Example. Estimating the speed of light

Simon Newcomb set up an experiment in 1882 to measure the speed of light. Newcomb measured the amount of time required for light to travel a distance of 7442 meters. A histogram of Newcomb's 66 measurements is shown in Figure 3.1. There are two unusually low measurements and then a cluster of measurements that are approximately symmetrically distributed. We (inappropriately) apply the normal model, assuming that all 66 measurements are independent draws from a normal distribution with mean μ and variance σ^2. The main substantive goal is posterior inference for μ. The outlying measurements do not fit the normal model; we discuss Bayesian methods for measuring the lack of fit for these data in Section 6.3. The mean of the 66 measurements is $\overline{y} = 26.2$, and the sample standard deviation is $s = 10.8$. Assuming the noninformative prior distribution $p(\mu, \sigma^2) \propto (\sigma^2)^{-1}$, a 95% central posterior interval for μ is obtained from the t_{65} marginal posterior distribution of μ as $[\overline{y} \pm 1.997 s/\sqrt{66}] = [23.6, 28.8]$.

The posterior interval can also be obtained by simulation. Following the factorization of the posterior distribution given by (3.5) and (3.3), we first draw a random value of $\sigma^2 \sim \text{Inv-}\chi^2(65, s^2)$ as $65s^2$ divided by a random draw from the χ^2_{65} distribution (see Appendix A). Then given this value of σ^2, we draw μ from its conditional posterior distribution, $\text{N}(26.2, \sigma^2/66)$. Based on 1000 simulated values of (μ, σ^2), we estimate the posterior median of μ to be 26.2 and a 95% central posterior interval for μ to be $[23.6, 28.9]$, quite close to the analytically calculated interval.

Incidentally, based on the currently accepted value of the speed of light, the 'true value' for μ in Newcomb's experiment is 33.0, which falls outside our 95% interval. This reinforces the fact that posterior inferences are only as good as the model and the experiment that produced the data.

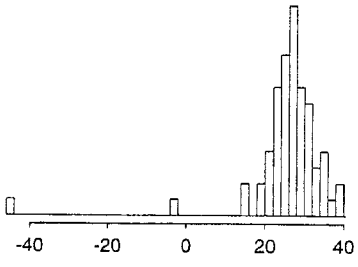

Figure 3.1 *Histogram of Simon Newcomb's measurements for estimating the speed of light, from Stigler (1977). The data are recorded as deviations from* 24,800 *nanoseconds.*

CHAPTER 6

Model checking and improvement

6.1 The place of model checking in applied Bayesian statistics

Once we have accomplished the first two steps of a Bayesian analysis—constructing a probability model and computing (typically using simulation) the posterior distribution of all estimands—we should not ignore the relatively easy step of assessing the fit of the model to the data and to our substantive knowledge. It is difficult to include in a probability distribution all of one's knowledge about a problem, and so it is wise to investigate what aspects of reality are *not* captured by the model.

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 to the data and the plausibility of the model for the purposes for which the model will be used. This is sometimes discussed as a problem of sensitivity to the prior distribution, but in practice the likelihood model is typically just as suspect; throughout, we use 'model' to encompass the sampling distribution, the prior distribution, any hierarchical structure, and issues such as which explanatory variables have been included in a regression.

Sensitivity analysis and model improvement

It is typically the case that more than one reasonable probability model can provide an adequate fit to the data in a scientific problem. The basic question of a *sensitivity analysis* is: how much do posterior inferences change when other reasonable probability models are used in place of the present model? Other reasonable models may differ substantially from the present model in the prior specification, the sampling distribution, or in what information is included (for example, predictor variables in a regression). It is possible that the present model provides an adequate fit to the data, but that posterior inferences differ under plausible alternative models.

In theory, both model checking and sensitivity analysis can be incorporated into the usual prior-to-posterior analysis. Under this perspective, model checking is done by setting up a comprehensive joint distribution, such that any data that might be observed are plausible outcomes under the joint distribution. That is, this joint distribution is a mixture of all possible 'true' models or realities, incorporating all known substantive information. The prior dis-

tribution in such a case incorporates prior beliefs about the likelihood of the competing realities and about the parameters of the constituent models. The posterior distribution of such an *exhaustive* probability model automatically incorporates all 'sensitivity analysis' but is still predicated on the truth of some member of the larger class of models.

