Bayesian Inference and Data Analysis of Blood Glucose levels

Alan Bouwman

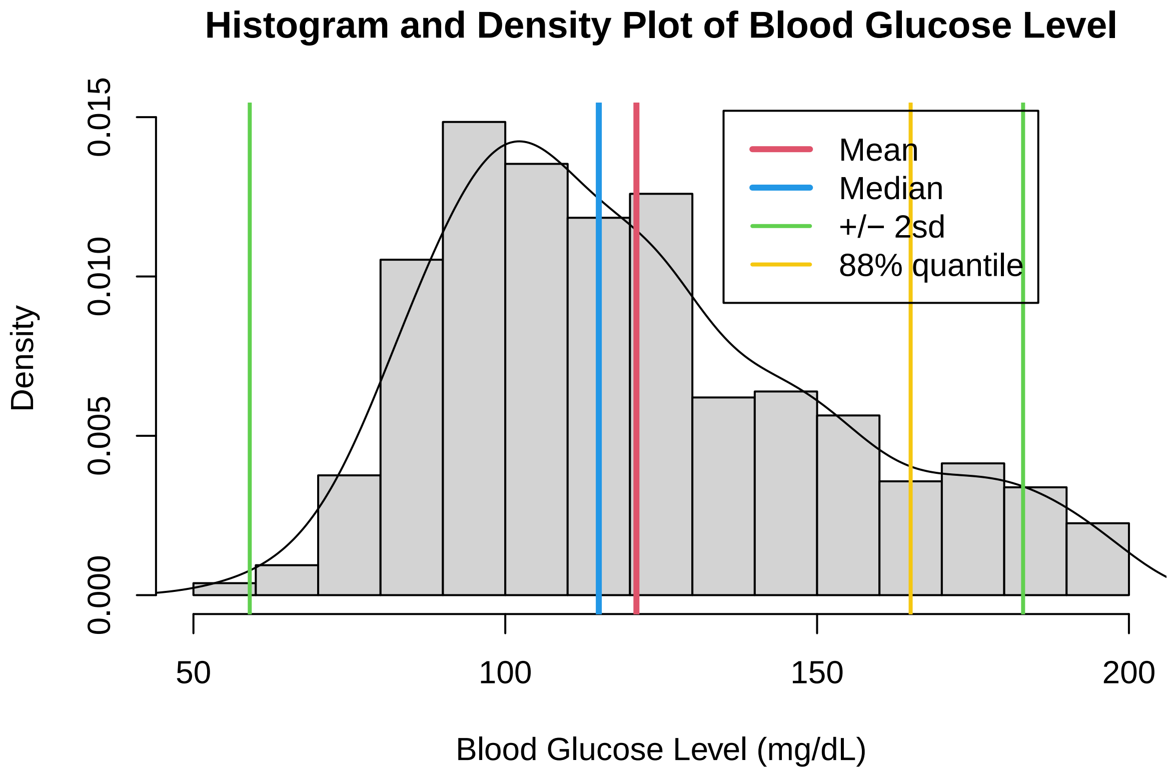
December 2, 2022

Section 1: Introduction

A population of 532 women living near Phoenix, Arizona were tested for diabetes. Other information was gathered from these women at the time of testing, including number of pregnancies, glucose level, blood pressure, skin fold thickness, body mass index, diabetes pedigree and age. Our goal is to determine if the statistical model presented in section 3 is adequate for describing blood glucose level of the 532 women. Figure 1 shows the histogram and density of the blood glucose levels for the 532 participants. The histogram indicates that we are probably looking at a mixture of normal distributions from 2 populations: women with normal blood glucose levels and women with high blood glucose levels (possibly diabetic).

Figure 1: Histogram and Density Plot of Blood Glucose Level

This is the histogram of the 532 participants. We also show the the mean (121 mg/dL), median (115 mg/dL), and ±2 standard deviations from the mean (one standard deviation is about 31 mg/dL). According to the Center of Disease control, about 12% of women in the united states have diabetes, so we also show the 88% quantile, which is about 165 mg/dL.



Section 2: Statistical Analysis

In this work, we use are doing Bayesian data analysis of the Blood Glucose levels. We derive the full conditional distributions and implement a Gibbs sampler to approximate the posterior distribution for the parameters $\theta_{(1)}$ and $\theta_{(2)}$ which represent the means of the normal and high blood glucose groups respectively.

Section 3: Statistical Model

Sampling and Prior Distributions

We choose to use a mixture model for the data. For each of the $n = 532$ study participants, we assign a group membership variable X_i such that

$$X_i = \begin{cases} 1 & \text{with probability } \pi \\ 2 & \text{with probability } 1 - \pi \end{cases}$$

Then the observed data Y_i is given the following density:

$$p(y_i|x_i) = \begin{cases} dnorm(y_i; \theta_1, \sigma_1^2) & x_i = 1 \\ dnorm(y_i; \theta_2, \sigma_2^2) & x_i = 2 \end{cases}$$

Note that the X_i are independent and the Y_i are independent given the X_i.

