Problemset 4

Nikolai German (12712506)

Exercise 4.1 - Bayesian Inference

An appliances company carries out a quality control check on the scales they produce. In particular, 27 scales are checked by using a standard weight of $100g$. We assume the weights measured by the scales to be normally distributed with variance $0.1g^2$. However, it is not known whether the scales measure without a bias.

(a)

We are interested in assessing the behavior of the mean weight μ. Specify the concept of a Bayesian credibility interval, elaborating on the underlying principle.

Using the posterior distribution $p(\mu|x)$, we can obtain the $\alpha/2$ and $1 - \alpha/2$ quantiles of this distribution. Thus obtaining a credibility intervall for the parameter μ to the level α.

(b)

Let us now assume that the posterior distribution $\mathbb{P}(\mu|x_1, ..., x_{27})$ has the following form: $\mathcal{N}(100.14g, 0.0049g^2)$. Specify the highest density 95% credibility interval (HDI) for μ. Do the scales conform to the standard of unbiasedness of the firm?

```
c("2.5%" = qnorm(.025, 100.14, 0.07),
  "97.5%" = qnorm(.975, 100.14, 0.07))
```

```
     2.5%     97.5%
100.0028 100.2772
```

There seems to be a bias, since the 95% credibility intervall is as a whole over 100g.

(c)

Company management asks you to provide a point estimate of μ. Which are the possibilities? Which number do you give them?

There are two options: either give the expected value of the posterior $\mathbb{E}(p(\mu|x)) = 100.14$ or the MLE $\hat{\mu}_{MLE} = \bar{x}$.

(d)

The prior distribution for   is assumed to be $\mathcal{N}(100g, 0.13g^2)$. Using results from the lecture, can you calculate the sample mean $\bar{x}$? If so, compute it, if not, explain why.

We know the posterior:

$$\mu|x \sim \mathcal{N}(100.14, (0.07g)^2)$$

and the prior:

$$\mu \sim \mathcal{N}(100, 0.13g^2)$$

We observe, that both prior and posterior are normal distributed. Therefore $p(\mu)$ is a conjugate prior to a unknown likelihood $f(x|\mu)$. Since there is only a finite amaount of conjugate priors, we know that $X_i|\mu \sim \mathcal{N}(\mu, \tau)$ must hold.

From Bayes' Theorem we know that:

$$p(\mu|x) \propto f(x|\mu)p(\mu)$$

with

$$f(x|\mu) \propto \prod_{i=1}^{N} exp(-\frac{1}{2}\frac{(x_i - \mu)^2}{\tau^2})$$

$$\propto exp(-\frac{1}{2\tau^2}\sum_{i=1}^{N}(\mu^2 - 2x_i\mu))$$

$$= exp(\frac{N\bar{x}}{\tau^2}\mu - \frac{N}{2\tau^2}\mu^2)$$

and

$$p(\mu) \propto exp(-\frac{1}{2}\frac{(\mu - 100)^2}{0.13^2})$$
$$\propto exp(-\frac{1}{2 \cdot 0.13^2}(\mu^2 - 200\mu))$$
$$= exp(\frac{100}{0.13^2}\mu - \frac{1}{2 \cdot 0.13^2}\mu^2)$$

we can obatin:

$$f(x|\mu)p(\mu) \propto exp\left((\frac{100}{0.13^2} + \frac{N\bar{x}}{\tau^2})\mu - \frac{1}{2}(\frac{1}{0.13^2} + \frac{N}{\tau^2})\mu^2\right)$$

And with some hard thinking (or the lecture slides) conclude:

$$\mu|x \sim \mathcal{N}\left(\left(\frac{N}{\tau^2} + \frac{1}{0.13^2}\right)^{-1}\left(\frac{N\bar{x}}{\tau^2} + \frac{100}{0.13^2}\right), \left(\frac{N}{\tau^2} + \frac{1}{0.13^2}\right)^{-1}\right)$$

