SDS_459_HW3_Coding

2025-04-11

Aidan Kardan SDS 459 HW 3

Question 1 done in separate pdf for convenience

Question 2

a)

```
library(HistData)
df1 <- HistData::Michelson

# Each observation such that x + 299,000 gives km/sec
x <- df1$velocity

# MLE for theta is the sample mean given model assumption
theta_hat <- mean(x)
theta_hat
```

```
## [1] 852.4
```

```
n <- length(x)
sigma2 <- 50
se <- sqrt(sigma2/n)

# 95% CI using large-sample approximation: theta_hat ± 1.96*se
CI_freq <- c(theta_hat - 1.96 * se, theta_hat + 1.96 * se)
CI_freq
```

```
## [1] 851.0141 853.7859
```

95% Frequentist CI is [851.0141, 853.7859].

b)

```
# Prior parameters
mu0 <- 800
tau2 <- 50

# Posterior calculations
posterior_variance <- 1 / (n/sigma2 + 1/tau2)
posterior_mean <- posterior_variance * (n*theta_hat/sigma2 + mu0/tau2)
posterior_mean
```

```
## [1] 851.8812
```

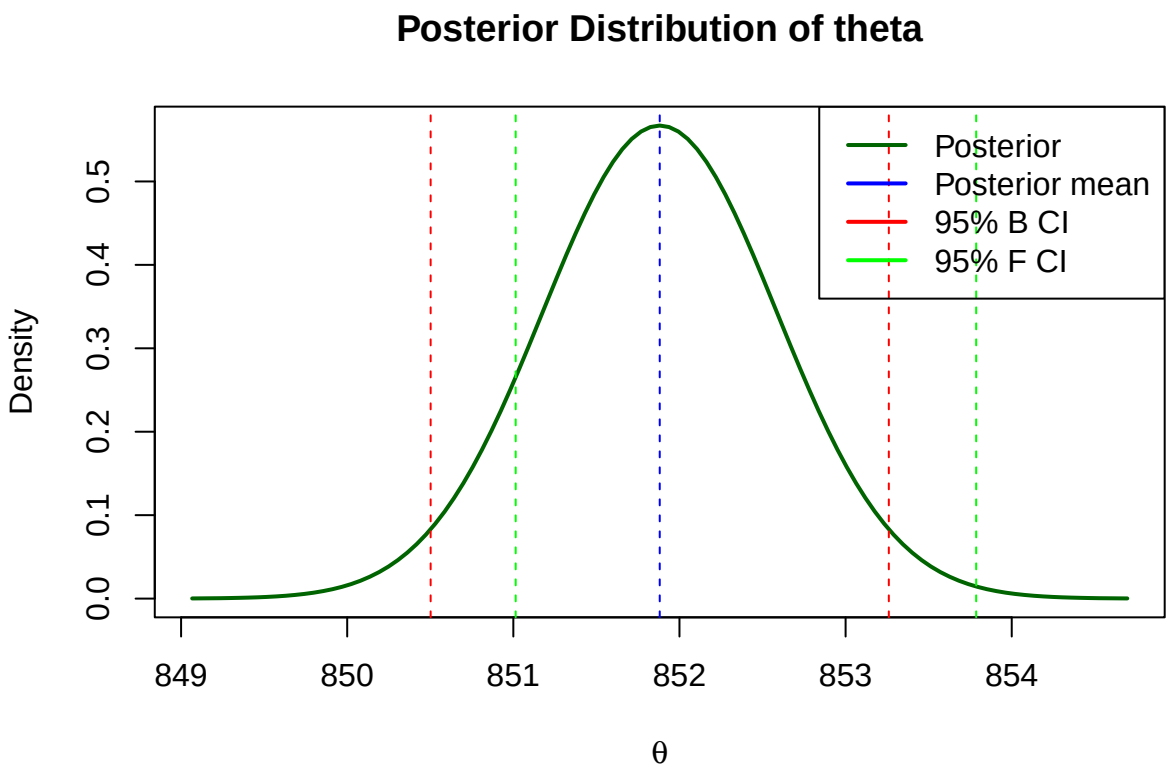
```
# 95% Bayesian equal-tailed credible interval for theta
CI_bayes <- c(posterior_mean - 1.96 * sqrt(posterior_variance),
              posterior_mean + 1.96 * sqrt(posterior_variance))
CI_bayes
```

```
## [1] 850.5021 853.2602
```

95% Bayesian equal-tailed credible interval for theta is [850.5021, 853.2602].

