Denis Ostroushko - PUBH 7440 - HW1

Problem 1

1 - A

According to a naive Google search, the chance of winning a lottery is about 1 in 300,000,000. We will use this as a prior chance of winning the lottery. We are playing the lottery for the first time, so we do not have data to make a guess about a chance of winning the lottery that is specific to us. But, if you are *feeling lucky*, you might say that if you were to play the lottery, then the chance of winning for you could be about 1 in 100,000,000. This means that the chance of winning the lottery on the fist try for someone who is feeling lucky and plays the lottery for the first time is between 1 in 100,000,000 and 1 in 300,000,000.

1 - B

Suppose that you are a good team that made it to the world series and won 65% of games through the regular season and the play-offs. But, suppose that you play the first game away, and the chance of winning such a game historically has been 45%.

This means that for such a team the chance of winning the first game in the world series will be somehwere between 45% and 65%.

Problem 2

2 - A

According to given parameters, prior distribution of θ is given by $\theta \sim N(0, 2)$

And the distribution of data, random variable Y is also normal, $y|\theta \sim N(\theta, 2)$

Using given derivation of posterior distribution, we obtain the distribution of θ given the observed values of data y.

Going forward, we fix $y = 4$, we let $B = \frac{2}{2+2} = \frac{1}{2}$. Then, $\theta|y \sim N(B \times \mu + (1 - B) \times y, (1 - B) \times \sigma^2) = N(\frac{1}{2} \times 0 + \frac{1}{2} \times 4, \frac{1}{2} \times 2) = N(2, 1)$.

So, $\theta|y \sim N(2, 1)$

To plot likelihood of observing y= 4, I will use the distribution of data $y|\theta$, I will vary the value of θ and keep variance of Y fixed at 2. Since θ is centered at 0 with variance 2, I will obtain values that correspond to prior distribution of θ

Desired plots for prior, likelihood, and posterior are shown in Figure 1

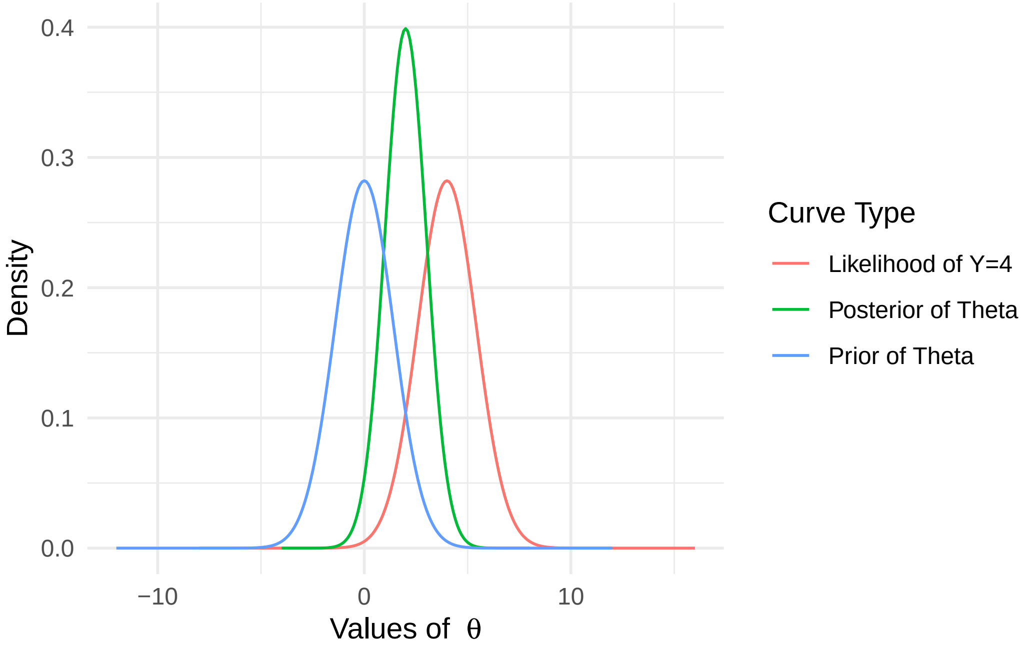


Figure 1: Plots for problem 2-A

2 - B

When $\tau^2 = 18$, $B = \frac{2}{2+18} = \frac{1}{10}$. The resulting distributions are then