In practice, however, setting up such a super-model to include all possibilities and all substantive knowledge is both conceptually impossible and computationally infeasible in all but the simplest problems. It is thus necessary for us to examine our models in other ways to see how they fail to fit reality and how sensitive the resulting posterior distributions are to arbitrary specifications.

Judging model flaws by their practical implications

We do not like to ask, 'Is our model true or false?', since probability models in most data analyses will not be perfectly true. Even the coin tosses and die rolls ubiquitous in probability theory texts are not truly exchangeable in reality. The more relevant question is, 'Do the model's deficiencies have a noticeable effect on the substantive inferences?'

In the examples of Chapter 5, the beta population distribution for the tumor rates and the normal distribution for the eight school effects are both chosen partly for convenience. In these examples, making convenient distributional assumptions turns out not to matter, in terms of the impact on the inferences of most interest. How to judge when assumptions of convenience can be made safely is a central task of Bayesian sensitivity analysis. Failures in the model lead to practical problems by creating clearly false inferences about estimands of interest.

6.3 Is the model consistent with data? Posterior predictive checking

159

If the model fits, then replicated data generated under the model should look similar to observed data. To put it another way, the observed data should look plausible under the posterior predictive distribution. This is really a self-consistency check: an observed discrepancy can be due to model misfit or chance.

Our basic technique for checking the fit of a model to data is to draw simulated values from the posterior predictive distribution of replicated data and compare these samples to the observed data. Any systematic differences between the simulations and the data indicate potential failings of the model. We introduce posterior predictive checking with a simple example of an obviously poorly fitting model, and then in the rest of this section we lay out

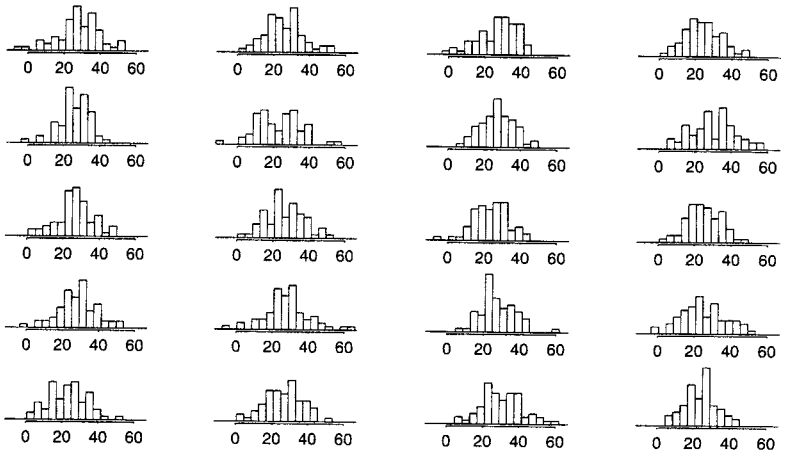

Figure 6.2 *Twenty replications, y^{rep}, of the speed of light data from the posterior predictive distribution, $p(y^{\text{rep}}|y)$; compare to observed data, y, in Figure 3.1. Each histogram displays the result of drawing 66 independent values $\tilde{y}_i$ from a common normal distribution with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

the key choices involved in posterior predictive checking. Sections 6.4 and 6.5 discuss graphical and numerical predictive checks in more detail.

Example. Comparing Newcomb's speed of light measurements to the posterior predictive distribution

Simon Newcomb's 66 measurements on the speed of light are presented in Section 3.2. In the absence of other information, in Section 3.2 we modeled the measurements as $\text{N}(\mu, \sigma^2)$, with a noninformative uniform prior distribution on $(\mu, \log\sigma)$. However, the lowest of Newcomb's measurements look like outliers compared to the rest of the data.

Could the extreme measurements have reasonably come from a normal distribution? We address this question by comparing the observed data to what we expect to be observed under our posterior distribution. Figure 6.2 displays twenty histograms, each of which represents a single draw from the posterior predictive distribution of the values in Newcomb's experiment, obtained by 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 quite different from the histogram of actual data in Figure 3.1 on page 78. One way to measure the discrepancy is to compare the smallest value in each hypothetical replicated dataset to Newcomb's smallest observation, -44. The histogram in Figure 6.3 shows the smallest observation in each of the 20 hypothetical replications; all are much larger than Newcomb's smallest observation, which is indicated by a vertical line on the graph. The normal model clearly does not capture the variation that Newcomb observed. A revised model might use an asymmetric contaminated normal distribution or a symmetric long-tailed distribution in place of the normal measurement model.

