STA623 - Bayesian Data Analysis - Practical 4

Marc Henrion

2024-10-31

Practical 4

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
- $\hat{T}$, $\hat{t}$ - 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

Let's revisit Exercise 5 from Practical 1&2.

We had 2 groups of women and we compared the number of children born to each women in the 2 groups. For each group we assumed a Poisson sampling model: $Y_{i,j} \sim \text{Pois}(\theta_i), i = 1, \dots, n_j, j = 1, 2$ and we found that the posterior distributions were:

1. Women without college degree: $\theta_1 \sim \Gamma(219, 112)$
2. Women with college degree: $\theta_2 \sim \Gamma(68, 45)$

We had computed $P(\theta_1 > \theta_2 | n_1, n_2, \sum_i y_{i,1}, \sum_i y_{i,2}) = 0.97$.

Use the Monte Carlo method to compute

$$P(\tilde{Y}_1 > \tilde{Y}_2 | n_1, n_2, \sum_i y_{i,1}, \sum_i y_{i,2})$$

For the group of women without college degree, remember that we found that the posterior predictive distribution was a negative binomial:

$$\tilde{Y}_1 | n_1, \sum_i y_{i,1} \sim \text{NegBin}(219, 112/113)$$

Compare this distribution with the empirical distribution of the raw data:

no. children per mother	number of mothers
0	20
1	19
2	38
3	20
4	10
5	2
6	2

Let $\mathbf{y} = (y_{1,1}, \dots, y_{n_1,1})$. Define $t(\mathbf{y})$ as the ratio of 2's in $\mathbf{y}$ to the number of $1's$. In this dataset we observe $t(\mathbf{y}) = 38/19 = 2$. Use the posterior predictive distribution for $\tilde{Y}_1 | n_1, \sum_i y_{i,1}$ and the Monte Carlo method to compute $P(t(\mathbf{Y}) \geq 2)$. What is your conclusion?