9

Using Gibbs Samplers to Compute Bayesian Posterior Distributions

In Chapter 8, we introduced the fundamental ideas of Bayesian inference, in which prior distributions on parameters are used together with data to obtain posterior distributions and thus interval estimates of parameters. However, in practice, Bayesian posterior distributions are often difficult to compute.

Gibbs Sampling is a computational method that uses Markov Chains, as discussed in Chapter 7, to approximate posterior distributions. The central idea is to use available information about a prior distribution and data to construct an ergodic Markov Chain whose limiting distribution is the desired posterior distribution. Then we simulate enough steps of the chain to obtain a good approximation to the limiting distribution

In this chapter, we consider several relatively simple Bayesian models, explicitly illustrating how to program suitable chains in R in order to approximate posterior distributions and obtain interval estimates of parameters. In Chapter 10, we show how WinBUGS software can simplify the programming to do inference for more intricate Bayesian models.

9.1 Bayesian Estimates of Disease Prevalence

In Section 5.2 we considered how one might use the properties and results of a medical screening test to estimate the prevalence of a disease. In particular, we assumed we know the sensitivity and specificity of a screening test,

$$\eta = P\{\text{Positive test}|\text{Disease present}\} = P\{T = 1|D = 1\}$$

and

$$\theta = P\{\text{Negative test}|\text{Disease absent}\} = P\{T = 0|D = 0\},$$

respectively. Based on these quantities, we sought to estimate the prevalence of the disease $\pi = P\{D = 1\}$, from the equation $\pi = (\tau + \theta - 1)/(\eta + \theta - 1)$, where $\tau = P\{T = 1\}$. If τ is estimated by $t = A/n$, which is the ratio of the

number of individuals with positive tests to the sample size, then replacing τ by t in this equation gives an estimate p of π:

$$p = \frac{A/n + \theta - 1}{\eta + \theta - 1} = \frac{t + \theta - 1}{\eta + \theta - 1}. \tag{9.1}$$

For the derivation see page 117. Endpoints of a confidence interval for τ can be plugged into this equation to obtain a confidence interval for π.

However, we have seen some circumstances in which such an estimate of π falls outside the interval $[0, 1]$. Even more often, the corresponding confidence interval for π can extend beyond this interval. Another difficulty with this method arises when the sample size is small and the proportion of "Successes" is near 0 or 1. Then binomial confidence intervals are known to be problematic and equation (9.1) may not provide a useful interval estimate of π.

In Section 6.4 we investigated the situation in which, for a particular population, we know the predictive power of a positive test $\gamma = P\{D = 1|T = 1\}$ and of a negative test $\delta = P\{D = 0|T = 0\}$ in addition to η and θ. Here the relationships $\gamma = \pi\eta/[\pi\eta+(1-\pi)(1-\theta)]$ and $\delta = (1-\pi)\theta/[\pi(1-\eta)+(1-\pi)\theta]$, follow from Bayes' Theorem. So if we knew π along with η and θ, we could compute γ and δ. Accordingly, it seems reasonable that we should be able to compute π if η, θ, γ and δ are known. In the examples of Section 6.4, we saw how this can be done either by simulation (simple Gibbs Sampler) or analytically (solving for the steady state of a Markov Chain). Unfortunately, these particular procedures are mainly of theoretical and pedagogical interest because data to estimate γ and (especially) δ are not typically available in practical situations.

Fortunately, in a framework with a Bayesian prior distribution, we can use a Gibbs Sampler to find useful estimates of π. With a prior distribution on π having support $[0, 1]$, the following example shows how to obtain a posterior probability interval for π based on data A and n, and on known values of the sensitivity η and specificity θ, with no need to make assumptions about the predictive values γ and δ. Because values of π outside of $[0, 1]$ are not contemplated in the prior, they have zero probability under the posterior.

Example 9.1. Suppose we use a screening test with sensitivity $\eta = 99\%$ and specificity $\theta = 97\%$, and among $n = 1000$ subjects we see $A = 49$ positive results. We use a beta prior distribution for π. That is, $\pi \sim \mathsf{BETA}(\alpha, \beta)$. Not claiming to have advance information about π, we choose the flat prior with $\alpha = \beta = 1$, so that the prior is $\pi \sim \mathsf{BETA}(1, 1) = \mathsf{UNIF}(0, 1)$.

Our Gibbs sampler starts with an arbitrary initial value π_1^* of π. From π_1^* and our knowledge of η and θ, we speculate as to the number X of the A test-positive subjects that may have the disease. Based on π_1^*, the probability that any one of these A subjects has the disease is equal to the predictive value of a positive test, $\gamma_1^* = \pi_1^*\eta/[\pi_1^*\eta+(1-\pi_1^*)(1-\theta)]$. So we use a binomial distribution with A trials and this "Success" probability γ_1^* to simulate X. In much the same way, we simulate the number Y of the $B = n - A$ test-negative

subjects that have the disease. Altogether, we now have $X + Y$ subjects out of n with the disease, and we can use this simulated total to update the beta distribution for π—as in the election polling examples of Chapter 8. From this updated distribution we simulate π_2^*, and iterate the procedure from there to get $\pi_3^*, \pi_4^*, \ldots$. In symbols, the partial conditional distributions relating the key quantities are

$$X|A, \pi \sim \mathsf{BINOM}(A, \gamma), \;\; Y|B, \pi \sim \mathsf{BINOM}(B, 1-\delta), \;\; \text{and}$$

$$\pi|X, Y \sim \mathsf{BETA}(\alpha + X + Y, \; \beta + n - X - Y),$$

where $\gamma = \pi\eta/[\pi\eta + (1-\pi)(1-\eta)]$, $\delta = (1-\pi)\theta/[\pi(1-\eta) + (1-\pi)\theta]$, and $B = n - A$. These relationships are used in the R code below. Simulated values of π shown with asterisks (*) above are elements of the vector `PI` (all capitals) in the code.

Because the distribution of $\pi_i^* =$ `PI[i]` depends only on known parameters and the previous value $\pi_{i-1}^* =$ `PI[i-1]` the values in `PI` simulate a Markov process with a continuous state space as in Chapter 7. It can be shown that the limiting distribution of this process is the posterior distribution of π based on the prior and the data.

```
# set.seed(1237)
m = 50000                        # iterations
PI = numeric(m);  PI[1] = .5     # vector for results, initial value
alpha = 1;  beta = 1             # parameters of beta prior
eta = .99;  theta = .97          # sensitivity;  specificity
n = 1000;  A = 49;  B = n - A    # data

for (i in 2:m)
{
  num.x = PI[i-1]*eta;  den.x = num.x + (1-PI[i-1])*(1 - theta)
  X = rbinom(1, A, num.x/den.x)
  num.y = PI[i-1]*(1 - eta);  den.y = num.y + (1-PI[i-1])*theta
  Y = rbinom(1, B, num.y/den.y)
  PI[i] = rbeta(1, X + Y + alpha, n - X - Y + beta)
}
aft.brn = seq(m/2 + 1,m)
mean(PI[aft.brn])
quantile(PI[aft.brn], c(.025, .975))
par(mfrow=c(2,1))
  plot(aft.brn, PI[aft.brn], type="l")
  hist(PI[aft.brn], prob=T)
par(mfrow=c(1,1))

> mean(PI[aft.brn])
[1] 0.02059591
> quantile(PI[aft.brn], c(.025, .975))
       2.5%       97.5%
0.007428221 0.035523630
```

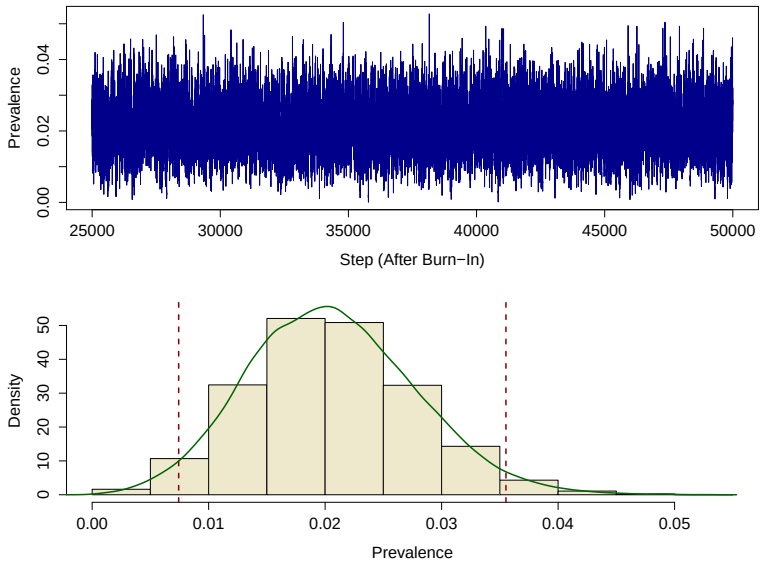

Figure 9.1. History plot (top) and histogram of 25,000 sampled prevalence values after burn-in. The plot shows good mixing of the Gibbs Sampler in Example 9.1. The histogram approximates the posterior distribution $\pi|A, B$; an estimated density curve is superimposed (see Problem 9.6). Dotted lines indicate the 95% Bayesian interval estimate of π. Compare with Figure 9.10 on page 231.

The histogram in the top panel of Figure 9.1 indicates the posterior distribution of π. Taking the mean of this distribution, we have the point estimate $\pi = 0.021$, and cutting off 2.5% from each tail of the this simulated distribution, we have the Bayesian interval estimate $(0.007, 0.036)$ for π. Problem 9.1 invites you to see how these results change when we use some informative prior distributions.

Essentially, Figure 9.2 is made using the following additional statements.

```
par(mfrow=c(1,2))
    acf(PI[aft.brn], ylim=c(0, .6))
    plot(1:m, cumsum(PI)/(1:m), type="l", ylim=c(.016, .024))
par(mfrow=c(1,1))
```

The three diagnostic graphs in the top panel of Figure 9.1 and in Figure 9.2 show that, in spite of some positive autocorrelation for neighboring values of $\pi_i^*|A, B$, the sampler mixes well and that running averages after burn-in converge smoothly to the point estimate. (See Problem 9.4 for more about running averages and burn-in.)

