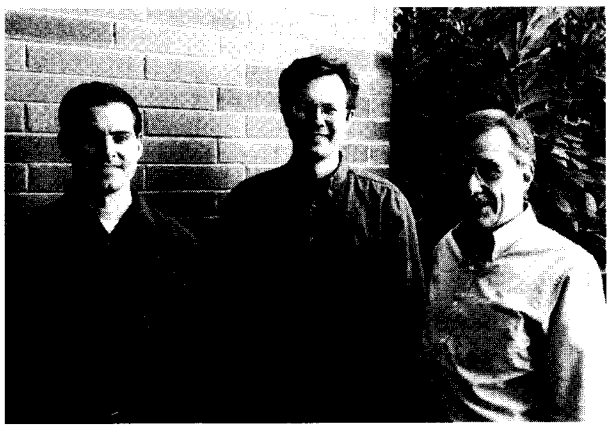
# Elementary Uses of the Gibbs Sampler: Applications to Medical Screening Tests



**Fraser, Suess, Trumbo**

Eric A. Suess, Christopher M. Fraser,
and Bruce E. Trumbo

Advances in computer technology have made a new generation of computationally intensive statistical methods available to applied statisticians. One of the most important of these new methods is the Gibbs sampler. Usually within a framework of Bayesian inference, it uses computer simulations and the convergence theory of discrete-time Markov processes to obtain results that would be difficult to obtain otherwise. [5, 12]

The purpose of this article is to show how Gibbs sampling methods can be used to make estimates of disease prevalence from medical screening test data. We include examples for which the theory can be understood at the undergraduate level. In Sections 1–3 we cover some basic facts about screening tests and establish notation needed throughout. Sections 4 and 5 introduce a simple Gibbs sampler and establish its relationship to a two-state Markov Chain. The remaining sections include the Bayesian context (Section 6), the description (Section 8), and simulation results (Section 9) of a more general Gibbs sampling process.

## ■ 1. Diagnostic Tests

Suppose that international public health officials want to determine the prevalence of a particular virus in donated blood at several sites throughout the world. Also suppose that a relatively inexpensive test is available to screen units of blood for this virus—an ELISA test. (ELISA stands for *enzyme-linked immunosorbent assay*. Specific ELISA tests detect antibodies to particular viruses, such as HIV, various types of hepatitis, etc.) Accordingly, the study will be based on the results of ELISA tests performed on randomly chosen units of blood donated at each place to be surveyed.

The proportion of ELISA tests indicating presence of the virus is not the same as the proportion of the sample actually contaminated with it. The ELISA test is useful, but not perfect.

**Sensitivity:** During its development, this ELISA test was performed on a large number of blood samples known to have come from subjects infected with the virus. Suppose that about 99% of these showed a positive result. That is to say, the ELISA test correctly detects the virus in 99% of infected units of blood. In terms of random variables and probabilities, we say that the *sensitivity* of the test is

$$\eta = \text{P(positive test | has virus)} = \text{P}(T=1|D=1) = 99\%.$$

Here $T$ is a random variable that takes the value 1 when a unit of blood shows a positive test, and $D$ is a random variable that takes the value 1 when the unit has the disease virus. (In this article we often express probabilities as percentages.)

**Specificity:** On the other hand, consider a group of units of blood known, from more costly, more accurate procedures than ELISA, to be free of the virus ($D = 0$). When administered to such units of blood, the ELISA test was found to give negative results ($T = 0$) for about 97% of them. That is, for some reason, ELISA incorrectly gave an indication of the virus in 3% of uncontaminated units of blood

*Eric Suess joined the faculty at California State University, Hayward in fall 1998 upon completion of his Ph.D. thesis at the University of California, Davis. Chris Fraser is a student in the MS program at CSU Hayward.* *The concept of this paper grew out of a student project funded by his grant from the CSU Hayward Associated Students.* *Bruce Trumbo is an ASA Fellow and the graduate advisor for the Statistics Department at CSU Hayward.*

(called "false-positive" results). We say that the *specificity* of the test is

$\theta$ = P(negative test | no virus)
= P($T$=0|$D$=0) = 97%.

**Hypothetical values:** The particular numerical values of $\eta$ and $\theta$ that we have given above, and continue to use throughout this article, are reasonable, but hypothetical values.[1] (See [4] for a discussion of sensitivity and specificity in several kinds of screening tests.)

Later in this paper we give several hypothetical prevalence values. In real life, actual prevalences range widely depending on the population and the disease. For example, in the US the prevalence of HIV in the donated blood supply is now essentially 0. (Pre-donation questionnaires used by blood banks tend to eliminate even donors likely to produce *false*-positive results with highly sensitive screening tests.) On the other hand, in clinical applications, screening tests are sometimes used where the prevalence of a disease exceeds 50%.

