Challenges of Teaching Bayesian Statistics to Undergraduates

Brian Reich

NC State

JSM 2020

Collaborators: Vianey Leos Barajas (U of Toronto)
Krishna Pacifici (NC State)

A motivating example for undergraduates

- A major challenge for teaching undergraduates is finding motivating applications
- In this talk, I will describe a shiny app that displays a Bayesian analysis of disease outbreak data
- This cursory example does not include any distributions or computing
- The lecture is meant to pique interest and give a high-level overview of Bayesian methods

This is a good motivating example

- Disease outbreak modeling is obviously topical
- Fiddling with curves is fun!
- There are only two parameters and they are interpretable
- Data is sparse early in the outbreak, so priors are crucial
- Leads to a good discussion of the role of prior and ideas for setting priors

Background

- The lecture would begin with a description of the data and model

- The SIR model is a differential equation, but it could be presented as a discrete time model (or without equations)

- Markdown document with background material

Description of the SIR model for a disease outbreak

Description of the data

We will use the *Influenza in a boarding school in England, 1978* in the packages *outbreaks* for illustration. The data are described in the packages as:

> These data comprise of a time series of influenza cases in a boarding school in England. The original data were available only in a figure with some additional data in the main text; hence, the exact numbers vary depending on the source. These data are from Chapter 9 of De Vries et al. (1996). The index case was infected by 1978-01-10, and had febrile illness from 1978-01-15 to 1978-01-18. 512 boys out of 763 became ill.

Example model fits are given below, and interactive model fitting can be performed at

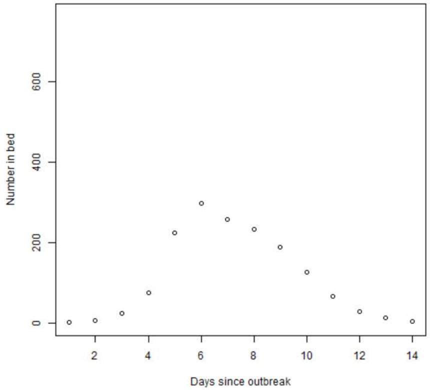
https://shiny.stat.ncsu.edu/bjreich/SIR/

```
library(outbreaks)
influenza_england_1978_school
```

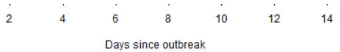
```
##          date in_bed convalescent
## 1  1978-01-22      3            0
## 2  1978-01-23      8            0
## 3  1978-01-24     26            0
## 4  1978-01-25     76            0
## 5  1978-01-26    225            9
## 6  1978-01-27    298           17
## 7  1978-01-28    258          105
## 8  1978-01-29    233          162
## 9  1978-01-30    189          176
## 10 1978-01-31    128          166
## 11 1978-02-01     68          150
## 12 1978-02-02     29           85
## 13 1978-02-03     14           47
## 14 1978-02-04      4           20
```

```
Y  <- influenza_england_1978_school[,2]
nt <- length(Y)      # Number of time sets
N  <- 763            # Population size
I0 <- 1              # Initial number of infected

plot(Y,ylim=c(0,N),xlab="Days since outbreak",ylab="Number in bed")
```



Description of the SIR model



Description of the SIR model

The Susceptible-Infected-Recovered (SIR) model is one of the most basic models of disease spread. At time t, let S_t, I_t and R_t be the number of people in each condition, with $S_t + I_t + R_t = N$ for all t. The states evolve according to the differential equations

$$\frac{dS_t}{dt} = -\beta S_t I_t / N \qquad \frac{dI_t}{dt} = \beta S_t I_t / N - \gamma I_t.$$