> It is easy to see why there is positive autocorrelation. If by chance at some step i in the iteration, we obtain a rather large value of π_i^*, then it is somewhat likely that unusually large values of X or Y or both will result at

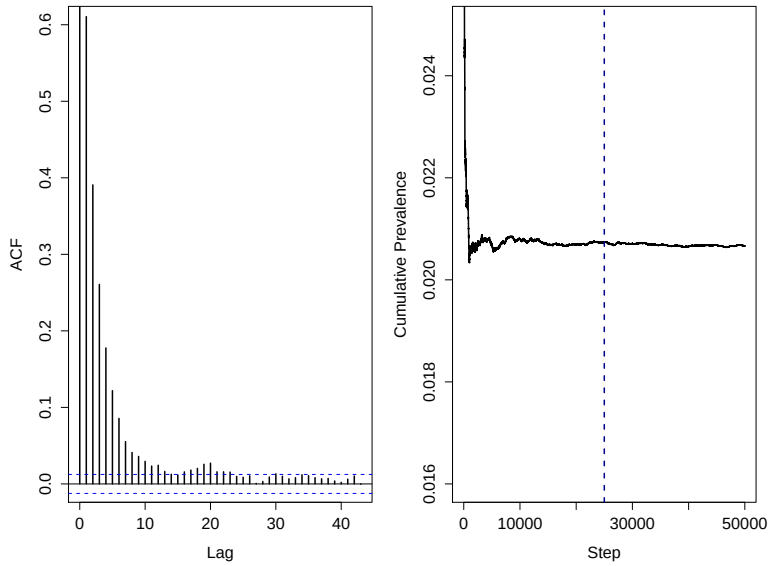

Figure 9.2. The ACF plot (left) of sampled values of π after burn-in for Example 9.1 shows autocorrelations of sampled values decaying to insignificance for lags above 25. The plot of cumulative averages of sampled values shows steady convergence to the Bayesian point estimate of π after burn-in (dotted vertical line.)

step $i+1$. Consequently, the first parameter of the beta distribution may be inflated, and along with it the expectation of the next value of π_i^*. However, the data and the prior exert an overall tendency towards appropriate values of π_i^*, so the process does not consistently "run away" towards ever larger values. A mirror of this argument holds in case we get an unusually low value of π_i^* at some step in the simulation. (See Problem 9.5 for more on autocorrelations.)

In this example, equation (9.1) gives the traditional point estimate $p = 0.020$ of π and corresponding confidence interval $(0.006, 0.034)$. [The Agresti-Couil estimates of τ give $p = 0.022$ and $(0.008, 0.036)$.] So in this situation where equation (9.1) works well, the Gibbs Sampler with a noninformative prior gives almost identical results. Moreover, Problem 9.2 illustrates that a Gibbs Sampler gives reasonable point and interval estimates of π, even in situations where equation (9.1) gives problematic negative estimates. $\diamondsuit$

Although there are many important applications in which equation (9.1) is not useful, the benefit of the Bayesian framework is not just to avoid absurd estimates outside the range of possible parameter values. In practice, one seldom encounters a situation where there is no prior information at all about prevalence. For example, if $\pi = 93\%$— or even $\pi = 30\%$—for a serious disease, the evidence of this public health catastrophe would be evident all around us

without reference to data from medical screening tests. Also, in practice one often encounters situations where there is very little data and a reasonable approach is to meld expert opinion with the bit of objective information that is available. For such reasons, it is fair to say that Gibbs Sampling, as illustrated in Example 9.1, has wide applicability in estimating prevalence from the results of screening tests across a broad spectrum of applications.

9.2 Bayesian Estimates of Normal Mean and Variance

In Chapter 8, we discussed separately (i) Bayesian estimation of the mean μ of a normal population when the variance is known and (ii) Bayesian estimation of the variance $\sigma^2 = \theta$ when the mean is known. In this section we use a Gibbs Sampler to provide Bayesian estimates of the normal mean and variance simultaneously. In the following example, we see that—when relatively flat priors are used—Bayesian results are very similar to those obtained from traditional methods based on Student's t and chi-squared distributions. In several problems, we explore the effect of informative priors.

Example 9.2. Changes in students' heights.

Heights of $n = 41$ young men at a boarding school are measured in the morning and also in the evening. For each student, the difference x_i, morning height minus evening height, is found. Considering these subjects to be a random sample from an appropriate population, our main purpose is to estimate the population mean μ of the change in height.

If we take differences in heights x_i, for $i = 1, 2, ..., 41$, to be normally distributed, this example is similar in some ways to Examples 8.3 and 8.7 (concrete beams) and Examples 8.4 and 8.8 (hemoglobin). Here we assume $x_i \sim \mathsf{NORM}(\mu, \sigma)$, where both parameters are unknown. The classical unbiased point estimators are $\bar{x} = \frac{1}{n}\sum_i x_i$ for μ and $s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$ for $\sigma^2 = \theta$. We seek Bayesian point and interval estimates for μ and σ.

Prior distributions. First, we choose a prior distribution for μ of the form $\mathsf{NORM}(\mu_0, \sigma_0)$, where $\theta_0 = \sigma_0^2$. Specifically, we choose $\mu_0 = 0$ because we have no reason to suppose heights differ systematically between morning and evening. Also, we want a reasonably flat prior because we claim no particular expertise in the matter of height changes, and we do not really know whether students might grow or shrink a little during the day. Thus, rather arbitrarily, we choose $\sigma_0 = 20mm$ (about 3/4 of an inch), so $\theta_0 = 400$.

Next, we choose a prior distribution for θ of the form $\mathsf{IG}(\alpha_0, \kappa_0)$, where α_0 and κ_0 are shape and rate parameters, respectively. We do not have much idea how accurately the measuring will be done, and differences involve two measurements. Also, if there are differences in heights during a day, those differences may be larger for some students than for others. Accordingly, we choose $\alpha_0 = 1/2$ and $\kappa_0 = 1/5$, which means we think the standard deviation of the differences is pretty sure to be between $0.3mm$ and $20mm$ (computed

from `sqrt(1/qgamma(c(.975, .025), 1/2, 1/5))`. This choice seems reasonable. As heights go, a millimeter is very small and it seems unlikely that measurements could be made much more precisely than that. Also, $20mm$ seems an unbelievably large amount of measurement error or variability in daily changes among students.

Data. From the data we find $n = 41$ differences x_i, for which the sample mean is $\bar{x} = 9.6mm$ and the sample variance is $s^2 = 7.48$, so that $s = 2.73mm$. (See Problem 9.11 for the 41 differences.) A Bayesian analysis will combine these data and our priors to give posterior distributions upon which we base our inferences.

Posterior distributions. Almost exactly as in Example 8.7, we have

$$\mu|\mathbf{x}, \theta \sim \mathsf{NORM}(\mu', \sqrt{\theta'}),$$

where the updated parameters (denoted with primes), reflecting the data, are $\mu' = \theta'(n\bar{x}/\theta + \mu_0/\theta_0)$ and $\theta' = (n/\theta + 1/\theta_0)^{-1} = \theta_0\theta/(n\theta_0 + \theta)$. Also, similar to our results in Example 8.8,

$$\theta|\mathbf{x}, \mu \sim \mathsf{IG}(\alpha', \kappa'),$$

where $\alpha' = \alpha_0 + n/2$ and $\kappa' = \kappa_0 + [(n-1)s^2 + n(\bar{x} - \mu)^2]/2$. The important change from Example 8.8 is the second term inside brackets in the expression for κ', needed here to take $\bar{x}$ into account because μ is not known. (See Problem 9.13 for some details of the derivation.)

Now, in order to find the posterior distributions of $\mu|\mathbf{x}$ and $\theta|\mathbf{x}$, we use a Gibbs Sampler to perform the required integrations:

$$p(\mu|\mathbf{x}) \propto \int p(\mu|\mathbf{x}, \theta)\, p(\theta|\mathbf{x})\, d\theta \;\text{ and }\; p(\theta|\mathbf{x}) \propto \int p(\theta|\mathbf{x}, \mu)\, p(\mu|\mathbf{x})\, d\mu.$$

Gibbs Sampler. Using the R code below, we simulate a bivariate Markov Chain with vectors denoted in the program as `MU` and `THETA`. The limiting distribution of this chain provides estimates of the posterior distributions of μ and θ, respectively, upon which Bayesian estimates are based. The simulation begins with known quantities: the parameters μ_0 and θ_0 of the normal prior distribution on μ, the parameters α_0 and κ_0 of the inverse gamma prior distribution on θ, the data $\bar{x}$ and s^2, and an arbitrary starting value `THETA[1]`.

Iteratively, at step i of the Gibbs Sampler, we generate values `MU[i]` and `THETA[i]` of the Markov Chain. We sample `MU[i]` from $\mathsf{NORM}(\mu', \sqrt{\theta'})$, where θ in the expressions for μ' and θ' is taken to be `THETA[i-1]`. Then we sample `THETA[i]` from $\mathsf{IG}(\alpha', \kappa')$, where μ in the expression for κ' is taken to be `MU[i]`.

