CHAPTER 16

Metric-Predicted Variable on One or Two Groups

Contents

It's normal to want to fit in with your friends,
Behave by their means and believe all their ends.
But I'll be high tailing it, fast and askew,
Precisely 'cause I can't abide what you do.[1]

In this chapter, we consider a situation in which we have a metric-predicted variable that is observed for items from one or two groups. For example, we could measure the blood pressure (i.e., a metric variable) for people randomly sampled from first-year university students (i.e., a single group). In this case, we might be interested in how much the group's typical blood pressure differs from the recommended value for people of that age as published by a federal agency. As another example, we could measure the IQ (i.e., a metric variable) of people randomly sampled from everyone self-described as vegetarian (i.e., a single group). In this case, we could be interested in how much this group's IQ differs from the general population's average IQ of 100.

In the context of the generalized linear model (GLM) introduced in the previous chapter, this chapter's situation involves the most trivial cases of the linear core of the GLM, as indicated in the left cells of Table 15.1 (p. 434), with a link function that is the

[1] This chapter describes data with a normal distribution, which is parameterized by its mean and precision. But data can have outliers, which demand descriptive distributions with high tails or skewed ends. The poem plays with alternative meanings of the words normal, fit, mean, end, high tail, skew, precise, and believe.

Doing Bayesian Data Analysis, 2nd Edition
http://dx.doi.org/10.1016/B978-0-12-405888-0.00016-7

identity along with a normal distribution for describing noise in the data, as indicated in the first row of Table 15.2 (p. 443). We will explore options for the prior distribution on parameters of the normal distribution, and methods for Bayesian estimation of the parameters. We will also consider alternative noise distributions for describing data that have outliers.

16.1. ESTIMATING THE MEAN AND STANDARD DEVIATION OF A NORMAL DISTRIBUTION

The normal probability density function was introduced in Section 4.3.2.2, p. 83. The normal distribution specifies the probability density of a value y, given the values of two parameters, the mean μ and standard deviation σ:

$$p(y|\mu,\sigma) = \frac{1}{Z}\exp\left(-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right) \tag{16.1}$$

where Z is the normalizer, which is a constant that makes the probability density integrate to 1. It turns out that $Z = \sigma\sqrt{2\pi}$, but we won't need to use this fact in the derivations below.

To get an intuition for the normal distribution as a likelihood function, consider three data values $y_1 = 85$, $y_2 = 100$, and $y_3 = 115$, which are plotted as large dots in Figure 16.1. The probability density of any single datum, given particular parameter values, is $p(y|\mu,\sigma)$ as specified in Equation 16.1. The probability of the entire set of independent data values is the multiplicative product, $\prod_i p(y_i|\mu,\sigma) = p(D|\mu,\sigma)$, where $D = \{y_1, y_2, y_3\}$. Figure 16.1 shows $p(D|\mu,\sigma)$ for different values of μ and σ. As you can see, there are values of μ and σ that make the data most probable, but other nearby values also accommodate the data reasonably well. (For another example, see Figure 2.4, p. 23.) The question is, given the data, how should we allocate credibility to combinations of μ and σ?

The answer is provided by Bayes' rule. Given a set of data, D, we estimate the parameters with Bayes' rule:

$$p(\mu,\sigma|D) = \frac{p(D|\mu,\sigma)\,p(\mu,\sigma)}{\iint d\mu\, d\sigma\, p(D|\mu,\sigma)\,p(\mu,\sigma)} \tag{16.2}$$

Figure 16.1 shows examples of $p(D|\mu,\sigma)$ for a particular data set at different values of μ and σ. The prior, $p(\mu,\sigma)$, specifies the credibility of each combination of μ,σ values in the two-dimensional joint parameter space, without the data. Bayes' rule says that the posterior credibility of each combination of μ,σ values is the prior credibility times the likelihood, normalized by the marginal likelihood. Our goal now is to evaluate Equation 16.2 for reasonable choices of the prior distribution, $p(\mu,\sigma)$.

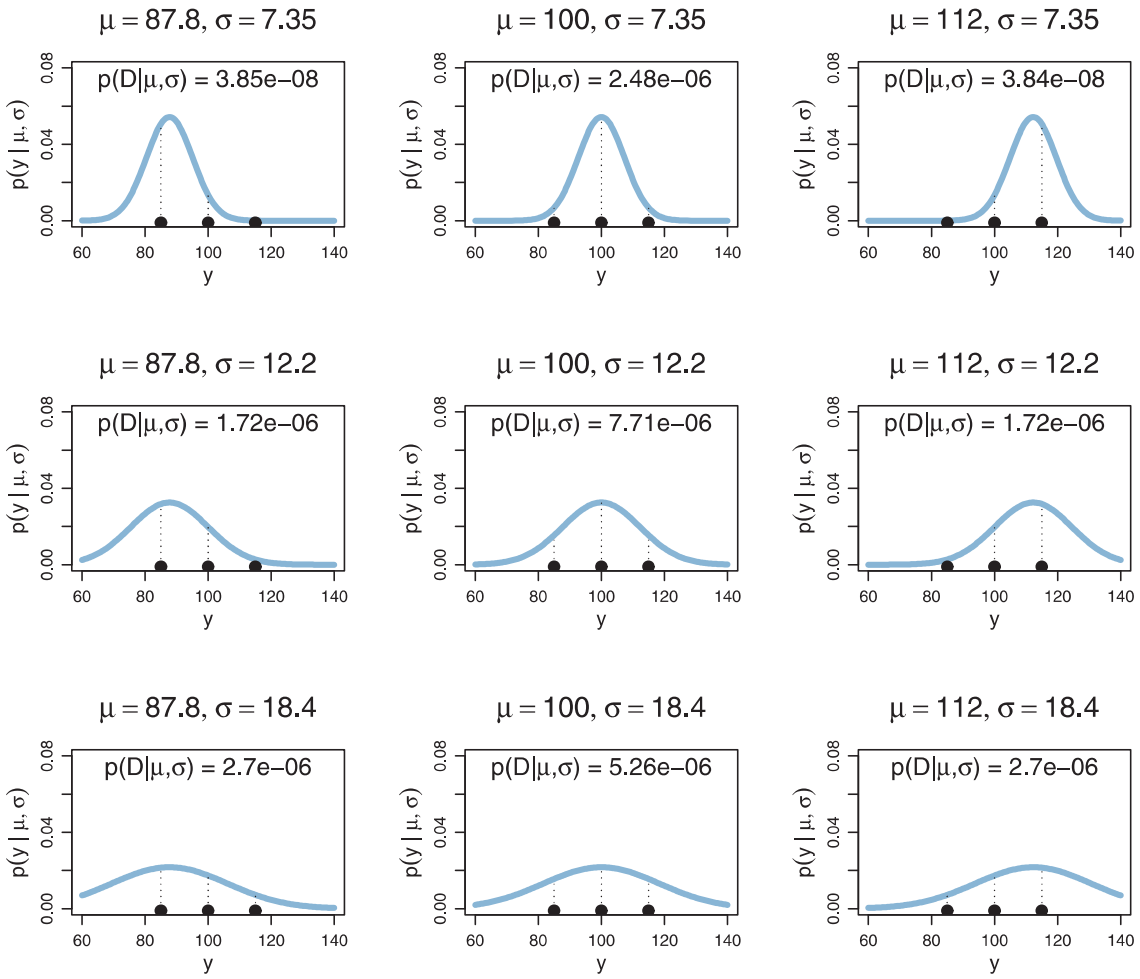

Figure 16.1 The likelihood $p(D|\mu,\sigma)$ for three data points, $D = \{85, 100, 115\}$, according to a normal likelihood function with different values of μ and σ. Columns show different values of μ, and rows show different values of σ. The probability density of an individual datum is the height of the dotted line over the point. The probability of the set of data is the product of the individual probabilities. The middle panel shows the μ and σ that maximize the probability of the data. (For another example, see Figure 2.4, p. 23.)

16.1.1. Solution by mathematical analysis

Because we are already familiar with JAGS and Stan, we could go directly to an MCMC solution. But this case is simple enough that it can be solved analytically, and the resulting formulas motivate some of the traditional parameterizations and priors for normal distributions. Therefore, we take a short algebraic tour before moving on to MCMC implementations.

It is convenient first to consider the case in which the standard deviation of the likelihood function is fixed at a specific value. In other words, the prior distribution on σ is a spike over that specific value. We'll denote that fixed value as $\sigma = S_y$. With

this simplifying assumption, we are only estimating μ because we are assuming perfectly certain prior knowledge about σ.

When σ is fixed, then the prior distribution on μ in Equation 16.2 can be easily chosen to be conjugate to the normal likelihood. (The term "conjugate prior" was defined in Section 6.2, p. 126.) It turns out that the product of normal distributions is again a normal distribution; in other words, if the prior on μ is normal, then the posterior on μ is normal. It is easy to derive this fact, as we do next.

Let the prior distribution on μ be normal with mean M_μ and standard deviation S_μ. Then the likelihood times the prior (i.e., the numerator of Bayes' rule) is

$$
\begin{aligned}
p(y|\mu,\sigma)\,p(\mu,\sigma) &= p(y|\mu,S_y)\,p(\mu) \\
&\propto \exp\left(-\frac{1}{2}\frac{(y-\mu)^2}{S_y^2}\right)\exp\left(-\frac{1}{2}\frac{(\mu-M_\mu)^2}{S_\mu^2}\right) \qquad \text{notice } \propto \text{ not } = \\
&= \exp\left(-\frac{1}{2}\left[\frac{(y-\mu)^2}{S_y^2}+\frac{(\mu-M_\mu)^2}{S_\mu^2}\right]\right) \\
&= \exp\left(-\frac{1}{2}\left[\frac{S_\mu^2\,(y-\mu)^2+S_y^2\,(\mu-M_\mu)^2}{S_y^2S_\mu^2}\right]\right) \\
&= \exp\left(-\frac{1}{2}\left[\frac{S_y^2+S_\mu^2}{S_y^2S_\mu^2}\left(\mu^2-2\frac{S_y^2M_u+S_\mu^2y}{S_y^2+S_\mu^2}\mu+\frac{S_y^2M_u^2+S_\mu^2y^2}{S_y^2+S_\mu^2}\right)\right]\right) \\
&= \exp\left(-\frac{1}{2}\left[\frac{S_y^2+S_\mu^2}{S_y^2S_\mu^2}\left(\mu^2-2\frac{S_y^2M_u+S_\mu^2y}{S_y^2+S_\mu^2}\mu\right)\right]\right) \\
&\quad\times\exp\left(-\frac{1}{2}\left[\frac{S_y^2+S_\mu^2}{S_y^2S_\mu^2}\left(+\frac{S_y^2M_u^2+S_\mu^2y^2}{S_y^2+S_\mu^2}\right)\right]\right) \\
&\propto \exp\left(-\frac{1}{2}\left[\frac{S_y^2+S_\mu^2}{S_y^2S_\mu^2}\left(\mu^2-2\frac{S_y^2M_u+S_\mu^2y}{S_y^2+S_\mu^2}\mu\right)\right]\right) \qquad (16.3)
\end{aligned}
$$