## ■ 2. First Attempts to Estimate Prevalence

At one of the sites under study, suppose that we estimate $\tau$ = P($T$=1) as $t$, the proportion of positive tests in a sample. As we have seen in Section 1, we cannot use $t$ directly as an estimate of the prevalence $\pi$ = P($D$=1). But can we somehow use $t$ indirectly to find an estimate $p$ of $\pi$?

One proposed method is to use the fact that $\tau$ and $\pi$ are related by the equation

$$\begin{aligned}\tau &= P(T=1) = P(D=1,\ T=1) + P(D=0,\ T=1)\\ &= P(D=1)P(T=1|D=1) + P(D=0)P(T=1|D=0)\\ &= \pi\eta + \pi^*\theta^*,\end{aligned}$$

where we write $\pi^* = 1 - \pi$ and $\theta^* = 1 - \theta$ (and similarly below for other Greek letters representing probabilities). Here we have partitioned all positive tests into true positives and false positives, applied the law of total probability, and used the definition of conditional probability. Solving this equation for $\pi$ and replacing $\tau$ by $t$, we obtain the estimate

$p = (t - \theta^*)/(\eta - \theta^*)$.

For example, suppose that we have a sample of $N$ = 1000 units of blood and that 49 of them test positive, that is, $A$ = #($T$=1) = 49. Then $t = A/N$ = 0.049 = 4.9%, and

$p$ = (4.9% - 3%)/(99% - 3%) = 1.98%.

The 95% confidence interval for $\pi$ based on the normal approximation is (3.56%, 6.24%), and the corresponding 95% confidence interval for $\pi$ is (0.58%, 3.38%).

Unfortunately, this method sometimes gives absurd estimates $p$ of $\pi$. For example, if we have a sample of $N$ = 215 units of blood and five of them test positive, then $t$ = 2.3% and $p$ = -0.73%. The problem here is that we expect 3% of the tests to be positive even if the prevalence is 0, but sampling variation has given us a value of $t$ less than 3%. In some applications, such estimates of prevalence that stray into negative territory can be quite common (see [8], pp. 130–132). If $\pi$ is very near 0, then $p$ will be negative about half the time. In different circumstances, this method can give absurd estimates of prevalence that exceed 100%.

## ■ 3. Additional Conditional Information

One possible path towards a better way to estimate the prevalence $\pi$ of infection is to perform a "gold standard" procedure on *some* of the units of blood. In concept, a gold standard provides an essentially 100% accurate determination as to whether or not the virus is present in a unit, but at a cost of administration that prevents its use on every unit of blood.[2] (If such a gold standard were inexpensive, why bother with imperfect ELISA tests for screening?)

As a hypothetical example, suppose that the unknown prevalence at a particular location is $\pi$ = 2%. Then, from the formula for $\tau$ in Section 2,

$\tau = \pi\eta + \pi^*\theta^*$
= (0.02)(0.99) + (0.98)(0.03) = 4.92%.

A common illustration of Bayes' theorem in basic probability texts is to compute

$\gamma$ = P($D$=1|$T$=1) = $\pi\eta/\tau$
= 0.0198 / 0.0492 = 40.24%.

This quantity $\gamma$ is called the *predictive value of a positive test.* Similarly, we compute

$\delta$ = P($D$=0|$T$=0) = $\pi^*\theta/\tau^*$
= 0.9506 / (1 - 0.0492) = 99.98%,

the *predictive value of a negative test.*

Unless the prevalence of the virus is of profoundly epidemic proportions at a particular location, the actual number of units with ELISA-*positive* tests found there may be small enough that we could check them all against the gold standard. Without knowing $\pi$, we could then estimate $\gamma$ for this site as #($T$=1, $D$=1)/#($T$=1); that is, the proportion of the ELISA-positive units proved by subsequent gold-standard procedures actually to have the virus.

Although we would not ordinarily be able to apply the gold standard to all units of blood that tested ELISA-*negative*, we might be able to check some of them against the gold standard to get an estimate of $\delta$ (if only to verify that $\delta$ really is very nearly 1, as would be the case if the prevalence is, say, below 5%).

If they were available, we shall see that reliable estimates of the conditional probabilities $\gamma$ and $\delta$ would provide the basis for an improved estimate of $\pi$. In the next section this possibility also provides a simple illustration of the important estimation technique known as the Gibbs sampler.

## ■ 4. The Gibbs Sampler—A Simple Discrete Case

If we knew the joint distribution of the random variables $D$ and $T$, then we could find the prevalence directly as the marginal probability

$$\pi = P(D=1)$$
$$= P(D=1, T=1) + P(D=1, T=0).$$

In real-world uses of screening tests, however, we often have no direct information about the joint distribution of $D$ and $T$. Instead, we may know (or have reasonable estimates of) the two conditional distributions $T \mid D$ and $D \mid T$. We have seen that the distribution of $T \mid D$ is determined by $\eta$ and $\theta$, and that the distribution of $D \mid T$ is determined by $\gamma$ and $\delta$. Based on this conditional information, the Gibbs sampler can be used to estimate prevalence by means of simulation.

