# Homework 4

STT 465, Bayesian Statistical Methods

Lowell Monis

October 11, 2025

## Question 1

The file `Crabs.txt` contains crab data from a study of female horseshoe crabs on an island in the Gulf of Mexico. After downloading the file onto your computer, import the data into R using `crab <- read.table("Crabs.txt", header = TRUE)`. Then, `crab$width` has data on carapace width of $n = 173$ female crabs. Use the carapace width data to answer the following questions.

```r
crab <- read.table("data/Crabs.txt", header = TRUE)
head(crab)
```

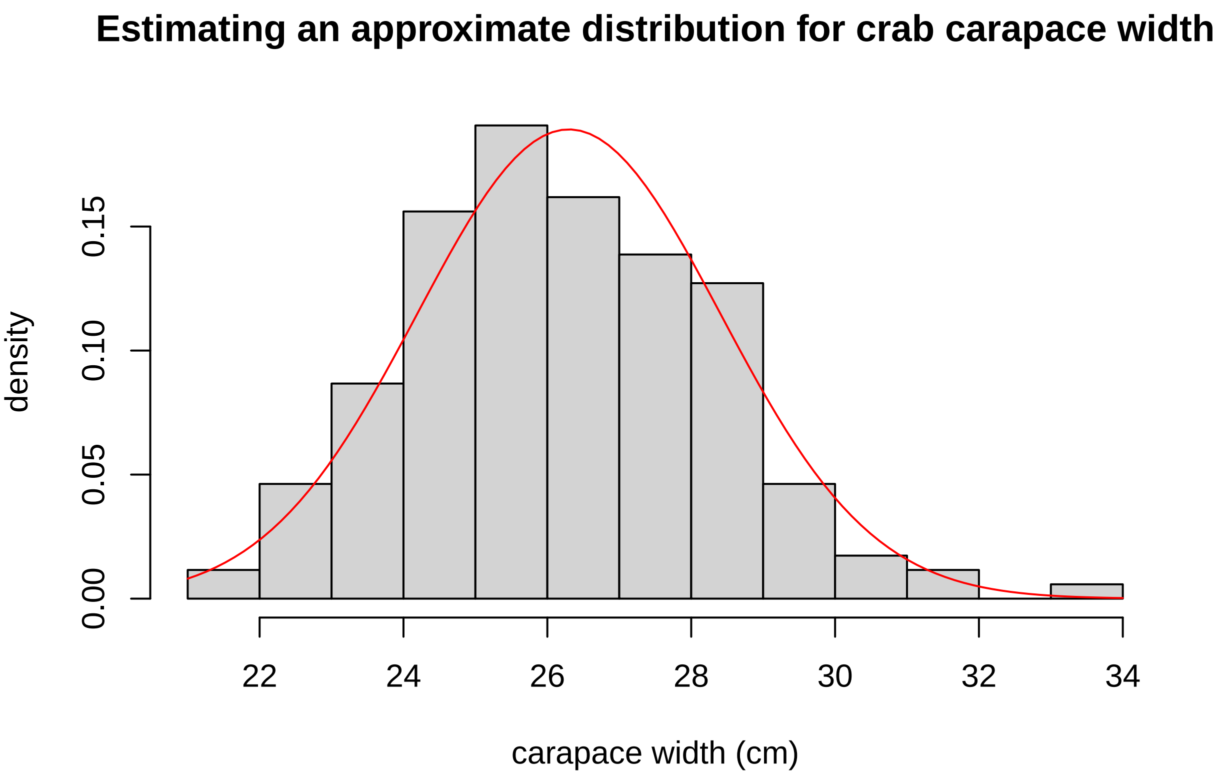
```
##   crab y weight width color spine
## 1    1 8   3.05  28.3     2     3
## 2    2 0   1.55  22.5     3     3
## 3    3 9   2.30  26.0     1     1
## 4    4 0   2.10  24.8     3     3
## 5    5 4   2.60  26.0     3     3
## 6    6 0   2.10  23.8     2     3
```

### (a) 10 points

Plot the histogram and the normal distribution $\mathcal{N}(\bar{y}, s^2)$ on a single plot, where $\bar{y}$ is the sample mean and $s^2$ is the sample variance. Is the normal distribution a good approximation to the data?

