Victor Agboli

STAT 8670: Computational Statistics

Abortion Data Analysis using Bayesian Inference

Introduction

Coming from a country where abortion is illegal the United States of America where abortion is legal (per state), hence, I assume that the legality of abortion will make the abortion rate high. In this project, I'm interested in studying abortion in the USA over the last few years. I will be using the Markov Chain Monte Carlo (MCMC) method to analyze this claim because the MCMC methods helps me to approximate the posterior distribution of the abortion rate by random sampling in a probabilistic space. In this project, I will study the differences in the abortion rate in the United States of America from 2015 – 2020 and determine if truly, there is a significant increase or decrease in successive years. This analysis will be used to validate or disprove my initial claim. The data for this analysis was gotten from Johnson's archive which feeds from the Center for Disease Control (CDC) database **[1].** The data contains the number of abortions in the 50 states in the USA plus the District of Columbia.

The questions that will be answered are:

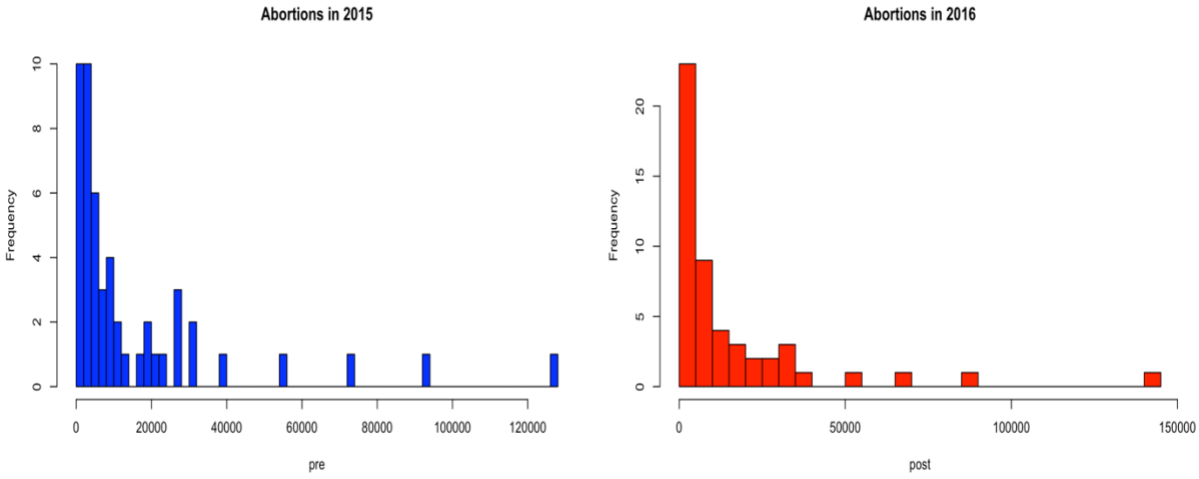
1. Is there a difference in the mean number of abortions in the USA from 2015 to 2016?
2. Is there a difference in the mean number of abortions in the USA from 2017 to 2018?
3. Is there a difference in the mean number of abortions in the USA from 2019 to 2020?

Methods

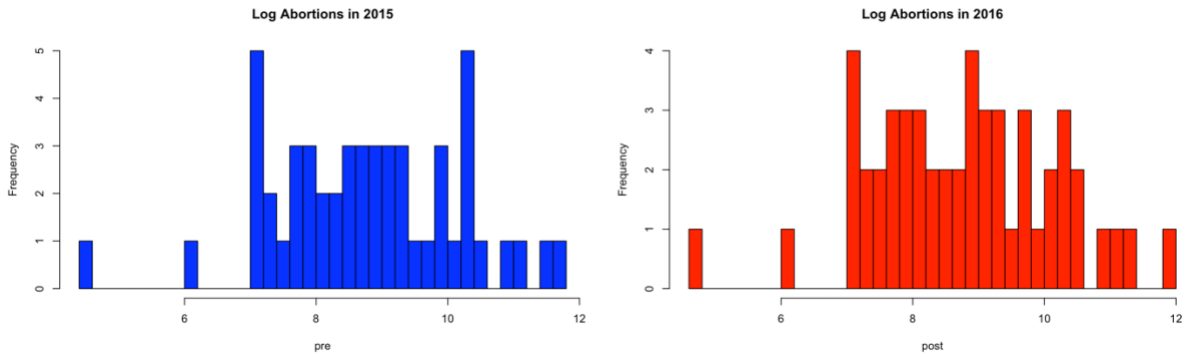
I will examine histograms of the abortion rate per year to understand the common characteristics. Next, I will hypothesize a distribution for each year and develop likelihood functions based on that distribution. I will also need to develop a prior distribution. Once I establish both of these, I will then create a posterior distribution. Once I have the posterior distribution, I will run a Markov chain using Metropolis. From that analysis, I will find an estimated posterior probability to determine the probability that the abortion rate increased.

2015-2016 Abortion Rate Analysis

Below are the histograms for the number of abortions in 2015 and 2016 respectively.



Examining the histograms, both are skewed to the right. Since I want to make my likelihood functions normal, I will take the log-transform of the number of abortions in 2015 and 2016.



Examining the histogram of the log-transform, I see that both of them are now normal.