**Step 1**. Here is how the simulation procedure works. Begin at step $m = 1$ with an arbitrary initial value of $D$, say, $D(1) = 1$.

**Step 2**. At step $m = 2$, condition on this value of $D(1)$ to simulate whether $T(1)$ is 1 or 0. That is, simulate $T(1) = 1$ with probability $\eta$ or $T(1) = 0$ with probability $\eta^*$. (Alternatively, if we had begun with $D(1) = 0$, then simulate using $\theta^*$ or $\theta$.)[3]

Next, simulate the value of $D(2)$ using information from $\gamma$ or $\delta$, as appropriate. For example, if we happened to get $T(1) = 1$, then we would simulate $D(2) = 1$ with probability $\gamma = P\{D(2)=1|T(1)=1\}$.

**Step 3.** In turn, at step $m = 3$, simulate $T(2)$ using either $\eta$ or $\theta$, and then simulate $D(3)$ using $\gamma$ or $\delta$, depending on the value of $T(2)$.

Notice that at step $m = 2$ we started with a value of $D(1)$ and went via a simulated value of $T(1)$ to obtain a simulated value of $D(2)$. At step $m = 3$ we start with this value of $D(2)$ to obtain a simulated value of $D(3)$.

**Iteration:** It can be shown that, upon iteration, this simulation process will stabilize to a limit. So repeat it to obtain $D(1)$, $D(2)$, ..., $D(M_1)$ during a "burn-in" period long enough to achieve stability, and then continue on to obtain values $D(M_1+1)$, ..., $D(M_2)$ at "steady state." Finally, we estimate $\pi$ as the proportion of steps at steady state for which $D(m) = 1$.

**Simulation results:** Throughout this article we are assuming that $\eta = 99\%$ and $\theta = 97\%$ for our ELISA screening test. At a particular site, suppose that reliable estimates of the other conditional probabilities are $\gamma = 40.24\%$ and $\delta = 99.98\%$. Based on these four numbers, we used an S-plus program to simulate 20 runs with $M_1 = 50{,}000$ and $M_2 = 100{,}000$. (See Appendix $D$ of [1] for the main simulation loop of our S-plus program implementing this procedure. The complete program is also available [1].)

Ten of these runs started with $D(1) = 0$ and ten with $D(1) = 1$. Then we estimated $\pi$ for each run from the last 50,000 values of $D$. Table 1 shows the results. In these circumstances, the results clearly show that $\pi$ must be near 2%.

**Table 1. Prevalence Estimated from 20 Simulation Runs (each based on 50,000 iterations at steady state)**

| | Initial Value | |
|---|---|---|
| | $D(1)=0$ | $D(1)=1$ |
| | 0.0204 0.0201 | 0.0216 0.0187 |
| | 0.0210 0.0218 | 0.0184 0.0192 |
| | 0.0212 0.0185 | 0.0200 0.0180 |
| | 0.0184 0.0194 | 0.0206 0.0222 |
| | 0.0194 0.0199 | 0.0210 0.0194 |
| **Mean** | 0.0200=2.00% | 0.0199=1.99% |
| **StDev** | 0.0011 | 0.0014 |

It would have been futile to try to simulate $\pi = P(D=1)$ directly from the conditional distributions of $D \mid T$. The probability $\gamma = P(D=1|T=1)$ is much too large and $\delta^* = P(D=1|T=0)$ is much too small. However, as the simulation goes through the process described above, the conditional distributions of $T \mid D$ come into play to ensure that in the long run each of these cases is simulated the appropriate proportion of the time.

**Questions:** Before we move on, it is worthwhile to ask several questions about the simulation process.

*First,* did the starting value, 0 or 1, of $D(1)$ make any difference in the values of $\pi$ obtained? Our results in Table 1 show that the starting value makes no significant difference.

*Second,* were our simulation runs long enough? The similarity of the results in Table 1 shows that our values of $M_1$ and $M_2$ were large enough to give reproducible results, but longer runs would have further reduced their variability. The running proportion of $D$'s taking the value 1 (up to each stage $m$) is plotted in Figure 1 for one of our runs. The stability shown there is typical of all our runs. (This particular example requires relatively long simulation runs to achieve satisfactory results because $\delta$ is nearly 1.)

*Third,* why did this converging process always settle down to values of $\pi \approx 2\%$? Although the 2%

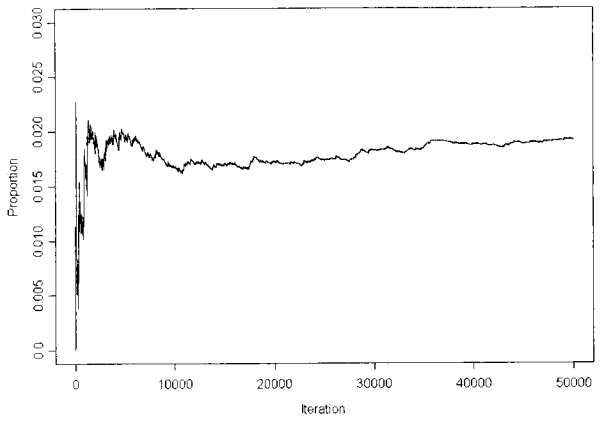



**Figure 1.** Running proportion of infected samples.

value of $\pi$ was not an explicit input value, it is consistent with the values of $\gamma$ and $\delta$ we assumed for our hypothetical site. (Recall the computations in Section 3.) The simulation has reclaimed this 2% value for $\pi$ from the conditional information we provided. In the next section we will see the theoretical basis for this.

