Computational Biology Group

D BSSE

ETH zürich

Introduction to Bayesian Statistics with R

4: Exercises

Jack Kuipers

28 November 2022

Exercise 4.1 - Credible intervals

The likelihood of data D made up of N observations of x: $x_i, i = 1, \ldots, N$, under a normal model with mean μ and sd σ is

$$P(D \mid \mu, \sigma) \propto \prod_{i=1}^{N} \frac{1}{\sigma} \mathrm{e}^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma^N} \mathrm{e}^{-\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma^N} \mathrm{e}^{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{N} x_i^2 - 2\mu\sum_{i=1}^{N} x_i + N\mu^2\right]}$$

so that, by precomputing the sum of x_i^2 and x_i we can evaluate the likelihood with the following functions

```
x2 <- sum(x^2)
x1 <- sum(x)
x0 <- length(x) # number of observations
g_like <- function(mu, sigma, x2, x1, x0) {
  exp(-(x2 - 2*mu*x1 + x0*mu^2)/(2*sigma^2))/(sigma^x0)
}
```

Note: for numerical accuracy we often log-transform and work with the log-likelihood.

For now we keep $\sigma = 1$ fixed and have a Student-t prior on μ with scale t_s and ν degrees of freedom

```
prior_t <- function(mu, t_s, nu) {
  dt(mu/t_s, nu)/t_s
}
```

Update the `basicMCMC` code from Exercise 3.2 to sample μ from its posterior distribution, depending on the data x, and the prior choices t_s and ν.

For data we will use Gosset's data on barley seed yield from his 1908 paper which introduced the Student-t distribution. The yields (in US cwt per acre) for 11 plots of land for normal seed and seed which has been dried in a kiln are in `seed_data.csv`.

- Read in the data, and extract the relevant quantity as x
- Choose prior parameters t_s and ν
- Sample from the posterior distribution of μ
- Visualise this distribution
- Obtain estimates for the 95% credible interval of μ