For the prior distribution, I'm going to assume that the log of the number of abortions in 2015 and 2016 is modeled by a normal distribution with $N(\mu, \sigma^2)$. Thus, the log(abortion rates) in 2015 comes from a $N(\mu_1, {\sigma_1}^2)$ and the log(abortion rates) in 2016 comes from a $N(\mu_2, {\sigma_2}^2)$. So, the parameter becomes $\theta = (\mu_1,\ \mu_2,\ \sigma_1,\ \sigma_2)$. Thus,

$$P(data \mid \theta) = P(log2015 \mid \theta)\ P(\log 2016 \mid \theta)$$
$$= \prod_{i=1}^{51} N(\log 2015_i \mid \mu_1, {\sigma_1}^2) \prod_{i=1}^{51} N(\log 2016_i \mid \mu_2, {\sigma_2}^2)$$

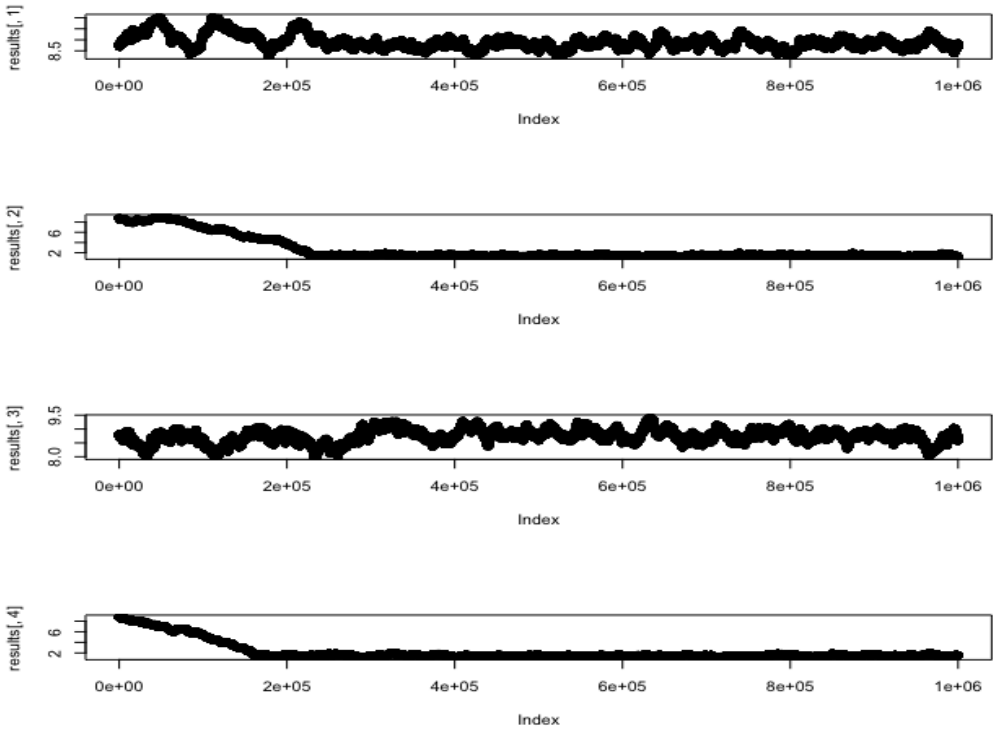
So, the distribution of μ_1 is $N(8.767\ 8.767^2)$ and the distribution of μ_2 is $N(8.768, 8.768^2)$. I also need to determine the prior for the means that does not presume which mean is bigger and to choose priors that are uniform over all plausible values. Let's assume that the distribution of σ_1 is exponential with a mean of 8.767 and I choose the distribution of σ_2 as exponential also with mean 8.768. Since all four variables are independent, I can calculate the prior as:

$$P(\mu_1,\ \mu_2, \sigma_1, \sigma_2)$$
$$= \frac{1}{\sqrt{2\pi}} \times (8.767)^{-0.5} \times e^{\frac{1}{2(8.767)^2}(\mu_1 - 8.767)^2} \times \frac{1}{\sqrt{2\pi}} \times (8.768)^{-0.5}$$
$$\times\ e^{\frac{1}{2(8.768)^2}(\mu_2 - 8.768)^2} \times \frac{1}{8.767} e^{\frac{-\sigma_1}{8.767}} \times \frac{1}{8.768} e^{\frac{-\sigma_2}{8.768}}$$

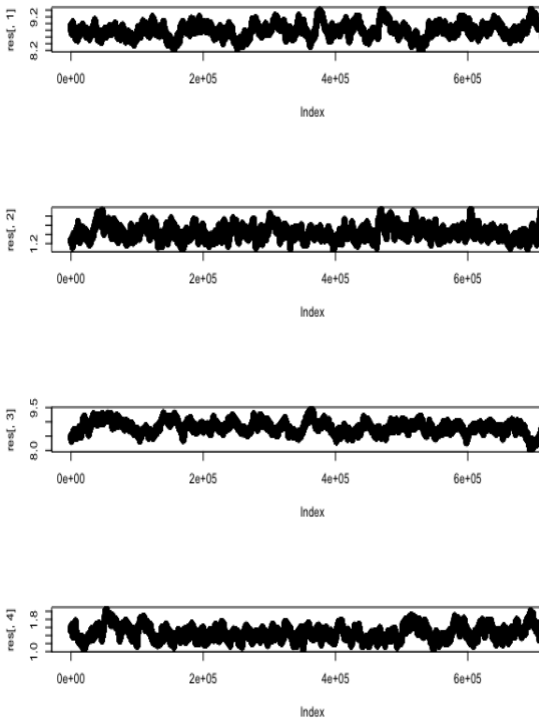
To obtain the posterior distribution, I will multiply the likelihood by the prior distribution (following from Bayesian Theorem), I will then run a Markov chain of 1000000 iterations to simulate a sample from the distribution. I choose this large number because 1000, 10000, and 100000 did not converge based on the trace plots. For this iteration, I need to state the starting vector. The vector is going to be the hypothesized means and standard deviations. Thus,