We use the following prior distribution for the model:

$$\begin{aligned}
p(\pi, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2) &= p(\pi)p(\theta_1)p(\theta_2)p(\sigma_1^2)p(\sigma_2^2) \\
pi &\sim beta(\alpha, beta) \\
\theta_j &\sim normal(\mu_0, \tau_0^2) \text{ for both } j = 1, 2 \\
\sigma_j^2 &\sim inverse - gamma(\nu_0/2, \sigma_0^2\nu_0/2) \text{ for both } j = 1, 2
\end{aligned}$$

Full Conditional Distributions

We derive the full conditional distributions for all of the variables. Let $n_1 = \sum_{i,x_i=1} 1$ and $n_2 = \sum_{i,x_i=2} 1$ and note that $n_2 = n - n_1$. Then, let $\bar{y}_1 = \frac{1}{n_1}\sum_{i,x_i=1} y_i$ and $\bar{y}_2 = \frac{1}{n_2}\sum_{i,x_i=2} y_i$

$$\begin{aligned}
p(X_i = x_i|\pi, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2, \boldsymbol{Y}, \boldsymbol{X_{-i}}) &\propto p(X_i, \pi, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2, \boldsymbol{Y}, \boldsymbol{X_{-i}}) \\
&\propto p(\boldsymbol{Y}|X_i, \pi, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2, \boldsymbol{X_{-i}})p(X_i|\pi)p(\pi)p(\theta_1)p(\theta_2)p(\sigma_1^2)p(\sigma_2^2) \\
&\propto p(Y_i|X_i, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2)p(X_i|\pi) \\
&\propto dbinom(x_i; n, 1, p_2/(p_1 + p_2)) + 1
\end{aligned}$$

where

$$\begin{aligned}
p_1 &= dnorm(y_i; \theta_1, \sigma_1^2), \\
p_2 &= dnorm(y_i; \theta_2, \sigma_2^2).
\end{aligned}$$

$$\begin{aligned}
p(\pi|\theta_1, \theta_2, \sigma_1^2, \sigma_2^2, \boldsymbol{Y}, \boldsymbol{X}) &\propto p(\pi, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2, \boldsymbol{Y}, \boldsymbol{X}) \\
&\propto p(\boldsymbol{Y}|\pi, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2, \boldsymbol{X})p(\boldsymbol{X}|\pi)p(\pi)p(\theta_1)p(\theta_2)p(\sigma_1^2)p(\sigma_2^2) \\
&\propto p(\boldsymbol{X}|\pi)p(\pi) \\
&\propto dbinom(X = n_1; n, \pi)dbeta(\pi; \alpha, \beta) \\
&\propto \pi^{n_1}(1-\pi)^{n-n_1}\pi^{\alpha-1}(1-\pi)^{\beta-1} \\
&= \pi^{n_1+\alpha-1}(1-\pi)^{\beta+n_2-1} \\
&\propto dbeta(\pi; \alpha + n_1, \beta + n_2)
\end{aligned}$$

$$\begin{aligned}
p(\theta_1|\pi, \theta_2, \sigma_1^2, \sigma_2^2, \boldsymbol{Y}, \boldsymbol{X}) &\propto p(\boldsymbol{Y}|\pi, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2, \boldsymbol{X})p(\boldsymbol{X}|\pi, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2)p(\theta_1) \\
&\propto p(\boldsymbol{Y}|\theta_1, \sigma_1^2)p(\theta_1) \\
&\propto dnorm(y_i; \theta_1, \sigma_1^2)dnorm(\theta_1; \mu_0, \theta_0) \\
&\propto dnorm\left(\theta_1; \frac{\frac{1}{\tau_0}\mu_0 + \frac{n_1}{\sigma_1^2}\bar{y_1}}{\frac{1}{\tau_0} + \frac{n_1}{\sigma_1^2}}, \frac{1}{\frac{1}{\tau_0} + \frac{n_1}{\sigma_1^2}}\right)
\end{aligned}$$

$$\begin{aligned}
p(\theta_2|\pi, \theta_1, \sigma_1^2, \sigma_2^2, \boldsymbol{Y}, \boldsymbol{X}) &\propto p(\boldsymbol{Y}|\pi, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2, \boldsymbol{X})p(\boldsymbol{X}|\pi, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2)p(\theta_2) \\
&\propto p(\boldsymbol{Y}|\theta_2, \sigma_2^2)p(\theta_2) \\
&\propto dnorm(y_i; \theta_2, \sigma_2^2)dnorm(\theta_2; \mu_0, \theta_0) \\
&\propto dnorm\left(\theta_2; \frac{\frac{1}{\tau_0}\mu_0 + \frac{n_2}{\sigma_2^2}\bar{y_2}}{\frac{1}{\tau_0} + \frac{n_2}{\sigma_2^2}}, \frac{1}{\frac{1}{\tau_0} + \frac{n_2}{\sigma_2^2}}\right)
\end{aligned}$$

$$\begin{aligned}
p(\sigma_1^2|\pi, \theta_1, \theta_2, \sigma_2^2, \boldsymbol{Y}, \boldsymbol{X}) &\propto p(\boldsymbol{Y}|\theta_1, \sigma_1^2)p(\sigma_1^2) \\
&\propto dnorm(y; \theta_1, \sigma_1^2)dinversegamma(\sigma_1^2; \nu_0/2, \nu_0/2\sigma_0^2) \\
&\propto dinversegamma\left(\sigma_1^2; \frac{\nu_0 + n_1}{2}, \frac{\nu_0 + n_1}{2}\frac{\nu_0\sigma_0^2 + \sum_{i,x_i=1}(y_i - \theta)^2}{\nu_0 + n_1}\right)
\end{aligned}$$

$$\begin{aligned}
p(\sigma_2^2|\pi,\theta_1,\theta_2,\sigma_1^2,\boldsymbol{Y},\boldsymbol{X}) &\propto p(\boldsymbol{Y}|\theta_2,\sigma_2^2)p(\sigma_2^2)\\
&\propto dnorm(y;\theta_2,\sigma_2^2)dinversegamma(\sigma_2^2;\nu_0/2,\nu_0/2\sigma_0^2)\\
&\propto dinversegamma\left(\sigma_2^2;\frac{\nu_0+n_2}{2},\frac{\nu_0+n_2}{2}\frac{\nu_0\sigma_0^2+\sum_{i,x_i=2}(y_i-\theta)^2}{\nu_0+n_2}\right)
\end{aligned}$$