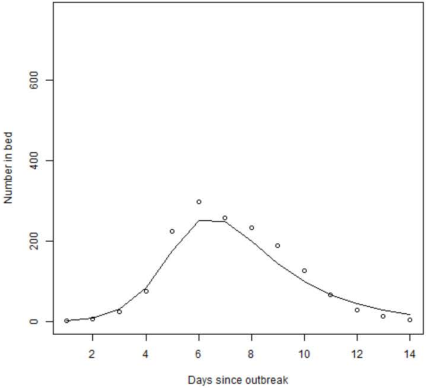
The parameter β controls the rate of new infections and the parameter γ controls the recovery rate. We will use a discrete approximation to these curves with hourly time steps, so $dt = 1/24$. Here are the curves for a few values of the parameters β and γ.

```
SIR <- function(I0,N,beta,gamma,nt,m){
   # Initial states
   i  <- I0
   s  <- N-I0
   dt <- 1/m
   I  <- rep(0,nt)

   for(t in 1:nt){
     for(j in 1:m){
        s <- s - dt*beta*i*s/N
        i <- i + dt*beta*i*s/N - dt*gamma*i
     }
     I[t] <- i
   }
return(I)}

# Reasonable fit
beta  <- 1.7
gamma <- 0.5
I     <- SIR(I0=I0,N=N,beta=beta,gamma=gamma,nt=nt,m=24)
plot(Y,ylim=c(0,N),xlab="Days since outbreak",ylab="Number in bed")
```

```r
# Reasonable fit
beta  <- 1.7
gamma <- 0.5
I     <- SIR(I0=I0,N=N,beta=beta,gamma=gamma,nt=nt,m=24)
plot(Y,ylim=c(0,N),xlab="Days since outbreak",ylab="Number in bed")
lines(I)
```



Larger infection rate

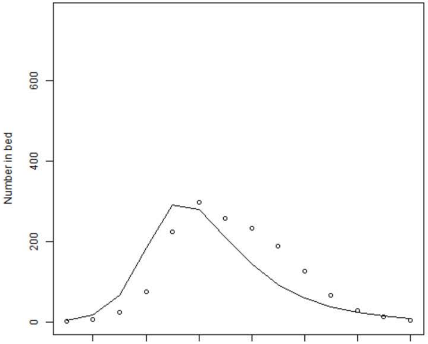
2 4 6 8 10 12 14

Days since outbreak

Larger infection rate

Increasing β gives a steeper curve because the disease spreads faster.

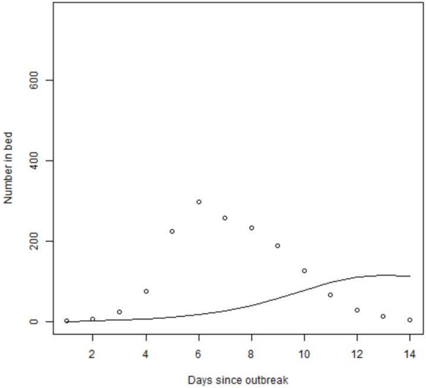
```
beta  <- 2.0
gamma <- 0.5
I     <- SIR(I0=I0,N=N,beta=beta,gamma=gamma,nt=nt,m=24)
plot(Y,ylim=c(0,N),xlab="Days since outbreak",ylab="Number in bed")
lines(I)
```



Smaller infection rate

Decreasing β flattens the curve.

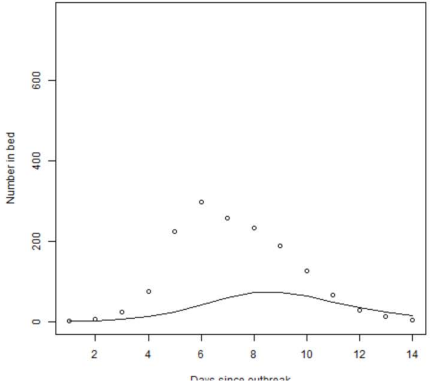
```
beta  <- 1.0
gamma <- 0.5
I     <- SIR(I0=I0,N=N,beta=beta,gamma=gamma,nt=nt,m=24)
plot(Y,ylim=c(0,N),xlab="Days since outbreak",ylab="Number in bed")
lines(I)
```

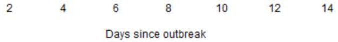


Larger recovery rate

Increasing γ gives fewer overall infections because people are infectious for a shorter amount of time.

```
beta  <- 1.7
gamma <- 1.0
I     <- SIR(I0=I0,N=N,beta=beta,gamma=gamma,nt=nt,m=24)
plot(Y,ylim=c(0,N),xlab="Days since outbreak",ylab="Number in bed")
lines(I)
```

