Victor Agboli

STAT 8310 – Applied Bayesian Statistics

Drug Overdose Deaths Analysis using Bayesian Inference

Introduction

In April, 2021, Rochelle P. Walensky, MD, MPH, a CDC Director said "Overdose deaths continue to rise across the United States. To bring an end to this crisis, collaborative public health and public safety solutions are needed within communities". In this project, I will use the Markov Chain Monte Carlo methods to analyze this growth claim. I will study the difference in drug overdose deaths in the United States of America from 2017 – 2020 and determine if there is a significant increase or decrease between successive years. I will use this analysis to validate or invalidate the claim by Dr. Walensky. The data for this analysis is found on the Centers for Disease Control and Prevention (CDC) website [1].

The data tabulates the number of deaths caused by drug overdose in the 50 states in the United States of America plus District of Columbia by opioids (synthetic and non-synthetic), psychostimulants such as methamphetamine, fentanyl, etc.

The questions I will answer are:

1. Is there a difference in the mean number of drug overdose deaths in the USA from 2017 to 2018?

2. Is there a difference in the mean number of drug overdose deaths in the USA from 2019 to 2020?

Methods

I will examine the histograms of the drug overdose deaths per year to look at 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 the prior and likelihood distributions have been established, I will create a posterior distribution. With this posterior distribution, I will run a Markov chain using Metropolis. From the analysis, I will find an estimated posterior probability to determine the probability that the drug overdose deaths increased.

Procedure

1. Define the parameters for the prior. I am going to assume that the log of the number of drug overdose deaths in 2017 and 2018 are modeled by a normal distribution with $N(\mu, \sigma^2)$. Thus, the log of drug overdose deaths for 2017 comes from $N(\mu_1\ ,\ {\sigma^2}_1)$ and the log of drug overdose deaths for 2018 comes from $N(\mu_2\ ,\ {\sigma^2}_2)$.

   The μ_1 follows $N(\mu_{1_0}, {\sigma^2}_{1_0})$. Similar, the μ_2 follows $N(\mu_{2_0}, {\sigma^2}_{2_0})$. Likewise, σ_1 follows an exponential distribution with mean (μ_{1_0}), σ_2 follows an exponential distribution with mean (μ_{2_0}).

   Thus, the prior function parameter vector should be $\theta = [\mu_1,\ \sigma_1,\ \mu_2,\ \sigma_2]$.

   The prior density is specified as follows:

$$Prior = N\left(\mu_{1_0}, {\sigma^2}_{1_0}\right) \times \exp\left(\mu_{1_0}\right) \times N\left(\mu_{2_0}, {\sigma^2}_{2_0}\right) \times \exp\left(\mu_{2_0}\right)$$

2. Define the parameters for likelihood.

$$P(data|\,\theta) = P(log\ Data_{2017}\,|\,\theta) \times\ P(log\ Data_{2018}\,|\,\theta) =$$

$$\prod_{i=1}^{51} N\left(\log Data_{2017_i}\,\middle|\mu_1,\ {\sigma^2}_1\right) \times \prod_{j=1}^{51} N\left(\log Data_{2018_j}\,\middle|\mu_2,\ {\sigma^2}_2\right)$$

