Name: Yuxuan Zhang

Date: September 16, 2024

HW2

1. (a) Since $\hat{\theta}_j = (1-w)\bar{y}_j + wc$, we have

$$\begin{aligned}(\hat{\theta}_j - \theta_j)^2 &= ((1-w)\bar{y}_j + wc - \theta_j)^2 \\ &= ((1-w)\bar{y}_j - \theta_j + w\theta_j - w\theta_j + wc)^2 \\ &= ((1-w)\bar{y}_j - (1-w)\theta_j + w(c-\theta_j))^2 \\ &= ((1-w)(\bar{y}_j - \theta_j) + w(c-\theta_j))^2 \\ &= (1-w)^2(\bar{y}_j-\theta_j)^2 + w^2(c-\theta_j)^2 + 2w(1-w)(\bar{y}_j-\theta_j)(c-\theta_j).\end{aligned}$$

Notice that

$$\mathbb{E}[2w(1-w)(\bar{y}_j-\theta_j)(c-\theta_j)] = 2w(1-w)(c-\theta_j)(\mathbb{E}[\bar{y}_j|\theta_j]-\theta_j) = 0.$$

Then, we have

$$\mathbb{E}[(\hat{\theta}_j-\theta_j)^2] = \mathbb{E}[(1-w)^2(\bar{y}_j-\theta_j)^2] + \mathbb{E}[w^2(c-\theta_j)^2].$$

We know that $\mathbb{E}[(\bar{y}_j-\theta_j)^2] = \mathbb{V}\mathrm{ar}(\bar{y}_j|\theta_j) = \frac{\sigma^2}{n}$. We also know that $\mathbb{E}[w^2(c-\theta_j)^2] = w^2(c-\theta_j)^2$. Then,

$$\mathbb{E}[(\hat{\theta}_j-\theta_j)^2] = (1-w)^2\frac{\sigma^2}{n} + w^2(c-\theta_j)^2.$$

Then, summing over all js, we have

$$\begin{aligned}\mathbb{E}[||\hat{\theta}-\theta||^2] &= \sum_{j=1}^{m}\left((1-w)^2\frac{\sigma^2}{n} + w^2(c-\theta_j)^2\right) \\ &= \frac{\sigma^2}{n}m(1-w)^2 + w^2\sum_{j=1}^{m}(c-\theta_j)^2.\end{aligned}$$

We then take partial derivatives to find the optimal w and c. For c, we have

$$\begin{aligned}\frac{\partial}{\partial c}\mathbb{E}[||\hat{\theta}-\theta||^2] &= \frac{\partial}{\partial c}w^2\sum_{j=1}^{m}(c-\theta_j)^2 \\ &= 2w^2\sum_{j=1}^{m}(c-\theta_j).\end{aligned}$$

Set this equal to zero, we have

$$\begin{aligned}2w^2\sum_{j=1}^{m}(c-\theta_j) &= 0 \\ mc &= \sum_{j=1}^{m}\theta_j \\ c &= \frac{1}{m}\sum_{j=1}^{m}\theta_j = \mu.\end{aligned}$$

For w, plugging in $c = \mu$, we have

$$\frac{\partial}{\partial w}\mathbb{E}[||\hat{\theta} - \theta||^2] = -2\frac{\sigma^2}{n}m(1-w) + 2w\sum_{j=1}^{m}(\mu - \theta_j)^2$$

Set this equal to zero, we have

$$\begin{aligned}\frac{\sigma^2}{n}m(1-w) &= w\sum_{j=1}^{m}(\mu - \theta_j)^2 \\ \frac{\sigma^2}{n}(1-w) &= w\frac{1}{m}\sum_{j=1}^{m}(\mu - \theta_j)^2.\end{aligned}$$

Let $\tau^2 = \frac{1}{m}\sum_{j=1}^{m}(\mu - \theta_j)^2$, we have

$$\begin{aligned}\frac{\sigma^2}{n}(1-w) &= w\tau^2 \\ w\frac{\sigma^2}{n} + w\tau^2 &= \frac{\sigma^2}{n} \\ w &= \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n} + \tau^2} \\ &= \frac{\frac{\sigma^2}{n}\frac{n}{\tau^2\sigma^2}}{\left(\frac{\sigma^2}{n} + \tau^2\right)\frac{n}{\tau^2\sigma^2}} \\ &= \frac{1/\tau^2}{n/\sigma^2 + 1/\tau^2}.\end{aligned}$$