$\theta \sim N(0, 18)$

$y|\theta \sim N(\theta, 2)$

Assuming y= 4, $\theta|y \sim N(0.1 * 0 + 0.9 * y, 0.9 * 2) = N(3.6, 1.8)$

Prior, marginal likelihood, and posterior distributions take on different shapes, given in Figure 2

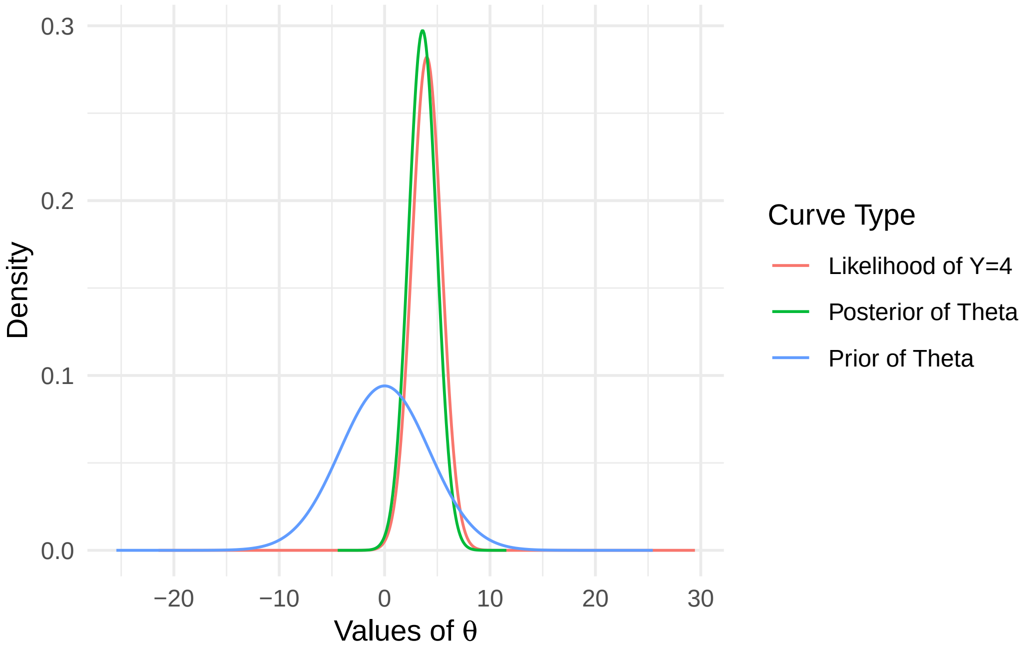


Figure 2: Plots for problem 2-B

The main difference from part A stems from the differences in prior distribution of θ. Larger variance of the prior distribution makes our beliefs about possible values of θ less certain, or less informative.

Figure 1 shows that the 'width' of prior and posterior distributions is similar. We can also see that the posterior distribution is being 'pulled' towards the likelihood of observing y=4. I suspect that due to the higher amount of information contained in the prior distribution, we have an updated distribution of potential values of θ, which still resembles prior.

On the other hand, Figure 2 shows what happens if we set a less informative prior. Visually we can see that once we have observed the value $y=$4, this is our best guess now. I suspect, as we keep sampling values of y, we will have a distribution centered somewhere where the actual population mean for Y would be.

Given this, I suspect that a frequencies statistician would prefer the method we used in part B, because the prior knowledge of θ would not have a strong influence on the final analysis results after sampling more and more values of Y.