Summarized the full conditional distributions are:

$$\begin{aligned}
p(X_i=x_i|\pi,\theta_1,\theta_2,\sigma_1^2,\sigma_2^2,\boldsymbol{Y},\boldsymbol{X_{-i}}) &\propto dbinom(x_i;n,1,p_2/(p_1+p_2))+1\\
p(\pi|\theta_1,\theta_2,\sigma_1^2,\sigma_2^2,\boldsymbol{Y},\boldsymbol{X}) &\propto dbeta(\pi;\alpha+n_1,\beta+n_2)\\
p(\theta_1|\pi,\theta_2,\sigma_1^2,\sigma_2^2,\boldsymbol{Y},\boldsymbol{X}) &\propto dnorm(\theta_1;\mu_{n,1},\sigma_{n,1}^2)\\
p(\sigma_1^2|\pi,\theta_1,\theta_2,\sigma_2^2,\boldsymbol{Y},\boldsymbol{X}) &\propto dinverse-gamma(\sigma_1^2;\nu_{n,1}/2,\tau_{n,1}^2\nu_{n,1}/2)\\
p(\sigma_2^2|\pi,\theta_1,\theta_2,\sigma_2^2,\boldsymbol{Y},\boldsymbol{X}) &\propto dinverse-gamma(\sigma_2^2;\nu_{n,2}/2,\tau_{n,2}^2\nu_{n,2}/2)
\end{aligned}$$

where

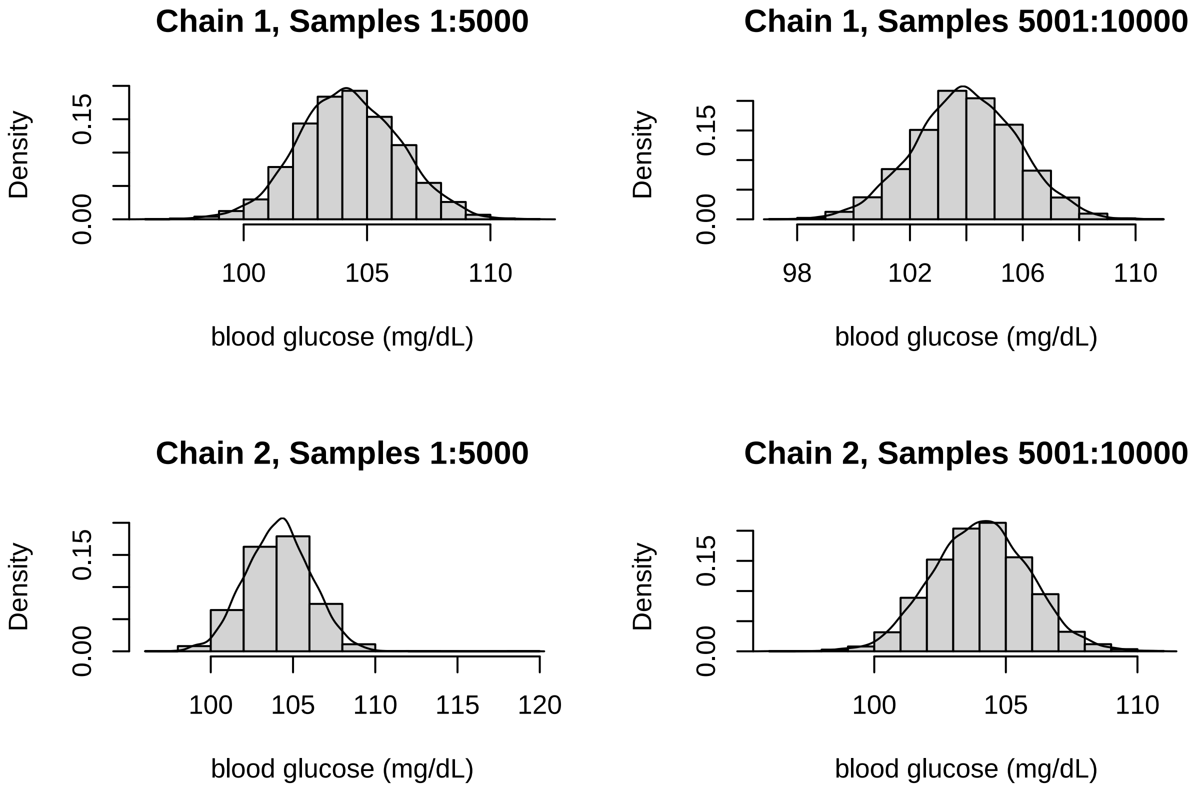
$$\begin{aligned}
p_1 &= dnorm(y_i;\theta_1,\sigma_1^2)\\
p_2 &= dnorm(y_i;\theta_2,\sigma_2^2)\\
\mu_{n,1} &= \frac{\frac{1}{\tau_0}\mu_0+\frac{n_1}{\sigma_1^2}\bar{y_1}}{\frac{1}{\tau_0}+\frac{n_1}{\sigma_1^2}}\\
\sigma_{n,1}^2 &= \frac{1}{\frac{1}{\tau_0}+\frac{n_1}{\sigma_1^2}}\\
\mu_{n,2} &= \frac{\frac{1}{\tau_0}\mu_0+\frac{n_2}{\sigma_2^2}\bar{y_2}}{\frac{1}{\tau_0}+\frac{n_2}{\sigma_2^2}}\\
\sigma_{n,2}^2 &= \frac{1}{\frac{1}{\tau_0}+\frac{n_2}{\sigma_2^2}}\\
\nu_{n,1} &= \nu_0+n_1\\
\tau_{n,1}^2 &= \frac{\nu_0\sigma_0^2+\sum_{i,x_i=1}(y_i-\theta)^2}{\nu_{n,1}}\\
\nu_{n,2} &= \nu_0+n_2\\
\tau_{n,2}^2 &= \frac{\nu_0\sigma_0^2+\sum_{i,x_i=2}(y_i-\theta)^2}{\nu_{n,2}}
\end{aligned}$$

