Solutions Chapter 14 - Gibbs Sampling

Alexandra Hotti

December 2019

14.1 The sensitivity and specificity of a test for a disease

Problem 14.1.1

Write down an expression for the likelihood, supposing that the prevalence for the disease is π. (Hint: multiply together the likelihoods corresponding to each of the cells in Table 14.1).

Answer: First, the posterior can be expressed as:

$$p(Y_1, Y_2, \pi, S, C|a, b) \propto p(a, b|Y_1, Y_2, \pi, S, C)p(Y_1, Y_2, \pi, S, C) \tag{1}$$

So, to create a gibbs sampler we first need an expression for the joint likelihood.

$$p(a, b|Y_1, Y_2, \pi, S, C) = (\pi S)^{Y_1} \cdot (\pi(1-S))^{Y_2} \cdot ((1-\pi)(1-C))^{a-Y_1} \cdot ((1-\pi)C)^{b-Y_2} \tag{2}$$

0.1 Problem 14.1.2.

Using the above expressions code up a working Gibbs sampler.

Answer: Now, lets derive the conditionals for the individual parameters within the joint based on the likelihood in eq. 2 and the priors given in the assignment description.

$$Y_1|a, \pi, S, C \sim (\pi S)^{Y_1} \cdot ((1-\pi)(1-C))^{a-Y_1} \equiv Binomial(a, \frac{\pi S}{(1-\pi)(1-C)+\pi S}) \tag{3}$$

We drop the conditioning variables that Y_1 is independent of.

Note that Y_1 is a count of the number of true positive disease cases, S is the proportion of disease positive individuals that test positive and π is the proportion of individuals that have the disease. Also, πS gives us the probability of true positives and $(1-\pi)(1-C)$ the probability of a false positive result. Thus, θ is the ratio of positive tests that actually are positive. The denominator ensures that this probability sums to 1.

A similar process is carried out to get the expression for Y_2

$$Y_2|b, \pi, S, C \sim (\pi(1-S))^{Y_2} \cdot ((1-\pi)C)^{b-Y_2} \equiv Binomial(b, \frac{\pi(1-S)}{(1-\pi)C+\pi(1-S)}) \tag{4}$$

Note that $(1-S)$ is the proportion of disease positive individuals that test negative and π is the proportion of individuals that have the disease. Thus, $\pi(1-S)$ is the probability of having the disease and getting a negative test and $(1-\pi)C$ is the probability of not having the disease and getting a negative result.

$$\pi|a, b, Y_1, Y_2 \sim \pi^{Y_1}\pi^{Y_2}S^{Y_1}(1-S)^{Y_2}(1-\pi)^{b-Y_2}(1-\pi)^{a-Y_1}(1-C)^{a-Y1}(C)^{b-Y_2} \cdot Beta(\alpha_\pi, \beta_\pi) \propto$$

$$\pi^{Y_1+Y_2+\alpha_\pi}(1-\pi)^{b+a-Y_2-Y_1+\beta_\pi} \equiv Beta(Y_1+Y_2+\alpha_\pi, b+a-Y_2-Y_1+\beta_\pi) \tag{5}$$

Note that π is a probability and that it has a beta prior. Thus, we could have arrived here by using the fact that the beta prior is a conjugate prior to the binomial likelihood.

Here, we again have a beta prior and we get:

$$S|Y_1, Y_2 \sim Beta(Y_1+\alpha_s, Y_2+\beta_s)$$

$$C|Y_1, Y_2, a, b \propto ((1-\pi)(1-C))^{a-Y_2} \cdot ((1-\pi)C)^{b-Y_2} \cdot C^{\alpha_c-1} \cdot (1-C)^{\beta-1} \propto C^{b-Y_2+\alpha_c-1} \cdot (1-C)^{a-Y_2+\beta-1} \equiv$$

$$\equiv Beta(b - Y_2 + \alpha_c, a - Y_1 + \beta) \tag{6}$$

Problem 14.1.3.

Suppose that out of a sample of 100 people, 20 of those tested negative and 80 positive. Assuming uniform priors on π, S and C, use Gibbs sampling to generate posterior samples for π. What do you conclude?

See Figure 1.