where the transition to the last line is valid because the term that was dropped was merely a constant. This result, believe it or not, is progress. Why? Because we ended up, in the innermost parentheses, with a quadratic expression in μ. Notice that the normal prior is also a quadratic expression in μ. All we have to do is "complete the square" inside the parentheses, and do the same trick that got us to the last line of Equation 16.3:

$$
\begin{aligned}
p(y|\mu,S_y)p(\mu) &\propto \exp\left(-\frac{1}{2}\left[\frac{S_y^2+S_\mu^2}{S_y^2S_\mu^2}\left(\mu^2-2\frac{S_y^2M_u+S_\mu^2y}{S_y^2+S_\mu^2}\mu+\left(\frac{S_y^2M_\mu+S_\mu^2y}{S_y^2+S_\mu^2}\right)^2\right)\right]\right) \\
&= \exp\left(-\frac{1}{2}\left[\frac{S_y^2+S_\mu^2}{S_y^2S_\mu^2}\left(\mu-\frac{S_y^2M_\mu+S_\mu^2y}{S_y^2+S_\mu^2}\right)^2\right]\right) \qquad (16.4)
\end{aligned}
$$

Equation 16.4 is the numerator of Bayes' rule. When it is normalized, it becomes a probability density function. What is the shape of the function? You can see that Equation 16.4 has the same form as a normal distribution on μ, such that the mean is $(S_\gamma^2 M_\mu + S_\mu^2 \gamma)/(S_\gamma^2 + S_\mu^2)$ and the standard deviation is $\sqrt{S_\gamma^2 S_\mu^2/(S_\gamma^2 + S_\mu^2)}$.

That formula is rather unwieldy! It becomes more compact if the normal density is re-expressed in terms of $1/\sigma^2$ instead of σ. The reciprocal of the squared standard deviation is called the *precision*. To get an intuition for precision, notice that a very narrow distribution is highly precise. When the standard deviation gets smaller, the precision gets bigger. Now, because the posterior standard deviation is $\sqrt{S_\gamma^2 S_\mu^2/(S_\gamma^2 + S_\mu^2)}$, the posterior precision is

$$\frac{S_\gamma^2 + S_\mu^2}{S_\gamma^2 S_\mu^2} = \frac{1}{S_\mu^2} + \frac{1}{S_\gamma^2} \tag{16.5}$$

Thus, the posterior precision is the sum of the prior precision and the likelihood precision.

The posterior mean can also be compactly re-expressed in terms of precisions. The posterior mean is $(S_\gamma^2 M_\mu + S_\mu^2 \gamma)/(S_\gamma^2 + S_\mu^2)$, which can be re-arranged as

$$\frac{1/S_\mu^2}{1/S_\gamma^2 + 1/S_\mu^2} M_\mu + \frac{1/S_\gamma^2}{1/S_\gamma^2 + 1/S_\mu^2} \gamma \tag{16.6}$$

In other words, the posterior mean is a weighted average of the prior mean and the datum, with the weighting corresponding to the relative precisions of the prior and the likelihood. When the prior is highly precise compared to the likelihood, that is when $1/S_\mu^2$ is large compared to $1/S_\gamma^2$, then the prior is weighed heavily and the posterior mean is near the prior mean. But when the prior is imprecise and very uncertain, then the prior does not get much weight and the posterior mean is close to the datum. We have previously seen this sort of relative weighting of prior and data in the posterior. It showed up in the case of updating a beta prior, back in Equation 6.9, p. 133.

The formulas for the mean and precision of the posterior normal can be naturally extended when there are N values of γ in a sample, instead of only a single value of γ. The formulas can be derived from the defining formulas, as was done above, but a short cut can be taken. It is known from mathematical statistics that when a set of values γ_i are generated from a normal likelihood function, the mean of those values, $\overline{\gamma}$, is also distributed normally, with the same mean as the generating mean, and with a standard deviation of $\sigma/\sqrt{N}$. Thus, instead of conceiving of this situation as N scores γ_i sampled from the likelihood, normal$(\gamma_i|\mu, \sigma)$, we conceive of this as a single score, $\overline{\gamma}$, sampled from the likelihood, normal$(\overline{\gamma}|\mu, \sigma/\sqrt{N})$. Then we just apply the updating formulas we previously derived. Thus, for N scores γ_i generated from a normal$(\gamma_i|\mu, S_\gamma)$ likelihood and a normal$(\mu|M_\mu, S_\mu)$ prior, the posterior distribution on μ is also normal with mean

$$\frac{1/S_\mu^2}{N/S_\gamma^2 + 1/S_\mu^2} M_\mu + \frac{N/S_\gamma^2}{N/S_\gamma^2 + 1/S_\mu^2} \overline{\gamma}$$

and precision

$$\frac{1}{S_\mu^2} + \frac{N}{S_\gamma^2}.$$

Notice that as the sample size N increases, the posterior mean is dominated by the data mean.

In the derivations above, we estimated the μ parameter when σ is fixed. We can instead estimate the σ parameter when μ is fixed. Again the formulas are more conveniently expressed in terms of precision. It turns out that when μ is fixed, a conjugate prior for the precision is the gamma distribution (e.g., Gelman et al., 2013, p. 43). The gamma distribution was described in Figure 9.8, p. 237. For our purposes, it is not important to state the updating formulas for the gamma distribution in this situation. But it is important to understand the meaning of a gamma prior on precision. Consider a gamma distribution that is loaded heavily over very small values, but has a long shallow tail extending over large values. This sort of gamma distribution on precision indicates that we believe most strongly in small precisions, but we admit that large precisions are possible. If this is a belief about the precision of a normal likelihood function, then this sort of gamma distribution expresses a belief that the data will be more spread out, because small precisions imply large standard deviations. If the gamma distribution is instead loaded over large values of precision, it expresses a belief that the data will be tightly clustered.

Because of its role in conjugate priors for the normal likelihood function, the gamma distribution is routinely used as a prior on precision. But there is no logical necessity to do so, and modern MCMC methods permit more flexible specification of priors. Indeed, because precision is less intuitive than standard deviation, it can be more useful instead to give the standard deviation a uniform prior that spans a wide range.

Summary. We have assumed that the data are generated by a normal likelihood function, parameterized by a mean μ and standard deviation σ, and denoted $y \sim \text{normal}(y|\mu, \sigma)$. For purposes of mathematical derivation, we made unrealistic assumptions that the prior distribution is either a spike on σ or a spike on μ, in order to make three main points:

1. A natural way to express a prior on μ is with a normal distribution, because this is conjugate with the normal likelihood when its standard deviation is fixed.
2. A way to express a prior on the precision $1/\sigma^2$ is with a gamma distribution, because this is conjugate with the normal likelihood when its mean is fixed. However in practice the standard deviation can instead be given a uniform prior (or anything else that reflects prior beliefs, of course).
3. The formulas for Bayesian updating of the parameter distribution are more conveniently expressed in terms of precision than standard deviation. Normal distributions are described sometimes in terms of standard deviation and sometimes in terms of precision, so it is important to glean from context which is being referred to. *In R and Stan, the normal distribution is parameterized by mean and standard deviation. In JAGS and BUGS, the normal distribution is parameterized by mean and precision.*

A joint prior, on the combination of μ and σ parameter values, can also be specified, in such a way that the posterior has the same form as the prior. We will not pursue these mathematical analyses here, because our purpose is merely to justify and motivate typical expressions for the prior distributions on the parameters, so that they can then be used in MCMC sampling.

Various other sources describe conjugate priors for the joint parameter space (e.g., Gelman et al., 2013).

16.1.2. Approximation by MCMC in JAGS

It is easy to estimate the mean and standard deviation in JAGS. Figure 16.2 illustrates two options for the model. The data, y_i, are assumed to be generated by a normal likelihood function with mean μ and standard deviation σ. In both models, the prior on μ is a normal distribution with mean M and standard deviation S. For our applications, we will assume a noncommittal prior that places M in the midst of typical data and sets S to an extremely large value so that the prior has minimal influence on the posterior. For the prior on the standard deviation, the left panel of Figure 16.2 re-expresses σ as precision: $\tau = 1/\sigma^2$ hence $\sigma = 1/\sqrt{\tau}$. Then a noncommittal gamma prior is placed on the precision. The conventional noncommittal gamma prior has shape and rate constants that are close to zero, such as $Sh = 0.01$ and $R = 0.01$. The right panel of Figure 16.2 instead puts a broad uniform distribution directly on σ. The low and high values of the uniform distribution are set to be far outside any realistic value for the data, so that the prior has minimal influence on the posterior. The uniform prior on σ is easier to intuit than a gamma prior on precision, but the priors are not equivalent.

How are the constants in the prior actually determined? In this application we seek broad priors relative to typical data, so that the priors have minimal influence on the posterior. One way to discover the constants is by asking an expert in the domain being studied. But in lieu of that, we will use the data themselves to tell us what the typical scale of the data is. We will set M to the mean of the data, and set S to a huge multiple (e.g., 100) of the standard deviation of the data. This way, no matter what the scale of the data is, the prior will be vague. Similarly, we will set the high value H of the uniform prior on σ to a huge multiple of the standard deviation in the data, and set the low value L to a tiny fraction of the standard deviation in the data. Again, this means that the prior is vague no matter what the scale of the data happens to be.