Section 4: Gibbs Sampler and MCMC Diagnostics

4.1 Sample Splits

We will example $\theta_{(1)}$, the mean of the normal blood glucose level group in mg/dL. The Sample Splits (seen in figure 2) between the chain indicate that the model is converging well. The 1-5000 chain looks fairly different from the other 3 chains. This is likely due to the way that x was initialized. Because x was assigned to group 1 or 2 (low glucose group or high glucose group) with a 50-50 chance, the mean for the first few iterations would be much closer to the population mean. The max value in the theta min for chain 1 is 111.6. There are only 6 (out of 10,000) sampled means in the theta min group whose mean is higher than 112, and they are all the first 6 values from the Gibbs sampler. So the model in chain 2 converging quite quickly, even though it is significantly different from the model with chain 1 for the first several iterations. If a burn in was used, we expect that all of these splits would look the same. We will test this claim later in the JAGS section by running JAGS on both chains with a burn in of 1000, and evaluate the distributions of the sample splits.

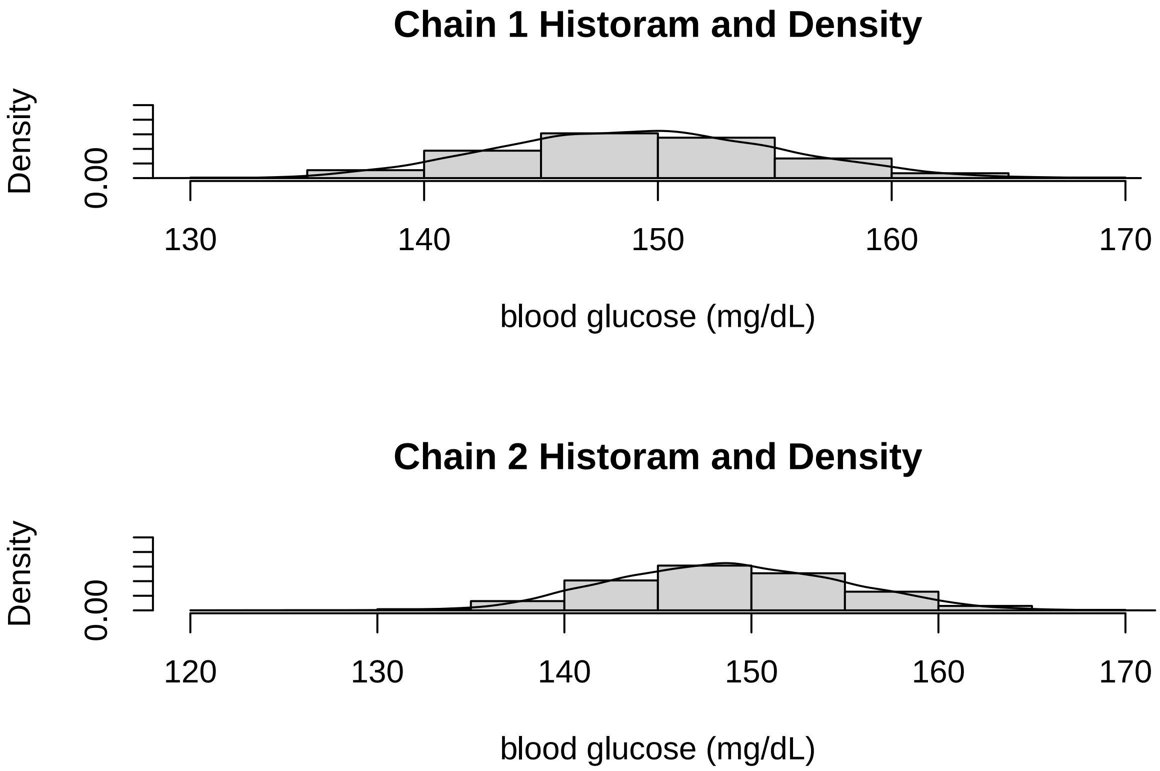
Figure 2: Sample Splits for multiple chains for normal blood glucose level group



4.3 Multiple Chains

In this section we will examine $\theta_{(2)}$, the posterior mean of the high blood glucose group. As both chains both quickly converge to the same posterior mean, we can be confident that the model is not too sensitive to initial conditions. Again, the second chain is skewed (this time in the other direction) due to the initialization of x. Due to this initialization, the first few means from both the normal and high blood glucose groups is very close to 120. After only a few iterations, the sampler converges to the mean.

