of the model.

- Graphs of residuals or other measures of discrepancy between model and data.

Direct data display

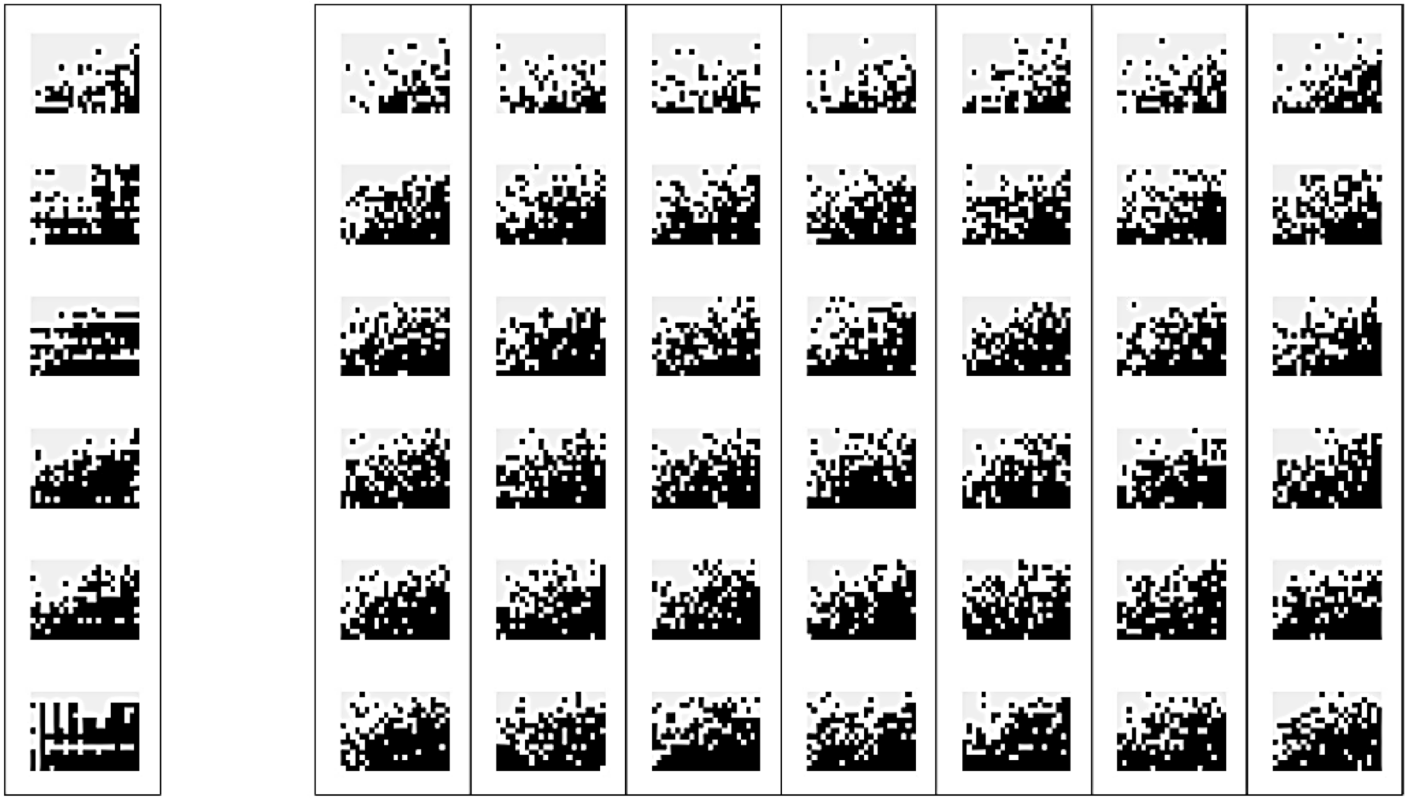


Figure 6.7 *Left column displays observed data y (a 15 × 23 array of binary responses from each of 6 persons); right columns display seven replicated datasets y^{rep} from a fitted logistic regression model. A misfit of model to data is apparent: the data show strong row and column patterns for individual persons (for example, the nearly white row near the middle of the last person's data) that do not appear in the replicates. (To make such patterns clearer, the indexes of the observed and each replicated dataset have been arranged in increasing order of average response.)*

Direct data display

- The left column of the figure displays binary data

for each of 6 persons, a possible ‘yes’ or ‘no’ to each of 15 possible reactions (displayed as rows) to 23 situations (columns)—from an experiment in psychology

- The right columns of the figure display seven independently simulated replications y^{res} from a fitted logistic regression model
- Before displaying, the reactions, situations, and persons have been ordered in increasing average response.
- the replicated datasets look ‘random’ compared to the observed data, which have strong rectilinear structures that are clearly not captured in the model.