We choose $m = 50\,000$ steps as the burn-in point. Accordingly, we consider values of `MU` and `THETA` from steps $i = m/2 + 1$ through m as simulated distributions of $\mu|\mathbf{x}$ and $\theta|\mathbf{x}$, respectively. Cutting off 2.5% from the tails of these simulated distributions gives us Bayesian interval estimates of μ and θ. From the interval estimate of $\theta = \sigma^2$, we obtain an interval estimate of σ.

```
# set.seed(1237)
m = 50000                                # iterations
MU = numeric(m);  THETA = numeric(m)     # sampled values
THETA[1] = 1                             # initial value
n = 41;  x.bar = 9.6;  x.var = 2.73^2    # data
mu.0 = 0;  th.0 = 400                    # mu priors
alp.0 = 1/2;  kap.0 = 1/5                # theta priors

for (i in 2:m)
{
  th.up = 1/(n/THETA[i-1] + 1/th.0)
  mu.up = (n*x.bar/THETA[i-1] + mu.0/th.0)*th.up
  MU[i] = rnorm(1, mu.up, sqrt(th.up))

  alp.up = n/2 + alp.0
  kap.up = kap.0 + ((n-1)*x.var + n*(x.bar - MU[i])^2)/2
  THETA[i] = 1/rgamma(1, alp.up, kap.up)
}

# Bayesian point and probability interval estimates
aft.brn = (m/2 + 1):m
mean(MU[aft.brn])            # point estimate of mu
bi.MU = quantile(MU[aft.brn], c(.025,.975));  bi.MU
mean(THETA[aft.brn])         # point estimate of theta
bi.THETA = quantile(THETA[aft.brn], c(.025,.975));  bi.THETA
SIGMA = sqrt(THETA)
mean(SIGMA[aft.brn])         # point estimate of sigma
bi.SIGMA = sqrt(bi.THETA);  bi.SIGMA

par(mfrow=c(2,2))
  plot(aft.brn, MU[aft.brn], type="l")
  plot(aft.brn, SIGMA[aft.brn], type="l")
  hist(MU[aft.brn], prob=T);  abline(v=bi.MU, col="red")
  hist(SIGMA[aft.brn], prob=T);  abline(v=bi.SIGMA, col="red")
par(mfrow=c(1,1))

> mean(MU[aft.brn])            # point estimate of mu
[1] 9.594313
> bi.MU = quantile(MU[aft.brn], c(.025,.975));  bi.MU
     2.5%     97.5%
 8.753027 10.452743

> mean(THETA[aft.brn])         # point estimate of theta
[1] 7.646162
> bi.THETA = quantile(THETA[aft.brn], c(.025,.975));  bi.THETA
     2.5%     97.5%
 4.886708 11.810233
> SIGMA = sqrt(THETA)
```

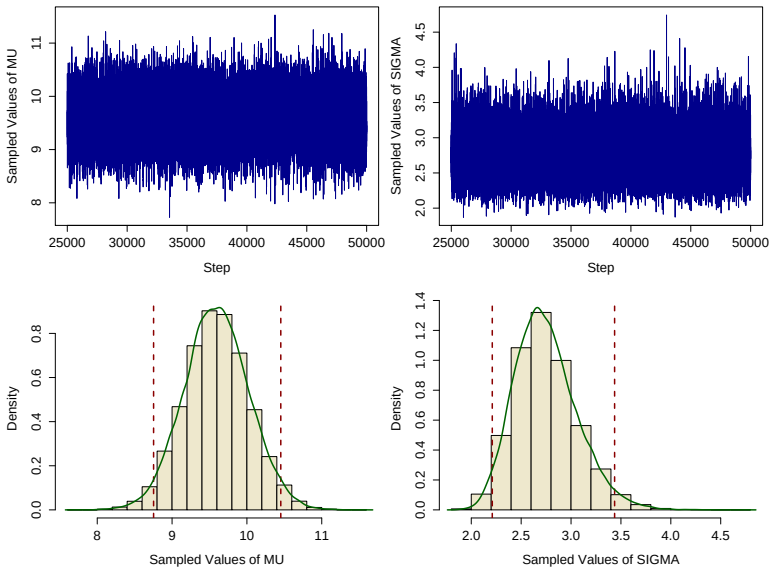

Figure 9.3. History plots (top) and histograms of sampled values approximating $\mu|\mathbf{x}$ (left) and $\sigma|\mathbf{x}$ in the Gibbs Sampler of Example 9.2. All values are after burn-in. Vertical dashed lines show 95% Bayesian interval estimates.

```
> mean(SIGMA[aft.brn])          # point estimate of sigma
[1] 2.747485
> bi.SIGMA = sqrt(bi.THETA);  bi.SIGMA
    2.5%    97.5%
2.210590 3.436602
```

Diagnostic graphs in Figure 9.3 (top) show good behavior of the Gibbs Sampler, so the numerical results from `MU` and `SIGMA` can be trusted accurately to represent the posterior distributions $\mu|\mathbf{x}$ and $\sigma|\mathbf{x}$. Also, because this is a bivariate Markov Chain we show, in Figure 9.4 on page 220, a scatterplot of the last 10 000 sampled pairs approximating $(\mu, \sigma)|\mathbf{x}$. (For additional diagnostic graphs, see Problem 9.8.)

The 95% Bayesian interval estimates are $(8.73, 10.44)$ for μ and $(2.22, 3.45)$ for σ. On average, it seems that from morning to evening the students shrink in height by about a centimeter ($10mm$ or about $3/8in$). Other studies have found similar decreases in height. A plausible explanation is that the cartilage between vertebrae is compressed during the day and expands during sleep.

Frequentist methods that use Student's t and chi-squared distributions give a 95% confidence interval $(8.74, 10.46)$ for μ and a 95% confidence interval $(2.24, 3.49)$ for σ (see Problem 9.9). The Bayesian probability intervals are slightly shorter than the corresponding frequentist confidence intervals, possibly because our prior distributions, even though diffuse, provide some useful

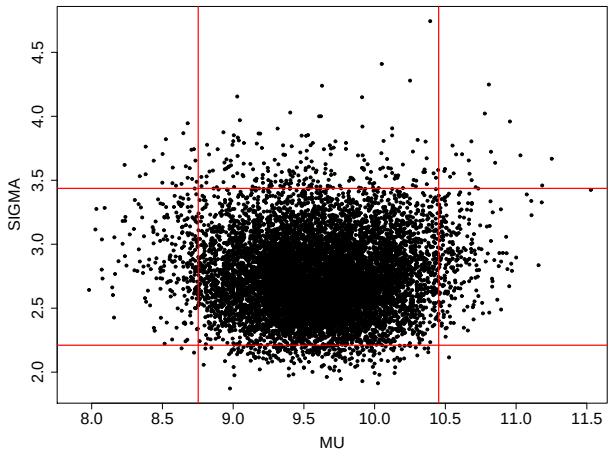

Figure 9.4. Scatterplot of the last 10 000 pairs in Example 9.2 simulating $(\mu, \sigma)|\mathbf{x}$. The prior distributions on μ and $\theta = \sigma^2$ are independent, as are the sample statistics $\bar{x}$ and s. Also, this plot shows no marked association between simulated values of $\mu|\mathbf{x}$ and $\sigma|\mathbf{x}$. Reference lines indicate 95% Bayesian interval estimates for μ and σ.

information about variability. But in this example, the effect of our prior distributions is relatively small because there is enough data to overwhelm the effect of priors that are not strongly informative. $\diamondsuit$

In general, if we make the prior parameter σ_0 very large and the parameters α_0 and κ_0 very small, then neither prior affects the posterior by much, and the Bayesian intervals are nearly equal to the corresponding frequentist confidence intervals. Specifically, one formulation of a noninformative prior gives the following posterior distributions

$$t = \frac{\mu - \bar{x}}{s/\sqrt{n}} \sim \mathsf{T}(n-1) \quad \text{and} \quad (n-1)s^2/\sigma^2 \sim \mathsf{CHISQ}(n-1)\,,$$

where μ and σ are random variables and $\bar{x}$ and s are observed values. These yield 95% Bayesian interval estimates for μ and σ that are numerically exactly the same as the respective traditional frequentist 95% confidence intervals.

Moreover, there are particular ways to formulate informative priors so that posterior distributions given $\bar{x}$ and s can be be expressed in closed form. Then Bayesian interval estimates can be found for μ and σ without the need for Gibbs sampling. (For discussions of more general priors see [BT73] and [Lee04].)

In practice, Gibbs Samplers are especially important in models with many parameters, to which Example 9.2 provides an important pedagogical bridge. In the next section, we consider a three-parameter model for which traditional methods may be especially inappropriate and for which a Gibbs Sampler is a practical way to compute useful Bayesian inferences.

9.3 Bayesian Estimates of Components of Variance

In Section 9.1 we saw that a Bayesian approach to estimating disease prevalence gave useful estimates in circumstances where traditional methods can give absurd results. In this section we look at one more practical situation in which a traditional frequentist approach often does not provide useful estimates and a Bayesian framework does.

Suppose a manufacturing process has two steps. Precursors of the finished items are made in batches, and then the batches are used to produce the individual items. If a key measurement on the final items shows excessive variability, the question arises whether this variability may arise mainly at the batch level or mainly at the final stage of the overall process. A logical step toward reducing variability is to try to understand where it arises. We want to estimate the two components of variance that contribute towards overall variance of individual items.

Assuming normal distributions for errors, we can write the measured value of the jth item from the ith batch as

$$x_{ij} = \mu + A_i + e_{ij},$$

where $A_i \sim \mathsf{NORM}(0, \sqrt{\theta_A})$, $e_{ij} \sim \mathsf{NORM}(0, \sqrt{\theta})$, and all A_i and e_{ij} are mutually independent. This implies that measurements on two items from two different batches are independent, but that measurements x_{ij} and $x_{ij'}$ on two items from batch i are correlated. Specifically, $\mathrm{V}(x_{ij}) = \mathrm{V}(x_{ij'}) = \theta_A + \theta$, $\mathrm{Cov}(x_{ij}, x_{ij'}) = \theta_A$, and $\rho_I = \rho(x_{ij}, x_{ij'}) = \theta_A/(\theta_A + \theta)$. The ratio ρ_I, called the **intraclass correlation**, is the proportion of the total variance that arises at the batch level of the manufacturing process.

Example 9.3. Consider a pilot project to manufacture a pharmaceutical drug in two steps as just described. Technicians want to know if variability among batches makes an important contribution to product variability. They assay $r = 10$ individual lots from each of $g = 12$ batches.