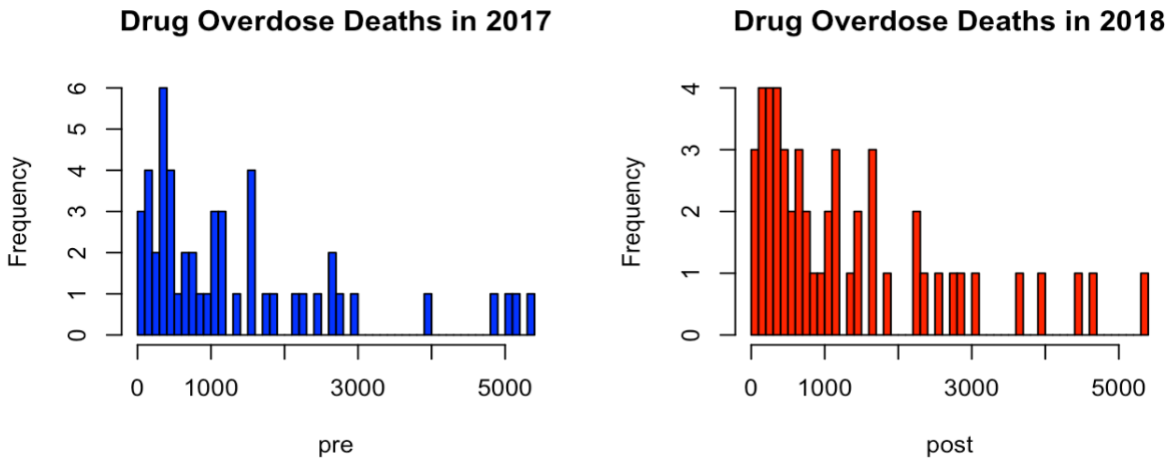
Where

- $Log\ Data_{2017}$ is the log transform of the drug overdose deaths in 2017.
- $Log\ Data_{2018}$ is the log transform of the drug overdose deaths in 2018.

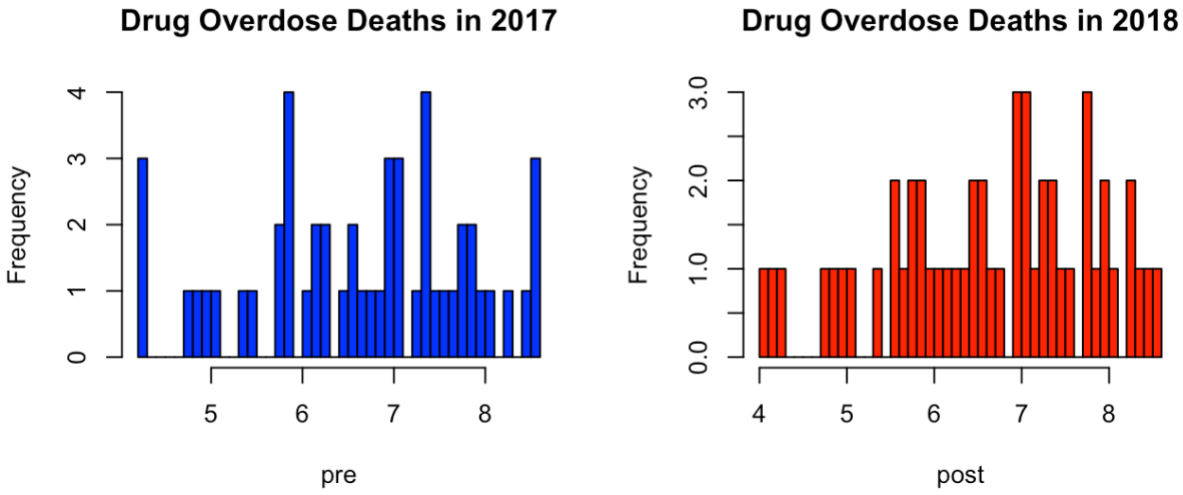
3. Generate the posterior function. The posterior function is proportional to the product of the prior and likelihood functions.
4. Decide the starting state of my prior distribution and run the MCMC models.
5. Accept or partial accept the new position based on the Metropolis-Hastings rule.
6. Run the iteration.
7. Reach a conclusion based on the analysis.
8. Repeat step 1 – 7, using the drug overdose deaths data for 2019 and 2020.
9. Make recommendations based on the conclusion in step 7.

2017 – 2018 Drug Overdose Deaths Analysis

Below are the histograms for the drug overdose deaths in 2017 and 2018 respectively



Examining the histograms, both are skewed to the right. Since I want the likelihood functions normal, I will take the log transform of the drug overdose deaths in 2017 and 2018.



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 drug overdose deaths in 2017 and 2018 is modeled by a normal distribution with $N(\mu,\ \sigma^2)$. Thus, the log (drug overdose deaths) in 2017 comes from $N(\mu_1,\ {\sigma_1}^2)$ and log (drug overdose deaths) in 2018 comes from $N(\mu_2,\ {\sigma_2}^2)$. So, the parameter becomes $\theta = (\mu_1,\ \mu_2,\ \sigma_1,\ \sigma_2)$. Thus,

$$P(data \mid \theta) = P(log\ Data_{2017} \mid \theta) \times P(log\ Data_{2018} \mid \theta) =$$

$$\prod_{i=1}^{51} N\left(\log Data_{2017_i} \middle| \mu_1,\ {\sigma^2}_1\right) \times \prod_{j=1}^{51} N\left(\log Data_{2018_j} \middle| \mu_2,\ {\sigma^2}_2\right)$$