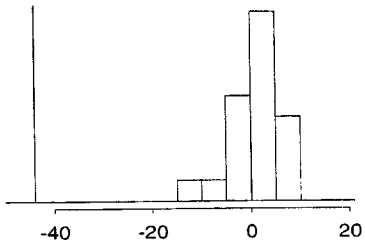

Figure 6.3 *Smallest observation of Newcomb's speed of light data (the vertical line at the left of the graph), compared to the smallest observations from each of the 20 posterior predictive simulated datasets displayed in Figure 6.2.*

Many other examples of posterior predictive checks appear throughout the book, including the educational testing example in Section 6.8, linear regressions example in Sections 14.3 and 15.2, and a hierarchical mixture model in Section 18.4.

For many problems, it is useful to examine graphical comparisons of summaries of the data to summaries from posterior predictive simulations, as in Figure 6.3. In cases with less blatant discrepancies than the outliers in the speed of light data, it is often also useful to measure the 'statistical significance' of the lack of fit, a notion we formalize here.

Notation for replications

Let y be the observed data and θ be the vector of parameters (including all the hyperparameters if the model is hierarchical). To avoid confusion with the observed data, y, we define y^{rep} as the *replicated* data that *could have been* observed, or, to think predictively, as the data we *would* see tomorrow if the experiment that produced y today were replicated with the same model and the same value of θ that produced the observed data.

We distinguish between y^{rep} and $\tilde{y}$, our general notation for predictive outcomes: $\tilde{y}$ is any future observable value or vector of observable quantities, whereas y^{rep} is specifically a replication just like 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, $\tilde{x}$.

We will work with the distribution of y^{rep} given the current state of knowledge, that is, with the posterior predictive distribution

$$p(y^{\text{rep}}|y) = \int p(y^{\text{rep}}|\theta)p(\theta|y)d\theta. \tag{6.1}$$

Test quantities

We measure the discrepancy between model and data by defining *test quantities*, the aspects of the data we wish to check. A test quantity, or *discrepancy*

measure, $T(y, \theta)$, is a scalar summary of parameters and data that is us as a standard when comparing data to predictive simulations. Test quantiti play the role in Bayesian model checking that test statistics play in classic testing. We use the notation $T(y)$ for a *test statistic*, which is a test quanti that depends only on data; in the Bayesian context, we can generalize te statistics to allow dependence on the model parameters under their posteri distribution. This can be useful in directly summarizing discrepancies betwee model and data. We discuss options for graphical test quantities in Sectic 6.4 and numerical summaries in Section 6.5.

Tail-area probabilities

Lack of fit of the data with respect to the posterior predictive distributio can be measured by the tail-area probability, or p-value, of the test quantity and computed using posterior simulations of (θ, y^{rep}). We define the p-valu mathematically, first for the familiar classical test and then in the Bayesia context.

Classical p-values. The classical p-value for the test statistic $T(y)$ is

$$p_C = \Pr(T(y^{\text{rep}}) \geq T(y)|\theta) \tag{6.2}$$

where the probability is taken over the distribution of y^{rep} with θ fixed. (The distribution of y^{rep} given y and θ is the same as its distribution given θ alone.) Test statistics are classically derived in a variety of ways but generally represent a summary measure of discrepancy between the observed data and what would be expected under a model with a particular value of θ. This value may be a 'null' value, corresponding to a 'null hypothesis,' or a point estimate such as the maximum likelihood value. A point estimate for θ must be substituted to compute a p-value in classical statistics.

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 because the test quantity is evaluated over draws from the posterior distribution of the unknown parameters. The Bayesian p-value is defined as the probability that the replicated data could be more extreme than the observed data, as measured by the test quantity:

$$p_B = \Pr(T(y^{\text{rep}}, \theta) \geq T(y, \theta)|y),$$

where the probability is taken over the posterior distribution of θ and the posterior predictive distribution of y^{rep} (that is, the joint distribution, $p(\theta, y^{\text{rep}}|y)$):

$$p_B = \int\int I_{T(y^{\text{rep}},\theta) \geq T(y,\theta)} p(y^{\text{rep}}|\theta) p(\theta|y) dy^{\text{rep}}\, d\theta,$$

where I is the indicator function. In this formula, we have used the property of the predictive distribution that $p(y^{\text{rep}}|\theta, y) = p(y^{\text{rep}}|\theta)$.