(b) From part (a), we know that $c = \mu$, $w = \frac{1/\tau^2}{n/\sigma^2 + 1/\tau^2}$. Plugging them in, we have

$$\begin{aligned}\mathbb{E}[||\hat{\theta}_j - \theta_j||^2] &= \frac{\sigma^2}{n}m(1-w)^2 + w^2\sum_{j=1}^{m}(c - \theta_j)^2 \\ &= \left(\frac{n/\sigma^2}{n/\sigma^2 + 1/\tau^2}\right)^2\frac{\sigma^2}{n}m + \left(\frac{1/\tau^2}{n/\sigma^2 + 1/\tau^2}\right)^2\sum_{j=1}^{m}(\mu - \theta_j)^2 \\ &= \left(\frac{n/\sigma^2}{n/\sigma^2 + 1/\tau^2}\right)^2\frac{\sigma^2}{n}m + \left(\frac{1/\tau^2}{n/\sigma^2 + 1/\tau^2}\right)^2 m\tau^2 \\ &= m\left(\frac{n/\sigma^2}{(n/\sigma^2 + 1/\tau^2)^2} + \frac{1/\tau^2}{(n/\sigma^2 + 1/\tau^2)^2}\right) \\ &= m\frac{n/\sigma^2 + 1/\tau^2}{(n/\sigma^2 + 1/\tau^2)^2} \\ &= \frac{m}{n/\sigma^2 + 1/\tau^2}.\end{aligned}$$

2. See below

Q2

```
path = "C:\\Users\\LEGION\\Downloads\\btrips2015-7-1-4.rds"
bike_data <- readRDS(path)
library(ggplot2)
library(lme4)
library(dplyr)