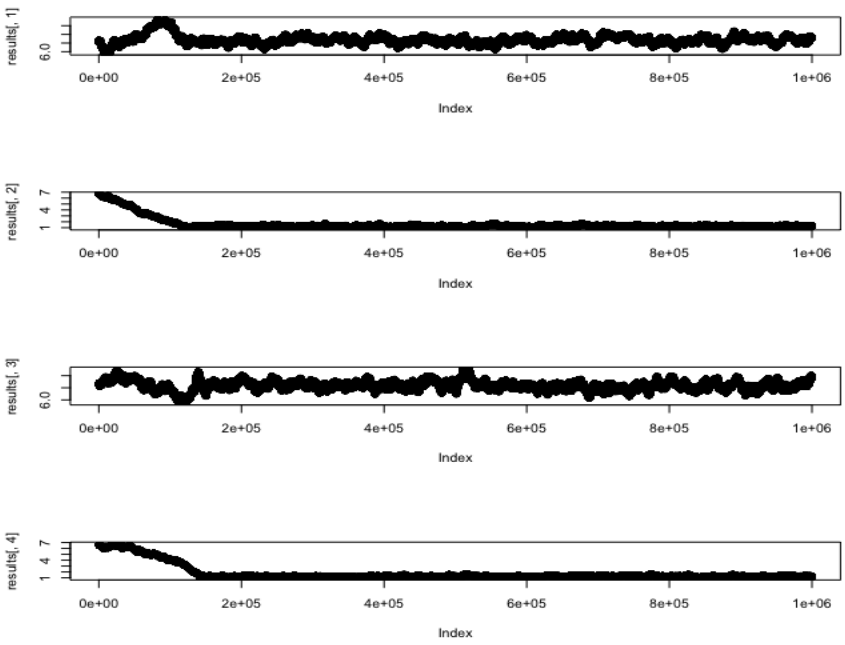
So, the distribution of μ_1 is $N(6.658, 6.658^2)$ and the distribution of μ_2 is $N(6.622, 6.622^2)$. I also need to determine the prior for the means that does not presume which mean is bigger and to choose priors that are uniform over all plausible values. Let's assume that the distribution of σ_1 is exponential with a mean of 6.658 and I choose the distribution of σ_2 as exponential with a mean of 6.622. Since all four variables are independent, I can calculate the prior as:

$$P\ (\mu_1, \mu_2, \sigma_1,\ \sigma_2)$$
$$= \frac{1}{\sqrt{2\pi}} \times (6.658)^{-0.5} \times e^{\frac{1}{2(6.658)^2}(\mu_1 - 6.658)^2} \times \frac{1}{\sqrt{2\pi}} \times (6.622)^{-0.5}$$
$$\times\ e^{\frac{1}{2(6.622)^2}(\mu_1 - 6.622)^2} \times \frac{1}{6.658} e^{\frac{-\sigma_1}{6.658}} \times \frac{1}{6.622} e^{\frac{-\sigma_2}{6.622}}$$