In practice, we usually compute the posterior predictive distribution using

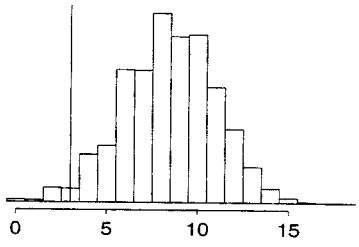

Figure 6.4 *Observed number of switches (vertical line at $T(y) = 3$), compared to 10,000 simulations from the posterior predictive distribution of the number of switches, $T(y^{\text{rep}})$.*

simulation. If we already have L simulations from the posterior density of θ, we just draw one y^{rep} from the predictive distribution for each simulated θ; we now have L draws from the joint posterior distribution, $p(y^{\text{rep}}, \theta|y)$. The posterior predictive check is the comparison between the realized test quantities, $T(y, \theta^l)$, and the predictive test quantities, $T(y^{\text{rep}\,l}, \theta^l)$. The estimated p-value is just the proportion of these L simulations for which the test quantity equals or exceeds its realized value; that is, for which $T(y^{\text{rep}\,l}, \theta^l) \geq T(y, \theta^l)$, $l = 1, \ldots, L$.

In contrast to the classical approach, Bayesian model checking does not require special methods to handle 'nuisance parameters'; by using posterior simulations, we implicitly average over all the parameters in the model.

Example. Checking the assumption of independence in binomial trials We illustrate posterior predictive model checking with a simple hypothetical example. Consider a sequence of binary outcomes, $y_1, \ldots, y_n$, modeled as a specified number of iid Bernoulli trials with a uniform prior distribution on the probability of success, θ. As discussed in Chapter 2, the posterior density under the model is $p(\theta|y) \propto \theta^s(1-\theta)^{n-s}$, which depends on the data only through the sufficient statistic, $s = \sum y_i$. Now suppose the observed data are, in order, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0. The observed autocorrelation is evidence that the model is flawed. To quantify the evidence, we can perform a posterior predictive test using the test quantity T = number of switches between 0 and 1 in the sequence. The observed value is $T(y) = 3$, and we can determine the posterior predictive distribution of $T(y^{\text{rep}})$ by simulation. To simulate y^{rep} under the model, we first draw θ from its Beta(8, 14) posterior distribution, then draw $y^{\text{rep}} = (y_1^{\text{rep}}, \ldots, y_{20}^{\text{rep}})$ as independent Bernoulli variables with probability θ. Figure 6.4 displays a histogram of the values of $T(y^{\text{rep}\,l})$ for simulation draws $l = 1, \ldots, 10000$, with the observed value, $T(y) = 3$, shown by a vertical line. The observed number of switches is about one-third as many as would be expected from the model under the posterior predictive distribution, and the discrepancy cannot easily be explained by chance, as indicated by the computed p-value of $\frac{9838}{10000}$. To convert to a p-value near zero, we can change the sign of the test statistic, which amounts to computing $\Pr(T(y^{\text{rep}}, \theta) \leq T(y, \theta)|y)$, which is 0.028 in this

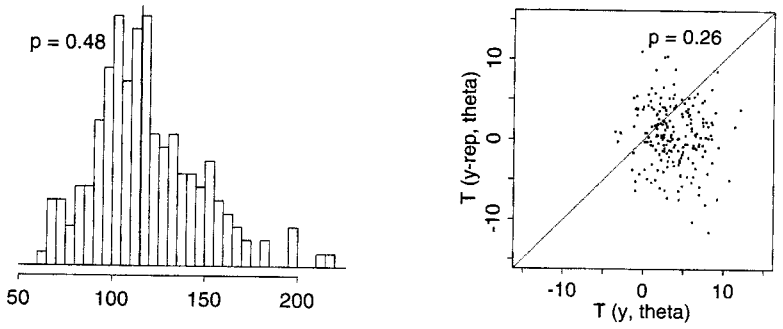

Figure 6.5 *Realized vs. posterior predictive distributions for two more test quantitie in the speed of light example: (a) Sample variance (vertical line at 115.5), compare to 200 simulations from the posterior predictive distribution of the sample variance (b) Scatterplot showing prior and posterior simulations of a test quantity:* $T(y,\theta) = |y_{(61)} - \theta| - |y_{(6)} - \theta|$ *(horizontal axis) vs.* $T(y^{\text{rep}},\theta) = |y^{\text{rep}}_{(61)} - \theta| - |y^{\text{rep}}_{(6)} - \theta|$ *(vertica 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 plot.*