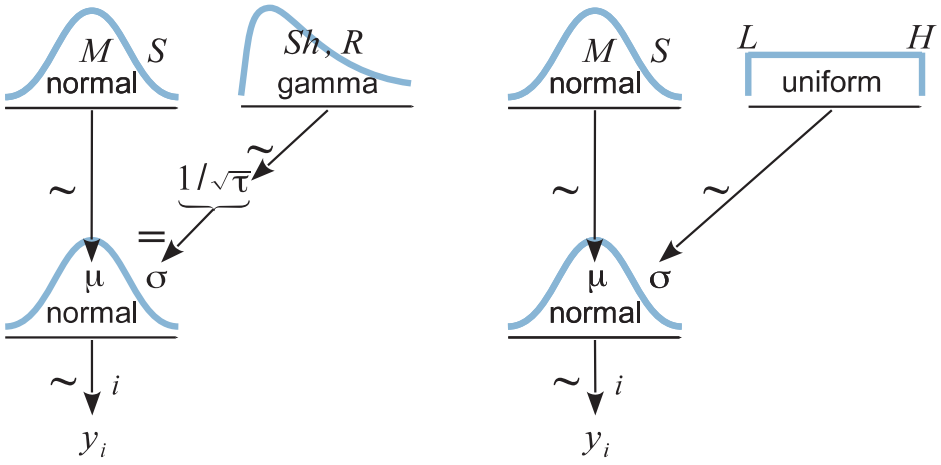

Figure 16.2 Dependencies of variables for metric data described by a normal distribution. The left panel puts a gamma prior on the precision $\tau = 1/\sigma^2$. The right panel puts a uniform prior on σ.

Before we examine the model specification for JAGS, consider the following specification of the data. The data themselves are in the vector named `y`. Then we define:

```
dataList = list(
  y = y ,
  Ntotal = length(y) ,
  meanY = mean(y) ,
  sdY = sd(y)
)
```

Notice above that the mean and standard deviation of the data are packaged into the list that gets shipped to JAGS, so JAGS can use those constants in the model specification. The model in the right side of Figure 16.2 is expressed in JAGS as follows:

```
model {
  for ( i in 1:Ntotal ) {
    y[i] ~ dnorm( mu , 1/sigma^2 )     # JAGS uses precision
  }
  mu ~ dnorm( meanY , 1/(100*sdY)^2 )  # JAGS uses precision
  sigma ~ dunif( sdY/1000 , sdY*1000 )
}
```

Notice that each arrow in Figure 16.2 has a corresponding line in the JAGS code. Notice that JAGS parameterizes `dnorm` by mean and *precision*, not by mean and standard deviation. Notice also that we can put the expression `1/sigma^2` in the argument of `dnorm`, instead of having to define precision as a separate explicit variable.

I have packaged the model and supporting functions in a program that is called from the high-level script, `Jags-Ymet-Xnom1grp-Mnormal-Example.R`. If the file name seems mysterious, please review the file name system explained back in Section 8.3, p. 206. The file name used here indicates that the predicted variable is metric (`Ymet`) and the predictor is nominal with a single group (`Xnom1grp`) and the model is a normal likelihood (`Mnormal`). The program produces graphical output that is specialized for this model.

For purposes of illustration, we use fictitious data. The data are IQ (intelligence quotient) scores from a group of people who have consumed a "smart drug." We know that IQ tests have been normed to the general population so that they have an average score of 100 and a standard deviation of 15. Therefore, we would like to know how differently the smart-drug group has performed relative to the general population average.

The aforementioned script loads the data file, runs JAGS, produces MCMC diagnostics, and graphs the posterior. Before considering the posterior, it is important to check that the chains are well behaved. The diagnostics (not shown here, but you can run the script and see for yourself) indicate converged chains with an ESS on both parameters of at least 10,000. Figure 16.3 shows aspects of the posterior distribution, along with a histogram of the data in the upper right panel.

The upper left panel of Figure 16.3 shows the posterior on the mean parameter. The estimated 95% HDI extends from 101.35 to 114.21. This HDI barely excludes a somewhat arbitrary ROPE from 99 to 101, especially if we also consider MCMC variability in the

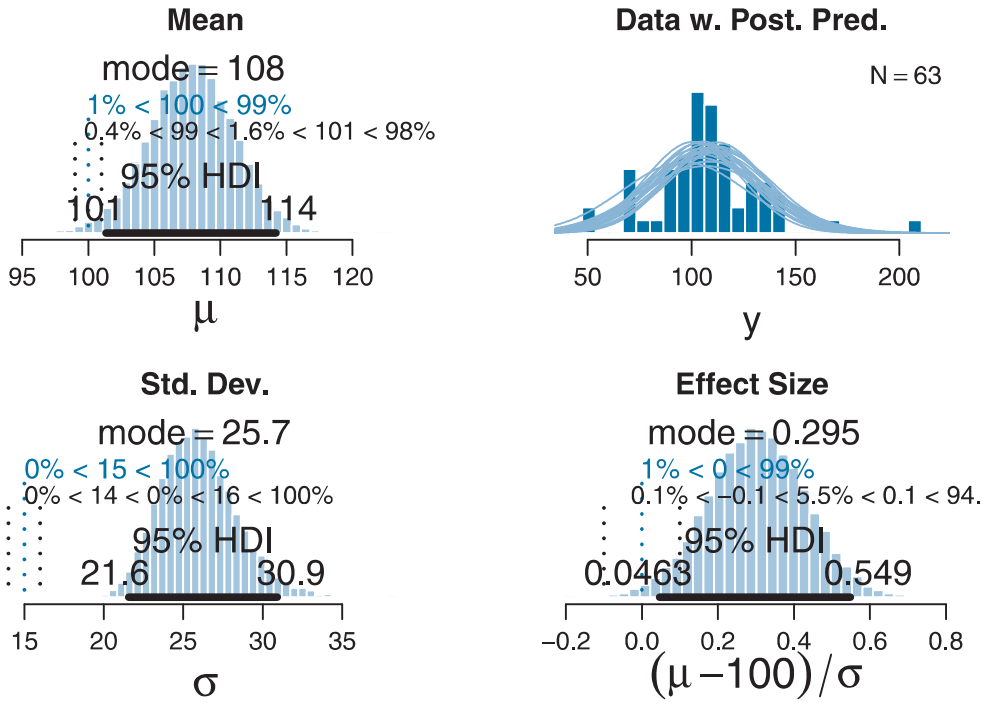

Figure 16.3 Posterior distribution of `Jags-Ymet-Xnom1grp-Mnormal-Example.R` applied to fictitious IQ data from a "smart drug" group.

HDI limit (recall Figure 7.13, p. 185). Thus it appears that the smart drug increases IQ performance somewhat, but the improvement is not large relative to the uncertainty in the estimate. Therefore, we might not want to make any strong decision, from these data, that IQ is credibly increased by the smart drug. This conservative conclusion is reinforced by considering the effect size in the lower right panel of Figure 16.3. *Effect size* is simply the amount of change induced by the treatment relative to the standard deviation: $(\mu - 100)/\sigma$. In other words, effect size is the "standardized" change. The posterior on the effect size has a 95% HDI barely excluding zero, and clearly not excluding a ROPE from −0.1 to 0.1. A conventionally "small" effect size in psychological research is 0.2 (Cohen, 1988), and the ROPE limits are set at half that size for purposes of illustration.

However, the lower left panel of Figure 16.3 suggests that the smart drug did have an effect on the standard deviation of the group. The posterior of the standard deviation is far outside any reasonable ROPE around the general-population value of 15. One interpretation of this result is that the smart drug caused some people to increase their IQ test performance but caused other people to decrease their IQ test performance. Such an effect on variance has real-world precedents; for example, stress can increase variance across people (Lazarus & Eriksen, 1952). In the next section we will model the data with a distribution that accommodates outliers, and we will have to modify this interpretation of an increased standard deviation.

Finally, the upper right panel of Figure 16.3 shows a histogram of the data superimposed with a smattering of normal curves that have credible μ and σ values from the MCMC

sample. This constitutes a form of posterior-predictive check, by which we check whether the model appears to be a reasonable description of the data. With such a small amount of data, it is difficult to visually assess whether normality is badly violated, but there appears to be a hint that the normal model is straining to accommodate some outliers: The peak of the data protrudes prominently above the normal curves, and there are gaps under the shoulders of the normal curves.

16.2. OUTLIERS AND ROBUST ESTIMATION: THE *t* DISTRIBUTION

When data appear to have outliers beyond what would be accommodated by a normal distribution, it would be useful to be able to describe the data with a more appropriate distribution that has taller or heavier tails than the normal. A well-known distribution with heavy tails is the *t* distribution. The *t* distribution was originally invented by Gosset (1908), who used the pseudonym "Student" because his employer (Guinness Brewery) prohibited publication of any research that might be proprietary or imply problems with their product (such as variability in quality). Therefore the distribution is often referred to as the *Student t* distribution.

Figure 16.4 shows examples of the *t* distribution. Like the normal distribution, it has two parameters that control its mean and its width. The standard deviation is controlled indirectly via the *t* distribution's "scale" parameter. In Figure 16.4, the mean is set to zero and the scale is set to one. The *t* distribution has a third parameter that controls the heaviness of its tails,

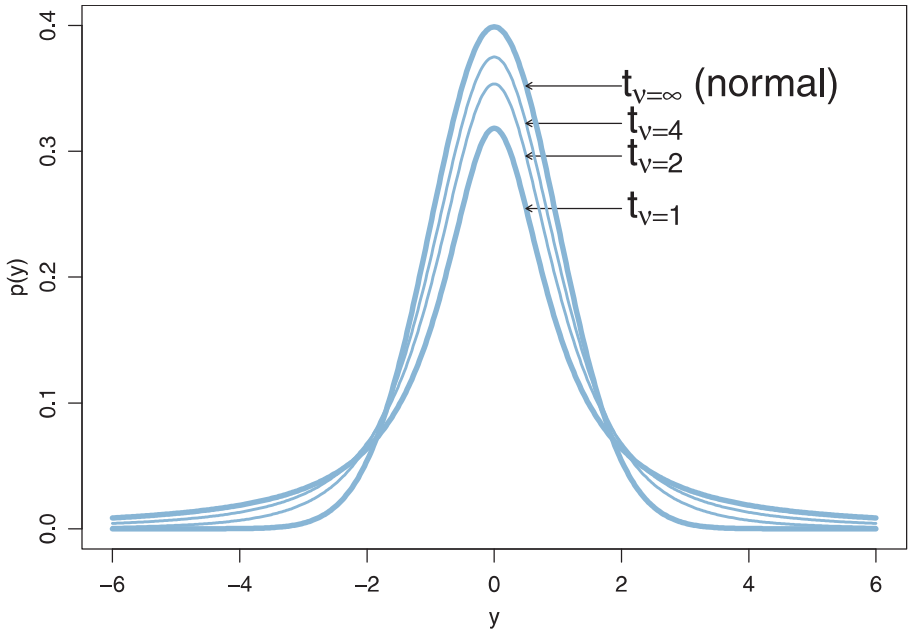

Figure 16.4 Examples of *t* distributions. In all cases, $\mu = 0$ and $\sigma = 1$. The normality parameter, ν, controls the heaviness of the tails. Curves for different values of ν are superimposed for easy comparison. The abscissa is labeled as *y* (not *x*) because the distribution is intended to describe predicted data.

which I will refer to as the "normality" parameter, ν (Greek letter nu). Many people might be familiar with this parameter as the "degrees of freedom" from its use in NHST. But because we will not be using the t distribution as a sampling distribution, and instead we will be using it only as a descriptive shape, I prefer to name the parameter by its effect on the distribution's shape. The normality parameter can range continuously from 1 to ∞. As can be seen in Figure 16.4, when $\nu = 1$ the t distribution has very heavy tails, and when ν approaches ∞ the t distribution becomes normal.