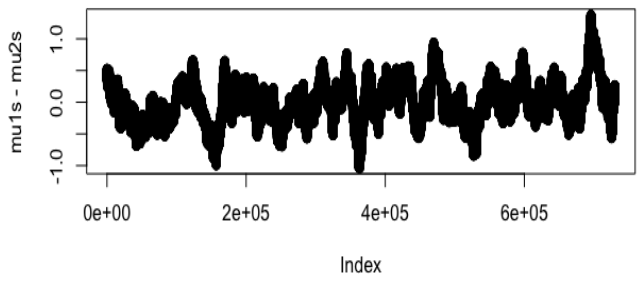
$$\theta_0 = (8.767, 8.767, 8.768, 8.768)$$

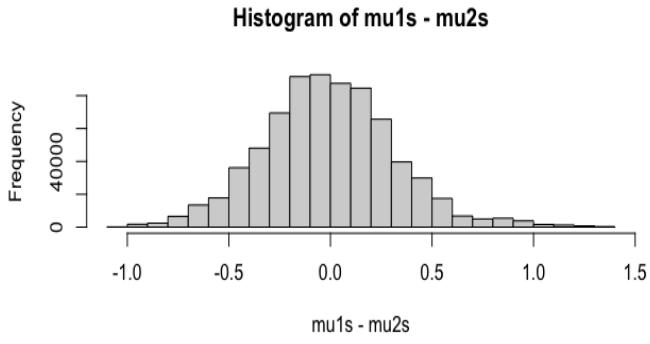
Let's us observe how the chain ran:



Looking at the chain run, I can see that the first few iterations affects the results, hence, not needed. So, I will let these iterations be my burn-in period and remove them. I will do the analysis and the inference from the remaining iterations.





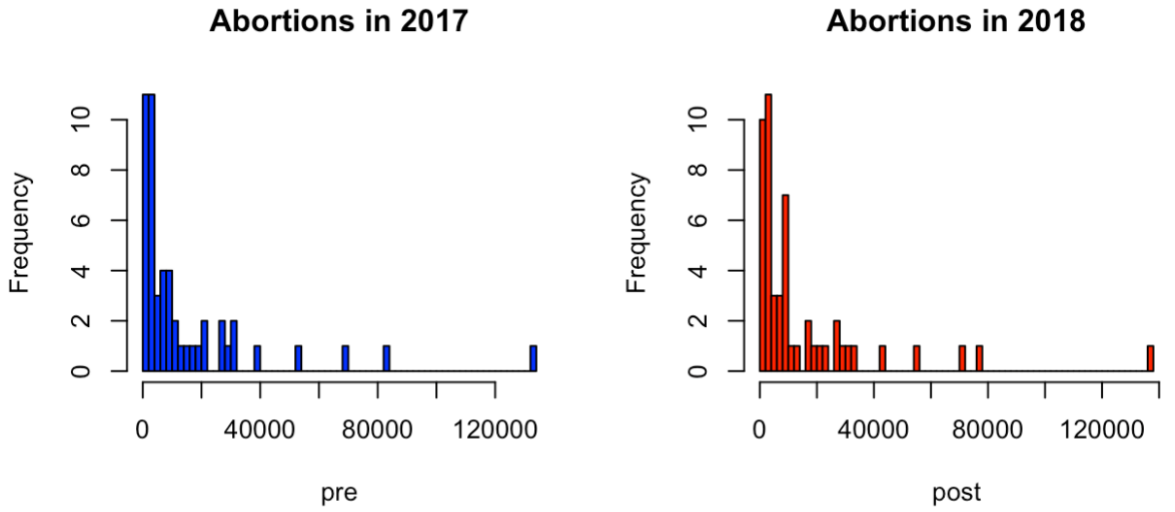


The above plot shows the trace-plots after the burn-in period have been removed.

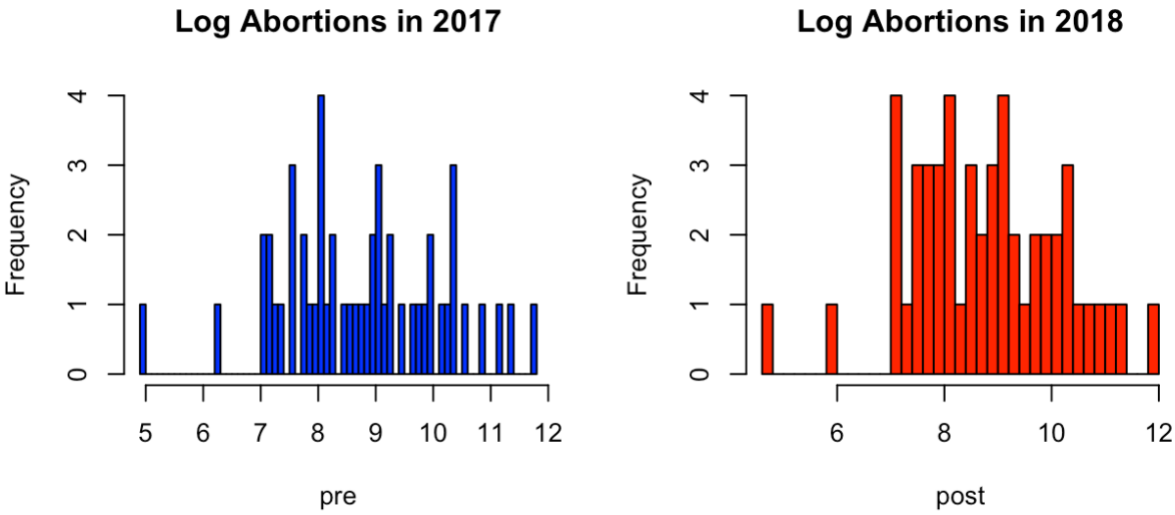
Thus, $P(\log(\mu_1) - \log(\mu_2) < 0) = P\left(\log\left(\frac{\mu_1}{\mu_2}\right) < 0\right) = P(\mu_1 < \mu_2) \approx 0.4716705$ – the probability that the number of abortions in 2015 is less than the number of abortions in 2016. This is not surprising because the two years under consideration have very close means.