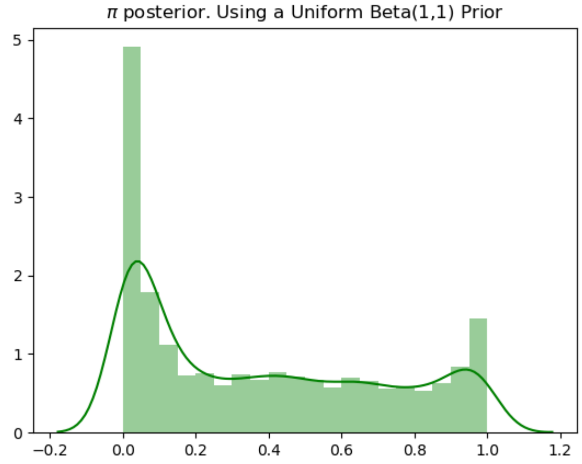

Figure 1: Posterior samples for π using uninformative priors for S and C.

Problem 14.1.4.

Suppose that a previous study that compares the clinical test with a laboratory gold standard concludes that $S \sim beta(10; 1)$ and $C \sim beta(10; 1)$ Use Gibbs sampling to estimate the new posterior for π. Why does this look different to your previously-estimated distribution?

See Figure 2. Since we have a reasonable knowledge of what the test sensitivity and specificity are, this means that we can now get a more concentrated estimate of what the disease prevalence is.

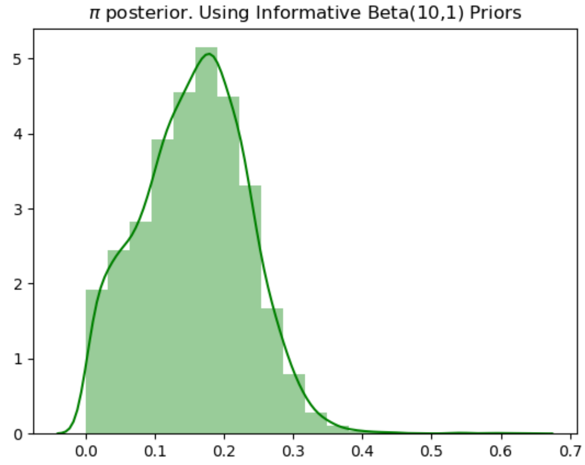

Figure 2: Posterior samples for π using informative priors for S and C.

Problem 14.1.5.

Suppose a previous analysis concluded that. $\pi \sim beta(1, 10)$. Using this distribution as a prior, together with uniform priors on S and C, determine the posterior distributions for the test sensitivity and specificity respectively. Why does the test appear to be quite specific, although it is unclear how sensitive it is?

See Figure 3 and 4. We have very little data about disease positive individuals. Thus, our posterior for S: which estimates ”the proportion of disease positive individuals that test positive” is still very uncertain. However, we have 80 data points were the patient got a negative result and thus we get a more certain posterior estimate for C.

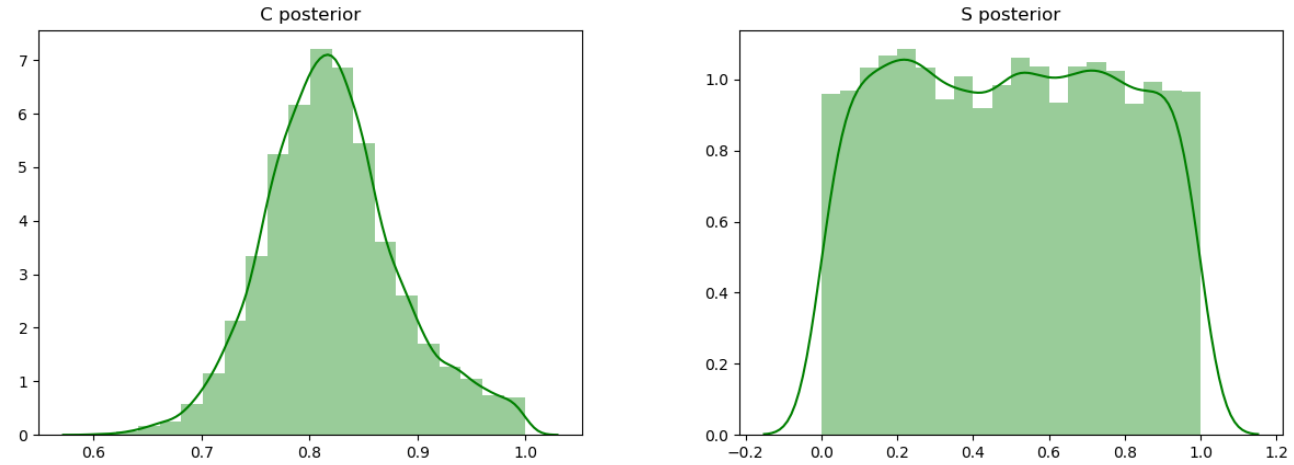

Figure 3: Posterior samples for right: test sensitivity (S), and left: specificity (C) assuming uniform priors on each of these parameters and $\pi \sim beta(1, 10)$.