Problem 3

3 - A

The chance from picking a ball from bucket 2 is governed by a fair six-sided die. We pick from bucket 2 when we roll 5 or 6, so, $P(Draw\ from\ 2) = \frac{1}{3}$

3 - B

Probability of drawing a blue ball is a weighted average of probabilities that correspond to the bucket that we draw from. $P(Blue) = P(Draw\ from\ 1) \times P(Blue|Draw\ from\ 1) + P(Draw\ from\ 2) \times P(Blue|Draw\ from\ 2) = \frac{2}{3} \times \frac{17}{17+35} + \frac{1}{3} \times \frac{37}{37+23} = 0.42$

3 - C

We need to use Bayes Rule to find this probability from available data $P(Draw\ from\ 2|Blue) = \frac{P(Draw\ from\ 2\ and\ Blue)}{P(Blue)} = \frac{P(Blue|Draw\ from\ 2) \times P(Draw\ from\ 2)}{P(Blue)} = \frac{\frac{37}{37+23} \times \frac{1}{3}}{0.42} = 0.49$

Problem 4

4 - A

Among 10000 draws from estimated posterior distributions, there were 80 cases where the relative frequency of counties with higher than average mortality rate in the Urban class was higher than the relative frequency of counties in the Rural class. This corresponds to the 0.8% chance that the true rate of counties with higher than average mortality in the urban class is above that or counties in the rural class. At the $\alpha = 0.05$ statistical significance level, we can conclude that the the rate of deaths in urban counties must be consistently lower. Distribution of potential differences is shown on Figure 3.

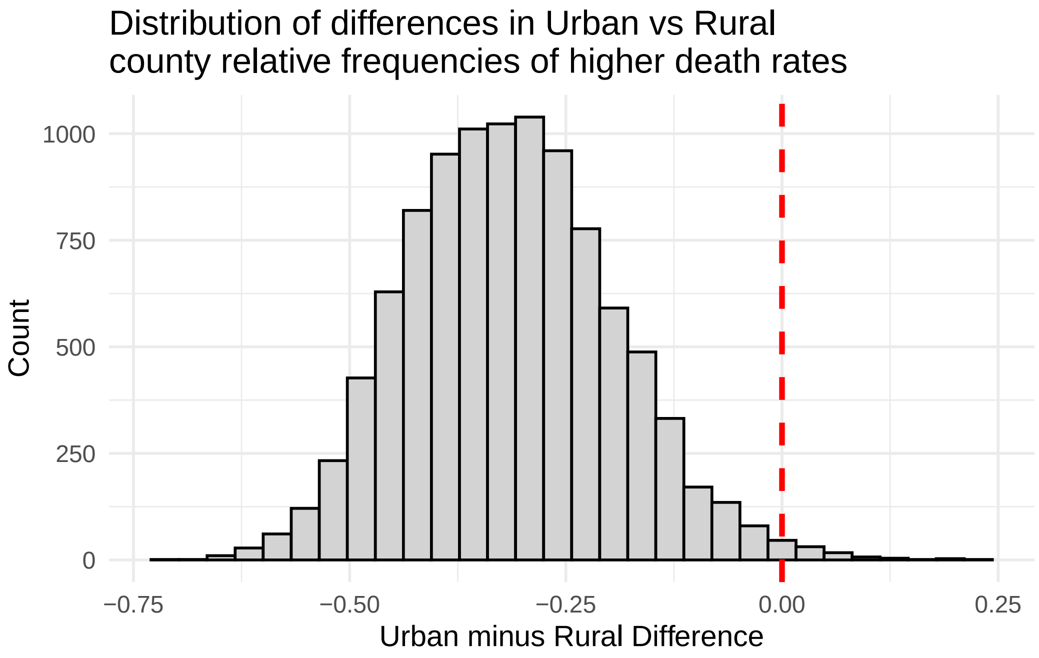


Figure 3: Problem 4-A supporting plot

4 - B

For this problem, I used data driven approach. I estimated death rate for the entire state of PA from the data, and set it to the deaths per 1,000 scale.

Among 10000 draws from estimated posterior distributions, there were 842 cases where the death rate per 1,000 people in the urban counties was higher than the death rate per 1,000 in the rural counties. This corresponds to the 8.42% chance that the true rate of deaths per 1,000 is higher in the urban counties is higher than that of rural counties.

While this chance is higher than the traditional 5% frequentist cut off rate for the significance level, estimated chance is still fairly slim. Distribution of differences supporting our decision is on Figure 4.

4 - C

The two analyses yield similar results. Both lead to the conclusion that the death rates associated with the heart disease are lower in the urban counties when compared with the rural counties.