Figure 3: Histrograms and Density for Multiple Chains for high blood glucose group


Chain 1 Historam and Density

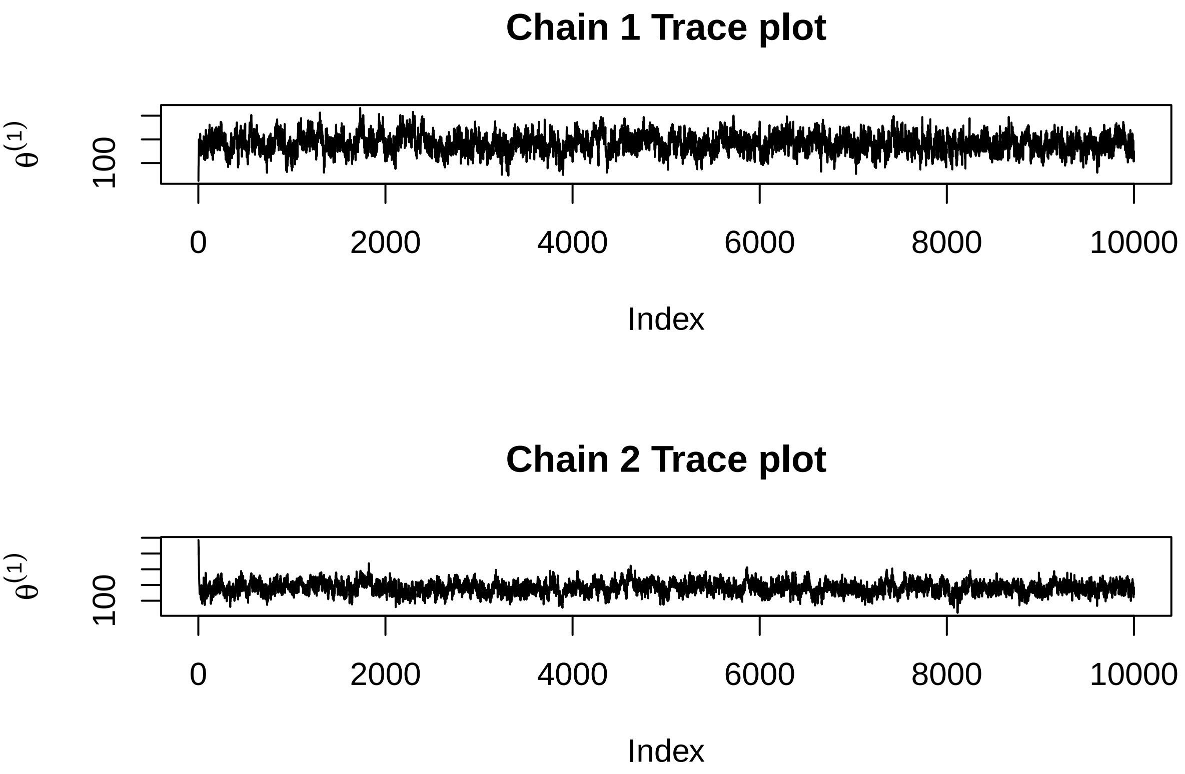
Chain 2 Historam and Density

4.4 Trace Plots

The trace plots for both chains indicate reasonable performance for the model. There do not appear to be multiple regions of stability. As discussed in the last section, the second chain starts quite a bit higher than the expected mean for the low blood glucose group due to random initialization, but converges very quickly.

Figure 4: Trace plots for multiple chains

Figure 4 shows the trace plots $\theta_{(1)}$, the posterior mean of the low blood glucose group, for chain 1 and chain 2.



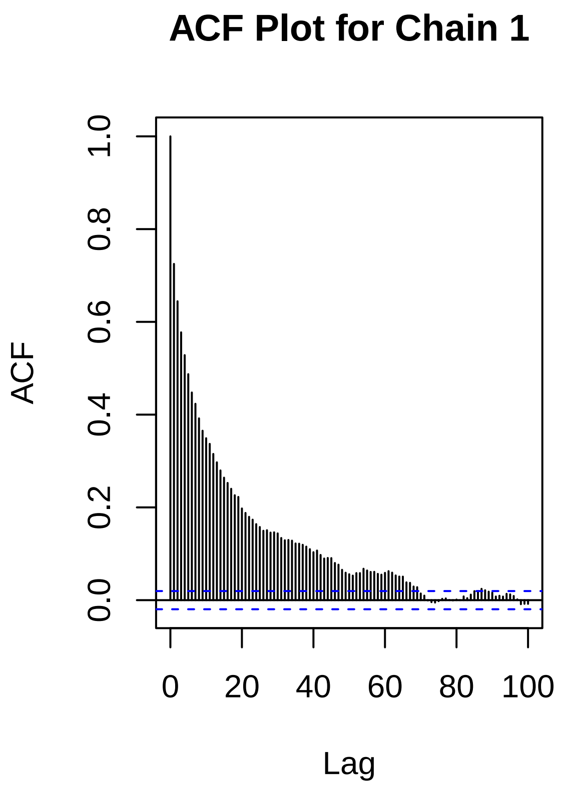
Chain 1 Trace plot

Chain 2 Trace plot

4.5 ACF and Effective Size

Figure 5: ACF Plots for Multiple Chains

Figure 5 shows the ACF plots for chain 1 and chain 2 for $\theta_{(1)}$, the posterior mean of the low blood glucose group. For chain 1, the samples are correlated up until about a lag of 70. That is, for about every 70 Gibbs samples, there is only about 1 independent sample. This will result in a small sample size. For chain 2 the samples are correlated for a lag of up to about 80.