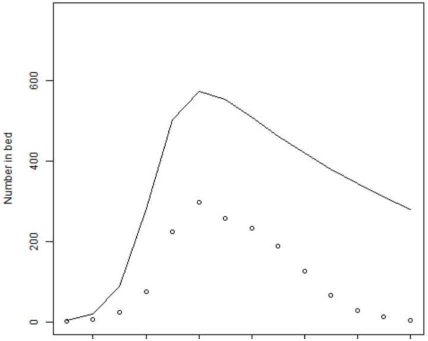




Smaller recovery rate

Decreasing γ gives more overall infections because people are infectious for a longer amount of time.

```
beta  <- 1.7
gamma <- 0.1
I     <- SIR(I0=I0,N=N,beta=beta,gamma=gamma,nt=nt,m=24)
plot(Y,ylim=c(0,N),xlab="Days since outbreak",ylab="Number in bed")
lines(I)
```



Definition of parameters

- The model *parameters* are β and γ
- Definition of parameters: fixed but unknown constants in the probability model
- If we knew the parameters we could:
  - Compare this outbreak to others
  - Forecast spread
  - Predict effects of interventions

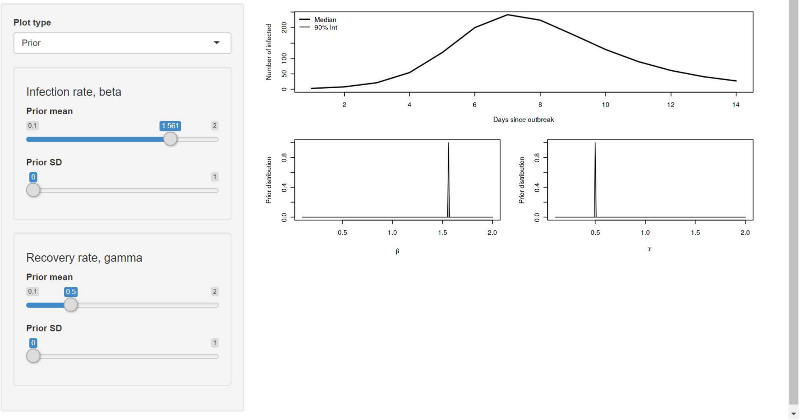
Discussion questions

- **Question 1**: At the onset of the outbreak, how would you estimate the parameters?
- **Question 2**: What are the consequences of using the wrong values?

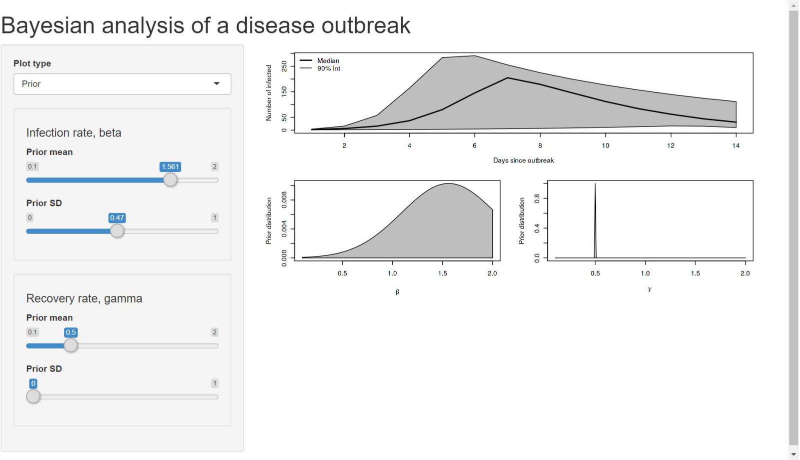
Definition of the prior distribution

- We rarely know the parameters at the onset of a study; this is why we collect data
- Rather than select a particular value, we place an uncertainty distribution on the parameters
- This is called the *prior distribution*
- You can explore priors and their effect on the curve here