In this example, so we can know whether our estimates are reasonable, we generate data with known parameter values $\mu = 100$, $\theta_A = 15^2 = 225$, and $\theta = 9^2 = 81$, so that $\rho_I = 225/306 = 0.7353$—values roughly modeled after proprietary data. These $gr = 120$ observations are plotted in Figure 9.5 and the procedure for generating them is shown in Problem 9.14. Because the data are normal, it is sufficient to look at the $g = 12$ batch means $\bar{x}_{i\cdot} = \frac{1}{r}\sum_j x_{ij}$ and variances $s_i^2 = \frac{1}{r-1}\sum_j (x_{ij} - \bar{x}_{i\cdot})^2$. Summary data by batch are shown in the printout below.

```
Batch      1      2      3      4      5      6
Mean    91.9  129.0  104.1   75.7  108.7  100.2
SD      9.96  10.07   4.98  12.16   5.06  10.65

Batch      7      8      9     10     11     12
X.bar   62.6  107.5   66.7  129.1  106.8   93.4
X.sd    6.52  11.05   9.90   8.39   8.99   8.14
```

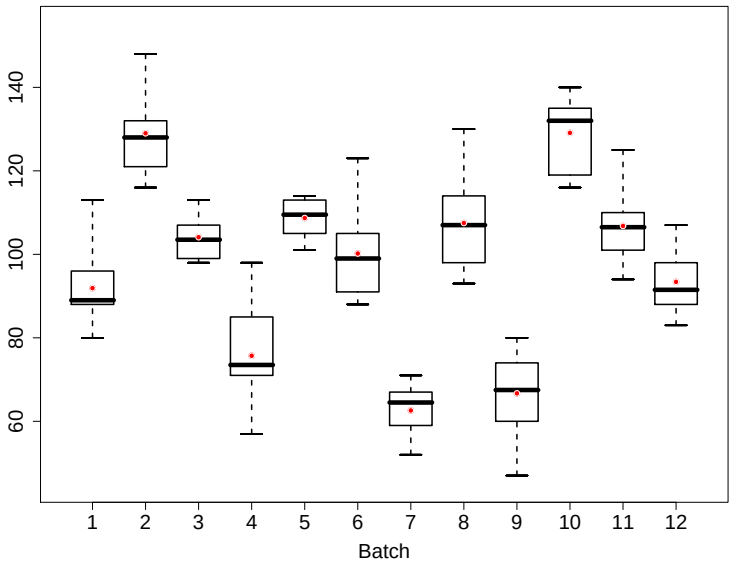

Figure 9.5. Boxplots of the data of Example 9.3. Dots show group means. Clearly, variance among batches contributes significantly to the variability of individual observations. Compare with Figure 9.6, which illustrates the data of Problem 9.17 where batch-to-batch variability is relatively much smaller.

In these circumstances, it is traditional to look at the following statistics.

$$\bar{x}_{\cdot\cdot} = \sum_{i=1}^{g}\sum_{j=1}^{r} x_{ij} = 97.975$$

$$\text{MS(Batch)} = \frac{r}{g-r}\sum_{i=1}^{g}(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2 = 4582.675$$

$$\text{MS(Error)} = \frac{1}{g(r-1)}\sum_{i=1}^{g}\sum_{j=1}^{r}(x_{ij} - \bar{x}_{i\cdot})^2 = \frac{1}{g}\sum_{i=1}^{g} s_i^2 = 82.68056$$

For normal data, these three statistics are independent, and the following distributions are useful for making confidence intervals.

$$(\bar{x}_{\cdot\cdot} - \mu)/\sqrt{\text{MS(Batch)}/gr} \sim \mathsf{T}(g-1)$$
$$(g-1)\text{MS(Error)}/(r\theta_A + \theta) \sim \mathsf{CHISQ}(g-1)$$
$$(gr-1)\text{MS(Error)}/\theta \sim \mathsf{CHISQ}(gr-1)$$
$$R = \text{MS(Batch)}/\text{MS(Error)} \sim \mathsf{F}(g-1, gr-1)$$

Unbiased point estimates are $\hat{\mu} = x_{\cdot\cdot} = 97.975$ (compared to the known $\mu = 100$), $\hat{\theta} = \text{MS(Error)}$, and $\hat{\theta}_A = [\text{MS(Batch)} - \text{MS(Error)}]/r = 449.999$.

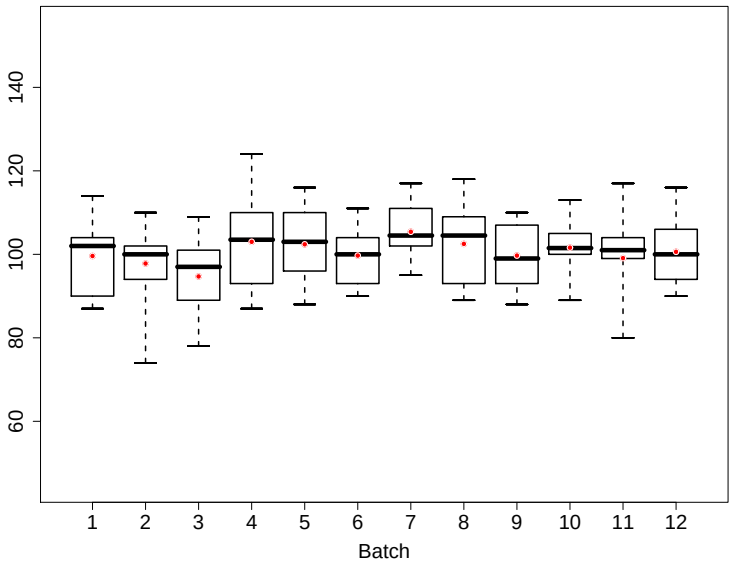

Figure 9.6. Boxplots of the data of Problem 9.17, drawn to the same scale as Figure 9.5 for easy comparison. Here batch-to-batch variance θ_A is so small that traditional methods of estimating it are problematic. A Gibbs Sampler yields a useful Bayesian probability interval (see the problem, page 237).

Taking square roots, we have estimates for the standard deviations $\hat{\sigma} = 9.09$ (compared to $\sigma = 9$) and $\hat{\sigma}_A = 21.21$ (compared to $\sigma_A = 15$). Of these, the estimate of θ_A can be problematic. Information about batch variability is entangled with information about item variability. Because $\hat{\theta}_A$ is found by subtraction, its value can be negative, even though θ_A is nonnegative.

Frequentist confidence intervals for μ, σ, and ρ_I can be obtained from t, chi-squared, and F distributions, respectively (see Problem 9.15). There is no such straightforward confidence interval for θ_A. Also, the confidence interval for ρ_I includes negative values when $\hat{\theta}_A < 0$. There are models in which intraclass correlation can legitimately be negative (see [SC80]), but ours is not one of them.

A Bayesian framework for this model is similar to that of Example 9.2, with one additional variance parameter. Our prior distributions are

$$\mu \sim \mathsf{NORM}(\mu_0, \sqrt{\theta_0}), \quad \theta_A \sim \mathsf{IG}(\alpha_0, \kappa_0), \quad \text{and} \quad \theta \sim \mathsf{IG}(\beta_0, \lambda_0)\,.$$

In order to have noninformative priors for this example with simulated data, we select a large value of θ_0 and small values of all four inverse gamma parameters.

Partial conditional distributions used in the Gibbs Sampler to compute the posterior distributions of μ, θ, and θ_A are as follows (see [GS85], page 405).

$$\mu|\theta_A, \mathbf{A} \sim \mathsf{NORM}(\mu', \sqrt{\theta'}), \;\; \theta_A|\mathbf{A}, \mu \sim \mathsf{IG}(\alpha', \kappa'), \;\; \text{and} \;\; \theta|\mathbf{X}, \mathbf{A} \sim \mathsf{IG}(\beta', \lambda'),$$

where

$$\mu' = (\mu_0\theta_A + \theta_0 \textstyle\sum_i A_i)/(\theta_A + r\sum_i A_i) \;\text{ and }\; \theta' = \theta_0\theta_A/(\theta_A + r\textstyle\sum_i A_i)$$

in the partial conditional for $\mu|\theta_A, \mathbf{A}$;

$$\alpha' = \alpha_0 + g/2 \;\text{ and }\; \kappa' = \kappa_0 + \tfrac{1}{2}\textstyle\sum_i (A_i - \mu)^2$$

in the partial conditional for $\theta_A|\mathbf{A}, \mu$; and

$$\beta' = \beta_0 + gr/2 \;\text{ and }\; \lambda' = \lambda_0 + \tfrac{1}{2}[(r-1)\textstyle\sum_i s_i^2 + r\sum_i (A_i - \bar{x}_{i\cdot})^2]$$

in the partial conditional for $\theta|\mathbf{X}, \mathbf{A}$. In the above, the g elements of $\mathbf{A}$ are

$$A_i \sim \mathsf{NORM}((r\theta_A\bar{x}_{i\cdot} + \theta\mu)/(r\theta_A + \theta),\; [\theta\theta_A/(r\theta_A + \theta)]^{1/2}).$$

The R code below shows how these relationships can be used in a Gibbs Sampler to simulate a 3-dimensional Markov Chain of sampled values in vectors denoted `MU`, `VAR.BAT`, and `VAR.ERR`. From them we can find Bayesian interval estimates of μ, θ_A, and θ, respectively. Each step of the sampler uses the prior distributions and the data.

- The sampler starts with an arbitrary initial value of `MU[1]`. We also require values of the random effects A_i, so-called **latent variables**, which are not directly observable as data. In the program these are denoted by the g-vector `a`. On the first pass through the loop, we use the group means $\bar{x}_{i\cdot}$ as initial values of `a`. On later passes, updated values of `a` are available.
- Next, the sampler uses values of `a` and `MU[1]` to sample `VAR.BAT[2]`, and `VAR.ERR[2]`, then it uses `VAR.BAT[2]` and `VAR.ERR[2]` to sample `MU[2]`.
- At the end of the loop, new latent values `a` are sampled using `MU[2]`, `VAR.BAT[2]`, and `VAR.ERR[2]`.
- The loop is iterated, at each pass using values of `MU[k-1]`, `VAR.BAT[k-1]`, `VAR.ERR[k-1]`, and the newest values `a` to sample elements of the vectors with index `[k]`.