ACF Plot for Chain 1

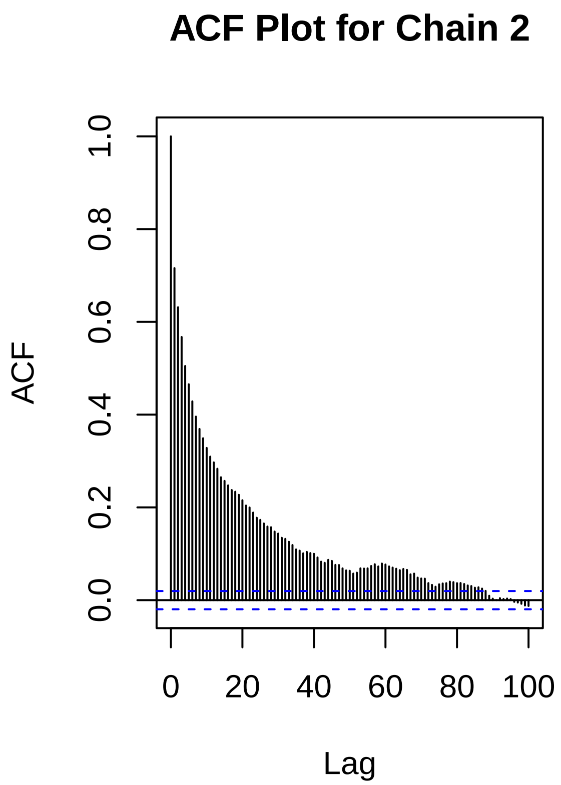

ACF Plot for Chain 2

Table 1: Effective Sizes for Multiple Chains

Table 1 shows the effective size for the parameters $\theta_{(1)}, \theta_{(2)}$ (mean blood glucose for normal and high glucose groups), and π (proportion of participants in each group). Although the effective sizes are pretty small, analysis on larger sample sizes leads to similar results in terms of posterior statistics.

parameter	chain1 effective size	chain 2 effective size
theta1 (normal blood glucose)	464	425
theta2 (high blood glucose)	227	225
pi	233	216

Bases on the results results, we believe that $S = 5000$ samples is enough for this analysis. Although the sample size is relatively small, we also tried larger sample sizes (up to 100,000) and found very similar results for posterior means and confidence intervals.

Summary statistics

Table 2: Summary Statistics for Gibbs Sampler

Table 2a and 2b show the summary statistics for the models. Interestingly, what each parameter in the model represents changes between the chains, and we reflect this by showing the proportion that the model itself captures (in chain 1, π captures the normal blood glucose group proportion, and in chain 2, π captures the high blood glucose group proportion). Chain 1 and 2 match closely, and the mean proportions between the chains sum to 1. #### Table 2a: Summary Statistics for Chain 1

variable	mean	2.5% quantile	97.5% quantile
normal blood glucose post. mean (mg/dL)	104.13	100.36	107.94
high blood glucose post. mean (mg/dL)	149.25	137.93	160.93
proportion in normal group	0.62	0.48	0.75
min of groups	104.13	100.36	107.94

Summary stats for chain 2. #### Table 2b: Summary Statistics for Chain 2

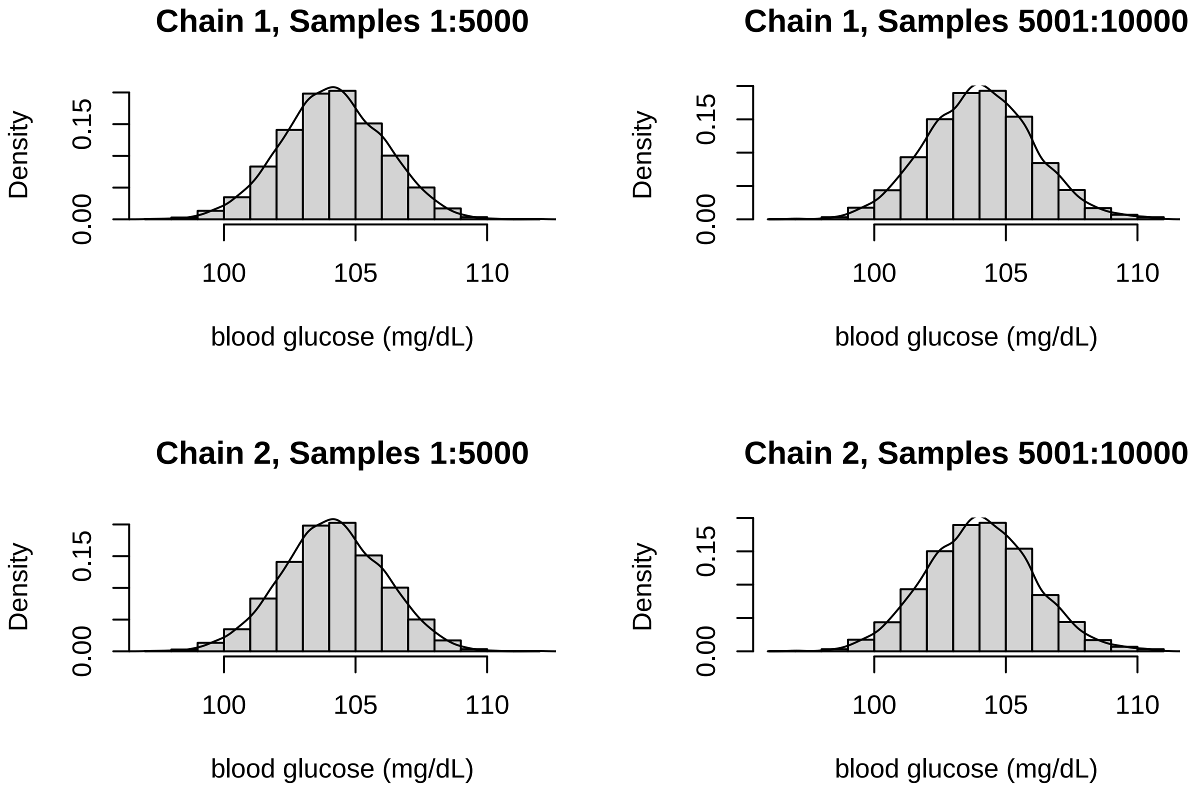
variable	mean	2.5% quantile	97.5% quantile
high blood glucose post. mean (mg/dL)	148.80	137.47	160.67
normal blood glucose (mg/dL) post. mean	104.10	100.46	107.85
proportion in high group	0.38	0.25	0.53
min of groups	104.10	100.46	107.85