We know that $\mathbb{V}(\mu|x) = (0.07)^2$ which implies

$$\frac{N}{\tau^2} + \frac{1}{0.13^2} = (0.07)^{-2}$$
$$\frac{N}{\tau^2} = \frac{1}{0.07^2} - \frac{1}{0.13^2}$$
$$\tau^2 = \frac{N}{\frac{1}{0.07^2} - \frac{1}{0.13^2}}$$
$$log(\tau) = \frac{1}{2}\left[log(N) - log\left(exp[-2log(0.07)] - exp[-2log(0.13)]\right)\right]$$

```
log_tau <- 0.5 * (log(27) - log(exp(-2*log(.07)) - exp(-2*log(.13))))

exp(log_tau)
```

```
[1] 0.4316509
```

and now with $\mathbb{E}(\mu|x) = 100.14$

$$\left(\frac{N}{\tau^2} + \frac{1}{0.13^2}\right)^{-1}\left(\frac{N\bar{x}}{\tau^2} + \frac{100}{0.13^2}\right) = 100.14$$

$$\frac{N\bar{x}}{\tau^2} + \frac{100}{0.13^2} = 100.14\left(\frac{N}{\tau^2} + \frac{1}{0.13^2}\right)$$

$$\bar{x} = \frac{\tau^2}{N}\left[100.14\left(\frac{N}{\tau^2} + \frac{1}{0.13^2}\right) - \frac{100}{0.13^2}\right]$$

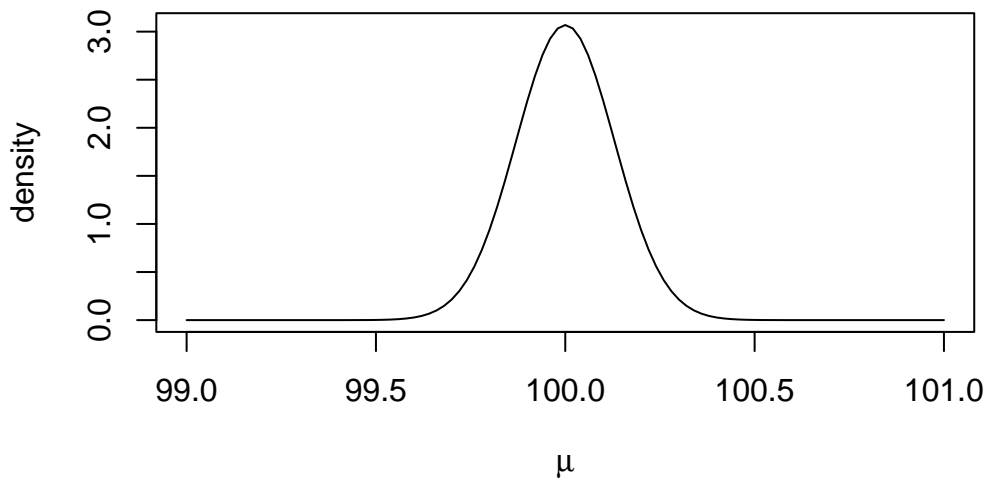$$\bar{x} = 100.14 + \frac{0.14}{0.13^2}\frac{\tau^2}{N}$$

```
100.14 + (exp(2*log_tau) / 27)*(.14/.13^2)
```

```
[1] 100.1972
```

(e)

Is the prior in subtask (d) an informative or uninformative prior, i.e. does it contain rather a lot or a little prior knowledge? Explain your thoughts.

The prior: $\mu \sim \mathcal{N}(100, 0.13g^2)$ is an informative prior. The variance is very small, meaning we are very sure about the mean.

Exercise 4.2 - Bayesian vs. Frequentist Inference

A quality control engineer tests $N = 500$ units of a newly developed safety-critical component. None of the tested units fail during the test period. That is, the number of observed failures is $x = 0$.

(a)

You are interested in estimating the unknown failure probability θ of a single unit. First, consider a frequentist perspective:

1. Estimate θ using the maximum likelihood estimator (MLE).

2. Attempt to construct a 95% confidence interval for θ. Discuss the challenges that arise in this situation.

The failure of each component gets treated as a binomial experiment:

$$X_i \sim Binom(\theta)$$

where $X_i = 1$ means failure. We compute the Likelihood of our sample:

$$L(\theta; x) = \prod_{i=1}^{N} \mathbb{P}(x_i; \theta) = \theta^x (1-\theta)^{N-x}$$

and the log-likelihood as

$$l(\theta; x) = x log(\theta) + (N-x) log(1-\theta)$$

for the score function $s(\theta) = \frac{dl(\theta)}{d\theta}$ we get:

$$s(\theta) = \frac{x}{\theta} - \frac{N-x}{1-\theta}$$

To obtain the MLE, it is neccessary to set $s(\theta) = 0$, if we do this we get:

$$x(1-\theta) = (N-x)\theta \implies \hat{\theta}_{ML} = \frac{x}{N}$$

Therefore the MLE for our data is $\hat{\theta}_{ML} = 0$.