c)

```
# Plot the posterior density for theta
curve(dnorm(x, mean = posterior_mean, sd = sqrt(posterior_variance)),
      from = posterior_mean - 4*sqrt(posterior_variance),
      to = posterior_mean + 4*sqrt(posterior_variance),
      xlab = expression(theta), ylab = "Density",
      main = "Posterior Distribution of theta",
      col = "darkgreen", lwd = 2)
abline(v = posterior_mean, col = "blue", lty = 2)
abline(v = CI_bayes, col = "red", lty = 2)
abline(v = CI_freq, col = 'green', lty = 2)
legend("topright", legend = c("Posterior", "Posterior mean", "95% B CI", '95% F CI'),
       col = c("darkgreen", "blue", "red", 'green'), lwd = 2)
```



For a normal (symmetric) posterior, the equal-tailed 95% credible interval is identical to the highest posterior density (HPD) interval. Thus, it is not possible to construct a narrower 95% non-equal-tailed credible interval. In a skewed posterior, however, a non-equal-tailed (HPD) interval can be shorter, but in our case the normality implies symmetry and no such improvement is possible.

Question 3

a) and b) done in separate pdf for convenience.

part c)

```
# For reproducibility
set.seed(123)
# given data
n <- 50
```

```r
lambda_true <- 3
# Generate a random sample from Exp(lambda_true)
sample_exp <- rexp(n, rate = lambda_true)

# MLE for lambda is 1/mean(x)
lambda_hat <- 1 / mean(sample_exp)
cat("MLE =", lambda_hat, "\n")
```

```
## MLE = 2.653996
```

```r
# Conjugate prior: choose hyperparameters
alpha0 <- 2
beta0 <- 1
# Posterior parameters for conjugate prior
posterior_alpha_a <- alpha0 + n
posterior_beta_a <- beta0 + sum(sample_exp)
posterior_mean_a <- posterior_alpha_a / posterior_beta_a

# Jeffreys prior: the posterior becomes Gamma(shape = n, rate = sum(x))
posterior_alpha_b <- n
posterior_beta_b <- sum(sample_exp)
posterior_mean_b <- posterior_alpha_b / posterior_beta_b

cat("Posterior Mean using conjugate prior =", posterior_mean_a, "\n")
```

```
## Posterior Mean using conjugate prior = 2.621032
```

```r
cat("Posterior Mean using Jeffrey's prior =", posterior_mean_b, "\n")
```

```
## Posterior Mean using Jeffrey's prior = 2.653996
```

d)

```r
# 95% credible interval for lambda under the conjugate prior
cred_int_a <- qgamma(c(0.025, 0.975), shape = posterior_alpha_a, rate = posterior_beta_a)
cred_int_a
```

```
## [1] 1.957512 3.379900
```

```r
# 95% credible interval for lambda under the Jeffreys prior
cred_int_b <- qgamma(c(0.025, 0.975), shape = posterior_alpha_b, rate = posterior_beta_b)
cred_int_b
```

```
## [1] 1.969847 3.438549
```

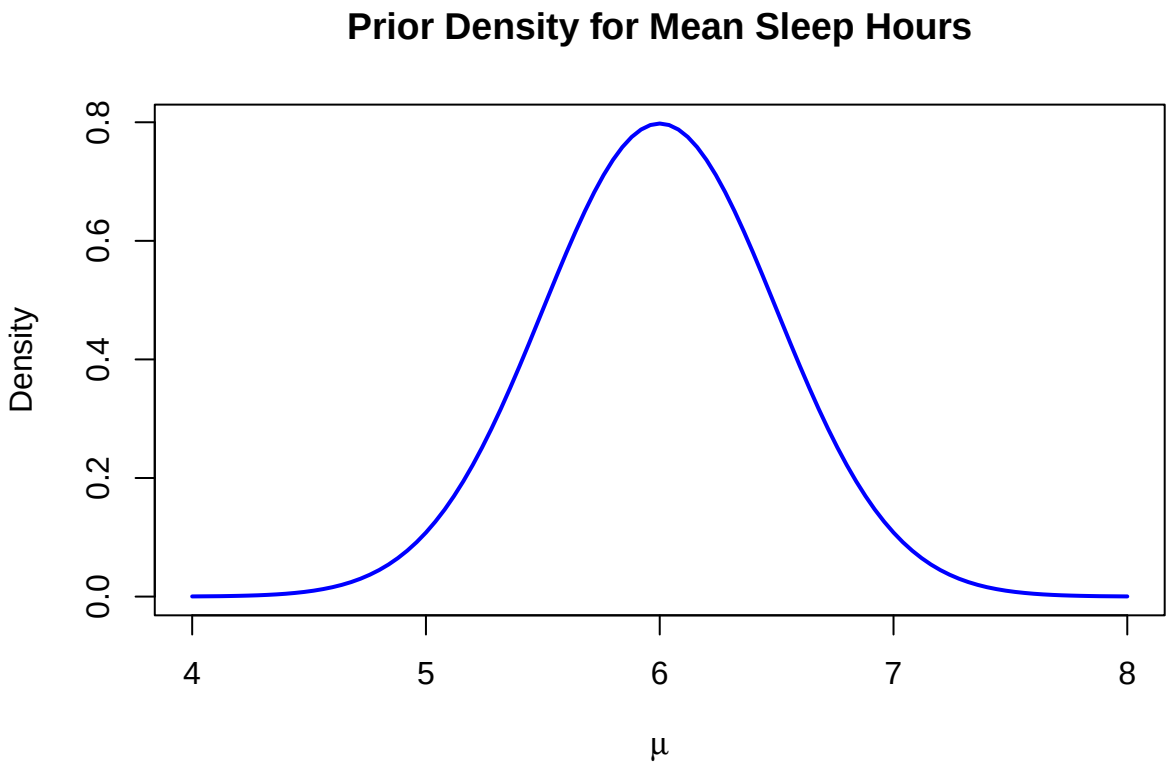
The interval from the informative conjugate prior is narrower because it injects additional prior belief via its hyper parameters that effectively shrinks the posterior variance relative to what the data alone would suggest via Jeffrey's prior.