```

(a)

```
bike_data$log_duration <- log(bike_data$duration)
anova_model <- aov(log_duration ~ station, data = bike_data)
anova_summary <- summary(anova_model)
anova_summary
```

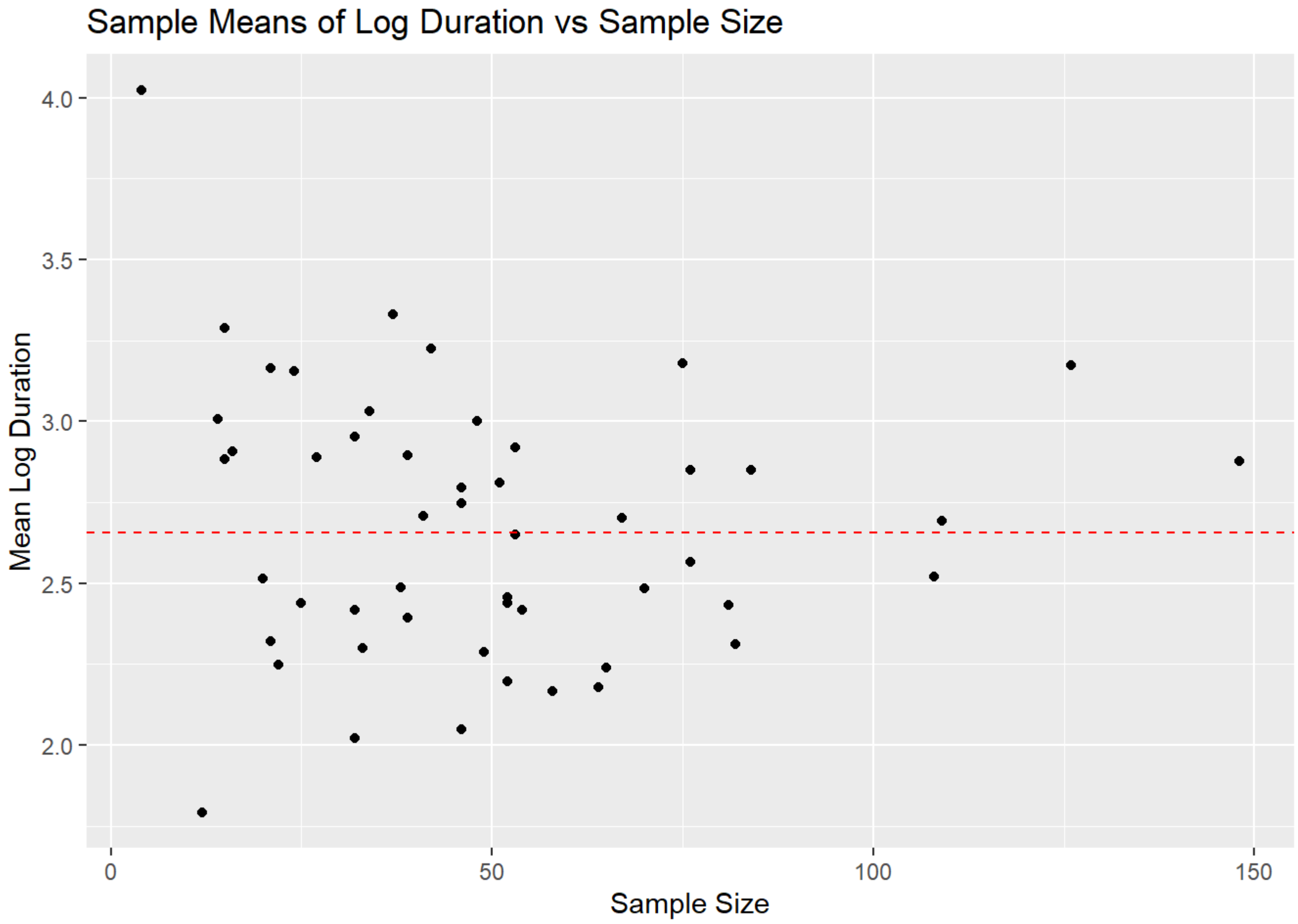
```
##               Df Sum Sq Mean Sq F value Pr(>F)    
## station       50  289.9   5.797   7.699 <2e-16 ***
## Residuals   2475 1863.6   0.753                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA analysis shows a high F-value of 7.699, with an associated p-value less that 2e-16, indicating strong statistical significance. We reject the null hypothesis that the mean log trip duration is the same for all stations. This suggests that there is substantial across-station heterogeneity in the log trip duration. The variation in trip duration is significantly different across stations, implying that the station of origin plays an important role n explaining differences in trip lengths.

(b)