Although the t distribution is usually conceived as a sampling distribution for the NHST t test, we will use it instead as a convenient descriptive model of data with outliers (as is often done; e.g., Damgaard, 2007; M. C. Jones & Faddy, 2003; Lange, Little, & Taylor, 1989; Meyer & Yu, 2000; Tsionas, 2002; Zhang, Lai, Lu, & Tong, 2013). Outliers are simply data values that fall unusually far from a model's expected value. Real data often contain outliers relative to a normal distribution. Sometimes the anomalous values can be attributed to extraneous influences that can be explicitly identified, in which case the affected data values can be corrected or removed. But usually we have no way of knowing whether a suspected outlying value was caused by an extraneous influence, or is a genuine representation of the target being measured. Instead of deleting suspected outliers from the data according to some arbitrary criterion, we retain all the data but use a noise distribution that is less affected by outliers than is the normal distribution.

Figure 16.5 shows examples of how the t distribution is robust against outliers. The curves show the maximum likelihood estimates (MLEs) of the parameters for the t and normal distributions. More formally, for the given data $D = \{y_i\}$, parameter values were found for the normal that maximized $p(D|\mu, \sigma)$, and parameter values were found for the t distribution that maximized $p(D|\mu, \sigma, \nu)$. The curves for those MLEs are plotted with the data. The upper panel of Figure 16.5 shows "toy" data to illustrate that the normal is strongly influenced by an outlier, while the t distribution remains centered over the bulk of the data. For the t distribution, the mean μ is 0.12, which is very close to the center of the cluster of five data points. The outlying datum is accommodated by setting the normality to a very small value. The normal distribution, on the other hand, can accommodate the outlier only by using a mean that is shifted toward the outlier and a standard deviation that is bloated to span the outlier. The t distribution appears to be a better description of the data.

It is important to understand that the scale parameter σ in the t distribution is not the standard deviation of the distribution. (Recall that the standard deviation is the square root of the variance, which is the expected value of the squared deviation from the mean, as defined back in Equation 4.8, p. 86.) The standard deviation is actually larger than σ because of the heavy tails. In fact, when ν drops below 2 (but is still ≥ 1), the standard deviation of the mathematical t distribution goes to infinity. For example, in the upper panel of Figure 16.5, ν is only 1.14, which means that the standard deviation of the mathematical t distribution is infinity, even though the sample standard deviation of the data is finite. At the same time, the scale parameter of the t distribution has value $\sigma = 1.47$. While this value of the scale parameter is not the standard deviation of the distribution, it does have an intuitive relation to the spread of the data. Just as the range $\pm\sigma$ covers the middle 68% of a *normal* distribution, the range $\pm\sigma$ covers the middle 58% of a t distribution when $\nu = 2$, and the middle 50%

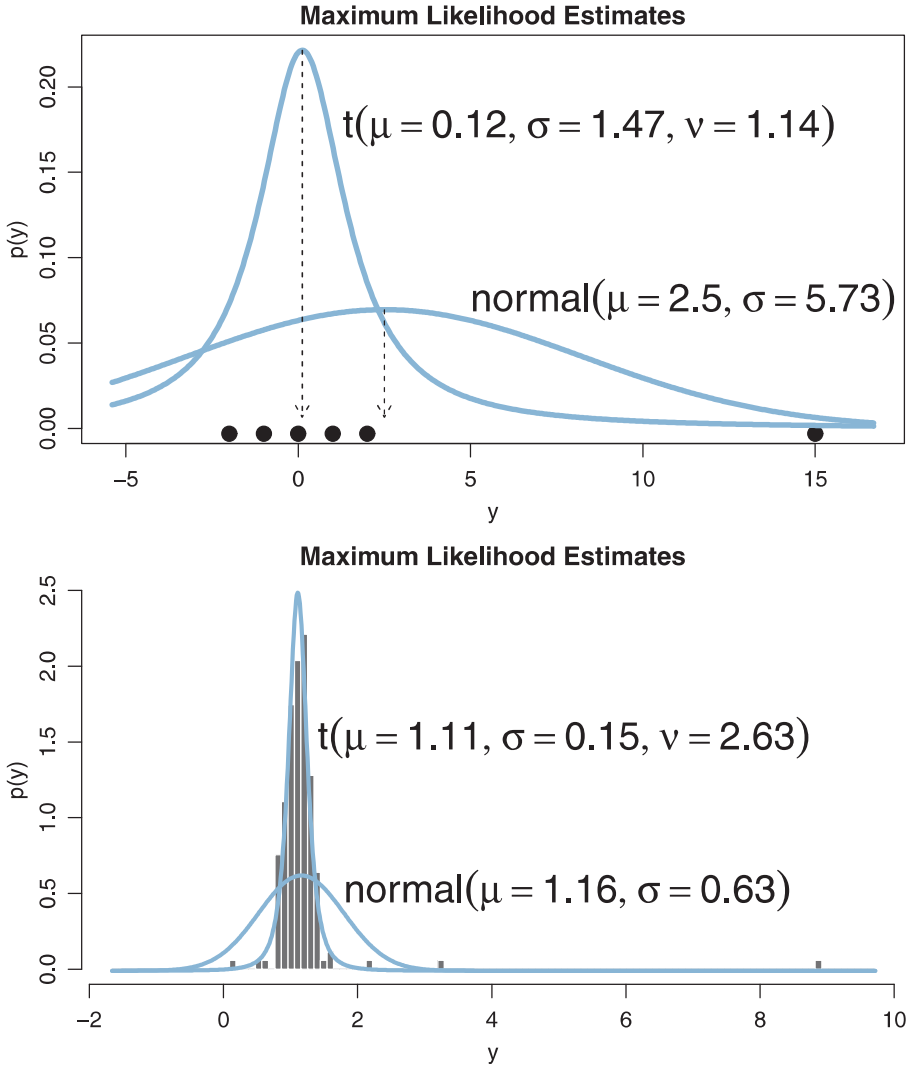

Figure 16.5 The maximum likelihood estimates of normal and *t* distributions fit to the data shown. Upper panel shows "toy" data to illustrate that the normal accommodates an outlier only by enlarging its standard deviation and, in this case, by shifting its mean. Lower panel shows actual data (Holcomb & Spalsbury, 2005) to illustrate realistic effect of outliers on estimates of the normal.

when $\nu = 1$. These areas are illustrated in the left column of Figure 16.6. The right column of Figure 16.6 shows the width under the middle of a *t* distribution that is needed to span 68.27% of the distribution, which is the area under a normal distribution for $\sigma = \pm 1$.

The lower panel of Figure 16.5 uses realistic data that indicates levels of inorganic phosphorous, measured in milligrams per deciliter, in 177 human subjects aged 65 or older. The authors of the data (Holcomb & Spalsbury, 2005) intentionally altered a few data points to reflect typical transcription errors and to illustrate methods for detecting and correcting such errors. We instead assume that we no longer have access to records of the original individual measurements, and must model the uncorrected data set. The *t* distribution accommodates the outliers and fits the distribution of data much better than the normal.