The extra parameters in the conjugate case add weight to the prior information, and when this prior information is reasonably in agreement with the data, it produces a posterior distribution with lower variance (greater precision) and as a result, a narrower 95% credible interval. Jeffrey's prior is non-informative and does not impose any extra structure, hence the resulting posterior reflects only the variability in the data, leading to a wider interval.

Question 4

a)

```r
mu0 <- 6
tau0_sq <- 0.25
curve(dnorm(x, mean = mu0, sd = sqrt(tau0_sq)), from = 4, to = 8,
      xlab = expression(mu), ylab = "Density",
      main = "Prior Density for Mean Sleep Hours",
      col = "blue", lwd = 2)
```



b)

```r
prior_quartiles <- qnorm(c(0.25, 0.5, 0.75), mean = mu0, sd = sqrt(tau0_sq))
prior_quartiles
```

```
## [1] 5.662755 6.000000 6.337245
```

The 25th percentile is approximately 5.66

The 50th percentile is approximately 6

The 75th percentile is approximately 6.33

c)

```r
prob_mu_gt7_prior <- 1 - pnorm(7, mean = mu0, sd = sqrt(tau0_sq))
cat("Prior Probability that the mean amount of sleep exceeds 7 hours:", prob_mu_gt7_prior)
```

```
## Prior Probability that the mean amount of sleep exceeds 7 hours: 0.02275013
```

d) Done in separate pdf for convenience.

e)

```r
n <- 24
sigma2 <- 2
ybar <- 7.688
posterior_variance_post <- 1 / (n/sigma2 + 1/tau0_sq)
```

```r
posterior_mean_post <- posterior_variance_post * (n*ybar/sigma2 + mu0/tau0_sq)

cat("Posterior mean given assumptions: ", posterior_mean_post, "\n")
```