Finally, when all iterations are completed, Bayesian interval estimates of μ, θ_A and θ are found from the values after burn-in of the three simulated vectors, and intervals for σ_A, σ, and ρ_I are found from information the Gibbs Sampler provides about $\theta_A = \sigma_A^2$ and $\theta = \sigma^2$.

```
#Assumes matrix X with g rows (batches), r columns (reps),
#Or provide g-vectors of batch means and SDs as the 2nd line.
# set.seed(443)
X.bar = apply(X, 1, mean);  X.sd = apply(X, 1, sd)
m = 50000;  b = m/4              # iterations;  burn-in
MU = VAR.BAT = VAR.ERR = numeric(m)
```

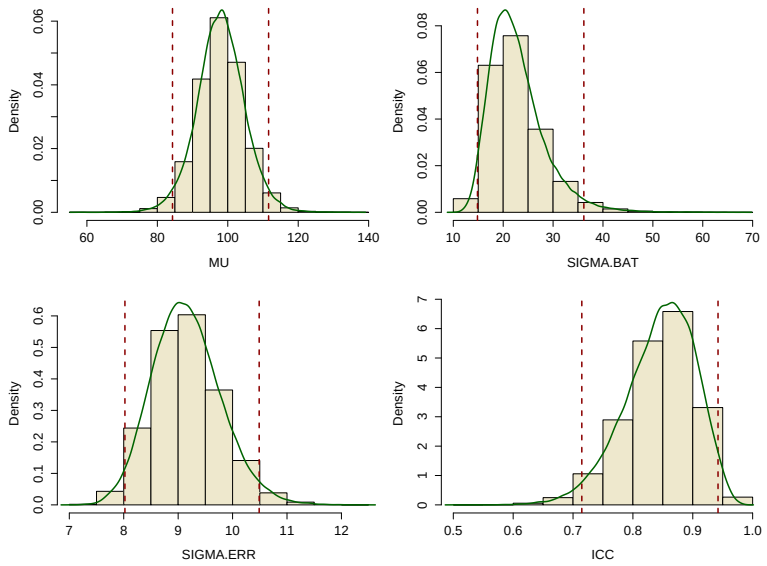

Figure 9.7. Histograms of simulated posteriors for Example 9.3. Results are from the vectors `MU`, `SIGMA.BAT`, `SIGMA.ERR`, and `ICC` of the Gibbs Sampler after burn-in. Indications of the 95% Bayesian interval estimate and the estimated posterior density are superimposed on each histogram.

```
mu.0  = 0;     th.0  = 10^10     # prior parameters for MU
alp.0 = .001;  kap.0 = .001      # prior parameters for VAR.BAT
bta.0 = .001;  lam.0 = .001      # prior parameters for VAR.ERR
MU[1] = 150;   a = X.bar         # initial values


for (k in 2:m)  {
  alp.up = alp.0 + g/2
  kap.up = kap.0 + sum((a - MU[k-1])^2)/2
  VAR.BAT[k] = 1/rgamma(1, alp.up, kap.up)

  bta.up = bta.0 + r*g/2
  lam.up = lam.0 + (sum((r-1)*X.sd^2) + r*sum((a - X.bar)^2))/2
  VAR.ERR[k] = 1/rgamma(1, bta.up, lam.up)

  mu.up = (VAR.BAT[k]*mu.0 + th.0*sum(a))/(VAR.BAT[n] + g*th.0)
  th.up = th.0*VAR.BAT[k]/(VAR.BAT[n] + g*th.0)
  MU[k] = rnorm(1, mu.up, sqrt(th.up))

  deno = r*VAR.BAT[k] + VAR.ERR[k]
  mu.a = (r*VAR.BAT[k]*X.bar + VAR.ERR[k]*MU[k])/deno
  th.a = (VAR.BAT[k]*VAR.ERR[k])/deno
  a = rnorm(g, mu.a, sqrt(th.a))  }
```

```
mean(MU[b:m]);  sqrt(mean(VAR.BAT[b:m]));  sqrt(mean(VAR.ERR[b:m]))
bi.MU = quantile(MU[b:m], c(.025,.975))
SIGMA.BAT = sqrt(VAR.BAT);  SIGMA.ERR = sqrt(VAR.ERR)
bi.SG.B = quantile(SIGMA.BAT[b:m], c(.025,.975))
bi.SG.E = quantile(SIGMA.ERR[b:m], c(.025,.975))
ICC = VAR.BAT/(VAR.BAT+VAR.ERR);
bi.ICC = quantile(ICC[b:m], c(.025,.975))
bi.MU;  bi.SG.B;  bi.SG.E;   bi.ICC

par(mfrow=c(2,2))
  hist(MU[b:m], prob=T);        abline(v=bi.MU)
  hist(SIGMA.BAT[b:m], prob=T); abline(v=bi.SG.B)
  hist(SIGMA.ERR[b:m], prob=T); abline(v=bi.SG.E)
  hist(ICC[b:m], prob=T);       abline(v=bi.ICC)
par(mfrow=c(1,1))

> mean(MU[b:m]); sqrt(mean(VAR.BAT[b:m])); sqrt(mean(VAR.ERR[b:m]))
[1] 98.00235
[1] 23.41598
[1] 9.177717

> bi.MU;  bi.SG.B;  bi.SG.E;   bi.ICC
     2.5%     97.5%
 84.30412 111.59195
     2.5%     97.5%
 14.84336  36.16719
     2.5%     97.5%
 8.022159 10.488738
     2.5%     97.5%
0.7146413 0.9421721
```

From the printouts for one run of the Gibbs Sampler, we see that the Bayesian point estimates (98.0 for μ, 23.4 for σ_A, and 9.2 for σ) are not much different from the traditional ones (98.0, 21.2, and 9.1, respectively). Also, the 95% Bayesian interval estimates of these parameters all happen to cover the known values we used to simulate the data (100, 15, and 9, respectively). Based on distributions stated earlier, traditional 95% confidence intervals are $(85.7, 110.2)$ for μ, $(8.0, 10.5)$ for σ, and $(0.71, 0.94)$ for ρ_I.

Figure 9.7 shows the approximate posterior distributions and 95% Bayesian interval estimates for μ, σ_A, σ, and ρ_I. Figure 9.8 shows diagnostic plots—all favorable—for the dimension of the sampler estimating σ_A, and we leave the remaining diagnostic plots to Problem 9.16.

The 95% Bayesian interval estimate of σ_A is very wide because we have information on only $g = 12$ batches. In contrast, we have much more information about σ and that information is not entangled with other effects, so the interval for σ is shorter. If all observations cost the same, it might be better to increase g at the expanse of r. But in practice, batches are often expensive. In our consulting experience, the number of batches has rarely exceeded 12.

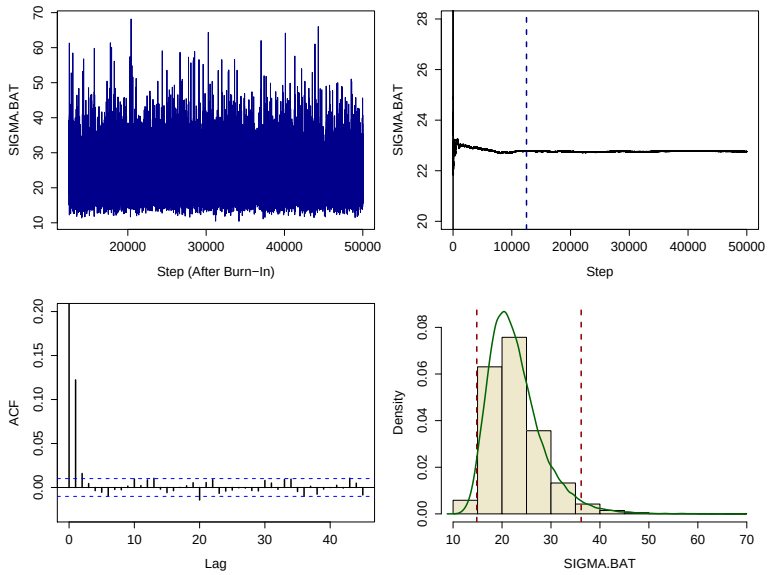

Figure 9.8. Diagnostic plots for the simulation of the posterior distribution of σ_A in Example 9.3. Evidently, the Gibbs Sampler converges smoothly to its limiting distribution. Only the plot of cumulative means shows all steps; the others use steps after burn-in. The histogram is also shown in Figure 9.7.

Faced with long interval estimates for the batch component of variance, some authors and practitioners use 90% intervals instead. (In this particular example, that would give an interval that doesn't cover 15.) In a Bayesian context where appropriate prior information is available, an informative prior on $\theta_A = \sigma_A^2$ might give a shorter and more useful interval estimate. $\diamondsuit$

In this example, traditional methods give useful answers. However, traditional methods become problematic when the batch component of variance is relatively small. Then the usual point estimate $\hat{\theta}_A$ of θ_A may be negative and the confidence interval for ρ_I can include negative values. As we see in Problem 9.18, this happens more than occasionally.

- One standard interpretation is to say this is an indication that θ_A must be "very small." Maybe so, but presumably we would not have chosen a model containing θ_A without reason to believe batches might make some contribution to overall variance, and this analysis leaves us with no idea how large θ_A might really be.
- A related traditional approach is to test the null hypothesis H_0: $\theta_A = 0$ against H_1: $\theta_A > 0$. What do we say if H_0 is accepted, as it surely will be when $\hat{\theta}_A < 0$? Again the interpretation is that θ_A is "very small." But then we would have to speculate about the power of the test, the probability

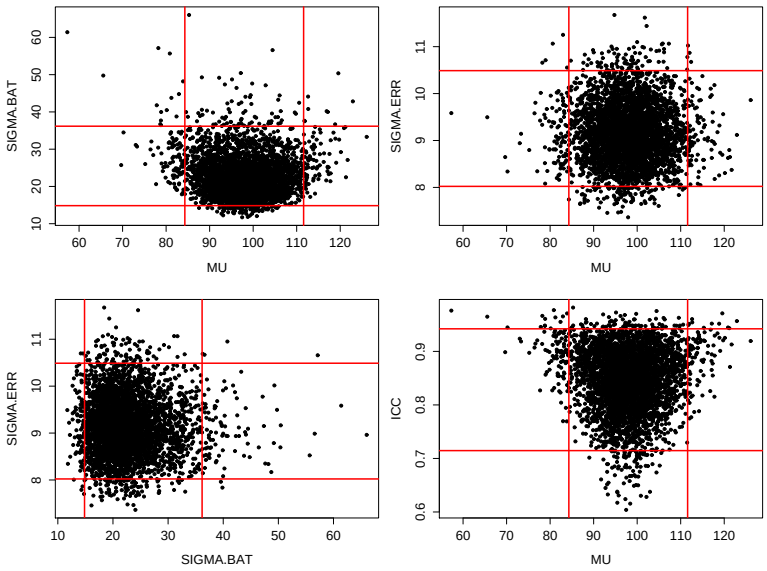

Figure 9.9. Bivariate plots from the Gibbs Sampler of Example 9.3. In each of the four panels, every 10th step after burn-in is plotted. The vectors `MU`, `SIGMA.BAT`, and `SIGMA.ERR` are not mutually independent. In particular, values of `MU` far from $\bar{x}_{..}$ tend to be associated with large values of `ICC`.

of accepting H_0, for various possible values of $\theta_A > 0$. We still have no idea how large θ_A might actually be. Perhaps this difficulty has been made more obscure by the terminology of hypothesis testing, but it has not gone away.

In Problem 9.17 we show data for which $\hat{\theta}_A < 0$, but the Gibbs Sampler of Example 9.3 gives useful Bayesian interval estimates for all parameters (see Figure 9.12 on page 236). Several additional problems show real data that result in $\hat{\theta}_A > 0$. For a comparison the models of Examples 9.2 and 9.3, see Problem 9.22.