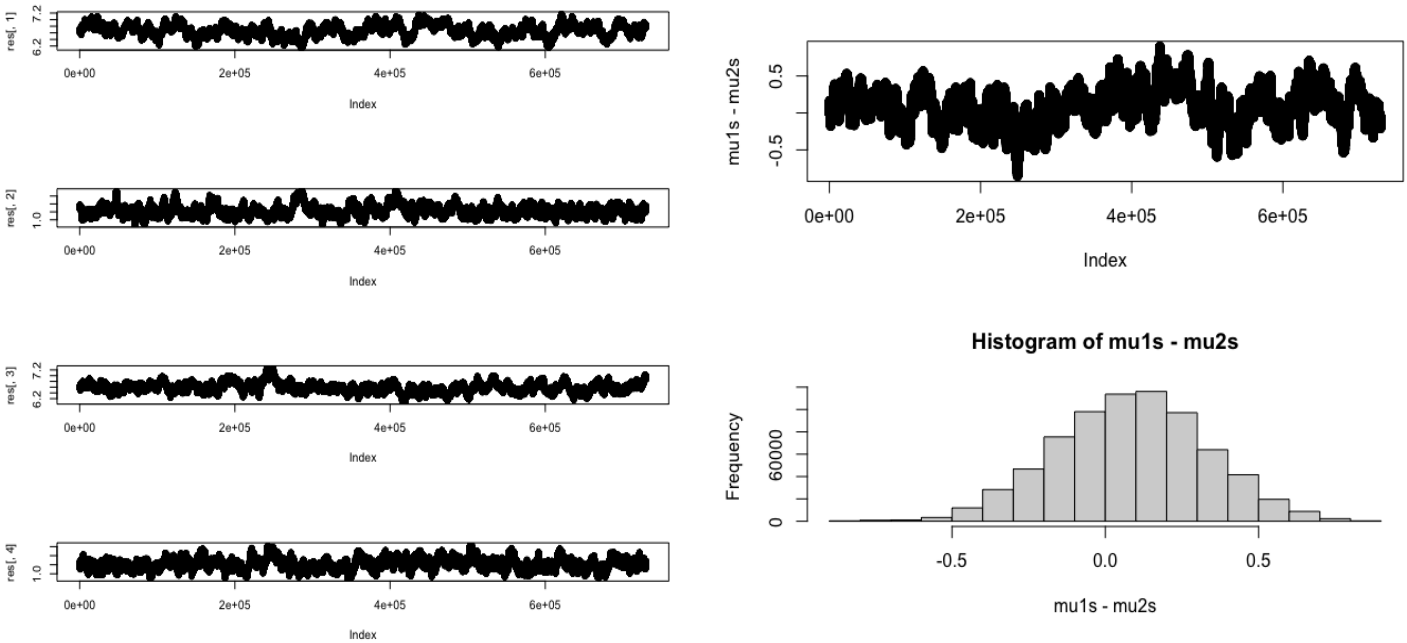
To obtain the posterior distribution, I will multiply the likelihood by the prior distribution (following Bayesian Theorem), I will then run a Markov Chain of 1000000 iterations to simulate a sample from the distribution. I choose this large number because 1000, 10000, 100000 did not converge based on 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,

$$\theta_0 = (6.658, 6.658, 6.622, 6.622)$$

Let's us observe how the chain ran:



Looking at the chain, I can see that the first few iterations affect 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.

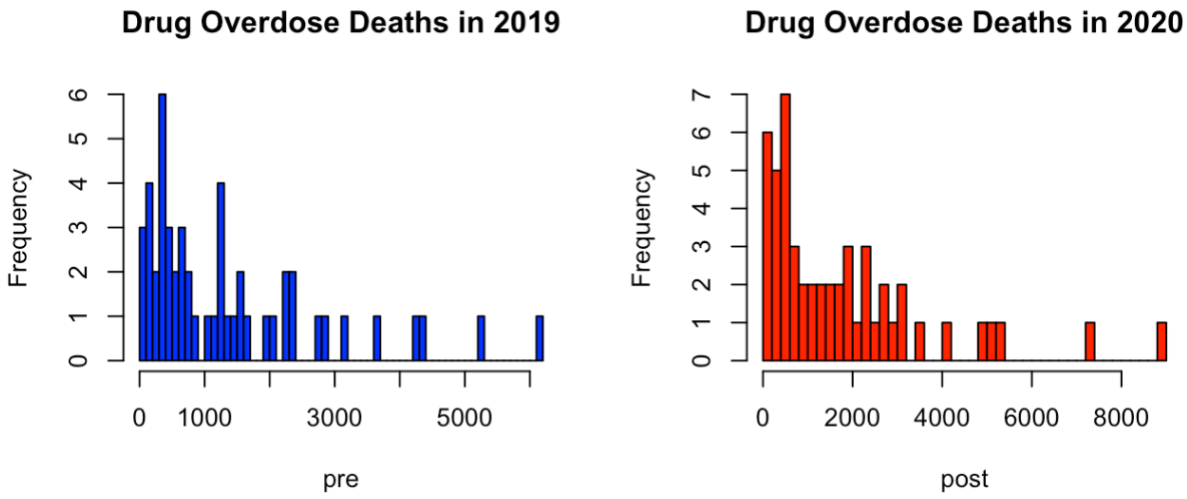


The above plot shows the trace-plots after the burn-in period has been removed.

