generates one vector of posterior simulations of all the uncertain quantities in the model.

9.3 Model evaluation

Despite our best efforts to include information, all models are approximate. Hence, checking the fit of a model to data and prior assumptions is always important. For the purpose of model evaluation, we can think of the inferential step of Bayesian data analysis as a sophisticated way to explore all the implications of a proposed model, in such a way that these implications can be compared with observed data and other knowledge not included in the model. For example, Section 6.4 illustrates graphical predictive checks for models fitted to data for two different problems in psychological research. In each case, the fitted model captures a general pattern of the data but misses some key features. In the second example, finding the model failure leads to a model improvement—a mixture distribution for the patient and symptom parameters—that better fits the data, as seen in Figure 6.9.

Posterior inferences can often be summarized graphically. For simple problems or one or two-dimensional summaries, we can plot a histogram or scatterplot of posterior simulations, as in Figures 3.3, 3.4, and 5.8. For larger problems, summary graphs such as Figures 5.4–5.7 are useful. Plots of several independently derived inferences are useful in summarizing results so far and suggesting future model improvements. We illustrate in Figure 14.2 with a series of regression estimates of the advantage of incumbency in Congressional elections.

Graphs (or even maps, as in Figure 6.1) are also useful for model checking and sensitivity analysis, as we have shown in Chapter 6 and further illustrate in many of the examples in Sections IV and V.

When checking a model, one must keep in mind the purposes for which it will be used. For example, the normal model for football scores in Section 1.6 accurately predicts the probability of a win, but gives poor predictions for the probability that a game is exactly tied (see Figure 1.1).

It is also important to understand the limitations of automatic Bayesian inference. Even a model that fits observed data well can yield poor inferences about some quantities of interest. It is surprising and instructive to see the pitfalls that can arise when models are automatically applied and not subjected to model checks.

Example. Estimating a population total under simple random sampling using transformed normal models

We consider the problem of estimating the total population of the $N = 804$ municipalities in New York State in 1960 from a simple random sample of $n = 100$—an artificial example, but one that illustrates the role of model checking in avoiding seriously wrong inferences. Table 9.2 presents summary statistics for the population of this 'survey' along with two simple random samples (which were the first and only ones chosen). With knowledge of the population, neither sample appears particularly atypical; sample 1 is very representative of the population

	Population ($N = 804$)	Sample 1 ($n = 100$)	Sample 2 ($n = 100$)
total	13,776,663	1,966,745	3,850,502
mean	17,135	19,667	38,505
sd	139,147	142,218	228,625
lowest	19	164	162
5%	336	308	315
25%	800	891	863
median	1,668	2,081	1,740
75%	5,050	6,049	5,239
95%	30,295	25,130	41,718
highest	2,627,319	1,424,815	1,809,578

Table 9.2 *Summary statistics for populations of municipalities in New York State in 1960 (New York City was represented by its five boroughs); all 804 municipalities and two independent simple random samples of 100. From Rubin (1983a).*

according to the summary statistics provided, whereas sample 2 has a few too many large values. Consequently, it might at first glance seem straightforward to estimate the population total, perhaps overestimating the total from the second sample.

Sample 1: initial analysis. We begin the data analysis by trying to estimate the population total from sample 1 assuming that the N values in the population were drawn from a $\mathrm{N}(\mu, \sigma^2)$ superpopulation, with a uniform prior density on $(\mu, \log \sigma)$. In the notation of Chapter 7, we wish to estimate the finite-population quantity,

$$y_{\text{total}} = N\overline{y} = n\overline{y}_{\text{obs}} + (N-n)\overline{y}_{\text{mis}}. \tag{9.1}$$

As discussed in Section 7.4, under this model, the posterior distribution of $\overline{y}$ is $t_{n-1}(\overline{y}_{\text{obs}}, (\frac{1}{n} - \frac{1}{N})s^2_{\text{obs}})$. Using the data from the second column of Table 9.2 and the tabulated Student-t distribution, we obtain the following 95% posterior interval for y_{total}: $[-5.4 \times 10^6, 37.0 \times 10^6]$. The practical person examining this 95% interval might find the upper limit useful and simply replace the lower limit by the total in the sample, since the total in the population can be no less. This procedure gives a 95% interval estimate of $[2.0 \times 10^6, 37.0 \times 10^6]$.

