Bayesian course - problem set 2 (lecture 3)

November 11, 2016

1 WHO's reported novel disease outbreaks

Suppose that you are interested in modelling the number of outbreaks of novel diseases that the WHO reports each year. Since these outbreaks are of new diseases, you assume that you can model the outbreaks as **independent** events, and hence decide to use a poisson likelihood; $X_t \sim \text{poisson}(\lambda)$, where X_t is the number of outbreaks in year t, and λ is the mean number of outbreaks.

Problem 1.1 *You decide to use a gamma(3,0.5) prior for the mean parameter (λ) of your poisson likelihood (where a gamma(α, β) is defined to have a mean of $\frac{\alpha}{\beta}$). Graph this prior.*

Problem 1.2 *Suppose that the number of new outbreaks over the past 5 years is $X = (3, 7, 4, 10, 11)$. Using the conjugate prior rules for a poisson distribution with a gamma prior, find the posterior and graph it.*

Hint: look at the table on this page, `https://en.wikipedia.org/wiki/Conjugate_prior` *and use the rule for a gamma(α,β) distribution, not for a gamma(k,θ).*

Problem 1.3 *Generate 10,000 samples from the posterior predictive distribution, and graph the distribution. To do this we first independently sample a value λ_i from the posterior distribution, then sample a value of X from a poisson(λ_i) distribution. We carry out this process 10,000 times.*

Hint: use R's `rgamma` *and* `rpois` *functions to draw (pseudo-)independent samples from the gamma nd poisson distributions respectively.*

Problem 1.4 *Compare the actual data with your 10,000 posterior predictive samples. Does your model fit the data?*

Problem 1.5 *(Optional) can you think of a better posterior predictive check to carry out on the data?*

Problem 1.6 *The WHO issues a press release where they state that the number of novel disease outbreaks for this year was 20. Use your posterior predictive samples to test whether your model is a good fit to the data.*

Problem 1.7 *By using your previously-determined posterior as a prior, update your posterior to reflect the new datum. Graph the PDF for this new distribution.*

Problem 1.8 *Generate posterior predictive samples from your new posterior and use it to test the validity of your model.*

Problem 1.9 *Would you feel comfortable using this model to predict the number of disease outbreaks next year?*

2 A fairground game

At a fairground a man advertises a gambling game that allows participants the chance to win a monetary prize, if they pay an entrance fee. The game sequence goes like this,

1. You pay £X.
2. The man flips a coin fairly (i.e. with an equal chance of the coin landing heads or tails up),
    - If the coin lands tails up, the game ends and you walk away with nothing.
    - If the coin lands heads up, he flips the coin a further two times and you receive the total number of heads across these two flips, H.

Problem 2.1 *Calculate the expected value of your winnings W if you participate, and hence determine the fair price of the game.*

Problem 2.2 *Create an R function that simulates a single run of the game, and use this to estimate the expected value of your winnings.*
Hint: use R's `rbinom` *and* `ifelse` *functions.*

Problem 2.3 *Suppose that you pay £1 for each game, and start with £10 in your pocket. By using your previously-created function, or otherwise, determine the expected number of games you can play before going broke.*

Problem 2.4 *Suppose you start with £10, and play the game 100 times (stopping only if your wealth is below the price of entry), each time paying £0.49. You want to insure against the risk of losing all your wealth. What is the fair price to pay for such an insurance scheme?*

In another game at the fair, the game's sequence goes like this,

1. You pay £X.
2. The man flips a coin fairly (i.e. with an equal chance of the coin landing heads or tails up),

- If the coin lands tails up, the game ends and you walk away with nothing.
- If the coin lands heads up, the game continues (see below).

3. Assuming the coin has landed heads up, the host then flips ten fair coins and records the number of heads H,
   - If there are five or more heads, you receive £H^2.
   - Otherwise you win nothing.

Problem 2.5 *By creating a function in R that generates random realisations of the above game, or otherwise, determine a fair value for X.*
Hint: use R's `rbinom` *and* `ifelse` *functions.*

Problem 2.6 *Estimate the probability distribution of game winnings for for an entry fee of £10.*

3 Sleep-deprived reactions

These data are from a study described in Belenky et al. (2003) [2] that measured the effect of sleep deprivation on cognitive performance. There were 18 subjects chosen from a population of interest (lorry drivers) who were restricted to 3 hours of sleep during the trial. On each day of the experiment their reaction time to a visual stimulus was measured. The data for this example are contained within "prob2_sleepstudy.csv", and contains three variables: Reaction, Days and Subject ID which measure the reaction time of a given subject on a particular day.