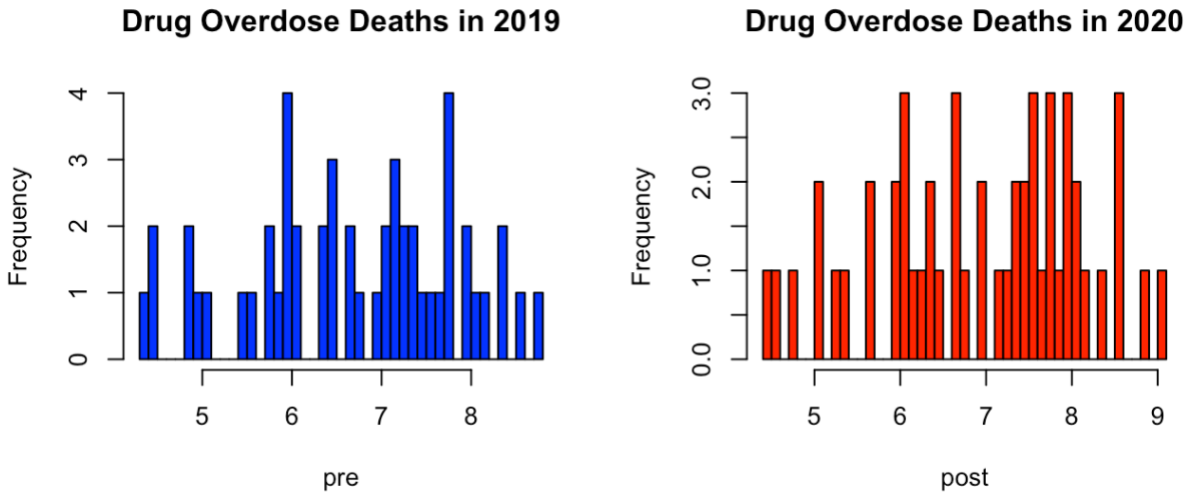
Thus, $P(\log(\mu_1) - \log(\mu_2) < 0) = P(\log(\frac{\mu_1}{\mu_2}) < 0) = P(\mu_1 < \mu_2) \approx 0.3203255$ – the probability that the drug overdose deaths in 2017 is less than the drug overdose deaths in 2018.

2019 – 2020 Drug Overdose Deaths Analysis

Below are the histograms of the drug overdose deaths in 2019 and 2020.



Both distributions are skewed to the right. Like before, I will take the log transform of 2019 and 2020 drug overdose deaths data to normalize them.



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

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

$$P(data \mid \theta) = P(log\ Data_{2019} \mid \theta) \times P(log\ Data_{2020} \mid \theta) =$$

$$\prod_{i=1}^{51} N\left(\log Data_{2019_i} \left| \mu_1,\ {\sigma^2}_1\right.\right) \times \prod_{j=1}^{51} N\left(\log Data_{2020_j} \left| \mu_2,\ {\sigma^2}_2\right.\right)$$

