STA623 - Bayesian Data Analysis

Practical 1 & 2

Marc Henrion

2024-10-29

Practicals 1 & 2

Notation

- X, Y, Z - random variables
- x, y, z - measured / observed values
- $\bar{X}$, $\bar{Y}, \bar{Z}$ - sample mean estimators for X, Y, Z
- $\bar{x}$, $\bar{y}, \bar{z}$ - sample mean estimates of X, Y, Z
- $\widehat{T}$, $\widehat{t}$ - given a statistic T, estimator and estimate of T
- $P(A)$ - probability of an event A occuring
- $f_X(.)$, $f_Y(.), f_Z(.)$ - probability mass / density functions of X, Y, Z; sometimes $p_X(.)$ etc. rather than $f_X(.)$
- p(.) - used as a shorthand notation for pmfs / pdfs if the use of this is unambiguous (i.e. it is clear which is the random variable)
- $X \sim F$ - X distributed according to distribution function F
- $E[X]$, $E[Y]$, $E[Z]$, $E[T]$ - the expectation of X, Y, Z, T respectively

Exercise 1

A company has developed a new diagnostic test, for a disease D with prevalence 10%, which can give a positive (T) or negative ($\bar{T}$) result. There is also an already existing test, which can also give a positive (T_{old}) or negative ($\bar{T}_{old}$) result.

The company is very proud that the new test has much better sensitivity ($P(T|D) = 0.99$) than the old one ($P(T_{old}|D) = 0.8$). However this increase in sensitivity comes at the cost of specificity: $P(\bar{T}|\bar{D}) = 0.9$ compared to the existing test $P(\bar{T}_{old}|\bar{D}) = 0.95$.

What is the probability of a patient having the disease if the test result for the new test is positive?

For the existing test?

We can assume that, given the disease state (D or $\bar{D}$), the two tests are independent as they use very different technologies and are not run on the same biological sample (each test takes a new sample from the patient).

What is the probability for a patient having the disease if both the new and the existing test give a positive result?

Hint: define a virtual test which is positive if both the new and existing tests return a positive results and gives a negative result otherwise. Compute the sensitivity and specificity for this compound test, then proceed as before.

Exercise 2

Suppose $\Pi \sim \text{Beta}(a, b)$. Compute $E[\Pi]$.

We have seen that if

$$\begin{cases} \Pi \sim \text{Beta}(a, b) \\ Y|\pi \sim \text{Bin}(n, \pi) \end{cases}$$

then $\Pi|Y = k \sim \text{Beta}(a + k, b + n - k)$.

Show that

$$E[\Pi|Y = k] = \frac{a+b}{a+b+n} E[\Pi] + \frac{n}{a+b+n} \bar{y}$$

Where $\bar{y} = k/n = \sum_i y_i/n$, the average of n Bernoulli(π) trials.

What does this tell you if $n \to \infty$?

Exercise 3

Suppose $Y_1, \dots, Y_n \sim_{\text{iid}} \text{Pois}(\lambda)$ and that you observe data $\{y_i\}_{i=1}^n$.

1. Derive the Jeffreys prior for this sampling model.
2. Is this a proper prior?
3. What posterior distribution do you get? Is this a valid distribution?

Exercise 4

Derive the posterior predictive distribution $p(\tilde{y}|y_1, \dots, y_n)$ for

$$\begin{cases} \Lambda \sim \Gamma(a, b) & \text{(prior)} \\ Y_i|\lambda \sim_{\text{iid}} \text{Pois}(\lambda) \quad i = 1, \dots, n & \text{(sampling model)} \end{cases}$$

Exercise 5

In the 1990s, the General Social Survey gathered data on the eductional attainment and number of cildren of 155 women who were 40 years old at the time of their participation in the survey.

We will compare women with college degrees to those without.

Let $Y_{1,1}, \dots, Y_{n_1,1}$ be the number of children for the n_1 women without a college degree and $Y_{1,2}, \dots, Y_{n_2,2}$ for those with degrees.

Assume the following sampling models and priors:

- $Y_{1,1}, \dots, Y_{n_1,1} \sim_{\text{iid}} \text{Poisson}(\theta_1)$
- $Y_{1,2}, \dots, Y_{n_2,2} \sim_{\text{iid}} \text{Poisson}(\theta_2)$
- $\theta_1, \theta_2 \sim_{\text{iid}} \Gamma(2, 1)$

The data from the survey can be summarised by:

- $n_1 = 111, \sum_{i=1}^{n_1} y_{i,1} = 217, \bar{y}_1 = 1.95$
- $n_2 = 44, \sum_{i=1}^{n_2} y_{i,2} = 66, \bar{y}_1 = 1.50$

Derive the posterior distributions for $\theta_1 | \sum_i y_{i,1}, \theta_2 | \sum_i y_{i,2}$.

Compute $P(\theta_1 > \theta_2 | \sum y_{i,1} = 217, \sum y_{i,2} = 66)$.

Derive and compare the posterior predictive distributions for each group.

Exercise 6

Let $\phi = g(\theta)$, where $g(.)$ is a monotone function, and let $h(.)$ be the inverse of $g(.)$ so that $\theta = h(\phi)$. If $p_\theta(\theta)$ is the probability density of θ, then the probability density of ϕ induced by p_θ is given by $p_\phi(\phi) = p_\theta(h(\phi)) \cdot \left|\frac{dh}{d\phi}\right|$.

Using this result, show that Beta(a, b) is conjugate for Bin(n, π) using exponential family notation.

Make this prior as weakly informative as possible for π.

Do you recognise this prior?

Exercise 7

Find the conjugate prior for

i. a Geometric(π) sampling model

ii. an Exponential(λ) sampling model

Exercise 8

Show that the conjugate prior for the mean parameter μ in a Gaussian distribution is itself a Gaussian distribution (assume that the variance σ^2 of the sampling model is known and fixed).