ST4234: Bayesian Statistics

Tutorial 1, AY 19/20

1. Suppose we are going to sample 100 individuals from a county (of size much larger than 100) and ask each person whether they support policy Z or not. Let $Y_i = 1$ if person i in the sample supports the policy, and $Y_i = 0$ otherwise.

   (a) Assume $Y_1, \ldots, Y_{100}$, are, conditional on θ, i.i.d. binary random variables with expectation θ. Write down the joint distribution of $\mathrm{P}(Y_1 = y_1, \ldots, Y_{100} = y_{100}|\theta)$ in a compact form. Let $Y = \sum_{i=1}^{100} Y_i$. Write down the form of $\mathrm{P}(Y = y|\theta)$ as well.

   (b) For the moment, suppose you believed that $\theta \in \{0.0, 0.1, \ldots, 0.9, 1.0\}$. Given that the results of the survey were $Y = 57$, compute $\mathrm{P}(Y = 57|\theta)$ for each of these 11 values of θ and plot these probabilities as a function of θ.

   (c) Now suppose you originally had no prior information to believe one of these θ-values over another, and so $\mathrm{P}(\theta = 0.0) = \mathrm{P}(\theta = 0.1) = \cdots = \mathrm{P}(\theta = 0.9) = \mathrm{P}(\theta = 1.0)$. Use Bayes' rule to compute $p(\theta|Y = 57)$ for each θ-value. Make a plot of this posterior distribution as a function of θ.

   (d) Now suppose you allow θ to be any value in the interval $[0, 1]$. Using the uniform prior density for θ, so that $p(\theta) = 1$, plot $p(\theta) \times \mathrm{P}(Y = 57|\theta)$ as a function of θ.

   (e) Write down the posterior distribution of θ when the uniform density on $[0, 1]$ is used as a prior for θ. Plot this posterior density as a function of θ. Discuss the relationships among all of the plots you have made for this exercise.

2. Suppose you are interested in the proportion of females in a large organization and that as a first step in your investigation you intend to find out the gender of the first 11 members on the membership list. Before doing so, you have prior beliefs which you regard as equivalent to 25% of this data, and your beliefs suggest that a third of the membership is female.

(a) Suggest a suitable prior distribution and plot this prior distribution as a function of θ. Find its standard deviation.

(b) Suppose that 3 of the first 11 members turn out to be female; find your posterior distribution. Plot this posterior distribution and find its mean, median and mode.

(c) Find the 50% highest posterior density region for θ and the 50% quantile-based confidence interval for θ. Compare these two intervals.

(d) Would it surprise you to learn that in fact 86 of the total number of 433 members are female?

3. Consider a pilot study in which $n_1 = 15$ children enrolled in special education classes were randomly selected and tested for a certain type of learning disability. In the pilot study, $y_1 = 2$ children tested positive for the disability.

(a) Using a uniform prior distribution, find the posterior distribution of θ, the fraction of students in special education classes who have the disability.

Researchers would like to recruit students with the disability to participate in a long-term study, but first they need to make sure they can recruit enough students. Let $n_2 = 278$ be the number of children in special education classes in this particular school district, and let Y_2 be the number of students with the disability.

(b) Find $\mathrm{P}(Y_2 = y_2|Y_1 = 2)$, the posterior predictive distribution of Y_2, as follows:

(i) Discuss what assumptions are needed about the joint distribution of (Y_1, Y_2) such that the following is true:

$$\mathrm{P}(Y_2 = y_2|Y_1 = 2) = \int_0^1 P(Y_2 = y_2|\theta)\ p(\theta|Y_1 = 2)\ \mathrm{d}\,\theta.$$

(ii) Now plug in the forms for $\mathrm{P}(Y_2 = y_2|\theta)$ and $p(\theta|Y_1 = 2)$ in the above integral.

(iii) Figure out what the above integral must be by using the fact that any proper probability density function must integrate to 1.

(c) Plot the function $\mathrm{P}(Y_2 = y_2|Y_1 = 2)$ as a function of y_2. Obtain the mean and standard deviation of Y_2, given $Y_1 = 2$.

(d) The posterior mode and the MLE (maximum likelihood estimate) of θ, based on data from the pilot study, are both $\hat{\theta} = 2/15$. Plot the distribution $\mathrm{P}(Y_2 = y_2|\theta = \hat{\theta})$, and find the mean and standard deviation of Y_2 given $\theta = \hat{\theta}$. Compare these results to the plots and calculations in (c) and discuss any differences. Which distribution for Y_2 would you use to make predictions, and why?