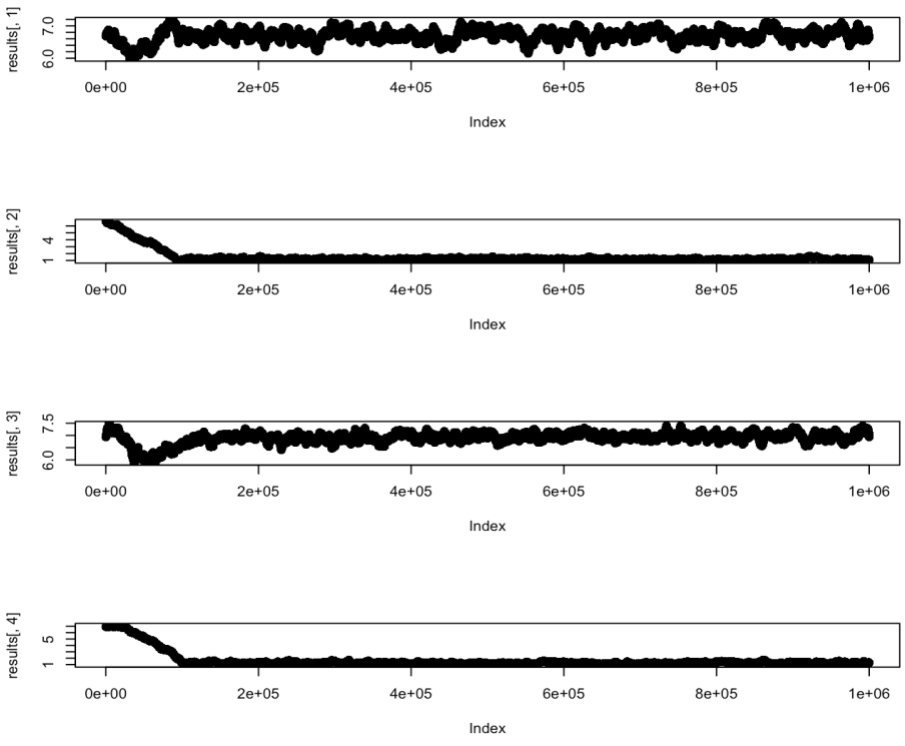
So, the distribution of μ_1 is $N(6.687, 6.687^2)$ and the distribution of μ_2 is $N(6.929, 6.929^2)$. I also need to determine the prior for the means that does not presume which mean is bigger and to choose priors that are uniform over all plausible values. Let's assume that the distribution of σ_1 is exponential with a mean of 6.687 and I choose the distribution of σ_2 as exponential with a mean of 6.929. Since all four variables are independent, I can calculate the prior as:

$$P\ (\mu_1, \mu_2, \sigma_1,\ \sigma_2)$$
$$= \frac{1}{\sqrt{2\pi}} \times (6.687)^{-0.5} \times e^{\frac{1}{2(6.687)^2}(\mu_1 - 6.687)^2} \times \frac{1}{\sqrt{2\pi}} \times (6.929)^{-0.5}$$
$$\times\ e^{\frac{1}{2(6.929)^2}(\mu_1 - 6.929)^2} \times \frac{1}{6.687} e^{\frac{-\sigma_1}{6.687}} \times \frac{1}{6.929} e^{\frac{-\sigma_2}{6.929}}$$

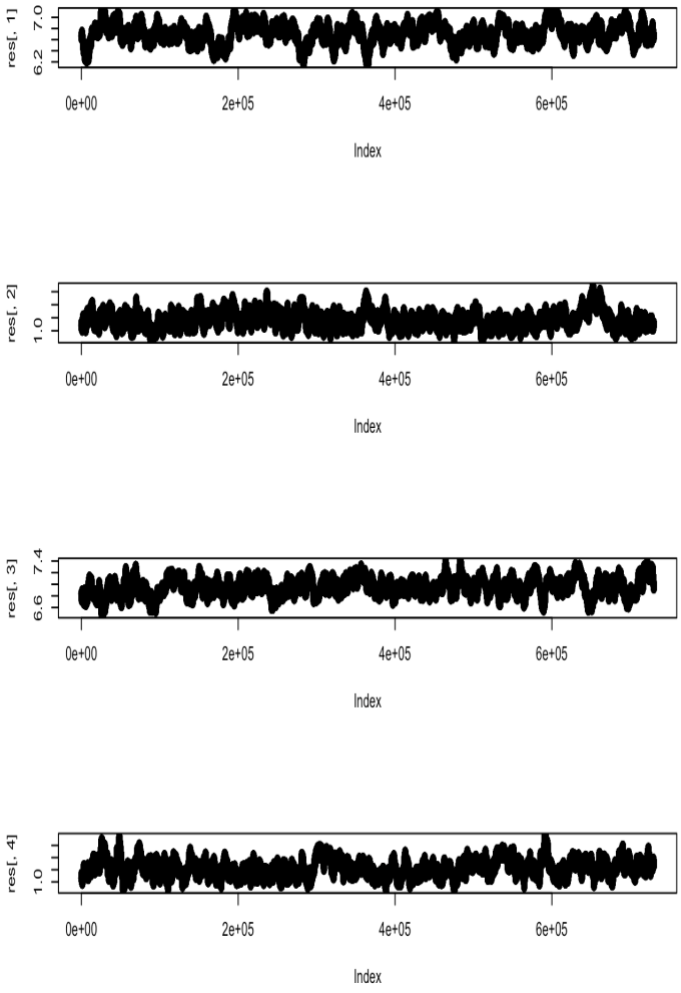
To obtain the posterior distribution, I will multiply the likelihood by the prior distribution (following Bayesian Theorem), I will then run a Markov Chain of 1000000 iterations to simulate a sample from the distribution. I choose this large number because 1000, 10000, 100000 did not converge based on 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,