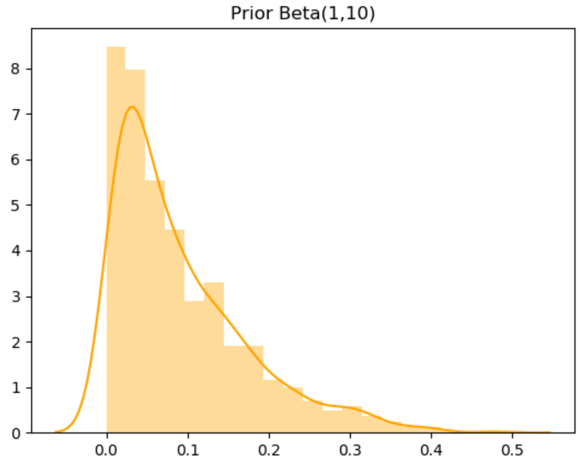

Figure 4: The Beta(1,10) Prior

Problem 14.1.6

Suppose that based on lab results you suppose that the test specificity $C \sim beta(10, 1)$, and $\pi \sim beta(1, 10)$, but the prior for S is still uniform. Explain the shape of the posterior for S now.

It is still pretty uninformative . Regardless of how specific our test is there are still a large number of ways we could find 20/100 people having the disease. Only some of those ways actually include the 10 people that likely have the disease out of the positively-testing sample!

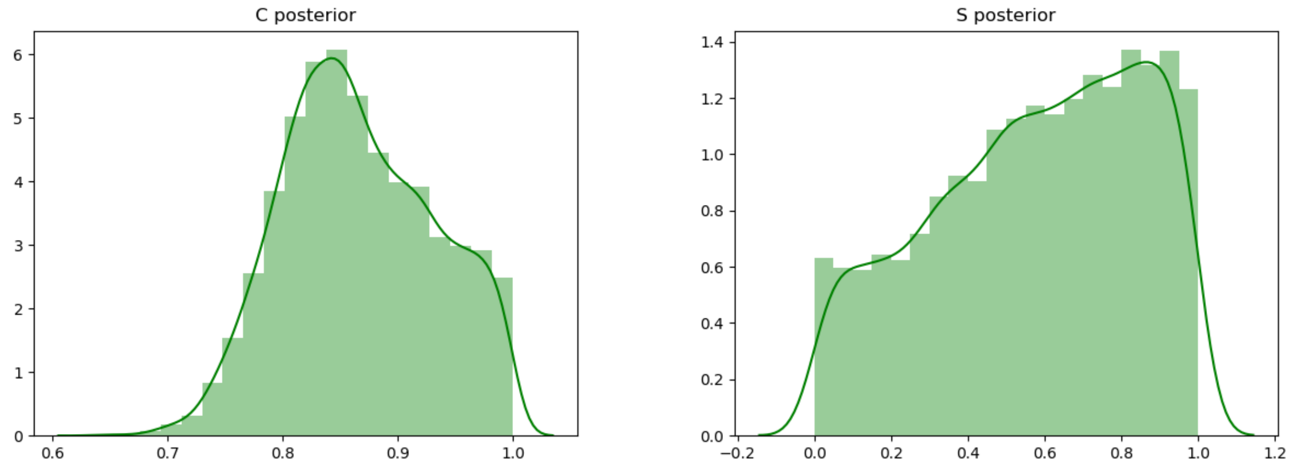

Figure 5: The Prior Predictive distribution estimated with Independent Sampling.

Problem 14.1.7

Now suppose that the sample size was 1000 people of which 200 tested positive. Using the same priors as the previous question, determine the posterior for S. What do you conclude about your test's sensitivity?

The results are still uninformative.