Note: We have seen that the traditional method of moments for estimating σ_A can give negative values. Computationally intensive methods are available to find approximate maximum likelihood estimates (MLEs) of σ_A. Except when the MLE of σ_A is small, these methods can also provide approximate confidence intervals. These MLE results are numerically similar to Bayesian results based on a noninformative prior from a Gibbs Sampler. When the MLE of σ_A is small, computational difficulties involving collinearity arise in finding MLE confidence intervals.

In this chapter we have seen situations in which a Bayesian approach has something to offer over a traditional one, and in which a Gibbs Sampler is a useful method for computing approximate posterior distributions. An inconvenience in using a Gibbs Sampler is the need to specify partial conditional distributions upon which to base the programming. In Chapter 10 we show how BUGS software can do Gibbs Sampling simply by specifying the model, but without having to write and explicitly program partial conditional distributions.

9.4 Problems

Problems for Section 9.1, Estimating Prevalence of a Disease
In working these problems, modify the program of the example as appropriate.

9.1 Estimating prevalence with an informative prior.

a) According to the distribution $\mathsf{BETA}(1, 10)$, what is the probability that π lies in the interval $(0, 0.2)$?
b) If the prior $\mathsf{BETA}(1, 10)$ is used with the data of Example 9.1, what is the 95% Bayesian interval estimate of π?
c) What parameter β would you use so that $\mathsf{BETA}(1, \beta)$ puts about 95% probability in the interval $(0, 0.05)$?
d) If the beta distribution of part (c) is used with the data of Example 9.1, what is the 95% Bayesian interval estimate of π?

Hints: c) Use `beta = seq(1:100); x = pbeta(.05, 1, beta); min(beta[x>=.95])`. Explain. d) The mean of the posterior distribution $\pi|X, Y$ is about 1.8%

9.2 In Example 5.2 on p118, the test has $\eta = .99$ and $\theta = .97$, the data are $n = 250$ and $A = 6$, and equation (9.1) on p212 gives an absurd negative estimate of prevalence, $\pi = -0.62\%$.

a) In this situation, with a uniform prior, what are the Bayesian point estimate and (two-sided) 95% interval estimate of prevalence? Also, find a one-sided 95% interval estimate that provides an upper bound on π.
b) In part (a), what estimates result from using the prior $\mathsf{BETA}(1, 30)$?

Comment: a) See Figure 9.10. Two-sided 95% Bayesian interval: $(0.03\%, 2.9\%)$. Certainly, this is more useful than a negative estimate, but don't expect a narrow interval with only $n = 250$ observations. Consider that a flat-prior 95% Bayesian interval estimate of τ based directly on $t = 6/250$ is roughly $(1\%, 5\%)$.

9.3 In each part below, use the uniform prior distribution on π and suppose the test procedure described results in $A = 24$ positive results out of $n = 1000$ subjects.

a) Assume the test used is not a screening test, but a gold standard test, so that $\eta = \theta = 1$. Follow through the code for the Gibbs Sampler in

Example 9.1, and determine what values of X and Y must always occur. Run the sampler. What Bayesian interval estimate do you get? Explain why the result is essentially the same as the Bayesian interval estimate you would get from a uniform prior and data indicating 24 *infected* subjects in 1000, using the code `qbeta(c(.025,.975), 25, 977)`.

b) Screening tests exist because it is not feasible to administer a gold standard test to a large group of subjects. So the situation of part (a) is not likely to occur in the real world. But it does often happen that everyone who gets a positive result on the screening test is given a gold standard test, and no gold standard tests are given to subjects with negative screening test results. Thus, in the end, we have $\eta = 99\%$ and $\theta = 1$. In this case, what part of the Gibbs Sampler becomes deterministic? Run the Gibbs Sampler with these values and report the result.

c) Why are the results from parts (a) and (b) not much different?

Hints: a) The Gibbs Sampler simulates a large sample precisely from $\mathsf{BETA}(25, 977)$ and cuts off appropriate tails. Why these parameters? Run the additional code:
`set.seed(1237); pie=c(.5, rbeta(m-1, 25, 977)); mean(pie[(m/2):m])`
c) Why no false positives among the 24 in either (a) or (b)? Consider false negatives.

9.4 *Running averages and burn-in periods.* In simulating successive steps of a Markov Chain we know that it may take a number of steps before the running averages of the resulting values begin to stabilize to the mean value of the limiting distribution. In a Gibbs Sampler it is customary to disregard values of the chain during an initial **burn-in period**. Throughout this chapter we rather arbitrarily choose to use $m = 50\,000$ iterations and take the burn-in period to extend for the first $m/4$ or $m/2$ steps. These choices have to do with the appearance of stability in the running average plot and how much simulation error we are willing to tolerate. For example, the running averages in the righthand panel of Figure 9.2 (page 215) seem to indicate smooth convergence of the mean of the π-process to the posterior mean after 25 000 iterations. The parts below provide an opportunity to explore the perception of stability and variations in the length of the burn-in period. Use $m = 50\,000$ iterations throughout.

a) Rerun the Gibbs Sampler of Example 9.1 three times with different seeds, which you select and record. How much difference does this make in the Bayesian point and interval estimates of π? Use one of the same seeds in parts (b) and (c) below.

b) Redraw the running averages plot of Figure 9.2 so that the vertical plotting interval is $(0, 0.5)$. (Change the `plot` parameter `ylim`.) Does this affect your perception of when the process "becomes stable"? Repeat, letting the vertical interval be $(0.20, 0.22)$, and comment.

c) Change the code of the Gibbs Sampler in the example so that the burn-in period extends for 15 000 steps. Compared to the results of the example, what change does this make in the Bayesian point and interval estimates of π? Repeat for a burn-in of 30 000 steps and comment.

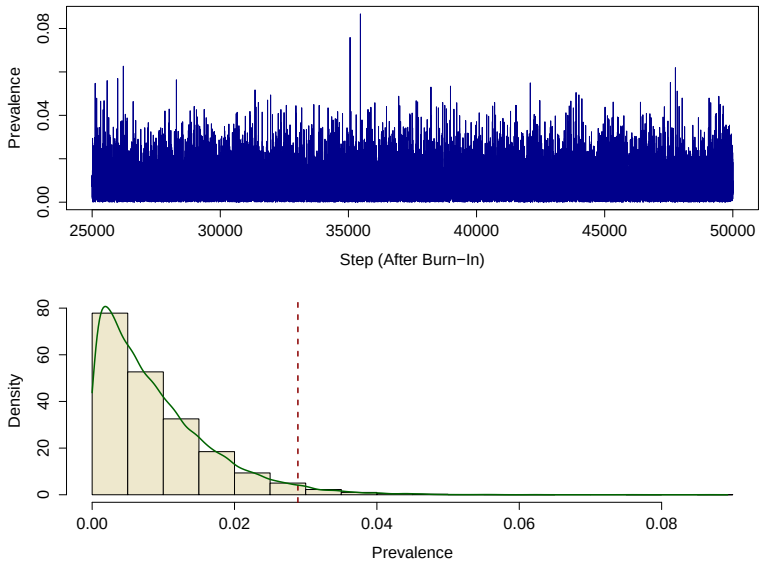

Figure 9.10. History plot (top) and histogram of 25,000 sampled prevalence values after burn-in for Problem 9.2. Here, traditional methods give a nonsensical negative point estimate of prevalence π. But a one-sided 95% Bayesian interval provides a useful upper bound on π (dotted line). Compare with Figure 9.1 on page 214.

9.5 *Thinning.* From the ACF plot in Figure 9.2 on p215, we see that the autocorrelation is near 0 for lags of 25 steps or more. Also from the right hand plot in this figure it seems that the process of Example 9.1 stabilizes after about 15 000 iterations. One method suggested to mitigate effects of autocorrelation, called **thinning,** is to consider observations after burn-in located sufficiently far apart that autocorrelation is not an issue.

a) Use the data and prior of Example 9.1. What Bayesian point estimate and probability interval do you get by using every 25th step, starting with step 15 000? Make a histogram of the relevant values of `PI`. Does thinning in this way have an important effect on the inferences?
b) Use the statement `acf(PI[seq(15000, m, by=25)])` to make the ACF plot of these observations. Explain what you see.

9.6 *Density estimation.* A histogram, as in Figure 9.1, is one way to show the approximate posterior distribution of π. But the smooth curve drawn through the histogram there reminds us that we are estimating a *continuous* posterior distribution. A Gibbs Sampler does not give us the functional form of the posterior density function, but the smooth curve is a good approximation. After the Gibbs Sampler of Example 9.1 is run, the following additional code superimposes an estimated density curve on the histogram of sampled values.

```
est.d = density(PI[aft.brn], from=0, to=1);  mx = max(est.d$y)
hist(PI[aft.brn], ylim=(0, mx), prob=T, col="wheat")
lines(est.d, col="darkgreen")
median(PI[aft.brn]);  est.d$x[est.d$y==mx]
```

a) Run the code to verify that it gives the result claimed. In the R Session window, type `?density` and browse the information provided on **kernel density estimation**. In this instance, what is the reason for the parameters `from=0, to=1`? What is the reason for finding `mx` before the histogram is made? In this book, we have used the mean of sampled values after burn-in as the Bayesian point estimate of π. Possible alternative estimates of π are the median and the mode of the sampled values after burn-in. Explain how the last statement in the code roughly approximates the mode.
b) To verify how well kernel density estimation works in one example, do the following: Generate 50 000 observations from $\mathsf{BETA}(2, 3)$, make a histogram of these observations, superimpose a kernel density-estimated curve in one color, and finally superimpose the true density function of $\mathsf{BETA}(2, 3)$ as a dotted curve in different color. Also, find the estimated mode and compare it to the exact mode 1/3 of this distribution.

9.7 So far as is known, a very large herd of livestock is entirely free of a certain disease ($\pi = 0$). However, in a recent routine random sample of $n = 100$ of these animals, two have tested positive on a screening test with sensitivity 95% and specificity 98%. One "expert" argues that the two positive tests warrant slaughtering all of the animals in the herd. Based on the specificity of the test, another "expert" argues that seeing two positive tests out of 100 is just what one would expect by chance in a disease-free herd, and so mass slaughter is not warranted by the evidence.

a) Use a Gibbs Sampler with a flat prior to make a one-sided 95% probability interval that puts an upper bound on the prevalence. Based on this result, what recommendation might you make?
b) How does the posterior mean compare with the estimate from equation (9.1) on p212?
c) Explain what it means to believe the prior $\mathsf{BETA}(1, 40)$. Would your recommendation in part (a) change if you believed this prior?
d) What Bayesian estimates would you get with the prior of part (c) if there are no test-positive animals among 100? In this case what part of the Gibbs Sampling process becomes deterministic?