To estimate the variance, we want to compute the Fisher Information of the MLE. First we compute $\frac{d^2 l(\theta)}{d\theta^2}$:

$$\frac{ds(\theta)}{d\theta} = -\frac{x}{\theta^2} - \frac{N-x}{(1-\theta)^2}$$

Now for the Fisher Information:

$$\mathcal{I}_N(\theta) = N\mathcal{I}(\theta)$$
$$= N \cdot (-\mathbb{E}_\theta(\frac{ds(\theta)}{d\theta}))$$
$$= N \cdot \mathbb{E}_\theta(\frac{x}{\theta^2} + \frac{1-x}{(1-\theta)^2})$$
$$= N \cdot \left[\frac{1}{\theta} + \frac{1}{(1-\theta)}\right]$$
$$= \frac{N}{\theta(1-\theta)}$$

Thus, obtaining the Variance as the inverse observed Fisher Info of the MLE.

$$\mathbb{V}(\hat{\theta}_{ML}) = \mathcal{I}_N(\hat{\theta}_{ML})^{-1} = \frac{\hat{\theta}_{ML}(1-\hat{\theta}_{ML})}{N} = 0$$

If we try to estimate the variance $S_x^2 = \frac{1}{n-1}\sum_{i=1}^n (\bar{x} - x_i)^2$ we obtain the same: $S_x^2 = 0$. If we then try to construct a CI of any level α around $\hat{\theta}_{ML}$ it will always be $[\frac{\bar{x}}{N}, \frac{\bar{x}}{N}] = \frac{\bar{x}}{N}$.

(b)

Now consider a Bayesian approach. Assume a Beta prior for θ: $\theta \sim Beta(\alpha, \beta)$.

1. Choose a reasonable prior, justifying your choice (e.g., a non-informative prior or a weakly informative prior that reflects safety-critical caution).

2. Using results from the lecture, state the posterior distribution for θ, given the observed data.

3. Compute a 95% Bayesian credibility interval (Highest Density Interval if possible) for θ.

1. Since we have no prior information, we take $\theta \sim Beta(1,1) = Unif(1,0)$, thus maximizing the entropy. The distribution is then given by $f(\theta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$

2. We take our usual setup: $p(\theta; x) \propto f(x; \theta)p(\theta)$

$$p(\theta; x) \propto (\theta)^x(1-\theta)^{N-x}\theta^{\alpha-1}(1-\theta)^{\beta-1} = (1-\theta)^{N-x+\beta-1}(\theta)^{x+\alpha-1}$$

and see that this looks familiar, thus: $\theta|x \sim Beta(x+\alpha, N-x+\beta)$. So in our case with $x=0$ and $N=500$: $\theta|x \sim Beta(1, 501)$

```r
get_CI <- function(level, n, alpha, beta, plot = FALSE) {
  checkmate::assertIntegerish(level,
                              len=1,
                              any.missing = FALSE,
                              lower = 1,
                              upper = 20)
  res <- matrix(nrow = n, ncol = 5,
                dimnames = list(NULL,  c("lower perc",
                                         "lower value",
                                         "upper perc",
                                         "upper value",
                                         "interval length")))
  res[,1] <- seq(0, level/100, length.out = n)
  res[,2] <- vapply(res[,1], function(x) qbeta(x, alpha, beta), numeric(1))
  res[,3] <- 1-level/100 + res[,1]
  res[,4] <- vapply(res[,3], function(x) qbeta(x, alpha, beta), numeric(1))
  res[,5] <- res[,4] - res[,2]

  if(plot) plot(res[2:(n-1), 1],
                res[2:(n-1), 5],
                type = "l",
                xlab = "lower percentile",
                ylab = "Length of Interval",
                main = "95% credibility interval length vs. lower percentile")

  return(invisible(res[which.min(res[,5]),]))
}
```
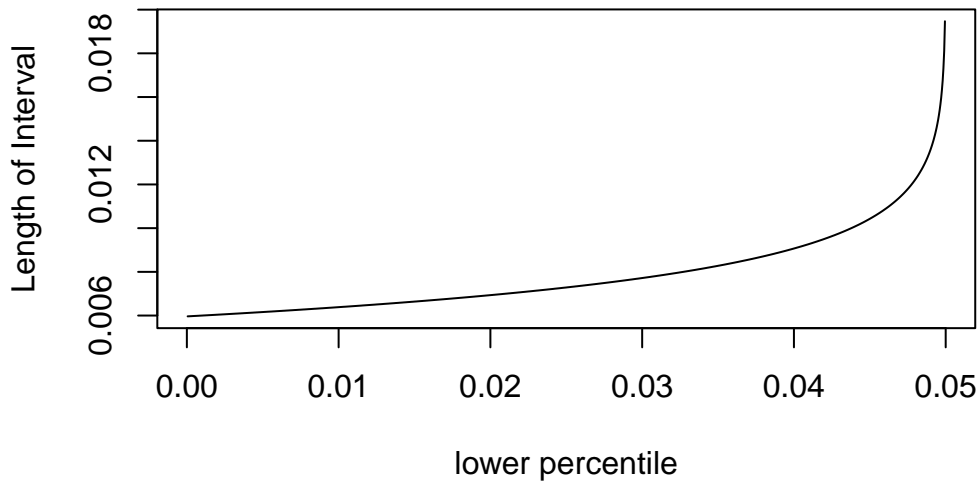
```r
get_CI(5, 1000, 1, 501, plot = TRUE)
```