A simple model that explains the variation in reaction times is a linear regression model of the form:

$$R(t) \sim N(\alpha + \beta t, \sigma) \tag{1}$$

where $R(t)$ is the reaction time on day t of the experiment across all observations.

Problem 3.1 *By graphing all the data here critically assess the validity of the model to the data.*

Problem 3.2 *Graph the data at the individual subject data using R's "lattice" package, or otherwise. What does this suggest about assuming a common β across all participants?*

Problem 3.3 *I have fit the above model to the data using MCMC, and 2,000 samples from the posterior distribution for (α, β, σ) are contained within the file "prob2_sleepPosteriors.csv". Generate samples from the posterior predictive distribution, and visualise them in an appropriate way.*

Problem 3.4 *How does the posterior predictive data compare to the actual data?*

Problem 3.5 *How (if at all) do the posterior predictive checks suggest we need to change our model?*

4 Independent sampling

An analysis results in a posterior with the following probability density function:

$$f(x) = \begin{cases} \frac{1}{1.335} \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}, & \text{if } x < 0.9735. \\ 0.186056, & \text{if } 0.9735 \leq x \leq 5. \end{cases} \tag{2}$$

Problem 4.1 *Verify that this is a valid PDF (hint: see R's numerical integration function).*

Problem 4.2 *Using independent sampling estimate the mean and variance of this distribution.*

Problem 4.3 *Construct uncertainty intervals around your estimates of the mean.*

Problem 4.4 *Verify your previous answer by calculating the mean and variance of this distribution.*

Problem 4.5

$$\begin{aligned} \mathrm{E(X)} &= \int x f(x) \mathrm{d}x \\ &= \int x \frac{f(x)}{g(x)} g(x) \mathrm{d}x \end{aligned}$$

On the basis of the above equation give another way we estimate the mean?

Problem 4.6 *Using the above method, find an estimate of the mean and an estimate of its uncertainty.*

Problem 4.7 *How should we choose $g(x)$ to yield estimators with the lowest variance? (Difficult)*

5 Discoveries data

The file "prob2_discoveries.csv" contains data on the numbers of "great" inventions and scientific discoveries in each year from 1860 to 1959 [1]. The aim of this problem is for you to build a statistical model that provides a reasonable approximation to this series. As such, you will need to choose a likelihood, specify a prior on any parameters, and go through and calculate a posterior. Once you have a posterior, you will want to carry out posterior predictive checks to see that your model behaves as desired.

6 Integration by sampling

Work out the following integrals by sampling.

Problem 6.1

$$\int_{-\infty}^{\infty} \frac{x^6}{\sqrt{2\pi}} \times exp(-\frac{x^2}{2})\mathrm{d}x \tag{3}$$

Problem 6.2

$$\int_{1}^{\infty} \frac{x^3}{\sqrt{2\pi}} \times exp(-\frac{x^2}{2})\mathrm{d}x \tag{4}$$

å

Problem 6.3

$$\int_{1}^{\infty} \frac{x^6}{\sqrt{2\pi}} \times exp(-\frac{x^2-4x}{2})\mathrm{d}x \tag{5}$$

Problem 6.4

$$\int_{1}^{10} x^6 \frac{e^{-\frac{x^4}{2}}}{\sqrt{2\pi}}\mathrm{d}x \tag{6}$$

Problem 6.5 *What is the uncertainty using sampling to evaluate integrals?*

7 Markovian coin

Consider a type of coin for which the result of the next throw - heads or tails - can depend on the result of the current throw. In particular if a "heads" is thrown then the probability of obtaining a "heads" on the next throw is $(\frac{1}{2}+\epsilon)$; if instead a tails is thrown then the probability of obtaining a "tails" on the next throw is $(\frac{1}{2}+\epsilon)$. To start we assume $0 \leq \epsilon \leq \frac{1}{2}$. The random variable $X = 0, 1$ if the coin lands "tails-up" or "heads-up" on a given throw.