Surely, modestly intelligent use of statistical models should produce a better answer because, as we can see in Table 9.2, both the population and sample 1 are very far from normal, and the standard interval is most appropriate with normal populations. Moreover, all values in the population are known to be positive. Even before seeing any data, we know that sizes of municipalities are far more likely to have a normal distribution on the logarithmic than on the untransformed scale.

We repeat the above analysis under the assumption that the $N = 804$ values in the complete data follow a lognormal distribution: $\log y_i \sim \mathrm{N}(\mu, \sigma^2)$, with a uniform prior distribution on $(\mu, \log \sigma)$. Posterior inference for y_{total} is performed in the usual manner: drawing (μ, σ) from their posterior (normal-inverse-χ^2) distribution, then drawing $y_{\text{mis}}|\mu, \sigma$ from the predictive distribution, and finally calculating y_{total} from (9.1). Based on 100 simulation draws, the 95% interval for y_{total} is $[5.4 \times 10^6, 9.9 \times 10^6]$. This interval is narrower than the original interval and at first glance looks like an improvement.

Sample 1: checking the lognormal model. One of our major principles is to apply posterior predictive checks to models before accepting them. Because we are interested in a population total, y_{total}, we apply a posterior predictive check using, as a test statistic, the total in the sample, $T(y_{\text{obs}}) = \sum_{i=1}^{n} y_{\text{obs}\,i}$. Using our $L = 100$ sample draws of (μ, σ^2) from the posterior distribution under the lognormal model, we obtain posterior predictive simulations of L independent replicated datasets, $y_{\text{obs}}^{\text{rep}}$, and compute $T(y_{\text{obs}}^{\text{rep}}) = \sum_{i=1}^{n} y_{\text{obs}\,i}^{\text{rep}}$ for each. The result is that, for this predictive quantity, the lognormal model is *unacceptable*: all of the $L = 100$ simulated values are lower than the actual total in the sample, 1,966,745.

Sample 1: extended analysis. A natural generalization beyond the lognormal model for municipality sizes is the power-transformed normal family, which adds an additional parameter, ϕ, to the model; see (6.14) on page 195 for details. The values $\phi = 1$ and 0 correspond to the untransformed normal and lognormal models, respectively, and other values correspond to other transformations.

To fit a transformed normal family to data y_{obs}, the easiest computational approach is to fit the normal model to transformed data at several values of ϕ and then compute the marginal posterior density of ϕ. Using the data from sample 1, the marginal posterior density of ϕ is strongly peaked around the value $-1/8$ (assuming a uniform prior distribution for ϕ, which is reasonable given the relatively informative likelihood). Based on 100 simulated values under the extended model, the 95% interval for y_{total} is $[5.8 \times 10^6, 31.8 \times 10^6]$. With respect to the posterior predictive check, 15 out of 100 simulated replications of the sample total are larger than the actual sample total; the model fits adequately in this sense.

Perhaps we have learned how to apply Bayesian methods successfully to estimate a population total with this sort of data: use a power-transformed family and summarize inference by simulation draws. But we did not conduct a very rigorous test of this conjecture. We started with the log transformation and obtained an inference that initially looked respectable, but we saw that the posterior predictive check distribution indicated a lack of fit in the model with respect to predicting the sample total. We then enlarged the family of transformations and performed inference under the larger model (or, equivalently in this case, found the best-fitting transformation, since the transformation power was so precisely estimated by the data). The extended procedure seemed to work in the sense that the resultant 95% interval was plausible; moreover, the posterior predictive check on the sample total was acceptable. To check on this extended procedure, we try it on the second random sample of 100.

Sample 2. The standard normal-based inference for the population total from the second sample yields a 95% interval of $[-3.4 \times 10^6, 65.3 \times 10^6]$. Substituting the sample total for the lower limit gives the wide interval of $[3.9 \times 10^6, 65.3 \times 10^6]$.