95% credibility interval length vs. lower percentile


```
get_CI(5, 1000, 1, 501)[c(2,4)]
```

```
lower value upper value
0.000000000 0.005961664
```

(c)

> Compare the frequentist and Bayesian intervals. Which one provides a more informative
> and realistic assessment of uncertainty in this context? Why?

The frequentist interval only takes the data into account, therefore yielding a point estimate
instead of a interval. The baysian approach provides us with a posterior distribution and a
small, yet interpretable interval. Since we think that a part has to break finally, the baysian
approach with a non-zero estimate for θ is a more realistic assessment of the uncertainty.

Exercise 4.3 - Conjugacy vs. MCMC

> One of the methods used by Astronomers to estimate the unknown mass μ of Stellar-
> mass black holes (1-100 solar masses) involves measuring X-rays from the accretion disc.
> However, actual measurements are costly and scientists thus have to work with as little
> data as possible. It is known that the measurements x_i are normally distributed with
> known measurement error (variance) $\sigma^2 = 15$:
>
> $$X_i \sim \mathcal{N}(\mu, \sigma^2) \qquad i = 1, ..., 5$$

We want to estimate μ of one such black hole with the following observed data (in solar masses):

$$x = (7.7, 8.3, 8.1, 8.1, 8.8)$$

We assume a prior distribution for μ with

$$\mu \sim \mathcal{N}(\mu_0, \tau_0^2)$$

with $\mu_0 = 8$ solar masses and $\tau_0^2 = 4(\text{solar masses})^2$. Under the Normal-Normal conjugate model, the posterior for μ is

$$\mu|x \sim \mathcal{N}(\mu_n, \tau_n^2)$$

with

$$\mu_n = \frac{\tau_0^{-2}\mu_0 + n\sigma^{-2}\bar{x}}{\tau_0^{-2} + n\sigma^{-2}} \qquad \tau_n^2 = \frac{1}{\tau_0^{-2} + n\sigma^2}$$

(a)

Assume that point estimates $m_1, m_2, ..., m_{64}$ from 64 previously measured black holes are available. Explain how these could have been used to determine the prior hyperparameters μ_0 and τ_0^2. Does this approach contradict the Bayesian idea?

We could use the point estimates to compute a posterior distribution, which then can be used as a prior distribution later on. This approach does not contradict the bayesian idea, on the contrary, we use information to obtain a posterior, and on getting hold of new information use this posterior as the prior.

(b)

How could empirical Bayes (methods) be used instead to specify the prior? Calculate.

Empirical Bayes uses the data to estimate the prior, which contradicts the bayesian idea, yet still yields usefull result frequently.

$$\mu_0 = \bar{x}\tau_0 = S_x^2$$

```r
x <- c(7.7, 8.3, 8.1, 8.1, 8.8)
c("mu_0" = mean(x), "tau_0_sq" = var(x))
```

```
    mu_0 tau_0_sq
    8.20     0.16
```

(c)

