University of Edinburgh
School of Mathematics

Bayesian Data Analysis

Exercise 1

Gene expression data analysis.

Expression (concentration) of each of 50 genes was observed in the pancreas tissue of two groups of people: in a case group of 12 people who have pancreatic cancer, and in a control group of 10 people who do not have pancreatic cancer (the groups are approximately matched by age and gender). The study was performed in the same hospital.
The question is to determine whether there is a difference between the gene expression in the pancreas tissue between the groups of people with and without the cancer.

Statistical analysis.

Introduce the following random variables. Denote the logarithm of the gene expression of the kth gene for individual i in the case group by X_{ik}, and for the jth individual in the control group by Y_{jk}, $k = 1, 2, \ldots, N = 50$, $i = 1, \ldots, n = 12$, $j = 1, 2, \ldots, m = 10$.
Possible likelihood:

$$X_{ik} \mid \mu_1, \sigma_1 \sim N(\mu_1, \sigma_1^2),\ i = 1, 2, \ldots, n \quad \textit{independently (given } \mu_1, \sigma_1^2),$$
$$Y_{jk} \mid \mu_2, \sigma_2 \sim N(\mu_2, \sigma_2^2),\ j = 1, 2, \ldots, m \quad \textit{independently (given } \mu_2, \sigma_2^2).$$

Also, X_{ik} and Y_{jk} are independent for all i, j, k.

1. Under the likelihood above, state the distributions of

$$\bar{X} = \frac{1}{nN}\sum_{i=1}^{n}\sum_{k=1}^{N} X_{ik}, \quad \bar{Y} = \frac{1}{mN}\sum_{j=1}^{m}\sum_{k=1}^{N} Y_{jk},$$
$$s_X^2 = \frac{1}{nN}\sum_{i=1}^{n}\sum_{k=1}^{N}(X_{ik} - \bar{X})^2, \quad s_Y^2 = \frac{1}{mN}\sum_{j=1}^{m}\sum_{k=1}^{N}(Y_{jk} - \bar{Y})^2.$$

2. The observed data can be summarised as follows: $\bar{x} = 4.03$, $\bar{y} = 2.59$, $s_X = 0.29$ and $s_Y = 0.11$. Discuss whether there is any loss of information by using only these data summaries for this likelihood.

3. Discuss the interpretation of μ_1 and μ_2, and of $\exp(\mu_1)$ and of $\exp(\mu_2)$. Propose a way to address the question of interest in terms of μ_1 and μ_2, whether there is a difference between the gene expression in the pancreas tissue between the groups of people with and without the cancer.

4. Assume σ_1 and σ_2 are known, for now fix them to be 0.3 and 0.1, respectively. Propose two 'non-informative' priors for μ_1 (that belong to a conjugate family, possibly in the limit) (Priors 1 and 2).

5. Now suppose that you have additional information from a study in several other hospitals that a similar Bayesian analysis of the log gene expression of the same 50 genes in control groups only produced the posterior distribution of μ_2 to be $N(2.38, 0.04^2)$, and use it as a prior. Discuss the implications.

6. Posterior analysis.

   (a) For each of the 3 priors for μ_2 (2 non-informative and 1 informative), determine the corresponding posterior distribution.

   (b) For each prior for μ_2, produce the posterior summaries: mean, median, standard deviation, and two-sided 90% credible interval, and compare them.

   (c) Check how each of the priors of μ_2 affects the inference by producing prior / likelihood / posterior plots. Discuss if there is any conflict between the informative prior and the likelihood, and if there is, discuss possible reasons. Choose the prior that has the least effect on the corresponding posterior for further analysis.

   (d) Use the same two non-informative priors for μ_1 you proposed in Question 4, and address (a)-(c).

   (e) For the priors for μ_1 and μ_2 chosen in 5(c), compare the posterior distributions of μ_1 and μ_2. Use them to address the question of interest (give the final conclusion in terms of the original question).

7. Discuss which assumptions on the likelihood may be questioned.