## ■ 5. A Markov Chain

How do we know for sure that a Gibbs sampler ever stabilizes, and what determines the limiting value of $\pi$? In our simple example we can show that the $D$'s form a two-state Markov chain known to have a limiting distribution. The key equation to this end uses the conditional distributions $T \mid D$ and $D \mid T$ to find the transition probabilities of this Markov chain:

$$\begin{aligned} &P\{D(m+1)=j|D(m)=i\} \\ &= \Sigma_k\, P\{T(m)=k,\, D(m+1)=j|D(m)=i\} \\ &= \Sigma_k\, P\{D(m+1)=j|T(m)=k\}\, P\{T(m)=k|D(m)=i\}, \end{aligned}$$

for $m = 1, 2, 3, \ldots$ and $i, j, k = 0, 1$. The last step above uses the definition of conditional probability (twice) and the Markov property

$$\begin{aligned} &P\{D(m+1)=j|D(m)=i,\, T(m)=k\} \\ &= P\{D(m+1)=j|T(m)=k\}, \end{aligned}$$

which holds because the earlier condition $D(m) = i$ is irrelevant once we are given the later information that $T(m) = k$.

With the same values of $\eta$, $\theta$, $\gamma$ and $\delta$ as in Section 4, this equation is equivalent to the matrix equation

$$\mathbf{P} = \mathbf{QR} = \begin{bmatrix} \theta & \theta^* \\ \eta^* & \eta \end{bmatrix} \begin{bmatrix} \delta & \delta^* \\ \gamma^* & \gamma \end{bmatrix}$$

$$\mathbf{P} = \begin{bmatrix} 0.97 & 0.03 \\ 0.01 & 0.99 \end{bmatrix} \begin{bmatrix} 0.9998 & 0.0002 \\ 0.5976 & 0.4024 \end{bmatrix} = \begin{bmatrix} 0.9877 & 0.0123 \\ 0.6016 & 0.3984 \end{bmatrix}$$

**P** is the transition matrix of the Markov chain $D(1)$, $D(2)$, ..., which has state space {0, 1}. For example, the upper-left element is

$$\begin{aligned} p_{00} &= P\{D(m+1)=0|D(m)=0\} = \theta\delta + \theta^*\gamma^* \\ &= 0.9877. \end{aligned}$$

Because **P** has all positive elements, one can show that the chain must have a limiting distribution.

In our simple case it was not really necessary to use simulation to find the limiting distribution $\boldsymbol{\lambda}$ of this chain. It is the solution of the matrix equation $\boldsymbol{\lambda}\mathbf{P} = \boldsymbol{\lambda}$ which, upon solving two equations in two unknowns, is seen to be $\boldsymbol{\lambda} = (\pi^*, \pi) = (0.98, 0.02)$. The simulated values in Section 3 came quite close to the true value of $\pi$. (See Appendix A of [1] for some background information on the convergence of a two-state Markov chain to the solution of this matrix equation.)

## ■ 6. Gibbs Sampling When Predictive Values Are Unknown

We have just seen an example in which the Gibbs sampler works, but in which simulation is not really necessary because we can easily find an analytic solution. It is not difficult to imagine cases—just slightly more complicated than our example—where classical analytic solutions are difficult or impossible to find. In such cases the Gibbs sampler may be a useful statistical tool.

In our hypothetical example it may be realistic to suppose that the conditional probabilities $\eta$ and $\theta$ in the **Q**-matrix are known. We have assumed some prior experience in using the ELISA test to screen blood for the presence of the virus.

Not quite so realistic is the assumption that the conditional probabilities $\gamma$ and $\delta$ in the **R**-matrix would be known accurately. Each place to be surveyed would be different, so the predictive values of positive and negative tests would have to be estimated for each particular population of interest. Also, it may be too difficult or costly to obtain gold-standard determinations.

Specifically, suppose that for a particular site we know the numbers of units testing positive and negative, $A = \#(T=1)$ and $B = \#(T=0)$, respectively, where the sample size is $N = A + B$. But we do not know $X = \#(T=1, D=1)$ and $Y = \#(T=0, D=1)$. Here we no longer have the direct estimates $X/A$ of $\gamma$ and $(B - Y)/B$ of $\delta$. Table 2 shows the theoretical counts in a sample distribution of $D$ and $T$ with this notation, where the known data consist only of $A$ and $B$.