I think that analysis employed in part B is more appropriate, as it looks at the actual death rates, because this analysis reflects the reality more accurately.

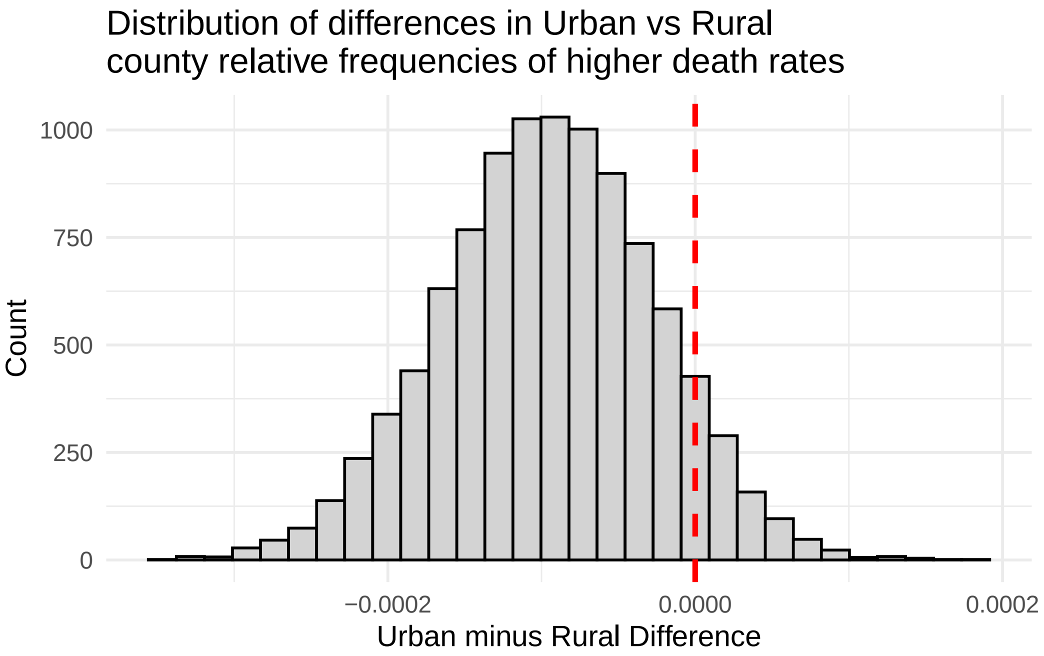


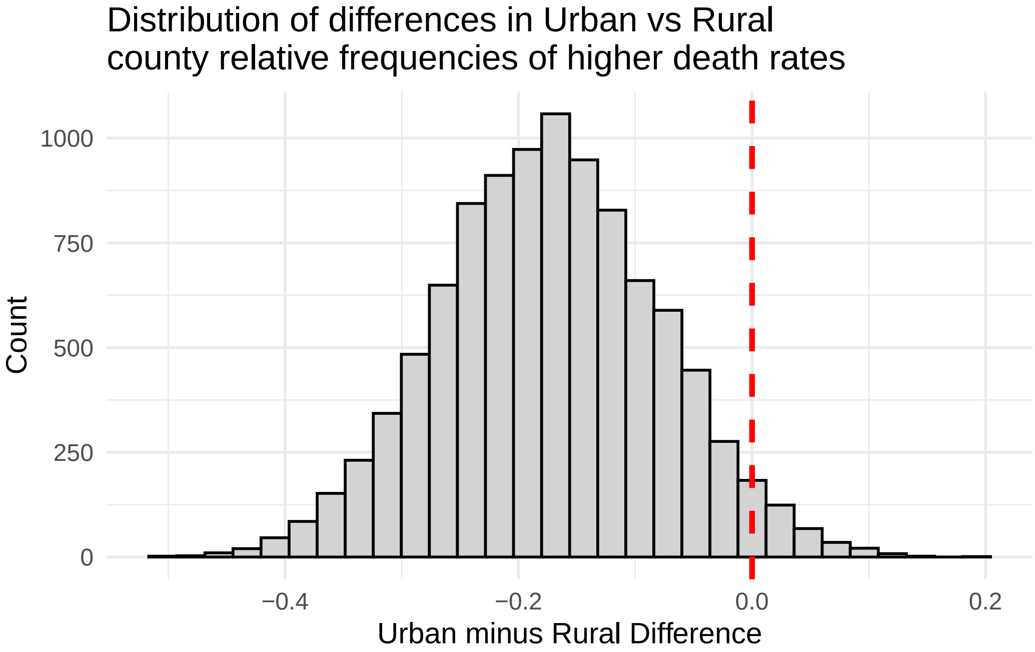
Figure 4: Problem 4-B supporting plot

While the binomial model tells us that there are less counties with less