Compute the posterior mean μ_n and variance τ_n^2 given the observed data and interpret the result.

```r
get_new_params <- function(X, mu_0, tau_0_sq, sigma_sq) {
  checkmate::assertNumeric(X, min.len = 1, any.missing = FALSE, finite = TRUE)
  checkmate::assertNumeric(mu_0, len = 1, any.missing = FALSE, finite = TRUE)
  checkmate::assertNumeric(tau_0_sq, len = 1, any.missing = FALSE, lower = 0)
  checkmate::assertNumeric(sigma_sq, len = 1, any.missing = FALSE, lower = 0)
  n <- length(X)
  mu_n <- (mu_0/tau_0_sq + sum(X)/sigma_sq) / (1/tau_0_sq + n/sigma_sq)
  tau_n_sq <- 1 / (1/tau_0_sq + n/sigma_sq)
  c("mu_n" = mu_n, "tau_n_sq" = tau_n_sq)
}
```

```r
params <- get_new_params(X = c(7.7, 8.3, 8.1, 8.1, 8.8),
                         mu_0 = 0,
                         tau_0_sq = 4,
                         sigma_sq = .2)
params
```

```
      mu_n    tau_n_sq
8.11881188 0.03960396
```
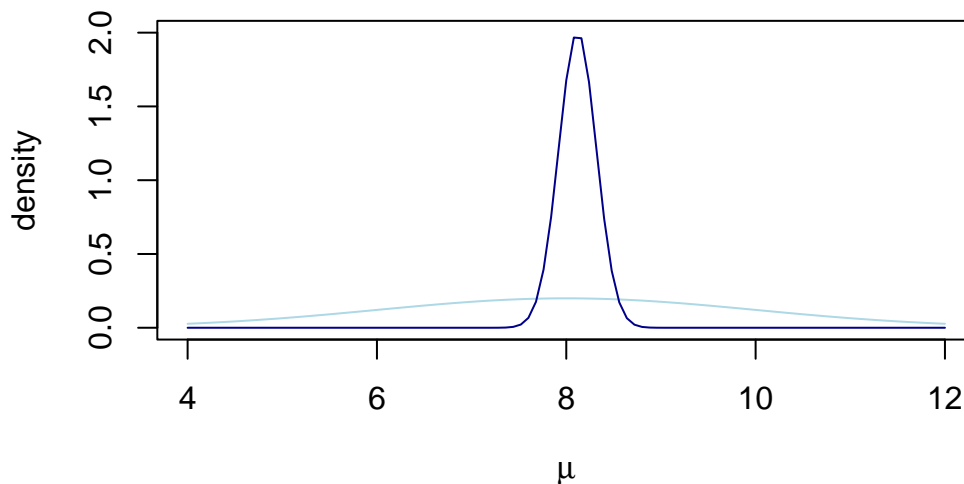

Through our observed data, we get significantly more certain about μ, the prior was almost non-informative in comparison to the posterior distribution.

(d)

Suppose that we chose the normal likelihood only for convenience and actually, the measurements follow a Laplace distribution, i.e.

$$X_i \sim Laplace(\mu, b)$$

with known $b = \sqrt{\frac{1}{10}}$.