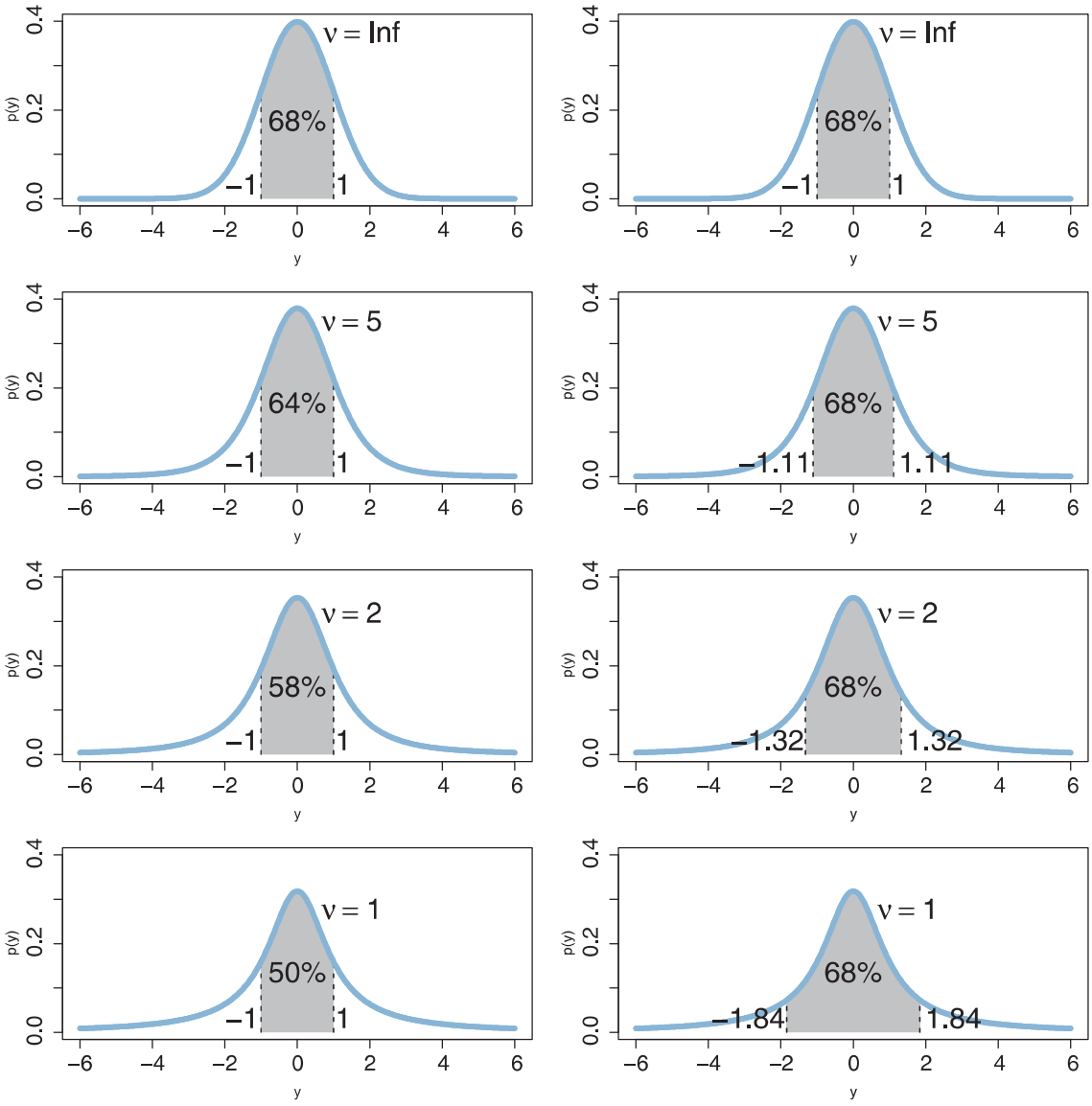

Figure 16.6 Examples of t distributions with areas under the curve. In all cases, $\mu = 0$ and $\sigma = 1$. Rows show different values of the normality parameter, ν. Left column shows area under the t distribution from $y = -1$ to $y = +1$. Right column shows values of $\pm y$ needed for an area of 68.27%, which is the area under a standardized normal curve from $y = -1$ to $y = +1$. The abscissa is labeled as y (not x) because the distribution is intended to describe predicted data.

The use of a heavy-tailed distribution is often called *robust estimation* because the estimated value of the central tendency is stable, that is, "robust," against outliers. The t distribution is useful as a likelihood function for modeling outliers at the level of observed data. But the t distribution is also useful for modeling outliers at higher levels in a hierarchical prior. We will encounter several applications.

16.2.1. Using the *t* distribution in JAGS

It is easy to incorporate the t distribution into the model specification for JAGS. The likelihood changes from `dnorm` to `dt` with the inclusion of its normality parameter, as follows:

```
model {
  for ( i in 1:Ntotal ) {
    y[i] ~ dt( mu , 1/sigma^2 , nu )    # JAGS: dt uses precision
  }
  mu ~ dnorm( meanY , 1/(100*sdY)^2 )
  sigma ~ dunif( sdY/1000 , sdY*1000 )
  nu <- nuMinusOne+1                    # nu must be >= 1
  nuMinusOne ~ dexp(1/29)               # prior on nu-1
}
```

Notice above that `dt` has three parameters, the first two of which are just like the parameters in `dnorm`. The third parameter is ν. The final lines of the model specification implement the prior on ν. The motivation for the prior is as follows. Look at the family of t distributions in Figure 16.4, and notice that nearly all the variation happens when ν is fairly small. In fact, once ν gets up to about 30, the t distribution is essentially normal. Thus, we would like a prior that gives equal opportunity to small values of ν (less than 30) and larger values of ν (greater than 30). A convenient distribution that captures this allocation is the exponential distribution. It is defined over positive values. It has a single parameter that specifies the reciprocal of its mean. Thus, the JAGS expression `nuMinusOne` $\sim$ `dexp(1/29)` says that the variable named `nuMinusOne` is exponentially distributed with a mean of 29. The value of `nuMinusOne` ranges from zero to infinity, so we must add 1 to make it range from 1 to infinity, with a mean of 30. This is accomplished by the penultimate line in the model specification. Figure 16.7, upper panel, shows the prior distribution on ν. Notice that it does indeed put considerable probability mass on small values of ν, but the mean of the distribution is 30 and large values of ν are also entertained. The lower panel of Figure 16.7 shows the same prior distribution on a logarithmic scale. The logarithmic scale is useful for displaying the distribution because it is extremely skewed on the original scale.

It should be emphasized that this choice of prior for ν is not uniquely "correct." While it is well motivated and has reasonable operational characteristics in many applications, there may be situations in which you would want to put more or less prior mass on small values of ν. The prior on ν can have lingering influence on the posterior even for fairly large data sets because the estimate of ν is influenced strongly by data in the tails of the distribution which, by definition, are relatively rare. To make small values of ν credible in the posterior, the data must contain some extreme outliers or many moderate outliers. Because outliers are rare, even in samples from truly heavy-tailed distributions, the prior on ν must have a fair amount of credibility on small values.

The script for running this model is called `Jags-Ymet-Xnom1grp-Mrobust-Example.R`. Figure 16.8 shows pair-wise plots of the three parameters. Notice that the credible values of σ are positively correlated with credible values of $\log10(\nu)$. This means that if the distribution is more normal with larger ν, then the distribution must also be wider with larger σ. This correlation is a signature of the data having outliers. To accommodate

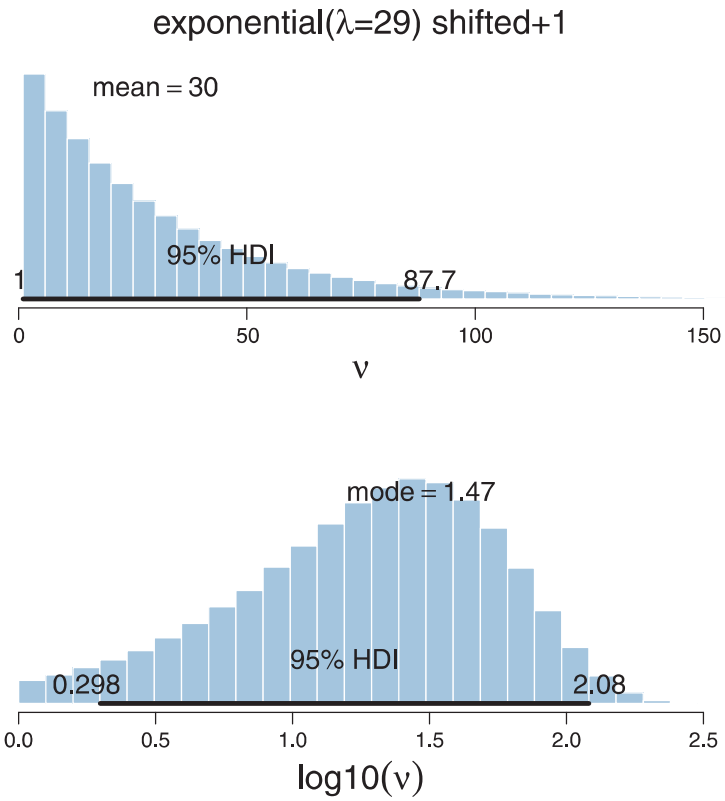

Figure 16.7 The prior on the normality parameter. Upper panel shows the shifted exponential distribution on the original scale of ν. Lower panel shows the same distribution on a logarithmic scale.

the outliers, either ν must be small to provide heavy tails or σ must be large to provide a wide distribution. We saw this same trade-off in the MLE examples of Figure 16.5.

Figure 16.9 shows other aspects of the posterior distribution. Notice the marginal posterior on the normality parameter in the lower left panel. Its mode (on the log10 scale) is only 0.68, which is noticeably reduced from the prior mode of 1.47 in the lower panel of Figure 16.7. This indicates that the data are better fit by small-ish values of ν because there are outliers.

Typically we are not interested in the exact estimated value of ν. We merely want to give the model flexibility to use heavy tails if the data demand them. If the data are normally distributed, then the posterior of ν will emphasize large values. Any value of ν greater than about 30 ($\approx$1.47 on the log10 scale) represents a nearly normal distribution, so its exact value does not have strong consequences for interpretation. On the other hand, small values of ν can be estimated accurately only by data in the extreme tails of the distribution, which, by their very nature, are rare. Therefore, we cannot anticipate much certainty in the estimate of ν, and we settle for broad statements about whether the posterior emphasizes small-ish values of $\log10(\nu)$ (e.g., under 1.47) or large-ish values of $\log10(\nu)$ (e.g., over 1.47).

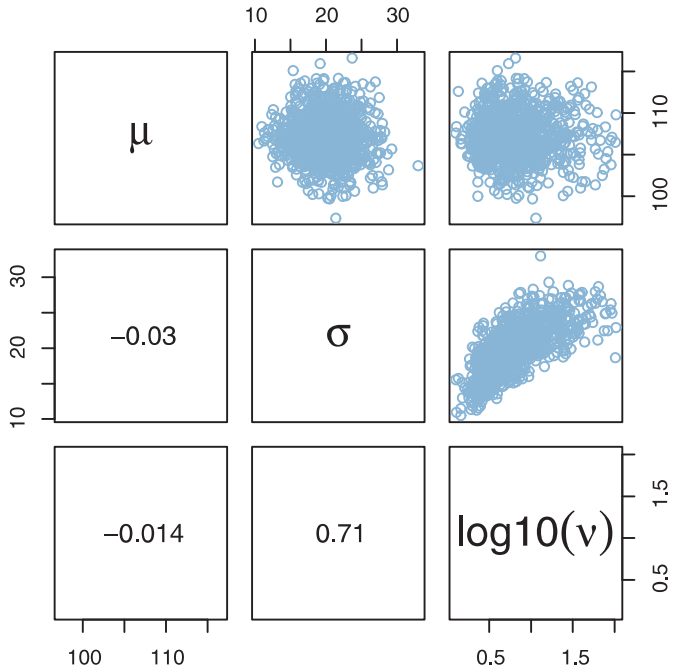

Figure 16.8 Posterior distribution of `Jags-Ymet-Xnom1grp-Mrobust-Example.R` applied to fictitious IQ data from a "smart drug" group. Off-diagonal cells show scatter plots and correlations of parameters indicated in the corresponding diagonal cells. Notice the strong positive correlation of σ and log10(ν).

The upper right panel of Figure 16.9 shows that the posterior predictive t distributions appear to describe the data better than the normal distribution in Figure 16.3, insofar as the data histogram does not poke out at the mode and the gaps under the shoulders are smaller.

Detailed comparison of Figure 16.9 with Figure 16.3 also reveals that the marginal posteriors on μ and effect size are a little bit tighter, with a little more of the distributions falling above the ROPEs. More prominently, σ in the robust estimate is much smaller than in the normal estimate. What we had interpreted as increased standard deviation induced by the smart drug might be better described as increased outliers. Both of these differences, that is, μ more tightly estimated and σ smaller in magnitude, are a result of there being outliers in the data. The only way a normal distribution can accommodate the outliers is to use a large value for σ. In turn, that leads to "slop" in the estimate of μ because there is a wider range of μ values that reasonably fit the data when the standard deviation is large, as we can see by comparing the upper and lower rows of Figure 16.1.