**Table 2. Joint Sample Distribution of *D* and *T* (all quantities are counts)**

| | Virus Present | | |
|---|---|---|---|
| Test | Yes (*D*=1) | No (*D*=0) | Total |
| Pos. (*T*=1) | *X* | *A-X* | *A* |
| Neg. (*T*=0) | *Y* | *B-Y* | *B* |
| Total | *X+Y* | *N-X-Y* | *N* |

Essentially, we are now back in the situation described in Section 2. In these circumstances the limit of the simulation process is no longer easy to find by traditional analytic methods. One reason is that this is an "under-specified" problem, in which we seem not to have quite enough information. Suppose that the sample size $N$ is known. Then the direct estimate of $\pi$ is $p = (X + Y)/N$, which requires us to know both $X$ and $Y$. But once $N$ is known, $A$ is the only *observed* value that provides information. In Sections 7 and 8 we will see that an approach using Bayesian estimation and the Gibbs sampler provides a useful estimate of $\pi$.

## ■ 7. A Bayesian Framework—Prior Distributions

In Bayesian statistical inference, population parameters are considered to be random variables. The distribution assigned to a parameter before data are collected is called a *prior* distribution. In many applications, expert opinion is used to pick the prior distribution. After data are available, the expert opinion in the prior is combined with the information in the data to find a *posterior* distribution of the parameter, which may be used to draw inferences about the parameter (for example, to find interval estimates). As the amount of data increases, the prior has a decreasing influence on the conclusions drawn. (See Appendix B of [1] for a very brief introduction to Bayesian inference and an example that does not involve screening tests; see the *STATS* article by Hal Stern [11] for a more extensive introduction.)

In the situation described in Section 6, it would be reasonable to take a Bayesian view in which the prevalence $\Pi$ is a random variable with a prior distribution based on the informed views of experts on blood bank epidemiology. In our case the level of expertise required is not great. For example, it would be enough to know that the true prevalence is likely to be much nearer to 0 than to 1 at a particular site.

Because $\Pi$ represents a probability that must take a value between 0 and 1, it is natural to use a prior distribution from the beta family. The shape of a beta density function is determined by two parameters $\alpha$ and $\beta$. Table 3 gives some information about four beta distributions that might be posed as candidates for a prior distribution. Priors (a) and (b) are two different ways to reflect the opinion that the prevalence is likely to be small, and (d) in contrast, reflects a judgment that the prevalence is considerably greater. (The values of the parameters $\alpha$ and $\beta$ were chosen to give desired values for the modes.) Prior (c), having a flat density with no mode, is "uninformative." That is, it represents a lack of definitive expert opinion about prevalence. Figure 2 shows plots of the density curves of these four choices. In general, we write $\Pi \sim$ Beta($\alpha$, $\beta$) for such beta prior distributions.

**Table 3. Four Prior Distributions for Prevalence**

| Description | $\alpha, \beta$ | Mode $(\alpha-1)/(\alpha+\beta-2)$ | Percentiles 2.5% | 97.5% |
|---|---|---|---|---|
| (a) Perhaps most realistic | 2,50 | 2% | 0.48% | 10.45% |
| (b) Somewhat more pessimistic | 2,20 | 5% | 1.17% | 23.82% |
| (c) Uniform, flat | 1,1 | none | 2.50% | 97.50% |
| (d) Questionable choice | 7,5 | 60% | 30.97% | 88.19% |

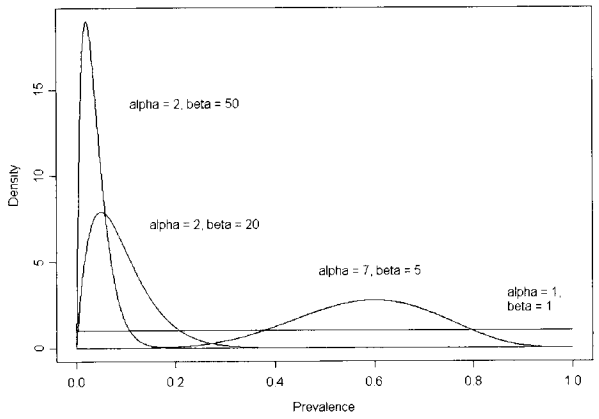



**Figure 2.** Beta prior distributions for prevalence.

Of course, experts may differ as to the choice of the prior distribution. One thing we hope for in a practical Bayesian application is to have an appropriate design and enough data so that the choice among reasonable priors has minimal influence on the conclusions we finally draw.

## ■ 8. Conditional Distributions and the Gibbs Sampler

Here we show how the Gibbs sampler works for the case where the predictive values of positive and negative test results are not known. This simulation process begins with two inputs: first, the beta prior distribution of $\Pi$ with parameters $\alpha$ and $\beta$ as discussed in the previous section, and second, the data, which consist only of the values $A$ and $B$ as shown in Table 2.

The relationships for our current situation that correspond to the key equation for the simpler situation of Section 5 are provided by the conditional distributions shown below. (Appendix C of [1] has details of the derivations.)