### Answer

```r
hist(crab$width, breaks = 12,
     freq = FALSE,
     main = "Estimating an approximate distribution for crab carapace width",
     xlab = "carapace width (cm)",
     ylab = "density")
curve(dnorm(x, mean(crab$width), sd(crab$width)),
      add = TRUE,
      col = 'red')
```



The histogram bars generally follow the shape of the normal curve. The distribution appears unimodal and roughly symmetric, with the mean slightly off-center. Based on the visual evidence, the normal distribution appears to be a reasonably good approximation to the distribution of the carapace width, although there is slight skewness visible, with a somewhat longer tail to the right, indicating positive skew.

### (b) 10 points

Assume the sampling model: $Y_1, \ldots, Y_n \mid \mu, \sigma^2 \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with $\sigma^2 = 4.8$. We consider a prior $\mathcal{N}(20, 1.2)$ for $\mu$. Find the posterior distribution $\mu$. Construct a 95% credible interval for $\mu$.

### Answer

Given prior $\mu \sim \mathcal{N}(\mu_0 = 20, \tau_0^2 = 1.2)$, and sampling model $Y_1, \ldots, Y_n \mid \mu, \sigma^2 \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2 = 4.8)$, we proceed to calculate the parameters for posterior $\mu \mid y \sim \mathcal{N}(\mu_n, \tau_n^2)$:

$$\frac{1}{\tau_n^2} = \frac{n}{\sigma^2} + \frac{1}{\tau_0^2}$$

$$\mu_n = \tau_n^2 \left( \frac{n\bar{y}}{\sigma^2} + \frac{\mu_0}{\tau_0^2} \right)$$

```r
n <- length(crab$width)
pop_var <- 4.8
sample_mean <- mean(crab$width)
prior_mean <- 20
prior_var <- 1.2
post_var <- 1/((n/pop_var)+(1/prior_var))
post_mean <- post_var*((n*sample_mean/pop_var)+(prior_mean/prior_var))
```

Using the code snippet above, we have computed the posterior parameters. I am using in-line code to print the parameters onto markdown.

The posterior mean $\mu_n = 26.1564972$, and the posterior variance $\tau_n^2 = 0.0271186$. Thus, the posterior distribution for the population mean $\mu$, given prior $\mathcal{N}(20, 1.2)$ is $\mu \mid y \sim \mathcal{N}($ 26.16, 0.03 $)$.

Next, I am going to compute the 95% credible interval for $\mu$. This can be done by computing the quantiles of the respective posterior normal distribution found above.

```
ci <- qnorm(c(0.025,0.975), post_mean, sqrt(post_var))
```

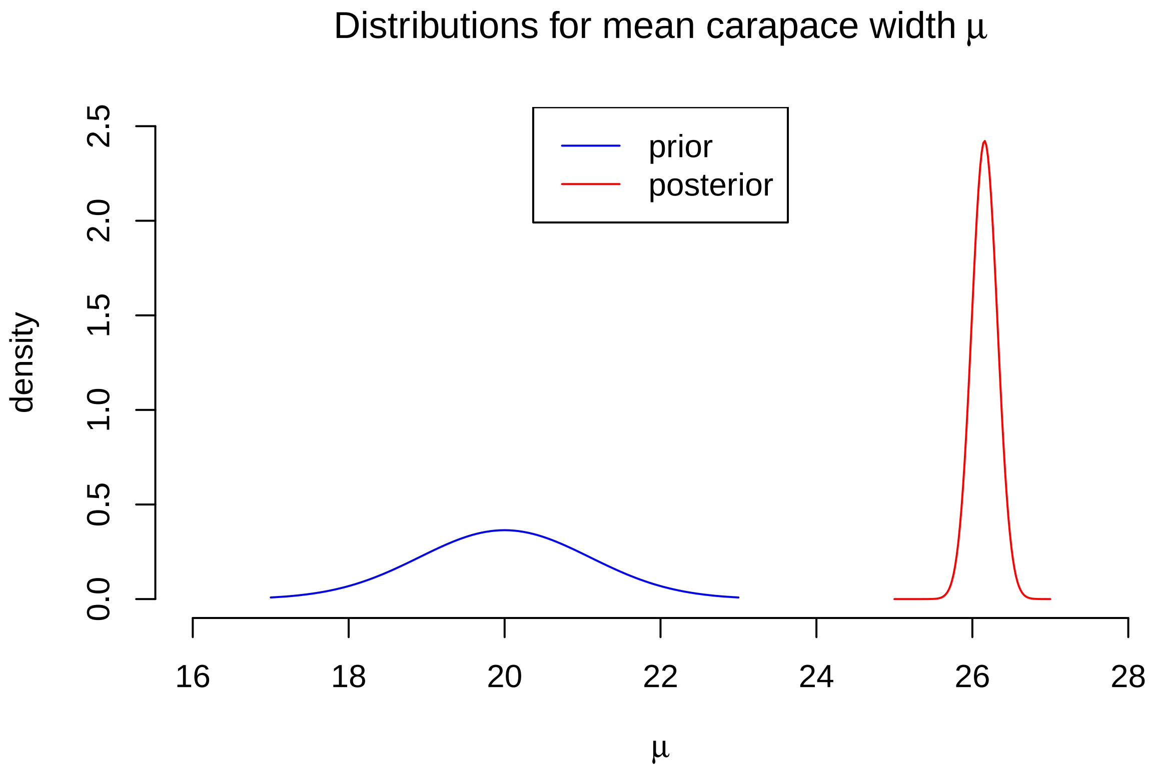
The 95% credible intervals of $\mu$ is [25.8337354, 26.4792589].