2017–2018 Abortion Rate Analysis

Below are the histograms for the number of abortions in 2017 and 2018.



Both distributions are skewed to the right. Like before, I will take the log transform of 2017 and 2018 abortion data to normalize them.



Examining the histogram of the log-transform, I see that both of them are now normal.

For the prior distribution, I'm going to assume that the log of the number of abortions in 2017 and 2018 are modeled by a normal distribution with $N(\mu, \sigma^2)$. Thus, the log(abortion rates) in 2015 comes from a $N(\mu_1, {\sigma_1}^2)$ and the log(abortion rates) in 2016 comes from a $N(\mu_2, {\sigma_2}^2)$. So, the parameter becomes $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2)$. Thus,

$$P(data \mid \theta) = P(log2017 \mid \theta)\, P(\log 2018 \mid \theta)$$
$$= \prod_{i=1}^{51} N(\log 2017_i \mid \mu_1, {\sigma_1}^2) \prod_{i=1}^{51} N(\log 2018_i \mid \mu_2, {\sigma_2}^2)$$

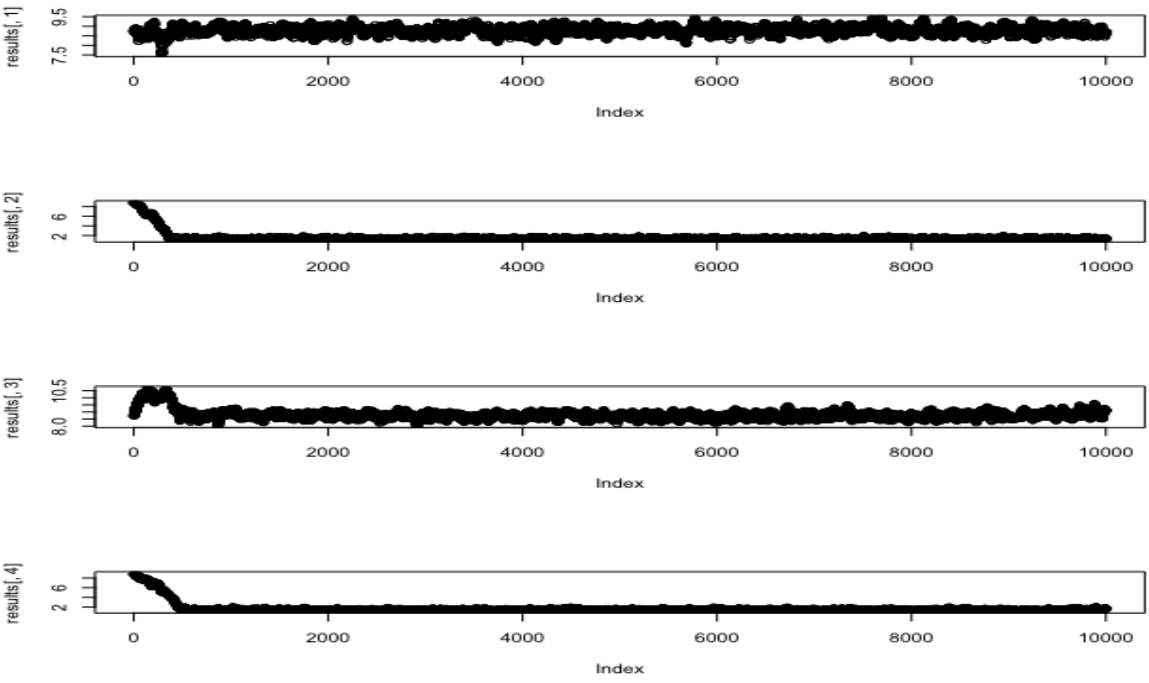
So, the distribution of μ_1 is $N(8.762, 8.762^2)$ and the distribution of μ_2 is $N(8.757, 8.757^2)$. I also need to determine the prior for the means that does not presume which mean is bigger and to choose priors that are uniform over all plausible values. Let's assume that the distribution of σ_1 is exponential with mean 8.762 and I choose the distribution of σ_2 as exponential also with mean 8.757. Since all four variables are independent, I can calculate the prior as:

$$P(\mu_1, \mu_2, \sigma_1, \sigma_2) = \frac{1}{\sqrt{2\pi}} \times (8.762)^{-0.5} \times e^{\frac{1}{2(8.762)^2}(\mu_1 - 8.762)^2} \times \frac{1}{\sqrt{2\pi}} \times (8.757)^{-0.5} \times e^{\frac{1}{2(8.757)^2}(\mu_2 - 8.757)^2} \times \frac{1}{8.762} e^{\frac{-\sigma_1}{8.762}} \times \frac{1}{8.757} e^{\frac{-\sigma_2}{8.757}}$$