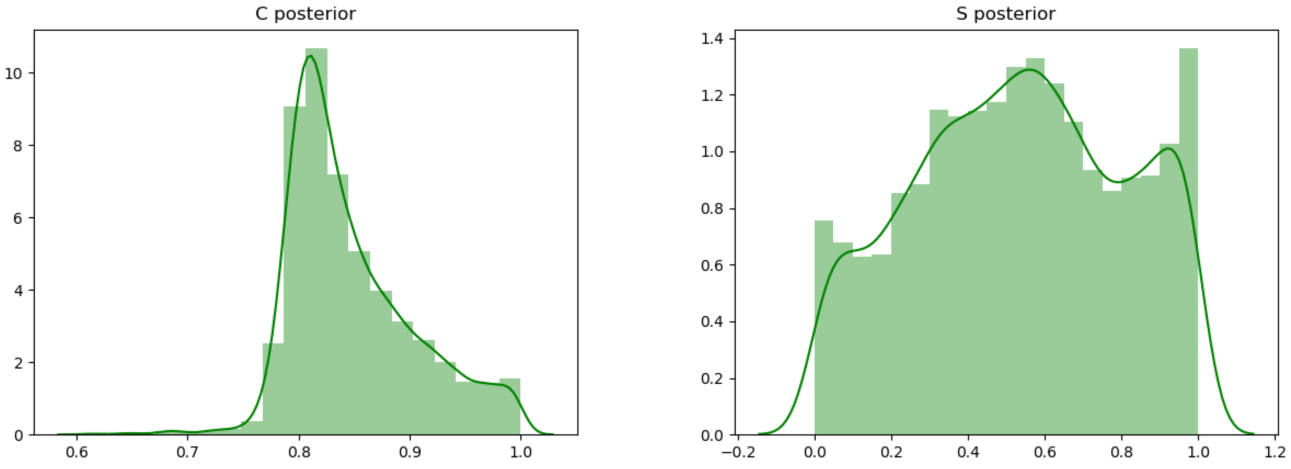

Figure 6: The Prior Predictive distribution estimated with Independent Sampling.

14.2 Coal mining disasters in the UK

Problem 14.2.1.

Graph the data over time. Around what year (n) does it appear the disaster rate occurred?

1880-1900

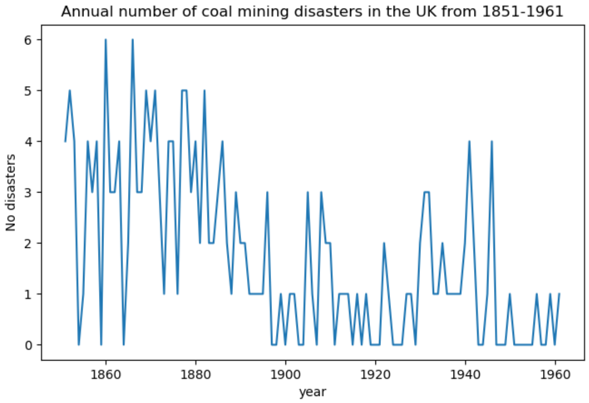

Figure 7

Problem 14.2.2.

Assuming the same $\lambda_i \sim (a;b)$ priors for $i = \{1,2\}$ and a discrete uniform prior for n between 1851-1861, determine out an expression for the full (un-normalised) posterior density.

$$p(\lambda_1, \lambda_2, n|x_{1:N}, a, b) \propto p(\lambda_1, \lambda_2, n, x_{1:N}, a, b) = p(x_{1:n}|\lambda_1, \lambda_2, n, a, b, x_{n+1:N}) \cdot p(\lambda_1, \lambda_2, n, a, b, x_{n+1:N}) =$$

$$p(x_{1:n}|\lambda_1) \cdot p(x_{n+1:N}|\lambda_1, \lambda_2, n, a, b) \cdot p(\lambda_1, \lambda_2, n, a, b) = p(x_{1:n}|\lambda_1) \cdot p(x_{n+1:N}|\lambda_2) \cdot p(\lambda_1|a, b) \cdot p(\lambda_2|a, b) \cdot p(n)$$

$$Poission(x_{1:n}|\lambda_1) \cdot Poission(x_{n+1:N}|\lambda_2) \cdot \Gamma(\lambda_1|a, b) \cdot \Gamma(\lambda_2|a, b) \cdot U(1851, 1861) \tag{7}$$

Problem 14.2.3.

Determine the conditional distribution for λ_1 by finding all those terms in the density that include λ_1, and removing the rest as constants of proportionality.