### (c) 10 points

Plot the prior and posterior distributions on a single plot, and explain what you observe.

### Answer

```
# Setting the plot field using a method from R docs
plot.new()
plot.window(xlim = c(16, 28), ylim = c(0, 2.5))
title(main = expression("Distributions for mean carapace width" ~mu),
      xlab = expression(mu),
      ylab = "density")
axis(1)
axis(2)
curve(dnorm(x, mean = prior_mean, sd = sqrt(prior_var)),
      add = TRUE,
      from = 17,
      to = 23,
      col = "blue")
curve(dnorm(x, mean = post_mean, sd = sqrt(post_var)),
      add = TRUE,
      from = 25,
      to = 27,
      col = "red")
legend("top",
       legend = c('prior', 'posterior'),
       col = c("blue", "red"),
       lty=1, lwd=1)
```



The distribution of the prior is flattened, and relatively wider with a variance of 1.2, centered around the prior mean of 20, and the distribution of the posterior is taller and rather narrow, with a variance of 0.03, which is a lot smaller than the prior variance The posterior distribution is centered at the posterior mean, 26.1564972, which is near the sample mean 26.2988439.

The data is highly informative with a larger sample size, thus improving the precision, so the posterior mean shifts significantly toward the sample mean, and the posterior distribution becomes much more precise, which we know from the large reduction in the variance, and indicator of lower uncertainty.

## Question 2

The files "school1.txt" and "school2.txt" contain data on the amount of time students from two high schools spent on studying or homework during an exam period. Download the files onto your computer, import the data into R using `s1 <- read.table("school1.txt"); s2 <- read.table("school2.txt")`. Analyze data from each of the two schools separately, using the normal model with a conjugate prior distribution in which $\{\mu_0 = 5, \kappa_0 = 1, a = 1, b = 4\}$. Use Monte Carlo approximations to compute the following.

```r
s1 <- read.table("data/school1.txt"); s2 <- read.table("data/school2.txt")
y1 <- s1[,1]; y2 <- s2[,1]
prior_mean <- 5
prior_precision <- 1
prior_shape <- 1
prior_scale <- 4
```

### (a) 10 points

Posterior means and 95% credible intervals for the mean $\mu$ of each school.

**Answer**

When both $\mu$ and $\sigma^2$ are unknown, we expect the conjugate prior to be the product of normal and Inverse-Gamma distributions.

The sampling model we will be using is $Y_i \mid \mu, \sigma^2 \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. The prior is given by $p(\mu, \sigma^2) = p(\mu \mid \sigma^2) \times p(\sigma^2)$, where $\mu \mid \sigma^2 \sim \mathcal{N}(\mu_0, \frac{\sigma^2}{\kappa_0})$ and $\sigma^2 \sim \text{Inverse-Gamma}(a, b)$.

The posterior distribution is given by $p(\mu, \sigma^2 \mid y) = p(\mu \mid y, \sigma^2) \times p(\sigma^2 \mid y)$, where $\sigma^2 \mid y \sim \text{Inverse-Gamma}(\tilde{a}, \tilde{b})$, and $\mu \mid \sigma^2, y \sim \mathcal{N}(\mu_n, \frac{\sigma^2}{\kappa_n})$, after the conjugate prior parameters are updated as follows:

$$\begin{aligned}
\kappa_n &= \kappa_0 + n \\
\mu_n &= \frac{\kappa_0 \mu_0 + n\bar{y}}{\kappa_n + n} \\
\tilde{a} &= a + \frac{n}{2} \\
\tilde{b} &= b + \frac{1}{2}\sum_{i=1}^{n}(y_i - \bar{y})^2 + \frac{n\kappa_0(\bar{y} - \mu_0)^2}{2(\kappa_0 + n)}
\end{aligned}$$