Following the steps used on sample 1, modeling the sample 2 data as lognormal leads to a 95% interval for y_{total} of $[8.2 \times 10^6, 19.6 \times 10^6]$. The lognormal inference is quite tight. However, in the posterior predictive check for sample 2 with the lognormal model, none of 100 simulations of the sample total was as large as the observed sample total, and so once again we find this model unsuited for estimation of the population total.

Based upon our experience with sample 1, and the posterior predictive checks under the lognormal models for both samples, we should not trust the lognormal interval and instead should consider the general power family, which includes the lognormal as a special case. For sample 2, the marginal posterior distribution for the power parameter ϕ is strongly peaked at $-1/4$. The posterior predictive check generated 48 of 100 sample totals larger than the observed total—no indication of any problems, at least if we do not examine the specific values being generated.

In this example we have the luxury of knowing the correct example (corresponding to having complete data rather than a sample of 100 municipalities). Unfortunately, the inference for the population total under the power family turns out to be, from a substantive standpoint, atrocious: for example, the median of the 100 generated values of y_{total} is 57×10^7, the 97th value is 14×10^{15}, and the largest value generated is 12×10^{17}.

Need to specify crucial prior information. What is going on? How can the inferences for the population total in sample 2 be so much less realistic with a better-fitting model (that is, assuming a normal distribution for $y_i^{-1/4}$) than with a worse-fitting model (that is, assuming a normal distribution for $\log y_i$)?

The problem with the inferences in this example is not an inability of the models to fit the data, but an inherent inability of the data to distinguish between alternative models that have very different implications for estimation of the population total, y_{total}. Estimates of y_{total} depend strongly on the upper extreme of the distribution of municipality sizes, but as we fit models like the power family, the right tail of these models (especially beyond the 99.5% quantile), is being affected dramatically by changes governed by the fit of the model to the main body of the data (between the 0.5% and 99.5% quantiles). The inference for y_{total} is actually critically dependent upon tail behavior beyond the quantile corresponding to the largest observed $y_{\text{obs}\,i}$. In order to estimate the total (or the mean), not only do we need a model that reasonably fits the observed data, but we also need a model that provides realistic extrapolations beyond the region of the data. For such extrapolations, we must rely on prior assumptions, such as specification of the largest possible size of a municipality.

More explicitly, for our two samples, the three parameters of the power family are basically enough to provide a reasonable fit to the observed data. But in order to obtain realistic inferences for the population of New York State from a simple random sample of size 100, we must constrain the distribution of large municipalities. We were warned, in fact, by the specific values of the posterior simulations for the sample total from sample 2: 10 of the 100 simulations for the replicated sample total were larger than 300 million!

The substantive knowledge that is used to criticize the power-transformed normal model can also be used to improve the model. Suppose we know that no single municipality has population greater than 5×10^6. To incorporate this information as part of the model we simply draw posterior simulations in the same way as before but truncate municipality sizes to lie below that upper bound. The resulting posterior inferences for total population size are quite reasonable. For both samples, the inferences for y_{total} under the power family are tighter than with the untruncated models and are realistic. The 95% intervals under samples 1 and 2 are $[6 \times 10^6, 20 \times 10^6]$ and $[10 \times 10^6, 34 \times 10^6]$, respectively. Incidentally, the true population total is $13.7 \cdot 10^6$ (see Table 9.2), which is included in both intervals.

Why does the untransformed normal model work reasonably well for estimating the population total? Interestingly, the inferences for y_{total} based on the simple untransformed normal model for y_i are not terrible, even without supplying an upper bound for municipality size. Why? The estimate for y_{total} under the normal model is essentially based only on the assumed normal sampling distribution for $\overline{y}_{\text{obs}}$ and the corresponding χ^2 sampling distribution for s^2_{obs}. In order to believe that these sampling distributions are approximately valid, we need the central limit theorem to apply, which we achieve by *implicitly* bounding the upper tail of the distribution for y_i enough to make approximate normality work for a sample size of 100. This is not to suggest that we recommend the untransformed normal model for clearly nonnormal data; in the example considered here, the bounded power-transformed family makes more efficient use of the data. In addition, the untransformed normal model gives extremely poor inferences for estimands such as the population median. In general, a Bayesian analysis that limits large values of y_i must do so explicitly.