than average death rates in the urban category, less extreme conclusions from the Poisson model tell us that those differences to the average death rate must be small.

Appendix

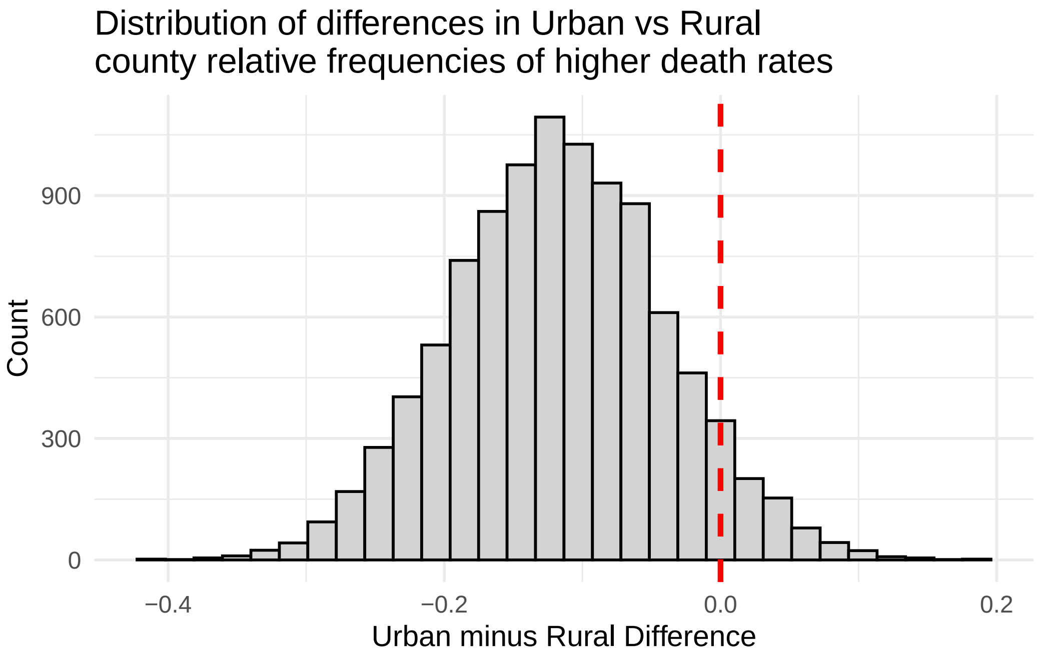
In this section I reproduce results from problem A with different priors. I want to see the impact of prior choice on the analysis results.