```
means_data <- bike_data %>%
  group_by(station) %>%
  summarize(mean_log_duration = mean(log_duration), sample_size = n())
grand_mean <- mean(bike_data$log_duration)

ggplot(means_data, aes(x = sample_size, y = mean_log_duration)) +
  geom_point() +
  geom_hline(yintercept = grand_mean, linetype = "dashed", color = "red") +
  labs(title = "Sample Means of Log Duration vs Sample Size",
       x = "Sample Size",
       y = "Mean Log Duration")
```



The red dotted line in the middle represent the grand mean of the data. We can see that with small sample size, the means of log trip duration are more spread around the grand mean, indicating high variance. When the sample size increases, the mean log duration tend to be closer to the grand mean.

(c)

```
hierarchical_model <- lmer(log_duration ~ (1|station), data = bike_data)
summary_hierarchical <- summary(hierarchical_model)
mu <- fixef(hierarchical_model)[1]
sigma_sq <- attr(VarCorr(hierarchical_model), "sc")^2
tau_sq <- as.data.frame(VarCorr(hierarchical_model))$vcov[1]
cat("Mu:", mu, "\n")
```

```
## Mu: 2.657768
```

```
cat("Tau^2:", tau_sq, "\n")
```

```
## Tau^2: 0.1148483
```

```
cat("Sigma^2:", sigma_sq, "\n")
```

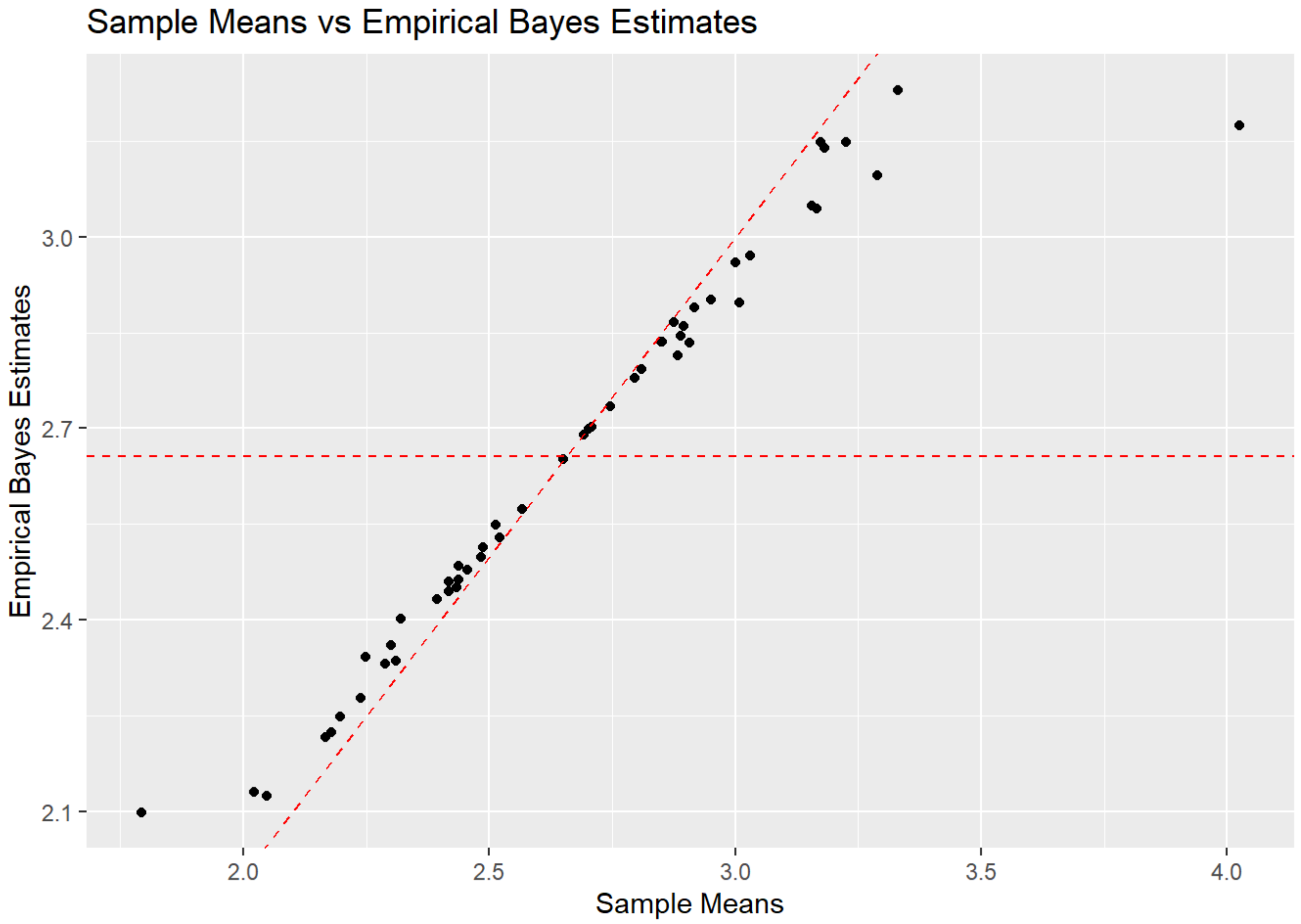
```
## Sigma^2: 0.7541967
```

```
summary_hierarchical
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: log_duration ~ (1 | station)
##    Data: bike_data
##
## REML criterion at convergence: 6561.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.2658 -0.6492 -0.0506  0.5580  4.1613
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  station  (Intercept) 0.1148   0.3389
##  Residual             0.7542   0.8684
## Number of obs: 2526, groups:  station, 51
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.65777    0.05167   51.44
```

(d)