Problem 7.1 *Find the mean and variance of the coin, supposing it starts with probability $\frac{1}{2}$ on each side.*

Problem 7.2 *Computationally estimate the mean of the coin by simulating 10, 20, and 100 throws for $\epsilon = 0$.*

Problem 7.3 *As $\epsilon \uparrow$ how does the error in estimating the mean change, and why?*

Problem 7.4 *When $\epsilon = \frac{9}{20}$ calculate the effective sample size of an actual sample size of 100. How does the effective sample size depend on ϵ?*

Problem 7.5 *Now assume that $\epsilon = -\frac{1}{3}$ - what is the effective sample size of an actual sample size of 100? Explain your result.*

8 Markovian die

Consider a type of die whose next value thrown can *depend* on the current value. The degree of dependence is specified by a parameter $0 \leq \epsilon \leq 1$ (see figure 1). If $\epsilon = 0$ then each separate throw of the die can be considered *independent* of the previous value. Another way of saying this is that each number has an equal probability of being thrown irrespective of the current value. If $\epsilon = 1$ then there is strong dependence from one throw to the next, where from a given number on a throw only neighbouring numbers are possible on the next. So $1 \to (6, 2)$, $2 \to (1, 3)$ etc. If $0 < \epsilon < 1$ we suppose that there is preference towards consecutive numbers, with the preference increasing in ϵ.

For all values of ϵ we assume that both the forward and backward steps are equally-likely, so $1 \to 2$ and $1 \to 6$ are of the same probability. When If $0 < \epsilon < 1$, we suppose that those transitions that are not neighbours are all of the same probability (which is less than the probability of consecutive numbers).

Specifically, we define ϵ in the following way:

$$Pr(X_{n+1}|X_n) = \frac{1}{6}(1-\epsilon) + \frac{\epsilon}{2} 1_{X_{n+1} \in \mathcal{C}(X_n)}$$

Where $1_{X_{n+1} \in \mathcal{C}(X_n)}$ is an indicator function which is equal to 1 if the next value of the die, X_{n+1} is in the neighbour set $\mathcal{C}(X_n)$ of the current value, X_n. (The above is just a fancy way of saying that we increase the probability of neighbours by an amount $\frac{\epsilon}{2}$ relative to the non-neighbours.)

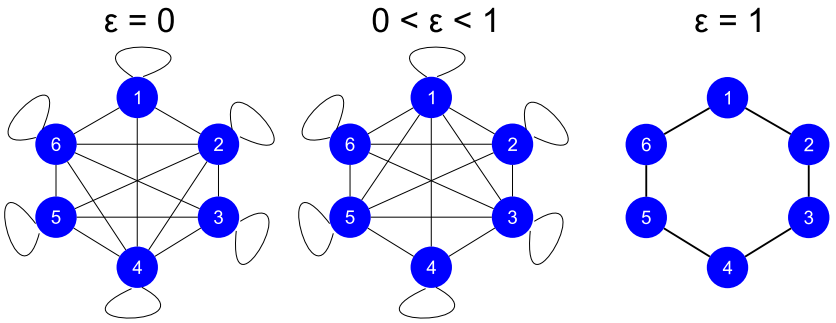

Figure 1: A Markovian die where ϵ determines the degree of dependence between throws of the die.

Problem 8.1 *Find the mean of the die across all values of ϵ assuming it starts on a randomly-selected side.*

Problem 8.2 *By simulating throws of the die, find an estimator of its mean.*

Problem 8.3 *Compute the error in estimating the mean as ϵ is varied at a sample size of 5, 10,and 100.*

Problem 8.4 *Find the effective sample size of 100 throws (when estimating the mean) for a die where* $\epsilon = 1$*. Comment on the effect of dependence on sampling.*

Problem 8.5 *Now suppose that the die starts always on side 2. Find the expectation of the die (not the running total, just the current value) at each time step. (Difficult!)*

Problem 8.6 *Following on from the last question find how long we need to leave the die before we are confident we are sampling from its unconditional distribution. (By "unconditional" here, we mean its probability distribution disregarding its start point.) (Difficult!)*

Problem 8.7 *Carry out the above investigations but for a die with* n *sides. How does* n *affect the results?*

References

[1] *The World Almanac and Book of Facts*. 1975.

[2] Gregory Belenky, Nancy J Wesensten, David R Thorne, Maria L Thomas, Helen C Sing, Daniel P Redmond, Michael B Russo, and Thomas J Balkin. Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: A sleep dose-response study. *Journal of sleep research*, 12(1):1–12, 2003.