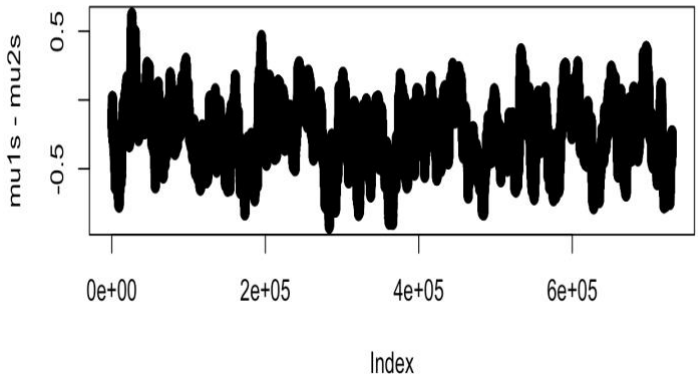
$$\theta_0 = (6.687, 6.687, 6.929, 6.929)$$

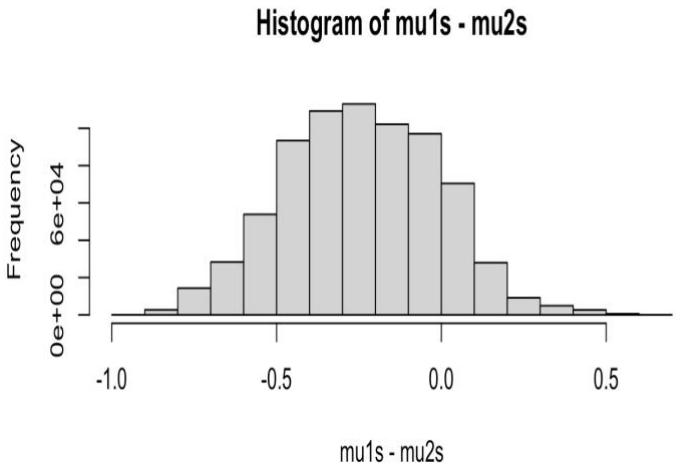
Let's us observe how the chain ran:



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, $P(\log(\mu_1) - \log(\mu_2) < 0) = P(\log(\frac{\mu_1}{\mu_2}) < 0) = P(\mu_1 < \mu_2) \approx 0.8413947$ – the probability that the drug overdose deaths in 2019 is less than the drug overdose deaths in 2020.

Conclusion

Based on the results of the analysis above, there's conclusive evidence to validate the claim by Dr. Walensky that the drug overdose deaths in the USA is on the high. The drug overdose deaths were probably higher in 2017 than the drug overdose deaths in 2018. This result is corroborated by the Center for Disease Control (CDC) drug overdose deaths reports in 2020. It says. "In 2018, there were 67,367 drug overdose deaths in the United States, 4.1% fewer deaths than in 2017 (70,237)" [2]. Similarly, the drug overdose deaths in 2019 were probably lower than the drug overdose deaths in 2020. This result is corroborated by the Center for Disease Control (CDC) drug overdose deaths reports in 2022. It says, "The age-adjusted rate of overdose deaths increased by 31% from 2019 (21.6 per 100,000) to 2020 (28.3 per 100,000)" [3].

As the years go by and more data becomes available, I can retest this claim made by Dr. Walensky.

References

[1] Death Rate Maps & Graphs | Drug Overdose | CDC Injury Center. (2022, June 2). https://www.cdc.gov/drugoverdose/deaths/index.html

[2] Drug overdose deaths in the United States, 1999–2018 - CDC. (n.d.). Retrieved April 17, 2023, from https://www.cdc.gov/nchs/data/databriefs/db356-h.pdf

[3] "Death Rate Maps & Graphs." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 2 June 2022, from https://www.cdc.gov/drugoverdose/deaths/index.html#:~:text=The%20age%2Dadjusted%20rate%20of,overdose%20deaths%20involved%20synthetic%20opioids

[4] Dr. Ying Zhang Project Sample and Lecture Notes