```
sample_means <- bike_data %>%
  group_by(station) %>%
  summarize(sample_mean = mean(log_duration))
eb_estimates <- ranef(hierarchical_model)$station[,1] + mu
sample_means$eb_estimates <- eb_estimates
ggplot(sample_means, aes(x = sample_mean, y = eb_estimates)) +
  geom_point() +
  geom_hline(yintercept = grand_mean, linetype = "dashed", color = "red") +
  geom_abline(slope = 1, intercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Sample Means vs Empirical Bayes Estimates",
       x = "Sample Means",
       y = "Empirical Bayes Estimates")
```



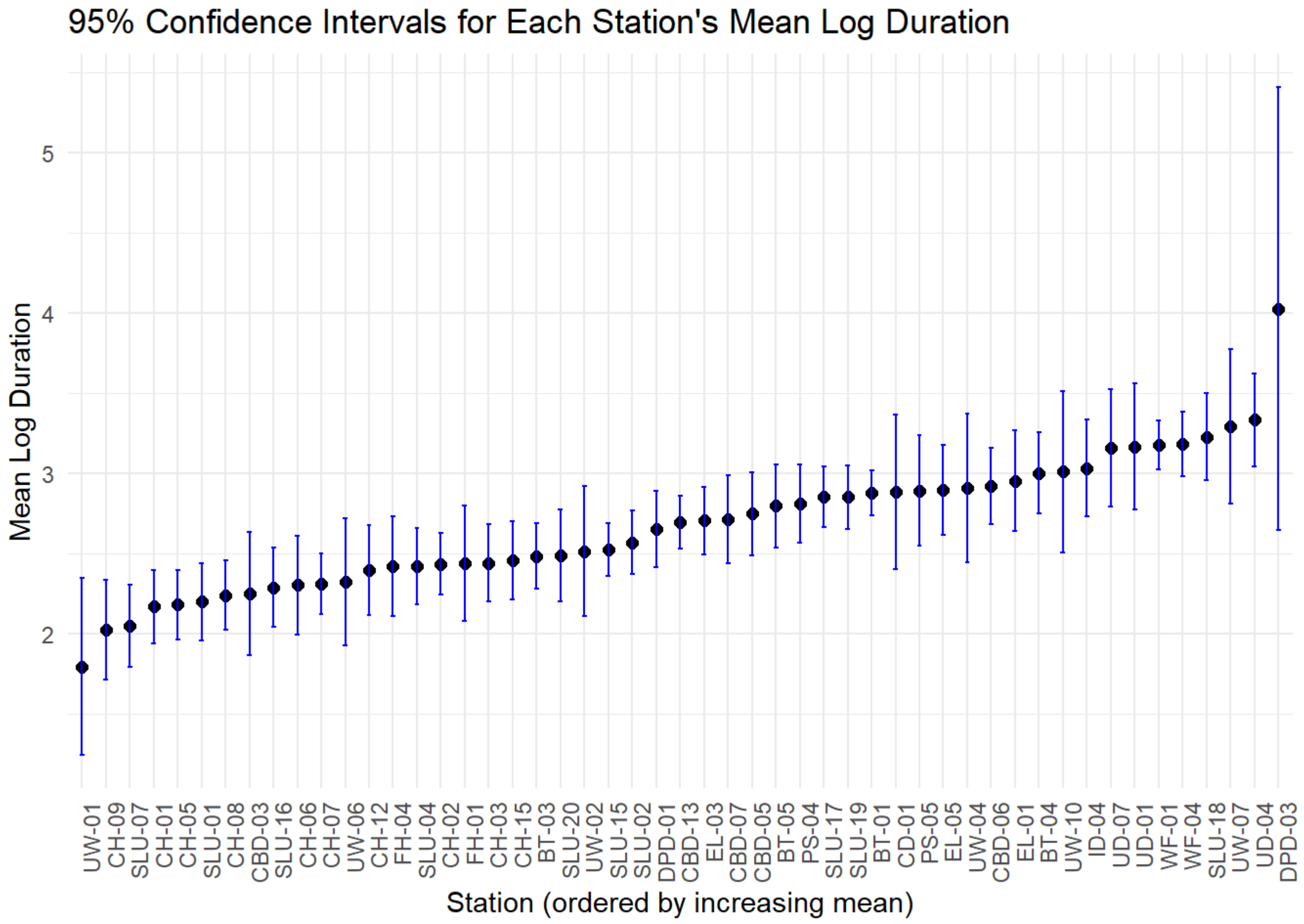
The plot demonstrates the effect of empirical Bayes (EB) shrinkage, where the EB estimates pull station-specific sample means towards the overall grand mean (indicated by the horizontal dashed line). We observe that the line formed by the points in the plot deviates from the 45-degree line (red dashed diagonal), indicating the shrinkage effect. The empirical Bayes estimates have a smaller range compared to the sample means, reflecting the tendency of EB to reduce variance by moving estimates towards the central tendency.

Regarding the amount of shrinkage, the two outliers in the upper-right and lower left of the plot experience significant shrinkage, as their sample means deviate substantially from the grand mean. In contrast, the points in the middle of the plot, where sample means are closer to the grand mean, exhibit less shrinkage.

(e)