The conditional posterior distribution of $X$, given $A$ and $\pi$, is binomial with $A$ trials and success probability $\pi\eta/(\pi\eta + \pi^*\theta^*)$. We write this as

$$X \mid A, \pi \sim \text{Bino}[A, \pi\eta/(\pi\eta + \pi^*\theta^*)].$$

Similarly, the conditional posterior distribution of $Y$, given $B$ and $\pi$, is

$$Y \mid B, \pi \sim \text{Bino}[B, \pi\eta^*/(\pi\eta^* + \pi^*\theta)].$$

The Gibbs sampler begins (step $m = 1$) by fixing an initial value $\pi$ of $\Pi$. Next, for step $m = 2$, it uses this value $\pi$, the known data $A$ and $B$, and the parameter values $\eta$ and $\theta$ to simulate values of $X$ and $Y$. These are used to obtain the conditional posterior distribution of $\Pi$, updated to take account of the values of $A$, $B$, $X$, and $Y$:

$$\Pi \mid X, Y, A, B \sim \text{Beta}(X+Y+\alpha, A+B-X-Y+\beta).$$

The form and parameters of this conditional distribution follow from a general version of Bayes' theorem. (Again here, see Appendix C of [1].) To complete step $m = 2$ of the simulation, we generate a new value of $\pi$ from this conditional distribution $\Pi \mid X, Y, A, B$.

For step $m = 3$, we plug our new value of $\pi$ into the expressions for the probability of success for each of the above binomial distributions, and sample new values of $X$ and $Y$ from these distributions. In turn, these values of $X$ and $Y$ yield yet another conditional posterior distribution of $\Pi$, from which we sample a value of $\pi$ to be used in step $m = 4$, and so on.

Upon iteration, this process simulates a Markov chain $\Pi(1), \Pi(2), \ldots$, which has the limiting distribution $\Pi \mid A, B$. We "burn in" the chain for enough steps to ensure that it has stabilized to this limiting distribution. Then we estimate the prevalence from the sample distribution of additional values of $\pi$ generated from the chain at steady state. (The main loop of an S-plus program implementing this process is shown in Appendix D of [1]; for the complete program, also see [1].)

The discrete-valued Markov chain of Sections 4 and 5 took only the values $D = 0$ and 1. Thus, its transitional behavior could be described in a $2 \times 2$ matrix. By contrast, the Markov chain $\Pi(1), \Pi(2), \ldots$ of this section can take values throughout the interval [0, 1], so the conditional distributions earlier in this section must be used to describe its transitional behavior. These conditional distributions play the role of $\gamma$ and $\delta^*$ in Section 5. In fact, at each step, the success probabilities of the conditional binomial distributions of $X$ and $Y$ have values $\gamma$ and $\delta^*$ (computed by plugging in the current value $\pi$ of $\Pi$).

It can be shown that such a continuous-valued Markov chain reaches a limiting distribution under conditions broad enough to make the Gibbs sampler a practical method for applied statistics. [2, 7]

## ■ 9. Gibbs Estimates of Prevalence

Here we show the results of the Gibbs sampler outlined in Section 8 for screening tests in which the values of $\gamma$ and $\delta$ are unknown. We used the values $\eta = 99\%$, $\theta = 97\%$, and $t = 4.9\%$ based on a sample size $N = 1000$, so that $A = 49$ and $B = 951$. We found results for all four of the priors suggested in Section 4, using simulation runs of length $M_2 = 20{,}000$ in each case.

For prior (a), Figure 3 shows a histogram of the sampling distribution after a burn-in of 10,000 values of $\pi$. This histogram estimates the marginal posterior distribution $\Pi \mid A,B$. Accordingly, it is used to make the point estimates (mean and median) and the interval estimate (2.5% and 97.5.% points) shown in the first row of Table 4. The next four rows of Table 4 show very similar results obtained from additional runs using the same prior.

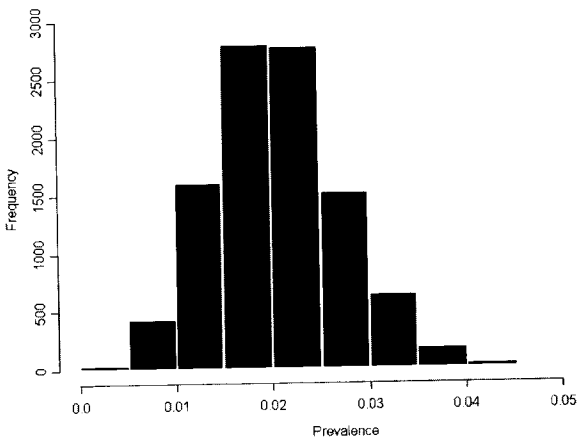



**Figure 3.** Sampled prevalence values (after burn-in).

The results of simulations with priors (b) and (c) are not much different. Prior (d), with a mode at 60%, gives estimates of prevalence that are only

about 1% higher than the others. In summary, we have enough data here that the choice of prior does not make a great deal of difference in the results.