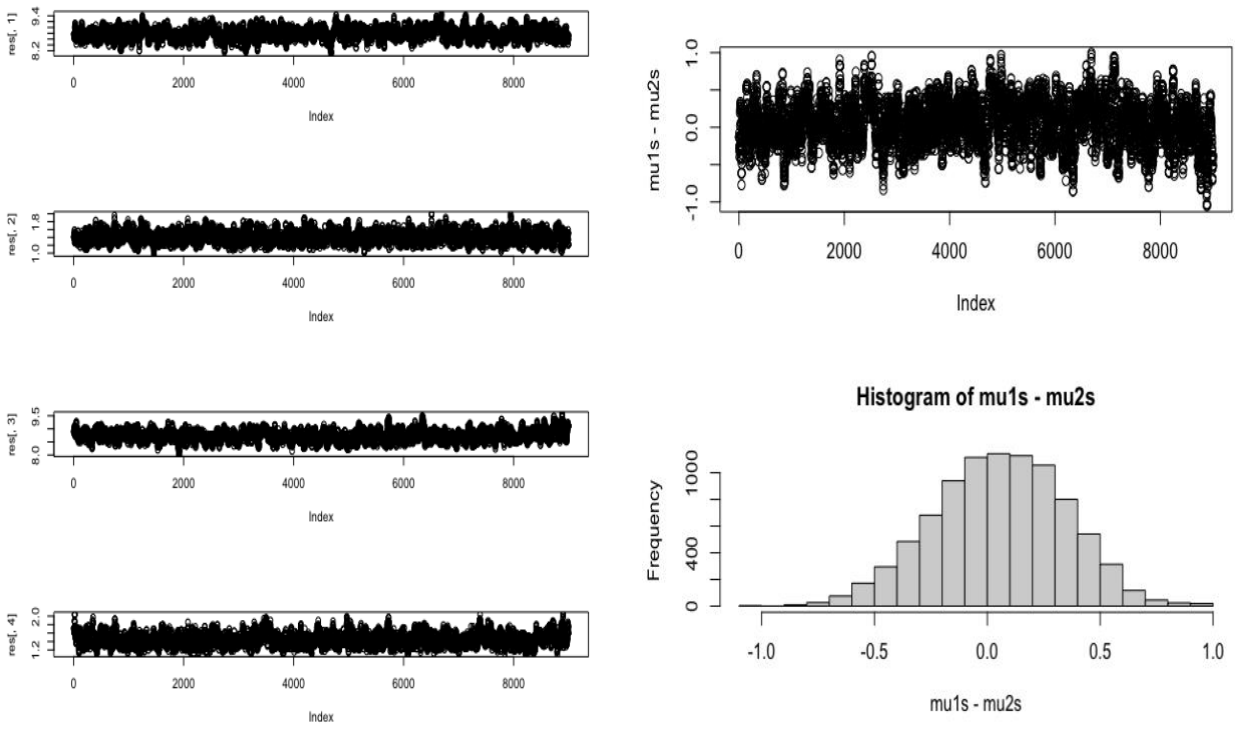
To obtain the posterior distribution, I will multiply the likelihood by the prior distribution (following from Bayesian Theorem), I will then run a Markov chain of 10000 iterations to simulate a sample from the distribution. For this iteration, I need to state the starting vector. The vector is going to be the hypothesized means and standard deviations. Thus,

$$\theta_0 = (8.762, 8.762, 8.757, 8.757)$$

Let's us observe how the chain ran:



As before, I need to remove the first few iterations (burn-in period) as they seem to affect the results. I will do the analysis and inference on the remaining iterations.

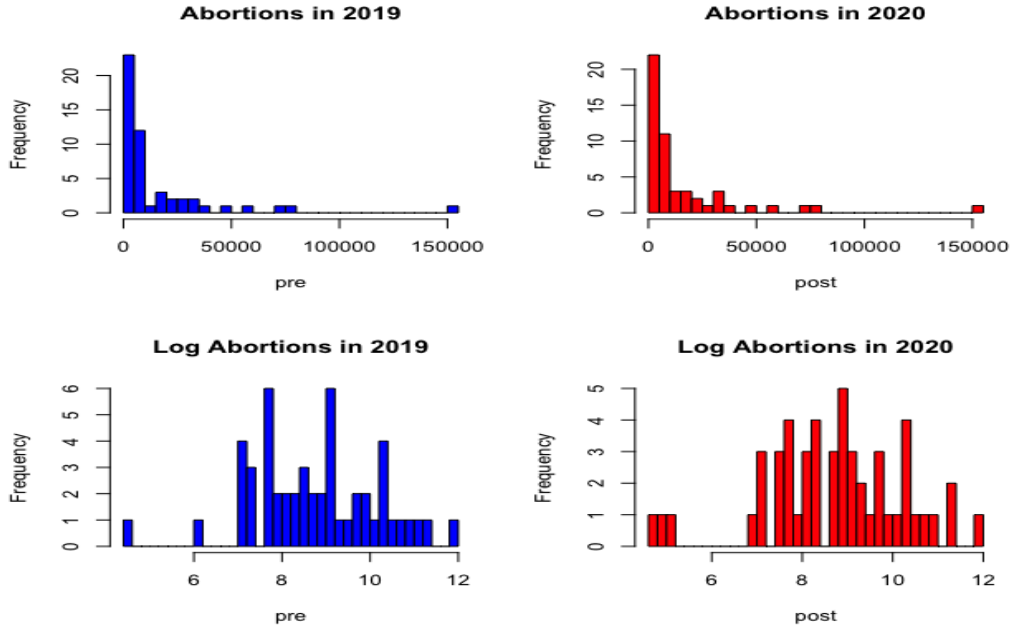


The above plot shows the trace-plots after the burn-in period have been removed.

Thus, $P(\log(\mu_1) - \log(\mu_2) < 0) = P\left(\log\left(\frac{\mu_1}{\mu_2}\right) < 0\right) = P(\mu_1 < \mu_2) \approx 0.5231111$ – the probability that the number of abortions in 2017 is less than the number of abortions in 2018. This is not also surprising because the two samples mean are very close to each other.

2019-2020 Abortion Rate Analysis

This is the histogram of the 2019 and 2020 abortion data, as well as the log-transformed data.



I can see the initial data is right skewed which made me do the transformation and it's also clear that the transformation made the data normal.