Section 5: MCMC Diagnostics

Sample Splits for JAGS

Figure 6: Sample Splits for JAGS normal blood glucose group

Figure 6 shows the sample splits for both JAGS chains for the normal blood glucose group. The main difference between these and the results in Figure 2 is Chain 2, Samples 1:5000. Earlier we saw that it took about 6 iterations for the chain 2 Gibbs sampler to reach a reasonable mean. Because a burn-in of 1000 is used for JAGS, the model had already converged. All 4 histograms and density plots here look very similar, indicating that the model is converging well.



Section 6: Model Checking and Summary Statistics

As the effective sample sizes from the previous section were small, we rerun the the model with 50,000 iterations. We will only focus on results from chain 1.

Table 3: Posterior Summary statistics

Table 3a and 3b show posterior summary statistics for both the Gibbs Sampler and JAGS. The Gibbs Sampler and JAGS models match very closely for means and confidence intervals. #### Table 3a: Summary stats for Gibbs Sampler. Table 3a shows the posterior mean for the Gibbs Sampler. The mean for the normal blood glucose group is about 104 mg/dL, which is roughly what would be expected from a fasted person without diabetes. The high glucose group has a mean of about 149. If a fasted person tested 126 mg/dL on two occasions, they would be diagnosed with diabetes. Thus, the two groups may represent no diabetes and diabetes.

variable	mean	2.5% quantile	97.5% quantile
normal blood glucose post. mean (mg/dL)	104.01	100.32	107.83
high blood glucose post. mean (mg/dL)	148.91	137.60	160.84
proportion in normal blood glucose group	0.62	0.47	0.75

Table 3b: Summary stats for JAGS.

variable	mean	2.5% quantile	97.5% quantile
normal blood glucose post. mean (mg/dL)	104.03	100.22	107.86
high blood glucose post. mean (mg/dL)	148.96	137.12	161.08
proportion in normal blood glucose group	0.38	0.25	0.54

Figure 7

Figure 7 shows the predictive posterior distribution histogram and density plot next to the histogram and density plot of the data. The overall shape largely matches. However, the data almost cuts off above about 200 mg/dL while the predictive distribution looks smooth at both ends. Also, the data is multi-modal with at least 3 peaks, while the predictive distribution appears unimodal.