Well-designed samples or robust questions obviate the need for strong prior information. Of course, extensive modeling and simulation are not needed to estimate totals routinely in practice. Good survey practitioners know that a simple random sample is not a good survey design for estimating the total in a highly skewed population. If stratification variables were available, one would prefer to oversample the large municipalities (for example, sample all five boroughs of New York City, a large proportion of cities, and a smaller proportion of towns).

Inference for the population median. It should not be overlooked, however, that the simple random samples we drew, although not ideal for estimating the population total, are quite satisfactory for answering many questions *without* imposing strong prior restrictions.

For example, consider inference for the median size of the 804 municipalities. Using the data from sample 1, the simulated 95% posterior intervals for the median municipality size under the three models: (a) lognormal, (b) power-transformed normal family, and (c) power-transformed normal family truncated at 5×10^6, are $[1800, 3000]$, $[1600, 2700]$, and $[1600, 2700]$, respectively. The comparable intervals based on sample 2 are $[1700, 3600]$, $[1300, 2400]$, and $[1200, 2400]$. In general, better models tend to give better answers, but for questions that are robust with respect to the data at hand, such as estimating the median from our simple random sample of size 100, the effect is rather weak. For such questions, prior constraints are not extremely critical and even relatively inflexible models can provide satisfactory answers. Moreover, the posterior predictive checks for the sample median looked fine—with the observed sample median near the middle of the distribution of simulated sample medians—for all these models (but not for the untransformed normal model).

What general lessons have we learned from considering this example? The first two messages are specific to the example and address accuracy of resultant inferences for covering the true population total.

1. The lognormal model may yield inaccurate inferences for the population total even when it appears to fit observed data fairly well.
2. Extending the lognormal family to a larger, and so better-fitting, model such as the power transformation family, may lead to less realistic inferences for the population total.

These two points are not criticisms of the lognormal distribution or power transformations. Rather, they provide warnings to be heeded when using a model that has not been subjected to posterior predictive checks (for test variables relevant to the estimands of interest) and reality checks. In this context, the naive statement, 'better fits to data mean better models which in turn mean better real-world answers,' is not necessarily true. Statistical answers rely on prior assumptions as well as data, and better real-world answers generally require models that incorporate more realistic prior assumptions (such as bounds on municipality sizes) as well as provide better fits to data. This comment naturally leads to a general message encompassing the first two points.

3. In general, inferences may be sensitive to features of the underlying distribution of values in the population that cannot be addressed by the observed data. Consequently, for good statistical answers, we not only need models that fit observed data, but we also need:

   (a) flexibility in these models to allow specification of realistic underlying features not adequately addressed by observed data, such as behavior in the extreme tails of the distribution, *or*

   (b) questions that are robust for the type of data collected, in the sense that all relevant underlying features of population values are adequately addressed by the observed values.

Finding models that satisfy (a) is a more general approach than finding questions that satisfy (b) because statisticians are often presented with hard questions that require answers of some sort, and do not have the luxury of posing easy (that is, robust) questions in their place. For example, for environmental reasons it may be important to estimate the total amount of pollutant being emitted by a manufacturing plant using samples of the soil from the surrounding geographical area, or, for purposes of budgeting a health-care insurance program, it may be necessary to estimate the total amount of medical expenses from a sample of patients. Such questions are inherently nonrobust in that their answers depend on the behavior in the extreme tails of the underlying distributions. Estimating more robust population characteristics, such as the median amount of pollutant in soil samples or the median medical expense for patients, does not address the essential questions in such examples.

Relevant inferential tools, whether Bayesian or non-Bayesian, cannot be free of assumptions. Robustness of Bayesian inference is a joint property of data, prior knowledge, and questions under consideration. For many problems, statisticians may be able to define the questions being studied so as to