Displaying summary statistics or inferences

- the model included two vectors of parameters, $\varphi_1, \ldots, \varphi_{90}$, and $\psi_1, \ldots, \psi_{69}$, corresponding to patients and psychological symptoms, and that each of these 159 parameters were assigned independent Beta(2, 2) prior distributions.
- Data were collected (measurements of which symptoms appeared in which patients) and the full Bayesian model was fitted, yielding posterior simulations for all these parameters.
- If the model were true, we would expect any single simulation draw of the vectors of patient parameters φ and symptom parameters ψ to look like independent draws from the Beta(2, 2) distribution.

Displaying summary statistics or inferences

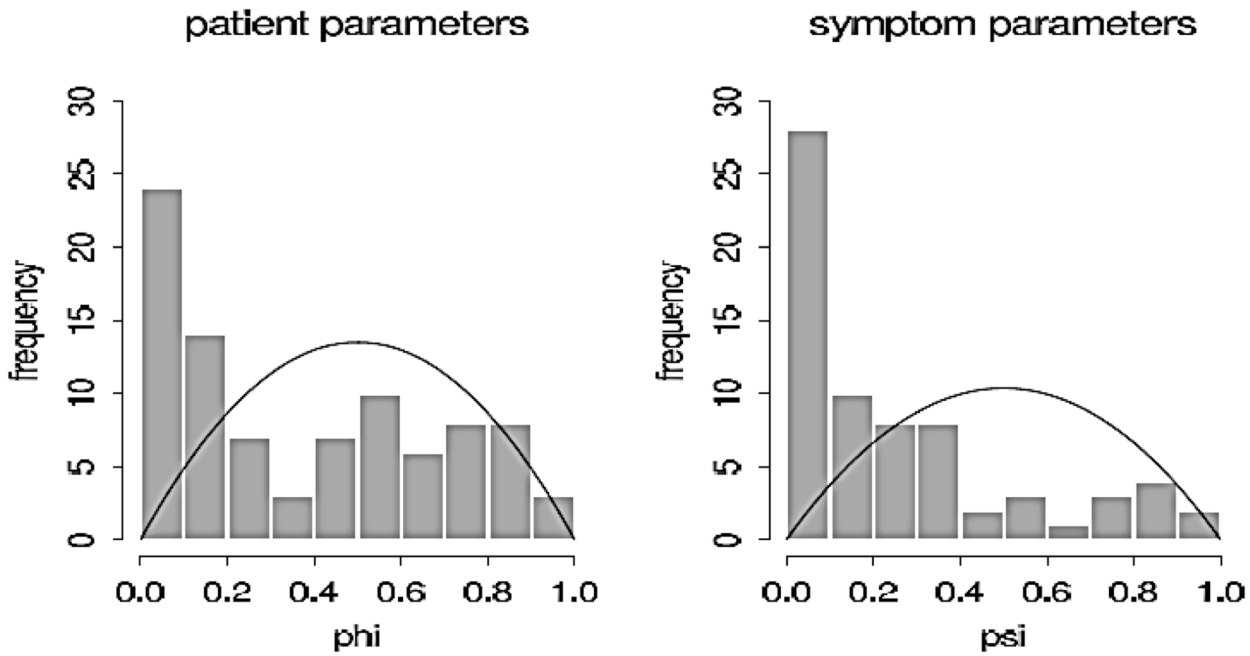


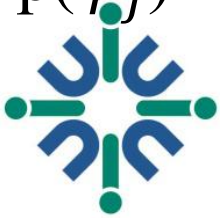
Figure 6.9 *Histograms of (a) 90 patient parameters and (b) 69 symptom parameters, from a single draw from the posterior distribution of a psychometric model. These histograms of posterior estimates contradict the assumed Beta(2, 2) prior densities (overlain on the histograms) for each batch of parameters, and motivated us to switch to mixture prior distributions. This implicit comparison to the values under the prior distribution can be viewed as a posterior predictive check in which a new set of patients and a new set of symptoms are simulated.*