16.2.2. Using the *t* distribution in Stan

When you run the JAGS program yourself, you will see that it uses many steps to produce a posterior sample for σ that has an ESS exceeding 10,000. You will also see that the ESS for

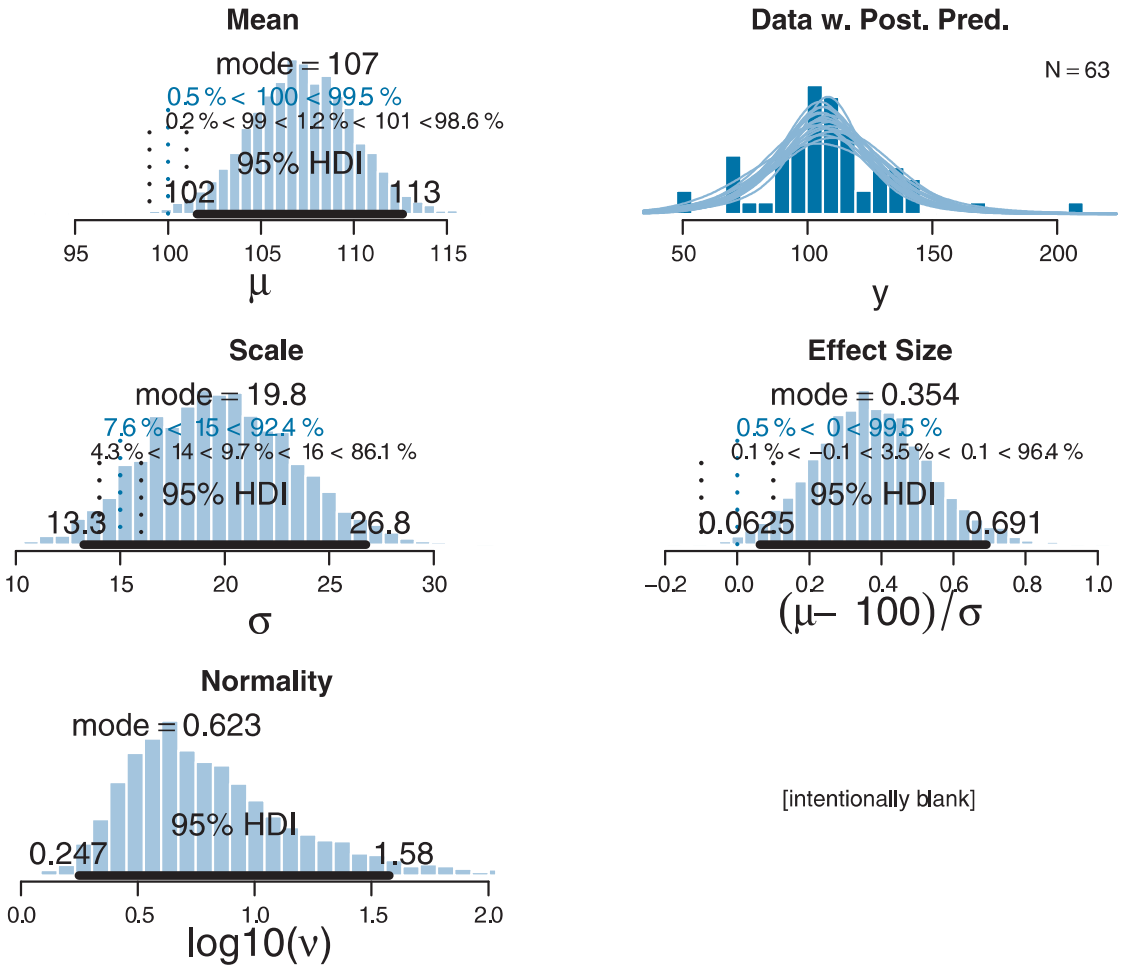

Figure 16.9 Posterior distribution of `Jags-Ymet-Xnom1grp-Mrobust` applied to fictitious IQ data from a "smart drug" group. Compare with Figure 16.3.

ν is less than 10,000 despite the long chain. In other words, there is high autocorrelation in the chains in JAGS.

We do not care that the chain for ν has a relatively small ESS because (a) we do not care about the exact value of ν when interpreting the posterior, as explained above, and (b) the exact value of ν has relatively little effect on the estimates of the other parameters. To be sure, the posterior sample of ν must be converged and truly representative of the posterior, but it does not need to be as finely detailed as the other parameters. Nevertheless, it would be less worrying if ν had a larger ESS.

The autocorrelation of the MCMC sampling in JAGS requires a long chain, which requires us to have patience while the computer chugs along. We have discussed two options for improving the efficiency of the sampling. One option is to run parallel chains on multiple cores using runjags (Section 8.7, p. 215). Another option is to implement the model in Stan, which may explore the parameter space more efficiently with its HMC sampling

(Chapter 14). This section shows how to run the model in Stan. Its results do indeed show a posterior sample of ν with higher ESS than JAGS.

In Stan, the sampling statement for a t distribution has the form `y ~ student_t(nu,mu,sigma)`. Notice that the normality parameter is the *first* argument, not the last as in JAGS (and BUGS). Notice also that the scale parameter is entered directly, not indirectly as the precision $1/\sigma^2$ as in JAGS (and BUGS). Here is the complete model specification in Stan, which begins with declaring all the variables for the data and parameters before getting to the model at the end:

```
data {
  int<lower=1> Ntotal ;
  real y[Ntotal] ;
  real meanY ;
  real sdY ;
}
transformed data { // compute the constants for the priors
  real unifLo ;
  real unifHi ;
  real normalSigma ;
  real expLambda ;
  unifLo <- sdY/1000 ;
  unifHi <- sdY*1000 ;
  normalSigma <- sdY*100 ;
  expLambda <- 1/29.0 ;
}
parameters {
  real<lower=0> nuMinusOne ;
  real mu ;
  real<lower=0> sigma ;
}
transformed parameters {
  real<lower=0> nu ;
  nu <- nuMinusOne + 1 ;
}
model {
  sigma ~ uniform( unifLo , unifHi ) ;
  mu ~ normal( meanY , normalSigma ) ;
  nuMinusOne ~ exponential( expLambda ) ;
  y ~ student_t( nu , mu , sigma ) ;  // vectorized
}
```

The script for running this model is called `Stan-Ymet-Xnom1grp-Mrobust-Example.R`. It uses a chain length and thinning that match the specifications for the corresponding JAGS script merely for purposes of comparison, but it turns out that Stan does not need such long chains as JAGS to produce the same ESS. Figure 16.10 shows the chain diagnostics for the ν parameter in JAGS and Stan. For both runs, there were 20,000 steps with a thinning of 5. The thinning was done merely to keep the saved file size down to a modest size for JAGS; thinning is not recommended if computer memory is not an

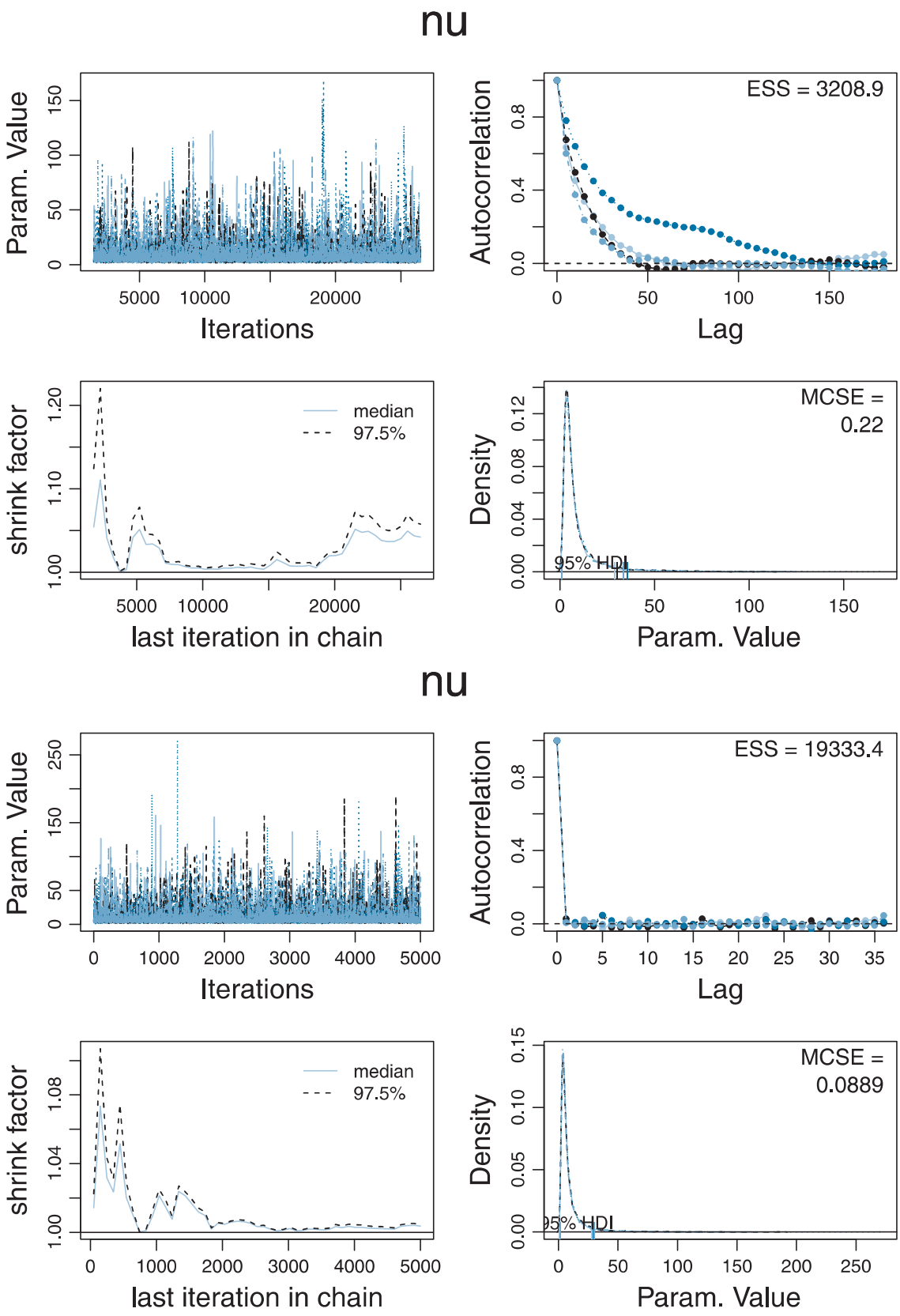

Figure 16.10 Chain diagnostics for JAGS (above) and Stan (below). Notice difference in autocorrelation in the upper right panels, and the resulting ESS.

issue. As you can see, Stan does an excellent job of sampling the normality parameter ν. Presumably this is because the Hamiltonian dynamics create proposed moves that efficiently jump around the parameter space. Stan also does a better job than JAGS in sampling σ and μ.