Increasing A, B to 12, 12 for 4-A



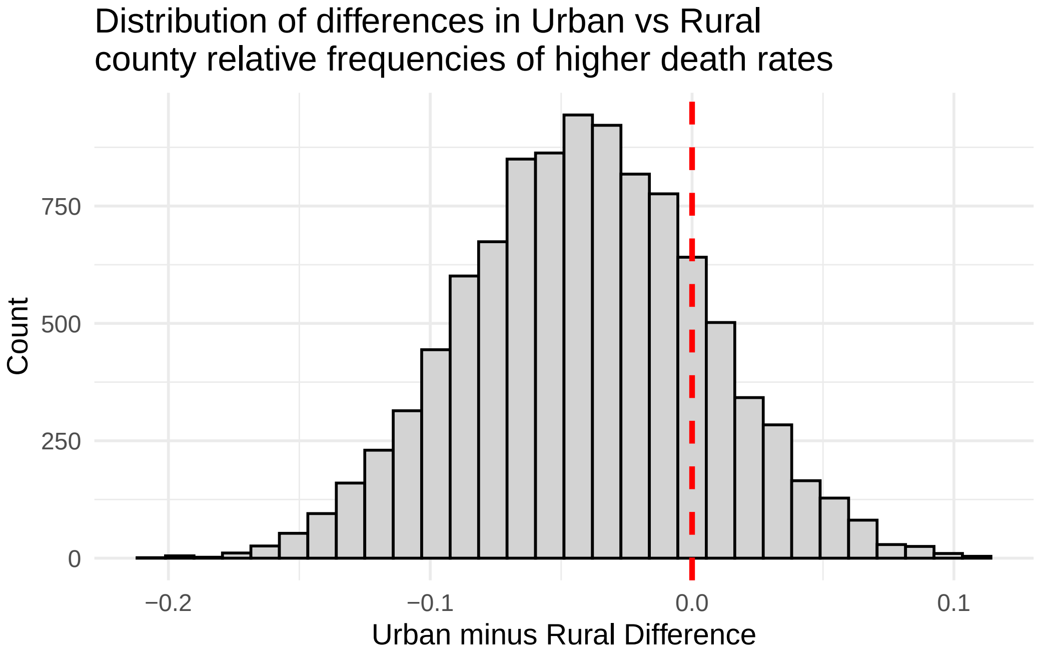
- New chance: 3.37%
- Estimated Difference: -0.17
- Estimated SD of distribution: 0.09

Increasing A, B to 25, 25 for 4-A



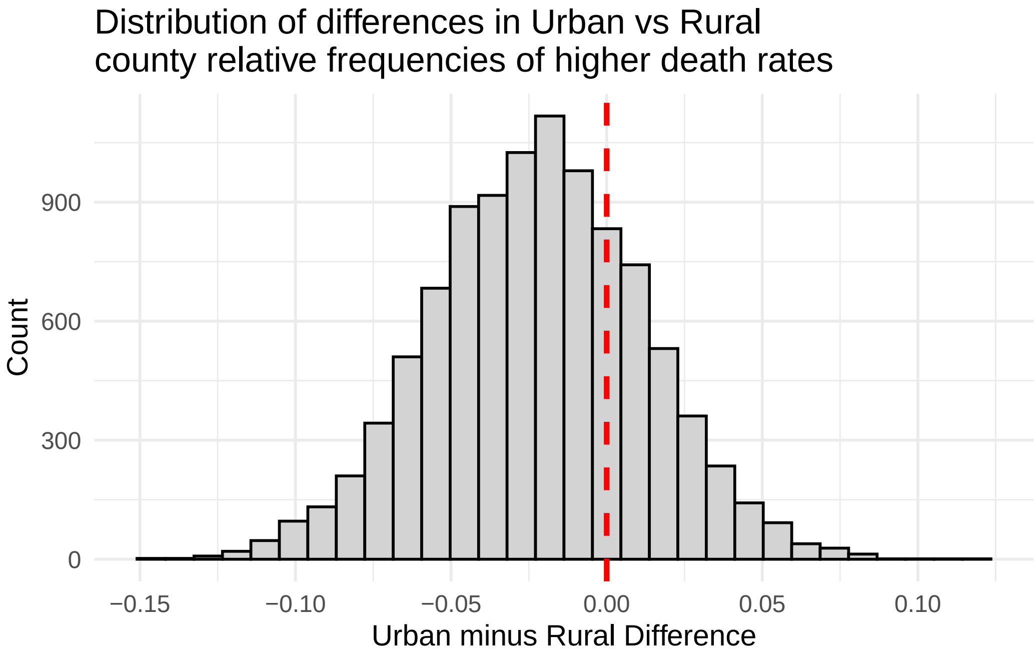
- New chance: 6.7%
- Estimated Difference: -0.12
- Estimated SD of distribution: 0.08

Increasing A, B to 100, 100 for 4-A



- New chance: 18.72%
- Estimated Difference: -0.04
- Estimated SD of distribution: 0.05

Increasing A, B to 200, 200 for 4-A



- New chance: 25.93%
- Estimated Difference: -0.02
- Estimated SD of distribution: 0.03