Using this new likelihood, use MCMC to estimate the posterior of μ via Metropolis-Hastings. Implement the algorithm in R (choose reasonable settings or parameters, if necessary) and plot the posterior sample you obtain. How do we obtain a posterior point estimate for μ?

```r
mcmc <- function(x,
                 mu_0,
                 tau_0,
                 b,
                 n = 1000,
                 sigma = 1) {
  checkmate::assertNumeric(x, min.len = 1, any.missing = FALSE, finite = TRUE)
  checkmate::assertNumeric(mu_0, len = 1, any.missing = FALSE, finite = TRUE)
  checkmate::assertNumeric(tau_0, len = 1, any.missing = FALSE, lower = 0)
  checkmate::assertNumeric(b, len = 1, any.missing = FALSE, lower = 0)
  checkmate::assertNumeric(sigma, len = 1, any.missing = FALSE, lower = 0)
  checkmate::assertIntegerish(n, len = 1, any.missing = FALSE, lower = 1)
  ### density of laplace RV
  dlaplace <- function(x, mu, s, log = FALSE) {
    if (log) {
      return(-abs(x - mu)/s - log(2*s))
    } else {
      return(exp(-abs(x - mu)/s) / (2*s))
    }
  }

  ### prop to posterior
  f.test <- function(mu, log = FALSE) {
    if (log) {
      sum(dlaplace(x, mu, b, log = TRUE)) + dnorm(mu, mu_0, tau_0, log = TRUE)
    } else {
      prod(dlaplace(x, mu, b)) * dnorm(mu, mu_0, tau_0)
    }
  }
}
```

```
  samp <- numeric(n)
  accepted <- 0
  mu <- mu_0
  for (i in seq_len(n)) {
    mu_star <- rnorm(1, mu, sigma)
    U <- runif(1)
    alpha <- min(1, exp(f.test(mu_star, log = TRUE) - f.test(mu, log = TRUE)))
    if(alpha >= U) {
      mu <- mu_star
      accepted <- accepted + 1
    }
    samp[[i]] <- mu
  }

  structure(samp, "acceptance rate" = accepted/n)
}
```

```
set.seed(123)
mcmc_sample <- mcmc(x = c(7.7, 8.3, 8.1, 8.1, 8.8),
                    mu_0 = 8,
                    tau_0 = 2,
                    b = sqrt(1/10),
                    n = 1000,
                    sigma = .5)
attr(mcmc_sample, "acceptance rate")
```
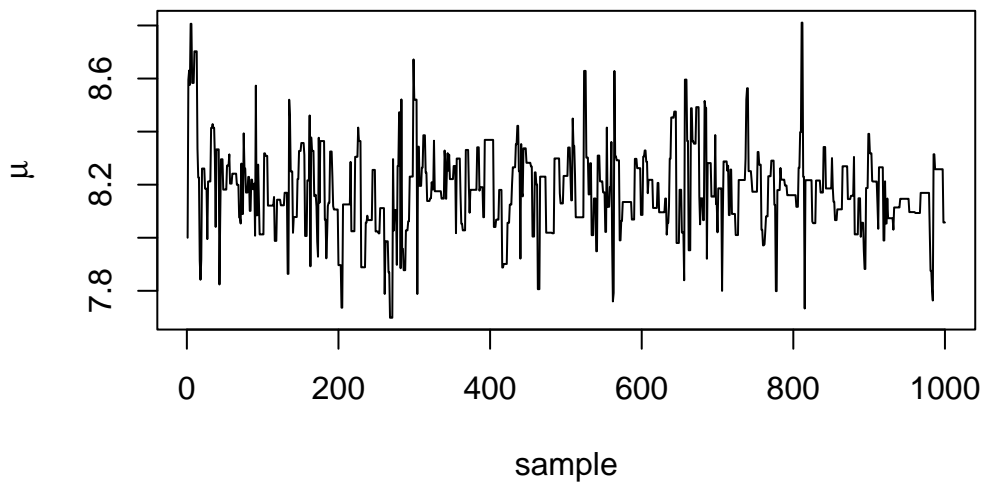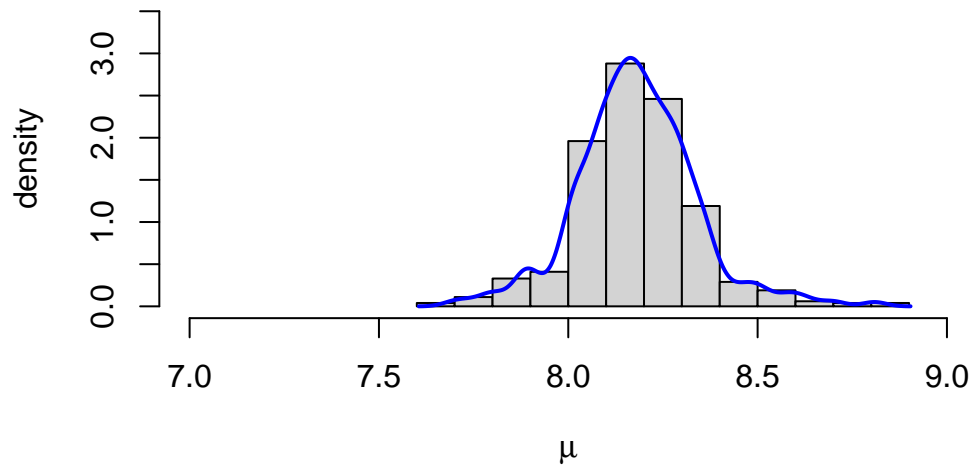
```
[1] 0.352
```

Posterior Sample

To get a posterior point estimate of μ, we want to compute $\mathbb{E}(\mu|x)$. Since we are not able to compute this expected value analytically, we do the next best thing and compute the mean of the sample we produced, using MCMC (we discard the first 50 values).

```
mean(mcmc_sample[50:1000])
```

```
[1] 8.172752
```