Figures 4 and 5 for prior (a) are typical of diagnostic graphics used in Gibbs sampling to judge whether the process stabilizes by the end of the burn-in period. Figure 4 shows all 20,000 of the simulated values of $\pi$ plotted in sequence. It appears that they fluctuate about a single value with an almost-constant variability. Figure 5 shows that the running average of these simulated values stabilizes. Together these figures show that the process stabilized and that our burn-in period was long enough. (Our simulations using the other three priors were also well-behaved. [1])

**Table 4.** Gibbs Estimates of Prevalence From 10,000 Simulated Values After Burn-In (data: 49 positive tests among 1,000 units of blood)

| Beta Prior On Prevalence | Simulated Values of $\pi$ Mean | Median | Gibbs Interval Estimate (2.5%,97.5%) |
|---|---|---|---|
| (a) $\alpha$=2, $\beta$=50 | 2.05% | 2.03% | (0.85%, 3.45%) |
| | 2.09% | 2.06% | (0.92%, 3.45%) |
| | 2.08% | 2.05% | (0.85%, 3.52%) |
| | 2.07% | 2.04% | (0.87%, 3.47%) |
| | 2.07% | 2.04% | (0.86%, 3.46%) |
| (b) $\alpha$=2, $\beta$=20 | 2.18% | 2.15% | (0.94%, 3.61%) |
| (c) $\alpha$=1, $\beta$=1 | 2.05% | 2.02% | (0.75%, 3.56% |
| (d) $\alpha$=7, $\beta$=5 | 3.23% | 3.20% | (1.90%, 4.76%) |
| None: (Sec. 2) | 1.98% | | (0.58%, 3.38%) |

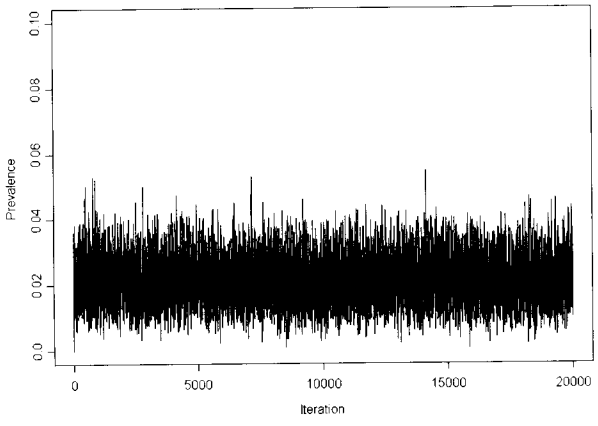



**Figure 4.** Sampled prevalence values in sequence.

We also looked at the results of the Gibbs sampler for the data $N$ = 215 and $A$ = 5, using the same values of sensitivity, specificity, run size, and burn-in as above. Table 5 summarizes our results for the four priors.

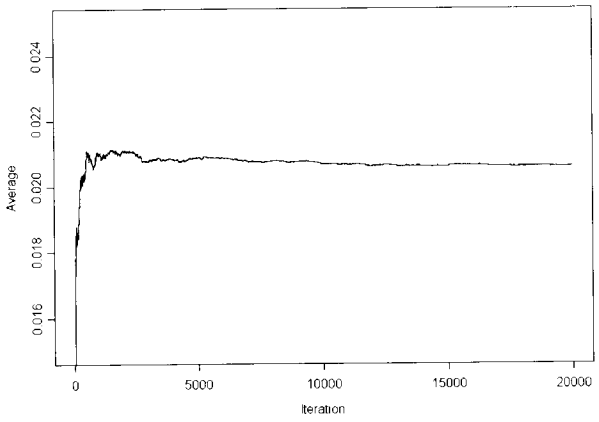



**Figure 5.** Running averages of sampled prevalences.

Clearly, the first three of these estimates are more sensible than the negative point estimate obtained in Section 2 from the same data. Here we have fewer observations ($N$ = 215), and the data we do have conflict somewhat with our assumed specificity (see Section 2). Therefore, the choice of prior has more influence on the results than in the example above ($N$ = 1000).

**Table 5.** Gibbs Estimates of Prevalence From 10,000 Simulated Values After Burn-In (data: 45 positive tests among 215 units of blood

| Beta Prior On Prevalence | Simulated Values of $\pi$ Mean | Median | Gibbs Interval Estimate (2.5%,97.5%) |
|---|---|---|---|
| (a) $\alpha$=2, $\beta$=50 | 1.27% | 1.11% | (0.18%, 3.16%) |
| (b) $\alpha$=2, $\beta$=20 | 2.51% | 1.34% | (0.21%, 3.74%) |
| (c) $\alpha$=1, $\beta$=1 | 0.95% | 0.75% | (0.03%, 3.19%) |
| (d) $\alpha$=7, $\beta$=5 | 4.42% | 4.28% | (1.99%, 7.61%) |

In particular, the results for prior (d) in Table 5 differ markedly from those obtained with the first three priors. Prior (d) declares values of $\pi$ less than 30% as quite unlikely and values near 0 as almost impossible (refer back to Table 3 and Figure 1), and we do not have enough data to totally overcome this declaration.