16.3. TWO GROUPS

An often-used research design is a comparison of two groups. For example, in the context of assaying the effect of a "smart drug" on IQ, instead of comparing the mean of the treatment group against an assumed landmark such as 100 (see Figure 16.3), it would make more sense to compare the treatment group against an identically handled placebo group. When there are two groups, we estimate the mean and scale for each group. When using t distributions for robust estimation, we could also estimate the normality of each group separately. But because there usually are relatively few outliers, we will use a single normality parameter to describe both groups, so that the estimate of the normality is more stably estimated.

Figure 16.11 illustrates the model structure. At the bottom of the diagram, the ith datum within group j is denoted $y_{i|j}$. The data within group j come from a t distribution with mean μ_j and scale σ_j. The normality parameter ν has no subscript because it is used for both groups. The prior for the parameters is exactly what was used for robust estimation of a single group in the previous section.

For two groups, there are only a few changes to the Stan model specification for a single group. The data now include a group identifier. The ith row of the data file specifies the IQ score as `y[i]` and the group identity as `x[i]`. In the data block of the model specification, the group membership variable `x` must be declared. The transformed data block is unchanged. The parameter block merely makes the mu and sigma variables vectors of 2 elements (for the

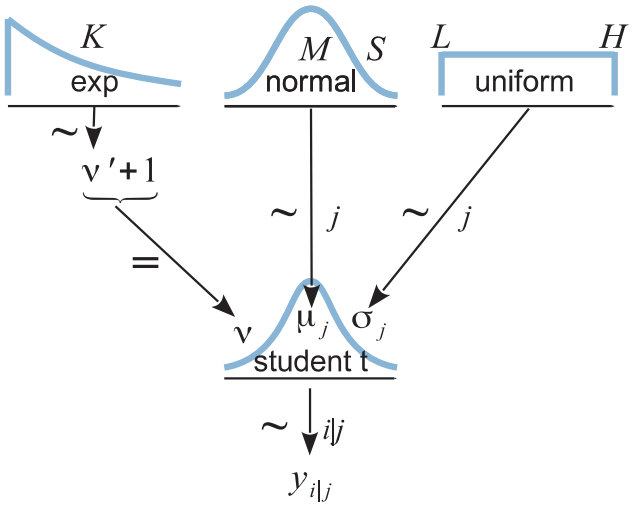

Figure 16.11 Dependency diagram for robust estimation of two groups. At the bottom of the diagram, $y_{i|j}$ is the ith datum within the jth group.

2 groups). Finally, in the model block, the likelihood function is put in a loop so that nested indexing of the group identifier can be used. Here is the complete model specification, with the changed lines marked by comments:

```
data {
  int<lower=1> Ntotal ;
  int x[Ntotal] ;                                    // group identifier
  real y[Ntotal] ;
  real meanY ;
  real sdY ;
}
transformed data {
  real unifLo ;
  real unifHi ;
  real normalSigma ;
  real expLambda ;
  unifLo <- sdY/1000 ;
  unifHi <- sdY*1000 ;
  normalSigma <- sdY*100 ;
  expLambda <- 1/29.0 ;
}
parameters {
  real<lower=0> nuMinusOne ;
  real mu[2] ;                                       // 2 groups
  real<lower=0> sigma[2] ;                           // 2 groups
}
transformed parameters {
  real<lower=0> nu ;
  nu <- nuMinusOne + 1 ;
}
model {
  sigma ~ uniform( unifLo , unifHi ) ;               // vectorized 2 groups
  mu ~ normal( meanY , normalSigma ) ;               // vectorized 2 groups
  nuMinusOne ~ exponential( expLambda ) ;
  for ( i in 1:Ntotal ) {
    y[i] ~ student_t( nu , mu[x[i]] , sigma[x[i]] ) ; // nested index of group
  }
}
```

Notice in the model block that there is essentially one line of code for each arrow in Figure 16.11. The only arrow that is not in the model block is the additive shift of ν by $+1$, which appears in Stan's transformed parameters block. When the diagram is implemented in JAGS, all the arrows appear in the model block.

A script for running the program is `Stan-Ymet-Xnom2grp-MrobustHet-Example.R`, and its format is minimally different from the other high-level scripts included with this book. The substantive changes are in the function definitions in `Stan-Ymet-Xnom2grp-MrobustHet.R`, which includes the Stan model specification (shown in the previous paragraph) and the specialized output graphs for this application.

The posterior distribution is shown in Figure 16.12. You can see the marginal posterior distributions on the five parameters (μ_1, μ_2, σ_1, σ_2, and ν) and the posterior distributions of the difference of means $\mu_2 - \mu_1$, the difference of scales $\sigma_2 - \sigma_1$, and the effect size, which is defined as the difference of means relative to the average scale: $(\mu_2 - \mu_1)/\sqrt{(\sigma_1^2 + \sigma_2^2)/2}$. The posterior distribution indicates that the difference of means is about 7.7 IQ points, but the 95% HDI of the difference barely excludes a small ROPE around zero. The posterior distribution on the effect size also suggests a nonzero difference, but the 95% HDI slightly overlaps the ROPE spanning ± 0.1. The difference of scales (i.e., $\sigma_2 - \sigma_1$) shows a credible nonzero difference, suggesting that the smart drug causes greater variance than the placebo.

Recall from Section 12.1.2.1, p. 340, that the differences of parameter values are computed from their jointly credible values at every step in the MCMC chain. This is worth remembering because the credible values of σ_2 and σ_1 are positively correlated in the posterior distribution, and therefore their difference cannot be accurately gleaned by considering their separate marginal distributions. The two scale parameters are positively correlated because they both trade off with the normality parameter. When the normality parameter is large, then both scale parameters must be large to accommodate the outliers in the two groups. When the normality parameter is small, then both scale parameters are better off being small to give higher probability to the centrally located data.

The posterior predictive check in the upper right panels of Figure 16.12 suggests that the t distribution is a reasonably good description of both groups. Neither group's data show clear departures from the smattering of credible t distributions.

16.3.1. Analysis by NHST

In traditional NHST, metric data from two groups would be submitted to a t-test. The t test is part of the standard R facilities; learn about it by typing `?t.test` at R's command line. When applied to the IQ data, here are the results:

```
> myDataFrame = read.csv( file="TwoGroupIQ.csv" )
> t.test( Score ~ Group , data=myDataFrame )

    Welch Two Sample t test
data:  Score by Group
t = -1.958, df = 111.441, p-value = 0.05273
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -15.70602585   0.09366161
sample estimates:
   mean in group Placebo mean in group Smart Drug
                100.0351                 107.8413
```

Notice that the p value is greater than 0.05, which means that the conventional decision would be *not* to reject the null hypothesis. This conclusion conflicts with the Bayesian analysis in Figure 16.12, unless we use a conservatively wide ROPE. The reason that the t test is less sensitive than the Bayesian estimation in this example is that the t test assumes normality and therefore its estimate of the within-group variances is too large when there are outliers.

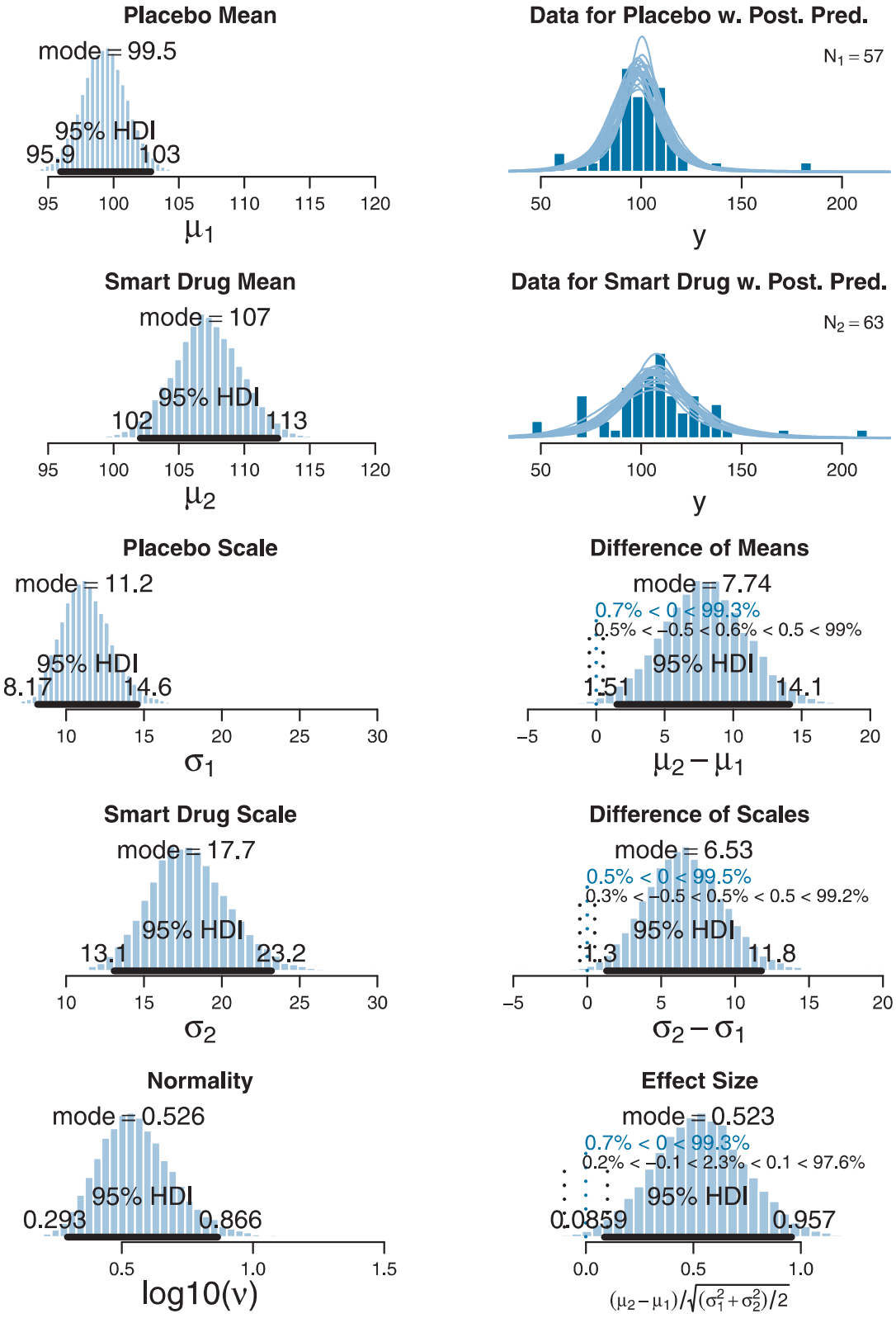

Figure 16.12 Posterior distribution for two groups.