case. The p-values measured from the two ends have a sum that is greater thai 1 because of the discreteness of the distribution of $T(y^{\text{rep}})$.

Example. Speed of light (continued)
In Figure 6.3, we demonstrated the poor fit of the normal model to the speec of light data using $\min(y_i)$ as the test statistic. We continue this example usin other test quantities to illustrate how the fit of a model depends on the aspect of the data and parameters being monitored. Figure 6.5a shows the observec sample variance and the distribution of 200 simulated variances from the posterio predictive distribution. The sample variance does not make a good test statistic because it is a sufficient statistic of the model and thus, in the absence of ar informative prior distribution, the posterior distribution will automatically b centered near the observed value. We are not at all surprised to find an estimatec p-value close to $\frac{1}{2}$.

The model check based on $\min(y_i)$ earlier in the chapter suggests that the norma model is inadequate. To illustrate that a model can be inadequate for some purposes but adequate for others, we assess whether the model is adequate except fo the extreme tails by considering a model check based on a test quantity sensitive to asymmetry in the center of the distribution,

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

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. The scatterplot in Figure 6.5b shows the test quantity for the observed data and the test quantity evaluated for the simulated data for 200 simulations from the posterior distribution of (θ, σ^2). The estimated p-value is 0.26, implying that any observed asymmetry in the middle of the distribution can easily be explained by sampling variation.

Defining replications

Depending on the aspect of the model one wishes to check, one can define the reference set of replications y^{rep} by conditioning on some or all of the observed data. For example, in checking the normal model for Newcomb's speed of light data, we kept the number of observations, n, fixed at the value in Newcomb's experiment. In Section 6.8, we check the hierarchical normal model for the SAT coaching experiments using posterior predictive simulations of new data on the same eight schools. It would also be possible to examine predictive simulations on new schools drawn from the same population. In analyses of sample surveys and designed experiments, it often makes sense to consider hypothetical replications of the experiment with a new randomization of selection or treatment assignment, by analogy to classical randomization tests.

6.4 Graphical posterior predictive checks

The basic idea of graphical model checking is to display the data alongside simulated data from the fitted model, and to look for systematic discrepancies between real and simulated data. This section gives examples of three kinds of graphical display:

- Direct display of all the data (as in the comparison of the speed-of-light data in Figure 3.1 to the 20 replications in Figure 6.2).
- 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 6.6 shows another example of model checking by displaying all the data. The left column of the figure displays a three-way array of 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 three-way array is displayed as 6 slices, one for each person. Before displaying, the reactions, situations, and persons have been ordered in increasing average response. We can thus think of the test statistic $T(y)$ as being this graphical display, complete with the ordering applied to the data y.

The right columns of Figure 6.6 display seven independently-simulated replications y^{rep} from a fitted logistic regression model (with the rows, columns, and persons for each dataset arranged in increasing order before display, so that we are displaying $T(y^{\text{rep}})$ in each case). Here, the replicated datasets look fuzzy and 'random' compared to the observed data, which have strong rectilinear structures that are clearly not captured in the model. If the data