Displaying summary statistics or inferences

- as a model check we can plot a histogram of a single simulation of the vector of parameters φ or ψ and compare to the prior distribution
- The lines in the figure show the Beta(2, 2) prior distribution, which clearly does not fit.
- For both φ and ψ, there are too many cases near zero, corresponding to patients and symptoms that almost certainly are not associated with a particular syndrome.
- Consider:

$$p(\varphi_j) = 0.5\ \text{Beta}(\varphi_j|1, 6) + 0.5\ \text{Beta}(\varphi_j|1, 1)$$

$$p(\psi_j) = 0.5\ \text{Beta}(\psi_j|1, 16) + 0.5\ \text{Beta}(\psi_j|1, 1).$$



Displaying summary statistics or inferences

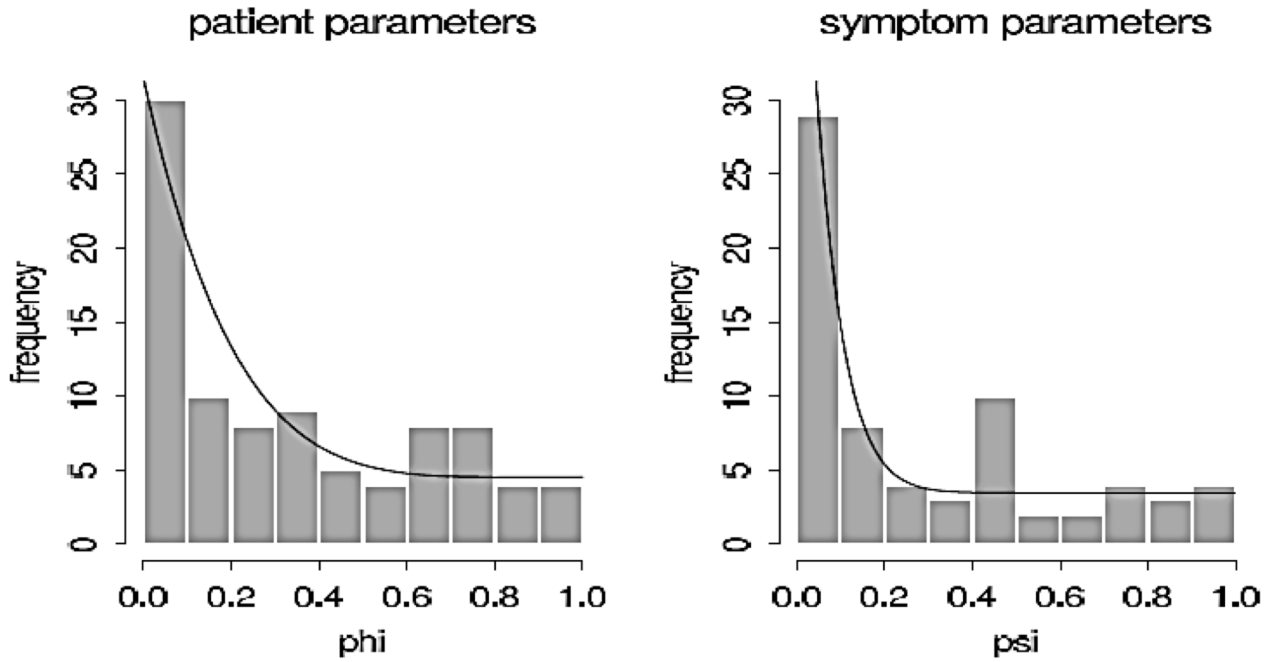


Figure 6.10 *Histograms of (a) 90 patient parameters and (b) 69 symptom parameters, as estimated from an expanded psychometric model. The mixture prior densities (overlain on the histograms) are not perfect, but they approximate the corresponding histograms much better than the Beta*$(2, 2)$ *densities in Figure 6.9.*

Thanks