University of Edinburgh
School of Mathematics

Bayesian Data Analysis

Assignment 1

Students' goals.

229 students (109 boys and 120 girls) aged 7-13 from 9 schools were asked whether popularity or sporting ability was most important to them. The outcome is summarised in the table below. The question is to determine whether there is a difference between the importance of popularity and of sporting ability for girls and boys.

	Sporting ability	Popularity
Boys	59	50
Girls	30	90

Statistical analysis.
Introduce the following random variables. Consider the ith boy, $i = 1, 2, \ldots, n = 109$, and set $X_i = 1$ if popularity is more important to him than sports, and $X_i = 0$ otherwise. Similarly, for the jth girl, $j = 1, 2, \ldots, m = 120$, set $Y_j = 1$ if popularity is more important to her than sports, and $Y_j = 0$ otherwise.
Possible likelihood:
$X_i \mid \theta_1 \sim Bern(\theta_1)$, $\theta_1 \in (0, 1)$, $i = 1, 2, \ldots, n$ independently given θ_1,
and $Y_j \mid \theta_2 \sim Bern(\theta_2)$, $\theta_2 \in (0, 1)$, $j = 1, 2, \ldots, m$ independently given θ_2.
Also, X_i and Y_j are independent for all i, j.

1. Under the likelihood above, state the distributions of $S_X = \sum_{i=1}^{n} X_i$ and $S_Y = \sum_{j=1}^{m} Y_j$. Discuss whether there is any loss of information by using only the data summaries for this likelihood. [1 mark]

2. Discuss the interpretation of θ_1 and θ_2. Propose a way to address the question of interest in terms of θ_1 and θ_2, whether there is a difference between the importance of popularity and of sporting ability for girls and boys. [2 marks]

3. Propose two 'non-informative' priors for θ_1 that belong to the family of conjugate Beta distributions. [1 mark]

4. Now suppose that you have additional information that a Bayesian analysis of data from a different boys-only school produced the posterior distribution of θ_1 to be $Beta(21, 10)$, and use it as a prior (the boys are in the same age group, 7-13 years old). (Discuss the implications of such choice of the prior.)

5. Posterior analysis.

   (a) For each of the 3 priors for θ_1 (2 non-informative and 1 informative), determine the corresponding posterior distribution. [1 mark]

(b) Check how each of the priors of θ_1 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 (for θ_1 only). Select one prior that least affects the posterior for further analysis. [2 marks]

(c) Use the same two non-informative priors for θ_2 you proposed in Question 3, and address (a)-(b). [1 mark]

(d) For the chosen priors for θ_1 and θ_2, produce the posterior summaries: mean, median, standard deviation, and two-sided 90% credible interval. [1 mark]

(e) For the priors for θ_1 and one for θ_2 chosen in 5(b), compare the posterior distributions of θ_1 and θ_2. Use them to address the question of interest (give the final conclusion in terms of the original question). [1 mark]

6. (Discuss which assumptions on the likelihood may be questioned.)