Unlike the estimate of Section 2, all of the estimates in Table 5 avoid negative values because the prior and conditional distributions of $\Pi$ do not allow negative values of $\pi$ at any step of the simulation.

## ■ 10. A Few Words of Caution

It would be a serious mistake to view the Gibbs sampler as some sort of statistical magic that can create new information. The only informational inputs to the Gibbs sampler are in the model, the prior, and the data. The simulation does not create

new information; rather it is a substitute for more traditional mathematical methods (algebra, integration, etc.) in converting the information we input to a possibly more meaningful form.

The examples we have chosen for this article are quite simple ones, and we hope they have helped you to understand what the Gibbs sampler is and how it works. But we do need to stress the importance of understanding—and using—the full array of diagnostic tools in every application of the Gibbs sampler. It is necessary to test whether there is a unique limit and whether the simulation has found it. Gibbs sampling is not a cure-all for every difficult problem of Bayesian inference. Because of its wide applicability to problems that are not feasible to treat by other means, it is inevitable that some attempts to use the Gibbs sampler will be unsuccessful. Diagnostic methods provide crucial warnings when that happens.

## ■ 11. More Information About Gibbs Sampling

So far we have used $\eta = 99\%$ and $\theta = 97\%$ as if they were precisely known values. For some applications of screening tests, the next step toward building a realistic probability model is to admit that these values are not known exactly. An extreme case occurs when there is no gold standard and knowledge about sensitivity and specificity is rather vague. This is the situation in [6], where data are presented for screening tests to detect intestinal parasites. To model the state of expert knowledge about $\eta$ and $\theta$, the investigators establish appropriate prior distributions. After reading our article, you should be able to follow the Gibbs sampling procedures used there to estimate prevalence.

See [1] for the complete annotated computer code for our simulations. We invite you to replicate our simulation results, to try different values of the screening test parameters, and to explore the effect of using different priors. Also, additional resources for the study of screening tests, Markov chains, and Gibbs sampling (including some exercises suitable for classroom use) are being accumulated in [1].

We have used S-plus in our illustrations of Gibbs samplers because this software is so widely available in colleges and universities. For many applications, it is more convenient to use software written specifically to support Gibbs sampling, for example [10].

A slightly more advanced treatment of Gibbs sampling than we present here can be found in [2], and a discussion of some successes and difficulties in using the Gibbs sampler can be found in [7].

## ■ Notes and References

Superscripts 1, 2, and 3 denote brief auxiliary comments currently available in the *Notes* section of [1]. Archival in [9] of selected items from [1] (for example, these Notes and Appendixes A–D) is anticipated.


[1] California State University, Hayward, Statistics Department web site, URL (case sensitive): www.telecom.csuhayward.edu/~stat/Gibbs

[2] Casella, G. and George, E., "Explaining the Gibbs sampler," *The American Statistician,* Vol. 46, No. 3, (August 1992), pages 167–174.

[3] Feller, W., *An Introduction to Probability Theory and Its Application, Vol. 1* (3rd ed.), 1950, Wiley, New York.

[4] Gastwirth, J.L., "The statistical precision of medical screening procedures: Applications to polygraph and AIDS antibody test data" (including discussion), *Statistical Science, Vol. 2,* No. 3 (1987), pages 213–238.

[5] Gelfand, A.E. and Smith, A.F.M., "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association,* Vol. 87 (1990), pages 398–409.

[6] Joseph, L., Gyorkos, T., and Coupal, L., "Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard," *American Journal of Epidemiology,* Vol. 141, No. 3 (1995), pages 263–271.

[7] Kass, R.E. (moderator), Carlin, B.P., Gelman, A., and Neal, R.M. (panelists), "Markov Chain Monte Carlo in practice: A roundtable discussion," *The American Statistician,* Vol. 52, No. 2 (May 1998), pages 93–100.

[8] Pagano, M. and Gauvreau, K., *Principles of Biostatistics,* 1993, Duxbury Press, Belmont, CA.

[9] STATS Web site, URL: www.amstat.org/STATS

[10] Spiegenthaler, D., Thomas, A., Best, N., and Gilks, W., "Bayesian inference using Gibbs sampling (BUGS)," MRC Biometrics Unit, Institute of Public Health, Cambridge, UK, 1997.URL:www.mrc-bsu.cam.ac.uk/bugs

[11] Stern, H., "A primer on the Bayesian approach to statistical inference," *STATS,* No. 23 (Winter 1998), pages 3–9.

[12] Tanner, M.A. and Wong, W., "The calculation of posterior distributions by data augmentation" (with discussion), *Journal of the American Statistical Association,* Vol. 82 (1987), pages 528–550.