Comments: In (a) and (b), the Bayesian point estimate and the estimate from equation (9.1) are about the same. If there are a few thousand animals in the herd, these results indicate there might indeed be at least one infected animal. Then, if the disease is one that may be highly contagious beyond the herd or if diseased animals pose a danger to humans, we could be in for serious trouble. If possible, first steps

might be to quarantine this herd for now, find the two animals that tested positive, and quickly subject them to a gold standard diagnostic test for the disease. That would provide more reliable information than does the Gibbs Sampler based on the screening test results. d) Used alone, a screening test with $\eta = 95\%$ and $\theta = 98\%$ applied to a relatively small proportion of the a herd seems a very blunt instrument for trying to say whether the herd is free of a disease.

Problems for Section 9.2, Estimating Normal Mean and Variance

9.8 Write and execute R code to make diagnostic graphs for the Gibbs Sampler of Example 9.2 showing ACFs and traces (similar to the plots in Figure 9.2). Comment on the results.

9.9 Run the code below. Explain step-by-step what each line (beyond the first) computes. How do you account for the difference between `diff(a)` and `diff(b)`?

```
x.bar = 9.60;  x.sd = 2.73;  n = 41
x.bar + qt(c(.025, .975), n-1)*x.sd/sqrt(n)
a = sqrt((n-1)*x.sd^2 / qchisq(c(.975,.025), n-1));  a;  diff(a)
b = sqrt((n-1)*x.sd^2 / qchisq(c(.98,.03), n-1));  b;  diff(b)
```

9.10 Suppose we have $n = 5$ observations from a normal population that can be summarized as $\bar{x} = 28.31$ and $s = 5.234$.

a) Use traditional methods based on Student's t and chi-squared distributions to find 95% confidence intervals for μ and σ.
b) In the notation of Example 9.2, use prior distributions with parameters $\mu_0 = 25$, $\sigma_0 = \sqrt{\theta_0} = 2$, $\alpha_0 = 30$, and $\kappa_0 = 1000$, and use a Gibbs Sampler to find 95% Bayesian interval estimates for μ and σ. Discuss the priors. Make diagnostic plots. Compare with the results of part (a), and comment.
c) Repeat part (b), but with $\mu_0 = 0$, $\sigma_0 = 1000$, $\alpha_0 = 0.01$, and $\kappa_0 = 0.01$. Compare with the results of parts (a) and (b), and comment.

Hints: In (a)-(c), the sample size is small, so an informative prior is influential. In (a) and (c): (21.8, 34.8) for μ; (3, 15) for σ. Roughly.

9.11 Before drawing inferences, one should always look at the data to see whether assumptions are met. The vector `x` in the code below contains the $n = 41$ observations summarized in Example 9.2.

```
x = c( 8.50,  9.75,  9.75,  6.00,  4.00, 10.75,  9.25, 13.25,
      10.50, 12.00, 11.25, 14.50, 12.75,  9.25, 11.00, 11.00,
       8.75,  5.75,  9.25, 11.50, 11.75,  7.75,  7.25, 10.75,
       7.00,  8.00, 13.75,  5.50,  8.25,  8.75, 10.25, 12.50,
       4.50, 10.75,  6.75, 13.25, 14.75,  9.00,  6.25, 11.75,  6.25)
mean(x)
var(x)
shapiro.test(x)
```

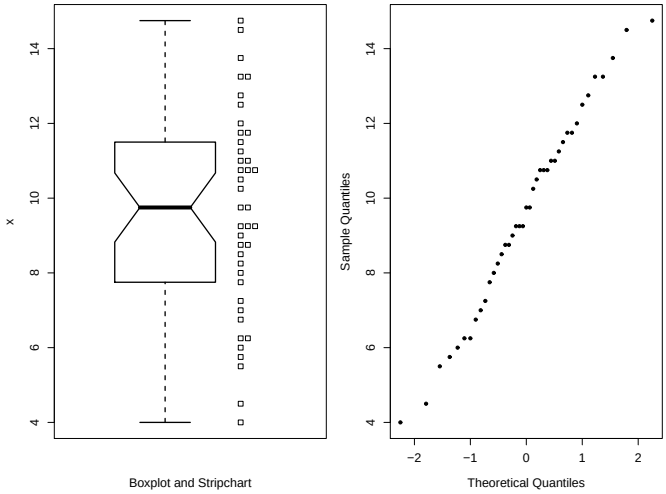

Figure 9.11. A boxplot, stripchart, and normal quantile plot of the height differences in Example 9.2. The overall impression is the data are consistent with the assumption of a normally distributed population. See Problem 9.11.

```
par(mfrow=c(1,2))
  boxplot(x, at=.9, notch=T, ylab="x",
    xlab = "Boxplot and Stripchart")
  stripchart(x, vert=T, method="stack", add=T, offset=.75, at = 1.2)
  qqnorm(x)
par(mfrow=c(1,1))
```

a) Describe briefly what each statement in the code does.
b) Comment on the graphical output in Figure 9.11. (The angular sides of the box in the boxplot, called **notches**, indicate a nonparametric confidence interval for the population median.) Also comment on the result of the test. Give several reasons why it is reasonable to assume these data come from a normal population.

Note: Data are from [MR58], also listed and discussed in [Rao89] and [Tru02]. Each data value in `x` is the difference between a morning and an evening height value. Each height value is the average of four measurements on the same subject.

9.12 Modify the code for the Gibbs Sampler of Example 9.2 as follows to reverse the order of the two key sampling steps at each passage through the loop. Use the starting value `MU[1]`$= 5$. At each step i, first generate `THETA[i]` from the data, the prior on θ, and the value `MU[i-1]`. Then generate `MU[i]` from the data, the prior on μ, and the value `THETA[i]`. Compare your results with those in the example, and comment.

9.13 (Theoretical) In Example 9.2, the prior distribution of the parameter $\theta = \sigma^2$ is of the form $\theta \sim \mathsf{IG}(\alpha_0, \kappa_0)$ so that $p(\theta) \propto \theta^{-(\alpha_0+1)} \exp(-\kappa_0/\theta)$. Also

the data $\mathbf{x}$ are normal with x_i randomly sampled from $\mathsf{NORM}(\mu, \sigma)$, so that the likelihood function is

$$p(\mathbf{x}|\mu, \theta) \propto \theta^{n/2} \exp\left\{-\frac{1}{2\theta}\sum_{i=1}^{n}(x_i - \mu)^2\right\}.$$

a) By subtracting and adding $\bar{x}$, show that the exponential in the likelihood function can be written as $\exp\{-\frac{1}{2\theta}\left[(n-1)s^2 + n(\bar{x}-\mu)^2\right]\}$.

b) The distribution of $\theta|\mathbf{x}, \mu$ used in the Gibbs Sampler is based on the product $p(\theta|\mathbf{x}, \mu) \propto p(\theta)\, p(\mathbf{x}|\mu, \theta)$. Expand and then simplify this product to verify that $\theta|\mathbf{x}, \mu \sim \mathsf{IG}(\alpha_n, \kappa_n)$, where α_n and κ_n are as defined in the example.

Problems for Section 9.3, Estimating Variance Components

9.14 The R code below was used to generate the data used in Example 9.3. If you run the code using the same (default) random number generator in R we used and the seed shown, you will get the same data.

```
set.seed(1212)
g = 12                                      # number of batches
r = 10                                      # replications per batch
mu = 100;  sg.a = 15;  sg.e = 9             # model parameters
a.dat = matrix(rnorm(g, 0, sg.a), nrow=g, ncol=r)
          # ith batch effect across ith row
e.dat = matrix(rnorm(g*r, 0, sg.e), nrow=g, ncol=r)
          # g x r random item variations
X = round(mu + a.dat + e.dat)               # integer data
X
```

```
> X
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,]  103  113   88   96   89   88   80   92   89    81
 [2,]  143  116  126  127  132  121  129  148  129   119
 [3,]  107  107   98  103  113  104   99  103   98   109
 [4,]   71   72   89   63   85   71   75   76   98    57
 [5,]  105  101  113  110  109  101  114  114  113   107
 [6,]   88   93  100   91   98  105  103   91  123   110
 [7,]   71   52   67   59   67   67   60   68   62    53
 [8,]  115  102   93  111  130  114   97  103  112    98
 [9,]   58   70   65   78   67   60   74   80   47    68
[10,]  133  119  130  136  133  116  131  118  140   135
[11,]  103  101   97  110  125  107  115  106  110    94
[12,]   83  106   86   91   88  107   92   98   88    95
```

a) Run the code and verify whether you get the same data. Explain the results of the statements `a.dat`, `var(a.dat[1,])`, `var(a.dat[,1])`, and `var(as.vector(e.dat))`. How do the results of the first and the second statements arise? What theoretical values are approximated (not very well because of the small sample size) by the last two statements.

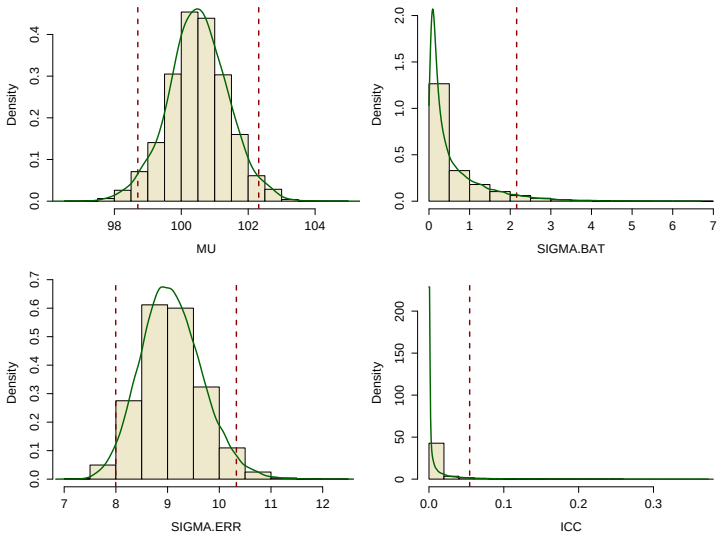

Figure 9.12. Histograms of simulated posteriors with estimated posterior densities for Problem 9.17. Results are from the vectors `MU`, `SIGMA.BAT`, `SIGMA.ERR`, and `ICC` of the Gibbs Sampler after burn-in. Indications of the 95% Bayesian interval estimate (one-sided for `SIGMA.BAT` and `ICC`). Compare with Figure 9.7 on page 225.

b) Explain why the following additional code computes MS(Batch) and MS(Error). How would you use these quantities to find the unbiased estimate of θ_A shown in the example?

```
X.bar = apply(X, 1, mean);  X.sd = apply(X, 1, sd)
MS.Bat = r*var(X.bar);  MS.Err = mean(X.sd^2)
```

Hints: a) By default, matrices are filled by columns; shorter vectors recycle. The variance components of the model are estimated.