The t test has other problems. Unlike the Bayesian analysis, the t test provides only a test of the equality of means, without a test of the equality of variances. To test equality of variances, we need to run an additional test, namely an F test of the ratio of variances, which would inflate the p values of both tests. Moreover, both tests compute p values based on hypothetical normally distributed data, and the F test is particularly sensitive to violations of this assumption. Therefore it would be better to use resampling methods to compute the p values (and correcting them for multiple tests).

I have previously written an extensive but self-contained article about this case of estimating two groups, with other examples comparing conclusions from NHST and Bayesian analysis (Kruschke, 2013a). The article includes programs for JAGS that have been translated into other formats by friendly enthusiasts; you can find links to their work at the Web site http://www.indiana.edu/kruschke/BEST/ (where BEST stands for Bayesian estimation). The article also explains Bayesian power analysis and some of the perils of p values, and, in an appendix, Bayes' factor approaches to assessing null values. If you are looking for a compact introduction to Bayesian data analysis, perhaps as a gift for a loved one, or as something to bring to the host of a party you are attending, that article might be just what you need.[2]

16.4. OTHER NOISE DISTRIBUTIONS AND TRANSFORMING DATA

When the data are not distributed like the assumed noise distribution, then the interpretation of the parameters can be problematic. For example, if the data have many outliers, and the assumed noise distribution is normal, then the estimate of the standard deviation parameter is artificially inflated by the outliers. If the data are skewed but the assumed distribution is symmetric, then the estimate of the mean parameter is artificially pulled by the skewed data values. In general, we want the noise distribution to accurately mimic the data, so that the parameters are meaningful.

If the initially assumed noise distribution does not match the data distribution, there are two ways to pursue a better description. The preferred way is to use a better noise distribution. The other way is to transform the data to a new scale so that they tolerably match the shape of the assumed noise distribution. In other words, we can either change the shoe to fit the foot, or we can squeeze the foot to fit in the shoe. Changing the shoe is preferable to squeezing the foot. In traditional statistical software, users were stuck with the pre-packaged noise distribution, and had no way to change it, so they transformed their data and squeezed them into the software. This practice can lead to confusion in interpreting the parameters because they are describing the transformed data, not the data on the original scale. In software such as JAGS and Stan, however, there is great flexibility in specifying various noise distributions (and higher level structure). We have seen one example in this chapter, in which an initially assumed normal distribution was changed to a t distribution. Many other distributions are available in JAGS and Stan, and we can also specify noise distributions by using the Bernoulli ones trick that was explained in Section 8.6.1, p. 214.

[2] Of course, the article can also be used for lining the bottom of bird cages, or for wrapping fish at the market.

As another example of non-normal noise distributions, consider models of response time. Response time data are typically positively skewed, because response times can only be so fast, but can often be very slow. There is a debate in the scientific literature regarding what sort of distribution best describes response times and why. As one recent example, Rouder, Lu, Speckman, Sun, and Jiang (2005) used the Weibull distribution in a hierarchical Bayesian model to describe response times. Whatever the preferred descriptive distribution is, probably it can be implemented in JAGS and Stan. For example, both JAGS and Stan have the Weibull distribution built in.

16.5. EXERCISES

Look for more exercises at https://sites.google.com/site/doingbayesiandataanalysis/

Exercise 16.1. [Purpose: Practice using different data files in the high-level script, with an interesting real example about alcohol preference of sexually frustrated males.] Shohat-Ophir et al. (2012) were interested in alcohol preferences of sexually deprived males. The procedure is illustrated in Figure 16.13, and was described as follows: "One cohort, rejected-isolated, was subjected to courtship conditioning; they experienced 1-h sessions of sexual rejection by mated females, three times a day, for 4 days. …Flies in the mated-grouped cohort experienced 6-h sessions of mating with multiple receptive virgin females (ratio 1:5) for 4 days. Flies from each cohort were then tested in a two-choice preference assay, in which they voluntarily choose to consume food with or without 15% ethanol supplementation. (Shohat-Ophir et al., 2012, p. 1351, citations and figure reference removed)" For each fly, the amount of each type of food consumed was

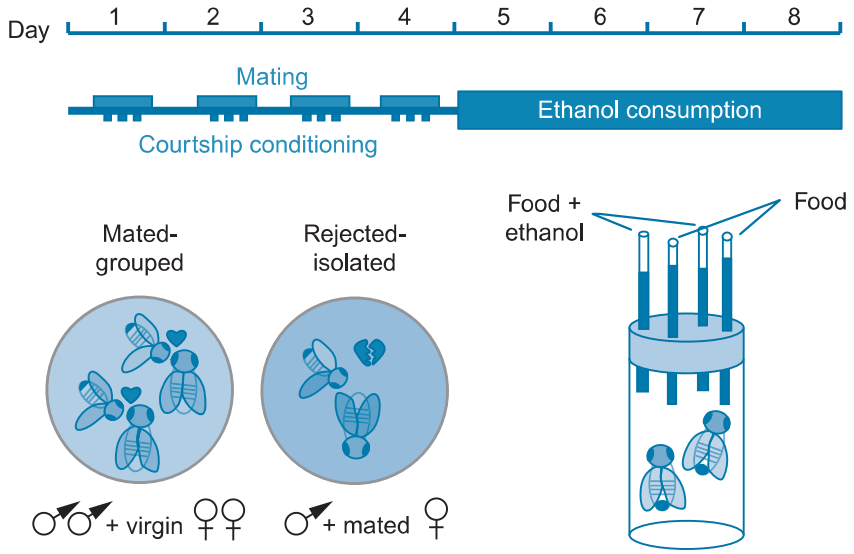

Figure 16.13 Procedure used for investigation of alcohol preference of sexually frustrated males (*Drosophila melanogaster*). From Figure 1A of Shohat-Ophir, Kaun, Azanchi, Mohammed, and Heberlein (2012). Reprinted with permission from AAAS.

converted to a *preference ratio*: the amount of ethanol-supplemented food minus the amount of regular food divided by the total of both. I constructed 3-day summary preference scores for each individual fruit fly by summing the consumption of ethanol and non-ethanol across days 6–8. The amounts of food consumed and the preference ratios are in the data file named ShohatOphirKAMH2012dataReduced.csv. My thanks to Dr. Galit Shohat-Ophir for providing the data.

(A) Run `Jags-Ymet-Xnom2grp-MrobustHet-Example.R` on the preference scores. Make sure that the ROPE on the means and standard deviation is scaled appropriately to the data. How big are differences between groups relative to the uncertainty of the estimate? What do you conclude? (If this result interests you, then you will also be intrigued by the results in Section 19.3.2, p. 563.)

(B) Instead of focusing on the *relative* amounts of ethanol and regular food consumed, we might also be interested in the absolute total amount of food consumed. Run the analysis on the total consumption data, which has column name GrandTotal in the data file. What do you conclude? In particular, would you want to make an argument to accept the null hypothesis of no difference? (Review Section 12.1.1, beginning on p. 336.)

Exercise 16.2. [Purpose: More practice using different data files in the high-level script, using a real example, with skewed data.] The typical lifespan of a laboratory rat that eats *ad lib* is approximately 700 days. When rats are placed on a restricted diet, their longevity can increase, but there is a lot of variability in lifespans across different individual rats. Restricting the diet might not only affect the typical lifespan, but restricting the diet might also affect the variance of the lifespan across rats. We consider data from R. L. Berger, Boos, and Guess (1988), as reported in Hand, Daly, Lunn, McConway, and Ostrowski (1994, data set #242), and which are available in the file named `RatLives.csv`.

(A) Run the two-group analysis on the rat longevity data. Use JAGS or Stan as you prefer (report which one you used). Report the code you used to read in the data file, specify the column names for the data, and the ROPEs appropriate to the scale of the data. Do the groups appear to differ in their central tendencies and variances? Does the value of the normality parameter suggest that the data have outliers relative to a normal distribution?

(B) The data within each group appear to be skewed to the left. That is, within each group, there are many rats that died relatively young, but there are fewer outliers on the high end. We could try to implement a skewed noise distribution, or we could try to transform the data so they are approximately symmetric within each group. We will try the latter approach here. To get rid of leftward skew, we need a transformation that expands the rightward values. We will try squaring the data. Read in the data and append a transformed data column like this:

```
myDataFrame = read.csv( file="RatLives.csv" )
myDataFrame = cbind( myDataFrame , DaysLiveSq = myDataFrame$ DaysLive^2 )
yName="DaysLiveSq"
```

Change the specification of the ROPEs to be appropriate to the transformed data. Do the groups appear to differ in their central tendencies and variances on the days-squared scale?

Does the value of the normality parameter suggest that the data have outliers relative to a normal distribution on the days-squared scale? Is the posterior effect size on the days-squared scale much different than the posterior effect size on the days scale from the previous part?

Exercise 16.3. [Purpose: For two groups, examine the implied prior on the effect size and on the differences of means and scales.]

(A) Modify the script `Stan-Ymet-Xnom2grp-MrobustHet-Example.R` and functions in `Stan-Ymet-Xnom2grp-MrobustHet.R` as needed to display the prior distribution. See Section 14.3.4, p. 412, for details. Explain what changes you made, and include the graphical output. Does Stan have convergence problems?

(B) Modify the JAGS version of the program to sample from the prior. Refer to Section 8.5, p. 211, for a reminder of how to sample from the prior in JAGS.

(C) From the previous two parts, is the prior on the effect size and differences suitably "noncommittal"? Briefly discuss.