Just like before, I'm going to assume that the log of the number of abortions in 2019 and 2020 are modeled by a normal distribution with $N(\mu, \sigma^2)$. Thus, the log(abortion rates) in 2019 comes from a $N(\mu_1, {\sigma_1}^2)$ and the log(abortion rates) in 2020 comes from a $N(\mu_2, {\sigma_2}^2)$. So, the parameter becomes $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2)$. Thus,

$$P(data \mid \theta) = P(log2019 \mid \theta)\, P(\log 2020 \mid \theta)$$
$$= \prod_{i=1}^{51} N(\log 2019_i \mid \mu_1, {\sigma_1}^2) \prod_{i=1}^{51} N(\log 2020_i \mid \mu_2, {\sigma_2}^2)$$

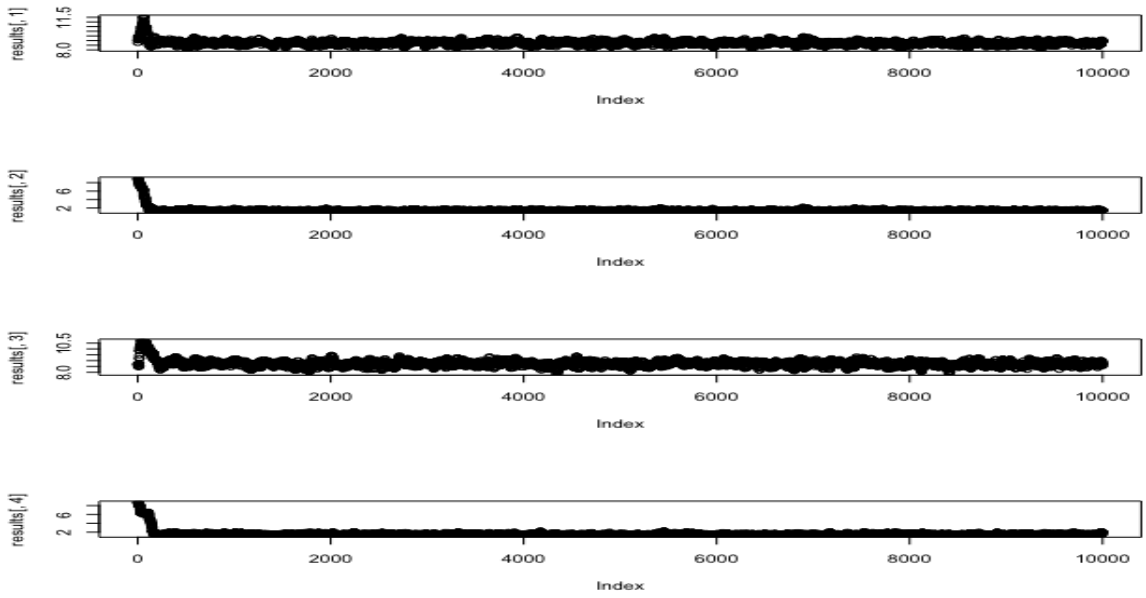
So, the distribution of μ_1 is $N(8.749, 8.749^2)$ and the distribution of μ_2 is $N(8.712, 8.712^2)$. I also need to determine the prior for the means that does not presume which mean is bigger and to choose priors that are uniform over all plausible values. Let's assume that the distribution of σ_1 is exponential with mean 8.749 and I choose the distribution of σ_2 as exponential also with mean 8.712. Since all four variables are independent, I can calculate the prior as:

$$P(\mu_1, \mu_2, \sigma_1, \sigma_2)$$
$$= \frac{1}{\sqrt{2\pi}} \times (8.749)^{-0.5} \times e^{\frac{1}{2(8.749)^2}(\mu_1 - 8.749)^2} \times \frac{1}{\sqrt{2\pi}} \times (8.712)^{-0.5}$$
$$\times\ e^{\frac{1}{2(8.712)^2}(\mu_2 - 8.712)^2} \times \frac{1}{8.749} e^{\frac{-\sigma_1}{8.749}} \times \frac{1}{8.712} e^{\frac{-\sigma_2}{8.712}}$$

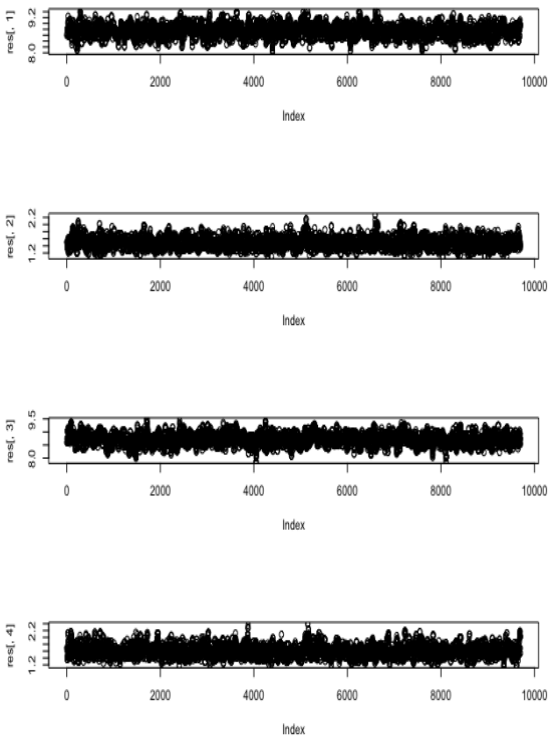
To obtain the posterior distribution, I will multiply the likelihood by the prior distribution (following from Bayesian Theorem), I will then run a Markov chain of 10000 iterations to simulate a sample from the distribution. For this iteration, I need to state the starting vector. The vector is going to be the hypothesized means and standard deviations. Thus,

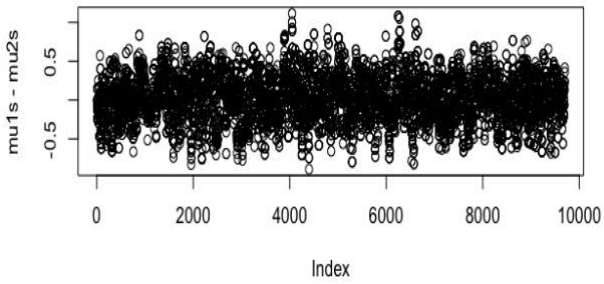
$$\theta_0 = (8.749, 8.749, 8.712, 8.712)$$