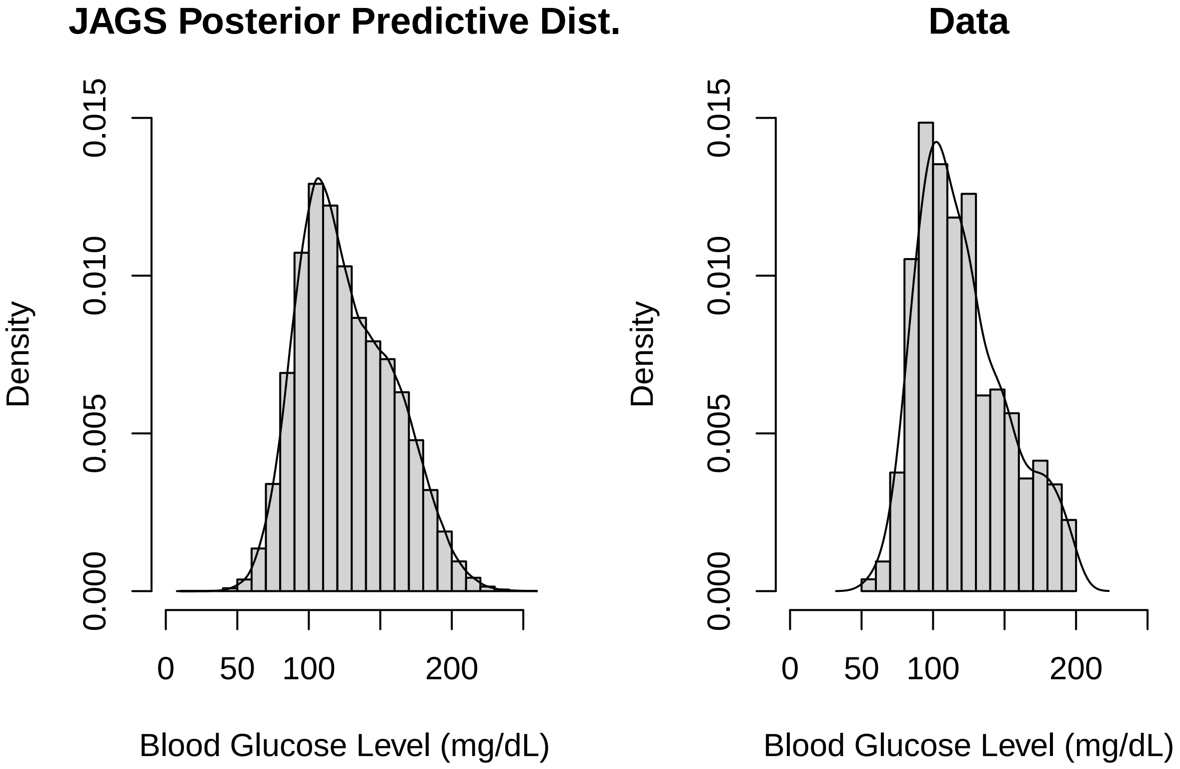
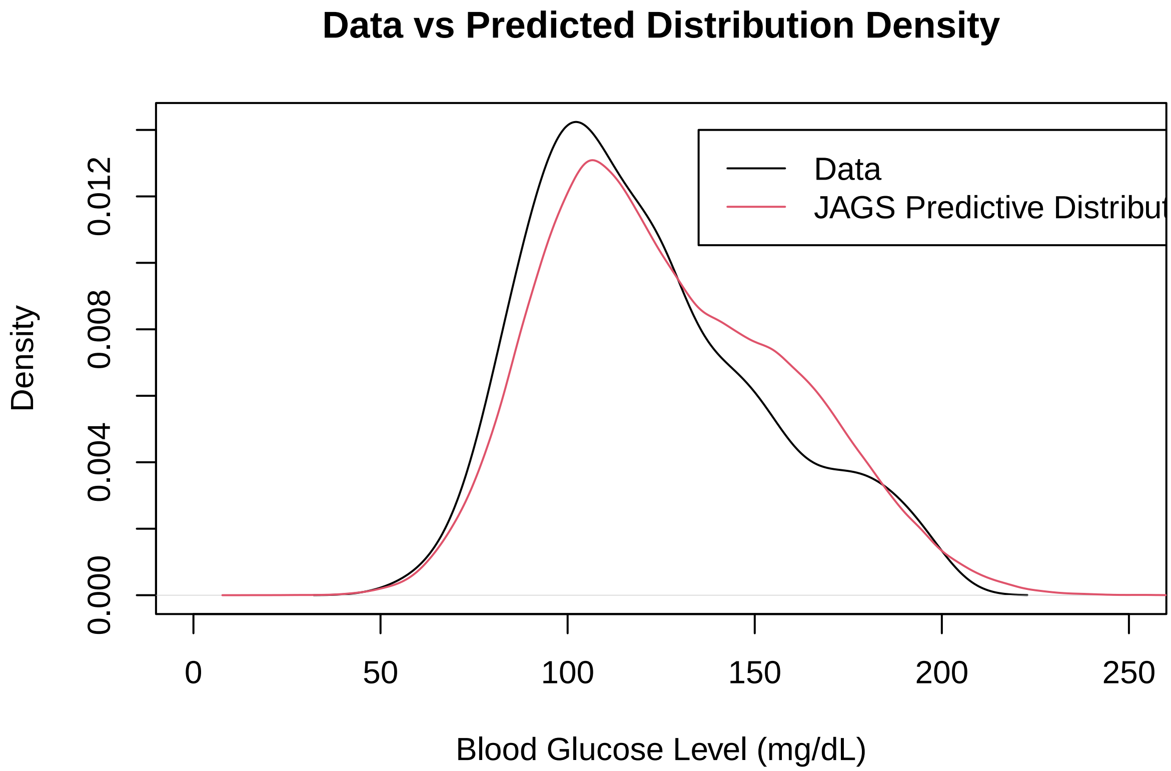


Figure 8: Density of data vs predicted distribution

Figure 8 shows the density plots of the Data (black) and the JAGS Posterior Predictive Distribution (Red). Again, the model matches the data pretty well.



Section 7: Conclusion and Discussion

Our statistical model was able to match the data fairly well. Multiple chains indicate that the model is not too sensitive to initial conditions, except in what each parameter in the model represents. Trace plots and sample splits indicate that the model converges fairly quickly. Trace plots also indicate that the model does not get stuck in any particular area. The samples are highly correlated, and every 70 or so samples only gives one independent sample. We were able to compensate for this by increasing the sample size. Sample sizes of $S = 10,000$ and $S = 50,000$ give similar results for posterior means. Also, the JAGS model and Gibbs sampler perform very similarly.

The model was able to capture different groups of participants, one group with a normal blood glucose of about 104 mg/dL, and the other group with a high blood glucose of about 149 mg/dL. Further, about 62% of participants were in the normal blood glucose group and the other 38% were in the high blood glucose group. We can say with near certainty that the participants in the first group do not have diabetes, as a fasted individual with diabetes will usually have a blood glucose level of at least 126 mg/dL (CDC). We cannot say very much about the second group. If we knew everyone who took the test was fasted, we could conclude the second group likely has diabetes. However, as we don't have this information it's possible that many of those participants simply had a snack before taking the blood test. 'The high glucose group included 38% about of the participants, which is much higher than the proportion of the US population (12%). However, the individuals who were tested came from the Pima Indians. The Pima Indians have a high rate of diabetes compared to the rest of the United States (NIH).

References:

CDC: https://www.cdc.gov/diabetes/basics/ . NIH: https://pubmed.ncbi.nlm.nih.gov/7468572/ .