Answer:

$$\lambda_1|x_{1:n}, n, a, b \sim Poission(x_{1:n}|\lambda_1) \cdot \Gamma(\lambda_1|a, b) \tag{8}$$

Since the Gamma is a conjugate prior to the Poisson we get:

$$\lambda_1|x_{1:n}, n, a, b \sim Gamma(a + \sum_{t=1}^{n} x_t, b + n) \tag{9}$$

Problem 14.2.4.

Using your answer to the previous problem write down the conditional distribution for λ_2.

Answer:

$$\lambda_2|x_{n+1:N}, n, a, b \sim Gamma(a + \sum_{t=n+1}^{N} x_t, b + N - n) \tag{10}$$

Problem 14.2.5.

By collecting the terms that depend on n, write down its density

$$p(n|x_{1:N}, \lambda_2, \lambda_1) = Poission(x_{1:n}|\lambda_1) \cdot Poission(x_{n+1:N}|\lambda_2) \cdot U(1851, 1861) \propto Poission(x_{1:n}|\lambda_1) \cdot Poission(x_{n+1:N}|\lambda_2) = \tag{11}$$

$$= \prod_{t=1}^{n} \frac{\lambda_1^{x_t}}{x_t!} e^{-\lambda_1} \cdot \prod_{t=n+1}^{N} \frac{\lambda_2^{x_t}}{x_t!} e^{-\lambda_2} \propto \lambda_1^{\sum_{t=1}^{n} x_t} e^{-n\lambda_1} \cdot \lambda_2^{\sum_{t=n+1}^{N} x_t} e^{-(N-n)\lambda_2} \tag{12}$$

Note that we can drop $x_t!$ since these terms will be constant.

0.2 Problem 14.2.10.

Combine all three previously created sampling functions to create a working Gibbs sampler. Hence estimate the change-point and its 95% central credible intervals.

Answer:

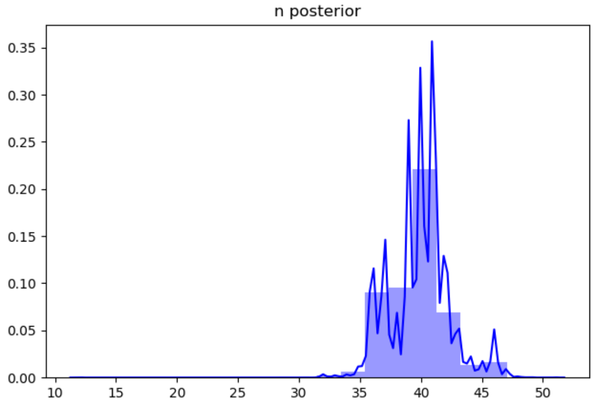

Figure 8: The posterior for n estimated via Gibbs Sampling.

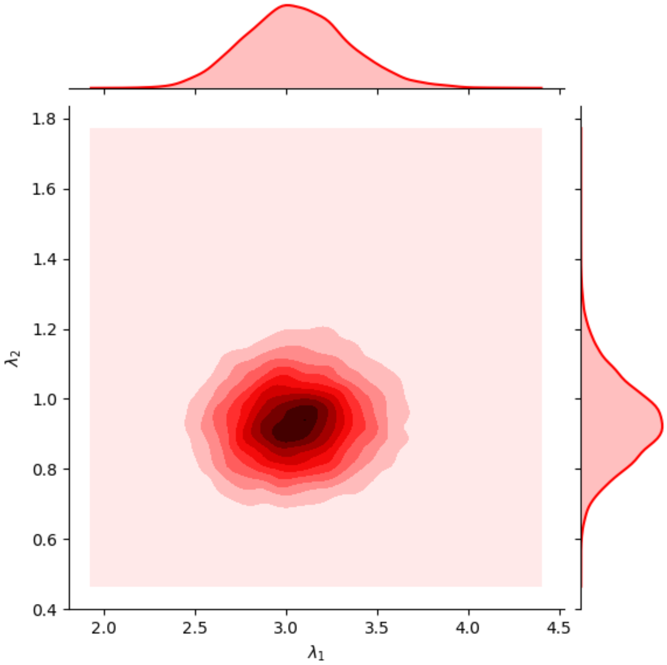

Figure 9: The posterior for λ_1 and λ_2 estimated via Gibbs Sampling.

The 95 % credible interval for n:

$$36 \leq n \geq 45 \tag{13}$$

[36 45]