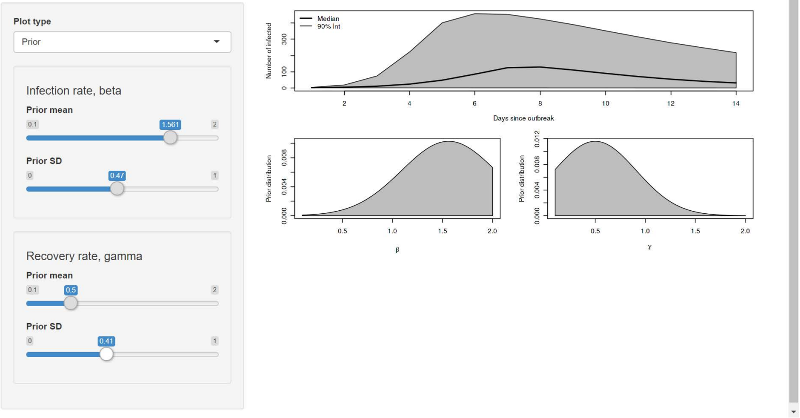
Bayesian analysis of a disease outbreak



Bayesian analysis of a disease outbreak



Bayesian analysis of a disease outbreak



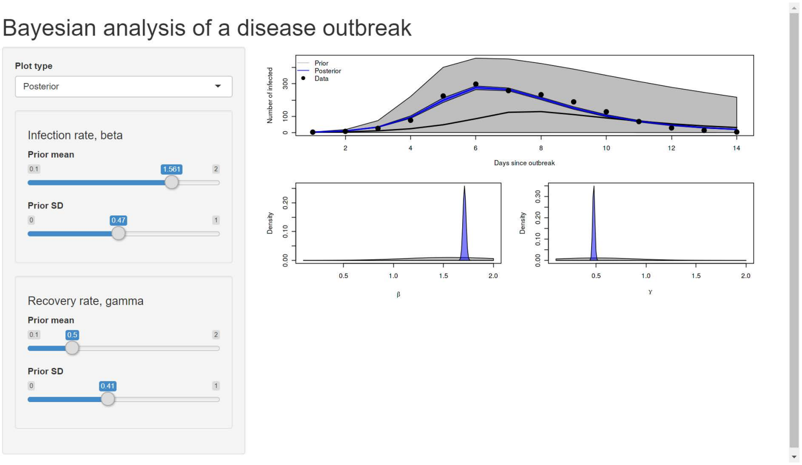
Discussion questions

- **Question 1**: At the onset of the outbreak, where can you get information about the parameters?

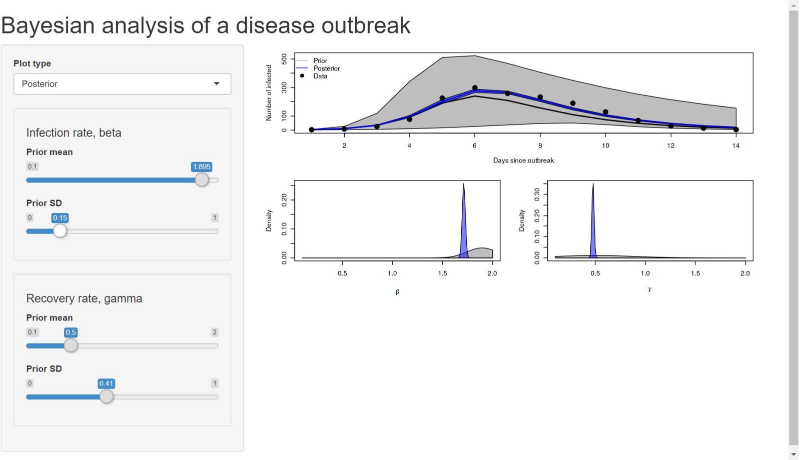
- **Question 2**: What would you do if there is no prior information?

Definition of the posterior distribution

- After the outbreak we have two sources of information: the prior and the data
- *Bayesian learning* combines these two sources of information
- The uncertainty distribution of the parameters after observing the data is the *posterior distribution*
- You can explore Bayesian learning at here


Bayesian analysis of a disease outbreak
Plot type
Posterior
Infection rate, beta
Prior mean
0.1
1.561
2
Prior SD
0
0.47
1
Recovery rate, gamma
Prior mean
0.1
0.5
2
Prior SD
0
0.41
1
Prior
Posterior
Data
Number of infected
Days since outbreak
Density
β
γ

Bayesian analysis of a disease outbreak



Discussion questions

- **Question 1**: Does the SIR model seem to fit the data well? What to do if it doesn't?
- **Question 2**: How much affect does changing the prior have on the posterior?

More information about the shiny app

- The material presented here can be found on the website for our book[1]
- The website is https://bayessm.org/ ("Link to Supplemental Materials" → "Bayesian modeling of a disease outbreak")
- We also have many datasets, worked examples, GUIs, beamer slides and solutions to homework problems
- Thanks!

[1] Reich and Ghosh (2019). Bayesian Statistical Methods. CRC Press.