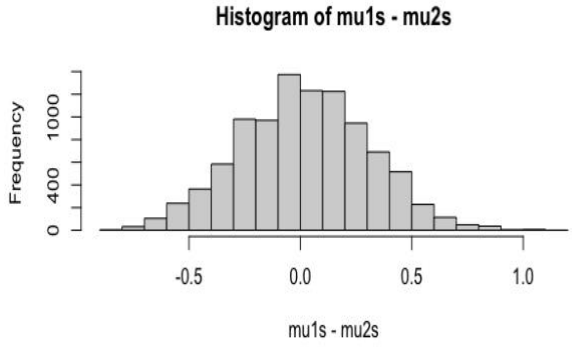
Let's us observe how the chain ran:



As before, I need to remove the first few iterations (burn-in period) as they seem to affect the results. I will do the analysis and inference on the remaining iterations.







The above plot shows the trace-plots after the burn-in period have been removed.
Thus, $P(\log(\mu_1) - \log(\mu_2) < 0) = P\left(\log\left(\frac{\mu_1}{\mu_2}\right) < 0\right) = P(\mu_1 < \mu_2) \approx 0.4794845$ – the probability that the number of abortions in 2019 is less than the number of abortions in 2020. This is not also surprising because the two samples mean are very close to each other.

Conclusion

Based on the results of the analysis above, there's no conclusive evidence to validate my claim that the abortion rate in the United Stated of America is high. The abortion rates of 2015 were probably higher than the abortion rate in 2016. This result is corroborated by the Center for Disease Control (CDC) abortion report in 2016. It says "from 2015 to 2016, the total number of reported abortions decreased 2% (from 636,902), the abortion rate decreased 2% (from 11.8 abortions per 1,000 women aged 15–44 years), and the abortion ratio decreased 1% (from 188 abortions per 1,000 live births)" **[2].** Similarly, the abortion rates of 2017 were probably lesser than the abortion rates in 2018. This result is also corroborated by the CDC abortion report in 2018. It says "from 2017 to 2018, the total number of abortions and abortion rate increased 1% (from 609,095 total abortions and from 11.2 abortions per 1,000 women aged 15–44 years, respectively), and the abortion ratio increased 2% (from 185 abortions per 1,000 live births)" **[3].** Lastly, the abortion rate of 2019 is probably higher than the abortion rate in 2020. This conclusion is also corroborated by the CDC in the 2020 abortion report. It says "from 2019 to 2020, the total number of abortions decreased 2% (from 625,346 total abortions), the abortion rate decreased 2% (from 11.4 abortions per 1,000 women aged 15–44 years), and the abortion ratio increased 2% (from 195 abortions per 1,000 live births)" **[4].** As the years go by and more data become available, I can retest my initial claim

References

[1] Abortion statistics and other data. (n.d.). https://www.johnstonsarchive.net/policy/abortion/index.html, accessed 11th November 2022,

[2] Jatlaoui, T. C. (2020, December 2). Abortion Surveillance — United States, 2016. Centers for Disease Control and Prevention. https://www.cdc.gov/mmwr/volumes/68/ss/ss6811a1.htm

[3] Kortsmit, K. (2021, December 7). Abortion Surveillance — United States, 2018. Centers for Disease Control and Prevention. https://www.cdc.gov/mmwr/volumes/69/ss/ss6907a1.htm

[4] Kortsmit, K. (2022, November 23). Abortion Surveillance — United States, 2020. Centers for Disease Control and Prevention. https://www.cdc.gov/mmwr/volumes/71/ss/ss7110a1.htm

[5] Dr. Brian Pidgeon Project Samples and Lecture Notes

Appendix (R Code)

#My code for Comp Stats. Project

```
setwd("/Users/victoragboli/Documents/Fall 22 Semester/Computational
Statistics/Projects/Comp Stats Project")
data1 = read.table("Abortion Data.txt")

#check data type
str(data1)

#2015 vs. 2016 Analysis
pre = data1[,1]
post = data1[,2]

#histogram
par(mfrow=c(2,2))
hist(pre,50,main="Abortions in 2015",col="blue")
hist(post,50,main="Abortions in 2016",col="red")

#log transform
pre = log(pre)
post = log(post)

mean(pre)
mean(post)

#histogram
hist(pre,50,main="Log Abortions in 2015",col="blue")
hist(post,50,main="Log Abortions in 2016",col="red")

#Likelihood Function
like=function(th){
  mu1=th[1]; sig1=th[2]; mu2=th[3]; sig2=th[4]
  prod(dnorm(pre, mean=mu1,sd=sig1))*prod(dnorm(post,mean=mu2,sd=sig2))
}

#prior Distribution
Prior=function(th){
  mu1=th[1]; sig1=th[2]; mu2=th[3]; sig2=th[4]
  if (sig1<=0 | sig2<=0) return(0)

dnorm(mu1,8.767,8.767)*dnorm(mu2,8.768,8.768)*dexp(sig1,rate=1/8.767)*dexp(sig2,rate=
1/8.767)
}

#posterior
Posterior=function(th){Prior(th)*like(th)}
```

```
#starting
mu1=8.767; sig1=8.767; mu2=8.768; sig2=8.768
th0=c(mu1,sig1,mu2,sig2)
nit=1000000
results=matrix(0,nrow=nit,ncol=4)
th=th0
results[1,]=th0
for (it in 2:nit){
  Cand=th + rnorm(4,sd=.003)
  ratio=Posterior(Cand)/Posterior(th)
  if (runif(1) < ratio) th=Cand
  results[it,]=th
}

#edit(results)
#getting the trace-plot
par(mfrow=c(4,1))
plot(results[,1])
plot(results[,2])
plot(results[,3])
plot(results[,4])


#removing the burns from the traceplots
res=results[2.7e+05:1e+06,]
par(mfrow=c(4,1))
plot(res[,1])
plot(res[,2])
plot(res[,3])
plot(res[,4])

mu1s=res[,1]
sig1s=res[,2]
mu2s=res[,3]
sig2s=res[,4]

par(mfrow=c(2,1))
plot(mu1s-mu2s)
hist(mu1s-mu2s)
mean(mu1s-mu2s<0)
```