```
alpha <- 0.05
sig_hat <- as.data.frame(VarCorr(hierarchical_model))$sdcor[2]
ci_data <- bike_data %>%
  group_by(station) %>%
  summarize(mean_log_duration = mean(log_duration),
            sample_size = n()) %>%
  rowwise() %>%
  mutate(
    t_crit = qt(1 - alpha / 2, df = sample_size - 1),
    margin_error = t_crit * sqrt(sig_hat^2 / sample_size),
    lower_ci = mean_log_duration - margin_error,
    upper_ci = mean_log_duration + margin_error
  )
ci_data <- ci_data %>% arrange(mean_log_duration)

ggplot(ci_data, aes(x = reorder(station, mean_log_duration), y = mean_log_duration)) +
  geom_point(size = 2) +
  geom_errorbar(aes(ymin = lower_ci, ymax = upper_ci), width = 0.2, color = "blue") +
  labs(title = "95% Confidence Intervals for Each Station's Mean Log Duration",
       x = "Station (ordered by increasing mean)",
       y = "Mean Log Duration") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



```
df_samplesize <- bike_data %>%
  group_by(station) %>%
  summarize(sample_size = n()) %>%
  arrange(sample_size)
head(df_samplesize)
```

```
## # A tibble: 6 × 2
##   station sample_size
##   <chr>         <int>
## 1 DPD-03            4
## 2 UW-01            12
## 3 UW-10            14
## 4 CD-01            15
## 5 UW-07            15
## 6 UW-04            16
```

The two outlying groups are schools ‘UW-01’ and ‘DPD-03’, as they have much larger confidence intervals, and their confidence intervals does not overlap with many other schools. We can see from the chart that these two schools have the lowest sample means, which may causes them to be the outliers.

3. (a) From question 1, we know that

$$\mathbb{E}[(\hat{\theta}-\theta)^2|\theta] = (1-w)^2\frac{\sigma^2}{n} + w^2(\mu-\theta)^2.$$

We also know that $\mathbb{E}[(\bar{y}-\theta)^2] = \frac{\sigma^2}{n}$. Then, assume $w > 0$, we have the inequality

$$\begin{aligned}
(1-w)^2\frac{\sigma^2}{n} + w^2(\mu-\theta)^2 &< \frac{\sigma^2}{n} \\
\frac{\sigma^2}{n} - 2w\frac{\sigma^2}{n} + w^2\frac{\sigma^2}{n} + w^2(\mu-\theta)^2 &< \frac{\sigma^2}{n} \\
-2w\frac{\sigma^2}{n} + w^2\frac{\sigma^2}{n} + w^2(\mu-\theta)^2 &< 0 \\
w^2\left(\frac{\sigma^2}{n} + (\mu-\theta)^2\right) - 2w\frac{\sigma^2}{n} &< 0 \\
w\left(\frac{\sigma^2}{n} + (\mu-\theta)^2\right) &< \frac{2\sigma^2}{n} \\
w &< \frac{\frac{2\sigma^2}{n}}{\frac{\sigma^2}{n} + (\mu-\theta)^2}.
\end{aligned}$$

(b) From question 1, we know that $w = \frac{1/\tau^2}{n/\sigma^2 + 1/\tau^2}$. Plug this in, we have

$$\begin{aligned}
\frac{1/\tau^2}{n/\sigma^2 + 1/\tau^2} &< \frac{2\sigma^2/n}{\sigma^2/n + (\mu-\theta)^2} \\
1/\tau^2 \cdot (\sigma^2/n + (\mu-\theta)^2) &< 2\sigma^2/n \cdot (n/\sigma^2 + 1/\tau^2) \\
1/\tau^2 \cdot (\mu-\theta)^2 &< 2 + 2\sigma^2/n\tau^2 - \sigma^2/n\tau^2 \\
(\mu-\theta)^2 &< 2\tau^2 + \frac{2\sigma^2}{n} - \frac{\sigma^2}{n} \\
(\mu-\theta)^2 &< 2\tau^2 + \frac{\sigma^2}{n}.
\end{aligned}$$