STA623 - Bayesian Data analysis - Assignment 2

22 - 26 September 2025

Marc Henrion

Assigment

Please email your typed or scanned solutions before 23:59 on Monday 24 November 2025 to BOTH `mhenrion@mlw.mw` and `biostat-unima@unima.ac.mw`.

Please include `STA623 - Assignment 2` in the subject line. Please include your code, model output and graphs. Please comment any submitted code.

Notation

Please try to use the following notation where possible.

- 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 occurring
- $f_X(.)$, $f_Y(.), f_Z(.)$ - probability mass / density functions of X, Y, Z
- $p(.)$ - used as a shorthand notation for pmfs / pdfs if the use of this is unambiguous
- $X \sim F$ - X distributed according to distribution function F
- $E[X]$, $E[Y]$, $E[Z]$, $E[T]$ - the expectation of X, Y, Z, T respectively

Table 1: Please use the random seed associated with your name / ID. Solutions using other data than those generated using your seed will not be accepted.

Student	ID	Seed
Eric Mangani	MSC/BIO/STAT/08/23	1899
Satiel Ngwira	MSC/BIO/STAT/17/23	1845
Ausbin Kutumani	MSC/BIO/STAT/J/01/25	1846
Chikondi Moyo	MSC/BIO/STAT/J/03/25	1608
Kenneth Kachiphaphi	MSC/BIO/STAT/J/04/25	1316
Steven Kaunda	MSC/BIO/STAT/J/06/25	1408
Felix Msamira	MSC/BIO/STAT/J/07/24	1005
Eliams Moyo	MSC/BIO/STAT/S/02/24	2616
Loveness Soko	MSC/BIO/STAT/S/04/24	2587
Filudi Nakutuwa	MSC/BIO/STAT/S/07/24	2472
Ephat Chitsulo	MSC/BIO/STAT/S/08/24	2100
Alex Kachitsa	MSC/BIO/STAT/S/09/24	1970
Steven Chiyembe	MSC/BIO/STAT/S/10/2024	2387
Charity Hamuza	MSC/BIO/STAT/S/12/24	2268
Cassim Nanyumba	MSC/BIO/STAT/S/13/24	1935
Hastings Malunga	MSC/BIO/STAT/S/14/24	1296
Osward Kaposa	MSC/BIO/STAT/S/15/24	1472
Seti Evance	MSC/BIO/STAT/S/16/24	1344
Edward Kamphongwe	MSC/BIO/STAT/S/17/24	2184
Chikondi Banda	MSC/BIO/STAT/S/19/24	2688
Steven Nanga	MSC/BIO/STAT/S/23/24	1920
Chikumbutso Banda	NA	1560

Exercise

For the exercise below, you will need to specify a seed value. You will be given individual seed numbers according to the table on the previous page. **You have to use your own individual seed value** – your data (and hence your results) will be unique to you and different from those of your colleagues.

Use the code below (downloadable as file `hospitalDeaths_generateData_2025.R` from GitHub) to simulate data on deaths in the A&E department for several hospitals.

```
set.seed(0000) # REPLACE 0000 with your individual seed value!
# Solutions using the seed value 0000 will not be accepted.

# Generate data
n<-rpois(n=1,lambda=2000)
hospRf<-rnorm(n=8,mean=0,sd=0.5)
hospRf<-hospRf-mean(hospRf)

ilogit<-function(x){
  res<-exp(x)/(1-exp(x))
  return(res)
}

dat<-data.frame(
  PID=paste(sep="","P",25000+1:n),
  sex=sample(c("M","F"),size=n,replace=TRUE,prob=c(0.5,0.5)),
  triage=factor(
    levels=c("Emergency","Priority","Queue"),
    sample(x=c("Emergency","Priority","Queue"),
           size=n,replace=TRUE,prob=c(0.1,0.25,0.65))
  ),
  hospital=factor(
    levels=paste(sep="","H",1:8),
    sample(x=paste(sep="","H",1:8),
           size=n,replace=T,prob=c(0.25,0.15,0.15,rep(0.09,5)))
  )
) %>%
  dplyr::mutate(
    hospRanEf=hospRf[as.integer(hospital)],
    died=rbinom(n=n,size=1,
                prob=ilogit(-4.5
                            +case_when(
                              triage=="Emergency"~rnorm(n=1,mean=1,sd=0.1),
```

```
                              triage=="Priority"~rnorm(n=1,mean=0.2,sd=0.02),
                              triage=="Queue"~0)
                            +hospRanEf)
  )
) %>%
dplyr::select(!hospRanEf)
```

The dataset you just simulated contains the following columns:

- `pid` - this is just an anonymised patient identification number
- `sex` - this records the biological sex of each patient
- `triage` - records the category that the patients were triaged into by an admission nurse (`emergency`, `priority` or `queue`); the idea is that emergencies get seen without delay, priority cases get seen more quickly than normal cases and then the third category is for all other cases
- `hospital` - this records an identification code for the hospital where each patient was seen
- `died` - this records whether the patient died (value 1) or survived (value 0)

Use `NIMBLE` (or another Bayesian software of your choice) to fit the following logistic regression model, choosing priors of your own choosing for each parameter, writing $Y_{i,j}$ for the mortality outcome for patient $i = 1, \dots, n$ seen in hospital $j = 1, \dots, k$:

$$Y_{i,j} \sim \text{Bernoulli}(\pi)$$

with π modelled as a function of patient sex, triage and hospital:

$$log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \cdot male\_sex_i + \beta_2 \cdot triage\_emergency_i + \beta_3 \cdot triage\_priority + \mu_j$$

where $\mu_j \sim \mathcal{N}(0, \sigma^2), j = 1, \dots, k$.

1. List the number of observations and the total number of deaths for your particular dataset. [5 marks]
2. Explain the choice of prior distributions for all model parameters $(\beta_0, \beta_1, \beta_2, \beta_3, \sigma^2)$. [15 marks]
3. Write `NIMBLE` model code to fit the model. [35 marks]
4. Fit the model, then show and summarise (as a point estimate + confidence interval) the posterior distributions for the various parameters. Explain your choice of Bayesian estimators you report. [15 marks]

5. Show trace plots and histograms for all model parameters and compute the effective sample size and Gelman-Rubin potential scale reduction factors. Discuss the results you are getting. [20 marks]

6. Discuss other model checks you could do. [10 marks]