#2017 vs. 2018 Analysis

```
#clear data before running
pre = data1[,3]
post = data1[,4]

#histogram
par(mfrow=c(2,2))
hist(pre,50,main="Abortions in 2017",col="blue")
hist(post,50,main="Abortions in 2018",col="red")
```

```
#log transform
pre = log(pre)
post = log(post)

#histogram of log
hist(pre,50,main="Log Abortions in 2017",col="blue")
hist(post,50,main="Log Abortions in 2018",col="red")


mean(pre)
mean(post)

#Likelihood Function
like=function(th){
  mu1=th[1]; sig1=th[2]; mu2=th[3]; sig2=th[4]
  prod(dnorm(pre, mean=mu1,sd=sig1))*prod(dnorm(post,mean=mu2,sd=sig2))
}

#prior Distribution
Prior=function(th){
  mu1=th[1]; sig1=th[2]; mu2=th[3]; sig2=th[4]
  if (sig1<=0 | sig2<=0) return(0)

dnorm(mu1,8.762,8.762)*dnorm(mu2,8.757,8.757)*dexp(sig1,rate=1/8.762)*dexp(sig2,rate=
1/8.757)
}

#posterior
Posterior=function(th){Prior(th)*like(th)}

#starting
mu1=8.762; sig1=8.762; mu2=8.757; sig2=8.757
th0=c(mu1,sig1,mu2,sig2)
nit=10000
results=matrix(0,nrow=nit,ncol=4)
th=th0
results[1,]=th0
for (it in 2:nit){
  Cand=th + rnorm(4,sd=.1)
  ratio=Posterior(Cand)/Posterior(th)
  if (runif(1) < ratio) th=Cand
  results[it,]=th
}


par(mfrow=c(4,1))
plot(results[,1])
plot(results[,2])
plot(results[,3])
plot(results[,4])

res=results[1001:10000,]
par(mfrow=c(4,1))
```

```
plot(res[,1])
plot(res[,2])
plot(res[,3])
plot(res[,4])

mu1s=res[,1]
sig1s=res[,2]
mu2s=res[,3]
sig2s=res[,4]
par(mfrow=c(2,1))
plot(mu1s-mu2s)
hist(mu1s-mu2s)
mean(mu1s-mu2s<0)
```

#2019 vs. 2020 Analysis

```
#remember to clear data before running
pre=data1[,5]
post=data1[,6]

#histogram
par(mfrow=c(2,2))
hist(pre,50,main="Abortions in 2019",col="blue")
hist(post,50,main="Abortions in 2020",col="red")

#log-transform
pre=log(pre)
post=log(post)

#log histogram
hist(pre,50,main="Log Abortions in 2019",col="blue")
hist(post,50,main="Log Abortions in 2020",col="red")

#mean
mean(pre)
mean(post)

#Likelihood Function
like=function(th){
  mu1=th[1]; sig1=th[2]; mu2=th[3]; sig2=th[4]
  prod(dnorm(pre, mean=mu1,sd=sig1))*prod(dnorm(post,mean=mu2,sd=sig2))
}

#prior Distribution
Prior=function(th){
  mu1=th[1]; sig1=th[2]; mu2=th[3]; sig2=th[4]
  if (sig1<=0 | sig2<=0) return(0)

dnorm(mu1,8.749,8.749)*dnorm(mu2,8.712,8.712)*dexp(sig1,rate=1/8.749)*dexp(sig2,rate=
1/8.712)
}

#posterior
```

```
Posterior=function(th){Prior(th)*like(th)}

#starting
mu1=8.749; sig1=8.749; mu2=8.712; sig2=8.712
th0=c(mu1,sig1,mu2,sig2)
nit=10000
results=matrix(0,nrow=nit,ncol=4)
th=th0
results[1,]=th0
for (it in 2:nit){
  Cand=th + rnorm(4,sd=.2)
  ratio=Posterior(Cand)/Posterior(th)
  if (runif(1) < ratio) th=Cand
  results[it,]=th
}

#edit(results)
#traceplot
par(mfrow=c(4,1))
plot(results[,1])
plot(results[,2])
plot(results[,3])
plot(results[,4])

#removing burnout
res=results[301:10000,]
par(mfrow=c(4,1))
plot(res[,1])
plot(res[,2])
plot(res[,3])
plot(res[,4])

#prob
mu1s=res[,1]
sig1s=res[,2]
mu2s=res[,3]
sig2s=res[,4]
par(mfrow=c(2,1))
plot(mu1s-mu2s)
hist(mu1s-mu2s)
mean(mu1s-mu2s<0)

#visualising of the data
par(mfrow=c(3,2))
plot(data1[,1], type = "l", main = "Abortions in 2015")
plot(data1[,2], type = "l", main = "Abortions in 2016")
plot(data1[,3], type = "l", main = "Abortions in 2017")
plot(data1[,4], type = "l", main = "Abortions in 2018")
plot(data1[,5], type = "l", main = "Abortions in 2019")
plot(data1[,6], type = "l", main = "Abortions in 2020")
```