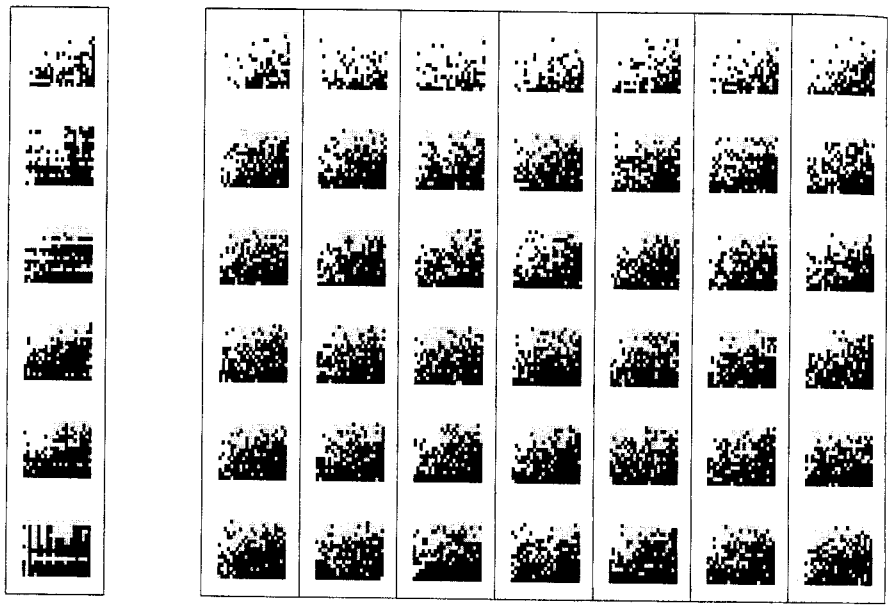

Figure 6.6 *Left column displays observed data y (a* 15×23 *array of binary responses from each of 6 persons); right columns display seven replicated datasets* y^{rep} *from a fitted logistic regression model. A misfit of model to data is apparent: the data show strong row and column patterns for individual 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.)*

were actually generated from the model, the observed data on the left would fit right in with the simulated datasets on the right.

Interestingly, these data have enough internal replication that the model misfit would be clear in comparison to a single simulated dataset from the model. But, to be safe, it is good to compare to several replications to see if the patterns in the observed data could be expected to occur by chance under the model.

Displaying data is not simply a matter of dumping a set of numbers on a page (or a screen). For example, we took care to align the graphs in Figure 6.6 to display the three-dimensional dataset and seven replications at once without confusion. Even more important, the arrangement of the rows, columns, and persons in increasing order is crucial to seeing the patterns in the data over and above the model. To see this, consider Figure 6.7, which presents the same information as in Figure 6.6 but without the ordering. Here, the discrepancies between data and model are not clear at all.

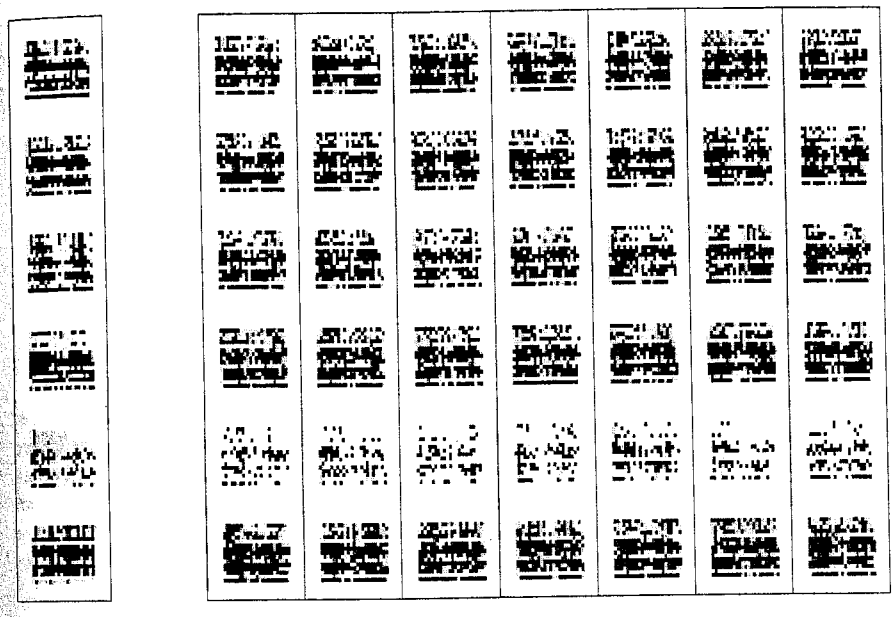

Figure 6.7 *Redisplay of Figure 6.6* without *ordering the rows, columns, and persons in order of increasing response. Once again, the left column shows the observed data and the right columns show replicated datasets from the model. Without the ordering, it is very difficult to notice the discrepancies between data and model, which are easily apparent in Figure 6.6.*