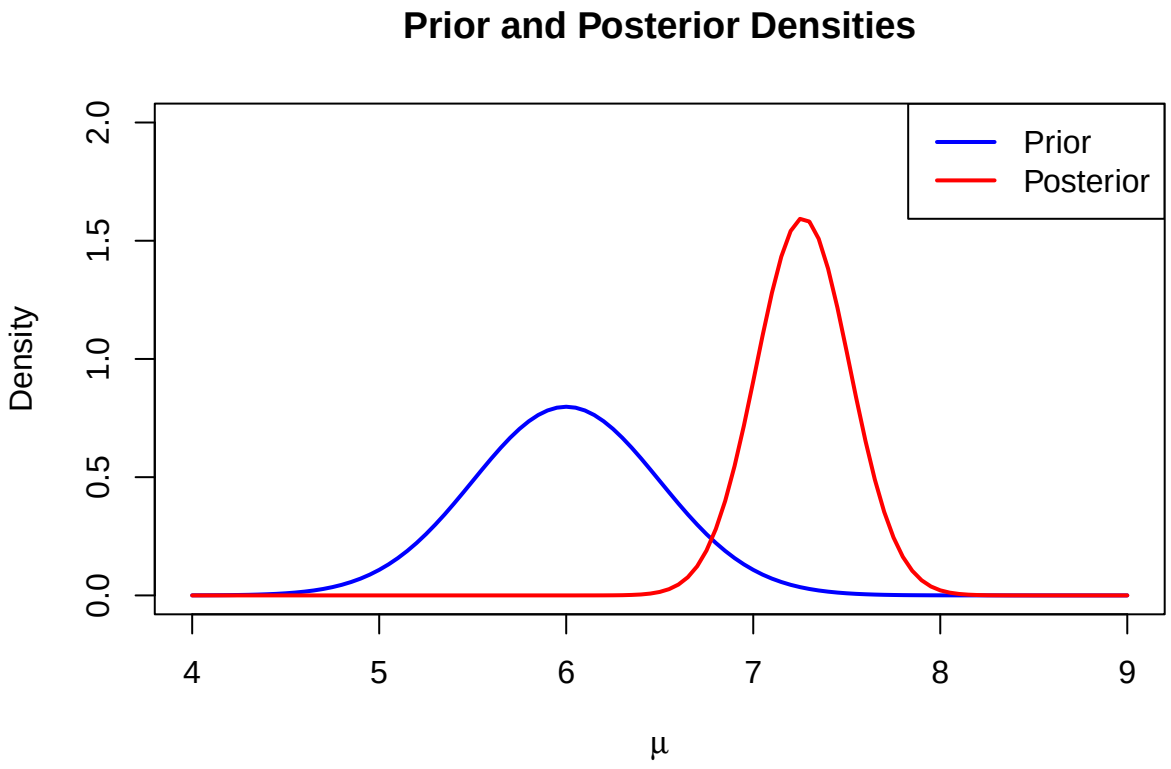
```
## Posterior mean given assumptions:  7.266
```

```r
cat("Posterior variance given assumptions:", posterior_variance_post)
```

```
## Posterior variance given assumptions: 0.0625
```

f)

```r
curve(dnorm(x, mean = mu0, sd = sqrt(tau0_sq)), from = 4, to = 9,
      xlab = expression(mu), ylab = "Density", ylim = c(0,2),
      main = "Prior and Posterior Densities",
      col = "blue", lwd = 2)
curve(dnorm(x, mean = posterior_mean_post, sd = sqrt(posterior_variance_post)),
      add = TRUE, col = "red", lwd = 2)
legend("topright", legend = c("Prior", "Posterior"),
       col = c("blue", "red"), lwd = 2)
```



g)

```r
z_val <- qnorm(0.96)  # 96th percentile for the upper bound (since lower tail is 0.04)
cred_int_post <- c(posterior_mean_post - z_val * sqrt(posterior_variance_post),
                   posterior_mean_post + z_val * sqrt(posterior_variance_post))
cat("92% Equal Tailed Posterior Credible interval:", cred_int_post)
```

```
## 92% Equal Tailed Posterior Credible interval: 6.828328 7.703672
```

h)

```r
prob_mu_gt7_post <- 1 - pnorm(7, mean = posterior_mean_post, sd = sqrt(posterior_variance_post))
cat("Posterior Probability that the mean amount of sleep exceeds 7 hours:", prob_mu_gt7_post)
```

```
## Posterior Probability that the mean amount of sleep exceeds 7 hours: 0.8563357
```

The prior probability in part c was extremely low, less than 0.05, while the posterior probability is extremely high, more than 0.85. This is because of the new information that was presented and the additional assumptions made in the analysis.

i)

In the frequentist framework, the true mean is regarded as a fixed, though unknown, value. When we construct a 95% confidence interval for the true mean, we are not claiming that there is a 95% probability that the interval contains the true mean. Instead, the interpretation is that if we were to repeat the process of collecting a sample and constructing a confidence interval many times, then approximately 95% of those intervals would capture the true mean. In any single instance, the true mean either lies within the interval or it does not; we do not assign a probability to the true mean in the context of that specific interval.

Because the true mean is fixed and not random, the frequentist approach does not allow us to directly say, for example, "there is an 85% probability that the true mean exceeds 7 hours." Instead, to address the question of whether the true mean is above 7 hours, a frequentist would conduct a hypothesis test.

For instance, one might test the null hypothesis that the true mean is less than or equal to 7 hours against the alternative that the true mean is greater than 7 hours. The resulting p-value would indicate whether there is statistically significant evidence to reject the null hypothesis. Although this testing framework does not yield a direct probability statement (like "the probability that the true mean exceeds 7 hours is 0.85"), it provides a method to infer whether the observed data support the claim that the true mean is above 7 hours.