I will now implement these Bayesian updations via an R function.

```r
bayesian_updation <- function(y, prior_mean, prior_precision,
                              prior_shape, prior_scale) {
  n <- length(y)
  sample_mean <- mean(y)
  var <- sum((y-sample_mean)^2)

  post_precision <- prior_precision+n
  post_mean <- ((prior_precision*prior_mean)+(n*sample_mean))/post_precision
  post_shape <- prior_shape+(n/2)
  post_scale <- prior_scale+(var/2)+
    n*prior_precision*(sample_mean-prior_mean)^2/(2*(prior_precision+n))

  return(c(post_precision, post_mean, post_shape, post_scale))
}
```

I will now compute the posterior parameters for school1 and school2.

```r
post_s1 <- bayesian_updation(y1, prior_mean, prior_precision,
                             prior_shape, prior_scale)
post_s2 <- bayesian_updation(y2, prior_mean, prior_precision,
                             prior_shape, prior_scale)
```

I will now conduct a Monte Carlo approximation with 10,000 simulations, at seed 465 for reproducibility. I sample the poster variance for that iteration from the Inverse Gamma distribution using the computed posterior parameters, and then use that variance to sample the posterior mean for that iteration from the normal distribution. The posterior mean for the true mean will be the mean of all the sampled posterior means. I then use quantiles from the sampled posterior means to calculate the 95% credible intervals.

```
set.seed(465)
var1 <- 1/rgamma(10000, shape = post_s1[3], rate = post_s1[4])
var2 <- 1/rgamma(10000, shape = post_s2[3], rate = post_s2[4])
mu1 <- rnorm(10000, mean = post_s1[2], sd = sqrt(var1/post_s1[1]))
mu2 <- rnorm(10000, mean = post_s2[2], sd = sqrt(var2/post_s2[1]))
post_mean_mc1 <- mean(mu1)
post_mean_mc2 <- mean(mu2)
ci1a <- quantile(mu1, c(0.025,0.975))
ci2a <- quantile(mu2, c(0.025,0.975))
```

Thus, for school1, the posterior mean of the mean amount of time spent on studying or homework during the exam period is 9.2769713, lying within a 95% credible interval of [7.787237, 10.8161473].

For school2, the posterior mean of the mean amount of time spent on studying or homework during the exam period is 6.9383299, lying within a 95% credible interval of [5.1208681, 8.7277344].

It looks like the average time spent on homework is estimated to be higher for school1 than for school2.

### (b) 10 points

Posterior means and 95% credible intervals for the standard deviation $\sigma$ of each school.

#### Answer

The prerequisites of the computation for this part of the question can be carried forward from (a). For computing the posterior mean of the standard deviation, I will use the posterior variances simulated above, take the square root of the simulated values, and then compute their mean.

```
post_sd_mc1 <- mean(sqrt(var1))
post_sd_mc2 <- mean(sqrt(var2))
ci1b <- quantile(sqrt(var1), c(0.025,0.975))
ci2b <- quantile(sqrt(var2), c(0.025,0.975))
```

Thus, for school1, the posterior mean of the standard deviation in the amount of time spent on studying or homework during the exam period is 3.9093155, lying within a 95% credible interval of [3.0069442, 5.1745082].

For school2, the posterior mean of the standard deviation in the amount of time spent on studying or homework during the exam period is 4.4017056, lying within a 95% credible interval of [3.356359, 5.9029138].

It looks like the variation in study times is larger between the students of school2 than those of school1.

### (c) 10 points

Posterior probability that $\mu_1 > \mu_2$, where $\mu_1$ is the mean of school1 and $\mu_2$ is the mean of school2.

#### Answer

I will use the paired Monte Carlo samples of the mean as computed in (a) for this part.

The theoretical concept I am using here is the indicator function for the given condition:

$$P(\mu_1 > \mu_2 \mid y) \approx \frac{1}{M} \sum_{i=1}^{M} I(\mu_1^{(m)} > \mu_2^{(m)})$$

where $M$ is the number of Monte Carlo samples.

```
mean((mu1>mu2)==TRUE)
```

```
## [1] 0.9734
```

The posterior probability that the mean time spent by students studying in school1 is greater than school2 is 97.34%.