9.15 (Continuation of Problem 9.14) Computation and derivation of frequentist confidence intervals related to Example 9.3.

a) The code below shows how to find the 95% confidence intervals for μ, θ, and ρ_I based on information in Problem 9.14 and Example 9.3.

```
mean(X.bar) + qt(c(.025,.975), g-1)*sqrt(MS.Bat/(g*r))
df.Err*MS.Err/qchisq(c(.975,.025), df.Err)
R = MS.Bat/MS.Err;  q.f = qf(c(.975,.025), g-1, g*r-g)
(R - q.f)/(R + (r-1)*q.f)
```

b) (Intermediate) Derive the confidence intervals in part (b) from the distributions of the quantities involved.

Hint: b) For ρ_I, start by deriving a confidence interval for $\psi = \theta_A/\theta$. What multiple of R is distributed $\mathsf{F}(g-1, g(r-1))$?

9.16 Figure 9.8 on page 227 shows four diagnostic plots for the simulated posterior distribution of σ_A in the Gibbs Sampler of Example 9.3. Make similar diagnostic plots for the posterior distributions of μ, σ, and ρ_I.

9.17 *Small contribution of batches to the overall variance.* Suppose the researchers who did the experiment in Example 9.3 find a way to reduce the batch component of variance. For the commercial purpose at hand, that would be important progress. But when they try to analyze a second experiment, there is a good chance that standard frequentist analysis will run into trouble. The code below is essentially the same as in Problem 9.14, but with the parameters and the seed changed. Group means and standard deviations, sufficient for running the Gibbs sampler of Example 9.3 are shown as output.

```
set.seed(1237)
g = 12;  r = 10
mu = 100;  sg.a = 1;  sg.e = 9
a.dat = matrix(rnorm(g, 0, sg.a), nrow=g, ncol=r)
e.dat = matrix(rnorm(g*r, 0, sg.e), nrow=g, ncol=r)
X = round(mu + a.dat + e.dat)
X.bar = apply(X, 1, mean);  X.sd = apply(X, 1, sd)
round(rbind(X.bar, X.sd), 3)

> round(rbind(X.bar, X.sd), 3)
       [,1]    [,2]   [,3]    [,4]   [,5]   [,6]
X.bar 96.90 103.700 97.300 100.900 95.100 95.900
X.sd  11.77  12.781  8.693  10.418  6.244  7.505
        [,7]  [,8]   [,9]  [,10]  [,11]   [,12]
X.bar 94.900 99.00 98.200 98.200 98.700 102.400
X.sd   9.871 10.76 10.304  6.356 11.146   8.289
```

a) Figure 9.6 shows boxplots for each of the 12 batches simulated above. Compare with Figure 9.5 on page 222. How can you judge from these two figures that the batch component of variance is smaller here than in Example 9.3?

b) Run the Gibbs Sampler of Section 9.3 for these data using the same uninformative priors as shown in the code there. You should get Bayesian interval estimates for $\sigma_A = \sqrt{\theta_A}$ and ρ_I that cover the cover the values used to generate the data `X`. Also, you should get Bayesian interval estimates for μ and σ that cover their "known" values. See Figure 9.12.

9.18 *Continuation of Problem 9.17.* Negative estimates of θ_A and ρ_I.

a) Refer to results stated in Problems 9.14 and 9.15. Show that the unbiased estimate of θ_A is negative. Also, show that the 95% confidence interval for ρ_I includes negative values. Finally, find 95% confidence intervals for μ and $\sigma = \sqrt{\theta}$ and compare them with corresponding results in from the Gibbs Sampler in Problem 9.17.

b) Whenever $R = \text{MS(Batch)}/\text{MS(Error)} < 1$, the unbiased estimate $\hat{\theta}_A$ of θ_A is negative. When the batch component of variance is relatively

small, this has a good chance of occurring. Evaluate $P\{R < 1\}$ when $\sigma_A = 1$, $\sigma = 9$, $g = 12$, and $r = 10$, as in this problem.

c) The null hypothesis H_0: $\theta_A = 0$ is accepted (against H_1: $\theta_A > 0$) when R is smaller than the 95th quantile of the F distribution with $g - 1$ and $g(r - 1)$ degrees of freedom. Explain why this null hypothesis is always accepted when $\hat{\theta}_A < 0$.

Hints: b) Exceeds 1/2. c) The code `qf(.95, 11, 108)` gives a result exceeding 1.

9.19 Calcium concentration in turnip leaves (% dry weight) is assayed for four samples from each of four leaves. Consider leaves as "batches." The data are shown below as R code for the matrix `X` in the program of Example 9.3; that is, each row of `X` corresponds to a batch.

```
X = matrix(c(3.28, 3.09, 3.03, 3.03,
             3.52, 3.48, 3.38, 3.38,
             2.88, 2.80, 2.81, 2.76,
             3.34, 3.38, 3.23, 3.26), nrow=4, ncol=4, byrow=T)
```

a) Run the program, using the same noninformative prior distributions as specified there, to find 95% Bayesian interval estimates for μ, σ_A, σ, and ρ_I from these data.

b) Suppose the researchers have previous experience making calcium determinations from such leaves. While calcium content and variability from leaf to leaf can change from one crop to the next, they have observed that the standard deviation σ of measurements from the same leaf is usually between 0. 075 and 0.100. So instead of a flat prior for σ, they choose $\mathsf{IG}(\alpha_0 = 35, \lambda_0 = 0.25)$. In these circumstances, explain why this is a reasonable prior.

c) With flat priors for μ, and θ_A, but the prior of part (b) for θ, run the Gibbs Sampler to find 95% Bayesian interval estimates for μ, σ_A, σ, and ρ_I from the data given above. Compare with your answers in part (a), and comment.

Note: Data are from page 239 of [SC80]. The unbiased estimate of $\theta_A = \sigma_A^2$ is positive here. Estimation of σ_A by any method is problematic because there are so few batches.

9.20 In order to assess components of variance in the two-stage manufacture of a dye, researchers obtain measurements on five samples from each of six batches. The data are shown below as R code for the matrix `X` in the program of Example 9.3; that is, each row of `X` corresponds to a batch.

```
X = matrix(c(1545, 1440, 1440, 1520, 1580,
             1540, 1555, 1490, 1560, 1495,
             1595, 1550, 1605, 1510, 1560,
             1445, 1440, 1595, 1465, 1545,
             1595, 1630, 1515, 1635, 1625,
             1520, 1455, 1450, 1480, 1445), 6, 5, byrow=T)
```

a) Use these data to find unbiased point estimates of μ, σ_A, and σ. Also find 95% confidence intervals for μ, σ, and ρ_I (see Problem 9.15).
b) Use a Gibbs Sampler to find 95% Bayesian interval estimates for μ, σ_A, σ, and ρ_I from these data. Specify noninformative prior distributions as in Example 9.3. Make diagnostic plots.

Answers: b) Roughly: (1478, 1578) for μ; for (15, 115) for σ_A. See [BT73] for a discussion of these data, reported in [Dav57].

9.21 In order to assess components of variance in the two-stage manufacture of a kind of plastic, researchers obtain measurements on four samples from each of 22 batches. Computations show that MS(Error) = 23.394. Also, *sums* of the four measurements from each of the 22 batches are as follows:

```
218  182  177  174  208  186
206  192  187  154  208  176
196  179  181  158  158  198
160  178  148  194
```

a) Compute the batch means, and thus $\bar{x}_{\cdot\cdot}$ and MS(Batch). Use your results to find the unbiased point estimates of μ, θ_A, and θ.
b) Notice that the batch standard deviations s_i, $i = 1, \ldots, 12$, enter into the program of Example 9.3 only as $\sum_i (r-1)s_i^2$. Make minor changes in the program so that you can use the information provided to find 90% Bayesian interval estimates of μ, σ_A, σ, and ρ_I, based on the same noninformative prior distributions as in the example.

Note: Data are reported in [Bro65], page 325. Along with other inferences from these data, the following traditional 90% confidence intervals are given there: $(43.9, 47.4)$ for μ; $(17.95, 31.97)$ for θ; and $(0.32, 1.62)$ for $\psi = \theta_A/\theta$. (See Problem 9.15.)

9.22 *Using the correct model.* To assess the variability of a process for making a pharmaceutical drug, measurements of potency were made on one pill from each of 50 bottles. These results are entered into a spreadsheet as 10 rows of 5 observations each. Row means and standard deviations are shown below.

Row	1	2	3	4	5	6	7	8	9	10
Mean	124.2	127.8	119.4	123.4	110.6	130.4	128.4	127.6	122.0	124.4
SD	10.57	14.89	11.55	10.14	12.82	9.99	12.97	12.82	16.72	8.53

a) Understanding from a telephone conversation with the researchers that the rows correspond to different batches of the drug made on different days, a statistician uses the Gibbs Sampler of Example 9.3 to analyze the data. Perform this analysis for yourself.
b) The truth is that all 50 observations come from the same batch. Recording the data in the spreadsheet by rows was just someone's idea of a convenience. So the data would properly be analyzed without regard to bogus "batches" according to a Gibbs Sampler as in Example 9.2. (Of course, this requires summarizing the data in a different way. Use

$s^2 = [9\text{MS(Batch)} + 40\text{MS(Error)}]/49$, where s is the standard deviation of all 50 observations.) Perform this analysis, compare with the results of part (a), and comment.

Note: Essentially a true story, but with data simulated from $\mathsf{NORM}(125, 12)$ replacing unavailable original data. The most important "prior" of all is to get the model right.